JADH 2018

"Leveraging Open Data"

September 9-11, 2018
Hitotsubashi Hall, Tokyo

https://conf2018.jadh.org

Proceedings of the 8th Conference of Japanese Association for Digital Humanities

Organized by:
Japanese Association for Digital Humanities

Hosted by:
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems

Co-organized by:
JSPS Grant-in-Aid Project (S) "Construction of a New Knowledge Base for Buddhist Studies" (15H05725)
International Institute for Digital Humanities

Sponsored by:

<oXygen/> XML Editor

AMANE LCC.

infocom

MetaInfo

NTT DATA Trusted Global Innovator

Supported by:
Japan Art Documentation Society
IPSJ SIG Computers and the Humanities
Japan Society for Digital Archives
Japan Association for English Corpus Studies
Japan Association for East Asian Text Processing
The Mathematical Linguistic Society of Japan
Japan Society for Information and Media Studies
Japan Society of Information and Knowledge

Proceedings of JADH Conference, vol. 2018

Proceedings of JADH Conference, vol. 2018

Edited by Chikahiko Suzuki

Published by:
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo
http://codh.rois.ac.jp

Online edition: ISSN 2432-3144 Print edition: ISSN 2432-3187

Table of Contents

JADH 2018 Committees

Program Committee:

- Toru Tomabechi, Chair (International Institute for Digital Humanities, Japan)
- Paul Arthur (Edith Cowan University, Australia)
- James Cummings (Newcastle University, UK)
- J. Stephen Downie (University of Illinois, USA)
- Øyvind Eide (University of Koeln, Germany)
- Makoto Goto (National Institute for Humanities, Japan)
- Shoichiro Hara (Kyoto University, Japan)
- JenJou Hung (Dharma Drum Institute of Liberal Arts, Taiwan)
- Jieh Hsiang (National Taiwan University, Taiwan)
- Akihiro Kawase (Doshisha University, Japan)
- Asanobu Kitamoto (National Institute of Informatics, Japan)
- Chao-Lin Liu (National Chengchi University, Taiwan)
- Maciej Eder (Pedagogical University of Kraków, Poland)
- Charles Muller (The University of Tokyo, Japan)
- Hajime Murai (Future University Hakodate, Japan)
- Kiyonori Nagasaki (International Institute for Digital Humanities, Japan)
- Geoffrey Rockwell (University of Alberta, Canada)
- Susan Schreibman (National University of Ireland Maynooth, Ireland)
- Masahiro Shimoda (The University of Tokyo, Japan)
- Raymond Siemens (University of Victoria, Canada)
- Donald Sturgeon (Harvard University, USA)
- Keiko Suzuki (Ritsumeikan University, Japan)
- Tomoji Tabata (Osaka University, Japan)
- Kathryn Tomasek (Wheaton College, USA)
- Christian Wittern (Kyoto University, Japan)
- Taizo Yamada (The University of Tokyo, Japan)

Local Organizers:

- Asanobu Kitamoto, Chair (Center for Open Data in the Humanities / National Institute of Informatics)
- Tarin Clanuwat (Center for Open Data in the Humanities / National Institute of Informatics)
- Mika Ichino (Center for Open Data in the Humanities / National Institute of Informatics)
- Kiyonori Nagasaki (International Institute for Digital Humanities)
- Masahiro Shimoda (The University of Tokyo)
- Chikahiko Suzuki (Center for Open Data in the Humanities / National Institute of Informatics)
- Akihiko Takano (National Institute of Informatics)
- Toru Tomabechi (International Institute for Digital Humanities)

Time Table

September 9 (Sun), Day 1

14:30-18:30	Workshop

September 10 (Mon), Day 2

9:00-9:30	JADH and TEI Joint Opening Session
9:30-11:00	Session 1: Open Scholarship (Room A1)
9:30-11:00	Session 2: Data Analysis (Room A2)
11:00-13:00	Lunch Break
13:00-16:15	JADH and TEI Joint Keynote Session
16:15-16:45	Coffee Break
16:45-19:00	JADH and TEI Joint Poster Session with Poster Slam
19:00-21:00	Conference Banquet

September 11 (Tue), Day 3

9:30-10:30	Session 3: Text Analysis (Room A1)
9:30-10:30	Session 4: Archiving (Room A2)
10:30-10:45	Coffee Break
10:45-12:15	Session 5: Data Analysis (Room A1)
12:15-13:45	JADH Annual General Meeting
13:45-15:15	Panel Session 1 (Room A1)
13:45-15:15	Session 6: Technical Development (Room A2)
15:15-15:30	Coffee Break
15:30-17:00	Session 7: Exploring History (Room A1)
15:30-17:00	Session 8: Collection and Curation (Room A2)
17:00-17:15	Coffee Break
17:15-18:45	Panel Session 2 (Room A1)
18:45-19:00	Closing (Room A1)

September 12 (Wed), Excursion Day

13:00-	Joint Excursions with TEI

[Workshop]

Machine Reading: Advanced Topics in Word Vectors

Eun Seo Jo[1], Javier de la Rosa[1]

Description

This half day workshop is an introduction to word vectors and text vectorization broadly. We will focus on building intuition of how word vectors work, incorporating visualization methods, using pre-trained vectors, and exploring applications of word embeddings. We will teach you both the high-level concepts and the practical usages of these widely used analytical tools for text analysis in digital humanities (DH). It is a hands-on workshop with practical activities for the participants starting with a review of word vectors byway of visualization, an overview of downloadable word vectors, and examining the potential pitfalls of using word vectors in humanistic analysis and the methods for mitigating these issues. Given the general applicability of machine learning models in real life, addressing issues concerning biased models, datasets, and algorithms, is of vital importance for correct interpretation of their applications.

We will provide a Python Jupyter Notebook and an accompanying text corpus that we will work through as a group. By the end of the workshop, the participants will have working knowledge of how and where to download or train word embeddings and the caveats of using them.

Relevance to the DH Community

Since the apparition of analytical approaches to distant reading and macro-analysis, popularized by Moretti and Jockers, and the possibility of access to huge amounts of textual data and long-term studies such as Culturomics, new tools were needed to tackle the increasing complexity of large corpora. Borrowing from advances in machine learning and computational linguistics, digital humanists have experimented with various methods of text quantification for interpreting macro contours of culture and language. In particular, word vectors have gained recognition for their versatility in DH studies. Scholars have used word vectors in a variety of tasks such as measuring similarity in word meaning (Caliskan et al., 2017), authorship attribution (Kocher et al., 2017), or dialogism in novels (Muzny et al., 2017).

This workshop is both a theoretical and practical introduction to humanist applications of these methods. Those interested in large scale text-analysis of any corpora will learn the basics of transforming textual data into numerical form.

Instructors

Eun Seo Jo researches the language of American foreign relations in historical contexts and applications of NLP and ML in history. She is a PhD candidate in history at Stanford University where she is also a member of the Literary Lab and a Digital Humanities Fellow. She has presented at various DH conferences and is a DH methodology consultant at Stanford.

Javier de la Rosa is a Research Engineer at the Center for Interdisciplinary Digital Research, a unit at the Stanford University Libraries focused on digital scholarship. He is an active member of the DH scholarly community at Stanford and regularly participates in conferences, professional organizations, and teaches workshops and tutorials to faculty and graduate students. He holds a Post-doctorate research fellowship and a PhD in Hispanic Studies at Western University, Ontario, where he also served as Tech Lead for the CulturePlex Lab. He completed both his MSc. in Artificial Intelligence and BSc. in Computer Engineering at University of Seville, Spain. His work and interests span from

[1] Stanford University

cultural network analysis and computer vision, to text mining and authorship attribution in the Spanish Golden Age of literature.

Target Audience and Prereqs
Post-docs, faculty, and advanced graduate students with Python prerequisites. Although the main concepts will be overviewed, knowledge of basic word embeddings and word2vec specifically would be desirable. In order to participate fully in all activities, participants must have working knowledge of basic programming concepts, the Python language, data structures, and the Numpy library.

- Technical Support: Microphones and Projector
- Proposed Length: Half-day (4 hours; 4 sessions)
- Medium: Notebook (Jupyter)
- Libraries: Numpy, Pandas, Textacy, SpaCy, Gensim, scikit-learn, matplotlib

Workshop Outline
The workshop is split into four 50 min sessions with 10 minutes breaks in-between. We teach several methods in each unit with increasing difficulty. The schedule is broken down below:

1) Understanding Word Vectors with Visualization
This unit will give a brief introduction of word vectors and word embeddings. Concepts needed to understand the internal mechanics of how they work will also be explained, with the help of plots and visualizations that are commonly used when working with them.

- 0:00 - 0:20 From word counts to ML-derived Word Vectors (SVD, PMI, etc.)
- 0:20 - 0:35 Clustering, Vector Math, Vector Space Theory (Euclidean Distance, etc.)
- 0:35 - 0:50 [Activity 1] Visualizations (Clustering, PCA, t-SNE) [We provide vectors]

2) Word Vectors via Word2Vec
This unit will focus on Word2Vec as an example of neural net-based approaches of vector encodings, starting with a conceptual overview of the algorithm itself and end with an activity to train participants' own vectors.

- 0:00 - 0:15 Conceptual explanation of Word2Vec
- 0:15 - 0:30 Word2Vec Visualization and Vectorial Features and Math
- 0:30 - 0:50 **[Activity 2] Word2Vec Construction [using Gensim] and Visualization (from part 1) [We provide corpus]**

3) Extended Vector Algorithms and Pre-trained Models
This unit will explore the various flavors of word embeddings specifically tailored to sentences, word meaning, paragraph, or entire documents. We will give an overview of pre-trained embeddings including where they can be found and how to use them.

- 0:00 - 0:20 Overview of other 2Vecs & other vector engineering: Paragraph2Vec, Sense2Vec, Doc2Vec, etc.
- 0:20 - 0:35 Pre-trained word embeddings (where to find them, which are good, configurations, trained corpus, etc.)
- 0:35 - 0:50 [Activity 3] Choose, download, and use a pre-trained model

4) Role of Bias in Word Embeddings
In this unit, we will explore an application and caveat of using word embeddings -- cultural bias. Presenting methods and results from recent articles, we will show how word embeddings can carry historical bias of the corpora trained on and lead an activity that shows these human-biases on vectors and how they can be mitigated.

- 0:00 - 0:10 Algorithmic bias vs human bias
- 0:10 - 0:40 **[Activity 4] Identifying bias in corpora (occupations, gender, ...) [GloVe] (Caliskan et al., 2017)**
- 0:40 - 0:50 Towards unbiased embeddings; Examine "debiased" embeddings
- 0:50 - 0:60 Conclusion remarks and debate

References

Caliskan, A., Bryson, J.J., Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186. https://doi.org/10.1126/science.aal4230

Kocher, M., Savoy, J., 2017. Distributed language representation for authorship attribution. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx046

Nanni, F., Dietz, L., Ponzetto, S.P., 2017. Toward a computational history of universities: Evaluating text mining methods for interdisciplinarity detection from PhD dissertation abstracts. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx062

Digital Open Scholarships in Heritage: The Archaeology of Portus Massive Open Online Course example

Eleonora Gandolfi[1,2], Graeme Earl[1]

The Archaeology of Portus Massive Open Online Course (MOOC) is a six week long online course hosted by FutureLearn. The course is structured around 4 hours per week of learner effort and has run six times since 2014. It was one of the first FutureLearn courses and was the focus of considerable experimentation, structured around the affordances of the social learning model and the opportunities to integrate a variety of digital interactions within a single course. The course aims to introduce learners to the archaeology of Portus, the Port of Imperial Rome, and provides a sense of the historical context, the work undertaken on site and the methods employed. It also uses Portus as a mechanism for introducing key research ideas in Roman archaeology. This paper discussed this course in relation to a broader "open scholarship spectrum" for Portus' heritage which includes mass broadcast media via citizens science, online tours and open education, through to individual, novel research with open access publications, open data and tools. It will explore the potential role for open education at the hearth of this process. In particular, the paper will analyse and examine the use of course materials in engaging with learners, including students in compulsory age education, and their impact on awareness of Roman Cultural Heritage in Italy.

The information provided as part of the MOOC and the complementary resources have been updated regularly following the Action Research method, which commonly appears in education and other fields of professional practice such as nursing. This methodology creates an *action,* or Intervention, which is carried out while *outcomes* are systematically 'researched' to improve teaching practice and inform policies. The evaluation conducted focuses on the impact of an intervention rather than systematically gathering evidence and data to measure outcomes (McNiff, 2013). This systematic improvement of the material offered, with the translation of most of the content in other languages (mainly Italian, with some French and Japanese) has increased the engagement with communities interested not exclusively in the heritage, but also in the learning of a second language or other transferrable skills.

The initial preparation stage included the alignment of the Portus MOOC's learning objectives to the Qualification and Credit Framework (QCF) levels (Ofqual, 2017), to the European Qualifications Framework (EQF; European Commission, 2017) and to Bloom's digital taxonomy (Anderson and Krathwohl, 2001; Churchs, 2008). The course units have then been mapped to the current English and Italian school curriculum to identify overlaps before applying the Simple Measure of Gobbledygook, also known as SMOG analyses (McLaughin, 1969), and the Gulpease index (Lucisano and Piemontese, 1988) to determine the complexity of English and Italian texts and to identify recurrent patterns that might have influenced the readability scores. The analyses conducted showed how technical terms might influence the readability scores and how the score gives a quantitative evaluation of the text and do not take into account the quality of the information provided. This means that despite classifying the text according to age groups and quality framework, it does not indicate how comprehensible the text amongst them is.

The engagement stage included the creation of Content and Language Integrated Learning (CLIL), blended learning and online tools in collaboration with teachers. The created tools have been used as part of the project to create educational cooperation between schools and Universities, increase access to education content, develop teachers' skills, promote cultural literacy in geographically dispersed student communities, develop future world citizens in both countries and provide important insights to understand how the digital component can used to integrated traditional education to shape our society.

[1] King's College London
[2] University of Southampton

The methodology presented aims to offer familiar notions in a different language to the Italian speakers, while offering non-familiar contents in a known language to the English schools. Small joint activities, such as the creation of 3d models or new content on Wikipedia, have been developed to facilitate the collaboration between the two groups of students creating a virtual international classroom. All results are then shared with the learners' community in an attempt to continue the data analyses in a sustainable and engaging way and to build new cultural identities. Students taking part to the project where asked to complete a survey at the beginning of the project and at the end to profile the students (mainly around knowledge of heritage and English language) and understand the evolution of their behaviors.

This paper will report on the different stages of development of the applied methodology, the difficulties encountered and strengths of the approach. It will also describe the results of the testing and formal evaluations carried out with secondary school students in comparison with online learners.

References

Anderson, L., Krathwohl, D. A. (2001). *Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* New York: Longman.

Churches, A. (2009). *Bloom's Digital Taxonomy*. Edorigami Wikispace, 1 April 2009. http://edorigami.wikispaces.com/file/view/bloom%27s%20Digital%20taxonomy%20v3.01.pdf/65720266/bloom%27s%20Digital%20taxonomy%20v3.01.pdf (last access 25/06/2018).

European Commission (2017). Descriptors defining levels in the European Qualifications Framework (EQF), Learning Opportunities and Qualifications in Europe. https://ec.europa.eu/ploteus/en/content/descriptors-page (last access 25/06/2018).

Lucisano P. and Piemontese M. E. (1988) *GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana, in Scuola e città*, XXXIX, nº 3, p. 110-24.

McLaughlin, G. H. (1969). "SMOG grading: A new readability formula." *Journal of Reading*, 12(8), 639-646.

McNiff, J. (2013). *Action Research: Principles and Practice*. Routledge.

Ofqual (2017). Get the facts: AS and A level reform. https://www.gov.uk/government/publications/get-the-facts-gcse-and-a-level-reform/get-the-facts-as-and-a-level-reform#introduction (last access 25/06/2018).

Capturing Literary Events at Metropolitan Scale: Open Data and 'One Book One Chicago'

John Shanahan[1]

This presentation describes methods and findings of a digital humanities project that combines open data and social media with library circulation figures in order to study literary reading at scale. The "Reading Chicago Reading" (RCR) project, supported by grants from the U.S. National Endowment for the Humanities Office of Digital Humanities, HathiTrust, and the Lyrasis Catalyst Fund, combines humanities, social sciences, and computer science expertise to create new visualizations and predictive models of reading behavior. Principal investigators are John Shanahan (DePaul University Department of English), Robin Burke (DePaul University School of Computing), and Ana Lucic (DePaul University Digital Scholarship Librarian). While it originates in the study of one large city, Reading Chicago Reading supplies lessons for scholars elsewhere who are interested in tools and methods for working with different types of open civic data for DH projects.

"Reading Chicago Reading" is a large-scale multi-disciplinary analysis of the popular and much-imitated "One Book, One Chicago" (OBOC) program of the Chicago Public Library (CPL). Since Fall 2001, the CPL has chosen books around which to organize city-wide public events, book discussions, and other creative programming. The chosen OBOC works aim to reflect the diversity of the city's residents, their cultural heritages, interests, and concerns. Our project began from the observation that the OBOC program might act as a natural experiment – data associated with each chosen work represents a time-stamped probe into library usage and, by extension, a window onto the reading behavior of the library patrons of a major American city. The program's annual repetition provides a means of studying reading behavior comparatively over time and in relation to other forms of civic activity when joined to large sets of open data from the City of Chicago data portal (https://data.cityofchicago.org/)

The project's guiding methods are to combine text characteristics, aggregate patron demographics, and promotional activities as variables that can be modeled to predict patron response to future OBOC titles. We know of no other project that combines these data sources to seek a predictive model of patron behavior. A predictive model of this type, combined with a user-friendly dashboard (a future direction for us), can be used by librarians to visualize and analyze the likely uptake of a prospective title by the reading public. While the "One Book One Chicago" program has provided a perfect launch-point for our project's methodology, our models, tools, and techniques can be applied to library holdings in city systems more generally. In recent years, large-scale data mining has become a common practice, but research shows that libraries rarely make resource decisions based on data-driven considerations. Indeed, sociologists of reading such as Wendy Griswold have found that public libraries often choose texts based on "gut" instincts rather than empirical data. "Reading Chicago Reading" is working to create tools to close this knowledge gap, creating a path for the use of large-scale data to achieve the broad goal of understanding reading behavior of library patrons across a metropolitan region.

To date, we have focused on connecting our historical circulation data with branch-level demographics. The present phase of our modeling work incorporates content-oriented aspects of the books measured against larger corpora of in-copyright work via HathiTrust, a large multi-institution digital library that contains roughly 16 million volumes scanned and digitized, with thousands of new titles arriving each month. In 2017, we received an Advanced Computing Support grant from HathiTrust, offering us the chance to perform additional large-scale text analysis across many in-copyright texts, something that would not otherwise be possible. From a longer list of over 300 works made from CPL's OBOC-associated recommended titles we have created a comparator set for

[1] DePaul University

ongoing text analysis. We use a variety of techniques to quantify the most salient formal features for modeling, including sentiment analysis, topic modeling, and type-token ratio analysis as well as the distribution of different types of verb classes across the texts. (Text measures and other visualizations and their code can be found on our project blog https://dh.depaul.press/reading-chicago/).

Social media data is another important source of information about books and readers through which we aim to understand the impact of "One Book One Chicago" programming on a diverse readership. Microblogs such as Twitter and other social media sources form a rich and at times nuanced reflection of some readers' responses to OBOC texts – punctual feedback data that supplements and contextualizes the raw numerical circulation data we have obtained from the library system. They demonstrate, for example, how discussions *around* "One Book" texts in the wider public sphere may drive interest in related books. While our CPL circulation statistics are anonymous (check in and check out records only, without personal information) our social media data does sometimes contain identifying elements and therefore we anonymize personal attributes in our social media data. At the same time, our project team has extracted a corpus of just over 25,000 reviews of the most recent OBOC titles from Goodreads.com. This data includes tags, reviews, and other user-generated data, and represents readers' voluntary responses to a text including emotional reactions both positive and negative. From this data, we are exploring sentiment analysis and topic modeling to build representations from these data sources that will support comparison across the texts.

Of particular interest of JADH researchers may be our newest direction of research: maps of book circulation data across the library system, and comparison maps for three recent OBOC choices that are set in Chicago and three that are not. For books with Chicago settings, we have prototyped interactive maps that join sentiment scores of sentences bearing Chicago location words with geo-located data from the Chicago city data portal. If literary form and real-world geography have ties to one another, as Franco Moretti and others have speculated, our project may be able to visualize such links.

Early Chinese Periodicals Online (ECPO) – from Digitization towards Open Data

Matthias Arnold[1]

Abstract

This paper presents the project "Early Chinese Periodicals Online (ECPO)". It introduces the database, and discusses two major directions of current development: 1) The installation of a cross-database agents' service to identify names, assign names to persons, and relate persons to authorities (GND, VIAF, Wikidata). 2) The conceptualization of a TEI module to expand the database with full texts functionality, thereby touching issues like semi-automatic page segmentation, use of non-Chinese speaking communities in crowd sourcing, and selecting of relevant TEI markup to encode Republican era publications.

This paper introduces the project "Early Chinese Periodicals Online (ECPO)"[1]. ECPO joins several important digital collections of the early Chinese press and puts them into a single overarching framework. To date, ECPO has focused on a body of rich but heretofore undervalued materials—women's and entertainment magazines. It is open to further additions: currently, we are adding a selection of literary, art and women's magazines, e.g. *Tianyi* 天義 (Tien yee), *Banyue* 半月(The Half Moon Journal), as well as western-language press published in China, e.g. *The Canton Press*.

The first building block for ECPO was several databases on early women's periodicals and entertainment publishing: "Chinese Women's Magazines in the Late Qing and Early Republican Period" (*WoMag*), "Chinese Entertainment Newspapers" (*Xiaobao*), and various databases hosted through the Academia Sinica in Taiwan.

WoMag[2] focuses on four influential women's magazines published between 1904 and 1937. It records all articles, images, advertisements, and related agents and assigned them to a complete set of scanned pages. This database is the model for what we have called the *intensive approach* within our database structure.

Xiaobao[3] provides basic publication data and characteristics of the contents of some 22 entertainment newspapers (*xiaobao*) from the late Qing and Republican periods. This database is the model for what we have called the *extensive approach* in our database structure.

The Academia Sinica, and in particular its Institute of Modern History, have in recent years digitized large parts of their collections of periodicals and built a database for the *Funü zazhi* 婦女雜誌 (The Lady's Journal)[4]. All these resources follow the model for what we call the e*xtensive approach*.

ECPO has begun to join these various materials in a second, ongoing phase of the project, which was first supported by a grant from the Chiang Ching-kuo Foundation (2012-2015) for collaboration between Heidelberg University and the Academia Sinica. Since 2015, the Institute of Chinese Studies and the Heidelberg Centre for Transcultural Studies (HCTS) at Heidelberg University have continued to support the project; technical development is coordinated through the Heidelberg Research Architecture (HRA). It is our aim to make the different collected materials accessible through a single search interface and to continue to acquire new publications.

As it currently stands, ECPO provides the research community with open access to more than 230 publications from the Early Republican period comprising over 250.000 pages of print. A key and unique aspect of the project is to make entire issues available, front-to-back, including illustrations, advertisements, and even blank pages. For approximately half of the publications, we also provide descriptions of individual items (articles, images, advertisements) and bibliographic metadata in Chinese with Pinyin transcription. These records also contain genre and column information, basic content

[1] University of Heidelberg

analysis (keywords), as well as the names and roles of agents associated with an item, including “mentioned in article” or “depicted in an image”. Overall, the project followed the five guidelines for the digital archiving of periodicals (Latham and Scholes 2006, p. 524).
To further increase the impact of ECPO and in order to sustain the information, ECPO has begun to enable the system to provide data for re-use as open data. We implemented a MODS XML API[5] to provide bibliographic information for all annotated items in the database, and installed a IIIF image service for all page scans.

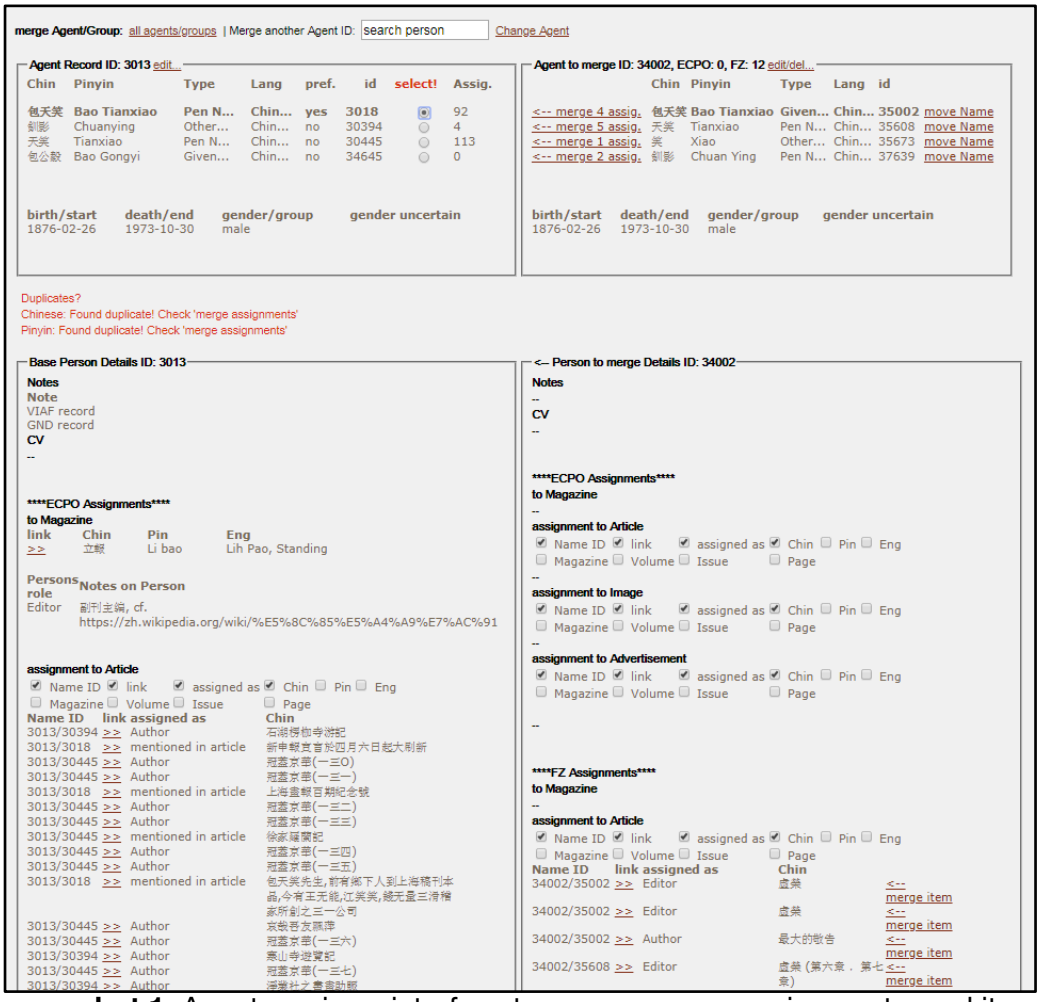
We are now working on the approximately 47.000 names recorded within the *WoMag* and ECPO databases. The aim is to produce a concise list of personnel occurring across the databases. We set up a cross-database agent service that distinguishes between all kind of names that occur within the data from actual persons, groups, or corporations. While some agents may have multiple names, some names may refer to different agents. The agents service allows us to: a) merge identical names across databases, b) identify agents and assigning names to them, and c) link agent records to authority data (GND, VIAF, Wikidata). Besides creating a curated list of agents occurring in the publications, we also aim to add missing persons to Authority files, using the German National Authority file (GND). (Screenshot1)

One aspect ECPO is now starting to focus on is full text capability. While some records occasionally feature full text passages in the metadata, for example some advertisements, this task is too big to solve manually – we need automated workflows. However, one cannot use OCR software out-of-the-box, for a number of reasons: document analysis fails to recognize complex newspaper layout, character recognition fails when it faces emphasis marks next to characters, and recognized passages have to be grouped in the right semantic order.

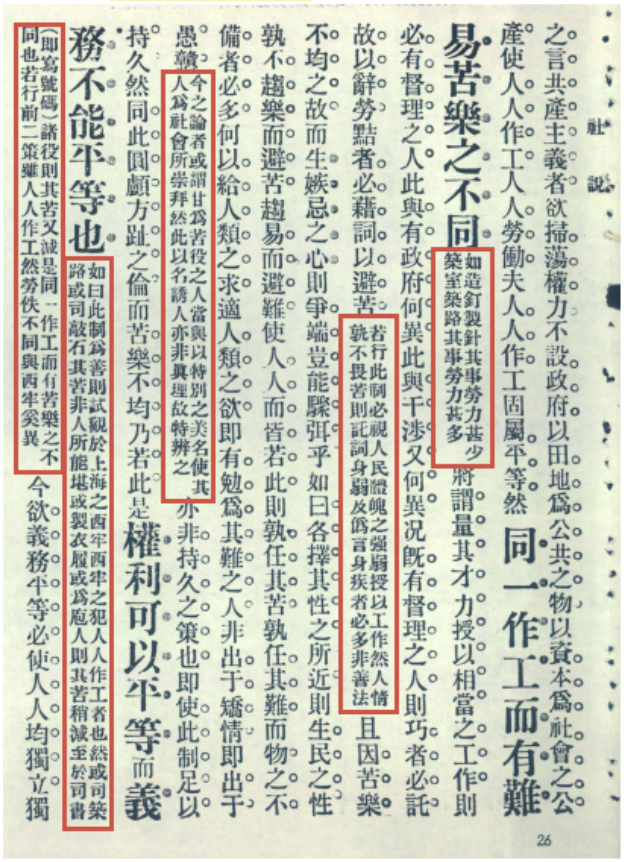
The paper will discuss a number of approaches to further exploring and analyzing the contents of collected publications, together with efforts to open the collection’s data for re-use. I will demonstrate workflows in the Agents service, which assists in curating agent records across databases and forms the basis for enhancing authority records. I will also present results from a crowd-sourced approach to newspaper segmentation to generate segments that can easier be OCRed. In addition, I will introduce our efforts to create a module for encoding text in TEI and relate it to the database.

Moreover, a proper mark-up still needs to be developed to encode full text in TEI XML for these early publications with their complex layout and design. Using *Tian yi* as example, I will discuss some of the problems encountered and present ideas about how to solve them[6]. (Screenshot2)

ECPO started as typical information-silo: a sophisticated data structure, but no “outside” connections. In recent years, we made a greater effort to open up database and data. We now provide our material open access and implemented technical interfaces for data re-use. In the future, we hope to expand this strategy also to agents and full text.
[975+23]



Screenshot 1: Agent service – interface to merge names, assignments, and items



Screenshot 2: Detail of Text analysis [6] for definition of TEI mark-up

References

Hockx, Michel, Joan Judge, and Barbara Mittler, eds. *Women and the Periodical Press in China's Long Twentieth Century: A Space of Their Own?* Cambridge: Cambridge University Press, 2018.

Sean Latham and Robert Scholes, “The Rise of Periodical Studies,” *PMLA: Publications of the Modern Language Association*, 121 (2006), 517-531.

Sung, Doris, Liying Sun and Matthias Arnold. “The Birth of a Database of Historical Periodicals: Chinese Women's Magazines in the Late Qing and Early Republican Period.” In *Tulsa Studies in Women's Literature* 33, no. 2 (2014): pp. 227-37. http://muse.jhu.edu/article/564237.

[1] http://ecpo.uni-hd.de
[2] http://womag.uni-hd.de
[3] http://xiaobao.uni-hd.de
[4] http://mhdb.mh.sinica.edu.tw/fnzz/
[5] http://kjc-sv034.kjc.uni-heidelberg.de/ecpo/api/mods
[6] Full document https://docs.google.com/document/d/1xsE4kavEe-LdL7JKwDpaXtGZQbXLsRLtzSJ8sM4MTbw/edit?ts=5ac5e21e

From Collection Curation to Knowledge Creation: Building a Bilingual Dictionary of Ming Government Official Titles through Expert Crowd-translation

Ying Zhang[1], Susan Xue[2], Zhaohui Xue[3]

Digital technologies have empowered librarians to fully engage in the entire cycle of scholarly communication, from research, data collection and analysis, authoring, peer review, publication to discovery and dissemination (ACRL, 2003). They provide an opportunities for librarians to step out from our unostentatious offices and collections and to explore a new area of facilitation and participation in the process of knowledge creation.

This presentation reports an international-level collaborative project among academic librarians, scholars, expert consultants, and information technologists. Funded by the Andrew Mellow Foundation under its Mellon-CEAL Innovation Program, several Chinese studies librarians in the U.S. initiated to develop a comprehensive bilingual dictionary of Ming government official titles. To develop the dictionary, we designed and created, with the support of information technologists, an online, expert-based crowd-translation system http://mingofficialtitles.lib.uci.edu/#/. The online system served as a virtual community where scholars around the world worked collaboratively to contribute English translations of official titles that had not been translated in existing publications, primarily Hucker's *A Dictionary of Official Titles in Imperial China* (1985). Four Ming scholars served as project consultants to provide expert advices and quality control.

The project has produced two open access (OA) products. The first is the online, crowd-translation system, of which the source codes have been made available at https://github.com/UCI-Libraries/Ming-Titles-Dictionary at no cost. Different from many other crowd-translation applications, our system has a built-in, triple quality-control method, allowing for credential authentication, anonymous peer-review, and expert judgment. It can be repurposed for compiling bilingual dictionaries of any subject domain, serving as "a digital platform for research also crowd work." (Blanke at al., 2016). On the platform, Ming scholars around the world contributed their knowledge collaboratively and collectively. Their contributions included, but were not limited to, submitting English translations of Ming government official titles, commenting on Chinese and pinyin titles, suggesting missing titles, and blind-reviewing others' contributions. By the official closing date of the project, we had registered and approved 45 scholars in the system. Among them, 17 made at least one contribution each.

Whereas the crowd-translation system serves as the digital research tool, the bilingual dictionary, the second OP product, fills a resource gap and "corrects the weaknesses" of library collections (Kennedy, 1983) for expressed needs from Ming scholars and a well-established, digital humanities project – China Biography Database (CBDB) http://projects.iq.harvard.edu/cbdb/collaborators. As the ultimate product, the first edition of *A Bilingual Dictionary of Ming Government Official Titles* *明代職官中英辭典* has been uploaded to eScholarship, the University of California's institutional repository for OA. At the permanent URL is https://escholarship.org/uc/item/2bz3v185, users can browse and search 3,147 Ming official titles or download the dictionary for offline use. The government official titles in the dictionary are organized and presented in a three-level, government hierarchical structure, developed by the librarian team in consultation with scholarly publications and expert consultants. Within its first six months of release, the online dictionary received over 5,300 hits and was downloaded nearly 1,400 times. Feedback from scholars has been overwhelmingly positive – "it is very impressive," and

[1] University of California, Irvine
[2] University of California, Berkeley
[3] Stanford University

"very, very helpful." In addition, the government official titles and hierarchical structure have been adopted by CBDB for enhancing its metadata infrastructure.

The bilingual dictionary could benefit not only the digital humanities research of the CDBD team, but people who are interested in different aspects of the Ming dynasty (1368–1644), including social life, military, economics, religion, education, and literature. Additionally, the English translations of Chinese historic official titles would allow western scholars, who may not necessarily study China, to conduct comparative research of government systems between the East and the West.

Overall, besides being highly collaborative, the project is innovative. First, through working closely with Ming scholars and expert constants, as well as reading related historical literature and research work, we learned a great deal about scholar needs and subject knowledge. This laid the groundwork for librarians to experiment with expanding our roles from curation to creation, as advocated by Zhang *et al.* (2015).

Second, taking advantage of Web 2.0 technology, we designed and deployed the expert-based crowd-translation system to collect English translations of over 1,500 government official titles, many of which are unusual and obsolete, within just a year. This remarkable accomplishment is in line with the findings of Anastasiou & Gupta (2011) in terms of the effectiveness (high quality for complex texts) and efficiency (short turnaround) of crowd-translation.

The third innovative element is librarians' global outreach efforts made to the world's scholarly community. Traditionally, librarians' connection with scholars has been limited within their own institutions. And humanities scholars tend to work alone rather than collaboratively (Stone, 1982), and to have their own tight-knit circles (Nicholas, 2015). For this project, we made tremendous outreach efforts, including attending and speaking at scholarly conferences in North America, Europe, East Asian and Australia, to ensure that more Ming scholars around the world to contribute and use the two OA products.

References

ACRL (2003). *Principles and Strategies for the Reform of Scholarly Communication*. Retrieved on June 22, 2018 from http://www.ala.org/acrl/publications/whitepapers/principlesstrategies.

Anastasiou, D. and Gupta, R. (2011) Comparison of crowdsourcing translation with Machine Translation. *Journal of Information Science*, 37: 637-659.

Blanke, T., Kristel, C. & Romary, L. (2016) Crowds for clouds: recent trends in humanities research infrastructures. In: Agiati Benardou et. al. (eds). *Cultural Heritage Digital Tools and Infrastructures*, Ashgate Publishing, Retrieved on February 15, 2016 at http://arxiv.org/pdf/1601.00533.pdf.

Hucker, C. (1985). *A Dictionary of Official Titles in Imperial China*. Stanford University Press.

Kennedy, G.A. (1983). The relationship between acquisitions and collection development. *Library Acquisitions: Practice and Theory*, 7 (1983): 225-232.

Stone, S. (1982). Progress in documentation: humanities scholars: Information needs and uses. *Journal of Documentation,* 38(4): 292–313.

Nicholas, D. (2015). *Using, Citing and Publishing Scholarly Content in the Digital Age: Case Study of Humanities Researchers.* Retrieved on December 15, 2015 at http://ciber-research.eu/download/20151005-Wykorzystanie_cytowanie_i_publikowanie.pdf.

Zhang, Y., Liu, S. & Mathews, E. (2015). Convergence of digital humanities and digital libraries. *Library Management*, 36(4/5): 362-377

Leveraging the Japanese Biographical Database as a digital resource for education and research

Leo Born[1]

Overview

The **Japanese Biographical Database** (JBDB) is a web-based resource (https://www.network-studies.org) intended to provide biographical information on Japanese historical figures and their personal, social, and political networks. This paper will focus on the database design, the development of the web application with particular emphasis on the visualization component, and hands-on results of using this resource in a university classroom setting.

Initially starting with research on Rai Shunsui (1746–1816) in 2012, the PostgreSQL database currently encompasses entries on ca. 5,500 individuals and ca. 7,000 events pertaining to these individuals and their interactions and is steadily growing. The database spans roughly 400 years, from 1524 to 1939, with most of the entries falling into the later Edo period (1600–1868). While some of the attributes are specific to the Edo period, or are at least optimized for it, the database and the web application can handle any time period with ease; in the future, the user interface will be more flexibly adapted to any given context for a new entry.

While the database architecture itself is built upon the architecture of the Harvard University **China Biographical Database** (Harvard University et al., 2018), we developed a new, modern web application to access the database. The tool is intended to be aimed at researchers and students alike, allowing to search all entries by date, social status, and other filters as well as visualize networks of interest in a dedicated visualization component. The web application is written in JavaScript on both client- and server-side, using an **Express.js**-based framework to create a RESTful API and to access the database, and **AngularJS 1** to serve the client. We will discuss how our API can be used to share data with third-party applications as well as how our application might benefit from external open data. Due to its modularity, we can easily extend the functionality of the application beyond what is currently possible: for example, any IIIF-standardized library can be used to include external visual data and a GIS component employing **Leaflet.js** is slated to be implemented in the near future to map the networks and constituting events on an interactive map.

Visualizing and analyzing historical social networks

The visualization component is built on top of **vis.js**[1] and the resulting visualizations are interactive, meaning that elements can be searched, re-arranged, highlighted, or hidden, either manually or automatically based on various filters, including a temporal filter (**time slider**) and an attributional one (e.g. based on gender or profession). All this happens on-the-fly and non-destructively, enabling users to freely manipulate and analyze data in the database. On top of that, the visualization tool allows basic graph-theoretical calculations and operations on the resulting networks, for example measuring centrality or finding paths or clusters, as well as featuring an export function for inclusion of the graphs in publications. The visualizations are based on the relations between persons – an easy induction given the inherent relational nature of our PostgreSQL database. Most relevant for the visualizations are events stored in the database – these are comprised of meetings and exchanges (of letters or of goods) – and kinship and non-kinship relations between people. A small example graph highlighting the interactions of Rai Shunsui between 1788.6.14 and 1790.5.26 based on meetings and exchange of letters is given in Figure 1.

[1] Heidelberg University

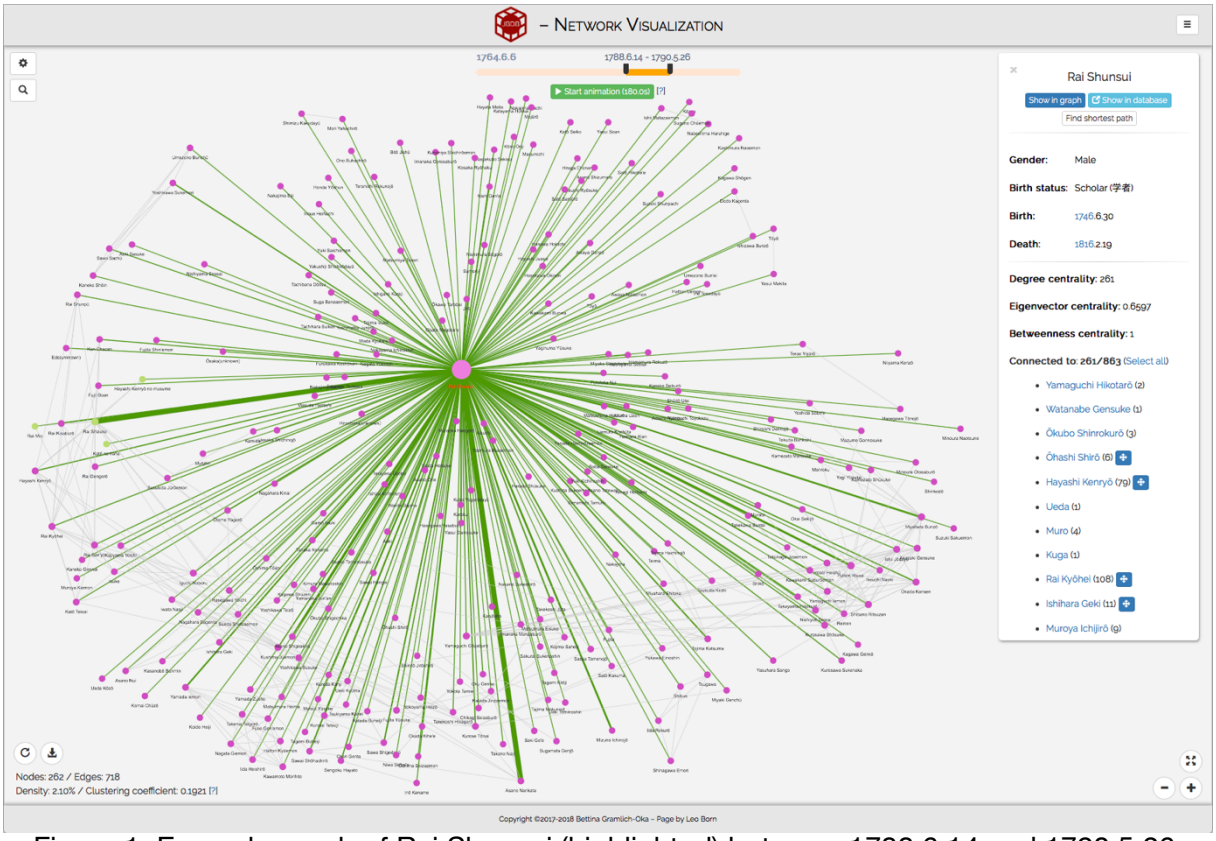


Figure 1. Example graph of Rai Shunsui (highlighted) between 1788.6.14 and 1790.5.26 (262/864 nodes, 718/2096 edges). Edge thickness indicates edge weight and nodes are scaled by degree centrality. The figure is best seen in color

With the addition of further biographical data, we aim to study patterns of societal changes that emerge out of familial and social relations among people, while also giving other researchers the ability to query the database in accordance with their own research questions. Up to this point, data has been input manually by a core team of project members, where biographical, event and kinship information has been extracted from the personal diaries of the persons under research (these are especially the *Shunsui nikki* for Rai Shunsui and the *Baishi nikki* for Rai Shizu, Shunsui's wife) and compiled from relevant biographical and historical lexica such as the *Nihon jinmei daijiten* or the *Kokushi daijiten*.

Our latest efforts encompass the automatic transfer of data from external resources to which we are granted access, such as the digitized person register of the diaries of Takayama Hikokurō (1747–1793)[2], and data input in the context of a graduate seminar at Sophia University's Graduate School of Global Studies. The seminar *Digital Humanities in the Classroom: Nineteenth Century Japan* starts off a new phase of employing and expanding the database and web application as an educational tool in the classroom, where we hope to provide a resource to easily query, contextualize, and visualize historical figures and their networks, while at the same time gathering new data for the database. This will also facilitate the students' ability to analyze and understand historical events from a prosopographical perspective by employing a **distant reading** (Moretti, 2013) methodology as well as enabling them to internalize and make use of a network-centric methodology for history studies.

We will also discuss potential enhancements to the core database entities in order to model, for example, organizations such as temples. This would allow for clustering networks by organizations or by representatives of organizations, enabling a more generalized analysis of the data by means of abstract concepts (e.g. "Did Rai Shunsui correspond with any temples at all and if so, when?"). Furthermore, as we strive to continuously improve the database and the application through an engaged dialogue with those who use them, we will also evaluate feedback from the students and discuss how

we can enhance the database to further our goal of a platform suited towards both education and research.

References

Harvard University, Academia Sinica and Peking University (2018). China Biographical Database. https://projects.iq.harvard.edu/cbdb .

Moretti, F. (2013). *Distant Reading*. London: Verso.

[1] visjs.org.

[2] This data set was compiled by the Ōta City *Takayama Hikokurō Kinenkan*.

Topic modelling as a Tool for Researching the Polish Daily Press Corpus ChronoPress of the Post-war Period (1945–1962)

Adam Tomasz Pawłowski[1], Tomasz Walkowiak[2]

The subject of the presentation is the use of topic modeling to generate key words characterizing the Polish press of the years 1945-1962. The press is considered here as a reflection of long-term civilizational trends, and also of specific phenomena distributed on the time axis that are in the main focus of public attention. We adopted an open-ended definition of keywords, combining both informational and cultural meanings. It is in the latter sense that Wierzbicka defines key words or collective symbols typical for various cultures (Wierzbicka, 2007; Panagl, 1998; Pisarek, 2003). Previous experiments indicate that automatic generation of topics can be used to discover concepts that reflect the fundamental and most common features of the culture in which the body of the text is produced (Maryl and Eder, 2017).

The goal of the study is to conduct an automatic analysis of the content of Polish press (1945-1962) on the basis of the ChronoPress corpus, using the topic modeling method. The specific objectives were: 1) to indicate topics that would adequately represent the main themes in a large set of press samples; 2) to create a synthetic representation of subjects found in the Polish press. The second goal is feasible given that one is able to determine an appropriate number of topics to generate, meaning that there are not too many and not too few of them. If the number of topics is too large, the result is a fragmentation of the description and random sets appear which do not bring valuable information. If this number is too small, the topics are merged into large collections that are inconsistent.

The study was conducted on 68,000 text samples from the ChronoPress corpus (1945–1962). An average sample length was 300 words. The samples were processed with a morpho-syntactic parser for the Polish language so that only nouns were analyzed. Metadata on the publication date of the periodical allowed us to generate chronological charts showing the variability of dominant topics over time.

The *Latent Dirichlet Allocation* (LDA) method was used for topic modelling. Its aim is to generate lexeme collections that are significant and specific to a given corpus on the basis of lexeme frequency and the scope of their presence in samples (Blei et al., 2003). Nouns which appeared in more than 80% of the documents were filtered out. This threshold is often referred as *max_df* (e.g. in the scikit library) and is used to remove terms that appear too frequently and do not convey valuable information.

In this study, topics of various sizes were generated and their value as potential key words was evaluated using the criterion of semantic coherence and cognitive value. It was assumed that there would be three categories of generated information clusters: coherent semantically and cognitively valuable, neutral (without significant errors but useless) and inconsistent or erroneous. The best results were obtained when the number of topics approached one hundred. Smaller collections revealed much worse results (they were semantically inconsistent). In addition, the topics were displayed on a chronological axis to show the evolution of the themes discussed over time.

The tests conducted allowed us to create a sort of semantic microphotography of the main themes that appeared in the Polish press during the period 1945-1962. It consists of ca 80 topics, every one being actually a small narrative. The most representative topics will be presented and discussed during the conference. However, displaying all of them would be a challenging task in a printed paper.

The resulting topics were grouped on the basis of lexical similarity in 39 intermediate subjects, and these were reduced in turn to 12 master ranges (labeled a posteriori).

[1] University of Wrocław

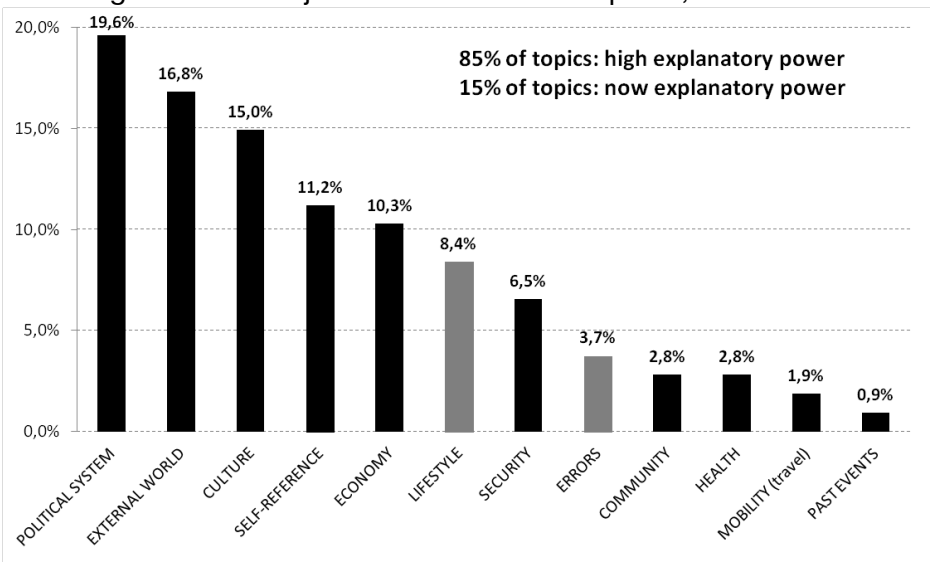
[2] Technical University of Wrocław

Tab. 1 Thematic scope of the Polish press, 1945–1962, as defined on the basis of topic analysis

Master range	Themes	Master range	Themes
POLITICAL SYSTEM	administration state politics management communism law	EXTERNAL WORLD AND SURROUNDINGS	countries cities farming nature land sea
COMMUNITY	Poland nation	HEALTH	health hygiene
ECONOMY	industry work transport	SELF-REFERENCE	language time numbers
SECURITY	war army	PAST EVENTS	history
CULTURE	arts religion education media science sport	LIFESTYLE	humans family house clothing food
MOBILITY	transportation travel	ERRORS	–

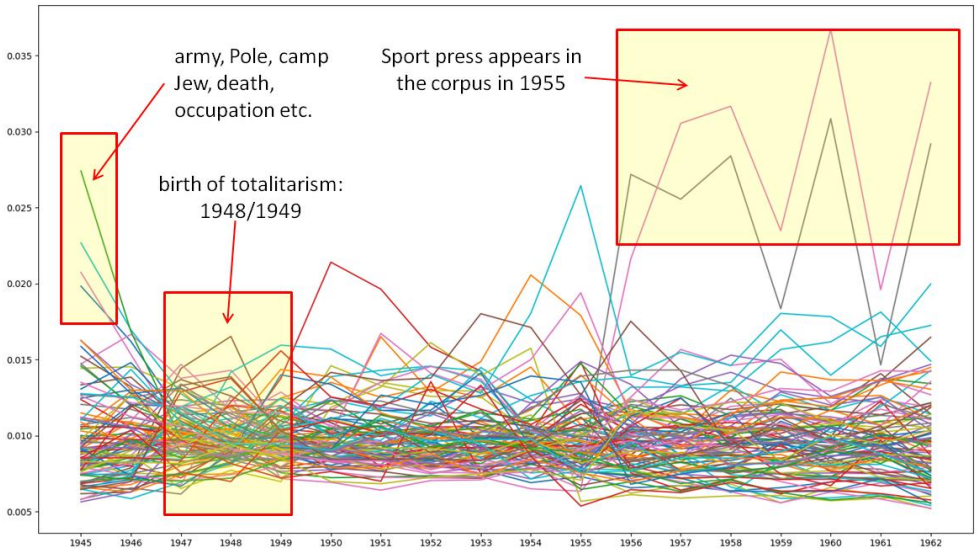
A survey conducted on a group of respondents found that out of the 100 topics analyzed over 85% had a high informational value (Figure 3). Only 3,7% were considered erroneous, while 11,2% expressed self-referential linguistic relationships between lexemes (they increase text coherence but do not convey semantic information).

Fig. 1 Broad subject areas in the Polish press, 1945-1962



An interesting result was provided by chronological analysis of topics (Pawłowski, 2016). It highlighted dominant themes in subsequent years. Figure 2 shows one hundred topics displayed as lines (the values correspond to the probability of drawing a given thematic range from the whole collection of samples in a given year).

Fig. 2 Chronological evolution of topics in the Polish press for the period 1945–1962



Test conducted prove that topic modeling is an effective text-mining tool when applied to a large body of short text fragments. With over 60,000 text samples, the generation of a set of one hundred topics can be considered as a good synthesis. Attempts to generate fewer topics caused information distortions. Topic modeling proved to be effective also in a chronological analysis of large text collections.

Keywords: *press, Polish, topic modelling, ChronoPress, 1945–1962, corpus linguistics*

Bibliography

Blei, D., Ng A. and Jordan, M. (2003). “Latent Dirichlet allocation.”, *Journal of Machine Learning Research* 3(4–5): 993–1022.

Maryl, M. and Eder, M. (2017). “Topic Patterns in an Academic Literary Journal: the Case of ‘Teksty Drugie’.” *Digital Humanities 2017: Conference Abstracts*. Montréal: McGill University and Université de Montréal, pp. 515–18.

Panagl, O. (ed.) (1998). *Fahnenwörter der Politik. Kontinuitäten und Brüche*. Böhlau Verlag: Wien, Köln, Graz.

Pawłowski, A. (2016). “Chronological corpora: Challenges and opportunities of sequential analysis. The example of ChronoPress corpus of Polish.” *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University and Pedagogical University, pp. 311–12.

Pisarek, W. (2002). *Polskie słowa sztandarowe i ich publiczność*. Kraków: Universitas.

Wierzbicka, A. (1997). *Understanding Cultures Through Their Key Words*. Oxford: Oxford University Press.

Studying Topics, Gender, and Impact in a Corpus of Czech Sociological Articles

Radim Hladík[1]

1. A corpus of Czech sociological articles

This paper will present a new corpus of Czech sociological articles assembled by the author and made available in the LINDAT/CLARIN repository. The articles were selected from the *Czech Sociological Review*, a journal that can be characterized as both a "core" – an important journal with long tradition and the only Czech sociological journal indexed in Web of Science – and a "generalist" – its aim and scope are not explicitly limited to any disciplinary specialty – publication outlet for Czech sociology. In this sense, the journal can be considered as representative of the discipline in the national framework. A corpus constructed from the journal therefore captures an important part of writing in social sciences in the Czech Republic.

At the beginning of the project, the journal website was scraped to obtain metadata and full-text versions of the articles. Missing texts were provided upon a request from the author to the editorial office of the journal, which kindly provided scanned and OCRed volumes. For concerns over OCR errors, digital-born material was preferred whenever possible. Only original scientific papers were to be included in the corpus. Therefore, available metadata on the articles' categories and manual inspection guided the selection. In borderline cases, features such an existing abstract, a list of references, or length were taken into account. Furthermore, all texts that were translated from languages other than Czech, regardless of the whether they were originally written for the journal or not, were also discarded from the collection. Ultimately, 522 articles were included in the corpus and cover the period 1993–2016. The journal publishes 4 issues in Czech language annually. When taking into consideration the above restrictions, choosing approximately 5 articles per issue seems to provide a good coverage of the peer-reviewed content. The raw PDF files had to be further processed to make them amenable to lemmatization and morphological tagging with the MorphoDiTa software (Straková et al., 2014), which is currently a standard tool in the construction of Czech language based corpora.

As a use-case scenario for the corpus, the author will introduce a study of Czech sociology as it is recorded in the texts of its primary journal. The research design is based on the combination of the textual corpus with additional sources of information: metadata, citation counts data, and topic models (Blei et al. 2001).

2. Modeling topics in the corpus

Topic modeling as a method has been used in many implementations, including, for example, a study of the history of ideas (Hall et al. 2008) or research on bibliometric impact in computer science (Mann et al. 2006). In digital humanities, Schöch (2017) employed it to identify types of topics and subgenres in the collection of French drama texts. Goldstone and Underwood (2014) used topic modeling to trace long-term trends in the evolution of literary studies as an academic discipline and showed how the method can be useful to provide new perspectives on the history of literary studies. In particular, they found topic modeling useful in revealing "patterns of representation that may not be visible to individuals participating in them" (Goldstone and Underwood 2014: 28). Maryl and Eder (2017) demonstrated the utility of the approach to describe concentric intellectual structure of topics in literary studies and, like Goldstone and Underwood, they traced temporal shifts in topics representation.

The use of topic models in this paper builds on this line of research and adds explicit comparison with the existing studies of the *Czech Sociological Review* that rely on content and bibliometric analyses (Janák and Kloboucký 2014; Skovajsa 2014; Vohralíková

[1] Institute of Philosophy of the Czech Academy of Sciences / National Institute of Informatics

2002; Nešpor 2014). Additional data on citation counts and the gender of the first authors allow describing differences in the scientific impact of topics and the position of women in various subfields of sociology. In contrast to previous studies of the journal, the computational approach effectively handles mixed-data and mixed-methods. The author argues that the main advantage of digital methods in this case is precisely the seamless integration of otherwise disparate techniques and the ensuing production of novel and more complex insights about the disciplinary history of Czech sociology.

Topic modeling was conducted using the standard Latent Dirichlet Allocation algorithm with several variations on the number of topics (20, 25, 30, 35, 40, 50, 60). With the increasing number of topics, interpretation has become more difficult not because the topics would become less semantically coherent, but it was becoming more difficult to align the topics with preexisting domain knowledge. Needless to say, American Sociological Association uses over 50 themes to organize its sections, so it could serve as an estimator of the upper boundary. In a previous content analysis of the journal, the authors (Janák and Kloboucký 2014) used 30 human-assigned topics. Eventually, the model with 35 topics was chosen to keep the analysis in a similar range. The small size of the corpus also does not allow for more generous approach. Figure 1 shows under the parameters of the model, many topics had less than 10 articles in them.

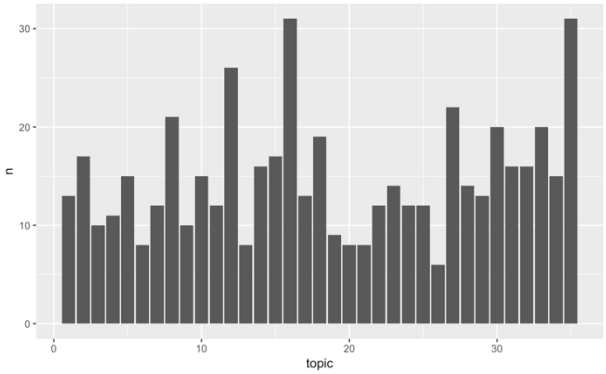


Figure 1: Number of articles per topic in the 35-topics model

Top rated words for each topic were inspected and interpreted as instantiations of a single category on the basis of available domain knowledge. The analysis of the topics based on the gender of the first author (Figure 2) shows a big variety among the topics on this criterion. Although women are generally underrepresented as authors in the corpus, some topics (related to health and gender) are predominately occupied by female authors. The analysis of citation counts (where n = 499 of records matched in WoS), reveals that topics with majority of female first authors are also among those that are generally less cited (Figure 3). Topic analysis thus not only provides insights into the intellectual structure of Czech sociology, but it also readily reveals that this structure is also strongly hierarchical.

