JADH2019

The 9th Conference of the Japanese Association for Digital Humanities

Localization in Global DH

http://conf2019.jadh.org/

2019.
8.29 THU → 31 SAT

KU-ORCAS International Symposium

East Asian Studies and DH

2019.8.30 FRI

KANSAI univ. Senriyama Campus

Supported by

ADHO
Alliance of Digital Humanities Organizations

KU-ORCAS DRAW YOUR ASIA
Kansai University Open Research Center for Asian Studies

Japan Art Documentation Society
IPSJ SIG Computers and the Humanities
Japan Society for Digital Archive
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

Proceedings of the 9th Conference of the Japanese Association for Digital Humanities
Edited by Nobuhiko Kikuchi and Hitoshi Ogawa

Published by:
Open Research Center for Asian Studies, Kansai University
3-3-35 Yamate-cho, Suita-shi, Osaka
http://www.ku-orcas.kansai-u.ac.jp/

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

Table of Contents

Keynote Session

Workshops

Session LP1: Developing DH

Session LP2: Analyzing society

Session LP3: Across internatonal borders

Session LP4: Challenging media

Panel session

Session SP1: Short paper session

Poster Session

JADH 2019 Committees

Program Committee:

- Paul Arthur (Edith Cowan University, Australia)
- Marcus Bingenheimer (Temple University, USA)
- Tarin Clanuwat (ROIS-DS Center for Open Data in the Humanities / National Institute of Informatics, Japan)
- James Cummings (Newcastle University, UK)
- J. Stephen Downie (University of Illinois, USA)
- Øyvind Eide (University of Cologne, Germany)
- Makoto Goto (National Museum of Japanese History, Japan)
- Shoichiro Hara (Kyoto University, Japan)
- Yuta Hashimoto (National Museum of Japanese History, Japan)
- JenJou Hung (Dharma Drum Institute of Liberal Arts, Taiwan)
- Jieh Hsiang (National Taiwan University, Taiwan)
- Akihiro Kawase (Doshisha University, Japan)
- Nobuhiko Kikuchi (Kansai University, Japan)
- Asanobu Kitamoto (ROIS-DS Center for Open Data in the Humanities / National Institute of Informatics, Japan)
- Chao-Lin Liu (National Chengchi University, Taiwan)
- Maciej Eder (Pedagogical University of Kraków, Poland)
- Yoko Mabuchi (National Institute for Japanese Language and Linguistics, Japan)
- A. Charles Muller (University of Tokyo, Japan)
- Hajime Murai (Future University Hakodate, Japan), Chair
- Kiyonori Nagasaki (International Institute for Digital Humanities, Japan)
- Satoru Nakamura (University of Tokyo, Japan)
- Chifumi Nishioka (Kyoto University, Japan)
- Ikki Ohmukai (National Institute of Informatics, Japan)
- Geoffrey Rockwell (University of Alberta, Canada)
- Martina Scholger (University of Graz, Austria)
- Susan Schreibman (National University of Ireland Maynooth, Ireland)
- Masahiro Shimoda (University of Tokyo, Japan)
- Raymond Siemens (University of Victoria, Canada)
- Donald Sturgeon (Harvard University, USA)
- Tomoji Tabata (Osaka University, Japan)
- Ruck Thawonmas (Ritsumeikan University, Japan)
- Toru Tomabechi (International Institute for Digital Humanities, Japan)
- Kathryn Tomasek (Wheaton College, USA)
- Ayaka Uesaka (Osaka University, Japan)
- Raffaele Viglianti (University of Maryland, USA)

- Christian Wittern (Kyoto University, Japan)
- Taizo Yamada (University of Tokyo, Japan)
- Natsuko Yoshiga (Saga University, Japan)

Local Organizers:

- Keiichi Uchida (Kansai University Open Research Center for Asian Studies) - Chair
- Takao Fujita (Kansai University Open Research Center for Asian Studies)
- Nobuhiko Kikuchi (Kansai University Open Research Center for Asian Studies)
- Hajime Murai (Future University Hakodate)
- Kiyonori Nagasaki (International Institute for Digital Humanities)
- Toru Tomabechi (International Institute for Digital Humanities)
- Hitoshi Ogawa (Kansai University Open Research Center for Asian Studies)

Time Table and Floor Map

August 29(Tue), Day 1

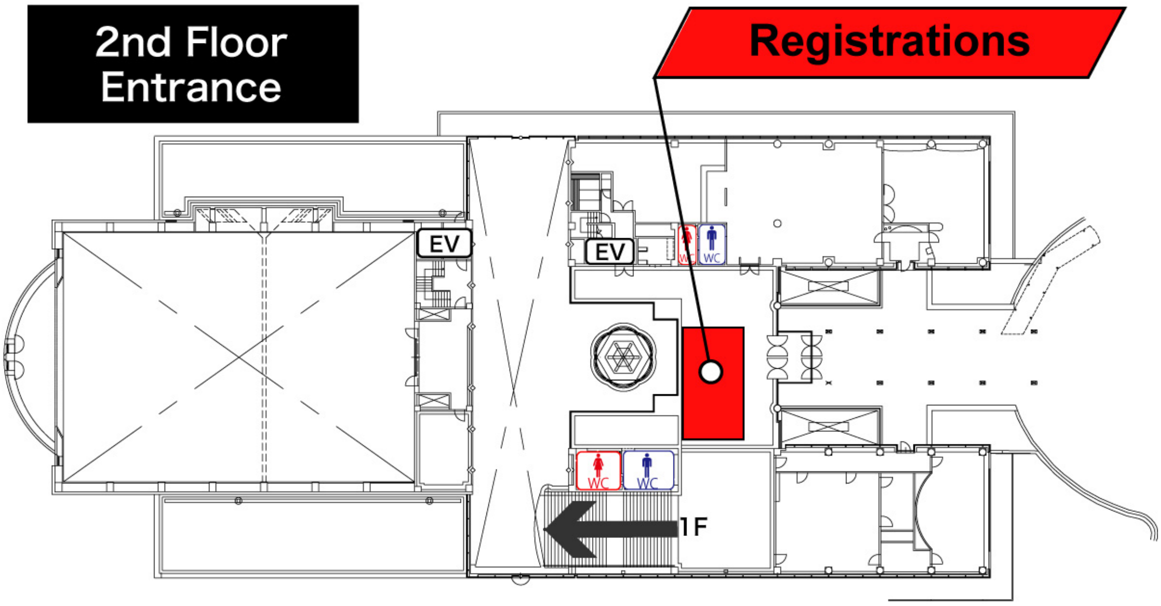
13:00-18:00	Workshop 1: TEI, Conference Room 3
13:00-16:15	Workshop 2: Morphological Analysis of Chinese Texts, Conference Room 4
13:00-16:15	Workshop 3: Japanese Map Warper, Conference Room 1

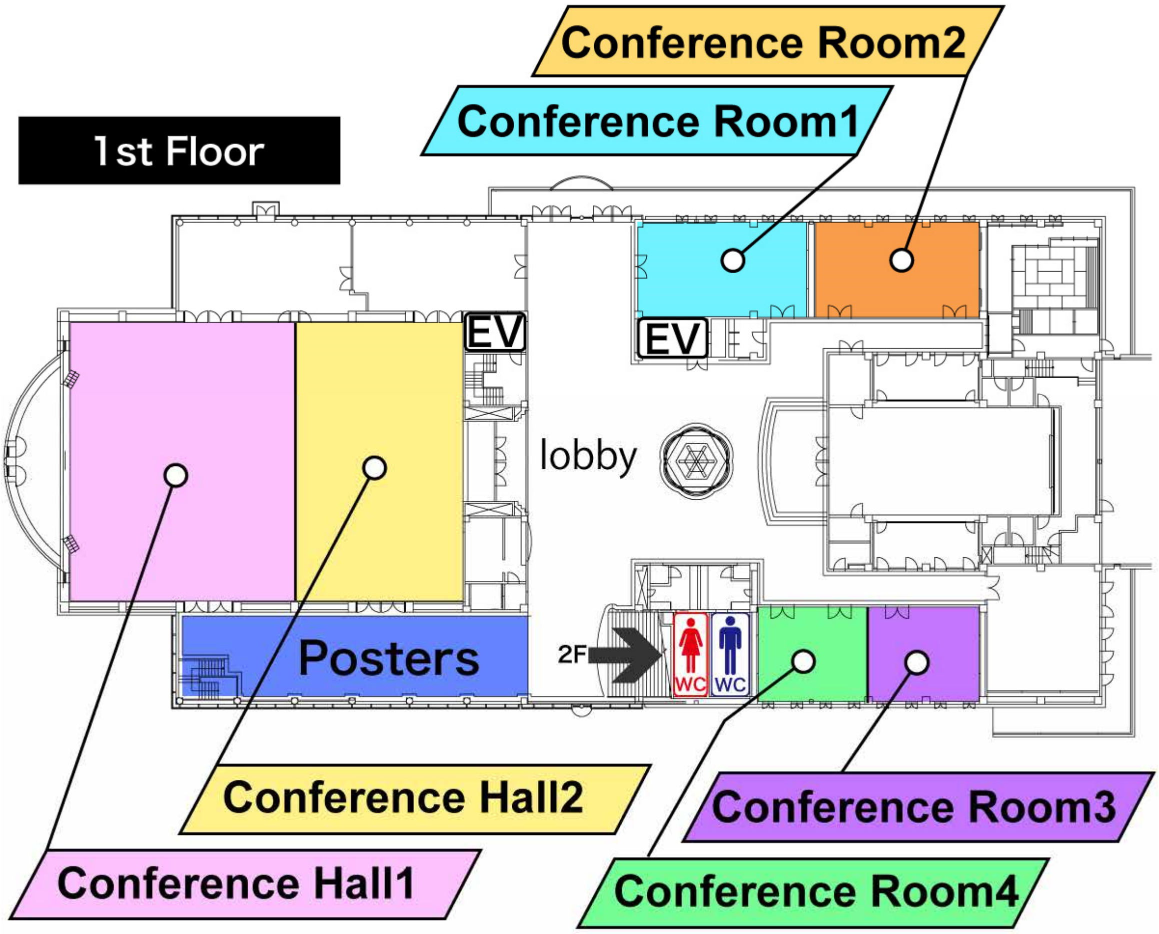
August 30(Fri), Day 2

9:30-9:45	Opening, Conference Hall 1
9:45-10:45	Session LP1 (2), Conference Hall 1
11:05-12:05	Session LP2 (2), Conference Hall 1
	Lunch break
13:30-15:30	JADH2019 Keynote Session and KU-ORCAS International Symposium, Conference Hall 1
16:00-17:30	Poster session with poster slam, Conference Hall 1
18:00-	Banquet, Conference Hall 2

August 31(Sat), Day 3

9:30-10:30	Session LP3 (3), Conference Hall 1
10:50-12:20	Session LP4 (2), Conference Hall 1
	Lunch on JADH Annual General Meeting, Conference Room 2
13:30-15:00	Panel session, Conference Hall 1
15:20-16:50	Short paper session (6), Conference Hall 1
16:50-17:10	Closing


2nd Floor
Entrance
Registrations
EV
EV
WC
WC
WC
WC
1F


1st Floor
Conference Room2
Conference Room1
EV
EV
lobby
Posters
2F
WC
WC
Conference Hall2
Conference Room3
Conference Hall1
Conference Room4

DRAW YOUR ASIA KU-ORCAS 関西大学 アジア・オープン・リサーチセンター

2017年度文部科学省私立大学研究ブランディング事業選定
「オープンプラットフォームが開く関大の東アジア文化研究」

関西大学 KANSAI UNIVERSITY

JADH2019 Keynote Session and
KU-ORCAS International Symposium

KU-ORCAS 国際シンポジウム

East Asian Studies and DH

参加費無料
事前申込要
当日参加歓迎

2019年8月30日(金)
13:30〜15:30
関西大学100周年記念会館ホール1

Program

13:30〜13:35
挨拶　内田慶市（KU-ORCAS）

13:35〜14:20
項　潔（台湾大学教授）
Harnessing Digital Resources
for Sinology Research

14:20〜14:30
〜休憩〜

14:30〜14:50
藤田高夫（KU-ORCAS）
Analysis of writing styles on wood slips
of the Han period

吉田壮（KU-ORCAS）
Image analysis for character region
extraction from wood slips

14:50〜15:10
菊池信彦（KU-ORCAS）
The KU-ORCAS's Digital Archives Project
for East Asian Studies

15:10〜15:30
小川 仁（KU-ORCAS）
Japanese Sources in the Vatican Library:
Takahashi Shomatsu and Ancient Shakyo

お申し込み
http://bit.ly/2LZTXw1

お問い合わせ
関西大学研究所事務グループ(以文館)
〒564-8680　吹田市山手町3-3-35
TEL：06-6368-1834　FAX:06-6368-0235
E-mail: ku-orcas@ml.kandai.jp

JADH2019 Keynote Session and KU-ORCAS International Symposium

East Asian Studies and DH

<Keynote Speaker>

- Harnessing Digital Resources for Sinology Research
  Jieh Hsiang (Distinguished Professor in Computer Science,
  Director of Research Center for Digital Humanities,
  National Taiwan University)

- Analysis of writing styles on wood slips of the Han period
  Takao Fujita （Professor, Faculty of Letters, Kansai University）

- Image analysis for character region extraction from wood slips
  Soh Yoshida (Assistant Professor, Faculty of Engineering Science, Kansai University)

- The KU-ORCAS's Digital Archives Project for East Asian Studies
  Nobuhiko Kikuchi (Project Associate Professor, KU-ORCAS, Kansai University)

- Japanese Sources in the Vatican Library: Takahashi Shomatsu and Ancient Shakyo
  Hitoshi Ogawa (Postdoctoral Fellow, KU-ORCAS, Kansai University)

[Workshop 1]
Workshop on TEI by SIG-EA/JP

Kazuhiro Okada, Satoru Nakamura, Kiyonori Nagasaki

Description

The workshop focuses on an introductory tutorial on textual markup according to Text Encoding Initiative (TEI), the international de-facto standard. As the tutorial is based on the activities of the Special Interest Group of East Asian/ Japanese, participants will learn about how to encode Japanese texts partially including characteristics of Chinese. After this workshop, participants will have skills of XML markup and structuring texts for humanities. (This workshop will be held mainly in English with some additional Japanese explanation).

Relevance to the DH Community

While TEI is a basic technology for digital humanities in the world, it has not spread enough among Asian countries. The workshop will provide an opportunity to bridge the gap.

Instructors

Kazuhiro Okada, Assistant professor, National Institute for Japanese Literature
Satoru Nakamura, Assistant professor, The University of Tokyo
Kiyonori Nagasaki, Senior fellow, International Institute for Digital Humanities

Target Audience and Prereqs

This workshop will offer the tutorial for people who are interested in how to structuralize text data for humanities especially for Japanese studies. Each participant will be expected to have a basic knowledge on HTML and/or XML.

Each participant must bring a laptop computer which enables to run Java environment, connect Wifi, and has USB type A slot to insert a USB memory.

Language

English

Workshop Outline

1. How texts should be structuralized for humanities?
 - Background: interoperability and sustainability
 - Diversity in the humanities

- Some applications for TEI
2. Markup for a letter (1)
 - Trying markup for a letter with an introductory tutorial
3. Markup for a letter (2)
- Trying markup for a letter with an introductory tutorial
4. Q&A

References

https://tei-c.org/ (in English)
https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html (in English)
https://doi.org/10.24576/jadh.1.0_3 (in Japanese)
https://github.com/TEI-EAJ/jp_guidelines/wiki (in Japanese)

[Workshop 2]
Morphological analysis and lexical indexing of Chinese

Yoshihiro Hino

Description

We will explain two themes: Chinese morphological analysis and creation of a full vocabulary index. With regard to morphological analysis, we will specifically explain services with relatively easy access to the web that can be used on the web, focusing on services such as NLPIR / THULAC / CTA / HanLP, and give lectures on how to use them. Furthermore, as one of the usages of the text subjected to morphological analysis, in order to extract a word included in a specific book and create an entire lexical index which can reveal how many times the vocabulary appears on which page We will introduce the available services and how to use them.

Relevance to the DH Community

It is related to DH communication because of the basic skills required to use text data obtained from books etc. as research resources.

Instructors

I am a lecturer at the Mejiro University, and a guest Researcher at Kansai University Open Research Center for Asian Studies. Specialized field is Chinese education and Chinese education history, Chinese analysis using a computer. The current research theme is a study on the history of Chinese education, focusing on the 19th to 20th century Chinese teaching materials. In recent years, we have started researching such as catalog creation and digitization of Chinese teaching materials, construction of a teaching material corpus using text data included in Chinese teaching materials of each era, extraction of basic vocabulary by morphological analysis, and the like.

Target Audience and Prereqs

People who are interested in Chinese morphological analysis and lexical indexing. All you need is basic computer knowledge. As you do the actual work, please bring a PC connected to the Internet.

Workshop Outline

The first half of the workshop will give an overview of Chinese morphological analysis by giving examples. In the second half we will look at some of the morphological analysis services we will introduce. In addition, a text service is marked up using a web service operated by a reporter called “Index-converter”, and the creation of an entire vocabulary index is practiced.

[Workshop 3]
Geo-referencing and mashup old Japanese maps with Japanese Map Warper

Ryo Kamata

Description

Map Warper is an online geo-referenceing tool developed by Mr. Tim Water in 2009. Map Warper is an open-source software and you can download the source code from GitHub.com and install it to your own server freely as long as you keep its license. With Map Warper, you can specify control points to warp uploaded map images via web browser. You can download or access to the geo-refferenced images in several format, such as KML, GeoTIFF, Tile URL and etc.

In the workshop, we will describe the background of the tool and an instance Japanese Map Warper, we have established as portal site of Japanese old maps. In the hands-on section, we will try to do geo-referencing with illustrated maps with Japanese Map Warper and utilize them with external services, such as Google Earth.

Relevance to the DH Community

The workshop describes:

- Why and how we have established such service
- cases of utilization of open-source software for DH research

Instructor(s)

- Developer of Geolonia (https://geolonia.com), a venture company which provides GIS services
- Developer of TaroSky inc.
- Part time teacher of Ritsumeikan University
- Expert of API and User Interface designing

Target Audience and Prerequisites

- Basic computer operation
- (optional) Please bring any old maps their copyrights and other rights are cleared and you want to try geo-reference with Map Warper if you have
- If you have a laptop, please bring it

Language

- Japanese

Workshop Outline

- Total: 2 classes
- Introduction of Map Warper (first half)
- Background
- System
- About Japanese Map Warper
- Hands-on (latter half)
- Registration
- Uploading Maps
- Geo-referenceing
- Mashup with external services

References

2017 年度　国際ワークショップ「日本の古地図ポータルサイト」
https://www.arc.ritsumei.ac.jp/GISDAY/2018/workshop.html

オープンプラットフォームによる日本の古地図オンラインの構築
矢野 桂司, 鎌田 遼
2017 年度日本地理学会春季学術大会
https://www.jstage.jst.go.jp/article/ajg/2017s/0/2017s_100313/_article/-char/ja/

ジオリファレンスソフトウェア Map Warper の導入事例紹介
鎌田 遼・矢野 桂司
FOSS4G 2017 KYOTO.KANSAI
https://www.osgeo.jp/events/foss4g-2017/foss4g-2017-kyoto-kansai-coreday-timetable

日本版 Map Warper の構築と活用
矢野桂司・鎌田遼
GIS 学会 2017 年度学術研究発表大会
http://www.gisa-japan.org/news/file/2017_abstract1029_ver.3.pdf

日本版 Map Warper を用いた旧版地形図の公開
GIS 学会 2018 年度学術研究発表大会
http://www.gisa-japan.org/file/1012_poster.pdf

Developing a Digital Humanities and Social Science Learning Environment in view of e-Learning

Chyi-Kuan Wang, Chih-Ming Chen, Chiao-Min Lin, Lin-Kuei Tsai, Ting-Wei Chiang, Wei-Yuan Fan[1]

With the improvement of digital technology, digital tools have become a transforming element for teaching and learning. The shift also affect humanities in many ways, including education in universities. Since 2010, Bunde and Engel have proposed the idea of partnering computing skills and humanities in undergraduate education. The rapid development of Digital Humanities has also represents the urgency of establishing a new education model, which merges humanities knowledge and the use of digital tools. However, in Taiwan, many people are unfamiliar with the field, it limits the development of potential talents. Moreover, the area of Humanities suffers from insufficient resources in developing a new education system, as well as the area of Social Science. In order to solve the problem, the Ministry of Education in Taiwan cooperates with National Chengchi University on "The Digital Humanities and Social Science Course Database and Forum Project." The main objective of the project is to build up a new learning environment that is able to support the need in teaching, self-directed learning, and improve communication between teachers and students.

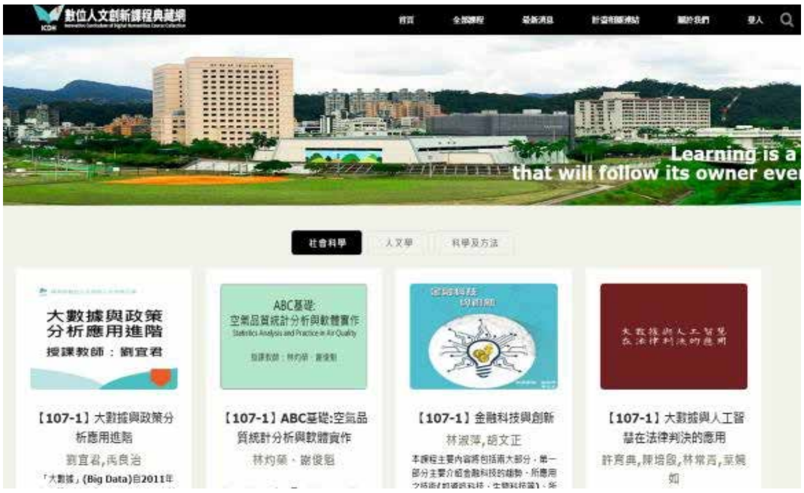
The purpose of the paper is to describe the building process of this new kind of e-learning environment for Digital Humanities based on ADDIE model. This model contains five stages: Analysis, Design, Development, Implementation, and Evaluation. In the Analysis stage, the project team found out that, although many teachers and students are eager to know more about the Digital Humanities, without a clear view of the field, it is hard for teachers to formulate new approaches in teaching. Thus, students are unable to understand the concept nor using certain digital tools properly.

Lack of applying digital resources in their research leads to low information literacy, and this is the reason why the development of a new learning environment becomes the first priority. In the Design and Development phase, we intended to develop a learning environment, which not only provide platforms for teaching and learning, but also allows people with humanities background to work with digital talents, as well as holding educational dataset for digital humanities.

Based on the first three stages, in the Implementation phase, we created five platforms: Innovative Curriculum of Digital Humanities Course Collection; Innovative Curriculum of Digital Humanities Courses in Moodle; Innovative Curriculum of Digital Humanities Courses Forum; Datasets for Curriculum of Digital

[1] National Chengchi University

Humanities Courses; and Talent Match in Digital Humanities Database. These platforms are capable to store multi-media resources, and also offer rich interactive components to user. Each database has its own distinct function. The Innovative Curriculum of Digital Humanities Course Collection and the Innovative Curriculum of Digital Humanities Courses in Moodle mainly focus on collecting the courses delivered by Digital Humanities specialists. These online course providers are served as customized systems, more specifically, they would be modified based on users' requests. They are also ask to hold the teaching materials offered by the specialists, which includes video lectures, handout, and other documents that are relevant to the courses, such as syllabus and a list of references. Therefore, the system designers use open source software as the systems' architecture, the former uses Dspace and the latter adopts Moodle. From February 2018 until June 2019, we have collected over 80 courses, and were categorized by three subjects, Social Science, Humanities and Science and Methodology. Because anyone can access these courses for free, we pay full attention on the copyright of the teaching materials. Through asking the specialists to join the Open Education program at Creative Commons works at the beginning of each semester, we are able to support CC mission through education.



【107-1】大數據與政策分析應用進階
劉宜君,禹良治
「大數據」(Big Data)自2011年

【107-1】ABC基礎:空氣品質統計分析與軟體實作
林灼榮、謝俊魁

【107-1】金融科技與創新
林淑萍,胡文正
本課程主要內容將包括兩大部分，第一部分主要介紹金融科技的趨勢，所應用之技術(如通訊科技，生物科技等)，第

【107-1】大數據與人工智慧在法律判決的應用
許育典,陳培殷,林常青,葉婉如

Figure1: The Homepage of Innovative Curriculum of Digital Humanities Course Collection

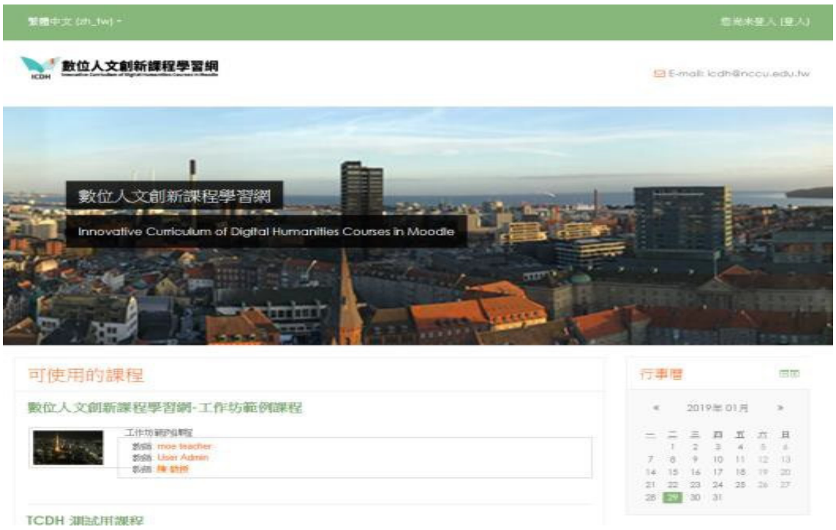
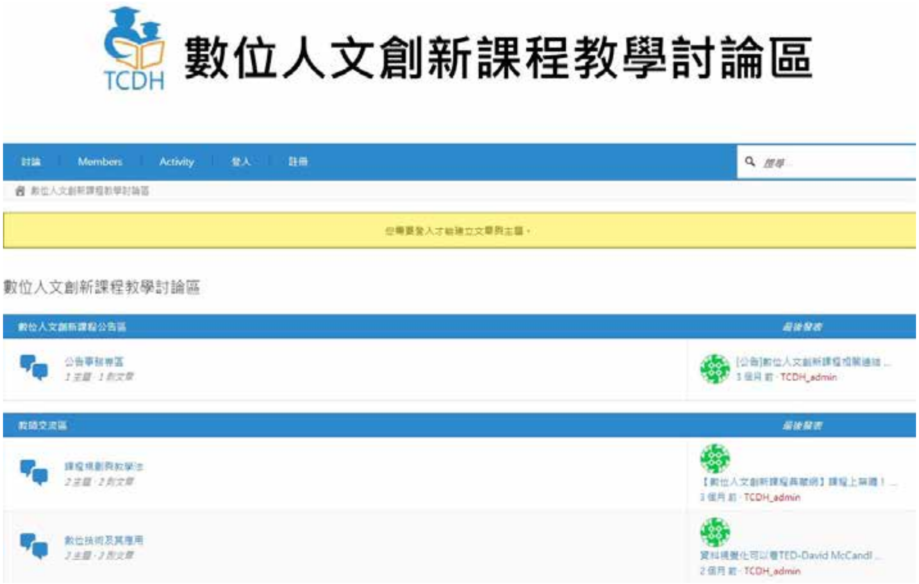


Figure 2: The Homepage of Innovative Curriculum of Digital Humanities Courses in Moodle

The later three databases are used to support teachers and students who are not familiar with Digital Humanities by providing them with a dataset of resources built by research institutes or local governments, and promote exchange between specialists and novices. As the paper mentioned above, these databases we developed are prone to build up on open source software, in order to cope with constant modification. The Innovative Curriculum of Digital Humanities Courses Forum applied Asgaros Forum, a free Wordpress plugin, as the system architecture. It helps junior faculties to overcome challenges in teaching. Experienced teachers in Digital Humanities field can share their teaching experience with others, and figure out solutions together. Moreover, experts in the field can also make suggestions to certain courses. Currently, the posts on the forum are classified in two categories, Announcement and Discussions. The construction of Datasets for Curriculum of Digital Humanities Courses is due to certain teachers and students' unfamiliarity with Digital Humanities, which we found in the analysis phase. Therefore, the dataset was designed to provide big data resources for teaching, and it combines the basic features of Moodle with MySQL relational database. Last but not least, the Talent Match in Digital Humanities Database is developed under Dspace architecture, and applies the features of PostSQL in designing the back-end database. This dataset aims to provide information of professors in the field of Digital Humanities, who can assist teachers, students, and corporations in searching for interdisciplinary collaborative partners. At the time of writing, the project is in mid-way to completion, and Assessment plan is ongoing.



數位人文創新課程教學討論區

討論 | Members | Activity | 登入 | 註冊 | 搜尋...

數位人文創新課程教學討論區

您需要登入才能建立文章與主題。

數位人文創新課程教學討論區

數位人文創新課程公告區	最後發表
公告事務專區 1 主題 · 1 則文章	[公告]數位人文創新課程相關連結 ... 3 個月 前 · TCDH_admin

教師交流區	最後發表
課程規劃與教學法 2 主題 · 2 則文章	【數位人文創新課程募職網】課程上架囉！ ... 3 個月 前 · TCDH_admin
數位技術及其應用 2 主題 · 2 則文章	資料視覺化可以看TED-David McCandl ... 2 個月 前 · TCDH_admin

Figure3: The Homepage of Innovative Curriculum of Digital Humanities Courses Forum

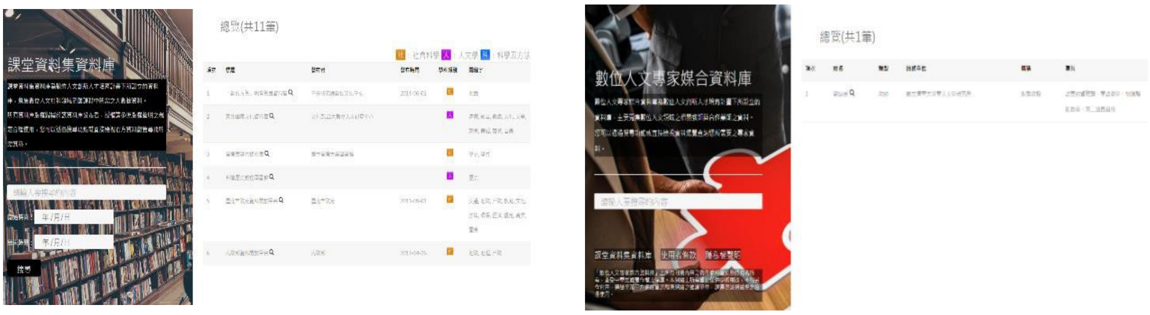


Figure4: Datasets for Curriculum of Digital Humanities Courses and Talent match in Digital Humanities Database

Digital Humanities is the key of cultivating interdisciplinary talents, and the e- Learning environment we built offers several benefits to teachers, learners, or even corporates. With the help of the tools and information technology specialists, the researchers in humanities can collect and analyze the data in a more effective way.

References

Bunde, J., Engel, D. (2010). Computing in Humanities: An Interdisciplinary Partnership in Undergraduate Education. Journal of Archival Organization, 8(2): 149-159.

The History of Australian Digital Humanities

Paul Longley Arthur[1]

This paper discusses the history of digital humanities in Australia, referring to major projects and events. The launch in 2002 of the Australian e-Humanities Network and Gateway was a milestone in the field's development. This database was the culmination of an Australian Research Council grant led by the Australian Academy of the Humanities. In its time the portal gave exposure and recognition to a vast variety of pioneering digital projects and examples of digitisation and scholarly efforts to create enhanced digital resources for preservation and access in the early web era.

From the beginning of the new century, conferences began to play a significant role. The conference Computing Arts: Digital Resources for Research in the Humanities, at the University of Sydney in 2001, has been regarded as the first major conference devoted to issues in humanities computing generally in the Australia-Pacific region. It was followed in 2004 by a further conference in the series, held at the University of Newcastle.

