Proceedings
of the

2019 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)

Regionality and Digital Humanities: South-South Connections

15 – 18 October 2019
Nanyang Technological University, Singapore

Editors
Prof. Jieh Hsiang & Prof. Michael Stanley-Baker

IEEE Catalog Number: CFP19M10-ART
ISBN: 9789869531726

2019 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)

IEEE Catalog Number: CFP19M10-ART
ISBN: 9789869531726

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com

Produced by IEEE eXpress Conference Publishing
For information on producing a conference proceedings and receiving an estimate, contact conferencepublishing@ieee.org
http://www.ieee.org/conferencepublishing

Pacific Neighborhood Consortium (PNC)

The mission of the Pacific Neighborhood Consortium (PNC) is to facilitate information exchanges among institutions of higher education in the Pacific Rim through computing and communications technology. PNC explores issues of information and technology exchange, interdisciplinary collaboration, and the development of the cultural knowledge contents. In fostering access to digitized data on the Pacific Rim, the PNC serves as a portal for access to digital research. It helps scholars to find the library, archive, and museum materials needed to support both teaching and research. The ultimate goal is to enable scholars to regard themselves, not as separated by vast distances, but as residents of a virtual neighborhood.

Established in 1993, the consortium was first initiated by the University of California, Berkeley in partnership with academic institutions in the Pacific Rim. Since 1997, the administrative operations of PNC have been transferred to the Academia Sinica in Taiwan.

Nanyang Technological University (NTU)

A research-intensive public university, Nanyang Technological University, Singapore (NTU Singapore) has 33,000 undergraduate and postgraduate students in the Engineering, Business, Science, Humanities, Arts, & Social Sciences, and Graduate colleges. It also has a medical school, the Lee Kong Chian School of Medicine, set up jointly with Imperial College London.

NTU is also home to world-class autonomous institutes – the National Institute of Education, S Rajaratnam School of International Studies, Earth Observatory of Singapore, and Singapore Centre for Environmental Life Sciences Engineering – and various leading research centres such as the Nanyang Environment & Water Research Institute (NEWRI) and Energy Research Institute @ NTU (ERI@N).

Ranked 11th in the world, NTU has been placed the world's top young university for the past six years. The University's main campus is frequently listed among the Top 15 most beautiful university campuses in the world and it has 57 Green Mark-certified (equivalent to LEED-certified) building projects, of which 95% are certified Green Mark Platinum. Apart from its main campus, NTU also has a campus in Singapore's healthcare district.

For more information, visit www.ntu.edu.sg

Welcome Message from Chair of Pacific Neighborhood Consortium

It is our greatest pleasure to welcome you to the 26th Pacific Neighborhood Consortium (PNC) Annual Conference and Joint Meetings. This year's event is jointly organized by Nanyang Technological University (NTU) and Academia Sinica. We also receive immense support from co-hosts and sponsors including Electronic Cultural Atlas Initiative (ECAI), IEEE Singapore Section, Ministry of Education, Taiwan and Singapore University of Technology and Design (SUTD). On behalf of Academia Sinica and PNC, I would like to express our sincere gratitude to all the sponsors and partners for their enthusiasm and support.

This will be the first time the PNC conference is held in Singapore. We are honored to hold the event at Nanyang Technological University, a leading global research university and is home to several world-class autonomous institutes. The conference brings together scholars and researchers from fields including information technology, humanities, and social sciences, to discuss the main theme of Regionality and Digital Humanities: South-South Connections. The conference will explore on a range of topics, such as heritage and conservation, emergent digital cultures, digital arts and literary studies, historical geography, and biodiversity.

The scientific program consists of thought-provoking keynote speeches, paper presentations, workshops as well as a poster session and demo. I would also like to encourage all participants to take part in the excursion and lunch banquet to explore the cultural diversity and the exotic cuisine of Singapore. Last but not least, we would like to thank the Local Organizing Committee and NTU staff for their dedication and efforts in organizing the event.

Thank you for joining us and I wish you all a stimulating and enjoyable time at the PNC 2019.

Professor Chin-shing Huang
Chair, Pacific Neighborhood Consortium
Vice President, Academia Sinica

Welcome Message from Professor Subra Suresh, NTU President

Established in 1991, NTU has roots that go back to 1981 when its predecessor institution, Nanyang Technological Institute, was set up on the grounds of the former Nanyang University as a teaching university. Today, NTU is recognised as one of the top global research universities for its impactful research, education and innovation. Over the years, NTU has also established deep collaborations and partnerships with more than 500 academic, industry and research institutions across the United States, Europe, the Asia-Pacific and beyond.

NTU is pleased to host this year's Pacific Neighbourhood Consortium conference to provide a platform for researchers from institutions of higher education in the Pacific Rim to discuss issues related to information and technology exchange, interdisciplinary collaborations, and the development of the cultural knowledge content.

Digital Humanities is fast spreading at tertiary institutions around the world as a powerful medium to bring together nuanced humanistic inquiry with large-scale statistical analysis. It offers a bridge between the humanities and the sciences, and between the quantitative and qualitative methods of inquiry. Now, more than ever, the disciplinary scalpel of the humanistic disciplines of history, philosophy, literature, and cultural studies must be brought to bear on the rapidly emerging field of data the sciences have made available to government and the private sector. The fields comprising the disciplinary cluster we now think of as the humanities have paved the way for the re-engineering of our public institutions to meet the standards of a modern society and a global citizenry. The invention of the printing press in 16th century Germany, the organization of the modern bureaucracy in the 17th century, French Enlightenment impetus to transition the hospital from place to die to a social locus for recovery of health in the 19th century, and the guiding principles and practical educational policies that inform 21st century multiculturalism—each of these institutional revolutions was generated by humanistic scholars set on mobilizing technology for the greater good of fellow citizens.

We strongly support the Digital Humanities at NTU, and are proud to partner with Academia Sinica in welcoming the annual Pacific Neighbourhood Consortium conference to Singapore. The Pacific Neighbourhood Consortium has had a long history since its founding in Berkeley in the 1990's, and through its transition to one of the premiere Sinological institutes in East Asia, the Institute of History and Philology at Academia Sinica, Taipei. Many prestigious institutions of higher education and research have hosted this conference in the past. We are proud to join their ranks in hosting researchers to come to Singapore and to our campus, to exchange knowledge and foster collaboration, and to share with you the great work being done in Digital Humanities at NTU. This conference marks the entry of NTU Digital Humanities onto the world stage as one of the leading centres and a resource for Digital Humanities in the South East Asian region.

I understand that in addition to the traditional keynote, paper presentations and exhibitions, there will also be hands-on workshops where scholars can exchange tools and techniques, as well as visits to the National Library's Rare Book Collection, Botanic Gardens, and the Asian Civilisations Museum. I am sure that the conference will be a stimulating and enriching experience, and I wish all of you a successful meeting and a pleasant stay in Singapore.

Professor Subra Suresh
President
Distinguished University Professor

The PNC 2019 Annual Conference and Joint Meetings is jointly hosted by Nanyang Technological University (NTU) and PNC, co-hosted and sponsored by Academia Sinica, Electronic Cultural Atlas Initiative (ECAI), Ministry of Education, Taiwan and Singapore University of Technology and Design (SUTD). PNC 2019 is technically sponsored by the IEEE Singapore Section. Proceedings will be submitted to the Xplore Digital Library.

Hosts

Technical Sponsor

Co-hosts and Sponsors

Scientific Program

Opening Ceremony & Plenary Sessions
09:00-12:00, October 15, 2019, Auditorium

09:00-10:00	Opening Ceremony / Group Photo
10:00-10:30	Coffee Break
Moderator:	William Chong Eng Keat, Research Associate, Nanyang Technological University
10:30-11:15	***A Cross-Discipliners Approach to Social Artificial Intelligence*** Vanessa Evers, Director, NTU Institute of Science and Technology for Humanity (NISTH), Nanyang Technological University.
11:15-12:00	***Technology Enhanced Adaptive Learning in Taiwan*** Bor-Chen Kuo, Director, Department of Information and Technology Education, Ministry of Education, Taiwan

Tracking Voices, Concepts and Knowledge Genres
13:30-15:00, October 15, 2019, Auditorium

Moderator:	Andrea Nanetti, Nanyang Technological University
13:30-14:00	***Materia Medica in Chinese Religious Sources: Towards a Critical Digital Philology for Modelling Knowledge Distribution in Early Chinese Texts*** Michael Stanley-Baker[1], William Eng Keat Chong[1], [1]Nanyang Technological University
14:00-14:30	***Women's Voices within the Dialectics of Control and Democracy in Taiwan's Cyberculture*** Wai Fong Cheang[1], [1]Chang Gung University
14:30-15:00	***A Comparison of a Concept 'Civilization' between Modern Korea and China: Based on the Methodology of Digital Humanities*** Jae Hak Do[1], Injae Song[1], [1]Hallym University

Digital Education
13:30-15:00, October 15, 2019, Lecture Room 1

Moderator: Peter X. Zhou, University of California, Berkeley

13:30-14:00 ***Using a Knowledge-Structure-Based Online Adaptive Learning System to Support Science Teachers' Inquiry and STEM Teaching: The Adaptive E-learning for Science Project (AESP) in Taiwan***
Ying-Tien Wu[1], [1]National Central University

14:00-14:30 ***Adding a Peer Agent? The Effectiveness of a Dialogue Based Intelligent Tutoring System in Learning Mathematics***
Huey-Min Wu[1], Bor-Chen Kuo[2], [1]National Taiwan Ocean University, [2]Ministry of Education

14:30-15:00 ***The Curriculum Development for Global AGILE Problem-Based Learning in Social Entrepreneurship in Global Teams***
Tosh Yamamoto[1], Christopher Pang[2], Benson Ong[2], Juling Shih[3], Hui-Chun Chu[4], Meilun Shih[5], [1]Kansai University, [2]Nanyang Polytechnic University, [3]National University of Tainan, [4]Soochow University, [5]National Taiwan University

ECAI Workshop: Heritage Conservation: from Taiwan to India
13:30-15:00, October 15, 2019, Lecture Room 2

Moderator: Tyng-Ruey Chuang, Academia Sinica

13:30-14:00 ***Connecting the Dots: Indo-Pacific Trade Networks***
David Blundell[1], [1]National Chengchi University

14:00-14:30 ***Redocumenting Wolfgang Franke's "Chinese Epigraphic Materials": Assessing Factors in the Loss of Chinese Cultural Heritage in Malaysia and Thailand***
Oliver Streiter[1], Yoann Goudin[1], Syan-Fei Shi[1], Mandy Manwai To[1], Hanna Yaqing Zhan[1], [1]National University of Kaohsiung

14:30-15:00 ***Ecological Temple Museum: Preserving Heritage and Enhancing Welfare of Local Communities***
Gauthama Prabhu[1], [1]Foundation of His Sacred Majesty

Computing Coding and Methods for Mining and Organising Data 15:30-17:00, October 15, 2019, Auditorium	
Moderator:	Cally Cheung Hiu Tung, Nanyang Technological University
15:30-15:50	***An Innovative Dynamic Visual Model for Digital Content Archive in Traditional Cultural Heritage – Evidence from Stone Weirs Fisheries Culture in Penghu, Taiwan*** Ju Chuan Wu[1], Jui Chi Wang[1], Shu Mei Lee[1], [1]Feng Chia University
15:50-16:10	***Coding to Decipher Linear A*** Niki Cassandra Eu Min[1], Duo Duo Xu[1], Francesco Perono Cacciafoco[1], [1]Nanyang Technological University
16:10-16:30	***Generating Derivational Relations for the Japanese WordNet: The Case of Agentive Nouns*** Francis Bond[1], Ryan Lim Dao Wei[1], [1]Nanyang Technological University
16:30-16:50	***A Python Library for Deep Linguistic Resources*** Michael Wayne Goodman[1], [1]Nanyang Technological University
16:50-17:00	***Q&A***

Applying DH Methods to Taiwanese Local Archives 15:30-17:00, October 15, 2019, Lecture Room 1	
Moderator:	Jieh Hsiang, National Taiwan University
15:30-16:00	***Making a Temporal-Spatial Cadastral Database: A Case Study of Taiwanese Plains Indigenous Peoples' Land Right*** Chung-Hsin Li[1], Hsiung-Ming Liao[2], [1]National Changhua University of Education, [2]Academia Sinica
16:00-16:30	***Exploring Guangxu-era Dan-Hsin Archives and Chinese Recorder through DocuSky*** I-Mei Hung[1], Chi-jui Hu[1], [1]National Taiwan University
16:30-17:00	***An HGIS Platform of Taiwanese Land Deeds*** Jieh Hsiang[1], Chih-Yang Huang[1], [1]National Taiwan University

ECAI Workshop: Community Mapping Projects
15:30-17:00, October 15, 2019, Lecture Room 2

Moderator: David Blundell, National Chengchi University

15:30-15:50 ***Communal Data Projects and Sustainable Digital Preservation***
Tyng-Ruey Chuang[1], [1]Academia Sinica

15:50-16:10 ***Participatory Mapping for Community-based Cultural Resources***
Jihn-Fa (Andy) Jan [1], [1]National Chengchi University

16:10-16:30 ***Early California in Time and Space***
Jeanette Zerneke[1], [1]ECAI, University of California, Berkeley

16:30-16:50 ***Sodeisha Avant-grade Ceramic: High Fidelity Digitization, Virtual Reality, and Interaction Design***
Zi Siang See[1], [1]University of Newcastle

16:50-17:00 ***Q&A***

Scientific Program

Keynote & Plenary Session
09:00-11:00, October 16, 2019, Auditorium

Moderator: Jieh Hsiang, Distinguished Professor, Department of Computer Science and Information Engineering, National Taiwan University

09:00-09:45 ***Digital Humanities and the Study of Singaporean Cultural History***
Kenneth Dean, Raffles Professor of Humanities and Head, Chinese Studies Department, National University of Singapore

09:45-10:30 ***Building Sustainable Society by Using Information and Communication Technology: Lessons from Disaster Management***
Yue-Gau Chen, Distinguished Research Fellow, Research Center for Environmental Changes, Academia Sinica

10:30-11:00 **Coffee Break**

Digital Scholarship I: Evolving the Role of Library for Digital Scholarship
11:00-12:30, October 16, 2019, Auditorium

Moderator: Schubert Foo, Nanyang Technological University

11:00-11:30 ***Academic Librarians and Scholarly Publishers: Partners in Digital Scholarship for National Evidence-based Research Performance***
Abrizah Abdullah[1], [1]University of Malaya

11:30-12:00 ***Taking a Journey on Research Data Management in Singapore***
Pin Pin Yeo[1], Danping Dong[1], [1]Singapore Management University

12:00-12:30 ***Using AI to Analyze Humanities Research Trends in Chinese and Taiwan Studies***
Shu-Hsien Tseng[1], Wen-de Huang[1], Jane Liau[1], [1]National Central Library

Spatialising Heritage and Community
11:00-12:30, October 16, 2019, Lecture Room 1

Moderator: Hedren Sum, Nanyang Technological University

11:00-11:30 ***Heritage Visualisation and Speculative Reconstruction: Using Digital Space to Facilitate Academic Work in an 'Unrecognised State'***
Michael Walsh[1], [1]Nanyang Technological University

11:30-12:00 ***City Stories: Mapping the Spatial Narratives of Singapore's Landscapes***
Kristy Kang[1], [1]Nanyang Technological University

12:00-12:30 ***Detection and Time Series Variation of Latent Topic from Diary in Northern and Southern Courts Period of Japan***
Taizo Yamada[1], Satoshi Inoue [1], [1]The University of Tokyo

ECAI Workshop: Early Buddhism
11:00-12:30, October 16, 2019, Lecture Room 2

Moderator: Alex Amies, Independent Scholar

11:00-11:30 ***Digitalization of Buddhist Sites***
Dayalan Duraiswamy[1], [1]Digitalization of Buddhist Sites

11:30-12:00 ***How much Buddhist was Bengal during the Pāla Dynasty Rule?***
Alexander Stolyarov[1], [1]Russian State University for the Humanities

12:00-12:30 ***Social Networks of Early Buddhism***
Margaret Meloni[1], [1]University of the West

Workshop on Spatiotemporal Knowledge
Part 1: Review of Studies about Sharing Spatiotemporal Information Resources
14:00-16:00, October 16, 2019, Lecture Room 1

Moderator: Shoichiro Hara, Kyoto University

14:00-14:47 ***Theme 1: Gazetteers and Maps***

- ***Possibility of Applying Historical Gazetteer to Knowledgebase of Japanese History***
  Makoto Goto[1], [1]National Museum of Japanese History
- ***Place Names and Ethnic Background in the Middle Basin of the Mekong River***
  Yoshikatsu Nagata[1], [1]Osaka City University
- ***Implementation of Reconciliation API for a Linked Data Gazetteer***
  Akihiro Kameda[1], [1]Kyoto University
- ***Datasets from Maps: Extraction and Remix***
  Tyng-Ruey Chuang[1], [1]Academia Sinica
- ***Q&A***

Moderator: Hsiung-Ming Liao, Academia Sinica

14:48-15:23 ***Theme 2: Temporal Information***

- ***HuTime Ontology as an Extension of OWL-Time***
  Tatsuki Sekino[1], [1]International Research Center for Japanese Studies
- ***Practice of Islamic Calendar in Malaysia***
  Nur Nafhatun Md Shariff[1], [1]Universiti Teknologi MARA
- ***Temporal Change Of Personal Name Based On Pre-Modern Japanese Historical Materials***
  Taizo Yamada[1], [1]The University of Tokyo
- ***Q&A***

Moderator: Tatsuki Sekino, International Research Center for Japanese Studies

15:24-16:00 ***Theme 3: Applications***

- ***Textual Analysis of Ancient Chinese Buddhist Canon with Spatiotemporal Framework***
  Howie Lan[1], [1]University of California, Berkeley
- ***The Cyber Infrastructure for GeoHumanities Studies***
  Hsiung-Ming Liao[1], [1]Academia Sinica
- ***Multi-Resource Data Integration of the Contemporary Museum***
  Wen-Cheng Shih[1], [1]National Museum of Taiwan History
- ***Q&A***

ECAI Workshop: History and Understanding of Buddhism
14:00-15:30, October 16, 2019, Lecture Room 2

Moderator: Margaret Meloni, University of the West

14:00-14:30 ***Demythologization Effort in New Buddhist Movements: The Trúc Lâm Thiền (Chan/Zen) Sect in Late 20th Century Vietnam***
Laura Loan Nguyen[1], [1]Independent Scholar

14:30-15:00 ***Buddhist against Empire***
Marju Broder[1], [1]Estonian Nyingma, Australia

15:00-15:30 ***My Understanding of Buddhism in Pictures***
Vello Vaartnou[1], [1]Estonian Nyingma, Australia

Digital Scholarship II: How Academic Organization Supports Digital Scholarship? Challenges and Opportunities

16:00-17:30, October 16, 2019, Auditorium

Moderator: Hsi-Yuan Chen, Academia Sinica

16:00-16:20 ***Exploiting Emerging Digital Scholarship Opportunities at Nanyang Technological University, Singapore***

Yew Boon Chia[1], Schubert Foo[1], [1]Nanyang Technological University

16:20-16:40 ***A Study on Exploring the Digital Scholarship Services in Taiwan's Academic Libraries***

Hao-Ren Ke[1], [1]National Taiwan Normal University

16:40-17:00 ***Exploring Service Concept for Digital Scholarship Support***

Xin S. Li[1], [1]Cornell University

17:00-17:20 ***The Landscape of Digital Scholarship in Academia Sinica: Projects and Collections Advancing Digital Humanities Research and Open Outreach***

Shu-Jiun Chen[1], [1]Academia Sinica

17:20-17:30 ***Q&A***

Workshop on Spatiotemporal Knowledge

Part 2: Current Status and Issues of Sharing Spatiotemporal Information Resources

16:00-18:00, October 16, 2019, Lecture Room 1

16:00-16:26 ***Reports of Current Status About Sharing Spatiotemporal Information Resources***

Hsiung-Ming Liao[1], Shoichiro Hara[2], and Tatsuki Sekino[3], [1]Academia Sinica, [2]Kyoto University [3]International Research Center for Japanese Studies

16:27-18:00 ***Discussion***

ECAI Workshop: Digital Buddhist Text Projects 16:00-17:30, October 16, 2019, Lecture Room 2	
Moderator:	Lewis Lancaster, University of California, Berkeley
16:00-16:20	***Identifying Keywords in the Buddhist Canon*** Alex Amies[1], Yashuo Deng[2], [1]Independent Scholar, [2]University of the West
16:20-16:40	***Building a Collaborative Translation with Technology*** Miao Guang[1], [1]Fo Guang Dictionary Translation Project
16:40-17:00	***Designing Knowledge Bases for Chinese Buddhist Textual Studies and Research*** You Zai[1], [1]Fo Guang Dictionary Translation Project
17:00-17:20	***Visualizing Bibliographies: Exploring Academic Resources through the NTI Visualizer*** William Chong[1] [1]Nan Tien Institute
17:20-17:30	***Q&A***

Scientific Program

Plenary Session
09:00-10:15, October 17, 2019, Auditorium

Moderator: Chang-Hung Chou, Academician, Academia Sinica

09:00-09:45 ***Agricultural Biodiversity Challenge and Agro-Ecology for Climate Change Resilience and Food Security in ASEAN***
The Anh Dao, Vice President of Vietnam Academy of Agricultural Sciences

09:45-10:15 **Coffee Break**

Texts in Space: Place Names and Epigraphy
10:15-11:45, October 17, 2019, Auditorium

Moderator: Sayan Bhattacharyya, Singapore University of Technology and Design

10:15-10:45 ***Combination of TEI and Python in Studies of Chinese Epigraphy in Singapore***
Duoduo Xu[1], Francis Charles Bond[2], [1]National University of Singapore, [2]Nanyang Technological University

10:45-11:15 ***Community Level Old Place Names in the Northeast of Thailand for a Historical Digital Gazetteer***
Yoshikatsu Nagata[1], [1]Osaka City University

11:15-11:45 ***Building a Corpus of Representations of China in English-language Novels, 1927-2007***
Graham John Matthews[1], Cally Cheung Hiu Tung[1], [1]Nanyang Technological University

Biodiversity in Agroecosystem and Sustainable Agriculture
10:15-11:45, October 17, 2019, Lecture Room 1

Moderator: Chang-Hung Chou, Academia Sinica

10:15-10:35 ***The Native and Cultivated Phalaenopsis Species from Biogeography to Horticulture***
Xiao-Lei Jin[1], Chi-Chu Tsai[2], Yu-Chung Chiang[1], [1]National Sun Yat-sen University, [2] Kaohsiung District Agricultural Research and Extension Station

10:35-10:55 ***Analysis of Transcriptome of Kogia Sima***
Hao-Ven Wang[1], [1]National Cheng Kung University

10:55-11:15 ***Land System Science and Socioecological Systems: the Case of Land-use and Land-cover Change Processes in Sumatra, Indonesia***
Janice Ser Huay Lee[1], [1]Nanyang Technology University

11:15-11:35 ***Silicon Improve Plant Growth and Increase Disease Resistance of Pitaya***
Siti Nordahliawate Mohamed Sidique[1], Xiao-Lei Jin[2], Hawa Masratul Mohd[3], Yu-Chung Chiang[2], [1]Universiti Malaysia Terengganu, [2]National Sun Yat-sen University, [3]Universiti Sains Malaysia

11:35-11:45 ***Q&A***

ECAI Workshop: ECAI Community Projects
10:15-11:45, October 17, 2019, Lecture Room 2

Moderator: Jeanette Zerneke, ECAI, University of California, Berkeley

10:15-10:45 ***Technological Approaches Applied in Recent Buddhist Text Projects***
Howie Lan[1], [1]ECAI, University of California, Berkeley

10:45-11:15 ***Radio Geographies (Aporias and Borderlands Studies)***
Linus Lancaster[1], [1]University of Plymouth

11:15-11:45 ***Atlas of Maritime Buddhism: A Report***
Lewis Lancaster[1], [1]ECAI, University of California, Berkeley

Closing Ceremony & Lunch Banquet
12:45-14:00, October 17, 2019, Bollywood Veggies

Scientific Program

Workshop
October 18, 2019, School of Humanities and Social Sciences (HSS)

Introduction to Digital Humanities
09:00-15:00, October 18, 2019, HSS-B1-08

Facilitator: Sayan Bhattacharyya, Lecturer in the Humanities, Singapore University of Technology and Design

Abstract: This workshop is an introduction to the principles and methods of digital humanities, an emerging interdisciplinary field that seeks to apply digital and computational methods to investigations in the humanities. The course will focus on how computational methods and tools can contribute to the interpretive activities that typically constitute humanistic inquiry, such as digital archive creation and organization and reading and interpretation of textual material and spatial data. Attendees will be made aware of the history of this young field, and will learn about the leading methodological issues and emerging epistemological challenges involved in digital humanities. No prior knowledge or prerequisites is required.

Introduction to Network Analysis for the Digital Humanities
09:00-15:00, October 18, 2019, HSS-B1-07

Facilitator: Miguel Escobar Varela, Assistant Professor, Faculty of Arts and Social Sciences, National University of Singapore.

Abstract: In this seminar we will introduce participants to several network-theoretical measurements and explain how they can be used in the digital humanities. We will explain how the open source software Gephi can be used to calculate many of these measurements and present an overview of DH projects that have used network analysis. Finally we will walk participants through one of our projects, a network analysis of traditional Indonesian stories (as used in a form of traditional theatre). The data and visualization for this project can be seen here.

Introduction to DocuSky
09:00-12:00, October 18, 2019, HSS-B1-12

Facilitator: William Chong, Nanyang Technological University and Chi-Jui Hu, National Taiwan University

Abstracts: DocuSky is a digital humanities academic research platform created to meet the research needs of humanities in organising and analysing research materials. Developed by National Taiwan University Digital Humanities Research Center, the platform provides an array of tools such as cloud database, GIS integration and database construction for the exploration and analysis of text-based data.

Searching for Patterns with Regular Expressions
13:00-16:00, October 18, 2019, HSS-B1-12

Facilitator: Michael Wayne Goodman, Post-doctoral Research Fellow, Digital Humanities Research Cluster, Nanyang Technological University

Abstract: Have you ever spent far too long scouring over your data to find patterns and examples, when basic search is just not flexible enough? Regular Expressions (regexes) can help, and are an indispensable tool for anyone who works with text data. They allow you to fine-tune your searches so you get back only the results you want. In this half-day tutorial you will learn what regexes are (and are not) good for, how to construct basic patterns, how to create flexible patterns, how to perform substitution, and what tools are available. At the end of the tutorial there will be a small regex-crafting workshop, so bring a laptop and your text data!

Poster, Demo & Artwork

Keynote Speaker

10:30-11:15, October 15

Prof. Vanessa Evers

Director, Institute of Science and Technology for Humanity (NISTH), Nanyang Technological University

Vanessa Evers is a chair and Professor of Human Media Interaction, University of Twente, the Netherlands. She is also vice-dean of research for the Faculty of Electrical Engineering, mathematics and computers Science and the Scientific Director and founder of the DesignLab, a centre for multidisciplinary projects with societal impact based on 'Science to Design for Society'.

Vanessa has studied Information Systems at the University of Amsterdam, Business Information Science at UNSW, Sydney and has a PhD from the Open University UK. She has worked for the Boston Consulting Group has been a visiting Scholar at Stanford University, and a part-time professorship at the University of Delft.

Her work exists at the intersection of Computer Science, Psychology, Design, Philosophy and Electrical Engineering and focusses on human interaction with artificially intelligent systems and cultural aspects of Human Computer Interaction. It covers design of Artificially Intelligent systems that are able to interpret human social behaviours and respond to people in a socially acceptable way as well as the evaluation of the impact of such technology on people and society. She is a frequent public speaker in the media and at international fora such as the World Economic Forum at Davos.

A Cross-Discipliners Approach to Social Artificial Intelligence

Prof. Vanessa Evers
Director, Institute of Science and Technology for Humanity (NISTH),
Nanyang Technological University

The current expectation is that artificially intelligent systems such as robots or personal voice agents will be integrated into every aspect of our lives be it home-life, work, leisure, care or education. To ensure that this process happens in a responsible and seamless way I pose the theory that robots must be able to learn socially from people. I will argue that social norms, embedded in people and the context of use must be taken into account when designing artificially intelligent technology and must be interpreted automatically.

Keynote Speaker

11:15-12:00, October 15

Prof. Bor-Chen Kuo

Director, Department of Information and Technology Education, Ministry of Education, Taiwan
Distinguished Professor, Graduate Institute of Educational Information and Measurement,
National Taichung University of Education, Taiwan

Professor Bor-Chen Kuo received the Ph.D. degree in electrical and computer engineering from the Purdue University, West Lafayette, IN, in 2001. He is currently a Distinguished Professor of the Graduate Institute of Educational Information and Measurement, National Taichung University of Education, Taiwan. Since 2015 He served as the President of Chinese Association of Psychological Testing. He was the Chief Editor of the Journal of Educational Measurement and Statistics (TSSCI), Taiwan from 2014 to 2017 and the Associate Editor of Educational Psychology (SSCI) since 2016. From 2019, Professor Kuo served as the Director of department of information and technology education, Ministry of Education, Taiwan.

Professor Kuo received an Outstanding Research Award from Ministry of Science and Technology, Taiwan in 2018 and an Outstanding and Excellence Research Award from The R.O.C Education and Research Society in 2009. His research interests include computerized adaptive learning and testing, cognitive diagnostic modeling, machine learning, and artificial intelligence in education.

Technology Enhanced Adaptive Learning in Taiwan

Prof. Bor-Chen Kuo
Director, Department of Information and Technology Education,
Ministry of Education, Taiwan
Distinguished Professor, Graduate Institute of Educational Information and
Measurement, National Taichung University of Education, Taiwan

In 2017, the Ministry of Education, Taiwan, launched an adaptive learning and assessment platform called Taiwan Competence based Adaptive Learning System (TCALS) (http://adaptive-learning.moe.edu.tw/). TCALS was developed based on knowledge structure with prerequisite relationship and uses information technology to deliver customized resources and orchestrate learning activities in order to adapt to the unique learning needs of individual learners. In TCALS, there are thousands of videos for micro-learning, items for instant diagnosing, interactive modules, intelligent tutoring agents and assessments of 21st century competences for supporting learning through scaffolds.

TCALS has been widely used in many primary and secondary schools, such as ICT instruction, flipped classroom, remedial instruction, and self-regulated learning. Research evidences in support of the effectiveness of the system for adaptive teaching and learning will be demonstrated. Additionally, data mining techniques were applied to analyze student's learning behaviors and records and the findings will be also presented.

Keynote Speaker

09:00-09:45, October 16

Prof. Kenneth Dean

Raffles Professor of Humanities and Head, Chinese Studies Department, National University of Singapore

Professor Kenneth Dean is Raffles Professor of Humanities and Head, Chinese Studies Department, National University of Singapore. He is the research cluster leader for Religion and Globalisation at the Asia Research Institute, NUS. His recent publications include *Secularism in South, East, and Southeast Asia*, NY: Palgrave, (2018) co-edited with Peter van der Veer, and *Chinese Epigraphy of Singapore: 1819-1911* (2 vols.), Singapore: NUS Press (2017), co-edited with Dr. Hue Guan Thye. He directed *Bored in Heaven: a film about ritual sensation* (2010*)*, on celebrations around Chinese New Year in Putian, Fujian, China. Other publications include *Ritual Alliances of the Putian Plain,* Leiden: Brill, 2010 (with Zheng Zhenman); *Lord of the Three in One: The spread of a cult in Southeast China,* Princeton: 1998; and *Taoist Ritual and Popular Cults of Southeast China*, Princeton 1993: *The Absolute State and the Body of the Despot*, NY: Autonomedia 1992 (with Brian Massumi) He has published nine volumes of stone inscriptions gathered in Fujian and Southeast Asia. His current project is the construction of an interactive, multi-media Singapore Historical GIS (SHGIS) and Singapore Biographical Database (SBDB) database. These projects can be viewed online at: shgis.nus.edu.sg and sbdb.nus.edu.sg

Digital Humanities and the Study of Singaporean Cultural History

Prof. Kenneth Dean
Raffles Professor of Humanities and Head, Chinese Studies Department,
National University of Singapore

The Chinese Studies Department of NUS has developed a Singapore Historical GIS (SHGIS) as well as a Singapore Biographical Database (SBDB). This talk will outline the features of these two on-line databases and discuss ongoing efforts to expand and link these databases. One set of new data we hope to add in is a TEI marked up version of Chinese Epigraphy of Singapore: 1819-1911, which provides the names of over 50,000 individuals who contributed to over 70 temples and huiguan. Another data set is the digitized version of the Bukit Brown Burial Record (1922-1972) with the names of over 68000 individuals. The talk will also discuss the possibility of developing a common digital platform for the Chinese diaspora across Southeast Asia.

Keynote Speaker

09:45-10:30, October 16

Prof. Yue-Gau Chen

Distinguished Research Fellow, Research Center for Environmental Changes, Academia Sinica

Executive Secretary, Center for Sustainability Science, Academia Sinica

Dr. Yue-Gau Chen is currently appointed as Distinguished Research Fellow and Executive Secretary at Research Center for Environmental Change and Center for Sustainability Science, respectively in Academia Sinica. He has also a co-appointment Professor in Department of Geosciences, National Taiwan University. He has long experience working on Environment Changes and Earthquake Geology; thus is familiar with scientific techniques on Stable Isotope Analysis, Quaternary Chronology, Morpho-tectonics, and Fault Kinematics. He was elected as fellow of Geological Society of America in 2008. He devoted himself to serve as Section President in Asia Oceania Geoscience Society from 2010 to 2012. In the meantime, he was also invited as the Executive Secretary of the Program on Applying Science and Technology for Disaster Reduction, National Science Council, Taiwan ROC. From 2012 to 2016, he spent his 4-year secondment as Director General of Department of Natural Science and Sustainability Development, Ministry of Science and Technology, Taiwan ROC, coordinating domestic as well as international scientific research programs. He therefore had a chance to serve in Steering Committee of Belmont Forum from 2016 to 2017. Recently he is still working for Belmont Forum to lead one of the Collaboration Research Actions (CRAs): Disaster Risk Reduction and Resilience (DR3).

Building Sustainable Society by Using Information and Communication Technology: Lessons from Disaster Management

Prof. Yue-Gau Chen

Distinguished Research Fellow,
Research Center for Environmental Change, Academia Sinica
Executive Secretary, Center for Sustainability Science, Academia Sinica

The modern society has become multi-functional to support individuals therein surviving and developing altogether. Given good and efficient functions, a society may expand beyond million individuals as large as cities and countries. To keep such a gigantic system running is a complicated daily course for the management level. The entire system is actually composed of numerous sub-systems, which rely on each other at some points. Anything goes wrong in one sub-system would affect the normal function of not only the system itself, but also other systems related. Such a dependence is commonly recognized as the risk in maintaining our society to be sustainable. It is also considered as the vulnerability when we assess the impacts of disaster's strikes. The overall societal performance actually depends on the vulnerabilities of all components that constitutes the society. Therefore to strengthen the societal resilience has to take all components into account. In terms of hazard themselves on the other hand, different natural hazard is caused by its own physics happening above or beneath Earth's surface. For the purpose of hazard mitigation and prevention of specific hazard, it requires relevant knowledge to lower the societal vulnerability for all subsystems including the central governance system. All related approaches for the action can be defined to societal resilience building.

One of above actions, of common in applications on all kinds of hazards, is the information distribution via efficient communication ways. It is of unequivocal essential in a modern society in addition to basic life-supporting systems. Some of the communication technologies, i.e., internet and cellphone, have provided us no-blind-spot and non-stop service, fulfilling high-end demand of well-developed society. For hazard managements they are also of critical. In particular during a disaster happening time, rapid and efficient information distribution can significantly strengthen the resilience and save possible loss. In the past, though we had radio or television, the hazard information were mainly collected by oral reporting. Nowadays imagery or video monitoring systems have become increasingly common than before. Such a big-data input needs auto-managing system to diagnose and analyze before reporting to the decision maker and then to the public. The broadcasting to the public has been also rapidly evolved and mainstreamed by new medias, i.e., internet webpage and social media platform. They can play even better role than traditional tools on information distribution while considering the effectiveness. Moreover, internet is a two-way communicating platform. Instantaneous information searching and grabbing can be utilized as a new input source, which helps quite a lot for information updates. Since the disasters are inevitable to a society, resilience building for all societal components is necessary, especially by using new information and communication technology.

Keynote Speaker

09:00-09:45, October 17

Dr. The Anh Dao

Vice President, Vietnam Academy of Agricultural Sciences

Dao The Anh is the Vice-President of Vietnamese Academy of Agricultural Sciences (VAAS) since 2017 in charge of international cooperation, post-graduation and policy communication. He was former Director of the Centre for Agrarian System Research and Development (CASRAD) and DDG of Field Crops Research Institute (FCRI) since 2006. With a first degree in Agronomy in Hanoi Agricultural University (1990), a master (1994) in Farming systems, a PhD (2004) in Agricultural, Rural and Food Economics at Montpellier SupAgro, France, he has developed an experience of 29 years in research and developement related to agricultural economics and farming system in Vietnam.

He stated his research in family farming diversification in Red river delta. Recently, he focuses on agro-ecology farming systems, adaptation and mitigation for climate change of production systems, livelihood diversification and biodiversity use such as Geographical Indications and other community brand names, food safety management and certification, cooperative and farmer organization development, branding for agricultural products, post-harvest loss, agri-food value chain governance for smallholder in Vietnam. Through the research and development works, he also involved actively in the capacity building, advocacy for small farmers and policy level in Vietnam with IFAD and in other countries in Africa and South America. He also has taken a sharp interest in the role of safe agriculture and Zero Hunger initiative based on improving the position of small farms in sustainable food system. He has been the Family farming focal point in Vietnam for the year of Family farming 2014, initiated by FAO.

He has authored and co-authored more than 30 peer-reviewed papers in journals and Congresses and managed many national and international research projects in agricultural economics. Through the research and development works, he also involved actively in the policy advocacy for provinces, for MARD and higher level. More recently, he has taken a sharp interest in the role of safe agriculture and in poverty alleviation strategies based on improving poor and small food producer access to markets and agro-ecological products. His latest work was realized with FAO, IFAD, WB, ADB, CIRAD, OXFAM, IRD, ACIAR, IFPRI, JICA, DFATD, SNV, GRET, VECO, CIAT, IRRI, CIP, CCAFS, GRIPS and different International Universities…

Recently, he has involved also in civil society works as Vice-President of Vietnam Science for Rural Development Association (PHANO) and Vice Editor in Chief of Vietnam Journal of Sciences for Rural Development since 2012. In 2019, he is recently nominated as Head of Vietnam Agriculture and Hydraulic Encyclopedia Edition Board. In 2018 he has received an Agriculture Merit Medal Grade Chevalier of French government, for his contribution of 25 years agricultural research.

Agriculture Biodiversity Challenge and Agro-Ecology for Climate Change Resilience and Food Security in ASEAN

Dr. The Anh Dao
Vice President, Vietnam Academy of Agricultural Sciences

The Southeast Asian comprises found to be one of the 25 global biodiversity hotspots which is the highest biodiversity centers in the world. All of 10 member states are Parties to the Convention on Biological Diversity (CBD). The Agro-biodiversity is expressed through diversity in ecosystems, species composition and genetic resources as well. Agro-biodiversity makes a significant contribution to each country economy, providing a basis for ensuring food security, maintaining genetic resources of animals and plants; and providing materials for construction, fuel and pharmaceutical resources.

Agro-biodiversity has changed significantly in ASEAN countries. Through agricultural intensification, many drivers affecting these changes including changes in land and water use and management, pollution and external inputs, over-exploitation and overharvesting, climate change, natural disasters, pests, diseases, alien invasive species, markets, trade, policies, population growth and urbanization, changing economic, socio-political, and cultural factors, advancements and innovations in science and technology. Of these, the pressure from the increasing human population combined with an increasing level of consumption which is resulting in overexploitation of biodiversity resources. In addition, land conversion and infrastructure construction has significantly reduced the area of natural habitats, increased ecosystem fragmentation, and degraded the habitats of many species of wild plants and animals. Natural resources, especially biological resources, are undergoing overexploitation. Of these, timber, non-timber and aquatic products are particularly vulnerable. Furthermore, alien species, environment pollution and climate change are all directly affecting the biodiversity. The agricultural intensification has created a lot of negative externalities in Vietnam and ASEAN.

There remain some challenges in achieving good management of biodiversity, including: (1) lack of effective inter-sectoral coordination mechanisms to respond to overlap in functions among relevant ministries and agencies; laws and regulations to protect biodiversity are still unsystematic and lacking in uniformity; (2) community involvement in biodiversity conservation has not been sufficiently mobilized, which leads to weak law enforcement; deforestation and illegal wildlife trade pose serious threats to biodiversity; (3) overall investments in biodiversity are insufficient, resulting in a lack of financial, human and technological resources. In the long run, Vietnam and other ASEAN countries have to find a balance between protectionism and sensible access to their national agriculture biodiversity to tackle challenges in biodiversity conservation, health issues, food security and climate change. In order to find the solution, the conventional agricultural intensification approach should be changed to agro-ecology, with ecological intensification principle, creating more positive externalities through eco-system services. The agro-ecology will find a trade-off between producing biomass for food security and protecting biodiversity for a climate resilience in our future.

PNC 2019 Conference Committee

Local Organizing Committee

Francis Bond - Nanyang Technological University
Andrea Nanetti - Nanyang Technological University
Michael Stanley-Baker - Nanyang Technological University
Tan Choon Keng - Nanyang Technological University
Miguel Escobar Varela - National University of Singapore
Sayan Bhattacharyya - Singapore University of Technology and Design

Program Review Committee (in alphabetical order)

Jieh Hsiang (Co-Chair) - National Taiwan University, Taiwan
Michael Stanley-Baker (Co-Chair) - Nanyang Technological University, Singapore
Sayan Bhattacharyya - Singapore University of Technology and Design, Singapore
Ta-Chien Chan - Academia Sinica, Taiwan
Su-bing Chang - National Taiwan Normal University, Taiwan
Shu-Jiun Chen - Academia Sinica, Taiwan
Shoichiro Hara - Kyoto University, Japan
Hsiung-Ming Liao - Academia Sinica, Taiwan
Jyi-Shane Liu - National Chengchi University, Taiwan
Andrea Nanetti - Nanyang Technological University, Singapore
Tatsuki Sekino - Research Institute for Humanities and Nature, Japan
Chuen-Tsai Sun - National Chiao Tung University, Taiwan
Miguel Escobar Varela - National University of Singapore, Singapore
Hedren Sum Wai Yuan - Nanyang Technological University, Singapore
Peter X. Zhou - University of California, Berkeley, USA

PNC Organizing Committee (in alphabetical order)

Ching-Ray Chang - National Taiwan University, Taiwan
Lih-Shyang Chen - National Cheng Kung University, Taiwan
Ling-Jyh Chen - Academia Sinica, Taiwan
Sophy Shu-Jiun Chen - Academia Sinica, Taiwan
Chang-Hung Chou - Academia Sinica, Taiwan
I-Chun Fan - Academia Sinica, Taiwan
Jieh Hsiang - National Taiwan University, Taiwan
Kuo-Hsing Hsieh - Academia Sinica, Taiwan

Chin-Shing Huang - Academia Sinica, Taiwan
Bor-Chen Kuo - Ministry of Education, Taiwan
Der-Tsai Lee - Academia Sinica, Taiwan
Feng-Tyan Lin - National Cheng Kung University, Taiwan
Fu-Shih Lin - Academia Sinica, Taiwan
Simon C. Lin - Academia Sinica, Taiwan
Cheng-Yun Liu - Academia Sinica, Taiwan
Ming-Yan Shieh - National Taiwan University, Taiwan
Shu-Hsien Tseng - National Central Library, Taiwan
Chung-Li Wu - Academia Sinica, Taiwan
Mi-cha Wu - National Palace Museum, Taiwan

PNC Steering Committee (in alphabetical order)

Ling-Jyh Chen - Academia Sinica, Taiwan
Royol Chitradon - Hydro and Agro Informatics Institute (HAII), Thailand
Shoichiro Hara - Kyoto University, Japan
Chin-Shing Huang - Academia Sinica, Taiwan
Lewis Lancaster - University of California, Berkeley, USA
Der-Tsai Lee - Academia Sinica, Taiwan
Hsiao-Ti Li - City University of Hong Kong
Chao-Han Liu - Academia Sinica, Taiwan
Tatsuki Sekino - Research Institute for Humanities and Nature, Japan
Hyun Seung Yang - Korea Advanced Institute of Science and Technology (KAIST), Korea
Jidong Yang - Stanford University, USA
Peter Zhou - University of California, Berkeley, USA

Secretariats

Pacific Neighborhood Consortium (PNC)
Catherine Liang / Peggy Yu

Nanyang Technological University (NTU)
Simin Leng

Table of Contents

Materia Medica in Chinese Religious Sources: Towards A Critical Digital Philology for modelling Knowledge Distribution in Early Chinese texts

Dr. Michael Stanley-Baker
History, Nanyang Technological University, Singapore
msb@ntu.edu.sg
orcid.org/0000-0001-6785-850

William Chong Eng Keat
History, Nanyang Technological University Singapore,
william.chong@ntu.edu.sg

Abstract— **The circulation of medical knowledge extends far beyond recognized "medical" fields in all cultures, but can be difficult to trace outside of canonical sources. This paper models the distribution of medical knowledge in Buddhist, Daoist and medical source from the early medieval period (Six Dynasties--up to the year 589) by charting the distribution of materia medica terminology. Surviving Shangqing Daoist records appear to contain the most materia medica vocabulary, and indicate the relatively higher use of drugs within the sect.**

The paper argues that such search results and visualisations constitute "qualified historical hypotheses" about the distribution of medical knowledge across religious sects in the period, and the relative saturation of medical knowledge in each genre of writing and sect. It details how search results for the terms are grounded in philologically rich meta-data about the textual corpora. The paper describes how, using post-search classification, a key function of DocuSky's TermStatsTools, how search results were produced using the enriched metadata, and then further refined to generate a readable graph and reliable statistics. These refinements are the "qualifications" of the hypothesis, and the analysis of how they impact the search results constitute what the paper refers to as a "critical digital philology."

The digital visualisations of DocuSky outputs allows researchers to model the distribution of knowledge and generate new research questions. Greater accuracy can be produced through semi-automatic markup and synonymy tables, i.e. name authority databases. The authors describe their current processes for producing a synonymy for Chinese materia medica, to account for the use of alternate names for specific materia medica. The relevant open access databases, metadata, term lists and name authority tables for the study of medicine across Daoist Buddhist and medical sources that the authors have published are all hyper-linked from the paper. The hypotheses of the paper can thus be reproduced, and the paper constitutes a full working model and datasets for the development of similar projects.

Keywords— DocuSky, visualisation, Chinese Medicine, Chinese Religion, authority tables, database, Digital Sinology

I. Pre-Amble: Categorial Problems with Religion and Medicine

In the study of medical history, there is a problem that different periods of time have different ideas of what "medicine" is and should consist of. When modern historians go to read early texts, they tacitly bring these assumptions about medicine to bear on what they read. This does not only affect the interpretation of texts, but also what gets included into the study and what does not. This is therefore not just a problem of interpreting the materials, but the very question of scope. When you exclude certain materials from the inquiry, this also excludes whole ranges of intellectual questions.

For example, the history of Chinese medicine has habitually been written about "doctors" *yi* 醫 and what they wrote and did. The same term *yi* stands both for intellectual and practical discipline of medicine and for the social role of the doctor. Thereby entire genealogies of doctors and of texts are produced, which centre around the delivery of, mainly, drugs, acupuncture, moxibustion, and then exercise, diet and massage. Many, many scholars have used this model to write about the changes in these genealogies across time, and what medicine looked like in different periods.

However, from the perspective of the medical marketplace, there were many more practitioners offering a wide variety of therapies. In fact, during the Six Dynasties period (220-589), the historical record preserves many *more* texts about healing in religious collections—the Daoist (*Zhengtong daozang* 正統道藏) and Buddhist canons (Taishō Shinshū Daizōkyō 大正新脩大藏經—than in so-called "secular" medical texts.[2] [3] All accounts point to religious actors as being more prolific than elite, scholarly doctors. Thus, it is worthwhile investigating these sources to understand the medicine of the time.