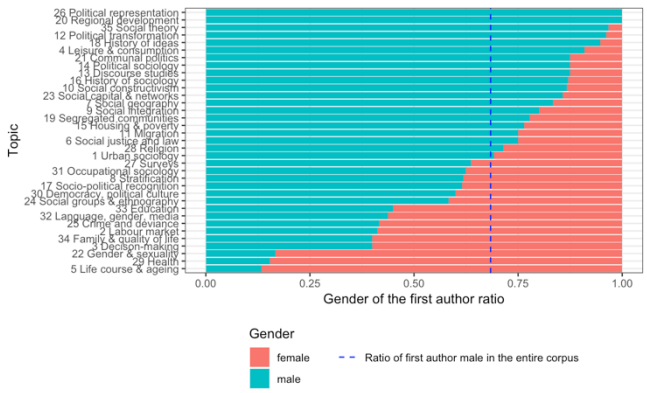


Figure 2: Distribution of genders among the first authors

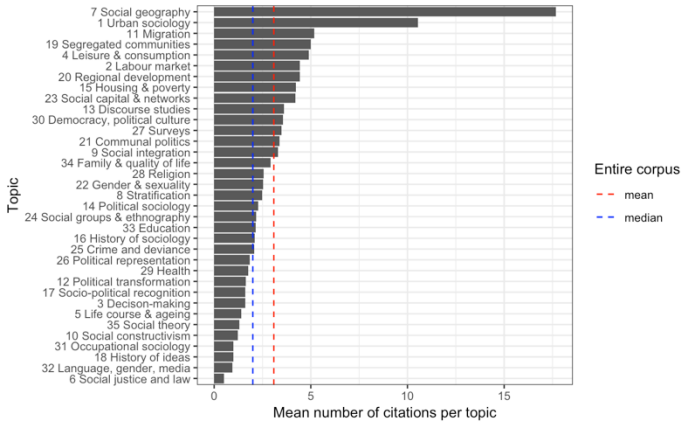


Figure 3: Average citation counts for articles in each topic

Bibliographical references

Blei, D., Ng, A. and Jordan, M. (2001). "Latent Dirichlet Allocation." In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 601–608. NIPS'01. Cambridge, MA: MIT Press.

Goldstone, A. and Underwood, T. (2014). "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45 (3): 359–384.

Hall, D., Jurafsky,D. and Manning. C. (2008). "Studying the History of Ideas Using Topic Models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 363–371. Association for Computational Linguistics.

Janák, D., and Klobucký. R. (2014). "What Would We Know about Sociology If We Only Read Sociologický Časopis and Sociológia? A Content Analysis of Two Journals of Sociology from the Velvet Revolution to the Present Day." *Czech Sociological Review* 50 (5): 645.

Mann, G., Mimno, D. and McCallum, A. (2006). "Bibliometric Impact Measures Leveraging Topic Analysis." In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, 65–74. JCDL '06. New York, NY, USA: ACM.

Maryl, M. and Eder, M. (2017). "Topic Patterns in an Academic Literary Journal: The Case Of Teksty Drugie." In *Digital Humanities 2017: Conference Abstracts*, 4. ADHO.

Nešpor, Z. (2014). "Padesát let české sociologie náboženství na stránkách Sociologického časopisu" *Czech Sociological Review* 50 (5): 735.

Schöch, C. (2017). "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama" 11 (2).

Skovajsa, M. (2014). "Celková a Zahraniční Citovanost 'Sociologického Časopisu': Výsledky Citační Analýzy." *Sociologický Časopis / Czech Sociological Review* 50 (5): 671–712.

Straková, J., Straka, M. and Hajič. J. (2014). "Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition." In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 13–18. Baltimore, Maryland: Association for Computational Linguistics.

Vohralíková, L. (2002). "O Čem Psali a Bádali Čeští Sociologové v Devadesátých Letech 20.Století" *Sociologický Časopis / Czech Sociological Review* 38 (1/2): 139–51.

What did Journalists Mention in the Russian Press?: Comparison of Articles about Yeltsin's Presidential Addresses to the Federal Assembly

Mao Sugiyama[1]

In 1994, Yeltsin's administration began to publicize its policies in the Presidential Address to the Federal Assembly in the Kremlin. According to the Russian Constitution, Section 84, the role of the Presidential Address is to inform the Russian public of the government's domestic and international policies. As the media plays an important part in spreading information about the policies to the country, this study focuses on how the Russian media reported on Yeltsin's presidential addresses by looking at articles from two popular publications, a tabloid and a broadsheet.

Ijima (2009) indicates that freedom of the press has been limited in Russia, especially after Putin's assumption of the presidency in 2000. Ognyanova (2010) investigated the existing practices of Internet control and censorship in Russia. She indicated that aspects of the culture and development of Russian society make it easy to ascertain a priority of patriotism and social solidarity exceeding individual rights and freedom of speech. Hakamada (2012) wrote a paper about domestic and international affairs such as falling population, immigration issues and ethnic problems in two Russian newspapers "Независимая газета" (Nezavisimaya gazeta) and "Аргументы и факты" (Argumenty i facty). These previous studies give us an understanding of the current situation of the media in Russia, however, as far as I know, there is no research investigating the different points of views between speakers, in this case the Russian President, and reporters. Therefore, this study compares the word usage between the speakers and the press.

The research questions of this study are i) to explore the different points of view between the Russian Presidential Addresses to the Federal Assembly, which were given by Yeltsin from 1994 to 1999, and the Russian press, and ii) to analyze the contrasting standpoints between the broadsheets "Nezavisimaya Gazeta" (NG) and the tabloid "Komsomolskaya Pravda" (KP). These publications have the highest number of articles about Yeltsin. "NG" is a respectable and independent newspaper, while "KP" is a tabloid, owned from 1925 by an oligarch, and from 1991 published as an independent news source. The articles were collected in the Russian State Library from microfilms of newspaper reports about Russian Presidential Addresses to the Federal Assembly given by Yeltsin from 1994 to 1999. After collecting the data, we transcribed it into text, making a Russian Press Corpus.

The study conducts analyses on the corpus, focusing on the three subcorpora: (1) Yeltsin's addresses; (2) broadsheet coverage; (3) tabloid reporting. The subcorpus of Yeltsin' addresses consists of 97,648 tokens and 7,454 types. The corpus sizes of the Russian Press are 21,291 tokens and 7,974 types for "NG", and 5,325 tokens and 2,398 types for "KP". By doing a correspondence analysis on these data using the top 100 frequent words in each subcorpus, the results show that there is a difference between the word usage in Yeltsin's public addresses and the media's coverage of the articles. To observe in more detail, a comparison has been made based upon the typical words in each text. The typical words from the result of correspondence analysis in Yeltsin's addresses are *law* (1994); *support*, *citizens* and *interests* (1995); *crisis* and *economy* (1996); *action*, *government* and *organization* (1997); *regional*, *budgetary*, *financial* and *business* (1998); *Russia*, *country*, *political* and *reform* (1999). The results of a correspondence analysis (Figure 1 and 2) show that even between these two publications, there was a difference in points of view.

[1] Osaka University

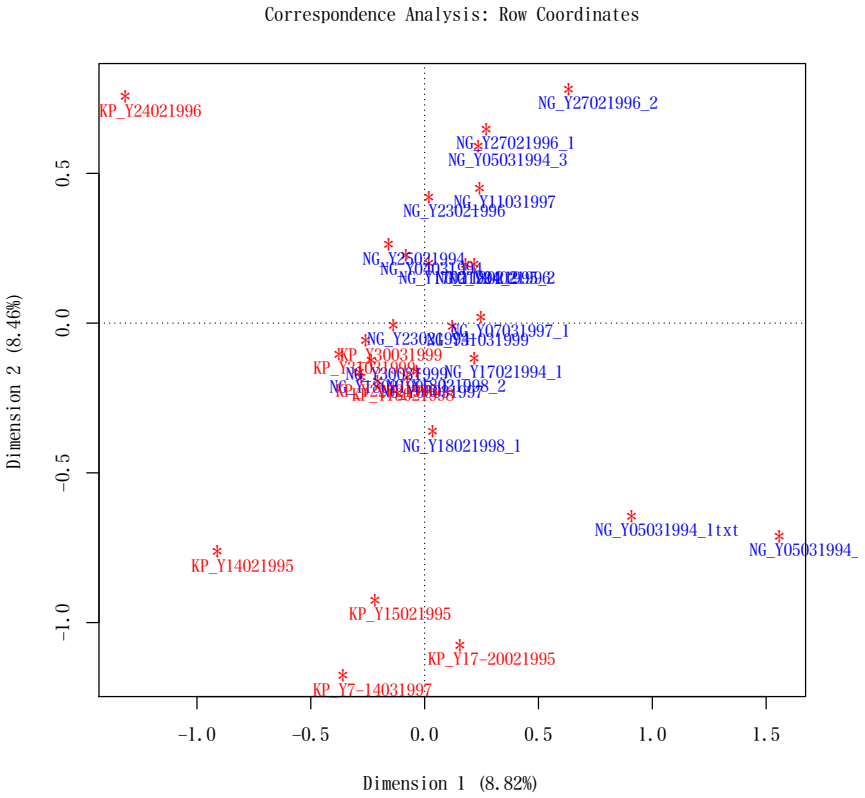


Figure1: The relationships of texts between “Nezavisimaya Gazeta” and “Komsomolskaya Pravda” (top 100 words)

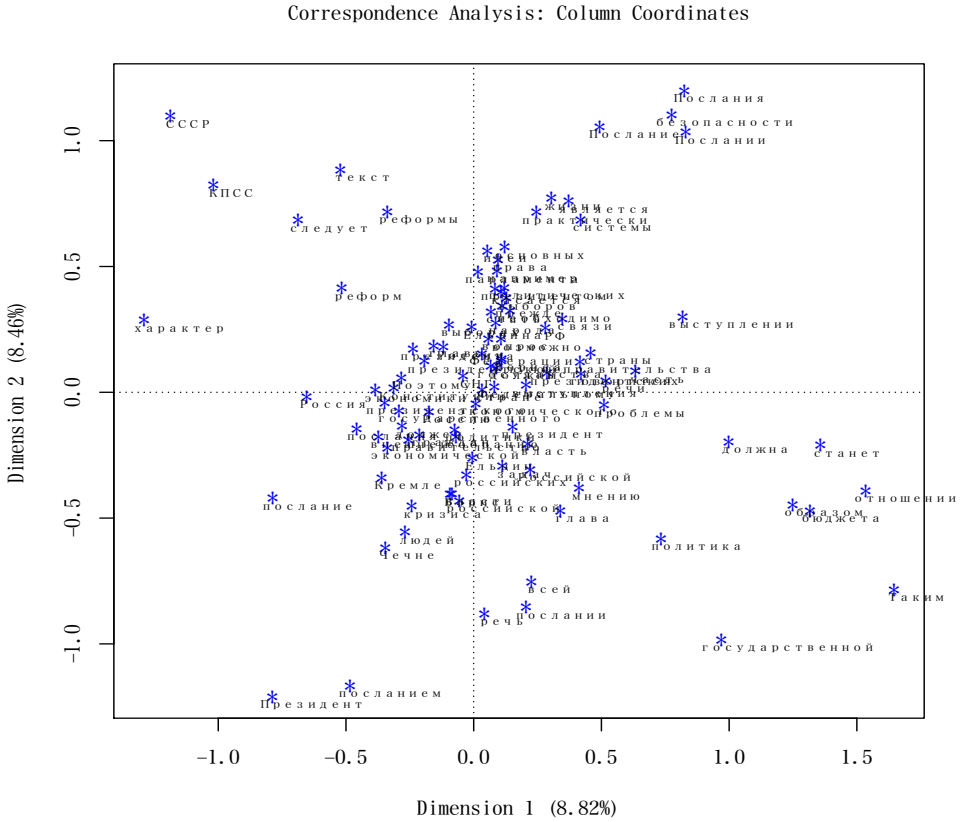


Figure 2: The relationships of words between “Nezavisimaya Gazeta” and “Komsomolskaya Pravda” (top 100 words)

Notice that there were no articles about Yeltsin's 94-year-address in "KP" in 1994. In "NG", journalists mentioned *reform* and *responsibility* in 1994, but this reform was not concerning Russian law, but a financial situation. Journalists in "KP" paid attention to relationships with the Chechen Republic in 1995. In 1996, "NG" reported about homeland security and "Komsomolskaya Pravda" reported on the opinion of an opposition party, but not on the crisis and economy of Yeltsin's address in 1996. In 1997, the tabloid "KP" pointed to the condition of Yeltsin's health, whereas the broadsheet "NG" took notice of the life of citizens, the financial system and Russia itself. The correspondence analysis shows the intersected part of these two coverages in 1998 and 1999. In these years, the two Russian presses mentioned the economy and Yeltsin.

Comparisons within each year's articles and Yeltsin's typical words in the address revealed both perception and similarities. The broadsheet "NG" tended to pay attention to the financial situation in Russia, whereas the tabloid "KP" quoted the other person's voice and critical opinion. A follow-up Random Forest classification experiment conducted using the R package in CasualConc using the top 1000 words shows an OOB error rate of 30% in classifying between "NG" and "KP". The key words, extracted based on the mean decrease in Gini, of "NG" are related with economy, duty, and situation, whereas the key words in "KP" are personal names.

This study indicates that Yeltsin informed many kinds of policies as a leader of the country, but the Russian people, or, at the least, journalists, critically evaluated Yeltsin's administration and considered the most important problems to be the economic situation and Yeltsin's word use, not the system of law or constitution of Russia.

References

Iijima, K. (2009) *Rosia no masumedia to kenryoku* [The Russian mass media and power], Eurasia booklet 133.

Imao, Y. (2017) CasualConc (Version 2.0.7) [Computer Software] URL: https://sites.google.com/site/casualconcj/download.

Hakamada, S. (2012) *Rosia no masumedia ha naisei・kokusai jyousei wo ikani ronji houjite iruka (2)* [How Russian media report internal affairs and international situations (2)], Aoyama journal of international politics, economics and business 86: 107-146.

Ognyanova, K. (2010) Careful What You Say: Media Control in Putin's Russia – Implications for Online Content, International Journal of E-Politics, 1(2): 1-15.

Building Oral Narrative Archives of Contemporary Events: Merits and Challenges of Open Data in Digital Social Sciences

David H. Slater[1], Flavia Fulco[1], Robin O'Day[2]

This presentation attempts to sketch out what might be called "Digital Social Sciences," using as illustration our 8 year project, the largest oral narrative archive of the 2011 triple disasters, **Voices from Tohoku**. While we have used many of the tools and approaches familiar in Digital Humanities, our focus on contemporary events and the digital recording and archiving of living subjects (narrators) has allowed us to re-think some of the goals, assumptions and practices in this field.

Our presentation has three main purposes:

1. to introduce the goals of oral narrative research within the context of a digital archiving project
2. to demonstrate collection methods, data management and dissemination practices for our archives
3. to addresses the challenges of privacy and ethical issues related to the archiving of living subjects

We will use examples from our archive **Voices from Tohoku** (東北からの声 https://tohokukaranokoe.org/) on the triple disasters of 2011, and smaller archives in progress on political activists, homeless and refugees in Tokyo.

1 Definition and Goals: The term "digital social sciences" (Spiro 2014), even more than digital humanities, is still evolving and often covers a relatively wide and heterogeneous range of different scholarly activities, including automated information extraction, social network analysis, geospatial analysis and complexity modeling. In our projects to date, we have focused on a restricted set of practices that include a) the collection/creation of new data through semi-structured oral narrative interview, b) that are then organized as digital objects with meta-data to search, compare and analyze the narratives in revealing ways, c) with the intent to make our research available to a wider public in a timely and accessible manner. Our goal is to create an archive of authentic voices (生の声), a chance for participants to bear witness to significant social events and contexts by telling their story in their own words.

2 Collection, Management and Dissemination: This project began as an advanced undergraduate methodology seminar at Sophia University on oral narrative when the 3.11 disaster hit in 2011. Building upon our volunteer relief efforts, we began holding informal talk sessions for residents in the affected areas. This turned into more formal interviews during the semester, as we developed student knowledge of the ethnographic context, interview techniques and technological skills. Through the semester, we created thematic tags and codes using grounded theory techniques to organize existing interviews and guide new interviews. As data accumulated over the years, we began using archiving platforms and retrieval methods that insured compatibility across sub-collections in different areas (eg, Koriyama, Rikuzentakata, etc). The result is both a closed scholarly archive of more than 500 hours of data, and as well as a completely open website 東北からの声 (https://tohokukaranokoe.org/) with more than 80,000 hits. We will discuss some of the pedagogical and scholarly challenges of this important arc of the research cycle.

[1] Sophia University
[2] University of North Georgia

3 Challenges: We will focus on the following challenges that stemmed from the fact that unlike most archives or digital humanities projects, we were working with living subjects--our narrators.

- **Ownership and control of data:** Focusing on the central role of our narrators, we have always asserted that "because it is their story, they should decide how it is best represented." This means that we had to find some way for our narrators to retain rights and control of their interviews, even as we constructed the archive and developed the website--often a delicate balance. In the selection of clips for the open website, we worked as collaboratively as possible, making sure that narrators have a central role in the selection and display of their own clips. All narrators retain full rights and indefinite control over their own interviews.
- **Representativeness:** The criteria of inclusion lies at the heart of any archive, and when the archive is collaborative project, this dynamic is ever more problematic. For example, some narrators wanted to tell the crying stories **（かわいそう**) while others wanted to focus on forward facing (**前向き**) narratives, where the ethnographic context is often more complicated than either or both. How are these sometimes conflicting concerns addressed within the context of an archive of limited scope? When the wishes of our narrators' and our own understanding of the full context are in conflict, how are we to proceed?
- **Privacy and Exposure:** In principle, preserving each narrator's rights to privacy more easily secured when you involve them in the post-production process as collaborators. In the context of the 3.11 disaster, our narrators, feeling distrustful and skeptical of the mass media's rendering of their situation, were eager to publicly share their stories with us. But in more recent projects on political activists, Tokyo homeless and foreign refugees in Japan, the question of revealing of identity and personal information has been much more sensitive. While our narrators share our goal of using their stories in a larger effort to let the public better understand their situation (otherwise they would not give us an interview), we are often struggling to find a compelling and responsible way to to tell their story while still making sure that we do not put them at risk.

By reviewing some of the design, collection practices and archival structure (and by showing some clips along the way) we ask, 'Can we begin to point to a set of "best practices" in the emerging field of Digital Social Science?'

References

Slater, D. H. and Veselic, M. (2014). Public Anthropology od Disaster and Recovery 'Archive of Hope' (希望アーカイブ), *Japanese Review of Cultural Anthropology*, vol.15.

Spiro, L. (2014). Defining Digital Social Sciences, dh+lib. 4 http://acrl.ala.org/dh/2014/04/09/defining-digital-social-sciences/ (accessed 24 April 2018).

Digital archiving vernacular records of natural disaster in Northern Thailand

Senjo Nakai[1]

This research is designed to collect and archive in the digital map the folkloric records of an ancient catastrophic event that is believed to have occurred in present-day Wiang Nonglom Wetlands of the northern province of Chiang Rai around 460 CE (Phrayaprachakitchakonrachak, 1898/2014: 197). The legendary city of Yonoknakaphan is said to have been inundated under present-day wetlands after the people killed a giant albino eel. Incidentally, a team of geologists discovered the presence of an active fault in the wetlands (Wood et al., 2004: 64). Through two rounds of fieldworks in 2011 and 2017, the disaster-lore of the lost city was collected. It is divided into six types: folk belief, performing arts, oral tradition, literary arts, place name, and modern folklore, according to the modified categories proposed in *Handbook of Folklore Survey* (Ueno et al., 1987). The collected images, interviews and documents regarding the disaster-lore were archived on an online map (https://drive.google.com/open?id=17-c-5wlnYbjbmbpZYNqABEGHhOW-2PRV&usp=sharing). The online archive has been accessible to anyone who is interested in the disaster-lore or natural disasters in the northern region since July 2017.

First, the local informants confided the continuation of legend-related beliefs (folk belief). These beliefs have given rise to other types of disaster-lore, i.e., taboo, ritual and shrine. The locals have enacted the legend of the sunken city as a drama and a song (performing arts), and orally recited the legend as a part of a story-telling contest organized by the local tourism authority (oral tradition). The disaster-lore is also recreated as a poem and a novel (literary arts). The poem was created by a local poet and is now displayed in a folk museum in the wetlands. Despite the drastic change in the wetlands, place names still remind the locals of the legendary disaster, such as 'Nonglom (the sunken swamp),' 'Ko Mae Mai (The widow's island),' 'Mae Ha (to catch in the Northern Thai dialect),' 'Mae Lak Creek (to drag),' 'Mae Lua Creek (to gut in the Northern Thai dialect),' and 'Mae Kok River (to chop in the Northern Thai dialect).' These names are derived from the plotline of the legend. In more recent years, the legend was replicated as a museum, statues, murals and a painting (modern folklore).

Historical data is crucial for informing policy-makers and other stakeholders of the risk of future natural disasters and allocating resources for the management. However, there is a paucity of the data in Thailand, especially outside the central region. After the northern kingdoms were fully annexed into Bangkok-centered Siam in the early twentieth century, the modern centralized administration system, including the systematic collection of meteorological data, were implemented in the northern region in 1942 (Northern Meteorological Center, 2018). Before the introduction of the centralized administration system to the northern region, historical records were locally remembered through oral recitation or recorded in the local languages or Pali on palm leaf manuscripts (Stratton, 2004: 93). Despite being told in a non-scientific discourse, such vernacular records once reminded the local population of the risk of natural disasters. Japanese historian Shōji Sasamoto, in his study of Japanese folklore of debris flows, argues that vernacular media is one of a few 'soft' technologies, which may help non-experts take necessary actions to save the lives and properties from natural disasters (1994, 369).

Both social and technological resources have been allocated disproportionally in the capital city and its vicinity. The post-2011 Flood recovery efforts, for example, focused on anti-flood infrastructure in the Greater Bangkok Area, as well as on financial support for the affected (Otomo, 2013: 224-225). Although being eventually expected to trickle down to the regional cities, the fruits of the development are not fully enjoyed by the peoples outside the political and economic center. In natural disaster prevention and mitigation, the imbalanced distribution of resources poses a great challenge in collecting

[1] Thammasat University

empirical data of natural disasters, particularly low-probability high-consequence' disasters, with which modern disaster management does not effectively deal. Thailand is not free from serious natural disasters despite the general complacency about the risk of natural disasters. In fact, the country's vulnerability has been exposed during the 2004 Indian Ocean Tsunamis and the 2011 Flood. In fact, it is only as recent as November 2007 when the government, in response to the 2004 tsunamis, enacted the Disaster Prevention and Mitigation Act and appointed Department of Disaster Prevention and Mitigation for the national level planning and coordination of relevant agencies.

In the resource-strained provincial areas, such local resources may compensate for the lack of the long-term data of natural disasters and anti-natural disaster infrastructure. They are locally relevant and self-sustainable resources that can be mobilized for natural disaster resilience-building efforts. However, the rapid urbanization and accompanying mobilization of local populations pose challenges in the sustenance of vernacular records as informants indicated. Such an invaluable cultural heritage can no longer be maintained locally and transmitted only through indigenous forms of communication channels. Exogenous initiatives are expected to play an important role for the revitalization of the disaster-lore; governmental organizations, civil society, business sector and even international organizations, despite different agendas, have started exploring the potential of the disaster-lore as part of natural disaster prevention and mitigation efforts. In fact, folklore has been utilized in development projects around the world. The development of digital communication technology further enhances the potential of folklore in development projects in resource-constrained societies, and digital humanities can facilitate the revitalization of disaster-lore by archiving and reintroducing it to the local communities. At the next stage of this project, the disaster-lore will be introduced to local stakeholders via the online archive and workshops in order to build natural disaster resilient communities in Thailand.

References

Nao, O. (2013). Disaster prevention policies in Thailand and "Disaster Prevention and Mitigation Act of B.E. 2550" [Tainiokeru bōsaiseisaku to "butsureki2550nen bōsai oyobi keigenhō"]. *Foreign Legislation* [*Gaikoku no rippo*], 251: 239-246. (in Japanese).

Northern Meteorological Center. (2018). *Prawatsunutuniyomvithayaphaknuea* [*The history of the meteorological center in the northern region*]. http://www.cmmet.tmd.go.th/met/history.php (accessed 19 June, 2018) (in Thai).

Phrayaprachakitchakonrachak [Bunnag, Chaem]. (1898/2014). *Yonok Chronicle* [*Phongsawadan Yonok*]. Bangkok: Sipanya. (in Thai).

Sasamoto, S. (1994). *The run-off of snake, others, tree spirits: The folklore of historical disasters* [*Januke, ijin, kodama: rekishisaigai to denshō*]. Tokyo: Iwatashoin.(in Japanese)

Stratton, C. (2004). *Buddhist scripture of Northern Thailand*. Chicago, Buppha Press.

Wood, S. H., Singharajwarapan, F. S., Bundarnsin, T., & Rothwell, E. (2004). Mae Sae basin and Wiang Nong Lom: radiocarbon dating and relation to the active strike-slip Mae Chan fault, Northern Thailand. In *Conference Proceeding from International Conference on Applied Geophysics*, November 26-27, 2004, Chiang Mai, Thailand, pp. 60-69.

Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories

Mohamed Khemakhem[1,2,3], Carmen Brando[4], Laurent Romary[1,2,5], Frédérique Mélanie-Becquet[6], Jean-Luc Pinol[7]

Whereas mapping systems, such as Google or Bing Maps, have become nowadays the common tools to geocode addresses or to browse neighborhoods on modern maps, browsing a legacy map representing a geographical snapshot of historical cities is far from being accomplished. The issue is related in the first place to the lack of data allowing a system to map a given address to a throwback location. Such information are abundantly available in dedicated paper resources, such as legacy address directories[1]. But even digitised, mining the content of these resources remains limited due to the ad-hoc employed information extraction techniques.

Time machine[2] is a major large scale project aiming to bridge this gap, among many others, by analysing and valorising the content of legacy documents for the ultimate purpose of redrawing the historical, social and economical heritage of Europe. In this context, we present our approach and first results of a state-of-the-art technique for extracting information from digitised address directories.

Our labour has been motivated by two emerging factors. First, the public release of several digitised versions in high-definition from the legacy address directories “Annuaires-almanach” of Paris, made available by the French National Library[3]. The directory series, which had been edited since the 18th century, carry a joint description of the commercial activities and postal information of the french capital. Second, a recently implemented approach by Khemakhem et al. 2017 and Khemakhem et al. 2018 has given an information extraction system, GROBID-Dictionaries, which has been designed to structure digitised dictionaric resources by using machine learning models. We have been struck by the similarities in the structures of dictionaries and address directories, where both resources share a semasiological representation. In fact, the latters could be perceived as encyclopedic resources where locations are described as unique concepts.

We have used Text Encoding Initiative (TEI) as a common modeling standard and proposed a first encoding of entries in an address directory. We distinguish between two categories of entries (see table 1). The first is reserved for each entry describing a single occupant in a unique or a shared address. In other terms, to each number in a street, one or many occupants could be assigned and for each one of them corresponds an entry. The second category of entries gathers the description blocs of a common street. An entry in this case encapsulates information like the name of the street, length, neighbouring street, etc.

[1] Inria ALMAnaCH
[2] Centre Marc Bloch
[3] Paris Diderot University
[4] CRH (EHESS / CNRS UMR 8558)
[5] BBAW - Berlin-Brandenburgische Akademie der Wissenschaften
[6] LATTICE, ENS / Paris 3 / CNRS UMR 8094
[7] LARHRA CNRS UMR 5180

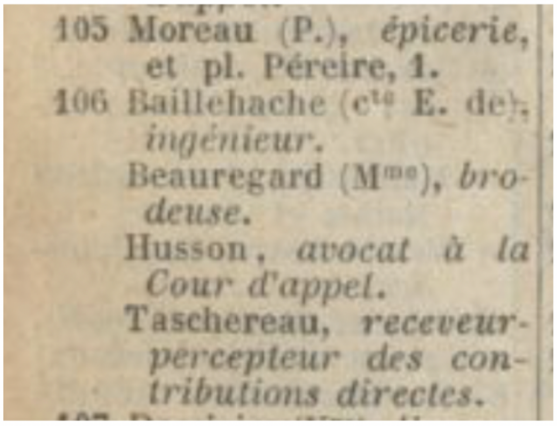
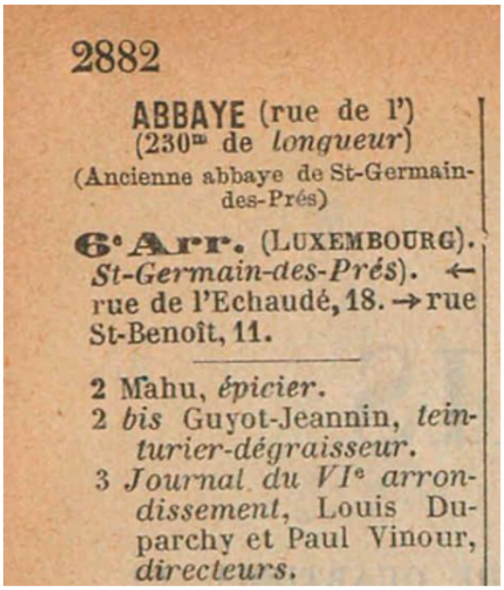
Digitised Sample	TEI Encoding
	<entry> <num>105</num> <form> <persName> <surname>Moreau</surname> <addName>(P.)</addName> </persName> </form> <pc>,</pc> <sense> <def>épicerie</def> <pc>,</pc> <lbl>et</lbl> <address>pl. Péreire, 1.</address> </sense> </entry>
	<entry> <form>ABBAYE (rue de V)</form> <sense>(230” de longueur) (Ancienne abbaye de St-Germain-des-Prés) 6* Arr. (Luxembodrg). St-Germain-des-Prés). <-rue de l'Echaudé, 18.-►rue St-Beno!t, 11</sense> </entry>

Table 1: Both images in lines 2 and 3 correspond respectively to excerpts of pages 3500 and 2882 of the 1901 release[4] of the Annuaires-almanach

The current architecture of GROBID-Dictionaries, based on cascading machine learning models, has been to a large extent able to support the presented encoding of the textual information and extract the macro structures. In fact, the first level of segmentation has the mission to differentiate between the different parts of a digitised page. The second level relies on a model for segmenting a page body to entries which will be further segmented in the third level to main semantic blocks.

Despite sometimes the noisy OCRised data (see table 1), till the third level, the only required adaptation of the system has been the implementation of a new label to mark the numbering of entries <num>. After this minor adaptation, a first experimentation of the system has shown interesting results for the first 3 segmentation levels, which we report in table 2.

Model	Annotated Data	F1-Score	
Dictionary Segmentation	10 Pages 7 training, 3 evaluation	**Micro Average**	**Macro Average**
		99.61	72.12
Dictionary Body Segmentation	319 Entries 270 training, 49 evaluation	**Micro Average**	**Macro Average**
		98.61	95.7
Lexical Entry	208 Entries 160 training, 48 evaluation	**Micro Average**	**Macro Average**
		90.31	91.36

Table2: Evaluation of the first three segmentation models

Although the models had given better results with dictionaries in previous experimentations, the current results are still considered impressive given the different nature of the address directories and the noise in the OCRs, especially for the first model. The outcome should be improved as soon as we annotate more data and further strengthen the selected features, if needed. To reach the complete encoding presented in table 1, we are investigating the creation of new models to be integrated in the existing architecture for processing the clusters of texts labels. Before considering building new models trained from scratch, the integration of models used for the same purpose in the GROBID[5] family projects is likely to be the most efficient solution, as for the parsing of addresses and person names. We are considering also to improve the OCRs for known entries such as the majority of street names, which could be checked against existing defined lists.

In conclusion, fueling a Time Machine with structured information extracted from legacy address directories does not seem to be an issue anymore thanks to the availability of the target digitised material and the advanced extraction techniques embedded in GROBID-Dictionaries. The existing architecture of the tool could be further improved by annotating more data, plugging in existing models or creating new ones to be applied in larger scale or on similar documents in other languages. Finally, our aim is to further retrodigitise releases of the Annuaires-almanach and geocode historical postal addresses listed there thereby to analyse commercial activity in old Paris taken from large amounts of historical sources as introduced by Kaplan and di Lenardo, 2017.

References

Kaplan, F. and di Lenardo I. (2017). “Big data of the past”. *Frontiers in Digital Humanities*, 4: 1-12.

Khemakhem, M., Foppiano L. and Romary L. (2017). “Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields”. *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Leiden. pp. 598-13

Khemakhem, M., Herold A. and Romary L. (2018). “Enhancing Usability for Automatically Structuring Digitised Dictionaries”. *Proceedings of GLOBALEX workshop at LREC.* Miyazaki. pp. 88-93

[1] Historical maps are evidently an important source of geolocalised information, our proposed approach aims to be complementary to well-known methods for georeferencing old maps and thus deals with a new kind of historical source.
[2] http://timemachineproject.eu
[3] http://gallica.bnf.fr/ark:/12148/cb32695639f/date
[4] http://gallica.bnf.fr/ark:/12148/bpt6k9763088f
[5] https://github.com/kermitt2

Matching methods: new approaches for the study of the *Online Dating* phenomena.

Jessica Pidoux[1]

Few studies in the social sciences and the humanities have examined the social issues, outcomes and implications of matching algorithms used by online websites and mobile dating applications (apps). While gaining increasing popularity, these dating applications are hidden behind "trade secrets" (Pasquale, 2015) that restrict the study of their programming practices to researchers. When scientists acquire some access to this data, research is often constrained by agreements negotiated and controlled by private companies. This paper introduces a mixed methodology combining qualitative and quantitative analysis, whose principal objective is to elucidate the matching process of dating apps and its impact on actors by analysing the features of their openly-accessible interfaces.

Matching algorithms have traditionally been the arena of computer scientists, with a focus on the selection of relevant features and the development of technologies (Jekel and Haftka, 2018; Xia et al., 2016; Yu et al., 2016; Tu et al., 2014). Deployed by dating companies, these algorithms and techniques, such as *machine learning*, make it possible to improve systems used for searching, classifying and recommending profiles on dating sites. They do not, however, address the underlying social implications of the information on which such programs rely and the decisions which they make.

Social scientists interested in the impact of quantification practices in the dating industry have been restricted in their study by problems of access and the proprietary nature of algorithms. In addition, approaches called *digital methods* (Venturini et al., 2018) or *digital ethnography* are often limited to the analysis of users (Vinck et al., 2018) and their personal data with legal and ethical implications. All of this, highlights the need for alternative methods to research the functioning and impact of digital platforms.

As part of my on-going doctoral dissertation, I propose a new perspective for the study of online websites and mobile dating apps. This paper focuses on one entry point, the Graphical User Interface (GUI) and its properties, without touching upon personal data. Openly accessible, interfaces contain features that can reveal the variables used by matching algorithms on dating sites. Indeed, the evolution of dating applications into "data-driven models" (Albury et al., 2017) has made the interface essential to collecting quantifiable data which can then be harnessed in mapping individual preferences.

More specifically, I study the "conceptual model" (Norman, 1988) used by dating companies. The model's design is essential for the user to understand the platform, as it allows them to create an appropriate mental image, or representation of the object, which makes sense of the operations of the app. Without a comprehensive conceptual model, users will not be able to recognize "affordances", that is the properties of an object and its relation with the subject that interacts with it (ibid.). For instance, I will focus on the user registration form, on the research bar, and other pages and functionalities without collecting user-generated content. Although GUIs only allow access to explicit features, they also offer rich possibilities for the analysis of algorithms in concert with complementary features. Characteristics of platforms, gathered through the manual cleaning and categorizing of data and the polishing of metrics lead to a clustering analysis. First, categories (centroids) of dating apps are found using similar features, and secondly the classification of multiple personæ conceived by developers and expressed in the platforms is analysed. By focusing on two case studies from a bottom-up perspective, I will demonstrate how researchers can analyse these openly-accessible features to understand how user experience is being shaped.

In a departure from popular viewpoints, such as the construction of online identity, couple formation or the improvement of technical systems, this discussion takes

[1] Digital Humanities Institute, EPFL.

a different perspective. It approaches the study, not from the angle of the user profile, but from that of the interface, to analyse human-computer interaction. This opens avenues for exploring the ways in which online websites and mobile dating experience is shaped, coloured and informed by dating applications through their platforms, leading to richer insights into the forging of modern relationships. From an academic viewpoint, this research *in progress* introduces methods, relying on open data, for the study of online dating without the need to access proprietary algorithms nor acquire sensitive private data.

References

Albury K., Burgess, J., Light B., Race K. and Wilken R. (2017). "Data cultures of mobile dating and hook-up apps: Emerging issues for critical social science research". *Big Data & Society*, July–December 2017: 1–11.

Jekel C. and Haftka R. (2018). "Classifying Online Dating Profiles on Tinder using Facenet facial embeddings". https://arxiv.org/pdf/1803.04347.pdf (accessed 26 June 2018)

Norman D. (1988). *The Psychology of Everyday Things*. New York: Basic Books.

Pasquale, F. (2015). *The Black Box Society. The Secret Algorithms that Control Money and Information*. London: Harvard University Press.

Tu K., Ribeiro B., Jensen D., Towsley D., Liu B., Jiang H. and Wang X. (2014). "Online Dating Recommendations: Matching Markets and Learning Preferences": *Proceedings of the fifth international workshop on social recommender systems, in conjunction with 23rd international world wide web conference*, Seoul, Korea, April 2014, pp. 787-792. https://dl.acm.org/citation.cfm?id=2579240 (accessed 26 June 2018)

Venturini, T., Bounegru, L., Gray, J. and Rogers, R. (2018). "A reality check(list) for digital methods". *New Media & Society*. http://doi.org/10.1177/1461444818769236 (accessed 26 June 2018)

Vinck D., Camus A., Jaton F. and Oberhauser P. N. (2018). "Localités Distribuées, Globalités Localisées : actions, actants et médiations au service de l'ethnographie du numérique". *Symposium*, 22:1, pp. 41-60.

Xia P., Shuangfei Z., Benyuan L., Yizhou S. and Cindy C. (2016). "Design of reciprocal recommendation systems for online dating". *Social Network Analysis and Mining*, 6, no 1: 32. https://doi.org/10.1007/s13278-016-0340-2 (accessed 26 June 2018)

Yu M., Zhang X. and Kreagery D. (2016). "New to Online Dating? Learning from Experienced Users for a Successful Match": *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, San Francisco, CA, USA, August 2016.

A Quantitative Analysis of Agatha Christie's Works Applying a Machine Learning Approach

Narumi Tsuchimura[1]

This study investigates stylistic changes in works by Agatha Christie using a machine learning method. Previous studies on style in Christie's works such as Lancashire and Hirst (2009), Le et al. (2011) and Inaki (2013) concern themselves with a limited number of works, so this study aims to analyse all of her works.

We analyse 66 of Christie's works published during 1920 and 1976. All the 66 texts are labelled as C1 to C66 in the chronological order of their publishing years. The last two novels, *Curtain* (C65) and *Sleeping Murder* (C66), while published in the 1970s, were written in the 1940s. We assume that the rest of her works were written just prior to their year of its publication. Short stories are excluded from this analysis, and in order to minimize difference between genres, this study deals only with mystery works. The whole dataset adds up to 4,183,485 words.

When we overview these data using a correspondence analysis using the frequency of the 200 most common words from the dataset, we can see most of Christie's earlier works on the left side of the plot, and later works on the right (see Figure 1).

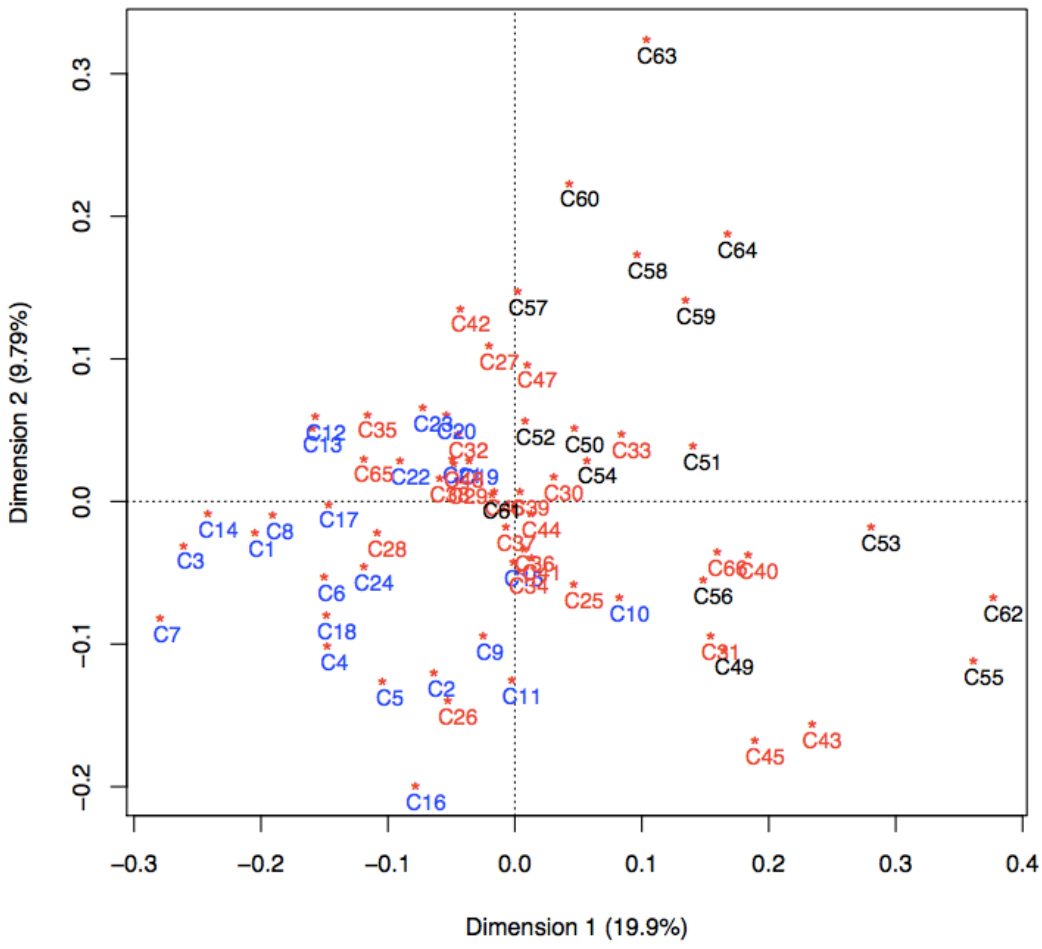


Figure 1 Plot of the result of correspondence analysis.

Figure 2 is the resulting dendrogram of a cluster analysis on the same data as Figure 1. We can see that most of the later works are classified into a cluster distinct from the earlier works.

[1] Osaka University

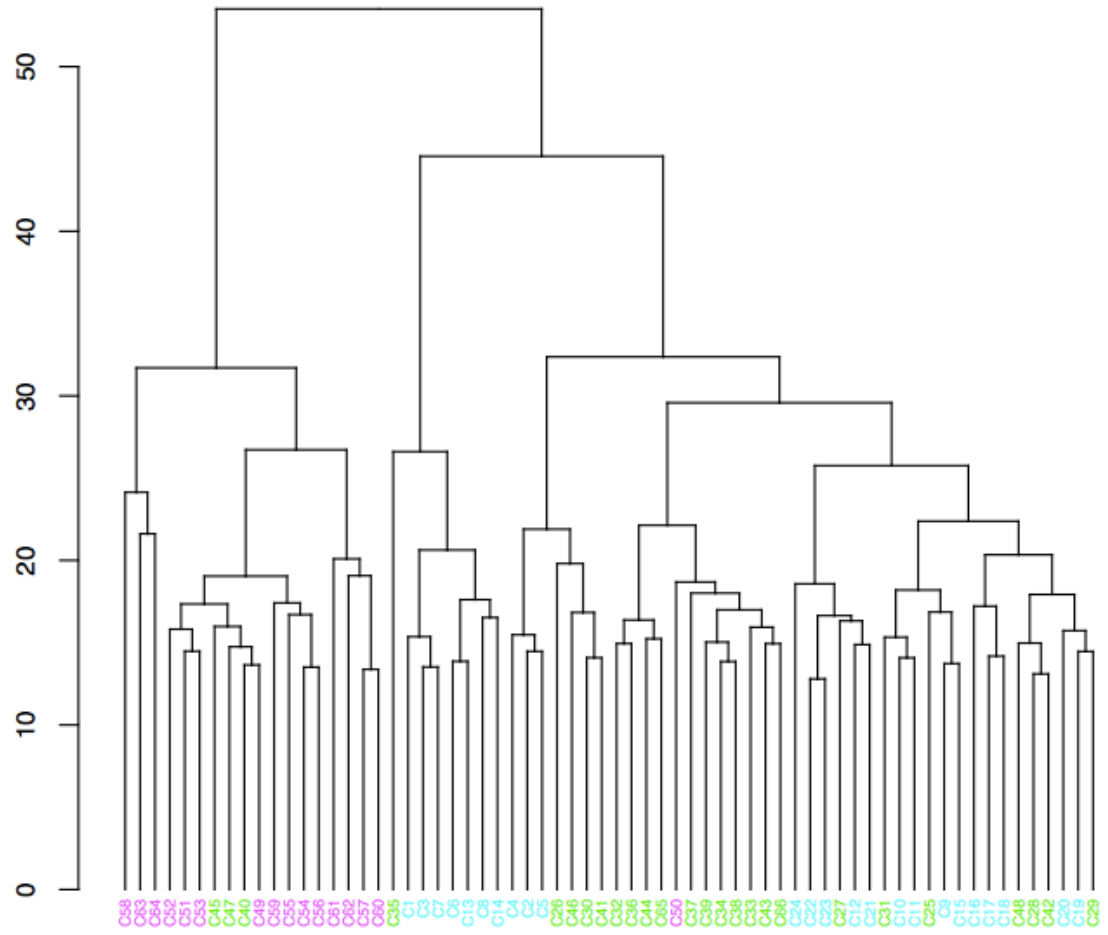


Figure 2 Resulting dendrogram of cluster analysis (distance: Euclidean, method: Ward)

Then questions arose as to whether we can distinguish earlier works from later ones using a machine learning method and extract key words for each group.

This study applies Random Forests (Breiman, 2001) for classification and extraction of key words. Tabata (2012) argues Random Forests overcome common problems in key word measures such as Log Likelihood or Chi-squared score, making them an attractive alternative. In this study, all of Christie's works are divided into three groups according to their year of publication; earlier (1920–1938: c1–c24), middle (1939–1956: c25–c48, c65, c66), and later (1957–1973: c49–c64). The three-way split was chosen after considering that classifying into two groups might be too coarse. Exceptions are c65 and c66; as mentioned above, they were published in the 1970s but written in the 1940s, so they are added to the middle group. The variables used in Random Forests are the most frequent 600 words, which these texts were classified the most accurately with. An example of the result of Random Forests is shown below (Figure 3 and 4). As we can see in Figure 3, Christie's earlier works are never classified as her later works, and vice versa.

```
Call:
 randomForest(formula = text.group ~ ., data = tbl, proximity = T,      importance = T, ntree = 10000)
               Type of random forest: classification
                     Number of trees: 10000
No. of variables tried at each split: 24

        OOB estimate of  error rate: 15.15%
Confusion matrix:
          1_earlier 2_middle 3_later class.error
1_earlier        21        3       0   0.1250000
2_middle          1       23       2   0.1153846
3_later           0        4      12   0.2500000
```

Figure 3 Result of Random Forests classification experiment on the three groups using the most frequent 600 words.

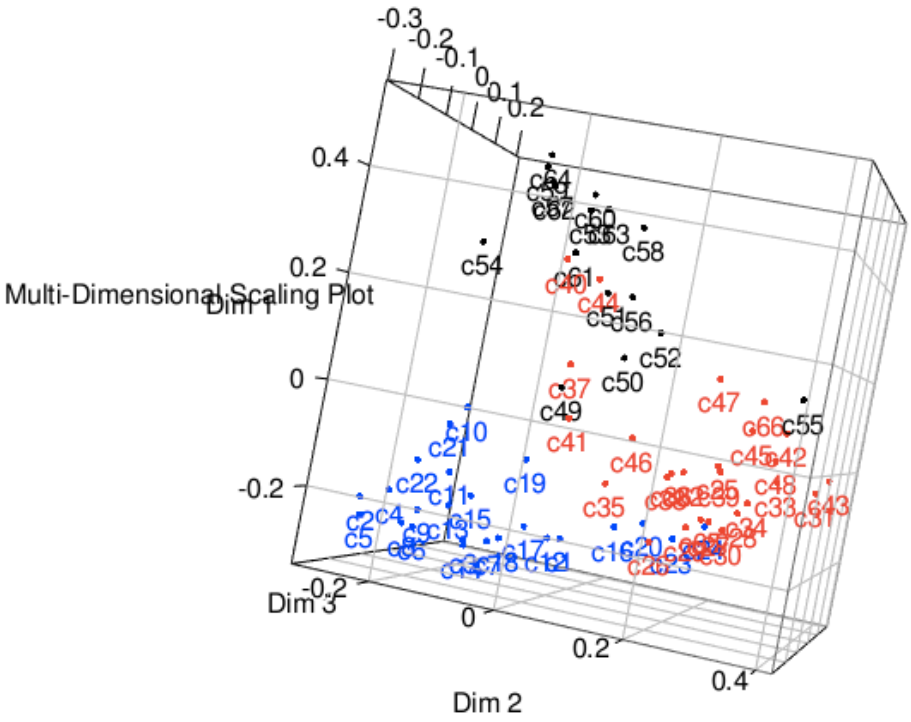


Figure 4 A multi-dimensional scaling plot based on the result of the Random Forests model.

Out of these top 600 words, the 100 most characteristic words are identified. Figure 5 shows the top 30 words which most contribute to the classification based on the mean decrease in the GINI importance, and Table 1 shows the 100 most characteristic words. We can see the word *something*, the fourth from the top of the figure, and we can see the word *things* in the table. Lancashire and Hirst (2009) investigates indefinite-term (*thing, something, anything*) usage in 14 Christie works, reporting that Christie's use of these words increases significantly with age, and they suggest that indefinite-term usage is a significant marker for Alzheimer's disease. The word *something* and *things* are also extracted as key words for her later works using Random Forests, and this result supports their argument.

In addition, there are many contracted forms such as *you've*, *wouldn't*, *I'd*, *she'd*, and *don't* as key words from the later works. This result might be owing to Christie's change of writing method in her novels. Le et al. (2011) says,

> "She wrote her earlier novels in longhand and then typed them on a typewriter …, but, on breaking her wrist in 1952, she began using a Dictaphone."

The result of this analysis also reflect her change of writing method.

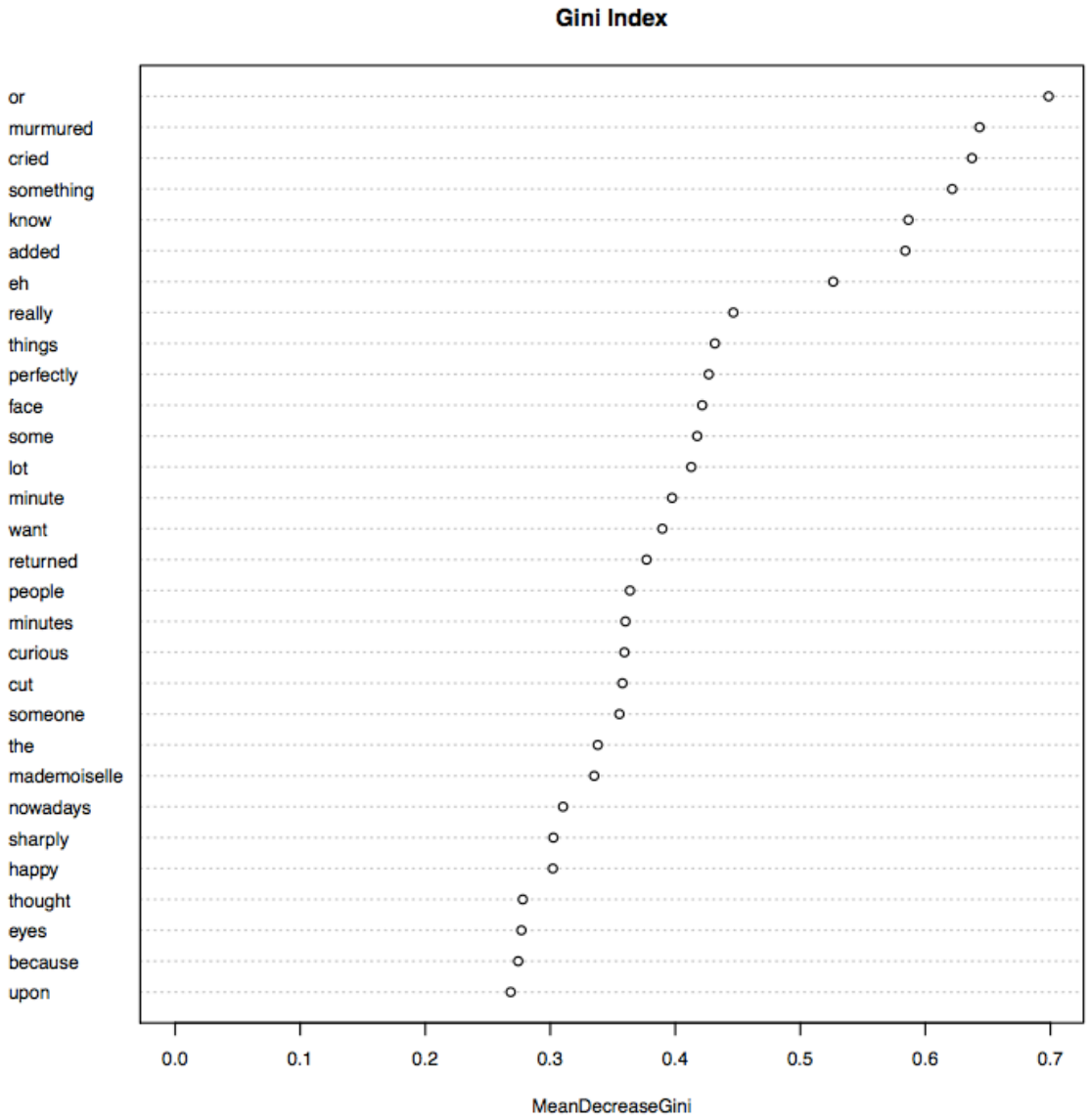


Figure 5 Variable Importance Plot of the result of Random Forests.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45: pp.5-23.

Inaki, A. (2013). Nazotoki no Kotoba-gaku: Agatha Christie no Eigo wo Tanoshimu. Tokyo: Eiho-sha.

Lancasire, I. and Hirst, G. (2009). *Vocabulary Changes in Agatha Christie's Mysteries as an Indication of Dementia: A Case Study*. Paper presented at the 19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice, Toronto, March 2009.

Le, X., Lancashire, I., Hirst, G. and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Literary and Linguistic Computing*, 26(4): 435-461.

Tabata, T. (2012). Stylometry of co-authorship: Charles Dickens and Wilkie Collins. *The Special Interest Group Technical Reports of Information Processing Society of Japan*, CH-93(3): pp.1-7.

Interpreting Visual Data in the Platformized Context: The Case of a Chinese Working-class Online Community

Jiaxi Hou[1]

Kwai, the video-clip sharing application is among the most popular mobile applications in China with over 700 million users, and a large proportion of them are consisted of working-class youths1. Situated within the rapid developments of mobile technologies, Kwai becomes the first video-based online community commonly accepted and shared by Chinese working-class youths, either living in the rural areas or working in metropolis as migrant workers. Being as the application and the community at the same time, Kwai has not only supported the vibrant visual expressions of the previously silent Chinese working-class youths, but has also served as a large data archive of these user-generated visual representations. Users utilize Kwai not only to record their daily living and working activities in physical world, but also share their originative and creative artifacts in the form of short videos, which have not yet gained sufficient investigation.

However, at the same time of "neutrally" archiving the visual data to for the users with technological affordances, Kwai is always criticized by the public for the level of vulgarity of these controversial user-generated content. There existed a large amount of video clips containing sexual connotations in offensive ways, or even to the level of child pornography and juveniles displaying self-torturing behaviors in the pursuit for fame and popularity. In addition, a large number of discussions are also triggered by Kwai about the "appropriate" norms for distributing algorithm, interactive behaviors, and creative artifacts of a participatory online community in the highly visual format. Therefore, the study aims to not only understand the cultural activities of Chinese working-class youths through the archived data on Kwai, but also endeavors to interpret them in a larger social and cultural context, in order to anatomize how the meanings of certain data in the visual form are constructed in the context.

Notably, the context here is a social and cultural specific environment deeply embedded in the process of platformization by digital technology. Previous scholars have found that abreast with the frequent use of the term, platform, in the commercial companies such as Facebook and YouTube for describing their systems, multifaceted aspects have been incorporated into the meaning of the term to constitute it as a useful paradigm to understand the current ecology of social media, for example, Burgess's (2015) investigation of YouTube's platformization process and Helmond's (2015) in Facebook. The assemblage of the Kwai's culture has incorporated different actors and developed its particular characteristics in organizing online participatory culture. In line with the platform perspective, the current study tries to interpret the user-generated videos, as a form of data, in the platformized context, by taking into consideration of multiple actors including the technological affordances of the application, its economic strategies, the state intervention, the public evaluation on the artifacts, as well as the user agency.

A combined method of online ethnography, data crawling, and discourse analysis will be utilized to conduct this analysis. First, a continuous participating observation of the Kwai application is needed, with the particular attention for the dynamic changes in its interface and algorithm design. Second, related data of the popular video clips are crawled to identify the representative artifacts of the community and then over fifty accounts and their archived video content are collected to understand and conclude the changing characteristics of its narratives, as the representative cultural artifacts of Kwai. Third, related articles addressing the platform are achieved through search engines. The origins of these published articles, the opinions, the structures, and how the readers express their opinions in the commenting area of these articles will be analyzed. These data from various sources will be utilized together to understand what are the cultural meanings of these video clips, as the form of archived data in the larger social context,

[1] The University of Tokyo

which is characterized by the contemporary platformizing process in China and also in the global world.

The meanings of Kwai's data in the form of videos have experienced four different phases in its transformation, which is structured complicatedly by technological affordance, users agency, the state power and the mobilized public. First, the original Kwai is positively embraced as an alternative social media for the working-class youths with an "objective" and "value-free" algorithm in "recording the world and you" (the slogan of Kwai). However, the implicit emphasis of social media in the pursuit for fame, together with users' agency, has paved the way for vulgar content including child pornography, which evoked severe criticism from middle-class and elite media, who have more power in defining the appropriate norms and ethics for online interaction compared to the working-class youths, the new group of netizens. At the same time of experiencing a quick user expansion from 400 million to 700 million, the platform responded to these criticisms with the refined versions in algorithm design, the additional reporting function, and the establishment of self-disciplinary committee. Though these actions have largely prevented the popularity of child pornography, but have not changed the images of the platform for being the headquarters of producing vulgar and kitsch visual content. Subsequently, the state power, accompanied by the anonymous public, infiltrated into the process in the name of protecting the juveniles from vulgarity, but with more attention in censorship and information control. As a consequence, the previous vulgar videos spontaneously produced by working-class have been purified in their meanings, and at the same time, the new group of netizens are socialized in the larger contexts to alter their self-representations with a "healthier" and more "appropriate" set of norms.

Notes

The data can be accessed from the Chinese news media, https://www.jiemian.com/article/1915543.html, on 20 May 2018.

References

Burgess, J. (2015). "From 'Broadcast Yourself' to 'Follow Your Interests': Making over Social Media." *International Journal of Cultural Studies*, 18(3): 281–85.

Helmond, A. (2015). "The Platformization of the Web: Making Web Data Platform Ready." *Social Media + Society*, 1(2): 10.1177/2056305115603080.

Cancelled

[Panel Session 1]

Digital Humanities Cyberinfrastructure: Integrating and Facilitating

Jieh Hsiang[1], Joey Hung[2], Chao-Lin Liu[3], Michael Stanley-Baker[4]

The development of digital humanities originates from the large-scale digitalization of scholarly or cultural repositories. As large quantities of digital resources become available online, humanities scholars start to have high expectations for and imagination of the applications of digital technologies to facilitate their research. Many digital tools have also been developed in response to this demand. However, For the humanists, there still exists a huge gap between the applicability of digital resources and digital tools, and the policies and technologies to implement them.