Around this time, projects started to demonstrate a more conscious commitment to interdisciplinarity and to the development of digital methods with a view to long-term sustainability of online resources. There were many centres, teams, projects and individuals that played a part in the long-term development of digital humanities in Australia.

Ground-breaking research in language and textual studies, for which Australia is internationally regarded, has been undertaken at the Centre for Literary and Linguistic Computing, established in 1989 at the University of Newcastle, where John Burrows pioneered early techniques in stylometry. The centre's continuing work has been led by Hugh Craig from 2001 to 2018. This foundational program has included development of techniques for stylistic analysis and authorship attribution including the Delta, Iota and Zeta methods.

The projects of the Archaeological Computing Laboratory (later Arts eResearch) at the University of Sydney, including TimeMap from 1995 and those built on the Heurist reference database system since 2005, have featured an emphasis on enabling end users to create their own data management solutions and on web-based mapping with a temporal dimension. They include the award-winning Digital Harlem, developed by University of Sydney historians to visualize and explore the spatial dimensions of everyday life in Harlem during its heyday, 1915–1930; and the Dictionary of Sydney from 2004.

[1] Edith Cowan University, Australia

AustLit was started in 1999 by a consortium of universities, led by the University of Queensland and ADFA at the University of New South Wales, initially with the National Library of Australia. The AustLit consortium has built upon this base to become what is now considered the world's most comprehensive record of a nation's creative writing across all forms and genres, and critical works associated with that output. AustLit has also developed a specialisation in Aboriginal and Torres Strait Islander writers.

Further projects initiated at the turn of the century included AusStage, the Australia Live Performance Database. The first phase of the project involved theatre scholars from eight Australian universities, with the Australia Council. Through successive grants, AusStage has evolved over a 20-year period, with a recent focus on linking to international collections and visualising Australian live performance venues.

The Consortium for Research and Information Outreach (CRIO, 2001–2009) at the Australian National University (ANU) had strengths in digital anthropology and filmmaking, forming the foundation for the Digital Humanities Hub (now Centre for Digital Humanities Research), established in 2010. The centre is now a leading research concentration indicative of the growing institutionalisation of the field nationally.

Paradisec (the Pacific and Regional Archive for Digital Sources in Endangered Cultures) began in 2003 with an original focus on the Asia-Pacific region but has since expanded to include materials from all over the world, thus providing citable research data as well as access for speakers of languages. With over eleven hundred languages included to date, in 2016 the resource was given a Special Commendation by the UK's Digital Preservation Awards.

The *Australian Dictionary of Biography* is the premier reference resource for significant lives in Australian history. The ADB online initially utilised the Online Heritage Resource Manager (OHRM) database system developed by the Australian Science and Technology Heritage Centre (Austehc) at the University of Melbourne, but since 2011 has been hosted independently at the Australian National University. Numerous other projects have utilised the OHRM including the Encyclopedia of Australian Science, Encyclopedia of Melbourne, Australian Women's Register and the Find & Connect Web Resource (jointly with the Australian Catholic University). Another premier reference resource and biographical project with a long history, Design and Art Australia Online was launched in 2007. It drew content from the *Dictionary of Australian Artists*, first published in 1984, but had its origins in the 1970s.

In Australia, as in other parts of the world, libraries have played a pivotal role in supporting the evolution of digital humanities. One of the best-known national digital projects is Trove. With its genesis in the late 1990s, Trove was formally planned in 2008 as a portal to the National Library of Australia's discovery services.

Trove has grown to become a full-text repository resource providing access to a vast array of information about Australia and Australians, an aggregator of diverse digital content and also a highly successful crowdsourcing platform for the correction of OCR-digitised content of Australian newspapers.

Through such initiatives it is clear that digital humanities points to a future in which researchers will be able to utilize comprehensive data from many different sources nationally and internationally, to form diverse and inclusive infrastructures and methods for data sharing and analysis, to further advance knowledge and understanding and to develop new skills in the broad field of humanities and social sciences.

References

Austlit (n. d.). Retrieved from https://www.austlit.edu.au

AusStage (2018). The Australian Live Performance Database. Retrieved from https://www.ausstage.edu.au

Australian Dictionary of Biography (1966–). Australian Dictionary of Biography. Retrieved from http://adb.anu.edu.au

Design and Art Australia Online (n.d.). About the DAAO. Retrieved from https://www.daao.org.au

Paradisec (the Pacific and Regional Archive for Digital Sources in Endangered Cultures). (n.d.). Retrieved from http://www.paradisec.org.au

Trove (n.d.). Retrieved from https://trove.nla.gov.au

From Tables to Graphs:
The Korean Yangban Network Data in Cytoscape and Neo4j

Javier Cha[1]

[1] College of Liberal Studies, Seoul National University

Equal Rights over Child Custody in Taiwanese Transnational Marriages: Natural Language Processing and Machine Learning in Legal Text

Hsuanlei Shao[1]
Siehchuen Huang[2]

There are getting more and more transnational (international) marriage cases in Taiwan today. We often name the foreigner of the marriage as “***New Immigrants* （新住民）**” who come from global area then becoming an important part in our local community. As an ethnic group, the *New Immigrants* are facing ***equality/ inequality*** of social right and legal right. The article would discover it by method of ***Digital Humanities (DH),*** and focus in their judge practice and legal text.

The topic we noticed in this article is the ***Child Custody Legal System.*** This system in Taiwan is a legal term which is used to describe the legal and practical relationship between a parent and a child in that person’s care, such as the right to make decisions on behalf of a child and the duty to care for and support the child. Usually the legal parent of the child has custody. If the child has two parents who are married to each other, both parents are entitled to child custody. This becomes a problem when the married couple decides to divorce, such that one parent will no longer live with the other and it becomes necessary to decide on a new child custody arrangement. In general, like all issues arising from divorce—including property division, child support, and alimony—child custody will either be decided by agreement between the divorcing couple or, in the case that an agreement cannot be reached, by the court.

Theoretically no matter where you come from and which nation you are belong to, you should be treated equal in court. But a situation can be seen in practice: there are many transnational marriages in Taiwan. Most of them are a Taiwanese man marry a woman who from Southeast Asia, is doing a labor work, often poor income, and weak social network. A “***new immigrant spouse***” who might have an unbalance social status to a Taiwanese one. It comes an instinct question: ***Is it equal a local/ foreigner person to use the legal system in Taiwan in practice? If yes/ no, how and why?***

In the other hand, The United Nation’s (U.N.) principles of equality and non-discrimination are part of the foundations of the rule of law. Taiwan government seems itself as a member states who noted in the ***Declaration of the High-Level Meeting on the Rule of Law***. The government would like to follow the principles in

[1] National Taiwan Normal University

[2] National Taiwan University

it, such as: “***’All persons, institutions and entities, public and private, including the State itself, are accountable to just, fair and equitable laws and are entitled without any discrimination to equal protection of the law***”(para. 2). They also dedicated themselves to respect the equal rights of all without distinction as to race, sex, language or religion (para. 3): ***The international human rights legal framework contains international instruments to combat specific forms of discrimination, including discrimination against indigenous peoples, migrants, minorities, people with disabilities, discrimination against women, racial and religious discrimination, or discrimination based on sexual orientation and gender identity.*** Equality of every citizen is a global issue.

We used the NLP and ML method on the legal text. We checked two main hypothesis: 1. Are there different “style” in local/ foreigner partner judge? 2. Are there word (concerned terms) in local/ foreigner partner judge? Regarding the research processing: 1. Collecting data: In this article, we processed 119 legal texts (judgements), which are about 3,000~10,000 Chinese characters. 2. NLP Method: we could input the raw data (unprocessed txt file), then pre-processing, segment, ifidf, then structured as a matrix automatically. 3. Then we use tree-based machine learning model to train and test the dataset. Mainly the random-forest model and gradient descent algorithm. 4. It could get average accuracy over 90 percent.

We can figure out ***“Aboard, Appear”*** are key factors which can differ the NI/TW the cases. As the bellowed table:

Table.1: Mean Tfidf of ***“Aboard, Appear”*** in NI/ TW jedgements

	Mean(TFIDF_aboard)	Mean(TFIDF_appaer)
New Immigrants	0.0371	0.0339
Taiwanese Both	0.000642	0.0208

The result shows the Judge concerns if NI is still in Taiwan. In the judgement, the Judge asks “if they aboard/ appear to the court?” and we can tell ***“Rent, live together”*** are key factors in NI and TW cases, too.

Table.2: Mean Tfidf of ***“Rent, live together”*** in NI/ TW jedgements

	Mean(TFIDF_rent)	Mean(TFIDF_live together)
New Immigrants	0.0155	0.0515
Taiwanese Both	0.00495	0.0341

And the research results are: 1. We cannot see specific style judges which can support us there is unequal judge in the corpus. 2.We can figure out some “key terms” can divide local/ foreigner partner judge, such as **“appear /aboard”**, ***“Rent,***

live together" etc.… which is about legal processing terms. It shows the Judge concerns if NI is still in Taiwan. The Judge asks "if they aboard/ appear to the court?" When we know these key factors more, we could give the judge or the party better advice.

Relating the equality issue of the new immigrants in Taiwan, we can find some difficult cases in former researches, interviews, reports and oral history. Peoples also aware there might be some problem in particular cases which often not to be seem in official statistics. But it is not a clear status up till the present moment. The article attempts to apply the method of ***Digital Humanities (DH)**,* to find if there is ***equality/ in equality of Custody Right in Taiwanese Transnational Marriages***. In detail, we will study hundreds real legal texts (judgement texts) by ***Natural Language Processing (NLP)*** and ***Machine Learning (ML)***. Then we can find a specific "terms" which can direct the legal reason in local/ foreigner judgement texts. It can prove how inclusive a new immigrants in Taiwan society nowadays by not a case studies, but a general way. This is an example that shows *local issue* (***Taiwanese New Immigrants***) could be solved by a global DH way (***NLP and ML***) which can be fitting the title of ***JADH2019: "Localization in Global DH"***. And it is also a ***global human right issue*** of Equality of social and legal right in a specific ethic groups in ***Taiwanese local community***.

Additionally, it is a fine practice in cross-domain work which have NLP ***Modern Chinese texts*** with special domain knowledges (**jurisprudence**) in theory and practice.

Keywords: Legal Analysis, Digital Humanities, Natural Language Processing(NLP), Machine Learning(ML), Equality, Custody, Transnational Marriages.

The Poetics of Scale:
The Convergence of Digital Humanities and World Literature

Youngmin Kim[1]

The Digital Humanities platforms provide us with the future orientation of our world literature studies which turns from close reading to distant viewing. In particular, when one reflects upon one's confronting the "other" literatures and cultures, one recalls the vortex of inbound authenticity and outbound hybridization in the big data of the literatures and cultures of the world. This dynamic vortex, when gathered and seen from the perspective of the structured data, will construct the database of world literature which will cover the ethnic, racial, cultural, and national border-crossing intermixtures. One can find mapping, digital reconstruction of social network, and large scale visualization as the potential representations of this vortex as we witness in such examples as Baldwin's Paris (baldwinsparis.com), Six Degrees of Francis Bacon (bit.ly/6-d-bacon), and On Broadway (on-broadway.nyc).

Technological innovations change quickly and the field of digital humanities expands, and more and more materials come online, through research projects. Other repositories or platforms combine technical, academic and cultural issues at a scale that is unprecedented. When one clicks one of the DH literary platforms, one can get the sense of the scales of these different projects. The image of geographic scale becomes more and more a set of distinct platforms upon which geopolitics and other social phenomena are performed. World systems theory posits the global sphere as the most important scale. Locality studies have privileged the local as the scale at which meaning or lived experience is constructed. The paradoxical positions taken on local, national, and global scales were starting points for much of the critical discourse on scale.

Cartographically "scale" implies the graphic scale bar or the representative fraction (RF) on a map. Richard Howitt applies the concept of the "scale" to representation of "glocalization," a double movement of the local and the global which contextualizes "the simultaneous and contested shift up-scale towards the global and down-scale to the local as a response to changing economic, political and cultural pressures" ("Scale," *A Companion to Political Geography,* 142). This concept of glocalization from the perspective of "up-scale" and "down-scale" is in fact zooming-in and zooming-out of the object of observation which ironically turning closely and going away distantly. Scale becomes the ground for an interactive map of world

[1] Professor, Department of English, Dongguk University, Seoul, Republic of Korea

literature studies to understand the dynamic interrelatedness of the whole of literature as in the so-called "The World Republic of Letters."

Casanova's rationale for the world republic of letters, her version of world literature, is based upon the scale of "aesthetic distance" of spatio-temporality. In another 2005 essay, "Literature as a World," Casanova provides a conceptual model of "world literary space," which she explains as "a body of literature expanded to a world scale, whose documentation and, indeed, existence remains problematic—but a space: a set of interconnected positions, which must be thought and described in relational terms" (277). To support her argument, Casanova provides the metaphor of the Persian rug. In Henry James's "The Figure of the Carpet," she deploys the beautiful metaphor of the Persian rug, which appears as "an indecipherable tangle of arbitrary shapes and colors." From the right angle and proper distance, the attentive observer can detect the carpet of "superb intricacy," when the observer can see "in their totality, in their reciprocal dependence and mutual interaction," but not when viewed "casually or too close up." Each figure in the carpet can be grasped only in terms of "the position it occupies within the whole, and its interconnections with all the others" (277). In fact, Casanova envisions each text in world literature on the basis of their "relative position within this immense structure," and Casanova's project is "to restore the coherence of the global structure within which texts appear, and which can only be seen by taking the route seemingly farthest from them; through the vast, invisible territory which [Casanova] called the 'World Republic of Letters.'" Casanova herself interprets her own conceptual tool of "world literary space," that each of the inhabitants or authors of the Republic of Letters are differently situated within their own national literary space, and each writer's position is dual: the position he/she occupies in a national space and place he/she occupies within the world space. This double position is inextricably national and international.

In fact, world literature reminds us of the big data of the literatures and cultures of the world. This big data, when gathered and seen from the perspective of the structured data and put in terms of the concept of "scale," can be seen in terms of "database." World literature as database is closely related to the concept of glocalization from the perspective of "micro-scale" and "macro-scale," since close reading and distant reading is, in fact, zooming-in and zooming-out of the data as we observe analytically and phenomenologically. What is at stake is the paradoxical/double positions of local, national, and global literary texts in world literature.

When literatures and cultures encounter those of the other counterparts in terms of the big data or statistics of a new reconfiguration, the tangential points of the borderland will be reduced to what W. J. T. Mitchell calls "a mere abstraction on a map," which nevertheless will provide the interstitial zone of "intersections, competition, and exclusions" for potential interpretation after one is equipped with the elaborated topic analysis of the literary data. Structured data beyond natural language processing (NLP) data is a crucial aspect of Digital Humanities work, providing the standards in data formats for the purpose of making it possible for digital humanists to extract, search, analyze, find, and style the data in files. Then, one needs to demonstrate some examples of how text analysis, text encoding, Mark-Up, TEI, GIS mapping, distant reading and cultural analytics, visualization, network analysis, semantic web, and other fields of digital humanities will contribute to the potential platform for the future convergent area of world literature and digital humanities.

References

Baldwin's Paris. baldwinsparis.com.
Casanova, Pascale (2004). *The World Republic of Letters*. Cambridge, Mass.: Harvard UP.
Casanova, Pascale (2013). "Literature as a World." In D'Haen T, César Domíngues, C, and Thomsen M. (eds), *World Literature: A Reader*. London and New York: Routledge, pp. 275-88.
Howitt, Richard (2003). "Scale." *A Companion to Political Geography.* In Agnew, M. and Toal, M. (eds), MA: Blackwell Publishing. p. 142.
Mitchell, W. J. T. (2013). "Border Wars: Translation and Convergence in Politics and Media." *English Language and Literature* 59.3 (2013): 343.
Montello, D. R. (2001). "Scale in Geography." In Smelser, N. J. and Baltes, B. (eds). *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier, 13501-13504. Web.
On Broadway. on-broadway.nyc
Six Degrees of Francis Bacon. bit.ly/6-d-bacon.

Who Studies Japan, What and When: Analysis of Journal of “Japanese Studies” by Machine Learning and Natural Language Processing

Hsuanlei Shao[1]

While studying *knowledge areas/ knowledge communities*, we researches often be limited by our resource, and shall make one's choice to cost/performance. But we could enhance our vision by modern technology. In this article, we try to analyze 2,237 research articles from *The Journal “Japanese Studies”* between 1991~2017, which is the journal published by *Institute of Japanese Studies, Chinese Academy of Social Sciences (JSCASS)* with methods of the Machine Learning and the Natural Language Processing.

In detail, we use the LDA(Latent Dirichlet allocation) model to cluster 2,237 research articles. The LDA is the one of the generative statistical models that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. It often used on English-based NLP studies, such as like: David M. Blei, John D.Lafferty, “A Correlated Topic Model of Science”, The Annals of Applied Statistics, Vol. 1, No. 1 (Jun., 2007), pp. 17-35 and John Laudun, Jonathan Goodwin, “Computing Folklore Studies: Mapping over a Century of Scholarly Production through Topics”, The Journal of American Folklore, Vol. 126, No. 502 (Fall 2013), pp. 455-475. But it still not so popular in Chinese-based NLP studies. That is why we would like to try in this article.

The process mainly can be described with five stages(*Figure 1. Research Processing*):

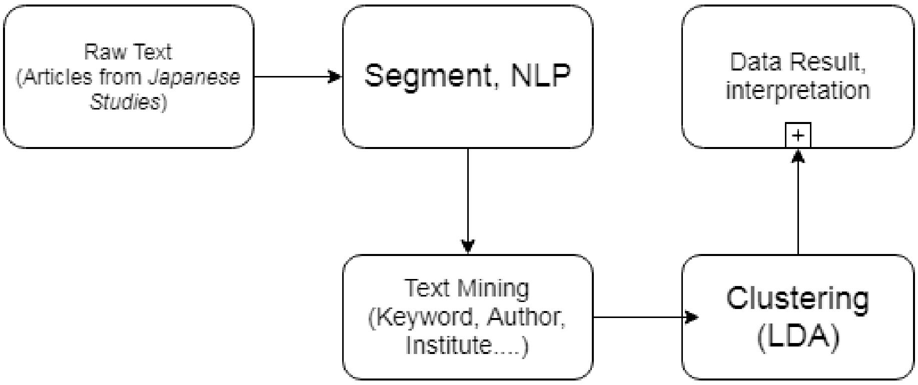

Raw Text
(Articles from *Japanese Studies*)
Segment, NLP
Data Result, interpretation
+
Text Mining
(Keyword, Author, Institute....)
Clustering
(LDA)

Figure 1. Research Processing

[1] National Taiwan Normal University

1. We collect the raw data (2,237 research articles) by from the well-established database. (CKNI, CSSCI…)
2. We make a corpus for this project, segment it.
3. We picked up some keywords, in order to filter information (like the institutes and authors)
4. Use LDA model to cluster every text.
5. We researchers can interpret (explain) the cluster meaning by our domain knowledge.

In this stage, I could show the table (Table 1. Keywords of each Cluster in 2,237 ″Japanese Studies″ articles) about step 4 and step 5, which shows the keywords (by Chinese) of the clusters after-LDAing.

Table 1. Keywords of each Cluster in 2,237 ″Japanese Studies″ articles

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**
Economics	*Politics*	*Managemen t*	*Ideology*	*Culture Society*	**IR**	*Social Economy*
金融	**憲法**	**企業**	**文學**	**文化**	**產業**	**日元**
邦交	**行政**	**技術**	**農業**	**意識**	**釣魚島**	**政黨**
經濟學	**戰爭**	**民間**	**資料**	**文明**	**對華**	**國際化**
哲學	**主義**	**公司**	**江戸**	**宗教**	**台灣**	**財政**
全國	**民族主義**	**科技**	**藍皮書**	**朝鮮**	**製造業**	**民主黨**
銀行	**軍國主義**	**工業**	**社會科學**	**婦女**	**安保**	**家庭**
金融危機	**靖國神社**	**經濟體制**	**小說**	**神道**	**海洋**	**資本主義**
貨幣	**小泉**	**價格**	**儒學**	**女性**	**產業結構**	**共同體**
東京	**思潮**	**立國**	**身份**	**佛教**	**能源**	**地方**
人民	**天皇制**	**日本式**	**農村**	**國民性**	**規制**	**政局**
城市	**天皇**	**法人**	**德川**	**藝術**	**地緣**	**物價**
國際交流	**右翼**	**市場經濟**	**武士**	**甲午戰爭**	**聯合國**	**匯率**
方針	**教科書**	**外國**	**幕府**	**形象**	**友好條約**	**人口**
馬克思主義	**保守主義**	**產業政策**	**輿論**	**帝國主義**	**夥伴關係**	**議員**
東北亞	**日本國**	**資產**	**差距**	**日語**	**沖繩**	**國債**
貨幣政策	**感情**	**範式**	**作品**	**思維**	**威脅論**	**政界**
文化交流	**策略**	**職能**	**投票**	**特性**	**主席**	**迴圈**
歐洲	**太郎**	**歐美**	**品質**	**福澤諭吉**	**競爭力**	**國會**
日本銀行	**倫理**	**科研**	**農產品**	**交流史**	**雙邊**	**景氣**
貸款	**議會**	**股票**	**工程**	**性格**	**福田**	**政治體制**
存款	**國家主義**	**汽車**	**韓國**	**史學**	**琉球**	**地震**

泡沫	家族	網路	天津	語言	空心化	勞動力
學術研究	信任	序列	規範	方法論	朝鮮半島	土地
利率	史觀	科學	大陸	關係史	中曾根康弘	居民
區域合作	侵略擴張	收支	獎評	列島	政經	社會保障
學派	歴史系	員工	朱子學	規律	石油	鳩山
人士	日本首相	經驗教訓	武士道	加藤	和約	稅制
面向	事變	法制	文學史	風土	國際法	債務
債權	權利	經貿	辦法	神話	日方	中央
日本學	三國	關係史	學術論文	印度	改革開放	石油危機

※ The raw text is Chinese-based corpus, so you provide it on it basically.

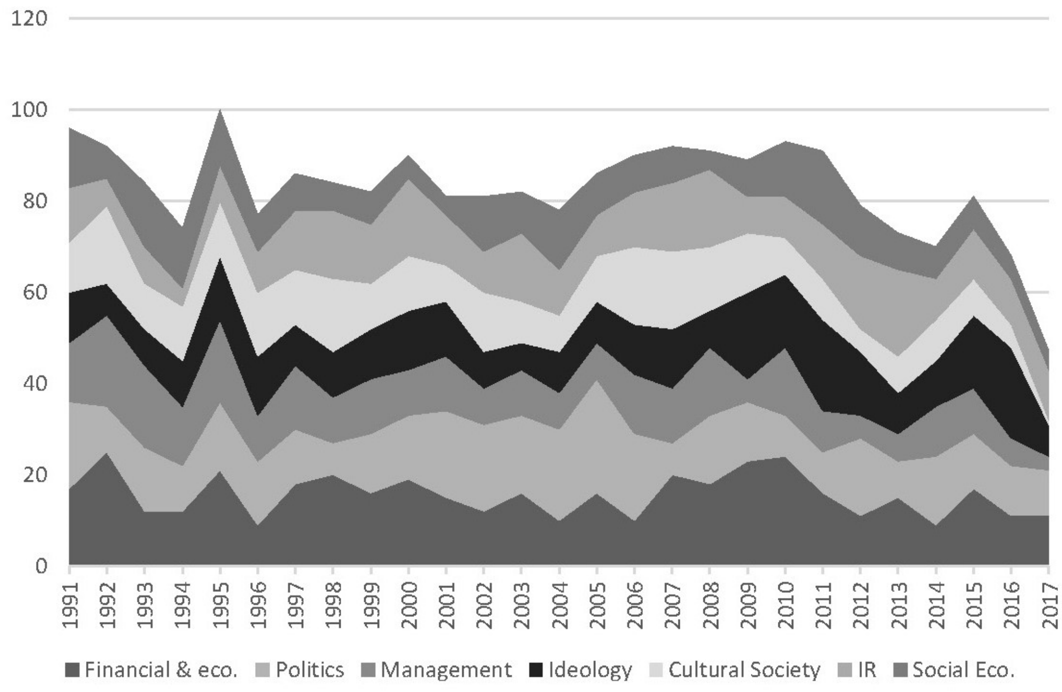
Then we could also calculate every cluster quantity by year, and visualize it. (table 2 and figure 2)

Table 2. Clustered Article of ”Japanese Studies” Number Per Year

LDA clusters	Financial & eco.	Politics	Management	Ideology	Cultural Society	IR	Social Eco.	Total
1991	17	19	13	11	11	12	13	96
1992	25	10	20	7	17	6	7	92
1993	12	14	18	8	10	8	14	84
1994	12	10	13	10	12	4	13	74
1995	21	15	18	14	12	8	12	100
1996	9	14	10	13	14	9	8	77
1997	18	12	14	9	12	13	8	86
1998	20	7	10	10	16	15	6	84
1999	16	13	12	11	10	13	7	82
2000	19	14	10	13	12	17	5	90
2001	15	19	12	12	8	11	4	81
2002	12	19	8	8	13	9	12	81
2003	16	17	10	6	9	15	9	82
2004	10	20	8	9	8	10	13	78
2005	16	25	8	9	10	9	9	86
2006	10	19	13	11	17	12	8	90
2007	20	7	12	13	17	15	8	92
2008	18	15	15	8	14	17	4	91
2009	23	13	5	19	13	8	8	89

2010	24	9	15	16	8	9	12	93
2011	16	9	9	20	9	12	16	91
2012	11	17	5	14	5	16	11	79
2013	15	8	6	9	8	19	8	73
2014	9	15	11	10	9	9	7	70
2015	17	12	10	16	8	11	7	81
2016	11	11	6	20	5	10	5	68
2017	11	10	3	7	1	11	4	47
Total	423	373	294	313	288	308	238	2237

Figure 2. Overlay of Clusters of ”Japanese Studies” Number Per Year



Then we could answer 1. Who/ what institutes study Japan. 2. What fields can be clustered in Japanese Studies, and their keyword. 3. The number flows of the fields by year. We have practiced to analyze thousands Chinese long text, and captured the key information semi-automatically, in order to support reader to simpler complex information efficiently. We could know highly productive writers, highly productive institutes, what they publish per year, and main fields of Japanese Studies, which are “*Economics, Politics, Management Science, Literature/Ideology, Culture/Society, IR, and Social Economy”* in this research progress. The research could provide a macro perspective in the *knowledge areas/ knowledge communities*.

I could use a borrowing idea from one of my DH colleague. It is called “long distant prospective". We can topic few or dozens essays by reading hard. But in a research field which have thousands or more articles? It only be possible by DH way. Therefore, to ask "Who Studies Japan? By what topics of a grand research field" can be a answerable question. This is I can see the way we can use this result, of course, also in other fields like DH.

Localization in the Context of Japan of a Large-Scale Relational Biographical Database on the Model of the China Biographical Database

Michael A. Fuller[1]
Bettina Gramlich-Oka[2]

The China Biographical Database (CBDB) draws on a wide range of historical data to model social experience in premodern China. Its integration of kinship structures, social networks, geographic networks, of the participation of individuals both in a range of social institutions and in institutional hierarchies allows users to explore large-scale interactions in the social, cultural, political, and religious history of premodern China in a way that has not been possible heretofore.

The more facets of historical experience that can be included in the database, the more powerful its relational analyses become. CBDB has been fortunate in having access to GIS data available from the China Historical GIS project, although even here there is significant continuing research to refine our understanding of historical administrative units across China. Much of the other historical information that serves as the foundation of the database has grown from research projects focused on particular aspects of premodern history. Scholars are still reconstructing the bureaucratic structures of the successive dynasties, and we still need a broader representation of the types of social institutions, such as temples and local academies, that played important roles in Chinese social activity.

This paper presents the challenges faced when localizing the CBDB to Japanese history. In early 2017, an older version of the current Japan Biographical Database (JBDB) was converted to the basic architecture of the CBDB. Directly applying the specific design of the Chinese model to the data of Japanese history did not work perfectly, owing to historical differences and particularities between the two cultures and societies. Even when the focus is only on premodern Japan, various issues had to and still must be overcome.

One challenge is to adapt the design of data structures to reflect differences in Japanese social experience. Although CBDB's basic concept of a database modeling the key entities in the structure of social experience remains as applicable to Japan as to China, in a database of Japanese prosopography the structures will need to be revised to better accord with the realities of Japanese historical experience. For example, in Japan there exist kinship relations that are rare in China and that will need additional strategies to be adequately reflected in the modeling. Another issue

[1] University of California, Irvine

[2] Sophia University

in data design is that the functioning of social institutions and official bureaucracies in Japan does not match the Chinese models and will require some rethinking of the tables to better represent these entities in the database. More recently, moreover, the JBDB added new tables that accommodate modern contexts, thus changing the database's original purpose of modeling social experience during the Tokugawa period (1600–1868).

A second challenge for developing a relational model for premodern (and also ideally modern) Japanese biographies is institutional rather than conceptual. Much information about people, places, and institutions already exists within the scholarly community in Japan, and it would be foolishly inefficient to recreate rather than integrate those different domains of data. Part of the challenge, then, is how to integrate the existing naming authorities for people, places, and institutions spread out among many heterogeneous databases. In Japan, NII and the Diet Library use different systems, but since VIAF is fed by both, JBDB is exploring issues around use their ID numbers. Without a unified coding of IDs for all of the entities in the database, the database cannot function properly. Thus such a project would require identifying possible participants who might contribute data and working approaches to data access. While there is a trend in Europe to design prosopographical databases with RDF and LOD standards to define entities in order to allow greater integration across databases, these paradigms remain cumbersome for developing JBDB APIs, given the structure of the JBDB entities, and like CBDB, JBDB for the moment plans to use a more ad hoc approach to assuring data exchange across databases within the Japan cyberinfrastructure. As we explore the possibility of integrating external data structures for GIS information, for additional biographical information, and for data on social institutions from repositories in Japan, we initially will examine the challenges of mapping between data structures on a case-by-case basis.

A third issue for a Japanese relational database for premodern biographies is institutional at a different level: where would the project be housed and how would it be funded and maintained? This is no small matter and depends very much on the local infrastructure for digital humanities in Japan. We hope to learn about possibilities and challenges from the experiences of other participants in the JADH conference.

Potentials and challenges of a data-driven perspective on videogame production, distribution and reception

Martin Roth[1]

This paper presents a data-driven perspective on Japan's videogame culture. Videogames have changed significantly throughout history. As one of the frontrunners of both technological development and industrialized entertainment culture, they are increasingly perceived as a global, pervasive culture. However, there is also considerable doubt about this globalization hypothesis among experts (Consalvo, 2016; Kerr, 2017). In my paper, I show that drawing on a wide range of metadata helps us to move beyond such general claims of universality and the global character of videogames, instead providing insight into the complexity at work in videogame production, distribution and reception. Focusing on the case of Japan is particularly interesting in this regard. Japan's videogame culture has been a very important influence on videogame cultures around the world since the early 1980s, producing widely popular content like the *Super Mario* or *Pokémon* franchises, as well as successful game consoles. At the same time, a significant share of professional and amateur (indie and *dōjin*) videogame production in Japan targets and remains within the confines of the domestic market.

Metadata, meaning data about data, exists in many forms, ranging from advertisements, teasers and manuals, to playthroughs on Youtube or Twitch, fan-driven wikis and walkthroughs published in print or online. For the purpose of this research project, we used data from diverse datasets ranging from fan- and commons-based data like Wikidata (https://www.wikidata.org/) or Mobygames (https://www.mobygames.com) to publicly funded, research-driven databases like the Japanese Media Arts Database (https://mediaarts-db.bunka.go.jp) or the Leipzig Wortschatz project (https://wortschatz.uni-leipzig.de/de). As I hope to show, the data available online offers a rich, fine-grained resource for videogame research. This serves to highlight the skill and effort fans and volunteers put into representing their favorite media.