It is when setting out on this research that fundamental problems arise. While there are now catalogues and fundamental studies that investigate the religious nature of these collections, their medical contents are for the main part unexplored. Religious scholars describe the materials according to that discipline—paying attention to sectarian formation, eschatology, ritual, philosophical and theological debates, hagiography. To come to an accounting of the canon from a medical history perspective means re-indexing the canons according to medical historical criteria – pharmaceuticals, disease terms, treatment regimes, lineages of

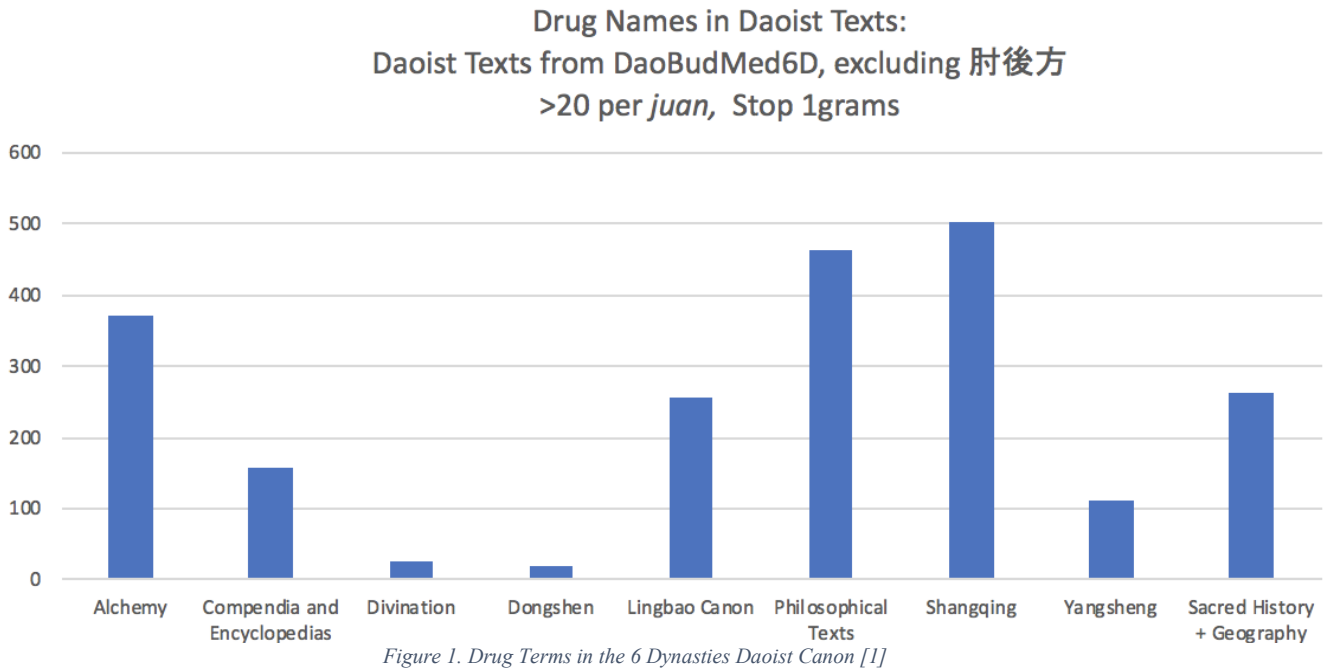


Figure 1. Drug Terms in the 6 Dynasties Daoist Canon [1]

practice, case histories. Again, we come to the problem of not just interpreting and describing contents, but of scope. Scholars may disagree about what is to be included in such an index. For example, a team of Japanese scholars produced a print index to the *Taisho* canon, in which they include a section on medical vocabulary. [4] While they indexed drugs and treatments, disease notions and what they referred to as "hygiene" (衛生), because of their unconscious modern bias which excludes "religious" practices, they excluded many treatments types that would have been primary treatments for Six Dynasties Buddhist practitioners: intercessory prayer, talismans, incantations and spells, rituals, meditation, and philosophical reflection on suffering and/or death. Yet later editors include precisely such elements in their sampling of Buddhist medicine. [5] A recent printed edition of so-called "Complete Writings on Buddhist medicine" [6] which followed tacit modernist principles similar to the editors of the *Taisho* index, have come under withering criticism for their lack of self-reflection on the scope of the medicine included. [7]

In Daoist studies, the situation is even more dire. Since the publication of an eponymous dissertation from Chengdu University on topic of "Daoist Medicine" in 2001, a wave of articles, books and dissertations claiming to represent this idealized typology has come into print. [8] These had the effect of propelling the category into canonical status as an historical reality, such that in 2015, the Daoist Association of China promised to print a collection of works on "Daoist medicine," which never came to fruition. [9] The basic framework of Gai Jianmin's study, and all those following in his wake, suffers from critical failures of category analysis, and constructs historical continuities where there were in fact deep differences, contradictions and at times conflicts. Stanley-Baker has argued elsewhere that the diversity of epistemological positions inherent in different forms of therapy calls for more critical historical reflection. [10, 11]

The question that faces the current study of medicine in early imperial China then, is how to respond to fluidity in the scope of what counts as medicine, and how to organize and account for it in the religious canons? This paper documents the development and use of new digital tools which make advances towards answering this question. Taking advantage of digital editions of the canons, and the ability to perform term-specific searches in sophisticated ways, this research offers new ways to entirely re-index the Buddhist and Daoist canons, according to researcher-driven questions. The method experimented with here was how to track the presence and possibly movement of materia medica across different genres of writing, and by implication, across different practice communities. Although the experiment is based on drugs, this approach can be re-used to search for and schematize any kind of knowledge which can be represented identified by a representative set of terms.x

The publication of open-access, high-quality digital transcriptions of pre-modern Chinese text collections has dramatically transformed the landscape of Digital Sinology. High quality editions of the Daoist Canon, the Buddhist Canon and the major literary collections *Siku quanshu* (四庫全書) and *Sibu congkan* (四部叢刊) as well as the dynastic histories can now be downloaded en masse and compiled into datasets for corpus-level analysis.[1]

These and other proprietary corpora have been complemented with the production of new digital tools at major

[1] See Kanseki Repository Catalog [12]. Their entire corpus is backed up on GitHub and can be extracted from there. The Chinese Text Project hosts the world's largest collection of pre-modern Chinese sources [13]. While it does not offer bulk download, there are many tools for corpus analysis within the site itself.

research centres like CHGIS and CBDB at Harvard [14] [15], MARKUS at Leiden [16], CCTS at Academia Sinica[17], DocuSky at National Taiwan University [18] and CBETA as well as the Buddhist Authority Database at Dharma Drum Institute for Liberal Arts [19, 20]. From place name and biographical authority databases, to text markup tools and corpora-wide vocabulary distributions, these services are closely integrated and operate as an interrelated ecology, with extensive coordinated development across the platforms. Constant inter-use and inter-development amongst these clusters ensures they are at the forefront of digital sinology, and will continue to produce new developments in the Sinological field.

DocuSky was developed by Tu Hsieh-chang 杜謝昌, partly during collaboration with Michael Stanley-Baker while he was PI of the Drugs Across Asia project at the Max Planck Institute for History of Science in Berlin, and then at Nanyang Technological University [21]. DocuSky allows users to build their own text databases, attach meta-data to the texts, and use these meta-data to parse searches for large term-sets using a method called "post-search classification" (houfenlei 後分類). This method was initially adopted in the Taiwan National University library, as well as the Taiwan Digital History Library, and now forms one of the primary features of DocuSky [22, 23, 24]. DocuSky also includes facilities for visualising historical GIS data within MARKUS files across historical maps.

Stanley-Baker and Chen Shi-pei compiled a corpus of 3830 chapters or fascicles (in Chinese juan 卷) of almost all surviving Buddhist, Daoist and medical texts up to the year 589, the end of the Six Dynasties period, a time when religious actors were at their most active in the healing arts [25]. The texts used were taken from Kanripo and elsewhere, and compiled into an open access Docuxml database with attached metadata. These are available for searching and analysis at the link in Ref [26]. The metadata for this corpus describes the text titles, canon numbers, genres of writing and sites of origin where known, and can be downloaded for inspection here [27]. The Buddhist metadata are from CBETA, and the Daoist Canon data was composed according to Schipper and Verellen's Daoist Canon [28]. Schipper and Verellen's chapter and section headings are used to structure the "Genre" categories, which perform important work in parsing out the circulation of knowledge, and are the basis for the textual categories in fig. 1.

Using DocuSky's post-search classification tool, called TermStatsTools, it was possible to search for 12,000 drug terms [19] in these sources, and model their distribution across these three corpora and multiple genres. The results of this search can be downloaded and studied from [1]. From these initial results, only chapters with 20 or more unique drug terms were selected, in order to increase probability that vocabulary terms reflected the topic. The higher the number of terms in a chapter, the higher the likelihood that they reflect the topic, i.e. drug names. 20 terms was taken as a baseline indicator for significance.

Using this method revealed that among the corpus, 151 chapters of Buddhist texts, 69 chapters of Daoist texts, and 28 chapters of medical texts contained significant drug data.

These distributions can be easily visualized by using pivot tables in Excel and simple social networking applications like Palladio, or more deeply analysed using Gephi.[29]

II. Modelling Knowledge: Critical Questions

Fig 1. models the frequency of drug terms which appear in Daoist texts that address drugs as a significant topic, and organises them according to different genres of writing. Some of these literary genres implicitly reference practice communities in the period, because the majority of these texts were esoteric and required initiation into a social network before one could possess them. These include Lingbao, Shangqing, Celestial Master, Alchemy, and less restrictively, Divination and Yangsheng. This is not a direct representation of the circulation of practice on the ground, because texts are not communities, and practices circulate in ways more fluid than the textual categories which attempt to capture them. Nevertheless, texts represent attempts by practitioners to control the flow of knowledge, to define orthopraxy, and to authoritatively communicate such knowledge within socially restricted channels. These genres thus form a qualified approximation of the distribution of knowledge.[2] Fig. 1 roughly approximates how much these communities knew about drugs. Further data-analysis can now be performed to discover which drugs circulated in which communities.

III. Modifying Search Results to Derive Readable Images

DocuSky cannot tell the semantic content of any words it finds, thus its results can only be taken as a probability that they refers to the correct meaning. The results need further verification by human readers that they in fact refer to relevant content. This modelling process which produced Fig. 1 was used to identify texts with the highest likelihood of drug term content. Two hundred chapters were then selected as part of the MPIWG Drugs Across Asia Project for semi-automatic markup and confirmation [21]. These marked texts are undergoing further refinement and will be the basis of more refined publications in future.

However, in order to derive Fig. 1, some fine-tuning of the results was required. We describe this here to explain how the graph was produced, and also as a methodological model for how to analyse DocuSky results.

First, Chinese does not use spaces for word segmentation, and thus cannot be used to recognize word-boundaries. Single-character drug names produce a high number of errors, so these were excluded (Stop 1gram).

Second, in order to provide better readability for the graph, only Daoist texts were selected.

Third, as mentioned earlier, chapters with a high concentration of terms (more than 20) were selected.

[2] On the relationship between practice and knowledge see [30] and on situating practice within Daoist knowledge communities see [31].

Fourth, as one text in the collection contains so many drug recipes, it concealed the variation of the other genres. Thus, even though it is contained in the *Daoist Canon*, the *Ge Xianweng Zhouhou Beiji Fang* 葛仙翁肘後備急方 DZ1306 was excluded from the graph.

These modifications were used to produce Fig. 1 as a model of the distribution of drug knowledge in the Six Dynasties period, but is an approximation derived from the above processes.

A pivot table derived from the data can be extremely useful for navigating and studying the data further. This format allows for closer inquiry into the data behind Fig. 1. For example, one can answer such questions as "Why does Philosophy contain so many drug terms?" Opening up the sub layers within the pivot table reveals sub-levels such as such as Genre, Text Title and Vocabulary Terms. Fig 2. shows the pivot table opened up to reveal successive levels within the Philosophy category. It shows that only three texts contain a significant number of drug terms (more than 20 per chapter), and how many terms appear in each text. The *Baopuzi neipian* 抱朴子內篇 DZ1185 is the primary culprit, containing 342 probable terms. The chapter with the second most number of drugs, is the *Huainan honglie jie* 淮南鴻烈解 DZ1184, with 77 terms. Opening another layer reveals which chapters make up this total, and how many terms per chapter. The *Honglie jie* contains two chapters containing 24 and 23 terms each. This is useful not only for identifying the individual drugs, but for understanding the initial graph in fig. 1. We can now see that the genre of philosophy is not concerned with drugs—the high frequency in this genre is because Schipper and Verellen, who finalized the *Daoist Canon* category, included the *Baopuzi* in philosophy—a text which many would argue has more to do with external alchemy than philosophy.

The important take away here is not that the graph is wrong. Schipper and Verellen's motivations for including the *Neipian* in Philosophy, not in Alchemy, no doubt stem from their

⊞ **Lingbao Canon**	**255**
⊟ **Philosophical Texts**	**463**
⊟ -	**463**
⊞ 南華真經	44
⊞ 抱朴子內篇	342
⊟ 淮南鴻烈解	77
⊟ **9**	**23**
女青;白虎;鯉魚;牽牛;苦菜;鹿角;半夏;蟋蟀;木香;火母;露水;鹿皮;鴛鴦;螻蛄;王瓜;栝樓;萎蕤;百舌鳥;含桃;君子;草藥;朝生;大室	23
⊟ **24**	**24**
君子;土龍;救火;百步;茯苓;雞頭;桑葉;五味;鮫魚;紫芝;慈石;將軍;百舌鳥;當道;松脂;女蘿;合歡;屈人;夜光;芳香;甑帶;雞足;無足;蜾蠃	24

Figure 2 Pivot Table of Drugs in Daoist Texts [1]

appreciation of the author Ge Hong's *ru*ist philosophical project in the much larger *Baopuzi waipian* 抱朴子外篇 [28, pp. 71-72], as well as an editorial desire on their part to catalogue both sections of his text together. Assuredly, the *neipian* has much more in common with alchemical than philosophical writings, but every editorial project has its constraints and limitations.

Rather than a simplistic critique of Schipper and Verellen's *Daozang* it is more important to take in the point that critical digital philology requires a keen understanding of how your data is structured, in order to explain curiosities and spot mistakes. Using DocuSky's termstats tools and a pivot table in combination, provides a very close visibility of how the metadata is related to and structures the search results. This transparency, the ease with which results can be analysed and interpreted, affords a philological rigor superior to the hidden algorithms of google-like search tools, unknown Gephi filters, or even topic modelling, which is based on mathematical proximity—not the human interpretive thinking which designed, or later analysed, the canons in the first instance.

These anaylses are not transparent just to the investigator. With digital humanities file-hosting DOI archives, this data can be made accessible to anyone. The entire search output from DaoBudMed6D produced by TermStatsTools, and the associated metadata, can be downloaded from the link at [1] including the pivot table and graph which are the basis of Fig. 1 and Fig. 2. Please download the file titled DaoBudMed6D_12kDrugs.Stop1.CategorizedFile.Result.xlsx and open the worksheet titled Daoist Texts Pivot.

By following this method, researchers can re-index large textual corpora according to the locati ons of any topic that can be modelled by a representative vocabulary, model the distribution of knowledge across time, corpuses and genres, and by extension, community, and critique the data results.

IV. Social Network Analysis for Comparing Text Contents

Further investigation can be done by exporting TermStatsTools results into the Social Network Application called Palladio.[29] DocuSky makes this simple with exports that can by directly uploaded into the site. The network graph function can visualise the degree of vocabulary overlap or distinctness between different genres, and across time. Fig. 3 displays a visualisation from the same base data as fig. 1 & 2, filtered to profile literature produced by the three main Daoist sects between 329-394 CE: Celestial Masters, *Lingbao* and *Shangqing*. This graph includes chapters with less than 20 drug terms per chapter. The three corners on the periphery show clusters of terms unique to the three sects. Smaller clusters situated between two or three of the clusters, linked by small lines, are vocabulary shared between two or three sects' literature. From the visualization, it is clear to see that *Lingbao* and *Shangqing* share much more vocabulary with each other

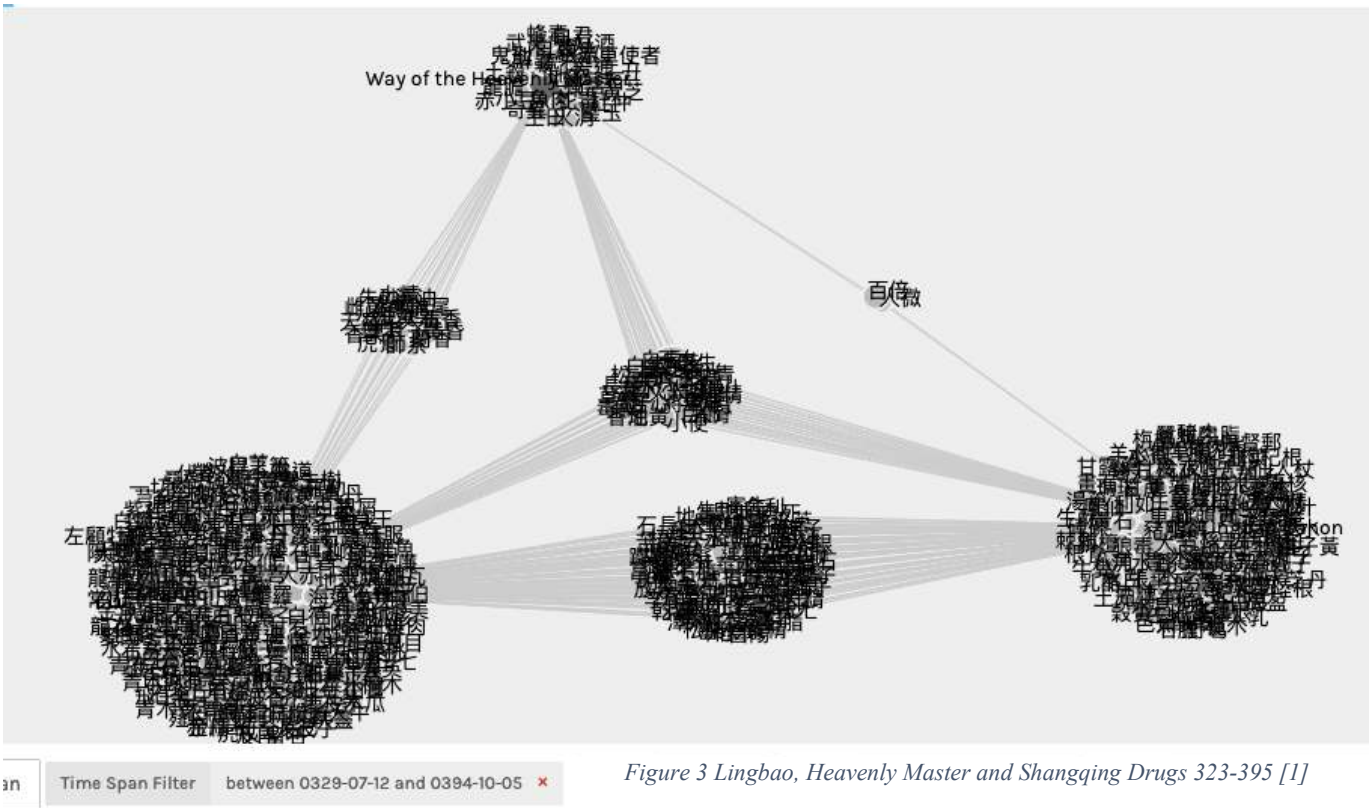


Figure 3 Lingbao, Heavenly Master and Shangqing Drugs 323-395 [1]

than with Celestial Masters, and that Celestial Masters literature shares much more vocabulary with *Shangqing* than *Lingbao*. The dataset and settings for this particular view can be downloaded from the repository linked here, and uploaded to Palladio for study, as well as to for further exploration. [1]

Making these comparisons, enables further questions arise. What are the drugs in the different clusters? Why do the Celestial Masters share more vocabulary with the *Shangqing* sect than with the *Lingbao* sect? What drugs or terms do they share in common? Do they exhibit other similar features – such as common drug pseudonyms, have similar drug properties, or drugs that grow in similar regions? Palladio does not allow extraction of the individual clusters, so further investigation requires going back to the excel file, or more powerful Social Network software, such as Gephi.

While researching how to produce and study these outputs, Stanley-Baker chose 200 chapters with the highest number of of drug terms for markup by a team in Taipei, under supervision of Chang Cha-jan of Fu-jen University and Joey Hong of DILA. These markups are currently undergoing editing at present for detailed analysis. When these are complete, these will allow for a comparison of the results from TermStatsTools modelling of the vocabulary and human markup. Our expectations are that the numbers of vocabulary items will be reduced as false hits are left out, but that the relative frequencies of drug terms will remain consistent. We use MARKUS, a semi-automatic text-marking software which automatically identifies pre-determined vocabulary sets and provides convenient interfaces for human readers to confirm or edit digitally-derived results.[16] It is integrated with some of the most significant Chinese language databases, enabling semi-automatic recognition and tagging of numerous types of vocabulary, such as personal names and historical placenames from biographical and historical databases, as well as traditional Chinese notation for time and date as well as official titles. It also allows markers to mark text automatically with pre-determined vocabulary sets, as well as using intuitive, simplified regex features to identify vocabulary in regular structures in highly structured texts. This toolset is part of a closely networked ecology of platforms and tools being developed at leading institutes for Digital Sinology at Harvard, National Taiwan University, Leiden University and the Max Planck Institute for the History of Science. This powerful institutional clustering through mutual development of these tools ensures that work done in DocuSky and MARKUS is at the core of the Digital Sinology movement and will have longevity that more peripheral platforms will not.

V. Reaching Beyond China: Building a Drug Term Synonymy

This ability to compare materia medica across different genres of writing opens up new possibilities of comparison. There is one problem remaining, which if resolved, may open up avenues to study traditional medicine in multiple languages from around the world. This problem is simply the issue of object recognition – the term search function in DocuSky has no way of telling when two different words in fact refer to the same material object – i.e. when it is capturing different synonyms for drugs. This is important because two different-looking recipes, may on close examination, turn out to be the same.

The solution to this problem begins in early imperial China. Since the 2nd Century CE, pharmacological scholarship has recorded alternate or common names for drugs, which can be based on variant terms, regional terminology or historical change over time. [32, 33, 34] These are now recorded in

modern dictionaries, which also often state the date and text where the name first entered into the record. By developing a synonymy table that tracks these equivalences, and integrating into DocuSky results, we plan to resolve this problem.

However, the synonymy function, which we refer to as a philology engine, offers many more possibilities for tracking equivalence. The function of tracking different terms for the same object in one language is basically the same for tracking it across multiple languages. If we can identify multiple Chinese terms for an object, we can aggregate to that solution terms in other languages. This would then allow us to compare materia medica and recipe texts in multiple languages. By identifying the dates of these texts, and studying when the drugs enter into the vocabulary of different regional medicines, we can come to a large-scale understanding of the migration of traditional medicines across the pre-modern world.

Thus the future phase of research will focus on how to search for common materia medica across different languages: e.g. Chinese, Malay, Sanskrit, Arabic, Persian, Hebrew, Latin, Greek and– languages of major medical traditions across the early and medieval world. In January 2019, Nanyang Technological University hosted a workshop titled *Medicines Across Eurasia* to bring together scholars working on medical history in these languages, with experts in Digital Humanities and computational linguistics. A key question which arose is "To what degree can we track common materia medica across these larger gulfs of language, epistemology and geographical distance?"

We address this question by using the function of the philology engine to extend beyond resolving name variance in Chinese, to studying terms in other languages. We can triangulate the terms through a variety of sources. Firstly, using dictionaries for each of the languages we can collect their listing of formal, common and regional names as well as the scientific botanical names for the plants. Secondly, we can cull ethnonyms from secondary scholarship, such as books, articles and published synonymies. Third we can cull equations made in pre-modern polyglot texts. We are currently partnering with the Medicinal Plant Name Service at the Royal Botanical Gardens at Kew to share data, and to update the botanical taxonomy of our dictionaries to the most recent standards.[35]

To kickstart this process, we first worked with publicly-available versions of four modern Chinese medical dictionaries, extracting fields of information that contained the alternative names listed under the primary drug name. [36, 37, 38, 39] This produced a total of 11241 primary drug names, of which 8533 were associated with a total of 35399 alternative names, which in some cases came with the sources the dictionary cited, and other information regarding the region or language from which the alternative name derives. Fig. 4 shows a sample of the primary drug and alternative names which have been obtained from these sources. These can be downloaded from [40].

No.	Drug Name 摘录	AltName	Remarks
780	斑竹壳	《中药大辞典》	斑竹衣	
782	报春花	《中药大辞典》	橡只玛尔布	（藏名）
783	报春花	《中华本草》	橡只玛尔布	
784	饱饭花	《全国中草药汇编》	米饭花	
784	饱饭花	《全国中草药汇编》	小叶珍珠花	
784	饱饭花	《全国中草药汇编》	乌饭子	
785	饱饭花	《中药大辞典》	米饭花	《四川常用中草药》
785	饱饭花	《中药大辞典》	乌饭子	
786	饱饭花果	《中华本草》	乌饭子	
789	包袱七	《全国中草药汇编》	铁骨散	
789	包袱七	《全国中草药汇编》	小八角莲	
789	包袱七	《全国中草药汇编》	半碗水	
790	包袱七	《中华本草》	铁骨散	
790	包袱七	《中华本草》	半碗水	
790	包袱七	《中华本草》	包袱莲	
790	包袱七	《中华本草》	一块砖	
791	宝盖草	《全国中草药汇编》	接骨草	[昆明]

Figure 4 A Sample of the Drug Name to Alternative Name Dataset [40]

We quickly found a troubling phenomenon: multiple drug names listed the same alternative name. As our objective is to enable users to use any of the names, whether drug or alternative, to search for the whole set of synonyms, having this many-to-one relationship would hamper the search. We drilled into the data and found that about 3000 alternative names were double-linked in this way to more than 4000 drug names. Our temporary solution was to first separate out these linking terms, which left us with 6525 primary drug names (i.e. 76% of 8433) and 24042 alternative names (68% of 35399). There is more work to be done to continue to clean these names, but this is the current state of progress.

To take the first trial step towards a search mechanism which could analyse material in multiple languages, we also captured the botanical names of the drugs in the online dictionaries. We received the kind gift of a digital edition of a critical edition of the twelfth-century CE Italian treatise on women's medicine, the *Trotula,* from the editor Monica Green.[41]. Stripping out the Latin plant names and their botanical attributions from the index, we used these to compare to our Chinese results.

We identified some 300 drugs common to both the modern Chinese dictionaries and the 12th-century C.E. Latin texts by comparing their modern botanical equivalents after some cleaning, anticipating some misspellings. Using this Latin-Chinese drug synonymy, we then compared the late fifth-century materia medica, *Bencao Jing Jizhu* 本草經集注 to the *Trotula*, and found fourteen drugs in common. Fig. 5 shows the extracted botanical names and the drugs they are attributed to,

	Drug Name	Dict Name	摘录	英文名	来源	MedNameFound
4	猕猴梨叶		《中华本草》		为猕猴	Actinidia arguta (Sieb.et Zucc.) Flarich.ex Miq.
5	子楝树根		《中华本草》		为桃金	Decadpermum gracilen-tum (Hance) Merr.et Perry
6	岗梅叶	中华本草	《中华本草》	leaf of Ro	为冬青	Ilex asprella (Hook.f.et Arn.) Champ.ex Benth.
7	鲈鱼	中华本草	《中华本草》	Japanese	为脂杉	Lateolabrax japonicus (Cuvier et Valenciennes)
8	山苍子叶		《中华本草》		为樟科	Litsea cubeba (Lour.) Pers.[Laurus cubeba Lour.]
9	白云瓜叶		《中华本草》		为旋花	Merremiayunnanensis (Church.et Gagnep.) R.c.fang
0	皮袋香		《中华本草》		为木兰	Michelia yunnanensis Franch.exFinet et Gagnep.
1	皮袋香根		《中华本草》		为木兰	Michelia yunnanensis Franch.exFinet et Gagnep.
2	锦香草叶		《中华本草》		为野牡	Phyllagathis cauleriei (Lev L. Et Van.) Guil L.
3	土大黄	中华本草	《中华本草》		为蓼科	Rumex obtusifolius L.[R.madaio auct.non Makino
4	土大黄叶	中华本草	《中华本草》		为蓼科	Rumex obtusifolius L.[R.madaio auct.non Makino
5	通骨消根		《中华本草》		为爵床	Thunbergia grandiflora (Roxb.ex Rottl L.) Roxb.

Figure 5 A Sample of the Latin-Chinese Drug Name Dataset

albeit without extensive regularisation or cleaning.

There is still much work to be done to regularize the spellings of their scientific names, or where historical scientific names have been used, to identify the current ones and to link

them to the International Plant Name Index and the Medicinal Plant Name Services. Our collaborators at Kew Gardens and the Singapore Botanical Gardens have offered much help to start the process.

The work is ongoing, and we look forward to further developments. We are looking forward to to track multiple languages in future expansions of the project, and have submitted applications to fund the coordination work with Kew and a pilot study comparing Malay medical manuscripts to a large corpus of Chinese texts with this method. We feel that the database of early imperial religious and medical texts is robust, and welcome users to try it out. It can be used not simply for searching materia medica, but for any type of knowledge that can be identified through a representative termset. It is our hope that the methodologies described here will assist others in their research on varied topics, and that the data we have produced will be of interest to historians of medicine and religion in China, and be of use to the field at large.

References

[1] M. Stanley-Baker "6D Drug Terms TermStatsTools Results table and Pivot," DR-NTU (Data), 2019 DOI: 10.21979/N9/DMLAW4 https://doi.org/10.21979/N9/DMLAW4

[2] "*Zhengtong daozang* 正統道藏," Shanghai: Shangwu yinshuguan, 1923.

[3] "*Taishō Shinshū Daizōkyō* 大正新脩大藏經," Tokyo: Taisho shinshu daizokyo kanko kai, 1924.

[4] Daizōkyō Gakujutsu Yōgo Kenkyūkai 大藏經學術用語研究會, "*Taishō shinshū Daizōkyō sakuin* 大正新脩大藏經索引," Tōkyō: Taishō Shinshū Daizōkyō Kankōkai, 1962-1990.

[5] C. P. Salguero, *Buddhism and Medicine: An Anthology of Premodern Sources*: Columbia University Press, 2017.

[6] "*Zhongguo fojiao yiyao quanshu* 中国佛教医药全书," Shi Yongxin 释永信 and Li Liangsong 李良松 eds., Beijing, 2011.

[7] D. Burton-Rose, "Desiderata for the Principles of Compilation of a Canon of Buddhism and Medicine," vol. 12, no. 1-2, pp. 203, 2017.

[8] Gai Jianmin 盖建民, "*Daojiao yixue* 道教医学," Beijing: Zongjiao wenhua chubanshe, 2001.

[9] "*Daoyi jicheng* 道医集成," Fan Jiping 范吉平 ed., Beijing: Zhongguo yiyao chubanshe, 2015.

[10] M. Stanley-Baker, "Daoism and Medicine: A review of Gai Jianmin's *Daojiao yixue* 道教醫學 (*Daoist Medicine*)," *Asian Medicine,* vol. 4, no. 1, pp. 249-255, 2009.

[11] M. Stanley-Baker, "Daoing Medicine: Practice Theory for Considering Religion and Medicine in Early Imperial China," *East Asian Science Technology and Medicine*, 2019.

[12] C. Wittern "Kanseki Repostirory [KANRIPO]." http://www.kanripo.org/

[13] D. Sturgeon "Chinese Text Project (Ctext)." http://http://ctext.org/

[14] Fairbank Center for Chinese Studies of Harvard University and the Center for Historical Geographical Studies at Fudan University "CHGIS [China Historical Geographic Information System] Version: 6," 2016. https://sites.fas.harvard.edu/~chgis/data/chgis/v6/.

[15] Harvard University, Academia Sinica, and Peking University "CBDB [China Biographical Database]," January 1, 2018. https://projects.iq.harvard.edu/cbdb

[16] B. Ho Hou-Ieong, and H. De Weerdt. "MARKUS Text Analysis and Reading Platform. ," 2014-, http://dh.chinese-empires.eu/beta/ Funded by the European Research Council and the Digging into Data Challenge.

[17] Academia Sinica. " Chinese Civilization in Time and Space, Version 1," http://ccts.ascc.net/searches.php?lang=en.

[18] Tu Hsieh-chang 杜協昌 "DocuSky," National Taiwan University Research Center for Digital Humanities. http://docusky.org.tw/DocuSky/ds-01.home.html

[19] "Chinese Buddist Electronics Text Association (CBETA)." http://cbeta.org

[20] Huimin Bhikṣu 釋惠敏, Tu Aming 杜正民, M. Bingenheimer, and Hung Jen-Jou 洪振洲. "Buddhist Studies Authority Database Project," https://authority.dila.edu.tw/.

[21] M. Stanley-Baker, Chen Shih-pei, and D. Schäfer. "Drugs Across Asia," https://www.mpiwg-berlin.mpg.de/page/drugs-across-asia. Max Planck Institute for the History of Science.

[22] Chen Kuang-hua 陳光華, and Wu Che-an 吳哲安, "Taiwan daxue jigou diancang xitong zhi jianzhi 台灣大學機構典藏系統之建置," Tushuguan xue yu zixun kexue 圖書館學與資訊科學, vol. 33, no. 2, pp. 33-47, 2007.

[23] Chen Shih-pei 陳詩沛, Tu Hsieh-chang 杜協昌, and Hsiang Jieh 項潔, "Shiliao zhengli fenxi gongju zhi mu hou—jieshao "Taiwan lishi shuwei tushuguan" de ziliao qianzhi chuli chengxu 史料整體分析工具之幕後——介紹「臺灣歷史數位圖書館」的資料前置處理程序," *Cong baocun dao chuangzao: Kaiqi shuwei renwen yanjiu* 從保存到創造: 開啟數位人文研究, Hsiang Jieh 項潔, Chen Yijun 陳怡君 and Cai Jiongmin 蔡炯民, eds., pp. 52-67, Taipei: Guoli Taiwan daxue chuban zhongxin, 2011.

[24] Tu Hsieh-chang 杜協昌, "On the Construction of a DocuSky Personal Database with Markups and Metadata," in Pacific Neihbourhood Consortium, Chengkong University, Tainan, 2017.

[25] M. Stanley-Baker, "Daoists and Doctors: The Role of Medicine in Six Dynasties Shangqing Daoism," PhD Diss., History, University College London, London, 2013.

[26] M. Stanley-Baker, Chen Shi-pei 陳詩沛, and Tu Hsieh-chang 杜協昌 "Daoist Buddhist Medical Texts from the Six Dynasties [DaoBudMed6D]," Research Center for Digital Humanities, National Taiwan University, 2018. DOI: 10.6681/NTURCDH.DB_DocuSkyDaoBudMed6D/Text. http://doi.airiti.com/LandingPage/NTURCDH/10.6681/NTURCDH.DB_DocuSkyDaoBudMed6D/Text

[27] M. 徐. Stanley-Baker, Hong Yimay 洪一梅, Hong Chen-chou 洪振洲, Lim Ding Xun, and Chew Shu Wen "Buddhist and Daoist Canon Metadata," DR-NTU (Data), 2018, Ver. V1 DOI: 10.21979/N9/REE4MJ https://doi.org/10.21979/N9/REE4MJ

[28] K. M. Schipper, and F. Verellen, *The Taoist Canon: a historical companion to the Daozang*, Chicago, Ill.: University of Chicago Press, 2004.

[29] D. Edelstein, N. Coleman, E. Jewett, E. Wells, G. Caviglia, and M. Braude. "Palladio," http://hdlab.stanford.edu/palladio/ Humanities + Design, Stanford University.

[30] J. Lave, and E. Wenger, *Situated learning : legitimate peripheral participation*, Cambridge [England]; New York: Cambridge University Press, 1991.

[31] M. Stanley-Baker, "Drugs, Destiny, and Disease in Medieval China: Situating Knowledge in Context," *Daoism: Religion, History and Society,* vol. 6, pp. 113-156, 2014.

[32] P. U. Unschuld, and Zheng Jinsheng, *Chinese Traditional Healing: The Berlin Collection*, Leiden, Boston: Brill, 2012.

[33] G. Métailié, *Science and Civilisation in China: Biology and Biological Technology—Traditional Botany: An Ethnobotanical Approach* Volume 6 Part 4: Cambridge University Press, 2015.

[34] "*Bencao jing jizhu* 本草經集注," Shang Zhijun 尚志鈞 and Shang Yuansheng 尚元勝 Beijing: Renmin weisheng chubanshe, 1994.

[35] B. Allkin "Medicinal Plant Name Servives," Royal Botanical Gardens at Kew (Ed.), 2019 DOI: https://mpns.science.kew.org/mpns-portal/

[36] Guo jia zhongyiyao guanliju zhonghua bencao bianwei hui 国家中医药管理局中华本草编委会, "*Zhonghua bencao* 中华本草," Shanghai: Shanghai kexue jishu, 1999.

[37] Jiangsu xin yixue yuan 江蘇新醫學院, "*Zhongyao da cidian* 中药大辞典," Shanghai, Hong Kong: Shanghai kexue jishu chubanshe, Shangwu yinshu guan, 1978.

[38] Bianxie zu 编写组, "*Quanguo zhongcaoyao huibian* 全国中草药汇编," Beijing: Renmin weisheng, 1996.

[39] Zhonghua Renmin Gongheguo wei sheng bu yao dian wei yuan hui 中華人民共和國衛生部藥典委員會, "*Zhongguo yao dian : Zhong yao cai se tu ji* 中國藥典 : 中藥彩色圖集 ", Hong Kong, 1991.

[40] W. E. K. Chong, and M. Stanley-Baker " Large List of Chinese Drug Names 40K," Nanyang Technological University, 2019 DOI: 10.21979/N9/V2XBOH https://doi.org/10.21979/N9/V2XBOH

[41] M. H. Green ed., The Trotula: An English Translation of the Medieval Compendium of Women's Medicine: University of Pennsylvania Press, 2002

Women's Voices within the Dialectics of Control and Democracy in Taiwan's Cyberculture

Wai Fong Cheang
Center for General Education
Chang Gung University
Taoyuan, Taiwan
cheangwf@mail.cgu.edu.tw

***Abstract*—The advent of digital technology, computers, Internet and cyberculture has brought forth changes in almost every aspect of material life as well as social life. This paper discusses the dialectics of control and democracy in cyberculture. It focuses on the issue of women's representation and women voices in the Internet. It begins with the optimistic utopian vision of cyberculture in Allvin Toffler's *The Third Wave*. Then it explores the pessimistic dystopian images of cyberculture in popular cultural productions such as movies and fictions. It draws attention to the phenomenon of women's voices reaching out far and wide in cyberculture. By using three contemporary Taiwanese woman Bloggers/YouTubers as examples, the paper reexamines the new possibilities for gender equality and women's empowerment within cyberculture.**

Keywords—cyberculture, control, democracy, gender equality, women's empowerment, Taiwan

I. Introduction: Computers, Changes, Optimism and Doubts

A. Optimism in Allvin Toffler's The Third Wave

The advent of digital technology, computers, Internet and cyberculture has brought forth changes in almost every aspect of material life as well as social life. In one of the best sellers published in 1980, *The Third Wave*, Allvin Toffler presents an optimistic vision of the changes which he terms the Third Wave: "The Third Wave shows us these new potentials…. It shows clearly and, I think, indisputably, that—with intelligence and a modicum of luck—the emergent civilization can be made more sane, sensible, and sustainable, more decent and more democratic than any we have ever known" [1]. The key to the Third Wave is, according to Toffler, the computer, with which we can construct "a new infosphere" [1].

Toffler's optimistic vision about computers, despite its appeal, has met skepticism. In reality, there have always been worries about what computers or their related technologies, such as artificial intelligence, may bring. There are worries that Internet users would be under surveillance and controlled by information fed to them on the Internet.

This paper discusses the dialectics of control and democracy in cyberculture. It focuses on women's representation in the Internet, and women voices, which are an essential part of democracy in the sense that all kinds of people and all kinds of gender should be presented. It discusses the restriction on women's representation and women's voices in general, after which it focuses on situations in Taiwan. It brings into attention three contemporary Taiwanese women's voices in cyberculture that have implications with gender equality and women's empowerment.

B. Doubts about the Third Wave Technology

In a collection of essays, *Computerization and Controversy: Social Conflicts and Social Choices*, Charles Dunlop and Rob King argue that "computerization" may cause problems which "foreshadow social problems of much greater significance", and one of the problems that they have predicted is "large scale unemployment in certain industries" [2]. Undeniably, this prediction about large-scale unemployment has turned out to be true in many places across the world, such as in Taiwan, where highway toll collectors lost their jobs in 2013 when electronic toll collection system was applied [3].

In 2016, news of AlphaGo, a computer go program, defeating Lee Se-dol, a world champion go player, brought further attention to the capability of computers and artificial intelligence [4]. More and more people began to be aware of the fact that computers may be capable of beating human intelligence, and more and more people worry where artificial intelligence would lead us.

C. Sci-fi Extrapolations of Disastrous Technological Development

Science fictions (abbreviated into sci-fis), which draw on what we know about science and technology, time and again play up our worries about computers and digital technology running out of control.

As early as 1984, just four years after Toffler published his optimistic views about computers in *The Third Wave,* a sci-fi movie presents a dystopian nightmare of computers. *The Terminator*, starring Arnold Schwarzenegger, dramatizes a robot with artificial intelligence getting back from the future

world to assassinate a woman, whose unborn son will be the leader of a war to come waged by humans against robots [5]. The movie was a blockbuster, and there were even several sequels starting from 1991. Despite the fact that these sci-fi movies can be readily dismissed as exaggerations, they symbolize our deepest worries about computers.

Interestingly, as early as 1949, George Orwell, an English writer, published a sci-fi titled *1984*, which depicts the surveillance and thought control of a totalitarian government led by a party leader, Big Brother [6]. Technology is what helps Big Brother to watch and control his people. Orwell's emphasis on the pressure of being watched coheres with a French philosopher, Michel Foucault's depiction of a powerful control system, the panopticon. Designed by Jeremy Bentham in the late eighteen century, panopticon is a prison building with inmates allocated in cells surrounding a watchman's place so that regardless of whether they are really being watched, they feel the pressure and are psychological forced to act appropriately [7].

Remarkably, Orwell's notion was appropriated by Apple Computers in a television commercial for its Mac computers released in 1984, which says, "1984 won't be like '1984'" [8]. Though the commercial was meant to suggest that Mac computers were free from the control of Big Brother, which was supposed to be IBM that had a dominant status in the market then, it echoed the fear of surveillance and thought control central in Orwell's famous sci-fi. Consequently, it reinforced the imaginative link between information technology development and Orwell's dystopian vision of technology.

D. Sci-fi Scenarios and Real Life Experiences

It is noteworthy that several high-tech devices depicted in Orwell's sci-fi, such as the speakwrite, a dictation system for converting voice to words on screen, and the telescreen, a bilateral screen serving mainly as television and surveillance camera, have become real in today's world. The telescreen is one of the main sources of psychological pressure for the protagonist. It has foreshadowed our current worries about surveillance by computers and modern technology.

Most of us are not unaware of the fact that search engines/browsers, social media and various apps are collecting data from us. The practice of search engines/browsers feeding us with information about whatever goods we have browsed earlier is a reminder that our digital footprints have been recorded.

The collection of data in the name of "Big Data" has inevitably incited suspicion despite all kinds of justification. "Is Big Data the Next Big Brother?"--to name just one among a host of similar discussions on the Internet, has attempted to address the issue [9]. Even though the main argument in this article is "the benefits [of Big Data] seemed to far outweigh the potential for harm" [9], I would argue that the suspicious link between Big Data and Big Brother remains rather strong for most people.

This suspicion, unfortunately, is constantly re-invoked by the similarity in the names--Big Data and Big Brother, both of which are a compound noun with a total of three syllables starting with the word big. Big Brother is definitely not a term unknown to the computer industry since in 1984, Apple Computers, as aforementioned, has referred to Orwell's 1984 in a television commercial [8]. Hence, had Big Data been named otherwise, its link to Big Brother, I argue, would probably be not as strong as it is now. After all, names do count.

II. CYBER AND ETYMOLOGY

A. The origin of "cyber" in Oxford English Dictionary

The term "cyber", which we use today to mean relating to computers and the Internet, is implicated with control in its etymology.

Oxford English Dictionary (OED), which defines the word "cyber" as an adjective meaning "relating to or characteristic of the culture of computers, information technology, and virtual reality", offers its origin as "1980s abbreviation of cybernetics" [10]. "Cybernetics" in OED is a plural noun meaning "the science of communications and automatic control systems in both machines and living things"; its origin is "1940s from Greek kubernētēs 'steersman', from kubernan 'to steer'" [11]. Thus seen, from its Greek root, "cyber" is connected to the idea of "control".

B. Cybernetics in Encyclopaedia Britannica

In *Encyclopaedia Britannica*, there is no entry for the word "cyber", but there is one for "cybernetics", which is "control theory as it is applied to complex systems" [12]. The origin of the term is from "the ancient Greek word kybernetikos ("good at steering"), referring to the art of the helmsman" [12]. Interestingly, Encyclopaedia Britannica mentions that as early as the first half of the nineteenth century, André-Marie Ampère, a French physicist, already suggested "the still nonexistent science of the control of governments be called cybernetics" [12]. In fact, Norbert Wiener, who is generally acknowledged to be the source of the word "cyber" [12], has named his book from which the word is said to be from, *Cybernetics, or the Control of the Animal and the Machine* [13]. The book's title manifests that control is the key to cybernetics. The control is not just of machines, but also of animals. When one thinks about animals, one cannot deny that human beings belong to that category. Hence, from the title of Wiener's book, we can infer that cybernetics is also the control of humans.

C. Other Explanations of "Cyber"

Harris Breslow and Aris Mousoutzanis wisely observe that "since their emergence, discourses, practices and theorisations

of cybercultures have been accompanied by questions of surveillance and activism, power and resistance, fixity and flow, a dialectic that was embedded in the connotations surrounding the prefix 'cyber-' since its very first use around the middle of the twentieth century" [14].

Therefore, I believe that it is not incorrect to conceptualize "cyber" as related to "control"--not just the control of information communication in the machines, but also of animals, as Wiener's book title indicates, and that includes human beings. The origin of the word "cyber" justifies our current worries about control mechanisms in cyberculture.

III. WOMEN AND CYBUERCULTURE

Most women in today's world, not unlike their male counterparts, engage themselves with computers and cyberculture.

A. Women's Rendezvous with Cyberculture in Taiwan

Nevertheless, women in Taiwan are still behind men in their rendezvous with computers and cyberculture. According to Survey on 2018 Individual/Household Digital Opportunity Survey in Taiwan Executive Summary, the latest survey report published in December 2018 by the National Development Council, an important government unit that can be understood as the helmsman of Taiwan's development--

> The overall online access rate of women still maintains 3.5 percentage points behind males. After entering the online world, women's online experience, digital footprint awareness, or self-assessment of information gathering and judgment ability is not much different from that of men, but overall, males' rate over the access to Internet devices and access to the Internet, and programming learning experience is higher than women [15].

This gender divide, though seemingly small, signifies that there is still room for improvement in Taiwan. Women today, not just in Taiwan, but also around the world, are consciously aware of the fact that throughout history, their gender has always been marginalized and underrepresented. The silencing of women's voices can be considered the best symbolic expression of the marginalization of women.

B. Representation of the Silencing of Women in Cultural Productions

One of the most horrifying images of men's silencing of women is probably the tongue-cutting of Lavinia in William Shakespeare's tragedy *Titus Andronicus*. Daughter to a powerful Roman general Titus, Lavinia is betrothed by her father first to a man and then to another for the father's political interest. Abducted by her father's enemies, Lavinia confronts them. One of her attackers says to her, "Nay, then I'll stop your mouth" [16]. She is raped after which her abductors have her tongue cut out and her hands cut off so that she will not reveal her victimizers. Such a horrifying dramatic image from Shakespeare signifies the powerlessness of women when their voices are silenced.

Many women in various places of the world are consciously aware of the fact that they have a small voice or are confined in silence. Anasuya Sengupta, a contemporary Indian poet, wrote in a poem, "Silence":

> Too many women
> in too many countries
> speak the same language
> of silence [17].

In 1995, Sengupta, then a student, gave this poem on a sheet of paper to Hillary Clinton, who was the First Lady of the United States visiting India. Clinton used it in her speech to a charity organization in India and later in various occasions. The poem found a place in Clinton's biography, *Hillary Rodham Clinton: A Woman Living History*. Karen Blumenthal, Clinton's biographer, writes that "these words captured feelings shared by women everywhere—that women have a right to be heard in their personal and public lives" [18].

By the same token, Bell Hooks (pen name of Gloria Jean Watkins), an American black female writer, informs us of the silence that suppresses women:

> For many women, it is not a simple task to talk about men or to consider writing about men. Within patriarchal society, silence has been for women a gesture of submission and complicity, especially silence about men…. This silence is often learned when we are young female children [19].

C. Women and Control Mechanism

What all these mean is that the silencing of women is within the family, the culture and the institutions. Thus seen, to free women from the control that has taken root deep in traditional family, culture and institutions, women need to find their own voices.

With the import from the West of feminist ideas into the early twentieth century China, Confucianism, which is believed to have upheld patriarchy and disempowered women for more than two millenniums, began to be rigorously scrutinized. Women are more and more aware of their own disempowerment and are trying to find voices for themselves.

D. Taiwanese Women and Social Expectations

Taiwanese women, not unlike many others of their gender in various parts of the world, are subordinated in social control mechanisms inherent in traditional practices and ideas. The age-old Confucian teaching about women's place being within the house, while men's being without, has a strong hold on

social practices even up to now. Moreover, the patriarchal practice of allocating a married woman to her husband's family alienates the woman from her original family and resources. With her in-laws, she is expected to be a good daughter-in-law, which means that she should adapt to the new family and be silent, which is a gesture of decency.

When one considers twentieth century Taiwan, one may divide it into two parts, the first of which is the Japanese colonization era, which began way back in 1895 and ended in 1945. As colonized, Taiwanese were under Japanese colonizers, and just like almost all other forms of social oppression, the colonization pushed women to the lowest level of social totem pole.

When the Kuomintang (KMT) government took over Taiwan from the Japanese in 1945, it "revived Confucianism to encourage people to obey its authority" [20]. It perpetuated the notion of the family as a base for the country, emphasizing women's role as caretakers for the family [20]. Women were assigned the realm of home.

Even though in recent years Taiwanese government has been proud of the high percentage of female representatives in her legislature—43 in a total of 113 seats in 2016 [21], women's voices are still small in comparison with men's voices, which have always been loud and prevalent, and which have dominated mass media.

IV. CYBERCULTURE AND WOMEN'S VOICES

Mass media in the twentieth century is generally taken to mean newspapers, magazines, radio broadcast and television, which are media able to reach to a mass of people. When one reads a newspaper, it is hard not to notice that the headline news is usually on politics or economics, something that is believed to be of great consequence. Discussions about cooking or family would usually accommodate no space or just a bit of room in the separate family and entertainment pages after the main pages. This organization reflects what is considered to be significant and insignificant. In other words, mass media in the last century focuses mainly on importan t political, military, economic and social issues, leaving no or rather little room for women, whose interests and concerns had been considered rather trivial because basically, they belong to the space of home.

In recent decades, various new forms of media that have appeared with the Internet, such as Blogs, Facebook, Instagram, and Twitter, which are called social media, together with YouTube, a popular video sharing website, are capable of reaching out to a mass of people just like traditional mass media. They are conceptualized as "self media". The word "self" in the term "self media" denotes its independence and freedom, which distinguishes itself from the traditional "mass" media.

One of the most fascinating features of self media is the fact that they can accommodate all kinds of alternative topics and issues. What has been traditionally considered to be significant by mass media is no longer the focal point of attention. In self media, all kinds of seemingly trivial matters flourish. Out of all these appear women's voices that have so often been ignored or marginalized by mass media.

A. Three Examples of Women's Voices in Self Media from Taiwan

Lisa Liu is a practicing female surgeon, who loves to write on her Blog about her busy life in the hospital [22]. Once she is off duty, as she informs her readers, she enjoys sitting in front of her computer to type her reflections on the day's work. She has countless fans that follow her Blog. Liu's postings are characterized by a touch of humor, not completely black but rather gray, and rather close to satire. She depicts the hospital where she works as a sweatshop. She offers her readers a glimpse into some life-or-death surgical procedures, and writes about her misery as a surgeon who has to witness inevitable deaths. With all the seemingly everyday records of her life, she brings attention to her challenges as a female surgeon in a profession traditionally dominated by male in Taiwan.

She describes how she suffered the pain from her menstruation, and had to ask her husband to inject via a syringe painkiller medication for her, after which she continued to work. When she was pregnant, her protruding tummy was a physical obstacle for her to approach her patient on a surgical bed. Her urge to urinate caused by her enlarged uterus pressuring against her bladder also obstructed her work in the operating room. Had she gone out to use the washroom, she would need hand scrubs and sterilization procedure to get back to the aseptic operating room. She criticizes the hospital's system for promotion, which does not accommodate maternity leave. Moreover, as a practicing surgeon who needed to spend long hours in the hospital, she laments that she did not have time to be with her baby.