In recent years, many digital platforms that are customized to adapt to the individual needs of humanities research have been developed. These platforms, in additional to offering heterogeneous services, should also be interlinked with other digital resources as well as allowing users the freedom and autonomy through their availability, usability and constant optimized capability.

Based on this panoramic perspective of the digital humanities cyberinfrastructure, this session discusses the current digital resources, digital tools and the integration and facilitation of open data on different digital platforms. The discussion topics will include issues of philosophy, concepts and technicality of establishing digital platforms. At the same time, we will also explore the expectations for and imaginations of such a cyberinfrastructure from the humanities researchers' perspectives. The purpose of this session aims to promote a better realization of digital humanities cyberinfrastructure.

Session Speakers Information

Speaker1: Prof. Joey Hung (洪振洲)

Topic: The Experience of using Open Data and Open API to Facilitate the Cross-System Collaboration of Digital Humanities Tools

Affiliation: Department of Buddhist Studies; Library and Information Center, Dharma Drum Institute of Liberal Arts, Taiwan.

Speaker2: Prof. Chao-Lin Liu (劉昭麟)

Topic: Service Agents for Accommodating Public and Private Data and Tool Resources for Researchers of Digital Humanities

Affiliation: Department of Computer Science, National Chengchi University, Taiwan.

Speaker3: Prof. Jieh Hsiang (項潔)

Topic: Connecting Open Texts and Tools through DoucSky

Affiliation: Research Center for Digital Humanities, National Taiwan University, Taiwan.

Speaker4: Prof. Michael Stanley-Baker (徐源)

Topic: Datamining Religious Sources for Medical History: Locating Materia Medica in the Daoist and Buddhist Canons

Affiliation: School of Humanities; College of Humanities, Arts, & Social Sciences, Nanyang Technological University, Singapore.

---

[1] National Taiwan University

[2] Dharma Drum Institute of Liberal Arts

[3] National Chengchi University

[4] Nanyang Technological University

“Cicerone”, a monuments’ guide plug-in for navigators: a proposal for a history- related software application to increase the value of cultural heritage historically with GIS and GPS open data.

Luigi Serra[1]

Abstract
Are satellite navigators complete to satisfy every need? This is a proposal for a software enhancement to live and visit monuments worldwide with satellite navigation systems in a different way. With their POI (Points Of Interest), navigators usually generate itineraries computing them essentially on GIS basis, instead of historical facts or themes. The advantage of this improvement I propose is to determine routes basing them on particular periods of interest, planning voyages with the powerful efficacy of GPS navigators, but keeping in mind a specific historic epoch and its related monuments with the help of both GIS and historical open data. As soon as this suggestion were welcomed by the market, it could be embedded into satellite navigation systems like Garmin™, Tom Tom™ or similar, or web based like Google Maps™ and others.

1. Introduction

The present paper aims at introducing a little new feature to be added into general-purpose navigators. In fact I noticed that for precise interrogations focused on specific interests, the current solutions we use every day have a little lack: we cannot chose particular periods of interest while adding POI or destinations to our journey. Those are variables not necessary for most users, but very useful for scholars and for the most demanding tourists.

2. History into navigators: an added value

A monument without a description and a narration that tells its history, its origins and why it is there in that context, could be seen as only a mass of matter without meanings. This is the role of scholars: analyze monuments and tell something authoritative about them helping us to make ourselves our own idea about them, beyond their pure aesthetic expression.

Standardizing the right criteria for the categorization and cataloguing of the monuments according to the scholars’ advices, it will help us in creating a useful open data container from which everyone can take the desired information. Engineers can solve technical problems related to the best practices, historians and geographers could benefit these tools adding other meanings to the tools themselves.

3. Navigator itineraries time-related by themes and periods: a possible scenario

The idea of the time-relation concept has born one day, when a Valencian friend, teacher at University, asked me to suggest him a travel plan in Sardinia (Italy) including some important monuments along the road belonging to a specific period. Trying to plan such kind of trip, I have experienced that it is possible to include some POI already present into navigators, but if I had wanted to select specific monuments related by a particular age or theme, I needed to study in deep the monumental panorama by hand, on the maps, because of the lack of such kind of information on the navigation systems. My idea would go beyond, changing the perspective: while a trip is almost ever place-based, I would like to plan a monument-centric trip, in which the routes comes up from a history-related monuments’ choice. E.g. if we would visit medieval castles in Sardinia, joined by their belonging to the Kingdom of Arboréa along its meridional border, we should know their building period, their geographical information, the certain belonging to that kingdom and so on. We should know ultimately, geography apart, their history. The castles fitting this

[1] National Research Council of Italy / Institute of History of Mediterranean Europe

information are six clockwise and their information and georeferenced positions are in turn (Place, Name, Latitude, Longitude, Building Period, Notes):

Laconi, Aymerich Castle, 39°51'19.24"N, 9° 3'18.80"E, 1053, Epigraph;
Las Plassas, Marmilla Castle, 39°40'57.63"N, 8°58'46.26"E, <1168, Partial ruins; Sanluri, Eleonora of Arboréa Castle, 39°33'47.72"N, 8°53'52.58"E, 1355, Certain period; Sardara, Monreale Castle, 39°35'41.78"N, 8°47'35.28"E, ~ 1275, Historical Sources; Arbus, Arcuentu Castle, 39°35'50.54"N, 8°32'48.04"E, <1168, Ruins;
Ales, Barumele Castle, 39°45'22.14"N, 8°48'45.31"E, 1385, Pre-existent to Arboréa.

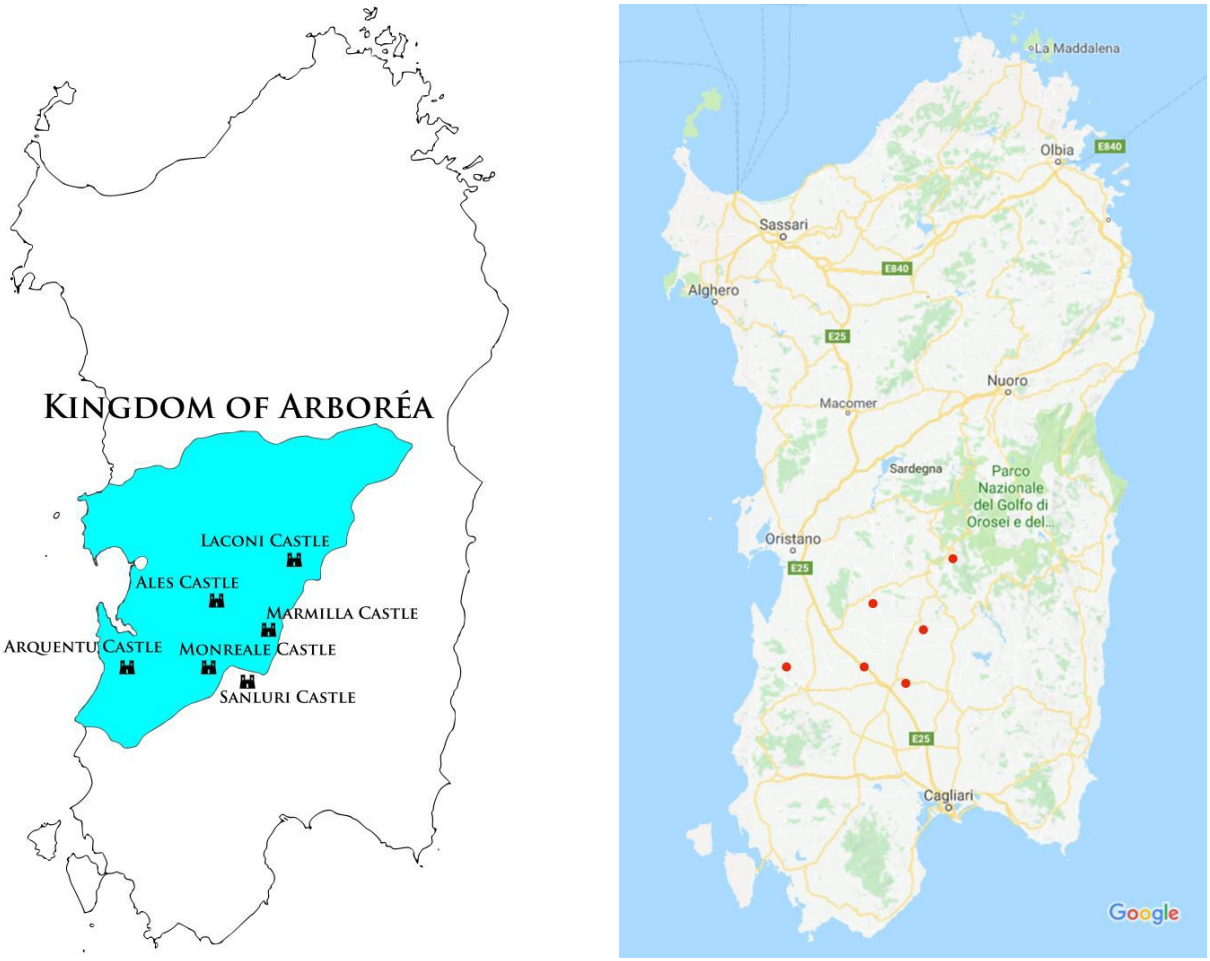


Fig. 1 – Past and present geographical position of the castles in the southern border of the Kingdom of Arboréa (Luigi Serra)

Looking at the shortest path using Google Maps, leaving from Cagliari and returning to Cagliari, the right trip order, counterclockwise, should be:

- Cagliari, 39°13'10.82"N, 9° 7'48.18"E (Departure)
- Marmilla Castle, 39°40'57.63"N, 8°58'46.26"E
- Aymerich Castle, 39°51'19.24"N, 9° 3'18.80"E
- Barumele Castle, 39°45'22.14"N, 8°48'45.31"E
- Monreale Castle, 39°35'41.78"N, 8°47'35.28"E
- Arcuentu Castle, 39°35'50.54"N, 8°32'48.04"E
- Eleonora Castle, 39°33'47.72"N, 8°53'52.58"E
- Cagliari, 39°13'10.82"N, 9° 7'48.18"E (Arrival)

At first glance, the software proposes additional passage points by simply concatenating the map locations (https://goo.gl/maps/o9wLNkKvjYP2). This behavior is evident because of the human input, but if the information were been already present in the system and correlated each other, the route would derive automatically based on the criteria chosen for the tourist tour. The aggregation could be possible only if an open database containing all the historical data, georeferenced and aggregated by period,

affinity and document evidences, was available, thus having time-related information to use linked with POIs.

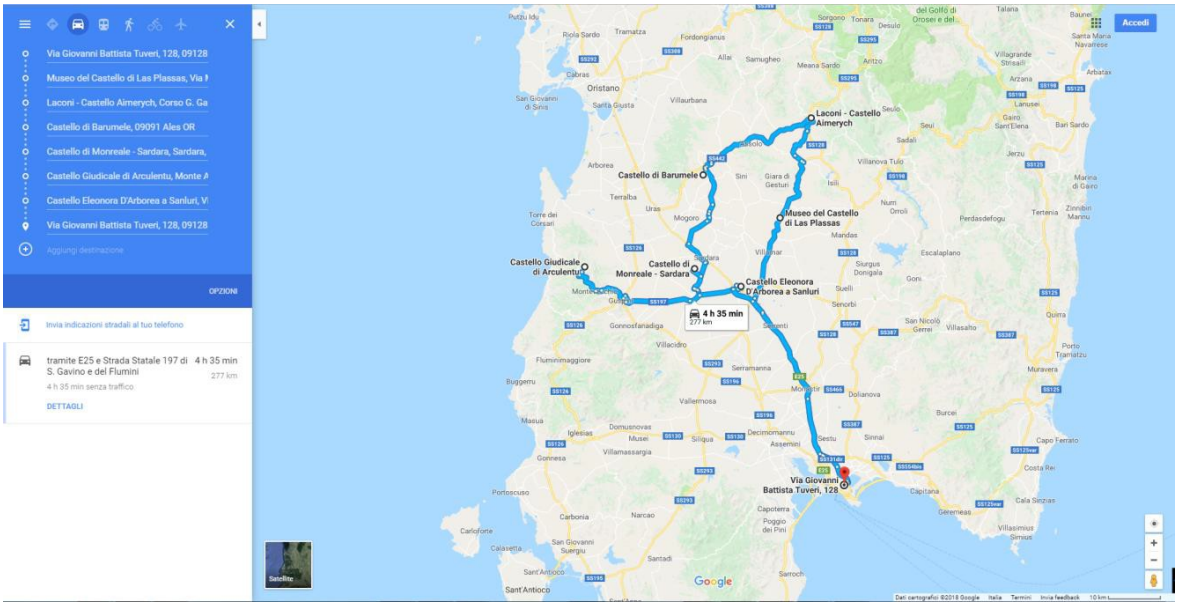


Fig. 2 – Simulation path, based on historical information of Sardinian Castle focused on their belonging to the southern border of the Arboréa's Kingdom on Google Maps (Luigi Serra)

4. Conclusions

Such plug-in could be very useful to scholars who are investigating about certain periods. In fact, it could aggregate different types of monuments related to a particular studied age based furthermore on institutions and statehood in their entirety: castles, fortresses, fortifications, religious buildings and any other kind of monument belonging to a definite State, place and time. The system could automatically propose a trip plan, matching the best routes depending on the roads or pathways present, the days available and other specific needs. It is a refining of the general-purpose POI suggestion mechanism today available in every GPS navigator. This simple proposal would be an opportunity for companies producing satellite navigation systems, to involve competent professionals in History, Geography, Archaeology, Architecture, Engineering, Computer Science and other scholars for a synergic multidisciplinary collaboration. Thus giving to the audience a useful tool to visit the world, starting from the landscapes' history, knowing not the only "where", to which satellite navigators are interested on, but the "why", the "who" and the "when" as the red thread that links the "all" through the places.

References

Allen, P. (1992). *Storia della Cartografia.* London: Marshall Editions.

Carta, M. and Spagnoli, L. (eds) (2010). *La ricerca e le istituzioni tra interpretazione e valorizzazione della documentazione cartografica*. Roma: Gangemi Editore.

Casula, F.C. (1997). *La terza via della storia: il caso Italia*. Pisa: ETS.

Dijkstra, E.W. (1959). "A note on two problem in connexion with graphs." *Numerische Mathematik*, 1:269–271.

http://galileognss.eu/ (accessed 23 April 2018)

http://qzss.go.jp/en/ (accessed 23 April 2018)

https://www.archives.gov/ (accessed 23 April 2018)

https://www.glonass-iac.ru/en (accessed 23 April 2018)

https://www.gps.gov/ (accessed 23 April 2018)

https://www.isro.gov.in/ (accessed 23 April 2018)

Harley, J.B. and Woodward, D. (eds) (1987). *The history of cartography, Vol.1 Cartography in Prehistoric, Ancient, and Medieval Europe and the Mediterranean.* Chicago and London: The University of Chicago Press.

Milanesi, M. (ed) (1990). *L'Europa delle carte.* Milano: Nuove Edizioni Grafiche Mazzotta.

Palagiano, C., Asole, A. and Arena, G. (1984). *Cartografia e territorio nei secoli.* Roma: La Nuova Italia Scientifica.

Quddus, M.A., Ochieng, W.Y. and Noland, R.B. (2007). "Current map-matching algorithms for transport applications: State-of-the art and future research directions." *Transportation Research Part C.* Elsevier Ltd, pp. 312–328.

Serra, L. (2017). "Relational and conceptual models to study the Mediterranean defensive networks: an experimental open database for content management systems." *International Conference on Modern Age Fortifications of the Mediterranean Coast FORTMED 2017: Conference Proceedings Vol. VI.* Alicante: Editorial Publicacions Universitat D'alacant, pp. 369- 376.

Serreli, G. (2015). "Il sistema difensivo del Regno di Arborèa tra il X e il XV secolo." *International Conference on Modern Age Fortifications of the Mediterranean Coast FORTMED 2016: Conference Proceedings Vol. IV.* Firenze: DIDAPRESS, pp. 433-440.

Steven, J.D. et al. (2015). *Historical Studies in the Societal Impact of Spaceflight.* Washington DC: NASA History Program Office.

Barber, P. and Istituto Geografico De Agostini (eds) (2001). *Segni e sogni della terra: il disegno del mondo dal mito di Atlante alla geografia delle reti.* Novara: De Agostini.

Tomlin, C.D. (1994). "Map Algebra: one perspective." *Landscape and Urban Planning 30.* Elsevier Ltd, pp. 3-12.

Tomlinson, R. F. (1967). "An Introduction to the Geo-Information System of the Canada Land Inventory." Ottawa: Canada Department of Forestry and Rural Development.

Why do I need four search engines?

Martin Holmes[1], Joseph Takeda[2]

Abstract

This presentation addresses the question of how to create digital editions and other online resources in forms that are likely to endure and remain functional over many decades. In particular, we focus on the dichotomy whereby the long-term robustness of a digital project, which is achieved by uncoupling it from transient server-side technologies and tools that require monitoring and maintenance, is undermined by the requirement to provide methods for users to search the collection. We present as a case-study the *Robert Graves Diary* project, which provides four separate search facilities using different approaches.

1. Introduction

Project Endings[1] is a collaboration of University of Victoria scholars, digital humanists and librarians whose aim is to address the progressive loss of digital scholarly resources due to failures in archiving, preservation, and documentation, and over-dependence on transient tools and technologies. The project is supported by a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC).

The project is working with a number of case-studies—digital edition projects already completed or nearing completion—and aiming to specify approaches, tools and technologies that can help researchers complete their projects and archive them in such a way that they have a strong chance of being available and functional for decades to come.

In previous work (Arneil and Holmes 2017, Holmes 2017, Holmes and Takeda 2017), we have argued strongly that likelihood that a digital edition project will survive and be usable over the long term depends on the selection of a small core set of technologies (HTML5, CSS and JavaScript), and the avoidance of server-side technologies that will require maintenance or replacement over time. Our case-study projects (among them *Le Mariage sous L'Ancien Régime and The Robert Graves Diary*) are constructed entirely in this way, with no server-side dependencies at all.

2. The problem of search

A digital edition consisting only of HTML, CSS and JavaScript can of course by rich and highly interactive. However, there is one important component of a website which generally requires some sort of server interaction: search. This is perhaps the most difficult challenge for the Endings project: how do we make a resource searchable without building in dependence on a server to host the index and respond to queries?

[1] University of Victoria
[2] University of British Columbia

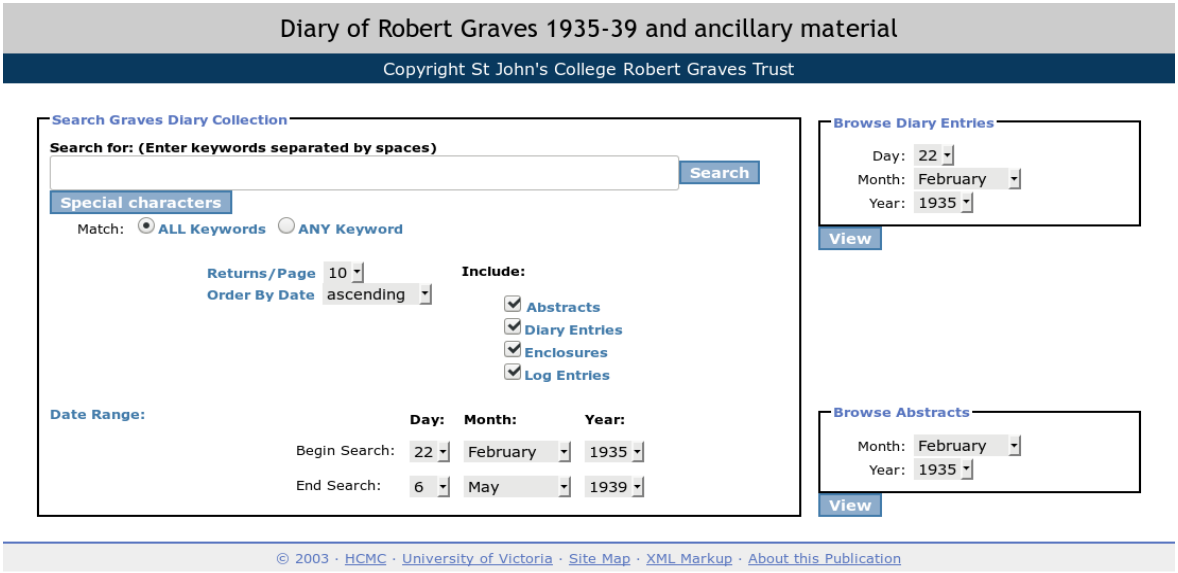


Figure 1. The first search interface of the Robert Graves Diary, which also serves as its home page.

Using the *Robert Graves Diary* project as a testbed, we have developed four distinct approaches, which will be described in this presentation:

1. Bite the bullet and accept the server dependency. We currently host the Graves project materials inside an eXist XML database, which allows us to provide rich faceted search functionality at the expense of a dependency that will inevitably be unsupported in the long term.
2. Enlist Google's help. We have built an additional Google Custom Search page into the site, allowing users to search in the interface which is probably most familiar to them. The obvious drawback here is that Google's terms, conditions and APIs change frequently, so we must expect this service to fail at some point when there is no active maintainer of the project.
3. Enlist the help of our Library. The long-term preservation of our project will ultimately be in the hands of the University Library, who run their own Solr server for searching their collections. As part of the project build, we now create JSON index files for Solr to ingest; we can then provide a search page which queries this index.
4. Provide a standalone search. For digital editions which are not too large, it is possible to create a JavaScript-only index, including stemming and relevance scoring, which is remarkably fast and requires no server support at all. This is the ultimate fallback when all else fails.

On the face of it, this level of redundancy may appear ridiculous, but in fact it provides a level of flexibility which we believe is essential for the survival of projects with no ongoing maintenance. In the best-case scenario, four different methods of searching the collection are available to the user, each with their own strengths and weaknesses. In the worst case, where the collection survives only as a simple collection of files on a drive somewhere, the standalone search will still work, while the other non-functional search interfaces provide evidence of the aspects of the collection thought to be crucial search facets.

Keywords: *archiving, digital editions, preservation*

Bibliography

Arneil, Stewart and Martin Holmes. 2017. "Archiving form and function: preserving a 2003 digital project." DPASSH Conference 2017: Digital Preservation for Social Sciences and Humanities, Brighton, UK, 14th June 2018.

Holmes, Martin. 2017. “Selecting Technologies for Long-Term Survival.” SHARP Conference 2017: Technologies of the Book, Victoria, BC, Canada, 10th June 2017. https://github.com/projectEndings/Endings/raw/master/presentations/SHARP_2017/mdh_sharp_2017.pdf.

Holmes, Martin and Joseph Takeda. 2017. “Beyond Validation: Using Programmed Diagnostics to Learn About, Monitor, and Successfully Complete Your DH Project.” Digital Humanities 2017 Conference, Montreal, Canada, 1th August 2017. https://dh2017.adho.org/abstracts/140/140.pdf.

[1] https://github.com/projectEndings, https://onlineacademiccommunity.uvic.ca/endingsproject/

Converting the Aozora Bunko into a corpus suitable for linguistic research

Bor Hodošček[1]

The Aozora Bunko project is a volunteer-driven Japanese digital library containing over 14,000 out-of-copyright and copyright-free works written by over 1000 different authors (Aozora Bunko, 2018). Predominantly consisting of long and short works of fiction and non-fiction, the collection is an invaluable language resource covering Japanese works of diverse genres from the late 19th to the middle of the 20th century—a timespan uniquely situated between existing corpora from the Meiji era, such as the Corpus of Historical Japanese (National Institute for Japanese Language and Linguistics, 2017), which contains magazines from the Meiji and Taishō eras, and the Balanced Corpus of Contemporary Written Japanese (Maekawa et al., 2013), which includes samples of various genres from 1976–2006. While the Aozora Bunko project began publishing works on its website from 1997, it's data is now also hosted in a version-controlled repository on GitHub, which makes it possible to programmatically subscribe to daily changes and refer to any past published data. The present work aims to provide continuously updated plaintext and TEI versions of works in the Aozora Bunko as well as a reworking of the available metadata that transform the Aozora Bunko from a reader-focused resource into a free, versioned, and accessible resource for reproducible linguistic study, with an initial focus on issues relevant for stylometric analysis and adoption by the wider non-Japanese speaking research community.

While the Aozora Bunko has been a focus of some research on its bibliographic metadata design (Chiba et al., 2006; Ochiai, 2013), discussion of its conversion into a format necessary for linguistic investigation, typically requiring main body text extraction and morphological analysis, is lacking and often relegated to undocumented ad-hoc extraction scripts that vary by study (Kanagawa and Okadone, 2017; Mochizuki and Suzuki, 2007). This preprocessing is complicated by Aozora Bunko's unique transcriber-friendly master text format, which is not formally defined and does not focus on linguistic structural metadata description, as well as several technical trade-offs, such as choice of SJIS as default encoding and recording of characters from the 3rd and 4th planes of the JIS X 0213 standard as textual descriptions and image files. The Himawari fulltext search Java application project offers the most complete handling of the above issues by providing a morphologically preprocessed and annually updated corpus of the Aozora Bunko as a set of import files for Himawari (Yamaguchi and Tanaka, 2005). While offering a shortcut to linguistic analysis, it omits versions of some works published using older writing styles, and does not consolidate Aozora Bunko's varying structural text metadata to clearly differentiate between main body and other text (preface, postscript, bibliography, etc.). Considering the aforementioned issues as common stumbling blocks impeding linguistic analysis, and allowing for the widest use of the resource as possible, the present system provides two versions for each of the TEI and plaintext formats, both with and without morphological parse information.

In order to map the Aozora Bunko master format into structured, UTF-8 encoded XML, the system uses a formal specification based on a contract system (Might et al., 2011) to parse regular metadata such as *furigana* (ruby), as well as more difficult natural language-style metadata such as textual descriptions of Chinese characters not in SJIS. These existing metadata tags from Aozora Bunko are complemented with additional metadata tags automatically extracted by the system, which currently include sentence boundaries and direct speech.

The use of Japanese text in linguistic analyses that make use of off-the-shelf and generic natural language processing techniques commonly requires the text to be split into tokens, which, in the case of Japanese, requires an additional morphological analysis

[1] Osaka University

preprocessing step when compared to whitespace-delimited languages such as English. Morphological preprocessing is conducted using MeCab (Kudo et al., 2004) and the Contemporary Written Japanese (CWJ) or Kindai variants of the morphological analysis dictionary UniDic (Oka, 2017), chosen based on the age and writing style of each work. Here the system also takes into account an issue common to earlier modern Japanese works that would otherwise cause morphological parsing difficulties, which is the intermittent use of *kanji katakana majiri bun* writing, wherein Chinese characters (kanji) are used in combination with katakana instead of hiragana. Accordingly, versions of the corpora containing morphological information can take into account these tags and provide correct readings from *furigana* ruby tags and fixed parses from *kanji katakana majiri bun* sentences. Of course, use of any tag is optional, as research requiring custom processing can choose to ignore any of the tags offered and use a different morphological or other type of parser. Finally, the naming convention for filenames is based on the one used by the popular stylometry package for R, Stylo (Eder et al., 2016), allowing the use of tokenized plaintext files in common analysis pipelines as-is.

Following the work of the Aozora Bunko Linked Open Data project by Ochiai (2013), the system provides updated conversion facilities for the metadata contained in Aozora Bunko's bibliography CSV. Similarly, integration is provided for relevant metadata, including author literary movement affiliations, work subject matter, and date of first publishing, that is contained within the Japanese National Diet Library's Web NDL Authorities and DBpedia LODs (Lehmann et al., 2015), the latter of which is extracted from the Japanese version of Wikipedia using the project's Information Extraction Framework. Based on this representation, a SPARQL endpoint and limited search functionalities are provided for access to relevant information in the formative stages of linguistic analysis, as well as providing for easier multi-faceted comparisons of extracted datasets.

In conclusion, the system provides preprocessed plaintext and TEI versions of the Aozora Bunko with updated metadata relevant for linguistic analysis. These corpus versions and their associated metadata are backed by a version-controlled repository that allows for the stable referencing and downloading of specific versions of the corpora at any time independent of changes occurring within the Aozora Bunko, as well as the extraction algorithms and metadata schema of the system.

References

Aozora Bunko (2018). Aozora Bunko https://www.aozora.gr.jp/ (accessed 5 June 2018).

Chiba, S., Iseki, S. and Chen, C. (2006). Aozora Bunko o gengo koopasu tosite tukaou: meetadeeta koutiku ni yoru rekisiteki/syakaigengogakutekikenkyuu e no kokoromi [Aozora Bunko as Corpus: Application of metadata construction for historic and sociolinguistic study] [in Japanese]. In *Proceedings of the 12th Annual Meeting of the Association for Natural Language Processing.* The Association for Natural Language Processing.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. *R Journal,* 8(1).

Kanagawa, E. and Okadone, T. (2017). Characterization and Similarity Analysis of Japanese Writers' Syntactic Structures by Kernel Method. *Transactions of the Japanese Society for Artificial Intelligence,* 32(3, F–G94_1). The Japanese Society for Artificial Intelligence: 1–15 doi:10.1527/tjsai.F-G94.

Kudo, T., Yamamoto, K. and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., et al. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2). IOS Press: 167–95.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. and Den, Y. (2013). Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*. Springer Netherlands: 1–27 doi:10.1007/s10579-013-9261-0.

Might, M., Darais, D. and Spiewak, D. (2011). Parsing with derivatives: A functional pearl. *In ACM SIGPLAN Notices*, vol. 46. (9). ACM, pp. 189–95.

Mochizuki, T. and Suzuki, Y. (2007). A trial of the writing style impression analysis in the novel. *IPSJ SIG Notes*(128(2007-MPS-067)): 179–82.

National Institute for Japanese Language and Linguistics (2017). Corpus of Historical Japanese, Meiji Era / Taishō Era Series I: Magazines http://pj.ninjal.ac.jp/corpus_center/chj/meiji_taisho.html (accessed 5 June 2018).

Ochiai, K. (2013). Aozora Bunko Linked Open Data http://mdlab.slis.tsukuba.ac.jp/lodc2012/aozoralod/index.html (accessed 5 June 2018).

Oka, T. (2017). UniDic for Morphological Analysis with reduced model size by review of CRF feature templates. In *Proceedings of the 2017 NINJAL Language Resources Workshop.* pp. 143–52.

Yamaguchi, M. and Tanaka, M. (2005). Design and implementation of full text search system for structured language resources [in Japanese]. *Journal of Natural Language Processing,* 12(4): 55–77 doi:10.5715/jnlp.12.4_55.

Methods of Meaning: Deciphering the History of "Literature" With Two Word Vector Approaches

Mark Algee-Hewitt[1], Alexandre Gefen[2], Eun Seo Jo[1], J.D. Porter[1], Marianne Reboul[3]

I. Introduction

We compare two methods for measuring the relationships between words as they are used in literary critical corpora. One of our fundamental premises is that a concept and its history can be discovered and understood on the basis of statistical relationships between words in written texts. Here, we show how two different ways of measuring these relationships, cosine similarity and K-means clustering, produce different interpretive results.

We expect our findings to be useful on three levels: First, they build on past work to show how the concept of literature has changed in French and British critical contexts from the eighteenth century to the twentieth. Second, the results illustrate the differences between cosine similarity and K-means clustering as ways to study word relationships. Finally, by comparing these two methods, we interrogate the notion of relational meaning that underlies much of digital textual analysis and literary criticism more broadly.

II. Corpora

Since the goal of the project is to investigate literature as an intellectual concept, we have focused on its presence in critical works. As a comparative project, we have corpora in two languages:

Language	**Source**	**Years**	**Word Count**
French	Academic articles	1700-1960	~150 Million
English	*British Periodicals Online*	1700-1960	~10 Billion

Tab.1: Overview of the two corpora for the experiments

Although these corpora vary with respect to their size and original audiences, they are similar in terms of their historical spans and in that they are nonfiction article-length prose. Past investigations have generated promising results. Recently, we created vectors using Word2Vec and analyzed changes in cosine similarity over time between "literature"/"littérature" and the other words in the corpora. The results indicated that over our period literature changed from a more functional concept (e.g., it was initially used similarly to "rhetoric") to a more aesthetic concept (e.g., it was finally used more similarly to "art") . (Gefen et al. 2017) In this project, we retain the vector model but examin the results yielded by K-means clustering rather than cosine similarity.

III. Methods

A. Cosine Similarity

Cosine similarity is a simple method of measuring distance between two multidimensional vectors according to the cosine of the angle between them. Word vector analysis frequently uses this method as a way to measure the similarity of usage between two different words. In the past, as we describe below, we have used cosine similarity to examine the terms that have been "closer to" or "farther from" literature/littérature in our

[1] Stanford University
[2] CNRS
[3] Paris III - Sorbonne Nouvelle

corpora. We measure the similarity between two words by calculating the distance between their word vectors across time. (Tan et al. 2006)

B. K-means

K-means is a widely employed unsupervised machine learning tool for clustering multidimensional vectors. For our purposes, the primary interest of K-means is its capacity to determine clusters given some number of properties of relation. Instead of operating on a word-to-word basis, like cosine similarity, K-means allows us to examine many-to-many relationships. In other words, instead of seeing how "literature" compares to some other word, we can find the group of words in which literature is a member, and examine that cluster relative to other semantic groups. (Tan et al. 2006)

IV. Results

Because our results are quite expansive, we have selected just one to demonstrate for this abstract.

Example Result: "Littérature" and "technique"

In the following figures we will show that the K-means clustering and the cosine similarity measurements both help us see the evolution of individual words, but also how the clusters that define them move and interact through time. We will verify a common theory in French literary criticism that sees the turning point towards a technical perception of literature at the end of the nineteenth century.

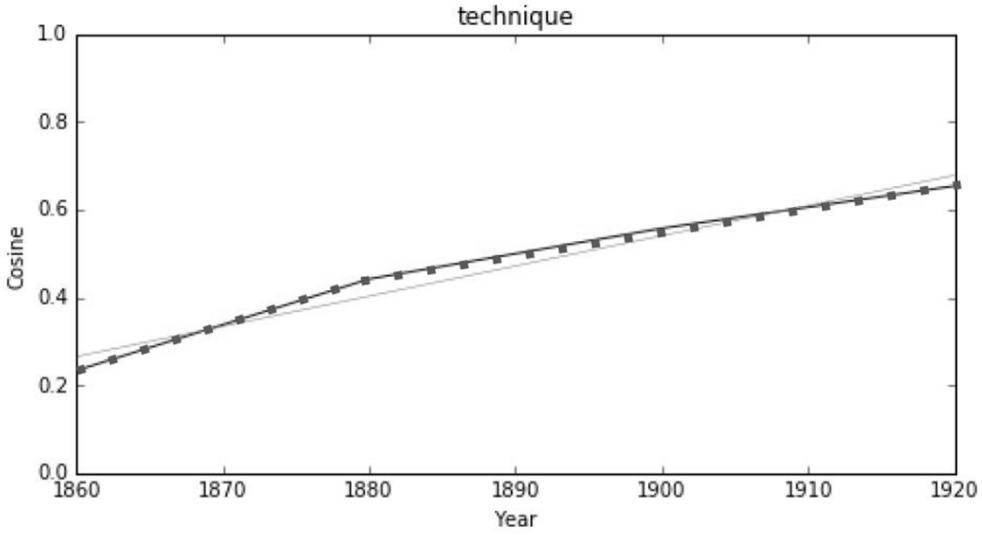


Fig.1 Cosine similarity with "littérature" over time for "technique"

In Fig. 1 we can see that the general similarity between "littérature" and "technique" is increasingly important through time, and that the peak for this evolution happens during the 1880s. This result that tends to verify the hypothesis.

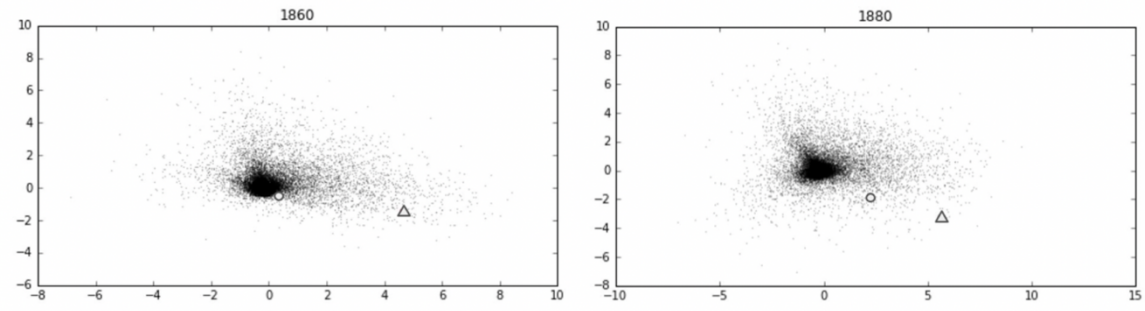


Fig.2

In Fig.2, we turn to cluster analysis. We can see that the cosine hypothesis is supported, since the cluster for “littérature” (red) moves closer to that for “technique” (blue). But we also see that both of them tend to disassociate themselves from the other clusters. That is to say, their meaning is getting more and more specific, as they are no longer near the mass of other clusters. It appears that the further out from this “center of gravity” a cluster gets, the more distinct its meaning becomes—in other words, these charts may show us the increasing monosemy of both literature and technicality as concepts.

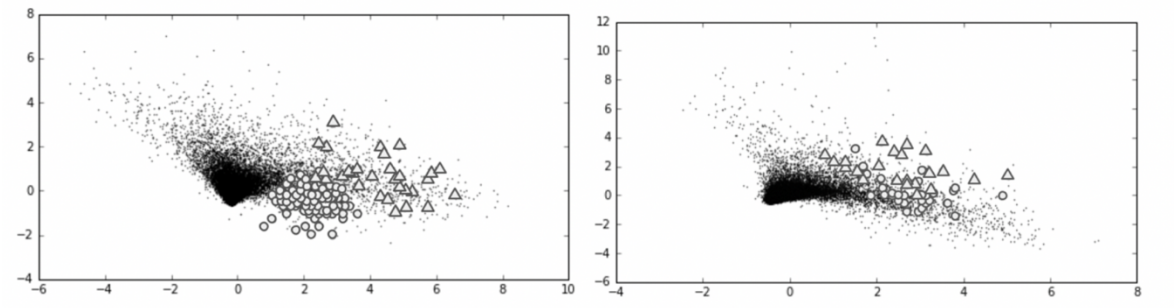


Fig. 3

In Fig.3 we compare how the clusters “technique” (blue) and “littérature” (red) belong to and see that both the predictions of the cosine and the clustering representation point to the same conclusions: first, “littérature” and “technique” share a lot more semantic space in the 1880s than in the 1860s; second, the term “technique” is attracted out of the central mass of clusters, and shares more space with “littérature”. The size of the cluster is bigger in the 1860s, signifying that the meaning was broader. Both terms became more specific and associated over time.

V. Conclusions

Our main conclusion is that the introduction of K-means clustering to our existing work with cosine similarity adds important nuance to our specific project of understanding the history of “literature” as a concept. Cosine similarity had already given us a sense of the ways that “literature” changed with respect to particular terms like “science” or “history”. But K-means clustering reminds us that “literature” does not travel alone. Rather, it has a semantic cohort, a cluster of words that move in relation to other words which are also configured in clusters. Thinking in these terms reconfigures our interpretation. As one specific example, we now have evidence that the literature concept was part of a suite of words that moved toward monosemy over the course of our period—that is, literature was not simply moving away from one meaning and toward another, but away from generality and toward specificity.

Select Bibliography

Alexandre Gefen, Mark Andrew Algee-Hewitt, David McClure, Frédéric Glorieux, Marianne Reboul, J.D. Porter, Marine Riguet. (2017). “Vector based measure of semantic shifts across different cultural corpora as a proxy to comparative history of ideas.” *Proceedings of the Japanese Association for Digital Humanities Conference 2017*, Kyoto, Japan.

Riguet, Marine, and Alaa Abi-Haidar. (2017). “Faire figure d’autorité : l’analyse de réseaux appliquée au discours”. In *Analyses et méthodes formelles pour les humanités numériques.*

Saussure, Ferdinand de. Ed. Perry Meisel, Haun Saussy. Tr. Wade Baskin. (2011). *Course in General Linguistics.* New York: Columbia University Press.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. (2006). *Introduction to Data Mining.* Boston : PeaArson Addison Wesley, pp. 75, 497.

Historical Big Data: Reconstructing the Past through the Integrated Analysis of Historical Data

Asanobu Kitamoto[1], Mika Ichino[1], Chikahiko Suzuki[1], Tarin Clanuwat[1]

1. Introduction

People of the present use smart phones to record their activity and send short messages (e.g. tweets) to communicate with other people. To take advantage of data-centric society, the "big data" approach focuses on reconstructing the real world in the cyber space, which is sometimes called "digital twin," through the integrated analysis of a variety of data, such as online data generated by people and sensor data observed in the world.

Then a question arises; is it possible to apply the same approach to reconstruct the world of 200 years ago? People recorded and communicated in many different ways in that age, such as writing diaries, letters and other documents on the paper. How can we reconstruct the world in the past even if we replace data sources from online dense sources of the present to offline sparse sources of the past?

This is a question we would like to ask in the "historical big data (HBD)" research. To fill the gap between dense data of the present and sparse data of the past, however, a challenge is to develop statistical or knowledge-based inference to turn sparse data into dense data suitable for quantitative analysis. Another challenge is to develop workflow to collect more high quality data for better reconstruction.

2. Concept of historical big data

HBD is the historical version of "big data" in the sense that we focus on the continuity of time between the past and the present to transfer algorithms, software tools, and frameworks developed for the present to the past. Our aim is to realize seamless big data platform that takes advantage of experiences and solutions for the present big data and takes historical perspectives into consideration.

Big data is typically analyzed by 4V, namely volume, variety, velocity and veracity, but relative importance of four concepts is different in a historical context. First, velocity is not a critical issue because we cannot make any actions on the analysis. Second, volume is not critical for historical structured data, but is not trivial for historical unstructured data such as digitized high resolution images and 3D models. Third, variety is probably the most important challenge. The workflow usually starts from digitization, cataloging, transcription of analogue data with the goal of creating structured data for "deep access" or machine-readable access to the content of historical documents. Here machine learning, such as character recognition, image tagging, and object detection, is expected to help structuring the variety of content. Fourth, veracity to evaluate the reliability of text is called source criticism, which is the core research challenge of historical studies[1].

In summary, present and historical big data research share similar challenges with different relative importance.

3. Types of Historical Big Data

The scope of historical big data ranges from natural phenomena such as weather and earthquake to societal phenomena such as market price, and human daily lives as well as crisis and disasters[2]. To turn a variety of unstructured data into records of structured formats, we study three types of historical big data as follows.

3.1 Historical situation record (HSR)

Situation record includes human sensory observations with, if any, spatio-temporal coordinates and writer's entity. Weather is a simple situation record because it is the result of human visual observation. Earthquake, on the other hand, may be a complex situation

[1] Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems / National Institute of Informatics

record. First, the earthquake itself can be observed as auditory information or tactile information, then as visual information. Second, the aftermath of the earthquake is observed as visual information such as collapse of the building or the death of people.

3.2 Historical activity record (HAR)

Activity record includes human actions that cause changes. For example, a human moves from a location to another location, a human buys or eats something at a shop, or a human makes a trip to a sightseeing spot. Difference between situation record and activity record is that the former is the description of the world, while the latter is the description of the human.

3.3 Historical transaction record (HTR)

Transaction record is the observation of factual data, such as market price, or the movement of commercial goods. They are the result of human activity, but not the activity itself.

Work is in progress to design a common metadata format for use cases such as weather records[3], with a workflow to fill the gap between data creators and users. Creators find weather description in old diaries and transcribe it as qualitative text, but users in paleoclimate study needs quantitative data. In addition, users prefer to have a reliability (quality) parameter for each record, but this is not a simple task for creators to estimate it. It is more likely that a reliability parameter is evaluated after studying relationship between a set of records with the help of visualization and comparison tools.

4. Related and Future Work

Similar projects have already started in Europe under the Time Machine FET Flagship consortium, which aims at building a large scale historical simulator mapping 2000 years of European History[4]. Most prominent project is Venice Time Machine[5], which provides a proof of concept of archival digitization and machine learning to reconstruct the shape of the city over its history. A similar project, Amsterdam Time Machine[6], has also started to create a hub for linked historical data, or the web of information on people, places, relationships, events, and objects in time and space through geographical and 3D representations. Inspired by those European projects, we could call our project as Edo Time Machine[7].

The initial step of our project focuses on people and places in Bukan Complete Collection (Figure 1), which covers almost 200 years of people in states (Daimyo) and the central government (Bakufu)[8]. Another focus is tourism such as the database of sightseeing spots in Edo (the old name of Tokyo). The hub of linked historical data will allow us to ask new types of research questions, and new answers to these questions will broaden our view on the history.

Figure 1: Bukan Complete Collection website (http://codh.rois.ac.jp/bukan/). Left: the list of Daimyo family emblems; right: animated visualization of spatio-temporal patterns of Daimyo trips (historical activity records).

References

1. **Asanobu KITAMOTO, Yoko NISHIMURA.** (2016). "Digital Criticism Platform for Evidence-based Digital Humanities with Applications to Historical Studies of Silk Road", Digital Humanities 2016: Conference Abstracts.
2. **Center for Open Data in the Humanities.** (2018). Sixth CODH Seminar: Historical Big Data - Challenges in Transforming Historical Documents to Structured Data for the Integrated Analysis of Records in the Past -, http://codh.rois.ac.jp/seminar/historical-big-data-20180312/ (accessed 8 May 2018).
3. **Mika ICHINO, Kooiti MASUDA, Asanobu KITAMOTO, Junpei HIRANO, Kenjiro SHO.** (2017). "Experience of historical climatology as a material in Digital Humanities", IPSJ SIG Computers and the Humanities Symposium 2017, pp. 139-146 (in Japanese).
4. Time Machine FET Flagship, https://timemachineproject.eu/ (accessed 8 May 2018).
5. **Alison Abbott.** (2017). "The 'time machine' reconstructing ancient Venice's social networks", Nature 546, 341–344, doi:10.1038/546341a.
6. Amsterdam Time Machine, http://www.create.humanities.uva.nl/uncategorized/amsterdam-time-machine/ (accessed 8 May 2018).
7. **Asanobu KITAMOTO, Hiroshi HORII, Misato HORII, Chikahiko SUZUKI, Kazuaki YAMAMOTO.** (2017). "Structuring Time-Series Historical Sources by Human-Machine Specialization: Toward the Construction of Edo Information Platform Referring to "Bukan"", IPSJ SIG Computers and the Humanities Symposium 2017, pp. 273-280 (in Japanese).
8. **Asanobu KITAMOTO, Hiroshi HORII, Misato HORII, Chikahiko SUZUKI, Kazuaki YAMAMOTO, Kumiko FUJIZANE.** (2018). "Differential Reading by Image-based Change Detection and Prospect for Human-Machine Collaboration for Differential Transcription", Digital Humanities 2018: Conference Abstracts.

A community based on data sharing and collaboration. The structure of the ZX Spectrum demoscene

Piotr Marecki[1]

As part of the presentation, the results of a two-year research project on the ZX Spectrum demoscene, which was carried out in the creative programming laboratory (UBU lab) at the Jagiellonian University in Krakow, will be presented using tools offered by media archeology and platform studies.

The Idea

The point of departure for our research and cooperation are the concepts of the third culture and the new renaissance proposed by John Brockman (Brockman, 1995). They refer to the lecture "The Two Cultures" from 1959 delivered by C.P. Snow, who initiated the discussion on the discrepancy between two environments: humanistic and scientific. As Snow proved, the intellectual life of Western societies is based on two worlds that have no connections and almost no dialogue between them. These two cultures develop without looking at each other, without understanding, and draw on separate dictionaries. "Intellectualists of literary provenance" and "scientists" use languages that are incomprehensible to the other, and the territories in which they move are treated as foreign worlds. John Brockman, referring to Snow's concept, proposed a model of the third culture, assuming the rejection of old divisions and building - as he called it - a new renaissance. In his understanding, in connection with the accelerated development of technology and its impact on life activities and expression, humanists are forced to learn and be able to name the secrets of science to be able to consciously participate in the modern world. The research presented below was carried out in a digital media laboratory, which brings together scientists, artists and programmers to implement the third culture principles into practice. According to the third culture, joint research on the demoscene is carried out by an expert on digital culture, artist and programmer.

The Research

The demoscene is mainly a European phenomenon. This is a community made up mostly of geeks and platform fans, who organize demoparties in order to present demos at them. The demo is a digital creative work which has at its sole purpose to demonstrate the capabilities of a given platform and the programmer. The name itself derives from the word "demonstration". Demosceners, therefore, have no thoughts or stories to convey, their works are more about a kind of praise. The demoscene brings together tens of thousands of people that deal both with old platforms (like, first personal computers or 8-bit consoles), but also with new platforms (modern computers or consoles). Since the 1980s, it has been a type of organized anarchy; parties as well as groups acting on the scene are not registered, and all demos that are created are immediately made available to other members of the community. Demosceners create a field of cultural production, and as it is understood by the French sociologist Pierre Bourdieu, it is treated as a part of a social structure which gathers together actors with similar aspirations (Bourdieu, 1996).

The demoscene is therefore one of the creative areas of the digital media field (other fields include video games, electronic music, media art or electronic literature). My talk will be devoted to the demoscene that is focused around the first European personal computers, ZX Spectrum, developed in Great Britain. Its use, however, in the country of its production will not be the area of my interest. I will focus on the creative use of the platform mainly in Russia, Poland, the Czech Republic and Slovakia. In these countries, the platform was adopted thanks to a specific approach to copyright issues. As part of this paper, I will present a select aspect of my research on the ZX Spectrum demoscene, referring to the structure of the demoscene and the issue of data sharing and archiving.

[1] Jagiellonian University

People working on the scene have their own functions; there is a certain structure, and specific rules for evaluating works during demoparties that build hierarchies in the field.

Relevant issues connected to identifying with the platform and building communities around it and structures within it will be presented. An important element is the creation of groups that consist of a graphic designer, coder, musician. Particular emphasis will be placed on an analysis of the role and function of the swapper, which in the pre-Internet era served as the person responsible for data distribution. The distribution strategies will be described, as well as the aesthetics of the distributed data (ways of creating floppy disk covers, paper as a data carrier, letters sent by swappers). Looking closely at the chosen community from the point of view of Bourdieu's field of cultural production theory, the ZX Spectrum demoscene will require determining the community's relationship with other scenes focused on other computers. Thus, the phenomenon of platform wars will be presented. In addition to Bourdieu's theory and media archeology (Parikka, 2012), methodologies such as platform studies and software studies have been used for the research. Especially distinguishable from the tradition of media history, media archeology allows us to look at known phenomena and narratives from a different perspective. The proposed research adds to the official history a little known side of platform development in Europe, mainly in countries behind the Iron Curtain. The findings are the result of in-depth interviews with the participants of the scene and ethnographic observations.

References

Bogost I., Montfort N. (2009). Platform Studies: Frequently Questioned Answers http://nickm.com/if/bogost_montfort_dac_2009.pdf (Accessed 03.05.2018)

Bourdieu P. (1996). *The Rules of Art: Genesis and Structure of the Literary Field.* Transl. Emanuel. S. Stanford, Calif.: Stanford University Press.

Brockman J. (1995). The Third Culture: Beyond the Scientific Revolution. New York: Simon and Schuster.

Parikka, J. (2012). What Is Media Archaeology? Cambridge: Polity

Polgár T. (2005). *Freax. The Brief History of the Computer Demoscene.* Winnenden: CSW Verlag.

Project financed by the program of the Polish Ministry of Science and Higher Education "National Programme for the Development of Humanities"

Towards Unifying Our Collection Descriptions: To LRMize or Not?

Jacob Jett[1], Katrina Fenlon[1], J. Stephen Downie[1]

Digital cultural heritage collections (hereafter referred to as collections) play an important role in digital humanities scholarship. The Meiji and Taisho Eras in Photographs (MTEP) special collection (http://www.ndl.go.jp/scenery_top/e/) created by the National Diet Library (Figure 1) is one example of an institutionally-built collection. There is also a growing interest in scholar-built collections, such as, HathiTrust Research Center's worksets (Jett et al. 2016), HathiTrust's user collections (https://babel.hathitrust.org/cgi/mb?colltype=updated), etc. Both of these kinds of collections are important to scholars because they provide resources ready for reuse in various research contexts. Regardless of whether a collection was built by an institution or a scholar, the challenge for many scholars is identifying all of the collections, or resources within them, pertinent to their research questions. Unfortunately while many collections are described, no singular unifying manner of describing them has emerged. Thus the identification and selection of pertinent resources remains a challenge. To alleviate this challenge we assert that digital cultural-heritage collections be treated as first-class bibliographic objects in their own right.

Figure 1: NDL's Meiji & Taisho Eras in Photographs Collection

Describing collections as first-class bibliographic objects in their own right is essential to facilitating their identification and use by scholars (and other users, scholarly or otherwise). In a perfect world a conceptual model like the International Federation of Library Associations' (IFLA's) Work-Expression-Manifestation-Item (WEMI) model (IFLA 2009) would provide a nucleus around which a unified model for describing aggregates like collections could be developed. The WEMI model is cornerstone of IFLA's Functional Requirements for Bibliograpic Records (FRBR). Using it would allow their descriptions to be integrated with our existing FRBR-ized catalog databases thereby making it easier for scholars to complete the common user tasks (summarized in Table 1 below) that WEMI is designed to facilitate (IFLA 2009, Riva et al. 2017).

[1] University of Illinois

Table 1: User Tasks Summary (excerpted from Riva et al. 2017, p 15)

Find	To bring together information about one of more resources of interest by searching on any relevant criteria
Identify	To clearly understand the nature of the resources found and to distinguish between similar resources
Select	To determine the suitability of the resources found, and to be enabled to either accept or reject specific resources
Obtain	To access the content of the resource
Explore	To discover resources using the relationships between them and thus place the resources in a context

Ideally treating collections as first-class bibliographic objects would make it easier for scholars to identify and exploit collections similar to NDL's MTEP collection. For example, a scholar using a in a FRBRized IR system that could recommend related collections such as the Widener Library's collection of Japanese Photographs of the Meiji Period (http://id.lib.harvard.edu/aleph/008800120/catalog), Nagasaki University Library's collection of Old Photographs from the Bakumatsu-Meiji Period (http://sepia.lb.nagasaki-u.ac.jp/jp/), or the Oxford's Pitt River's Museum's collection of Meiji-era photographs (http://pittrivers-photo.blogspot.com/2017/07/picturing-japan-meiji-era-photographs.html). These collections could be brought together via through the FRBR works' property *has subject*, since they all have the same or related topicalities. Furthermore adopting a FRBR-ized view of these collections—which are carefully curated specifically to cultivate the context among the objects being gathered into them (Palmer 2004, Palmer et al. 2010)—affords scholars the ability to differentiate among different versions of the collections. This affordance is important because digital collections wax and wane over time according to the contextual needs of their users (Fenlon 2017). Thus a FRBRized view of collections enables scholars to identify and select the particular version of a collection most suitable to their research needs.

However, recent work reconciling FRBR and the FRBR-related series of conceptual models set forth by IFLA (Riva et al. 2017) in the form of the IFLA Library Reference Model (LRM) may negate the apparent gains of FRBRizing collection descriptions. Specifically, LRM employs the controversial (Tillet et al. 2014) findings from the FRBR Working Group on Aggregates (2011) to set forth a trio of new definitions for the media type—*Aggregates*. The underpinnings of these new definitions set forth the position that all *aggregates*, including collections, are not describable at the WEMI *expression* and *work* levels. Instead, we are told, "An aggregate is defined as a *manifestation embodying multiple expressions*." – Riva et al. 2017, p 93. In this world-view digital cultural heritage collections could never be first-class bibliographic objects.

Among the problems associated with the LRM model for aggregates is the inability to express the topicality of digital collections. According to the WEMI model (which LRM uses), only *works* possess the *has subject* property; *manifestations* do not (Renear & Choi 2006). The demotion of collections from works to manifestations in the LRN model is problematic because the topicality of a collection (and other aggregates) is not necessarily (and perhaps not even often) derived from the topicality of the individual works gathered into them (Fenlon 2017; Jett & Dubin 2018). Consequently, if an existing standard like the HTRC's workset ontology (Jett et al 2016) were to adopt the LRM model, much of the metadata currently recorded by the standard—properties like *has criterion, has research motivation, has intended use, has subject*—would all need to be removed from the standard (and from all existing workset/collection descriptions).

For instance a collection of photographs that were taken during the Meiji-era might be topically about Meiji-era life and culture but the individual photographs will have their own topicality as *works* in their own right. The loss of descriptive power associated

with the collection as a whole bibliographic unit presents a formidable obstacle for scholars attempting to complete FRBR's *identify* and *select* user tasks using IR-systems based on the LRM model.

To summarize, we find that a FRBRized view of collections as first-class bibliographic objects presents many benefits for scholars by providing a method for linking collections together through topicality. However adopting the LRM model impedes scholarly research by mandating the removal of key parts of collection descriptions thereby making important user tasks, like *identify* and *select* much harder by breaking topical links.

References:

Fenlon, K. (2017). *Thematic research collections: Libraries and the evolution of alternative scholarly publishing in the humanities* (Doctoral dissertation). Retrieved from: https://www.ideals.illinois.edu/handle/2142/99380

FRBR Working Group on Aggregates (FRBR-WGA). (2011). *Final report of the Working Group on Aggregates.* The Hague: IFLA.

IFLA Study Group on FRBR (IFLA). (1998). *Functional requirements for bibliographic records: Final report* [revised 2009]. München: K.G. Saur Verlag.

Jett, J., Cole, T. W., Maden, C., & Downie, J. S. (2016). "The HathiTrust Research Center workset ontology: A descriptive framework for non-consumptive research collections." *Journal of Open Humanities Data,* 2, p e1. DOI: http://doi.org/10.5334/johd.3

Jett, J. & Dubin, D. (2018). "How are dependent works realized?" Paper presented at *Balisage: The Markup Conference 2018,* Washington D.C., 31 July – 3 August, 2018.

Palmer, C. L. (2004). Thematic research collections. In Schreibman, S., Siemens, R., and Unsworth, J. (Eds.) *A Companion to Digital Humanities.* Oxford: Blackwell Publishing.

Palmer, C. L., Zavalina, O. & Fenlon, K. (2010). "Beyond size and search: Building contextual mass in aggregations for scholarly use." *Proceedings of the 73rd ASIS&T Annual Meeting* (Pittsburgh, PA, 22-27 October 2010).

Renear, A. H. & Choi, Y. (2006). "Modeling our understanding, understanding our models: The case of inheritance in FRBR." *Proceedings of the 69th ASIS&T Annual Meeting* (Austin, TX, 3-8 November 2006).

Riva, P., Le Boeuf, P., & Žumer, M. (2017). *IFLA library reference model: A conceptual model for bibliographic information.* The Hague, Netherlands: IFLA.

Tillett, B. B., Kuhagen, J. A., Cato, A. & Murtomaa, E. (2014). "Letter to the editor." *Cataloging & Classification Quarterly,* 52(3): 359-61.

Exploring the Implications: Open Access Repositories and Social Media

Luis Meneses[1], Alyssa Arbuckle[1], Hector Lopez[1], Belaid Moa[2], Richard Furuta[3], Ray Siemens[1]

Vannevar Bush, in his pioneering 1945 essay "As We May Think," (Bush, 1945) envisions a time when the world's knowledge is accessible by machine, and the connections that describe the higher level relationships among sources are themselves shareable objects of scholarship. We can see this today on the Web in the utility of investigation mechanisms such as Walden's Paths (Bogen et al., 2011), where users can build linear narrative structures from online resources. This is a natural side effect of collaboration and cooperation. As the problems to be solved in the Humanities grow beyond the technical abilities of an individual scholar, and as social media become more embedded into our work practices, the presence of resources that situate knowledge into the broader environment will also become more prevalent.

The methods for representing documents and disseminating knowledge are changing. In recent years we have witnessed an increase of social media, which challenges how scholarship is processed and distributed. While it is true that that the definition of social media is very nuanced, it does emphasize its relevance and its potential to transform the scholarly communication system (Sugimoto et al., 2016). More so, mechanisms are not in place to assimilate the discussions that are happening within social media into the workflow of a digital collection.

We have developed a framework to address these challenges that extends the functionality of an Open Access Repository by implementing processes to incorporate the ongoing trends in social media into the context of a digital collection. We refer to these processes collectively as the Social Media Engine. This engine and its underlying framework aim to instigate public engagement, open social scholarship, and social knowledge creation by matching readers with publications. This framework relies on the gathering and analysis of corpora harvested, indexed, and rendered through Open Access and academic materials —which were influential towards our technical choices.