Integrating heterogeneous datasets offers a partial solution to a perceived lack of reliable data in the field of videogames (Kerr, 2017: 31). However, such diversity has its challenges for the researcher. In my paper, I introduce some of these challenges related to access, integration, analysis and visualization from the perspective of a media culture researcher. Introducing several concrete research cases, I show that the character of the used databases promotes asking certain

[1] Leipzig University and Stuttgart Media University

questions over others; the scope of the data determines the possible scope of the research; their ontologies (in the information science sense of the categories and relations between concepts, data and entities) determine the compatibility with other data and influence the research perspective in many ways.

More broadly speaking, the data-based perspective on videogame culture I develop in this paper is an attempt at approximating the complex character of the videogame medium. As I hope to show, the working solutions my research group (https://diggr.link) has developed shed light on this complexity and allow tracing the multilayered local and global entanglements Japan’s videogame culture has been built on since the early 1980s. This methodology can potentially be applied in a comparative study of videogame cultures around the world. At the same time, the paper emphasizes some of the challenges a data-driven perspective on media culture confronts both cultural researchers and data scientists with. Doing so, I hope to contribute to a discussion about the potentials and limitations diverse data sources have for a more adequate understanding of the medium videogame. Analyzing the various classification systems created in scholarly and fan community practice furthermore serves as a starting point for discussing how the politics of classification (Bowker and Star, 2000) are informed by the complexity of the medium and, in turn, influence videogame cultures across the globe.

Bibliography

Bowker, G. C. and Star, S. L. (2000). *Sorting Things out: Classification and Its Consequences*. First paperback edition. (Inside Technology). Cambridge, Massachusetts London, England: The MIT Press.

Consalvo, M. (2016). *Atari to Zelda: Japan’s Videogames in Global Contexts*. Cambridge (Mass): The MIT Press.

Kerr, A. (2017). *Global Games: Production, Circulation and Policy in the Networked Era*. New York & London: Routledge.

Predicting Book Circulation Across a Large Public Library System

John Shanahan[1]

This paper explains findings and tools that link book circulation data, social media data, city data, and textual features data from several years of a large city-scale literary program. The Reading Chicago Reading (RCR) project is a digital humanities project studying the One Book One Chicago program (OBOC) hosted by the Chicago Public Library (CPL) and repeating annually city-wide since fall 2001. Our research group recently completed its first predictive models to discover relationships between text measures, book circulation over time, and social media.

While our data comes from one city (Chicago), our techniques and model can be applied elsewhere. Our project offers researchers an opportunity to formulate repeatable and testable hypotheses about reading culture at city-scale. The project has grown from some core digital humanities tasks such as text processing to quantify and then compare textual features data in a corpus of in-copyright texts by using the secure data capsule of the HathiTrust digital library.

This paper will first cover inputs of our predictive model, for example the 80-branch CPL book circulation data:

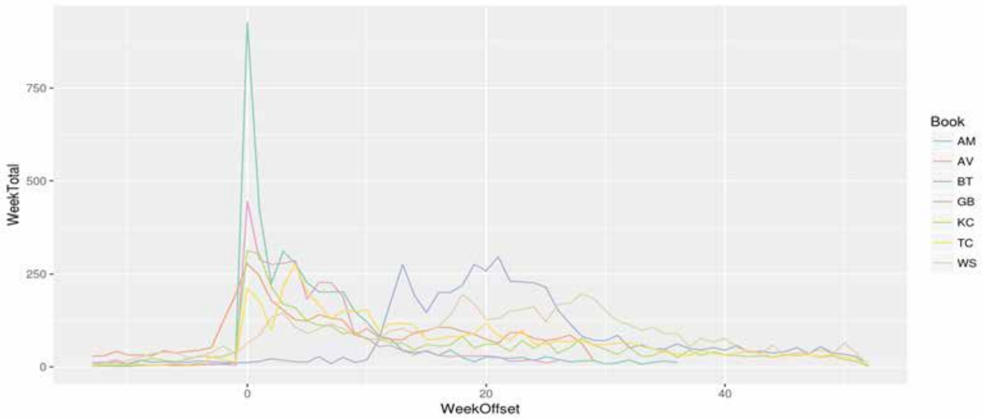


Figure 1: book circulation (all branches) for seven seasons (2011-2017) with CPL launch date as zero on the x axis.

The time series in Figure 1 allows us to grasp the effects of promotion by CPL. One can see that in most cases after an initial burst of interest (i.e. checkouts) created by the launch date, the checkout totals decline sharply. However, notably books present different shapes of secondary and tertiary spikes for their city-wide checkout totals.

Figure 2 shows a (logarithmic) visualization of normalized circulation at each

[1] Department of English, DePaul University, Chicago, USA

branch for each book.

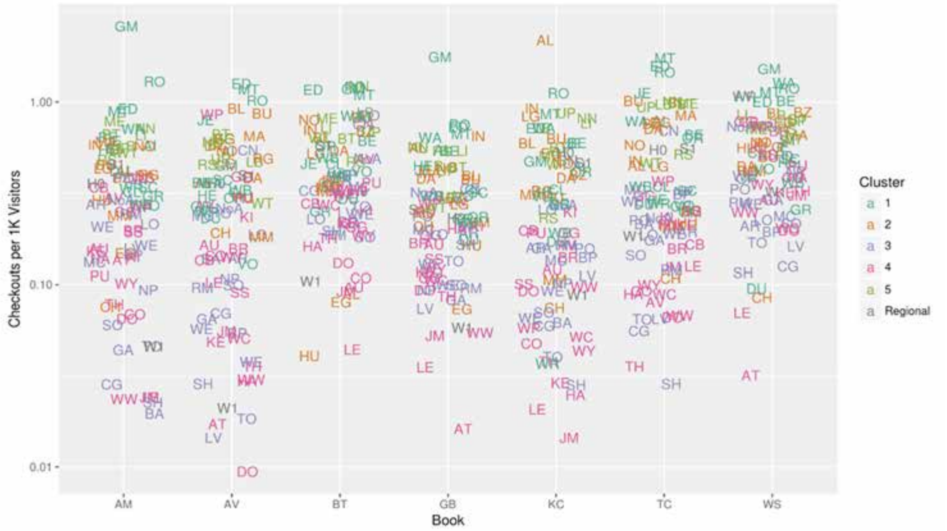


Figure 2: Normalized branch circulation by book, colored by cluster.

The cluster colors of branches indicate how six different classes of neighborhood demographics might correlate to interest in OBOC books.

The dependent feature of our circulation models is normalized circulation value: that is, checkouts at a given branch per 1000 visitors. The independent variables are combined reading difficulty score, degree of promotion (number of events at a branch), and locality. The values are combined with the variables from a previous model: three demographic variables and the number of holdings. For the first model, MPrior, we added each book’s prior circulation (previous 90 days) at that branch.

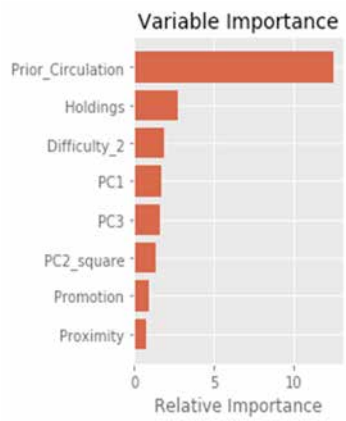


Figure 3: Average feature importance for MPrior circulation models

With prior circulation as an input variable, we enable MPrior to predict impact of the library’s selection of a particular book. In other words, what is the change in circulation pattern induced by a book’s selection? The model without prior circulation (MCirc, below) is interesting to us for a different reason: with this model, we predict *de novo* what the match between demographic and book characteristics might say about a book’s popularity at a given branch given a particular level of investment by the library system (in promotions and book holdings).

We added three principal components drawn from city neighborhood demographic data. The model type we chose is a boosted regression tree, an appropriate model for learning the relationship between a numeric output variable and a diverse set of input variables. Our model was trained and evaluated using a cross-validation technique. In each step of evaluation, one OBOC season was omitted and the model was trained on the other six seasons. Then the model was used to predict the missing year and the error calculated. This was repeated across all seven years, and averages computed across all years. The average mean absolute error for MPrior was 0.017. (Recall that all values were normalized between 0 and 1, including the circulation.) This average 1.7% corresponds to about 7-8 checkouts in a given branch. The corresponding value for MCirc is 6%, or about 3.5 times as high. This is not surprising as this model has much less information to work with. However, it is still within 10% of the actual checkout total. The importance of prior circulation is very high, and reflects the interest of a particular set of patrons in particular text. The other variables in descending order of importance are Holdings, Difficulty, PC1, PC3, PC2X, Promotion, and Proximity.

Figure 4 contains similar importance values for MCirc, in which prior circulation is not considered. Here we see that the roles are reversed between the set of demographic variables and the holdings/reading difficulty measure. The prior circulation variable in MPrior to some extent builds in the baseline appeal of the book to the patrons of a particular branch, and when this variable is removed demographic factors become a stronger element. Here we see effects of principal components 1, 3, and 2X, similar to those found in our multi-level model. Once we get past these effects, we have a consistent pattern relating holdings, reading difficulty, promotional events, and proximity.

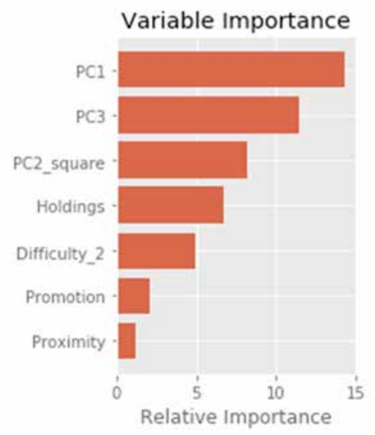


Figure 4: Average feature importance for MCirc models.

Figure 5 shows the average feature importance for an MBinary model. The results for MBinary are quite important as they demonstrate that high accuracy in toponym attribution is not essential to making use of locality in our circulation modeling. It is sufficient to label a book as Chicago-connected or not, and we expect that this will be possible without the manual effort required to achieve high accuracy for each geographic label.

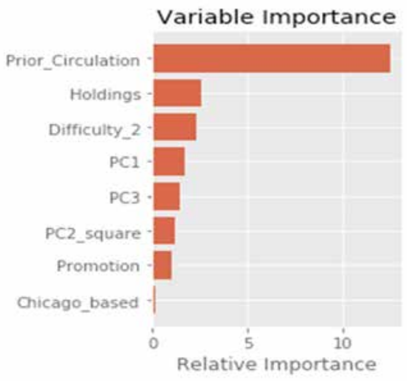


Figure 5: Average feature importance for MBinary models.

One of the key findings of our predictive model is that prior circulation makes the largest predictive contribution for the circulation of OBOC selected works. It is possible, as we have shown, to do similar types of predictions without prior circulation data, but with significantly lower accuracy. Our ability to quantify the effect with these models and forthcoming "dashboard" will enable library staff to reason about the tradeoffs inherent in choosing texts already circulating well in the system as opposed to "importing" choices from outside the system in the name of expanding readers' horizons.

Models of Context Discovery Systems for Different Archives

Jieh Hsiang[1], Chijui Hu[2], Chih-Yang Huang[3], I-Mei Hung[4]

After decades of digitization effort, large quantities of historical resources are now available on the Web. Systems that are designed to use these materials in research are highly anticipated by historians. However, different material requires different thinking in how they can be utilized. In this panel we present four different system designs, each tailors to a different type of digital historical archives. All of them were developed by the Research Center for Digital Humanities of National Taiwan University and have been heavily used by scholars.

Although each system caters to the specific archival material, a common design principle is that they are not based on the precision/recall retrieval model. Instead, they regard a search result as a meaningful set and try to find the textual context such as contexts based on metadata or statistics (e.g. word co-occurrence) among documents in the set. We find this methodology much closer to a scholar's need than precision/recall.

The four papers in this panel are :

"Context Discovery with the Digital Library of Local Councils Journals (DLLCJ)", presented by Jieh Hsiang.

"The Border in Missionaries' Imagination: Through the Chinese Recorder Index Searching Engine (CRISE)", presented by Chijui Hu.

"A user-oriented HGIS platform of Taiwanese Land Deeds " by Chih-Yang Huang.

"Dan-Hsin Archives on DocuSky: an application to Hakka Studies", presented by I-Mei Hung.

Session Speakers Information

Speaker1: Prof. Jieh Hsiang

Topic: Context Discovery with the Digital Library of Local Councils Journals (DLLCJ)

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

Speaker2: Dr. Chijui Hu

Topic: The Border in Missionaries' Imagination: Through the Chinese Recorder Index Searching Engine (CRISE)

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

[1] Research Center for Digital Humanities, National Taiwan University, Taiwan.

[2] Research Center for Digital Humanities, National Taiwan University, Taiwan.

[3] Research Center for Digital Humanities, National Taiwan University, Taiwan.

[4] Research Center for Digital Humanities, National Taiwan University, Taiwan.

Speaker3: Mr. Chih-Yang Huang
Topic: A user-oriented HGIS platform of Taiwanese Land Deeds
Affiliation: Research Center for Digital Humanities, National Taiwan University, Taiwan.

Speaker4: Dr. I-Mei Hung
Topic: Dan-Hsin Archives on DocuSky: an application to Hakka Studies
Affiliation: Research Center for Digital Humanities, National Taiwan University, Taiwan.

Pre-modern Japanese Books as Data of Humanities: Finding Image of Edo Famous Place from Meisho-Ki 名所記 and Meisho-Zue 名所図会 using IIIF Curation Platform

Chikahiko Suzuki[1], Asanobu Kitamoto[2]

Introduction

Media such as publications have functions to recognize and enhance urban images, and functions to create images also. This paper attempts to retrieve information of Edo city from pre-modern Japanese books and try to examine the effects of publications on the formation of famous places and urban images in Edo period (17th-19th century Japan). For this purpose, we use IIIF Curation Platform (ICP) to deal with the information contained in the digitized pre-modern Japanese books. As an interim conclusion of our analyses, we focus on plants. We find the difference in tendency between famous places related to the cherry blossom 桜 and pine 松.

Materials and method: Curation "Illustration of Edo Famous Places"

A publication is collection of information. The information is organized in a certain context. We considered that it is possible to take information separated from context for analyze across multiple publications. In particular, non-text information such as illustrations can be important data comparable to text information.

Images of Edo famous places are provided by various publications such as novels and Ukiyo-E (Ando,2005 Yamamoto,2005). We focus on the Meisho-Ki 名所記 and Meisho-Zue 名所図会 that have been published throughout the Edo period. They contain information about famous places. For examples, *Edo-Meisho-Ki* 江戸名所記 and *Edo-Meisho-Zue* 江戸名所図会 have been used for studies about famous places (Okano et al,2002). *Edo-Meisho-Zue* is considered to be the Edo image shown by Edo residents (Suzuki,2001).

Many Meisho-Ki and Meisho-Zue are digitized by the National Diet Library and NIJL-NW project and provided on IIIF (International Image Interoperability Framework). We consider the illustrations on these books as data about image of Edo famous places. We pick up 1255 illustrations from 6 titles (21 volumes), *Edo-Meisho-Ki* (1662), *Edo-Meisho-Hyakunin-Isshu* 江戸名所百人一首(1663), *Ehon-Edo-Zakura* 絵本江戸桜(1795), *Ehon-Edo-No-Mizu* 絵本江都の見図(1795), *Edo-Meisho-Zue* (1834-

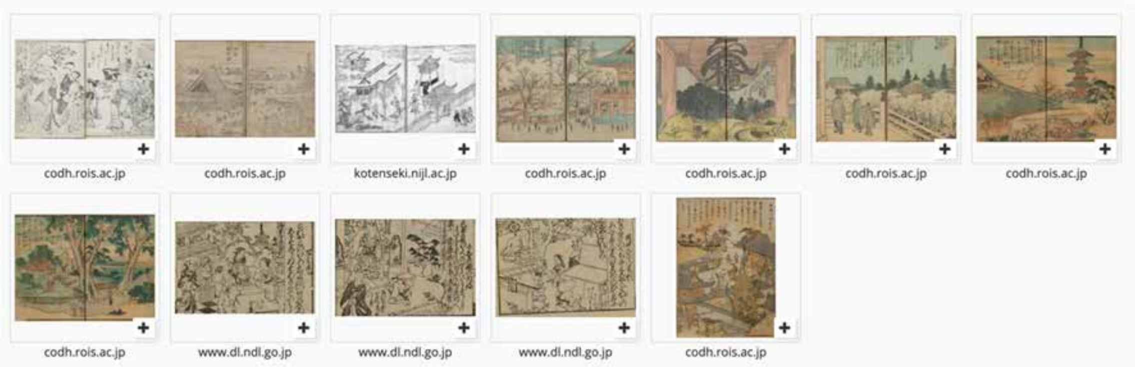
[1] ROIS-DS Center for Open Data in the Humanities, National Institution for Informatics

[2] ROIS-DS Center for Open Data in the Humanities, National Institution for Informatics

1836) and *Ehon-Edo-Miyage* 絵本江戸土産(1850-1867). In the Edo period, many Meisho-Ki and Meisho-Zue have been published. At this stage, we select two titles from each century.

We use the ICP for organizing illustrations as data. Then we add metadata such as place names and keywords. Place names were input as described in each document. We make name identification on "Rekishi Chimei Data" created by National Institutes for the Humanities, and input as a "controlled place name". At that time, the id on "Rekishi Chimei Data" was also input, so that information such as latitude and longitude was available. The results are published as the "Curation: Illustration of Edo famous places" (Fig.1).

Fig.1: Illustration of Edo famous places: Senso-Ji(an important temple in Asakusa)



from each book

Results: Difference between cherry blossoms and pine

We aggregated the metadata to understand the tendency. There are various famous places such as temples, bridges and playgrounds. In this time, we focused on the plants.

As a plant in the famous place, cherry blossoms have the largest appearance frequency with 43 cases. They are linked to specific places such as Kanei-Ji 寛永寺, Asuka-Yama 飛鳥山 and Senso-Ji 浅草寺. The frequency is high because there are multiple illustrations in some books. The number of pines is slightly reduced with 38 cases, but it is pictorialized as a ubiquitous planting in various places (Table.1).

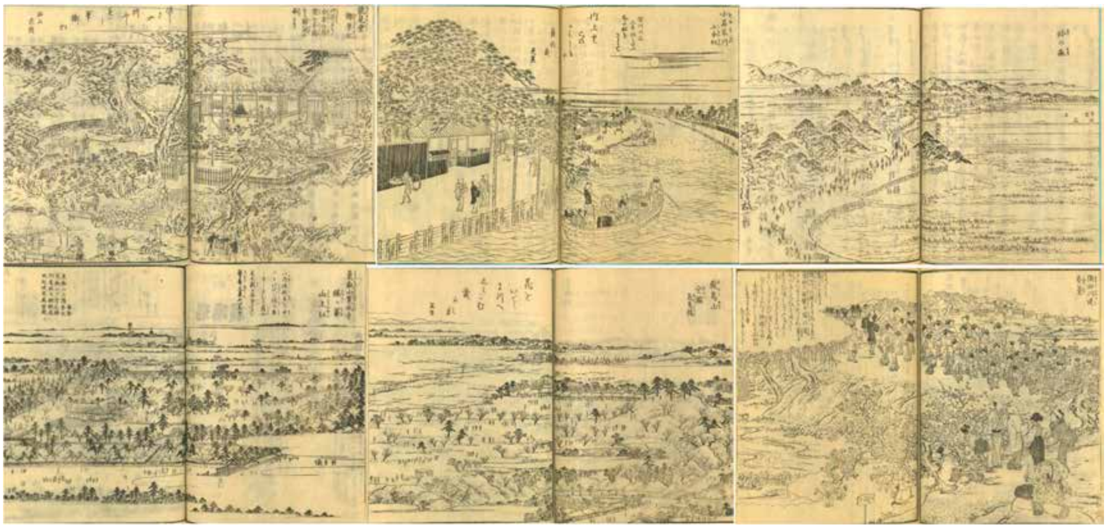
Table.1: Number of illustrations about plants and number of famous places related

Plant	Number of Illustrations	Number of Famous Plases
Cherry 桜	43	14
Pine 松	38	28
Plum 梅	13	9
Maple 紅葉	8	6
Wisteria 藤	5	2

(more than 5 appearances)

In the early Edo period, cherry blossoms and pines are listed as specific single tree such as Konno-Sakura 金王桜 and Azabu-Ippon-Matsu 麻布一本松. The later, the single cherry tree are replaced by the masses of trees. On the other hand, single pine remains as landmarks (Fig.2). Moreover, the number of single pine tree increases. The famous places have been extended gradually to the suburbs (Iwabuchi,2016). This trend can be seen strongly for pines.

Fig.2: Single pines: Fude-Sute-Matsu 擲筆松, Gohon-Matsu 五本松, Isonare-Matsu 磯馴松（upper）
Masses of cherry trees: Kanei-Ji, Asuka-Yama, Sumida-Gawa-River 隅田川（lower）
Edo-Meisho-Zue, from Dataset of Pre-Modern Japanese Text (10.20730/100249896)



Discussion: Image of famous places with publications

In the previous research on tourism and planting in Edo, the relation between the development of famous places by Shogunate and cherry blossoms is pointed out (Iioka,2006). The first example is the maintenance of Kanei-Ji by Shogun Tokugawa Iemitsu 徳川家光. After that Yoshimune 吉宗 maintained the outskirts of Edo, such as Sumida River and Asuka-Yama. As confirmed by our result, the cherry blossoms in the publications published after middle Edo period are masses of trees such as in Kanei-Ji, Sumida River and Asuka-Yama.

On the other hand, we can see various plants in previous research about landscapes. After the middle Edo period, plants in the garden of a farmer or temple in the suburbs became famous places. Also, outing to the suburbs were recreational (Hida,1991). Under these circumstances, famous places with pines were created and found. They were depicted such as Isonare-Matsu on the Suzuga-Mori 鈴ヶ森 coastline, and Fude-Sute-Matsu with Myokeno-Do 妙見堂 at Kanazawa-Hakkei 金沢八景.

This paper is in an early stage which confirmed the tendency and tried to connect it to the previous studies. Our purpose is to take information separated from context for analyze across multiple publications. It becomes possible to look back on the Edo period like a time machine by using abundant digitalized materials about Edo (Okuno, 2014). In next step, we will try to retrieve information of Edo city like the point of interest (POI) from publications not limited to famous places. We also should confirm the influence between publications. It is also necessary to enrich data curated from the publications.

References

Ando, Yuichiro. (2005). *Kankou Toshi Edo no Tanjo,* Shinchosha

Hanyu, Fuyuka. Okano, Shoichi. (2003). "A Study on the Characteristic of Traditional Sights of Edo, and on its Transition after the Meiji Restoration" *Journal of the Institute of Landscape Architecture*, 66 (5): pp.457-460

Iioka, Masaki. (2006). "Edo-Bakufu no Kankou Kourakuhci Kaihatsu to Keiei" *Bulletin of ARACS*, 9: pp.103-122

Iwabuchi, Reiji. (2016). "Haruka naru Edo" *Bibliology*, 9: pp.16-26

Hida, Norio. (1991). "Teien Shokusai no Rekishi 20, Edo-Jidai no Shokusai 3" *Nihon Bijutsu Kougei*, 635: pp.82-89

Okano, Shoichi. Toyoda, Akira. Hanyu, Fuyuka. (2002). "A Study of the Change in Edo City Suburban Sights as Seen in the Guidebook to Sights of the Edo Era" *Journal of the Institute of Landscape Architecture*, 65 (5): pp.797-800

Okuno, Takuji. (2014). *Edo "Media Hyoushou" Ron*. Iwanami Shoten

Suzuki, Akio. (2001) *Edo no Meisho to Toshi Bunka*. Yoshikawakoubunkan.

Yamamoto, Mitsumasa. (2005). *Edo Kenbutsu to Tokyo Kanko*, Rinsen Shoten

Curation: Illustration of Edo famous places http://www.ch-suzuki.com/icpedo/finder/ (View 2019-06-26)

IIIF Curation Platform http://codh.rois.ac.jp/icp/ (View 2019-06-26)

Rekishi Chimei Data https://www.nihu.jp/ja/publication/source_map (View 2019-06-26)

A study on narrative structure and change factors of emotions through analysis of literary emotions - Focusing on Korean Classic Novel *The Cloud Dream of the Nine(九雲夢)* –

Kang Woo-kyu[1], Kim Ba-ro[2]

Is emotion universal or cultural? From the evolutionary view, emotion is a product of evolution and universal. Eckmann's six basic emotions and Plutchik's Wheel of Emotions represent an evolutionary view of the universality of emotions. From an anthropological view, emotions are seen as historical and cultural products. But both viewpoints are all about the universality and locality of emotions, and they are not mutually exclusive. The universality and locality are the same as the relationship between Yin(陰) and Yang(陽). If so, what aspects of emotion are universal, and which aspects are cultural? This paper addresses this question.

Toward this end, this paper analyzes Korean and English versions of Korean classic novel *The Cloud Dream of the Nine*. We also compare the emotional narrative structures of the two versions using a both Korean-based and English-based digital emotion dictionary. Based on the comparison result, this research aims to examine the universality of emotions and the factors that change emotions according to culture.

The Cloud Dream of the Nine is a Korean representative classic novel, and there are numerous editions. The contents of *The Cloud Dream of the Nine* analyzed in this paper is the Korean version(*Jung, Byung-Sul's translation)* and the English version(*Gale's translation)*. Both versions are assumed to be derived from *eulsabon(乙巳本,* Chinese version*)*.

The Korean-based digital emotion dictionary, the HK + project of Chung-Ang University, was used by 594 Korean modern literary works with 49463 texts and their emotions using the deep-learning method. The English-based digital emotion dictionary is used by the NRC Word-Emotion Association Lexicon (EmoLex), which extracts the intensity of emotion words using collective intelligence using Amazon's Mechanical Turk.

[1] Chung-ang University

[2] Chung-ang University

This paper has been established the following analysis processes for comparing the emotional narrative structures of Korean and English versions of *The Cloud Dream of the Nine*.

1) Building and refining data of ① Korean version, ② English version, ③ automatic translation English version (automatic translation of Korean version through Google translator) and ④ automatic translation Korean version (automatic translation of English version through Google translator)
2) Through the Korean-based emotion dictionary, emotional analysis and cross-validation of text ① and ④
3) Through the English-based emotion dictionary, emotional analysis and cross-validation of text ② and ③
4) Analysis of emotional narrative structures between Korean version and English version based on results of 2) and 3) emotional analysis cross validation

Table 1 example of 3)

type	text	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
②	Your ladyship has already spared this head of mine that was doomed, and now you wish to serve me.	0	0	0	1	0	1	0	1	2	0
②	How shall I repay you?	1	1	0	0	1	0	0	1	0	1
②	My wish is to bind you to me by the endless contract of marriage.	1	1	0	1	2	1	0	2	1	2
③	How can a man save his life and grant his body so that he will repay his grace?	1	2	0	0	3	0	0	3	0	4
③	I will make it a hundred years with my son."	0	0	0	0	0	0	0	0	0	0

Through these processes, we try to understand the emotional narrative structures of Korean and English versions of *The Cloud Dream of the Nine* and compare cultural emotions in the translation of literary works. Also, we compare the analysis results of the Korean-based emotion dictionary and the English-based emotion dictionary to investigate the difference of the emotion analysis results according to the method of building the emotion dictionary. Based on this study, we will establish a basic methodology for comparing novel emotion. After that, we will expand the study to a large number of Korean and English novels.

Hoshi: A Japanese Morphological Adorner for TEI XML

Jerry Bonnell[1], Mitsunori Ogihara[2]

Morphological adornment of text in TEI-encoded XML can be useful for studies in textual analysis (John Unsworth and Martin Mueller, 2009) (C. M. Sperberg-McQueen, 2018). For English texts, the principal tool for facilitating such functionality is demonstrated by Northwestern University's MorphAdorner (Philip R. Burns, 2013). While an indispensable resource for DH scholarship, its handling of branching text, e.g. when <choice> appears, modifies the input presentation and, consequently, requires an additional preprocessing step to obtain the desired results. Motivated by this issue, this project purposes to answer the following questions: (1) how can branching text be handled with minimal corruption of the input TEI XML, (2) treating parsing software available online as a "black box," can an application be written to address these delicate issues using an output format similar to MorphAdorner, and (3) can a tool be developed to adorn text in other languages? If the answers are in the affirmative, such a tool can scale to efficiently and reliably provide adornment for large exhibits of texts in the target language. To enlighten us about the feasibility of such an application, we introduce a tool to adorn TEI-encoded XML for Japanese text. Unlike English, Japanese poses a unique challenge to parsing: words lack space delimiters and sentences do not require punctuation to terminate. Furthermore, software like MorphAdorner does not exist for the language to-date. Japanese parsers, however, abound and can be used to handle (and abstract) the parsing logic needed for supplying the adornment. Three are selected for this project: Kuromoji, MeCab, and Kagome (atilika) (Taku Kudou) (ikawaha). The standard IPADIC dictionary is used to build the model for each and, once constructed, a 9-piece annotation is returned for each token in the input. This contains useful morphological information such as levels of parts of speech, inflections, spellings, readings, etc. To encode this in TEI XML, we extend TEI with a conformant custom schema to include new attributes in the standard <w> tag. To present this information in the output document, each tag will begin on a newline and indentation will be increased for every open tag and decreased for every closing tag seen. Applying this design also allows for recovery of the input XML; a tool need only to strip all <w> tags, newlines, and tab characters.

To implement the processing work for this application, we require that the input document be complete and its sentences terminate with the "kuten" maru

[1] Department of Computer Science, University of Miami

[2] Department of Computer Science, University of Miami

(“。”). It is also assumed that the input is free of any overlapping markup. List data structures are then used to process each sentence, and parts within the sentence are characterized as tag (e.g., <w>, <saidwho>, <choice>) or non-tag (e.g., メロスは激怒した。) units. Tag units are printed directly to the output document without modification, and non-tag units are collected into a buffer and passed to the parser for adornment. The parser returns the tokens in the sentence with its corresponding annotations and, using a processing loop, these are aligned with the original text. While this strategy is effective for most cases, it is challenged by text that is segmented by some tag unit(s) and branching text, issues common with MorphAdorner. To overcome these, we propose that adornment be given only to the first character(s) of the segmented token and, following MorphAdorner, tags like <sic> be ignored when a branch is given. An additional “token” attribute containing the full token is introduced and appended to the custom attribute set if the token is segmented. This approach is taken in mind with the goal of preserving the input document integrity and allowing an input TEI XML to be submitted with minimal preprocessing, e.g. the addition of punctuation delimiters. If an input document lacks such punctuation and its length exceeds the capacity of what the parsers can receive at once, the tool may exhibit undefined behavior; an alternative parser that can scale to large input sizes would be needed to handle this use case. Should such software exist, its incorporation into the pipeline is straightforward thanks to the flexibility of our application.

While this tool is useful for outputting a TEI XML that can be queried by some morphological criteria, the applications of the methods outlined here extend further. Of these, our tool can be used to adorn any section of a TEI XML and to supply adornment for Japanese poetry. Moreover, the adorning techniques proposed here can be generalized to other languages, so long as a parser exists in the target language. This can be encouraging to DH scholars to make use of adorning techniques in their target language and, consequently, help to enhance TEI scholarship that prioritizes the use of morphological information.

References

atilika *Kuromoji: Japanese Morphological Analyzer*. Tokyo, Japan: atilika http://www.atilika.org/.

C. M. Sperberg-McQueen (2018). *Page, Text, and Sentence Markup for the ATMO Project: Interlocking TEI Customizations*. Technical report http://uyghur.ittc.ku.edu/2018/05/atmo-schemas-PTS.xml.

ikawaha *Kagome: Self-Contained Japanese Morphological Analyzer Written in Pure Go*. https://github.com/ikawaha/kagome.

John Unsworth and Martin Mueller (2009). *The MONK Project Final Report*. Technical report Indiana University, University of North Carolina at Chapel Hill, University of Virginia, and Northwestern University http://monkproject.org/MONKProjectFinalReport.pdf.

Philip R. Burns (2013). *MorphAdorner v2: A Java Library for the Morphological Adornment of English Language Texts*. Evanston, IL. Northwestern University. https://morphadorner.northwestern.edu/morphadorner/download/morphadorner.pdf.

Taku Kudou *MeCab: Yet Another Part-of-Speech and Morphological Analyzer*. http://taku910.github.io/mecab/.