These problems, which are typical to the female sex with her natural physical state involving menstruation and childbirth, have appealed to readers on the Internet. Her popularity eventually attracted publishers, so she converted her Blogs into books.

Another example of women's voices in the Internet is a famous Blogger Jia Nu Siao Hong ("Jia Nu", in Chinese, means "home" and "women;" it is a term meaning a woman who is reluctant to step out of her home. "Siao Hong" means "Little Red") [23]. She records and comments liberally on what happens in her life as an ordinary woman having to work and having to take care of her family. She focuses mainly on the conflict between her and her parents-in-law. Mother-daughter-in-law relation is commonly believed to be one of the main sources of friction in Taiwanese family life. Such a

friction originates from the conflict between traditional Taiwanese concepts and new feminist ideas.

Traditional Taiwanese concepts expect a married woman's subordination to her husband and to her husband's parents. This subordination means showing filial piety and taking care of them. It is a concept that incurs psychological rejection from many younger generation Taiwanese women, who are aware of women's equality and the disadvantageous situation caused by reallocation to a different family by marriage.

Most of the topics Jia Nu Siao Hong covers seem to be trivial, such as her parents-in-law's different opinion about parenting. She writes: "After I have two kids, I start to understand why people say, with kids, the relationship between you and your parents-in-law will be tense" [24]. She discusses the consequences of speaking out her mind and remaining quiet. She also criticizes men's reluctant attitude in the sharing of household chores. She has a job as a purchasing agent in a company, but she does most of the housework. Her husband, just like many other husbands in Taiwan, does not readily offer help. Even after she has asked her husband blatantly to give her a hand in some routine household chores, such as to do the dishes, he would procrastinate.

Self proclaimed as "the lighthouse of the daughters-in-law circle" [24], Jia Nu Siao Hong, not unlike Lisa Liu, also ventured from cyberspace to book publishing. Her fame in cyberspace got her invited to write columns for several magazines and newspapers, such as *China Times*, a Chinese language newspaper in Taiwan.

A third example of women's voices in the Internet is Li Ke Tai Tai (Evelyn Chen) [25]. She is honored as one of the most popular YouTubers in Taiwan in 2018 [26]. She began broadcasting her video shows in June 2018. Within a few months, she has more than 680,000 subscribers [26]. Her video shows are mostly about household chores and things in everyday life, such as good ways for doing certain household chores or for cooking. She uses science to theorize the ways she suggests, and her explanations are simple and understandable to the general public.

The two Chinese characters "Li Ke," as in her YouTube name "Li Ke Tai Tai," means "science subject". It is a subject commonly believed in Taiwan to be what females are not good at. With a master's degree in biomedical engineering from Columbia University, Li Ke Tai Tai contradicts the stereotype of women being weak in science. The two other Chinese characters, "Tai Tai" in her YouTube name means "wife" in Chinese. Unlike many YouTubers who try to appeal by their beauty, she presents herself as an ordinary married woman. It is commented that her "expressionless" face and flat tone, to many people's surprise, appeal to her audience [26]. The ordinary woman image she presents has proven that a woman's voice, despite how flat, and despite how trivial a subject it is about, can reach out to many and can touch many.

B. Preferences in Mass Media and Women

For the most part of the last century, when the major means for information dissemination to the general public were television, radio channels, newspapers, magazines and books, only a small or a selected portion of people's voices could reach out through these means. The number of television and radio channels is limited, and whatever programs broadcast were the products of not an individual, but a group of people, if not the government or certain interest groups, whose preferences and ideologies could have influenced them. Similarly, newspapers, magazines and books involve a group of people, such as reporters, writers, editors and publishers, in their production. All these people could have become censors. They might disapprove certain perspectives or arguments based on their own preferences and interests. Under such a circumstance, woman's voices have been ignored or suppressed. Trivial subjects such as cooking or household chores have been generally considered to be of no significance when compared to wars, international relations, oil issue, trade and economic problems.

C. Canon of Great Books and Women

In *Great Books: My Adventures with Homer, Rousseau, Woolf, and Other Indestructible Writers of the Western World* by David Denby in 2013, out of the 28 writers hailed as having great influence, only four are women [27]. In *50 Plus One: Great Books You Should Have Read (and Probably Didn't)* by George Walsh in 2006, the 51 books beginning from the Bible, Homer, Confucius, to Nelson Mandela and Yukio Mishima are all works by male authors [28]. There is nothing by female author. In *The Great Books Reader: Excerpts and Essays on the Most Influential Books in Western Civilization* by John Mark Reynolds in 2011, out of a list of twenty five books from Homer, Plato, Aristotle to more recent writers such as G. K. Chesterton, only one is female--Jane Austen [29].

These are just three examples out of the canon of great works that has always overlooked women's voices. One might argue that women have always been less productive when it comes to literary output. This is, in reality, the result of gender inequality throughout the ages. A smaller percentage of women were literate in the past, which led directly to this result.

The renowned English female writer, Virginia Woolf, delineates the plights of women in her Judith Shakespeare story. She explains why Judith, sister to William Shakespeare, that is, if he had one, would not be able to make it to the theatre like her playwright brother even if she was equally talented--from the very beginning, she would not be allowed to attend school [30]. Moreover, back in Renaissance England, women were not allowed to act on the stage. Throughout history, there have always been apparent and hidden obstacles for women to cultivate their talents and to enunciate their voices.

Therefore, the basic cause for women's literary production to be far behind men's in quantity in the past lies in the fact that the percentage of men being literate beats that of women throughout history.

Even though there have been increasing calls of gender equality in most places of the world, there is still a gap between reality and ideal. According to Global Gender Gap Report 2018 published by World Economic Forum, "although average progress on gender parity in education is relatively more advanced than in other aspects, there are still 44 countries where over 20% of women are illiterate" [31]. This high illiteracy rate of women can be used to explain why women's literary productions have occupied only a small space in the canon of great books discussed above.

The Global Gender Gap Report covers 149 countries, but does not include Taiwan. Remarkably, Taiwan wrote reports on the Implementation of the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) adopted by the United Nations, and has recorded in her third report published in December 2017 that in the "2016 legislative election, women accounted for 38.1% of elected legislators (compared to 33.6% in 2012, an increase of 4 percentage points)" [32]. It is obvious that though there has been great progress over the years, there is still a gap, and this actually coheres with the so-called "Global Result" in the Global Gender Gap Report above-mentioned, which states that: "the gap is still large across most of the 149 countries assessed…to date, no country has achieved [gender] parity" [31].

D. Cyberculture and Turning the Tables

Cyberculture with its availability and accessibility to Internet users has allowed more and more women's voices to be heard. As above discussed, more and more people are increasingly alarmed about their digital footprints being tracked. They worry that this may lead to surveillance resulting in control and disempowerment--just as George Orwell's sci-fi extrapolates. Hence comes the question—Are women's voices in cyberculture advantageous?

Not unlike the controversy about whether Big Data is beneficial or harmful, whether cyberculture means control or democracy, the question about whether women's voices in cyberculture are advantageous also involves doubts about surveillance and freedom. When we consider the fact that women have been voiceless as they have been marginalized way back in history by traditional concepts that expect them to be silent, we would more likely see cyberculture as an opportunity to turn the tables.

For women whose voices have been marginalized by mass media, self media in cyberspace appears to be a new form of democracy. What is of great value in self media is its capacity to accommodate all regardless of how trivial they are. Theoretically, only those that explicitly disrupt public order and contradict good morals, such as videos of pornography and violence, racist/sexists talks or death threats, might be withdrawn or censored by media service providers or website administrators. Yet the authority of the providers or administrators may still incur doubts about control. These doubts are worsened by the news of the practice of certain countries censoring certain social media, video sharing channels, or apps, and releasing fake news. Beyond a shadow of a doubt, we are consciously aware that cyberculture can be abused or misused.

Fortunately, in current Taiwan, cyberculture is more of a space of democracy than control. Freedom of expression has been manifested in a great variety of ways. Satirical criticisms of government policies and of political figures, for instance, have become ingredients for popular video shows in the Internet, such as Brian Night Night Show, Eye CTV and Catino Mad News. With a large number of audiences and subscribers, these video shows are able to create revenue from commercials, and they are generally believed to be independent from government control. Cyberculture in Taiwan, as proven by these video shows, is a virtual space for all kinds of contending voices. It is a space reflecting our democratic state.

Within this virtual space, the above-mentioned three woman Bloggers/YouTubers, Lisa Liu, Jia Nu Siao Hong, and Li Ke Tai Tai, have empowered their own voices. They speak about the subjects relevant to their own lives regardless of how insignificant they may seem when compared to the subjects covered by traditional news stories or television programs in the age of the mass media. Women's private issues, such as the discomfort of period pain, a daughter-in-laws' unhappy friction with her parents-in-law, and woman's complaint about her husband's reluctance to share household chores, which in the last century were subjects of embarrassment not to be discussed in public, become interesting stories in self media. What used to be considered trivial chores such as how to cook a certain dish, or how to vacuum clean one's house, have also become interesting subjects.

Despite the fact that there is still a gender divide in Taiwan according to the above-mentioned Survey on 2018 Individual/Household Digital Opportunity, which shows that women are behind men in their rendezvous with cyberculture, women's voices are no longer small and suppressed. Cyberculture has offered the possibility for women's voices to reach far and wide. For the time being, this newly acquired freedom for women has outweighed the worry about the possible control of cyberculture.

V. CONCLUSION: A New Possibility for Women

Orwell's sci-fi *1984*, which emphasizes the stress from being watched by Big Brother, and which coheres with Michel Foucault's theory of the psychological pressure from surveillance, has already warned us of the capability of modern technology that may be abused to control and condition us. Undoubtedly, the notorious manipulation of information on the Internet by certain countries has increasingly alarmed us.

Nevertheless, I believe that it is difficult for any form of control to be completely watertight. New talents and new needs will create new possibilities out of what appeared to be impossible in the past. VPN services, which provide virtue pathways that may serve as proxy, are an example of the new possibility for resisting control.

What remains true is the fact that there is an ongoing dialectics of control and democracy in cyberculture. Within this dialectics, women's issue has acquired a new space for discussion. What has already accumulated in our history resulting in a heavy psychological burden is women's voicelessness. Hence, the flourishing of women's voices in cyberspaces in contemporary Taiwan, I believe, is a phenomenon advantageous to women's empowerment. In the new self media made possible by digital technology and computers, even small and trivial voices may reach out clearly to their audiences. Though there are worries about surveillance and control, the worries, I argue, are far less weighty than women's age-old plight of voicelessness. The primary concern for women, figurative speaking, is to be relieved from the predicament of Shakespeare's Lavinia whose tongue and arms are cut.

What happens to Lavinia, one might wonder if one is not familiar with Shakespeare? Tongue-less and hand-less, she is till able to reveal her victimizers. She takes a staff in her mouth, guides it with her amputated stumps, and writes on a sandy land her victimizers' names [16]. Though at the end, Lavinia meets her ill fate of being killed by her father, who says this is to end her "shame" and his "sorrow" [16], what she has done by using her mouth and stumps to write signifies that there is no silencing of women.

It is important to note that had Lavinia been illiterate, she would not be able to write any name. Thus seen, what really empowers her is literacy. By the same token, literacy would be able to empower many other women. In the present era of cyberculture, besides the ability to read and write, Internet literacy, the ability to read and write on the Internet, would serve to further empower women.

In *The Third Wave* discussed at the very beginning of the paper, Allvin Toffler points out that the new form of civilization brought forth by computers might be "more decent and more democratic" [1]. Even though there are worries and fears about what "cyber" culture, which from its very etymological root is implicated with "control", might bring us, in view of women's voices that Taiwan's cyberculture accommodates, I believe that we are heading towards Toffler's vision of a "more decent and more democratic" world. But of course, to get there, we would still need what Toffler's says--"intelligence and a modicum of luck" [1], without which it would be quite difficult to stay optimistic.

****

REFERENCES

[1] A. Toffler, *The Third Wave*. Morrow, 1980.

[2] C. Dunlop and R. King, *Computerization and Controversy: Social Conflicts and Social Choices*. San Diego: Academic Press, 1991.

[3] L. Z. Wang, "500 unemployed toll collectors demand FarEast to keep her promise to "relocate all" (王立柔, 500名失業收費員 要遠通履行「全數安置」). *The Storm Media*. Mar. 11, 2014. Accessed Jun. 29, 2019.
https://www.storm.mg/article/28422

[4] "Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol", *BBC News*. Mar. 12, 2016. Accessed: Jun. 20, 2019.
https://www.bbc.com/news/technology-35785875

[5] *The Terminator*, movie, directed by James Cameron. 1984.

[6] G. Orwell, *1984*. Harcourt, 1949.

[7] M. Foucault, *Discipline and Punish: The Birth of the Prison*. Vintage Books, 1995.

[8] 1984 Apple's Macintosh Commercial, YouTube video. Accessed: Jun. 20, 2019.
https://www.youtube.com/watch?v=VtvjbmoDx-I

[9] F. Gallagher, "Is Big Data the Next Big Brother?" *Wired*. Accessed Jun. 20, 2019. https://www.wired.com/insights/2013/03/is-big-data-the-next-big-brother/

[10] "Cyber", *Oxford Dictionaries*—English. Accessed: May 20, 2019.
https://www.lexico.com/en/definition/cyber

[11] "Cybernetics", *Oxford Dictionaries*—English. Accessed: May 20, 2019.
https://www.lexico.com/en/definition/cybernetics

[12] "Cybernetics", *Encyclopaedia Britannica*. Accessed: May 20, 2019.
https://www.britannica.com/science/cybernetics

[13] M. Wiener, *Cybernetics, or the Control of the Animal and the Machine*. New York: The Technology Press, 1948.

[14] H. Breslow and A. Mousoutzanis, "Introduction", *Cybercultures: Mediations of Community, Culture, Politics*. Ed. H. Breslow and A. Mousoutzanis. New York: Rodopi, 2012. p. vii-xviii.

[15] "Survey on 2018 Individual/Household Digital Opportunity Survey in Taiwan Executive Summary (December 2018)". Entrusted by: The National Development Council, Executed by: United Marketing Research Co., Ltd.
https://ws.ndc.gov.tw/001/administrator/11/relfile/5813/32110/078c2deb-b441-4c77-a033-d1543d40de2e.pdf

[16] W. Shakespeare, *Titus Andronicus*. Revised Edition. Ed. J. Bate. Bloomsbury Publishing, 2018.

[17] A. Sengupta, "Silence", quoted from *Hillary Rodham Clinton: A Woman Living History*, by K. Blumenthal. Feiwel & Friends, 2016. p. 225.

[18] K. Blumenthal, *Hillary Rodham Clinton: A Woman Living History*. Feiwel & Friends, 2016.

[19] B. Hooks, *Talking Back: Thinking Feminist, Thinking Black*. South End Press, 1989.

[20] F. L. Chen, *Working Women and State Policies in Taiwan: A Study in Political Economy*. Springer, 2000.

[21] "Female legislator numbers hit record high in Taiwan", *Taiwan Today*. Jan. 28, 2016. Accessed: May 20, 2019.

https://taiwantoday.tw/news.php?unit=2,23&post=3835

[22] @Drlisaliu (小劉醫師-劉宗瑀 Lisa Liu 粉絲團). Accessed: May 20, 2019.
https://www.facebook.com/Drlisaliu/

[23] @siaohong. Accessed: May 20, 2019.
https://www.facebook.com/SIAOHONG/

[24] Jia Nu Siao Hong, "Different attitudes towards mother-in-law, mother and people in the streets", *Educating, Parenting, Family Lifestyle* (宅女小紅：對待婆婆、媽媽與路人心態大不同. 親子天下雜誌) issue 88, Apr. 1, 2017. In Chinese. The translation is mine.
https://www.parenting.com.tw/article/5073722-%E5%AE%85%E5%A5%B3%E5%B0%8F%E7%B4%85%EF%BC%9A%E5%B0%8DD%E5%AA%BD%E8%88%87%E8%B7%AF%E4%BA%BA%E5%BF%83%E6%85%8B%E5%A4%A7%E4%B8%8D%E5%90%8C/

[25] @liketaitai. Accessed: May 20, 2019.
https://www.facebook.com/liketaitai/

[26] "In 87 days Li Ke Tai Tai became Internet celebrity. Her expressionless face was the style for her fame", 僅花 87 天就爆紅！理科太太靠「面癱」風格 成最狂網紅. EBC New. Jan. 10, 2019. Accessed June 20, 2019. In Chinese.
https://https:https://news.ebc.net.tw/News/Article/147663//news.ebc.net.tw/News/Article/147663

[27] D. Denby, *Great Books: My Adventures with Homer, Rousseau, Woolf, and Other Indestructible Writers of the Western World*. Simon and Schuster, 2013.

[28] G. Walsh, *50 Plus One: Great Books You Should Have Read (and Probably Didn't)*. Encouragement Press, LLC, 2006.

[29] J. M. Reynolds, *The Great Books Reader: Excerpts and Essays on the Most Influential Books in Western Civilization*. Bethany House, 2011.

[30] V. Woolf, *A Room of One's Own*. Broadview Press, 2001. (First published in 1929.)

[31] "Global Gender Gap Report 2018" published by World Economic Forum. Accessed: May 20, 2019.
http://www3.weforum.org/docs/WEF_GGGR_2018.pdf

[32] "Implementation of the Convention on the Elimination of All Forms of Discrimination against Women, Third Report Submitted under Article18 of the Convention, Republic of China (Taiwan), Convention-specific Document", December 2017. Accessed May 20, 2019.
file:///Users/wf/Downloads/ROC(Taiwan)+CEDAW+3rd+REPORT_Convention+specific+Document.pdf

Building a Corpus of Representations of China in English-language Novels, 1927-2007

Graham Matthews
School of Humanities
Nanyang Technological University
Singapore
gmatthews@ntu.edu.sg

Cally Cheung Hiu Tung
School of Humanities
Nanyang Technological University
Singapore
hiutungc001@e.ntu.edu.sg

***Abstract*—We present a database of representations of China in English-language novels with the goal of revealing the construction of China in the Western cultural imaginary. We have collected 9,383 passages from 6,840 literary works. The database is released under an open license.**

Keywords—keywords, corpus linguistics, literature, China

I. Introduction

English-language novels frequently include passing reference to China but these instances are usually too minor to contribute to the major themes of the novel and pass without comment in traditional forms of literary scholarship. Nevertheless, when collected together, these references are indicative of the ways in which the cultural perception of China has fluctuated and changed over time. Literature and culture do not simply reflect our perception of a nation but help to shape it. Literary texts offer unique perspectives on China for three main reasons: (i) it offers privileged insight into the nation as seen and experienced by an individual person; (ii) novel-writers stand at the forefront of public reflection, debate and awareness — on numerous past occasions, novel-writers were first to note and scrutinise the incipient changes of track or new trends in the challenges that their contemporaries faced and struggled to tackle (Bauman and Mazzeo, 2016); (iii) the novel and the newspaper provided the technical means for representing the imagined community that comprises a nation (Anderson, 2016).

Literary scholars typically make claims based on a few exceptional texts. Such projects offer valuable insight into individual texts but would benefit from supplementary research into broader cultural trends. Consequently, this project examines passages from 6,840 English-language novels across a period of 80 years in order to trace broad patterns and shifts in the literary representation of China. This process reveals fresh connections and perspectives on the complex interplay of power, culture, and the subjective experience of China throughout the twentieth and early-twenty-first centuries.

Western interest in China dates back hundreds of years, starting with some of the earliest recorded contact during the Age of Exploration in the 15th century. However, the image of China as a nation-state is highly fluid within the Western cultural imaginary. Roy Porter refers to the propensity of Westerners to flatten the idea of "Chineseness" into a merely aesthetic arena for the surface play of insubstantial signs (Porter 2001). And throughout much of the seventeenth and eighteenth centuries, China functioned as blank slate for Western fantasies and imaginations.

We have built a database of 9,383 passages that include the words **CHINA** or **CHINESE** as a preliminary step to identifying keywords and tracing the shifting representation of China in English-language novels. The database currently covers the years 1927 to 2007. This project extends Raymond William's famous Keywords project using digital search technology and seeks to inspire further debate about the cultural representation of China (Williams 2014). Our large collection of passages from English-language novels makes it possible to explore questions such as the following:

- To examine the representation of China using the literary field as a whole through vastly larger sampling than the tiny number of novels that comprise the canon traditionally studied by literary scholars. Although each passage is a small part of each novel, treated en masse they reveal broad trends in the cultural construction of China.
- To analyse and interrogate the history of particular conceptions of China by generating a coding scheme to identify keywords. We can then trace historical shifts in the representation of China. For instance, we can determine which years China was predominantly associated with Communism or foot-binding in the Western cultural imaginary.
- To establish the groundwork for a broader study of the cultural representation of China in English-language novels and to support existing research in this field. The database offers empirical evidence for claims made about the representation of China and will help determine whether individual texts are part of a general trend or outliers.

The database has the potential to contribute to nearly every pre-existing and ongoing study of the cultural representation of China by offering empirical data that either supports, nuances or debunks current research.

II. The China in English-Language Novels Database

The database contains the following records:

- Quotation
- Keywords
- Character Involved
- Character's Gender
- Chapter, Page from Book (Ch, Pg)
- Title of Book
- Author
- Nationality

- Year of Original Publication
- Book's Edition
- Remarks (sources of full texts; items of note; duplicate entries from anthologies)

For example, for the first quote of out five from Eric Ambler's *Cause for Alarm* (1938) the data would be:

- We were eating in a Chinese place, and I have heard that the Chinese are a very difficult race to astonish; but I seem to remember seeing the cook, a Cantonese with a figure like a water butt, goggling incredulously at us through the service door.
- Low Service Jobs (cook); Restaurant
- Nicky Marlow
- Male
- 10
- Cause for Alarm
- Eric Ambler
- British
- 1938
- Penguin, 2009
- N/A

The database currently lists keywords in a single column but future editions will separate them into three columns to aid data queries.

A. Selection Criteria

The acquisition of data was guided by the development of a list of literary novels determined by the Wikipedia pages entitled "[year] in Literature". These articles present lists of the literary events and publications in a particular year. Research assistants (RAs) extracted the list of novels from each article and excluded children's literature, Young Adult literature, drama, poetry, non-fiction, and novels that are not written in English. They then listed full metadata for each text in a separate database called Literature List. Additional data highlights relevance to the China project by indicating whether the novel contains the terms **CHINA** or **CHINESE** and where searchable text can be located. The latter criteria is intended to aid future projects that seek to search English-language novels for terms unrelated to China.

The Literature List database contains the following records:

- Title of Novel
- Author
- Nationality of Author
- Year of Publication
- Reference to **CHINA** or **CHINESE**
- Google Books
- Amazon (US)
- Amazon (UK)
- Internet Archive
- Remarks (link to searchable text)

For example, for *The Cradle Will Fall* (1980) by Mary Higgins Clark the data would be:

- The Cradle Will Fall
- Mary Higgins Clark
- American
- 1980
- Y
- N
- N
- N
- Y
- https://archive.org/details/cradlewillfallclar00clar

We chose to exclude texts that have more than 10 instances of the terms **CHINA** or **CHINESE** because they were deemed to explicitly thematise China. Rather than rehearsing stereotypes or reproducing cultural assumptions, these texts tend to offer more considered and nuanced representations of China and consequently stand as outliers. In addition, the high frequency of the terms **CHINA** and **CHINESE** means that these texts would dominate the database. The RAs later returned to the dataset for quality control: removing repeat entries, double checking the accuracy of the metadata, and normalizing the data.

The RAs located digital copies of each text using resources such as Googlebooks, Project Gutenberg, and Internet Archive. They also made use of databases such as Eighteenth Century Collections Online (ECCO), Nineteenth Century Collections Online (NCCO), and Proquest's Literature Online (LION), which contains over 350,000 digitised (and searchable) works by key novelists writing between 1782 and 1903. As far as possible they are linking the text to standard authority controls. The Library of Congress Control Number (LCCN) will be used in future editions of the database to ensure accuracy when recording metadata. The key barrier to employing artificial intelligence (AI) to search novels for references to China is the lack of a unified open-source catalogue of novels.

Texts published from 1927 onwards fall under copyright law and are unavailable unless purchased. However, we are able to search texts using Snippet View, which shows information about the book and a few snippets that can be directed using search terms. The snippet view reveals a few sentences for each instance of the search term and is consequently entirely suitable for this project. Texts are nearly always available for snippet search via Google Books, Amazon or Internet Archive. Amazon UK and Amazon US appear to employ their own digitisation method and search functionality; if a text is unavailable on one, it is likely to be available on another. We also accessed alternative databases

such as the HathiTrust Digital Library, eBooks@Adelaide, and Scribd for rare texts. The few remaining texts will be purchased via Kindle or other ebook providers; these editions are fully searchable.

Keywords are determined through collocation or proximity to the words **CHINA** or **CHINESE**. We analysed each reference using a set of qualitative research questions divided into two stages. Stage 1 consists of the collection of the raw data on lexical words within a certain proximity of the search terms, and leads to the identification of a list of keywords for analysis (Stage 2).

Stage 1: Identification of keywords:

1) Identify the topic of the citation.
2) Identify lexemes (basic units of meaning) within a specified proximity of the search terms **CHINA/CHINESE**.
3) Explain how these lexemes generate meaning by reading the surrounding text.
4) Compile a list of these lexemes and record their frequency in order to generate a list of keywords.

Stage 2: Analysis of keywords:

1) Identify the sense in which a keyword is used and check if any semantic shifts (changes in meaning) have taken place over the course of the data. The Online Oxford English Dictionary (OED) will be consulted to check for additional meanings.
2) Describe how each keyword intersects with other keywords.
3) Identify the duration and time frame of the keyword.
4) Record trends or associations that accumulate in geographically specific locations.

The coding scheme is based on definitions from the OED. A list of 40 keywords was generated by examining clusters of quotations and taking into account the authors's nationalities and the publication years. We derived definitions for each keyword and identified sub-categories based on frequency. For instance, the keyword **TORTURE** has the following definition and has the sub-category **WATER TORTURE**:

Deliberate act of cruelty; inflicting pain to get information; tool of warfare; various forms of physical, mental, and emotional torment; Chinese water torture is a routine where water slowly and persistently leaks onto the prisoner's forehead

Identification of keywords enables us to detect patterns in the data and to record frequency of key conceptions of China over time. Keywords derived from quotations from novels are always provisional and contingent upon the reader's interpretation.

B. The State of the Database

There are currently 9,383 records in the database although not all fields are complete. Sometimes the search delivers only a short quotation and in these instances the RAs conduct additional searches for low-frequency words to pick out further context and lengthen the quotation. This context helps determine the keyword for a passage.

The most common keywords are shown in Table 1 and the frequency of the terms **CHINA** and **CHINESE** is shown in Table 2. Interest in China waxes and wanes over the twentieth and twenty-first centuries and significant peaks appear in the late 1930s, late 1950s, and the early 2000s. These dates coincide with the Second Sino-Japanese War, the establishment of the People's Republic of China, and a period of significant economic growth. A significant trough appears from the mid-1960s to mid-1970s, which coincides with the Cultural Revolution.

The most common keyword is Chinoiserie by an extremely wide margin. Keywords such as War, Food, and Women are very common. The keyword Outland refers to instances when China is referred to as a country very far away, on the other side of the world, or as an index of foreigness. This data indicates that China is typically associated with distance, war, food, femininity, crime, mystery, and restaurants in the Western cultural imaginary. More granular analysis of individual keywords or specific time periods will reveal more significant findings. Definitions of the most common keywords are provided in Table 3.

# Cites	Keyword
3665	Chinoiserie
1036	Outland
801	War
744	Food
680	Women
674	Profile
428	Language
304	Crime
296	Mystique
276	Restaurant

Table 1: The most common keywords

Table 2: Frequency of **CHINA** and **CHINESE**

C. Access

The China in English-language Novels database is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which allows you to share and adapt in any medium or format for any purpose, so long as you give appropriate credit, provide a link

to the license, and indicate if changes were made. A snapshot is currently available at https://doi.org/10.21979/N9/2YMOAK and future releases will be made available there.

III. Future Work

In future work, we intend to both increase the size and richness of the database, further analyze it and add support for visualizing patterns. First we intend to extend the database to the present day and to extend backwards in time to trace long-term shifts in the representation of China. Earlier texts are free from copyright and are accessible through digital archives, which will significantly increase the collection rate. We will also fill in as much missing information as possible. This work will be completed within two years. The revised database will be published as an open-access online resource available to other scholars and researchers as well as members of the general public.

We wish to take advantage of linked open data to link the works, through ISBN, to further metadata, with the help of NTU's librarians. This helps both with normalization and checking of the data. Having an ISBN number allows us to link to the library catalogue's controlled vocabulary. We also aim to display computer-generated visualisations of the history of representations of China in English-language novels in a manner accessible to a lay audience at venues such as the British Library and the Art-Science Museum in Singapore.

We are developing more granular analysis of particular keywords. We have researched the representation of Chinese food in the Western cultural imaginary and are currently investigating the keywords Women and Health. Frequency is displayed in Tables 4 and 7. Issues related to Traditional Chinese Medicine (TCM), doctors in China, infection and sickness sharply rise from the mid-1990s onwards. By contrast, the depiction of Chinese women remains consistently high over the course of the twentieth century.

The frequency of the keyword for Family (Table 5) sharply declines in the early 1960s and does not rise again until the twenty-first century. Western writers tend to emphasise filial piety and loyalty among Chinese people. One possible explanation for the decline is that Chinese people began to place the concept of national allegiance above their families following the onset of the Cultural Revolution. The keyword Mystique (Table 6), signifying incidents of mystery, obscurity, or the incomprehensibility of China from a Western perspective, peaks following the end of the Cultural Revolution. Whereas China had previously been an inaccessible topic, now it appeared unpredictable and volatile to Western observers.

Finally, the keyword Food steadily increases in frequency (Table 7). This trend correlates with the popularisation of Chinese take-away. Food is increasingly used as a shorthand for an entire culture. Alimentary images present foreign cultures in sensory terms, rather than run-of-the-mill descriptions and facts. Further analysis of individual keywords will reveal fresh perspectives on the representation of China in English-language novels.

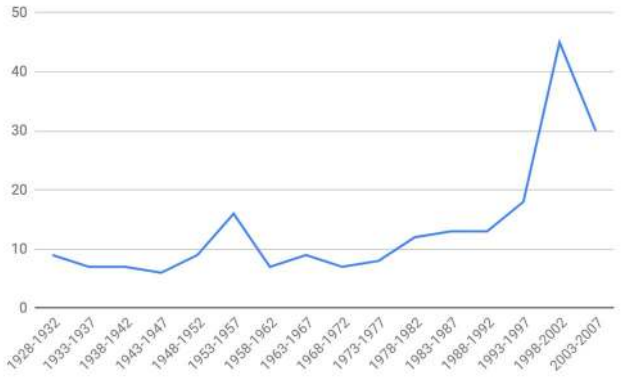


Table 4: Frequency of **HEALTH**

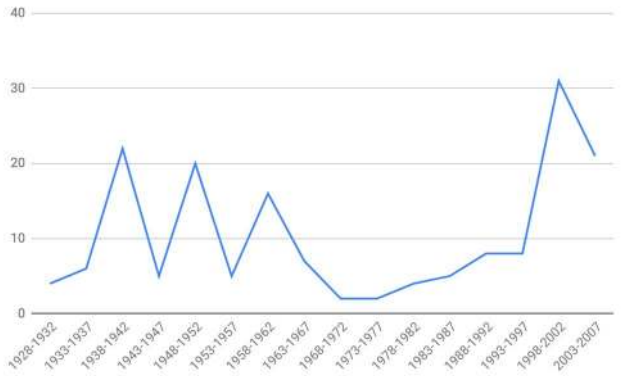


Table 5: Frequency of **FAMILY**

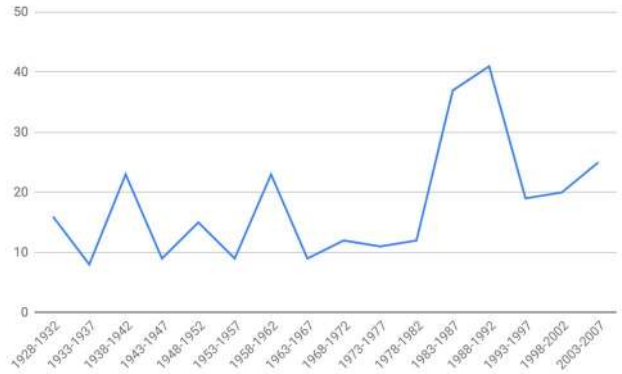


Table 6: Frequency of **MYSTIQUE**

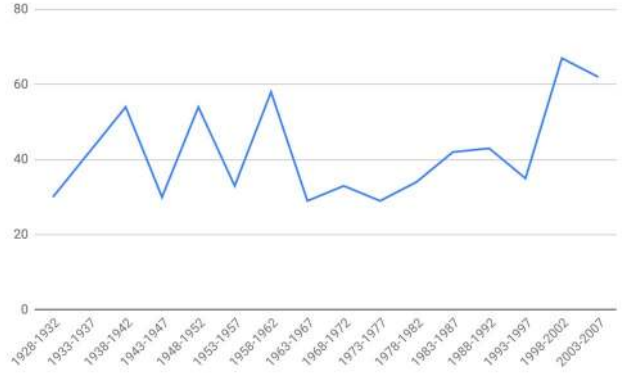


Table 7: Frequency of **WOMEN**

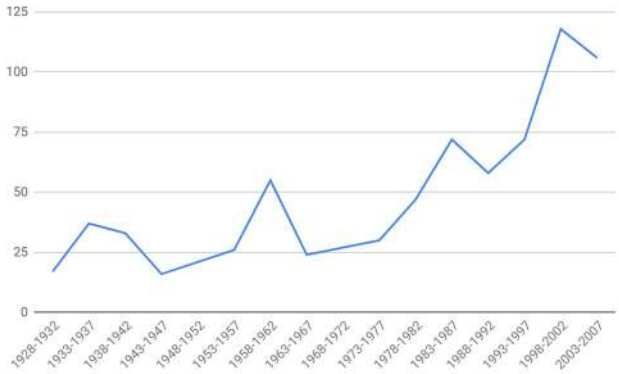


Table 8: Frequency of **FOOD**

IV. Conclusions

We have created a database of 9,383 quotations from 6,840 English-language novels, which we released under an open license. It allows us to study how the cultural representation of China has evolved over an 80 year period.

Acknowledgment

This research was supported by the AcRF Tier 1 Grant, Digital Mapping the Literary Epigraph: Quantitative analysis of literary influence using network theory and thousands of epigraphs (M4011754) and the NTU CoHaSS Cluster on Digital Humanities. We would like to thank the reviewers for their insightful comments.

References

1. Z. Bauman and R. Mazzeo, In Praise of Literature. Cambridge: Polity, 2016.
2. B. Anderson, Imagined Communities: Reflections on the Origin and Spread of Nationalism, revised edn. London: Verso, 2016.
3. R. Williams, Keywords: A Vocabulary of Culture and Society, 2nd edn. Oxford: Oxford University Press, 2014.

Keyword	Definition(s)	Sub-categories (if any)
Chinoiserie	Objects bearing decorative techniques; elements, or motifs deemed "Chinese"; articles made from porcelain and explicitly called "china"; chinaware; pottery; textiles; furniture; architecture; qualities; comestibles; not a measure of "cultural authenticity"	Vase, Rug, Gong, Lanterns, Pagoda
Crime	Illegal acts under Western law; convicts; felony; gangs; pirates; black market; natives; transgression; victims; sitting target; sacrifice; abuse; intense situations of peril; unstable or unfair power relations	Underground
Food	Dishes that constitute a meal; meagre sustenance; Asian cuisine; Chop suey; wanton soup; dumplings; suspicion of animal cruelty; weak tea leaves; tea without milk; subpar quality when compared to Western cuisine; delicious mystery	Fruit, Tea
Language	Characters; verbal noises; substandard communication; mimicry; chicken talk; bizarre; vulgarity; broken system of representation (English); outlandish interpretations; "Do I sound like I am speaking Chinese?"	
Mystique	Unsettling sense of mystery; ghostly or eerie happenings; subjects of enigma; secrecy; folklore about witches and princesses; an intricate and obscure construct; ingenuity of the plan; difficult to comprehend	Puzzle, Chinese Box
Outland	Outlying regions of a country; provinces; overseas; an inaccessible location; appropriated as an extreme description to signify ambition or idealistic dreams of travel and/or connections via trade routes	Great Wall
Profile	Common adjectives and/or descriptors that describe Chinese people, e.g. slanted eyes, yellow skin, big forehead, slow responses, emaciated frame, pointy cheekbones, shuffling or suspect gait. Another phrase that illustrates clumsiness: "Bull in a china shop"	Sage, Insult, Degrading, Bull
Restaurant	A family's livelihood, usually operated by the father manning the cashier or kitchen, while the children wait on tables. Might also be: a cover for dens or an extension of parlours or other underground venues; the cheapest and/ or most cost-effective meal option; accessible provision	Food, Front
War	Waging battles against natives; Western oppressors; World Wars; struggles of a nation; armed conflicts; senseless violence; topics relating to politics and global decisions that affect world neutrality	Politics
Women	Mothers; daughters; wives; antithetical to Western woman because she is demure, docile, and puts the pleasure of men before her own; prostitutes; submissive; pitiful; silent; representative of her culture	Eroticised, Pitiful

Table 3: Definitions of the most common keywords

A Comparison of a Concept 'Civilization' between Modern Korea and China Based on the Methodology of Digital Humanities*

Jaehak Do**
Hallym Academy of Sciences
Hallym University
Chuncheon, Republic of Korea
djhgood@hallym.ac.kr

Injae Song***
Hallym Academy of Sciences
Hallym University
Chuncheon, Republic of Korea
tanksong@hanmail.net

***Abstract*—This study aims to investigate a concept 'civilization' that is an essential concept that represents a modern transformation of East Asia, and to discuss the relationship or similarities and differences between Korea and China. Methodologically, an approach of digital humanities is adopted for extracting and analyzing data from historical materials. Thus the advent, settlement, and change of 'civilization' in Korea and China could be explained and contrasted mutually.**

Keywords—civilization, culture, modernity, corpus linguistics, conceptual history, digital humanities, Korea, China

I. OVERVIEW

Along with 'culture,' 'civilization' is an essential concept that represents a modern transformation of East Asia. Since these concepts were diffused from the West to East Asia, various actions such as interpretation, translation, evaluation, adoption, and rejection were attempted and tangled up by different subjects of socio-political activities. Then, the concepts 'civilization' and 'culture' of those days can not be understood reasonably by founding on a definition in the current dictionary. Therefore, it is necessary to seek an alternative research methodology that analyzes usages and contexts of modern times from a perspective of conceptual history[1, 2, 3].

Not a few researchers have examined in detail the historical fact that a concept of 'civilization' casts an anchor to each country of East Asia and spreads out on the society at large. Most precedent studies, however, focused on the regional relationship between the West and the East(Japan or China) on the one hand, and between Japan and Korea on the other hand. In particular, it is verified that activities of Western missionaries, contacts with the outside of Guō Sōngtào, Kāng Yǒuwéi and Liáng Qǐchāo, a personal connection of Fukuzawa Yukichi and Yu Kil-chun, *etc*. played critical and pivotal roles in the formation of the concept of 'civilization' in each country. However, with regard to 'civilization,' there were few discussions about the relationship between Korea and China.

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A6A3A01022568).
** First author
*** Corresponding author

At first, a sketchy explanation will be discussed for the advent of 'civilization' based on annual appearance frequency information and a list of co-occurring words. Furthermore, a comparison with data for 'culture' will be provided to improve our understanding of linguistic and historical contexts. Nextly, considering the connotation and development phase of a concept 'civilization' in Korea and China, it is possible to figure out the similarities and differences in terms of processes of introduction, settlement, and change.

This study has considerable importance and significance in the way to supplement a weak point of discussion about the conceptual formation in modern East Asia. Especially, to make use of digital databases of modern Korea and China in parallel is a novel method in research history for modern concepts.

II. ANALYSIS & DISCUSSION

A. Approach and Data

- Conceptual history and corpus linguistics.

This study, in a different stance from precedent researches, compares the concept of 'civilization' in modern Korea and China.

Methodologically, an approach of digital humanities is adopted for extracting and analyzing data from historical materials.

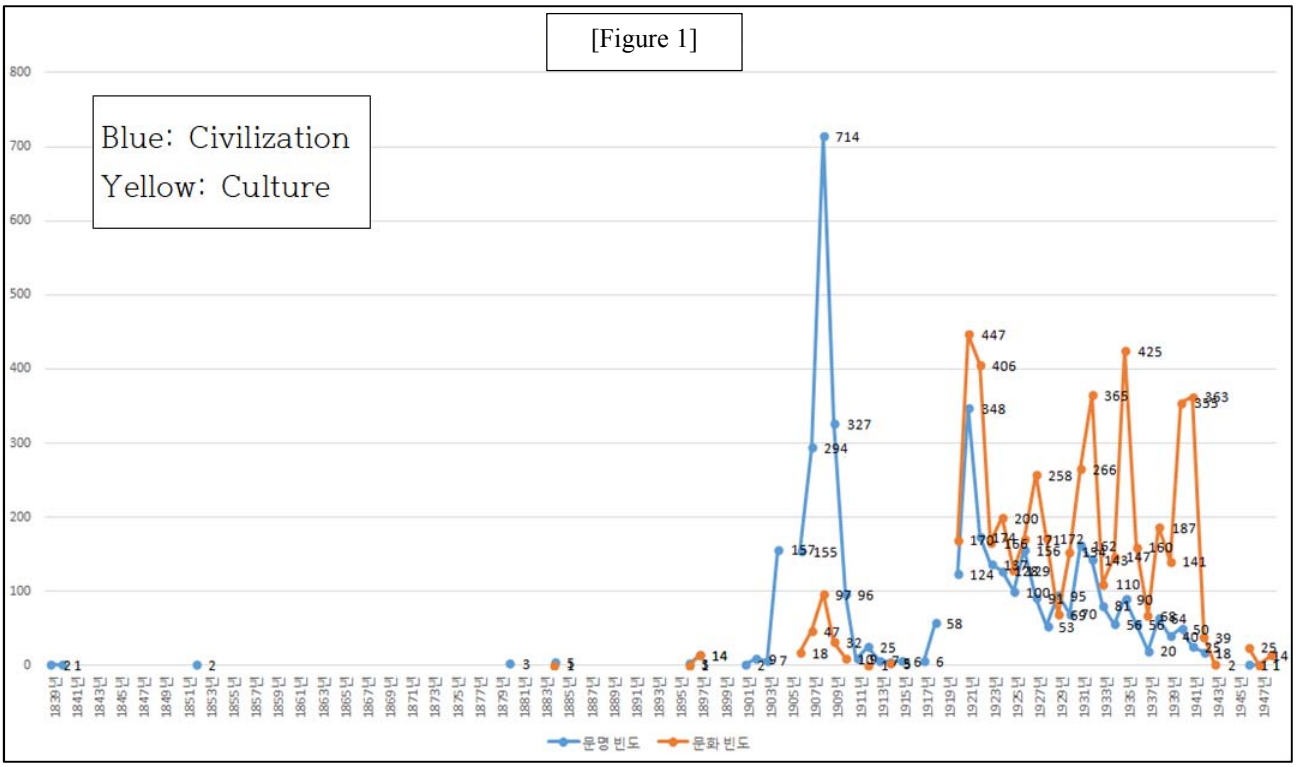
To be specific, data from corpora are appearance frequencies of the word 'civilization,' co-occurring words, and example sentences. Through analyzing these data, it is possible to investigate the validity of precedent studies and discover newer issues related to the concept of 'civilization' inductively.

- Korean and Chinese corpus.

For Korean, data extracted from a corpus of 20 kinds of modern journals housed in Hallym Academy of Science and the historical corpus of the 21st Century Sejong Project.

For Chinese, data extracted from the database for the study of modern Chinese thought and literature(1830-1930) constructed by National Chengchi University in Taiwan.

[Figure 1]



B. A Frequency of 'Civilization' and 'Culture' in Korea[Figure 1].

- A frequency of 'culture' exceeded 'civilization' after the 1920s.

While a frequency of 'civilization' showed a decreasing tendency, a frequency of 'culture' appeared an increasing tendency.

This point has something in common with precedent research, which presents the argument that 'culture' was more and more occupying or invading the semantic domain of 'civilization' in the 1920s[4].

- A quick and wide spread of 'Civilization Supervising theory[文明指導論]'

A very high frequency of 'civilization' from 1907-1909 is worthy of notice. 'Civilization' appeared 294 times in 1907, 714 times in 1908, 327 times in 1909.

Before the Korea-Japan Annexation Treaty, there were widespread preliminary works that insist and emphasize that civilized Japan should lead uncivilized Korea.

'Civilization' was widely used regardless of text types(genre) such as newspapers, journals, textbooks for children, books for current affairs and enlightenment, the new-style novels, *etc*.

- Time of first appearance of 'civilization' as a modern meaning

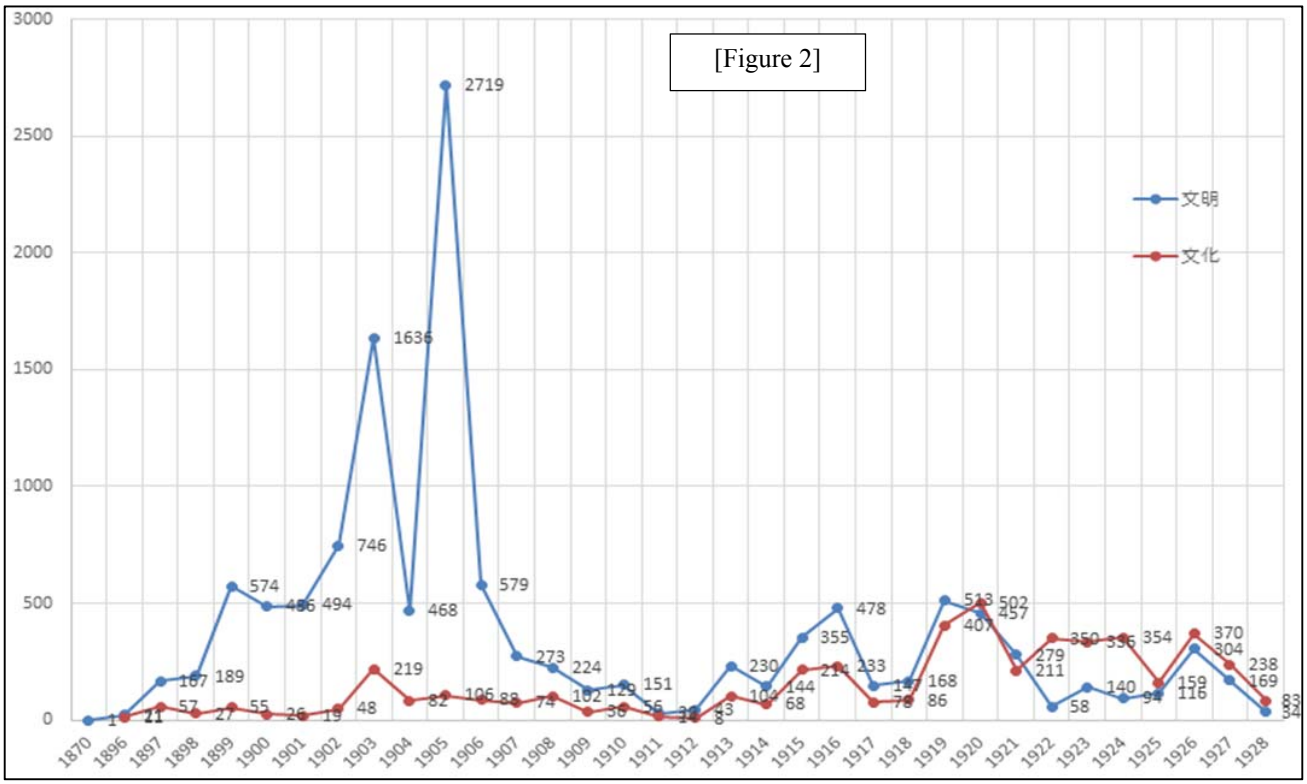
In 1839, 'civilization' used in <the Royal Decrees against Catholic Evils[論中外大小民人等斥邪綸音, so-called 斥邪綸音]>. Though it is evident that this usage had a Confucian moralization, it would be reasonable to think that 'civilization' was not a strange word to intellectuals in those days.

Yu Kil-chun(1856-1914) who went Japan for study and learned from Fukuzawa Yukichi(1835-1901) used at the very first a 'civilization' as a modern meaning on his persuasive writing of a newspaper in the early part of his studying abroad. Then after returning to the homeland, his usage of 'civilization' began in earnest in various texts such as a commemoration of the first publication of <Hansung Sunbo(漢城旬報)> in 1883.

- Coexistence of predominant 'civilization' and slight 'culture' until 1920, an extension of 'culture' after 1920.

As is generally known, many people, including Yu Kil-chun used 'civilization' together with 'culture.' Interestingly, an editor who is known as Park Yeong-hyo(1861-1939) even corrected the word 'civilization' of Yu Kil-chun into 'culture'[5].

'Civilization' and 'culture' were mixed in <Korean Annotation of Ieon[易言諺解]>(1884). <Ieon(易言)> is an enlightening book that handles strenuous efforts for China(Qing Dynasty) written by Zhèng Guān Yīng.



C. A Frequency of 'Civilization' and 'Culture' in China[Figure 2]

- A frequency of 'culture' exceeded 'civilization' after the 1920s.

In the Database for the Study of Modern Chinese Thought and Literature 1830-1930, 'civilization' is appeared at 1870 firstly, and the whole frequency is 12,616 times. Meanwhile, a similar concept 'culture' is used firstly in 1896, and the whole frequency is 4,821 times.

A frequency of 'civilization' surpasses 'culture' until the 1910s, but a gap between 'civilization' and 'culture' considerably shrank in 1909, then eventually 'culture' passes over 'civilization' in the 1920s.

- A sudden increase of 'civilization' by the media from Japan.

In the 1900s, there was a crucial influence of Japan in terms of media. Journals by the Chinese students studying in Japan published in large quantity in 1903, around the same time Liáng Qǐchāo who sought refuge in Japan mentioned 'civilization' many times in <Xīn Mín Cóng Bào(新民叢報)>, and <Wài Jiāo Bào(外交報)> that was delivering foreign news talked about 'civilization' constantly.

A frequency of 'civilization' increased gradually after 1911 and 1921. 'Civilization' was mainly mentioned in the journal that was raising a question of China itself.

Increasing 'culture' and decreasing 'civilization' in this time suggest a reducing tendency of the influence of 'civilization' in the public sphere.

- Time of first appearance of 'civilization' as a modern meaning.

In China, the word 'civilization' came and settled via Japan from 1870 to 1899. Specifically, it is foreign news texts that had an essential role by dealing with the existing state of international affairs. Japanese such as Kozyo Teikichi(古城貞吉) translated articles about the international situation in those days.