The fundamental concepts behind our framework can be explained using three points. First, our framework yields a list of topics related to individual entries and articles in the corpus by applying textual analysis techniques and topic modeling. Second, our engine connects readers and publications by monitoring social media for trending topics and returning links to Open Access publications —which are then used to feed and enrich the discussion due to their stability and persistence. Finally, our engine identifies trending papers on social media by making inferences on the number of times that papers on social science topics are shared, saved, liked, or commented on. Altogether, our framework uses social media to reorganize the contents of an Open Access Repository by suggesting keyworks, reordering rankings and notifying users about relevant resources.

Our previous presentations have focused on the computational aspects behind our framework (Meneses et al., 2017b) (Meneses et al., 2017a). In this abstract, we propose to elaborate on the findings that we have obtained from implementing a prototype, its resulting implications, and our plans for future work. This analysis gains special relevance when taking into account that our framework changes some preconceived notions by making repositories more dynamic —and consequently changing the patterns of interaction in Open Access Repositories.

This change in the patterns of interaction that we are pursuing brings forth several implications that must be addressed within the context of the Humanities —mostly stemming from the underlying technologies that our framework employs and the

[1] Electronic Textual Cultures Lab, University of Victoria
[2] University of Victoria
[3] Texas A&M University

characteristics of our document corpus. For instance: our framework has three main components: 1) a keyword extraction module, 2) a social media mining component and 3) a search engine (Apache Software Foundation, 2017b). These components rely on a complex set of technologies —that include metadata harvesting protocols, parallel programming languages (Apache Software Foundation, 2017a), topic modeling (Blei et al., 2003) and keyword extraction techniques. Additionally, our framework incorporates external APIs and services: we use Altmetric.com (Altmetric LLP, 2017) to monitor social media, identify trending topics and facilitate the matching of readers with publications that are of their interest.

In terms of the Humanities, we had to devise specific methods to process our corpus —that differ from approaches in other disciplines. We can point out that the multiple languages in the documents and the lack of standardized metadata influenced our workflow and methods. As examples, we found that 91% of the documents were in French, 8.6% in English and the remaining 0.4% in other languages. We also found the metadata that was provided to us was incomplete: 19% of the documents in our collection had descriptions and full text that were in the same language. Understanding these nuances in the Humanities allowed us to grasp an overall understanding of the collection and set a foundation towards implementing solutions that can deal with the characteristics of the documents.

However, we believe that one of the more complicated aspects of our study is its assessment. We performed ranking calculations using topic modeling, term frequency–inverse document frequency (Tf-Idf), and a combination of topic modeling and Tf-Idf. Using a simple ranking function, Tf-Idf provided better results over topic modeling. This was expected to some extent, given that the terms come from the documents themselves. On the other hand, the classification from the topic modeling also come from the documents, but provide an overall context for arranging the collection. Ultimately, we believe that a more thorough evaluation is needed to assess our framework —which can be obtained by running user studies. We have scheduled two rounds of user studies for the near future.

All in all, our framework is composed of a diverse set of complex technologies that are evaluated with equally intricate evaluation schemes. Why did we choose this specific set of technologies? How are they impacting our analysis and evaluation? Taking into account that the intended use of these technologies diverges from their actual application in our framework, their resulting implications are worthy of study and analysis. We will address these two questions and their implications in the longer version of our abstract.

References

Altmetric LLP (2017). Altmetric https://www.altmetric.com/ (accessed 20 October 2017).

Apache Software Foundation (2017a). Apache Spark: Lightning-fast cluster computing http://spark.apache.org (accessed 11 April 2017).

Apache Software Foundation (2017b). Apache Solr http://lucene.apache.org/solr/ (accessed 20 October 2017).

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3: 993–1022.

Bogen, P. L., Pogue, D., Poursardar, F., Li, Y., Furuta, R. and Shipman, F. (2011). WPv4: a re-imagined Walden's paths to support diverse user communities. *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries.* 1998164: ACM, pp. 419–20 doi:10.1145/1998076.1998164.

Bush, V. (1945). As We May Think. *The Atlantic Monthly.*

Meneses, L., Arbuckle, A., Lopez, H., Moa, B. and Siemens, R. (2017a). Social Media Engine Paper presented at the Open Cyberinfrastructure for the Humanities and Social Sciences Workshop 2017, Montreal, Canada.

Meneses, L., Arbuckle, A., Moa, B., Furuta, R. and Siemens, R. (2017b). Towards a more Context Aware Digital Library: Implications in the Humanities Paper presented at the Joint CSDH/SCHN & ACH Digital Humanities Conference 2017, Toronto, Canada.

Sugimoto, C. R., Work, S., Larivière, V. and Haustein, S. (2016). Scholarly use of social media and altmetrics: a review of the literature. *ArXiv:1608.08112* [Cs] http://arxiv.org/abs/1608.08112 (accessed 3 May 2017).

Towards unified descriptive practices for Japanese classical texts: TEI, IIIF, and the UCLA Toganoo Collection of Esoteric Buddhism

Tomoko Bialock[1], Dawn Childress[1], Hiroyuki Ikuura[2], Kiyonori Nagasaki[3]

Using UCLA Library's Toganoo Collection as a case study, this paper discusses the description and presentation of Japanese classical texts in a TEI / IIIF ecosystem.

In this study, the presenters explore the implications of using TEI in the new networked environments in which we are increasingly operating. For digital catalog / corpora projects like the Toganoo collection, the data standards and systems used serve a number of different functions, from searching and faceted browsing to detailed encoding of features, content, and context, and now interoperability and exchange of project data between systems. There exists no single system or standard that meets the needs of all functions, but rather than being limited to one system, we work with modular technologies that co-exist with one another. Thanks to the principles of linked data, we needn't limit ourselves to a single standard such as Dublin Core, MODS, or TEI. When another standard meets some of our descriptive needs in ways that TEI might not, we can envision a project space where we adopt multiple schemas at once, using a variety of namespaces within a single XML document and making use of the elements from each schema that best serve the materials. Also, where we customize and create new project-specific fields, we can to make these elements reusable as linked data URIs that can then be easily adopted by similar projects.

The UCLA Library Toganoo Collection of Esoteric Buddhism comprises 342 titles in 968 volumes of classical and modern Japanese texts[1]. In addition to the bibliographic data about each text, the collection is rich with provenance data and serves as a valuable resource for tracing the provenance of Japanese texts, book distribution histories, and uncovering patterns in the history of Japanese books. As part of the UCLA Library Digital Collections, bibliographic metadata for the collection is recorded using a MODS-based schema to facilitate efficient and scalable search and browse of digital collections[2]. Standards such as there are ubiquitous and necessary for interoperability between systems and for search and browse functions. The project also relies on the IIIF APIs and standards for viewing and annotation of high-resolution images and to promote sharing and interoperability between institutions and researchers.

While both MODS and IIIF are integral to the project, both standards fall short when it comes to cataloguing and describing manuscripts and early printed works. These materials require more nuance in their description if they are to be accurately identified and described, especially if the data is to be useful in the context of new linked data environments and data-inflected research methods beyond discovery and access[3]. In the case of the Toganoo Collection, achieving "thick" bibliographic description is necessary to expose the data for the study of Japanese classical texts[4].

To meet the bibliographic description and interoperability needs, the project looks to the Text Encoding Initiative (TEI) guidelines. TEI, a standard for the representation of texts in digital form, has been used with much success to describe manuscripts and early printed books where detailed bibliographic descriptions are desired. TEI provides structured descriptions of texts at a variety of different levels and, as an XML technology, lends itself to readily to machine processing. Of specific interest is the TEI "manuscripts" module. Here, TEI provides detailed specifications for describing manuscripts and similar materials, allowing for much fuller description of manuscript characteristics. List of

[1] University of California, Los Angeles
[2] Waseda University
[3] International Institute for Digital Humanities / The University of Tokyo

institutions using TEI for describing manuscripts and early printed texts is long, including the University of Cambridge Library and Penn Libraries, both employing TEI for description of East Asian texts as well. The TEI files underlie discovery and the metadata display, and are openly available for direct download and reuse by scholars[5].

During the short talk, presenters will make the case for using TEI as the central data store and demonstrate how they are making use of the TEI encoded data throughout the project, from generating derivative data (MODS, IIIF) to analysis and visualization. The team will also examine the roles and functions of TEI, IIIF, linked data, and other technologies as they converge and work in tandem to support analysis, presentation, and sharing of the Toganoo and related materials.

[1] The Toganoo Collection was curated by Toganoo Shōun (1881-1953), scholar of modern Esoteric Buddhism who served as president and chief librarian of Kōyasan University, and purchased by UCLA in 1962-63.

[2] See also, MARC and MODS, developed by the Library of Congress and Dublin Core, developed by OCLC for digital collections

[3] Childress, D. “Beyond Access: Critical Catalog Constructions.” DH 2017, Montreal. http://dawnchildress.com/2017/07/20/speccat/

[4] *Collectors’ Seal Database*: http://base1.nijl.ac.jp/~collectors_seal

[5] Cambridge Digital Library: https://cudl.lib.cam.ac.uk/ ;
OPenn: http://openn.library.upenn.edu.
For TEI examples, see https://cudl.lib.cam.ac.uk/view/PR-FJ-00734/1 and Zengxian Li , Hyōsen Kōshi Kego, https://cudl.lib.cam.ac.uk/view/PR-FB-00769-00001/1

A TEI Markup for the Contents of Tang Poems

Yan Cong[1], Masao Takaku[1]

Currently, many Tang-dynasty poems are used in the student textbooks of Japanese junior and senior high schools. To help students understand the meaning of a poem, multiple annotations of kanji (Chinese characters) are provided around certain kanji in the textbooks. These multiple annotations, which are called *kunten* (訓点), not only indicate a pronunciation or an interpretation of the kanji, but also order the sentence correctly in Japanese. However, when converting the fulltext with annotations into a digitized text, there is no simple standard for transcribing elements such as *kunten*. This makes it quite difficult to digitize the original Tang poems with *kunten*.

In this study, we marked up the contents of Tang poems that were included in the textbooks of Japanese junior and senior high schools in 2016 according to TEI: P5 Guidelines[1]. We primarily marked up fulltext of Tang poems along with the *kunten* as punctuated texts (*kundokubun*; 訓読文).

The textualization of Tang poems plays an important role in helping poetry learners search for related information in various ways. This related information can be used and shared by everyone and help learners understand related materials across the fulltexts more easily. Digitization of Tang poems provides machine-readable sentences, including specific *kunten* in the written sentences.

TEI: P5 Guidelines[1] have standard elements for content markup. TEI markup is used for machinereadable material and can be used in combination with general-purpose XML. With TEI: P5 Guidelines[1], the related information of Tang poems is marked up as follows.

1) **The title of a Tang poem and the author's name.** The titles include the forms of both punctuated text and reading text.
2) **The fulltext of a Tang poem and the types of different content expressions.** There are four types of fulltext expressions: unpunctuated text (*hakubun*; 白文), punctuated text (*kundokubun*; 訓読文), reading text (*kakikudashibun*; 書き下し文), and translation text (*honyakubun*; 翻訳文) [2].
3) **A line of Tang poem and the number of the line.** The verse line contains a single line in a Tang poem, and the line number indicates the line of the poem in which this part of the verse is included.
4) **The *kunten* information in the fulltext.** The focus is on the punctuated texts used in textbooks, and we particularly mark up the fulltext of the Tang poem with *kunten* information.

TEI Guidelines[1] and XML are used to mark up Tang poems. First, the element <head> that contains any type of heading is used to indicate a Tang poem's title. The element <persName> indicates the author's name of the Tang poem.

Second, the element <lg> is an abbreviation for "line group," which contains the complete fulltext of the Tang poem as a formal unit. A type attribute of the element <lg> indicates the fulltext expressions. As an attribute value of the type, the name of the fulltext expression is used. Tang poems are divided into four kinds of fulltext expressions due to the content's different reading order. They are named as unpunctuated text (*hakubun*; 白文), punctuated text (*kundokubun*; 訓読文), reading text (*kakikudashibun*; 書き下し文), and translation text (*honyakubun*; 翻訳文)[2]. The specific names of attribute values used are unpunctuated, punctuated, reading, and translation.

Third, the element <l> is a part of the element <lg> and indicates a single line of the Tang poem. We use attributes n = 1, 2, ... to indicate the line number in the Tang poem.

[1] University of Tsukuba

Finally, the element <seg> denotes an arbitrary segment and describes the *kunten* information that is adjacent to certain kanji. *"Okurigana"* and *"kaeriten"* are usually used in textbooks to depict *kunten*. *Kaeriten* indicates the order of the adjacent Kanji, which are marked by *re* (レ), *one* (一), *two*(二) and should be reversed and read in Japanese order. Since TEI Guidelines do not include elements that can explain the *kunten* information to be given around a particular character, we decided to name the types of *kunten* information with the Roman Latin spelling in Japanese as its own attribute value. In other words, the attribute value of an attribute "type" of element <seg> will be set as "*okurigana*" and as *"kaeriten"* in punctuated text.

Table 1 Example of the fourth line of "Thoughts in a tranquil night" by Li Bai

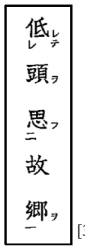
Fourth line of original fulltext	 [3]
Marked up content	<head>静夜の思ひ</head> <persName>李白</persName> <lg type="punctuated"> … <l n="4"> 低<seg type="okurigana">レテ</seg><seg type="kaeriten">レ</seg> 頭<seg type="okurigana">ヲ</seg> 思<seg type="okurigana">フ</seg><seg type="kaeriten">二</seg> 故郷<seg type="okurigana">ヲ</seg><seg type="kaeriten">一</seg> </l> </lg>

Table 1 illustrates how the fourth line of a Tang poem named "Thoughts in a tranquil night" by Li Bai[3] was marked up. As described before, the elements <head> and <persName> describe the title and author's name of the Tang poem. The element <lg> was used to mark up the contents of the Tang poem, and the attribute type described a fulltext expression with punctuated text. The element <l> indicates that the fourth line of the fulltext is marked up.

When a kanji is marked with *kunten* information in the content with either *"okurigana"* or *"kaeriten,"* the element <seg> is used to mark up the *kunten* and indicate the type of *kunten*. For example, the kanji 低 has explanation marks with both "*okurigana*" *rete* (レテ) and "*kaeriten*" *re* (レ).

This paper discussed special expressions in Tang poems and proposed an approach to creating an appropriate markup for Tang poems in Japanese. In future research, we will attempt to 1) find an appropriate way to represent the other annotations used in Tang poems, and 2) develop an application for learning Tang poems.

Acknowledgment:
This work was partially supported by JSPS KAKENHI Grant Number JP16H02913.

References:

[1] **The Text Encoding initiative**. "TEI: P5 Guidelines". http://www.teic.org/Guidelines/P5/, (accessed 2018-04-24).
[2] **Yan CONG; Masao TAKAKU**. "Prototype of Linked Open Data Model for Tang Poems". Japanese Association for Digital Humanities Conference 2017 (JADH2017), Kyoto, Japan, pp.50-52 (2017-09).
[3] **Junya Noji, et al.** 2015. "Cyu Gakko Kokugo3 (中学校国語 3) ", Gakko Tosho. Inc. p.176.

The Digital Curation Project- Popularization of Democracy in Post-War Japan – virtual reunification of dispersed materials hidden in the Hussey Papers Archival collection

Keiko Yokota-Carter[1]

This short paper introduces how the University of Michigan Library Online Exhibit *Popularization of Democracy in Post-War Japan* (University of Michigan Library, 2018) has evolved as a digital curation project, taking the "Linear and Goal-Oriented Approach" (Punzalan, 2014) in reconstructing dispersed materials into 'a small collection' within one archived collection, *The Alfred Rodman Hussey papers*.

The *Alfred Rodman Hussey Papers* (1945-1948) is the collection that Commander Hussey gathered during his work with the Government Section, Supreme Commander for the Allied Powers (SCAP), during the Allied occupation of Japan following the World War II, and later while he was in the Central Intelligence Agency. It contains 3,650 titled documents including: correspondence, memoranda, orders, reports, official and unofficial policy papers, drafts of legislation, other writings, slides, and audiotapes. Below we describe significant discoveries by the project.

One of the unique collections is a box of 16 color slides. The content tells the story of the birth of the New Constitution of Japan. Originally there were two boxes of slides, but the 'No.2 Box' only exits at University of Michigan Library. A project team was organized by the University of Michigan's Japanese Studies Librarian with a Gentō Media scholar, and a Japanese Studies graduate student. Three narration booklets "hidden" in the *Hussey Papers* were also found. The three pamphlets had been mentioned in *the Jigyō Hōkokusho* (Kenpō Fukyukai, 1947: 42-43). One of the pamphlets titled *Jinken Sengen* (32 episodes) by Hidezō Kondō was identified as the narration for 32 of the slides mentioned in the *Rōdō kyōiku tenrankai kankei shiryō* (Chuō Rōdō Gakuin, 1947: 20-1). The art style of Kondō found in his column in the Yomiuri Shinbun (Kondō, 1940) also matched the one used in the slides. The scholar determined that the narration with the slides was the one that had won the Promoting Constitution Contest held by the Constitution Popularization Society (Kenpō Fukyukai, 1947: 43). To secure semi-permanent preservation the slides were digitized and displayed as *Alfred Hussey Collection: Japan's Constitution Slides* in the library's Digital Collections and preserved in the Hydra/Fedora library repository system.

A graduate student and librarian worked on curating 'a digital small collection' for the library's Online Exhibit's Omeka Platform by reunifying the dispersed slides and the narration text in the *Hussey Papers*. It also includes as added value historical background, translated texts and links to related open access data held in various repositories such as the United Nation Treaty Collection, the National Archives of Japan, and the National Diet Library, and other institutions.

The original Online Exhibit was debuted at the European Association of Japanese Resource Specialist Annual Meeting and at Doshisha University (Yokota-Carter, 2018) in 2017 and at the Council of East Asia Libraries Annual Meeting in 2018. In collaboration with the Digital Design Team at the University of Michigan Library, we are now in the process of improving the interface design to increase the usability for diverse users, including the visually impaired, researchers and educators, as well as in general public.

Beagrie and Punzalan provided the theoretical framework for this digital curation project. Beagrie defines digital curation as including preservation, maintenance, management, and the future use of the digital data as well as "the capacity to add value to data to generate a new source of information and knowledge" (Beagrie, 2004: 7). This definition of "digital curation" corresponds with the Punzalan's concept of 'virtual

[1] University of Michigan Library

reunification' as a strategy to gather together dispersed archival materials "to a single origin or common provenance." (Punzalan, 2014).

Museums, libraries, and public/private institutions have been producing a massive amount of data by digitizing analog resources while 'born-digital' resources have emerged in volumes. This data has become a part of libraries and museums collections. Libraries have started the online exhibitions around themes by using digitized texts and images that are dispersed in various open access data archives, scattered among different institutions around the world, through web links. Digital curation has become a part of a library's broader collection development. As data is collected, reused and transformed for education and research, new knowledge emerges. Data as a collection provides materials for digital scholarship.

Another important idea addressed by this project is the concept of "virtual repatriation," a controversial topic among the libraries, archives, museums (LAM) community (Punzalan, 2014). The above mentioned slide set and narration text were brought to the United States from Japan as a result of the postwar occupation of Japan. We ask the question, "Can this project be considered a 'virtual repatriation' to the Japanese community, who originally produced these materials in their effort to promote the new Japanese Constitution?" Yokota-Carter's past presentations at meetings of the European Association of Japanese Resource Specialist and Council of East Asia Libraries on this digital curation project has given librarians in North America and Europe a model for planning new virtual reunification projects of Japan's World War II and postwar documents that are physically scattered around the world.

References

Beagrie, Neil. (2004). "The Digital Curation Centre." *Learned Publishing*, 17(1):7-9. https://onlinelibrary.wiley.com/doi/epdf/10.1087/095315104322710197 (Accessed 17 June, 2018)

Hussey, Alfred Rodman. *The Alfred Rodman Hussey Papers* (1945-1948). https://mirlyn.lib.umich.edu/Record/012858405 (Accessed 17 June, 2018)

Kenpō Fukyukai. (1947). *Jigyō gaiyō hōkokusho*. https://www.digital.archives.go.jp/das/image/F0000000000000331751 (Accessed 17 June, 2018)

Kondō, Hidezō. (1940). "'Nigiyakana mon' o tataku 'Kodakara butai' hōmon." *Yomiuri Shinbun Yomidas Rekishikan*, October 19, 1940. https://database.yomiuri.co.jp/rekishikan/ (Accessed 21, 2018)

Punzalan, Ricardo L. (2014). "Understanding virtual reunification." T*he Library Quarterly: Information, Community, Policy,* 84(3): 294-323. https://www.journals.uchicago.edu/doi/10.1086/676489 (Accessed 17 June, 2018)

Rōdō shō and Chuō Rōdō Gakuen (eds). (1947). *Rōdō kyōiku tenrankai kankei shiryō.* Chuō Rōdō Gakuen. http://dl.ndl.go.jp/info:ndljp/pid/1439500 (Accessed 17June, 2018)

University of Michigan Digital Collections. *Alfred Hussey Collection: Japan's Constitution Slides.* https://quod.lib.umich.edu/h/hussey1ic (Accessed 17June, 2018)

University of Michigan Library (2018). *Popularization of Democracy in Post-War Japan* (Original). https://deepblue.lib.umich.edu/handle/2027.42/144505 (Accessed 17 June, 2018)

Yokota-Carter, Keiko. (2018). "Popularization of Democracy in Post-War Japan Online Exhibit Project – Making History Alive Again." *Toshokangaku nenpō*, 43: 5-18. https://doors.doshisha.ac.jp/duar/repository/ir/26096/021000430002.pdf (Accessed 17 June, 2018)

Archive as Data: Reading *Kisho Shushi* to Follow Meteorology and the Boundary of the Empire in Meiji Japan

Ryuta Komaki[1]

As the *Santa Barbara Statement on Collection as Data* (Always Already Computational - Collection as Data Project Team, 2017) declares, "any digital material can potentially be made available as data that are amenable to computational use. Use and reuse is encouraged by openly licensed data in non-proprietary formats made accessible via a range of access mechanisms that are designed to meet specific community needs." This paper explores the concept of "collection as data" by engaging with an existing digital archive from Japan - the digitized collection of *Kisho Shushi* (*Journal of the Meteorological Society of Japan*), Series I - using (semi-)computational approaches. The collection was intentionally chosen to assess what it means to design "access mechanisms" that "meet specific community needs" - in this case the needs that pertain to an inquiry based disciplinarily on the history of science.

Meteorology has simultaneously been an international science and a science of the empire. While the former prioritizes collaboration, sharing and mobility, the latter prefers a dominance over - and domination by - knowledge. The coupling of meteorological observation and colonial ambitions of the British Empire is well documented by historians of science (Williamson, 2015; Mahony, 2016). Through this study, I intend to argue that, similar to British sciences, the development of modern meteorology in Meiji Japan was deeply tied to the nation's imperial mission, while at the same time being a science dependent on a cross-border exchange of ideas, personnel, goods and data.

Kisho Shushi is the official scientific journal of Tokyo Kisho Gakkai (Meteorological Society of Tokyo, later renamed Dai Nihon Kisho Gakkai, and again in 1941 to Nihon Kisho Gakkai). Series I of the journal was published during the first forty years of the society, from 1882 to 1883, and after a five-year hiatus, from 1888 to 1922. The scientific articles from the journal run are fully digitized and made available on J-STAGE (https://www.jstage.jst.go.jp/browse/jmsj1882/). Using this digitized archive, and reading it with an assistance of the computer as a record of activities of meteorologists in Meiji and Taisho Japan, the study aims to trace and visually map the movement of the European science of meteorology into Japan's center accompanied by its own network of weather stations, meteorologists and communication technology, which then moved on to Hokkaido, Taiwan and Korea following - and sometimes preceding - the expanding boundary of the Japanese Empire.

In *Africa as a Living Laboratory*, Tilley (2011) argues that "The layer of institutions established to meet the needs of the empire occupied an interstitial space that was neither national nor international" (9). In the case of meteorology in Meiji Japan, too, national, imperial and international were deeply intertwined. My findings show that the transfer of the European science of meteorology was aided by the international aspect of meteorology, which had standardized personnel training, equipment, measurement and sharing of reports, as well as established, while limited, network of observatories in East Asia. The standardization and the availability of data from existing observatories made it possible for Japanese meteorologists to quickly theorize (albeit insufficiently, particularly compared to the state of meteorology and climatology today) East Asia and Japan, and move their knowledge and procedures to the empire's new territories in Hokkaido, Taiwan and Korea when opportunities arose. At the same time, the geography of Japanese meteorology was inseparable from national and imperial interests. The data the first generation of Japanese meteorologists relied on to make sense of Japan and East Asia were those gathered from

[1] Washington University in St. Louis

observatories situated in ports and cities that were strategically important to European powers, and processed and distributed through calculation centers in the Metropoles and colonial hubs. And as the Japanese Empire and Japanese meteorology expanded, meteorological knowledge and meteorologists often followed, or was followed by, the interests of the empire.

Following the geography of the Japanese Empire and Japanese meteorology also suggests that the elements that made up meteorology had "differential mobilities" (Sheller and Urry, 2006) which at times became apparent as they intersected with the "mobility" of science (Livingstone, 2003), as well as with the moving boundary of the empire. Scientific data, personnel and equipment each required different infrastructure and protection to move around. The geography of knowledge and the geography of power overlap; however, in the process of achieving that overlap, differential mobilities created many disruptions and disconnections.

My engagement with the digitized archive of *Kisho Shushi* also tested the concept of "collection as data" and teased out challenges and limitations of using an archive of "scientific communication" for a social scientific/humanistic inquiry. The archive in the digital and open-access format offer several opportunities, particularly the ease of access and the opportunity to use optical character recognition (OCR) technologies to convert text to data. The latter opens up possibilities for scholars to apply computerized and computational methods to study its content. The way the digitized archive of journal run is currently provided, however, also poses challenges to scholars approaching the archive with digital methods and social scientific/humanistic questions. These challenges include the quality of digital scan, the OCR's ability to read Meiji texts, as well as the omission of non-scientific communications. The quality of scanned pages seems to vary depending on the quality of original issues (both the type of the paper and typefaces used), how well they have been preserved, and when the scans were made and added to the digital archive. Tesseract open source OCR engine (https://github.com/tesseract-ocr/tesseract), not specifically trained, returned OCRed text with 40-50% accuracy per page, with most errors stemming from katakana characters. (Early *Kisho Shushi* articles, as with many Meiji documents, employed katakana where hiragana is normally used in modern writing). The digitized archive on J-STAGE also does not include most of "miscellaneous" pages present in the original issues. Omitted pages include announcements of members' new appointments, transfers and retirements, which may not be essential to trace scientific discourses, but are still integral to answering questions regarding the network and mobility of early Japanese and colonial meteorologists.

References

Always Already Computational - Collection as Data Project Team (2017). *Santa Barbara Statement on Collection as Data.* https://collectionsasdata.github.io/statement/ (accessed June 23, 2018).

Livingstone, D. N. (2003). *Putting Science in Its Place: Geographies of Scientific Knowledge.* Chicago: University of Chicago Press.

Mahony, M. (2016). "For an empire of 'all types of climate': meteorology as an imperial science." *Journal of Historical Geography*, 51: 29-39.

Sheller, M. and Urry, J. (2006). "The new mobilities paradigm." *Environment and Planning* A, 38: 207-226.

Tilley, H. (2011). *Africa as a Living Laboratory: Empire, Development, and the Problem of Scientific Knowledge, 1870-1950.* Chicago: University of Chicago Press.

Williamson, F. (2015). "Weathering the empire: meteorological research in the early British Straits Settlements." *British Journal for the History of Science,* 48(3): 475-492.

[Panel Session 2]

Broadening Perspectives of Historical Researchers: From a Case of Interdisciplinary Workshop organized by Graduate Students in Japan

Satoru Nakamura[1], Masato Fukuda[1], Jun Ogawa[1], Sho Makino[1], Ayano Sanno[2], Shohei Yamasaki[1]

In recent years, useful information for various fields of historical research can be easily obtained on the web. Along with this trend, practical implementation of digital history, which is history research applying digital technology, has been required. This necessity led to the launch of an interdisciplinary workshop "Tokyo Digital History (ToDH for short)" organized by mainly graduate students majoring historical research, archivists, and engineers.

There are several reasons why the young researchers, mainly graduate students, take the lead in practice of digital history. The reasons are as follows.

1: While interest in Digital Humanities has been increasing, there are still few movements related to Digital Humanities in the field of historical research, and it is necessary to increase such practical examples.
2: Enhancement of the minor education program including the University of Tokyo has led to an increase in the number of historical researchers of graduate students who practice Digital Humanities.
3: Young researchers need their strength required for career development in response to the recent trend which humanities are at stake.
4: They assume becoming a position to train researchers with advanced information literacy accompanying mandatory programming education.

Regarding the career development of humanities researchers in particular, it is now necessary to think about how researchers themselves can evaluate their research contents. While the evaluation of historical research applying digital technology has been discussed, it seems relatively weak in Japan. Therefore, the main purpose of this workshop is to explore how young researcher's own practice can guide the evaluation criteria and to consider what it can have meaning to career development.

Graduate students gathered from the background mentioned above prepared a fixed time and place every week and held seminars for mastery of digital technology and discussion on participants' research contents. As a seminar, ToDH have conducted study sessions for Python and TEI several times. Furthermore, ToDH held a symposium on April 15, 2018, and nearly 90 participants gathered at the venue, 10 people participating with video conference systems from the UK, Germany and France. The movement of ToDH is a good example of a humanities and information science project that requires interdisciplinary collaboration, and it is one of epoch-making attempts in the field of historical research in Japan.

In this panel, 5 students present practical examples of each research theme, and mention what kind of things they learned through ToDH activities and how it helped his own research, including changes in analysis viewing angle.

Masato Fukuda will mention the web scraping technology to obtain data for research, and discuss the visualization of hierarchical structure of a historical source and the metrological analysis with the availability of acquired data. To be precise, he conducted web scraping

[1] The University of Tokyo
[2] Ochanomizu University

from the digital archive of National Archives of Japan, in order to obtain metadata of about 110,000 historical materials for the government historical materials in the Meiji Era. By using the huge amount of acquired data, it enabled him to conduct cross-governmental / temporal analysis.

Jun Ogawa, who specializes in history of ancient Roman provinces, analyzed Latin text. He acquired the TEI / XML file of Caesar's “Commentarii de Bello Gallico” using the API provided by Perseus Digital Library. He visualized text information in the form of a frequency table and co-occurrence network, and interpreted text from a viewpoint that could not be obtained only by character information. This kind of analysis and careful reading of historical materials will make it possible to understand historical materials deeper and more diversely than ever before.

Sho Makino focuses on Irish aspects of the English Revolution in the 1640s. He uses British History Online and 1641 Depositions: especially the Depositions is one of the most important sources for the 21st century Irish historians. Scrutinizing the gaps between the real state of affairs of the 1641 rebellion and discussions of the parliament in London, he aims to understand the Wars of the Three Kingdoms throughout its time, since the historical narrative tends to be fragmented in political occasions.

Ayano Sanno aims to deepen prosopographical research on Index biographique de l'Académie des sciences in the 18th century Paris. She considers the methodology of structuring the description of various membership information in compliance with TEI and refers to the possibility of contributing to building the foundation of related research.
For example, she will present some prospects of the collaboration with the Japanese research group focusing on Encyclopédie and Enlightenment in 18th century France.

Shohei Yamasaki shows the usefulness of batch processing using a programming language in data cleansing and processing as a case example of correction of prefectural boundary change in Japan in the Meiji Era. Specifically, he shows that writing and publishing he work procedure as code of the programming language and batch processing are useful not only for saving task effort but also for increasing reproducibility of research by third parties.

As a summary of this panel, **Satoru Nakamura** describes the prospect of how ToDH activities can contribute to the development of Digital Humanities in Japan. While he is one of ToDH members, he practices Digital Humanities from an informatical standpoint, such as building digital archives at the University of Tokyo library. Furthermore, he will review current status and issues of Japan on research evaluation with examples such as data journals and TAPAS project.

Collaborative approaches to implement Science as a service in an Open Innovation in Science framework: Japanese Diaspora Studies on the example of Thomas Higa

Yoshiyuki Asahi[1], Eveline Wandl-Vogt[2], Jose Luis Preza Diaz[2]

Movement across borders is increasingly exponentially and taking new forms of impact directly on a nation's traditional sense of itself in our fast changing global arena. In this paper, the team introduces a concept for an innovative knowledge system, designed to increase open collaboration between various actors in global society and improve understanding of the influence of human journeys and displacement of people across borders.

In the approach we apply in the project and describe here, methods and practices from Open Innovation applied to Science become inherent to the creative and innovation driven research process. While originally created within the realm of Business Management, the principles of Open Innovation are expanding to a broad range of academic fields (Chesbrough 2003). In our work we go with Bogers and Chesbrough 2014, defining Open Innovation as a distributed innovation process based on purposively managed knowledge flows across organizational boundaries, using pecuniary and non-pecuniary mechanisms.

In the process described, we focus on the application of Lead user experiments in a global scale. We introduce lead user method, who are the experts connected to our work and how we go for building the framework and linking people. The Japanese diaspora studies serve as an example, embedded into international initiatives such as the Dariah-EU working group "Analysing and linking biographical data" as well as the UNESCO group on "Human journeys in the global Era" (application process ongoing).

The design is technically introduced by implementing Science as a services infrastructure on the example of discovering innovative diaspora studies and biographical narratives exemplified on Thomas Taro Higa.

The project is designed as a collaboration project between NINJAL, Microsoft Research and ACDH-ÖAW and is embedded into the project "NIHU International Collaborative Project on Japan-related materials overseas". The NINJAL project started in 2010 and it investigated Nikkei-related materials created through the Japanese American history in the US since 1960s. Our target collection was magnetic audio/audio-visual tapes as well as photos and documents and we have digitized them through the cooperation from the local institutions in Japan. Taro Higa, a second Okinawa-born Japanese American, was one of the most active figures based both in Hawaii and west coast of the US, commit himself to better understand the history of Japanese diaspora primarily in the US. During his life, he has written a series of newspaper articles and academic papers in Hawaii, and he was interviewed by a number of historian on Japanese American studies. Currently, we have a large amount of unstructured information available in various formats, such as audio, visual and pictures (jpg), which is going to be made (openly) available via a proposograpical information system. The infrastructure developed in the project framework of APIS (Schlögl and Lejtovicz 2017) at ACDH, is going to be tested and applied for the application.

This poster aims to introduce the management of knowledge (flows) within our project, including workflow design and (technical) architecture as well as issues of data licences. It aims to analyse and give a deep understanding on the social processes taking place when connecting globally, cross-sectoral and cross-cultural.

In doing so, we aim to offer a case study on the cultural change and (social) challenges leveraging open data in a global collaborative setting implicates beyond pure software

[1] National Institute for Japanese Language and Linguistics
[2] Austrian Academy of Sciences

development and give a brief example on the possible “collaborative turn” (Spiegel 2015) open innovation in science might stimulate or reply to.

References:

Henry Chesbrough (2003) Open Innovation: The New Imperative for Creating and Profiting from Technology. Boston, MA: Harvard Business School Press.

Henry Chesbrough, Marcel Bogers (2014) Explicating Open Innovation : Clarifying an Emerging Paradigm for Understanding Innovation. New Frontiers in Open Innovation. ed. / Henry Chesbrough; Wim Vanhaverbeke; Joel West. Oxford University Press, 2014. p. 3-28.

Ian T. Foster, Ravi K. Madduri (2013) Science as a service: how on-demand computing can accelerate discovery. In: Science Cloud '13 Proceedings of the 4th ACM workshop on Scientific cloud computing Pages 1-2. doi>10.1145/2465848.2480345.

Alisa Goikhman, Roberto Theron, Eveline Wandl-Vogt (2016) Designing Collaborations: Could Design Probes Contribute to Better Communication Between Collaborators? In: Conference: Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality. At: Salamanca, Spain. doi>10.1145/3012430.3012431.

Matthias Schlögl, Katalin Lejtovicz (2017) A Prosopographical Information System (APIS). Eveline Wandl-Vogt and Lejtovicz, Katalin. In: Eveline Wandl-Vogt, Katalin Lejtovicz (Ed.) Biographical Data in a Digital World 2017. A conference in the framework of the project APIS, 6–7 November 2017. Abstracts. [Wien].

Peter Spiegel (2015) WeQ more than IQ. oekom.

Yoshiyuki Asahi (2013) Kibei's ways of speaking three languages, Japanese, Ryukyuan and English: Evidence from Thomas Taro Higa. A paper read at 112 American Anthropological Association conference.

Yoshiyuki Asahi (2017) Detecting and mining biographical data from audio/audio-visual magnetic tapes: A case of the Japanese American collections in the US. A paper read at Biographical Data in Digital World conference 2017.

Philograph: Textual Analysis Tools in the Digital Humanities

Jerry Bonnell[1]

The genesis of this project is a curiosity about new paths of scholarship in the Digital Humanities, specifically in the subfield of text analysis. Literary historian Franco Moretti, one of today's standard bearers of the discipline, passionately advocates for a rethinking of our involvement with and study of texts. His purpose: to create a long content-timeline of literature – the **longue durée** – to distinguish, exclusively, similarities and/or differences across genres and centuries of published texts (Moretti, 2013: 85). While the results have been promising and offer new insight into literary analysis, less interest has been shown at the micro-level. This is to say: our understanding of the machine's capabilities in reading individual texts – not libraries – is significantly limited. The questions framed in this project propose to enlighten us about the machine's capabilities on this level, i.e., can the machine assist the individual scholar with the task of textual analysis? Furthermore, can the machine operate in tandem with methodologies that have long been dominant in the discipline of literary studies, i.e., close reading?

This project assembles 20 sermons from the archives of the 18th-century minister Jonathan Edwards. These are at the Jonathan Edwards Center at Yale University and the Jonathan Edwards Collection maintained by the Bible Bulletin Board. The metaphors of language that breathe life into religious texts, as well as their manageable lengths, were the top factors that led to their selection. A close reading of the individual sermons was conducted with special attention paid to keywords and themes – such as theology, forgiveness, saints, sinners, and community – that framed the message and purpose of the texts. To validate the individual interpretation of the source material, established scholarly authorities in history and literature were consulted. Finally, two technologies from the field of Machine Learning were applied: (1) k-means, an unsupervised learning technique, with the purpose of identifying similar structural content among the corpus, and (2) Support Vector Machines (SVM), a supervised learning algorithm, to evaluate if the machine can output the appropriate category of the sermons. Because these technologies operate at maximum efficiency when learning from large datasets, the sermons were split into segments and a sliding window – the number of overlapped lines between these segments – was used for contextualization. This process was the cornerstone for the insightful results generated.

In its first assessment, the machine was tasked with clustering the collection of sermons into groups using k-means. By analyzing the words characteristic of each cluster, we can visualize the structure of language used by Jonathan Edwards in his writing. This is to say, Edwards' understanding of the duality of religion is manifest in the oscillation of his language between a celebration of God and a condemnation of man. It would be time-consuming to arrive at this conclusion without aid from the machine's clustering of words.

In its second assessment, the machine was tasked with using the SVM model to classify the sermons in the corpus by theme. While it is noteworthy to report the high accuracy rate of the classifier in its testing, it is fascinating to reflect instead on its few failures. Most notably, its erring in labeling the parts of the theology sermon *Christian Happiness* with forgiveness, and the parts of the forgiveness sermon *The Value of Salvation* with theology. Its misclassifications can be interpreted as the machine's advice to the scholar that the categories of forgiveness and theology are not mutually exclusive. It may be wise then to consider the intersection between them. In so doing, the machine invites an interpretation of how Jonathan Edwards understood the interaction between the constituent themes of Puritanism; in this study, the connection between the rules of Puritanism (theology) and its promise for redemption (forgiveness). The machine's inability

[1] University of Miami

to discriminate between these are not errors, but links in Edwards' cosmology: connections that may escape the eyes of a scholar.

Indeed, the machine is not the author of these conclusions. It is understood that the interpretation of the data can only happen with an academic understanding of colonial New England and Jonathan Edwards. Nevertheless, the lessons noted here support the machine's capability in outputting data that directs a scholar to these conclusions. In so doing, the machine's partnership gifts the scholar new lenses with which to read the sources. He, then, has the practical advantage of connecting with many more sources, especially when the desire is to paint a landscape of an individual or theme. This conclusion does not discredit Moretti's vision of the **longue durée** but suggests the possibility of the democratization of the use of computation in the humanities (Moretti, 2013: 85). The data from Jonathan Edwards' sermons, and their analyses, positions the machine not as a parting point to new methodologies and conclusions (like Moretti), but as a partner and friend in research. Its success in generating the data is confidence that the future of the Digital Humanities rests not on extremes, but in its dissemination for use to all scholars.

References

Moretti, F. (2013). *Distant Reading.* London ; New York: Verso.

Representing digital humanities collections: A preliminary analysis of descriptive schema

Katrina Fenlon[1], Jacob Jett[2], J. Stephen Downie[2]

The *collection* is a familiar form of production in the digital humanities. Among emergent genres of digital scholarship, the collection is one of the most commonly recognized (Fenlon, 2017; Flanders, 2014; Palmer, 2004; Unsworth, 2000). Digital humanities collections are created as scholars select and gather thematically related digital primary sources and related materials and publish them online to support research and learning. Well known exemplars include the *Walt Whitman Archive*, the *Valley of the Shadow Archive*, and the *Dickinson Electronic Archive*. But even beyond these large-scale, long-running initiatives, there are hundreds of digital humanities collections on the web.

Despite their proliferation and significance, complex digital humanities projects – such as these collections – tend to rise and decline rapidly and invisibly on the web, compromising the integrity of both the cultural and the scholarly records. Unlike books, journal articles, and other genres of digital scholarship, collections are rarely preserved in library collections; they are rarely discoverable in indexes or directories; and they are rarely subject to formal evaluation.

Digital humanities collections are disadvantaged in part because we lack common data models for representing and describing them. Most digital humanities collections – though they may work by similar structural logic – are built idiosyncratically. Even where they employ standards for the representation of *items* within the collection – such as TEI-XML for representing texts – no such standardization exists for the representation of collections as wholes.

As a starting point for addressing this question, this paper offers a preliminary analysis of three extant collection-description schemas from other domains to assess their adequacy for representing digital humanities collections: the Dublin Core Collections Application Profile (DC-CAP); the Europeana Data Model Collection Profile (EDM-CP), and the HathiTrust Research Center (HTRC) Workset Ontology (Jett et al., 2016). Each of these three schemas was developed to represent and describe collections in different contexts. Together they may offer a foundation for the representation of digital humanities collections.

Through standardized description and representation, collections may become more deeply useful to a wider variety of researchers: to digital humanities scholars working across disciplinary boundaries, to scholars seeking to discover and reuse open data from different domains, and to scholars seeking to forge links between related resources on the web.

The most prominent schemes for the description of collections come from libraries and cultural heritage institutions. The DC-CAP offers a set of metadata terms for describing collections, which it defines as aggregations of physical or digital resources of any type. The EDM-CP enables the description of collections within the Europeana Data Model, which underlies the massive Europeana cultural heritage aggregation. EDM-CP draws heavily on the DC-CAP but adds properties to support Europeana-specific functionality.

We can also look to ontologies, useful for representing collections as open data. The HTRC Workset Ontology supports the representation of one specific type of collection in one specific context: within the HTRC computational environment for doing research using texts from the HathiTrust Digital Library. The HTRC Workset Ontology describes a hierarchy of classes of collection-types (see Figure 1). It aims to represent collections at the lowest level of this hierarchy, collections that are specifically intended for computational analysis within a non-consumptive research paradigm. Most digital humanities collections, in contrast, might be understood as inhabiting the level above the

[1] University of Maryland
[2] University of Illinois at Urbana-Champaign

workset in this model – as *research collections*, demonstrating curatorial selection and a specific research motivation.

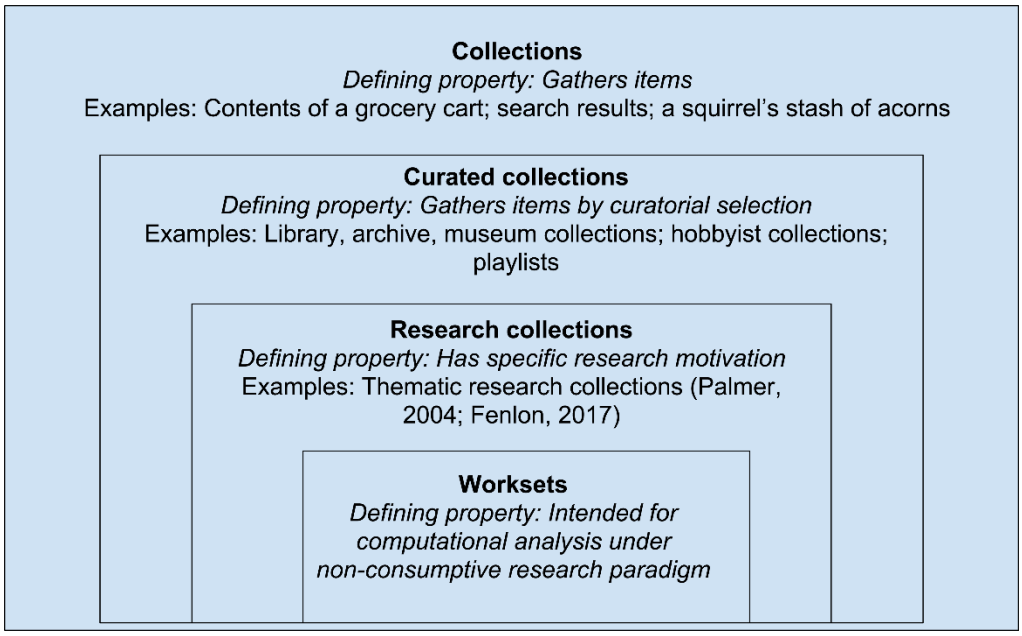


Figure 1: Hierarchy of classes of collections in HTRC Workset Ontology

Prior work (Fenlon, 2017) identified a set of properties that are common to digital humanities collections, based on a literature survey. These properties are listed in Table 1.

Table 1: Properties of digital humanities collections, organized into three related clusters

Cluster	Categories of analysis
Context	Theme; Purposes; Impact; Creators; Audience; Documentation; Provenance; Related collections; Related projects and publications; Review; Funding; Developmental stage; Host; Rights; Sustainability and preservation plans; Method
Content	Items; Diversity; Size; Narrativity; Quality; Language; Completeness; Density; Spatial coverage; Temporal coverage; Interrelatedness
Design	Data models; Navigation; Infrastructural components; Interface design; Interactivity; Interoperability; Openness; Identification and citation; Modes of access and acquisition; Accessibility; Flexibility

Table 2 offers a snapshot of our preliminary attempt to map the identified properties of digital humanities collections into the available collection-description schema. Here we show just three properties, all of which are essential to the representation of collections:

- Theme: What a digital humanities collection is about;
- Purposes: The intended purposes of a collection; and
- Completeness: The ideal of completeness toward which a collection is being developed – e.g. does a collection aim to be a comprehensive, definitive source on a subject, or does it aim to gather just enough evidence to answer a specific research question?

Table 2. Snapshot of preliminary mapping of properties to extant description schemas

Property	Potential mapping to…			
	DC-CAP	**HTRC-WO**	**EDM-CP**	**Assessment**
Theme	Subject [dc:subject] ; Spatial Coverage [dcterms:spatial] ; Temporal Coverage [dcterms:temporal]	see DC-CAP	see DC-CAP	Elements are not sufficient to describe Theme, which must be characterized in more complex ways than by repeated, discrete subject/coverage terms.
Purposes	N/A	Research motivation [htrc:hasResearchMotivation]	N/A	Element may not be sufficient to describe Purposes.
Completeness	N/A	N/A	N/A	There are no elements to describe Completeness.

This poster will give an extended version of this table, representing a more complete set relevant processes and articulating analytic processes.

Our preliminary analysis of extant collection-description schemas suggests that they are not adequate to represent even a few essential properties of digital humanities collections, not to mention the full complexity and range of collection information. This paper is intended to lay groundwork for developing frameworks for representing collections and digital humanities projects more generally, with the ultimate goal of increasing the discoverability, use, share-ability, and sustainability of digital humanities scholarship.

References:

Fenlon, K. (2017). *Thematic research collections: Libraries and the evolution of alternative scholarly publishing in the humanities* (Doctoral dissertation). University of Illinois at Urbana-Champaign.

Flanders, J. (2014). “Rethinking Collections.” In Arthur, P. and Bode, K. (eds), *Advancing Digital Humanities.* Palgrave Macmillan UK, pp. 163-74.

Jett, J., Cole, T., Maden, C., and Downie, J. (2016). “The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections.” *Journal of Open Humanities Data* 2(0).

Palmer, C. L. (2004). “Thematic Research Collections.” In Schreibman, S., Siemens, R., and Unsworth, J. (eds), *A Companion to Digital Humanities.* Blackwell.

Unsworth, J. (2000). “Thematic Research Collections.” Modern Languages Association Annual Conference, Washington, D.C.

entity-fishing: a DARIAH entity recognition and disambiguation service

Luca Foppiano[1], Laurent Romary[1]

This paper presents an attempt to provide a generic named-entity recognition and disambiguation module (NERD) called entity-fishing as a stable online service that demonstrates the possible delivery of sustainable technical services within DARIAH, the European digital research infrastructure for the arts and humanities. Deployed as part of the national infrastructure Huma-Num in France, this service provides an efficient state-of-the-art implementation coupled with standardised interfaces allowing an easy deployment on a variety of potential digital humanities contexts. The topics of accessibility and sustainability have been long discussed in the attempt of providing some best practices in the widely fragmented ecosystem of the DARIAH research infrastructure.

The history of entity-fishing has been mentioned as an example of good practice: initially developed in the context of the FP9 CENDARI (Lopez et al., 2014), the project was well received by the user community and continued to be further developed within the H2020 HIRMEOS project where several open access publishers have integrated the service to their collections of published monographs as a means to enhance retrieval and access. entity-fishing implements entity extraction as well as disambiguation against Wikipedia and Wikidata entries.

The service is accessible through a REST API which allows easier and seamless integration, language independent and stable convention and a widely used service oriented architecture (SOA) design. Input and output data are carried out over a query data model with a defined structure providing flexibility to support the processing of partially annotated text or the repartition of text over several queries. The interface implements a variety of functionalities, like language recognition (Nakatani, 2010), sentence segmentation and modules for accessing and looking up concepts in the knowledge base. The API itself integrates more advanced contextual parametrisation or ranked outputs, allowing for the resilient integration in various possible use cases.

The entity-fishing API has been used as a concrete use case to draft the experimental stand-off proposal (Banski et al., 2016), which has been submitted for integration into the TEI guidelines. The representation is also compliant with the Web Annotation Data Model (WADM). In this paper we aim at describing the functionalities of the service as a reference contribution to the subject of web-based NERD services.

In order to cover all aspects, the architecture is structured to provide two complementary viewpoints. First, we discuss the system from the data angle, detailing the workflow from input to output and unpacking each building box in the processing flow. Secondly, with a more academic approach, we provide a transversal schema of the different components taking into account non-functional requirements in order to facilitate the discovery of bottlenecks, hotspots and weaknesses. The attempt here is to give a description of the tool and, at the same time, a technical software engineering analysis which will help the reader to understand our choice for the resources allocated in the infrastructure.

Thanks to the work of million of volunteers, Wikipedia has reached today stability and completeness that leave no usable alternatives on the market (considering also the licence aspect). The launch of Wikidata in 2010 have completed the picture with a complementary language independent meta-model which is becoming the scientific reference for many disciplines. After providing an introduction to Wikipedia and Wikidata, we describe the knowledge base: the data organisation, the entity-fishing process to exploit it and the way it is built from nightly dumps using an offline process.

We conclude the paper by presenting our solution for the service deployment: how and which the resources where allocated. The service has been in production since Q3 of 2017,

[1] ALMAnaCH, Inria

and extensively used by the H2020 HIRMEOS partners during the integration with the publishing platforms. We believe we have strived to provide the best performances with the minimum amount of resources. Thanks to the Huma-num infrastructure we still have the possibility to scale up the infrastructure as needed, for example to support an increase of demand or temporary needs to process huge backlog of documents. On the long term, thanks to this sustainable environment, we are planning to keep delivering the service far beyond the end of the H2020 HIRMEOS project.

References

Banski, P., Gaiffe, B., Lopez, P., Meoni, S., Romary, L., Schmidt, T., Stadler, P. and Witt, A. (2016). *Wake up, StandOff!.* https://hal.inria.fr/hal-01374102.

Lopez, P., Meyer, A. and Romary, L. (2014). *CENDARI Virtual Research Environment & Named Entity Recognition Techniques.* Einstein-Zirkel Digital Humanities https://hal.inria.fr/hal-01577975.

Nakatani, S. (2010). *Language Detection Library for Java.* https://github.com/shuyo/language-detection.

Collocation Patterns of Pitch-Class Sets: Comparing Mozart's Symphonies and String Quartets.

Michiru Hirano[1], Hilofumi Yamamoto[1]

1 Introduction

The constructional differences of the string sections between symphonies and string quartets are unknown. Although the string sections are identical for both symphonies and string quartets, consisting of two violins, a viola, and a bass part, they differ in terms of player number: more than one player plays each string part for a symphony, while there is only one player per part in the case of string quartets. Symphonies and string quartets also differ in that the former may include wind and percussion sections in addition to a string section, while the latter only consists of the four string parts (Figure. 1). The features which reflect these differences may be found on the score.

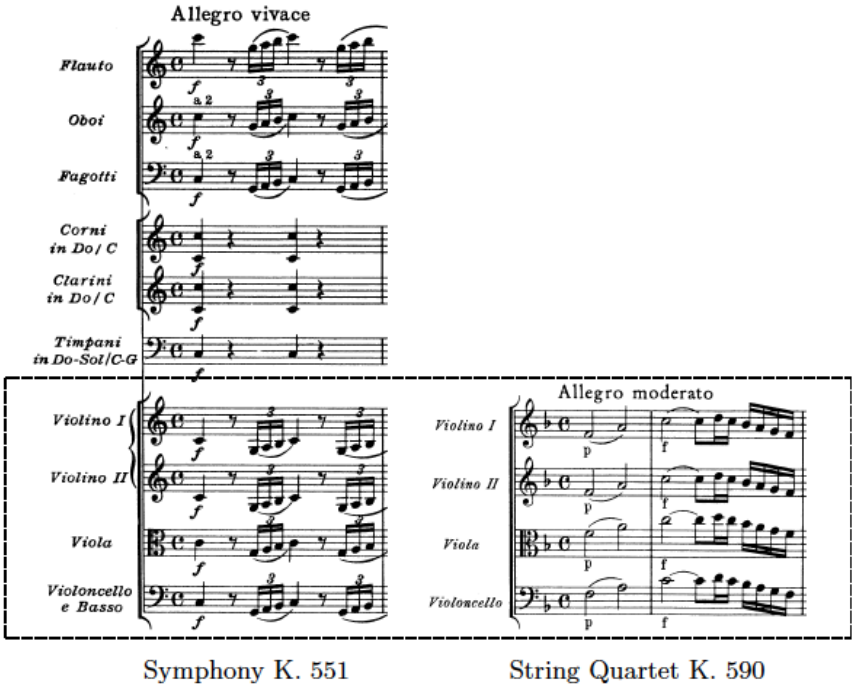


Fig. 1 Examples of a symphony and a string quartet: The beginning of the scores for Symphony K. 551 (left) and String Quartet K. 590 (right) composed by W. A. Mozart

We aim to elucidate the different constructions of the string sections in Mozart's symphonies and string quartets by examining the collocation patterns of Pitch-Class sets (PC sets). More specifically, we investigate which patterns frequently occur and are statistically significant. It has been demonstrated that the frequencies of particular PC sets vary for the string sections of Mozart's symphonies and string quartets (Hirano and Yamamoto, 2017). The collocation patterns of PC sets, especially in terms of bi-gram frequencies, is assumed to reflect more detailed harmonic features, given that classical composers regarded harmony progressions as important. Wolfgang Amadeus Mozart (1756-1791), a typical classical composer of 18th century, composed at least 39 symphonies and 23 string quartets during his life.

2 Methods

2.1 Materials

We target Mozart's 62 initial movements, of which 39 are for symphonies and 23 are for string quartets. The scores were converted into MusicXML, which is a machine readable format.

[1] Tokyo Institute of Technology

2.2 Data structure

In particular, we utilize the notion of a PC set, referring to the set of distinct integers that represent pitch classes (Forte, 1973: 3). Thus, we extract all the pitch classes located across four string parts of a measure and convert them into PC sets as summarized patterns, using computer programs of our own making (Figure 2).

The procedure for determining the PC sets for each measure of a work are as the followings: 1) each pitch corresponding to one of 12 distinct pitch classes (i.e. 12 steps C/B$^{\sharp}$, C$^{\sharp}$/D$^{\flat}$, D, ..., A$^{\sharp}$/B$^{\flat}$, B/C$^{\flat}$) is replaced by an integer from 0 to 11; 2) all pitches across the four string parts that appear within a measure are combined to form a PC set, with repetitions eliminated; 3) the PC sets are then sorted by "normal order" based on the minimum differences determined by subtracting the first value from the last; and 4) the PC sets are transposed on to a "prime form" (Forte, 1973: 3), where the first integer is 0. Through this procedure, a measure containing a major triad (such as a chord constructed from C, E and G) would, for instance, be represented with the notation of {0, 4, 7}, regardless of its root pitch.

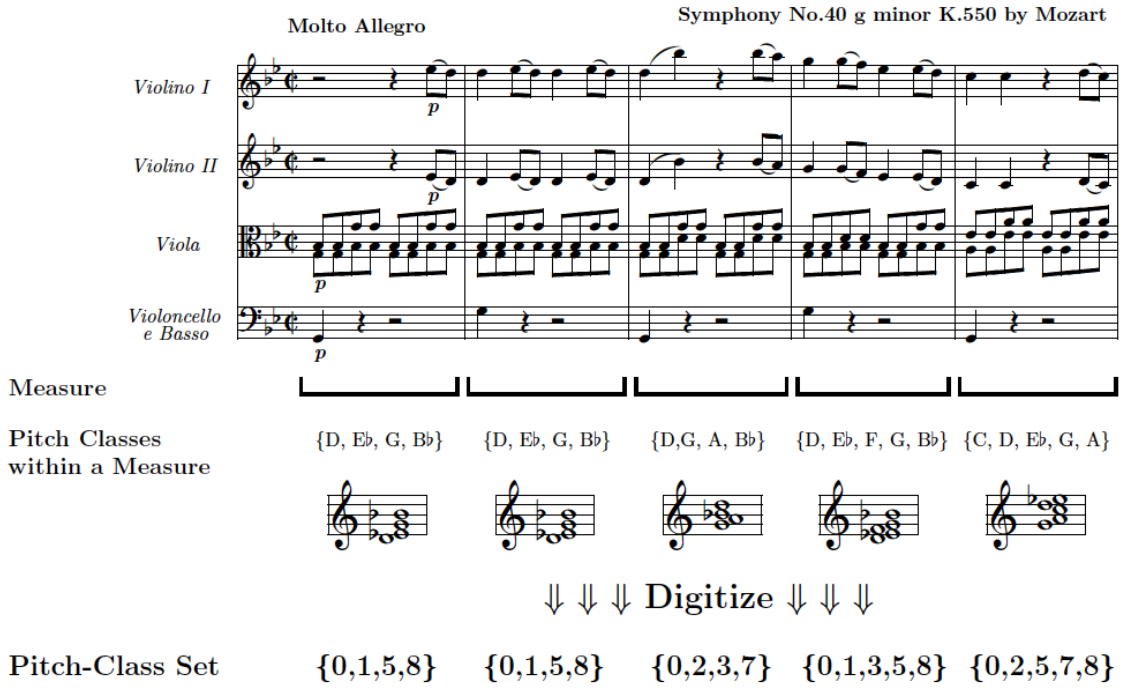


Fig. 2 An illustration of molding PC sets, where pitch classes across the four string parts of a measure are digitized according to a predefined rule

3 Results

First, we computed the frequencies of individual PC sets throughout the compositions and labeled the PC sets according to their ranking, in the form of P^{i} where ^{i} is its ranking (i.e. {0, 1, 3, 5, 6, 8, 10} P1, {0, 4, 7} P2, etc.). The top 10 PC sets are shown in Table 1.