The Bibliography is not Flat: Crowd-sourced Visualizations, Glocal Knowledge

Chong Eng Keat[1]

While global DH efforts have led to many fruitful outputs meaningful for the entire humanities community, local applications and use at the level of individual researchers remain limited to only a segment. Bridging the two requires that we rethink our approach to DH as an academic enterprise, and what new approaches need to be developed to bring about glocalized – globally-informed yet local – knowledge. At the same time, this paper proposes that DH scholars should consider and implement measures to make their work immediately impactful to other scholars who might be less engaged in the field.

Through the lens of the Humanistic Buddhism Bibliography Project (HBBP) initiated by Nan Tien Institute, Australia, and funded by the Hsing Yun Education Foundation, we explore how the compilation of English academic publications specifically related to Humanistic/Engaged/Applied Buddhism can simultaneously achieve the global, local and glocal aspects of DH – building a shared pool of records, identifying meaningful keywords, annotating them automatically, then building network visualizations of them to aid scholars in quickly building a bibliography suitable and specific to their research question.

The HBBP was initiated with the mandate to bring all these materials together in a single bibliographic record in order that 1) researchers can conveniently locate publications related to their topic of interest and 2) to present a full picture of the state of research in Humanistic Buddhism. After completing its first phase, the HBBP thus contributed its entries to the H-Buddhism Zotero Project, resulting in an addition of 200 new items. At present, the Humanistic Buddhism bibliography currently features 335 books (chapters) and 407 articles, a result of compiling from other bibliographies, library records and online database searches. A quick review of the list shows that topics covered are many – from ethics and law to economics and environmentalism.

Yet despite the small number of records, researchers will probably find it difficult to filter it down to a list which is directly relevant to their research interest. This is due to two possible scenarios: 1. they do not fully understand what is contained in the entire bibliography, small as this current one is at just over 700 entries; and, 2. the bibliography exist as a flat record which is not readily grasped. To resolve these, and as a first step, the HBBP compiled keywords from the keyword

[1] Nan Tien Institute, Australia

database of the "Digital Library and Museum of Buddhist Studies[2]," totaling 737 covering the broad categories of general introduction, sect, history, dharma-gate, practitioner, buddha and bodhisattva, sutra, appellation and others.

By identifying these keywords in the titles and abstracts of the HBBP entries, a bimodal network of titles and their keywords is then built, which is then presented in a browser-based application which can be easily accessed by users keen on using the HBBP itself.[3] This helps researchers quickly understand the overall state of research in a particular field, while still retaining the traditional functionality of the search function so that the bibliography can be navigated and filtered down, which we have demonstrated in other webpages with specific keywords selected (e.g., Humanistic Buddhism, Engaged Buddhism, Pure Land, Taixu, Chinese Buddhism, Bodhisattva, Master, Compassion, Spirit, Mind, Meditation).[4]

In many cases, researcher are not interested in only singular topics or keywords but what the publications they are linked to discuss. In this case, using a ego-network with value of two, we can show what other keywords are related to a specific one. Consider here that of the keyword "Meditation,"[5] which is found in the abstracts of 31 publications. To further filter down the list to a more manageable one, the researcher might find he is interested in "virtue" and thus selecting finds that the bibliography he is looking at lists 4 entries.[6] In doing so the researcher has quickly ascertained a select and pertinent set of publications he should consider reading.

While traditional boolean search allows for such functionality, the visualization of a set of defined keywords as occurs in the publication abstracts make the existing keywords visible and ready to be used for filtering of the HBBP into a usable list.

[2] See http://buddhism.lib.ntu.edu.tw/DLMBS/en/search/search_key.jsp

[3] See http://dlinkup.com/02_KeywordAbstractHBEBIntersection/

[4] See respectively http://dlinkup.com/03a_KeywordAbstractHumanisticBuddhism/; http://dlinkup.com/03b_KeywordAbstractEngagedBuddhism/; http://dlinkup.com/03c_KeywordAbstractPureLand/; http://dlinkup.com/03d_KeywordAbstractTaixu/; http://dlinkup.com/03e_KeywordAbstractChineseBuddhism/; http://dlinkup.com/03f_KeywordAbstractBodhisattva/; http://dlinkup.com/03g_KeywordAbstractMaster/; http://dlinkup.com/03h_KeywordAbstractCompassion/; http://dlinkup.com/03i_KeywordAbstractSpirit/; http://dlinkup.com/03j_KeywordAbstractMind/; http://dlinkup.com/03k_KeywordAbstractMeditation

[5] See http://dlinkup.com/04a_KeywordsAbstract_Meditation_Ego2/

[6] See http://dlinkup.com/05a_KeywordsAbstract_Meditation_Virtue/

A similar procedure could be applied to publications of other themes, whether in Buddhism, to perhaps “Zen” or “Japanese Buddhism,” so that together with existing networks, such visualizations may be made more comprehensive and interesting. If we were to extend the same to fields outside of Buddhism, it should work similarly, with the caveat that there exist similar keyword databases for that field.

The future holds much more possibilities. With the application of AI in the field of automated news summaries, we can perhaps look forward to similar technologies helping scholars with their literature reviews for their papers, and uncovering sources and materials which were previously unknown to us.

Establishment of the Ukiyo-e Similarity Database by a Game with a Purpose

Zhenao Wei[1], Yuntian Ma[2], Shizhe Wang[3], Nogc Cuong Nguyen[4], Pujana Paliyawan[5], Ruck Thawonmas[6], Tomohiro Harada[7], Keiko Suzuki[8], and Masaaki Kidachi[9]

This work presents a new game with a purpose (GWAP) to collect image similarity data. Few existing databases contain image similarity data. In particular, there exists no database of Japanese artwork ukiyo-e images containing such data. Inspired by existing studies on GWAP, we propose a solution that utilizes players' interests in games to help us obtain ukiyo-e similarity data.

Keywords - GWAP, ukiyo-e, database, image similarity

1. Introduction

We introduce a game with a purpose (GWAP) for acquiring similarity data of ukiyo-e images. Ukiyo-e is a genre of Japanese art which presents a distinctive hue and richness in color and has always been popular with the public. An ukiyo-e database operated by the Art Research Center (ARC)[1], Ritsumeikan University, one of the largest ukiyo-e databases, has 157,125 images. However, this ARC database does not provide a recommender system for users. As a result, users can hardly find data that suit their preferences in a short time.

Our long-term goal is to build a recommender system for the ARC’s ukiyo-e images. To do so, we need user preferences and information such as similarity data between images. Most reliable similarity data can be obtained by humans, but it costs a lot of human efforts and time. Therefore, to solves this issue, we propose a GWAP which can reach a large number of players and allow them to have fun while our system is obtaining said similarity data.

2. Previous Work

[1] Ritsumeikan University
[2] Ritsumeikan University
[3] Ritsumeikan University
[4] Ritsumeikan University
[5] Ritsumeikan University
[6] Ritsumeikan University
[7] Ritsumeikan University
[8] Ritsumeikan University
[9] Ritsumeikan University

2.1 Recommender systems

Recommender systems are roughly divided into two categories: collaborative filtering (CF) [1] and content-based filtering (CBF) [2]. A CF-based system recommends items to a user, based on historical data of users that have similar usage behaviors with the current user. A CBF-based system recommends items to a user, based on the predefined features of items and the usage behaviors of the current user. CF-based and CBF-based systems are often combined into hybrid recommender systems.

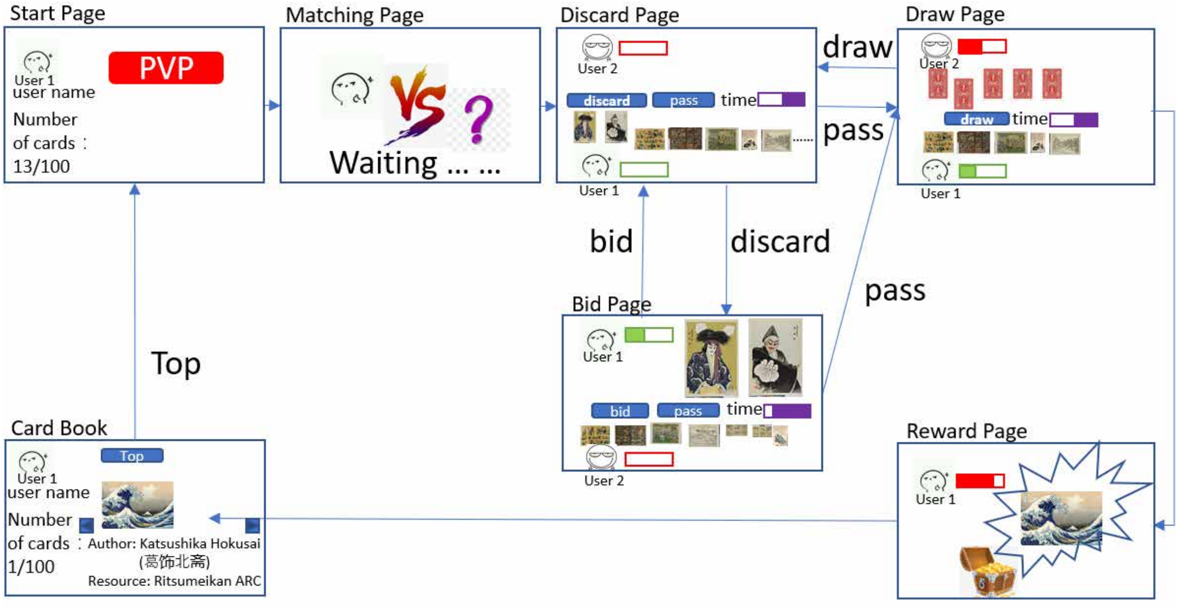


Figure 1. Flow chart of the game

Each type of recommender systems has strengths and weaknesses. CF-based systems require a large amount of user decision data to make accurate recommendations, and thus in the beginning face with a so-called "cold start problem [3]". Although CBF-based systems need less amount of data to start, only items similar to the ones that the current user has watched are recommended. This is a reason why most recommender systems nowadays use the hybrid approach [4].

2.2. GWAP

The concept of GWAP [5] is based on the idea of using games to enhance the enthusiasm of players to solve tedious and complex problems. Many web games use this kind of concept to obtain image labels, through keywords about those images that the players type [6]. However, as Steinmayr et al. [7] stated, existing GWAP systems do not produce comprehensive image descriptions.

Our solution is to design a card game similar to Old Maid to obtain human subjective feelings about image similarity and to ensure the quality of answers through mutual supervision of players.

3. System Design

Our game is divided into two major parts: 1) main gameplay which is a card-matching game and 2) a gallery (Card Book) for card collection. Players obtain cards for their gallery as a reward from playing the card-matching game. Progress on card collection is based on their scores in the game. The purpose of having this gallery is to promote players' motivation. The whole game system is shown in Fig. 1.

3.1 Rules of the system

The current game version has only a two-player mode. The game randomly pairs players and starts gameplay by selecting seven pairs of similar cards—determined by a typically used similarity metric explained in 3.3—from the database and randomly assigning seven of the cards to each player. The players can only see their own cards.

In each turn, a player can discard a pair of hands that the player thinks are most similar, and the system will give a "score" based on the aforementioned similarity metric. On the other hand, when the player thinks there is no similar pair at all, the player can pass the turn. In addition, at the beginning of each turn, except for the first turn, the player must draw one of the opponent's cards.

The game is over when all cards are discarded, or game play time runs out.

3.2 Prevent cheating

A player can choose to "bid" each time the opponent discards if she or he can find better pairs for the discarded cards. The two discarded cards will go to the bidder's hand with the cost of scores equal to that the discarder gained. This mechanism allows players to get more points from capturing the opponent's mistakes or cheating by randomly discarding cards. In order to prevent players from hoarding cards, we set the hand limit for each player to nine cards.

3.3 Unpolished similarity metric

During the "cold start" period, i.e., when the amount of collected image similarity data is still low, we use a similarity metric called cosine similarity, presented in our previous work [8], to calculate image similarity. This is considered as an "unpolished similarity metric," which will be improved later by using similarity data measured by humans obtained from this game.

3.4 Reward

After each game, a player has a chance to get a reward ukiyo-e card to their gallery. The higher the score, the greater the chance. Initially, all the cards in the gallery are gray. When a card is obtained, it will be colorized.

4. Conclusions and Future Work

We proposed an interesting game to collect subjective feelings from its players to establish an ukiyo-e similarity database.

We intend to improve the game, for example, by developing multiple modes, such as a single-player mode and a multiplayer mode.

References

[1] P. Resnick and H. R. Varian. Recommender systems. Comm. ACM. 1997. pp. 56–58.

[2] R.J. Mooney and L. Roy. Content-Based book recommending using learning for text categorization. ACM SIGIR’99. 1999. pp. 195−204.

[3] M. Elahi, F. Ricci and N. Rubens. Active learning in collaborative filtering recommender systems. Springer. 2014. pp. 113–124.

[4] R. Hoekstra. The knowledge reengineering bottleneck. Semant Web. 2010. pp. 111–115.

[5] L. von Ahn and L. Dabbish. Designing games with a purpose. Comm. ACM. 2008. pp. 58–67.

[6] L. von Ahn and L. Dabbish. Labeling images with a computer game. CHI04. 2004. pp. 319–326.

[7] B. Steinmayr, C. Wieser, F. Kneißl and F. Bry. Karido: A GWAP for telling artworks apart. CGAMES. 2011. pp. 193-200.

[8] Z. Wei, L. Xiong, K. Mori, T. D. Nguyen, T. Harada, R. Thawonmas, K. Suzuki and M. Kidachi. Deep Features for Image Classification and Image Similarity Perception. JADH. 2017. pp. 60-62.

Image-Based Content Indexing for Books with Iconographic Elements - the Case of Bukan Complete Collection

Hakim Invernizzi[1], Asanobu Kitamoto[2], Frédéric Kaplan[3]

1. Introduction

This work proposes a machine learning based pipeline with the goal of indexing the content of books through their iconographic content. It is argued that an image-based approach to content indexing can be relevant when there is an abundance of iconographic elements and when text-based solutions do not yield optimal results or are more challenging to implement.

At least one similar pipeline has been proposed, with a focus on the enrichment of the metadata of book and newspapers illustrations through deep learning techniques (Moreux and Chiron, 2018). However, the idea of using iconography as a proxy for textual content is, to our knowledge, innovative. It must be noted that such an approach is effective only if the iconography in the book is highly representative of its content. Examples of promising candidates include ancient illustrated books such as illuminated manuscripts or incunabula, as well as books containing heraldic achievements or flags.

The goal is to design a general pipeline that can be applied to any collection of books with iconographic elements. As a first step, the pipeline will be tested on the bukan collection (CODH, 2017). The collection currently consists of 381 bukan books, owned by the National Institute of Japanese Literature (NIJL). A bukan is a woodblock-printed book from the Edo period, which contains information of administrative nature about people and institutions of the time. Its purpose could be compared to today’s college yearbooks: providing distinctive information about a person or a group at a particular time.

The “Bukan Complete Collection” (CODH, 2018) is the digital database of bukan created and curated by ROIS-DS Center for Open Data in the Humanities (CODH). The project aims at creating a complete time-series database of the people, daimyō and bakufu officers described in the bukan collection (Kitamoto et al., 2018). However, the indexing of the collection brings some challenges. First, its large size demands for an automated approach. Moreover, text-based indexing is difficult to

[1] École polytechnique fédérale de Lausanne (EPFL), National Institute of Informatics

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

[3] École polytechnique fédérale de Lausanne (EPFL)

implement because bukan are written in pre-modern Japanese, a language for which OCR research is still undergoing (Clanuwat et al., 2018). Therefore, automated solutions for image-based indexing are the focus of this work.

2. Proposed pipeline

Image-based content indexing is the mapping of index terms to their occurrences in the collection by using image features obtained from iconographic material. The pipeline to extract and classify the features of iconographic elements is composed of the three following steps.

1. segmentation of the iconographic elements
2. classification of the segmented elements into distinct iconographic classes
3. identification of index terms based on iconographic classes

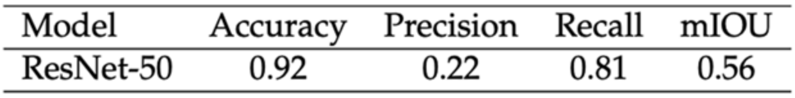
In the following section, the three-steps pipeline is described in detail.

3. Methods and results

3.1 Segmentation

The goal of the segmentation step is to extract all iconographic elements in the input image. It is performed using dhSegment (Oliveira et al., 2018), a document segmentation framework developed by EPFL's digital humanities lab (DHLAB). dhSegment leverages a state-of-the art convolutional neural networks (CNN), Resnet-50 (He et al., 2016), pretrained on the ImageNet dataset (Deng et al., 2009) in order to carry out the segmentation of the document.

The first experiment focuses on segmenting daimyō family emblems. As Figure 1 shows, the segmentation model takes a bukan page as input and produces a mask of family emblems as output. In terms of input for the segmentation model, 161 pages from 4 bukan constitute the training data. The resulting model has been evaluated on 70 pages from 3 other bukan. Table 1 summarizes performance of the model. The model achieves high recall, which means that it is able to detect the large majority of pixels belonging to the emblems.



Model	Accuracy	Precision	Recall	mIOU
ResNet-50	0.92	0.22	0.81	0.56

Table 1: results of the segmentation step

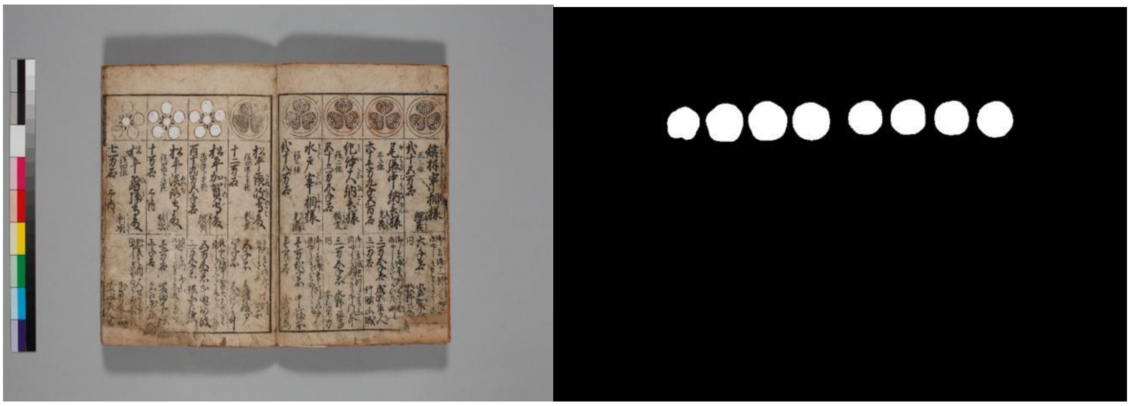


Figure 1: Input (left) and output (right) of the segmentation step

3.2 Classification

The goal of the classification step is to label the segmented elements into distinct iconographic classes. The definition of these classes is performed by the user and depends on the kind of iconographic category being processed. It is argued that the classification of iconographic material from books is a few-shot problem, since often few samples per class are available. In light of this remark, a few-shot technique called “prototypical networks” (Snell et al., 2017) is used for the classification. This model can be trained on a set of classes and tested on another, reducing the need for a large number of samples per class and allowing for a transfer learning approach.

In the case study, the definition of the emblem classes posed a challenge. A naive approach would be to create a class for each family emblem. However, the analysis of studies on Japanese heraldry (Phillips et al., 2018; Dower, 1990) highlights why this methodology is unsuitable. This is due to the existence of countless emblems, many of which are very similar visually. Instead, each class will represent a heraldic charge. Charges are clearly defined visual elements suitable both for image classification and for family identification, since daimyo family tended to use a limited set of charges. Examples of heraldic charges can be seen in Figure 2. The results of the classification step using prototypical networks, trained on 1077 emblems from 23 classes, and tested on 1011 emblems from 55 classes (the 23 seen classes plus 32 new classes) are presented in Table 2. The model achieves solid performance.

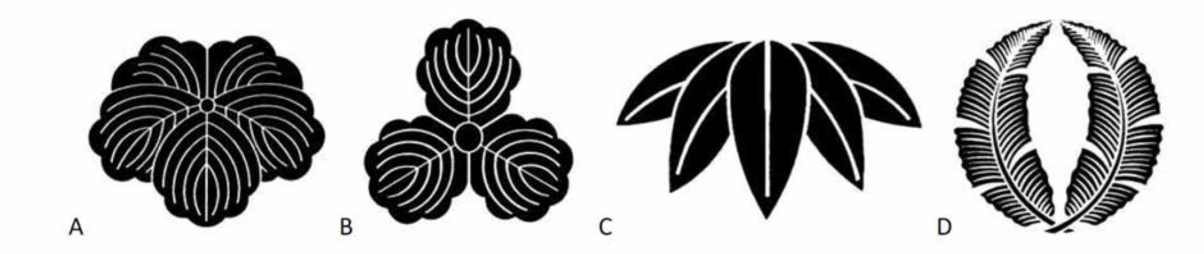
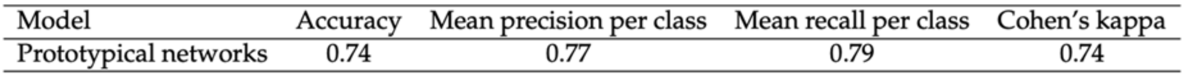


Figure 2: Heraldic charges. Left to right: ivy, oak, bamboo, plantain. (Dower, 1990)



Model	Accuracy	Mean precision per class	Mean recall per class	Cohen's kappa
Prototypical networks	0.74	0.77	0.79	0.74

Table 2: results of the classification step

3.3 Identification

The goal of the identification step is to define which index terms are present on a page of the collection, based on which iconographic classes are represented. To do so, a probabilistic approach to content indexing will be introduced, where the presence of an iconographic class estimates the probability of a given index term being present.

In the case study, a probability will be associated to each possible combination of heraldic charge and family name.

4. Conclusion

The first experiment produces promising results with respect to the first two steps of the pipeline. The identification step is currently being tested. Future plans include testing the pipeline on other iconographic categories present in the bukan collection.

Acknowledgments

The authors thank Prof. Kazuaki Yamamoto and Prof. Kumiko Fujizane of NIJL for their help in the analysis and understanding of the bukan collection.

References

Moreux, J.-P. and Chiron, G. (2018). Hybrid Image Retrieval in Digital Libraries: A Large Scale Multicollection Experimentation of Deep Learning Techniques. In Méndez, E., Crestani, F., Ribeiro, C., David, G. and Lopes, J. C. (eds), *Digital Libraries for Open Knowledge*, vol. 11057. Cham: Springer International Publishing, pp. 354–58 doi:10.1007/978-3-030-00066-0_39.

Center for Open Data in the Humanities (2017). Dataset of Pre-modern Japanese Text http://codh.rois.ac.jp/pmjt/book/?武鑑 (accessed 21 June 2019).

Center for Open Data in the Humanities (2018). Bukan Complete Collection http://codh.rois.ac.jp/bukan/ (accessed 21 June 2019).

Kitamoto, A., Horii, H., Horii, M., Suzuki, C., Yamamoto, K. and Fujizane, K. (2018). Differential Reading by Image-based Change Detection and Prospect for Human-Machine Collaboration for Differential Transcription. In Palau, J. G. and Russell, I. G. (eds), *Digital Humanities 2018, DH 2018, Book of Abstracts, El Colegio de México, UNAM, and RedHD, Mexico City, Mexico, June 26-29, 2018*. Red de Humanidades Digitales A. C., pp. 414–415.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K. and Ha, D.

(2018). Deep Learning for Classical Japanese Literature.

Oliveira, S. A., Seguin, B. and Kaplan, F. (2018). dhSegment: A generic deep-learning approach for document segmentation. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*: 7–12 doi:10.1109/ICFHR-2018.2018.00011.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–78 .

Deng, J., Dong, W., Socher, R., Li, L., Kai Li and Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 248–55 doi:10.1109/CVPR.2009.5206848.

Snell, J., Swersky, K. and Zemel, R. (2017). Prototypical Networks for Few-shot Learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R. (eds), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pp. 4077–4087.

Phillips, D. F., Valerio, E. and Kariyasu, N. (2018). *Japanese Heraldry and Heraldic Flags*. 1st edition. Danvers, Massachusetts: Flag Heritage Foundation.

Dower, J. (1990). *Elements Of Japanese Design: Handbook Of Family Crests, Heraldry & Symbolism*. Revised ed. edition. Boston: Weatherhill.

A Quantitative Approach to a New Digital Platform of Ancient and Medieval Japanese Sutras

Aiko Aida[1]

Our research projects, in collaboration with the Art Research Center at Ritsumeikan University (ARC) and the National Museum of Japanese History, have been compiling a digital database of old Japanese Buddhist manuscripts. The database is modeled on the framework of the ARC digital resource database, ARC ArtWiki. It was designed to provide basic information (including title, period, medium, dimensions, and owner), infrared digital photographs of manuscripts, and explanations of the Buddhist painting motifs. We feel that the database contributes to the promotion of interdisciplinary studies and analyses of metadata. There are exceptionally few venues for researchers to directly examine these black-on-white or gold-on-blue manuscripts from the Nara, Heian, and Kamakura periods (eighth to the fourteenth century), even though there are many web sites dedicated to Buddhist manuscripts. The SAT Daizōkyō Text Database (http://21dzk.l.u-tokyo.ac.jp/SAT/) and the Collection of Old Buddhist Manuscripts in Japanese (https://koshakyo-database.icabs.ac.jp) are already quite renowned.

One of the most significant features of our database is the infrared digital photography. The digital images of these sutras reveal elements such as characters, inscriptions, pictures, and signatures previously hidden beneath the navy blue dye of some of the manuscripts. Until now, these kind of web site was nothing. Another significant feature is that the entire database provides explanations of the painting motifs found on these manuscripts, with special attention given to over 60 motifs found on the Lotus Sutra (Sanskrit, Saddharma Puṇḍarīka Sūtra). As a result, information about and images of Buddhist narrative paintings will be readily available.

Furthermore, this study explores the relationship between period or age and the quality of the paper of these ancient and Medieval scriptures. A quantitative survey is the principal component of the study. The survey assigns numerical values to a manuscript based on its “period” and “paper quality.” The latter refers mainly to the size of various elements of the paper in order, which aids in gaining deeper insight into the variety of old Japanese Buddhist scriptures copied on indigo, purple, white, and other decorative papers. We modeled the relationship between the size of the manuscripts’ various elements, the existence/non-existence of hidden signatures, and their ages using principal component analysis. Our results showed that the score of the principal elements distributed more in the thirteenth and fourteenth centuries’

[1] Japan Society for the Promotion of Science

grope, and the existence of hidden signatures tended to correlate with the size or age of the paper.

In conclusion, the group of manuscripts from the thirteenth and fourteenth centuries tended to have the widest spread, and the ancient group also indicated a wide score. Nonetheless, the data suggest that the relationship between the size of the paper and the size of the ruled lines is justified in all ages. This study suggests that the ratio is more significant than any individual rule, such as is found in volume twelve of Engishiki (延喜式), written in 927. This means that our database contributes to the analysis of metadata concerning the old Japanese Buddhist manuscripts.

A Preliminary Consideration on Designing Domain Ontologies on Local Foods and Their Real Stories

Ikue Kawamura[1], Shun Shiramatsu[2]

1. Background and Purpose

Stories on local foods handed down is important to understand the characteristics of the regional cuisine culture. Local foods are made from foodstuff with local characteristics and have been handed down in each area. Therefore, if there is no one to convey it, details of the cooking method may not remain. In addition, even similar local foods often have subtle differences and contexts that only local people can understand. Therefore, in this research, we have developed a web map of local foods to organize and visualize such local food information [1]. This map displays the location and attributes of mocha (Popular local food in Japan) on a Japanese map, and is a system to know regional characteristics from local food understanding.

In this study, we aim to handle more detailed information by adding an ontology to the web map[3] we have created before. Therefore, we think that important information is included in the stories of people in the area where each local food travels, and we study the local food ontology for storing detailed information, including subtle differences in local food. In addition, in considering this ontology, we are considering ways of thinking that the difference between the past and the present is important for preservation and utilization of food history and culture.

2. Proposed Method

We designed a domain ontology for recording stories on local foods. Our ontology is represented by the resource description framework (RDF) standardized by W3C and IMI common vocabulary standardized by IPA, a governmental agency in Japan [2].

In this research, we especially focus on “Mochi”, which is the most popular local food in Japan, when designing the local food ontology. We firstly try to construct an instance dataset about Mochi stories, which is we actually heard, to consider what kind of classes and properties are needed. After that, we design two ontologies, one is about local food and another is about local food stories, according to the instance dataset. There are many foods named "Mochi" that are made with similar materials

[1] Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology University

[2] Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology University

[3] http://phirip.com/t/mochimap/e/index.html

and methods, and names and materials may differ slightly depending on the region.

3. Result

We focused on the stories about difference between similar Mochis that have different names when designing the ontology. Such viewpoint is important to understand regionality from local foods. As a result, we found that we need to prepare original properties such as material and how to eat.

4. Conclusion

We designed a domain ontology for local food stories by referencing actual stories about Mochi. We found that we need to prepare original properties such as material and how to eat through considering difference between similar two local foods. We currently reconsider the design of the ontology for local food stories in order to represent historical details. As a future work, we are planning to apply our ontologies to build a Mochi dataset of various actual cases and apply the dataset to e-learning system for local food culture with visualization mechanisms.

References

[1] Ayao Okumura: Mochi, Zouni, Kiki-sho - Home cooking in hometown, vol. 5, 2002. (in Japanese)

[2] IMI common vocabulary project: Practice! Hackathon and Workshop on IMI common vocabulary, https://peatix.com/event/298493?lang=ja, 2018. (in Japanese)

An Analysis of the Differences Between Classical and Contemporary Poetic Vocabulary of the Kokinshū

Hilofumi Yamamoto[1], Bor Hodošček[2]

1. Introduction

The purpose of the project is to clarify the relationship between literal elements and non-literal elements of an ancient language. We will examine whether the translations of the Kokinshū use the same words as in a poem or whether they use words not corresponding to words in a poem. To specify elements written only in the translations, we subtract the elements of original poems (OP: the Kokinshū) from the elements in their contemporary translations (CT), and analyze the residual elements. The differences, therefore, may include two kinds of elements: 1) elements resulting from chronological differences in language; 2) elements added for interpretation.

2. Methods

We will use the corpus of the Kokinshū by Nakamura et al. (1999). Poems are separated into tokens using the classical poem tokenizer, kh (Yamamoto 2007). We convert the tokens into meta-codes, then using the meta-codes, subtract the elements of the original from the elements of their translations. We examine the length of the portion of meta-codes between the two elements (Figure 2). As an algorithm for matching the elements of CT and OP, we use Longest Common Subsequence (Traum and Habash 2000). An example of subtraction processing with code2match.c (Yamamoto 2005) is shown in Figure 3.

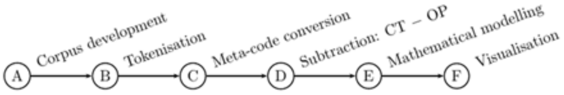
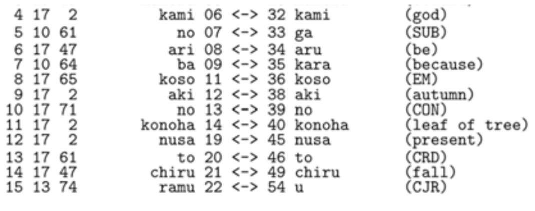


Figure 1: Flowchart of data processing



```
 4 17  2          kami 06 <-> 32 kami          (god)
 5 10 61            no 07 <-> 33 ga            (SUB)
 6 17 47           ari 08 <-> 34 aru           (be)
 7 10 64            ba 09 <-> 35 kara          (because)
 8 17 65          koso 11 <-> 36 koso          (EM)
 9 17  2           aki 12 <-> 38 aki           (autumn)
10 17 71            no 13 <-> 39 no            (CON)
11 17  2        konoha 14 <-> 40 konoha        (leaf of tree)
12 17  2          nusa 19 <-> 45 nusa          (present)
13 17 61            to 20 <-> 46 to            (CRD)
14 17 47         chiru 21 <-> 49 chiru         (fall)
15 13 74          ramu 22 <-> 54 u             (CJR)
```

Figure 3: An example of the alignment of the matched elements between OP$_{(298)}$ and CT$_{(298,\ \mathrm{koma})}$. Each line consists of the matched pair ID number (1), the matching level indicated by the value (17), ID number of POS (11) which indicates a place name, OP element (***tatsutahime**), ID number of OP element, ID number of CT element, CT element (***Tatsutahime**), and the glossary; * written in different kanji.