D. A Comparison of the frequency of 'civilization' in Korea and China

- In common, 'civilization' was dominant until 1920. However, 'culture' passed over and replaced 'civilization' in the 1920s.

There was a so-called 'New Culture Movement [新文化運動]' in Korea and China around the same time.

Under Japanese colonial rule of Korea since 1910, there were still consistent exchanges of personal and material resources between China and Korea, especially in the aspect of socio-cultural parts.

- There is a different certain period that the frequency of 'civilization' was very high in the 1900s.

In Korea, the period was 1907-1909. It is concerned with 'Civilization Supervising theory[文明指導論].'

In China, the period was 1903-1905. It is related to the influence of students studying in Japan and Liáng Qǐchāo, *etc*.

[Table 1]

rank	until 20's			after 20's		
	word	t-score	frequency	word	t-score	frequency
1	I[我]	12.77	168	today[현대]	12.24	150
2	world[世界]	11.83	141	we[우리]	10.84	122
3	country[國]	11.60	138	world[세계]	8.56	74
4	today[今日]	11.31	130	language[말]	8.24	75
5	wealth and power[富强]	9.73	95	development[발달]	8.17	67
6	people[人]	9.48	99	science[과학]	8.12	66
7	country[國家]	9.34	90	age[시대]	7.71	60
8	development[發達]	9.06	83	mankind[인류]	7.60	58
9	degree[程度]	8.92	80	life[생활]	7.53	58
10	advance[進]	8.50	73	people[사람]	7.11	57
11	education[教育]	8.37	72	society[사회]	7.07	51
12	advance[進步]	7.97	64	machine[기계]	6.99	49
13	age[時代]	7.36	55	culture[문화]	6.53	43
14	I[吾]	6.97	52	nationality[민족]	6.44	42
15	countries[諸國]	6.84	47	modern[근대]	6.32	40
16	Japan[日本]	6.82	48	Choseon[조선]	6.20	40
17	the people[國民]	6.67	46	degree[정도]	5.97	36
18	spirit[精神]	6.47	43	advance[진보]	5.91	35
19	Korean[韓]	6.31	41	the West[서양]	5.65	32
20	each countries[各國]	6.29	40	The East[동양]	5.65	32
21	the East[東洋]	6.22	39	materials[물질]	5.56	34
22	liberty[自由]	6.09	38	construction[건설]	5.46	30
23	countryman[同胞]	6.06	38	today[금일]	5.36	29
24	the people[民]	5.79	35	Europe[歐州]	5.27	28
25	today[現今]	5.79	34	today[오늘날]	4.98	25
26	society[社會]	5.78	35	century[세기]	4.79	23
27	thought[思想]	5.75	34	capitalism [資本主義]	4.69	22
28	import[輸入]	5.73	33	city[도시]	4.67	22
29	century[世紀]	5.72	33	regime[제도]	4.67	22
30	work[事業]	5.68	33	thought[사상]	4.65	22

E. Analyzing the list of high-rank co-occurred words with 'civilization' in Korea[Table 1].

- General and comprehensive perspective.

There are five types of major semantic categories.

① Terms for nation, country, and region: world[世界], nation[國家], the East[東洋], the West[西洋], Europe[歐州], *etc.*
② Terms for the ideal state: wealth and power[富强], education[教育], development[發達], science[科學], thought[思想], *etc.*
③ Terms for evolution theory of civilization: advancement[進步], degree[程度], barbarism[野蠻], *etc.*
④ Terms for the present time: today[今日], today[현대, 오늘날],
⑤ 1st personal pronoun: I[我, 吾], we[우리]

It seems that those words reflect the properties of the civilization theory of intellectuals in those days.

① To distinguish countries(or nations, regions) of the world by a degree of civilization.
② To diagnose the current state of our own.
③ To state a target point that we have to pursue.

- Differences between until the 1920s and after the 1920s

In an aspect of the character, high-rank co-occurred words are Sino-character until the 1920s, whereas Hangeul after the 1920s.

In an aspect of vocabulary, there are more modern neologisms such as 現代, 科學, 人類, 社會, 物質, *etc.* after the 1920s than until the 1920s.

In an aspect of the number of co-occurred words, there are more related words until the 1920s than after the 1920s. It seems that this reflects the result of extension and invasion of 'culture' after the 1920s.

① In the highest rank 30, words of until the 1920s are of high frequency and t-score as compared with after the 1920s.

② In terms of t-score 4.5 point, while there are 54 words for until the 1920s,
35 words for after the 1920s.

[Table 2]

rank	1870-1899		1900-1913		1914-1928	
	word	frequency	word	frequency	word	frequency
1	Japan[日本]	231	China[中国]	1969	China[中国]	806
2	China[中国]	225	world[世界]	1429	we[我们]	787
3	China[支那]	161	barbarism [野蛮]	1162	material[物质]	695
4	world[世界]	159	Japan[日本]	1051	society[社会]	688
5	country[国]	150	today[今日]	1022	world[世界]	599
6	today[今日]	131	the people [国民]	948	mankind[人类]	538
7	Europe[欧洲]	121	country[国]	935	material civilization [物质文明]	494
8	people[人]	118	each country [各国]	750	spirit[精神]	488
9	the people [国民]	113	liberty[自由]	726	the West[西洋]	427
10	barbarism [野蛮]	92	advancement [进步]	725	country[国家]	400
11	the West[泰西]	86	Europe[欧洲]	718	life[生活]	394
12	world[天下]	85	society[社会]	680	thought[思想]	384
13	globe[地球]	80	so-called [所谓]	668	ism[主义]	383
14	tast[事]	73	nation[民族]	617	science[科学]	355
15	advancement [进步]	67	government [政府]	615	they[他们]	351
16	politics[政治]	63	country[国家]	602	present[现在]	326
17	country[国家]	59	task[事]	554	Europe[欧洲]	315
18	country[其国]	58	age[时代]	537	advancement [进步]	304
19	so-called [所谓]	55	people[人]	534	nation[民族]	292
20	world[世]	55	education [教育]	497	today[今日]	279
21	each country [各国]	54	development [发达]	486	age[时代]	279
22	Confucius [孔子]	52	the people [国人]	481	so-called [所谓]	276
23	countries [诸国]	51	politics[政治]	475	the people [国人]	274
24	the people [人民]	51	the people [人民]	464	barbarism [野蛮]	264
25	liberty[自由]	50	India[印度]	462	culture[文化]	261
26	great[之大]	50	own country [我国]	451	the East[东方]	251
27	thousand years[千年]	48	ism[主义]	429	country[国]	235
28	Westerner [西人]	48	thousand years[千年]	398	nature[自然]	234
29	evolution [进化]	47	degree[程度]	396	development [发达]	215
30	school[学校]	46	countries [诸国]	393	liberty[自由]	208

F. Analyzing the list of high-rank co-occurred words with 'civilization' in China[Table 2].[1)]

- An aspect is similar with Korea. Three periods could be divided.
- The first, 1870-1899 as an introductory period.

China(中國+支那) and Japan were the most frequent words. It is possible to think that 'civilization' used in the context of Chinese self-evaluating.

Various words are referring to geographical range such as world[世界, 天下], country[國], Europe[欧洲], the West[泰西], globe[地球], *etc*. This point suggests that 'civilization' in this period was related to the different situations of the world.

- The second, 1900-1913 as a peak period.

Barbarism[野蠻] usually contrasted with 'civilization' has the 3rd rank. It shows that the context of 'civilization' changed from regionality to hierarchy(or degrees of development) in this period.

5th word of the list today[今日] that denote current time is significant because this mirrors their diagnosis of the circumstances in those days.

- The third, 1914-1928 as a declining and final fling period.

It is worthy of notice that 'materials[物质], spirit[精神], mankind[人类]' represent the features of 'civilization.' It implies that the properties of 'civilization' became the primary concern in the discussion.

1) The t-scores for Chinese are not available. However, frequency is enough for our discussion.

- Words by total frequency[Table 3].

There are 55 words.

The words only in the first period(1870-1899) are China[支那], the West[泰西], world[天下], globe[地球], country[其国], world[世], Confucian[孔子], great[之大], Westerner[西人], evolution[进化], school[学校].

The words emerged in the second period(1990-1913) are society[社会], nation[民族], age[时代], ism[主义], the people[国人], development[发达], government[政府], education[教育], India[印度], own country[我国], degree[程度].

The words appeared newly in the third period(1914-1928) are we[我們], material[物质] mankind[人类], material civilization[物质文明], spirit[精神], the West[西洋], life[生活], thought[思想], science[科学], they[他们], present[现在], culture[文化], the East[东方], nature[自然].

Though Japan[日本], the people[国民], each country[各国] showed a high frequency in the early days, they disappeared soon after 1914. The traditional terms about the geographical range such as China[支那] world[天下] Westerner[西人], country[其国] went down steadily according to lexical change for modernity. By contrast, the words that increased after the 1900s mostly denote the properties of 'civilization.' It suggests that the center of mass of discussion shifted from the geographical space or situation of each country to 'civilization' itself.

To sum up, 'civilization' in China in those days made Chinese people realize that they were not a whole world but a partial region of the world. A concept of 'civilization' from Japan was discussed among various modern values and let China in the 20th century learn the spirit of the age and reform itself.

III. Conclusion

The concept 'civilization' is a significant element that is composing the so-called 'modernity' or 'modern thought.'

This study showed annual appearance frequencies and a list of co-occurring words in modern Korea and China, then try to explain and contrast those data. As a result of this, some specific information that could not have known is identified; for example, the exact timing and rough tendency of the words' usage, the explicit conceptual relation between 'civilization' and 'culture,' and consistent interaction between modern Korea and China.

This study could be regarded as a novel attempt for research of conceptual history in the respect of adopting corpus linguistics(in broad terms, digital humanities).

Lastly, it is a quite interesting point that the status of 'civilization' in Korea and China today is prominently different. In Korea, 'civilization' is not an influential word that represents the zeitgeist of society[6]. It is hard to find in everyday language. In China, however, it is considered as a significant value in the present Chinese society as we can see by the word-usage, such as 文明城市, 文明單位, 文明用餐, 文明言語, 文明型國家[7, 8].

[Table 3]

rank	word	1870 1899年	1900 1913年	1914 1928年	total
1	China[中国]	225	1969	806	3000
2	world[世界]	159	1429	599	2187
3	barbarism[野蛮]	92	1162	264	1518
4	today[今日]	131	1022	279	1432
5	society[社会]		680	688	1368
6	country[国]	150	935	235	1320
7	Japan[日本]	231	1051		1282
8	Europe[欧洲]	121	718	315	1154
9	advancement[进步]	67	795	304	1096
10	country[国家]	59	602	400	1061
11	the people[国民]	113	948		1061
12	so called[所谓]	56	668	276	999
13	liberty[自由]	50	726	208	984
14	nation[民族]		617	292	909
15	age[时代]		537	279	816
16	ism[主义]		429	383	812
17	each country[各国]	54	750		804
18	we[我们]			787	787
19	the people[国人]		481	274	755
20	development[发达]		486	215	701
21	material[物质]			695	695
22	people[人]	118	534		652
23	task[事]	73	554		627
24	government[政府]		615		615
25	mankind[人类]			538	538
26	politics[政治]	63	475		538
27	the people[人民]	51	464		515
28	education[教育]		497		497
29	material civilization [物质文明]			494	494
30	spirit[精神]			488	488
31	India[印度]		462		462
32	own country[我国]		451		451
33	thousand years[千年]	48	398		446
34	countries[诸国]	51	393		444
35	the West[西洋]			427	427
36	degree[程度]		396		396
37	life[生活]			394	394
38	thought[思想]			384	384
39	science[科学]			366	366
40	they[他们]			351	351
41	present[现在]			326	326
42	culture[文化]			261	261
43	the East[东方]			251	251
44	nature[自然]			234	234
45	China[支那]	161			161
46	the West[泰西]	86			86
47	world[天下]	86			86
48	globe[地球]	80			80
49	country[其国]	58			58
50	world[世]	55			55
51	Confucian[孔子]	52			52
52	great[之大]	50			50
53	Westerner[西人]	48			48
54	evolution[进化]	47			47
55	school[学校]	46			46

REFERENCES

[1] J. Fisch, "Zivilisation/Kultur," O. Brunner, W. Conze and R. Kosellek (eds.) *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*, Vol. 7, 1992.(Translated by Ahn S. H., *A Dictionary for Conceptual History of Kosellek: Civilization and Culture*, Seoul: Purunyeoksa, 2010.

[2] A. Yanabu, Translated by Park Y. S., *One Word Dictionary: Culture*, Seoul: Purunyeoksa, 2013.

[3] D. H. Noh, *Civilization*, Seoul: Sohwa, 2010.

[4] S. Hur, "Language Network Analysis of Munmyeong and Munhwa in Early 20th-Century Korea," *Concept and Communication*, Vol. 22, 2018, pp. 241-279.

[5] Y. S. Ha, "World Politics of Civilization: The Conceptual History of "Civilization" in 19th Century Korea," *World Politics*, Vol. 24, 2002, pp. 319-344.

[6] S. M. Jang, "The Conceptual Network of Civilization-Culture-Religion in Colonial Korea," *The Critical Review of Religion and Culture*, Vol. 28, 2015, pp. 215-240.

[7] I. J. Song, "Finding and Interpreting 'Civilization' and the vision of China," *The Study of Confucian Philosophy and Culture*, Vol. 48, 2012, pp. 169-194.

[8] Y. C. Jung, "Consideration given to 'Wenming Yuyan'," *The Journal of Study on Language and Culture of Korea and China*, Vol. 32, 2013, pp. 41-63.

The Curriculum Development for Global AGILE Problem-Based Learning in Social Entrepreneurship in Global Teams

Tosh Yamamoto
Center for Teaching and Learning
Kansai University
Osaka, Japan
ctltosh@kansai-u.ac.jp

Chris Pang
School of Business Management
Nanyang Polytechnic University
Singapore
chris_pang@nyp.edu.sg

Benson Ong
School of Business Management
Nanyang Polytechnic University
Singapore
benson_ong@nyp.edu.sg

Juling Shih
Department of Information and Learning Technology
National University of Tainan
Tainan, Taiwan
juling@mail.nutn.edu.tw

Hui-Chun Chu
Department of Computer Science and Information Management
Soochow University
Taipei, Taiwan
carolchu@csim.scu.edu.tw

Meilun Shih
Center for Teaching, Learning, & Development
National Taiwan University
Taipei, Taiwan
mshih63@gmail.com

***Abstract*— This paper deals with a progress report on the execution of the sense making in the curriculum development for global AGILE Problem-Based Learning incorporating computational thinking enhanced with ICT, which has been based on the collaborative endeavors between School of Business Management at Nanyang Polytechnic University in Singapore (NPU) and Center for Teaching and Learning at Kansai University (KU), intending to foster the Vision 2020 skills as well as the future work skills defined by Institute of the Future. Although Problem-Based Learning has been ubiquitous in the realm of the face-to-face onsite learning environment, the project is based on PBL in which project team members with common interests in entrepreneurship from both universities organize several teams to aim for startup business plans with simulation in the virtual learning environment. The paper will walk readers through the rationale behind such curriculum as well as the entire process of the curriculum development from the initial preparation to the final product including the assessment. The key factors of such curriculum development are elaborated in the conclusion.**

Keywords*— *global AGILE learning, ICT-enhanced education, virtual collaborative learning environment, social entrepreneurship

I. Introduction

PBL in teams in a classroom has advantages to enhance active learning in a closed learning environment to foster consensus building through discussion [1]. There is no question about it. On the other hand, when it comes to incorporating the future educational skills defined by Horizon 2020 as well as Institute for the Future (IFTF), the classroom version of such learning model has limitations. Horizon 2020 emphasizes that the collaborative education to foster creative thinking in the constructive paradigm is essential in terms of the personalized education for individual learners.

Further, envisioning the future education, Institute for the Future (IFTF) defines the essential future ten skills: Sense Making, Social Intelligence, Novel & Adaptive Thinking, Cross-Cultural Competencies or Global Awareness & Collaboration, Computational Thinking, New Media Literacy, Transdisciplinarity, Design Mindset, Cognitive Load Management, Virtual Collaboration. It follows that the future education must incorporate all these fundamental concepts into the educational paradigm to give rise to successful learners. Refer to Figure 1.

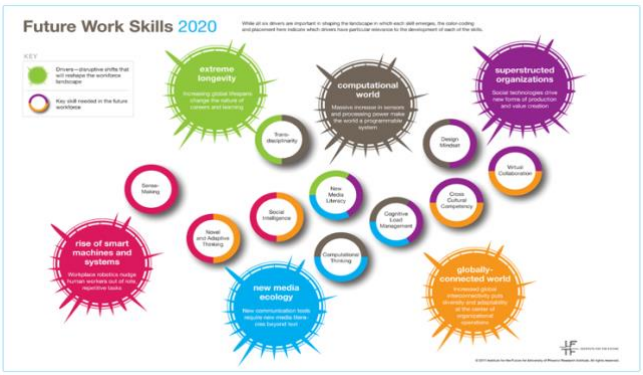


Fig. 1. Future Work Skills 2020. Source: http://www.iftf.org/

Thus, the mission of the future education must incorporate all such skills in the educational paradigm. In other words, it is the sense making of the educational curriculum on the one hand, with the future design of education enhanced with ICT on the other. In here, the consensus among the global team members through communication in the virtual learning environment is the key.

II. Active Learning

What underlies here is, of course, the movement toward the proactive learning. As the Bloom's Taxonomy Matrix in Figure 2 shows that learning is initiated from the stage of memorizing and recalling, and then, to the stage of understanding, followed by the stages of applying, analyzing, evaluating, and creating in sequence [2]. The vertical axis displays the levels and quality of learning provided by the curriculum. On the vertical axis for the learning opportunity provided to the learner, there are, from top to bottom, factual information, conceptual information, procedural information, and opportunities for the metacognitive reflection. See Figure. 2.

Fig. 2. The Bloom's Taxonomy Matrix.

The global education is designed in the way that the entire grids of the Bloom's Taxonomy Matrix must be included in the realm of active learning in "global teams" [3].

III. Learner-Centered Education

At the initial stage of the curriculum development for global active learning in 2017, the goal was to enhance students' liberal arts skills for PBL in global teams through collaboration between Kansai University and National Taiwan University, Asia University, and National University of Tainan. The project was called Collaborative Online International Learning (COIL) courses, targeting at undergraduate as well as graduate students across the border of the campuses. The major learning objective for the COIL courses was to nurture the students' fundamental academic skills such as basic research skills, critical thinking skills with graphic organizers, writing skills, and presentation skills to enhance their comprehensive global communication skills [4]. The courses were conducted in the blended learning approach [5], in which students in global teams conducted active learning synchronously as well as asynchronously through the Internet with mobile devices [6].

From the experience for conducting COIL courses, it was learned that PBL in global teams was not enough to foster students' future skills defined by Vision 2020 and the future skills defined by IFTF [10], [11], [15]. The realm of learning must go beyond the border of the virtual classroom involving the society where the future education resides. In other words, in order to nurture the global mindset equipped with the global future skills, the realm of social entrepreneurship including SDGs must be incorporated into the curriculum [7], [8].

In what follows, the general description of the online course for social entrepreneurship is elaborated. The pilot curriculum for global AGILE Problem-Based Learning incorporating computational thinking enhanced with ICT was conducted with the collaboration of School of Business Management at Nanyang Polytechnic University in Singapore (NPU) and Center for Teaching and Learning at Kansai University (KU), involving the total of 50 students [12], [13].

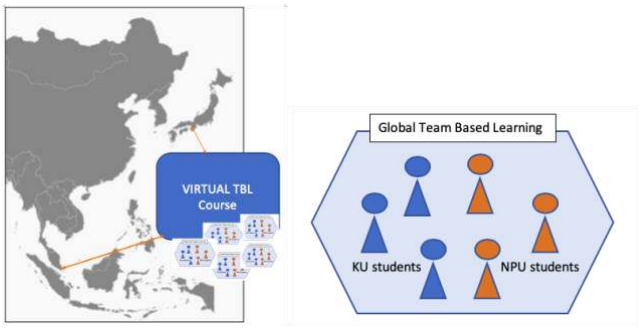


Fig. 3. Global AGILE Problem-Based Learning enhanced with ICT.

IV. Course Structure

The proposed curriculum was designed to foster the ICT-enhanced global and AGILE learning in unison with the collaborative efforts between the institutions located in different countries. The course was designed through face-to-face meetings and online video meetings [14]. The general course information is given below, summarizing the grave points in the syllabus.

A. Course Title

The course title is "Social Entrepreneurship for Asian Students". The subtitles are: "From the Preservation of our Cultural Heritage to a new paradigm: Sustainable Development Goals & From your ideas to products in the market through innovation".

B. Course Description

This is a course enhanced with ICT for virtual learning environment targeting at Asian Coalition Universities for fostering Global AGILE Learning through PBL/TBL.

C. Course Goals

This course offers a great opportunity for students to learn about their own cultural heritages and to visualize them to share with students from other Asian countries. By the same token, they will also learn about the cultural heritages and values from the neighboring countries in Asia.

The fundamental learning activities include the following:

- Explore your own culture and visualize findings in rich media.

- Learn to master skills to visually craft the findings in rich media.

- Work in an international team to share the findings and learn other team members' cultural heritages.

- Through the discussion of the diversity in the different values, appreciate and respect the culture other than their own.

- Build trust among the team members through advanced communication skills.

- Archive and preserve the cultural heritages for the benefit of the future generation in rich media.

- Through global team activities, learn to acquire skills for PBL through TBL, i.e., the essential future skills for the future.

- Acquire the situational leadership skill in AGILE learning model.

D. Objectives

This course will inspire students to develop their skills to be ready for the future citizens to serve as lifelong and life wide active leaders in Asia. Using well-designed content, methods, and self-developed models for assessment and analysis, students will learn how to enhance their presentation/communication skills, by sharing empathy and mindset to build trust, through consensus building in a team to see and act globally.

At the initial stage, the course will walk the students through the entire map of the learning processes as well as tutorials for the ICT tools for visualization in rich media [16].

It is hoped that the learning experience in the course will inspire the learners' innovative minds to make changes in better ways to their academic life as well as their career development [17].

With this course, students will experience how to identify problems in the society effectively and then to build a consensus through discussion to come up with optimal solutions for the benefit of the future.

V. Course Structure: In-Class Learning Tasks & Outside-Class Learning Tasks

In order to compensate for the difference in the spatial and temporal distances among learners, the learning activities the entire course was structured in two-fold: (i) In-Class Tasks conducted by local students and global students online and (ii) Outside-Class Global Team Based Tasks/Assignments conducted synchronously or asynchronously online. In this way, all students from both campuses are on the same page of the learning in and outside of the classroom throughout the course, where ICT plays the crucial role to make such learning environment practically possible.

VI. All Team Members on the Same Page

Now that all the global team members can share the same learning environment for their active learning, it was designed that all of them must be on the same page of their project management mindset throughout the course.

In order to have them on the same page of learning, some strategies have been considered. In this course, in order to accomplish such requirement, an ICT tool called Padlet® was employed. It is noted that Padlet® can be accessed 24/7 from any network device such as smartphones, pad PCs, and notebook PCs.

The entire learning activities were categorized to several sprints, each of which includes a few weeks of learning objectives. See Figure 4. Figure 4 shows at the beginning of a Sprint for the team's final presentation. The link in red will lead viewers to the team's final presentation. The students can view any team presentations and write their constructive comments, which are shown in Figure 5. When all the students finish viewing other teams' presentation and giving comments, each team will discuss and write the reflective remarks on their own project, which is shown in Figure 6.

In this way, all team members as well as the entire class can stay on the same page of the entire activities in the Sprint of learning. It should be noted that the students conducted peer learning through peer evaluation followed by the team-level reflective discussion, which was documented for the artifact of the meta-cognitive skill.

Fig. 4. A Sprint of Learning on Padlet®: at the beginning

Fig. 5. A Sprint of Learning on Padlet®: at the end of the same Sprint

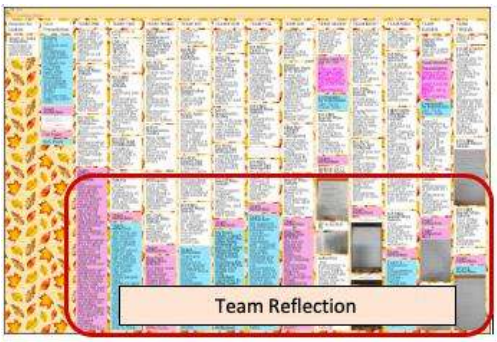


Fig. 6. A Sprint of Learning on Padlet®: Team Reflection shown in the red square

This can be presented in a metaphor of a three-tier stone bridge below. The entire bridge symbolizes the global collaborative course as a whole. Bridge tiers symbolize individual learners' contribution to the team, glocal team members or series of in-class tasks, and outside-class team-level learning activities. Every piece of stone contributes to the whole. A loss of any piece will ruin the entire construction. Each arch symbolizes a sprint of learning. A series of arches will lead to the goals of the global collaborative course that are elaborated above.

Fig. 7. Three-Tier Stone Bridge
Image source: https://commons.wikimedia.org/wiki/File:Pont_du_gard_panoramique.jpg

VII. Pilot Curriculum

KU students enrolled in Social Entrepreneurship in 2018 and NPU business major students from Social Entrepreneur

Course participated in the pilot program. The students were divided into global teams, each of which consisted of three to five Japanese or international students at Kansai University and three to five Singaporean students. None of the students had met face-to-face before or during the course. They all worked in teams in the virtual learning environment. After the orientation in the first synchronous online class, the students introduced themselves to the entire class. Since such a method was not enough for team building with empathy, all the students were asked to create their own "selfy" introduction video to be shared with the rest of the team members. A smartphone APP called "FlipGrid", was employed for the easy video making, editing, uploading, and sharing of the two-minute introduction video on Padlet®. See Figure 9. Once all the team members got to know each other, each team prepared for their first ZOOM meeting communicating with LINE or text messaging SNS service. In this way, each team had team sessions for the learning activities and tasks assigned by the instructors.

In order to encourage the students to develop the project planning and management mindset through the series of team discussions, all teams were required to prepare for the progress report in the form of video during the midterm and then for the final presentation for the final assessment. Presentation materials were all prepared in the online team folder using Google Slides so that all members can work together on the team file "always on the same page". See Figure 8. In addition, each team was given a team folder in which a template for the progress report, a template for the final report, the Gantt chart for the team project management, and the team activity log were stored throughout the semester.

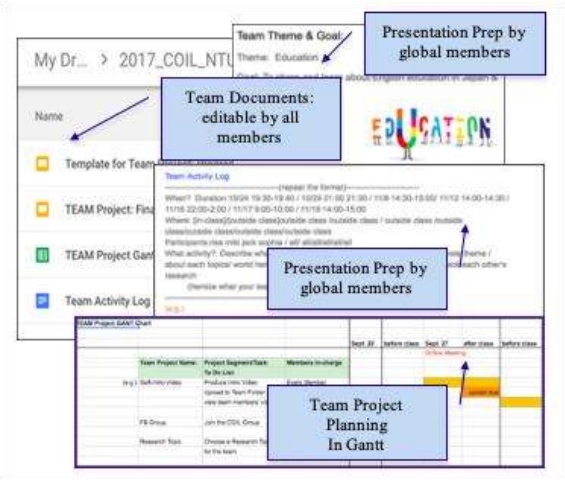


Fig. 8. Team Activity Folder on Google Drive

Furthermore, all teams' progress reports in video, the final presentations, promotional videos for the team presentations, the abstracts of the team research papers, team reflective writing as well as individual students' reflective writing, and all other work were showcased on Padlet®, another smartphone APP, in which all team members viewed other teams' research reports and gave constructive comments. It should be noted, by passing, that Padlet® has the feature of generating a PDF version of the entire learning activities as a document. See Figure 9.

Fig. 9. Padlet® Canvas for all team Members to be on the Same Page

VIII. Final Products/Accomplishments: Leanstack Canvas

The leanstack® canvas was employed to summarize the team's learning activities throughout the course, which gave the students opportunities to develop their comprehensive reflective skills. For all members to be on the same page of the entire learning process, the leanstack® canvas of the team was linked to the Padlet® together with their final presentation slides. Figure 10 below shows the beginning stage of the feedback session.

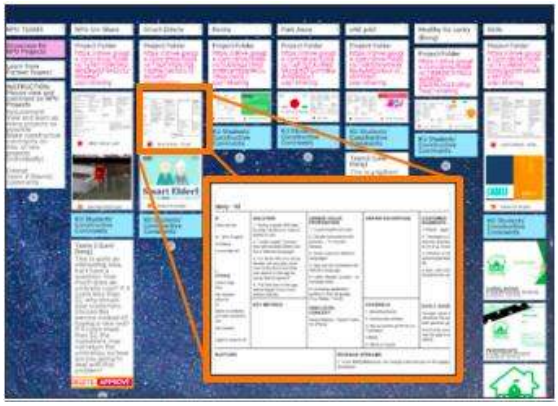


Fig. 10. Feedback Session with leanstack canvas on Padlet®

IX. Rich Media Literacy

In addition to the leanstack® canvas, each team worked on the branding and the product design of their projects in rich media. For example, some teams worked on the promotional video making use of the cloud application called binumi® while some other teams used VivaVideo®, a smartphone APP. Through such learning activities, the students had a chance to reflect the processes of their entire project meta-cognitively as well as comprehensively.

X. Maintaining Motivation of Learning

For the students to maintain their motivation during the course and to continue working in the global team without dropping out in the middle, the team was given the responsibility to choose the research theme. Meanwhile, the students were given instructions or mini-lectures for the use of visual organizers for brainstorming, organizing ideas, and conducting analysis with thinking tools in order for them to visually view and share their progress in the project.

XI. Assessment Strategies

Finally, a reflection session was conducted at the end of the course. Each team reflected upon their activities in the semester and created the team activity/motivation graph and wrote comments of the turning points in the dynamics of the team activities. When the team graph was completed, each team member drew their own learning activity graph over the team graph. See Figures 11 & 12.

Fig. 11. Reflection Session by Motivation Graph

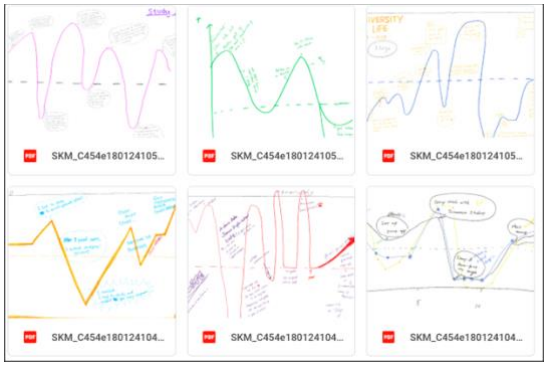


Fig. 12. Visual Presentation of Reflection Using Motivation Graph by the individual Team Members Overlaid with the Motivation by the entire team.

In addition, the reflective writing session was conducted at the end of the course. To trigger productive reflective feedback, the following probing cues listed in Figure 13 were given to the students.

A. Look at your course learning pledge/plan sheet you wrote at the beginning of the class.

Did you accomplish the goal you set at the beginning of the class?

B. What was the most memorable learning activity in the course?

Write also why it was the most memorable learning?

C. Does this course help you prepare for your plan to study abroad in the future?

D. Feel free to write anything that you can think of after learning in this course. (e.g. happy moments, sad moments, worries, future plan, etc.)

Fig. 13. Reflective Writing: A List of Probing Questions

When finished writing, the students were asked to take a snapshot of your reflection paper and post it to the course Padlet! See Figure 14 below.

In general, most students found difficulties in communicating with their team members due to the language barrier since English was the common language. And yet, they consider such experience as valid for their future goal of studying abroad. Making international friends and acquiring the skill to conduct research were most common positive experience in the course.

Fig. 14. Course-Final Reflection Paper

For the evaluation for the course, it should be pointed out that the grade for each student was based on: (i) 50% of learning attitudes at the individual level and (ii) the other 50% of PBL in TBL and the situational leadership skill at the team level.

XII. Conclusion

This paper summarized the pilot curriculum development for global AGILE Problem-Based Learning incorporating computational thinking enhanced with ICT between Kansai University (KU) and School of Business Management at Nanyang Polytechnic University in Singapore (NPU) in the social constructive paradigm in the academic year of 2017. The curriculum made full use of active learning, PBL in Global TBL in the social constructive paradigm, in which the team was composed of global members with various values and viewpoints. It was proved that the team members who had never met before the course were able to conduct active learning in PBL with the solid team building with empathy at the beginning of the course. Bjerede, Atkins, & Dede (2010) claimed that ICT components such as Web 2.0 interactive media, immersive interfaces, collaborative knowledge creation and sharing, multi user virtual environment, mobile wireless interfaces, information technology, play an important role as the robust tool for the actively learning mind. There is no question about their claim. The pilot curriculum, indeed, went beyond their claim.

The proposed curriculum provided the students with learning opportunities to gain self-efficacy to conduct the global-level PBL with a global team, which eventually makes them ready for a long-term studying abroad program as well as the lifelong as well as life wide journey of learning in the global career.

Admitting that one semester-long course is not long enough for students to have the global teamwork mindset. It is the right direction for the benefit of the future generation to develope a series of courses in the global curriculum by sensemaking with Vision 2020 as well as Future Work Skills defined by IFTF.

References

[1] Anderson, T. (2008). The theory and practice of online learning (pp. 45–74). Retrieved January 20, 2019, from http://www.aupress.ca/books/120146/ebook/99Z_Anderson_2008Theory_and_Practice_of_Online

_Learning.pdf.
[2] Davis, L. (2011). Revised Bloom's Taxonomy. Retrieved January 20, 2019, from https://www.slideshare.net/LauraDavis/blooms-taxonomy-made-easy.
[3] Boling, E. C., Hough, M., Krinsky, H., Saleem, H., & Stevens, M. (2012). Cutting the distance in distance education: Perspectives on what promotes positive online learning experiences. Internet and Higher Education, 15, 118-126.
[4] Bonk, C. J., & Graham, C. R. (2006). The handbook of blended learning environments: Global perspectives, local designs. Retrieved January 20, 2019, from https://books.google.com.mx/books?isbn=1118429575.
[5] Engeström, Y. (1999). Activity theory and individual and social transformation. In Y. Engeström, R. Miettinen, & R.-L. Punamäki (Eds.), Perspectives on activity theory (pp. 506–518). Cambridge, England: Cambridge University Press.
[6] Flavin, M. (2016). Disruptive conduct: The impact of disruptive technologies on social relations in higher education. Innovations in Education and Teaching International., 53, 3-15.
[7] Fry, N., & Love, N. (2011). Business lecturer's perceptions and interactions with the virtual learning environment. International Journal of Management Education, 9, 51-56.
[8] Glazer, H. R., & Wanstreet, C. E. (2011). Connection to the academic community: Perceptions of students in online education. Quarterly Review of Distance Education, 12(1), 55-62.
[9] Graham, C. R. (2006). Blended learning systems. Definition, current trends, and future directions. In C. J. Bonk & C. R. Graham (Eds.). The handbook of blended learning: Global perspectives, local designs, 3–21. Retrieved January 20, 2019, from https://books.google.com.mx/books?isbn=1118429575.
[10] Horizon 2020. (2018) Funding, Tenders. Retrieved January 20, 2019, from https://ec.europa.eu/programmes/ horizon2020/en/what-horizon-2020.
[11] INSTITUTE FOR THE FUTURE. (2011) FUTURE Work Skills 2020. Retrieved January 20, 2019, from http://www.iftf.org/uploads/media/IFTF_FutureWork SkillsSummary_01.gif
[12] Kozma, R. B. (Ed.). (2003). Technology, innovation and educational change: A global perspective. Eugene: International Society for Technology in Education International Association for the Evaluation of Educational Achievement.
[13] Lajoie, S. P., Hmelo-Silver, C. E., Wiseman, J. G., Chan, L. K., Lu, J., Khurana, C., et al. (2014). Using online digital tools and video to support international problem-based learning. Interdisciplinary Journal of Problem-Based Learning. https://doi.org/10.7771/1541-5015.1412.
[14] Law, N., Pelgrum, W. J., & Plomp, T. (2008). Pedagogy and ICT use in schools around the world: Findings from the IEA SITES 2006 study (CERC Studies in Comparative Education). Hong Kong: Springer, Comparative Education Research Centre.
[15] VISION 2020. (2014) Vision 2020 – Education. Retrieved January 20, 2019, from http://www.planningcommission.gov.in/reports/genrep/bkpap2020/14_bg2020.pdf.
[16] Waddoups, G. & Howell, S. (2002). Bringing online learning to campus: The hybridization of teaching and learning at Brigham Young University. International Review of Research in Open and Distributed Learning, 2(2). Retrieved Month day, year, from http://www.irrodl.org/index.php/irrodl/article/view/52/108.
[17] Yamamoto, T., Watanabe, M., & Okunuki, M. (2017). Academic writing as corpus for assessment of ePortfolio. 2017 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC). Retrieved January 20, 2019, from http://ieeexplore.ieee.org/document/8203 518.

Detection and Time Series Variation of Latent Topic from Diary in Northern and Southern Courts Period of Japan

Taizo YAMADA
Historiographical Institute
The University of Tokyo
Tokyo, Japan
t_yamada@hi.u-tokyo.ac.jp

Satoshi INOUE
Historiographical Institute
The University of Tokyo
Tokyo, Japan
inoue@hi.u-tokyo.ac.jp

***Abstract*—In the study, we detected the latent topics by Latent Dirichlet Allocation(LDA) and analyzed the texts through the analysis of the topics, with the text of "後愚昧記"(Gogumai-ki) which is an old diary written in Northern and Southern courts period of Japan. We analyzed time-series variations of the topics and considered transitions of the topics in the text. In addition, we have tracked time series variations of the topics.**

***Index Terms*—topic model, Japanese history, old diary, word segmentation**

I. Introduction

In recent years, the use of databases published by various organizations as research materials has begun to be recognized as a matter of course in Japanese history. Since the Historiographical Institute(HI) the University of Tokyo started publishing the database (we call "SHIPS DB" [1]) on the Internet in 1997, the number of users has gradually increased. The number of published databases SHIPS DB is 30, and SHIPS DB obtained about 4.4 million searches in last year (from February 2013 to January 2014). SHIPS DB has various types of data as follows; catalogs, texts, figurines such as portraits and sketches, personal name, character, and so on.

As of March 2019, full-text databases indispensable for the study of Japanese ancient and medieval history such as the following have been released;

- The Dai Nihon Shiryo[2] Unified Database
- The Full-text Database of the Old Japanese Diaries (ODFT)
- The Komonjo [3] full-text database
- The Nara Period Komonjo full-text database
- The Heian Ibun ("平安遺文")[4] full-text database
- The Kamakura Ibun ("鎌倉遺文")[5] full-text database

Each full-text database had about 4.5 times the number of searches in 2018 compared to 2005. For example, in the case of the ODFT, in 2018, there were 243,966 searches (in 2005 57,403 searches). We believe that the increase in the amount of text published in the last few years, and the fact that text search and use are beginning to be recognized in Japanese history research may also be a factor.

Most of text databases provide functions such as string matching search for texts and KWIC (keyword in context), and the users can compare the strings before and after the hit part of the search query with other historical materials. Some of search systems are some which provide a means to access historical material images (such as original historical materials or editorial historical images) that correspond to the hit location. However, there are few systems that provide following text mining functions for analysis of a text and detection of useful information from the text.

- What kind of terms appear in texts,
- What kind of relations between the appearing terms,
- What kind of groups or classes the terms are in.

With such the functions, it is considered possible to provide more sophisticated and inherently necessary historical materials as search results in response to a user's search request for historical materials.

In the paper, we introduce a method to detect the time-series variation of appearing terms for "後愚昧記" (Gogumai-ki) which is an old diary in Northern and Southern Courts period of Japan. Therefore, we will implement 1) term extraction from the text of "gogumai-ki", 2) detection of topics of the terms, and 3) detection of time-series changes in the topics. In order to do step 1), we extracted the text about "Gogumai-ki" from the ODFT, performed word segmentation, and generated Bag-of-Words. The methods are described in Section 3. The number of kinds in extracted terms in the text was 23,928. About 80% of the extracted terms do not appear on the texts

[1]Database system of Shiryohensan-jo(HI; "史料編纂所") Historical Information Processing System

[2]The Dai Nihon Shiryo ("大日本史料") series is an ongoing and comprehensive attempt to cover all major events mentioned in the existing records between the years 887 and 1867 in Japan.

[3]It means old document

[4]collection edited with the aim of covering the old documents of the Heian period

[5]collection edited with the aim of covering the old documents of the Kamakura period

Fig. 1. Article of 貞治五年八月 (Sep 1366) in "Gogumai-ki" (original) held by Historiographical Institute, the University of Tokyo (東京大学史料編纂所)

of other dates. Therefore, it is not so useful to analyze the time series change of the terms as it is. Therefore, in step 2) we describe a method to detect latent topics using the co-occurrence relations of the terms (in Section 4), and in step 3) we describe detection of the topic and the time series variations of topics detected from the text (in Section 5).

II. Target Resource

A. Outline of the Resource

The author of "Gogumai-ki" is 三条公忠 (Sanjo Kintada; 1324 - 1383) who was one of the Japanese aristocratic class (Kuge; 公家), was an active part during the Northern and Southern Courts Period of Japan. He was well aware of 有職故実 (Yusokukojitsu) which consists of ritual, mores, habit, clothing and so on of Imperial Council in Japan. "Gogumai-ki" is a historical material which consists of articles about Imperial Council's ceremony, letters about Yusokukojitsu and so on from 延文 6 年 (1361) to 永徳 3 年 (1383) excluding 貞治元年 (1362), 貞治 4 年 (1365), 永和元年 (1375) and 康暦 2 年 (1380 年). Therefore, it is one of important historical materials when conducting research on the Northern and Southern Courts Period of Japan. Figure 1 shows the article of 貞治五年八月 (Sep 1366) of an original of "Gogumai-ki". HI compiled "Gogumai-ki" and published it as "大日本古記録 [6] 後愚昧記" (Dai-Nihon-Kokiroku Gogumai-ki). Figure 2 shows the compile edition of the article of 貞治五年八月 (Sep 1366).

[6]It is one of compilation series by HI, its target is old diaries of Japan

B. Text, Database

The text of the publication is stored in ODFT. ODFT has 1,545 records for the text of "Gogumai-ki". ODFT manages text in units of days.

III. Term Extraction

Term extraction for Japanese historical texts is a very difficult task. First of all, it is necessary to perform term segmentation, but no method has been established to realize it. For modern sentences, morphological analyzers such as chasen [7] and mecab [8] are available, but they do not work well for pre-modern Japanese text. The result of the term segmentation of text "節会参仕人関白坊城中納言内弁別当侍従宰相右兵衛督等也" which is the beginning of the article of 延文六年正月十六日 using the mecab is as follows:

節 | 会 | 参 | 仕 | 人 | 関白 | 坊 | 城 | 中納言 | 内 | 弁別 | 当 | 侍従 | 宰相 | 右 | 兵衛 | 督 | 等 | 也

Some parts such as "中納言" and "侍従" can be segmented correctly, but most of them are divided into single characters and it can not be said that it is functioning effectively. As in [1], development of a dictionary for morphological analysis

[7]http://chasen.naist.jp/hiki/ChaSen/

[8]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

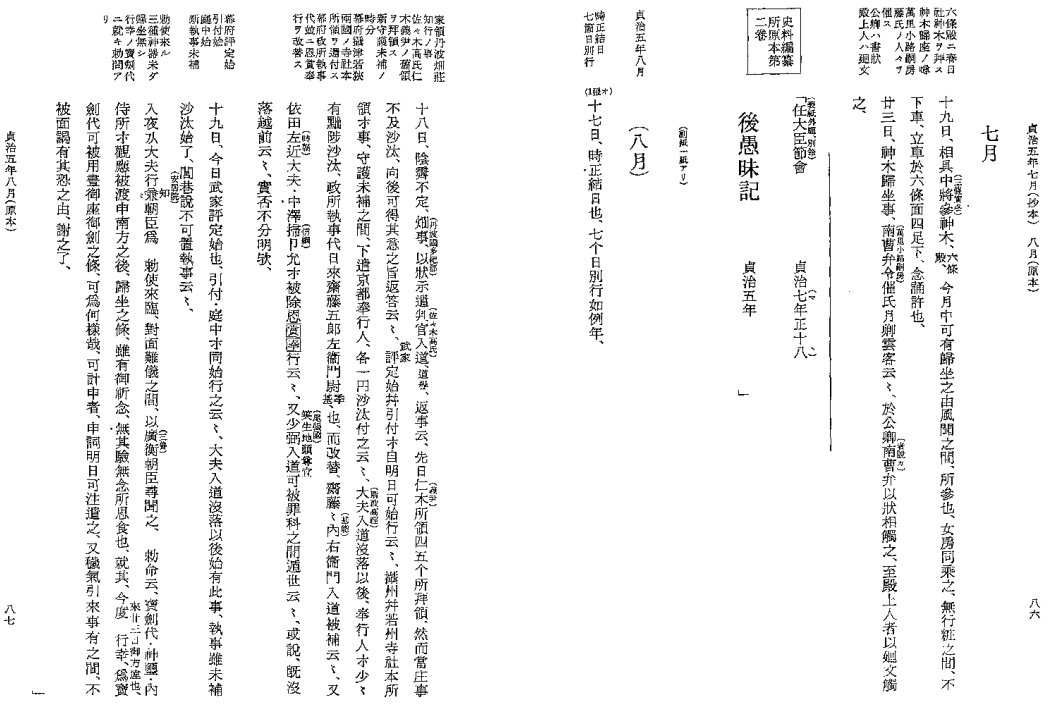


Fig. 2. Article of 貞治五年八月 (Sep 1366) in "Dai-Nihon-Kokiroku Gogumai-ki" compiled by HI

of classical texts is in progress, but unfortunately, it is still difficult to apply to Japanese historical materials which have mixture of Kanji style and Kana style and the grammar is different from modern Japanese. Looking at the beginning of the article "延文六年正月十六日", there are many nouns for appearing part-of-speech, and few other parts of speech such as verbs. The matter is the same tendency in the other day's articles. So, in this research, we decided to treat the results of term segmentation as terms.

As far as we know, the method of term segmentation for the texts of Japanese historical materials is few except for the [2]. The method of [2] is can estimate terminology to calculate with MCMC (Markov Chain Monte Carlo) method and dynamic programming. The language model of the method is one of n-gram model and based on nonparametric Bayesian method called NPYLM. Figure 3 shows the result of term segmentation using the method. The result shows that it is close to the word division of the correct one though it is not perfect. The accuracy rate of the term segmentation in the whole is about 0.55 and it seems low rate at first glance. However, we think that the segmentation method is available to applying to the text because most of the mistakes are the segmentations relate to dates and particles.

In the term segmentation method, Forward filtering-Backward sampling method is used to estimate the optimal term segmentation. In this case, the transition probability from a sentence to other sentences is determined by dynamic programming. The transition probability is calculated using an n-gram language model called NPYLM. The term segmentation is dicided by sampling with dynamic programming called Gibbs Sampler [3]. At this time, Forward filtering-Backward sampling method is applied to all sentences, and the language model and parameters for term segmentation are updated based on the term segmentation obtained from the matter. The fuction is repeated until various parameters converge or their fluctuations become smaller than threshold.

Text:	十六日節会参仕人関白坊城中納言内弁別当侍従宰相右兵衛督等也
Result:	十六日|節会|参仕人|関白|坊城|中納言|内弁|別当|侍従|宰相|右兵|衛督|等也|
Correct?:	十六日|節会|参仕人|関白|坊城|中納言|内弁|別当|侍従|宰相|右兵衛督|等|也|

Fig. 3. Example of the result of term segmentation

In the section, we briefly explain Forward filtering-

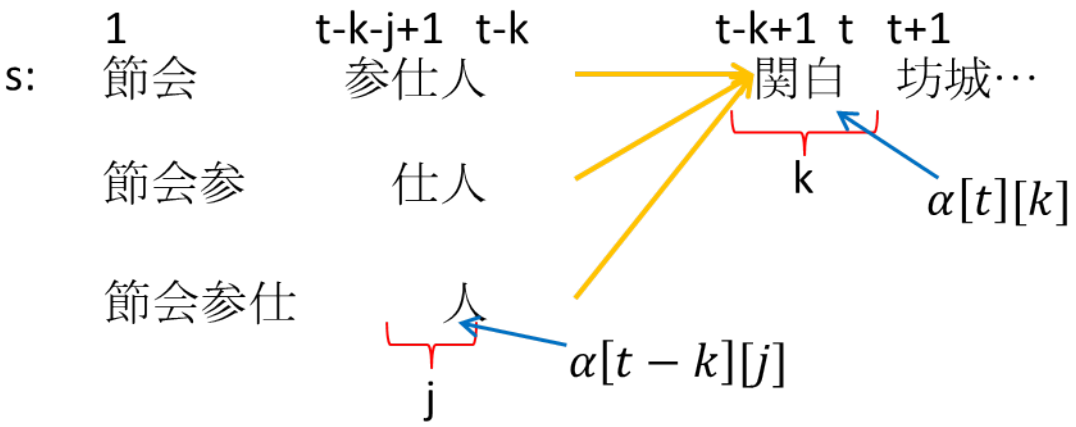


Fig. 4. Illustration of Forward filtering

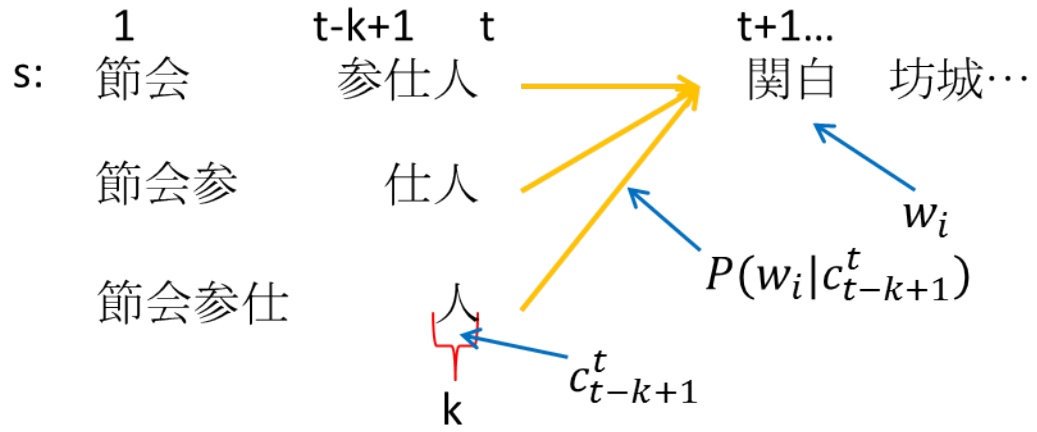


Fig. 5. Illustration of Backward sampling

Backward sampling and NPYLM. Note that refer to [2] for details.