Next, we computed bi-gram frequencies among the top 5 PC sets within the symphonies and the string quartets, respectively (Table 2 and 3). The underlined values in bold letters within Table 2 and 3 are significantly large compared with the other according to chi-square test and residual analysis (Haberman, 1973).

Figure 3 is network model for the significant sequences of PC sets across the symphonies (solid lines) and the string quartets (dashed lines).

Table 1 Top 10 PC sets according to frequencies within the 62 compositions

Rank	Frequency	PC set	An Example of Contents
P1	1,368	{0, 1, 3, 5, 6, 8, 10}	C, D, E, F, G, A, B
P2	1,110	{0, 4, 7}	C, E, G
P3	889	{0, 1, 3, 5, 6, 8}	C, D, E, F, G, B
P4	420	{0, 3, 6, 8}	G, B, D, F
P5	379	{0, 2, 4, 5, 7}	C, D, E, F, G
P6	342	{0, 1, 3, 5, 8}	C, D, E, G, B
P7	313	{0, 2, 4, 5, 7, 9}	C, D, E, F, G, A
P8	224	{0, 2, 4, 6, 7, 9}	C, D, F, G, A, B
P9	216	{0, 3, 7}	A, C, E
P10	205	{0}	C

Table 2 The bi-gram frequencies among the top 5 PC sets within the symphonies

From \ To	P1	P2	P3	P4	P5
P1	240	**107**	63	10	18
(EF)	(325.0)	(91.1)	(84.4)	(17.4)	(22.7)
ASR	−9.2	2.9	−4.1	−3.1	−1.7
P2	46	**253**	**62**	70	24
(EF)	(44.9)	(196.3)	(50.2)	(67.6)	(20.1)
ASR	0.2	7.5	2.9	0.5	1.5
P3	63	**84**	147	10	36
(EF)	(76.3)	(74.3)	(143.4)	(12.7)	(35.5)
ASR	−2.7	1.9	0.5	−1.3	0.1
P4	18	62	19	33	14
(EF)	(16.0)	(60.3)	(18.0)	(32.8)	(11.3)
ASR	0.8	0.3	0.3	0.0	1.3
P5	31	**20**	29	11	**62**
(EF)	(28.8)	(14.7)	(32.8)	(8.0)	(46.2)
ASR	0.7	2.3	−1.1	1.8	4.0

Note: EF = expected frequency, ASR = adjusted standardized residual.

Table 3 The bi-gram frequencies among the top 5 PC sets within the string quartets

From \ To	P1	P2	P3	P4	P5
P1	**245**	29	**63**	**16**	16
(EF)	(159.9)	(44.8)	(41.5)	(8.5)	(11.2)
ASR	9.2	−2.9	4.1	3.1	1.7
P2	21	40	13	31	6
(EF)	(22.0)	(96.6)	(24.7)	(33.3)	(9.8)
ASR	−0.2	−7.5	−2.9	−0.5	−1.5
P3	**51**	27	67	9	17
(EF)	(37.6)	(36.6)	(70.5)	(6.2)	(17.4)
ASR	2.7	−1.9	−0.5	1.3	−0.1
P4	6	28	8	16	3
(EF)	(7.9)	(29.6)	(8.9)	(16.1)	(5.6)
ASR	−0.8	−0.3	−0.3	−0.0	−1.3
P5	12	2	20	1	7
(EF)	(14.1)	(7.2)	(16.1)	(3.9)	(22.7)
ASR	−0.7	−2.3	1.1	−1.8	−4.0

Note: EF = expected frequency, ASR = adjusted standardized residual.

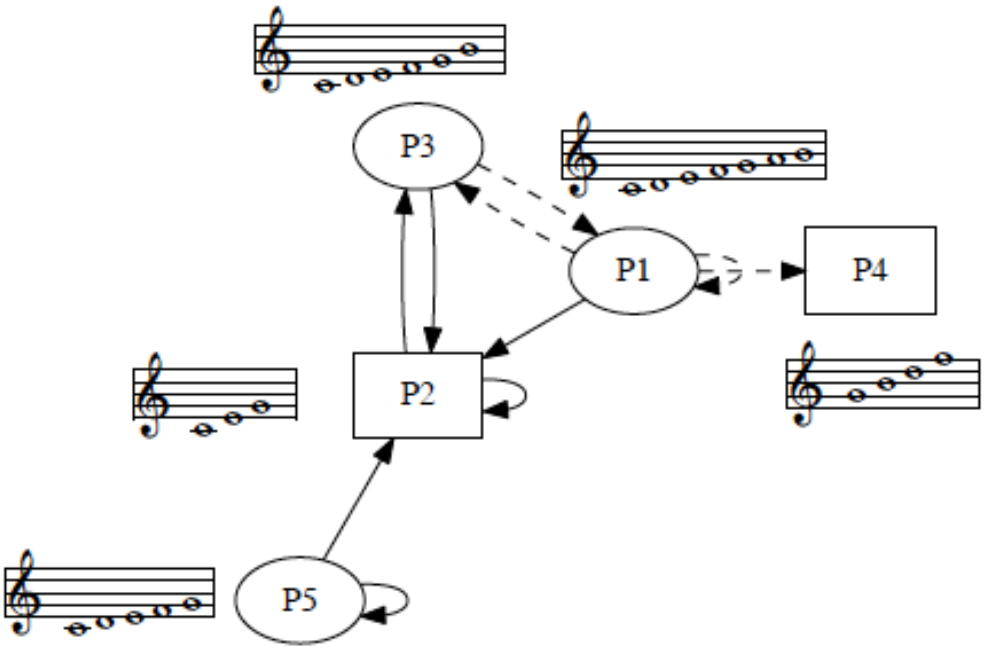


Fig. 3 Network model for the significant sequences of PC sets across the symphonies (solid lines) and the string quartets (dashed lines) with an example of contents of each PC set beside the correspond node

4 Discussions

Among the top 5 PC sets, P1, P3, and P5 consist of at least five components whose arrangements correspond to the whole or part of diatonic scale, while P2 and P4 have three or four components with intervals of skips. The notable points are: P2 and P4 correspond to the components of a major triad, a chord of the root note with a major third and a perfect fifth above, and a dominant seventh chord, a chord of a major triad with a minor third above, respectively. Therefore, we regard P1, P3, and P5 as melodic patterns and P2 and P4 as harmonic patterns.

The results indicate that the frequencies at which the melodic patterns (P1, P3, and P5) are followed by a major triad (P2) are significantly higher for symphonies than for string quartets. The high frequency of the repetition pattern of P2 for symphonies is also remarkable. These results suggest that P2, i.e. the major triad, has the central role among the sequences of PC sets for symphonies because of its stable sonority, for which various melodic patterns and the major triad itself tend to be headed.

We found two characteristics in the analysis of string quartets: 1) melodic patterns in terms of P1 and P3 tend to be continuously used; and 2) P1 tends to proceed to the other harmonic pattern, P4, i.e. the dominant seventh chord, which has dissonant and unstable sonority. String quartets do not require the stable sonority of P2 so much in general.

5 Conclusion

The present study has compared the differences between the symphonies and the string quartets composed by W. A. Mozart, in terms of analyzing bi-gram patterns and their frequencies of PC sets. The findings indicate that collocation patterns for the PC sets of symphonies vary from those of string quartets, such as the patterns where a melodic pattern proceeds to a major triad are more frequent for symphonies than for string quartets.

References

Forte, A. (1973). T*he Structure of Atonal Music*: Yale University Press.

Haberman, S. J. (1973) “The Analysis of Residuals in Cross-Classified Tables,” *Biometrics*, 29(1): 205-220.

Hirano, M. and Yamamoto, H. (2017) “Harmonic Analyses for Mozart’s Symphonies and String Quartets using Pitch-Class Set,” *Proceedings of Jinmoncom 2017*, pp. 83-88.

"Spots of Time" and Space: Mapping the Present, Past, and Atemporal Spaces in Charlotte Smith's *Beachy Head*

Holly Horner[1]

Project premise

This project closely interrogates the layering of physical and imaginary geography, time, and national identity in Charlotte Smith's (1794-1806) *Beachy Head*. The intermingling of physical and imaginary geography and time appear to criticize the politics of French and British imperialism, which Smith further illuminates through her extensive footnoting. Scholars generally agree upon the significance of the Smith's fusing landscape and time in this text—as eighteenth-century literature scholar Michael Wiley indicates, *Beachy Head* "addresses the geography of England and Europe…[and] the geography of the world and an extra-geographical, fanciful and visionary space" (Wiley, 2006: 64). Smith's treatment of time and geography, I argue, draws upon the quintessential Romantic notion of "spots of time," a phrase Wordsworth later writes in the *Prelude* (Wordsworth, 2008: 258-276). The moment where Smith reflects upon the "Haunts of [her] youth!" culminates with the intersection of the past, present, and imaginary, which also can be characterized as a spot of time. This instance demonstrates the interweaving of the geographical and temporal layers seen repeatedly throughout the poem: ranging from the present as the speaker gazes upon the Beachy Head rock formation, to the "vast concussion" as the British Isle separates from the mainland during antiquity and, finally, to the poet's imaginings of a nameless shepherd (Smith, 2017: 163, 174).

Methodology

For the purposes of this project, I refer to three distinct categories of time as Smith treats it in this text: 1) the present moment in the poem as narrated by the speaker, 2) the historical past, and 3) the atemporal, or imaginative, scenes of the pastoral that Smith locates outside of time itself. These temporal layers allow for Smith to simultaneously critique the consequences of the French Revolution outside and within Britain and to retreat from the current historical moment via spots of time.

Building upon previous scholarship on *Beachy Head*, I use geo-coding to map the latitudes and longitudes of these actual and imagined locations to illustrate how Smith layers this complex of geography and time to construct a critique the French Revolution's aftermath in Britain. This system was conducted in R (a programming language and software) through ggplot 2. Each unique place mentioned in *Beachy Head* counted as its own data-point and was assigned rough geographical coordinates. In the poem, Smith is quite clear with the general locations (i.e. she identifies Gallica, or modern France), but she does not always provide the precise geographical locations. Consequently, the coordinates I provided are speculative in nature and based on clues from the text. Although this project strives for geographical accuracy, it is impossible to accurately deduce all the specific locations Smith envisioned for this text.

Discussion

Although rendering visualizations of Romantic writings through mapping is not a new phenomenon, most scholars do not fully consider the relationship between temporality and space in these projects. For instance, The Byron Online Project offers a different mapping endeavor on the British Romantics that tracks the frequency of named locations in Byron's correspondence and in *Childe Harold's Pilgrimage* Cantos I-II. This sort of project remains rooted in Byron's contemporary historical reality and is a popular method for mapping the Romantics. The map from the Byron Online Project does perform important work for visualizing the relationship between these two types of writing, but it does not fully consider

[1] Florida State University

the relationship between time and space. This, in part, could be due to the difference between Byron's and Smith's personal writing agendas.

Implications

Nevertheless, reading the spaces in *Beachy Head* in terms of temporality offers a new perspective of approaching Romantic texts. The temporal oscillation between past and present in the poem remains entrenched in the geographical location of Beachy Head itself as it stands representative of past and present threats of French invasion. In these moments, Smith critiques past and contemporary imperial ideology by invoking traumatic instances held in British national memory and detailing the effects of a global empire upon the rural individual.

Then, through juxtaposing historical and pastoral landscapes, Smith moves away from the traumatic consequences of globalization and into the atemporal landscape of the pastoral. This temporal shift into the imaginary culminates in the spots of time, which allows Smith to distance herself geographically and imaginatively from the current political turmoil.

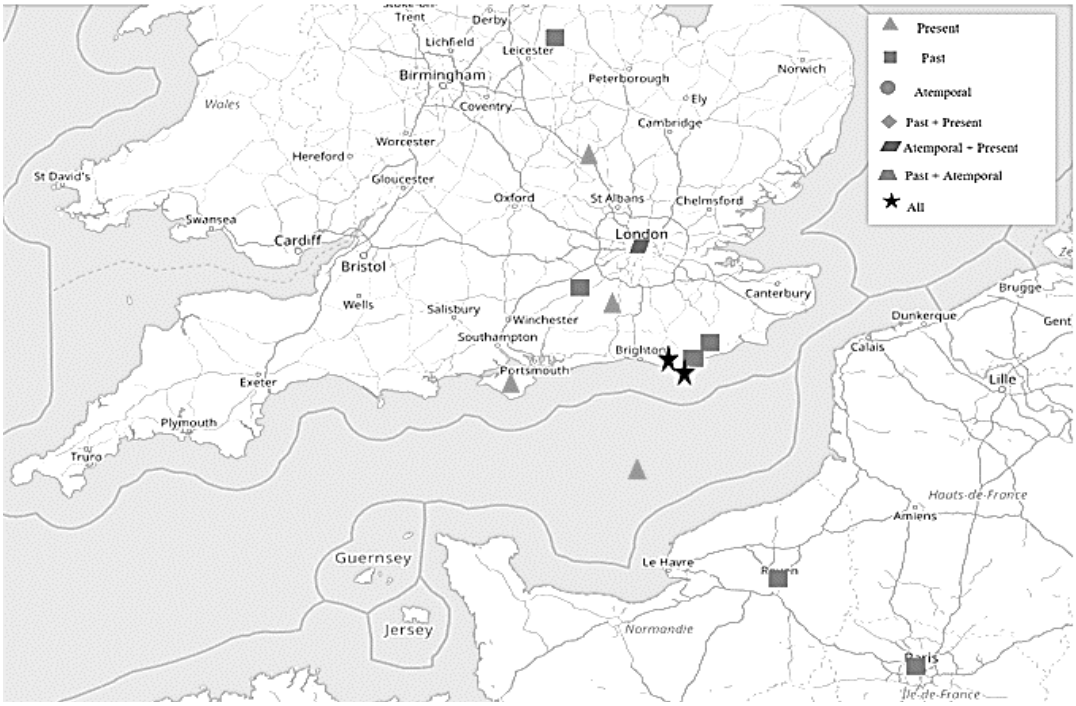


Fig. 1 This map illustrates the primary geographical focus of Beachy Head: the coasts of the British mainland and France divided by the English Channel.

As seen from the above visual, there are only two moments in the text where all three classified temporalities occur at once (i.e. "Haunts of my youth!" and "But from thoughts like these") located on the South Downs of Eastern Sussex—the same region as the Beachy Head rock formation (Smith, 2017: 173, 179). These two moments illustrate how Smith creates her own spot of time by layering different classifications of temporalities. During these moments, she is located simultaneously in past, present, and nowhere, which depicts the development of the spot of time. By mapping Smith's endeavors to create distance from the contemporary political tension, it is possible to see Romantic ideology playing out across the landscape as Smith retreats within the imaginary. This system of mapping reaffirms Smith's status as an early British Romantic poet because she lays the foundation for the spots of time as a Romantic literary tradition, later made famous by Wordsworth. Due to *Beachy Head's* geographic nature, it's necessary and illuminating to create a visualization of the poem's movements across space and time to understand how it fits into a larger literary tradition.

References

Smith, C. (2017). Beachy Head. In Knowles, C. and Ingrid, H. (eds), *Charlotte Smith: Major Poetic Works.* Peterborough: Broadview Press, pp. 163–90.

Wiley, M. (2006). The Geography of Displacement and Replacement in Charlotte Smith's The Emigrants. *European Romantic Review*, 17(1): 55–68.

Wordsworth, W. (2008). The Prelude: Book Eleven. In Gill, S. (ed), *William Wordsworth: The Major Works.* Oxford: Oxford University Press, pp. 258–76.

(2016). Mapping Byron's Mediterranean Letters and Childe Harold's Pilgrimage I-II: By the Numbers *Byron Project Online* http://byrononlineproject.com/neatline/show/byrons-mediterranean-letters-and-childe-harolds-pilgrimage-i-ii-by-the-numbers.

The Brontës in the World: Creating a Digital Bibliography to Expand Access to Single-Language Sources

Matthew Hunter[1], Judith Pascoe[1]

This poster outlines our experience with using Zotero, a free and open-source citation management tool, to make Japanese translations and adaptations of Emily Brontë's classic novel *Wuthering Heights* more accessible to scholars and fans who do not have command of Japanese. Leading this project are Professor Judith Pascoe, the George Mills Harper Professor of English at Florida State University (FSU), and Matthew Hunter, the Digital Scholarship Technologist at FSU Libraries. Prof. Pascoe's research interests include Romantic-era literature and cross-cultural adaptation. Mr. Hunter's work centers on applications of emerging technology in humanities scholarship and pedagogy.

"The Brontës in the World" is the first iteration of a collaborative, multidisciplinary project carried out by undergraduate researchers at Florida State University under the direction of Pascoe and Hunter. The work is enabled by a partnership with the FSU Undergraduate Research Opportunity Program (UROP), which encourages undergraduate students to discover and explore their own research interests with mentorship from university faculty. We designed this project to build on Pascoe's Brontë research, but also to allow undergraduate researchers to track the Brontës' legacy in a variety of cultural contexts. Although we have focused on the Brontës in Japan for this first iteration of our project, the project will develop in keeping with the foreign language strengths and particular research interests of subsequent generations of student researchers.

We chose Zotero as the vehicle for this project because of its ability to gather, organize, and augment bibliographic metadata. Especially as compared to other citation management platforms, Zotero allows users to freely draw on and reconfigure open source bibliographic data. We set out to compile and enrich open data culled from library catalogs and catalog aggregator sources, such as OCLC's WorldCat and the National Diet Library Search. We do so in order to create a new contact point for enriched bibliographic data, a reference site that illuminates how Western literature has been transformed through translation and adaptation in non-Western contexts, and that makes information about these adaptations more broadly accessible.

Our poster also outlines how Zotero functions as a pedagogical tool useful for interrogating digital scholarship methodologies. In producing this bibliography, we have been forced to grapple with how bibliographic structures fail to accommodate non-Western cultural markers. For example, our students have noticed that some adaptations' multiple creator roles (artists, editors, directors, storyline adapters, inkers, etc.) are not reflected in "standard" bibliographic categories, and that non-Western naming conventions are often not easily represented.

Together with our students, we are also engaging with Zotero as a hermeneutic device that helps us think about the organizational structures imposed by current cataloguing systems. As our research team adds bespoke tagged and relational data to our library, we see how connections among our sources enable some forms of relationship-building but delimit others. In other words, tagging is meaning-making. While interacting with this tool, our students and we have, by necessity, questioned how we access and compartmentalize knowledge.

Our poster then summarizes our experience using a Zotero bibliography as a teaching tool, a research activity, and a mode of scholarly humanistic inquiry into digital hermeneutics. "The Brontës in the World" stands as an effort to showcase the transmission of the Brontës' work, but also as a meditation on data organization that, we hope, will fuel

[1] Florida State University

conversations in the international DH community about the affordances and limitations of current resource management infrastructure.

We are happy to share how Zotero, nominally a citation management tool, has served as the foundation for both our research and pedagogy. It has allowed us to build a database that will serve researchers interested in translation and adaptation studies, and to establish a hub for ongoing student explorations of data collection and citation practice. To supplement the poster presentation, we provide an illustrated two-language (English and Japanese) handout that highlights our discoveries and future plans.

The Metadata Hub for Interdisciplinary Knowledge Sharing of Historical Situation Records

Mika Ichino[1], Junpei Hirano[2], Kooiti Masuda[3], Asanobu Kitamoto[1], Hiroyuki Den[4]

Introduction

"Historical Situation Records" (HSR) are the records containing various information of historical events such as earthquakes, weather patterns, supernovas, the blooming of cherry blossoms, famines, social activity, and other notable incidents. HSR has been used as a source of data for research in numerous fields, including seismology, climatology, astronomy, sociology, and history. These can be found on different historical materials like paper notes, chapters in wood, stone monuments, images in photographs and paintings, and so much more.

Promoting the use of HSR in interdisciplinary fields is one of the component measures of Historical Big Data (HBD)[1]. The processes of using HSR data (HSR-workflow) not only includes the identification, obtaining, transcribing and reading of HSR, but also structuring, analyzing and integrating, and sharing the HSR data. Unfortunately, each process of HSR-workflow has difficulties which do not appear with the born digital data. Furthermore, scientists and engineers using HSR data for research are not always familiar with historical documents, making it challenging for them to manage HSR workflow by themselves.

Until now, this research has been using HSR data in individual fields by such HSR workflow. Occasionally researchers extract HSR data from the same material. For example, the Ishikawa Diary[2], written by a farmer family since 1720 in Tokyo, contains not only daily weather conditions, but also important records such as meteorite and earthquake events.

Under these circumstances, sharing knowledge and experience associated with HSR (HSR metadata) can be useful for research and technologies associated with main HSR-workflow operations. This can help avoid repeating processes that have been done by others, and reduce difficulties in HSR-workflow, notably identification and obtaining HSR data. We have piloted and continue to develop an improved system for the metadata hub. All of these are included and introduced in this paper.

Concept of the Metadata Hub System

The basic idea of the metadata hub system is to reuse various information from previous research in each field as HSR metadata with the users' mutual corporation. The HSR metadata contains information about not only something's bibliography and location, but also its HSR and various descriptions like the state of the materials (paper material, image, transcribed text data, published book, digital text data, or structuring data). Table 1 shows a sample of HSR metadata of a published historical diary. The HSR metadata can also contain various information about HSR such as errors and reliability of HSR data, which previous research have acquired on the HSR-workflow. The system can thus help researchers to avoid repeating the same processes of HSR-workflow which have been done by others.

If a document has a description about HSR, a user can register these subjects as HSR metadata items in Table 1. Then, other users can search by temporal and spatial range and type of phenomena and obtain HSR metadata such as the items in Table 1. Fewer mandatory metadata items and adding flexibility to optional metadata items and

[1] Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems / National Institute of Informatics
[2] Teikyo University
[3] Tokyo Metropolitan University
[4] Academic Express

descriptions would attract more registration and participation. At the least it would inform others of the existence of the various state material and data as HSR metadata. This includes data information of material still partially undergoing research, or with restricted access.

Table 1. A sample of HSR metadata items users can obtain
*: required, +: optional, -: inputted by the system

Metadata Items (＊: required, +: optional, -: inputted by the system)			samples of descriptions
ID	Identifier	ー	
登録日時	Data	ー	
書誌情報	Bibliographic resource	＊	www.cneas.tohoku.ac.jp/news/2012/publication04_2.html#38
書誌名称	Title	＊	東北アジア研究センター　叢書　第38号　佐藤大介編著『18～19世紀仙台藩の災害と社会別所万右衛門記録』「天保凶作日記（一）～(五）」
登録者	Creator	＊	平野
著作権等	Rights	＋	CC BY NC
観測値	Observation Point	＋	宮城県仙台市
記録開始（和暦）	Time of starting to record	＋	天保4年
記録開始（西暦）	Time of starting to record	ー	1833/02/20
記録終了（和暦）	Time of ending to record	＋	天保14年
記録終了（西暦）	Time of ending to record	ー	1844/02/17
観測地の変遷情報	History of observation points	＋	記録地の変遷なし
歴史的状況記録	Information of HSR (type of phenomena such as earthquake, the weather, disasters, etc)	＋	天気
		＋	相場（米、大豆、大麦、小麦など）
		＋	地震
		＋	水害
		＋	昆虫（蚊、蛍など）の記述あり
		＋	植物季節（桜）の記述あり
記録の連続性	Continuity of recording	＋	9割ほど
天気の詳細に関する情報	Level of detailed weather	＋	時間変化あり、寒暖の記述あり、降水程度の記載あり、風の記述あり、雲の記述詳細
間接的な記録（社会的な記録）	Social information	＋	他領地米の購入、到着などの記述あり、天候祈願の記述あり
整理状況（翻刻済など）	State of the material	＋	翻刻済、出版されている

The required functions of the metadata hub system are the following.

1. Registering a few required items(the mark "＊") and at least one state of the HSR or material (the mark "+") in Table 1
2. Searching HSR metadata items, especially types of phenomena, and obtaining various HSR metadata in some data formats such as csv, json, and pdf.
3. Search by temporal and spatial range
   This needs to handle numerical types of data and use APIs that have been developed for historical temporal and spatial data.
4. Revising and adding more information to the registered data by other users

It will allow HSR information to be managed, by allowing experts such as history academics to add reliable information.

Developments of the Metadata Hub System

A first prototype has adopted the open source software Omeka Classic version 2.6[3] as a platform. Even though it has simple functions including sharing HSR metadata, its adaptability is not enough to fulfill the requirements as described above, particularly the 3rd and 4th requirements. An upcoming system thus utilizes a spreadsheet on Google drive[4] for registration and needs to develop some functions such as search by temporal and spatial range on. Moreover, improvement of an increasing usability, metadata schema, linking sources, displaying the results, user interface, and increasing the data number towards the release are in great need.

Conclusions and Future Works

We have attempted to develop and improve a system of sharing HSR metadata. This does not only help reduce these difficulties in HSR-workflow, but also creates new interdisciplinary collaboration between researchers. A significant role is to describe the reliability of information in the system. Although the administrators can take these quality controls of all descriptions currently, acquiring more participants can advance reliability through cross checks in the future. Additionally, to secure its use in the future, HSR metadata needs to link to other data and databases by collaboration using new data sharing technologies.

References

1. Historical Big Data: http://codh.rois.ac.jp/historical-big-data/ (accessed 4 July 2018).
2. **Hachioji Board of Education.** (1988). *Ishikawa Family's diaries*, Hachioji Local Museum.
3. Omeka Classic: https://omeka.org/classic/ (accessed 8 May 2018)
4. Google Drive: https://www.google.com/drive/ (accessed 3 July 2018)

Construction of NINJAL media resources collection for searching and previewing sound and video data

Yuichi Ishimoto[1], Takumi Ikinaga[2], Tomokazu Takada[1]

For more than 70 years, the National Institute for Japanese Language and Linguistics (NINJAL) has carried out Japanese language research such as investigations of dialects, vocabularies, language life, and corpora. The research results are completed as academic reports and papers and made public. In addition, the intermediate products of research (e.g., spreadsheets and index cards) and primary sources (e.g., sounds, videos, questionnaires, original magazines for vocabulary research, research plans, and minutes) are preserved at the research materials room of NINJAL [1-3]. However, it is not easy to browse these materials, especially sounds and videos, because they are stored in old media such as reel-to-reel tape, cassette tapes, and DAT that need the vanishing media players in the present day. In this paper, we introduce the digital cataloging of research materials preserved in NINJAL and describe web-based systems of resource collection for searching the materials and previewing sounds and videos.

Language research resources were once stored on paper. Then it became possible for sound and video resources to be stored in magnetic tapes along with recording technology since modern times. Thus, we can access the media resources at a later date. However, the problem with magnetic tapes is that data may be lost as the tapes deteriorate over time. In recent years, we have been engaged in converting the data from the old media to digital data on HDD. At the same time, we are constructing a digital catalog consisting of a list of research materials collected or created by various research projects of NINJAL over the past 70 years and their abstracts. The catalog, *Research Materials on NINJAL*, is available to the public on the website [4] (Fig. 1), and is registered to the NINJAL Research Library OPAC.

Figure 1: Digital catalog of research materials on NINJAL

[1] National Institute for Japanese Language and Linguistics
[2] Tokyo Denki University / National Institute for Japanese Language and Linguistics

Researchers searching for suitable data for their studies have been required to watch and listen to the materials stored in NINJAL. However, some of the media resources include sensitive personal information; therefore they cannot be open to everyone. We can also play the data on the standard PC by digitalizing the sound and video data, but it takes time to confirm whether the data have the characteristics we want because there are large files that last for more than two hours. Accordingly, we constructed the system, *NINJAL Media Resources Collection*, for quickly previewing sound and video resources via the web (Fig. 2). This collection has 18,054 sound files and 278 video files as of May 2018. The system restricts the download of sound and video files and permits streaming playback by using a specialized player in it (Fig. 3). It is only accessible from the local network of the NINJAL, and the outsiders have to come to the NINJAL library to access the system. Thus, we manage both users' accessibility to media data and prevention of data leakage via the web.

Figure 2: NINJAL Media Resources Collection

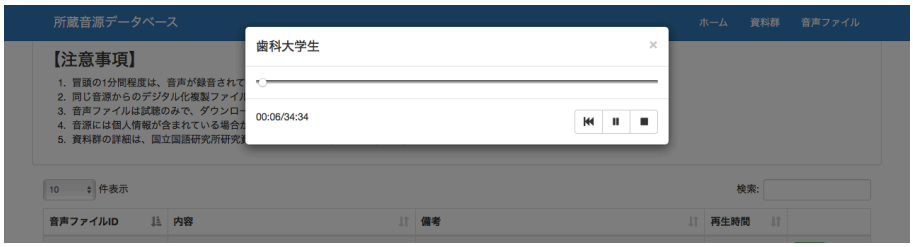


Figure 3: Media player implemented on the web

The *NINJAL Media Resources Collection* includes rare sounds and videos. For example, a study titled "Research in the colloquial Japanese" [5] attempted to survey characteristics of Japanese from various points of view in the 1950s and includes everyday conversations between ordinary people of former days. Also, a study titled "Some aspects of honorific expressions: In special reference to discourse" [6] investigated honorific expressions spoken in a local community in the 1960s and includes conversations recorded at home in 24 hours. Nowadays people seldom have an opportunity to listen to the everyday conversations before the 1960s because it was the age before portable recording devices were widely available. We believe that the systems will help various researchers engage in the study of the Japanese language.

References

[1] **S. Morimoto,** "Developing EAD-based archival description at the National Institute for Japanese Language," Journal of the Japan Society for Archival Science, No. 4, pp. 92-102, 2006. (in Japanese).

[2] **H. Terashima,** "Practical Uses for Research Materials Owned by NINJAL," NINJAL Research Papers, No. 10, pp. 245-263, 2016. (in Japanese).

[3] **R. Yamaguchi, M. Sekikawa,** "An improvement of access to research materials in the National Institute for Japanese Language and Linguistics," IPSJ Symposium Series, Vol. 2016, No.2, pp. 51-56, 2016. (in Japanese).

[4] Research Materials on NINJAL, http://rmr.ninjal.ac.jp/

[5] **The National Language Research Institute,** "Research in the colloquial Japanese," The National Language Research Institute Research Report; 8, 1955. (in Japanese).

[6] **The National Language Research Institute,** "Some aspects of honorific expressions: In special reference to discourse," The National Language Research Institute Research Report; 41, 1971. (in Japanese).

Developing a Block Puzzle Game for Studying Ryukyuan Language Phonetic System

Takayuki Kagomiya[1], Yuto Niinaga[1], Nobuko Kibe[1]

National Institute for Japanese Language and Linguistics (NINJAL) and National Museum of Japanese History (Rekihaku) started The Mobile Museum Project. The aim of this project is developing a compact and movable exhibition system like a travelling funfair, and to contribute social pedagogy by using the travelling exhibition (Figure 1). As a part of this project, we developed a block puzzle game for studying Ryukyuan (Okinawa) language phonetic system.

Figure 1: A sample of The Mobile Museum exhibition kit.

For most of Japanese people who brought up out of Okinawa region, Ryukyuan language sounds very different from Japanese language and hard to understand. However, between Japanese and Ryukyuan language, systematic phonological correspondence rule is observed. Thus, many basic vocabulary words of Japanese are able to be translated into Ryukyuan by replacing phonemes according phonological rule. For example, Japanese phoneme /o/ is realized as /u/, /ki/ corresponds /ʧi/ (Ono and Shibata eds. 1977; Iitoyo, Hino and Sato eds 1984) . Thus Japanese /kojomi/ (calendar) is able to be translated as /kujumi/, /kimo/ (viscera) corresponds /ʧimu/ (National Institute for Japanese Language ed. 1963). Target of our game-style studying material is to understand this phonological rule and to have interest in analyzing language system with fun.

To make studying more fun, a game style teaching material is effective. Thus, our teaching material designed as a quiz whose rule is translation of Japanese words into Ryukyuan's. The studying material consists of two parts: Ryukyuan phonological rule instruction and block puzzle. The instruction includes correspondence table of Japanese and Ryukyuan phonemes (Figure 2). People who want to play the game read the instruction first and learn how to translate Japanese words into Ryukyuan's. The block puzzle also divided into three components: block chips, answer board and control unit. On the top of each block chip, a Japanese mora (ex. [ku], [ʧi] etc.) is printed respectively (Figure 3). On the answer board, Japanese words (ex. [kojomi], [kimo] etc.) are printed. Beside a Japanese word, a column for answer is located. The answer columns have sockets into which the block chips are able to be inserted (Figure 3). A player reads the Japanese word and translate into Ryukyuan, then the player should insert appropriate block chips into a

[1] National Institute for Japanese Language and Linguistics

sockets and complete Ryukyuan word. If the answer is correct, Ryukyuan speech sound of the word is played. The answer is judged by computer which is set in the control unit. Playing sound is also a function of the control unit.

Figure 2: The correspondence table of Japanese and Ryukyuan phonemes

Figure 3: The answer board (upper) and block chips (below).

Another feature of this teaching material is portability. As described above, the teaching material is a part of the Mobile Museum. Thus, the teaching material must be transport with the Mobile Museum system. The answer board and the block chips are able to be stored in the control unit case (Figure 4). All components can be packed in

portmanteau-size container. Setting the material is also easy. The teaching materials can be installed with few steps.

Figure 4: Whole system installed in a Mobile Museum kit. Control unit box is shown below table.

We conducted short-term exhibition using the Mobile Museum includes this teaching material in a university. Spectators played the game and studied about Ryukyuan language. After playing the game, the spectators evaluated the teaching material. The results of the evaluation indicated the game was fun and useful, users had positive impression not only from the material but also from studying linguistics.

References:

Shuzen Hokama (1977). “Okinawa-no gengo-to sono rekishi (Ryukyuan language and its history,” In **Susumu Ono and Takeshi Shibata (eds),** *Iwanami Koza Nihongo 11 Hougen (Iwanami Handbook of Japanese Vol.11, Dialects)*, Tokyo: Iwanami Shoten.

Masachie Nakamoto and Takeo Nakamatsu (1984). “Nanto-hogen no gaisetsu (Abstracts on Ryukyuan Language),” In **Kiichi Iitoyo, Sukezumi Hino and Ryoichi Sato (eds),** *Koza Hougengaku 10, Okinawa Amami-no Hougen (Handbook of Japanese Dialects Vol.10, Ryukyuan Dialects),* Tokyo: Tosho-Kanko-kai.

National Institute for Japanese Language (ed) (1963). *Okinawa-go Jiten (Dictionary of Okinawa language)*, Tokyo: Printing Bureau, Ministry of Finance.

Comparisons of Pitch Intervals in Japanese Popular Songs from 1868 to 2010

Akihiro Kawase[1]

1. Aim of the study

The aim of this study is to quantitatively describe the characteristics of the melodic pattern of Japanese popular songs of each era by examining the basic statistical data obtained from a musical corpus. Music is an art established in which various elements are complicatedly related, and it is difficult to grasp the whole body objectively. However, when listening to music, people understand musical genres and styles based on some characteristics. Aspects of musical structure, such as meter, phrase structure, contrapunctual structure, pitch spelling, harmony, and key, are well known and understood by many music studies, and thus, are frequently taken for granted as musical facts. However, one question that has yet to be answered is what process underlies the inference of such structures (Temperly 2001).

In the previous research, we converted 120 Japanese popular songs of all within the top three domestic sales from 1970 to 2010 and extracted the tendency of change in tonality over 40 years of Japanese music culture using the KeyScape algorithm (Sapp 2011). Based on the findings, we predicted that there are three tendencies: (1) as time goes by, songs on the minor keys decrease, and songs on the major keys increase; (2) periodic appearance of songs with a lot of modulation (key change); (3) considering the above two, the trend of Japanese popular songs can be divided into five periods.

However, the history of Japanese music is long, and if we consider the end of the 19th century, when Western music was imported into Japanese culture, as the beginning of popular songs, there will be a history of over 140 years. Therefore, since study in order to grasp the long-term history of Japanese popular songs has not yet conducted so far, in this research, we aim to analyze popular songs over 140 years in Japan and grasp the periodic changes more precisely regarding tonality changes by era.

2. Procedure

We analyze all 2,136 songs included in all nine volumes of 'Nihon no Uta' (*Songs of Japan*), a collection of musical scores from 1868 to 2010. We digitized all the songs from each subcorpus (see Table 1) and generated sequences that contain interval information from the song melodies.

In order to achieve the purpose of this study, we converted all the songs in each volume into MusicXML file format and constructed subcorpus by classifying the song every ten years. However, due to the small amount of songs from 1868 to 1930, the period was divided into two from 1868 to 1910 and 1911 to 1930, respectively. By extracting the pitch intervals for ten subcorpus and comparing the results every decade, we found a characteristic melody pattern of each era.

[1] Doshisha University

Table 1: Basic statistics of song data classified every ten years

Era	Songs
1868-1910	132
1911-1930	163
1931-1940	237
1941-1950	167
1951-1960	221
1961-1970	278
1971-1980	289
1981-1990	219
1991-2000	282
2001-2010	148

The procedures are as follows: (1) we digitized all the songs from each sub corpus and generated sequences that contain interval information from the song melodies; (2) extracted transition frequencies for every subcorpus separately, and create a 25-dimensional data from interval of -12 to +12 with 10 samples (eras); and (3) applied hierarchical cluster analysis and correspondence analysis to identify pitch height usages in the data, and to highlight their similarities and differences.

We devised a method of digitizing each note in terms of its relative pitch by subtracting the next pitch height for a given MusicXML. It is possible to generate a sequence T that carries information about the pitch to the next note: $T = (t_1, t_2, \dots, t_i, \dots, t_n)$. An example of the corresponding pitch intervals for t_i can be written as shown in Table 2. We treat sequence T as a categorical time series and execute unigram to capture transitions and their trends.

Table 2: Corresponding pitch intervals

t_i	Pitch intervals	t_i	Pitch intervals
0	perfect unison	7	perfect fifth
1	minor second	8	minor sixth
2	major second	9	major sixth
3	minor third	10	minor seventh
4	Major third	11	major seventh
5	perfect fourth	12	perfect octave
6	aug.fourth/dim.fifth	13	minor ninth

3. Results and Discussions

Figure 1 is a mosaic plot showing the relationship between era and pitch interval. In the figure, mX (m12, m11, ..., m1) represents the pitch interval in descending order (t_i<0), and pX (p1, p2, ..., p12) represents the pitch interval in ascending order (t_i>0), respectively. The pm0 represents the same pitch transition (t_i=0). From Figure 1, we can confirm that pm0 is used most frequently in any era, mX and pX are almost synchronized with each other, and have a bilaterally symmetric distribution centered on pm0.

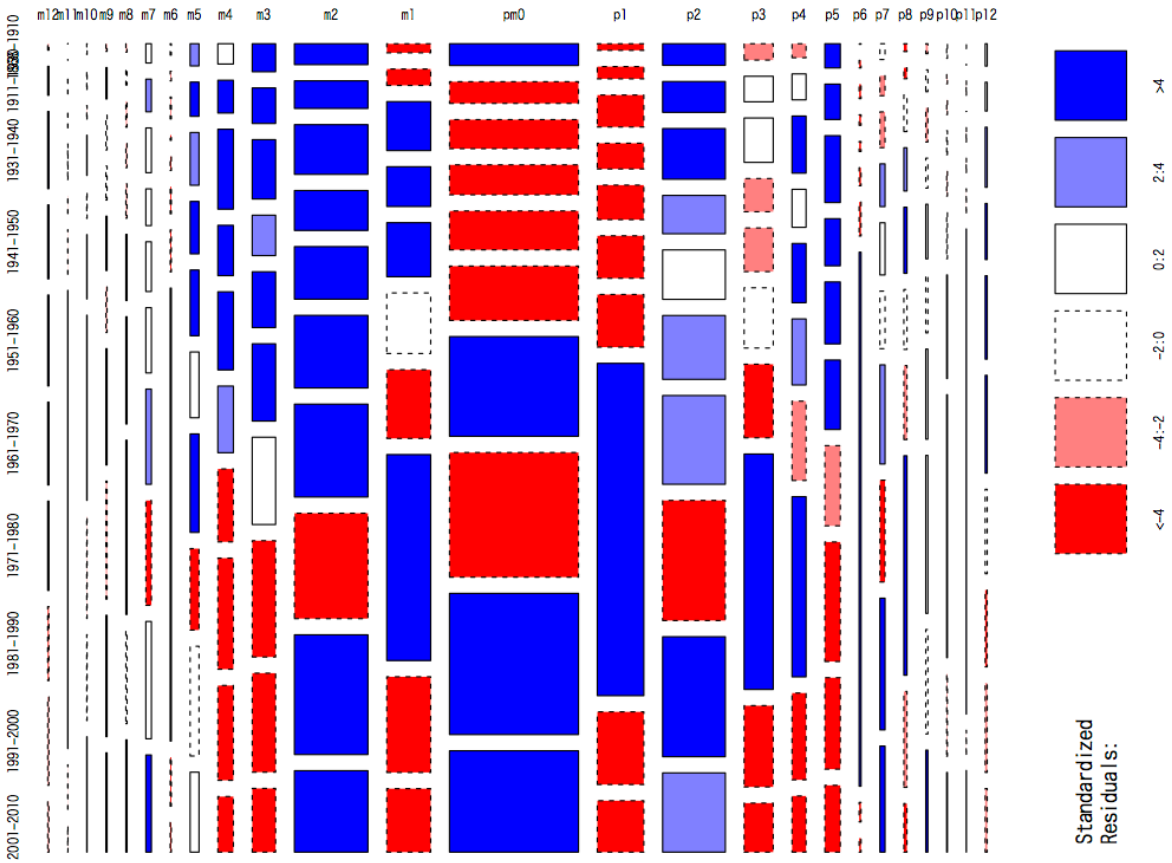


Figure 1: Mosaic plot of pitch interval frequencies for every ten years

Figure 2 is the result of a hierarchical cluster analysis using the unigram of the pitch intervals as input variables. We see that, except for the 1970's and the 1980's, the cluster formed by the decades. From this fact, it can be assumed that only the data of the 1970's and the 1980's have a different tendency of pitch intervals from the preceding and the following eras.

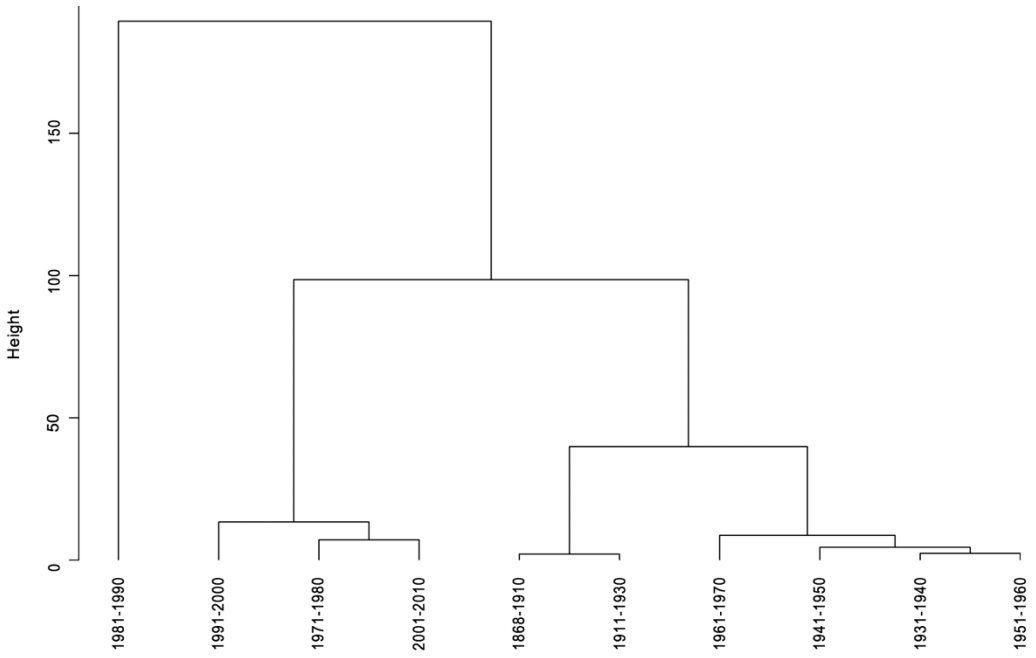


Figure 2: Result of cluster analysis

In order to distinguish and grasp the trend of the pitch interval used for each era, we carried out a correspondence analysis. Figure 3 shows the result of analysis using 25

pitch intervals (e.g., -12, -11, …, -1, 0, +1, +2, …, +12), and Figure 4 shows the result of analysis using 13 variables that summarizes the ascending and descending intervals into one (e.g., 0, ±1, ±2, …, ±12,).

As shown by these two results, popular songs in the 1980's, the influence of songs using pitch intervals of minor second pm1 (±1) was stronger than in other eras, and as in the music from the 2000's onwards, songs in the 1970's was confirmed to be in a different position from other eras because of the stronger tendency of using perfect unison pm0 (±0) and major sixth pm9 (±9).

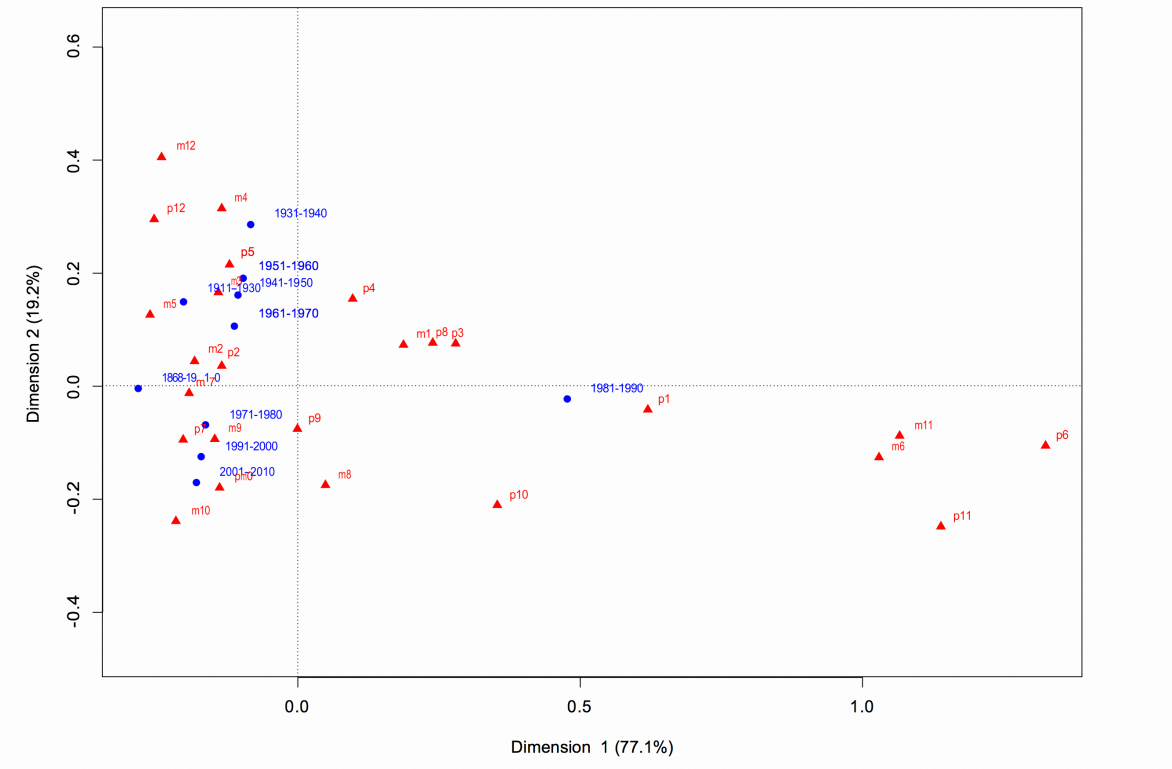


Figure 3: Results of correspondence analysis using 25 pitch intervals

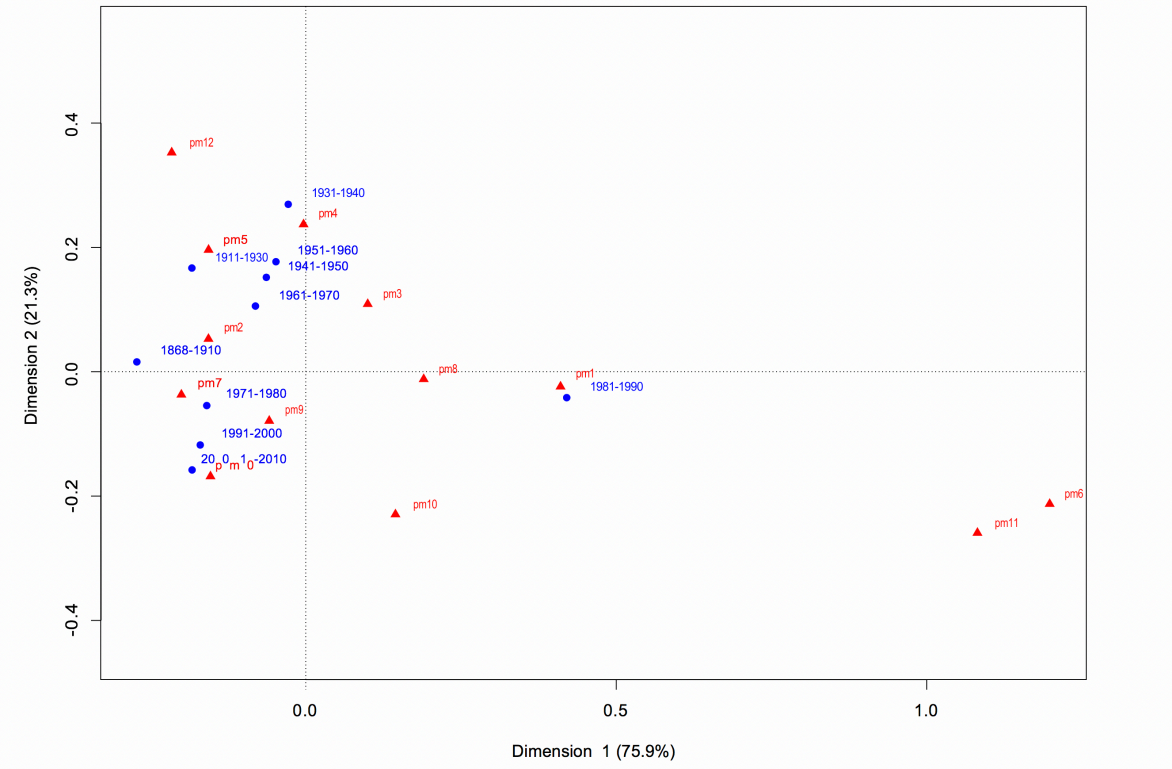


Figure 4: Results of correspondence analysis using 13 pitch intervals as input variables

4. Conclusion

In this research, in order to explore the transition of the musical characteristics of Japanese popular songs, we executed multivariate analysis and compared the musical trend in terms of pitch intervals for 2,136 songs from 1868 to 2010. We revealed that the tendency of the pitch intervals which influences music differs according to each era. Moreover, it was confirmed that the use trend of the pitch intervals is similar between near decade, but the songs in the 1970's (1971-1980) and the 1980's (1981-1990) did not follow this rule.

Although we were able to extract the pitch interval information of music and quantitatively analyzed the secular change of the trend of music, we did not grasp the local features such as pitch transition patterns and rhythm patterns. As future tasks, in order to highlight the transition of Japanese music culture, we will focus on transition patterns of both pitch and rhythm information by analyzing the secular change of features.

References

[1] **Sapp, C. S.:** *Computational Methods for The Analysis of Musical Structure* : Stanford University, 2011.
[2] **Temperley, D.:** *The cognition of basic musical structures*, The MIT Press, 2001.
[3] MusicXML, http://www.musicxml.com/for-developers/ [accessed 26 November 2017].

KU-ORCAS: Trans-Border Digital Archives Project for East Asian Cultural Studies

Nobuhiko Kikuchi[1]

1. Introduction

Kansai University Open Research Center for Asian Studies (KU-ORCAS) is a project focused on building digital archives for East Asian Cultural Studies. It was selected as one of the Research Branding Projects of the Ministry of Education, Culture, Sports, Science and Technology in 2017.

2. Purpose of the project

KU-ORCAS aims to serve as an international research hub for East Asian Cultural Studies by constructing digital archives and an open platform that will provide openly-licensed digital images and resources.

It should be noted that East Asian Cultural Studies in this context does not only assume specific national frameworks such as Chinese history or Korean cultural research, but also has a trans-border aspect. This is because Kansai University has a long history of East Asian Cultural Interaction Studies and KU-ORCAS sets this field as a central research theme. The two meanings of trans-border are trans-border from the national research framework and trans-border from academic research fields. Therefore, we must imagine both users whose research themes are the cultural relationships across national and/or regional boundaries and users who do not have expertise in East Asian Cultural Studies. In short, we need to consider who our users are and how they will use our digital archives in order to seek the best way of providing data and designing the data usage environment.

The significance of this project is to propose functions of digital archives for supporting trans-border studies, which are a current research trend in the humanities, such as in the field of Global History.

3. Materials provided by KU-ORCAS

We plan to digitize and openly provide specific and abundant resources. The resources are roughly divided into three groups. The first group comprises pre-modern resources translated into Asian languages such as dictionaries, grammar text books, and missionary reports. The second group is the Hakuen Bunko archives which is the collection formerly possessed by Hakuen Shoin(泊園書院), which is one of the origins of Kansai University. In addition to that, Kansai University's pre-modern Osaka Art Collection(大阪画壇) will also be added to this group. The third group is composed of materials such as excavation data and drawings relating to ancient Asuka and Naniwazu studies, which have been promoted by Kansai University.

4. Three concepts of Openness and an Open Platform

We will provide the materials from the standpoint of three concepts of openness on our open platform.

The first openness is the opening of research resources. This refers to the digitization and free provision of materials in KU-ORCAS's digital archives. Kansai University is a member of the IIIF Consortium, and we will release images complying with the IIIF standard (see Figure 1).

[1] Kansai University

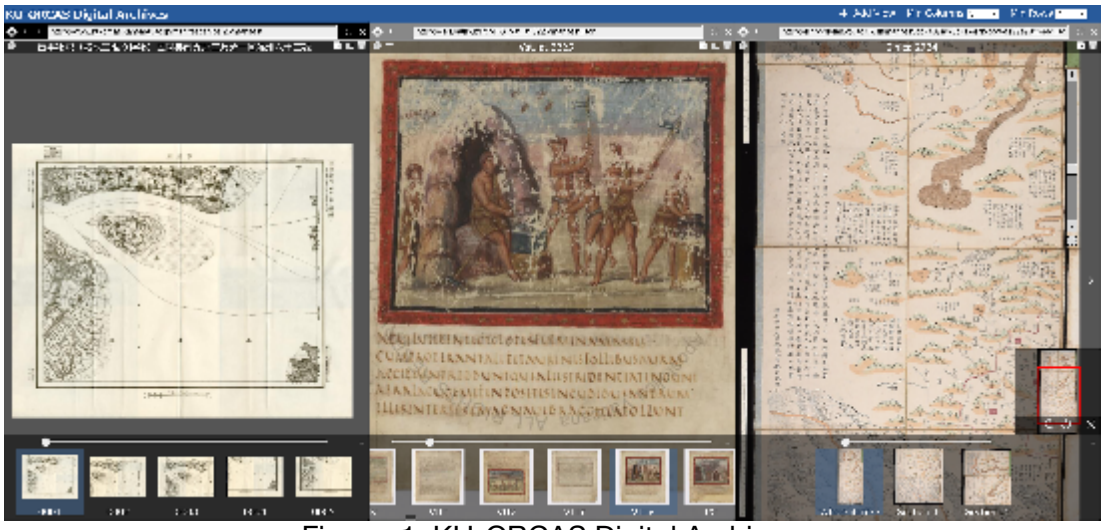


Figure. 1: KU-ORCAS Digital Archives

The second is the opening of research groups. This assumes cooperation with researchers inside and outside of Kansai University, academic societies, educational institutions, and citizens. In particular, we plan to develop a crowdsourcing system for transcribing digitized materials through which citizens can easily participate in our research.

The third is the opening of research know-how. We will build a website to disseminate technical information and know-how accumulated through the construction and operation of KU-ORCAS's digital archives. On that site, users will be able to exchange knowledge about how to use data from the digital archives.

Finally, the open platform will employ the functions of the above three concepts of openness and provide a portal search engine. Since we will convert the bibliographic data of the digital archives to the Linked Open Data format, users will be able to expand their search range and find unexpected search results.

5. Functional requirements as trans-border digital archives

In this final chapter, we will explain the requirements for our digital archives to acquire trans-border functionality to support East Asian Cultural Studies. As we noticed before, the requirements are based on both trans-border from the national research framework and trans-border from academic research fields.

Regarding the former, it goes without saying that interfaces and help screens must be multilingual. In addition, we must provide functions with which users can read the texts and document titles in their native languages. In other words, we should offer not only English translation and romanization of material titles but also auto-translated texts through AI image recognition or transcribed texts through the crowdsourcing system.

Regarding the latter, it is necessary to provide reference resources for various subjects. This can be accomplished by linking and displaying so-called reference materials such as dictionaries and encyclopedias in our digital archives. In addition, it is possible to include functions with which multiple researchers in different fields can interpret materials so that multiple users are able to read digital materials jointly. It will be also effective for users to add their annotations to others' annotations or to respond to each other in thread form in the digital archives.

KU-ORCAS now shifts to the developing phase to realize the functions discussed in this paper. We will lead the innovation of East Asian Cultural Studies through digital archives.

Acknowledgements:

This paper was originally published in a research report of 117th IPSJ SIG Computers and the Humanities. The original report was written by Nobuhiko Kikuchi, Keiichi Uchida and Kiyonori Nagasaki. Both Dr. Uchida and Dr. Nagasaki allowed Kikuchi to submit this paper as an individual to JADH2018.

References:

[1] **Nobuhiko Kikuchi, Keiichi Uchida, and Kiyonori Nagasaki**. (2018) . “越境する」デジタルアーカイブの機能要件を考える －KU-ORCAS が備えるべきものー.” *Research Report of 117th IPSJ SIG Computers and the Humanities*, May 2018.

[2] *関西大学アジア・オープン・リサーチセンター「KU-ORCAS」*. [online] Available at: http://www.ku-orcas.kansai-u.ac.jp/ [Accessed 2 Jul. 2018].

Cancelled

Alignment Table between UniDic and 'Word List by Semantic Principles'

Asuko Kondo[1], Makiro Tanaka[2], Masayuki Asahara[1]

Word-sense annotated Japanese corpus is an important resource for both linguistic research and natural language processing. A systematic sense hierarchy, used to express the semantic categories, is necessary in order to investigate Japanese lexicon semantically. The corpus, which is annotated with the word senses, should be a representative and a balanced one. Moreover, annotation should be performed on all the words in the corpus. If we have an alignment table between a morphological analyser lexicon and the thesaurus, then the table enables us to extract all possible word senses by using the morphological analyser. This paper presents the project to develop the alignment table between UniDic and 'Word List by Semantic Principles' (hereafter WLSP). UniDic is a largescale lexicon for the Japanese morphological analyser, MeCab. WLSP is a thesaurus that has four classes of syntactic categories and hierarchical semantic categories. Both lexicons are open language resources that are used for academic purpose. The constructed alignment table enables us to extract all possible senses in the registered entry.

Below, we present the design of the alignment table between the UniDic lexeme and the entries of WLSP. We set the domain of the alignment relational table as WLSP vocabulary entries and the range as UniDic lexemes.

A lexeme of UniDic is defined a group of surface words derived from the same origin. This is the most appropriate layer to align based on word senses. UniDic lexeme can be uniquely identified by 'lexeme', 'lexeme reading', 'lexeme subtype', 'class', and 'word type'. We judge whether the two entries from WLSP and UniDic are the same word by the following conjunctive conditions (AND):

(A) the vocabulary entry in WLSP and the lexeme in UniDic are agreed.
(B) the reading of the vocabulary entry in WLSP and the reading in UniDic are agreed.
(C) the class (類) of vocabulary entry in WLSP and the class in UniDic are agreed.

The class is a superclass of POS. Nominal class (体), verbal class (用), modifier class (相), and others (その他) are defined in WLSP. Though the class is also defined in the UniDic lexeme, the label set is different between WLSP and UniDic lexeme. The table shows the alignment table between the two.