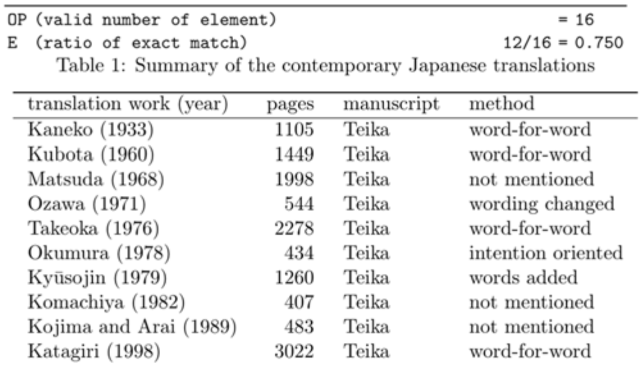
[1] Tokyo Institute of Technology

[2] Osaka University

3. Results

Table 2 shows a calculation of the components of OP(298) as an example. 12 elements out of 16 (75 percent) are matched in CT(298, koma). If we assume that matched elements at all the three levels are expressed in CT(298, koma), then 15 elements (94 percent) of OP(298) are expressed as the elements in CT(298, koma). The remaining 6 percent of elements of OP(298) do not match against any elements in CT(298, koma). None of the ten translations could be fully expressed with the ancient language. The amount of added information was 80 percent higher than the original (Table 4).

Table 2: Result of subtracting the elements of OP(298) from those of CT(298, koma): it indicates the ratio of the ingredients of OP(298).

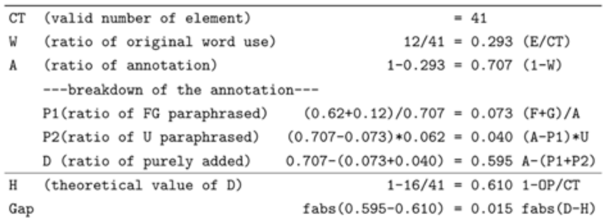


OP	(valid number of element)	= 16
E	(ratio of exact match)	12/16 = 0.750

Table 1: Summary of the contemporary Japanese translations

translation work (year)	pages	manuscript	method
Kaneko (1933)	1105	Teika	word-for-word
Kubota (1960)	1449	Teika	word-for-word
Matsuda (1968)	1998	Teika	not mentioned
Ozawa (1971)	544	Teika	wording changed
Takeoka (1976)	2278	Teika	word-for-word
Okumura (1978)	434	Teika	intention oriented
Kyūsojin (1979)	1260	Teika	words added
Komachiya (1982)	407	Teika	not mentioned
Kojima and Arai (1989)	483	Teika	not mentioned
Katagiri (1998)	3022	Teika	word-for-word

Table 3: Component of CT in case of KKS *298* by Komachiya (1982): fabs(D-H) stands for the function of the absolute value of the practical value, D, minus the theoretical value, H.

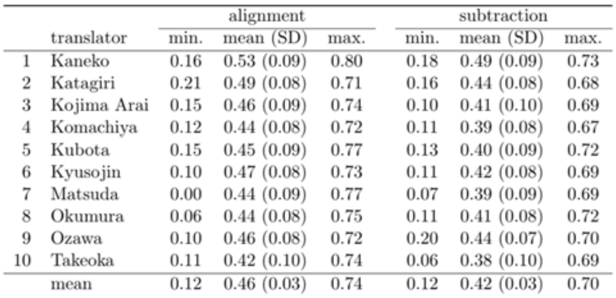


CT	(valid number of element)	= 41	
W	(ratio of original word use)	12/41 = 0.293	(E/CT)
A	(ratio of annotation)	1-0.293 = 0.707	(1-W)
	---breakdown of the annotation---		
	P1(ratio of FG paraphrased)	(0.62+0.12)/0.707 = 0.073	(F+G)/A
	P2(ratio of U paraphrased)	(0.707-0.073)*0.062 = 0.040	(A-P1)*U
	D (ratio of purely added)	0.707-(0.073+0.040) = 0.595	A-(P1+P2)
H	(theoretical value of D)	1-16/41 = 0.610	1-OP/CT
Gap		fabs(0.595-0.610) = 0.015	fabs(D-H)

4. Discussion

Based on the differences between the two, we assume that translators attempted to express some cultural elements unfamiliar to modern people. CT(298, koma) uses the same 12 elements as in OP(298) (Table 3). The total number of elements of CT(298, koma) is 41; thus 29 percent of CT(298, koma) is calculated as the component of OP(298). The rest of CT(298, koma), Ratio A, 71 percent, is considered as added annotated text and it should be deconstructed into three kinds of components: 1) P1 estimated from the field match F and the group match G; 2) P2 estimated from the unmatched elements which are assumed to be somehow translated into CT; and 3) the purely added component, D estimated from the ratio of the annotation minus P1 and P2 (Figure 4). If the estimation is correct, the practical value D can be close to the theoretical value H, and the validity of the operation will be supported.

Table 4: Amount of added information (N=1000)



		alignment			subtraction		
	translator	min.	mean (SD)	max.	min.	mean (SD)	max.
1	Kaneko	0.16	0.53 (0.09)	0.80	0.18	0.49 (0.09)	0.73
2	Katagiri	0.21	0.49 (0.08)	0.71	0.16	0.44 (0.08)	0.68
3	Kojima Arai	0.15	0.46 (0.09)	0.74	0.10	0.41 (0.10)	0.69
4	Komachiya	0.12	0.44 (0.08)	0.72	0.11	0.39 (0.08)	0.67
5	Kubota	0.15	0.45 (0.09)	0.77	0.13	0.40 (0.09)	0.72
6	Kyusojin	0.10	0.47 (0.08)	0.73	0.11	0.42 (0.08)	0.69
7	Matsuda	0.00	0.44 (0.09)	0.77	0.07	0.39 (0.09)	0.69
8	Okumura	0.06	0.44 (0.08)	0.75	0.11	0.41 (0.08)	0.72
9	Ozawa	0.10	0.46 (0.08)	0.72	0.20	0.44 (0.07)	0.70
10	Takeoka	0.11	0.42 (0.10)	0.74	0.06	0.38 (0.10)	0.69
	mean	0.12	0.46 (0.03)	0.74	0.12	0.42 (0.03)	0.70

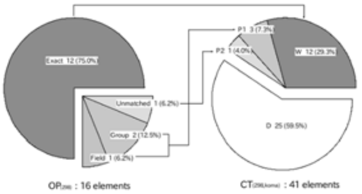


Figure 4: Pie-charts illustrating the components of OP(298) and CT(298, koma): the ratio of purely added components is estimated based on the number of elements in common in OP and CT.

5. Conclusion

The current paper discussed the differences between the original poems of the Kokinshū and their translations. We attempted to classify the components of both OP and CT to examine whether or not CT includes added elements, which are the non-literal elements of OP. After subtracting the matched elements between OP and CT from CT, the presence of a residual indicated that CT includes newly added elements. It shows that it is impossible to convert the contents in the ancient language into only their equivalents in the modern language.

References

Hasumi, Yoko (1991) “Dōitsu jōhō ni motozuku bunshōhyōgen ni tsuite no bunseki / Difference of expressions on the same information”, *Mathematical Linguistics*, Vol. 18, No. 3, pp. 136–144.

Kaneko, Motoomi (1933) Kokinwakashū Hyōshaku: Shōwa Shimban, Tokyo: Meijishoin.

Katagiri, Yoichi (1998) Kokinwakashū Hyōshaku Jō, Chū, Ge, Tokyo: Kodansha.

Kojima, Noriyuki and Eizō Arai (1989) Kokinwakashū, Vol. 5 of Shin-Nihon bungaku taikei (A new collection of Japanese literature), Tokyo: Iwanami shoten.

Komachiya, Teruhiko (1982) Gendaigo yaku taishō Kokinwakashū (Kokinwakashū with modern Japanese translations), Obunsha Bunko Taiyaku Koten Series, Tokyo: Ōbunsha.

Kubota, Utsubo (1960) Kokinwakashū Hyōshaku (Vol. 1, 2, 3), Tokyo: Tokyodo shuppan.

Kyūsojin, Hitaku (1979) Kokinwakashū Zen'yaku Chū (Comprehensive annotations of the Kokinwakashū), Vol. 1–5 of Kodansha Gakujutsu Bunko: Kodansha.

Matsuda, Takeo (1968) Shinshaku Kokinwakashū Vols.1 and 2, Tokyo: Kazama Shobo.

Miyazima, Tatuo (1979) "Kyōsantō Sengen no yakugo (Translated terms in the Communist Manifesto)", in Gengogaku Kenkyūkai ed. *Gengo no Kenkyū* (Study of language), Tokyo: Mugi Shobo, pp. 425–517.

(1980) "Jodōshi' to 'Hojodōshi' (Auxiliary verbs and subsidiary verbs)", in **Society of Modern Language ed**. *Kindaigo kenkyū* (Study of contemporary vocabulary), Vol. 6, Tokyo: Musashino shoin, pp. 455–468.

Nakamura, Yasuo, Yoshihiko Tachikawa, and Mayuko Sugita (1999) Kokubungaku kenkyū shiryōkan dētabēsu koten korekushon "*Nijūichidaishū*" Shōhobanbon CD-ROM (Database Collection by National Institute of Japanese Literature "*Nijūichidaishū*" the Shōho edition CD-ROM): Iwanami Shoten.

Okumura, Tsuneya (1978) Kokinwakashū, Shinchō Nihon Koten Shūsei, Tokyo: Shinchō sha.

Ozawa, Masao (1971) Kokinwakashū, Vol. 7 of Nihon Koten Bungaku Zenshū, Tokyo: Shōgakkan.

Suzuki, Tai (1988) "Weirando Shūshinron no Kanji (Kanji in the "Elements of moral science" by Francis Wayland)", in *Gengo no Kenkyū* (Study of language), Vol. 8 of Kindai Nihongo to Kanji (Contemporary Japanese and Kanji), Tokyo: Meijishoin, pp. 128–164.

Takeoka, Masao (1976) Kokinwakashū Zen Hyōshaku Jō Ge (the complete annotated edition of Kokinwakashū, Vols. 1 and 2), Tokyo: Yubun Shoin.

Traum, David and Nizar Habash (2000) "Generation from lexical conceptual structures", in *NAACL-ANLP 2000 Workshop on Applied interlinguas*, pp. 52–59, Morristown, NJ, USA: Association for Computational Linguistics.

Hilofumi Yamamoto (2005) "A Mathematical Analysis of the Connotations of Classical Japanese Poetic Vocabulary", Ph.D. dissertation, Australian National University.

Hilofumi Yamamoto (2007) "Waka no tame no Hinshi tagu zuke shisutemu / POS tagger for Classical Japanese Poems", *Nihongo no Kenkyu / Studies in the Japanese Language*, Vol. 3, No. 3, pp. 33–39.

An innovative dynamic visual model for digital archive in traditional cultural Heritage – Evidence from Stone fish weirs Culture in Penghu, Taiwan

Ju-Chuan Wu[1], Jui-Chi Wang[2]

Development with technology and multimedia, digital archives are made to expand historical research and recreate new value to many valuable contents. Unlike traditional displaying way that historical context is limited paper margin, using dynamic visual model can reveal traditional outcomes and innovative presentation. Hence, most researchers adopt digital archive visual methods to collect, display, analyze, and recreate value-added historical data. Led to focus on preserving the local traditional culture of the country and humanities scholars using digital collections. The knowledge level of human sharing and multiple data value-added applications to promote the local life and culture. To let these precious cultural productions can be sustainable maintained through the form of digital collections, for example, the preservation and presentation of the early stone weir culture which only exists in Penghu nowadays.

This study has observed the relevant literature and historical data which related to the stone weirs preservation status and methods around the world in recent years. There are still several improvements, such as (1) the relevant literatures of the stone weirs haven't been digitalized completely; (2) the digitalized data is too fragmented and incomplete; (3) the historical data categories are not clear, it is difficult to reorganize the historical context and to reproduce the pattern and appearance of the stone culture; (4) the system presents mostly forms of static content, similar to the early paper collection of the presentation design, the lack of time axis guideline and the event series and other dynamic presentation; (5) the sustainable management issue of digital collections. The current situations can be summarize into two levels factors: (a) content presentation: the relevant factors in the activities of the stone weirs haven't been fully summarized yet, it is difficult to understand the pattern between the stone weirs and the local life, and has become the barrier for humanistic researchers to investigate the changes of the stone weirs (b) information and communication technology needs: the past research results show that the man-machine interface is one of the important factors in the use of digital collections, so the man-machine interface design is particularly important, and the multiple resources of the stone weirs are not effectively integrated and applied.

[1] Department of Business Administration. Feng Chia University, Taiwan

[2] Enterprise Information System Center, Feng Chia University, Taiwan

This study aimed to focus on protecting origin appearance of stone weirs culture and dynamic displaying Penghu ancient life story in community. First, summarize the elements of the stone weirs , then separate the activities dimensions, interface, service and process. Second, use the concept of hypermedia to visualize and connect the interface elements, and assemble the interface layer, the service logic layer, the data layer under the service-oriented architecture for the process integration. Third, build the historical pattern of the original stone culture, and manage the stone weirs knowledge through the digital form to provide users to investigate the local humanities and the environment under different time and space and to understand the social and economic relations and implications of the ancient Penghu community. Users can accordance with needs to assemble multi-media resource by itself to show dynamic diversification of stone weir and to return culture visualization.

This study proposed and developed dynamic visualization models for the display of stone weirs and Penghu culture provided for two efforts: (1) Researchers of humanities can follow mechanism of stone weirs culture to analyze features about lifestyle and religion in Penghu community in at the past from built style and year of stone weirs. (2) Application of value-adding data is for marketing and popularizing, to provide general users to realize Penghu local culture and to experience cultural story and life behavior of mutualism of local forefather and ocean.

Keyword: Stone weirs, Information Communication Technology(ICT)、Service Oriented Architecture(SOA), Service as a Service (SaaS),Visualization.

Analyzing Cybersecurity-related Articles from World’s Major Newspapers

Piyush Ghasiya, Koji Okamura

Securing nations citizen, critical infrastructure, and internet-of-things (IoT) devices from cyber-attacks is the primary concern of governing authorities of almost every country of the world. Every day there is news of some kinds of data breach, hacking, and backdoors. As the importance of cybersecurity is growing, the attention given to it by media is also increasing. In this scenario, it would be interesting to analyze how some of the most prominent newspapers (online) from different countries are reporting about cybersecurity. Doing this would not only help in understanding the principal concern of a nation in the field of cybersecurity but also useful in grasping the similarity and the difference between countries priorities to the critical events under cybersecurity area. The countries that are considered for this research are Japan, the U.S., the U.K., Australia, Canada, and India. English language website of newspapers will be used for this research. The period chosen to collect articles is from April 2018 to March 2019.

This research can be divided into four main steps:

1) *Step One – Data Acquisition*

   This step will collect cybersecurity-related articles from at least three newspapers from each country. For that, Python’s BeautifulSoup package will be used.

2) *Step Two – Content Analysis*

   Content Analysis of collected articles will be performed using KH Coder. KH Coder is an open source software for quantitative content analysis or text mining. As the main objective of this research is to understand how newspapers are reporting cybersecurity related issues in different countries, this step would help to find critical patterns from the collected corpus.

3) *Step Three – Sentiment Analysis*

   After finding patterns from step two and categorizing the corpus accordingly, sentiment analysis would be performed on only articles related to the specific issue. Like content analysis, sentiment analysis is also a content-based analysis. However, sentiment analysis categorizes the text into positive or negative. Python’s TextBlob package will be used for performing sentiment analysis.

4) *Step four – Critical Analysis*

   This step would critically analyze the specific issue or issues that content analysis showed from step two.

First, all four steps would be performed for each country, and the comparative analysis between different countries is performed to understand the difference.

Analysis of Japan’s Newspaper

As this researcher has already finished the analysis of Japan’s newspapers. Below are the results of each step.

1. *Step One - Data Acquisition*

   182 Cybersecurity-related articles from three newspapers: The Japan Times, Asahi Shimbun, and Mainichi Shimbun are collected between the period mentioned above.

2. *Step Two – Content Analysis*

   The content analysis used term-frequency analysis and co-occurrence analysis for finding patterns. These analyses showed that contention between the U.S. and China over Huawei’s 5G technology or network is the crucial issue. Forty-five articles (25%) of the collected corpus are related to Huawei. This can explain the importance of the ongoing tussle over 5G for Japan.

3. *Step Three – Sentiment Analysis*

   From the perspective of International Relations (IR), the Huawei issue is a negative development. Even the media is also using the word such as tech-war, race, contention, and battle. In this setup, it would be interesting to look at how sentiment analysis categorizes Huawei-related articles. 89% of 45 articles are categorized as positive. This shows the limitation of the content-based analysis. This point can be further pondered upon, as this research move further in analyzing newspapers from other countries.

4. *Step Four – Critical Analysis*

   Critical analysis of the Huawei issue showed that Huawei is one of the leaders in 5G technology. The U.S. allegations of a backdoor in Huawei product and spying for the Chinese government is only theoretical till now. Huawei is just a pawn in this battle of tech supremacy between the two most powerful nations of the world. Japan, the biggest ally of the U.S. in Asia, also acted against Huawei and banned it. This whole issue can create a deep fissure in Japan-China relation, which was showing a sign of normalization after a long time.

Future research would analyze cybersecurity-related newspaper articles from the countries mentioned above. As this is an ongoing research, the author is still in the process of collecting data from other countries starting from the U.S. The analysis of newspaper articles from the U.S. and comparative study of Japan and the U.S. would be included in the final poster.

Keywords – Content Analysis, Sentiment Analysis, Critical Analysis, Newspaper, Cybersecurity

Approach to develop Digital Collection for Small Organization considering Sustainability and Reusability with IIIF and Static File

Satoru Nakamura[1]

1. Introduction

Digital collection plays an important role as a research resource for digital humanities. In particular, the recent introduction of IIIF has greatly improved the interoperability and reusability of image resources. On the other hand, one of the challenges in digital collection is the sustainability of the system. Eschenfelder[1] introduces a nine dimensional framework for organizational sustainability in the digital cultural heritage community; Technology, Management, Relationships, Revenue, Cost, Valued product, service, Disaster planning, Legal/policy, and Metrics/assessment. In this research, we focus on three of those dimensions, which are Technology, Management, and Cost. In concrete, maintaining staff and budget for system operation and maintenance over the years is difficult. To address this issue, Egusa[2] proposes a way to develop digital collections using only static files such as HTML, CSS, JavaScript files. Her approach, or generic SSG (Static Site Generators), offers better security and reduce server related requirements such as server side language and database software. This reduces maintenance and migration costs. In this study, IIIF is added to her approach, and a methodology is proposed to develop digital collections that take into account both sustainability and reusability with IIIF and static files.

2. Proposed method

The proposed method consists of the following four steps.

- Upload image data to server
- Create metadata with spreadsheet software
- Convert data to HTML and JSON files
- Provide browsing environment using online tool

In the first step, the images in the collection are uploaded to the server and can be accessed by a URL. IIIF Image service is desirable but not required in IIIF Presentation API[3], therefore this approach also does not require IIIF-compatible image server.

Next, metadata about the item is described using spreadsheet software, such as MS Excel and Google Spreadsheet. Two types of data are required. One is a list of metadata for each item. Metadata fields specified in spreadsheet software header

[1] The University of Tokyo

lines are represented by URIs, and this data can be converted to RDF data. For example, the title field is represented by "dcterms:title". In addition, required fields and optional items are provided as metadata. For example, licenses (dcterms:rights) and attribution (sc:attributeLabel) of items are provided as required fields, and these data are used to generate IIIF manifests.

The third step is data conversion from spreadsheet to IIIF manifest (JSON file) and landing page (HTML file) of each item. This data conversion uses a scripting language such as Python. In addition, RDF data (such as XML or JSON file) and IIIF collection (JSON file) are also generated for bulk data download. Especially for landing pages, RDF metadata is embedded by RDFa to improve machine readability. By uploading these data to the server, each item can be accessed by URL.

The fourth step provides a browsing environment using various online tools. Online Document Viewer[4] is used to browse the list of metadata in MS Excel. Users can browse images with the landing pages and IIIF viewers such as Universal Viewer that loads the IIIF manifest. In addition, Image Annotator[5] is used to list thumbnails and browse images based on the IIIF collection.

The proposed methodology applies to the development of several digital collections. These sites are basically published using GitHub Pages, which is a static file hosting service. This realized the improvement of data reusability with IIIF, while improving the sustainability by developing web sites with only static files.

3. Discussion

This study does not propose a whole new approach. Several organizations have managed metadata in a spreadsheet and converted to IIIF manifests using tools such as [6, 7, 8]. Regarding publication of websites using GitHub Pages, it has been applied to different websites as guided by Visconti[9]. The difference from these existing works is that this proposed method covers the end-to-end process, from uploading images to developing a digital collection system with JavaScript. We also need to think about updates of JavaScript, but the data with standard formats such as RDF data and IIIF Collection will help to reduce the cost to handle those updates.

Furthermore, for the purpose of developing larger digital collections, a search engine such as Apache Solr should be adopted. On the other hand, the proposed methodology is an approach which aims at developing a small digital collection that can browse the list of items by spreadsheet software. Although the functionality of the search system in this proposed methodology is limited, this approach provides high data reusability and interoperability. Thus, data such as the IIIF manifest may be easily harvested by others or applications, and the inconvenience of the developed system can be solved by them. This is the utility of the proposed methodology that takes into consideration the sustainability and reusability of digital collections.

[1] Eschenfelder, K. R., Shankar, K., Williams, R. D., Salo, D., Zhang, M., & Langham, A. A nine dimensional framework for digital cultural heritage organizational sustainability: A content analysis of the LIS literature (2000–2015), Online Information Review, Vol.43, No.2, pp.182-196, 2019.

[2] 江草由佳. 移行しやすく使いやすいデジタルアーカイブの構築：教育図書館貴重資料デジタルコレクションの経験から, 情報知識学会誌, Vol.28, No.5, pp.367-370, 2018.

[3] https://iiif.io/api/presentation/2.1/#image-resources

[4] Online Doc Viewer | View MS Office Documents Online, https://products.office.com/en/office-online/view-office-documents-online

[5] Image Annotator, https://www.kanzaki.com/works/2016/pub/image-annotator

[6] https://github.com/ndlib/marble-manifest-pipeline

[7] https://github.com/ColbyMuseum/export_align

[8] http://georgetown-university-libraries.github.io/File-Analyzer-Test-Data/iiif/demo8

[9] Visconti, A. Building a static website with Jekyll and GitHub Pages, The Programming Historian, 2016.

Building Models from Topic Models

Radim Hladík[1]

The method of automated topic analysis known as LDA – Latent Dirichlet Allocation (Blei et al., 2003) – can process large number of texts as well as detect topics – defined as distributions of words across documents – that researchers would not necessarily find with traditional content analysis. Although it has not been always welcomed enthusiastically by digital humanists (Schmidt, 2012), its sheer utility has nevertheless made it a widely-used tool (e.g. Schöch, 2017; Tabata, 2017; Goldstone and Underwood, 2012). From the digital humanities perspective, the fact that LDA provides ordered lists of words makes it superior to methods such as latent semantic indexing (LSI), a type of dimension reduction technique for document-term matrices, which is readable by machines but noninterpretable by humans. In addition to topics in the corpus, LDA also calculates descriptions of documents as mixtures of topics. This is often desirable, as usually readers do not expect texts to be about one and single topic only. Another potentially advantageous attribute of LDA is the assignment of words to topics based on probabilities, no matter how infinitesimal it may be. This result matches well with the intuition that the same word can appear in more contexts from which it will ultimately derive a specific meaning.

The LDA method has also its drawbacks, some of which are overcome in its more advanced variations. For example, the mutual independence of topics in the traditional LDA violates the common-sense assumption that some topics are more likely to be discussed together. The correlated topic models (CTMs) makes room for correlations among individual topics (Blei and Lafferty, 2007). Structural topics models (STMs) go a step further and allow for specification of external metadata with which the topics should correlate. Effectively, this means that STMs make it possible to trace associations between topics and other variables (Roberts et al., 2013). However, because these associations are constructed in the very process of building the topic model, STMs cannot be used for inferential claims.

The presentation will address another shortcoming of LDA that is the property of the compositional nature of the Dirichlet distribution (or other normalized distributions) used under the hood of the technique. The mixtures of topics on document level cannot be used in regression models, where they would introduce multicollinearity. Alternative regression models for compositional data exist, but rely on transformations that make interpretation of effects challenging (Filzmoser et al., 2010). Discarding some topic loadings based on arbitrary thresholds (Antons et al.,

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

2018) can allow for models to work computationally, but such hacks do not respect the nature of compositional data, where a change in one value must be accompanied by a corresponding change in at least one other value. In short, while LDA-based topic models allow researcher to quantify semantic information, using these models beyond the document classification tasks is cumbersome at best.

Alternative solutions to topic modeling can come from modeling of semantic networks (Leydesdorff and Nerghes, 2017). A recent method, TopSBM model, proposes hierarchical stochastic models in bi-partite networks made up of words and documents (Gerlach et al., 2018) as a network alternative to LDA that shares mathematical properties with probabilistic topic modeling. It does not require setting of hyperparameters and, perhaps more importantly, the topic models based on TopBSM have non-probabilistic, empirical distributions. The technique porposed by Gerlach et al. (2018) normalizes topic distributions per document, but the existence of empirical distributions and non-overlaping topic categories actually does not necessitate this step. Instead, we propose a different implementation that allows us to express the keyness (Pojanapunya and Watson, 2018) of topics in a document as log odds ratios instead of probabilities and thus overcome the multicollinearity issues associated with LDA and LDA-inspired methods. Such topic loadings can be interpreted as conditional probabilities of a document belonging to a topic given the overall prevalence of the topic in the corpus. Consequently, we are able to treat each topic as a fully independent covariate to be used in other models, i.e. as features in a regression. In this sense, despite having its own limitations, such as the inability to account for heterogenous meanings of the same word, TopSBM is a welcomed addition to the methodological arsenal for modeling semantic structures in corpora.

The proposed approach is illustrated by showcasing a regression model of citation counts of sociological papers. The model explores the question whether the topic of a paper can have effect on citations. If citations reflected only quality, we would expect citations to be distributed randomly across all topics. However, it is also conceivable that some topics (perhaps through a funding intervention by the government or thanks to the involvement of prominent scholars) are more “popular” among academic authors than others. To examine the effect of sociological topics on citations, we use the available data on Czech sociology. The data consists of articles published between 1993–2016 in the *Czech Sociological Review* (CSR), which is a “core” and generalist journal of the Czech sociology. The corpus of 522 articles was used for building a topic model with the TopSBM method. Only nouns appearing more than 2 times were retained to enhance both the computational and semantic efficiency of modeling (Martin and Johnson, 2015). For 499 of the articles, citation data were matched in the Web of Science with their citation records. Control variables include binary author-level data: single vs. team authorship, man vs. woman lead author, the age of publication, and lexical richness calculated as type/token ratio from random samples of 250 terms per document. Table 1 conveys the results of the regression model for cumulative citations after stepwise selection of topic variables (pseudo-R2 = 0.26). The ability to discard some topics with the stepwise or other methods of feature selection is one of the advantages of our approach. Topics themselves are reported through their five highest-ranking words.

We can see that quantitative (topic 19) and network-based (topic 22) approaches along with urban sociology (topic 6) yield advantage in citations. On the contrary, articles with a focus on social theory (topic 17) or with a penchant for philosophical reflection (topic 25) have negative effect on citations. The highest negative effective is associated with research on political decision-making (topic 30), which may be of more interest to qualitative sociology. An interesting contrast exists between this topic and the slightly positive effect of another political topic that, however, betrays preoccupation with more quantitative aspects of political life, such as voting behavior and surveys (topic 8). The opposite effects for theoretical (17, 25) and methodological topics (19,22) suggest a mismatch between theory and empirical research in Czech sociology. Perhaps the most important insight is that accounting for the topics of articles in the model of citation counts eliminates the significance of control variables, including the sex of authors and the team vs. single authorship. If future research arrives at the same conclusions, it will be strong evidence about the moderating effect of research topics on bibliometric indicators. The demonstration of the predictive utility of topical feature space can also inspire other applications.

Table 1:

Variable	Coefficent	Sig
(Intercept)	0.819 (0.651)	
1st author sex (female = 0, male = 1)	0.028 (0.119)	
Age (Years since publication)	0.006 (0.008)	
Collaborative authorship (single = 0, team = 1)	0.15 (0.124)	
Lexical richness (types/tokens ratio)	0.237 (0.966)	
Topic 01 school phase transition origin dynamics	- 0.087 (0.048)	.
Topic 06 city resident municipality house housing	0.115 (0.045)	*
Topic 07 question society case measure job	0.34 (0.22)	
Topic 08 election government democracy reform vote	0.094 (0.044)	*
Topic 13 sociology sociologist image order spirit	0.132 (0.077)	.
Topic 17 theory principle action law rule	- 0.161 (0.082)	*

Topic 19 number data information category factor	0.26 (0.09)	**
Topic 22 network interaction community integration exclusion	0.147 (0.045)	***
Topic 23 centre facility transportation countryside concentration	0.103 (0.06)	.
Topic 24 position demand claim conflict public	0.103 (0.077)	
Topic 25 world thing science culture critique	-0.224 (0.097)	*
Topic 26 sign communication intention media representation	-0.154 (0.055)	**
Topic 30 majority member support program decision	-0.29 (0.102)	**
Topic 33 sex old-age ageing senior doing	-0.108 (0.053)	*

*Note: . p < 0.01, * p < 0.05, ** p < 0.01, *** p < 0.001*

References

Antons, D., Joshi, A. M. and Salge, T. O. (2018). Content, Contribution, and Knowledge Consumption: Uncovering Hidden Topic Structure and Rhetorical Signals in Scientific Texts. *Journal of Management*: 0149206318774619 doi:10.1177/0149206318774619.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1): 17–35 doi:10.1214/07-AOAS114. http://arxiv.org/abs/0708.3601.

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

Filzmoser, P., Hron, K. and Reimann, C. (2010). The bivariate statistical analysis of environmental (compositional) data. *Science of the Total Environment*, 408(19): 4230–38 doi:10.1016/j.scitotenv.2010.05.011.

Gerlach, M., Peixoto, T. P. and Altmann, E. G. (2018). A network approach to topic models. *Science Advances*, 4(7): eaaq1360 doi:10.1126/sciadv.aaq1360.

Goldstone, A. and Underwood, T. (2012). What can topic models of PMLA teach us about the history of literary scholarship. *Journal of Digital Humanities*, 2(1): 39–48.

Leydesdorff, L. and Nerghes, A. (2017). Co-word maps and topic modeling: A comparison using small and medium-sized corpora (N < 1,000). *Journal of the Association for Information Science and Technology*, 68(4): 1024–35 doi:10.1002/asi.23740.

Martin, F. and Johnson, M. (2015). More Efficient Topic Modelling Through a Noun Only Approach. *Proceedings of Australasian Language Technology Association Workshop*: 111–15.

Pojanapunya, P. and Watson, T. R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1): 133–67 doi:10.1515/cllt-2015-0030.

Roberts, M. E., Tingley, D., Stewart, B. M. and Airoldi, E. M. (2013). The Structural Topic Model and Applied Social Science. *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation.*: 4.

Schmidt, B. M. (2012). Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities*, 2(1).

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

Tabata, T. (2017). Mapping Dickens s Novels in a Network of Words, Topics, and Texts: Topic Modelling a Corpus of Classic Fiction. In, *Proceedings of the 7th Conference of Japanese Association for Digital Humanities 'Creating Data Through Collaboration'*. Kyoto: Doshisha University, pp. 73–79.