A. Forward filtering-Backward sampling

We describe the Forward filtering-Backward sampling method to estimate the term segmentation $w(s)$. The method has two phases, Forward filtering and Backward sampling.

1) Forward filtering: $\alpha[t][k]$ means the transition probability from substring to others and expressed by bigram. Here, sentence s consists of $c_1, c_2, \ldots, t-1, t, t+1, \ldots$. As shown in Figure 4, α [t] [k] indicates the probability in which the substring which last k length substring of substring $c_1, \ldots, c_t$ of the sentence s is generated. The probability can be calculated by the following.

$$\alpha[t][k] = \sum_{i=0}^{t-k} p\left(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}\right) \cdot \alpha[t-k][j] \tag{1}$$

Here $\alpha[0][0] = 1$. The right side of Equation (1) indicates marginalization for the transitions from all possible segmentations before the target substring to the target substring.

2) Backward sampling: Backward sampling takes an optimal k from $\alpha[t][k]$ obtained by Forward filtering. Backward sampling takes k backwards from the end of the sentence. So we start with k from $\alpha[T][k]$ (where T is the length of the sentence s). As shown in Figure 5, k is taken out in proportion to the following equation and repeated until the beginning of the sentence.

$$k \propto p\left(w_i | c_{t-k+1}^{t}, \Theta\right) \cdot \alpha[t][k] \tag{2}$$

Here, w_i indicates a term that has already been extracted, and Θ indicates a language model. After the procedure, $w_i, w_{i-1}, \ldots, w_1$ can be obtained as the result of the term segmentation. In the study, NPYLM is used as the language model.

B. Nested Pitman-Yor Language Model

NPYLM is an n-gram language model that extends the hierarchical Pitman-Yor language model (HPYLM) [4] and expresses with the combination the character n-gram and the term n-gram. So, HPYLM can be defined as an n-gram

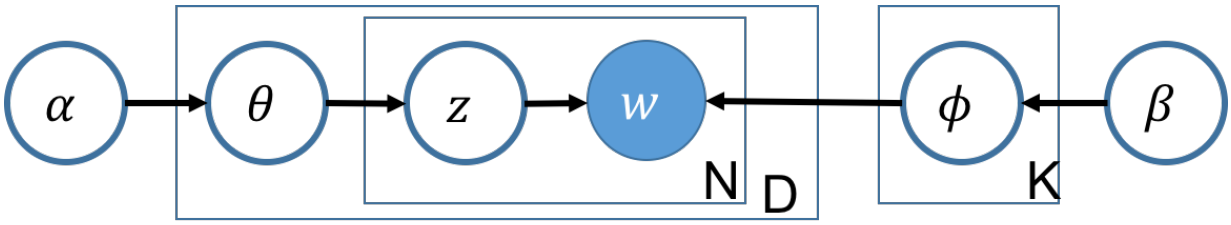


Fig. 6. Graphical model for LDA

language model based on hierarchical Pitman-Yor process. If the term is w and the history of w is h, the conditional probability $p(w|h)$ of n-gram in HPYLM can be obtained by the following equation.

$$p(w|h) = \frac{\theta + c(h)}{c(w|h) = d \cdot t_{hw}} + \frac{\theta + c(h)}{\theta + d \cdot t_h} \cdot p(w|h') \qquad (3)$$

Here, $c(\cdot)$ indicates frequency, h' is a one-dimensional reduced history of h, d and θ are parameters in the Pitman-Yor process, $t_h w$ is a parameter in HPYLM, and t_h is $\sum_w t_{hw}$. NPYLM holds HPYLM for each of the terms n-gram and letter n-gram, and uses letter n-gram as the 0-gram probability distribution in the term n-gram.

IV. Topic Detection

The historical materials have some topics in the described content. In many cases, these topics are not specified in historical materials, but can be detected and understood by reading comprehension semantically. The potential subjects can be called latent topics. In topic detection we used Latent Dirichlet Allocation(LDA) [5] which is one of the topic models. LDA treats a set of terms which are subject to statistical co-occurrence as latent topics. In the topic model using LDA, it is assumed that there are multiple topics in one document. The LDA models the distribution of the topics.

Figure 6 shows the graphical model representation of the LDA used in the paper. Here, the blue circle indicates the observation variable, the white circle indicates the unknown variable, the rectangle indicates repeating, and the lower right numerical character indicates the number of the repeating times represented by the rectangle. w indicates the result of the term extraction mentioned above and is the only observed variable in here. z is the topics, θ is the topic distributions for articles, and ϕ is the term distributions. α and β are θ and ϕ parameters and indicate hyperparameters of LDA. When the number of documents is D and the number of terms in document d is N_d, θ_d and ϕ_k are generated by

$$\begin{aligned} \theta_d &\sim Dir(\alpha) \qquad (d = 1, \ldots, D), \\ \phi_k &\sim Dir(\beta) \qquad (k = 1, \ldots, K). \end{aligned} \qquad (4)$$

Here, $Dir(\cdot)$ represents the Dirichlet distribution. The topic $z_{d,i}$ is generated by

$$z_{d,i} \sim Multi(\theta_d) \qquad (i = 1, \ldots, N_d). \qquad (5)$$

Here, $Multi(\cdot)$ represents a multinomial distribution. In addition the term $w_{d,i}$ is generated by

$$w_{d,i} \sim Multi(\phi_{z_{d,i}}) \qquad (i = 1, \ldots, N_d). \qquad (6)$$

1. Initialize α and β
2. Initialize z
3. Set S : the number of sampling
4. for s = 1, …, S do
5. for d = 1, … , D do
6. for i = 1, … , N_d do
7. Sample $z_{d,i}$
8. Update $N_{d,z_{d,i}}$
9. end for
10. end for
11. Update α and β
12. end for

Fig. 7. Procedure of collapsed Gibbs sampling

There are variational Bayesian method(VB) [5] and gibbs sampling (GS) [6] as mainly well-known inference methods of LDA model. Collapsed gibbs sampling (CGS) [6] which is improved version of GS. Because it is not necessary to calculate θ_d and ϕ_k if the method used, the caluclation cost can be reduced greatly. Moreover, according to [7], the prediction performance of CGS (by perplexity) is better than VB. There is another inference method called collapsed variational Bayesian (CVB) method [8]. CVB is an improved version of VB which can perfome marginalization of θ_d and ϕ_k like CGS in the inference. However, CVB algorithm is complex and requires a large amount of memory. In the paper we detected topics from newspaper data using CGS .

A. Sampling

The procedure of CGS used in this paper is shown in Figure7. Here, $N_{d,k}$ indicates the number of terms assigned

Topic	Total	Kinds	Topic	Total	Kinds	Topic	Total	Kinds
Topic1	549	530	Topic11	1280	997	Topic21	2121	1198
Topic2	1469	1129	Topic12	470	439	Topic22	719	653
Topic3	1426	1056	Topic13	5002	2017	Topic23	1002	878
Topic4	613	557	Topic14	865	774	Topic24	7720	2320
Topic5	1086	937	Topic15	918	846	Topic25	729	668
Topic6	1518	1124	Topic16	583	524	Topic26	703	665
Topic7	2754	1636	Topic17	760	663	Topic27	695	634
Topic8	871	744	Topic18	3496	973	Topic28	746	519
Topic9	963	828	Topic19	3018	1856	Topic29	689	559
Topic10	10799	2766	Topic20	938	835	Topic30	663	625

Fig. 8. Summary of topic detection by LDA

topic k in Document d. $z_{d,i}$ can be sampled by

$$z_{d,i} \sim Multi(p(z_{d,i}|W, Z_{\backslash d,i})) \quad (7)$$

$$\propto (N_{d,i} + \alpha)\frac{N_{k,w_{d,i}} + \beta}{N_k + \beta V} \quad (8)$$

Here, W indicates all documents (whole article data). $Z_{\backslash d,i}$ indicates a set of topic except for $z_{d,i}$. V indicates the number of kinds of terms.

α and β are hyperparameters in LDA, and are parameters in Dirichlet distribution. In general, when each value of α is not uniform ($\alpha_k \neq \alpha_l, k \neq l$) and the value of β is uniform ($\beta_1 = \beta_2 = \ldots = \beta V$), the performance of the LDA improves [9]. We decided to use the hyperparameter setting. The hyperparameters can be inferred by maximizing marginal likelihood. When fixed point iteration can be used for the iteration,

$$\alpha_k^{new} = \alpha_k \frac{\sum_D \Psi(N_{d,k} + \alpha_k) - D\Psi(\alpha_k)}{\sum_d \Psi(N_d + \sum_{k'} \alpha_{k'}) - D\Psi(\sum_{k'} \alpha_{k'})}, \quad (9)$$

$$\beta^{new} = \beta \frac{\sum_k \sum_v \Psi(N_{k,v} + \beta) - KV\Psi(\beta)}{V \sum_k \Psi(N_k + \beta V) - KV\Psi(\beta V)}. \quad (10)$$

Here $\Psi(\cdot)$ indicates digamma function.

B. Sampling

V. Experiment

A. Latent topics in "Gogumai-ki"

We tried to detect topics by the LDA in "Gogumai-ki" text. The temporal range is from 康安 1 年 1 月 1 日 (Feb 6, 1361) to 永和 4 年 12 月 29 日 (Jan 18, 1379). The number of iterations of Gibbs sampling in the term segmentation is 200, and the number of samplings in marginalization in each iteration is 20. As a result, the kind number of extracted terms is 23,929, and the total number was 892,101. The number of topics in LDA was 30, and the number of iterations for the collapsed Gibbs sampling was 1,000.

Figure 8 shows summary (contains the total number and the kinds of terms for each topic) of the LDA result. The total number of terms belonging to Topic 10 is 10,799 and is the largest total number in the whole, followed by Topic 24, 13, 18, and 19. Topic 12 is the smallest at 470.

Topic 10 has a list of terms that includes "云々 (and so on)", "之由 (for the reason)", "之間 (during)", "御 (means respect)", "又 (and)", "可為 (in order to)", "可被 (means passive expression)", "一 (one, first or meaning of itemization's item)", "云'(say)', "之処 (where)", "不可 (can not)", "如此 (like this)", "沙汰 (result)", "為 (for)", "内府 (one of an ancient office in the Japanese Imperial)", "今度 (this time)". These are terms that frequently appear in historical materials of the Northern and Southern Courts period of Japan. On the other hand, we focused on Topic12 which is the smallest topic overall. Topic 12 has a list of topic that includes "廿九日'(29th)', "付女房 (女房 means introverted female servant who served the people of the royal court.)", "返事自是之 (reply for it)", "是非候", "又無出御 (not go out)", "可詠之 (詠 is a performance of reciting a Japanese poem or a Chinese poem read in Japanese,)", "城中 (in the city)", "師元", "手負 (hurt)", "旨示了 (show)". We did not know what Topic 12 represents by the terms. As an example, we confirmed the article "応安 6 年 1 月 7 日" which has frequently terms belong to Topic 12. The text is as follows:

晩頭、経重使者青使〔侍〕来尋云、自北小路上土門前内府出了、彼門永不用之由、有其説、或又五旬以後可開之由、有申之仁、可為何様哉云々、答云、一廻以後可開之、但可替閾也、常俗説如此歟者、凡年始凶事問題、不解機嫌歟、然而先日葬礼儀ハ不獲止事也、此門事、雖不為今日、不可有子細、無故実之至也、

The terms belonging to Topic 12 in this sentence are: "節会", "使者", "青使来", "尋云", "北小路", "彼門", "永不用", "或又五", "可開", "答云", "可開之", "替閾也", "常俗説如此歟者", "不解", "儀ハ", "不獲止事也", "此門事". The article shows the dialogue between 三条公忠 and the messenger of 経重 (勧修寺経重; Kajuji Tsuneshige). This is a query on how to proceed with the ritual, as 勧修寺経顕 (Kajuji Tsuneaki) who is the father of 勧修寺経重 died on the New Year. From the article and terms and other examples, we could understand that Topic 12 shows one of the dialogue about 有職故実 (Yusokukojitsu).

Topic 6 has a list of terms that includes "神輿 (portable shrine)", "山門 (Buddhist temple or the gate of the temple)", "衆徒'(many priests or monks)', "入洛 (entering Ky-

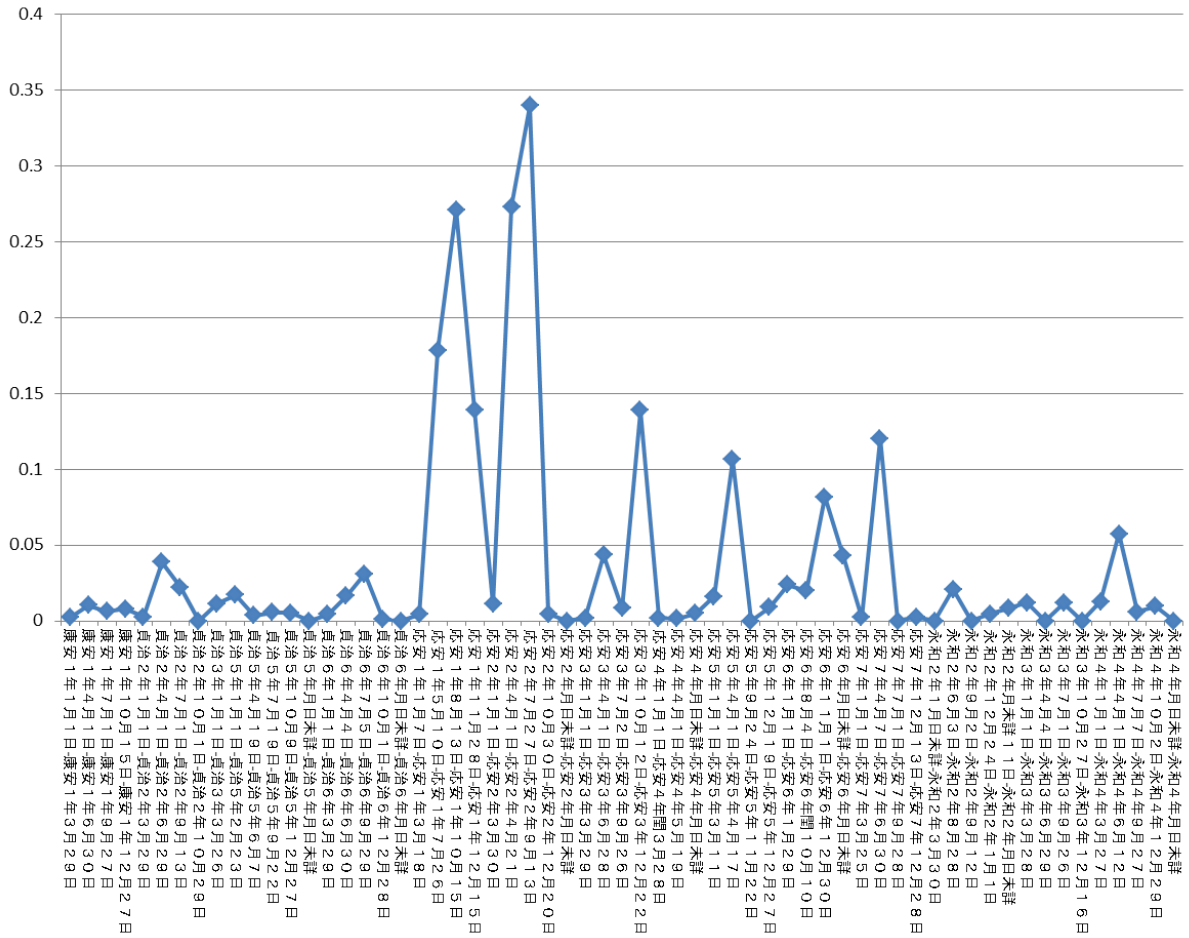


Fig. 9. Time series variation of Topic 6

oto)”, “南禅寺 (temple’s name)”, “賀茂 (place name)”, “日吉 (place name)”, “門跡 (the kind of temple)”, “青蓮院 (temple’s name)”, “一条 (place name or family name)”, “武士等 (Bushi or Samurai)”. This topic can be understood easily for who have studied Japanese history. The topic indicates 強訴 (Goso) which means the appeals and requests made by the group of monks to show off the authority of the Buddha and to the court and the shogunate. Maybe the topic can be understood easily for who have studied Japanese history. Here we analyze the time series variation of the topic. The article of “Gogumai-ki” in the ODFT is from 延文 6 年 1 月 1 日 to 永和 4 年 12 月 29 日. We summarized the daily articles in units of three months Figure 9 shows the time series variation of Topic 6. It can be seen that related terms frequently appear for one year of 応永元年 and during from 応永 2 年 4 月 and 応永 2 年 9 月, and then peaks(maximul frequencies) appear several times, but gradually disappear.

VI. Conclusions

In the study, we detected the latent topics by LDA and analyzed the texts through the analysis of the topics, with the text of “Gogumai-ki”. We analyzed time-series variations of the topics and considered transitions of the topics in the text. Here we have tracked time series variations.

In the future we would like to track spatial variations as well. In addition, we would like to advance analysis of historical events using analysis of the topics, taking into consideration the relationship with items such as personal names, place names, times, and so on. Through these, we would like to lead to promotion of Japanese history research by the usability improvements of texts by statistical analysis methods rather than just full text search. By contrasting with related articles, we confirmed that the frequency of the topic and the activity of the Goso were proportional.

Acknowledgment

A part of this work was supported by JSPS KAKENHI Grant Number 18H03576, 18H03588, 18H05221, 17H00773 and 16H01897.

References

[1] Toshinobu Ogiso, Mamoru Komachi, Yasuharu Den and Yuji Matsumoto, “UniDic for Early Middle Japanese: a Dictionary for Morphological Analysis of Classical Japanese,” Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012), pp.911–915, 2012.

[2] Daichi Mochihashi, Takeshi Yamada, Naonori Ueda, "Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling," Proceedings of ACL-IJCNLP 2009, pp.100-108, 2009.
[3] Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. "Markov chain Monte Carlo in practice," Chapman & Hall, 1st ed edition, 1996.
[4] Teh, Y. W. "A Hierarchical Bayesian Language Model based on Pitman-Yor Processes," Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 985, 2006.
[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.
[6] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National Academy of Sciences of the United States of America, vol.101, pp.5228–5235, 2004.
[7] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," Proceedings of UAI ' 09: Proceedings of the International Conference on Uncertainty in Artificial Intelligence, pp.27–34, 2009.
[8] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation," Proceedings of NIPS 19, pp.1353–1360, 2007.
[9] H. M Wallach, D. M Mimno, A. McCallum, "Rethinking LDA: Why priors matter,"Proceedinsg of NIPS2009, pp.1973–1981, 2009.
[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," Journal of the American Society of Information Science, Vol. 41, No. 6, pp. 391–407, 1990.
[11] T. Hofmann: Probabilistic Latent Semantic Indexing, Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50–57, 1999.

Coding to Decipher Linear A

Niki Cassandra Eu Min
Linguistics and Multilingual Studies
Nanyang Technological University
Singapore
NEU001@e.ntu.edu.sg

Duo Duo Xu
Department of Chinese Studies
National University of Singapore
Singapore
chsxd@nus.edu.sg

Francesco Perono Cacciafoco
Linguistics and Multilingual Studies
Nanyang Technological University
Singapore
fcacciafoco@ntu.edu.sg

***Abstract*- This paper discusses the program logic for an attempt at using coding to aid in the decipherment of Linear A, the writing system of the Ancient Minoan Civilization. Using Python, a system can be created to compare Linear A strings to lexical lists and dictionaries of languages from a compatible time period.**

Keywords—Linear A, Python, Programming, Historical Linguistics

I. Introduction

Linear A is the writing system of the Ancient Minoan Civilization, a Bronze Age Aegean civilization that flourished in Crete and several other Aegean islands, and was used between 1700-1450BCE before being replaced [1]. In-between the use of Linear B by the Mycenaeans for Mycenaean-Greek, there was also a period of Cypro-Minoan, used by the pre-Greek people of Cyprus [2]. Discovered alongside Linear B samples by Sir Arthur Evans in 1886, Linear A samples have been found in a variety of locations including Cyprus, Aegean Islands like Kea, Kythera, Melos and Thera [3], and mainland Greece and Turkey [4].

Since its discovery, many researchers have tried attributing a language family relation to Linear A, or have tried deciphering the language of the writing system, with limited success. Various languages (and language families) have been attributed to Linear A, but a large-scale attempt has not yet been made to process all the samples against dictionaries and lexical lists of various languages at a time. In order to vastly expand the search for a potential language family, this paper proposes the use of a program coded in Python to carry out the search faster, and help narrow down the list of potential candidates for in-depth analysis.

II. Literature Review

Linear A has around 90 signs/ symbols in regular use, 80% of which are unique when compared to Linear B and have been found to be used as individual signs as in combination [1]. Found on a variety of artefacts, it is generally agreed that a majority of the inscriptions found on tablets, roundels and seals denote economic transactions or were used for a stocktaking purpose. This conclusion was reached based on two sets of evidence: first, internal analysis found that a large number of tablets bore logograms in addition to regular signs, denoting commodities such as figs or olives, and preceded numbers. This, Linear A inscriptions have been found on stone vases (some inscribed, others painted), on stucco architectural features, libation tables, metal objects and other items. A vast number of the samples in Linear A are made up of roundels (see Figure 1: Roundel KH We 2057, from GORILA Vol. 3), a clay disc with one or more impressions. They were used as the "conveyance of a commodity, either within the central administration or between the central administration and an external party" [5].

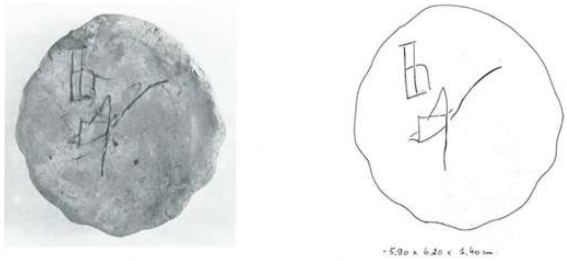


Figure 1: Roundel KH We 2057, from GORILA Vol. 3

The difficulty in deciphering Linear A begins with the number of samples available for analysis: The corpus of Linear A is, as of now, very small. There are 1427 artefacts with Linear A inscriptions, with signs appearing around 7400 times [5]. This may seem sufficient, but it is a small amount when compared to the Linear B corpus which appears on more than 4600 artefacts with signs occurring around 57000 times.

Most decipherment attempts begin by provisionally assigning Linear B phonetic values to Linear A signs which are graphically similar. This is done because Linear B, deciphered by Michael Ventris in 1952, was modelled on Linear A- a conclusion which can be drawn due to the shared signs between the writing systems. By doing this, we are able to 'read' the signs of Linear A, but the reliability of doing so is compromised. For one, the time difference between the use of the two writing systems is very vast, and Linear B encodes Mycenean Greek. Attempts to link Linear A to Greek have been thus far inconclusive, producing meaningless 'words'[6], with an additional issue being that 80% of Linear A signs are unique [7] and thus have no clear phonetic equivalent. These unsuccessful attempts to link Linear A to Greek have also affected its likelihood of being an Indo-European language. The 'Minoan' that the writing system encodes appears unrelated to any language, allowing a vast number of languages to be proposed as possible relations. Among the proposed language relations have been Greek [8], Etruscan [9], Sanskrit [10], and various Semitic Languages [11], [12].

Following the graphic similarity to Linear B, many scholars have postulated different hypotheses to identify the language. The first attempt to decipher Linear A was proposed by Vladimir I. Georgiev in 1957 to bear a Greek relation, and Gregory Nagy would continue this line of inquiry in 1963, presenting arguments on the phonetic-graphemic, lexical, and semantic levels. Nagy shows evidence of "varying worth" [8] of these Greek-like, or Indo-European, elements. In his paper, he claims consonant clusters such as I-JA-TE (which he had identified in Graffito II 12 of Phaistos, as *ne-ma i-ja-te*), and can be observed in Greek, are of "clearly Indo-European origin", and he speculated later in his paper (209) that Luwian, a language belonging to the now-extinct Anatolian branch of the Indo-European family, was the language in question. Despite proposing several examples that show this Indo-European connection, Nagy's claim is weakened because similar findings have been highlighted in other language families, a point that will be discussed later. Adopting Youngers' stance on this method of determining language family relationships, the weakness of Nagy's method of decipherment lies in the use of vocabulary to identify a language. This is because vocabulary is prone to being borrowed and the examples given by Nagy may not actually be from Linear A / Minoan [7]. Towards the end of his paper, Nagy speculates that there could be a relation between Linear A and Luwian, an Indo-European language. This would then be developed by later scholars.

Based on the Linear B phonetic values, Leonard R. Palmer theorized that Linear A could be the writing of an Anatolian language, possibly Luwian, or a Cretan variant of Luwian. Palmer posited this mostly because he believed that "Greece and Crete were twice invaded by Indo-European people during the second millennium BC" [13], an event which would have sparked a mass migration to Crete. This theory was based on two elements: the presence of "Minyan" ware in Beycesultan (Western Anatolia), and a Linear A inscription that he interpreted as "Mount Parnassos" and, according to him, was based on Luwian. In Luwian, 'Parnassos' means '(place) of the temple'. He also based this theory on the inscriptions found on "vessels of Minyan shapes",and claims to have recognized Luwian deity names in Linear A on, for example, the Pylos tablets. One such example is Pylos FR1227, where Palmer claims to read *wa-na-so-i pa-se-da-o-ne*, believing it to mean 'The two Queens and Poseidon'. This Indo-European link found another supporter in was Gareth Alun Owens, who released a collection of essays entitled *Kritika Daidalika* [14], and suggested a similar relation to Luwian but as an archaic relative. Using Linear B phonetic values, he detailed 50 words of the Minoan language which he 'deciphered' [10] Owens attributed, to two Linear A inscriptions, the values *ja-di-ki-te* and *i-da*. These he linked to the two holy mountains in Crete (Dikte and Ida). He stated that these words had "good Indo-European etymology". *I-da*, in particular, he proposed, was very similar to *i-na-/i-ja-*, and suggested that these words "(come) from the same root and indicate 'holy'" – a root which he connected to *ieros* in Greek and *isirah* in Sanskrit, 'proving' that Linear A must encode an Indo-European language. Owens postulated that Linear A represents a language from the *Satem* branch of the Indo-European family with "closer lexicographical characteristics with Greek and Sanskrit, more than with Hittite".

The theories of a Luwian (or Indo-European) connection were proposed various times over the years, but never gained consensus among the academic community. Palmer is first criticized for the heavy reliance on his interpretation of the tablets, which can have varying interpretations because of an incomplete understanding of the orthography. Immerwahr's 1963 critique of Palmers' work also touches on the issue of the Minyan ware, expressing how "few prehistoric archaeologists will accept this premise that the Minyans were Luwians and that the Indo-European migration that marked the end of the Early Helladic was not yet a Greek migration", also citing a shortage of archaeological evidence. Mylonas (1962) also challenges the Luwian theory on various grounds, echoing Immerwahr's view that there is a large amount of doubts about whether the Beycesultan people were Luwians, citing that Palmer's evidence of the Minyan Ware was not sufficiently qualified [15]. Minyan Ware is identified based on characteristics such as colour and form features and as a result of this, establishing the development of the pottery is difficult. Palmer's theory also relies on an invasion in 1700BC, which coincides with the use of Linear A and resulted in the naming of a mountain, *Parnassos*. However, there is still a lack of concrete evidence of when Mount Parnassos was named and, additionally, no archaeological evidence that the Luwians used the area at all as a place of worship (recall that Parnassos is supposed to mean '(place) of the temple'). A mountain could not have possibly be named after a temple that did not yet exist, which further bolsters the lack of physical evidence to support Palmer's linguistic evidence. There is, all in all, very little evidence pointing to anything other than trade contact between Luwians and Minoans and, therefore, it is unlikely that their languages would be related. Additional reasons for the rejection of the connection include the small states along the Western coast of Asia Minor that would have been natural barriers to the contact between the Luwians and Minoan Crete, and no remarkable resemblance between Minoan and Luwian morphology [16].

The second major language family of interest for Minoan is the Semitic language family, first proposed to be connected with Linear A by Cyrus H. Gordon (1966). Like most scholars working on the topic, Gordon applied the phonetic values of Linear B to the Linear A samples and found give words identified by Ventris and Chadwick [17]: *su-po*, *ka-ro-pa*, *pa-pa*, *su-pa-ra*, *pa-ta-qe* (all accompanied with pot signs), as well as the commonly found *ku-ro* at the end of administrative tablets. Gordon, who had extensive knowledge of the Semitic languages and worked specifically with Ugaritic, recognized that three of these vessel words show consonantal roots that exist in Ugaritic: *sp*, *krpn*, and *spl* (matching the first, second, and fourth words listed previously). Following this success, Gordon would continue to identify words in Linear A that were recognizable in various languages belonging to the Semitic language family, like Akkadian and Hebrew, eventually believing Linear A was connected specifically to West Semitic. Western Semitic is a good candidate for relation, as dialects of it were spoken along the Mediterranean seaboard, an area which is geographically close to Crete. In a lecture based on Gordon's initial findings, Maurice Pope (1958) gave a lecture that bolstered the possibility of Semitic as the language not only by corroborating some of the words that Gordon had identified, but also pointing out certain Semitic grammatical features, such as the presence of a copula on tablets 117a, and 122a & b, where *u-* can be found at the beginning of the second word consistently. This is important because in Akkadian and ancient Hebrew where 'and' is denoted by *u* and *waw*, showing a possible connection to Semitic. Of course, this is by no means conclusive; however, the presence of grammatical inflection [18] on top of this identification seemed to only further promote the connection. The word *kuro* is also commonly raised as an indicator of, at a basic level, some Semitic influence – it is the only word in Linear A whose meaning is the most probable under the Semitic theory, meaning 'total'. Present archaeological evidence does not rule out Semitic influence, but, at the same time, it does not fully support Semitic influence either. Jan Best (1972) would continue Gordon's attempts, presenting a controversial paper promoting Linear A as the script of a Semitic language, closely related to Ugaritic [19].

Language contact with Semitic is, at the very least, a possibility. Minoans traded all over the Eastern Mediterranean, and there has been evidence of cultural contact in places like Cyprus, Canaan (located in present-day Lebanon, Syria, Jordan, and Israel), and the Levantine Coast. Minoan-Style wall paintings were also discovered in 2009 in Tel Kabri in Israel. In Tel Kabri, remains of a Canaanite city from the Middle Bronze age (2000-1550 BC) coincide with the time Linear A was in use, and Canaan is a Semitic-speaking region. Kamares Ware (a distinct type of Minoan Pottery that reached its peak in popularity around MMIIB, about 1750 BC) has also been found in many Egyptian sites including the Delta, Middle Egypt, and Aswan in Upper Egypt [20]. Evidence of Middle Minoan pottery (dating 2100BC to 1500BC) can also be found in the Aegean Islands, the Near East (the countries of the Arabian peninsula), Mesopotamia, and Anatolia, showing how much the Minoans traded in the surrounding regions, increasing the possibility for language contact.

No theory comes without controversy- much like the Luwian hypothesis, many scholars reject Semitic as a possibility for the language of Linear A. Gordon found approximately 50 words, and the reliability of these matches is compromised because all the words identified are vocabulary items. As mentioned previously, vocabulary items are not considered a reliable means of identifying a possible family connection. Packard (1974) also points out the difficulty in connecting the five words (*su-po*, *ka-ro-pa*, *pa-pa*, *su-pa-ra*, *pa-ta-qe*) with Ugaritic names because of how vowels are ambiguous in Semitic writing [21]. Additionally, because trade was so prominent, these word strings, found on administrative tablets, could have just been loanwords from the surrounding regions. Chadwick also rejected the Semitic theory, stating that "if the vowels are ignored we are leaving out half the information presented by the script". This is because, in Semitic languages, vowels could be considered 'semi-vowels' with a specific 'colour'. The common criticism of Gordon's work also stems from the fact that he linked various elements to not one Semitic language, but several – Canaanite, some Aramaic, some Akkadian, and so on. This apparent lack of any specific Semitic language prompted the view by many scholars that Gordon's work was not successful in establishing a Semitic link.

More recent decipherment attempts have turned to algorithmic approaches, in the hope that computerized, automated efforts would be more efficient in generating more matches [22]. Revesez (2017) proposed that the language of Linear A was connected to the Uralic family, and unlike previous attempts, presented an algorithm which would "find the syllabic values of the Linear A symbols". Taking these values, Revesez then uses the proposed Linear A values to build a Uralic-Minoan dictionary which then is used to 'translate' twenty-eight Linear A documents from GORILA. This novel new approach allowed Revesez to 'read' close to 30 sets of inscriptions and propose a dictionary. However, the problem of biased interpretation may remain for two reasons. Firstly, Revesz explicitly set out to prove the hypothesis that the Minoan language could be linked to the Uralic family. The determination of the Syllabic values of Linear A were carried out specifically with other proposed languages of the family and were based entirely on the graphic similarity. Cross-family

comparisons were not made to evaluate the relative likelihoods of the Minoan's language relation to one family over another's. Next, words seem to fit in the most plausible positions, but this has been done without any consideration of the provenance of the artefacts that contain the clusters. Such an approach is incomplete as many contextual clues which can help evaluate the relative validity of interpretations, and hence debunk certain interpretations which might at first seem tenable, were not considered. It is also interesting to note that the examples of translation he had used in his paper were restricted to the libation tables and objects, and no attempt was made using the economic tablets that can be found in GORILA volumes 1 and 3, which have a known and agreed upon context.

III. Methodology

A. Document Preparation

Documents for comparison needed to be created and sorted for input into the program. Two major lists were created: one that contained all usable Linear A samples and another for the dictionary or lexical list it was being compared to. Samples were drawn from Godart and Olivier volumes 1, 3. Volume 2 was excluded because it included mostly individual signs, which could not be formed into strings for comparison. Volume 4 was left out because it records libation tables and could use a ritual language, a version of the language that would otherwise not be used outside of its purpose [23]. Various dictionaries and lexical lists also needed to be converted into a digital format in order to be used by the program.

B. Intended Program Logic

There are various elements and variables that need to be considered when designing the program and the logic it runs on in order to give us lists that can then be used for a manual translation attempt. In a previous paper, I attempt a manual translation and matching via root comparison: the shortfalls of this method have since been accounted for. The considerations and details for each step have been included in the methodology.

Two excel files are prepared: one containing samples of Linear A from GORILA 1, 3 and 4, and another with words pulled from the various dictionaries listed above in **Error! Reference source not found.**.

The program then draws from Linear A master list that contains the samples, and **splits word strings into 2, 3 and 4 phone long chunks.** Phones are defined in Linear A according to the individual symbol and its Linear B phonetic equivalent. For example, IO ZA 1 from GORILA 4 was initially transcribed with the Linear B phones as per below, with dashes separating the phones of individual symbols, and x's demarcating places where symbols were missing or unclear, due to the age of the sample:

A-TA-I-A301-WA-JA x JA-DI-KI-TU x JA-SA-SA-RA-[x x x]-SI x I-PI-NA-MA-x

In order to make it suitable for the program, we reformat the string into something like this:

A-TA-I-[A301]-WA-JA[]JA-DI-KI-TU[]JA-SA-SA-RA-[]-SI[]I-PI-NA-MA

Gaps between 'words' or missing signs are demarcated with []. One of the biggest problems on applying this method to Linear A is the presence of symbols that do not have a phonetic equivalent in Linear B. Ideally, the program would be able to apply the search and consider any dictionary entry which matches other elements of the string a positive match.

A-TA-I-A301
A-TA-I
A-TA
TA-I-A301-WA
TA-I-A301
TA-I
I-A301-WA-JA
I-A301-WA
I-A301
A301-WA-JA
A301-WA

All the above should generate as a part of the splitting of word strings for the string A-TA-I-A301-WA-JA. The program should not loop to the first syllable and should do this for all of the words In the list.

Next, the program then needs to make adjustments in the Dictionary List so that it can be compared properly. After accounting for variables, a new dictionary list is generated for comparison. For example, in the Hamito Semitic Dictionary, capital letter V stands for a variable vowel. This means: abVnan → abanan → abenan → abonan etc.

For the book which suggests Basque as Proto-Indo-European, C is for a variable consonant.

The unique variables of each dictionary or language need to be accounted for. Some languages also feature use of 'special characters' (i.e. θ or ð), and so the program must be able to read those. That said, we have encountered no dental fricatives in the data. An important consideration for each individual test language are the C and V's that we use as variables- for example, not all consonants of the English language are valid consonants in other languages, and as such, these differences need to be accounted for by the program.

The program then compares items from (1) with items from (2) and outputs a file which shows all the matches. These matches can then be taken and manually processed based on their probability, obtained from the number of matches we got from the program.

The basic structure of the program is implemented to compare two spreadsheets for similarities. One spreadsheet contains Linear A transcriptions, while the other one the entries from dictionaries. In Python context, the module "pandas" is often imported for this purpose. The module helps to check if the two dataframes have the same shape and elements. Then the module "numpy" should be imported to find out the index of the cells where the value is "True". Alternatively, the module "pandas.DataFrame.equals" helps to find out the elements with exactly the same values.

C. Frequency Analysis Based on Online Corpus of Linear A

In the recent months, the Linear A corpus has been digitalized online by Robert Hogan, called the Linear A Explorer. Recording basic information of some of the Linear A samples, it also features commentary (where available) by John Younger. The explorer is able to provide a frequency

analysis by matching recurring clusters, and is an incredibly useful tool that can now be used to cut down the time it takes to identify recurrences- hovering over a word cluster informs the user of any matches throughout the rest of the corpus. While useful, it is important to note that its largest limitation is that clusters with any kind of variation are ignored by its search program. For example, in GORILA Volume 4, a common cluster amongst the libation tables is string A-TA-I-A301-WA-JA. On the explorer, there are a recorded 7 instances of this exact string. However, manual analysis produces additional results such as A-TA-I-A301-U-JA and A-TA-I-A301-WA-E that have similar if not identical preceding strings to their A-TA-1-A301-WA-JA counterpart. This does not mean the frequencies showed are unreliable, rather, they must be selected with care. Samples from GORILA 1 and 3, as mentioned previously, are record the transaction of commodities. As such, these vocabulary items are more easily identified for their length (by cluster) and are likely to vary less. As such, strings with higher frequencies that we see on the explorer can be reliably used for experiment with the program and to double check the results generated by the program itself.

IV. Conclusion

The application of this program to Linear A represents a key move away from previous, philological attempts. A large majority of studies so far have relied on outward resemblances with words from other languages. It attempts a method similar to that of Revesez (2017), while using the phonetic values from Linear B. The current program, while simple, allows Linear A to be compared to a variety of languages from the Semitic family, expanding the matching process beyond word recognition. In addition, changes to the program can be made to accommodate comparison with other language families. The program offers the opportunity to narrow the candidate for a language family through a larger, statistics-based process instead, taking into account variables and the full corpus of comparison language. This, of course, is not perfect- to quote Yves Duhoux from March 1998, "The conclusion must be that even if one can find casual resemblances between words in both languages (remember this MUST statistically happen) …they are probably structurally different." The overall phonological and more importantly, morphological system must be resolved before a complete conclusion can be drawn. The program can be eventually adjusted to make matching other language-families possible. This program aims to hasten the process by 'brute force', but would aid in future research efforts.

References

[1] G. Cadogan, *Palaces of Minoan Crete*. Routledge, 1976.

[2] B. Davis, "Introduction to the Aegan Pre-Alphabetic Scripts," University of Melbourne, Australia, Melbourne, Australia, Paper 1, 2010.

[3] F. Perono Cacciafoco, "Linear A and Minoan The Riddle of Unknown Origins," Nanyang Technological University, 17-Jan-2014.

[4] Anadolu Agency, "Mycenean artifacts found in Bodrum," *Hurriyet Daily News*, Turkey, 10-Nov-2014.

[5] I. Schoep, "Social and Political Organisation on Crete in the Proto-palatial Period: The Case of Middle Minoan II Malia," *J. Mediterr. Archaeol.*, vol. 15, Jun. 2002.

[6] L. Godart, "Du Lineaire A au Lineaire B," in *Aux origines de l'hellénisme: La Crète et la Grèce.*, vol. 15, Paris: Publications de la Sorbonne, 1984, pp. 121–128.

[7] J. Younger, "Linear A Texts: Homepage," *Linear A Texts and Inscriptions in Phonetic Transcription & Commentary*, 30-Nov-2000. [Online]. Available: http://people.ku.edu/~jyounger/LinearA/#7c. [Accessed: 14-Dec-2017].

[8] G. Nagy, "Greek-like elements in Linear A," *Greek Roman Byzantine SStudies*, vol. 4, no. 4, pp. 181–211, 1963.

[9] G. M. Facchetti and M. Negri, *Creta minoica: sulle tracce delle più antiche scritture d'Europa*. L.S. Olschki, 2003.

[10] G. A. Owens, "The Structure of the Minoan Language," *J. Indo-Eur. Stud.*, vol. 27, pp. 15–56, 1999.

[11] C. H. Gordon, *Evidence for the Minoan Language*. Ventor Pub, 1966.

[12] J. G. P. Best, *Some preliminary remarks on the decipherment of Linear A*. Amsterdam : Hakkert, 1972.

[13] L. R. Palmer, *Mycenaeans and Minoans; Aegean Prehistory in the Light of the Linear B Tablets*. Faber & Faber, 1961.

[14] G. A. Owens, *"Kritika Daidalika": Evidence for the Minoan Language: Selected Essays in Memory of James Hooker on the archaeology , epigraphy and philology of Minoan and Mycenaean Crete*. Hakkert, 1997.

[15] G. E. Mylonas, "The Luvian Invasions of Greece," *Hesperia J. Am. Sch. Class. Stud. Athens*, vol. 31, no. 3, pp. 284–309, 1962.

[16] F. Perono Cacciafoco, "Linear A and Minoan: Some New Old questions," *Analele Univ. Din Craiova Ser. Ştiinţe Filol. Lingvistică*, no. 1–2, pp. 154–170, 2017.

[17] M. Ventris and J. Chadwick, *Documents in Mycenaean Greek*, 2nd ed. Cambridge [Eng.]: University Press, 1973.

[18] M. Pope, "On the Language of Linear A," University of Capetown, 1958.

[19] G. A. Rendsburg, "'Someone Will Succeed in Deciphering Minoan': Cyrus H. Gordon and Minoan Linear A," *Biblic. Archaeol.*, vol. 59, no. 1, p. 36, Mar. 1996.

[20] P. Bradley, *The Ancient World Transformed*. Cambridge University Press, 2014.

[21] D. W. Packard, *Minoan Linear A*. University of California Press, 1974.

[22] P. Revesz, "Establishing the West-Ugric Language Family with Minoan, Hattic and Hungarian by a Decipherment of Linear A," *WSEAS Trans. Inf. Sci. Appl.*, vol. 14, pp. 306–335, Nov. 2017.

[23] W. T. Wheelock, "The Problem of Ritual Language: From Information to Situation," *J. Am. Acad. Relig.*, vol. 50, no. 1, pp. 49–71, 1982.

Generating Derivational Relations for the Japanese WordNet: The Case of Agentive Nouns

Francis Bond and Ryan Lim Dao Wei
Linguistics and Multilingual Studies
Nanyang Technological University
Singapore
bond@ieee.org RLIM040@e.ntu.edu.sg

***Abstract*—This paper presents work in progress on the development of derivational links for the Japanese WordNet, with a focus on the retrieval, validation and elaboration of nouns and verbs linked by the agentive noun derivation. 2,340 such links are generated, of which we validated 833 such pairs. We briefly discuss some challenges in determining valid link pairs as well as their morphosemantic natures. We also consider the possibilities and challenges of automating the discovery of morphosemantic links, by linking our results with current theoretical issues in agentive nominals. In addition, we are currently corroborating these Japanese agentive derivations with English counterparts from the Princeton WordNet and intend to perform a more rigorous cross-lingual comparison**

Keywords—Wordnet, Japanese, derivational morphology, morphosemantic relations, agent nouns

I. INTRODUCTION

Most WordNets originate as a loose collection of what is essentially separate nets split according to parts of speech. The development of the Japanese Wordnet (**wnja**) [1] is no exception – networks of verbs and nouns respectively form its trunk, supported by smaller ones covering adjectives and adverbs. Developmental efforts are concentrated on expanding **wnja**'s lexical coverage, such as the addition of synsets currently absent in the Princeton English WordNet (from which **wnja** takes its relational structure), as well as the correction of erroneous entries.

One feature is absent from **wnja** is the presence of links between derivationally related forms of the same word. These were originally omitted due to a lack of resources. Information about derived forms of words can be isolated within the affixes and as such they do not necessitate separate entries in the WordNet [2]. With the development of many Wordnets globally now striving towards the addition of derivational relations, many have also realized the potential of discovering morphosemantic links [3] One key motivation concerns the ability of Wordnets to facilitate automated information retrieval tasks, and the capacity to make textual inferences.

This project seeks to generate derivational relations for Japanese WordNet, specifically accounting for the agentive noun derivations (such as between the appropriate senses of the verb *work* and the noun *worker*). Agentive constructions have been observed cross-lingually [4] and for lexical resources like WordNet, such agentive constructions form a crucial interface between mostly separate verb and noun networks.

Although the term agent noun is often used refer to such constructions, a growing consensus suggests that across independent language families, a non-negligible portion of such nouns are not agents at all. This paper therefore refers to such constructions as the agentive derivation, whereas the term agentive noun (AN) is used to refer to the output.

II. AGENTIVE NOUNS IN JAPANESE

The three main derivational processes in Japanese behind the formation of agentive nouns are given by [5]:

(1) addition of an agentive suffix: e.g., 労働-する *roudou-suru* 'work' > 労働-者 *roudou-sha* [work-person] 'worker'; 書く *kaku* 'write' > 書き-手 *kaki-te* [write-hand] 'writer'

(2) compounding between nouns and infinitive verbs: 酒 *sake* 'alcohol' + 飲む *nomu* 'drink' > 酒飲み *sake-nomi* 'heavy drinker'

(3) conversion from verbs to nouns: 見張る *miharu* 'watch' > 見張り *mihari* 'watchman, guard'

The creation of agentive nouns through process (ii) and (iii) is largely restricted to native Japanese words. More importantly, derivation through compounding and conversion is significantly less productive than suffixation, with the majority of agentive nouns in the Japanese lexicon being formed through suffixation. As such, this study looks only at agentive nouns of the suffixation type.

This study is concerned with the behaviour of seven agentive suffixes. (1-6) refer to humans involved in the event of the base, while (7-8) refer to implements that facilitate the action denoted by the base.

(1) 者 *sya* 'person'
(2) 人 *nin* 'person'
(3) 手 *shu* lit. 'hand' (metonymical extension)
(4) 員 *in* 'personnel'
(5) 家 *ka* lit. 'house' (metonymical extension
(6) 士 *shi* lit. 'warrior' (metaphorical extension)
(7) 機 *ki* 'machine'

(8) **器** *ki* ‘gadget’

At this stage, a further semantic distinction can be made amongst the suffixes denoting humans. (1-2) largely refer to persons performing an action at a specific place and time, although this is not without exception. (3-6) denote persons involved in actions which need not necessarily have happened, but instead events that regularly occur (3). Alternatively, suffixes convey information about the person, with (4) marking a person operating in a professional capacity, (5) indicating specialized knowledge and expertise and (6) indicating a qualification. A tentative distinction can also be made between (7-8) despite them sharing the same phonetic value *ki*. -器-suffixed objects tend to have a simplex operation, often mechanical in nature or requiring human control as in 緩衝-器 *kanshou-ki* ‘damper/shock absorber’ as well as 消火-器 *shouka-ki*, ‘fire extinguisher’. In contrast, -機 agentive nouns facilitate the action of the verbal stem in comparatively complex and automated ways, such as 洗濯-機 *sentak-ki* ‘washing machine’ or 販売機 *hanbai-ki* ‘vending machine’.

III. METHODOLOGY

Most existing wordnets adopt either one of three strategies to introduce derivational links. The Princeton WordNet 3.0 automatically identifies suitable noun pairs from base verbs, as well as from existing pairs which share similar semantics. Several wordnets intending to develop derivational relations import the PWN’s set of relations and align with it, for instance BulNet [6] Alternatively, new synsets are suggested by deriving new words from existing ones, sometimes with the help of morphological analysers, such as the Wordnet Bahasa [7]

Our method largely resembles that of the PWN. A program was written in Python to generate links between verb and noun synsets. All lemmas of **wnja**’s verb synsets were extracted and analysed. Decomposing the lemmas required two distinct processes, depending on its syntactic status as a native-Japanese verb or Sino-Japanese verbal noun. Verb lemmas were first converted into their infinitive form (*kaku* > *kaki*); the infinitive form was then suffixed with the relevant agentive suffixes (*kaki-te*). For verbal noun lemmas such as *roudou-suru*, the support verb *-suru* was first discarded, and the remaining verbal noun was then suffixed as well. If the resulting agentive construction could be matched to an identical lemma in any noun synset, a link between the synsets was automatically generated. In total, 2340 candidate links were generated between 991 unique verb synsets and 486 unique noun synsets. Within the noun synsets, there were 697 unique lemmas.[1]

The links were then manually checked and determined to be either valid or invalid. We made use of a simple entailment relation between the noun synset and verb synset. A link was judged valid if the activity, action or state suggested by the noun synset entailed the activity, action or state represented by the verb synset. For instance, synset 03251766-n ‘an appliance that removes moisture’ (乾燥機 *kansouki* “dryer”) was determined to have a valid agentive link with 00218475-v ‘remove the moisture from and make dry’, but not with 00212790-v ‘preserve by removing all water and liquids from’, since removing moisture does not imply the intention of preservation. This also follows the general expectation that agentive nouns, particularly those formed through productive means, tend to be more abstract than the verb senses from which they originate.[2]

If a given link between two synsets was found to be valid, then the same link was deemed valid for all proposed sense pairs. This was driven by two considerations. The first was a general assumption that the lemma provided within each synonym set are attested, though glaringly erroneous entries were flagged out and designated for removal from **wnja**. The second concerns the purpose of this exercise, which is the enhancement of **wnja** by establishing as many links between related senses, while bearing in mind that existing lexical relations (hypernymy, hyponymy etc.) in WordNet are formed between synsets and not lemmas.