[1] National Institute for Japanese Language and Linguistics
[2] Meiji University

WLSP class	UniDic lexeme class
体 (nominal)	体 (nominal)
	固有名 (proper noun)
	人名 (person name)
	姓 (surname)
	名 (name)
	地名 (place)
	国 (country)
	数 (numeral)
	接尾-体 (nominal suffix)
用 (verbal)	用 (verbal)
	接尾-用 (verbal-suffix)
相 (modifier)	相 (modifier)
	接尾-相 (modifier-suffix)
他 (others)	他 (others)

The following figure shows an example of matching rule:

WLSP				UniDic Lexeme		
Entry	Lex. Reading	Class		Lexeme	Lex. Reading	Class
事	こと	体	→	事	こと	体
koto		nominal		*koto*		nominal

We also define the following three exception rules:

(A') even if the condition (A) is not satisfied, we check the agreement between the article of WLSP and the examples of the UniDic lexeme.

The following example shows the aligned pair as defined by the condition (A').

WLSP					UniDic Lexeme		
Entry	Lex. Reading	Class	Article		Lexeme	Lex. Reading	Class
これ	これ	体	こそあど	→	これ	これ	体
Kore		nominal	demonstrative		*koto*		nominal

(B') even if the condition (B) is not satisfied, we check the agreement between the readings of WLSP and UniDic.

The following example shows the aligned pair as defined by the condition (B').

WLSP				UniDic Lexeme			
Entry	Lex. Reading	Class		Lexeme	Lex. Reading	Class	Reading
依存	いそん	体	→	依存	イゾン	体	イソン
ison		nominal		*izon*		nominal	*ison*

(C') even if the condition (C) is not satisfied, we check the agreement between the class of WLSP and the examples of UniDic POS.

The following example shows the aligned pair as defined by the condition (C').

WLSP				UniDic Lexeme			
Entry	Lex. Reading	Class		Lexeme	Lex. Reading	Class	POS
リアル	リアル	相	→	リアル	リアル	体	名詞-普通名詞-形状詞可能
riaru		modifier		*riaru*		Nominal	Adjective-Noun

We annotate the same word relation from all the entries of WLSP to UniDic lexemes. Some relations between the WLSP entries and the UniDic lexemes are not one-to-on relations, i.e. neither injective (n-to-one) nor functional (one-to-n).

The following example shows relations that are not injective (n-to-one). In total, 10,683 UniDic lexemes are assigned to multiple WLSP entries. The maximum number of WLSP entries for one UniDic lexeme is 13.

WLSP						UniDic Lexeme		
Entry	Reading	Class	Article Number	Article		Lexeme	Reading	Class
出す	だす	用 Verbal	2.3832	出版・放送 Publish	→	出す	ダス	用
出す	だす	用	2.3770	授受 Exchange				
出す	だす	用	2.1531	出・出し Put out				
出す	だす	用	2.1521	移動・発着 Movement				
-出す	だす	用	2.1502	開始 Start				
出す	だす	用	2.1211	発生・復活 Occur				
出す	だす	用	2.1210	出没 Appear				

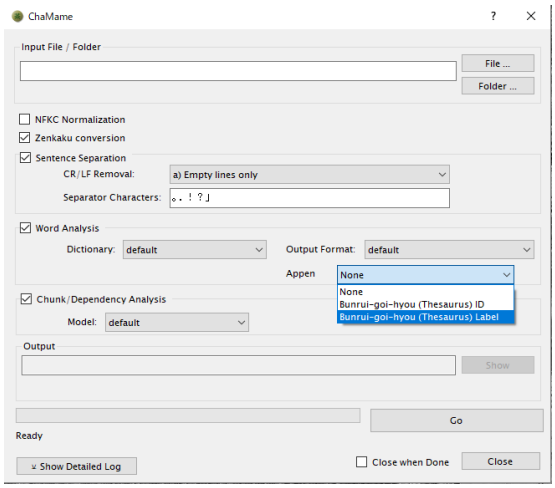
The following example shows relations that are not functional (one-to-n). Nonfunctional relations, ranging from one WLSP to multiple UniDic lexemes, are limited to

270 WLSP entries. The maximal number of assigned UniDic lexemes is limited to 2.

WLSP						UniDic Lexeme		
Entry	Reading	Class	Article Number	Article		Lexeme	Reading	Class
小じゅうと	こじゅうと	体	1.2140	兄弟 Brother		小舅 Brother-in-law	コジュウト	体 Nominal
Sibling-in-law						小姑 Sister-in-law	コジュウト	体 Nominal

Finally, the alignment table is available: https://github.com/masayu-a/wlsp2unidic.

We also developed Windows wrapper GUI `ChaMame' for a morphological analyser MeCab to extract all possible WLSP article numbers. We demonstrate the GUI tools at the poster presentation.



	A	C	D	E	F	G	H	I
1	本日	体-関係-時間-現在						
2	,							
3	第	相-関係-類-等級・系列						
4	1	体-関係-量-数記号（一二三）;体-関係-量-数記号（一二三）;体-関係-量-順位記号（甲乙丙）;相-関係-類-異同・類似						
5	8	体-関係-量-数記号（一二三）						
6	9	体-関係-量-数記号（一二三）						
7	回	体-関係-量-助数接辞						
8	国会	体-主体-機関-議会						
9	の							
10	開会	体-関係-作用-開始;体-活動-交わり-集会						
11	式	体-関係-類-類・例;体-活動-言語-表・図・譜・式;体-活動-生活-行事・式典・宗教的行事						
12	に							
13	臨み	用-関係-時間-時機;用-関係-空間-方向・方角						
14	,							

A pilot study on the museum visitors interest by using eye tracking system

Emi Koseto-Horyu[1]

Introduction

Today, museums face many challenges, one of these is how to build a relationship between museum exhibitions and the visitors. Currently, understanding the exhibitions' effect to the visitors, questionnaires are often used. However, the number of visitors filling in the questionnaire is a specific type, and the result is considered not to show the overall trend of visitors. Therefore, an idea has come to the author that it could be possible to know what factors or contents affect and retain the visitors by using eye tracking system, since eye tracking gives the data that the visitors are looking for with unconsciously.

Previous eye tracking studies in museums

History of studies on eye tracking analysis is back to the early 20th century. Many studies using eye movements to investigate cognitive processes have appeared from mid-1970s. Rayner was reviewed about studies of eye movements in reading and other information processing tasks, such as music reading, typing, visual search, and scene perception (Rayner, 1998). However, in the field of museums, research using gaze measurement has become popular since the 21st century. In recent years, the number of researches using eye tracking measurement is increasing in museums to see how visitors see paintings. I introduce some examples here. Wooding and colleagues installed eye tracking system and collected data from over 5000 subjects looking at arts from the National Gallery (Wooding, 2002; Wooding *et al*.,2002). Though they used non-mobile eye tracking system, but nowadays, due to the spread of mobile sys tracking device, there are many studies in museums. Mayr showed a case study in a small museum exhibition to discusses the suitability of tracking visitors' eye movements as a method to explore mobile learning in museums. (Mayr et al.,2009). Massaro revealed that the movement of the viewer's gaze is different by the theme of paintings (Massaro *et al*.,2012). Ahmad, even the arts were displayed on the computer monitor, studied how the audience sees famous paintings, like "Mona Lisa" by Leonardo Da Vinci or so (Ahmad 2015). and Walker examined the eye movement behavior of children and adults looking at five Van Gogh paintings in the Van Gogh Museum, Amsterdam. (Walker *et al.*,2017).

Following these previous studies, in this study, I set two goals, one is to know what factors or contents are affective to the visitors by using eye tracking, the other is to know the data of eye tracking has any relevance to the result of the questionnaires. The examinations were done for the portal exhibitions (called as "the mobile museum") developed by the National Museum of Japanese History (NMJH) and the National Institute of for Japanese Language and Linguistic (NINJAL), aiming for contribute the social education. Since the portal exhibition system can be disassembled, becomes compact enough to be able to send by delivery service, it can be used at University, libraries, public spaces or so on for people don't have a chance to visit museums. This portal exhibitions were first held from 7th of May to 25th of May in the Kanagawa University with the aim of educating college students and the examination of eye tracking was also first used at there.

Method

The device of eye tracking that used is developed by Fujitsu, sold as "Eye Expert", small (length1.2cm, width 7.1cm and height 1.2cm) enough to make people unconscious of its existence. This eye tracking is based on measurement method called corneal reflex method, and by analyzing movement of eyes by irradiating the cornea with near infrared rays, we can know "where and how it looks". The devices were placed on the bottom of

[1] National Museum of Japanese History / National University of SOKENDAI (The Graduate University for Advanced Studies)

four panels used for the portable exhibitions, the geographic panel of the exhibition" Taiwan and Japan – the earthquake in modern era- ", the Ukiyo-e panel for the exhibition "The big earthquake in Edo period and ukiyo-e", the map panel for the exhibition "Japanese direct" and the panel for "Interactive language map". When people stood in front of the panel and looked at it, the device automatically found out the eye moving and send the data to PC. A questionnaire of the portal exhibitions was also placed to know the effect and to compare the data of eye tracking.

Result and future research

The results are shown as Figure 1 (below). The eye tracking heat map of" Taiwan and Japan – the earthquake in modern era- "indicates that the visitors are closely watching the place described in its explanation. However, in the heat map of "Interactive language map", it is known that viewers are paying attention to judicial precedents and Japanese maps rather than reading explanation. In addition, interestingly, although the result of the questionnaire indicates that the viewer is more interested in "Interactive language map" than" Taiwan and Japan – the earthquake in modern era- ", the viewer's gaze is closely watching" Taiwan and Japan – the earthquake in modern era- "longer, rather than "Interactive language map".

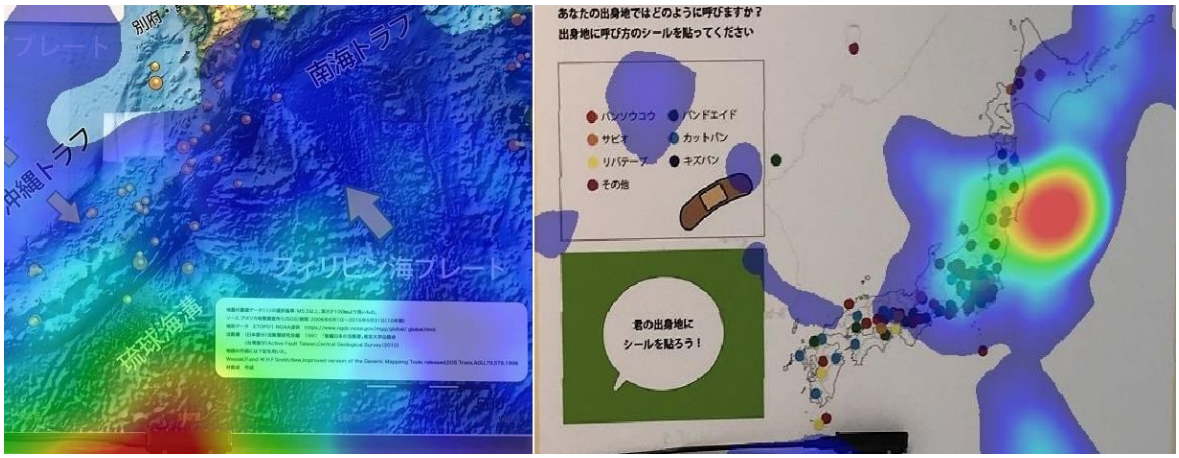


Figure 1. Heat map of eye movement. Lest is the map of Taiwan and Japan – the earthquake in modern era- "and right is the map of "Interactive language map"

Although, the amount of data is small at the present, better results can be obtained by accumulating data in the future. I will plan to be held the portable exhibition and the Osaka Museum of History in July. I also plan to make the examination for the permanent exhibitions of NMJH. By adding the data of these exhibitions and by comparing both the questionnaires and the eye tracking data, we could know the museum visitors interest and improve the exhibitions. I am convinced that it will contribute not only to exhibition but also social education.

Bibliography

[1] **Rayner, K.** (1998). "Eye Movements in Reading and Information Processing: 20 Years of Research." *Psychological Bulletin*, *124,* pp. 372-422.

[2] **Wooding, D. S., Muggelstone, M. D., Purdy, K. J., Gale, A. G.** (2002). "Eye movements of large populations: I. Implementation and performance of an autonomous public eye tracker." *Behavior Research Methods, Instruments and Computers*, *34*(4), pp.509-517.

[3] **Wooding, D. S.** (2002)." Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps." *Behavior Research Methods, Instruments and Computers, 34*(4), pp. 518-528.

[4] **Mayr, E., Knipfer, K., Wessel., D** (2009) "In-Sights into Mobile Learning an Exploration of Mobile Eye Tracking Methodology for Learning in Museums." *Researching mobile learning*" *Frameworks, methods, and research designs*, pp189-204

[5] **Massaro, D., Savazzi, F., Di Dio, C., Freedberg, D., Gallese, V., Gilli, G., Marchetti, A** (2012). "When Art Moves the Eyes: A Behavioral and Eye-Tracking Study. " *PLoS one, vol. 7, issue 5*, p. e37285

[6] **Ahmad,G.** (2015)." Eye Fixation curves along Analogical Thinking in Scene Viewing Eye Fixation curves along Analogical Thinking in Scene Viewing." *International Journal of Engineering and Industries (IJEI) Volume 6, Number 1* pp54 -62

[7] **Walker, F., Bucker, B., Anderson, NC., Schreij, D., Theeuwes J** (2017) "Looking at paintings in the Vincent Van Gogh Museum: Eye movement patterns of children and adults." *PloS one, vol,12 issue 6*, p. e0178912

In nihilum reverteris – retro text game

Robert Hellboj Straka[1], Yerzmyey[2], Piotr Marecki[3]

The proposed poster is devoted to the *In nihilum reverteris* (2018) retro text game developed for an 8-bit ZX Spectrum computer. The work represents an old school sci-fi text adventure game comparable with former text games from the 80' era. The game itself behave as interactive book with different pathways, the player can take during the gameplay. We used the assembly language to code the game engine. This low-level approach was used mainly to allow 64x24 characters per screen, faster user-game interaction and to play in-game music on the AY-3-8912 chip. The text is accompanied with graphical screens which are loaded during the game progress. Due to enormous size of the whole project and to minimize tape/disk operations, the game was developed for 128K versions of ZX Spectrum only and divided to two parts. Port of the game to the Raspberry PI (RPI) and other linux enabled architectures for which SDL library can be installed or compiled, AmigaOne, Atari 8-bit, Apple II (with enhanced graphics) is also part of the project. Several other ports are in preparation: Commodore 64, Atari ST. RPI version is based on the low-level SDL (Simple DirectMedia Level) library and developed in the C language. It is available as an open-source package to enable its compilation on every linux platform using provided Makefile. No graphics acceleration is needed and the game will run on the minimal resources where just frame buffer and sound subsystems are available and accessible to the SDL libraries. The artists used ZX Spectrum computer to create an advanced digital work in 2018, because of its constraints (it is said ZX Spectrum is the most limited computer among 8-bit platforms). The game is a collaboration beetween demosceners: Yerzmyey (music, text, graphic design) and Hellboj (code). Piotr Marecki is a producer of the work. *In nihilum reverteris* is one of the of the digital works on retro platforms developed in the UBU lab at the Jagiellonian University. The lab primarily produces digital works that can function in a few fields of the demoscene, electronic literature, video games and media art. The research conducted in the lab focuses on, among other things, local phenomena in the digital media field, e.g. strategies for cloning platforms in Central and Eastern Europe, as well as the digital genres and their specific features in Central and Eastern Europe. The artists, programmers and scholars affiliated with the lab develop new genres and communication practices (technical reports, open notebook science) to describe the creative process in its widest definition in the era of digital textuality. The project has been made possible through the support of the Polish Ministry of Science and Higher Education "National Programme for the Development of Humanities".

References

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

Christie, T. A. (2016). *The Spectrum of Adventure A Brief History of Interactive Fiction on the Sinclair ZX Spectrum.* Extremis Publishing.

[1] AGH University of Science and Technology
[2] independent artist
[3] Jagiellonian University

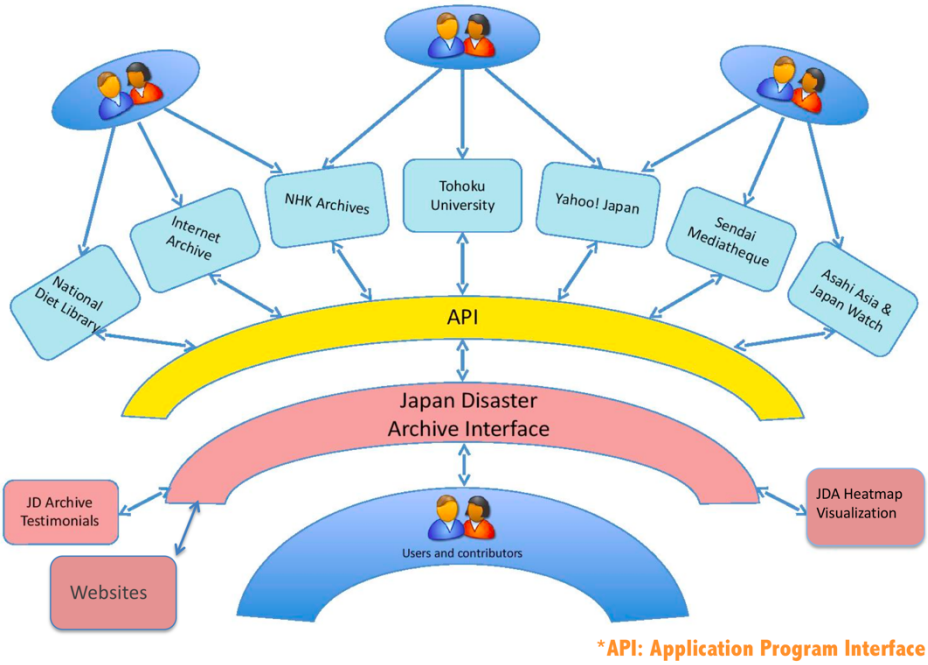
The Possibilities of a Participatory Digital Humanities Platform: A Case Study of the Japan Disasters Archive (JDA)

Andrew Gordon[1], Katherine Matsuura[1]

When triple disaster struck Japan in 2011 (earthquake, tsunami, and nuclear meltdown), the Reischauer Institute of Japanese Studies at Harvard University was one of many organizations that sprang into action through the capture and preservation of information. As outlined in a 2011 Reconstruction Report to Japan's Prime Minister, these steps to "record the disaster for eternity" have been both an attempt "to remember and honor the many lives that have been lost," as well as "draw lessons that will be shared with the world and passed down to posterity" (Reconstruction Design Council, 2011: 2). This has resulted, as of 2018, in a total of more than 60 disaster archives in Japan.

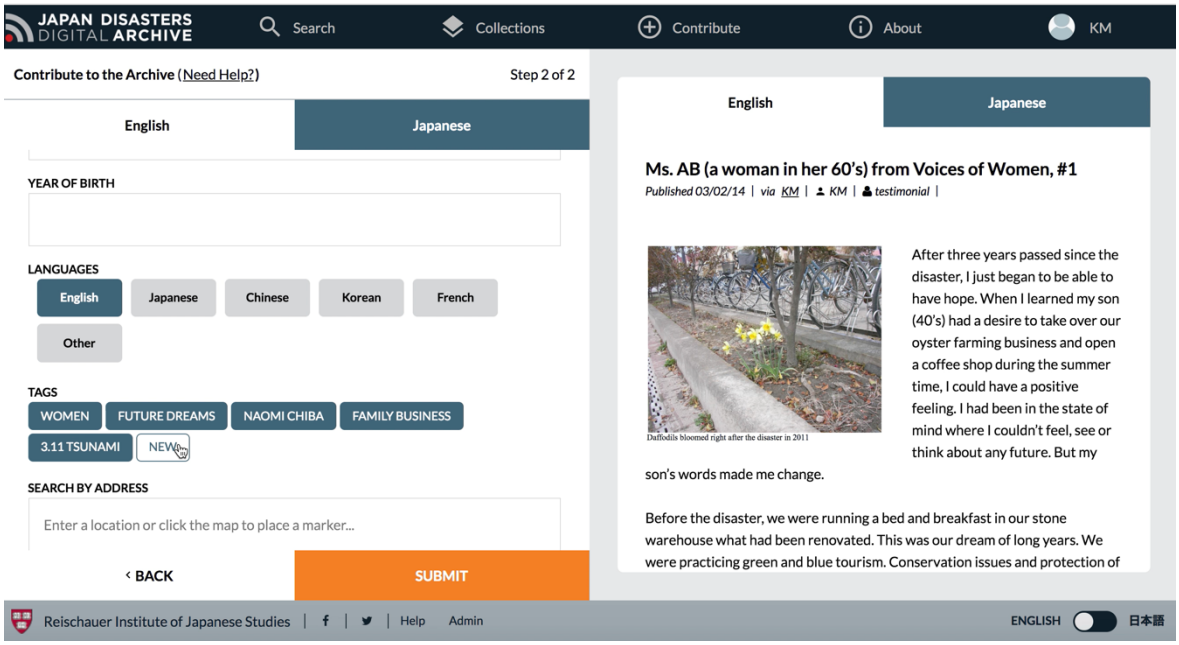
Given that Harvard University is physically remote from both Japan and the impacted Tohoku region, it was initially unclear whether a meaningful contribution was possible or realistic. As a project of Reischauer Institute, the Japan Disasters Digital Archive (JDA) had been archiving and preserving a number of tweets, testimonials, and full-text English news articles, but ultimately, it is not this content that make the project unique. Instead, the JDA has been focused on building a federation of existing Japanese digital archives and promoting various shared usages of the massive dataset being collected on the ground in Japan.

As of June 2018, the Japan Disasters Archive allows users to access more than 1.6 million items drawn from websites, images, video/audio content, documents, news headlines, and tweets in collaboration with more than a dozen institutions and organizations in Japan. The number of partners continues to grow even today, but JDA does not actually host the majority of this content. Instead, materials are linked via shared API in partnership with other archives. Although all partners benefit from an international platform and wider accessibility, the real appeal and value of the Japan Disasters Archive is the public space it provides for information sharing, collaboration, and conversation for citizens, researchers, students, and policy makers. It is an interactive space that encourages and thrives on user participation.

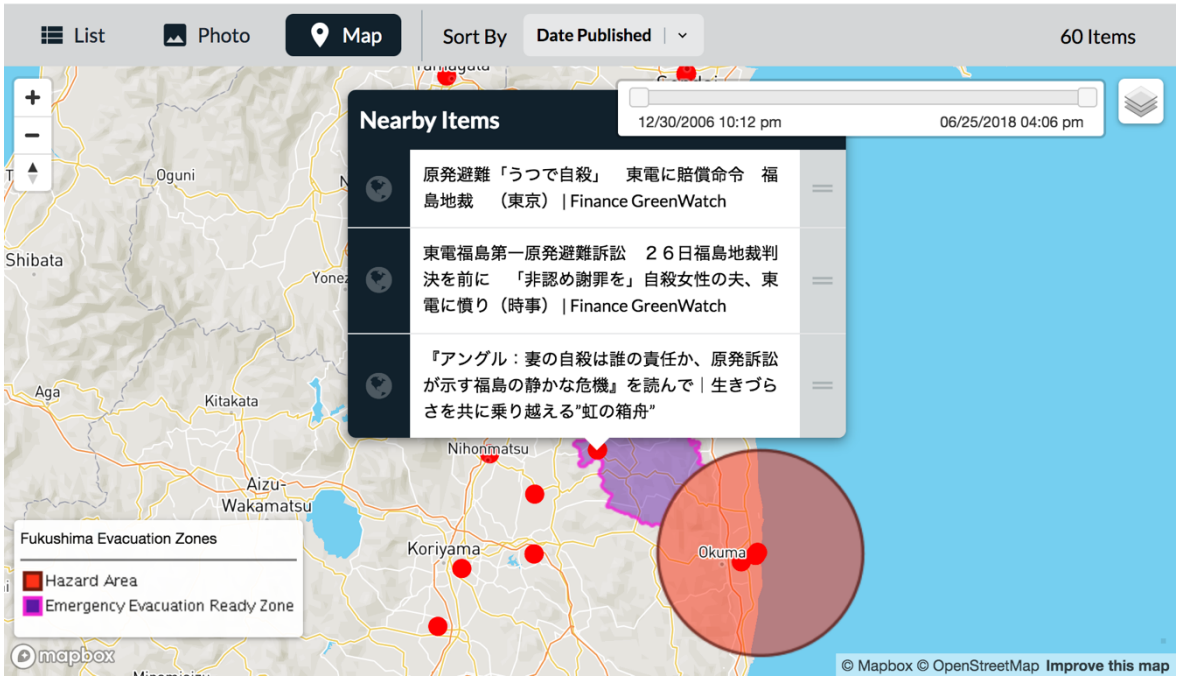


[1] Harvard University

The JDA works with crowd sourcing in a number of ways. The first feature is the ability for users to contribute materials and directly submit resources to the archive. This includes websites, videos, photographs, or testimonials related to personal experiences of disaster and its aftermath. Contributions to the archive can be made in multiple languages, using a submission form or bookmarklet. Users are also encouraged to add their own metadata and translate existing descriptions of the archived items into either English or Japanese.

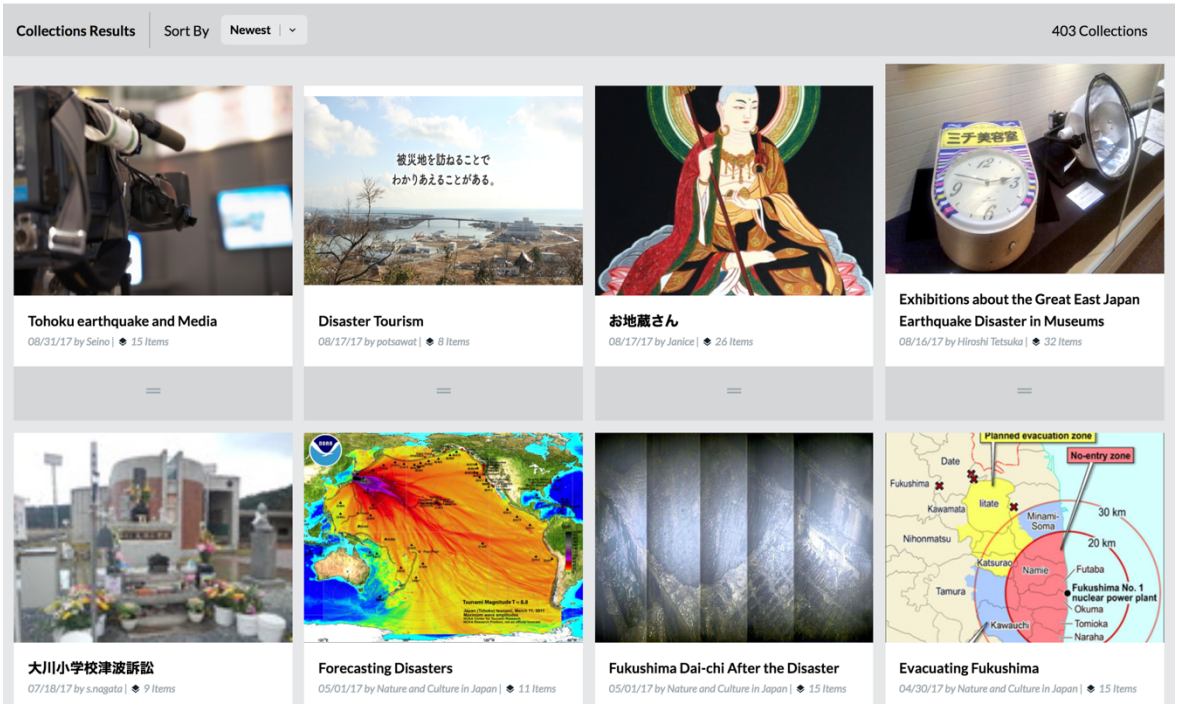


Discovery of JDA items is generated through keywords and user-generated tags, and it is also facilitated by an innovative heat map feature that visualizes all materials that are tagged with geographic information in real time.



Another way to participate with the JDA is through the creation of personal collections. Once a user creates an account, they are able to compile, curate, and annotate digital materials along a specific theme or area of interest, and this collection can then be shared with a wider audience if the user wishes. Through the evolution of this project, it has become clear that these “collections” are often part of an assigned course, and many

of the 400 public collections have been created by students both in Japan, as well as the United States.



The most recent feature is an enhancement of these collections and is an attempt to strengthen the pedagogical use of the JDA and any attempt to use its content for disaster prevention and awareness. Launched in time for the March 11, 2018 commemoration, any user is now able to showcase their collection as a full screen interactive, multimedia presentation. With the addition of this new feature, the Japan Disasters Digital Archive is attempting to incorporate the learning process used in coursework and take a more wholistic approach when thinking about the goals and purpose of a digital archive.

different types of O-Jizo-San

Published 09/16/2017 via Janice

References

Reconstruction Design Council. (2011). *Towards Reconstruction: Hope Beyond the Disaster.* [pdf] Tokyo: Ministry of Foreign Affairs of Japan. Available at https://www.mofa.go.jp/announce/jfpu/2011/7/pdfs/0712.pdf (accessed 25 Jun. 2018).

Digitizing Zeami

Hanna McGaughey[1]

Zeami (ca. 1363 - ca. 1443) was an actor, troupe leader, playwright, and composer who wrote treatises about the performing art known today as noh. These treatises relate his unique approach to the artistic process, focusing on practice as cultivation and on the need for understanding interpersonal relationships as the basis for successful communication. He describes performers' unique challenges and appropriate tasks for different stages in their lifetimes, thereby demonstrating his focus on cultivation[1]. His formulation "view from a removed perspective" (*riken no ken*) refers to the need for performers to relativize their own understanding of performance situations and consider audiences' perspectives[2]. These ideas suggest that Zeami considered artists' self-conscious grasp of their own contingency essential to successful artistry. However, modern Japanese scholarship since the early twentieth century heralds Zeami as an artistic genius worthy of representing Japanese culture, albeit an alternative to the predominant aristocratic culture of his time, and modern critical editions incorporate this view in their expositions. In digitizing influential modern commentary and examples from the pre-modern manuscript tradition, I aim to not only create an online critical edition of Zeami's work, but also to highlight and question the implicit views concerning art and culture that have informed previous editions.

Because Zeami's treatises were secretly transmitted by his heirs in various performer families, they were not widely available until Yoshida Tōgo's initial publications in 1908 and 1909[3]. In the century since then, Zeami's texts have been published in many critical editions. Nose Asaji presented the first annotated edition in two volumes published in 1940 and 1944[4]. Nose's commentary influenced Omote Akira's annotations in the current standard academic edition, volume twenty four in the *Nihon shiso taikei* series. About the relationship between expertise and success as a performer Zeami writes,

> In all matters, some people have a certain kind of proficiency and are recognized for it. Sometimes others, even though they have attained a superior artistic rank, are unable to manage this particular thing[5].

Nose claims that this "certain kind of proficiency" refers to "inborn natural talent" (*umaretsuki no tenpin*) and Omote agrees that it refers to "inborn, personal advantage" (*umaretsuki mi nitsuite iru chōsho*)[6]. Both notes suggest that certain individuals are born with unique artistic aptitude. Notions of artistic creativity at the time of the treatises' discovery were influenced by the Hegelian notion that an artist with genius could ensure the inherent unity in a work of art that scholars such as Ernest Fenollosa introduced to Japan as a modern, scientific norm[7]. I want to know to what extent the reasons for Zeami's appeal in early twentieth century Japan and international scholarship shaped the consequent analysis of his work. While a digital critical edition will enable machine searches of secondary as well as primary text, the principle advantage of a digital critical edition will be that it opens the field for alternative annotations by a community of researchers and students.

I have begun transcribing Nose's edition and supplementary manuscripts and marking up both using TEI. To enable thorough machine searches of the texts, I need to either mark all character and script variations or use a search engine with a fluctuating search function (Jp. *yuragi kensaku kinō*). As I proceed, I intend to collate the texts and make them available through an online user interface. To facilitate community collaboration and discussion about alternative annotations, I will initially build an online forum before considering further options for text manipulation such as administrative status for select individuals. At the beginning of this project, various issues concerning text ownership and copyright have presented themselves. Because Omote's academic edition is covered by

[1] University of Trier

copyright until 2060, Nose's edition provides the initial basis for the project. The performer families who still own most of the important manuscripts carefully protect access even in online archives because copyright does not directly protect their own artistic expertise[8]. Use of manuscript images will require carefully navigating relationships with Japanese researchers, archivists, and performers. Because the copyright status of all early translations of these texts is murky at best, I welcome suggestions for building a system similar to Google books that provides text snippets as search results. Including collated translations would clarify the broader consequences of editorial decisions in writing annotations and would make the texts accessible to a wider audience. This poster will introduce the motivation and initial design of the project and key challenges in its implementation.

Bibliography

Hare, Tom, trans. *Zeami Performance Notes.* New York: Columbia University Press, 2008.

"Kōekishadan hōjin Nōgakukyōkai | Chosakuken kanren nitsuite." Accessed June 25, 2018. http://www.nohgaku.or.jp/copyright/index.html.

Nose Asaji, ed. Zeami jūroku bushū hyōshaku. 2 vols. Tokyo: 1940 and 1944.

Omote Akira, ed. *Zeami Zenchiku.* Tokyo: Iwanami, 1974. See esp. "Fūshikaden dai ichi nenrai keiko jōjō," 15-20, and "Kakyō," 84-109.

Rimer, J. Thomas. "Hegel in Tokyo: Ernest Fenollosa and His 1882 Lecture on the Truth of Art." In *Japanese Hermeneutics: Current Debates on Aesthetics and Interpretation.* Honolulu: University of Hawai'i Press, 2002.

Yokoyama Tarō. "Zeami hakken: Kindai nōgakushi ni okeru Yoshida Tōgo *Zeami jūroku bushū* no igi nitsuite," *Interdisciplinary cultural studies* 9 (2004), 19-45.

Yoshida Tōgo, ed. *Zeami jūroku bushū.* Tokyo: Nōgakukai, 1909.

[1] Omote Akira, ed., "Fūshikaden dai ichi nenrai keiko jōjō,"in *Zeami Zenchiku* (Tokyo: Iwanami, 1974), 15-20.

[2] Cf. Omote Akira, ed., "Kakyō" *in Zeami Zenchiku* (Tokyo: Iwanami, 1974), 88.

[3] Yoshida Tōgo, ed., *Zeami jūroku bushū* (Tokyo: Nōgakukai, 1909).

[4] Nose Asaji, ed., *Zeami jūroku bushū hyōshaku* (Tokyo: 1940 [vol. 1] and 1944 [vol. 2]).

[5] Tom Hare, trans., *Zeami Performance Notes* (New York: Columbia University Press, 2008), 42.

[6] Nose, ed., *Jūroku bushū hyōshaku* 98 and Omote, ed., *Zeami Zenchiku* 32.

[7] Yokoyama Tarō, "Zeami hakken: Kindai nōgakushi ni okeru Yoshida Tōgo *Zeami jūroku bushū* no igi nitsuite," *Interdisciplinary cultural studies* 9 (2004), 37-38 and J. Thomas Rimer, "Hegel in Tokyo: Ernest Fenollosa and His 1882 Lecture on the Truth of Art" in *Japanese Hermeneutics: Current Debates on Aesthetics and Interpretation* (Honolulu: University of Hawai'i Press, 2002): 105.

[8] "Kōekishadan hōjin Nōgakukyōkai | Chosakuken kanren nitsuite," accessed June 25, 2018, http://www.nohgaku.or.jp/copyright/index.html.

Building Linguistically and Intertextually Tagged Coptic Corpora with Open Source Tools

So Miyagawa[1], Amir Zeldes[2], Marco Büchler[3], Heike Behlmer[1], Troy Griffitts[4]

* This work has been supported by joint funding from the National Endowment for the Humanities (NEH grant HG-229371) and Deutsche Forschungsgemeinschaft (DFG project 273503199), and funded by Deutsche Forschungsgemeinschaft's Collaborative Research Centre 1136 "Education and Religion in Cultures of the Mediterranean and Its Environment from Ancient to Medieval Times and to the Classical Islam," B05 "Biblical Interpretation and Educational Traditions in Coptic-speaking Egyptian Christianity of Late Antiquity: Shenoute, Canon 6."

Coptic is the last stage of the Egyptian language. Before Coptic, Ancient Egyptian was written in Hieroglyphs, Hieratic, and Demotic scripts. Starting in the third century CE (excluding "Old Coptic"), Coptic used an alphabet based on the Greek and several added Demotic letters. A large but understudied corpus of literary texts exists in Coptic, including important Gnostic, monastic and Manichaean texts, as well as early Biblical translations. Efforts to build a digital Coptic corpus are still in their initial phases. In this paper, we present the most recent work in a partnership of Digital Humanities projects. Coptic SCRIPTORIUM (Schroeder and Zeldes, 2016) is a major initiative endeavoring to put corpora online which are linguistically and philologically annotated (i.e. supporting grammatical, paleographical and literary annotations), while projects in Göttingen are producing digital editions of Coptic texts focusing on philological standards and critical editions: A project at the Göttingen Academy of Sciences and Humanities is preparing a complete digital edition of the Coptic Old Testament (Behlmer and Feder, 2017), and in a project of Collaborative Research Centre 1136 "Education and Religion" digital diplomatic editions of selected works of Shenoute and Besa, 4th-5th century abbots of the White Monastery in Upper Egypt, are being prepared for text reuse research (see below). Based on our experiences, we have schematized workflows for building Coptic corpora with linguistic and literary information by using open source programs, merging data from OCR (Optical Character Recognition) and transcription sources, Natural Language Processing (NLP) tools, and manual annotation interfaces allowing for the correction of automatic tool output.

Digital transcriptions of Coptic texts are acquired in several ways, taking care to target either out-of-copyright editions or diplomatic transcriptions of manuscripts, both of which can be made freely available under Creative Commons licenses. For data not yet available in digital transcription, we adopted OCRopus, an open-source, language-independent neural network-based OCR program first developed by Thomas Breuel (Bulert et al., 2017). Our OCRopus model, trained for Coptic print editions, achieves a high accuracy rate close to 97%. The open-source data is available at the GitHub repository of the KELLIA project (https://github.com/KELLIA/CopticOCR, accessed 6 July 2018).

In addition to OCR data, we are working to offer consistent representations of already digitized texts. Pioneers of Coptic DH such as Tito Orlandi have accumulated digital transcriptions using old ASCII-based fonts that display the Latin alphabet in a Coptic font. Since 2013, Coptic SCRIPTORIUM has undertaken to convert and re-publish such

[1] University of Göttingen / Deutsche Forschungsgemeinschaft, Collaborative Research Centre 1136 "Education and Religion in Cultures of the Mediterranean and Its Environment from Ancient to Medieval Times and to the Classical Islam," B 05 "Biblical Interpretation and Educational Traditions in Coptic-speaking Egyptian Christianity of Late Antiquity: Shenoute, Canon 6"
[2] Georgetown University
[3] University of Göttingen
[4] Göttingen Academy of Sciences and Humanities

data, as well as digitizing new texts in Unicode. A Unicode converter for old ASCII font encodings is available on the SCRIPTORIUM website. Using converted texts, OCR data and new transcriptions, a broad collection of digital Coptic texts has been produced.

To validate the results of both conversions and of new digital transcriptions, we use two freely available annotation interfaces: the Virtual Manuscript Room developed by Troy Griffitts (VMR, https://vmrcre.org/, accessed 6 July 2018, see Griffitts, 2017), and GitDox (Zhang and Zeldes, 2017, https://corpling.uis.georgetown.edu/gitdox/, accessed 6 July 2018), an annotation environment optimized for correcting linguistic annotations.

The VMR editor enables a team to produce online diplomatic and critical editions, using digital manuscripts images. In this phase, we can correct the errors of the OCR, the digital or the original transcriber. Moreover, one can tag philological information appropriate for Coptic and Greek manuscripts and export the data in a TEI XML format. The GitDox interface offers XML validation options as well, but also includes a spreadsheet-based interface, which makes it easy to view aligned annotations at the levels of word forms, phrases, and sentences. It is being used to correct automatic part-of-speech (POS) tagging, lemmatization, and morphological analysis, among other things.

Coptic Scriptorium's NLP pipeline (Zeldes and Schroeder, 2016) provides automatic linguistic analyses, including a morphological tokenizer for the highly agglutinative complex word forms used in Coptic, a lemmatizer linked to the Coptic Dictionary Online (https://corpling.uis.georgetown.edu/coptic-dictionary/, accessed 6 July 2018), automatic POS tagging, language of origin detection for Greek loan words, and a syntactic dependency parser which outputs annotations in the Universal Dependencies scheme (http://universaldependencies.org/, accessed 6 July 2018). All of these tools are trainable, meaning they could also be used to benefit automatic analysis in other languages with similar challenges. The Coptic texts of the manuscripts are in *scriptio continua*. Many modern editions, however, insert spaces between phrases known as bound groups, similarly to the analysis of complex space-delimited word forms in modern Arabic and Hebrew, or related ancient language varieties, such as Biblical Hebrew, Classical Arabic, or Syriac. Thus, the tokenization and "word"-segmentation are a key component for Coptic NLP. Coptic SCRIPTORIUM's tools currently achieve an average of 98.82% correct boundary detection, or 94.87% perfectly segmented bound groups in tokenization. Based on this tokenization, we tag the linguistic categories mentioned above, i.e. POS tagging, syntactic parsing, lemmatization and language of origin, which can then undergo manual correction. The data is exported in a number of formats, including EpiDoc XML (Bodard and Stoyanova, 2016), TreeTagger SGML (Schmid, 1994) and Paula XML (Dipper, 2005), all of which are well documented open standards. Finally, one can visualize the corpus with linguistic and philological tags using ANNIS (Krause and Zeldes, 2016), a search and visualization platform for richly annotated corpora which is currently in use for a variety of Digital Humanities corpora.

With annotated corpora at hand, we have focused on applications such as text reuse detection and visualizing intertextuality among Coptic monastic texts. The latter incorporates quotations from other texts considered authoritative, especially from the Coptic translation of the Bible. The eTRAP research group of the University of Göttingen has developed TRACER (Büchler et al., 2018 forthcoming), a program to detect text reuse, especially in classical or historical languages. Using TRACER we have found previously undetected quotations of the Bible in selected works of the abbots Shenoute and Besa. The information gained is visualized using the TRAViz program (Jänicke et al., 2015) and can be represented in ANNIS.

The means to build a Coptic corpus using only open data and tools are currently only available for Sahidic, the main literary dialect of Coptic in Late Antiquity. Extending this work to other dialects, we ultimately hope to provide standards of Natural Language Processing for the entire Coptic literary corpus.

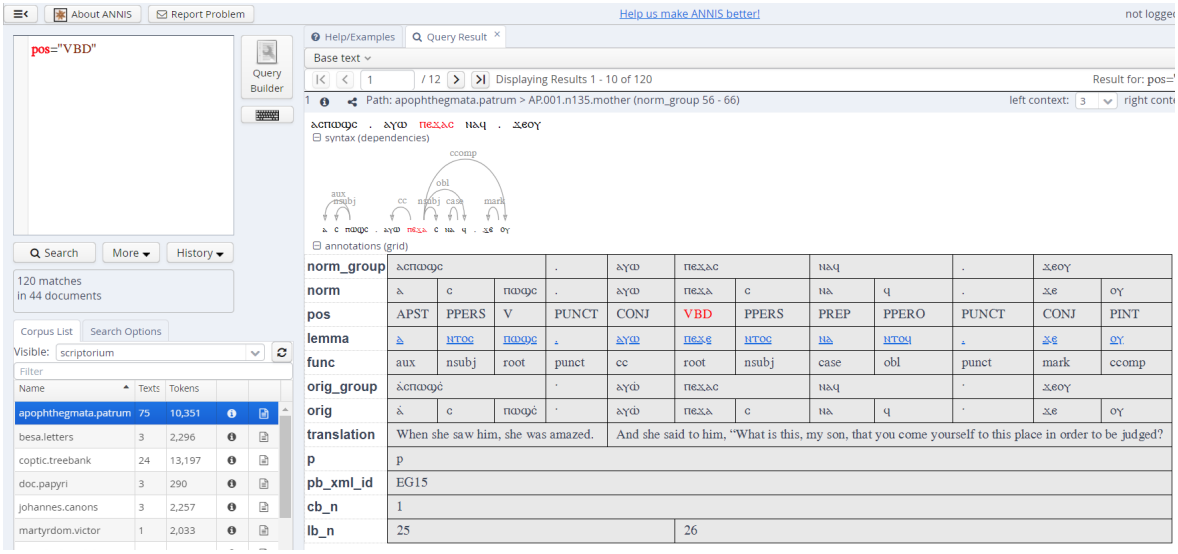


Figure 1: Coptic XML corpora visualized by ANNIS from Coptic SCRIPTORIUM

References

Behlmer, H. and Feder, F. (2017). "The Complete Digital Edition and Translation of the Coptic Sahidic Old Testament. A New Research Project at the Göttingen Academy of Sciences and Humanities." *Early Christianity*, 8: 97–107.

Bodard, G. and Stoyanova, S. (2016). "Epigraphers and Encoders: Strategies for Teaching and Learning Digital Epigraphy." In Bodard, G. and Romanello, M. (eds) *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange & Public Engagement.* London: Ubiquity Press, pp. 51–68.

Büchler, M., Franzini, G., Franzini, E., Moritz, M. and Bulert, K. (2018 forthcoming). "TRACER - a Multilevel Framework for Historical Text Reuse Detection." *Journal of Data Mining and Digital Humanities* (Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages).

Bulert, K., Miyagawa, S. and Büchler, M. (2017). "Optical Character Recognition with a Neural Network Model for Printed Coptic Texts." In Rhian L. et al. (eds) *Digital Humanities 2017 Conference Abstracts.* pp. 657–9.

Dipper, S. (2005). "XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation." In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, pp. 39-50.

Griffitts, T. (2017). *Software for the Collaborative Editing of the Greek New Testament.* Ph.D. Thesis, University of Birmingham.

Jänicke, S, Geßner, A, Franzini, G., Terras, M., Mahony, S. and Scheuermann, G. (2015). "TRAViz: A Visualization for Variant Graphs." *Digital Scholarship in the Humanities* (Digital Humanities 2014 Special Issue).

Krause, T. and Zeldes, A. (2016). "ANNIS3: A new architecture for generic corpus query and visualization." *Digital Scholarship in the Humanities 2016*, 31. http://dsh.oxfordjournals.org/content/31/1/118 (accessed 6 July 2018).

Schmid, H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of International Conference on New Methods in Language Processing.* Manchester.

Schroeder, C. T. and Zeldes, A. (2016). "Raiders of the Lost Corpus." *Digital Humanities Quarterly,* 10(2). http://www.digitalhumanities.org/dhq/vol/10/2/000247/000247.html (accessed 6 July 2018).

Zeldes, A. and Schroeder, C. T. (2016). "An NLP Pipeline for Coptic." In Proceedings of LaTeCH 2016 - *The 10th SIGHUM Workshop at the Annual Meeting of the ACL.* Berlin, pp. 146–55.

Zhang, S. and Zeldes, A. (2017). "GitDOX: A Linked Version Controlled Online XML Editor for Manuscript Transcription." In *Proceedings of FLAIRS 2017, Special Track on Natural Language Processing of Ancient and other Low-resource Languages*. Marco Island, Florida, pp. 619–23.

Transitions of Plot Elements in a Japanese Detective Comic

Hajime Murai[1]

Introduction

In the field of literary studies, research on narratives structures and story patterns is called "narratology." Various studies conducted in the field of narratology have been based on humanities studies. For instance, Propp insisted that the functions of character roles in stories can be categorized based on a few patterns in specific genre (31 story functions and 7 character roles in Russian folktales) (Propp, 1968). Greimas hypothesized that the structures of general stories can be divided and symbolized by certain elements. In addition, these elements can be categorized into few symbols (Greimas, 1966). Other than that, Genette analyzed the stylistic and semantic differences between the sequences in the story narration and their chronological order (Genette, 1972).

Moreover, based on those research results, there have been some attempts to generate stories automatically by utilizing computers and various other information technologies. In order to generate stories, it is necessary to evaluate the naturality and interestingness of the sequences of the plot elements. To resolve this issue, the What If Machine Project developed a database of common consequences for various situations and made stories by selecting logically connectable events (The What-If Machine Project, 2013). In another case, the Kimagure AI Project made templates of general story structures to generate various easily understandable stories (Toyosawa et al., 2018). However, there is only few case studies of empirical data about the sequences based on existing story works. Therefore, it is not clear how the sequences in stories can be made natural and interesting.

In this study, the detective story genre was selected in order to analyze plot sequences quantitatively. Detective stories are highly patternized, include many short stories, and are suitable for the purpose of collecting many sample works. Therefore, detective stories seemed to be adequate for this study. The plot elements from the detective stories were symbolized in order to analyze the frequencies and patterns of plot elements in the sequences. By describing plots as sequences of machine-readable symbols, the plots transitions, initial and final situations, and whole narrative structures could be analyzed. By using these quantitative data, it is possible to extract frequent patterns to compute the transition probability of plot elements. This could be one of the foundations for automatic plot generation by computers in the future.

Target Contents

In this research study, the Japanese detective comic series "Case Closed" (the Japanese title is "Meitantei Conan") was targeted for analysis (Aoyama, 1994). "Case Closed" is a best-selling detective story series and is famous as a successful multimedia property in Japan. "Case Closed" has been serialized in the comic magazine "Weekly Shōnen Sunday" since 1994, its episodes' number more than 1000, and the latest comic book installment was volume 94 (published in December 2017). The total number of copies printed are estimated to be more than 100 million. This work has been aired as an ongoing TV anime series for more than 20 years (since 1996). "Case Closed" includes many homages to traditional detective stories such as those of Arthur Conan Doyle and Agatha Christie. The plot style of each case follows that of traditional detective stories. Moreover, because the main reader demographic of the series is young boys, the story structure tends to be very simple and easily understandable, with explicit explanations of the tricks and mysteries employed. For this research study, the comics of volumes 1 to 45 were analyzed as the data source (volumes in 10 years). Because the number of episodes differed from the

[1] Future University Hakodate

number of actual cases, 134 cases out of 461 episodes were targeted for analysis. On average, 3 to 4 episodes were found to correspond to one case.

Categorization of Plot Elements

Although there are many methods to categorize story plots based on certain elements, this research study focused on the general functions in detective stories in order to describe plot elements. The macro-level functions of stories were utilized for dividing and categorizing plot elements manually. The story functions for 134 cases were then described and synthesized on the basis of the frequently appearing typical functions of detective stories. As a result, the study depicted 134 cases with 47 types of plot elements. The 47 types of plot elements are shown in table 1. A total of 1123 plot elements were obtained. Each case includes 8.4 plot elements on average.

Patterns of Plot Transition

In order to investigate the sequential relationships between each plot element, the transitions between those plot elements were visualized as a direct network (Figure 1) by utilizing Graphviz (Ellson et al., 2003). For visualization purposes, the numbers of sequential transitions for two different plot elements were aggregated in 134 cases, and frequent transitions (more than 3) were utilized as edges in the network. Each plot element was a node for the network. The nodes' font sizes were changed in proportion to the frequency of each plot element.

Table 1. Types and number of appearances of plot elements

Investigation	145	Preliminary notice (of crime)	10
Reasoning	144	Explosion	9
Confession	98	Surrender	8
Appearance of Cadaver	95	Intimidation	8
Introduction of People	65	Reinforcement	8
Past Cases	57	Closed Circle	7
Arrest	54	Forgiveness	6
Discovery	35	Witness	6
Request of Investigation	33	Abduction	6
Fight	32	Strange Incident	6
Incursion	32	Vigilance	6
Escape	32	Theft	5
Invitation	26	Serious Condition	4
Travel	25	Contest of Detective	4
Pleasures	19	Arson	3
Murder	18	Attempted suicide	3
Exposed Identity	17	Ambush	3
Induction	16	Concealment	3
Chase	15	Appearance of detective	2
Dining	11	Visiting	1
Death	11	Shopping	1
Unaccounted	11	Suicide	1
Confinement	11	Accident	1
Quarrel	10		

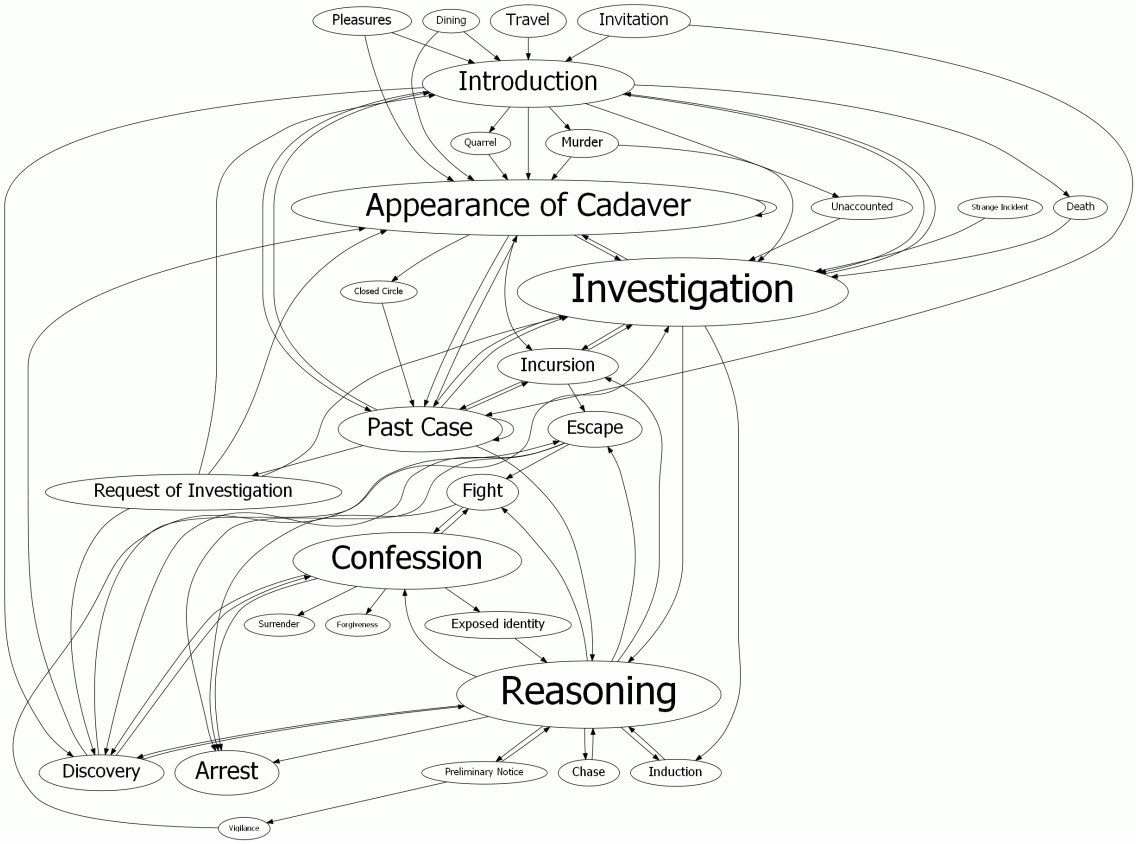


Figure 1. The transition network of plot elements in the detective story

Figure 1 shows that the fundamental narrative patterns can be symbolized as the transitions of typical functions in detective stories. Based on frequently appearing plot elements, it is clear that the common story pattern in the famous Japanese comic detective series “Case Closed” can be described as a sequence consisting of “Introduction,” “Appearance of Cadaver,” “Investigation,” “Past Case,” “Reasoning,” “Confession,” and “Arrest.” On the other hand, based on the quantitative structure of this transitional network, human-like plot sequence generation could be enabled by utilizing mathematical methodologies such as the random walks in the Markov chain model.

Conclusions and Future Work

This study conducted the data description of 134 detective stories and listed 47 types of plot elements in order to realize a computational narratological analysis. Moreover, the transitions between various plot elements were visualized as a network. Although the objectivity of symbolization needs to be verified by other coder, this computational narratological result could be one of the foundations of automatic story generation by artificial intelligence in the future.

Aoyama, G. (1994). *Case Closed*, Weekly Shōnen Sunday, Shogakukan.

Ellson, J., Gansner, E., Koutsofios, E., North, S., and Woodhull, G. (2003). "Graphviz and dynagraph - static and dynamic graph drawing tools," *Graph drawing software*, Springer-Verlag, pp. 127-148.

Genette, G. (1972). *Discours du recit in Figures III*, Èditions du Seuil.

Greimas, A. J. (1966). *Sémantique structurale : recherche de méthode*, Larousse.

Propp, V. (1968). *Morphology of the Folktale*, University of Texas Press.

The What-If Machine Project. (2013). “The What-If Machine,” http://ccg.doc.gold.ac.uk/research/whim/resources/poster_UC.pdf, Last accessed 12 Apr 2018.

Todorov, T. (1977). “The Typology of Detective Fiction,” *The poetics of prose*, pp. 42-52.

Toyosawa, S., Kudou, H., Ishita K., Endou, S., Kawase, R., Kikuchi, R., Kudou, K., Kurihara, M., Sakurai, K., Satou, Y., Tamaki, H., Nemoto, Y., Harashima, M., Hisano, T., Hirata, I., Murai, H., Tsubakimoto, Y., Sumi, K., and Matsubara, H. (2018). "Development and Evaluation of an Interactive System That Automatically Generates and Visualizes Detective Novel Plots," *SIG Technical Report Computer and Humanities*, 2018-CH-116(13): 1-5.

Open Data as the Essentials of Teaching and Textual Research

Susan Allés-Torrent[1], Mitsunori Ogihara[1]

This paper arises from our experience in teaching digital humanities (Brier, 2012), at the undergraduate level, in an initiative shared among participants from Modern Languages and Literature, English, and Computer Science at the University of Miami in Florida, USA. We will share some of the challenges we have encountered, and stress the usefulness of open data for conducting research in linguistic corpora and teaching students the necessary skills for the research, including text mining and topic modeling. In the field of digital humanities, an important question is the universality of the method used, that is, the applicability of an approach developed for studying a corpus to the study of another corpus. Furthermore, the scalability of the approach used for much larger data sets (how much resources the approach consumes as the data size grows) is a question. Although large textual corpora with appropriate text encoding do exist, large datasets that serve the purpose are hard to come by. Thus, to study the question of scalability, one may try assembling various data sets. The need for combining smaller data sets to generate a large data set raises a question of adapting existing, possibly heterogeneous open data sets to a certain data format into one aggregate data set. Unfortunately, the field lacks open textual data sets, specifically in non-English languages, and this fact hinders both our research and teaching.

While the use of free-access data sets is attractive, scholars have been careful about making their scholarly data available to anyone to access for free as well as using data collected by others (Borgman, 2012). The hesitation to share data with other scholars stems both from the difficulty in harvesting and formatting data and from the lack of standard practices for acknowledging the effort of others. The hesitation to use someone else's data comes from the fear and the difficulty of using data of unknown format and of unknown quality. To promote open data scholarship in humanities, the field must develop practices that benefit the creators as well as the users.

One of the key issues that we face is data formatting. Important points in data formatting are as follows:

- *Preservation of the source of the raw data.* For future changes in data formatting and for verifying reproducibility, it is important to keep a copy of the raw. However, most of the times raw data itself cannot be published due to copyright issues.
- *Preservation of the strategies and tools used for processing raw data.* In some research areas, software may quickly come and go. For reproducibility purposes again, it is very important to save a copy of the software tools and preserve platforms on which the software tools could be run. The strategies are not like tools and it is important to document the process implemented in processing the data.
- *Using a generic scheme for formatting the open data component.* Using an XML format may come natural to the humanists, but in many cases, it is important that the definitions of the attributes and the data structure are made very clear.
- *Text encoding.* For non-English corpora, characters with diacritics and special symbols must be chosen with some careful standardizations. Also, there are multiple character codes that represent quotation marks and punctuation marks. Creators should state which symbol scheme is used, and stick to the scheme. The use of well-defined data formats makes it easier for the users to incorporate the data for their own analysis.

[1] University of Miami

Another important issue is data harvesting. If multiple possible sources exist for raw data, a source must be chosen to minimize the effort and maximize the amount of data obtained. The reasoning and the process for choosing the data source must be well documented.

Digital humanities scholars must be fully aware of these issues when we take upon the task of developing and/or using open data sets. We question, as educators, if we must teach our students the awareness of these issues? When we use data sets in our, we always pose questions such as:

- Where does the raw data come from?
- If there is a data set on the Internet that can be downloaded by hand or using a computer program, can a scholar go ahead and download it? If not, how should a scholar approach the owner of the raw data?
- What if the owner of the raw data refuses to permit the use of the data?
- If the raw data can be acquired or generated for a fee, how should a scholar garner funding for that?
- What is the benefit of sharing the data with others that was obtained with funding and/or with substantial human effort?
- If research has been carried out with non-open data sets, how much of the research process and findings can be disclosed, and how much can be shared with others?