Deep Learning based Japanese Early Books Understanding

Bing Lyu[1], Hiroyuki Tomiyama[2], Lin Meng[3]

Early books and literatures record lots of information as culture heritage, which help us understand the politics, history, culture etc. However, lots of these early books and literatures are unlocked which limits the currently research.

This paper introduces an in-work-progress, focus on understanding early Japanese books which are recoded in the Database of Ritsumeikan University [1]. In the Early Japanese Books Database, the early Japanese books are scanned and stored as digital images. When we try to lock these literatures by image processing, following technical problems exist.

1) Some of literature images have lots of noises due to the aging process, which makes the recognition became difficult.

2) The figure and character areas are ununiformed in the literature image.

3) The pictures and the texts are mixed together, increasing the difficulty of understanding the literature.

4) Lots of literatures are described by the cursive style which is very difficult to be recognized, because of lots of variation. And only few specialists can read the style.

In this paper, we propose a deep-learning based method for recognizing the characters and help preserving and protecting these cultural heritages.

In detail, for solving the problems of noise, Gaussian filtering and binarization are applied for noise reduction, firstly. The threshold is set by the Otsu method. For extracting the figure and character area, the pixel number of vertical axis and horizontal axis in binarized image are counted, and a statistical method are proposed for extracting the figure and character area.

About the character recognition, many image processing methods have been proposed. Currently, researchers find that deep learning methodes such as GoogLeNet, AlexNet etc. are better than the conventional image processing methods [2]. However, the problem of these models is the images should be cut previously.

In this research, we try to apply and extend the deep learning model of SSD (SingleShot MultiBox Detector) for extracting the characters from the literature images directly. Unfortunately, SSD is not fit for more than thousand objects detection. Furthermore, a larger number of characters exist in the literatures, and one words have several styles, and one styles have several variations especially in the case of

[1] Graduate School of Science and Engineering, Ritsumeikan University

[2] Graduate School of Science and Engineering, Ritsumeikan University

[3] Graduate School of Science and Engineering, Ritsumeikan University

cursive style. If the several type characters are mixed in the same training database, overfitting may happen.

Here, we plan to apply three network mode for solving this problem. Firstly, we try to apply a smaller network for recognizing the character types. Then, characters typed SSD and AlexNet combined networks are applied for detecting and recognizing every types character, respectively. SSD is used for character detection, which may give a bounding box to every detected character, and the detected characters are forwarded to AlexNet which is used for character recognition.

Acknowledgments

Part of this work was supported by a JSPS KAKENHI Grant-in-Aid for Early-Career Scientists (No. 18K18337) and Art research center of Ritsumeikan University.

[1] https://www.arc.ritsumei.ac.jp/en/database.html (2019/4/28 accessed)

[2] Lin Meng, Naoki Kamitoku and Katsuhiro Yamazaki, “Recognition of Oracle Bone Inscriptions Using Deep Learning based on Data Augmentation,” 2018 IEEE International Conference on Metrology for Archaeology and Cultural Heritage (IEEE MetroArchaeo 2018), Oct. 2018. (In Cassino, Italy)

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy: SSD: SingleShot MultiBox Detector, ECCV 2016: Computer Vision – ECCV 2016, pp.21-37,2016.

Emotional Effect on Utterance Styles in Fiction Writing

Hajime Murai[1], Naoko Matsumoto[2]

Introduction

Among the various elements in story texts, conversational sentences are a challenge for automatic processing. Colloquial language often includes irregularities, reflecting daily usage of omissions and idiomatic expressions. Therefore, it is difficult to process irregular word sequences using natural language processing techniques.

Moreover, in novels or general story texts, each character is differentiated based on their manner of speech; it is a popularly used technique to help readers understand each character's personality (Kinsui, 2003). Some readers can identify various attributes (e.g., gender, age, temperament, and social status as in actual daily conversations (Okamoto, 1997 and 1999)) of characters in a text based on the characteristics of each character's dialogs. Moreover, even if the speaker's identity is not elaborated through descriptive sentences, most readers can accurately identify the speaker through conversational sentences. Basic types of those Japanese utterance styles in story texts have been clarified based on usage of particles and auxiliary verbs (Murai, 2018).

In addition to those basic types that are related to speakers' various attributes, emotional states could have an effect not only on tones of voice (Siegman and Boyle, 1993), but also utterance styles. Therefore, this paper investigated the differences in utterance styles that are dependent on emotional styles in Japanese story texts.

Results of the scientific analyses can provide objectification and falsifiability to the interpretation of narratives. Moreover, they can clarify effective features for identifying characters' emotional states.

Target contents

To analyze relationships between attributes and conversational sentences, a tagged dialog corpus of Japanese novels was employed (Murai, 2016). This corpus is based on a random sampling of the texts of Japanese novels within the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). The Japanese texts, included in the Nippon Decimal Classification class number 913, were extracted from the library-based corpus in the BCCWJ, and 100 texts were randomly selected (Appendix). Although there is also a BCCWJ speaker information corpus that covers

[1] Future University Hakodate

[2] Future University Hakodate

the texts of Japanese novels in BCCWJ, that corpus includes only gender and age attributes.

Conversational sentences within the selected texts were extracted and attributes of the speaker (name, gender, occupation) and listener (name), relationship between the speaker and listener, and situations (e.g., family, office, criminal investigation…) were manually added to each utterance. A total of 5,632 utterances from 100 Japanese novels were tagged.

Tagging and categorization of emotion

Tags about the emotional state of the speaker were added for each utterance in the 100 Japanese texts based on consultation with two researchers. Those emotional states were inferred on grounds of the utterance and surrounding descriptive sentences. As a result, emotional tags were added to 554 utterances. The numbers of resultant tags are shown in Table 1. Those emotional states were categorized in four major categories (“Positive,” “Negative,” “Aggression,” and “Unexpected”) to fulfill the requirements of statistical analysis.

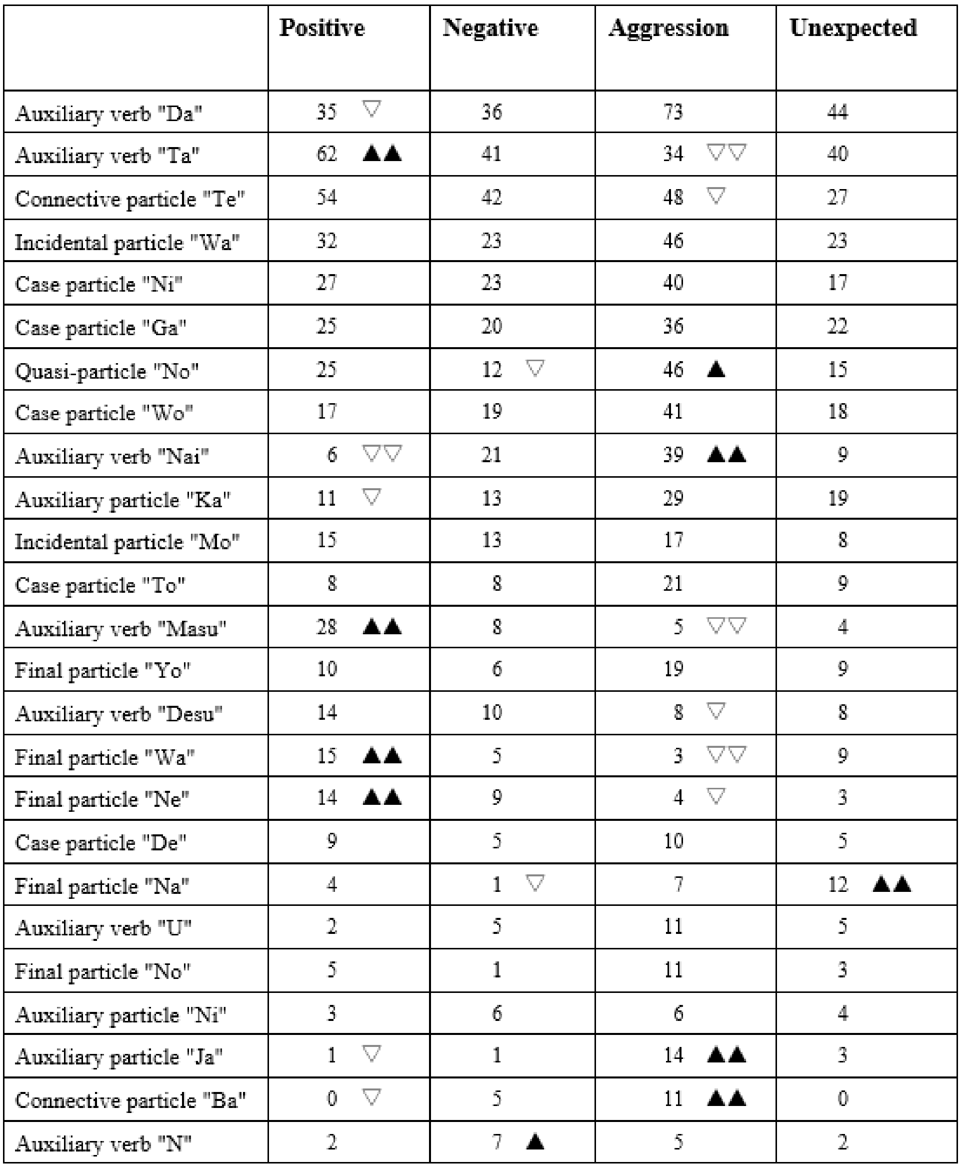
Table 1. Large and small categories for emotions within utterances

Positive		**148**
	Joy	36
	Confidence	4
	Admiration	12
	Gratitude	76
	Love	5
	Reassurance	14
	Nostalgia	1
Negative		**107**
	Sadness	12
	Regret	28
	Lack of confidence	10
	Feelings of guilt	40
	Compassion	17
Aggression		**173**
	Anger	117
	Disgust	56
Unexpected		**126**
	Surprise	61
	Bewilderment	24
	Fear	40
	Restlessness	1

Five hundred and fifty-four utterances were analyzed by Japanese parts of speech tags, and functional words (particles and auxiliary verbs) were extracted. For frequently appearing functional words, a chi square test was conducted (Table 2). In

Table 2, ▲▲ shows many at a 1% significance level, ▲ shows many at a 5% significance level, ▽▽ shows few at a 1% significance level, and ▽ shows few at a 5% significance level. Those cells meet Cochran’s rule.

Table 2. Chi-squared test of frequencies of particles and auxiliary verbs in four categories



	Positive	Negative	Aggression	Unexpected
Auxiliary verb "Da"	35 ▽	36	73	44
Auxiliary verb "Ta"	62 ▲▲	41	34 ▽▽	40
Connective particle "Te"	54	42	48 ▽	27
Incidental particle "Wa"	32	23	46	23
Case particle "Ni"	27	23	40	17
Case particle "Ga"	25	20	36	22
Quasi-particle "No"	25	12 ▽	46 ▲	15
Case particle "Wo"	17	19	41	18
Auxiliary verb "Nai"	6 ▽▽	21	39 ▲▲	9
Auxiliary particle "Ka"	11 ▽	13	29	19
Incidental particle "Mo"	15	13	17	8
Case particle "To"	8	8	21	9
Auxiliary verb "Masu"	28 ▲▲	8	5 ▽▽	4
Final particle "Yo"	10	6	19	9
Auxiliary verb "Desu"	14	10	8 ▽	8
Final particle "Wa"	15 ▲▲	5	3 ▽▽	9
Final particle "Ne"	14 ▲▲	9	4 ▽	3
Case particle "De"	9	5	10	5
Final particle "Na"	4	1 ▽	7	12 ▲▲
Auxiliary verb "U"	2	5	11	5
Final particle "No"	5	1	11	3
Auxiliary particle "Ni"	3	6	6	4
Auxiliary particle "Ja"	1 ▽	1	14 ▲▲	3
Connective particle "Ba"	0 ▽	5	11 ▲▲	0
Auxiliary verb "N"	2	7 ▲	5	2

Table 2 shows that polite (“Desu”) and feminine (“Wa” and “Ne”) utterance styles (Murai, 2018) were frequently used and the negative expression (“Nai”) does not appear frequently in positive emotions. More than half of the positive emotion was composed of “gratitude” and Japanese females more frequently express “gratitude” than Japanese males. Therefore, positive emotions are related to a feminine utterance style.

On the other hand, a negative utterance style frequently appeared, and polite or feminine styles were not used in a state of aggression. Moreover, the final particle “Na” was frequently used in unexpected situations.

Conclusions and future work

The results clarified that Japanese utterance styles in fiction writing change depending on emotional state. However, changes of utterance styles related to detailed emotional states have not been investigated because of the limitation of data size.

The detailed examination of relationships between emotional states and utterance styles could be enabled if a larger corpus, which includes more emotional expressions, could be utilized in the future.

Kinsui, S., (2003). *Virtual Japanese: Mystery of Functional Words*, Iwanami Shoten, Tokyo.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y., (2014). “Balanced corpus of contemporary written Japanese,” *Language Resources and Evaluation*, 48(22): 345−371.

Murai, H., (2016). “Towards agent estimation system for story text based on agent vocabulary dictionary,” IPSJ Symposium Series, 2016(2): 209–214, 2016. (In Japanese).

Murai, H., (2018). “Factor analysis of utterances in Japanese fiction writing based on BCCWJ Speaker Information Corpus,” *Advances in Human-Computer Interaction*, 2018 Article ID 5056268, 9 pages.

Okamoto, S., (1997). “Social context, linguistic ideology, and indexical expressions in Japanese”, *Journal of Pragmatics*, 28(6): 795–817.

Okamoto, S., (1999). “Situated politeness: Manipulating honorific and non-honorific expressions in Japanese conversations”, *Pragmatics*, 9(1): 51–74.

Siegman A.W., and Boyle, S. (1993). “Voices of fear and anxiety and sadness and depression: The effects of speech rate and loudness on fear and anxiety and sadness and depression,” *Journal of Abnormal Psychology*, 102(3): 430–437.

End to End word spotting network for modern Japanese magazines

Anh Duc Le[1]

With the development of open data in humanities, an enormous amount of historical documents has been available electronically on the Internet for humanities researchers. Such as large documents can be accessed efficiently if researchers can search and extract necessary text. The traditional approach for this task is making an index for documents manually. Since the manual approach is expensive, automatize indexing process will reduce costs. On our current project, a large of modern Japanese Magazines have been available for research communities. However, their indexes do not exist, researchers have to look for necessary keywords manually. In this research, we aim to propose a new method to spot keywords effectively on modern Japanese magazines.

For Japanese documents which have a complex layout and a large vocabulary, the state of the art methods for word spotting such as PHOC descriptors [1], PHOCNet [2], HWNet [3] are inappropriate, since they require that documents have been segmented into words. Moreover, the accuracy of OCR for modern Japanese magazines is still low. It is very challenging to make indexing from OCR results. To overcome the above challenges, we propose a new method for keyword spotting, which predicts locations of an input keyword from document images without any pre-processing.

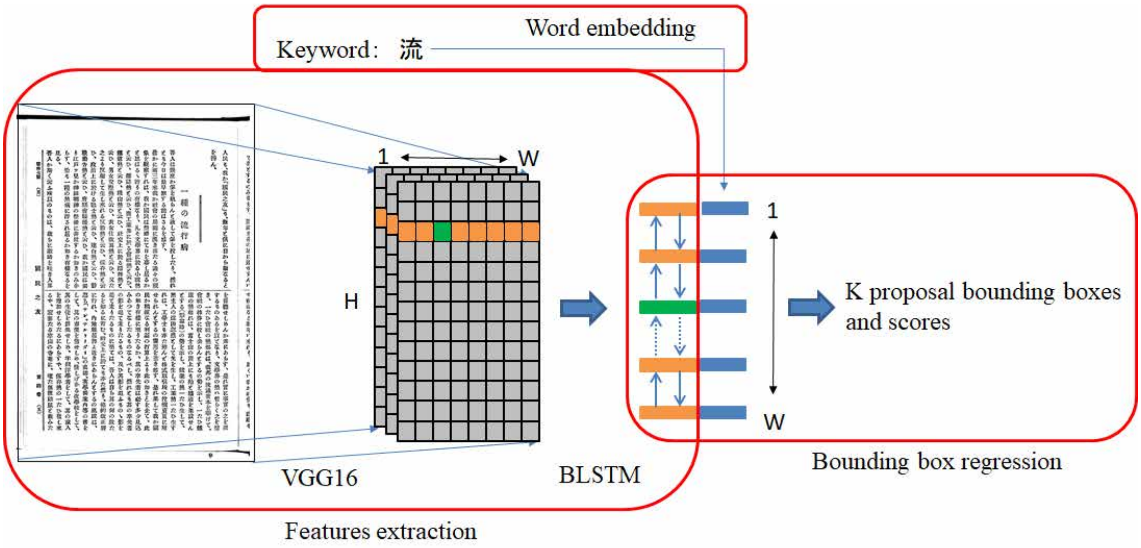


Figure 1. The architecture of the word spotting network.

The architecture of the word spotting network is shown in Figure 1. The

[1] Center for Open Data in The Humanities, Tokyo, Japan

network is inspired by Single Shot Detection [4] on object detection and Connectionist Text Proposal Network [5] on text detection. The network has three parts: features extraction, keyword embedding, and bounding box regression. For feature extraction, we employ VGG16 to extract features from an input image. Then, we employ a Bidirectional Long Short Term Memory (BLSTM) to explore meaningful context information of keyword. For keyword embedding, we convert a keyword character to an embedded vector. The embedded vector is concatenated with each extracted features. Finally, the bounding box is predicted from the concatenated vector. Figure 2 shows the result from the word spotting network for keyword “流”. There are three sequential bounding boxes (red, green, and blue boxes). The result location is determined by merging sequential bounding boxes to a single box for the input character.

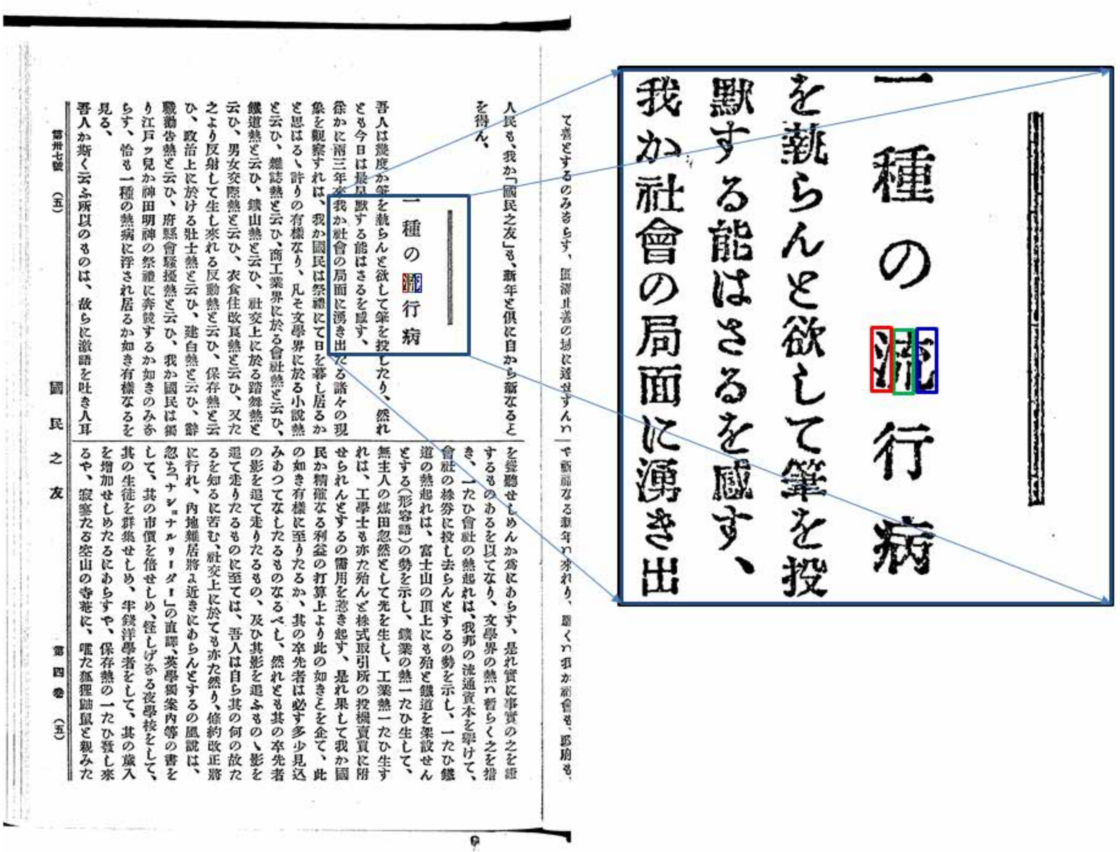


Figure 2. The result of word spotting for keyword “流”

This research is work-in-progress. Currently, we are developing the network on the Tensorflow framework. We plan to employ annotated documents from previous work [6, 7] for evaluating the end to end word spotting network. The dataset contains 1000 pages of historical magazines from 1870 to 1945. We randomly select 80% of pages for training, 10% of pages for validation and the rest for testing. We plan to do initial experiments by selecting 100 characters for keywords. Then, we will increase the vocabulary of keywords to 1000 and 3000 characters. The experimental results will be presented on the poster at JADH2019.

[1] J. Almazan, A. Gordo, A. Fornes, E. Valveny, Word spotting and recognition with embedded attributes, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (12) (2014) 2552-2566.
[2] S. Sudholt and G. A. Fink, PHOCNet: A deep convolutional neural network for word spotting in handwritten documents, in ICFHR, 2016.
[3] Praveen Krishnan · C.V. Jawahar, HWNet v2: An Efficient Word Image Representation for Handwritten Documents. https://arxiv.org/abs/1802.06194.
[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, SSD: Single Shot MultiBox Detector, ECCV 2016.
[5] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, Detecting Text in Natural Image with Connectionist Text Proposal Network, ECCV, 2016.
[6] 永野雄大, 幡谷龍一郎, 持橋大地, 増田勝也, CNN を用いた近代文献画像からのテキスト領域抽出, PRMU 2018
[7] Anh Le Duc, Daichi Mochihashi, Katsuya Masuda, Hideki Mima, An attention-based encoder-decoder for recognizing Japanese historical documents, PRMU 2018.

Facilitating Digital Humanities Research in Central Europe through the Advanced Speech and Image Processing Technologies

Jan Švec[1], Petr Stanislav[2], Marek Hrúz[3],
Aleš Pražák[4], Josef V. Psutka[5], Pavel Ircing[6]

In the recent decades, there is a constantly growing amount of multimodal data being collected and stored in order to be preserved for a future use. These data include – among other things – videotaped oral history interviews, archived footage of various TV broadcasts and a plethora of scanned hand-written and typed documents and photographs. The resulting archives present invaluable resources for many branches of the humanities (history, linguistics) and social sciences (political science, communication studies).

However, almost all of such archives share the same problem; unless they are really thoroughly equipped with the detailed metadata (describing e.g. topics of the individual documents, names and place appearing in the documents and/or being mentioned there, etc.) – and it is only rarely the case – it is almost impossible to find the desired information in the vast amount of data.

Our research lab has been participating in several projects (in some of them as the leading partner) that applied first the speech processing technologies and later also the image processing ones to facilitate the access to the information content of

[1] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

[2] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

[3] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

[4] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

[5] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

[6] Department of Cybernetics & NTIS – New Technologies for the Information Society Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

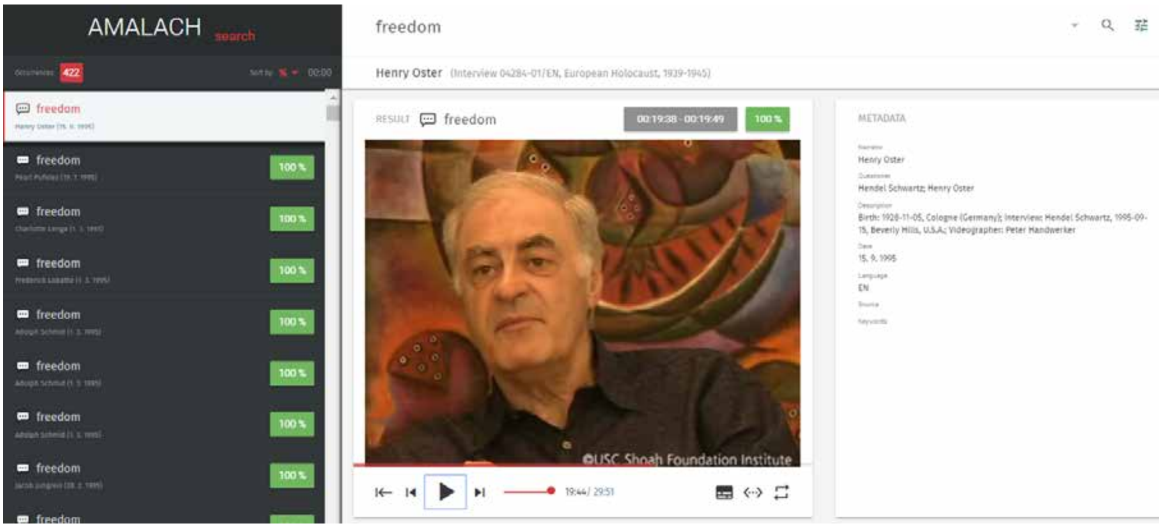
the archives.

First, we started with building the systems for automatic speech recognition (ASR) for the Czech, Slovak, Russian and Polish recordings from what is nowadays the USC Shoah Foundation Visual History Archive (http://vhaonline.usc.edu). The original plan was to transcribe the audio track of the video recordings into the plain text format and then use the traditional information retrieval techniques to find elaborately crafted search topics. This has proven to be problematic due to a relatively poor recognition accuracy (more than 30% Word Error Rate) on the challenging data and we have resorted to using an approach called *spoken* **term detection (STD)** instead.

In STD, the task is to find any occurrence of a specified word or a short phrase (**the query**) in the archive and return the exact time point where such occurrence takes place. Ideally, there is also a graphical interface that allows user to play the relevant segment from the recording instantly, as is shown in Figure 1.

Since the user does not need to read the text produced by the ASR as he/she watches directly the original passage from the recordings, we can use more than just the usual sequence of words that the ASR engine deems most probable (so called one-best transcription). Instead, we can take into account the entire set of recognition hypotheses stored in the form of a lattice – a directed acyclic graph where nodes represent the time instants and edges "carry" words together with a confidence score that represents a probability that a given word actually occurs within the given time interval.

In order to allow efficient searching, the lattices are processed into the data structure called the **(inverted) index** using the following procedure: The individual edges of the lattice are subject of a two-stage pruning. The first stage takes place at the beginning when all the edges whose confidence scoreis lower than a threshold $\theta_w = 0.05$ are discarded. Each of the remaining edges is represented by a 5-tuple (*start_t, end_t, word, score, item_id*) where *start_t* and *end_t* are the beginning and end time, respectively, *word* is the ASR lexicon item associated with the edge, *score* is the aforementioned confidence score and finally *item_id* is the identifier of the original video file (*start_t* and *end_t* represent the offset relative to the beginning of this file). The index is further pruned by removing similar items; that is, if there are two edges labeled with the same word that are either overlapping or are being less than $\Delta t_w = 0.5\,s$ apart, only the edge with the higher score is retained. It follows from the description that the indexing procedure omits the structural properties of the original lattice but, on the other hand, makes a compact and efficient representation of the recognized data. The resulting index is stored in a MongoDB database.


AMALACH search
422
freedom
Henry Oster (Interview 04284-01/EN, European Holocaust, 1939-1945)
RESULT freedom
00:19:38 - 00:19:49
100 %
METADATA
©USC Shoah Foundation Institute
19:44/ 29:51

Henry Oster

Hendel Schwartz; Henry Oster

Birth: 1928-11-05, Cologne (Germany); Interview: Hendel Schwartz, 1995-09-15, Beverly Hills, U.S.A.; Videographer: Peter Handwerker

15. 9. 1995

EN

Figure 1 - The latest version of the user interface for STD in the USC Shoah Foundation Visual History Archive

The current version of the search system depicted in Figure 1 is able to search in all the Czech and Slovak recordings from the USC Shoah Foundation Visual History Archive (approx. 1 000 hours and 800 hours, respectively) and about 2 000 hours of the English ones (in this case, it amounts to only about 4% of all the available interviews). All Czech, Slovak and English versions use the same core ASR, indexing and visualization technologies but of course there is also a substantial amount of localization to the individual languages – from the obvious ones such as different ASR phoneme sets and lexicons to the completely distinct approach to lemmatization used for Czech and Slovak on one side and English on the other.

In the last decade, we have started developing similar search engines for Czech institutions that have recorded similar audiovisual archives. The processing of the audio track of the video recordings was the main task in all cases but there were also some additional "data streams" to be explored. The first of the archives – the collection put together by Institute for the Study of Totalitarian Regimes (https://www.ustrcr.cz/en/) – includes also a substantial amount of scanned written documents that we wanted to make searchable in the similar manner as the audio content. The second archive – the ever-growing collection of the main news broadcast of the Czech public television broadcaster – contains the video recordings only but the broadcaster asked us to develop a technique that would allow to search also in the "visual" track. That is, to look for a specific person appearing in the video and to find the occurrence of words showing on the screen (running headlines, text documents shot by the camera or even banners in the crowd).

Thinking about those requirements carefully, we have found out that all of them can be rather easily indexed as in the case of audio-only indexing. Once we have a sequence (or lattice) of words, we can create a searchable index in the same manner as described above for the speech data. The challenge is then reduced to getting the text representation from all the modalities – this can be done using the **optical character recognition (OCR)** when dealing with scanned documents, **face detection and recognition** when searching for faces and a so-called **reading text in the wild** techniques when processing the text from the video footage. That way we can create multiple indices and have the application to search in all of them. Note that for the faces and the text captured in the video footage the individual components of the index 5-tuple have the same meaning as for the speech track, in the case of the words recognized in the scanned documents, we use the coordinates of the word bounding boxes instead of *start_t* and *end_t* labels.

Honkoku2: Towards a Large-scale Transcription of Pre-modern Japanese Manuscripts

Yuta Hashimoto[1], Yasuyuki Kano[2]

Introduction

In this poster presentation, we will describe and demonstrate the next version of *Minna de Honkoku* (https://honkoku.org/), a crowdsourcing platform for the transcription of pre-modern Japanese manuscripts. This was originally developed by members of the Historical Earthquake Study Group (HESG) at Kyoto University. HESG is a joint group of seismologists and historians at Kyoto University, including the present authors, who have been studying pre-modern earthquake records for seismic research and disaster prevention. The original *Minna de Honkoku* was a project to transcribe a vast number of historical records stored at the Earthquake Research Institute (ERI) of the University of Tokyo.

Although crowdsourcing has become a major technique for transcribing large volumes of historical manuscripts in the last 10 years, it has been claimed that this technique is not applicable to pre-modern Japanese materials, as reading *kuzushiji*, classical calligraphic renderings of Japanese characters, is prohibitively difficult for non-trained volunteers. *Kuzushiji* was common both for publishing and handwriting in medieval and early-modern Japan. However, owing to the drastic change in the writing system that occurred at the end of the 19th century, 99% of modern Japanese people are unable to read *kuzushiji*.

The authors' approach to this challenge was to integrate crowdsourcing with the online learning of *kuzushiji* [1]. In other words, we designed our crowdsourcing system as an online learning service for *kuzushiji*, so that users can participate in the project as a continuation of their learning. We developed the first version of *Minna de Honkoku* (Honkoku1) based on this approach, and launched it in January 2017. The result surpassed our expectations. By March 2019, 4,887 people had participated in the project, and the transcription of all 499 records (16,076 pages) stored at the ERI had been completed. The total number of characters transcribed by volunteers reached six million.

The next version of *Minna de Honkoku* (Honkoku2) aims to extend the approach described above to make it possible to transcribe a broader range of manuscripts on an even larger scale. To this end, two technologies will play key roles: the International Image Interoperability Framework (IIIF) and Handwritten Text

[1] National Museum of Japanese History

[2] Earthquake Research Institute, The University of Tokyo

Recognition (HTR).