IV. RESULTS

Of the 2,340 candidate links checked, 833 links were deemed valid and suggested for addition into Japanese WordNet. The links were distributed across 426 unique noun synsets, which covers 87.85% of the total noun synsets linked. Table 1 shows the effectiveness of our link generation:

Total unique lemma	**697**
Error in lemma 圧搾器 ‘a dense crowd of people’; -器 suffix for human referent	13
Not agentive noun 押し手 *oshi-te* ‘left hand’; analyzed as 押す ‘push’ modifying 手 ‘hand’	5
Unique agentive noun lemma	679
Linked to at least one verb sense 印刷+する *insatsu-suru* ‘reproduce by printing’ to 印刷-機 *insatsu-ki* ‘machine used for printing’	610
No suitable link found	69

Table 1: Breakdown of unique lemma

[1] A unique lemma is defined as a lexical item with a distinct sense. For example, a polysemous word like 勤労者 *kinrousha* ‘a person who works at a specific occupation’ (09632518-n), ‘a member of the working class’ (10481711-n) constitutes two unique lemma

[2] The average noun has 1.23 senses, and the average verb has 2.16 senses [8]

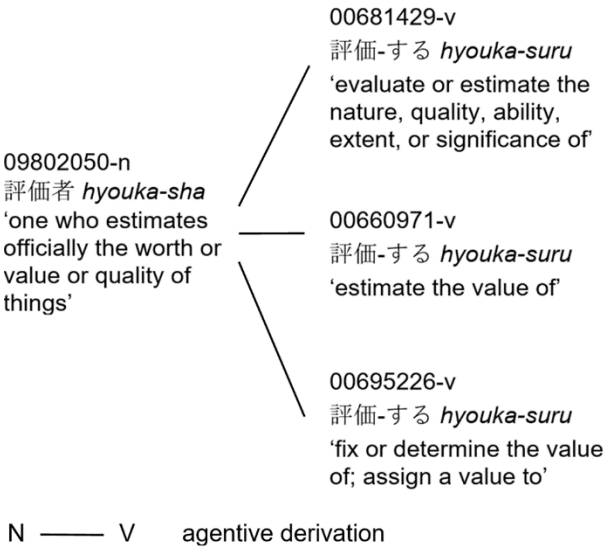
833 agentive links were generated for 610 unique lemmas, yielding a link-lemma ratio of 1.37. This is a satisfactory result, for it is not desirable to bloat lemmas with an excess of links for a single derivation type. Yet, neither should we expect to achieve a perfect one-to-one mapping between verbs and agentive nouns. At the same time, we observed a marked difference between the link-lemma ratios of certain noun supertypes.

	Links generated per lemma				Link-lemma ratio
	1	2	3	≥4	
ARTIFACT	98	9	3	0	1.1364
PERSON	320	129	23	9	1.4277

Table 2: Link generation in ARTIFACT and PERSON-supertype nouns

Though we did not predict differences between these noun supertypes with regards to their link-taking behavior, that ARTIFACT supertype nouns tend to produce nearly one-to-one mappings is easily explainable. ARTIFACT supertype nouns almost always denote machines, which are built with specific functions in mind and referred to as such. Their definitions can therefore be expected to contain a more restricted (singular) set of entailments in terms of the actions, states or events their referents participate in.

It is possible to also extend this and suggest that PERSON supertype agentive nouns have a looser relationship with their referents, and therefore entail a larger set of actions, states or events from which agentive constructions derive from. At the same time, the potential of a non-negligible number of lemmas being congested with derivational links (six being the highest) is an unsatisfying one. For instance, consider the agentive noun below and the links deemed valid:

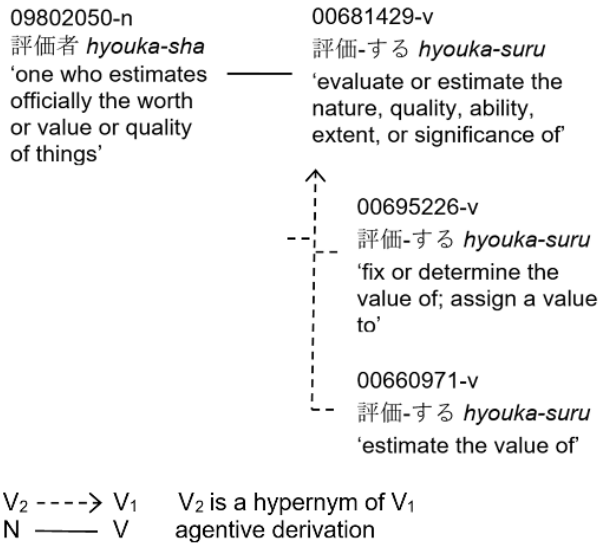


It becomes apparent that this is not a case of a verb with three heterogenous senses, because these senses demonstrate the potential to be organized along hypernymy/hyponymy relations within the verb supertype of COGNITION. Therefore, generating three derivational links seems unwieldly and has the potential to be summarized. We formulated a brief rule to simplify the process of link generation in such instances:

Given that

1. an agentive nominal sense AN has a valid link with *n* verbal senses V_1, V_2, … V_n
2. There is a hyponymy relation such that $V_1 < V_2 < \ldots < V_n$,
3. V_1, V_2, … V_n all belong to an identical verb supertype[3],

An agentive link between AN and V_1 can be extrapolated to also hold true for V_2 through V_n, as illustrated below:



One immediate benefit is that it becomes necessary to posit only one derivational link (to the smallest hyponym amongst the set of verbs), significantly reducing the link-lemma ratios and decongesting lemmas such as the ones given above. In addition, modeling agentive derivations using hierarchical verb relations (as compared to linear ones) also reinforces the ability to make automated textual inferences such as the following, one of the goals identified by [3]

[3] There are several verb senses belonging to different supertypes that nevertheless demonstrate very plausible hypernym/hyponym relationships. For instance, 訪問+する *houmon-suru* 'come to see in an official or professional capacity' (MOTION) can be regarded as a hyponym of 訪問+する 'pay a brief visit' (SOCIAL). However, we recognize that the former represents a fuzzy example of MOTION rather than a defect of verb supertypes, and that a conservative approach to positing internal verb relations is largely beneficial for maintaining transparent agentive derivations.

Among the 161 PERSON supertype lemmas with more than one derivational link, 102 of them, including the two examples above, can be organized using such a hierarchical approach. The link-lemma ratio also reduces itself to 1.2058, a figure more consistent with the other major ARTIFACT supertype. What this also reveals is that PERSON and ARTIFACT supertype nouns do not differ significantly in terms of their granularity or specificity, as our initial ratios would lead us to believe. Rather, a greater number of sense distinctions are being made for verbs with PERSON-supertype arguments, most likely to specify different contexts, mannerisms and intentions (which we do not expect to observe in their ARTIFACT-supertype counterparts). These sense distinctions can in fact be arranged according to hypernym/hyponym relationships, from which we can distil a minimal number of agentive derivations which hold true for hypernymous verb senses.

Of the 69 lemmas without a suitable agentive link found, it is interesting that 25 (36.23%) of them originate from the native Japanese verb stratum. While this number may seem trivial, it must be considered alongside the fact that native-stratum lemmas form a much smaller proportion of the overall unique lemma (10.76%). Examples of Japanese lemmas with no suitable verb to form a derivational link with include:

落ち-人 *ochi-bito* 'an exile who flees for safety'

変り-者 *kawari-mono* 'a person with an unusual or odd personality'

掛かり-者 *kakari-mono* 'a person who relies on another for support (especially financial)[4]

These agentive nouns tend to have highly idiomatic meanings or have undergone some degree of conventionalization. In either case, it was not possible to trace the agentive senses back to their original verbal sense(s), demonstrating the opaque relationship between the two. Furthermore, locating the referents of native-stratum ANs within the argument structures of their verbal stems is a highly irregular task. Consider the case of synset 10765679-n 'someone who lives a wandering, unsettled life', with lemma 放浪者 *hourou-sha* and 渡り者 *watari-mono* from the Sino-Japanese and native stratums respectively. While the following syntactic inferences can be made for *hourou-sha*, the same inference is unnatural for *watari-mono:*

ジョン は <放浪-者/渡り-者> です
jon wa <hourou-sha / watari-mono> desu
'John is a <wanderer / wanderer>'

彼は<放浪する / ?渡る>人という意味です
kare wa <hourou-suru / ?wataru> hito to iu imi desu
'This means he is someone who <wanders / wanders>'

This is not due to *hourou-suru* and *wataru* having inherently divergent verbal senses. Rather, *wataru* requires other arguments to be present (either explicitly or contextually) in order to disambiguate the specific, intended sense of 'wander about' from its more general sense of 'traverse/cross (as in a river)'. In turn, the agentive noun *watari-mono* takes on a more idiomatic meaning, since these other arguments inherent in *wataru* become less retrievable. While we should not expect agentive nouns and their verbal senses to have identical granularities, simply generating links based on lexical similarity does not contribute much to facilitating machine translation and automated inference tasks.

V. MORPHOSEMANTIC RELATIONS

The nature of each valid link was then inspected and a morphosemantic relation was assigned. The types of morphosemantic relations were adopted from a categorization of agentive derivations, formulated by Fellbaum (2009) to account for the similar suffix *-er* found in English.

While morphosemantic relations across different languages and WordNet are shared to a large extent [9]; our Japanese data did not motivate as extensive a categorization as Fellbaum's. While Fellbaum's categorizations cover a variety of semantic relations, including INSTRUMENT, INANIMATE AGENT /CAUSE, PURPOSE/FUNCTION, VEHICLE, LOCATION, UNDERGOER/PATIENT, RESULT/ CAUSE, an overwhelming majority of the valid links in our data set could be classified as either:

AGENT: Agentive nouns of this semantic relation tend to be animate, and their participation in the action, state or event is usually volitional or intentional. For instance:

出品-人 *shuppin-nin* 'someone who organizes an exhibit' < 出品-する *shuppin-suru* 'give an exhibition'
密猟-者 *mitsuryou-sha* 'someone who hunts or fishes illegally' < 密漁-する *mitsuryou-suru* 'hunt illegally'

INSTRUMENT: Besides being non-animate artifacts, instruments are also unlike agents as they require the control or operation of an agent:

濾過-器 *roka-ki* 'bowl-shaped strainer' < 濾過+する *roka-suru* 'remove by passing through a filter'
印刷-機 *insatsu-ki* 'a machine that prints' < 印刷-する *insatsu-suru* 'reproduce by printing'

UNDERGOER/PATIENT: These denote the undergoer of a transitive action performed or initiated by a distinctly different agent.

[4] The agentive suffixes -人 and -者 here are realized as *-bito* and *-mono* respectively, reflecting the native Japanese reading as compared to their Sino-Japanese readings *-nin* and *-sha*. We treat *-bito/-nin* and *-mono/-sha* as allomorphs of the same suffix, although there is by no means any consensus on this issue. A fuller discussion of this is beyond the scope of this paper.

仕置-者 *shioki-mono* ‘someone who has been legally convicted’ < 仕置-する *shioki-suru* ‘impose a penalty on’
登録-者 *touroku-sha* ‘a person who is formally entered in a register’ < 登録-する *touroku-suru* ‘record in writing’

Table 3 demonstrates the breakdown of morphosemantic links between synsets, according to suffix type

	AGENT	INANIMATE AGENT/ BODY PART[5]	INSTRUMENT	UNDERGOER
者 *sya* ‘person’	680	2		18
人 *nin* ‘person’	113			4
手 *shu* lit. ‘hand’	55		3	1
員 *in* ‘personnel’	29			1
家 *ka* lit. ‘house’	63			
士 *shi* lit. ‘warrior’	10			
機 *ki* ‘machine’			64	
器 *ki* ‘gadget’		2	56	

Table 3: Distribution of morphosemantic links

These results are as expected and affirm the productivity as well as regularity of agentive suffixation in general. One should note however, that these categories are useful in describing the morphosemantics of suffixation thus far, and that accounting for agentive forms derived via compounding and conversion may necessitate more categories. For instance, [10]’s proposed classification for noun-verb agentive compounds includes the categories of LOCATION and RESULT[6].

[5] We collapsed the morphosemantic categories INANIMATE AGENT and BODY PART, following Fellbaum’s (2009) suggestion and the observation of practically 2 examples in each of these categories within the data. Unlike instruments, agentive nouns of this category do not require the control of agents to effect actions, states or events. However, they lack volition and intention, for example 発声器 *hassei-ki* ‘an organ involved in speech production’.

[6] As discussed above, compounding is restricted to items from the native Japanese stratum, and our previous observation that native-stratum derivations tend to be more idiomatic also explains its distribution across a wider range of morphosemantic categories.

We can for example generate simple rules to determine the morphosemantic nature of agentive nouns (produced by agentive suffixation): suffixes -家 *ka* -士 *shi* produce AGENTS while -機 *ki* produces INSTRUMENTS, and so on.

However, Table X also demonstrates a need to account for exceptions to potential rules. It is here that some light can also be shed on a wider debate surrounding agentive nominals in Japanese:

A. Non-agent humans

The presence of patient-readings for human agentive suffixes has been noticed by [11] and [12]. These unexpected occurrences largely occur with the general-purpose suffixes -者 *sha* and -人 *nin*. [5] adopts a pragmatic approach to explain the ambivalent interpretations of nouns such as 仕置-者 *shioki-mono* and 登録-者 *touroku-sha* above. These readings are derived from the lexical meaning of the verbal stems, which clearly indicate a certain directionality and transitivity in their actions – 仕置-する *shioki-suru* ‘to convict’ has relatively fixed arguments. Knowing who is the one performing the conviction is not pragmatically interesting, since it is usually understood to be someone of legal power like a judge. On the other hand, the person being convicted can be anybody before a court of law. Thus, there is more information to be delivered. Ono’s argument thus suggests that the semantics of agentive nouns can be predicted by the pragmatic focus of the agentive nouns and suffix.

Ono’s conclusions prove extremely useful in our description of 者 and 人. These two suffixes are general purpose suffixes which convey little information about the referent beyond its status as a human [13] – it is also for this reason that they are the most productive and attach to much more verbal stems than the other human suffixes. The pragmatic focus of -者 and -人 agentive nouns is then located within the verbal stem, and this produces variance between agent and patient/undergoer readings depending on the conventional situations ascribed by these verbs.

As discussed above, suffixes -手 *shu*, -家 *ka*, -員 *in* and -士 *shi* differ from -者 *sha* and -人 *nin* because they convey additional information about their referents beyond the fact that they are human. Broadly, these suffixes indicate a regularly occurring event (such as a job), expertise, professional capacity and qualifications respectively. For example, agentive nouns suffixed with -家 *ka* and -士 *shi* tell us something about the ability to perform said verbal event, which can only be co-referent with an AGENT.

The lone instance of a UNDERGOER/PATIENT human referent[7] within these suffixes appears in 商議-員 *shougi-in* 'someone who gives advice about problems', derived from 商議-する *shougi-suru* 'get or ask advice from'. Pragmatically, the lexical nature of *shougi-suru,* where an act of seeking advice is usually directed towards a professional, would lead us to predict that *shougi-in* should be co-referent with the person seeking advice (the agent): the set of referents for people asking for advice is greater than that of people to seek advice from and more information can be delivered there. However, the suffix -員 then necessarily pivots the referent towards the professional.

What these instances suggest is that the semantic information embedded in human agentive suffixes cannot be separated from the semantic information embedded in their verbal stems, and the combination of semantics is not purely an additive process. The additional information provided by specialized human suffixes (ability, frequency) imposes certain selectional restrictions on the semantic arguments of their verbal stems. Because general-purpose suffixes -者 *sha* and -人 *nin* do not convey such semantic information, these restrictions are absent; argument disambiguation then happens at a pragmatic level wholly within the verbal stem.

VI. CONCLUSIONS & FUTURE WORK

The main deliverable from this project has been the introduction of 833 agentive derivational links, linking 692 unique agentive nouns within 486 unique nominal synsets to their verbal stems. In addition, the hyponymy/hypernymy relation proposed will be ideal not only for tidying up a potential glut of derivational links from unique lemma, but also assist in further enhancing the **wnja**'s verb subnet. At the same time, errors in lemmatization as well as inappropriate or redundant verb senses which contributed to false hits and high link-lemma ratios will be recommended for removal from **wnja**. We hope this can contribute to strengthening the overall ontology of **wnja**. The links will be incorporated into the next release of the Japanese Wordnet.

Following the inspection and manual categorization of the initial set of 2,340 Japanese links, we intend to follow this up with a further inspection of the Morphosemantic Database on Princeton WordNet [3], identifying new derivational links to be introduced. We also plan to perform a cross-lingual comparison between the Princeton database and our nascent database, investigating how consistently these links are replicated across multilingual synset entries.

For this exercise, our adoption of the morphosemantic relations used by [3] is partially motivated by the desire to further integrate **wnja** into the Open Multilingual WordNet (OMW). At the same time, it is also desirable for these morphosemantic relations to reflect certain language-dependent features found in Japanese and to also integrate the morphosemantic categories proposed for other agentive noun subsets. We hope to further motivate these morphosemantic relations by developing more rigorous syntactic tests, collapsing or expanding them where necessary.

[7] There is another UNDERGOER/PATIENT agentive noun in 読み手 *yomi-te* 'one in a series of texts for students learning to read', but this is an idiomatic formation (note the derivation from the native Japanese stratum) and decidedly a non-human agentive noun.

VII. ACKNOWLEDGEMENT

We wish to acknowledge the funding support for this project from Nanyang Technological University under the Undergraduate Research Experience on CAmpus (URECA) programme.

VIII. REFERENCES

[1] Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In Sixth International conference on Language Resources and Evaluation (LREC 2008), Marrakech.

[2] Christine Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database. Boston: MIT Press.

[3] Christiane Fellbaum, Anne Osherson, and Peter E. Clark. 2009. Putting semantics into WordNet's "morphosemantic" links. In Proceedings of the Third Language and Technology Conference, Poznan, Poland. Reprinted in: Responding to Information Society Challenges: New Advances in Human Language Technologies. Springer Lecture Notes in Informatics, volume 5603, pages 350–358.

[4] Hans C. Luschützky and Franz Rainer. 2011. Agent-noun polysemy in a cross-linguistic perspective. *Language Typology and Universals* 64(4):287-338.

[5] Naoyuki Ono. 2016. Agent nominals. In Taro Kageyama and Hideki Kishimoto, editors, *Handbook of Japanese Lexicon and Verb Formation*, pages 599-629. Berlin: De Gruyter Mouton.

[6] Kiril Simov and Petya Osenova. 2010. Constructing of an ontology-based lexicon for Bulgarian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

[7] Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open Wordnet Bahasa. In Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25), pages 258–267, Singapore

[8] Dan Jurafsky and James H. Martin. 2018. Speech and Language Processing, 3rd Edition. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

[9] Orhan Bilgin, Ozlem Cetinoglu and Kemal Oflazer. 2004. Morphosemantic relations in and across Wordnets – a study based on Turkish. In *Proceedings of the 2nd Global Wordnet Conference (GWC 2004),* pages 60–66. The Global WordNet Association.

[10] Yoko Yumoto. 2016. Conversion and deverbal compound nouns. In Taro Kageyama and Hideki Kishimoto, editors, *Handbook of Japanese Lexicon and Verb Formation*, pages 311-345. Berlin: De Gruyter Mouton.

[11] Yoko Sugioka. 1986. Interaction of derivational morphology and syntax in Japanese and English. (The University of Chicago dissertation, 1984). New York: Garland.

[12] Taro Kageyama. 2002. Dōsashumeishi ni okeru goi to tōgo no kyōkai [The boundary between lexicon and syntax in agentive nominals]. *Kokugogaku* [Japanese Linguistics], 53: 44–55.

A Python Library for Deep Linguistic Resources

Michael Wayne Goodman
School of Humanities
Nanyang Technological University
Singapore
goodmami@uw.edu

***Abstract*—This paper describes PyDelphin: an open-source software library, coded in Python, for working with the resources and results of Deep Linguistic Processing with HPSG (DELPH-IN). These resources and processing results offer rich syntactic and semantic information from linguistically-informed grammars but the original software stack for working with them presents significant technical hurdles for installation and use, particularly for new users and for single experiments. PyDelphin is quick to install and its thorough documentation makes it easy for newcomers to get started. It contains packages that implement a number of DELPH-IN technologies for working with semantic representations, grammar descriptions, tokenization, and corpus databases. It also has client interfaces for related external processors which make it straightforward to integrate these tools into a single workflow. The library has been successfully used in several research projects in topics such as abstractive summarization, machine translation, and natural language generation.**

***Index Terms*—research software, linguistics, semantics, HPSG, computational linguistics, natural language processing, open source software**

I. Introduction

The resources produced by the DELPH-IN Consortium[1] are rich with syntactic and semantic information in the form of treebanks containing hand-selected linguistic analyses and the linguistically-informed grammars that describe the rules used to produce the analyses. The LOGON[2] suite of software is the original and most complete repository of code for interacting with these resources, but getting started with this repository may be too heavy a commitment for some researchers. The LOGON installation is nontrivial and has a large footprint; it requires a Linux or macOS environment and, for some uses, the Emacs editor; and many parts are without up-to-date and easily discoverable documentation (or any at all) which does little to offset the already steep learning curve.

This paper therefore describes PyDelphin,[3] a relatively lightweight Python library implementing a number of the structures and algorithms necessary for using DELPH-IN resources in research. This paper is not intended to document specific ways the way the code is used, a task better left to the API documentation,[4] but to introduce the project and its goals, capabilities, and suitability for research in computational linguistics and digital humanities.

[1]http://www.delph-in.net/
[2]http://moin.delph-in.net/LogonTop
[3]https://github.com/delph-in/pydelphin/
[4]https://pydelphin.readthedocs.io/

II. Deep Linguistic Resources with HPSG

The DELPH-IN Consortium is an international collaboration of researchers centered around the development and application of hand-built grammars in the Head-driven Phrase Structure Grammar (HPSG) [1] framework. Numerous technologies and resources have been developed to support DELPH-IN's mission. Some of the major developments include Type Description Language (TDL) [2], [3] for encoding the types, rules, and instances of the hand-built grammars; Minimal Recursion Semantics (MRS) [4] and derivative representations [5]–[8] for representing linguistic meaning; test suite database construction and related methodologies ([incr tsdb()]) [9] for organizing training and test corpora and for tracking grammar performance on these corpora; a preprocessor and tokenizer based on regular expressions (REPP) [10], efficient algorithms for parsing text to semantics [11] and realizing semantics as text [12], broad-coverage grammars for English (ERG) [13], Norwegian (Norsyg) [14], and Japanese (Jacy) [15], among others; and treebanks containing hand-selected (gold) parses for English (Redwoods) [6] and Japanese (Hinoki) [16], among others.

Some of the most crucial tools in DELPH-IN are the **processors** which read grammar descriptions and use them to parse or generate from language inputs. The LKB [3] is the original and the most useful for grammar development, but for general parsing tasks there are more efficient options: PET [17], ACE,[5] *agree* [18], and LKB-FOS.[6] ACE stands out for also having support for realizing sentences from semantic inputs, and *agree* is the only processor built to work in a Windows environment. LKB-FOS is a "fully open source" version of the LKB that has been disencumbered by its non-free software dependencies and also has generally better performance.

These technologies and resources have been developed and improved upon by different groups with different priorities at different points in time, but they are all anchored to the hand-built grammars and the production or consumption of parse results using them.

III. PyDelphin: Development Philosophy

The development philosophy behind PyDelphin prioritizes that its code is as follows:

[5]http://sweaglesw.org/linguistics/ace/
[6]http://moin.delph-in.net/LkbFos

1) correct and grammar-agnostic
2) documented
3) user-friendly
4) tested
5) understandable
6) efficient

That is, above all the code should correctly implement the structures and algorithms of DELPH-IN resources. Correctness includes independence from any particular grammar implementation, electing instead to follow (or establish, if necessary) conventions that are applicable to any DELPH-IN grammar. Secondly the code should have discoverable documentation for all available functions. If possible the functions should be intuitive and easy to use. Ideally each function is tested against a variety of inputs to ensure correctness. The code itself follows common style guidelines and does not make use of overly complex solutions so that users with some programming experience can inspect the code without much trouble. Finally, the implementations make use of data structures and optimizations to improve computational efficiency, but only as long as these optimizations do not conflict with more prioritized goals.

One additional point of the development philosophy is that PyDelphin should not package any data, which includes database schemas and partial grammar descriptions. Instead, it should implement only the empty structures that can be filled with values from a given data resource. In practice, however, this point is not entirely strict, as the practicality of including near-universal data descriptions sometimes beats the purity of a fully grammar-independent codebase.

IV. PyDelphin Capabilities

PyDelphin implements a subset of the representations and processes developed by the DELPH-IN Consortium. Particular attention is paid to semantic representations and corpus databases, but the other modules were also carefully developed to provide correct implementations.

A. Semantic Representations and Models

Three representations of DELPH-IN-style semantics are implemented in the library: Minimal Recursion Semantics (MRS) [4], Elementary Dependency Structures (EDS) [6], [7], and Dependency Minimal Recursion Semantics (DMRS) [8]. **MRS** is the original representation, and is notable for its treatment of scope, for using several kinds of underspecification to efficiently encode sources of ambiguity, and for having a straightforward interpretation into logical form (LF). MRS is also the semantic representation output by DELPH-IN grammars. Fig. 1 shows a graph visualization of an MRS for *A child kicked the ball.*[7] **EDS** and **DMRS** are transformations of MRS that present the predications of MRS as semantic dependencies, which is often a more convenient form for some kinds of semantic analysis. They differ in that EDS discards scope information in favor of representational simplicity while DMRS preserves it in favor of interconvertibility with MRS.[8]

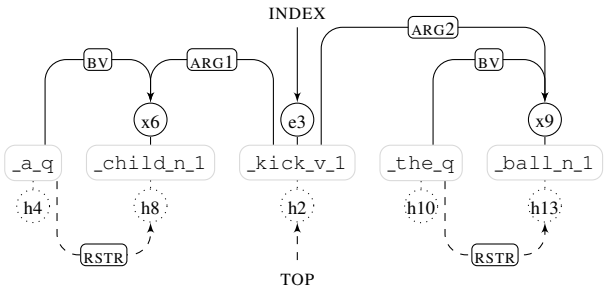


Fig. 1. MRS for *A child kicked the ball.*

In PyDelphin, each representation gets its own documented module and data structures to ensure it is natural for a user to inspect and manipulate instances of the representations. At the same time, each implements a small set of common methods which make it possible to have functionality that works with two or more representations. For example, the module for scope operations works with either MRS or DMRS structures, as these both implement the methods required for scoping representations. These methods also allow for operations that interact with two representations at the same time—namely, converting from one representation to another—without creating a mutual code dependency between the modules.[9] Listing 1 shows how to convert an MRS object to a DMRS object in Python code using PyDelphin.

```
from delphin import mrs, dmrs
m = mrs.MRS(...)  # a hypothetical MRS
d = dmrs.from_mrs(m)
```

Listing 1. Converting from MRS to DMRS

In addition to the three main representations, PyDelphin also implements modules dealing with the conventional and semi-formal semantic components used in DELPH-IN. These components include predicate symbols, predications, variables, scope constraints (used for encoding underspecified scope), individual constraints (used for phenomena such as *topic* and *focus*), and surface alignments that link semantic structures to character spans in the original sentence. Some of these components are relatively minor, but as they are defined within DELPH-IN they each get a documented and tested code object (module, class, function, etc.). Users generally do not need to worry about the details of these components as PyDelphin handles them behind the scenes.

The above representations and components are designed to be "grammar-agnostic"—that is, not tied to the details of

[7]Not shown are morphosemantic properties, such as the tense of `_kick_v_1` or the number of `_ball_n_1`, or surface alignments.

[8]Interconvertibility is a useful property because the realization task uses MRS as its input, and it is thus beneficial to be able to return to that representation after working with the dependency form.

[9]Code dependencies are when one module requires another to operate. Mutual dependencies are a purely technical limitation as they can introduce subtle bugs.

any particular grammar that produces them, and many operations, such as conversion or simple comparison, are possible without any resources external to the semantic representations themselves. Some tasks, such as semantic composition (e.g., parsing), synthesis (e.g., transfer-based machine translation or summarization), or more sophisticated comparison using subsumption, however, may require information about the entities allowed in the semantic model for a grammar. Loading a full grammar to accommodate these tasks is not ideal as it involves many complicated and time-intensive operations that are not directly relevant. Instead, PyDelphin provides a module for loading the **semantic interface**, which is a resource produced by a grammar that provides a distilled view of the semantic model.

PyDelphin also provides support for nearly all serialization formats ("codecs") for MRS, EDS, and DMRS, and for most of these it provides bidirectional read/write support. Some of these codecs are mainly historical, such as the Prolog and XML views of MRS, while others are special-purpose, such as the JSON codecs used for web applications. The current list includes the following: (1) SimpleMRS, (2) MRX, (3) IndexedMRS, (4) MRS-JSON, (5) MRS-Prolog, (6) EDS, (7) EDS-JSON, (8) EDS-PENMAN, (9) SimpleDMRS, (10) DMRX, (11) DMRS-JSON, and (12) DMRS-PENMAN.

B. Grammar Descriptions

PyDelphin has limited support for working with grammar descriptions. It has a complete and correct parser for **TDL** [3] which allows for static analysis of grammar files, programmatic type definition editing, and reformatting of TDL code. These features have found use in grammar development, grammar documentation [19], and computer-assisted language learning applications. In this release of PyDelphin, TDL parsing does not include grammar compiling, a task which involves constructing a well-formed type hierarchy and unification of feature structures. For some grammars, regular-expression preprocessors are part of the grammar description, and these are supported by PyDelphin, described below in Section IV-C.

Listing 2 shows how to load a TDL lexicon file and create a mapping of lexical entry identifiers to their orthographic forms. In the ERG, these forms are lists (e.g., the entry for *ad hoc* has a list with two parts, *ad* and *hoc*, although most entries only have one part), so they are joined together into one string.

```
from delphin import tdl
lexfile = '~/erg/lexicon.tdl'
lexmap = {}
parse_events = tdl.iterparse(lexfile)
for event, entry, lineno in parse_events:
    if event == 'TypeDefinition':
        parts = entry['ORTH'].values()
        form = ' '.join(map(str, parts))
        lexmap[entry.identifier] = parts
```

Listing 2. Mapping ERG lexical entries to orthographic forms

Derivation trees are highly-specific parse trees that describe the rules in a grammar that were applied, and the order of their application, to achieve a parse for some input. In DELPH-IN, these trees are encoded in the UDF format[10] (e.g., for parse reconstruction) or as JSON objects (for web applications). PyDelphin handles both formats.

C. Tokenization

Tokenization is a conceptually simple task—break a sentence string into word substrings—but as it is one of the first tasks done in a language-processing pipeline, the behavior of tokenization can have significant effects on the final result [20]. Many DELPH-IN grammars come with their own tokenization rules defined as a system of regular expressions called **REPP**, or a regular-expression preprocessor, which was found to be significantly more accurate than other off-the-shelf tokenizers at the task of replicating the Penn Treebank [21] tokens [10]. REPP systems may define certain classes of rules (e.g., for a domain or document type) as modules that can be activated and deactivated for different parsing tasks.

PyDelphin is able to read REPP files and use them to preprocess input strings, as shown in Listing 3. In contrast to the REPP engines provided by ACE and the LKB, it is able to preserve the surface alignments exactly as in the reference implementation. Finally, PyDelphin provides a "trace" mode that logs each rule application on a given input, which is useful for grammar developers in debugging REPP systems.

```
from delphin import repp
repp_cfg = '~/erg/pet/repp.set'
r = repp.REPP.from_config(repp_cfg)
r.tokenize("Abrams didn't chase Browne")
```

Listing 3. Tokenizing with REPP

REPP introduced a method of **surface alignment**[11] to trace the tokens (which may have been rewritten and are thus different from their original form, such as tokenizing `won't` as `will n't`) to their character span in the source string. These surface alignments are called **lnk** values in DELPH-IN representations, and are also used in semantic representations to show which words correspond to which predications. Lnk values may contain other kinds of alignment beyond character spans, but the distinctions are not relevant here as most users only use character spans. PyDelphin implements lnk structures that handle all alignment types.

D. Corpus Databases

In DELPH-IN, databases of sentences with related information are called **test suites** as one primary use of them is for testing a grammar's competence and performance. When the results of parsing the sentences are included they are also called **profiles** and contain a wealth of information about the sentences, including their syntactic derivation, MRS semantic representation, number of tokens, the time and memory needed

[10] http://moin.delph-in.net/ItsdbDerivations

[11] [10] calls this *characterization*.

to obtain a parse, the number of alternative analyses, and sometimes information about linguistic phenomena exhibited by the sentences. For each sentence there may be multiple parse results which collectively exhibit the range of ambiguity for the sentence allowed by the grammar.[12] Profiles with hand-selected, or "gold", parses are particularly valuable for data-driven tasks, such as training a model for automatic parse-selection, and for discovering linguistic patterns.

Test suites are encoded in a textual file-based database system called TSDB [9]. The system and interface built to support working with these databases is called [incr tsdb()] [9]. A custom query language based on SQL, called TSQL,[13] is used for complex queries of the databases. PyDelphin provides both low-level operations for reading and writing the textual TSDB database files as well as higher-level operations, similar to what is offered by [incr tsdb()], for processing test suites. In addition, it provides a module implementing TSQL data-selection queries.

```
from delphin import itsdb, tsql
ts = itsdb.TestSuite('~/erg/tsdb/gold/mrs')
tsql.select(
    'i-input where mrs ~ "_try_v_1"',
    ts)
```

Listing 4. Using TSQL to find inputs with a specific predicate in their MRSs

E. Processor Client Interfaces

PyDelphin does not perform parsing or generation by itself, but it is useful to be able to do so within the library so users can have one cohesive workflow for working with data. Therefore PyDelphin provides two client interfaces for external processors: one for the web-based parsing API[14] and one for the ACE processor. PyDelphin manages the communication protocol between these processors and itself so that users can submit sentences and receive results without having to understand the underlying mechanics. These processor clients can also be plugged into the API for corpus databases, discussed above in Section IV-D, to process all the items in a database, as shown in Listing 5 shows how to parse a corpus database using an ACE processor.

```
from delphin import itsdb, ace
erg = 'erg-2018-x86-64-0.9.30.dat'
ts = itsdb.TestSuite('new_profile')
with ace.AceParser(erg) as cpu:
    ts.process(cpu)
```

Listing 5. Using an ACE processor to parse

V. Example Baseline Experiments

In this section PyDelphin is used in some example research applications to demonstrate its utility. These examples are not meant to yield state-of-the-art results, but to illustrate the ease with which PyDelphin can be employed to tackle some research problems.

A. Parse Selection by Semantic Complexity

DELPH-IN grammars are developed to capture, rather than avoid, ambiguities as any grammatical sentence should be licensed by the grammar. This practice means, however, that the first result returned by a grammar is not necessarily the best result, and furthermore the combinatorics of these ambiguities leads to very high numbers of readings for moderately long sentences. Grammars with treebanks can use trained parse-reranking models [22] to select the most likely reading, but grammars without treebanks, or without treebanks in the relevant domain, will find the accuracy of the reranker suffers from a lack of data. An example of valid ambiguity is sentences with causative alternations, as in (1), which have at least two readings licensed by the ERG. One (generally preferred) reading for (1) has *open* represented by a unary event (an intransitive verb in the syntax) with *door* as its argument. The alternative reading is where the causative *open* is used in a relative clause, i.e., *the door that was opened*, and the main verb in the sentence is omitted (see (2) for an example with a main verb). The MRSs for these readings are shown in Fig. 2 and Fig. 3, respectively.

(1) The door opened.

(2) The door [that was] opened fell off its hinges.

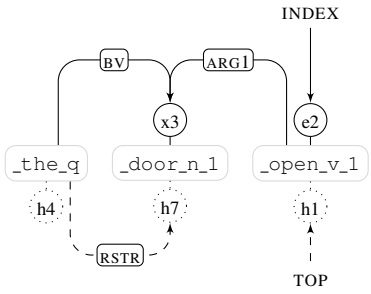


Fig. 2. MRS for *The door opened* (preferred reading)

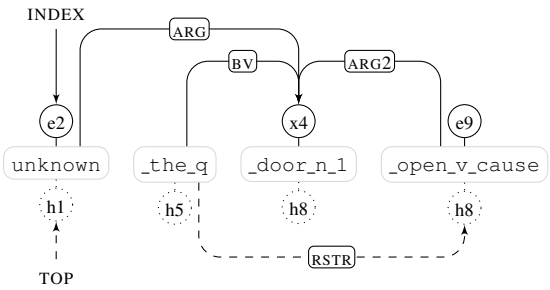


Fig. 3. MRS for *The door opened* (alternative reading)

For (1), the grammar managed to construct the alternative reading by hypothesizing two null elements: the omitted main

[12]If the grammar is correct, this is the range of valid ambiguity, otherwise some ambiguities indicate errors in the grammar to the grammar developer.

[13]For *Test Suite Query Language*, not to be confused with *Transact-SQL*.

[14]http://moin.delph-in.net/ErgApi

verb and the causer of *open*.[15] After seeing several examples like these, one might question if readings that are semantically less complex are more likely to be preferred. If so, grammars without treebanks can craft a baseline parse reranker using semantic information. To test this idea, the sentences of a section of the Redwoods treebank are reparsed and the results are compared against the gold reading. The results are first randomized (to avoid the effects of the default reranker), then sorted according to various criteria of the MRS graph, as defined in Listings 6–9.

```
def order(m):
    return len(m.rels)
```

Listing 6. Graph order (number of predications)

```
def size(m):
    args = m.arguments(types='exh')
    return sum(map(len, args.values()))
```

Listing 7. Graph size (number of arguments)

```
def density(m):
  o = order(m)
  if o <= 1:
      return 0.0  # avoid division by zero
  else:
      return size(m) / (o * (o - 1))
```

Listing 8. Graph density ($\frac{size}{(order*(order-1))}$)

```
def unknown_ratio(m):
    unks = [ep for ep in m.rels
            if ep.predicate == 'unknown']
    return len(unks) / order(m)
```

Listing 9. Ratio of `unknown` predications to others

With these functions as criteria for sorting MRSs, the position of the gold MRS in the sorted list can be found by checking for isomorphism, as shown in Listing 10.

```
def index(gold, ms):
    for i, m in enumerate(ms):
        if mrs.is_isomorphic(gold, m):
            return i
    return -1
```

Listing 10. Function to determine parse rank

The results are shown in Tab. I, where an average rank of 0.00 would mean that the first (0-indexed) reranked result is always isomorphic (graph-equivalent) with the preferred reading. All the rerankers except the unknown-ratio one did better than the randomized order, with the MRS graph order being the best of this set. This shows that semantic complexity does inversely correlate with the preferred analysis, although the effect is very slight. These naïve rerankers, however, were nowhere close to the performance of a statistical reranker trained on in-domain data.

[15]The causer entity does not appear as a predicate in the MRS but as an unbound ARG1 on _open_v_cause which is not shown in Fig. 3.

Ranking Method	Average Rank
Random	12.63
Order	**10.46**
Size	10.75
Density	10.65
Unknown Ratio	13.51
Trained Reranker	2.11

TABLE I
RESULTS OF SEMANTIC PARSE RERANKERS

This example experiment took about an hour to code and perform using PyDelphin's corpus database queries, ACE processor client, and MRS modeling modules. Future directions might investigate testing larger semantic constructions for well-formedness or other heuristics based on basic semantic components.

VI. Gender Distribution by Semantic Role

Modern techniques in corpus linguistics, along with improving cultural awareness, have made it easier to show gender bias in language use. For example, it has been shown that in syntactic examples of linguistics literature, male participants are generally more common and distributionally more likely to be syntactic subjects (semantic agents, experiencers, themes, etc.) than female participants [23]. One question is whether these patterns exist in other domains of literature. For example, it is known that these gender biases are present in the training data used in artificial intelligence systems, leading to systems that reflect these biases [24].

In this section PyDelphin is used to investigate these findings using the semantic roles of MRS in several domains of the Redwoods corpus. MRSs from the following domains are used: (1) Tourism, from Norwegian hiking guides; (2) Tech, from *The Cathedral and the Bazaar*, a technical essay by Eric S. Raymond; (3) Finance, from sections of the Wall Street Journal corpus; and (4) Fiction, from Sherlock Holmes stories by Arthur Conan Doyle. MRSs are first converted to DMRSs as this makes the resolution of the targets for scopal arguments more straightforward. For each (source, argument, target) triple within a domain, the argument's role and target's morphosemantic property for gender are tallied. The core portion of this routine is shown in Listing 11, operating on a DMRS object d. Tab. II show the number of gendered pronouns in the data for each domain. Fig. 4 shows the results for each domain.

```
for link in d.links:
    tgt = d[link.end]  # target node
    if (link.role in ('ARG1', 'ARG2')
            and tgt.type == 'x'):
        gender = tgt.properties.get('GEND')
        if gender in ('f', 'm', 'n'):
            totals[link.role][gender] += 1
```

Listing 11. Tallying the gender of argument targets

While the gendered pronouns in the Tourism and Tech domains are completely dominated by masculine pronouns, these domains have relatively few gendered pronouns per

Domain	Sentences	ARG1	ARG2
Tourism	5,816	70	124
Tech	715	27	25
Finance	1,436	219	308
Fiction	1,742	520	516

TABLE II
COUNTS OF SENTENCES AND GENDERED PRONOUNS PER DOMAIN

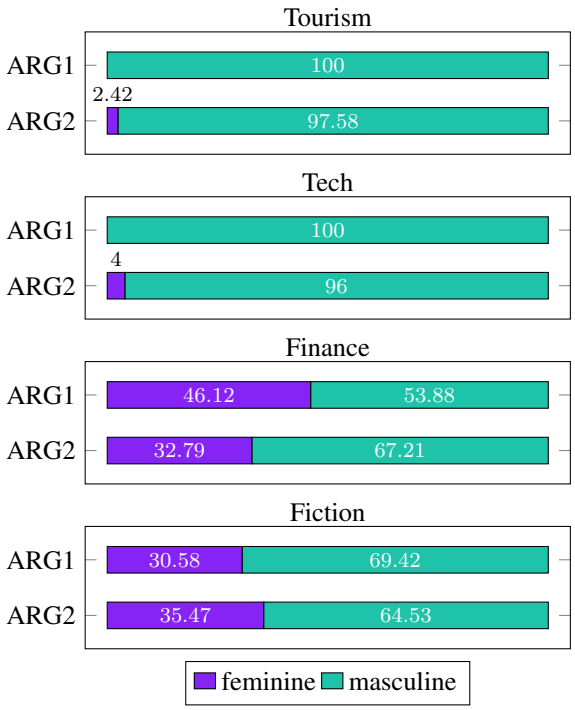


Fig. 4. Percentage of feminine vs masculine pronouns

sentence, as shown in Tab. II, so the finding may not be very accurate. The other domains have many more examples and a more robust distribution, but nevertheless masculine pronouns outnumber feminine pronouns roughly two to one. The effect of female participants being more likely to occur in object than subject roles does appear in all domains but Finance, but the effect is slight.

This experiment has a number of deficiencies, such as ignoring gendered proper names, non-gendered pronouns, and semantic roles other than ARG1 and ARG2, but it suffices for an example. Future work might expand in these directions as well as integrate a lexical semantic or sentiment model to examine the kinds of contexts in which male and female participants occur.

VII. Usage in Practice

PyDelphin has been used for a variety of published research applications, sometimes in the crux of an experimental algorithm and sometimes at the edges, performing pre- or post-processing. Some examples of its use include:

- Detection of errors in implemented grammars [25]
- Extraction of semantic features for improving sentiment classification [27]
- Statistical semantic realization [28]
- Regression testing of implemented grammars for Chinese [29], Indonesian [30], and for the Grammar Matrix Customization System [31]
- Sentence chunking based on semantic cohesion [32]
- Proposition-based abstractive summarization [33]
- Semantic composition and Functional Distributional Semantics [34], [35]
- Semantic representation conversion, inspection, and manipulation for machine translation [36]
- Detection grammar errors in research articles [26]
- DMRS-based neural semantic parsing [37] and generation [38]

In addition, PyDelphin also assists authors by preparing visualizations of semantic representations[16] for use in research articles.

VIII. Conclusion

This paper describes PyDelphin, a Python library for working with the rich linguistic resources produced by the DELPH-IN Consortium. The original LOGON software repository, while much more featureful, presents a number of challenges for researchers who are not already invested in this workflow. PyDelphin, therefore, aims to reduce the barriers to entry into research using DELPH-IN's HPSG implemented grammars and treebanks. It implements Minimal Recursion Semantics, Elementary Dependency Structures, and Dependency Minimal Recursion Semantics as well as conversion between these representations and a variety of serialization formats. The Type Description Language (TDL) of grammar descriptions is implemented for static grammar analysis and reformatting. DELPH-IN's REPP-based tokenization method is accurately implemented for direct use or in a trace-mode for REPP tokenizer development. PyDelphin also implements TSDB corpus databases and queries thereon, making it easy to perform tasks on large corpora or treebanks. The utility of PyDelphin is demonstrated by an intrinsic parse reranking task and an extrinsic task to extend a sociolinguistic claim.

PyDelphin is released under the open-source MIT license and is available at https://github.com/delph-in/pydelphin.

Acknowledgment

I thank the three anonymous reviewers of this article for their comments. I also thank the contributors[17] and users of PyDelphin and the DELPH-IN community for their help in creating and improving the software.

References

[1] C. J. Pollard and I. A. Sag, *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press, 1994.

[2] H.-U. Krieger and U. Schäfer, "TDL: a type description language for constraint-based grammars," in *Proceedings of the 15th conference on Computational linguistics*, vol. 2. Association for Computational Linguistics, 1994, pp. 893–899.

[16] Along with the related delphin-viz project: https://github.com/delph-in/delphin-viz

[17] https://github.com/delph-in/pydelphin/graphs/contributors

[3] A. Copestake, *Implementing typed feature structure grammars*. CSLI publications Stanford, 2002, vol. 110.

[4] A. Copestake, D. Flickinger, C. Pollard, and I. A. Sag, "Minimal Recursion Semantics. An introduction," *Research on Language and Computation*, vol. 3, no. 4, pp. 281–332, 2005.

[5] A. Copestake, "Robust Minimal Recursion Semantics," *unpublished draft*, 2004.

[6] S. Oepen, D. Flickinger, K. Toutanova, and C. D. Manning, "Lingo redwoods," *Research on Language and Computation*, vol. 2, no. 4, pp. 575–596, 2004.

[7] S. Oepen and J. T. Lønning, "Discriminant-based MRS banking," in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006, pp. 1250–1255.

[8] A. Copestake, "Slacker semantics: why superficiality, dependency and avoidance of commitment can be the right way to go," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 1–9.

[9] S. Oepen and D. P. Flickinger, "Towards systematic grammar profiling. test suite technology 10 years after," *Computer Speech & Language*, vol. 12, no. 4, pp. 411–435, 1998.

[10] R. Dridan and S. Oepen, "Tokenization: Returning to a long solved problem — a survey, contrastive experiment, recommendations, and toolkit —," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 378–382. [Online]. Available: http://www.aclweb.org/anthology/P12-2074

[11] Y. Zhang, S. Oepen, and J. Carroll, "Efficiency in unification-based n-best parsing," in *Trends in Parsing Technology*. Springer, 2010, pp. 223–241.

[12] J. Carroll and S. Oepen, "High efficiency realization for a wide-coverage unification grammar," in *IJCNLP*. Springer, 2005, pp. 165–176.

[13] D. Flickinger, "On building a more efficient grammar by exploiting types," *Natural Language Engineering*, vol. 6, no. 1, pp. 15–28, 2000.

[14] P. Haugereid, "Increasing grammar coverage through fine-grained lexical distinctions," *Bergen Language and Linguistics Studies*, vol. 8, no. 1, 2017. [Online]. Available: https://bells.uib.no/index.php/bells/article/view/1334

[15] M. Siegel, E. M. Bender, and F. Bond, *Jacy: An Implemented Grammar of Japanese*, ser. CSLI Studies in Computational Linguistics. Stanford: CSLI Publications, Nov. 2016.

[16] S. Fujita, T. Tanaka, F. Bond, and H. Nakaiwa, "An implemented description of Japanese: The Lexeed dictionary and the Hinoki treebank," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Sydney, Australia: Association for Computational Linguistics, July 2006, pp. 65–68. [Online]. Available: http://www.aclweb.org/anthology/P06-4017

[17] U. Callmeier, "PET–a platform for experimentation with efficient HPSG processing techniques," *Natural Language Engineering*, vol. 6, no. 1, pp. 99–107, 2000.

[18] G. C. Slayden, "Array tfs storage for unification grammars," Master's thesis, University of Washington, 2012.

[19] C. Hashimoto, F. Bond, and D. Flickinger, "The lextype db: a web-based framework for supporting collaborative multilingual grammar and treebank development," in *International Workshop on Intercultural Collaboration*. Springer, 2007, pp. 76–90.

[20] A. Fokkens, M. van Erp, M. Postma, T. Pedersen, P. Vossen, and N. Freire, "Offspring from reproduction problems: What replication failure teaches us," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 1691–1701. [Online]. Available: https://www.aclweb.org/anthology/P13-1166

[21] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," 1993.

[22] K. Toutanova, C. D. Manning, D. Flickinger, and S. Oepen, "Stochastic hpsg parse disambiguation using the redwoods corpus," *Research on Language and Computation*, vol. 3, no. 1, pp. 83–105, 2005.

[23] M. Macaulay and C. Brice, "Don't touch my projectile: Gender bias and stereotyping in syntactic examples," *Language*, pp. 798–825, 1997.

[24] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[25] M. Goodman and F. Bond, "Using generation for grammar analysis and error detection," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Suntec, Singapore: Association for Computational Linguistics, Aug. 2009, pp. 109–112. [Online]. Available: https://www.aclweb.org/anthology/P09-2028

[26] D. Flickinger, M. Goodman, and W. Packard, "UW-Stanford system description for AESW 2016 shared task on grammatical error detection," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego, CA: Association for Computational Linguistics, Jun. 2016, pp. 105–111. [Online]. Available: https://www.aclweb.org/anthology/W16-0511

[27] J. Kramer and C. Gordon, "Improvement of a naive Bayes sentiment classifier using MRS-based features," in *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, Aug. 2014, pp. 22–29. [Online]. Available: https://www.aclweb.org/anthology/S14-1003

[28] M. Horvat, A. Copestake, and B. Byrne, "Hierarchical statistical semantic realization for minimal recursion semantics," in *Proceedings of the 11th International Conference on Computational Semantics*. London, UK: Association for Computational Linguistics, Apr. 2015, pp. 107–117. [Online]. Available: https://www.aclweb.org/anthology/W15-0116

[29] Z. Fan, S. Song, and F. Bond, "An HPSG-based shared-grammar for the Chinese languages: ZHONG [—]," in *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 17–24. [Online]. Available: https://www.aclweb.org/anthology/W15-3303

[30] D. Moeljadi, F. Bond, and S. Song, "Building an HPSG-based Indonesian resource grammar (INDRA)," in *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 9–16. [Online]. Available: https://www.aclweb.org/anthology/W15-3302

[31] E. M. Bender, S. Drellishak, A. Fokkens, M. W. Goodman, D. P. Mills, L. Poulson, and S. Saleem, "Grammar prototyping and testing with the LinGO grammar matrix customization system," in *Proceedings of the ACL 2010 System Demonstrations*. Uppsala, Sweden: Association for Computational Linguistics, Jul. 2010, pp. 1–6. [Online]. Available: https://www.aclweb.org/anthology/P10-4001

[32] E. Muszyńska, "Graph- and surface-level sentence chunking," in *Proceedings of the ACL 2016 Student Research Workshop*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 93–99. [Online]. Available: https://www.aclweb.org/anthology/P16-3014

[33] Y. Fang, H. Zhu, E. Muszyńska, A. Kuhnle, and S. Teufel, "A proposition-based abstractive summariser," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 567–578. [Online]. Available: https://www.aclweb.org/anthology/C16-1055

[34] G. Emerson and A. Copestake, "Semantic composition via probabilistic model theory," in *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*, 2017. [Online]. Available: https://www.aclweb.org/anthology/W17-6806

[35] ——, "Functional distributional semantics," in *Proceedings of the 1st Workshop on Representation Learning for NLP*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 40–52. [Online]. Available: https://www.aclweb.org/anthology/W16-1605

[36] M. W. Goodman, "Semantic operations for transfer-based machine translation," Ph.D. dissertation, University of Washington, Seattle, 2018.