Presently at the University of Miami, teaching of these issues is taking place not in multidisciplinary settings, but in domain specific settings (that is, in foreign language and in computer science courses and projects). The questions mentioned in the above are asked at appropriate points in student project: when the students start on their digital humanities scholarly projects tasks and when their projects reach certain milestones. This is quite challenging, since the set of questions to be addressed is not common among the projects, and thus, for the students to be exposed to a fuller set of issues, they will need multiple projects. In the present teaching environment, an effective way to build awareness of these issues appears to be through forums in which the students share their experience with others.

References

Borgman, C. L. (2012). The Conundrum of Sharing Research Data, Advances in Information Science, 63(6):1059-1078.

Brier, S. (2012). Where's the Pedagogy? The Role of Teaching and Learning in the Digital Humanities, Gold, M. K. (ed), Debates in the Digital Humanities. Minneapolis: University of Minnesota Press.

The Italian reception of the English Novel. A digital enquiry on Eighteenth Century literary journalism

Andrea Penso[1]

This short paper aims at showing the first results of the ongoing collaborative research project *The reception of the English novel in the Italian literary press between 1700 and 1830: a transcultural enquiry into the early shaping of the modern Italian literary and cultural identity*. The research focuses on an existing corpus of data relative to the publication, dissemination, translations, critical reviews, and editorial advertisements of English novels in Italian literary newspapers and journals of the time. The main purpose of the project is to uncover how the English novels were introduced to the Italian readership through literary journalism with the application of Digital Humanities methodologies of investigation. One of the project goals is in fact to create a methodological paradigm that may be extended to the study of the reception of English novels in the literary journalism of other nations. The present paper therefore has two primary objects:

a) To show for the first time to the public the first research output: an open access, bilingual, and annotated digital repository, which consists of a Drupal-based software for corpora, and represents an immediate way to develop the research. The first step of the project has been the cataloguing, analysis, and digitization of the corpus of reviews. This preliminarily created digital database allows the subsequent computational, textual and critical surveys. The creation of the database allows also to understand the "genealogical dimension" of the Italian reviews, i.e. the comparative analysis of reviews that were taken from French or English periodicals and made their way into the Italian press. These sources have already been identified, and will allow to understand the extent of the influence French and English journalism had on the Italian press and to outline the specific Italian input. The text encoding of the reviews makes possible to point out the elements that are original and innovative with respect to the foreign reviews of the time which the Italian press copied from, often adapting the contents. The relational database makes also possible to focus on the re-interpretations of Italian reviewers who drew on the Italian literary tradition but challenged its subjects, genres and linguistic structures.

b) To illustrate the two main lines of approach that will be applied in order to digitally explore the corpus. The first consists of a stylistic and linguistic analysis of the reviews, which will be pursued equalizing and comparing stylistic and lexical constellations belonging to different discursive practices from a number of periodicals and journalist. Digital stylometry, word frequency and statistical analyses tools such as *R, MiniTab* and *Intelligent Archive* will be used during this phase. The study of the readers' response to the contents, spread by the novels via the reviews, is deeply connected to the stylistic analysis of the reviews. In fact, the outlining of the reviews' stylistic features is crucial to understanding in which ways the contents were revealed to the public, and how the audience was influenced in the perception of the moral values and the social messages of the novels. The second line of approach will concern the spatial analysis of the data, which will be mapped thanks to *GIS* (Geographical Information System) digital tools integrated with Geo-criticism. The analysis spatial analysis allows the visualization of popular reading trends in 18th and early 19th century Italy, and will be supported by a careful survey of sources and their circulation, with a methodological approach combining Material Culture and Philology.

[1] Vrije Universiteit Brussel

Sustainable Metadata Management for Cultural Heritage Image Data using XMP

Oliver Pohl[1]

When managing large quantities of data, it is a common solution to utilize a centralized data management software to forge a connection between metadata and the data objects themselves. In case of text-based objects without any attached metadata, it is easy for humans to contextualize these objects by recognizing patterns such as filenames, titles, authors etc. This task becomes a challenge when dealing with non-text-based objects like images in the cultural heritage domain. Without metadata or expert knowledge, it becomes difficult to estimate the creation date of a painting or tell the name of its painter. Thus, the ability to contextualize data depends on whether there is a working connection between the metadata store and the data object itself. This connection fails as soon as the file is moved on the file system without having these changes also applied in the corresponding data base, or when the file is shared without a reference to its original location. This paper presents an approach to overcome that type of co-dependency by utilizing XMP to embed cultural heritage metadata directly into image files to ensure their location-independent long-term preservation. The "Corpus Vitrearum Medii Aevi" Germany (CVMA) project serves as an example use-case.

CVMA Germany is a joint project by the Berlin-Brandenburg Academy of Sciences and Humanities and Academy of Sciences and Literature Mainz, with teams sitting in Berlin, Potsdam, Mainz and Freiburg. The focus of CVMA is to catalog and document stained glass windows situated in German churches built throughout the medieval and early modern period. Furthermore, the project publishes collections of stained glass photographs including metadata in an online image archive (CVMA Deutschland, 2018). The most fundamental data are uncompressed high-resolution photographs of the church windows. For proper documentation of these photos, the CVMA Germany developed its own metadata schema (CVMA Deutschland, 2016), building upon Dublin Core (Weibel et al., 1998) and IPTC (2014), and implementing its own CVMA namespace, which focuses on stained glass research in art history. When entering metadata, the project also makes use of authority files such as the GND (Deutsche Nationalbibliothek, 2018) for storing information about creators and donors, or geonames to avoid any unambiguous data and create compatibility with the linked open data cloud.

To maintain readability of its image files in the future, the CVMA embeds metadata directly into the files, thereby ensuring their readability across platforms without running into any co-dependency with an additional management layer, thus creating a union of the image and its metadata.

Regarding long-term preservation, the CVMA is guided by patterns employed in the domain of text-based document preservation of PDF files. When creating PDF documents, any fonts, images and additional metadata are embedded into the document itself so users can access the contents as intended without the use of third party software or data (Adobe Systems, 2006).

In both cases, CVMA and PDF, metadata are embedded using the eXtensible Metadata Platform (XMP) (Adobe Systems, 2012). When using XMP, it is possible to define your own metadata schemata and embed metadata as RDF/XML into any file without having any effect on its readability (Bright, 2006). For instance, when embedding metadata into a TIFF file, image viewers with XMP capabilities can show these additional data, while viewers without these capabilities remain unaffected and show the image as if it would not contain any additional data. XMP has seen a surge of usage in digital asset management systems (Regli, 2009) because of the easy way of jointly transporting files and metadata.

There are different workflows on how to embed and read XMP metadata throughout the different project branches in Germany. The Freiburg team utilizes

[1] Berlin-Brandenburg Academy of Sciences and Humanities

exiftoolGUI (2018), a graphical user interface wrapper for the EXIF- and XMP manipulation command line tool exiftool (Phil Harvey, 2018). Both are free open source software products and can be configured to make use of custom metadata standards. However, these tools are only helpful when editing small sets of data or when searching across a small pool of files, since they lack any management capabilities.

The team in Potsdam uses the proprietary application FotoStation (Fotoware, 2018). In contrast to exiftoolGUI, FotoStation comes with management and search features. Moreover, it is very easy to employ custom metadata standards and configure metadata input forms by drag and drop, making it very user friendly. Unfortunately, there are only few options when it comes to software that is capable of embedding XMP metadata into files with the option of utilizing custom metadata standards.

When it comes to online display, the images are sent to the CVMA digital online image archive. On ingest, the back-end extracts metadata into a relational data base for efficiency purposes.

On the one hand the utilization of XMP comes with some challenges and obstacles. For instance, modifying just a single metadata field in a large TIFF file can be very resource intense, since the whole image needs to be re-written. Additionally, although the metadata is now embedded into the images, it takes strict file management guidelines that determine how to name and where to store files in order to avoid creating a messy file collection. Moreover, it is still a challenge to integrate XMP-image-files into research data stores that handle heterogeneous types of data from varying domains of the sciences and humanities (Grunzke et al., 2016).

Utilizing XMP has proven to be resourceful with the CVMA project for multiple reasons. First, ensuring long-term preservation has become a matter of just backing up the files. Secondly, due to its RDF/XML based nature, XMP makes it easy to extract and map metadata into different metadata schemata for easier re-use in other online image archives. Additionally, it also becomes easily feasible to expose the metadata to the linked open data cloud. As for now, XMP had a very positive effect on the workflows within the CVMA project by consolidating the (re-)use of existing metadata vocabularies, by using authority files and creating a new metadata standard for church windows for other related projects to re-use.

Bibliography

Adobe Systems (ed). (2006). PDF Reference: sixth edition. https://wwwimages2.adobe.com/content/dam/acom/en/devnet/pdf/pdf_reference_archive/pdf_reference_1-7.pdf (accessed 4 May 2018).

Adobe Systems (ed). (2012). XMP Specification Part 1: Data Model, Serialization, and Core Properties. https://wwwimages2.adobe.com/content/dam/acom/en/devnet/xmp/pdfs/XMP%20SDK%20Release%20cc-2016-08/XMPSpecificationPart1.pdf.

Bright, J. (2006). First steps: XMP. *Journal of Digital Asset Management*, 2(3–4): 198–202 doi:10.1057/palgrave.dam.3650025.

CVMA Deutschland (2016). Corpus Vitrearum Deutschland: XMP Metadatenspezifikation Version 1.1 http://www.corpusvitrearum.de/cvma-digital/spezifikationen/cvma-xmp/11.html (accessed 7 May 2018).

CVMA Deutschland (2018). Digitales Bildarchiv - Öffentliche Betaversion: Corpus Vitrearum Deutschland http://www.corpusvitrearum.de/cvma-digital/bildarchiv.html (accessed 7 May 2018).

Deutsche Nationalbibliothek (2018). Gemeinsame Normdatei http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html (accessed 7 May 2018).

exiftoolGUI (2018). exifToolGUI http://u88.n24.queensu.ca/~bogdan/ (accessed 7 May 2018).

Fotoware (2018). FotoStation | Blazing fast photo management software https://www.fotostation.com/ (accessed 7 May 2018).

Grunzke, R., Hartmann, V., Jejkal, T., Prabhune, A., Herold, H., Deicke, A., Hoffmann, A., Schrade, T., Meinel, G. and Herres-Pawlis, S. (2016). Towards a metadata-driven multi-community research data management service. Rome, Italy.

IPTC (2014). *IPTC - NAA Information Interchange Model Version 4*. International Press Telecommunications Council http://www.iptc.org/std/IIM/4.2/specification/IIMV4.2.pdf (accessed 7 May 2018).

Phil Harvey (2018). ExifTool https://www.sno.phy.queensu.ca/~phil/exiftool/ (accessed 7 May 2018).

Regli, T. (2009). The state of digital asset management: An executive summary of CMS Watch's Digital Asset Management Report. *Journal of Digital Asset Management*, 5(1): 21–26 doi:10.1057/dam.2008.49.

Weibel, S., Kunze, J., Lagoze, C. and Wolf, M. (1998). Dublin Core Metadata for Resource Discovery. http://www.rfc-editor.org/info/rfc2413 (accessed 4 May 2018).

The Visualization of Academic Inheritage in Historical China

Yong Qiu[1], Jun Wang[1], Hongsu Wang[1]

Introduction

The visual analysis of academic inheritage of scholars and their students' features can help researchers to understand the evolution of their impact quickly. Many current humanities researches focus on the qualitative sampling analysis. Therefore, we proposed and designed an academic analysis platform that users are able to search scholars' successors of multiple generations and study their basic features in a flexible and quantitative way.

Data and Method

The biographical data of the scholars and their relationships with successors come from the China Biographical Database Project[1] (CBDB). CBDB is a freely accessible relational database consisting of Chinese Historical biographical information up to 417,000 individuals, primarily from the 7th through 19th centuries. We extracted the Teacher-Student relationship from the Scholarship catalog in CBDB social association types, including Menren 門人, Student and Disciple 弟子. Those students followed their teacher's ideas in a formal way which are convinced to be called academic successors.

The academic visualization platform is organized in four parts:

Firstly, we used tree map to represent Teacher-Student relationship of all generations until no successors can be searched, so that we can organize all the relationships in a strict hierarchy. However, there were some challenges. The first challenge is that some individuals had multiple teachers. The second challenge is that, we found a very special case. In this case, two scholars apprenticed with each other so much. Finally, they became both teachers and students with each other. This relationship can't be presented by the tree map, so we only represented the relationship once. Another way to illustrate Teacher-Student relationship is network map. This kind of graph loses strict hierarchy while comparing to tree diagram. However, the advantage is that those individual with multiple teachers can be easily picked up from network diagram. (see Figure. 1)

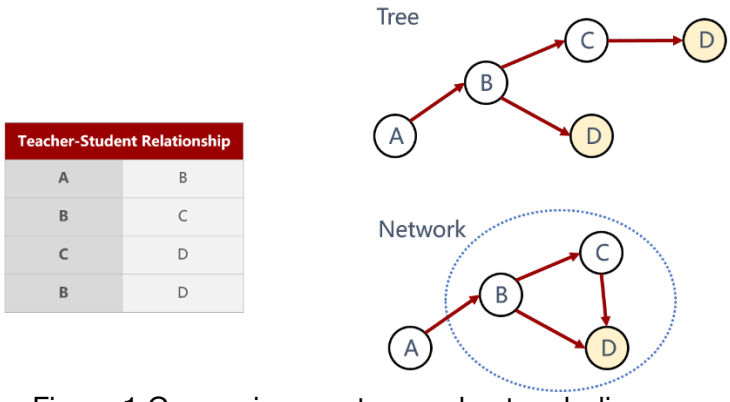


Figure 1 Comparison on tree and network diagram

Based on the construction of academic tree map and network map, we added two useful functions. The first function is that the nodes in academic tree diagram can be clicked to collapse and expand if next-generation successors exists. The second function is that the nodes in academic network diagram can be clicked to show the corresponding biographical information which offered by CBDB API. These two functions provide our users with an interactive way to explore their own interests.

[1] Peking University

Secondly, we visualized the geographical distribution of academic successors. Geographical distribution can reflect the geographical changes of scholars' impact. All the successors with basic address affiliation were encoded with scatters on map and they can be filtered by generation. The color of scatter corresponds to its generation from cool colors to warm which helps us to see that whether scholars' academic prosperity is close to his contemporary era.

Another group feature is the posting rank distributions. In premodern China, the students with better performance in academia are more competitive in their official career. Due to the complex and detailed bureaucratical system, Dr.Hu from department of history of Peking University divided the official titles to five levels by official ranks. Finally, we illustrate it with bar charts.

Results

We set up a website to let users to explore the visualization of the Scholars' academic successors in historical China(http://dh.kvlab.org/cbdb_vis/part2_dizinet_eng.html). The website is programmed with Bootstrap.js, Echarts.js and Python.

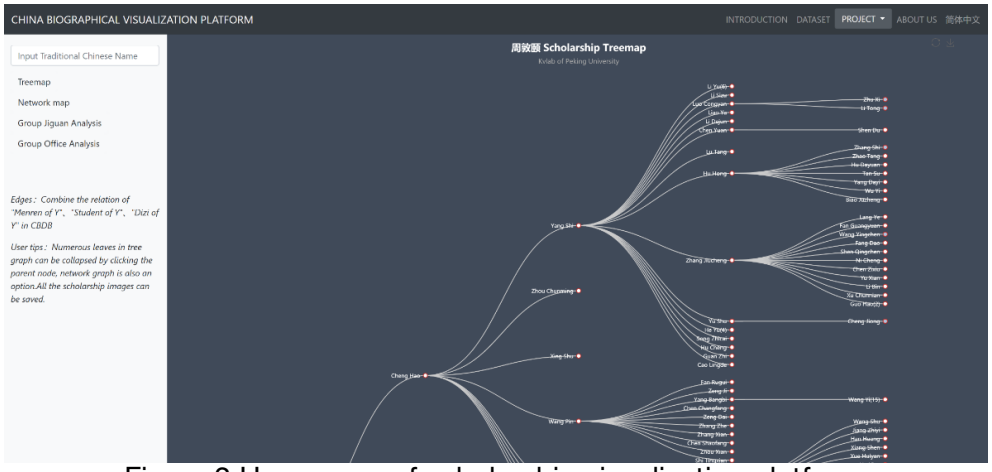


Figure 2 Homepage of scholarship visualization platform

Take Zhou Dunyi（周敦頤）as an example, as a famous and important new confucianist in the Song Dynasty, his academic ideas successfully continued to 9th generation and started to boom from 4th generation because of Zhu Xi（朱熹） who is one of the most famous Confucians in Chinese history.

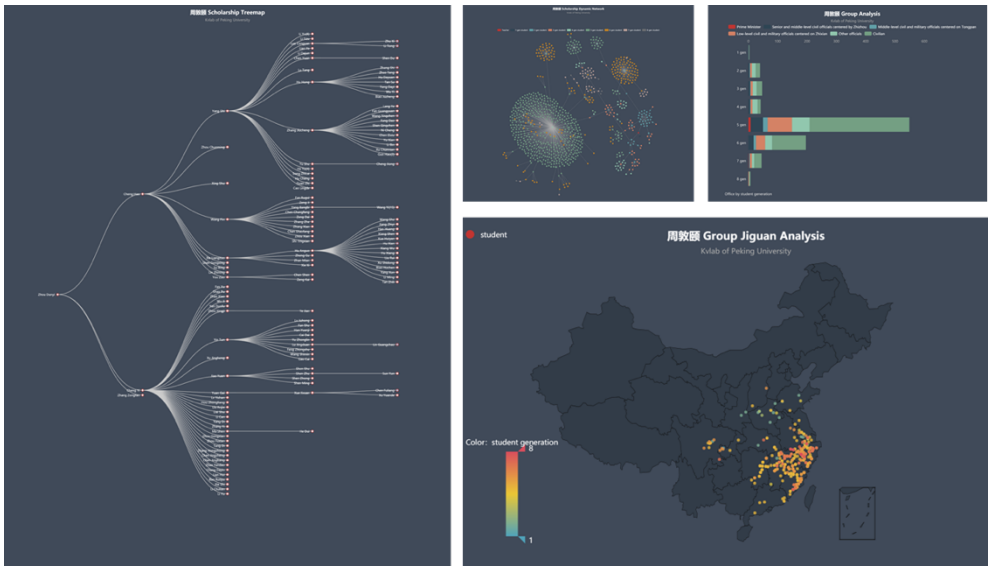


Figure 3 Academic Visualization of Zhou Dunyi

Conclusions and Future Work

In this paper, we introduce our work on visualizing the Scholars' academic successors in premorden China. In the future, we will discover the relation between academic and relatives or some other social relations.

References

[1] **Harvard University, Academia Sinica, and Peking University** (2018), China Biographical Database, https://projects.iq.harvard.edu/cbdb.

The Visualization of the Historical People's Migration in Tang Dynasty

Yong Qiu[1], Jun Wang[1]

Introduction

This paper discussed how to apply large-scale visualization and quantitative dynamical tools to reproduce the evolution of cultural and political centers of the Tang dynasty. We build a dynamical migration network to visualize the migration traces of the historical people from their birthplace to their death place across the Tang dynasty of China. To assess the ancient city's influence in old age, we propose a new calculating index based on the PageRank algorithm.

Data and Method

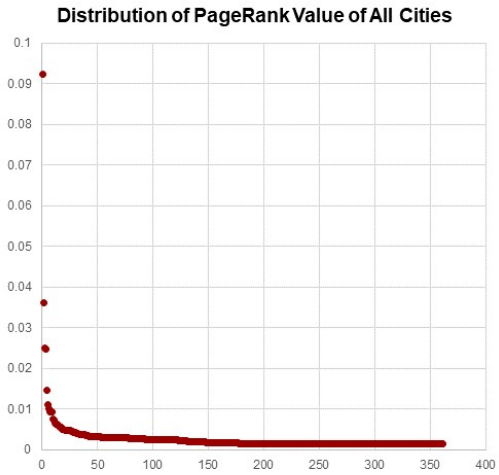
The biographical data of the people in Tang Dynasty comes from the China Biographical Database Project[1](CBDB). The CBDB is a freely accessible relational database consisting of biographical information up to 417,000 individuals, primarily from the 7th through 19th centuries. The biographical data recorded in the CBDB system are mainly extracted from published indices, such as Wang Deyi's revised *Index to Biographical Sources for Song Figures* and similar works; Some are from online databases or from studies of text sources.

The outline of the visualization work is as follows.

Firstly, we select the persons with birth and death information available in CBDB. Among those 854 individuals, the male to female ratio is about 8:2 and the official to non-official is about 3:7.

Secondly, we connected 361 cities with 854 directed curves, and each of which represents the migration path of a historical individual from his birthplace to his death place. The death place not only indicates its cultural attraction but also reflects the political dominance to notable people.

We then measure the influence of the ancient cities based on the structure of the migration network. We calculate the PageRank value of each node in the network. The PageRank[2] is an algorithm used by Google Search to rank websites in their search engine results. The PageRank works by counting the number and quality of links to a page to determine a rough estimation of its importance. Similarly, the influence of a city can be calculated by the number and quality of cities linked. Here is the top 20 influential cities in Tang Dynasty. (see Figure. 1)



Top20 Influential Cities by PageRank Value

Rank	NodeId	PRValue	CityName	CityName_CHN
1	14693	0.0923	Luoyang	洛陽
2	14494	0.0362	Chang'an	長安
3	14714	0.0251	Heyang	河陽
4	14492	0.0249	Wannian	萬年
5	15118	0.0146	Luzhou	潞州
6	14510	0.0112	Jingyang	涇陽
7	15692	0.0102	Jiangdu	江都
8	14937	0.0093	Zhencheng	甄城
9	15218	0.0093	Linlv	林慮
10	15890	0.0077	Hongzhou	洪州
11	15128	0.0073	Xiangxian	鄉縣
12	15501	0.0067	Xiangyang	襄陽
13	15445	0.0064	Jianglingfu	江陵府
14	16374	0.0064	Guangzhou	廣州
15	15963	0.0061	Shezhou	歙州
16	400616	0.0059	Yongchang	永昌
17	15348	0.0056	Cangzhou	滄州
18	15758	0.0056	Wuxian	吳縣
19	14788	0.0051	Zhengzhou	鄭州
20	15948	0.0051	Xuanzhou	宣州

Figure 1 Calculating PageRank Value of Each City

After the above calculation, we made the visual encoding on the migration network and cities. Migration paths are represented as directed curves on the Tang map

[1] Peking University

of A.D. 741. For those people whose death year are not available, their index years are used instead. The Index year (the “sixtieth year of age”) is a figured value that aims to help researchers locate a person in time for computational purposes. Beside migration paths, the size of dot representing a city in the map is in proportion to the number of people died there minus the number of people born there. Those places with more deaths are colored in orange, those with more births are colored in blue.

At last, we produce an animation by drawing the migration paths one by one with the shift of time from 620 to 940. The interactive and playable timeline enable audiences to view the historical migration in different time scale. (see Figure. 2)

Results

We set up a website to let users to explore the visualization of the historical migration of Tang Dynasty (http://dh.kvlab.org/cbdb_vis/part1_migrate_tang_eng.html). The website is programmed with Bootstrap.js, Echarts.js and Python. For the first century and a half the Tang Dynasty (from 618 to 756), the empire was expanding outward from its base in Chang'an and its secondary capital in Luoyang. It is clearly showed in the migration network that Luoyang was more attractive than Chang'an.

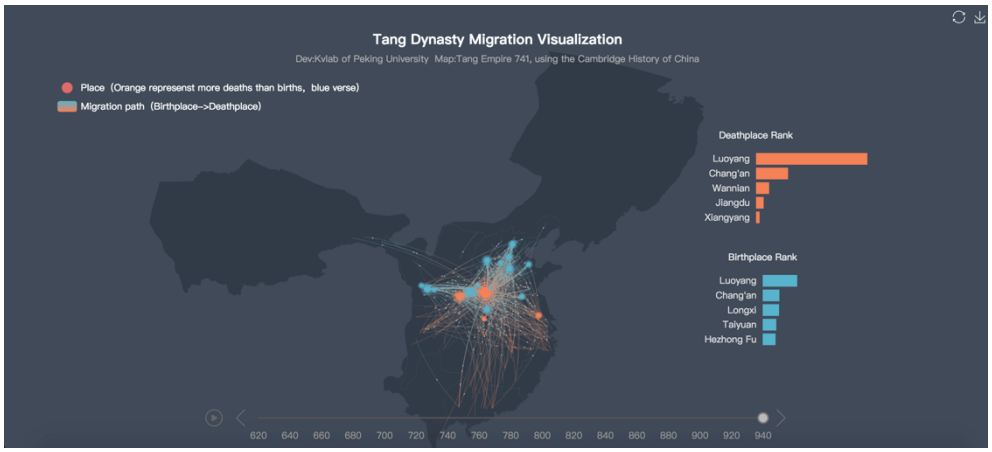


Figure 2 Migration Network of Tang Dynasty

It can be watched in Figure 2 that the area below the Yangzi River (Inc. Jiangnan Dao, Huainan Dao and Jiannan Dao) attracted a large amount of notable individuals. It corresponds to the fact that Jiangdu was the start of canal in Tang dynasty.

Conclusions and Future Work

In this paper, we introduce our work on visualizing the migration of historical people in Tang dynasty, and provide some explanation to the visualized pictures with historical facts. There are certainly a lot more important Tang persons in history. Due to the limited number of visualized Tang persons, we will absorb more representative data to support for a historical conclusion.

References

[1] **Harvard University, Academia Sinica, and Peking University** (2018), China Biographical Database, https://projects.iq.harvard.edu/cbdb.

[2] **Page, L.** (1998). The pagerank citation ranking : bringing order to the web. Stanford Digital Libraries Working Paper, 9(1), 1-14.

Machine learning approaches for background whitening and contrast adjustment of digital images

Wataru Satomi[1], Toru Aoike[1], Takeshi Abekawa[2], Takanori Kawashima[1]

INTRODUCTION

The National Diet Library (NDL) has a large collection of digitized materials[1] and is actively involved in researching way to improve usability of these materials. Presently, we are focusing on the development of techniques that utilize machine learning for enhancing digitized images.

In this study, we present two approaches to enhancing the readability of digitized materials.

BACKGROUND

Some digitized materials have poor readability due to the condition of the original materials, limitations of the digitization process, or a combination of both. For example, images digitized from microfilm are generally processed as grayscale images, but their background color is prone to be gray rather than white. Similarly, the paper used in printed matter often turns yellow or exhibits other discoloration. The low contrast inherent in such materials reduces their readability. Even newer materials are often difficult to digitize when it is necessary to preserve the color gamut for pictures. In such cases, the text tends to appear gray rather than black even when the background is white.

Manual adjustment of the contrast or brightness of an image is always possible, but the optimal parameters are not always obvious and, particularly for inherently low contrast images, can only be found by trial and error.

We examined two techniques for addressing these issues. The first was the use of generative adversarial networks (GAN) to whiten the background of digitized images, and the second was the use of semantic segmentation to effect contrast adjustments automatically.

Whitening the background of digitized images using GAN

METHOD

We used pix2pix[2], a general-purpose solution for image-to-image translation that utilizes GAN. Given pairs of input and output images as training data, the pix2pix model learns to mapping from input to output, and is then capable of generating an output image from an input image based on the learned mapping. For our purposes, the input images are low contrast images, and the output images are high contrast ones, which means that background is lighter and text is darker than the input image. It is worth noting that this technique differs from binarization, which often results in text with jagged edges.

Oversized input images are split into smaller tiles of 256 × 256 or 512 × 512 pixels, which are used as input tiles. The output tiles are then concatenated to form an output image.

Although we used color images as training data, our target images are primarily monochrome printed pages, so the output was set to grayscale.

TRAINING DATA

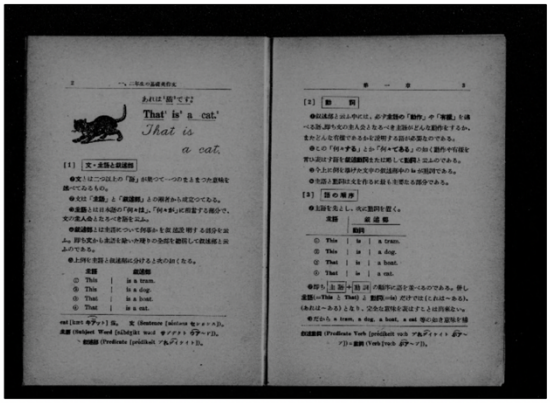
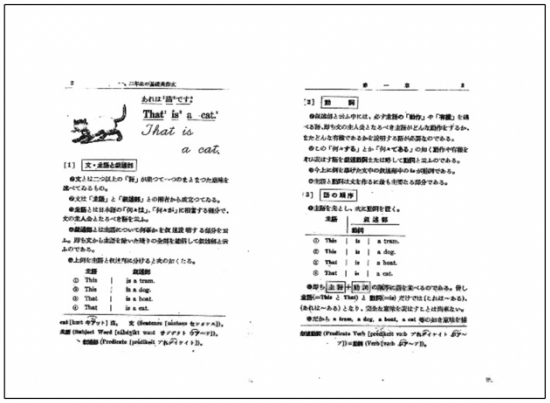
A total of 4,555 images from 26 library materials for which copyright had expired were selected for use as input images for training. The contrast and brightness of the input images were adjusted manually to create output images for training.

[1] National Diet Library
[2] National Institute of Informatics

RESULT

We tested 89 images that were not used for training to confirm that this technique successfully enhanced legibility and print quality. We further confirmed that no text was deleted unintentionally nor was any false text generated. As shown in Table 1, however, this technique is not an effective means for correcting diagrams or photographs.

Table 1: Whitening the background of digitized images using GAN

Input image	Output image
	
	

Adjusting the contrast of digitized images using semantic segmentation

METHOD

We adopted a two-step approach, in which we first used DeepLab V3+[3] to extract text regions from digitized materials per semantic segmentation. This was achieved by having the semantic segmentation network label each pixel as one of multiple predefined types. For our purposes, the pixels were labeled text, image, or background. The second step involved application of a grayscale transformation to the text regions, after which discriminant analysis[4] is used to determine the threshold of binarization. This threshold value is then used to perform a sigmoid function that adjusts the contrast. By implementing this process in two steps, it is possible to find optimal parameters, free of the influence of diagrams or background.

TRAINING DATA

Our document image dataset comprised 210 images extracted from 143 randomly selected magazines from our digitized materials. These images were classified into three types of regions: illustrations, text, and background. Training was performed with 180 training images and 30 validation images by randomly cropping 300 × 300 pixel patches from the images for use in training DeepLab V3+.

RESULT

Contrast adjustment can be performed more quickly than the regeneration of an enhanced image, and therefore is well suited for enhancing readability when patrons are viewing digitized images via a web service.

As can be seen in Fig. 1, we were able to identify text regions in digitized materials with an accuracy of roughly 85% in 55 images from the dataset of the ICDAR2009 Page Segmentation Competition[5].

Fig. 2 shows how the text regions can be enhanced using different techniques. There are, however, also some materials that are too dark overall, indicating that there is still room for improvement in the functionality and parameters used for contrast adjustment.

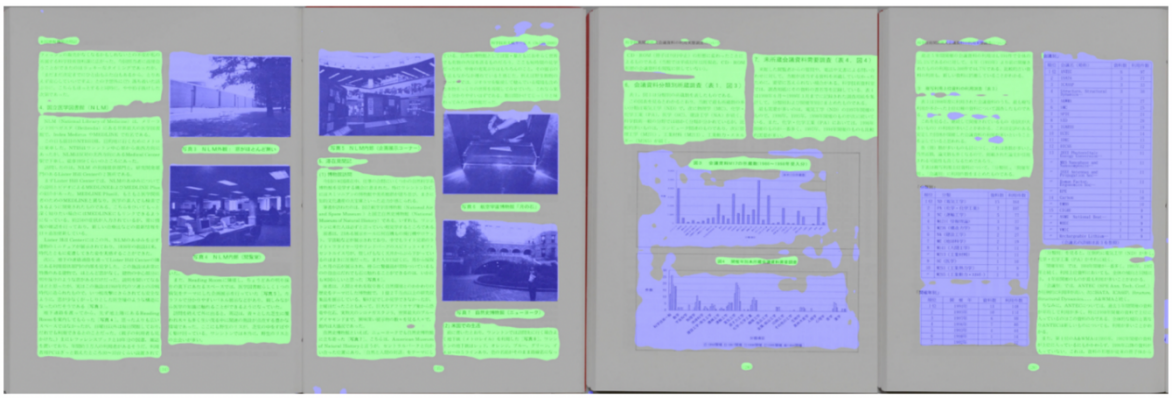


Figure 1: Results of segmentation using DeepLab V3 plus
Blue areas are illustrations and green areas are text.

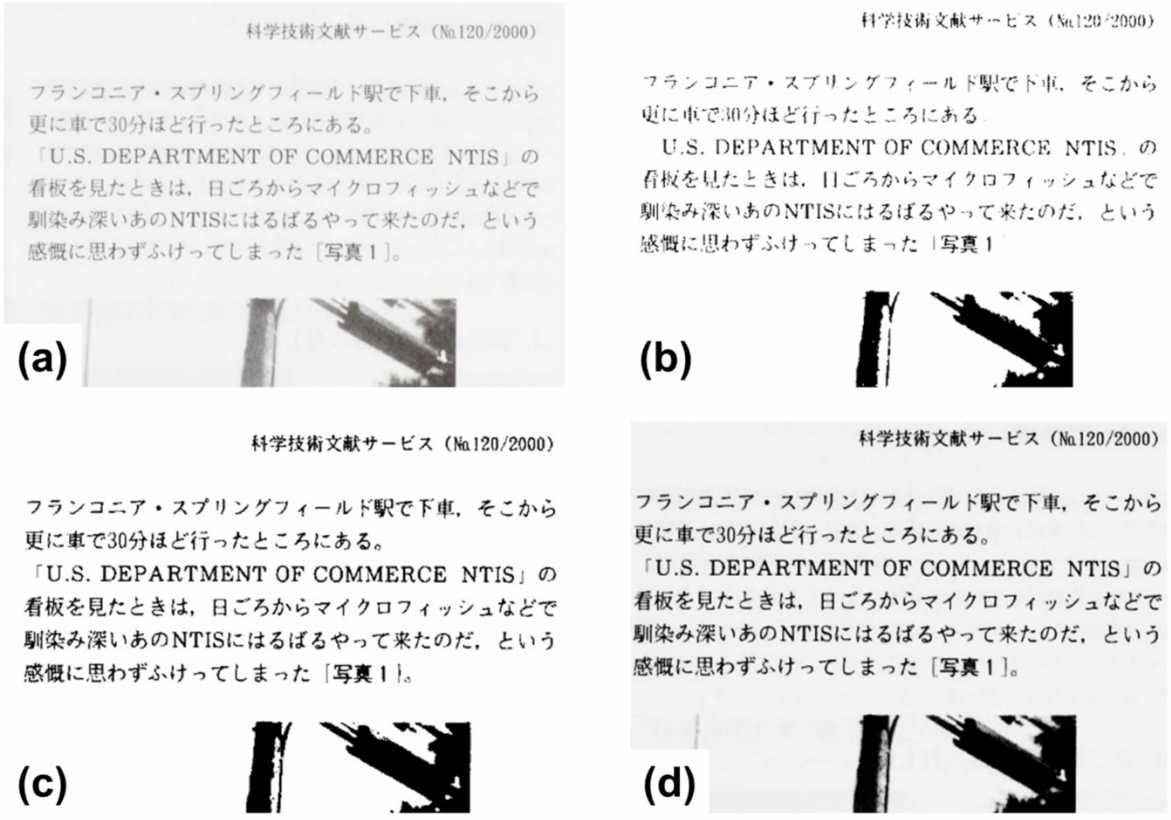


Figure 2: A typical result
(a) Raw image (b) Binary image adjusted using the entire page (c) Binary image adjusted using only text areas (d) Contrast adjusted image using our proposal.

CONCLUSION AND FUTURE WORK

We have examined two methods to improve readability of digitized materials and both of them work properly for most images. In the future, it will be necessary to solve problems related to the enhancement of illustrations as well as to handle color images, which will require more extensive training data. Also, it would be more effective to exclude graphics prior to application of whitening when using semantic segmentation.

REFERENCES

[1] National Diet Library Digital Collections, http://dl.ndl.go.jp/

[2] **P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros.** Image-to-image translation with conditional adversarial networks. In CVPR, 2017

[3] **Chen, Liang-Chieh, et al.** "Encoder-decoder with atrous separable convolution for semantic image segmentation." arXiv preprint arXiv:1802.02611 (2018)

[4] **Otsu, Nobuyuki.** "A threshold selection method from gray-level histograms." IEEE transactions on systems, man, and cybernetics 9.1 (1979): 62-66.

[5] **Antonacopoulos, Apostolos, et al.** "ICDAR 2009 page segmentation competition." Document Analysis and Recognition, 2009. *ICDAR'09. 10th International Conference on.* IEEE, 2009.

A Collaborative Approach for GIS Historical Maps Metadata Project

Naomi Shiraishi[1], Haiqing Lin[1]

Historical spatial analysis is recently gaining popularity among digital humanities researchers. Many major academic libraries have large collections of historical maps which are critical for historical spatial studies. However, making those historical maps discoverable and accessible to digital humanities projects is a challenge. Traditionally, historical maps are descriptively organized by library cataloging systems by following the cataloging rules. However, library catalogs lack systematic frameworks to extract significant geographical features in machine readable formats. This research proposes a new approach to describe geographic features of historical maps and publish them in machine readable formats to support researchers' use of historical maps through GIS applications.

In the changing landscape of digital research and open access, the roles of technical services librarians are not limited to traditional cataloging. One of the new roles we envision is supporting digital humanities research by organizing, managing, and providing access to data sets via metadata creation and management.

Traditionally, catalog records for print materials created by technical services librarians have helped searching, retrieving, and identifying resources and organizing information by using controlled vocabularies. Metadata does the same or more for digital materials (resources and projects) in the digital tools and linked data environment.

But when it comes to metadata for digital projects, it is rare for researchers to involve technical services librarians during the process of their projects. Also, some institutions (including our own) consider using outsourcing venders for metadata creation due to their speedy processing or budgetary reasons. As a result, such metadata do not fully utilize tools for information organization and management.

Metadata is important in making digital humanity research more discoverable and accessible. For geospatial resources, there are different metadata standards available, but no one standard can cover all materials. Some specialized projects may need experienced technical services librarians' help for creating effective metadata. Historical maps are unique in that they may not necessarily fit into regular geospatial metadata standards well and this is where technical services librarians can utilize their knowledge and experience. As a case study, we created a mock digital map project that compares Japanese historical maps by using digital humanities tools to show how technical services librarians can play an important role in digital humanities research.

In this project, we compared Aou Tokei's Kokugun Zenzu 國郡全圖 (1837) with Nagakubo Sekisui's Kaisei Nihon Yochi Rotei Zenzu 改正日本輿地路程全図 (1779). In the preface of Kokugun Zenzu, the author says his maps in this atlas were created based on Nagakubo Sekisui's Kaisei Nihon Yochi Rotei Zenzu. However, it is not easy to see how they are related since one is an atlas and the other is a single sheet map. So, to compare them visually, we decided to layer the maps together with a contemporary map by using digital tools. Layering would reveal similarities and differences among those three maps more clearly and make it easier to see if there are specific traits that show any influences of Kaisei Nihon Yochi Rotei Zenzu on Kokugun Zenzu.

When comparing maps, it is important that maps are rectified into a common coordinate system. To rectify these maps, ground control points are needed. This is one of the processes where a technical librarian's skills are useful since finding ground control points in historical maps may require consulting gazetteers and other reference tools. In this particular case, mountains, lakes and capes are selected since other geographical points, such as villages, may have changed or disappeared overtime.

[1] University of California, Berkeley

When we create metadata for such a GIS project, information about ground control points should be included. We believe that such information is integral to reusability of data, which is a key to successful digital humanities. We are currently working on developing metadata guidelines that include the minimum number of ground control points and types of geographic features. These guidelines also discuss how to trace historical geographic name changes and record the source of information. We recommend encoding such information in GeoJson, KML, or CVS to directly support GIS applications. Technical services librarians can help researchers by figuring out consistent vocabularies and metadata standards.

Furthermore, the linked data technology enables the data in the metadata to link to other resources or projects. For example, the key access points, such as the titles of the maps or the ground control points can be linked to other research projects or resources involving those maps or place names, contextualizing the project in the world of digital research and leading to new discoveries. If librarians and researchers can create metadata collaboratively, such process can be done more effectively.

This research aims to examine the functional requirements of metadata services for digital humanities research by analyzing a GIS project involving East Asian historical maps. By exploring some metadata solutions, we intend to demonstrate the role of controlled vocabulary as a means to improving the access to and usability of spatial data. We hope to show the various possibilities that arise from a collaborative approach when creating metadata for digital humanities projects.

Cell Phone City: Pedestrians' Mobile Phone Use and the Hybridization of Space in Tokyo

Deirdre Sneep[1]

Abstract

Cities all over the world are rapidly changing due to a mobile communication technology revolution. All around us, screens are getting bigger and internet is getting faster. Although the increase in mobile internet use is a global phenomenon, there are several urban agglomerations in East-Asian countries that rank particularly high in mobile internet data consumption. Among those is Tokyo, the cradle of mobile Internet technology. The change from static to mobile internet use is bound to have large impact on how people experience urban space, because it blurs the boundaries between physical and virtual space. The constant connectedness to internet transforms the city: it makes streets into places for diverse social interaction, metros into workspaces, and crossroads into arcade halls. QR codes and URLs that are pasted everywhere prompt the mobile phone user to visit the online. Furthermore, the rapid recent developments of mobile Internet use have made it difficult for researchers of urban studies to keep their framework contemporary. As the urban dweller is becoming increasingly 'digital' and human behavior is increasingly influenced by use of mobile internet, researchers of urban culture are faced with a new kind of digital/analog 'hybrid' pedestrian, one that is challenging traditional meanings of urban space.

The concept of digital/physical space hybridization as a result of internet has been studied before, and studies often rightly emphasize the social-cultural environments in which mobile phone technologies operate (Katz and Aakhus, 2002; Katz, 2003; Ito, Okabe and Matsuda, 2005b; McLelland, 2013). However, these leading theories on digitalization of the city have yet to consider the degree of mobile internet that is now showcased in Tokyo and other major East-Asian urban agglomerations. Japan in this regard is an especially interesting case. Those living in Japanese cities such as Tokyo have been using mobile phones since the 1990s, and, as Japan was the first country to implement a nation-wide mobile phone internet network, have been connected to mobile phone internet since as early as 1999. This resulted in Japan having, from an early stage, sprouted a culture of internet-heavy mobile phone use – that is, being connected to the flow of information anytime, anywhere – that has come to be hard-coded into the urban system and the movement and behavior of its pedestrians, which makes it a particularly interesting case study for research of digitalization of urban life.

Mobile phone use in Tokyo has been shown to have had interesting and lasting effect on the way pedestrians move and interact with others as well with their environment (Fujimoto, 2005; Ito, Okabe and Matsuda, 2005a; Cui *et al*., 2007; Baron and af Segerstad, 2010), effects which have sometimes been deemed negative as mobile phone users are often seen as 'disconnected' or dissociative (Takao, Takahashi and Kitamura, 2009; Ikeda and Nakamura, 2014), and seen as potentially dangerous to those around them due to a lack of awareness of surrounding environment, especially when the mobile phone is used on the streets (Masuda and Haga, 2015; Obara, Kashiwagi and Nakamura, 2016). While these are observations that seem to be following global trends of fears and concerns surrounding the rise of mobile phone technologies (see i.e. Choi et al., 2012; Lamberg and Muratori, 2012), there are other forms of mobile phone behavior that seem more specific to the Japanese case, which show how technology use is influenced by cultural patterns. It is these aspects of mobile phone use that can tell us more about a society and its stance towards digitalization and technology. Fujimoto (2005, 2006) previously pointed out the specific way in which Japanese use their mobile phones as ways to 'shield off' interaction with people around them, calling the mobile phone a 'territory machine'. On the other hand, however, the mobile phone has been said to be contributing towards Japan becoming a

[1] University of Duisburg-Essen

'ubiquitous' society in which mobile internet is a part of each and all aspects of daily lives ('ubiquitous computing', Sakamura 2002), especially with regard to social interaction (Sakamura, 2002; Murakami, 2003; Srivastava, 2004; Tawara, 2008). On the one hand being portrayed as a tool for reclusion or isolation, while on the other hand being portrayed as an all-pervasive tool for social interaction with society, the mobile phone is in between two extremes. There is a significant gap, however, in testing and further developing these two opposing theories in real-life situations in Japanese cities, as most research thus far has been driven by a risk-focused discourse that portrays the mobile phone as disrupting urban life, instead of as an expression of it.

This paper critically re-visits the existing theoretical framework on the influence of the digitalization of the pedestrian through mobile phone use on the city, to answer the question what characterizes mobile phone-using pedestrian's behavior in Tokyo and their interaction with the city and environment. It compares findings with anthropological fieldwork the author conducted by observing, monitoring and mapping smartphone use and behavior among pedestrians in the center of Tokyo for a period of eight months. The research points out, how there are new forms of pedestrian behavior emerging due to mobile phone use, which adds new, semi-virtual layers of meaning to the existing meanings of urban space. Placing this virtual/physical 'hybrid' space at the base for mobile phone using pedestrians' interaction with the environment and those around them, this research analyzes the spatial behavior of mobile phone using pedestrians in Japan from a socio-cultural perspective, opposing Fujimoto's theory of the 'territory machine' to the theory of 'ubiquitous computing', establishing a new theory on social behavior of mobile phone-using pedestrians, and on urban life in Japan. The results of this research show how both theories exist simultaneously, and how the mobile phone user of Tokyo has become a 'hybrid' between isolated as well as pervasive behavior. Eventually, this paper adds relevant findings about an increasingly important facet of our 'cell phone cities', while simultaneously giving new insights into the digitalization of Japanese society, from a socio-cultural perspective.

Keywords
Mobile Phones, pedestrian behavior, space hybridization, social meaning of space, Tokyo, ubiquitous society

References

Baron, N. S. and af Segerstad, Y. H. (2010). Cross-cultural patterns in mobile-phone use: public space and reachability in Sweden, the USA and Japan, *New Media & Society*, 12(1), pp. 13–34.

Choi, H.-S., Lee, H.-K. and Ha, J.-C. (2012). The influence of smartphone addiction on mental health, campus life and personal relations - Focusing on K university students, *Journal of the Korean Data and Information Science Society*. Korean Data and Information Science Society, 23(5), pp. 1005–15.

Cui, Y. *et al*. (2007). A Cross Culture Study on Phone Carrying and Physical, in Aykin, N. (ed.) *Usability and Internationalization Part 1.* Springer-Verlag Berlin Heidelberg, pp. 483–92.

Fujimoto, K. (2005). The Third-Stage Paradigm: Territory Machines from the Girls' Pager Revolution to Mobile Aesthetics, in Ito, Mizuko; Okabe, Daisuke; Matsuda, M. (ed.) *Personal, Portable, Pedestrian: Mobile Phones in Japanese Life.* Massachusetts Institute of Technology, pp. 77–102.

Fujimoto, K. (2006). Anti-Ubiquitous 'Territory Machine', in Matsuda, M., Ito, M., and Okabe, D. (eds) *Keitai no aru Fuukei Tekunorojii no nichijōka wo kangaeru [Mobile Phone Environment: Technology becoming part of Daily Life].* Kyoto: Kitaohji Syobo.

Ikeda, K. and Nakamura, K. (2014). Association between mobile phone use and depressed mood in Japanese adolescents: A cross-sectional study, *Environmental Health and Preventive Medicine,* 19(3), pp. 187–93.

Ito, M., Okabe, D. and Matsuda, M. (2005a). *Personal, Portable, Pedestrian: Mobile Phones in Japanese Life.* Cambridge MA: MIT Press.

Ito, M., Okabe, D. and Matsuda, M. (2005b). Technosocial Situations: Emergent Structuring of Mobile E-mail Use, in Ito, M., Okabe, D., and Matsuda, M. (eds) *Personal, Portable, Pedestrian: Mobile Phones in Japanese Life.* Cambridge MA: MIT Press, pp. 257–73.

Katz, J. E. (2003). *Machines That Become Us: The Social Context of Personal Communication Technology.* Transaction Publishers.

Katz, J. E. and Aakhus, M. A. (2002). Conclusion: Making Meaning of Mobiles - a Theory of Apparatgeist, in Katz, J. E. and Aakhus, M. A. (eds) *Perpetual Contact: Mobile Communication, Private Talk -Public Performance.* Cambridge University Press, pp. 301–18.

Lamberg, E. M. and Muratori, L. M. (2012). Cell phones change the way we walk, *Gait and Posture,* 35, pp. 688–690.

Masuda, K. and Haga, S. (2015). Effects of Cell Phone Texting on Attention, Walking, and Mental Workload: Comparison between the Smartphone and the Feature Phone, *JES Ergonomics.* Japan Ergonomics Society, 51(1), pp. 52–61.

McLelland, M. J. (2013). Socio-cultural Aspects of Mobile Communication Technologies in Asia and the Pacific: a Discussion of the Recent Literature, in Goggin, G. (ed.) *Mobile Phone Cultures.* Routledge, pp. 124–34.

Murakami, T. (2003). Establishing the Ubiquitous Network Environment in Japan: From e-Japan to U-Japan, *NRI Papers,* 66, pp. 1–20.

Obara, T., Kashiwagi, S. and Nakamura, M. (2016). Measurement of Angles of Smart phones at 'texting while walking', *Proceedings of the IEICE Engineering Sciences Society/NOLTA Society Conference.* The Institute of Electronics, Information and Communication Engineers, 2016, p. 335.

Sakamura, K. (2002). *Yubikitasu konpyūta kakumei: jisedai shakai no sekai hyōjun. Kadokawa,* Tokyo.

Srivastava, L. (2004). Japan's ubiquitous mobile information society, *info,* 14(4), pp. 234–51.

Takao, M., Takahashi, S. and Kitamura, M. (2009). Addictive Personality and Problematic Mobile Phone Use, *CyberPsychology & Behavior,* 12(5), pp. 501–07.

Tawara, Y. (2008). Introduction: Working toward Realizing the Ubiquitous Network Society, *The Journal of the Institute of Electronics, Information and Communication Engineers.* The Institute of Electronics, Information and Communication Engineers, 91(7), pp. 563–68.

A Case Study on Digital Pedagogy for the Style Comparative Study of Japanese Art History Using "IIIF Curation Platform"

Chikahiko Suzuki[1], Akira Takagishi[2], Asanobu Kitamoto[1]

Introduction

This paper describes the work-in-progress challenge of digital pedagogy for the style comparative study in Japanese art history. The style comparative study is a basic research method in art history where researchers observe numerous artworks and arrange these styles systematically. Novice students do not have many opportunities to observe artworks, and possess little systematic information regarding styles. Therefore, it is difficult to train them to compare and understand styles.

Digital technology is an effective training tool for beginner students of art history, especially regarding artworks that involve comparing several elements, like Emaki (illustrated scroll). As we discuss later, art history constantly employs technology. Our method is an innovative step forward.

Case study

(1) Purpose

Our method provides comparative training on styles for beginner students of Japanese art history. A method is required that enables students to make basic comparisons, comprehending how expert scholars do so, despite possessing minimal information regarding art style. Therefore, we designed the pedagogy using the **IIIF Curation Platform**. With this method, beginners can personally experience expert scholars' approaches, learning how and where they focus their attention.

(2) Software

We designed the method using the IIIF Curation Platform, an extension of the International Image Interoperability Framework (IIIF) developed by the Center for Open Data in the Humanities (CODH). The IIIF Curation Platform provides the function making lists of images across multiple IIIF Manifests. The **IIIF Curation Viewer** and **IIIF Curation Finder** are built on this platform. IIIF Curation Viewer can easily create and share curation – a list of canvases with metadata. We can search and identify interesting canvases across multiple curations and create derivative curations with IIIF Curation Finder (Figure 1) (CODH, 2018).

The IIIF Curation Platform enhances not only education but also research in art history. Sharing all evidential images as curations and citing URLs from academic papers, enhances shareability and reusability of research, allowing other researchers to verify papers' results effortlessly (Suzuki et al., 2017).

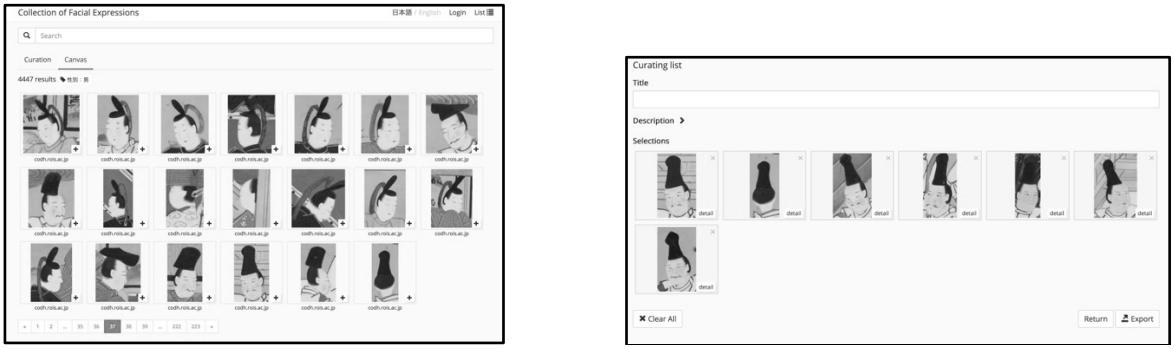


Figure 1: Example of search with metadata "man" on IIIF Curation Finder, and new curation focus on headgear

[1] Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems / National Institute of Informatics
[2] The University of Tokyo

(3) Material

We used Ishiyama-dera Engi Emaki vol. 5 (Miraculous origins of the Ishiyama-dera temple, 『石山寺縁起絵巻』第五巻) produced during the Muromachi-era (15th century) as experiment material. Masahiko Aizawa examined the facial expression styles in this artwork in detail (Figure 2 left) (Aizawa, 2016).

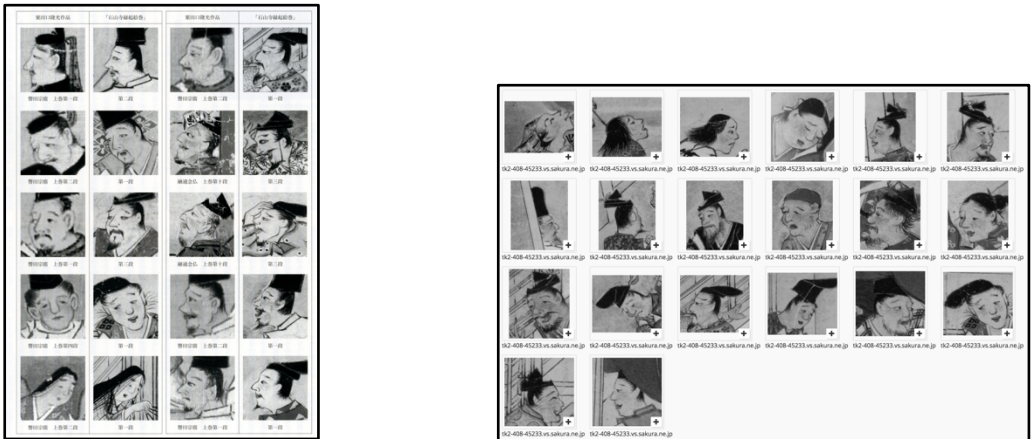


Figure 2: Comparison of facial expression styles from Ishiyama-dera Engi Emaki and other Emaki by Masahiko Aizawa (left). Sample curation of all facial expressions from Ishiyama-dera Engi Emaki (right)

(4) Lecture plan

1. Lecturer creates a curation of all facial expressions from Ishiyama-dera Engi Emaki vol. 5 with the IIIF Curation Viewer (Figure 2 right).
2. Lecturer provides basic metadata of all the facial expressions such as gender and face direction.
3. Lecturer imports curation (created in 4-2) to the IIIF Curation Finder.
4. Lecturer shows students the facial expressions that Aizawa identified from other artworks to compare with Ishiyama-dera Engi Emaki.
5. Students search and identify facial expressions that they think are similar to Aizawa's choices (in 4-4) with the IIIF Curation Finder, and create a new curation.
6. Students explain their curation.
7. Lecturer and students discuss the curation, while referring to Aizawa's choices.

(5) Results

Currently, this method is a work-in-progress. We expect beginners to personally experience expert scholars' approaches. We create a curation of all facial expressions so that students can identify the elements experts rejected. Essentially, they learn to understand the thinking process.

We intend to conduct a practical lecture during the history of art course at the University of Tokyo in July 2018, where we will include the lecture results and findings.
In addition, during the preparatory phase itself we realized that creating a curation itself is effective training on comparatively studying styles because the creator must observe digital images of artworks meticulously. Given the interoperability of IIIF, this method can be applied to other artworks on IIIF. We intend to use this method on medieval Japanese Emaki (12th – 16th century) that comprise around 600.

Discussion

This is an effective training method for beginner students and this case study provides a pedagogical background of art history through mechanical reproductions. Careful appreciation of original artworks is the most important aspect of research and education. However, it is difficult to conduct lectures with original artworks every time. Therefore, reproducing artwork images is important too.

Analog photographs and photographic slides are important tools currently. Pedagogy with mechanical reproductions especially photographic slides was established in the late 19th century in German universities (Nelson, 2000). Analog photographs and slides are common property of research units, and play important roles in research and education (Kawaguchi, 2014). In addition, publications with photographs of artworks were common in the 19th century. *Kokka* (『国華』), the Japanese journal of oriental art, has published detailed woodprints and photographs of important artworks since 1889. This publication established appreciation through print media in Japan (Okazuka, 2001).

Using prints, photographs, and photographic slides in a lecture are common today. This is the result of art history constantly adopting new technology. Nowadays, it is also common practice to conduct lectures with digital images. Digital images make it convenient to pan and zoom. However, there is no major pedagogical alteration with photographic slides.

Digital images are more than a substitute for photographic slides with IIIF and curation on the IIIF Curation Platform. As indicated in this case study, this new technology introduces interactive lecture plans and shareability of curation. Inevitably, art history will adopt the convenience of digital images and IIIF.

References

Aizawa, M. (2016). "Ishiyamadera engi emaki shoukai." *Ishiyamadera engi emaki shusei*, Chuokouron Bijutsu Shuppan: Tokyo. (in Japanese)

CODH. (2018). IIIF Curation Platform, http://codh.rois.ac.jp/iiif-curation-platform/ (accessed 26 April 2018).

Kawaguchi, M. (2014). "Bijutsushi ni okeru gazou no chikara to digital gijutsu." *DHjp: Digital Humanities Jp,* (2): 52–59. (in Japanese)

Nelson, R. S. (2000). "The slide lecture or the work of art 'history' in the age of mechanical reproduction." *Critical Inquiry,* 26(3): 414–34.

Okazuka, A. (2001). "Meiji ki no bijutsu shashin shuppanbutsu." *Bijutsu Forum 21,* 4: 39–43. (in Japanese)

Suzuki, C., Takagishi, A. and Kitamoto, A. (2017). "IIIF curation viewer brings about 'detail' and 'reproducibility' for art history a study of Eiribon / Emaki at Edo period with IIIF." *IPSJ Symposium.* Vol. 2017, No. 2, pp. 157–164. (in Japanese)

Detecting Unknown Word Senses in Contemporary Japanese Dictionary from Corpus of Historical Japanese

Aya Tababe[1], Kanako Komiya[1], Masayuki Asahara[2], Minoru Sasaki[1], Hiroyuki Shinnou[1]

Word sense disambiguation (WSD) involves identifying the senses of words in sentences when the word has multiple senses. Various techniques for WSD have been proposed in the field of natural language processing. However, a variety of studies have been investigated about WSD for contemporary Japanese corpora but few studies have been investigated for historical Japanese corpora.

When we classify the word senses of words in historical Japanese corpora into the word senses in a contemporary Japanese dictionary, a number of words cannot be classified because they are not contemporary words. In addition, it often happens that they have word senses that are not used in the present day even if the words themselves are used now.

Therefore, in the current study, we proposed a system to classify the word senses of words in a Japanese historical corpus to determine the word senses that are not listed in a dictionary in the present day. We used Corpus of Historical Japanese (CHJ) as a Japanese historical corpus and Word List by Semantic Principles (WLSP) (Kokuritsukokugokenkyusho, 1964), which is a Japanese thesaurus of contemporary words, as a contemporary Japanese dictionary.