Methods

Honkoku1 was designed to only deal with the digital images stored by the ERI. The introduction of IIIF will remove this limitation, and make it possible for our platform to collaborate with multiple cultural institutions. The IIIF is an international standard allowing the interoperability of images and collections of images, developed by a community of academic and cultural institutions [2]. Increasing numbers of institutions in Japan, including the National Diet Library (NDL) and National Institute of Japanese Literature (NIJL), have adopted the IIIF as a framework to publish their digital collections online. Honkoku2 will be IIIF-compatible; that is, it will be able to import images and their metadata from any digital archive that supports the IIIF via APIs, and conduct crowdsourced transcriptions of these images (see Fig. 1).

The last 10 years have witnessed rapid progress in HTR, mostly owing to the breakthroughs in deep neural network (DNN) technologies. A number of papers on the automatic recognition of *kuzushiji* using DNNs have been published in recent years [3,4]. However, HTR remains a computationally difficult problem, and it is still difficult to imagine that fully automated HTR of *kuzushiji* with 100% accuracy will become possible in the near future. Our idea is to employ HTR to augment rather than replace human transcribers, for more productive and accurate transcriptions. Honkoku2 will come with an automated recognition system for *kuzushiji.* Given a rectangle for a character, this will display candidate characters (see Fig. 2). We expect this feature to make it easier for transcribers, especially beginners, to decipher *kuzushiji*, thus leading to wider participation.

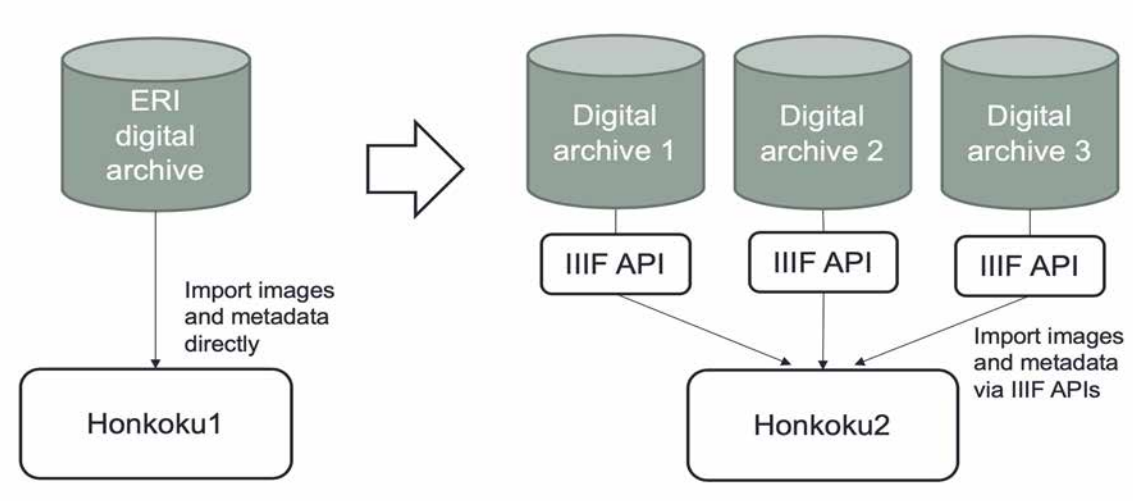
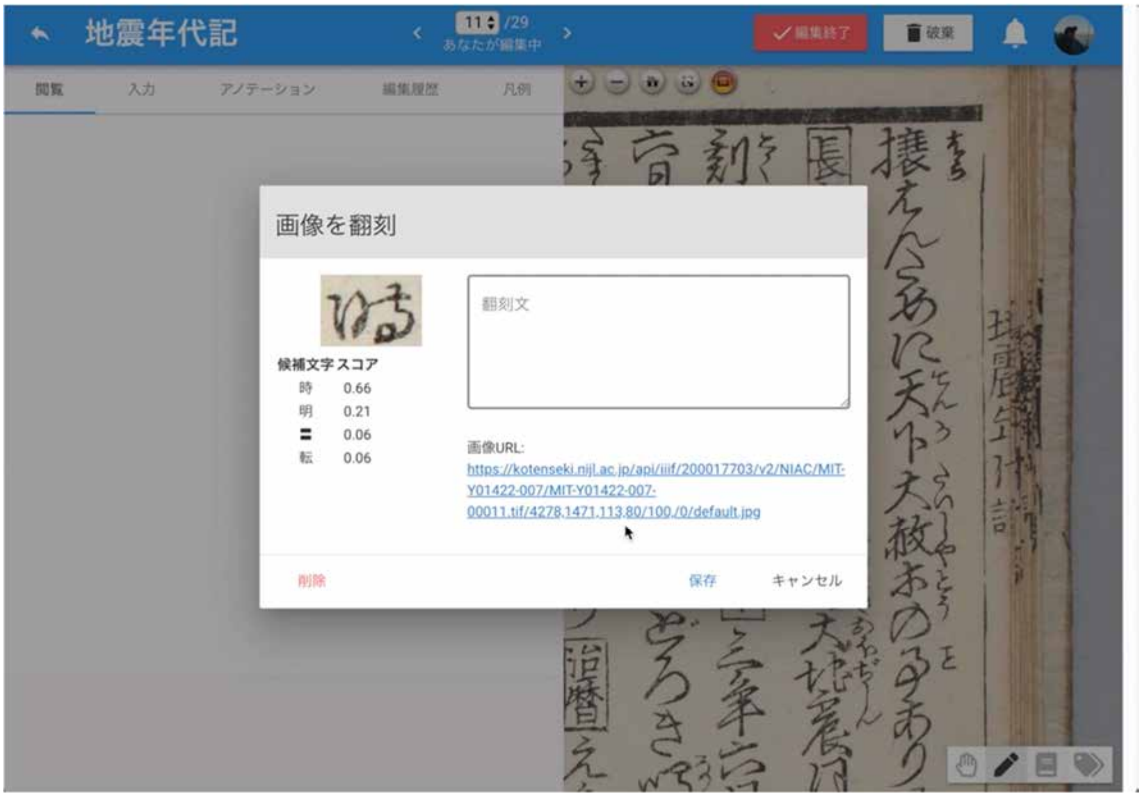


Fig. 1 A comparison of image handling between Honkoku1 and Honkoku2.



候補文字	スコア
時	0.66
明	0.21
〓	0.06
転	0.06

画像URL:
https://kotenseki.nijl.ac.jp/api/iiif/200017703/v2/NIAC/MIT-Y01422-007/MIT-Y01422-007-00011.tif/4278,1471,113,80/100,/0/default.jpg

Fig. 2 Automatic recognition of *kuzushij*. The program lists four candidate characters with their probability scores.

Conclusion

In contrast to Honkoku1, which began as a project to transcribe the historical records stored at the ERI, Honkoku2 aims to provide an inter-institutional crowdsourcing platform with the support of machine learning technology, for larger-scale transcriptions of pre-modern Japanese materials. Honkoku2 will officially be released at the start of July 2019. Thus, we believe that we will be able to report on the initial reactions of the public to Honkoku2 at the conference.

Bibliography

1. Yuta Hashimoto, et al. (2018). Minna de Honkoku: Learning-driven Crowdsourced Transcription of Pre-modern Japanese Earthquake Records. In *Proceedings of Digital Humanities 2018,* pp. 207-210, Mexico City, 2018.
2. International Image Interoperability Framework. https://iiif.io/
3. Hung Tuan Nguyen, et al. (2017). Attempts to Recognize Anomalously Deformed Kana in Japanese Historical Documents, In *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, pp. 31-36, 2017.
4. Tarin Clanuwat, et al. (2018). Deep Learning for Classical Japanese Literature, arXiv preprint, https://arxiv.org/abs/1812.01718.

Musical Pitch Expansion by Spectral Peak Shifting for Japanese Traditional Music Box

Shoji Ueda[1], Misaki Otsuka[2], Takahiro Fukumori[3],
Takanobu Nishiura[4], and Ryo Akama[5]

Digital archives have recently attracted attention for preserving tangible and intangible cultural heritage including art, sounds, and production processes. Japan has a wealth of cultural heritage, providing knowledge about the country’s history, which is important to preserve for future generations. Digital archives enable long-term preservation without time-related deterioration and can be disclosed to people via the Internet.

In this research, we focused on digital archives for the sounds produced by a Japanese traditional music box called *shikokin*. It is mechanical musical instrument that was popular during the middle of the Meiji era (from 1884 to the middle of 1897). It is difficult to determine what the music fashion was during the Meiji era from historiography; thus, we investigated the music fashion during the Meiji era to learn more information about Japanese culture. We also wanted to enable many people to learn about the *shikokin* by exhibiting the results of our research via an accessible system. To achieve this, we reproduced the sounds of the *shikokin* through digital signal processing [1]. For ease of access to our results, we also expanded the capacity of sound reproduction such as playing style and playable pitch. We also created a new musical piece by a combining traditional and modern music.

The *shikokin* has a crank on the side and fourteen holes on top, as shown in Fig. 1. Inside each hole, reeds are positioned with different lengths. When the crank is turned, the reed is vibrated by air flow through the hole, producing sound with the pitch related to the reed length. To play a melody, a rolled music score made of Japanese paper is required, as shown in Fig. 2. It has holes to pass the air flow, and a pattern of holes represents the notes of the melody. The sounds are produced only when the holes on the score overlap with the holes on the body of the *shikokin*. The score must be hand made with the holes in the proper position. Editing the melody requires modifying the holes by hand again. Therefore, it takes a long time to make a score. In digital data, the melody can be easily edited and is accessible to many

[1] Ritsumeikan University

[2] Ritsumeikan University

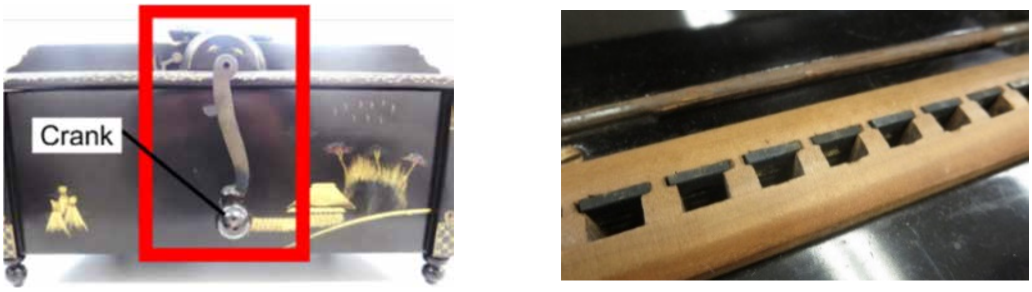
[3] Ritsumeikan University

[4] Ritsumeikan University

[5] Ritsumeikan University

people. In our previous research, we digitally reproduced the sounds, which are pitches that can be played with the *shikokin*. We also constructed an accessible system by expanding the pitch range to play contemporary music.

We developed a method of reproducing the sounds of the *shikokin* with the pitches of an 88-key piano. We first acquired the original sound played with the *shikokin*. However, squeaky noise was also included due to turning the crank. To reduce this noise, we estimated the power spectral peaks of the noise and subtracted the estimated peaks from the power spectrum of the acquired sound. We then reproduced the sound of the *shikokin* from the acquired sound with the 88 pitches. Previous studies on sound reproduction conducted concatenative synthesis [1, 2]. This can create pitch-changed sound by concatenating the stretched sound from the original sound. It can sufficiently reproduce sounds with the pitch range near the pitches playable with the *shikokin*. However, at extremely higher or lower frequencies, the pitch-changed sounds are close to electronic sounds. This indicates that it insufficiently reproduces sounds because it enhances the unnecessary spectral peaks depending on the spectrum of the original sound.



(a) Overall (b) Holes on top of *shikokin*

Figure 1: Japanese traditional musical box *shikokin*

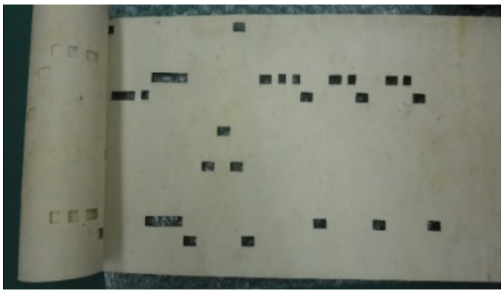


Figure 2: Rolled musical score

We expanded the range of the pitches from the original sound by spectral peak shifting. Figure 3 shows the flows of musical-pitch expansion. We first convert the waveform of the original sound to the spectrum. In this spectrum, the intervals between the contiguous spectral peaks contain the pitch information. Hence, the pitch is changed by shifting the spectral peaks. We can easily modify the power on the unnecessary spectral peaks because this method is used in the frequency domain.

We can then obtain the pitch-changed sound by converting the peak-shifted spectrum to the waveform. On the other hand, this method insufficiently reproduces the fluctuation in the sound pressure by turning the crank. The sound pressure of the *shikokin* fluctuates related to the rotational speed of the crank. To simulate this fluctuation, we measured the rotational speed. According to the measured speed, we applied the process that the sound pressure fluctuates in a random period to simulate it more naturally. We carried out a subjective evaluation to confirm the effectiveness of our method. This method reproduced the sound of the *shikokin* more smoothly and naturally than concatenative synthesis.

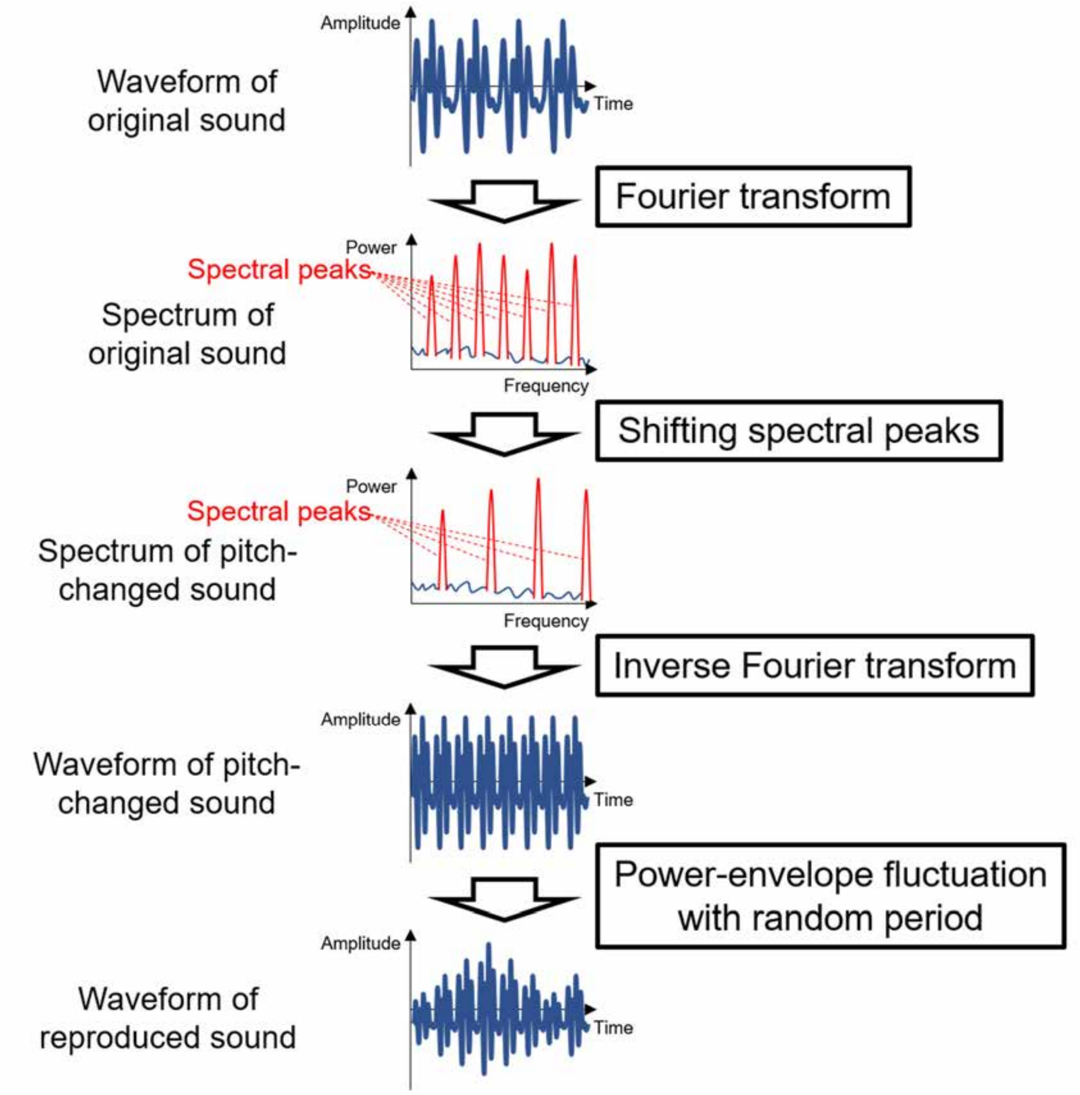

Waveform of original sound
Amplitude
Time
Fourier transform
Power
Spectral peaks
Spectrum of original sound
Frequency
Shifting spectral peaks
Power
Spectral peaks
Spectrum of pitch-changed sound
Frequency
Inverse Fourier transform
Amplitude
Waveform of pitch-changed sound
Time
Power-envelope fluctuation with random period
Amplitude
Waveform of reproduced sound
Time

Figure 3: Flows of pitch expansion

Based on above efforts, we constructed a web-based system to exhibit the results of our research. With this system, we can listen to the original and reproduced sounds of the *shikokin* and observe its operation. It also has a synthesizer to create a melody played with the *shikokin*. This synthesizer selects the reproduced sounds based on the melody inputted by the user and concatenates them by concatenative synthesis. We can listen to well-known contemporary music with traditional sound with this system.

Acknowledgement
This work was partly supported by Art Research Center, Ritsumeikan University.

References

[1] M. Otsuka, S. Okayasu, T. Fukumori, T. Nishiura, and R. Akama, “Sound reproduction by concatenative synthesis for Japanese traditional music box,” Culture and Computing, pp. 153-154, 2017.

[2] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a larger speech database,” ICASSP, pp. 373-37

Potentials of Games With a Purpose and Audience Participation Games for Descriptive Data Collection in Humanities Research

Ngoc Cuong Nguyen[1], Pujana Paliyawan[2], Ruck Thawonmas[3],
Hai V. Pham[4], Harada Tomohiro[5], Keiko Suzuki[6], and Masaaki Kidachi[7]

This abstract provides a survey on Games With a Purpose (GWAPs) and Audience Participation Games (APGs). We also discuss potentials in combining GWAPs and APGs on live-streaming platforms. Nowadays, many deep learning and machine learning techniques have been developed, which require a large amount of high-quality data in the training process. However, for certain applications, particularly those in the humanities, which need descriptions of cultural items such as images or artworks, such descriptive data are usually sparse, insufficient or low in quality. Usually, human experts are required to create the data, but this task is tedious, demanding and costly. GWAPs have demonstrated how games can harness human intelligence to generate more quality descriptive meaningful and relevant data without any pain but fun and enjoyable [1]. APGs are based on an idea of game live streaming that allows audiences to not only watch but also participate in part of play. It would be, therefore, interesting to combine a GWAP and an APG to utilize their advantages for descriptive data collection in humanities research.

GWAPs have proven successful in many different domains from computer vision, natural language processing (NLP) to web accessibility. For example, the ESP game was the first successful GWAP implemented by Ahn and Dabbish to annotate arbitrary images [2], in which a player guesses what his or her partner would type as keywords for a given image to achieve a matched label for the image. In another game called Karido [3], players play to obtain specific labels for artwork images. Dziedzic [4] proposed a GWAP to annotate natural language data in the NLP domain. The Phetch game by Ahn et al. [5] copes with the problem of attaching descriptive sentences to images on the web, which is beneficial in web accessibility. In the aforementioned GWAPs, the players have to concentrate on playing them, without doing anything else. Therefore, it is difficult to reach a massive number of players

[1] Ritsumeikan University, Japan

[2] Ritsumeikan University, Japan

[3] Ritsumeikan University, Japan

[4] Hanoi University of Science and Technology, Vietnam

[5] Ritsumeikan University, Japan

[6] Ritsumeikan University, Japan

[7] Ritsumeikan University, Japan

willing to spend time only on such games for a long time. Meanwhile, live-streaming APGs can be a perfect choice for implementing GWAPs as in-game mini games to reach massive audiences and have them do valuable tasks while enjoying watching the main contents.

Game live streaming is currently booming, with a large and ever-increasing number of people watching the contents on services such as Twitch. One of the reasons is because they want to interact with others and participate in streaming communities of their interests [6]. Therefore, with the increase of interaction-oriented audiences, the role of the audiences has changed. This idea of audience-participation gaming has blurred the line between audiences and players by allowing the former to impact gameplay conducted by the latter in a meaningful way [7]. The most well-known example of APGs is Twitch Plays Pokémon, a result of a crowdsourced attempt to play the Pokémon video game by parsing, processing, and executing commands sent by audiences through the channel's chat room [8]. However, the purpose of this kind of games was just for fun or to promote social interactions between audiences and streamers or audiences themselves, not for obtaining valuable data through audiences' participation in the games.

We are convinced that using GWAPs on live-streaming platforms can reach mass audiences who not only watch game live streaming but would also be willing to participate in playing mini games provided. For example, Nguyen et al. [9] introduced a mechanism for GWAPs on live-streaming platforms such as Twitch to obtain informative and descriptive sentences for ukiyo-e images by letting audiences play a mini game by using chat messages in addition to watching live streaming of gameplay.

In summary, this abstract presented an introduction to the literature on GWAPs and APGs and proposed to combine them in order to create high-quality descriptive data for humanities research and new applications. The combination of a GWAP and an APG is promising as it allows audiences to have more fun, build social interactions and create valuable descriptive data at the same time.

References

[1] Von Ahn, Luis, and Laura Dabbish. "Designing games with a purpose." Communications of the ACM 51, no. 8 (2008): 58-67.

[2] Von Ahn, Luis, and Laura Dabbish. "Labeling images with a computer game." In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 319-326. ACM, 2004.

[3] Steinmayr, Bartholomäus, Christoph Wieser, Fabian Kneißl, and Fracois Bry. "Karido: A GWAP for telling artworks apart." In 2011 16th International Conference on Computer Games (CGAMES), pp. 193-200. IEEE, 2011.

[4] Dziedzic, Dagmara. "Use of the Free to Play model in games with a purpose: the RoboCorp game case study." Bio-Algorithms and Med-Systems 12, no. 4 (2016): 187-

197.
[5] Von Ahn, Luis, Shiry Ginosar, Mihir Kedia, Ruoran Liu, and Manuel Blum. "Improving accessibility of the web with a computer game." In Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 79-82. ACM, 2006.
[6] Hamilton, William A., Oliver Garretson, and Andruid Kerne. "Streaming on twitch: fostering participatory communities of play within live mixed media." In Proceedings of the 32nd annual ACM conference on Human factors in computing systems, pp. 1315-1324. ACM, 2014.
[7] Seering, Joseph, Saiph Savage, Michael Eagle, Joshua Churchin, Rachel Moeller, Jeffrey P. Bigham, and Jessica Hammer. "Audience Participation Games: Blurring the Line Between Player and Spectator." In Proceedings of the 2017 Conference on Designing Interactive Systems, pp. 429-440. ACM, 2017.
[8] Ramirez, Dennis, Jenny Saucerman, and Jeremy Dietmeier. "Twitch Plays Pokémon: a case study in big g games." In Proceedings of DiGRA, pp. 3-6. 2014.
[9] Ngoc Cuong Nguyen, Zhenao Wei, Pujana Paliyawan, Hai V. Pham, Ruck Thawonmas, Tomohiro Harada, "Using GWAP to Generate Informative Descriptions for Artwork Images on a Live Streaming Platform," The 4th International Conference on Consumer Electronics Asia 2019 (ICCE-Asia 2019), Thailand, Jun. 1

Putting the Official History in Personal Memory: TEI and LOD

Sho Makino[1]

This study aims to make *History of the Irish Confederatio*n by Richard Bellings (written in 1660s and published in seven volumes from 1882 to 1891) marked up with TEI (Text Encoding Initiative) and connect to LOD (Linked Open Data). As one of the leading politicians of the Irish Confederation that fought the **British Civil War** in Ireland in the mid-seventeenth century, which was a fatal phenomenon, only Bellings composed the official historical writing of the Confederation. Therefore, his work is crucial for the interpretation of the seventeenth century Irish history.

Historians has regarded *the History of the Confederation* as an important primary source in the seventeenth century Irish history in two senses: public and personal. The former is that the book was written using a lot of official documents which now no longer exist because the archive of them was burnt down in the early eighteenth century. The other is, although it has been less mentioned comparing to the first one, that the work is composed of a personal memory of the author. Even though he composed the book with official documents, it has inevitable personal biases on his memory; it is assumable that the source is also fit for considering about a mental world of the seventeenth century Ireland. Therefore, considering the importance of the historical source, to formatise the text with TEI is a important work for the seventeenth century Irish historiography.

First of all, it is important to note that, as to develop this study for building a bigger model later, I am going to sample the first two volumes out of, in total, seven where each of them has over a hundred pages. Next, the way how to handle with the text will be three-fold: to obtain, clean, and put in TEI format. Initially, thanks to Google Books, the primary source already not only has images but also text data. Yet, as the text may not still be perfect, due to an accuracy of OCR (Optical Character Recognition), I will need to clean it up for marking up with TEI. Next, in order to clean the text-data automatically, making use of Regular-Expression with Python would be helpful. Thirdly, I am going to put them into TEI format, then, focusing especially on “person’s names” (<persName> tagging) and “place names” (<placeName> tagging) to connect them to LOD (Linked Open Data) technology such as VIAF (The Virtual International Authority File), DOI (Digital Object Identifier) given to biographical entries, and historical maps accordingly (O’Hara, 2013; 小風, 2019).

Regarding to the extracting names, Recogito should be taken into consideration for the NER (Named Entity Recognition) for automation of the

[1] Graduate School of University of Tokyo

procedure (Simon et al., 2015). Also, historical maps Recogito prepares is easier to plot places on a visual map. However, it will be necessary to modify the parson's names by hand to some extent in the end. This is because it is presumable that there might be a problem for the names; Irish names tend to have "Mc", "Mac" or "Ó" before surnames and the names were not set in a particular form yet in the seventeenth century.

In sum, it is possible to consider that this study prospects roughly two features. Simply speaking, first, the result of this study will make the text easier to specify who and where through tagging with TEI. Second, more importantly, specifications of person's names and place names connected to LOD enables to refer to the particular person's biography and his bibliography. Furthermore, as *the History* often describes where battles happened and how they went on, which implies that geographical information tells how the war progressed as time passed. Analysing the person's and place names, this study is going to clarify what emphasis Bellings intended to put on about the wars. Hence, putting *the History* in TEI will enable to regard it as a text of personal experience and thought as well as the official record of the Confederation.

Bibliography

Gilbert, J. (1882-1891). *The History of the Irish Confederation, and the War in Ireland,* 7 vols. Dublin.

Gillespie, R. (2000). "Political Ideas and their Social Contexts in Seventeenth-Century Ireland" in **O'Hara, L. T.** (2013). "Cleaning OCR'd text with Regular Expressions." *The Programming Historian* 2, <https://programminghistorian.org/en/lessons/cleaning-ocrd-text-with-regular-expressions>.

Ohlmeyer, J. (ed.), *Political Thought in Seventeenth-Century Ireland*. Cambridge University Press: 107-127.

Rankin, D. (2009). *Between Spenser and Swift: English Writing in Seventeenth-Century Ireland.* Cambridge: Cambridge University Press.

Simon, R. et al (2015). "Linking Early Geospatial Documents, One Place at a Time: Annotation of Geographic Documents with Recogito." *e-Perimetron*,10-2: 49–59.

小風尚樹 (2019).「Python によるセマンティック TEI マークアップのためのガイドライン」『「TEI で青空文庫勉強会」参考資料』、1–23 頁。

Query-Based Mashups of Historical Live Music Recordings

Florian Thalmann, Thomas Wilmering, Mark B. Sandler[1]

Navigating and searching large archives of audio or musical material can be tedious due to the temporal nature of the content. Whereas in visual archives one can gain a reasonably reliable overview of search results at a glance simply by scrolling through collated images, with audio material one has to go through search results by listening to individual recordings one by one – a process that may take a considerable amount of time and effort. In this paper we investigate different strategies of automatically creating sound collages based on user-defined queries enabling the user to gain a quick overview of a large number of results. We demonstrate the principles in a prototypical application based on the Grateful Dead collection of the Live Music Archive (LMA)[2].

Background

In recent years, advancements in music information retrieval have led to an improved organization of large digital music collections. In addition to extra-musical metadata, items in musical archives can now be automatically annotated with content-based information ranging from simple audio descriptors to high-level musical analyses. With these annotations one can now satisfyingly query and filter archive content, and quantify or visualize desired characteristics of the results (Elias et al., 2002, Bechhofer et al., 2017). However, for certain tasks it may still take a long time to make sense of these outputs, which can often only be done by listening to the discovered examples one by one.

To address this problem, we propose the concept of *query-based mashups* which can be generated once the material of an archive is annotated with appropriate audio features. These features can be used to align and reorganize relevant fragments of the material and present them to the users as coherent collages facilitating the exploration of the audio results in an interactive way. Query-based mashups are based on a three-step process:

- *Selection*: the user designs a search query to find a subset of the database's content. For example, all the recordings of a particular performance, all the versions of one song within a given time span, or all the performances in a particular place. The queries may also be based on audio descriptors such as

[1] Centre for Digital Music, Queen Mary University of London

[2] https://archive.org/details/GratefulDead

tempo, timbre, or musical structure.

- *Organization*: the users decide how the resulting material is to be organized or ordered, e.g. by recording date, tempo, or based on content similarity
- *Mashup Parameters*: finally, the users define how the material is to be mixed together: how much of each result is used (e.g. in number of bars or seconds), what part is used (e.g. the loudest or most varied segment), by how much the segments overlap, etc.

Demo Application

Our prototype is part of a Web platform focusing on live concerts of the band *The Grateful Dead* (Benson, 2016) of which a considerably large number of recordings survive in the LMA - over 12,000 from the years 1965 to 1995. The platform allows users to explore the band's concert history in the form of an audiovisual experience. The audio recordings are linked with content from several Semantic Web resources such as LMA Linked Data (Bechhofer et al., 2013) or DBpedia[3], as well as data from other Internet resources such as the Grateful Dead Archive Online at UCSC[4]. The data include information about venues, locations, setlists or lineups, and scans of artefacts such as tickets, posters, photos, or fan mail, which are aggregated and transferred into structured RDF data based on several ontologies (Thalmann et al., 2018). Audio feature extraction results are linked via the Computational Analysis of the LMA (CALMA) dataset (Bechhofer et al., 2017). In earlier publications we demonstrated how Semantic Web technologies are useful for interlinking various Cultural Heritage resources and searching them in a joint manner (Wilmering et al., 2016, Thalmann et al., 2018).

The platform allows exploring different entities such as shows, venues, locations, songs, tours, or musicians with their aggregated metadata as well as creating custom playlists of items found along the way. In addition to this we experimented with different alternative ways in which the musical material can be explored. One application automatically mixes the different recordings of one particular concert into an immersive experience by aligning and resampling the recordings to match the different tape speeds, and by clustering them into a multidimensional space based on their average distance from each other over time (Wilmering et al., 2016).

In our latest addition (shown in Figure 1) one can choose between different sets of songs, orderings, and parameter settings to create mashup compositions from the archive material. Mashups are particularly interesting in this context given that one of composer John Oswald's *plunderphonic* albums is based on selected Grateful

[3] https://wiki.dbpedia.org

[4] https://www.gdao.org

Dead recordings (Oswald 1996). The prototype is based on an in-browser automatic DJ mixing framework based on semantic audio technologies (Thalmann et al., 2018.). One can for example create a diachronic timelapse mashup that traces the evolution of one song through the history of the band, each version playing for the duration of a given number of bars. Alternatively, one can choose to organize the chosen recordings by similarity and overlap them for a significant amount of time in order to create a coherent sound collage reminiscent of John Oswald's. Figure 1 shows a mashup of the song *Me and My Uncle*, currently playing an excerpt of a concert at *The Matrix* and showing poster of the event. During playback, each fragment is complemented with corresponding visual artifacts from the collection. The metadata or artifacts can be clicked on to explore further and to reach the original recordings currently heard. The primary audience of this application are music listeners and fans of the band and the queries and mashups parameters are thus presented as a form of simple presets. However, following the principles outlined above, the platform could be easily adapted to enable digital humanities scholars and musicologists to answer specific questions such as for example finding the most unusual rendition of a particular song, or getting a sense of the audience and atmosphere at a particular venue over time.