[37] J. Buys and P. Blunsom, "Robust incremental neural semantic graph parsing," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1215–1226. [Online]. Available: https://www.aclweb.org/anthology/P17-1112

[38] V. Hajdik, J. Buys, M. W. Goodman, and E. M. Bender, "Neural text generation from rich semantic representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2259–2266. [Online]. Available: https://www.aclweb.org/anthology/N19-1235

Taking a journey on research data management in Singapore

Pin Pin Yeo
Libraries
Singapore Management University
Singapore, Singapore
ppyeo@smu.edu.sg
https://orcid.org/0000-0001-5331-1046

Danping Dong
Libraries
Singapore Management University
Singapore, Singapore
dpdong@smu.edu.sg
https://orcid.org/0000-0002-2229-6709

***Abstract*— In recent years, there has been a growing need and importance for Research Data Management (RDM), in line with open science, FAIR principles, transparency, and especially with the policies of funders and journal publishers. Institutions worldwide have been responding to these needs.**

The context of RDM in Singapore and its impact on universities will be covered. We started by examining the RDM policies of other institutions both locally and worldwide. Collaborating with the SMU Office of Research and Tech Transfer (ORTT), we contributed to a draft RDM policy covering issues including ownership, retention, and archiving. We interviewed 18 faculty researchers from different disciplines to understand their research process and data produced at different stages, their understanding on reproducible research and disciplinary differences, and current practices in RDM. Concurrently, feedback on the policy was sought at various levels in SMU, which led to further changes in the policy. In the next stage, SMU plans to set up the institutional infrastructure to support RDM activities of SMU researchers.

We reflected on the lessons learnt from our journey and found that policies were critical drivers and that engagement at different levels was important.

Keywords—research data management, RDM, Singapore, university policy, data interview, data repository, research office, academic library

I. Introduction

The National Academy of Sciences' Committee on Science, Engineering, and Public Policy published a report in 2009 that stated: "Developing the policies, standards, and infrastructure needed to ensure the integrity, accessibility, and stewardship of research data is a critically important task. It will require sustained effort on the part of all stakeholders in the research enterprise" [1]. The stakeholders included researchers, research institutions, funding agencies, professional societies, and journals. The National Institutes of Health issued a Statement on Sharing Data in February 2003 and the policy became effective in October 2003 [2].

The United Kingdom has a Concordat on open research data [3] supported by a multi-stakeholder group in 2016. Carr [4] from Wellcome Trust discussed the issues and benefits of sharing data. Wellcome Trust supports open research [5] and has put in place a platform called Wellcome Open Research, available at https://wellcomeopenresearch.org, which "provides all Wellcome researchers with a place to rapidly publish any results they think are worth sharing. All articles benefit from immediate publication, transparent peer review and the inclusion of all source data." The UK Research and Innovation (formerly Research Councils UK) is "committed to opening up research data for scrutiny and reuse, to enable high-quality research, drive innovation and increase public trust in research" [6].

The European Commission issued their guidelines on data management in 2016, with the intent to make "research data findable, accessible, interoperable and reusable (FAIR)" and acknowledged that "good research data management is not a goal in itself, but rather the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse" [7]. The European Open Science Cloud made a declaration in October 2017, followed by an action list that covered data culture, skills, FAIR, data repositories, and governance [8].

For Australia, the Australian Code for the Responsible Conduct of Research, first issued in 2007 was updated in June 2018. For research data, the code states that institutions are to "provide access to facilities for the safe and secure storage and management of research data, records and primary materials and, where possible and appropriate, allow access and reference" and that researchers are to "retain clear, accurate, secure and complete records of all research including research data and primary materials, where possible and appropriate, allow access and reference to these by interested parties" [9]. Recipients of funding from National Health and Medical Research Council and Australian Research Council must compile with the code. Other Australian research agencies are encouraged to adopt this code as a mandatory requirement.

The confluence of open science, transparency, FAIR principles, have been encouraging the stakeholders towards more systematic governance of research data management (RDM).

II. Literature Review

Pinfield, Cox and Smith noted that "university libraries have moved into this space and are increasingly seen as major contributors to RDM activity in general and in the design of research data services (RDS) in particular" [10]. LIBER made recommendations for libraries to get started with research data management in 2012 [11]. How have academic libraries dealt with these changes?

Tenopir, Birch and Allard [12], and Reznik-Zellen, Adamick and McGinty [13] did environment scans of institutions in the United States. Cox, Pinfield and Smith [14]

interviewed libraries in the United Kingdom on how they were coping with RDM. Tenopir, et al [15] studied the RDM activities of European academic research libraries. Cox, Kennan, Lyon and Pinfield [16] did an international study covering Australia, Canada, Germany, Ireland, the Netherlands, New Zealand and the United Kingdom.

These studies found evidence of increasing maturity of RDM activities and service in the countries studied, especially for having policies in place or policies being planned, and for RDM training. They also found that development in the areas of advanced research data services and advocacy had been mixed.

III. Changes in the Singapore Policy Landscape

Incidents of breaches in research integrity by researchers affiliated with Singapore research institutions resulted in the retraction of articles. According to the Retraction Watch database (http://retractiondatabase.org/), there were 130 cases of retractions between 2004 to 2018 with at least one author affiliated to Singapore, with a peak of 20 cases in 2010. Research reproducibility and research integrity became high priorities and contributed to the focus on RDM initiatives.

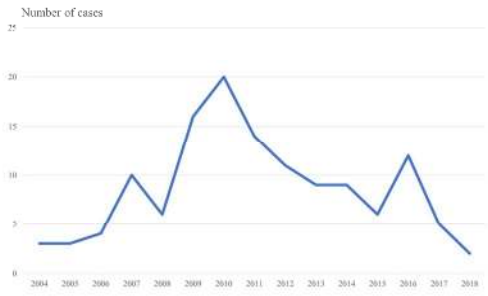


Fig. 1. Number of cases in Retraction Watch for Singapore

The Singapore Statement on Research Integrity available at https://wcrif.org/guidance/singapore-statement was developed as part of the 2nd World Conference on Research Integrity, 21-24 July 2010 held in Singapore, as a global guide to the responsible conduct of research. Four local institutions signed the statement, namely the Agency for Science, Technology and Research, Nanyang Technological University, National University of Singapore and Singapore University of Technology and Design.

The six major research funders in Singapore, which includes the Agency for Science, Technology and Research, Ministry of Education, and National Research Foundation, have a normalized clause with a requirement for open access of publications within 12 months after publication since May 2015. The National Medical Research Council added the requirement for sharing of data in November 2015 for the grants they funded, as seen on their webpage http://www.nmrc.gov.sg/policy-guideline/data-sharing. The Social Science Research Council and the Institute for Adult Learning had subsequently required the sharing of data for the grants they funded.

The National University of Singapore (NUS) implemented their research data policy in 2012. In November 2017, NUS launched their data repository, using the same DSpace platform as their institutional repository called ScholarBank available at https://scholarbank.nus.edu.sg/.

The Nanyang Technological University (NTU) implemented their research data policy in September 2015. This was followed by the requirement for their researchers to submit data management plans before they were able to access their research funds. In November 2017, NTU launched their data repository called DR-NTU (Data) using the Dataverse platform, available at https://researchdata.ntu.edu.sg/.

IV. Early Indicators for RDM at SMU

SMU is a social sciences university focusing on business and management, hence its researchers make use of datasets in social sciences, business and finance. Prior to 2010, the SMU Libraries had been receiving requests from faculty for datasets. There were some overlaps in requests from different faculty. The library staff wanted to study the issue further and conducted a simple survey to understand faculty needs for datasets.

The survey partially revealed some faculty needs for data. SMU Libraries engaged Terrence Bennett, from the College of New Jersey, as a consultant to study these needs further. Bennett interviewed 34 faculty members. The completed report was shared with relevant stakeholders. Bennett [17] identified five tiers of service relevant for RDM, as seen in Fig. 2.

Tier 1	Identify and recommend existing data resources to meet researcher needs
Tier 2	Instruction and training in finding and using data
Tier 3	Recommend acquisition of specialized data sets to meet researcher and institutional needs
Tier 4	Consultations for effective use of data sets Assistance with preparing data for analysis
Tier 5	Data curation for institution-specific repository or shared repository

Fig. 2. Research Data Service Continuum

The library staff have been handling basic queries relating to data and data sources and did some instruction on finding and using data from our subscribed resources, but the more advanced research data services identified in Tier 4 and Tier 5 were not offered.

The report by Bennett recommended that the SMU Libraries "should initiate a distinct research data services function that provides, at minimum, identification and dissemination of data resources, instruction in finding and using data, and some technical assistance with the use of specialized software" [17].

As a result of the report recommendations, SMU Libraries proposed to add a part-time headcount to work on research data. A postgraduate student was hired for 12 months from October 2012, with the job title of Data Service Specialist. The work done by the Data Service Specialist was reviewed and it was decided to convert the position to a full-time headcount. A librarian was hired for the position of Research Data Librarian from April 2014 to August 2015.

The Research Data Librarian focused on creating guides and developing information literacy workshops on research data. At the same time, the librarian tried to engage faculty to share their research data. The existing institutional repository InK available at https://ink.library.smu.edu.sg was used as a data repository for a start, as the volume was expected to be low. It was good experience for the librarian to start learning about metadata for datasets. However, the progress was slow. At the end of 2015, there were less than 25 datasets in the repository.

A new Scholarly Communication Librarian was hired to in April 2016 with broader responsibilities in scholarly communication but still focused on research data. The new librarian updated the existing guides and created bite-sized classes on RDM targeted at post-graduate students.

V. Review of RDM policies

SMU Libraries has been collaborating closely with the Office of Research & Tech Transfer (ORTT) and were aware that SMU intended to draft a research data policy. It was an opportunity for SMU Libraries, especially the Scholarly Communication team and the liaison librarians to participate and to work together with ORTT on the policy.

Prior to drafting the SMU policy, the policies of other institutions in the United Kingdom, United States, and Australia were studied [18], [19], [20]. The review was useful as it highlighted the commonalities in the policies across these countries.

ORTT and SMU Libraries also contacted their counterparts in NTU and NUS to learn how they implemented their policies and how they were handling research data management in their institutions. Both institutions started with the policy followed by setting up the infrastructure to support storing of datasets. The study of both the local and overseas policies set the foundation for our institution's RDM policy.

VI. Evolution of SMU RDM Policy

The owner of the RDM policy was ORTT and they drove the form and direction of the policy. SMU Libraries provided input to various clauses including ownership, retention, and archiving. In the policy, SMU Libraries would have a role to play as the custodian of the data and the data repository, as well as being the liaison with faculty and providing RDM advisory services.

The library staff also studied the data policies of journals where SMU faculty tend to publish in, and it was discovered that some journals included mandatory data sharing requirements, while some encouraged data sharing or required authors to state the availability of underlying data. Journal policies, together with the changes in funder policies, formed part of the motivation for developing RDM initiatives at SMU.

SMU planned to sign the Joint Statement on Research Integrity relating to Scholarly Publications in October 2018. In the Joint Statement, for purposes of reproducibility "research personnel must maintain accurate and detailed research records of procedures and results (for a minimum of 10 years), to allow others to replicate the work, and ensure reproducibility of one's experimental results."

The proposed RDM policy was presented to SMU's Council of Deans in July 2018 for their endorsement. ORTT also engaged the Faculty Senate, the Deans and Vice-Deans of Research of all the six schools in SMU for their feedback on the draft policy. SMU Libraries participated in these engagement sessions to hear first-hand the issues raised by the Faculty Senate and the schools. It was planned that the policy will become effective in January 2020 to give time for feedback and revision.

VII. Data Repository Platforms

SMU Libraries undertook an environmental scan and feasibility study on data repository solutions from June to August 2018. Given that SMU is a medium-sized university and that the IT department has a lean team, an open-source solution supported in-house is not the preferred approach. Hence in the feasibility study, emphasis was placed on hosted commercial solutions, i.e. Software as a Service (SaaS).

The use of the current institutional repository InK as a data repository was reviewed. The current platform had some drawbacks, as it did not have the different types of access control for researchers to manage access to their data, no collaboration features, and large files took a long time to upload. Hence, we explored other options for a dedicated data repository.

Repository solutions used by institutions in Singapore and worldwide were identified. NUS and NTU, the other two major universities in Singapore with a data repository, have adopted DSpace and Harvard Dataverse respectively. Other solutions were identified from a literature search and re3data.org. We analyzed the strengths and weaknesses of these solutions with recommendations for further study and evaluation for the next stage of the project. The team also proposed high-level objectives of the data repository project and came up with a set of preliminary functional requirements.

It is of key importance that the solution had high usability and user-friendliness, as nothing deters users more than a clunky, poorly designed interface and confusing features. A user-friendly solution would encourage adoption, or at least does not discourage it. As RDM is mostly non-mandatory, it is important that the technology is simple to use and offers value-add or "carrots" in the research process. It is desirable to have features for research collaboration (e.g. data sharing among project members), so that data management is integrated into the early stages of research. In more realistic terms, adoption would not depend on technology alone, with many other factors at play, e.g. cultural practice in the discipline, policy and mandates and incentives for RDM.

Common data management features and compliance with standards are also important factors. Other key features include DOI registration, version control, data citation, large file size support, flexible access control and data sharing, and availability of open APIs. It is also important to follow standards to enable interoperability, and to conform to FAIR (Findability, Accessibility, Interoperability and Reusability) principles to be future proof as the RDM landscape evolves.

The feasibility study was a useful exercise to understand the technology aspects and available platforms on the market, in preparation for a possible future acquisition of a data repository system.

The project to implement a data repository was given the green light in April 2019. The project team included representatives from Integrated Information Technology Services (IITS), ORTT and SMU Libraries. The plan is to launch the data repository in the first quarter of 2020 to support the RDM Policy.

VIII. Data Interviews

Towards the end of 2018, SMU Libraries conducted interviews with selected faculty from six Schools at SMU. We adopted face-to-face interview approach, which promoted open-ended communication and more in-depth discussion rather than surveys or focus groups.

The objective was to understand the nature of research data produced from the major disciplines of SMU (business, finance and management, computer science, social sciences, economics, accounting and law) and current data management practices of SMU researchers. It was thought that understanding user behavior and research workflow was an important step before implementing the data repository project. It was also a good opportunity to engage the faculty on the draft policy and to gather their feedback.

Interviews were conducted with 18 faculty researchers who were representative of most broad research disciplines at SMU. As RDM was highly discipline specific, it was important to cover the range of disciplines in SMU. The invitation was via email broadcast to the SMU faculty and via personalized invitations by liaison librarians to pre-identified researchers who did more data-intensive research.

A semi-structured interview approach was adopted, simplifying and modifying an interview template from ANDS [21] to fit the SMU context and the specific objectives of the project. The interviews lasted 30-60 minutes, jointly conducted by the liaison librarian for the School and the RDM Librarian. Interviews were recorded in audio when the interviewee gave consent, and the transcripts were coded for analysis.

The findings were categorized into 7 main themes: data characteristics; data documentation; storage, backup and archiving; data ownership; data repository; reproducibility; data sharing. Some of the key learnings were summarized below:

A better understanding on the characteristics of SMU research data, as well as how they transform and evolve over the course of a research project was gained. There was a myriad of raw data sources used by SMU researchers, including primary and secondary data sources. Occasionally some researchers may aggregate data from multiple sources to form the raw dataset. It was recognized that different types of raw data used may have legal and contractual implications on data archiving and preservation at the end of a research lifecycle, which needed to be dealt with on a case-by-case basis.

Fig. 3 is a simple visualization of research outputs generated at different stages of a data-intensive research project as summarized from what interviewees shared about their research process. The research data were broadly categorized into three types, based on the stage of research they are in:

- **Raw data**: Data in its original form without any processing or cleaning.
- **Processed data**: Data obtained after raw data has been processed, cleaned, extracted, coded, transformed, or organized and ready to be used for analysis.
- **Final outputs**: The final research outputs that exist at the last stage of the research project. Examples include results and outputs from tools and software, codes/scripts/syntax for analysis and any other supplementary materials.

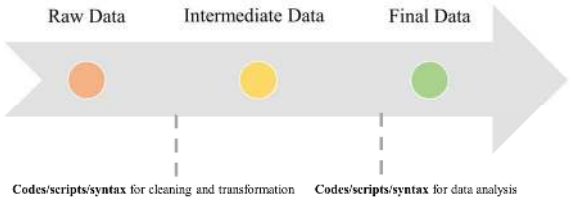


Fig. 3. Typical outputs generated at different stages of a research project

One of the important objectives of the RDM policy at SMU is to support research integrity and facilitate reproducibility or replication of research findings. The two terms reproducibility and replication are often confused from each other and sometimes misused. Here we take the definition from the American Statistical Association, that reproducibility is achieved when you take the original data and code from the author and are able to reproduce all findings from a study, whereas replicability is defined as the act of repeating a study without use of the original data from the author, but generally using the same method [22]. In the context of SMU's draft RDM policy, the focus is more on tackling the issues of reproducibility, being mindful of risks in institutional reputation in any possible case of research misconduct.

The answers to the question of what outputs should be archived for reproducibility were mixed, depending on the researcher's research area and the type of research, i.e. experimental, modelling, simulation or programming. Some researchers recognized the importance of codes/scripts/syntax to replicate an analysis, some felt that it was important to keep raw data, and others pointed out that archiving only the final data might be more practical and require less effort. The opinions varied according to discipline, type of research methodology, and possibly influenced by subjectivity such as individual researcher's preference and current practice. It is recognized that there was no standard answer. Reproducibility should be treated as a spectrum instead of two extreme states of polarity. Due to the lack of well recognized standards and widely adopted practices, especially in the research areas at SMU, currently it was largely left to the researchers' own preference for the selection and curation of data for preservation.

Data documentation is important for reproducibility and understandability of research data. Keeping documentation and descriptive materials is a recommended best practice for research data management and is indispensable for data

preservation [23]. A spectrum of data documentation practices among SMU researchers was observed. As seen in Fig. 4, which is a simplified representation of typical personas that surfaced out of the interviews, representing increasing rigor and quality of data documentation practices from left to right.

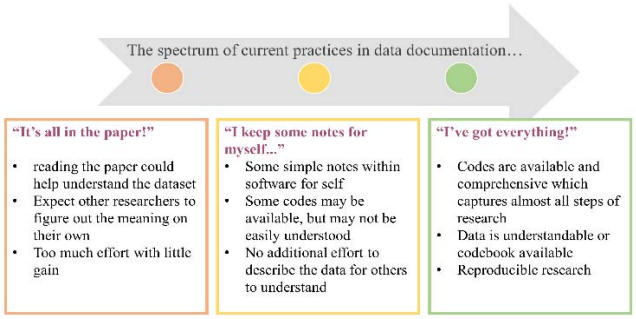


Fig. 4. Spectrum of data documentation practices

It was found that many of the interviewees regarded data documentation as a tedious and time-consuming task with little benefit for their actual research. Some interviewees also pointed out that it was practically impossible to track everything that happens to a dataset, especially when it involves manual steps or minor changes. Therefore, keeping the final dataset might be relatively more feasible. Some researchers have formed the habit of keeping simple notes or information that does not require a lot of extra effort, e.g. codes for analysis. They felt that doing so may help in case they need to revisit the data in the future. Most of them are reluctant to put the additional effort to make their data understandable to other researchers.

Through the interviews, a few researchers stood out. They were relatively young and early in their academic career, and already adopted best practices in data management and saw these practices as an innate part of their research work. One of them is an advocate of Open Science and shared almost all his projects and data via the Open Science Framework. A typical comment from these researchers was: "if one forms the habit of good data management and performs it consistently, it will not take much time and effort. On the contrary, it will help one achieve overall higher efficiency."

The interviews were a good opportunity to get faculty feedback on the draft RDM policy. Some interviewees showed high levels of interest in the policy, some expressed concern on issues such as ownership and sensitive data management. Some researchers exhibited a strong sense of ownership over their data. Some researchers were concerned that they might lose access to their research data under certain circumstances.

The relatively high level of engagement and display of interest from faculty with SMU's RDM policy differed from some of the other institutions where academics displayed a lack of interest in RDM policy development despite attempts to obtain their feedback [10].

Besides achieving a better understanding of the current state of RDM at SMU, the interviews also played a role in raising awareness of the RDM draft policy and upcoming data repository project. Some researchers have been identified for engagement on the implementation of the RDM policy and the data repository.

Reflecting on the project, it was important to involve liaison librarians. Liaison librarians at SMU maintain close ties with the Schools, and they are the first point of contact for faculty with regards to libraries services. Their relationship with faculty helped with getting a good spread of participants, and their subject knowledge contributed to tailored questions and in-depth discussions.

IX. Summary of the SMU Journey

The SMU journey could be visualized using the maturity model shown in Fig. 5 which was adapted from Cox, Kennan, Lyon and Pinfield [16].

At the basic level, the liaison librarians were already offering basic advisory services to faculty about data sources for their research needs. The librarians undertook surveys and interviews to learn about faculty needs for data.

At the second level, the librarians created a guide for data sources in 2012 which is available at http://researchguides.smu.edu.sg/datasources, followed by a guide on RDM in 2014 which is available at http://researchguides.smu.edu.sg/dmg. It was useful for showing best practice and sources for RDM and raising awareness of RDM. SMU embarked on drafting an RDM policy in 2018, and sought feedback on the policy in 2019.

At the third level for capacity building, SMU Libraries had been investing in its staff for RDM training and data literacy. This was done through library staff spending time at conferences, seminars, talks, library visits and the Digital Curation Centre. The librarians used the new knowledge gained to create and to teach classes on these topics targeted at our students. The librarians also kept abreast of the trends in RDM, changes in funder requirements and journal requirements.

At the fourth level, SMU Libraries together with ORTT and IITS are implementing a data repository which is planned for launch in early 2020. This is in preparation of the stewardship role for data within SMU. SMU Libraries would be planning for the metadata and curation support needed for datasets.

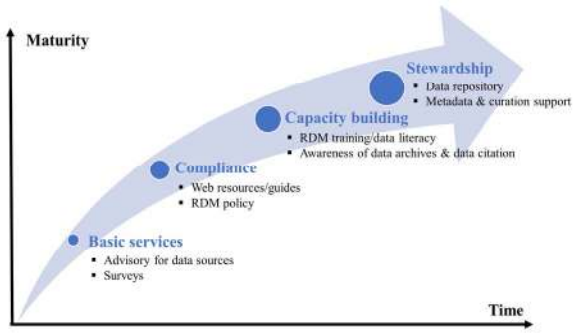


Fig. 5. RDM journey at SMU

X. Conclusion

The changes in the research and policy landscapes created an impetus for stronger governance in RDM, which galvanized the institutions in Singapore to act. In SMU, it created an opportunity for deeper collaboration across the institution to put in place the policies and the infrastructure for RDM, which is a strategic and university-wide initiative.

Once SMU decided to implement an RDM policy, SMU Libraries was able to collaborate with ORTT. There were opportunities to present the background of the trends and reasons for the policy to senior management, and deans of the schools. As part of the process to get buy-in for the policy, there was further engagement with the Faculty Senate, Vice-Deans of Research, and faculty. The engagement done raised awareness of RDM and there were discussions about the issues relating to RDM and the policy.

The RDM policy also led to the recognition of the need to provide the infrastructure to support the policy. Hence, approval was given for the implementation of a data repository. The plan was to launch the data repository when the policy comes into effect.

Looking at the road ahead, there will be another part of the RDM journey, for our faculty to be following best practices in research data management. Looking at the road ahead, there will be more faculty storing their datasets in the repository or other subject repositories more suitable for their disciplines. Looking at the road ahead, there will be opportunities for the roles of the librarians to be strengthened in RDM and in research. It is hoped that all these efforts by individual institutions and other stakeholders for good research data management will be the "key conduit leading to knowledge discovery and innovation" [7].

References

[1] Committee on Science, Engineering, and Public Policy, "Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age." Washington, DC: National Academy of Sciences, 2009. Available: https://www.ncbi.nlm.nih.gov/books/NBK215264/

[2] National Institutes of Health, "NIH Data Sharing Policy," 2007. https://grants.nih.gov/grants/policy/data_sharing/ (accessed May 30, 2019).

[3] Concordat on Open Research Data, July 2016. https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/ (accessed May 30, 2019).

[4] D. Carr, and K. Littler, "Sharing research data to improve public health: A funder perspective," *Journal of Empirical Research on Human Research Ethics*, vol. 10, no. 3, pp. 314-316, Jul. 2015. https://doi.org/10.1177/1556264615593485

[5] Wellcome Trust, "Open Research." https://wellcome.ac.uk/what-we-do/our-work/open-research (accessed May 30, 2019).

[6] UK Research and Innovation, "Data policy." https://www.ukri.org/funding/information-for-award-holders/data-policy/ (accessed on 9 May 2019).

[7] European Commission, "H2020 Programme: Guidelines on FAIR data management in Horizon 2020." Brussels: The Commission, 2016. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[8] European Open Science Cloud, "Declaration action list," Oct. 2017. Available: https://ec.europa.eu/research/openscience/pdf/eosc_declaration-action_list.pdf

[9] Australian Code for the Responsible Conduct of Research, 2018. Available: https://www.nhmrc.gov.au/sites/default/files/documents/attachments/grant%20documents/The-australian-code-for-the-responsible-conduct-of-research-2018.pdf

[10] S. Pinfield, A. M. Cox, and J. Smith. 2014. "Research data management and libraries: Relationships, activities, drivers and influences," *PLoS ONE*, vol. 9, no. 12, e114734, Dec. 2014. https://doi.org/10.1371/journal.pone.0114734

[11] LIBER, "Ten recommendations for libraries to get started with research data management," Jul. 2012. Available: https://libereurope.eu/wp-content/uploads/The%20research%20data%20group%202012%20v7%20final.pdf

[12] C. Tenopir, B. Birch, and S. Allard, "Academic libraries and research data services," ACRL White paper, 2012. Available: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf

[13] R.C. Reznik-Zellen, J. Adamick, and S. McGinty, "Tiers of research data support services," *Journal of eScience Librarianship,* vol. 1, no. 1, e1002, Feb. 2012. https://doi.org/10.7191/jeslib.2012.1002

[14] A.M. Cox, S. Pinfield, and J. Smith, "Moving a brick building: UK libraries coping with research data management as a 'wicked' problem," *Journal of Librarianship and Information Science*, vol. 48, pp. 3-17, Mar. 2016. https://doi.org/10.1177/0961000614533717

[15] C. Tenopir, S. Talja, W. Horstmann, E. Late, D. Hughes, D. Pollock, B. Schmidt, L. Baird, R. Sandusky, and S. Allard, "Research data services in European academic research libraries," *LIBER Quarterly*, vol. 27, no. 1, pp. 23-44, Feb. 2017. http://doi.org/10.18352/lq.10180

[16] A.M. Cox, M.A. Kennan, L. Lyon, and S. Pinfield, "Developments in research data management in academic libraries: Towards an understanding of research data service maturity," *Journal of the Association for Information Science and Technology*, vol. 68, no. 9, pp. 2182-2200, Sept. 2017. https://doi.org/10.1002/asi.23781

[17] T. Bennett, "Research data services at Singapore Management University: Engagement summary report," Jul. 2010. Available: https://ink.library.smu.edu.sg/library_research/5/

[18] L. Horton, and Digital Curation Centre, "Overview of UK institution RDM policies," Aug. 2016. Available: https://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies

[19] S. Hodson, and L. Molloy, "Current best practice for research data management policies" (report commissioned from CODATA by the Danish e-Infrastructure Cooperation and the Danish Digital Library), May 2014. Available: https://doi.org/10.5281/zenodo.27872

[20] ANDS, "Institutional policies and procedures," 2017. https://www.ands.org.au/working-with-data/data-management/data-management-policy (accessed on 9 May 2019).

[21] ANDS, "Research data interviews," 2012. Available: https://www.ands.org.au/working-with-data/data-management/institutional-dm-frameworks/research-data-interviews

[22] American Statistical Association, "Recommendations to funding agencies for supporting reproducible research," Jan. 2017. Available: https://www.amstat.org/asa/files/pdfs/POL-ReproducibleResearchRecommendations.pdf

[23] UK Data Archive, "Managing and sharing data," May 2011. Available: https://ukdataservice.ac.uk/media/622417/managingsharing.pdf

Using AI to Analyze Humanities Research Trends in Chinese and Taiwan Studies

Shu-hsien Tseng
National Central Library
Taipei, Taiwan
director@ncl.edu.tw

Wen-de Huang
National Central Library
Taipei, Taiwan
Wende@ncl.edu.tw

Jane Liau
National Central Library
Taipei, Taiwan
liaujane@ncl.edu.tw

Abstract—The Humanities Academic Trends System (https://trends.ncl.edu.tw/) was created by National Central Library (Taiwan) in 2018. It combines Artificial Intelligence, text mining and meta-analysis techniques to work over the web pages and social media of Chinese Studies institutions worldwide as well as over 2000 academic papers published in Center for Chinese Studies. The system surveys research in the fields of Taiwan Studies and Chinese Studies, analyzes popular academic trends, and exploring the spatio-temporal distribution of prevalent academic concepts.

The system homepage features interactive analytical charts and uses Responsive Web Design to arrange content to fit different screen sizes. The system uses Chinese and English text mining to automatically analyze web pages for the most recent 3 months, producing keywords, top 50 vocabularies, hot topic analysis, post volume, spatial distribution analysis, and popular article ranking charts. Clicking on these interactive charts opens relevant data lists and displays links to the original articles. The system can also perform an analysis in depth based on user-defined keywords.

In future, the system will gradually increase the number of websites surveyed. The aim is that Chinese Studies and Taiwan Studies scholars worldwide can use this system to examine academic trends in literature, history, and philosophy, to accumulate research experience, and to discover new research directions.

國家圖書館於 2018 年建置之文史哲學術趨勢系統（https://trends.ncl.edu.tw/），利用文字探勘與 AI 巨量分析技術，分階段觀察全球重要漢學機構網頁及社群，目前已擷取 300 個世界漢學學術相關網站、85,000 多筆資料，以及漢學研究中心出版品全文資料庫的 2010 筆學術文章，以 AI 文字探索分析出熱門學術關鍵詞及議題資訊。系統每日更新最新網頁資訊，藉以從龐大的資料中找出使用者所關注、需要的資訊，進而創造智慧化的資料分析。以提供漢學與臺灣研究領域觀察，分析熱門學術趨勢，以探索學術觀念流行時空分布。

系統設計以互動式分析圖表呈現活潑的首頁設計。採用響應式網頁設計（RWD），可供不同螢幕大小切換內容排版，利用中英文文字探勘技術，自動分析近 3 個月 300 個漢學機構網頁的關鍵字、Top50 的熱門詞彙、熱門議題分析、聲量趨勢分析、空間分布分析及熱門文章排行等圖表，圖表為互動式，點選圖表可連結進資料列表頁和原始文章連結。使用者可也以自定義自己喜歡或需要的主題關鍵字，系統可以針對使用者喜歡的主題做深度分析，甚至是可以選定一個時間軸去觀察，進行有憑（透過 AI 的神經網路，社群討論一把抓）有據（數據圖表分析）的呈現。

期望此一系統之建立，可匯整國際漢學研究歷程、累積研究經驗，激發人文學者發掘在人文研究上的新面向。有效掌握全球漢學與臺灣研究文史哲研究的發展情形，觀察分析熱門學術趨勢，以利提供國內外學界參考，並持續與其他各種數位人文研究服務結合，以發揮功能相乘的效果。

Keywords—AI, Digital Humanities, Academic Trends, Text Mining

I. INTRODUCTION

The development and popularization of the internet has helped make the world's scattered Chinese studies research institutions much more accessible. The great ease of use of the internet has advanced the dissemination and utilization of online Chinese studies resources. Not only has this allowed these institutions to more directly interact with users, but it has also expanded the domain of the services they provide. The Humanities Research Trends System allows scholars to find out what research is being conducted by the world's sinologists. The system uses text mining techniques to analyze the content of international sinology research institutions' web pages, identifying keywords that frequently appear, and therefore which fields are comparatively lacking in research, and which fields of research are more popular. A better understanding of the developmental trends in international Chinese Studies research will expand the international horizons of Taiwan's sinology research community, while also allowing Taiwan's cultural strength to be fully realized. This will also enhance the visibility of Taiwan-themed Chinese studies research in the international research community, and further attract and increase the participation of scholars worldwide.

In 2018, National Central Library obtained financial assistance from the Ministry of Education to build the Humanities Research Trends System (https://trends.ncl.edu.tw/). The aim of the project was to provide scholars with an online digital platform that would enable them to analyze and explore the content of international sinology websites and help guide them to discover new research topics. The system utilizes software that gathers information from web pages, automatically monitoring the websites and social media web pages of Chinese studies institutions around the world. AI algorithms

and text mining technology are used to extract the latest news and event reports on these downloaded web pages and saves them in a database. The system then uses big data analytics to automatically filter the desired research event related information. This is then fed into a Geographical Information System (GIS), which shows the academic events occurring in each region on a world map and provides analyses of trends. [1]

At present the system analyzes a selection of 300 web pages, social media, and discussion forums associated with research institutions based in Taiwan and overseas, comprising 210 official websites and 90 social media pages. Of these, 105 are based in Taiwan, 40 in China, 70 in the Americas, 50 in Europe, and 30 across Africa, Oceania, and the remainder of Asia. The collected institution web page content is mainly either latest news or event reports, while the social media pages are mainly research institutions' pages on Facebook or Weibo. Discussion groups come in a complex variety of forms, for example Kam-a-tiam (歷史學柑仔店) is in blog format; Story Studio is an article-based platform; and the PTT bulletin boards are forum-based. The main website languages are English, Traditional Chinese, and Simplified Chinese. The system draws on include publicly accessible websites; the structure of the data on community websites differs from the information contained on the institution's webpage, but typically larger quantities of information can be mined, providing the right to privacy is not violated. Apart from this, the content that appears on social media is more of an entertaining, instant nature compared to formally published articles and dissertations, and so is able to more directly reflect research themes of more immediate interest to society.

Until June 15, 2019, the system has captured in total of approximately 85,000 news items, averaging 2,000 items per month. This project is a four-year project. It is estimated that by 2020 the system will be capturing 500 web pages and have stored over 200,000 news items, making it even more able to effectively assist humanities and social science research scholars in creating a greater multiplicity of directions for research.

II. System Design

The system homepage's lively design makes use of interactive analytical charts, while Responsive Web Design techniques have been employed to arrange content for different screen sizes. The charts include "Hot Keyword Analysis," "Research Topic Analysis," "Volume Analysis," "Spatial Distribution Analysis," and "Ranked Popular Articles." Clicking on these interactive charts opens relevant data lists and displays links to the original articles.

A. Hot Keyword Analysis

This chart displays keyword(s) captured using the system's semantic analysis, set by default to the 60 most frequently appearing keywords over the last three months. There are two different modes of analysis: popular and trending. Volume statistics are displayed inside varying sized bubbles. Different colored bubbles can also be used according to the attribute of the word(s) in question (e.g., person, location, group/organization, business/brand, keyword, etc.). Clicking on a keyword bubble takes you to the list of articles and clicking on an article in this list pulls up further detail on that article.

The total matches can be further filtered by origin (official website or social media website); language (Traditional Chinese, Simplified Chinese or English); author; keywords mentioned; persons mentioned; locations mentioned; or businesses/brands mentioned.

B. Research Topic Analysis

Hot topic analysis includes two analysis modes, popular and trending, and displays the top five bubbles by default. Through word(s) captured by the system's semantic analysis, users can rapidly find connected or associated keywords, carry out automatic hot topic matching, and be prompted with keywords.

Fig. 1 The System Front Page

C. Post Volume Analysis

Post volume analysis identifies increasing or decreasing trends in the number of articles posted on webpages and communities, news pages, and forums. Clicking on a peak node enables the user to carry out an advanced search to examine the content of that peak. Growth trends can also be examined for each of five major regions: Taiwan, China, the Americas, Europe, and Asia-Africa-Oceania.

D. Spatial Distribution Analysis

The spatial distribution analysis chart integrates GIS technology to present the spatial distribution data on a world map. The map is zoomable: the top layer shows the world divided into five major regions with the total items for each region, while the second layer shows the totals for each country. Statistics can be presented visually for any location for a chosen time period.

E. Ranked Popular Articles

This chart provides a ranked list of popular trending content or academic research articles for a given time period, as a reference for scholars. Clicking on an item in the list shows the content in more detail.

III. AI AND TEXT MINING

In the age of big data, 80% of all data is unstructured. Though it is difficult to analyze huge volumes of highly diverse data using traditional rules, this data may carry some useful information. This system uses AI and text mining techniques to analyze very large quantities of data, in this case the content of web pages, to extract useful information. [2]

A. Text Mining Techniques

Scholars proposed that web mining could be divided into three types: web content mining, web structure mining, and web usage mining. Of these three types, the emphasis in web content mining is on extracting data from the content of retrieved web sites, including text, links, and index structure. Web structure mining mainly uses graphical methods to describe web page structure and show patterns of links, which can then be used to categorize web sites, whereas web usage mining is used to find patterns of web browsing and access. [3] This study focuses on mining the textual content of web pages themselves using data mining techniques to deal with large quantities of web page textual content. 300 web pages are automatically analyzed to identify the most frequently appearing keywords. These keywords are analyzed to find any trends, connections, and new topics, and then.

Similar to the need to perform internet opinion analysis to understand the market trend before commercial brands go to public, researchers also need to conduct analysis to understand the latest trend or research topics on the internet aside from traditional literature review. This information cannot be obtained simply by using Google. The purpose to build a system like this is to use the AI to analyze the data and explore the connection between them after crawling the websites and text mining is performed.

Most of the pages captured by the system are written in Chinese, which means that the language that appears most in the charts is Chinese. Tokenization of Chinese is innately more difficult than English. A search on English keywords results in a list of English web pages. For example, a search on the Chinese keyword ”Digital Humanities” in the past two years of data, produces 478 results. The top related keywords as below.

The top “Hot keyword” is Digital Humanities Lab which is a laboratory in social hacking in Taiwan. The Facebook of this lab published nearly 1000 digital humanity related news during 2018 to 2019. Key words ranked from the 2nd to the 10th place shows locations that has ever held academic activities by Digital Humanities lab.

B. Post-Category Filters

The post-category filters can be used to further refine the search results. They include the Category Statistics/Filters in the panel on the on the left-hand side of the webpage and the regional filter. These filters enable the user to find connected topics of interest.

- Category Statistics/Filters: Post-Category Filters include origin, language, author, keywords mentioned, persons mentioned, locations mentioned, and group or organization mentioned. These can be used to select the conditions for a cross category filter of the results. Take key word “digital humanities” for example, the system is capable to define filters to show information are of the user’s interest, e.g. to show only related digital archives, related techniques…etc. When applying technique filter, the system will show keywords like cloud computing, big data, artificial intelligence…etc.
- Region Filter: The Region Filter bar is located directly above the charts or listed search results. After checking individual checkboxes for the five regions: Taiwan, China, the Americas, Europe, and Asia-Africa-Oceania, the region filter is activated by clicking on the ‘send’ button. The filter is preset to all regions by default.

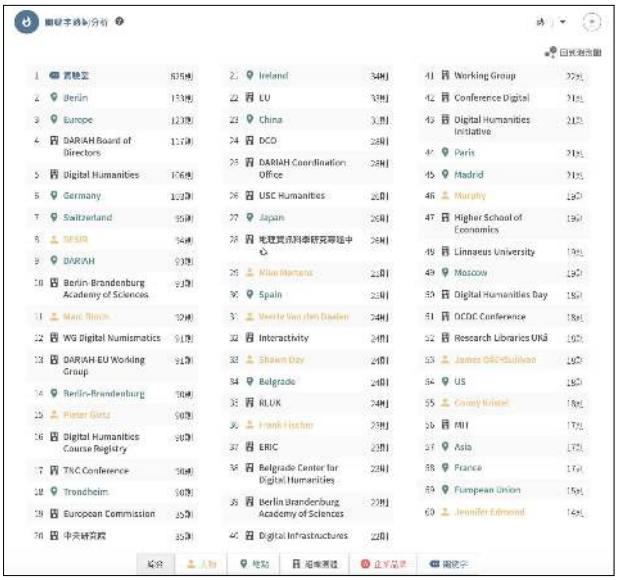


Fig. 2 The 60 most frequently appearing keywords

C. *AI Analysis of Articles*

The brief information shown for each article includes its index number, title, website of origin, author, time posted, region of origin, and a brief snip of the page's content. Clicking on the original URL link takes the user to that article's original web page. Additionally, clicking on 'Share' allows the user to copy the URL of either the original article or the current page to the clipboard.

The system displays 50% of the original article content, with the matched keywords shown in a different color. The user must click on the link to the original URL to display the entire article. Any comments to the article are also captured and shown below the main body of the article in the order that they appear with the original. The comments header shows the total number of comments. The respondent's account, body of the comment, and date and time are captured for each comment. The system automatically captures the keywords and various tags from the article, including any keywords, persons, locations, or groups and organizations it mentions.

An example news item, "Poets, Artists, Game Makers and New Media," is shown in the image below. It was posted on November 17, 2018 by Cornell University Department of Asian Studies. Besides providing a link to the original article, the system has automatically captured its keywords, such as persons mentioned (Brett M. de Bary, Andrew P. Campana), locations mentioned (Japan) and organizations and groups mentioned (Massachusetts Institute for Technology).

Long term tracking and correction are required to further optimize and improve the accuracy of the system's data analysis. With input from users and scholars in building and debugging the system, using sustained and automated methods, the AI word recognition will gradually strengthen and enrich the existing system content and data retrieval mechanism, while increasing its efficacy. It is hoped that humanities scholars will discover new research topics, and that the fruits of these discoveries will be seen far and wide.

IV. Future Plans

The aim of the Humanities Research Trends System is to try to archive and summarize international Chinese Studies research, accumulate research experience, and provide tools with which humanities scholars can discover new research directions. The goal of next stage of this project is to gradually increase the number of web pages analyzed, and by analyzing the full text of news items posted by sinology research institutes around the world, rapidly provide the E-Newsletter for Research in Chinese Studies [4] with the news it needs, avoiding the need to carry out time consuming searches for event news. At the same time, statistics and analyses produced using the new tool will enrich the E-Newsletter's content. Despite the fact that in the current stage the system does not perform text mining well enough to recognize all relavent keywords, we believe though continuous training the system by entering new academic keywords to improve the AI model, the system will gradually evolve and improve over the next 1-2 years. The final goal is to help scholars understand the fluidity between the online exposure of academic ideas, research presentation, and publication.

Fig. 3 Analysis of Articles

REFREENCES

[1] 黃文德、廖箴,"以人工智慧探索人文研究趨勢—漢學研究中心「文史哲學術趨勢分析系統」介紹 in Newsletter for Research in Chinese Studies, Vol. 38, No. 1, pp.43-46.

[2] 賴彥文, "數據資料擷取與分析探討: 以 Google Trends 為例," 國立中興大學應用數學系所碩士論文,2018.

[3] R Cooley, B Mobasher, J Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web. ictai 97, 558-567

[4] The website of E-Newsletter for Research in Chinese Studies is https://ccsnews.ncl.edu.tw/

An Innovative Dynamic Visual Model for Digital Content Archives in Traditional Cultural Heritage – Evidence from Stone Weirs Fisheries Culture in Penghu, Taiwan

Ju Chuan Wu
Department of Business Management
Feng-Chia University
Taichung,Taiwan
katejcwu@mail.fcu.edu.tw

Jui Chi Wang
Enterprise Information System Center
Feng-Chia University
Taichung,Taiwan
peggywang@mail.fcu.edu.tw

Shu Mei Lee
Department of Business Management
Feng-Chia University
Taichung,Taiwan
lineamylee22229@gmail.com

***Abstract*—History is the local life culture developed from the observation to environment and the environmental characteristics by the local residents. And now become the reference for humanity researcher to inference the local early life and culture. The rapid development of information technology and content digitization in recent years has brought many cross-field technical support and applications, such as the use of digital collections in the field of historical humanities research. The use of information technology to enhance the degree of professional knowledge and digitization of industry and enhance culture role in the economic status. 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.

Therefore, this study takes the stone culture of Penghu as an example, and carries on the dynamic presentation design of the multi-source in history and culture from the perspective of digital humanities. First, summarize the elements of the stone weirs activities depend on Activity Theory, 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. Through the digital form of sustainable management of stone weir knowledge, it can be used in other historical and cultural digital collections construction design, and use the information and communication technology in multi- dimensional observation and research, such as the age, geographical environment, and humanities and social changes and other relations. To effectively assist in its exploration the potential meaning and value in historical context.

Keywords: Stone weirs, Digital Humanities, Activity Theory (AT), Hypermedia, Serviced-Orientation Architecture (SOA)

I. INTRODUCTION

A. Background and Motivation

Overview at the country, region, and the local history, local people usually evolved out their own living culture depend on the environment and living conditions. Humanistic researchers nowadays could inference early local life and culture by those remains. With the development trend of information technology and digitalization in recent years, it brings many cross- domain innovative solutions to the field of industry and academia. In the field of history and humanities, in view of the preservation of culture and the needs of researchers, the concept and technology of digital collections create a new way of storing, presenting, and educating, effectively reducing the preservation and presentation of the current cultural preservation and humanities research. The quality and quantity of Taiwan in the digital collections have reached a certain level, and its affiliated digital content industry has the characteristics of developing digital and knowledge economy. By using information technology to enhance the professional knowledge and digitization of industry, and to improve the connotation of the quality and competitiveness of the important basic. And highlights the role of culture in the economic status, led to focus on preserving the local traditional culture of national and humanistic scholars using digital collections, the local life, and the culture. In the form of digital preservation and presentation, hope through the digital collection for the knowledge level of human sharing and multiple data value-added applications to promote the local culture of life, so that these precious cultural intellectual properties could be managed through the form of the digital collections sustainable, for example: the early stone culture which only exists in Penghu nowadays preservation and presentation.

However, 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: first of all, the past research results show that the man- machine interface is one of the important factors in the use of digital collections[1], so the man-machine interface design is particularly important, and the multiple resources of the stone weirs are not effectively integrated and applied; secondly, because the historical data of the stone weirs is not complete and has many omissions, it needs to spend many time, manpower and resources to collect and construct, for example: field survey and visit nearly 600 stone weirs and their owners, the digital database import, integration and presentation, and even the maintenance management all need support and assistance. As the stone weirs humanities, environment, and community elders dying, the stone culture and techniques which heritage by mouth and ears are gradually lost. The stone cultural preservation and maintenance has become the urgency. Due to the ICT application gap and the lack of historical data, how to save, organize, present, and maintain these valuable history and culture has derived the motivation of this research and the current considerations of establishing the digital collections.

Therefore, this study takes the Penghu stone culture as an example, and carries on the design of the dynamic presentation structure of the historical and cultural sources from the view of the digital humanities. First, summarize the elements of the stone weirs activities depend on Activity Theory, 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 through the dynamic presentation design of the multi-source in history and culture.

B. Purpose

The purpose of this study is to save the original pattern of history and culture, and to present dynamically via multi-source combination. The case study of the stone culture in Penghu combines digital contents, digital humanities, Activity Theory, Hyper- media, and Service-Orientation Architecture. Different from the digital collection presentation in the past, the integration of multi-source from digital media, visual interface, parameter filtering, and retrieval services provides users to view multidimensional perspective and multi-source digital information of stone weirs. Dynamic intuition presents a multitude of stone weirs culture and appearance, and uncovers the potential value. Integrated above thesis, the purpose of this study is as follows:

1) Meet the needs of digital collections in history and culture contents: Including human-*oriented and integrated structure of information history and culture (Digital humanities), summarize the composition elements framework of the activities of the stone culture (Activity Theory), and construct the multi-pattern data model of the stone culture.*

2) Meet the need of digital collections in ICT: Establish *the integration system of multi- source digital media resource, content pattern, and interface elements of the stone weirs (Hyper-media), and establish the dynamic presentation system of the stone culture (Service-Oriented Architecture)*

II. THEORETICAL BACKGROUND

This study reviews the stone culture, digital content industry, hypermedia and service orientation related researches. Then summarizes the elements of the stone culture, the patterns of the stone weirs, interface visualization and connection, and service assemble to deduce the dynamic presentation design of the multi-source in history and culture.

A. Stone weirs in Penghu

Stone weir fishing is an ancient world fishing activities. The first stone weir originated in the Neolithic Age, throughout Korea, Japan, Ryukyu, Taiwan, Philippines, Thailand, Polynesia, Melanesia and other Pacific Islands have the stone weirs distribution. According to the research, the stone weirs group of Penghu is one of the world's most densely populated and most intertidal large-scale structures, which is one of Taiwan's most potential entries for the world's cultural heritage. Origin time has been unable to study, the original literature recorded in the 1720 A.D. Taiwan County Notes, the stone weir estimated history has

been more than 315 years. The fishermen were mostly in the intertidal zone engaged in fishing activities, and later the use of the seahore vine plants to woven into a net to encircle the fish, and to observe the characteristics of the fish stocks due to the tide, and gradually develop the local unique fishery model - stone weir. As the construction of stone weir need to spend a lot of manpower and time, the sponsor raise funds to gather the working population within the community. Relying on economic capacity to adopt the amounts of shares. Use basalt (black stone) and coral stone (white stone) as the main building materials, and use bamboo raft to carry building materials to fixed- point, then use the "stone method" to fill the stone weir, gradually evolved into a bail-out of the fishery economy developed by the nature of the natural fishing traps. Therefore, Penghu County Marine Culture Association professor Lin Wen-Zhen has considered the Penghu stone weirs have the three kinds of value which are the early economic lifeline, the earliest ecological work, and the community overall construction.