In the WLSP, the article numbers or concept number indicate shared synonyms. In the WLSP thesaurus, words are classified and organized by their meanings and each WLSP record contains the following fields: record ID number; lemma number; record type; class; division; section; article; article number; paragraph number; small paragraph number; word number; lemma (with explanatory note); lemma (without explanatory note); reading; and reverse reading. Each record has an article number, which represents four fields: class; division; section; and article. For example, the word“犬”(inu ,meaning spy or dog) has two records in the WLSP, and therefore has two article numbers, 1.2410 and 1.5501, indicating that the word is polysemous. We can use the article numbers in WLSP with words as word senses, because we can treat a pair of a concept and a word as a word sense. We have CHJ with the article numbers, which is a word-sense-tagged corpus that is in its infancy, and used it for the experiments.

We automatically classified all the words in CHJ with the article numbers into three classes. The first class is the words that have the word senses listed in WLSP. In other words, the article number of the word must be listed in WLSP and the word itself must have an entry in WLSP as the example of the word that has the article number. The second class is the words that have the word senses whose article number is listed in WLSP of contemporary words but the word does not have the meaning any more in the present day. The third class is the words that have the concept not listed in WLSP of contemporary words. In this case, WLSP has no concept that the word describes.

Generally, unknown word senses in contemporary corpora are mostly new usages that have few examples. Therefore, it is difficult to detect new word senses with classification models. However, we have a number of unknown word senses in CHJ because it is a historical corpus. Therefore, we used support vector machine to classify them.

We used orthographic tokens, pronunciation tokens, readings, lemmas, original texts, parts of speech, conjugation types, and conjugation forms as the basic features for a classifier. In addition, we introduced the dictionary information features. They are (1) if the word has one entry in WLSP or not, (2) if the word has more than one entries in WLSP or not, and (3) if the word has no entry in WLSP or not. Please note that even if the word

[1] Ibaraki University
[2] National Institute for Japanese Language and Linguistics

has one or more entries, it is not sure that the word in CHJ has the concept in WLSP. In addition, we used word embeddings created by word2vec, which is a vector that can measure the similarities between words and is proposed by (Mikolov et al, 2013), as the features. We tested three dimensionalities of word2vec, 50, 100, and 200, and also developed a model without word2vec. In addition, we investigated the performances according to the condition of the basic features (1) all the basic features were used and (2) only orthographic tokens, lemmas, and parts of speech were used.

The experiments using Taketori-monogatari, Hojoki, Tsurezure-gusa, Tosa-nikki, and Toraakira-bon corpora showed that the accuracy was the best when all the basic features and 200-dimension word2vec were used (80.53%). They also showed that accuracies with all the basic features were always better and word embeddings are effective for classification. The accuracy increased with the growing of the dimensionality of word embeddings.

This work was supported by JSPS KAKENHI Grants Numbers 18K11421 and 17H00917 and was a project of the Center for Corpus Development, NINJAL.

[Mikolov, 2013] **Mikolov, T., Chen, K., Corrado, G. and Dean, J.**: Effcient Estimation of Word Representations in Vector Space, ICLR Workshop paper (2013).

Verifying the Authorship of Saikaku Ihara's *Arashi ha Mujyō Monogatari* Using a Quantitative Approach

Ayaka Uesaka[1]

1. Introduction

This study aims to focus on *Arashi ha Mujyō Monogatari ("The Tale of Transient Popular Kabuki Actor Arashi's Life"; 1688)*, a novel from the early modern Japanese literature, written by Saikaku Ihara (1642–93) using principal component analysis (PCA) and cluster analysis (hierarchical clustering). It is a first work of Kabuki actor's life in Japan (Kabuki is a traditional stage arts performed exclusively by male actors with the accompaniment of live music and songs). Saikaku was a national author whose novels were published in 17th century in Japan. One recent hypothesis has stated that he wrote twenty-four novels, however it remained unclear which works were really written by Saikaku except *Kōshoku ichidai otoko ("The Life of an Amorous Man"; 1682)*, *Shōen Ōkagami ("The Great Mirror of Female Beauty"; 1684), Kōshoku ichidai onna ("The Life of an Amorous Woman"; 1686), Kōshoku gonin onna ("Love Stories about Five Women"; 1686)*, while research on his works has proceeded, these fundamental doubts about his authorship remain.

2. Previous Studies

Noma found and introduced *Arashi ha Mujyō Monogatari* in 1941. He mentioned that the novel was actually written by Saikaku, for the following reasons (Noma, 1941 and 1964). (1) The handwriting of the novel belongs to Saikaku; and (2) He found a similar writing error in *Arashi ha Mujyō Monogatari* and Saikaku's work.

The handwriting is not crucial in deciding if they are Saikaku's novels. According to Emoto *et al*. (1996), among his twenty-four novels, the handwriting of nineteen works does not belong to Saikaku. Moreover, Saikaku made a fair copy of other writer's draft such as *Kindai Yasa Inja ("The story of a hermit"; 1686)* by Kyōsen Sairoken (? -?) and *Shin Yoshiwara Tsurezure ("The book of commentary on the licensed quarters of a certain area"; 1689)* by Sutewaka Isogai (? -?).

Mori (1955) has argued that Saikaku's novels are an apocryphal work mainly written by Dansui Hōjō (1663-1711) except *Kōshoku ichidai otoko*.

As he gained a national audience, Saikaku was pressured to write on demand and in great volume. At first he wrote only one or two novels a year, however in the two years from 1687 to 1688 he published twelve books, with a total of sixty-two volumes. Saikaku's style and approach also changed at this point (Shirane, 2004).

There is possibility that Saikaku had some assistant (Nakamura, 1969). *Arashi ha Mujyō Monogatari* was published in this period. Moreover, *Arashi ha Mujyō Monogatari* does not have a preface, epilogue, signature, namely it is not specified that it was written by Saikaku. Despite the authorship problem of *Arashi ha Mujyō Monogatari* remains unanswered; little work has been done about it. For that reason, this study re-examines the authorship of *Arashi ha Mujyō Monogatari* using a quantitative approach.

In our previous studies, we have analyzed Saikaku and Dansui's novels, and have clarified the following points by extracting their writing style using PCA and cluster analysis: (1) A comparison of the Saikaku and Dansui's novels showed ten prominent features: the grammatical categories, words, nouns, particles, verbs, adjectives, adverbs, adnominal adjectives, grammatical categories bigrams and particle bigrams (Uesaka, 2015, 2016); and (2) Using these features, we analyzed Saikaku's four posthumous novels (many researchers have raised questions about the authorship, because these novels were edited and published by Dansui after Saikaku's death). We found these four posthumous works indicated same features of Saikaku's novel, therefore we concluded that most part of these four posthumous novels belonged to Saikaku (Uesaka・Murakami,2015ab, Uesaka, 2016).

[1] Osaka University

Furthermore, we compared *Arashi ha Mujyō Monogatari* to Saikaku and Dansui, as authenticated novels of them using PCA and cluster analysis to see the differences in each novels. We found that *Arashi ha Mujyō Monogatari* was different from Saikaku and Dansui's works (Uesaka, 2017). In order to clarify *Arashi ha Mujyō Monogatari*'s author, it is necessary to add the data of other writers with the possibility of the author of *Arashi ha Mujyō Monogatari*. Thus, we digitized and added two novels; Nishiwaza Ippu's *Shinshiki Gokansho(1698)* and Yashioku Jibun's *Kōshoku Mankintan (1694)*.

3. Data for This Study

(1) We digitized all the text of 120 works of Saikaku (24 novels, 80 poem books, etc.); (2) Since Japanese sentences are not separated by spaces, we built the rule with early modern Japanese researchers, who were editors of *Shinpen Saikaku Zenshu ("The new complete works of Saikaku")*; and (3) Based on this rule, we added spaces between the words in all of the sentences. In addition, grammatical categories' information was added.

We also made the database of Dansui's novels *Shikidō Ōtuzumi* (*"The Great Drum of Love"*; 1687), *Chuya yōjin ki* (*"The Night and Day of Precaution"*; 1707), *Budō hariai Ōkagami* (*"The Great Mirror of Martial Arts"*; 1709), Ippu's novel *Shinshiki Gokansho* and Jibun's novel *Kōshoku Mankintan*, using same methods and rules of Saikaku's database.

Table1. Title and Length

Title	Length
S:*Kōshoku ichidai otoko*	36,781 words
S:*Shōen Ōkagami*	45,753 words
S:*Kōshoku ichidai onna*	20,184 words
S:*Kōshoku gonin onna*	26,581 words
D:*Chuya yōjin ki*	24,589 words
D:*Budō hariai Ōkagami*	22,328 words
D:*Shikidō Ōtuzumi*	12,760 words
I:*Shinshiki Gokansho*	27,346 words
J:*Kōshoku Mankintan*	21,402 words
Arashi ha Mujyō Monogatari	9,386 words

4. Analysys and Results

In this study, we compared *Arashi ha Mujyō Monogatari* to Saikaku, Dansui, Ippu and Jibun by twelve prominent features(the grammatical categories, words, nouns, particles, verbs, auxiliary verb, adjectives, adverbs, adnominal adjectives, grammatical categories bigrams, particle bigrams and auxiliary verb bigrams) using PCA and cluster analysis to see the differences in each novels.

We conducted PCA and *Arashi ha Mujyō Monogatari* depicted independently in eight features (the grammatical categories, words, nouns, verbs, adjectives, adverbs, grammatical categories bigrams and auxiliary verb bigrams) , depicted with Saikaku's novels in three features (the particles, auxiliary verb, and particle bigrams) and depicted with Jibun's novels in one feature(the adnominal adjectives) (see Figure 1). Furthermore, we conducted a cluster analysis. When calculating distances between each novels, we normalized the frequency of each words, and used the Euclidean distance, Euclidean Square distance, Manhattan distance, Minkowski distance, Canberra distance, Maximum distance, Cosine distance and Kullback–Leibler divergence and the algorithm from the Ward method. Furthermore, we obtained 96 results; 34% of the result in the mixed-up cluster, 29% in Saikaku's cluster, 20% in Dansui, Ippu and Jibun's cluster and 16% made only by *Arashi ha Mujyō Monogatari* (see Figure 2).

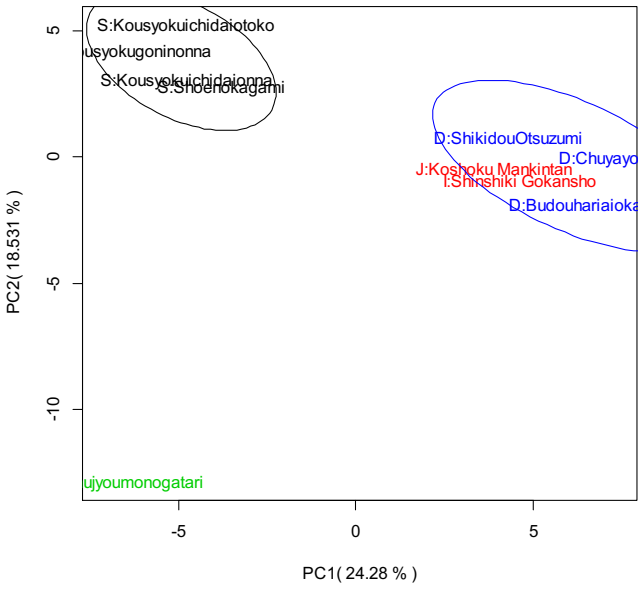


Figure 1. PCA results for the adverbs

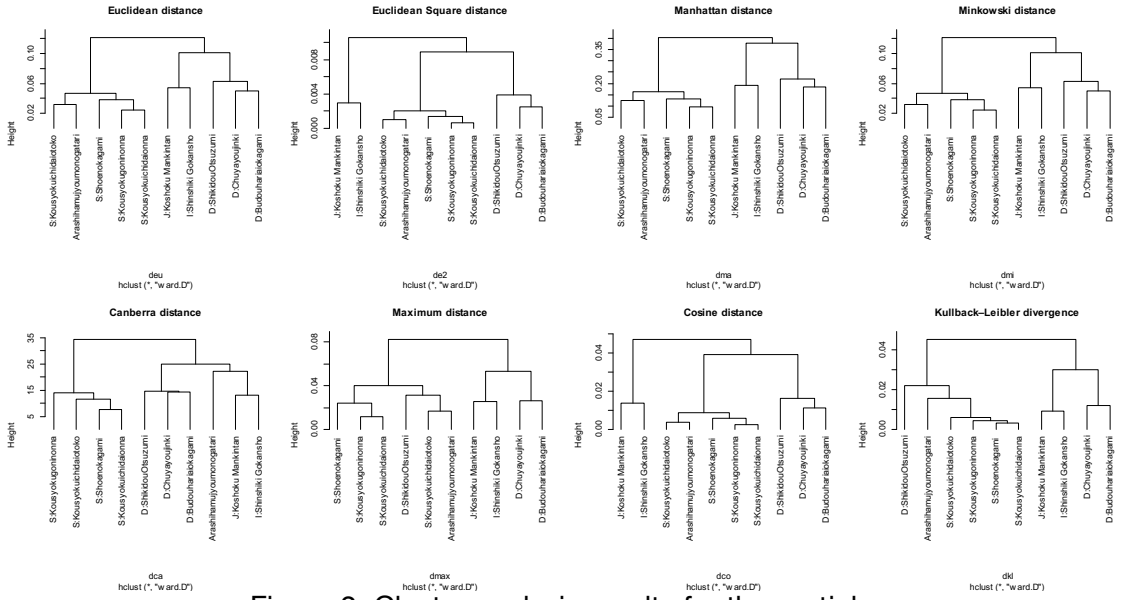


Figure 2. Cluster analysis results for the particles

5. Discussion and Conclusion

When comparing twelve prominent features using PCA and cluster analysis. While Saikaku and Danaui's made each groups, *Arashi ha Mujyō Monogatari* was not clearly classified by one author; Saikaku, Dansui, Ippu and Jibun. No fully different with Saikaku, nor fully same with Saikaku. In order to clarify this point, we need to add more data of other writers and we will do comparisons in the future study.

Acknowledgements

We would like to thank Professor Masakatsu Murakami, Professor Hidekazu Banno, Professor Takayuki Mizutani and Professor Yusuke Inaba for their help on our research. This work was mainly supported by JSPS KAKENHI Grant Number 17K12799.

References

[1] **Shirane, H.** (2004). Early Modern Japanese Literature: An Anthology 1600–1900. New York: Columbia University Press.

[2] **Noma,K.** (1941). Arashi ha Mujyō Monogatari (“The Tale of Transient Popular Kabuki Actor Arashi’s Life-Explanation and Understaning”). In: Saikaku Shin Shinkō (“Saikaku New Article”;1981). pp231-290. Tokyo:Iwanami Publishing.

[3] **Noma,K.** (1964). Sairon Arashi ha Mujyō Monogatari (“Re-explanation of the Tale of Transient Popular Kabuki Actor Arashi’s Life”). In: Saikaku Shin Shinkō (“Saikaku New Article”;1981). pp291-313. Tokyo:Iwanami Publishing.

[4] **Emoto, Y. and Taniwaki, M.** (1996). Saikaku Jiten (“A Saikaku Dictionary”). Tokyo:Ouhu Publishing.

[5] **Mori, S.** (1955). Saikaku to Saikaku Bon (“Saikaku and Saikaku’s Novel”). Tokyo:Motomotosha Publishing.

[6] **Nakamura, Y.** (1969).Saikaku Nyumon(“The Introduction of Saikaku’s Research”). In:Kokubungaku kaishaku to kansho(“Japanese literature-Explanation and Appreciation”) 34(11). pp.10-25. Tokyo: Shibundo Publishing.

[7] **Uesaka, A.** (2015). “A Quantitative Comparative Analysis for Saikaku and Dansui’s Works.” Japan-China Symposium on Theory and Application of Data Science. pp.41-46. Kyoto:Doshisha University Faculty of Culture and Information Science.

[8] **Uesaka, A.& Murakami, M.** (2015a). “Verifying the Authorship of Saikaku Ihara’s Work in Early Modern Japanese Literature: A Quantitative Approach.” Digital Scholarship in the Humanities. 30(4). pp.599～607. Oxford: Oxford University Press.

[9] **Uesaka, A.& Murakami, M.** (2015b). “A Quantitative Analysis for the Authorship of Saikaku’s Posthumous Works Compared with Dansui’s Works.” Digital Humanites2015: Conference Abstracts. Sydney: The University of Western Sydney. pp. 359–60.

[10] **Uesaka, A.** (2016). Saikaku Ikōshu no Chosha no kentō (“Verifying the Authorship of Saikaku’s Posthumous Works”). pp187-263. In: The Computational Authorship Attribution. Tokyo: Bensei Publishing.

[11] **Uesaka, A.** (2017). “Verifying the Authorship of Saikaku Ihara’s *Arashi ha Mujyō Monogatari* in Early Modern Japanese Literature: A Quantitative Approach.” Digital Humanites2017: Conference Abstracts. Montreal: McGill University and the Université de Montréal.

Predicting Prose that Sells: Issues of Open Data in a Case of Applied Machine Learning

Joris van Zundert[1], Marijn Koolen[1], Karina van Dalen-Oskam[1,2]

Unfortunately there are target areas for digital humanities (DH) research where we will not always be able to work with open data. We present such a case where we apply computational methods to revitalize the ability of publishing companies to produce revenue in the extremely difficult market of literary publishing. We argue that although open data is on principle preferable, in certain cases we need to accept a suboptimal position with regard to the openness of data to demonstrate the applied potential for DH methods and to further the development of such methods.

More books are published than ever before (Segura, 2017), yet publishing companies experience difficult times. Dutch market analysis showed that often revenue from just a few bestsellers has to cover losses incurred by a large number of non selling titles (Buss, 2015). A Dutch publishing company was interested in our proposal to try to improve revenue by exploring the capability of computational methods to predict selling potential of literary materials. Obviously if we could predict the selling capability of a title reliably this would enhance the ability of publishers to produce a profit. However, apart from Jodi Archer's and Matthew Jocker's widely acknowledged The Bestseller Code (2016) there seems to have been almost no work done in this area.

We cast the problem of predicting selling capability as a binary classification problem: can an algorithm given a large enough training and test set distinguish between known bestsellers and non selling literary fiction? To establish a baseline we have created a training set of 400 novels (Dutch and translated works of literary fiction) for which sales numbers for the years 2010–2016 are known; 200 of these are bestsellers (over 12,000 copies sold), 200 have sold few copies (0 to 100 copies). In the same way we created a test set of 50 bestsellers and 50 non sellers. The text of each novel was used to construct a term frequency–inverse document frequency (tf–idf) matrix that for each document represents the relative importance of a word's occurrence within a text. This tf–idf matrix, plus for each novel the knowledge if it was a bestseller or non seller (expressed respectively as 1 or 0), served as the input for a feed forward multi-layer perceptron (MLP, see Beam, 2017) with just a minimum of three layers (input, hidden, and output). This approach resulted in a success rate of ~80% accuracy. That is: after training the neural network model was able in 80% of cases to predict correctly if a novel from the test set, which it had not previously seen, was a bestseller or a non seller.

To understand how stable and solid our predictions are we subsequently cross validated our results for different training set sizes and for a different algorithm. A particular interesting model to compare the MLP model with appeared to be the Ružička or MinMax metric (Schubert and Telcs 2014), which is a similarity metric recently applied in stylometry exercises with impressively accurate results (Kestemont et al., 2016). To study the impact of training set size and to cross validate, both models are trained on sets of N_{train} = 40 (20 top, 20 bottom), N_{train} = 100 (50 top, 50 bottom) and N_{train} = 200 (100 top, 100 bottom) novels. The validation or test set in each setting is 40 novels (20 bestsellers, 20 non sellers). The top 120 novels and bottom 120 novels in terms of sales figures are randomly sampled and split across 20/20 novels for validation and N_{top} = 20, 50, 100, N_{bottom} = 20, 50, 100 for training. We used repeated random sub-sampling of the training and validation sets, with 10 iterations for each of the experimental settings. The results yielded are shown in figure 1.

[1] Royal Netherlands Academy of Arts and Sciences
[2] University of Amsterdam

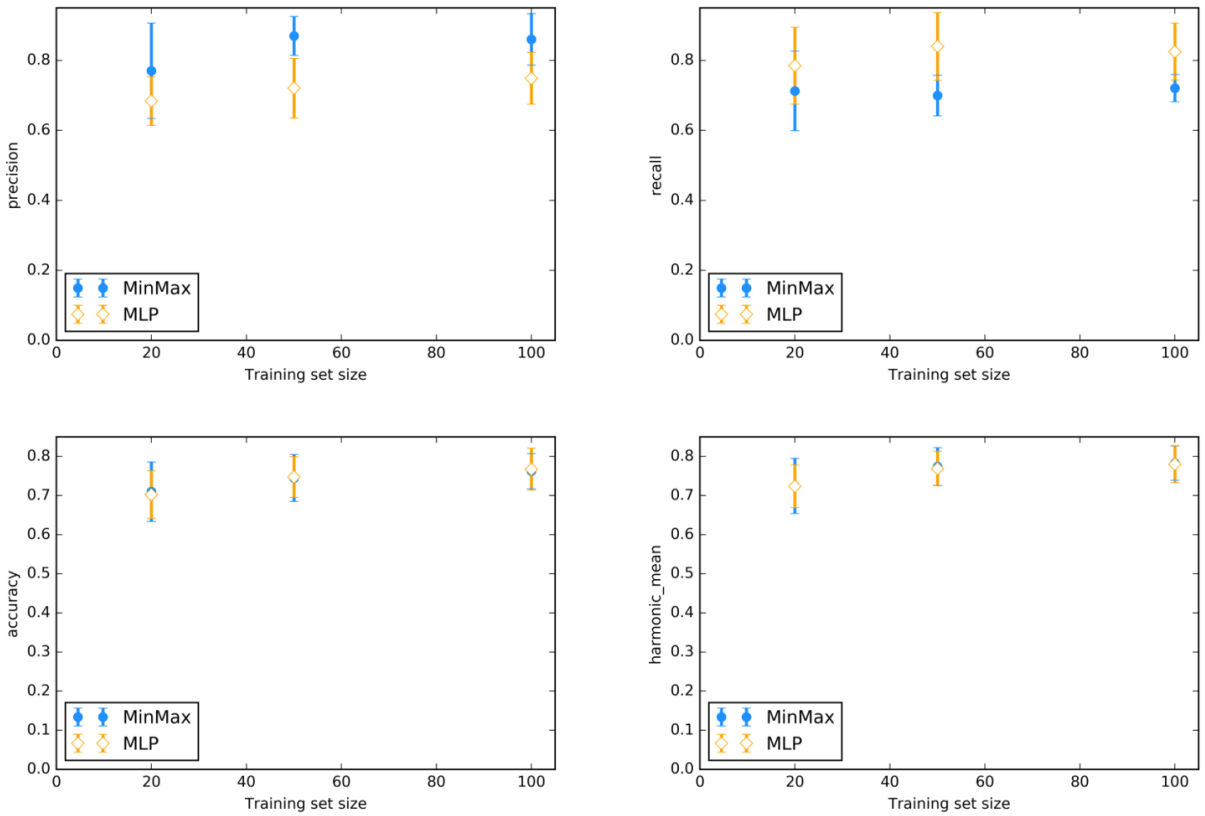


Figure 1: Evaluation of different training set sizes using MLP and MinMax algorithms; measures are Precision, Recall, Accuracy and F1 score; error bars indicate mean and standard deviation based on 10 iterations for each set size.

Our results show that both algorithms perform equally well on the task of classifying bestsellers and non sellers. Furthermore performance of both models is stable and improves only slightly with training set size. This means that indeed we would be able to support publishers in judging the selling capabilities of new materials proposed by authors—not so much maybe to benefit bestsellers but primarily to once more examine carefully manuscripts for which the algorithm predicts low selling numbers.

Both publishers and we however are not solely interested in how well a novel will sell: primarily we are keen to know what textual features are associated with high in-demand literature. We are able to start gauging such features by comparing the vocabulary of bestselling and non selling titles. To this end we explored literary vocabulary by determining which words are used relatively more often by top selling titles—by taking the top 1,000 terms used in titles with a more than 80% probability of being a well selling title and comparing their relative frequencies with the relative frequencies of these terms in non selling titles. When we discard character names and function words results reveal—as we will demonstrate in our presentation—that words appearing relatively more in literature that sells tend to be noticeable 'masculine' in nature. As is for instance corroborated in research by one of our colleagues (Koolen, 2018) this suggests that there is a still strong cultural tendency to prefer masculine themes and motives in literature.

For both ethical and commercial reasons we could not disclose data about authors, titles, and sales numbers used in our research. However, by engaging across institutional borders with a commercial partner we were both able to progress our methods and our understanding of key features of literary fiction. Our collaboration also resulted in the establishing of an experimental research environment in the Dutch National Library to enable research towards this closed corpus. We thus argue that, notwithstanding a principled preference for open data, closed data sometimes must be defensible to further the aims of DH research.

Acknowledgements

We would like to thank the people at WPG Publishers (www.wpg.nl) who kindly provided us with the sales data that made this research possible. We also would like to extend our gratitude to our colleagues at the National Library of the Netherlands (www.kb.nl) who provided the digital research corpus and secured environment that was used to execute this research. We thank also Hermann Buss and Michel Blaauw of Driven By Data (driven-by-data.com) and Emile den Tex, who all contributed significantly to our research.

References

Archer, Jodie, and Matthew L. Jockers. 2016. *The Bestseller Code: Anatomy of the Blockbuster Novel.* New York: St. Martin's Press, September. isbn: 978-1-250-08827-7.

Beam, Andrew L. 2017. *Deep Learning 101 - Part 2: Multilayer Perceptrons.* Online book, February. Accessed November 20, 2017. https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part2.html.

Buss, Hermann. 2015. *Bestsellers en Badsellers: Naar andere strategiën voor het uitgeven van boeken* [in Dutch]. Desset Publishers. isbn: 987-90-823869- 0-5.

Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016. "Authorship Verification with the Ruzicka Met- ric." In *Digital Humanities 2016: Conference Abstracts,* 246–249. Kraków: DH Benelux, July. Accessed November 16, 2017. http://hdl.handle.net/20.500.11755/5f7d08c9-f8fe-44b0-be8a-49364f390d7b.

Koolen, C.W. 2018. "Reading Beyond the Female: the relationship between perception of author gender and literary quality." Phd, University of Amsterdam. Accessed April 27, 2018.

Schubert, András, and András Telcs. 2014. "A note on the Jaccardized Czekanowski similarity index." *Scientometrics* 98, no. 2 (February): 1397–1399. issn: 1588-2861, accessed November 20, 2017. doi:10.1007/ s11192-013-1044-2.

Segura, Jonathan. 2017. "Print Book Sales Rose Again in 2016." *Publishers Weekly* (January). Accessed November 20, 2017. https://www.publishersweekly.com/pw/by-topic/industry-news/bookselling/article/72450-print-book-sales-rose-again-in-2016.html.

Retouching Our Food in Digitized Era: A Case Study of Hong Kong Foodie Critics

Wong Hei Tung[1]

This paper explores how our food is translated into visual text on social media, with the focus on digital design and editing process on *Instagram*. In digitized era, food is one of the most frequently shared cultural object on social media, which attracts foodies and other media users to participate in visual-oriented writing of foodie criticism. By foodies and foodie criticism, this paper seeks to argue the potency of foodies in retouching food stylistically, which defines their distinctive status as professional foodie critic, rather than being passionate food lover (amateur) only. With available features on *Instagram* to promote and enable easy retouching of food, foodie criticism has reflected how modern taste and visual culture have played an important role in affecting our perception of taste and constructing the fashion of food. As inherited to the studies of affect and image machine (Wissinger, 2007; Carah and Shaul, 2016), this paper provides critical discussion on how image-oriented design has embodied on *Instagram* and how these designs are used by foodie critics in Hong Kong to retouch food in a stylistic manner, which in turn, gaining them the popularity and authority on food media circuit (Rousseau, 2012; Portwood-Stacer, 2013). Theoretically put, how the simple usage of *Instagram* design in editing pictures with our mobile phone helps facilitated image production and 'stimulat[e] and captur[e] the productive activity of producing, circulating, and attending to images' (Carah and Shaul, 2016: 71) will be closely examined to unravel the impact of image-oriented design in digitized era. This line of argument, however, might tend to imply technological determinism that shapes our food criticism and sharing practices. Thus, the present examination of foodie critics helps extend these studies by taken into account how human agents (foodie critics) adopts and resists the affection of 'image machine' in digitized era. Through two-year ethnographic observation and semiotic analysis of foodie critics in Hong Kong (May 2016 – May 2018), my finding suggests how foodie critics have stylized pictures to control 'the condition of emergence of emotions', (Wissinger, 2007: 251) while at the same time being affected 'on a level below consciousness awareness' (Wissinger, 2007: 250) that implies their simulation to the digitized practices on *Instagram*.

In line of the above, this paper first proceeds to review the literature of foodies and food criticism akin the context of Hong Kong, with a view to outlining their specificities and relations in shifting the writing practices of food criticism: from literal to visual; from knowledge to retouching in digitized era. Then, I will introduce the design of *Instagram* and explain why this social platform has become one of the most popular sharing platforms of images among other contemporary digital media, pertaining to issues of digital perception, retouching functions, open excess and nature of digital sharing. In addition, this paper steps forward to understand digital writing in relation to senses and colour. As I shall demonstrate with three examples of retouching food by foodie critics, stylizing colour helps explain how our human senses are more and more relying on digitized tonality and affection. Finally, this paper concludes with the implication of foodie criticism and digitized sharing practices in modern mythmaking of taste.

References

Wissinger, E. (2007). 'Modelling a Way of Life: Immaterial and Affective Labour in the Fashion Modelling Industry,' *Ephemera: theory & politics in organization* 7.1: 250-69.

Rousseau, S. (2012). *Food and Social Media: You Are What You Tweet.* US: Rowman Altamira.

Portwood-Stacer, L. (2013). 'Foodie Culture'. [Online]. https://repsub13.wordpress.com/projects/joy/ (Accessed 6 May, 2018)

Carah, N. and Shaul, M. (2016). 'Brands and Instagram: Point, tap, swipe, glance,' Mobile *Media & Communication* 4.1: 69-84.

[1] Hong Kong Baptist University

A study on the distribution of cooccurrence weight patterns of classical Japanese poetic vocabulary

Hilofumi Yamamoto[1], Bor Hodoscek[2]

Introduction

The present study focuses on ongoing work exploring the threshold values dividing words in classical Japanese text into three groups: content, functional, and in-between. Content or semantic based analyses usually employ some techniques of data cleansing, such as eliminations of tags, punctuations, or symbols, as a preprocessing step. Stop words are also a type of token to be eliminated since they contain comparatively less meaning for content analysis. In general, it can be said that the most frequent words will be common words such as 'the' or 'and,' which help build ideas but do not carry any significance themselves (Rajaraman and Ullman, 2012: 8). Lists of stop words are commonly used, but have some problems: 1) it is necessary to compile them in advance; 2) they necessarily change depending on the domains of analyses; and 3) it is not clear which words should be included when analyzing classical texts.

Our previous study grouped modern Japanese words into low-, mid-, and high-range groups according to their information content given by their term frequency-inverse document frequency (*tf-idf*) and found that low-range words corresponded to infrequent and highly topical words, and high-range words corresponded to functional words expressing the grammatical relations between words. The study did not find an automatic method capable of classifying tokens into low-, mid-, and high-range. Furthermore, we found that previous research almost exclusively ignored the properties of the mid-range (Hodoscek and Yamamoto, 2013).

One of the methods used in Hodoscek and Yamamoto (2013) exploited the occurrence not of individual words but of pairwise or cooccurrence patterns such as 'fragrance–flower' relationships and revealed that the distribution of cooccurrence weights in modern Japanese texts approximately fitted a Gaussian curve. In this study, we will attempt to expand this analysis to classical texts by utilizing the characteristics of the Gaussian distribution to automatically group words into three clusters of cooccurrence patterns.

Methods

We use the *Hachidaishu* as the material of the present study, which comprises the eight anthologies compiled under order of the Emperors (ca. 905–1205) and contains about 9,500 poems. We developed the corpus and a method of cooccurrence weighting similar to the *tf-idf* method, *cw* (Yamamoto, 2006), which calculates the weight of patterns of any two words occurring in a poem sentence (Spärck Jones, 1972; Robertson, 2004; Manning and Schutze, 1999; Rajaraman and Ullman, 2012).

$$
\begin{aligned}
w(t,d) &= (1+\log\ tf(t,d)) \cdot idf(t) \\
cw(t_1,t_2,d) &= (1+\log\ ctf(t_1,t_2,d)) \cdot cidf(t_1,t_2) \\
cidf(t_1,t_2) &= \sqrt{idf(t_1) \cdot idf(t_2)} \\
idf(t) &= \log \frac{N}{df(t)}
\end{aligned}
$$

Where *w* is a weight, *t* a token, and *N* the number of tokens. The function *idf* is called the "inverse document frequency" (Spärck Jones, 1972; Robertson 2004; Manning

[1] Tokyo Institute of Technology
[2] Osaka University

and Schutze, 1999). The function *cw* is called the "cooccurrence weight," which allows us to examine the patterns of poetic word constructions through mathematical modeling.

As in Figure 1, there is a concept (Losee, 2001: 1019) of terms located in each layer being effective query terms. Luhn (1968) cuts the top and bottom words of the frequency and uses the mid-range vocabulary for the development of an automatic outline generation system (Figure 1). Nagao (1983: 28) also mentioned mid-range vocabulary to be effective in generating automatic abstracts. Nagao's viewpoint is slightly different from Luhn (1968) in that it allocates the distribution of word lengths around the Gaussian curve. The positions of both the upper cutoff and the lower cutoff are, however, assumed to be empirical; it is not discussed where to cut them off.

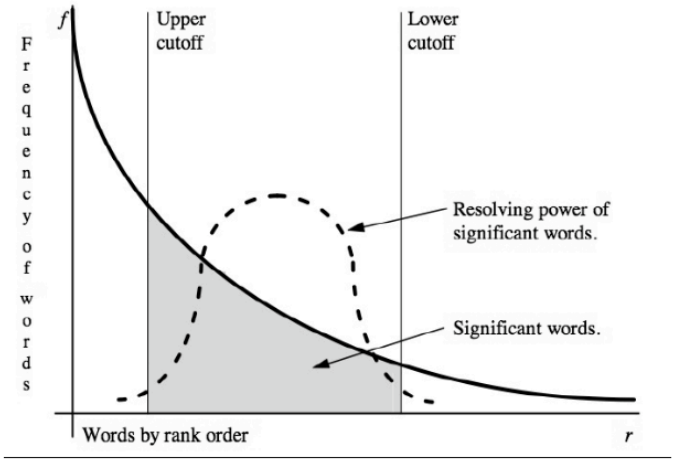


Fig. 1: Hyperbolic curve relating occurrence frequency with rank order; adapted from (Luhn 1968: 120)

Results

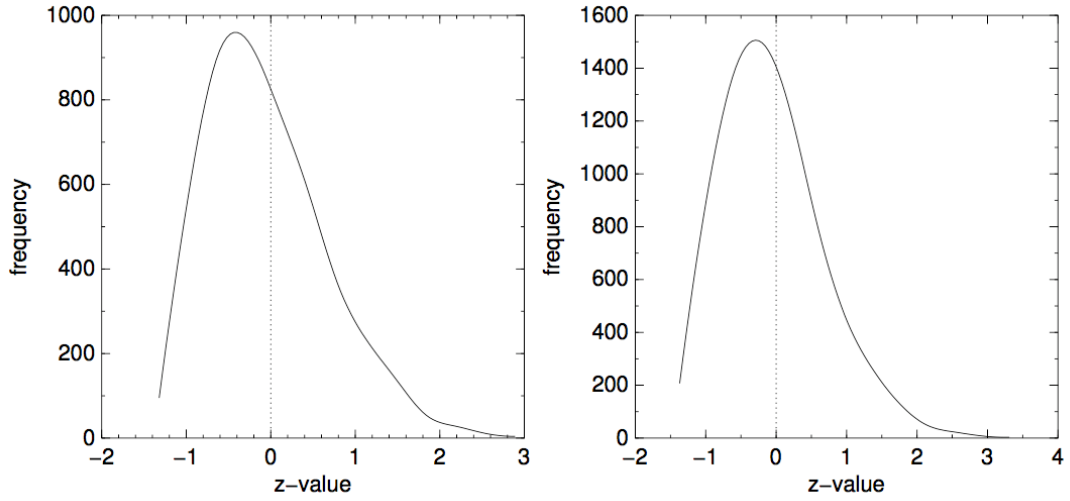


Fig. 2: The distribution of cw values *ume* (plum; left) and *sakura* (cherry; right) in Hachidaishū; The statistics of *ume* (plum): N=7016, min=-1.370, mean=0.138, max=3.700, SD=0.740, SE=0.009, CV=534.012%, Reliable interval low - upper = 0.116 - 0.161 (95%), skew=0.737, kurtosis=3.567, and that of *sakura* (cherry): N=4734, min=-1.320, mean=0.132, max=3.240, SD=0.716, SE=0.010, CV=544.116%, Reliable interval low - upper = 0.104 - 0.159 (95%), skew=0.740, kurtosis=3.345 indicate both approximately fit a Gaussian curve

The distribution of *cw* values is taken from the network model of both *ume* (plum) and *sakura* (cherry) and their curves belong to Gaussian curve as well as in classical texts (Figure 2). Therefore we will attempt to divide this shape into three layers by inflection points.

The cooccurrence patterns of *sakura* (cherry) under -0.9 (near -1) *cw* value are adjacent patterns comprising function words, and over 1 *cw* value are patterns with

content words as we expected (Table 1 and 2). As for the upper cutoff, we used an under -0.9 (near -1) σ value of *cw*, which could extract patterns of functional tokens: almost all patterns included functional words, while as lower cutoff, we used over 1 σ values, which could extract patterns of content tokens: almost all patterns included content words. Both under -1 and over 1 σ are regarded as inflection points which have mathematically interesting properties.

Discussion

Inflection points are defined as the points on the curve where the curvature changes its sign while a tangent exists (Bronshtein et al., 2004: 231). We consider the threshold values that separate upper cutoff, mid-range, and lower cutoff not as coincidental but as evidential points. It is, however, necessary to conduct further experiments and continue to discuss the mathematical traits behind the distributions of cooccurrence weights.

In terms of removing the low-range (upper cutoff) and extracting the high-range (lower cutoff) from poetic texts, we found that we do not need to use any filters to eliminate terms, since *cw* values returned semantically cooccurring patterns. Apart from low-range and high-range, the characteristics of the mid-range lexical layer are still unknown.

Table 1: Upper cutoff patterns of *ame* (sakura): *cw* = co-occurrence weight; *z* = z-value (normalized value of frequency). word annotations: ari(be), ba(cond.), ha(topic.), hana(flower), hito(human), keri(past.), ki(past.), koso(emphatic.), miru(see), mo (also), nasi(no exist), nu(neg.), o(obj.), omou(think), ramu(aux.will), su(do), te(p.), to(and), ware(we), zo(emphatic.), zu(neg.)

	cw	*z*	pattern		*cw*	*z*	pattern		*cw*	*z*	pattern
1	0.62	-0.91	mo–keri	11	0.59	-0.96	nasi–ha	21	0.52	-1.05	nu–o
2	0.62	-0.92	hana–o	12	0.57	-0.98	o–ramu	22	0.52	-1.05	o–zo
3	0.62	-0.92	o–koso	13	0.57	-0.98	mo–ramu	23	0.52	-1.05	miru–o
4	0.60	-0.94	zu–keri	14	0.57	-0.98	ha–ki	24	0.48	-1.09	ba–mo
5	0.60	-0.94	su–ha	15	0.56	-1.00	zu–mo	25	0.48	-1.09	o–keri
6	0.60	-0.94	to–ba	16	0.56	-1.00	o–te	26	0.43	-1.16	zu–ha
7	0.59	-0.96	ari–ha	17	0.55	-1.01	hito–mo	27	0.43	-1.16	to–o
8	0.59	-0.96	ari–mo	18	0.54	-1.02	zu–te	28	0.43	-1.16	te–ha
9	0.59	-0.96	ware–mo	19	0.52	-1.05	zo–ha	29	0.34	-1.27	o–ha
10	0.59	-0.96	nasi–o	20	0.52	-1.05	omou–o	30	0.34	-1.27	o–mo

Table 2: Lower cutoff patterns of *ame* (sakura) in Kokinshū: 30 out of 164 patterns extracted; *cw* = co-occurrence weight; *z* = z-value (normalized value of frequency) word annotations: ba(cond.), bakari(only), besi(should be), chiru(fall), fukakusa(deepgreen), hana(flower), isa(already), kakusu(hide), katu(win), koku(pull), komoru(go deep inside), magiru(mix), makasu(entrust), maku(wind up), manimani(as it is), masi(as), mazu(mix), me(eye), minami(south), miyako(city), mono(thing), nagara(even if), sakura(cherry), si(emphasic.), sumi(black ink), tatu(start,stand), tazumu(being around), tu(past.), uturou(change), watasu(give), yamakaze(mountain wind), yamu(stop), yanagi(willow), yononaka(world)

	cw	*z*	pattern		*cw*	*z*	pattern
1	3.86	3.18	yamu–manimani	106	2.38	1.31	si–fukakusa
2	3.75	3.04	minami–magiru	107	2.38	1.31	sakura–hana
3	3.67	2.93	minami–maku	108	2.38	1.31	sakura–isa
4	3.61	2.86	maku–magiru	109	2.38	1.31	sakura–ba
5	3.42	2.62	yanagi–koku	110	2.38	1.30	sakura–me
6	3.38	2.57	yamu–makasu	—			
7	3.38	2.56	mazu–koku	155	2.17	1.04	chiru–katu
8	3.27	2.43	yanagi–mazu	156	2.17	1.04	bakari–sumi
9	3.26	2.42	sakura–yamu	157	2.16	1.03	maku–besi
10	3.25	2.40	minami–yamakaze	158	2.16	1.03	tatu–maku
–				159	2.16	1.03	tatu–tazumu
101	2.40	1.33	uturou–komoru	160	2.16	1.03	tazumu–tu
102	2.40	1.33	sakura–watasu	161	2.16	1.03	miyako–sakura
103	2.40	1.33	katu–nagara	162	2.16	1.02	kakusu–si
104	2.39	1.32	sakura–masi	163	2.14	1.00	yononaka–sakura
105	2.39	1.31	sakura–makasu	164	2.14	1.00	mono–sakura

Conclusion

Using the distribution characteristics of cooccurrence weights, we were able to classify cooccurrence patterns into three layers of cooccurrence patterns: high-, mid-, and low-range patterns.

We found that 1) the distribution of classical texts fits a Gaussian curve as well as in modern texts; 2) the *cw* value can separate patterns into three layers (low-, mid-, and high-range) using inflection points (-1σ and 1σ); 3) of the three layers, the high-range could be extracted without a list of stop words; 4) the mid-range lexical layer might include mathematical traits not yet revealed in the present study.

References

Bronshtein, I.N., Semendyayev, K.A., Musiol, G., and Muehlig, H. (2004). Handbook of Mathematics: Springer-Verlag, 4th edition.

Hodoscek, B. and Hilofumi Y. (2013). "Analysis and Application of Midrange Terms of Modern Japanese", in Computer and Humanities 2013 Symposium Proceedings, No. 4, pp. 21–26.

Losee, Robert M. (2001). "Term dependence: A basis for Luhn and Zipf models", Journal of the American Society for Information Science and Technology, Vol. 52, No. 12, pp. 1019–1025.

Luhn, H. P. (1968). HP Luhn: Pioneer of Information Science: Selected Works: Spartan Books.

Manning, C.D. and Schutze, H. (1999). Foundations of statistical natural language processing, Cambridge, Massachusetts: The MIT press.

Nagao, M. (1983). Gengo kogaku (Language Engineering), Jinkochino sirizu 2 (Series of Artificial Intelligence): Shokodo.

Rajaraman, A. and Ullman, J.D. (2012). Mining of massive datasets, Cambridge: Cambridge University Press.

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

Spärk Jones, K. (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", Journal of Documentation, Vol. 28, pp. 11–21.

Yamamoto, H. (2006). "Konpyuuta 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.

Construction of Japanese Historical Hand-Written Characters Segmentation Data from the CODH Data Sets

Tang Yiping[1], Kohei Hatano[1], Emi Ishita[1], Tetsuya Nakatoh[1], Toshifumi Kawahira[1]

Introduction

Techniques for character recognition are of key components in the digital humanities. These days, there are more digital images of historical documents made, due to the development of scanners and computers. Although huge amount of such digital images are available, typical OCR (optical character recognition) systems for modern characters cannot be directly applicable to pre-modern character recognition problems. The hardness depends on languages. In particular, for Japanese, the difficulties of recognizing pre-modern hand written characters with computers are that (i) such documents are written by brushes and many characters are often connected, not separated by spaces, (ii) several different symbols (e.g., Chinese and Japanese ones) are used for meaning the same character, (iii) some characters are simplified or abbreviated. Therefore, it is still a challenge to recognize Japanese pre-modern texts from their images.

The recognition task can be divided into two phases, segmentation of sentences to single characters and recognition of single characters. Given an image of a single character, it is now an easy task to recognize the character, say, by using machine learning techniques such as the deep neural networks. For example, Nguyen et al. reported that their system can recognize single characters with accuracy 97% (Nguyen et al., 2017). On the other hand, segmenting an image of sentence to those of single characters is a bottleneck. Nguyen et al. also reported that the accuracy of their system for three consecutive characters is about 88%. So, a good segmentation algorithm will further increase the accuracy of recognition systems.

The goal of this work is to construct data sets of Japanese pre-modern text with the information of segmentation of sentences, for which researchers and developers could test their segmentation algorithms. Our data sets will be available through the QIR, the institutional repository of Kyushu university.

The CODH data sets and their variants

In 2017, the Center for Open Data in the Humanities (CODH) published open data sets of Japanese pre-modern characters and literatures (CODH, 2017a). The data sets consist of images of pre-modern Japanese books and their transcriptions as well as data sets of images of 403,242 individual characters of 3,999 different types. The data sets are released under the license of CC-BY-SA which can be freely used and modified if an appropriate citation is added.

Based on the data sets, the PRMU (PRMU, 2017), hosted the programming context of recognizing Japanese pre-modern hand-written texts (called the “Kuzujishi Challange") (CODH, 2017b) in 2017. In this context, they posted about 230,000 pages of Japanese kana characters data from the 15 Japanese ancient books from the CODH data sets, released it on the web site, so that it can be freely used by any potential participants. The contest data sets consist of tuples of an original image of particular text, the characters of interest (one to about six characters) which correspond to the “true" recognition result, and information of positions of characters expressed by the coordinates of the rectangle enclosing them.

The task of the contest is, given the image and the coodinates of the enclosing rectangle, to recognize the characters. In particular, the contest data sets has three types of data, the level 1, 2 and 3 depending on the hardness. The level 1 data set consists of

[1] Kyushu University

240,000 single characters. The level 2 data set consists of 80,000 sets of three consecutive characters. The level 3 data set consists of sets with multiple characters (more than 3). An illustration of these data sets is shown in Figure 1.

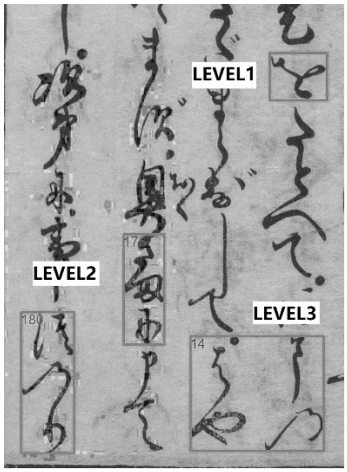


Figure 1: An example of contest data of the PRMU.

Construction of our data sets

Our technical work is to construct segmentation data sets by reforming the level 1,2, and 3 data sets from the PRMU contest. Our simple observation is that all the level 2 and 3 data sets contain single characters appearing in the level 1 data set. Therefore, by adding the information of enclosing rectangles of single characters in the corresponding, we can construct segmentation data sets of three or more characters. As a result, we construct 78,940 segmentation data sets of three consecutive characters and 12,583 multiple characters, respectively. More precisely, the each data is the tuple of the original image, information of a large rectangle enclosing three or more characters (the X and Y coodinates of the top-left corner, width and height), as well as small rectangles enclosing single characters within the large rectangle. Figure 2 represents examples of our data set.

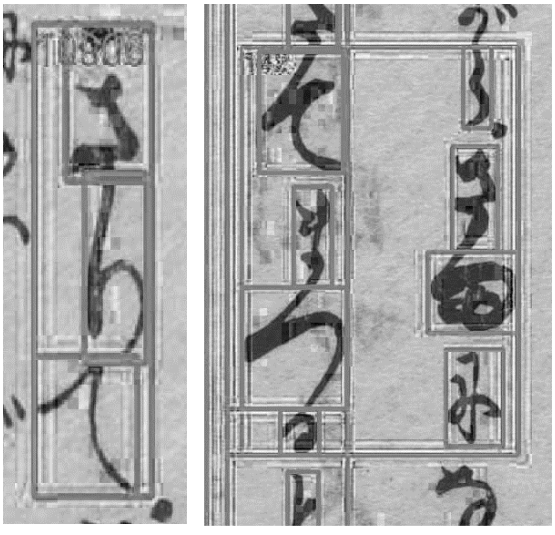


Figure 2: An example of our data set.

Furthermore, we also verified manually all of the constructed data. People checked the data set are 8 students specializing Japanese literatures, who can recognize Japanese historical characters. Through the manual check, we corrected 104 instances of the level 2 data set and 548 tuples of the data are excluded from our data set since students could not recognize them. Similarly, we corrected 95 instances of the level 3 data. Examples of difficult instances are shown in Figure 3.

Conclusions and future work

In this work, we constructed segmentation data sets of Japanese pre-modern characters from the CODH data sets and the PRMU contest. Our data set will be available under the license of CC-BY-SA at the web site. We will show some preliminary results of various segmentation methods over our data sets in the poster session.

Acknowledgement

We thank the support of the Tsubasa project at Kyushu University and JSPS KAKENHI Grant Number JP18K18508.

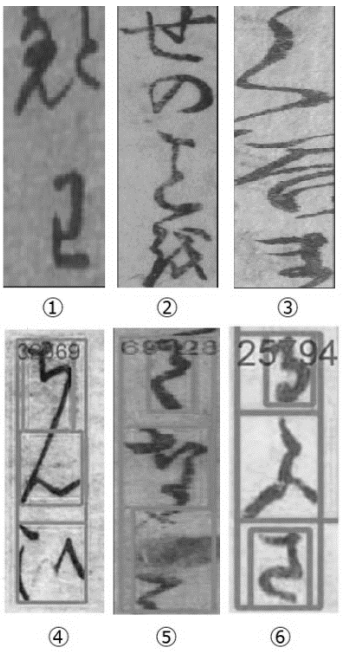


Figure 3: Examples of abnormal instances excluded from our data sets.

References

CODH. (2017a). http://codh.rois.ac.jp/char-shape/.

CODH. (2017b). http://codh.rois.ac.jp/old-char-challenge.

Nguyen, H. T., Ly, N. T., Nguyen, K. C., Nguyen, C. T., & Nakagawa, M. (2017). "Attempts to recognize anomalously deformed Kana in Japanese historical documents." In *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing* (HIP2017), 31–36.

PRMU. (2017). https://sites.google.com/view/alcon2017prmu.

How to Critically Utilise Public-sourced Open Data? A Proof-of-Concept: Enrich the SOAS Authority Datasets with Wikidata and VIAF

Fudie Zhao[1]

In the movement towards Open Data in the library sector, two types of sources have emerged: authoritative data sources initiated by academic institutions, like VIAF (Virtual International Authority File), and public-sourced open data, like Wikidata. The former type is better at the depth and quality of knowledge provided. The latter has an advantage in terms of the breadth of knowledge. How to critically utilise the collective public intelligence while avoiding its pitfalls remains debatable in Digital Humanities (Davidson, 2012). The proposed poster intends to contribute to the discussion upon this issue by sharing SOAS (School of Oriental and African Studies, University of London) library's proof-of-concept practice for leveraging VIAF and Wikidata to enrich its authority dataset in order to enhance its discovery service.

1) Why do we choose to use Open Data?

The library catalogue provides a united search across its four constituent bibliographic datasets (print, digital service, archive and institutional repository). However, the four datasets have different metadata standards and schemas. On top of that, since SOAS specialises in Asia, Africa, and Middle East studies, the datasets are multilingual and cross-script. These require an authority system to facilitate the united search (Mendias, 2017).

SOAS has a local authority dataset, however, it is outdated in several ways: 1) originally from the print's side, it does not fully meet the needs for the other three datasets; 2) it is not in keeping with an emerging trend in library sector to reconcile local authority files with other institutions and share a unique identifier for the same person/corporate; 3) some files are inadequate in multilingual and cross-script information.

SOAS library thus seeks to enrich its local authority dataset with external unique identifiers, as well as multilingual and cross-script data. SOAS library digital service team discovers that instead of manually adding the information one at a time, Open Data, like VIAF and Wikidata, may provide a better enrichment in quantity. Before the implementation, SOAS conducts a proof-of-concept to understand:

1) Which sources of Open Data should be selected?
2) What tools and methods are available for leveraging Open Data?
3) How effective is Open Data in serving SOAS's specific purpose?

2) Which sources of Open Data should be selected? VIAF or Wikidata?

As the quantity of Open Data available is increasing, how to choose appropriate sources become a question. SOAS has two options. VIAF is an institutionally supported resource, which combines name authority files from national libraries and other partners into a single name authority service. Wikidata, on the other hand, is sourced from a wider and more general public. Its initial datasets are derived from Wikipedia and other Wikimedian resources.

We have tried both VIAF and Wikidata reconciliation services in OpenRefine with our 25,472 sample entries for person names, and the results are as shown below (Fig.1 and Fig.2):

[1] University College London / SOAS Library Digital Service

Fig.1

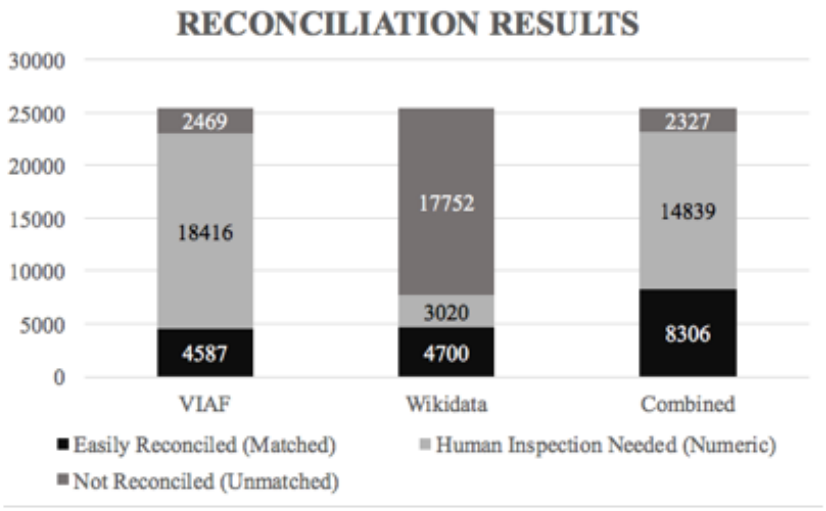
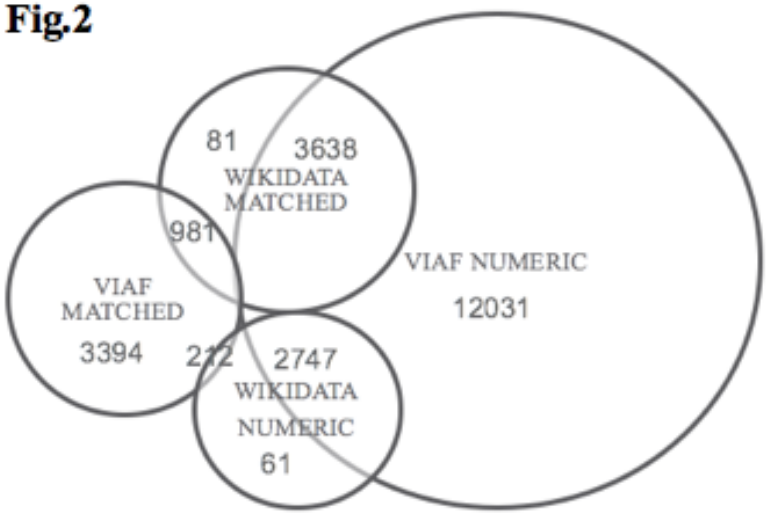


Fig.2



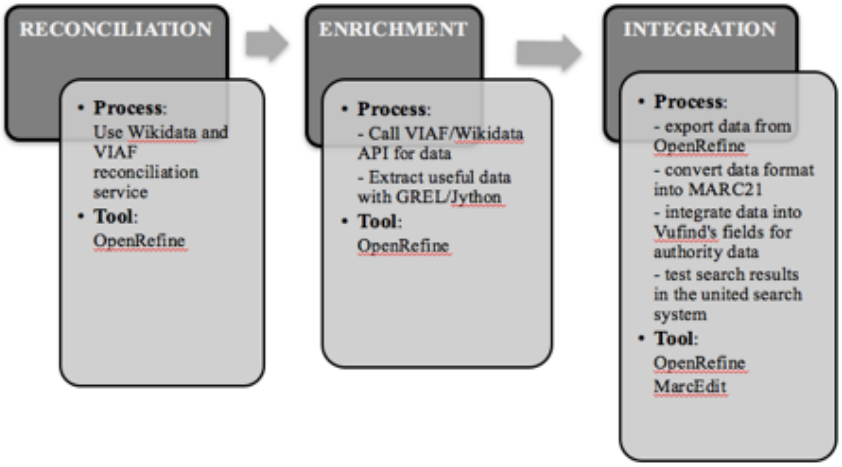
OVERLAPS OF RECONCILED ENTRIES

As is demonstrated in the stacked bar chart, on the whole, the sample dataset reconciles better with VIAF. However, after a further analysis, we discover that only 20% of VIAF's matches overlap with Wikidata's, which indicates the two sources are complementary. Therefore, a combined use of both sources provide the best reconciliation results for SOAS's dataset.

3) What tools and methods are used for leveraging Open Data?

SOAS uses free tools like OpenRefine and MarcEdit, and adopts a low-tech method to leverage Open Data. As there is still a significant amount of manual work required, it is better to follow a low-tech way, so that the library staff can pick up the skills quickly and work collaboratively. The whole process is as follows (Fig.3):

Fig.3



4) How effective is Open Data in serving SOAS's specific purpose?

Since the project is at an initial stage, metrics for measuring effectiveness is yet to be developed. Based upon the work we have done so far, data enrichment capability of VIAF and Wikidata can differ greatly according to individuals and regions.

Name entries for person in the library's local authority files range from worldly well-known people to those who have only one publication recorded. The former group have adequate information in both VIAF and Wikidata, while the latter lacks it. However, in terms of search enhancement, the former group weigh more, as they are normally related to more publications in the library and bring more trouble to the library's discovery service. Therefore, VIAF and Wikidata enable relatively automatic enrichment approach for those who affect the search most, while leaving the long tail to manual inspection.

The major factors influencing search returns differ in regions. For example, search problems caused by variations in writing systems in CJK regions (China, Japan, and Korea) share similarities in dealing with Chinese characters. However, they also have differences, which lead to different approach to measurement (Table 1). A specific region needs to be evaluated on customised standards to determine whether data provided by VIAF and Wikidata can facilitate the resource discovery.

Table 1

Possible Metrics Based Upon Differences in Writing Systems						
	Chinese		Japanese		Korean	
Similarity	chinese characters (hanzi, kanji, hanja)					
Differences	romanisation	script	romanistaion	script	romanisation	script
	Pinyin	simplified Chinese	Hepburn	kana (syllabic Japanese scripts)	revised romanisation of korean	hangul (Korean alphabet)
	Wade-Giles	traditional Chinese	Kunrei	Japanse-made Kanji	McCune-Reischauer	Korean-made Hanja

References

Mendias, C. (2017). "Project Charter: SOAS ADset". Project Proposal from SOAS Library and Information Services.

Davidson, C. (2012). Humanities 2.0: Promise, Perils, Predictions. In Gold M. (Ed.), *Debates in the Digital Humanities* (pp. 476-489). University of Minnesota Press. Retrieved from http://www.jstor.org/stable/10.5749/j.ctttv8hq.31