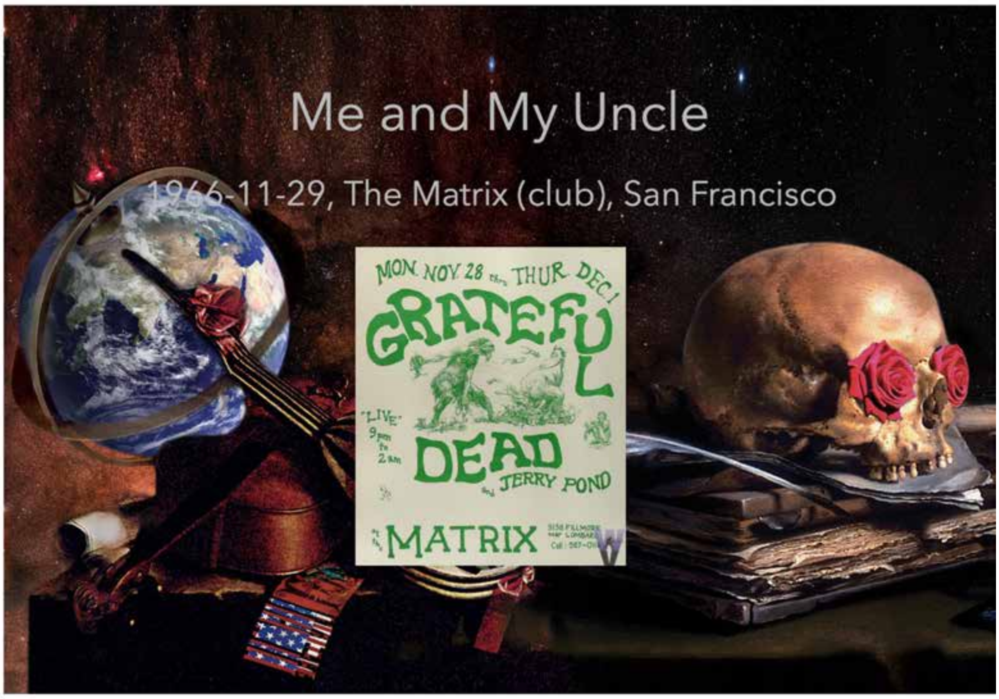

Me and My Uncle
1966-11-29, The Matrix (club), San Francisco
MON NOV 28 thru THUR DEC 1
GRATEFUL
DEAD
"LIVE" 9pm to 2am
and JERRY POND
MATRIX

Figure 1: The plunderphonics prototype during a mashup of *Me and My Uncle*.

References

Bechhofer, Sean, Kevin Page, and David De Roure. "Hello Cleveland! Linked Data Publication of Live Music Archives," in Proceedings of WIAMIS, *14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services*, 2013.

Bechhofer, Sean, Kevin Page, David M. Weigl, György Fazekas, and Thomas Wilmering. “Linked Data Publication of Live Music Archives and Analyses.” *Lecture Notes in Computer Science The Semantic Web – ISWC 2017*, 2017, pp. 29–37.

Benson, Michael. *Why the Grateful Dead Matter*. ForeEdge, an Imprint of University Press of New England, 2016.

Oswald, John. “Grayfolded” [CD]. Artifact Music, 1996.

Pampalk, Elias, Andreas Rauber, and Dieter Merkl. "Content-based organization and visualization of music archives." In *Proceedings of the tenth ACM international conference on Multimedia*, pp. 570-579. ACM, 2002.

Thalmann, Florian, Lucas Thompson, and Mark Sandler. "A User-Adaptive Automated DJ Web App with Object-Based Audio and Crowd-Sourced Decision Trees." In *Proceedings of the 4th Web Audio Conference*, Berlin, 2018.

Thalmann, Florian, Thomas Wilmering, and Mark B. Sandler. "Cultural Heritage Documentation and Exploration of Live Music Events with Linked Data." In Proceedings of the 1st International Workshop on Semantic Applications for Audio and Music, ISWC 2018, Pacific Grove, pp. 1-5. ACM, 2018.

Wilmering, Thomas, Florian Thalmann, and Mark B. Sandler. “Grateful Live: Mixing Multiple Recordings of a Dead Performance into an Immersive Experience.” *Audio Engineering Society (AES) 141st Convention*, 2016.

Rubbing Character Recognition base on Deep Learning and Lexical Analysis

Zhiyu Zhang[1], Yuxi Chen[2], Hiroyuki Tomiyama[3], Lin Meng[4]

Rubbing is a reproduction technique for the descriptions which are scribed on the bone, metal, stone etc., by placing a piece of paper over these subjects and rubbing the paper with ink. Also, Rubbings are among the oldest ancient literatures and potentially contain a lot of knowledge yet to be discovered. However, as of now, the rubbing characters are recognized manually which takes an enormous amount of time and effort, because there are so many different character styles, variations and lots of variations which are only legible to a small number of specialists. Still, the aging process, which may damage characters or add noises, further increases the difficulty of rubbing character recognition.

Currently, researchers try to recognize the rubbing characters for understanding the rubbing descriptions. Some studies have been shown that deep learning methods bring an exciting accuracy of rubbing character recognition, much better than the conventional image processing methods [1,3,4]. However, there are two problems limiting the accuracy improvement of rubbing character recognition. One is the shortage of the training data, resulting to models with insufficient training and then mis-recognition. The other is the broken characters caused by aging process. Some parts are lost or very blur, increasing the difficult of recognition process, which indicates that in these cases, deep learning might not be an optimal solution to good accuracy.

In this project, we propose a method by combining deep learning and lexical analysis for increasing the accuracy of rubbing character recognition. The method utilizes deep learning for obtaining the confidence of candidates, and uses lexical analysis method for re-recognizing the low-confidence and unclear characters. This work is work-in-progress, and the experimentation uses rubbing character database which is developed by Kyoto university. This database is the most complete database of the rubbing images and preserves about 5,000 rubbing images from China’s Han period (starting in 206 B.C.) to Qing period (finishing in 1912 A.D.) [2].

In detail implementation of the proposed algorithm, a conventional deep learning

[1] Graduate School of Science and Engineering, Ritsumeikan University
[2] Graduate School of Science and Engineering, Ritsumeikan University
[3] Graduate School of Science and Engineering, Ritsumeikan University
[4] Graduate School of Science and Engineering, Ritsumeikan University

Table. 1 Recognition results of deep learning

Target characters	Recognition Results				
	Rank1	Rank2	Rank3	Rank4	Rank 5
崇	崇(99.40)	崇(0.46)	宗(0.1)	茅(0.0)	堂(0.0)
哉	哉(99.83)	裁(0.09)	天(0.05)	災(0.02)	戌(0.01)
天	天(100)	大(0.0)	更(0.0)	戸(0.0)	未(0.0)
柱	怒(90.45)	柱(3.75)	妃(1.89)	桂(1.49)	掾(0.92)
逈	避(66.58)	堂(7.54)	怛(6.08)	逈(2.93)	哩(2.61)
出	出(99.97)	山(0.02)	土(0.01)	之(0.0)	二(0.0)
孤	孤(67.72)	弧(32.18)	豫(0.04)	軫(0.01)	鞘(0.01)
亭	亭(99.88)	戸(0.03)	粟(0.03)	茅(0.02)	宗(0.01)

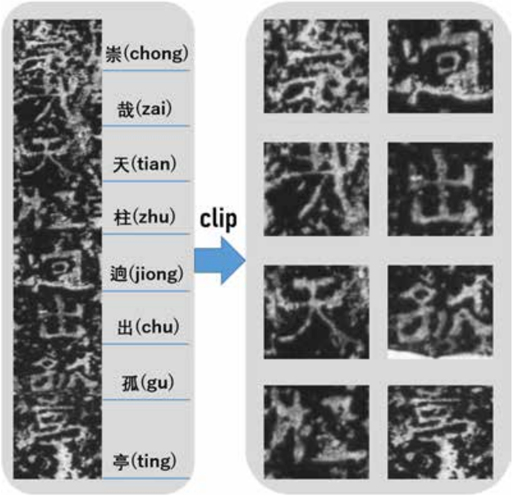


Fig1. Rubbing image example

method (AlexNet)[3] is first used for recognizing the characters and obtaining the confidence of the candidates (classes). Then, rubbing descriptions for clear character images are generated by selecting the recognition results with high-confidence. The rubbing description dose not complete here, and the character with blurry images or low-confidence candidates will be re-recognized. The third step is lexical analysis which uses statistical methods such as the appearance frequency of words, N-Gram etc., for predicting the unclear or low-confidence characters. Certainly, in the lexical analysis step, these unclear and low-confidence characters are not used, on the contrary, the weights of these candidates are calculated. At last, the weights and the confidences of these candidate are used for the calculation and decision of the final re-recognition results.

Table 1 shows recognition results of the rubbing image of Fig. 1. The characters are less legible. In Table 1, the column of target character shows the Kanji of the rubbing image according to the rubbing character database provided by Kyoto University [2], and the ranking shows the candidates obtained through deep learning and the confidence of them. The results prove that deep learning is an effective method for rubbing character recognition, however its accuracy needs to be improved.

Therefore, we propose to re-recognize the rubbing characters using lexical analysis. Hence, combining the deep learning and lexical analysis for achieving high accuracy recognition becomes an important future issue.

Acknowledgments

Part of this work was supported by a JSPS KAKENHI Grant-in-Aid for Early-Career Scientists (No. 18K18337) and Art research center of Ritsumeikan University. The authors wish to acknowledge Institute for Research in Humanities, Kyoto University and Dr. Koichi Yasuoka, Professor of Digital Humanities, Kyoto University, for providing the data and research advices.

[1] Lin Meng, Masahiro Kishi, Kana Nogami, Michiko Nabeya and Katsuhiro Yamazaki, "Unlocking Potential Knowledge Hidden in Rubbing: Multi-style Character Recognition using Deep Learning and Spatiotemporal Rubbing Database Creation," EuroMed2018 7th International Conference on DIGITAL HERITAGE, LNCS11196, pp.741-751, 2018.
[2] Rubbing characters Database, Kyoto University. http://coe21.zinbun.kyoto-u.ac.jp/djvuchar. (2019/4/30 accessed)
[3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), 2015.
[4] Alex Krizhevsky, IIya Sutskever, and Geoffrey E. Hinton, G.E, "ImageNet Classification with Deep Convolutional Neural Networks," Neural Information Processing Systems 25 (NIPS 2012), 2012

The Hidden Link -- A Clue to Explore the Relationships among Poets, Capitalists and Social Activists Through Analytical Tools

Su-Bing Chang[1], Hao-Ren Ke[2], Shun-Hong Sie[3]

Taiwan Biographical Database (TBDB) is a database for Taiwanese historical figures. TBDB comprises a database and a set of analytical tools according to the characteristics of modern Taiwanese history and historical figures and the needs of historians. One of the major purposes of developing TBDB is to build a model for exploring Taiwanese figures, either through the prosopography, social network analysis (SNA), or geographic information system (GIS).

In the initial phase, TBDB includes 888 figures from the Personage Biographies in the revised Local Gazetteers of ChangHua County, all authors of which participate in the creation of TBDB. TBDB employs the concept of object-oriented databases for the properties of figures and Lucent as the search engine for biographic full-texts, and it is equipped with several tools such as text mining, SNA, and GIS.

One important tool is the SNA tool, which is designed to help both scholars and those who do not have the digital humanities background to easily perform basic social network analysis.

In modern Taiwanese history, group- or poetry-society- participation was a popular and potentially important activity for Taiwanese historical figures. Therefore, when designing the database, we have been paying special attention to such new historical phenomenon. Similar to CBDB (China Biographical Database), we attempt to populate TBDB automatically by named entity recognition (Bol, Hsiang and Fong, 2012; Liu, Huang, Wang, and Bol, 2015).

The proposed approach is a semi-supervised approach. Historians build a preliminary list of poetry societies, and store them in the peripheral database. Three parameters can be computed from the preliminary list: 1) the average length of poetry society names, **l**, 2) the length of the longest common strings between poetry society names, **n**, and 3) the respective longest common strings, **W**. Given a biographic full text **S**, if **W** appears in the $\mathbf{p}^{th}$ position of **S**, then **substring(S, p-l+n, l-n)+W** and **W+substring(S, p+n, l-n)** may be new poetry society names. Fig. 1

[1] Graduate Institute of Taiwan History, National Taiwan Normal University, Taiwan, R.O.C.

[2] Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taiwan, R.O.C.

[3] Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taiwan, R.O.C.

shows that if **n** =2, **l**=4, and **W**=吟社, four poetry societies, 菱香吟社, 螺溪吟社, 興賢吟社, 香草吟社, can be recognized (Sie, Ke and Chang, 2017) . This approach is simple and effective, and can be extended to other types of named entities such as organizations and schools.

.....他被推選為菱香吟社首任社長, 擅長傳統古詩……亦多次於北斗螺溪吟社、員林興賢吟社、二林香草吟社擔任課題詞宗.....。

Fig1. An example of the recognition of poetry societies

Based on these results, TBDB could draw a map about relationships between members and poetry societies. As Fig. 2 shows, when combining members, poetry societies and locations, TBDB provides a start point for researchers to further investigate. We have to admit the fact that because TBDB lacks the information on the years of poetry-society participation of the historical figures, it can only give a clue for researchers to explore. Despite this, the SNA tool has the ability to analyze the social networks of the members of poetry societies. With an initial analysis, we have already found various interesting points that are worth further investigation. First, there were very few overlap of membership among different poetry societies. Second, two sets of poetry societies (each included two poetry societies) have a significant number of overlapping members. Third, among the four largest poetry societies, three were well connected in the social network, but one by and large was secluded from the network. All the three points are worth further investigation to uncover their historical significance.

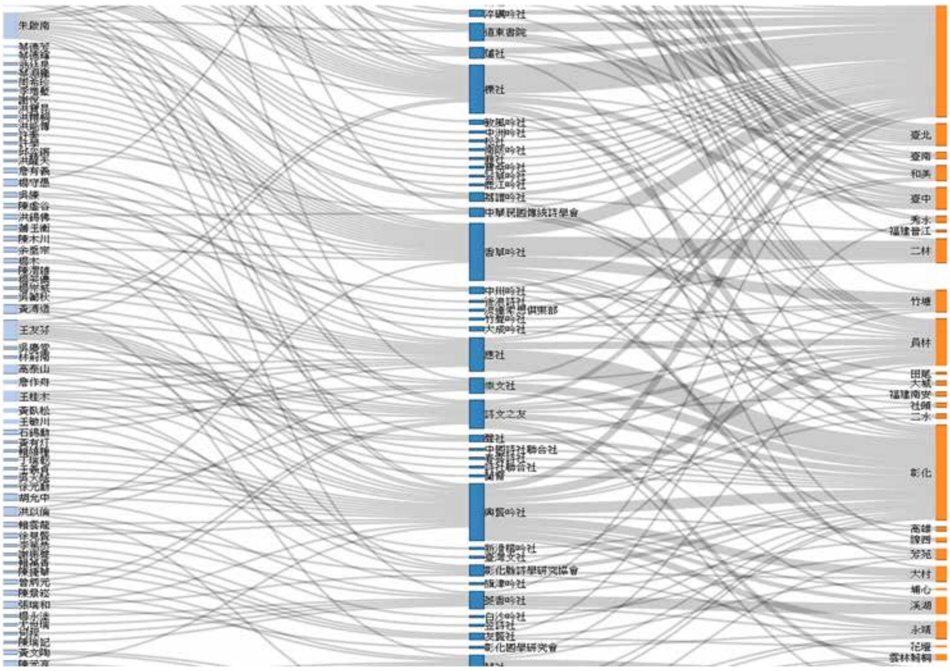


Fig 2. An example of poetry societies and members' relationship including historical figure names (left), poetry society names (middle), and place names (right)

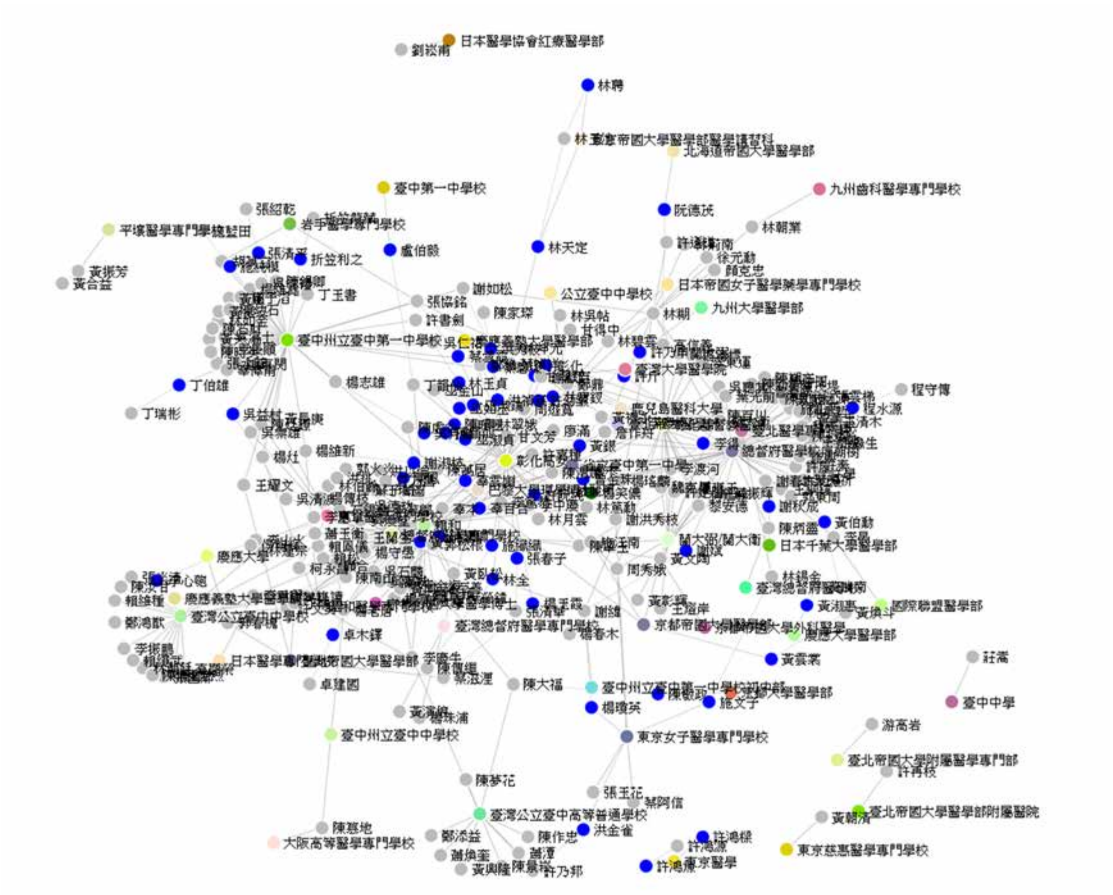


Fig 3. Combining more properties into single social network

More interestingly, when we combine the social networks of poetry society with those of other categories, some hidden relations were revealed. Although none of the major bank shareholders participated in Taiwanese parliament movement directly, many of them participated in cultural associations, many members of which were also members of poetry societies. This suggests that capitalists (major bank shareholders) actually also cared about politics, but in order not to confront with the Japanese government directly, they chose to influence the politics in an indirect way, participation in cultural activities and poetry societies to a certain extent concealing their political engagement. As Fig 3 shows, we try to combine more properties such as education background and organization into a single social network, which might provide more hidden information in it.

Thus, we develop a method to find hidden relationships and make connections with different groups or people automatically, by using extra information with the proposed Automatic Concept Hierarchy Generation algorithm, which is based on hierarchical agglomerative clustering (Jain and Dubes, 1988). As Fig 4 shows, the name marked as red were connected to another one by the same detected information.

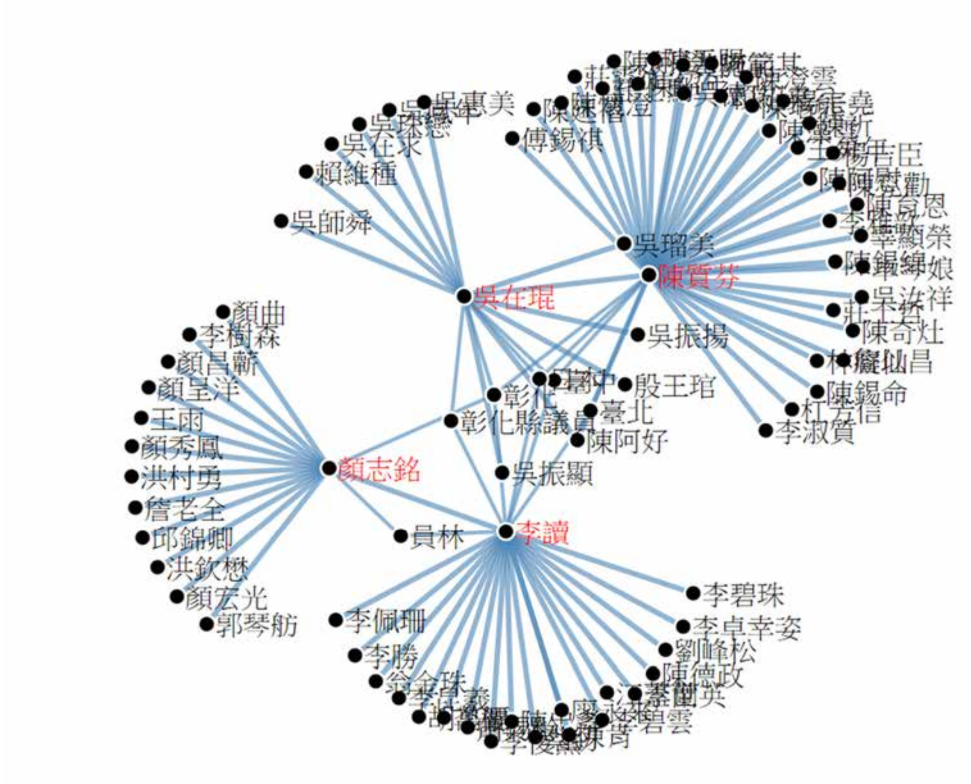


Fig 4. An example of Automatic Clustering in the Construction of Relationships between Historical figures

The above example is clear to demonstrate the importance of analyzing group and poetry-society social networks in modern Taiwanese history. The current challenge lies in continuing to develop the tool for enabling users to choose any combination of categories of social networks and overlap them, and to visualize the resultant social network in a clear and clean format.

In the future, this project will continue to expand the content of the database by adding information extracted from biographies of other gazetteers, dictionaries of historical figures, Taiwanese biographies and diaries. In the process of development, the database will adjust dynamically, such as adding more fields for new attributes, in order to respond to new types and new kinds of data.

Reference

Bol, P.K., Hsiang, J. and Fong, G. (2012). Prosopographical Databases, Text-mining, GIS and System Interoperability for Chinese History and Literature, Proceedings of the 2012 International Conference on Digital Humanities.

Jain, A. K. **and Dubes, R. C.** (1988). *Algorithms for clustering data*.Prentice Hall.

Liu, C. L., Huang, C. K., Wang, H. and Bol, P. K. (2015) Toward Algorithmic Discovery of Biographical Information in Local Gazetteers of Ancient China, In 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29), Shanghai, China, October 30 – November 1, 2015.

Sie, S. H., Ke, H. R. and Chang, S. B. (2017, November). Development of a Text Retrieval and Mining System for Taiwanese Historical People. In *Pacific Neighborhood Consortium Annual Conference and Joint Meetings* (PNC), 2017 (pp. 56-62).

Thou and *You* in Emily Dickinson’s Poems Using Topic Modeling: Reconsideration of Interjections

Miki Okabe[1]

1. Introduction

The present study attempts to reveal interrelationship between the meaning of poems written by Emily Dickinson and second person pronouns (2PP) which appear in her works.

American poet Emily Dickinson (1830-86) mainly uses traditional themes in her poems: NATURE, LOVE, DEATH, ETERNITY, and GOD. More than half of her poems use “*I*”, and the stories go on as *I* narrate. With regard to 2PP in her poems, the frequencies of two pronouns are almost comparable[2]: she uses “*thou*”, which has almost decayed in Modern English, as well as “*you*”, which is generally used nowadays.

You		*Thou*
address to social superiors	← - →	address to social inferiors
address to social equals: upper ranks	← - - - - - - →	address to social equals: lower ranks
address in public	← - →	address in private
formal or neutral address	← - →	familiar or intimate address
respect, admiration	← - →	contempt, scorn

Figure 1 The main uses of *thou* and *you* during the medieval period (adapted from Wales 1983:116)

In Old English, old *thou* meant singular, and *you* was used as plural pronoun. This distinction began to disappear in the late Middle English, when people used *you* as reverential or polite address and *thou* as familiar or contempt address (See Figure 1). “*You*-forms increasingly encroached on the territory of the *thou*-forms, ” Wales (1996:75) says, “so that a distinction between ‘polite’ and ‘familiar’ usage […] came gradually to be replaced by the turn of the sixteenth/seventeenth centuries by […]

[1] Graduate School of Language and Culture, Osaka University

[2] 229 poems use *you*-family (total frequency of pronouns is 545) and 193 poems use *thou*-family (444).

‘unmarked’ (*you*) versus ‘marked’ (*thou*).” *Thou*-forms remain in conventional poetic and dramatic address with an ‘elevated’ stylistic function, and also in ‘liturgical’ discourses, except some regional dialects.

Considering the diachronic transformation of 2PP, it is noteworthy that a lot of *thou* are found in Dickinson’s poems written in 19th century. Then, how Dickinson uses *you* and *thou* separately? Is there any interrelationship with poetic motifs? In order to answer the questions, the present study applies topic modeling to Dickinson corpus with a view to analyzing what kind of words make up topics featuring 2PP.

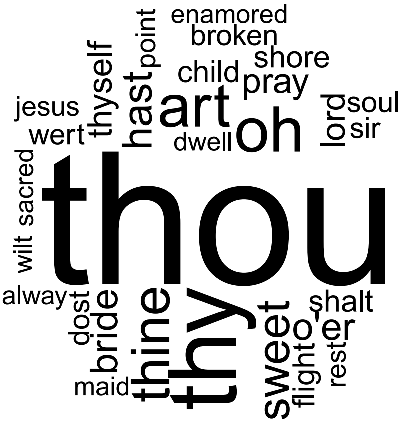
2. Applying topic modeling to the Dickinson corpus

Topic modeling, which is one of machine learning methods, helps discover hidden topics from a corpus of text. Here, Topic means word clusters which are likely to co-occur across texts, in other words, subject or category of document. The method is applied to various literary data, such as novels (Jockers & Mimno 2013), dramas (Schöch 2017), and poetry (Navarro-Colorado 2018).

In the present study, LDA (Latent Dirichlet Allocation) topic modeling (Blei et al. 2003) has been applied to the Dickinson corpus using MALLET[3] after pre-processing as follows; (1) all the 1785 poems are labeled as D1 to D1785 in the chronological order of their production years[4]; (2) set the number of topics as 40; (3) filter by stop-words including function words instead of lemmatization. Concerning (2), the appropriate number of topics varies depending on the size of a corpus as well as the granularity of the analysis. Therefore, several tests have been done using various amounts of topics (20, 40, 60, 80, and 100), and it is concluded that 40 is the best among them to get coherent topics. Moreover, the unique point of the present study is, with reference to (3), that 2PP are not added to stop-words deliberately, although pronouns are usually excluded in topic modeling experiments.

[3] A Java-based toolkit for machine learning applications to text.

[4] The corpus is made from Johnson T. H. (ed.) (1960). *The Complete Poems of Emily Dickinson.*



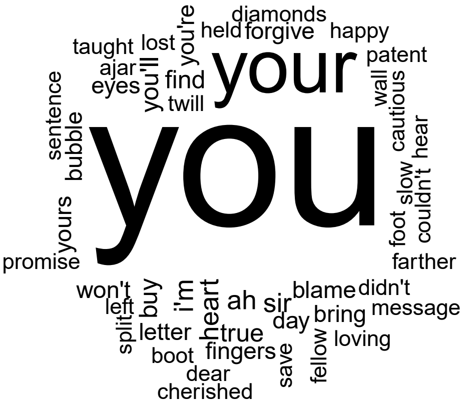


Figure 2 word cloud of topic 23 Figure 3 word cloud of topic 3

As a result, topics numbered 0 to 39 and word clusters composing each topic are obtained. Among the 40 topics, topic 23 features *thou* (Figure 2) and topic 3 highlights *you* (Figure 3). As analyzing topic 23, it is found that there are mainly three characteristics describing *thou* topic; first, archaic wordings are included in word clusters such as *art* (are)*, hast* (have)*, dost* (do), which appear in concord with the archaic pronoun; second, words related to Christianity such as *soul, pray, jesus, lord, sacred,* and *bride* (meaning bride of God) are also main keywords of this topic, and it is understood that *thou* is used in religious contexts; third, *thou* also co-occurs with words as *maid* and *sweet*, where *thou* is employed as an affectionate term of address like medieval period. The third one is surprising a little, because that usage is not typical in modern poetry.

On the other hand, topic 3 is so complicated that it's difficult to interpret. However, two features have been found so far. First, wordings related to correspondence are found such as *dear, letter, message,* and *hear*, indicating that *you* are preferred to be used when poetic motif is letter. Also, there are words of body parts such as *foot, eyes,* and *fingers*, which are used to describe actions of characters in poems.

3. 2PP and interjections

Interestingly, topic 23 (*thou*) and topic 3 (*you*) include different interjections, *oh* and *ah* respectively. Table 1 shows the number of four patterns of poems. According to the table, the number of poems including *thou* & *oh* are much more than *thou* & *ah* poems, although there is only slight difference between *you* & *ah* poems and *you* & *oh* poems.

	You	*Thou*
Ah	12	6
Oh	9	25

Table 1 number of poems including *ah* and *oh*

The reason why specific interjection relates with a pronoun might be due to a theme of a poem. For instance, as previously noted, main key words of *thou* topic are archaic and religious wordings, which give poems solemn impression. In those poems, *thou*, who *I* (a narrator) talks to, is described as unfamiliar person or *I* thinks *thou* is beyond reach of *me*. Then, the poems tend to depict the narrator's surprise at such existence by using *oh* rather than *ah*, because *ah* suggests that received information has connected with information in a speaker's mind, resulting in his understanding or remembrance of something.

The impression of minor *thou* & *ah* poems, on the other hand, is quite different from *thou* & *oh* poems. In those poems, *thou* is close to *me* like a lover or a friend. Here, the pronoun is obviously used as familiar *thou,* which is mentioned in Section 2. However, this usage of *thou* seems peripheral. The same explanation could be possible about the relation between *you* and interjections, while it has still ambiguous grounds and needs more research.

References

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

Brown, R. & Gilman, A. (1960). The Pronouns of Power and Solidarity, Sebeok, T. A. (ed.), *Style in Language*, 253-276.

Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics* 41.6: 750–769.

Johnson, T. H. (Ed.) (1960). *The Complete Poems of Emily Dickinson*. Little Brown and Company.

Navarro-Colorado, B. (2018). On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry, *Frontiers in Digital Humanities*. Aveilable at https://www.frontiersin.org/articles/10.3389/fdigh.2018.00015/full (Access date 2019-6-25)

Schöch, C. (2017). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama, *Digital Humanities Quarterly,* 11.2. Aveilable at http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html (Access date 2019-6-25)

Wales, K. (1983). Thou and you in Early Modern English: Brown and Gilman re-considered*, Studia Linguistica* 37, 107-25.

———— (1996). *Personal Pronouns in Present-day English*. Cambridge.

———— (2004). Second Person Pronouns in Contemporary English: The End of a Story or Just the Beginning?, *Franco-British Studies* 33, 172-85.

Walker, T (2007). *Thou and You in Early Modern English Dialogues*. John Benjam