1) *The early economic lifeline: the total catch of Penghu in 1950 A.D., stone weirs output accounted for 77% of the total. It was the most important fishery way in the time that the mechanical power ship has not yet developed. The island which is lack of resources like Penghu, if there is no stone weir is equal to no living resources, it is difficult to keep a family and cannot be established. At that time, stone weir and field are also play the role of the property, it can be borrowed, mortgaged, traded, and even as the heritage, but also the measurement of the benchmark for wedding. This feature has last until 1960s, after the industrial structure changes and mechanical technology is mature, then gradually change.*
2) *The earliest ecological work: As mentioned above, the piles of the stone weir doesn't use any artificial adhesive, through the rain and the construction of stone chemical reaction to solidify the stone, so there will not be artificial pollution, and the surface of stone weir is rough and leaves a gap, thus becoming a hotbed of epiphytic shells (e.g. stone oysters, barnacles) and algae, and thus increasing the strength of the structure. Because of the stability of food and pores, it also become the best place for the growth of juvenile fish.*
3) *The community overall construction: As the construction of stone weirs need a lot of manpower integration, so teacher Hong Guo- Xiong has pointed out that the construction and division of labor in the stone weirs can be divided into selected shareholders, selected location, collected stone materials, delivered stone materials, construction time, construction rules, stone weir maintenance, and worship stone weir these eight dimensions. These works are required to share with all the owners of stone weir, so the stone weir often has the close relationship with the clan, temple, or even village. And form the system of a learning, inheritance and self- management mechanism as a feature of the island culture at that time.*

Overview at the above appointments, stone weir has become the economic lifeline of the early residents and the social enterprise model, evolved out a set of "public production" of the joint venture heritage. And the stone weir equity will transfer with the marriage, trading, leasing relationships by donating or shifting the existing fishery forces. No matter the distribution of the economic and social status and the labor force for the entire marine community, the allocation of fishery resources, or the economy and life within the community (e.g. marriage, faith, etc.), there is far- reaching impact, and its own stories and the spirit. Those stories were spread by the shareholders and their descendants in "word of mouth" so far. So, these stone weirs around Penghu is the best embodiment of Penghu's life and culture, and has the value of history and preservation for humanistic researchers in the study of Penghu early life and cultural level of reference value based on its development. However, the related cultural assets are dying year by year. Since the information and communication technology and digital preservation technology matures, the concept of stone weir digital collections plans to implement, such as digital collections - the stone weir form and culture in Penghu.

B. Digital content and digital humanity

With the development of knowledge economy and information and communication technology, the demand for big data driven, the digital economy has become an important indicator of future industrial development. Therefore, the digital content industry has become one of the world's most dynamic and growing industries [2]. In addition to the development of digital content indicators, it also has the meaning of knowledge economy indicators. Taiwan Ministry of Culture defines the content and the scope of digital content industry as "Digitalize images, characters, films, voice and other information and integrate the use of technology, products or services. The various types of content material after digital technology production processing, transform from the traditional data into digital format, and become a new application form. It contains the following advantages, such as easy to access, interactive, transmission, copy, search, edit and reuse. Through the Internet, mobile communication network, wireless and cable television, satellite communications, movies, digital broadcasting and other "media carrier", then by networking, smart TV, smart phones, personal computers, tablet PCs, MP3, AR And VR system equipment, transmission to the consumer or institutional users to use, that is, the formation of a complete digital content industry structure." . Therefore, all the digital publishing, digital games, computer animation, digital learning, digital audio and video, mobile applications, network services and content software are digital content core industry.

Overview at the application of digital content in the field of education, marketing and tourism, and the digital value collection of cultural assets, it comprehensively affects people's life, work, study, entertainment, knowledge, culture and education. And the accumulation and application of digital content has become the basis of digital humanities. From the early form of text analysis, and gradually turned to hypertext, digital knowledge and multimedia integration [3], the advantages for the use of science and technology to restore the previous paper interval caused by the historical context, breaking the situation, time and space, and past media form. The diversity of knowledge products and digital collection system is no longer regarded as the "additional items" in the process of archiving, but to think as design-oriented. According to the characteristics of the contents, combined with the context of time, space, and event which construct the elements of history the context, and then organize the digital humanities system.

Therefore, the digital humanities should begin with the theory of traditional humanities, and focus on the integration of knowledge related to information and communication technology and professional fields. According to the characteristics of these materials, to organize and to show the meaning of multiple context. Use the application of computing technology for humanity investigation and its significance for the humanities [4], for example, the combination of GIS space-time application of Taiwan's landforms history; therefore, this study uses the activity theory to summarize the cultural elements of the stone culture.

C. Activity Theory

Activity Theory development is based on the Cultural-History Theory. The interaction between the subject and the object is carried out by the tool [5], and then extended by the scholar Leont'ev (1978) who put forward the concept of "activity" and derivative the framework of the second-generation activity theory [6]; the activities related to the object, the purpose of the project, the application of tools, and language-related projects are set in the framework [7], and contradiction is the core of activities analysis. Through three major intermediaries to explore the complex process of the relationship between the activities and further verify the activities of the relevant framework and analysis of the units: (1) tool unit on the subject - object intermediary path; (2) rules unit to the community - subject main path;(3)division of labor unit to the community - object intermediary path. The other additional paths may have secondary mediating relationships [8], which can help to understand complex work and social activities [9].

The activity theory is mainly used in the teaching, organization, human-machine interface (HMI) or system-building issues. Recent developments in the field of information technology have increased the trend [10], for example, to synthesize the framework of adaptive digital learning system, the exploration of the topic is focused on the system and behavior integration. For the field of digital humanity research, activity theory can provide an integrated framework to compute the stone weirs activities elements, and show the complete stone culture.

Therefore, this study uses the framework ofactivity theory proposed by scholar Engestrom (2000)[11] and refers to the Eight-Step-Model [12] proposed by scholar Mwanza (2001), and based on the elements of the second-generation theoretical framework to go through the clarification of the eight open questions for the establishment of each elements in stone weirs activities and as follows:

1) *Tools mediating the activity:*
   Construction of stone weir (stone weir structure): latitude and longitude, construction type, stone materials, delivery methods, the number of wells, the direction, Lu-lian, the area, the number of teeth, naming, construction time.
   Stone weir fishing: main fishing tools, fishing types, tide.
2) *Subjects in this activity:*
   Sponsor and shareholders raise funds and build the stone weir together.
3) *Objects:*
   Early residents build stone weir to fishing for family food and clothing beyond the development of a limited degree of marine economy, and the formation of mutual benefit life style with the ocean.
4) *Outcome:*
   fishing, develop a limited marine economy, form a “mutual benefit” life style with the ocean
5) *Activity derivative:*
   stone weir stories, stone weir appearance (photos/films), social relationship
6) *Rules and regulations:*
   fishery license, property map, fishing areas, using status (in use/ decline)
7) *Community:*
   stone weir owner, shareholders
8) *Division of labor:*
   rules for fishing, stone weir flow, stone weir recoding book, stone weir ceremony, company flow, number of share holds, equity details

D. Hypermedia

The concept of Hypermedia is extended from Multimedia, which not only links logical and semantic- related information in Network-Form, but also provides various digital media-type resources provided by multimedia with content-oriented Hyper Text information related facilities [13]. And transform the information nodes from the text elements into multimedia elements, and independently presented in text (Textual, Graphic, Video (Video) and audio media (Audio Media) [14,15]. Through the link in the node, non-linear cascade into the mesh data link structure as information presentation and presentation system, and with the four major advantages which are non-linear, relevance, flexibility, and efficiency [15].

Widely used in teaching and business marketing activities, hypermedia can organize the system logic according to the characteristics of knowledge, but also provide users with knowledge according to their own needs to organize. To provide users with rich and dynamic human-computer interface, and to attract users to explore from the man-machine interface interaction, but also provide a simple graphical operation. Allowing users to combine their own browsing needs, starting a node of the special link to reach the destination node for a wide range of reading, communication and information work culture [16]. Be able to display a number of specific topics for the specific theme, compared to the linear form, more conducive to complex things reasoning [17], and can be checked for the activities of the stone weir in the context of each view. And provided humanities researchers the diversity of exploration.

Therefore, this study uses the concept of hypermedia to integrate multi-source digital resources, content context and interface connection as reference basis, and to construct the dynamic presentation architecture of the media data link series. The main interface presented in this study is stone weir knowledge map, the others include time axis (time dimension), the latitude and longitude (organ dimension), the screening condition (event dimension) as the triggering interface elements to change the rules. According to the triggered filter conditions, showing the stone weir map and each detail information (stone weir details, stone weir story, stone weir photos, and stone weir films). In order to effectively combine the interface of hypermedia components, retrieval services, and system processes, this study uses the service- oriented architecture to present the structure of the internal logical design.

E. Service-Oriented Architecture

Under the competitive environment of the global economy, how to respond quickly to market demand, effectively integrate the upstream and downstream cooperation manufacturers business processes, and in accordance with the needs of enterprises to adjust the flexibility of business rules, vital to the survival and development of enterprises. The application of the network has been able to link the enterprises, customers, suppliers, and channels, but the extension of inside and outside enterprise information system is still not easy. The reason comes from each system may be used different platforms, technologies, protocols, data definitions, and security mechanisms, with the increasingly complex enterprise IT system architecture and the enterprise environment changes, the demand for different systems integration is increasing, so the concept of Service-Oriented Architecture (SOA) born. SOA is an emerging system architecture model. The main concept is started from the point of view of service science. In a complete business automation behavior, the need for a number (or numbers) of different systems to provide the "service components" for service. And the decision to use the order of business processes and all service components [18]; therefore, the combination elements usually include three parts: software components, services, and processes [19,20]. To manage the business process and code hierarchically and to build SOA components, assembly and application, so SOA

components also have the characteristics of reusability, low coupling, composability, and permeability.

Liu, Wu et al. (2009) pointed out that SOA provides an opportunity for enterprises to reorganize or to build an open and resilient technology architecture that integrates internal and external applications with rapid response to enterprise environment changes [21]. Through service-oriented applications, when the industry rules and business processes change, it only need to replace several SOA components to achieve interactive software applications between the low coupling [22] and rapid response to the changes and challenges in demand. SOA has the features and flexibility to provide this study integration of multi- source components, multiple contexts interface and services and information layer, while the future of human resources to explore the expansion of demand for structural adjustment.

Therefore, this study starts from the point of view of service design and innovation, combined with the application of information and communication technology, through the concept of innovative service design to integrate the dynamic presentation design of the multi-source in history and culture structure of digital humanities, stone weir activities, and hypermedia to show the historical dynamic appearance and characteristics. Leading the user into the historical situation of stone weir which through the simple intuitive elements of the combination to show a multi- oriented stone culture and to provide a more accurate analysis to meet the user needs initiatively.

III. METHODS

This service system development tool is Microsoft Visual Studio 2017. The developing language is HTML5 (Javascript, Jquery, CSS), and use Bootstrap3 technology to build the interface. Use the logic of C # to construct the logic service process, and use IIS to set up a service website. Use SQL syntax to obtain Microsoft SQL Server data.

The service system in the system analysis and design stage use Visual Paradigm for UML 3.0 for system analysis and design. This study uses the system spiral development method. Visit the stone weirs owners, collect secondary resources, review the relevant stone weirs literature as the first phase for capturing the demand. Establish the initial needs, users (viewers and managers), and the use of features: dynamic content presentation and background maintenance and management. According to the functions divide into two execution stages are as follows:

First stage – Design and develop the dynamic content presentation structure:

1) *Demand capture and system analysis: The service system collects the current resources of the stone weirs, visits the owners and environment, summarizes the activities elements of stone weirs by Activity Theory, and understands the characteristics of the stone weirs data.*
2) *System design: The service system dismantles and assembly the stone culture (activities) context, and do the data regularization, resulting in data related graph (fact and dimension data table), category map (search service), and interface blueprint (interface elements), then construct the stone data model.*
3) *Interface visualization and connection: The service system uses the concept of hypermedia to visualize the interface, to connect the elements, and to establish the interface blueprint and vocabulary (the control type).*
4) *Dynamic process service construction: The service system design based on the service- oriented architecture. Assemble and connect the interface layer (interface blueprint), service logic layer (interface category, data category) and data layer (data associated graph), and establish the use of service processes.*
5) *Construct the dynamic presentation structure of multi-source pattern from stone weirs: the service system digitalizes the previous data, and transacts the stone weirs data into testing process after the system development. Prototype system will be test and feedback by the humanities researchers, and based on the actual feedback to adjust the system.*

Second stage – Design and development of background management:

1) *Demand capture and system analysis: Interview with relevant cultural website managers to capture the demand, and in accordance with the dynamic presentation structure of the organs, events, time context and fact data sheet, and divide the use of cases and service process.*
2) *System design: According to the use of cases and service process, and based on the dynamic presentation architecture design principles, drawing the using case map, flow chart, tough map, sequence diagram, interface blueprint, category map data association graph.*
3) *Background management functions development: Integrated the first stage dynamic presentation architecture for background management functions development. Prototype system will be test and feedback by the humanities researchers, and based on the actual feedback to adjust the system.*

Based on the activity theory framework, the composition factors of the stone weir activities are the basis of data integration, and the three elements of history are used to establish the element dimension of the event. Data Warehouse storage architecture - multi- dimensional data cube for the stone weir activities of the composition, and then deduce the rules of multiple context (dynamic business logic) and the corresponding presentation interface; the richness and diversity of the interface, so that users can easily enter the context to find the potential implication, and the past results also show that the human-computer interface for the use of digital collections in the use of important factors [37]. The design of the man-machine interface is particularly important, focusing on the interaction between users and the system of human-computer interaction, providing a simple way for users to find and imply the potential meaning in the history. On the other hand, how to integrate and present multi-source digital resources under the premise of visualized interface; therefore, this study adopts the concept of hypermedia to carry on the multi-source digital media resource interface element composition, contextual content and interface integration design to assist in the presentation of the dynamic structure of multiple contexts.

1) *Multi-source stone weir related resources collection: resources in paper form, currentdigital collections, field study resource, owner interview*
2) *Multi-source stone weir resources analysis and collation: summarize the stone weir activities elements by activity theory, and make sure the completion of the resources. Construct the data base after 3NF.*
3) *Construct the data model (information storage): multiple context division.*

The basic elements to establish the history: organ context (stone weir basic information), time context (evolution time), event context (related parameter), fact data table (the details of stone weir, the story, the film, the photo).

This digital collection system takes organ, time, and event context as the dimension of data model.

4) Build the multi-layer structure:

- Build the multiple context filter interface: Time dimension (time line), Event dimension (parameter filtering), Organ dimension (latitude and longitude, map), Fact data table (the details of stone weir, the story, the film, the photo)
- Build the searching rules: Organ -Time (Searching stone weir information through evolution age), Organ - Time - Event (Searching stone weir information through evolution age and condition parameter), Time - Event (Searching stone weir allocation through evolution age and condition parameter), Organ - Event (Searching stone weir information through the stone weir parameter)

5) Front-end application interface

IV. Results

A. Service System Concept and Architecture

The service system architecture can be divided into stone weir knowledge map and background management. Stone weir knowledge map is mainly used for inquiries, and background management is the management of stone weir knowledge map parameters. The relevant system concepts are as follows:

1) The historical elements for the design: In the process of historical observation and research, it should be guide and interpret the complete story to lead the researchers into history, and then converted into the detailed observation between the historical context. Time, events, and organs context is the basic elements of history. These are the indispensable factors while humanistic researchers engaged in the past human studies, combined with historical elements: organs, time, event context to organize the stone culture history and to show the dynamic characteristics. Leading the user into the historical situation of the stone weir. Through the simple combination of intuitive elements, to show a multi-oriented stone culture and to provide a more accurate analysis to meet the user needs initiatively.

2) The link between multi-source content: in accordance with the historical attributes (such as: picture of stone weir appearance; text and picture of stone weir story) using different digital media resources as the stone weir history presenting tools. Connect he series of digital media resources and stone weir history for dynamic content to present the real history of stone culture life.

3) The combination of digital humanities and digital collections: Most of the previous collections show historical content in static design, which makes the users has limited in the research and observation of digital collections. Users can only use the data retrieval and query function for human studies. Digital humanities is based on the humanistic research point of view, through the application of information technology and digital resources, the use of digital collection of digital work processes, the establishment of a context between the historical relationship between the presentation system to fill the lack of historical data itself and to provide more research and observation oriented. And take the initiative to close with the needs of users. The system is based on the design of the digital collection system. Users can follow the demand combination to observe, establish, and explore the context between the relationship of stone weir content.

4) A service-oriented presentation architecture: The service system architecture can provide users with information technology vehicles such as mobile devices or computers to view the distribution of the stone weir in Penghu on the map and to expand the structure in response to future requirements.

B. Service System Function and Interface Introduction

The service system users in accordance with the use of authority is divided into administrator and general user, administrator can maintain the relevant parameters of stone weirs details, stories, multimedia resources, and the general user can browse the stone address and view the stone weir history and culture through the screening mechanism, and the corresponding function of the user as follows:

1) Administrator

- Parameter data management: administrator can manage the parameters through the parameter data management function, add, modify, inquire, deactivate the map in the filter parameters.
- Media data management: administrator can manage the media resources through the media data management, add, modify, inquire, deactivate the films and photos in the map.
- Stone weir data management: administrator can manage the stone weirs data through the stone weir data management, add, modify, inquire the stone weir data.
- Stone weir story management: administrator can manage the story through the stone weir story management, add, modify, inquire the stone weir story.

2) General user

- View the stone weir knowledge map: the user can directly click on the stone weir in the map, and view the latest information of stone weir general information, story, film collection, and photo collection.
- Inquire the parameter condition: the user can set the conditions through the screening query for specific conditions under the stone weir knowledge map, and also can be further selected the stone weir query of the information, story, film collection, photo collection.

Fig. 1 Main Interface

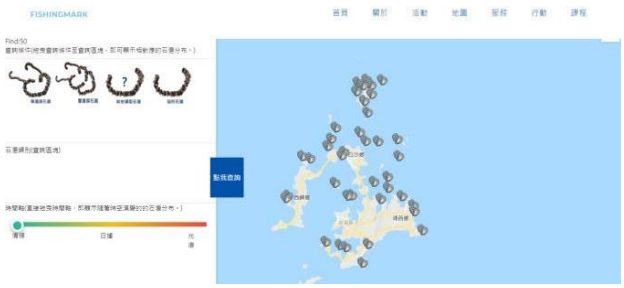


Fig. 2 Stone weir knowledge map (searching interface)

Fig. 3 Stone weir knowledge map (filtered subquery)

V. Conclusion

This study presents a dynamic design of multi- source in history and culture, which provides humanistic researchers a historical knowledge map which can be combined autonomously to meet their research needs, and integrates multi-source digital resources. From the perspective of digital humanities, the design structure of historical and cultural characteristics can provide users a multi-angle view of historical and cultural implications, to effectively assist in the exploration of the potential meaning and value in historical context, and can be applied to other digital collections in history and culture demand.

In this Penghu stone weirs case study, the application of digital collections preserves and presents the local community elders' life stories with the stone weirs, and inherit the local seniors culture and life knowledge. To decrease the risk of internal oral culture and the loss of knowledge, the results of this study provided a value-added solution to the local culture and education institutions to help set up the related courses of the stone culture. Through the link between the stone weirs map and the dynamic images, users can cultivate cultural literacy of ocean, and arouse the awareness of marine culture preservation. To maintain the stone weirs together and to fulfill the social responsibility that the ancestors teach descendants the "mutual benefit" concept through the stone culture.

References

[1] Hong, J. C., Hwang, M. Y., Hsu, H. F., Wong, W. T., & Chen, M. Y. (2011). Applying the technology acceptance model in a study of the factors affecting usage of the Taiwan digital archives system. Computers & Education, 57(3), 2086-2094.

[2] González-Rojas, O., Correal, D., & Camargo, M. (2016). ICT capabilities for supporting collaborative work on business processes within the digital content industry. Computers in Industry, 80, 16-29.

[3] Dalbello, M. (2011). A genealogy of digital humanities. Journal of Documentation, 67(3), 480–506. doi 10.1108/00220411111124550

[4] Sula, C. A. (2013). Digital humanities and libraries: A conceptual model. Journal of Library Administration, 53(1), 10-26.

[5] Miettinen, R., et al. (2009). "Re-turn to practice: an introductory essay." Organization Studies 30(12): 1309-1327.

[6] Engeström, Y. (1987). Learning by Expanding: An Activity-Theoretical Approach to Developmental Research.Helsinki: Orienta-Konsultit.

[7] White, L., et al. (2016). "Understanding behaviour in problem structuring methods interventions with activity theory." European Journal of Operational Research 249(3): 983-1004.

[8] Georg, G., Mussbacher, G., Amyot, D., Petriu, D., Troup, L., Lozano-Fuentes, S., & France, R. (2015). Synergy between Activity Theory and goal/scenario modeling for requirements elicitation, analysis, and evolution. Information and Software Technology, 59, 109-135.

[9] Karanasios, S., et al. (2015). "Information Systems Journal Special Issue on: Activity Theory in Information Systems Research." Information Systems Journal 25(3): 309-313.

[10] Allen, D. K., et al. (2013). "How should technology-mediated organizational change be explained? A comparison of the contributions of critical realism and activity theory." MIS quarterly 37(3): 835-854.

[11] Engestrom, Y. (2000). Activity theory as a framework for analyzing and redesigningwork. Ergonomics, 43(7), 960-974.

[12] Mwanza, D. (2001). "Changing tools changing attitudes: effects of introducing a computer system to promote learning at work."

[13] Hardman, L., Bulterman, D. C., & Van Rossum, G. (1994). The Amsterdam hypermedia model: adding time and context to the Dexter model. Communications of the ACM, 37(2), 50-62.

[14] Akscyn, R. M., McCracken, D. L., & Yoder, E. A. (1988). KMS: a distributed hypermedia system for managing knowledge in organizations. Communications of the ACM, 31(7), 820-835.

[15] Liu, M., & Reed, W. M. (1995). The relationship between the learning strategies and learning styles in a hypermedia environment. Computers in human behavior, 10(4), 419-434.

[16] Paterson, B., Winschiers-Theophilus, H., Dunne, T. T., Schinzel, B., & Underhill, L. G. (2011). Interpretation of a cross-cultural usability evaluation:

A case study based on a hypermedia system for rare species management in Namibia. Interacting with Computers, 23(3), 239-246.

[17] Kornmann, J., Kammerer, Y., Zettler, I., Trautwein, U., & Gerjets, P. (2016). Hypermedia exploration stimulates multiperspective reasoning in elementary school children with high working memory capacity: A tablet computer study. Learning and Individual Differences, 51, 273-283.

[18] Bieberstein, N. (2006). Service-oriented architecture compass: business value, planning, and enterprise roadmap: FT Press.

[19] Krafzig, D., Banke, K., & Slama, D. (2005).

Exploring Service Concept for Digital Scholarship Support

Xin Li
Library of Cornell University
Cornell University
Ithaca, New York, U.S.A.
xin.li@cornell.edu

***Abstract*— Over the past two decades, research libraries in the U.S. have advanced steadfastly from following academic discourse about digital humanities to providing a suite of services to support scholar's practices. The services require different integration of technical, pedagogical, subject expertise, as well as movements between physical and virtual spaces across campuses. Many U.S. research universities have digital scholarship labs or centers now, and many libraries have digital humanities or data librarians on staff. However, the support structure has been formed opportunistically, resulting in fragmented workflow that wastes library's resources and risks library's ability to provide sustainable services. This paper discusses the takeaways from Cornell University Library's work, including user studies and immersion programs for graduate students in sciences, humanities, and social sciences, to explore building blocks for an intentional service model.**

Keywords—academic libraries, digital scholarship, research support, library services, service model

I. Introduction

This paper examines the type of library work like plumbing. The pipelines are invisible when they work well. Their importance only shows when one leaks. The conference organizer asked the presenters to share our own institution's work in digital scholarship support. Hence, the discussion below focuses on my home library (Cornell University Library) and orients toward digital scholarship.

Abby Smith Rumsey noted that, "[d]*igital scholarship* is the use of digital evidence and method, digital authoring, digital publishing, digital curation and preservation, and digital use and reuse of scholarship." [1] What should the digital scholarship support look like in a library service menu? Experts in operation management science tell us that the service concept plays an important role in the process of service design. "From the service organization's perspective, designing a service means defining an appropriate mix of physical and non-physical components. But do customers define a service as a sum of components? Or do customers define a service as a singular outcome they are seeking when they obtain or purchase the service?... Regardless of how the service organization defines their service and how customers perceive the service, a delivered service should function seamlessly for customers to perceive it correctly (i.e. as designed)." [2] My library does not see users as customers, with whom we have a transactional relationship. We see them as our community, to whom we ourselves belong and with whom we partner to deliver the university's educational mission. This relationship does not change the importance of using a service concept to drive our design.

From this standpoint, I include the support for digital scholarship within the realm of library's support to research. I propose that we, in the service context, lower the semantic emphasis of digital scholarship's uniqueness for library's own sake. Wherever the term, data, is used, I refer to the broadest range, be it numerical, visual, or textual, which would include collection as data. Wherever "scholarship" is mentioned, I refer to all practices, including digital humanities, and thus I use collection and data interchangeably. I also use "researcher", "scholar", and "user" interchangeably, although I understand that faculty in different disciplines prefer one term over the other. Focusing on the service concept is important because of our users. We agree with the findings in reports like the one by the Educause Center for Analysis and Research that the digital humanities scholars may or may not self-identify as such [3]. To them, it is all about searching, thinking, reading, and writing [4]. For them, there is no demarcation between digital and non-digital scholarship, although it always has been understood that some scholarships result from computation and others do not.

The user observations mentioned below come from several local sources:

- The Cornell University Library conducts user study regularly. A survey of faculty (2014, with 1,670 participants) and another of graduate students (2016, with 2,400 participants) are especially relevant to this discussion[1].
- A team of Cornell Library staff conducted a research project that observed twenty-one undergraduate, graduate students, and faculty from various disciplines [4] .
- My library has been offering several immersion programs[2] and a Summer Graduate Fellowship in Digital Humanities[3]. Library staff who have taught the sessions shared their experiences through staff presentations and with the author directly.

[1] https://ac.library.cornell.edu/data (Some reports are restricted to staff use.)
[2] http://guides.library.cornell.edu/olingrad and https://guides.library.cornell.edu/ScienceImmersionProgram
[3] https://digitalscholarship.library.cornell.edu/education

The EDUCAUSE Center for Analysis and Research working group has described four representative models to support digital humanities: centralized, hub-and-spoke, mesh network, and consortia [3, pp. 12-16]. Many research libraries in the U.S., adapted them to suite their institutional needs. Agnostic of the library-level service model, operational elements need to be examined. I highlight a few different areas below with the full awareness that the list is not exhaustive. Many digital scholarship issues, such as open access, open data, publishing, and other scholarly communication topics are not discussed.

II. Collection as Data

User observations: The data needs from our researchers differ significantly from discipline to discipline. Some don't need library's help at all as they have well-formed communities with datasets, standards, and archiving protocols. Others start with "where do I begin?". Many datasets that my library acquired for users are not big data. In our 2014 survey, faculty identified lack of time, funding for research, and research skills (what is out there and how quickly can I get it?) as their top academic challenges, whereas graduate students named scholarship, competing demands, and finding and managing information as their top academic challenges.

There has been a lot of great work done on operational, technical, and cultural issues of collection as data. The Always Already Computational: Collections as Data project sponsored by the Institute of Museum and Library Services is such an example [5].

In Cornell's case, we have actively negotiated for text mining terms in our licenses. But not all agreements have text mining rights. Where we have secured them, whether the data is hosted by the vendor or sent to us on a physical medium, we have no standard workflow to describe the terms and make them searchable and usable by users. We have no current-awareness mechanism to make our own liaisons and subject librarians aware of their existence and thus lose the opportunity to compensate, even if only by word-of-mouth, for the non-searchable system flaw.

When we scan our own holdings, we continue to create digital collections for reading not for mining due to the daunting task of funding and maintaining a new infrastructure to support them. We are not alone. Padilla et al pointed out, "… community needs are constantly changing, collections as data are varied in implementation. Efforts to meet these needs benefit from collaboration across multiple communities of practice." [5, p. 16] Prudent evaluation needs to be made as to whether providing locally created digital collections for computational research is a library priority. If it is, how to make this a reality, with whom, and at what scale? Currently, we rely on our consortial partner, HathiTrust, to make our locally created digital collections minable through its effort to support computational research.

If each academic research library were to add up the investment we have already made to acquire, license, and create minable collections that are lying hidden, the waste is significant. Steven Wheatley compared the challenge of new scholars and new disciplines to that of entrepreneurs. "Every scholarly career is something of a start-up enterprise" [1, p. 163]. Not making them findable and accessible is costing scholars' opportunities and libraries' relevance.

To address this, we are building functions into our new library management system to document text mining terms so they can be fed into the discovery system for access. We aim to negotiate for mining rights in licenses that lack them and develop workflow for processing, and communication channels for making staff aware of their availability. Where users have their own established community and do not need the library to acquire data, we focus on understanding the services gaps.

III. Metadata as Connector

User observations: The 21-user research project revealed that "…scholars seek other experts not just as a form of networking, but also as a crucial element in their research processes… Creating communities of practice around topics of interest researchers already have, or connecting academics in different fields and allowing them to discover shared interests that might lead to interesting research opportunities are two areas of possible focus." [4, p. 37]

The richness of metadata is not always thought as bridges for networking. Metadata is contextual. The scope of a context extends or confines an exploration. Even a well-staffed library cannot claim that it has all the tacit knowledge on staff but the community, to which the collections belong or about which the collections are, can. Without the community's input, our descriptions are limited. By opening metadata creation for special collections to external communities, the library may help researchers make new connections with communities and open new discovery paths. Cornell has been an active participant in linked data projects for this end goal. For metadata created by staff, we have experimented with a community engagement approach in a limited fashion[4] and has been looking to other libraries, such as the Washington State University for inspiration[5]. To open a traditionally closed process is no small feat as it puts us in a vulnerable position. Our standing as information management authority may be challenged. But the gain for scholars and us would be enormous. We would have the opportunity to enable the marginalized communities to have an equal voice. We get to learn to work with internal and external parties, paid employee and volunteers, local or remote workforces. We also get to have exposures to diversity at a different level than we knew before. Why do this matter to digital scholarship support? Because metadata is a vital link between evidence and scholarship creation.

IV. Tailoring without Algorithm

User observations: Students in our immersion and fellowship programs told us that their needs for skills/tools are idiosyncratic, driven by his/her project and field. More tailored training, networking opportunities for expertise sharing and for emotional support are valued. All user groups in our surveys, including faculty, expressed a need for building their library research skills.

One might argue that the immersion programs are not digital scholarship support, while our intensive Summer Graduate Fellowship in Digital Humanities[6] is, because the former trains the students to be better researchers. Our librarians observed that the applications to join our programs are becoming more sophisticated due to exposure to new tools and new forms of scholarship in recent years. But students' skills have not seen significant improvement. The immersion programs help

[4] https://dcaps.library.cornell.edu/projects/seneca-haudenosaunee-archaeological-materials-circa-1688-1754 (Accessed on August 3, 2019)

[5] https://plateauportal.libraries.wsu.edu (Accessed on August 3rd, 2019)

[6] http://blogs.cornell.edu/sgfdh/

build skills in the research lifecycle, in which users do not distinguish where digital scholarship begins and ends.

Large academic research libraries like mine have always strived to have both broad and deep reach. Tailoring a service is challenging for all businesses, not just the library. Industries collect individual's data and use algorithms to achieve customization. Libraries like mine, however, hold users' privacy sacrosanct. We make a conscious choice to forgo the benefits of mining users' personal data. We try to find different ways to provide relevant services.

The trajectory of the immersion programs at Cornell is one of customization. The programs began in 2012, initially for the humanities, but soon expanded to include social sciences graduate students. In 2017, a team of librarians launched an additional program targeting students in natural and physical sciences. In 2019, another one for the sciences and engineering students was piloted. The immersion programs range from two to four days while the Summer Graduate Fellowship in Digital Humanities is a six-week, intensive training which often leads to continued collaboration long after the program ends. Through a study of humanists conducted by a group of Cornell library staff, we learned that communities of practice enable transformative learning [6]. Our immersion programs librarians make a deliberate effort not to use the programs to demonstrate personal mastery of knowledge. They aim to create an extendable co-learning experience for all participants. Aside from disciplinary-specific methods and tools, some offerings have been highly valued by the users: The overall focus on project-agnostic, transferrable skills that cross the disciplinary boundaries; the introduction to open source tools and platforms to sustain users' research after they graduate from Cornell. Most recently, the journal editor's panel, formed by our own faculty, arranged by our science and engineering immersion program team, was very popular. Posner's words best describe our intention which is,"…investing in people's long-term potential as scholars" [7].

Tancheva et al suggested, "…the library of the future will need to focus on individualized services, apps, and integrations as opposed to a one-size-fits-all approach" [4, p. 41]. We have segmented our immersion programs by discipline but have not yet been able to tailor to the individual level at scale. The participants have given all of our programs high marks. But the librarians continue to struggle between making the offerings relevant to all participants, scaling the program up, and maintaining the quality within our means [8]. We are often confronted with the questions: Is scalability a measure of success? Is boutique service the future of research libraries? Is our organization nimble enough to transform at the speed of our users' needs? This brings me to the organization and staff.

V. Moving from Hierarchical to Relational Practice

> User Observations: "Serious researchers have "good enough" systems that work for them in information seeking and knowledge production. When they fail them, they seek the expert. Consequently, library services will be expert, smart services: personalized, flexible, and portable" [4, p. 40].

How do we deliver an expert, smart service? At Cornell, we have more than a dozen unit libraries that are in the facilities of a department or college. Most of our departments are organized hierarchically, clustered by the physical location of staff. The subject liaison model that we have cultivated carefully in the recent years also encourages vertical, subject specialization. Cross-functional collaboration depends on the liaison's own confidence, motivation, and bandwidth. This is not to say such a structure does not have merits. It does, for relationship building with departments; but it is not conducive for cross-functional, cross-unit service portability.

My library has made a lot of progress building capacities, although often opportunistically, when an existing position becomes vacant. We support digitization projects initiated by faculty research or courses. A virtual library team, with expertise in digitization, metadata creation, online delivery, repository, and digital preservation, supports these projects. On the service front, we have staff with expertise in digital humanities, social science data, geospatial data, audiovisual production, copyright, and digital publishing. In addition, we have a campus partnership, the Research Data Management Service[7], that provides research, computational, and information management support. The staff are from several units on campus, including the library. Over the years of experience-building, we have been inspired by many peer libraries. Models proposed by Jennifer Vinopal[8] and Bethany James Nowviskie[9], for example, are used by our librarians to assess gaps. Others[10] influence how we shape our overall support structure.

An expert-service library needs to be as open as possible to lower the barriers for user entry. The library has limited control over where users seek our support. But the library has a lot of control over how to provide a consistent user experience. This requires a dynamic formation of subject and functional expertise, including utilizing partner's expertise on campus. The practitioners in my library say that changing the reporting structure may not be possible or desirable, because what to include or exclude in an organizational unit depends on how we want to define the service menu, i.e., being a digital humanities support would call for a different team than being a scholarly communication support, although both are part and parcel of digital scholarship from users' perspective.

A more achievable solution might be to create a sharing culture in which the highest level of library leadership views their staff as collectively shared talents. They cross organizational boundaries to invite or deploy staff to do work. They function interchangeably to minimize multiple approvals and speed up decisions. This needs to be supported with frequent communication and transparent decision-making. The job descriptions, including for the highest-ranking leaders, codify responsibilities and accountability, but also reward behaviors that support fluid movement of expertise. Instead of a select group being the designated liaisons, every library staff should function as a node on the library's service network.

Many of us have existing internal mailing lists, to which staff send users' requests. We have some expert staff behind these lists but we have a lot more whose jobs end with sending a user request. Such practice is no longer good enough because we create a divide between "knows and don't knows" among in our organization. The staff who perform the referral function are as critical as the subjects and functional experts. They not only need to be aware of where to send users, they need to have a general understanding of the services the subject and functional librarians provide. To deliver expert and smart services, the library needs to elevate the overall level of current

[7] https://data.research.cornell.edu/
[8] http://vinopal.org/2014/01/09/services-for-all-areas-of-knowledge/
[9] http://nowviskie.org/2011/a-skunk-in-the-library/
[10] See excellent work at University of Virginial Library's Scholar's Lab (https://scholarslab.lib.virginia.edu/about/#what-is-the-scholars-lab) and Duke Libraries Digital Scholarship Services (https://library.duke.edu/digital)

awareness and digital scholarship knowledge of its general staff to minimize breakdown points due to staff's lack of preparedness.

This relational operation behind the scene needs to be paired with designated physical spaces to accommodate different learning styles. Many libraries already have digital scholarship centers or hubs. The findings of Malpas, Schonfeld, et al described that physical spaces are so important for engaging campus community that some library leadership views them not only as facilities but as a service [9, p. 42]. For staff, such space encourages teamwork and service portability especially when staff offices are spread out across campus, when they report to different supervisors.

VI. Deliberate Branding

> "[C]ustomers have an image of the service concept regardless of whether it has been defined by word-of-mouth or other sources of information or from real service experiences….Before, during, and after service delivery, service organizations set customer expectations. These expectations relate to the nature of the service package, as well as to the nature, duration, and customer flexibility" [2, p. 122].

It may not be surprising to hear that there is no agreement among librarians if we asked whether digital scholarship is a distinctive branch of scholarship. This state of affairs manifests itself in the ways we present our services to users. On our website, we list "digital scholarship services" but promote digital humanities, not the other research support equally. Through our own mouths, we say, "The scientists find the digital humanities tools very helpful," projecting an exclusion between the two. If we asked our staff whether, for example, systematic review or citation management training is supporting digital scholarship, we may very likely hear conflicting answers. Although all of these may be semantically intended, the effect is that we are fragmenting our own service concept and reducing our own effectiveness in articulating an identity that we desire—we are professionals who support scholarship creation, however users define it, with deep expertise in subject, technology, and information management, dissemination, and preservation. Compared to the other challenges noted earlier, the branding might be the easiest one to address. We can revise our online presence, call our services with names that resonate with users, and train staff to use consistent vocabulary to promote what we do. This is an area Cornell plans to improve as well.

VII. Conclusion

This paper aims to hit the pause-button for a moment so we can take stock of areas needing attention. As Anne et al pointed out, "It is possible that in a decade or less, the term *digital humanities* will be obsolete and that the methodologies, tools, and technologies discussed… will be fully integrated into humanities disciplines and the culture of academic institutions. " [3, p. 40] We need to plan for this real possibility with a robust service design. A successful support to digital scholarship needs to be simple to use. Users need to be able to find it when and where they need it. The users should be able to pick from a service menu without having to adjust to how a library operates. An academic research library designs such an experience with a strong service concept in mind, intentionally organizes our collections, people, and workflows, codifies responsibilities and accountabilities, and places them within a collaborative culture, all for delivering relevant services in a timely manner without breakdowns. "Scholars' needs for digital scholarship support are inherently diverse; in attempting to meet them without considering scale and sustainability, we risk developing narrowly focused or short-lived solutions that are difficult to maintain over time and with infrastructure that cannot be repurposed to benefit other projects." [10] It takes strong discipline to focus on basic building blocks that are not glamorous. A library does not have a future in digital scholarship unless its plumbing can sustain the flow of needs of its academy.

Acknowledgement

The author thanks the librarians at the Cornell University Library for sharing their invaluable insights: Tre Berney, Eliza Bettinger, Jasmine E. Burns, Jeremy P. Cusker, Dianne Dietrich, Kate Ghezzi-Kopel, Erica M. Johns, Jesse Koennecke, Wendy A. Kozlowski, Peter McCracken, Michelle Paolillo, Jill Powell, and Zoë Wilkinson Saldana.

References

[1] A. S. Rumsey, "Scholarly Communication Institute 9: New-model scholarly communication-road map for change," University of Virginia Library, 2011.

[2] S. M. Goldstein, R. Johnston, J. Duffy and J. Rao, "The service concept: the missing link in service design research?," *Journal of Operations Management,* vol. 20, no. 2, pp. 121-134, 2002.

[3] K. Anne and e. al, "Building capacity for digital humanities: a framework for institutional planning. ECAR working group paper," ECAR, Louisville, 2017.

[4] K. Tancheva and e. al, "A day in the life of a (serious) researcher: envisioning the future of the research library," Ithaka S+R, New York, 2016, March 8.

[5] T. Padilla, L. Allen, H. Frost, S. Potvin, E. Roke and S. Varner, "Always already computational: collections as data," 2019. [Online]. Available: https://zenodo.org/record/3152935#.XVVtp-hJFPZ. [Accessed 13 8 2018].

[6] G. Castro-Gessner, S. Newberry, D. Schmidle and K. Tancheva, "Humanists in the house of learning: academic research libraries' role in fostering communities of practice," in *ACRL 2013 Proceedings*, Washington, D.C., 2013.

[7] M. Posner, "Commit to DH people, not DH projects," 18 March 2014. [Online]. Available: http://miriamposner.com/blog/commit-to-dh-people-not-dh-projects/. [Accessed 13 8 2019].

[8] E. R. Eldermire, E. M. Johns, S. Newberry and V. A. Cole, "Repackaging library workshops into disciplinary bootcamps: Creating graduate student success," *College & Research Libraries News,* vol. 80, no. 7, p. 394, 2019.

[9] M. Constance, R. Schonfeld, R. Stein, L. Dempsey and D. Marcum, "University futures, library futures: aligning library strategies with institutional directions," OCLC Research, Dublin, OH, 2018.

[10] J. Vinopal and M. McCormick, "Supporting digital scholarship in research libraries: scalability and sustainability," *Journal of Library Administration,* vol. 53, no. 1, pp. 27-42, 2013.

Identifying Keywords in the Buddhist Canon

Alex Amies
independent scholar

Yashuo Deng
PhD student, University of the West

***Abstract*—This paper discusses approaches to keyword identification in Buddhist texts. The paper first defines what is meant by a keyword and how it differs from terminology and other related concepts. Systematic approaches to keyword extraction with natural language processing methods are described as well as visualization techniques to help researchers identify keywords. Examples from the *Heart Sūtra* are explored followed by extension to the larger body of the *Taishō Tripitaka*. The use of term frequency, co-occurrence networks, and collocations is also discussed. The results show that term frequency filtered by terms labelled as Buddhist gives a reasonable set of keywords with subjective assessment based on linguistic and stylistic arguments. Ranking of collocations can indicate predominant word senses. Finally, potential future areas of application of the tools and methods are discussed.**

***Index Terms*—Buddhism, natural language processing**

I. Introduction

The body of Buddhist literature is immense with many commentaries extending to the present day. This body of literature is densely packed with terminology. Labelling texts with keywords can help readers navigate this huge body of literature. A **keyword** may be regarded as a word that is representative of a specific text. [21] The term keyword is also sometimes defined as a word that indicates a "cultural preoccupation."[7] A third sense of the term keyword is a word used in a library index used to find relevant books or documents.

In contrast to keywords, **terminology** is a set of terms specific to a specialist domain. [4] For example, Buddhist terminology or Confucian terminology. Categories, topics, and subjects refer to domains or branches of domains, such as Buddhism or Mahāyāna Buddhism.

One reason to pay attention to keywords is that translation of keywords may have a disproportionate impact on the translation of a document. In their text on translation, Vinay and Darbelnet write, "Every text is dominated by a number of key words which are usefully identified at the outset."[20] How can keywords be extracted from the source text itself? The way that a word is used can indicate its status as a keyword. For example, first person narrative is more personal than third person and may be a cue for keywords, say as part of a personal judgement. [19] The main characteristic examined in this paper will be repetition.

In the context of library science and information retrieval, Baeza-Yates and Ribeiro-Neto define keywords or index terms as "a preselected group of words that represents a key concept or topic in a document." [3] Keywords in library indexes and other information systems are often chosen from a controlled vocabulary. A **controlled vocabulary** is a limited set of terms representing clearly defined concepts. The Library of Congress Subject Headings (LCSH) is an example of a controlled vocabulary. [10] Use of a controlled vocabulary may restrict keywords to those more expected by users. However, one difficulty with controlled vocabularies is that it is very time consuming to establish an extensive controlled vocabulary and expand it as it is used to label a large number of documents.

Traditional library systems relied on keywords listed in subject indexes. However, modern search methods, such as the full text search methods described

by Zhai and Massung, [23, pp. 92-100] do not require keyword extraction. All the terms in the documents to be searched are stored in an index and any terms desired can be used in queries. This prompts the question, are keywords still needed? Mai describes the user experience of a student searching for information in a library. [11] Users are challenged with an overwhelming amount of information and do a substantial amount of exploration and iterative query refinement to discover materials. So, while keywords are not required, users often choose to search using keyword queries. Keywords may also be useful in organizing information and presenting search results to users.

The best way to find keywords may be for an expert human curator to read a text and decide on the most relevant terms. However, the size of the body of Buddhist literature makes this infeasible, so a central focus of this paper is evaluation of methods for automatic keyword extraction. An important aspect of the evaluation is what a machine might miss in comparison to a knowledgeable human curator. The second aim of this paper is to make the identification of keywords simpler for humans though software tools.

II. Literature Review

In their text on information retrieval and text mining, Zhai and Massung describe term extraction using vector space models, including term frequency and TF-IDF, in the context of text search. [23, pp. 92-100] They do not specifically include keyword extraction but similar techniques may be able to be used for keyword extraction. **Term frequency-inverse document frequency** (TF-IDF) is a formula that computes a weight for terms in a source document using term frequencies and comparing with the presence of the terms in other documents in the corpus. This gives a greater weight to terms that not only occur more frequently but also that are used more specifically in the text being studied. The method gives less weight to terms commonly used throughout the corpus. [12, pp. 108-110] The **document frequency** for a term is defined as the number of documents in the corpus containing that term. This will be described in more detail below.

Xie and Matusiak describe user tagging as an emerging and popular method for keyword assignment. [22] In this method the readers of the material assign keywords, also called tags, and those are shared with all the users of the system. This method has been most popular on social media platforms but also considered by digital library system designers.

Social network analysis (SNA) has been used in digital humanities to analyze social structures based on networks analysis and graph theory. [24] SNA models and visualizes the networked structures which consists of the nodes and the edges between them. In addition to the application of social studies, network analysis can also be used in linguistic studies.

Rose, et al. describe keyword extraction from documents using a keyword scoring formula based on term frequency and word degree. [18] **Word degree** is defined as the number of edges incident to a node in a network constructed from word co-occurrences in the text. The authors' technique also filters out stop words. The results are benchmarked against human assigned keywords from a corpus of journal abstracts evaluating precision and recall. The method does not require the use of a controlled vocabulary of candidate keywords but their evaluation of accuracy did require the use of a large set of documents with keywords assigned by a human curator.

Mihalcea and Radev also describe methods for keyword extraction using graph-based methods [13, pp. 163-165] and explain the underlying theory in the book *Graph Based Natural Language Processing and Information Retrieval*. A **graph** is a set of nodes connected by edges and is the term generally used in the mathematics and computer science literature for network. Mihalcea and Radev make a distinction between graphs as artificially generated and **networks**, as naturally occuring, such as social networks and network constructions for natural language. [13, p. 174].

The authors demonstrate some elegant visualizations with word networks. [13, pp. 89-97] In addition to their utility in visualization, graph-based methods are also appealing because they can be used to analyze natural language structures with linguistic models [13, pp. 131-157]. This contrasts to the purely numerical machine learning methods mentioned below that are divorced from linguistic models. The authors also discuss recent theories on random networks and how this relates to natural language [13, pp. 61-97]

Mihalcea and Radev describe keyword extraction with a co-occurrence network. [13, pp. 87-88] A **co-occurrence network** may be defined as a network of edges of pairs of words where the two words appear next to each other or in close proximity in a text. This methodogy will be described in more detail below and followed to construct a co-occurrence network and extract keywords from the *Heart Sūtra*.

Kenderdine, et. al. describe visualization of the analysis of the *Tripitaka Koreana*. There are a number of very interesting points about the approach described. [8] Firstly, the methods do not presuppose the existence of a controlled vocabulary. Rather the methods enable users to explore texts visually. In addition, the analysis was performed on the entire *Tripitaka Koreana*, demonstrating the applicability of the methodology at scale. Several visualizations are described in the paper: social networks in the *Gaoseng Zhuan*, exploration of search results with a 'blue dots' visualization to assist in identification of clusters of terms, and visualization of structures in texts, such as chiasmic structures and ring compositions. The authors do not describe keyword identification.

Lee, et. al. describe a machine-learning technique for keyword extraction for Chinese texts, with a literary Chinese corpus including Chinese Buddhist Texts. [9] The technique described uses a support vector machine (SVM) algorithm. [12, pp. 293-299] SVM is a kind of supervised training algorithm, a class of algorithms where a set of labels that is defined in advance is used to train a model. The method supposes the existence of a controlled vocabulary of labels and a large training data set. The authors trained the SVM model and reported good correlation with predicted values.

The current state of the art in solving many natural language problems is concentrated around artificial neural networks. Frome et. al. describe a visual-semantic word embedding model based on unsupervised learning. [5] The output of the analysis is a collection of word embeddings. A **word embedding** is a numerical representation of a word that is optimal in its ability to predict the previous and following words in a unlabelled corpus. The concept of a word embedding is a purely numerical construct. The authors argue that word embeddings do have semantic meaning, although they do not describe a linguistic basis. However, word embeddings have a visual representation in a vector space, which helps to explore connections between words. The distances between the points representing the words appears to generally correlate with closeness of meaning. The authors trained their model on a 5.7 million document, 5.4 billion word corpus extracted from Wikipedia, which certainly demonstrates ability to scale.

III. Methodology

The methodology in this paper is exploratory since the authors possess neither a controlled vocabulary of candidate keywords nor a set of documents with keywords assigned by a human curator that could be used for training and quantitative evaluation of accuracy.

The methodology used is based on term frequency and network analysis. **Term frequency**, loosely referred to as word frequency, is the number of occurrences of a term in a document or group of documents. Multiword expressions may be considered terms if they are a logical grouping. In order to compute the term frequency, a text is first segmented into individual terms in a process called word segementation, and then the occurrences summed to find the term frequencies.

A list of terms is a perhaps too simple of a structure to reduce a text to. The relationships between

terms is also important. The next simplest structure beyond a single term is a **bigram**, which is a combination of two terms that occur next to each other. Bigrams will be used to construct a network representations.

A. Quantitative text analysis and visualization

Word clouds may be used as a visual aid to identify the keywords in texts. Each term is shown by specific font size or color depending on its frequency of occurrence. This method provides an overview at a glance, reducing a text down to the terms ordered by their frequencies.

The pairs of high frequency terms and their collocated terms can be extracted from the texts and put into a matrix to designate the relationship and distribution of the significant pairs of words in the text.

Based on the matrix, the third step is network analysis. The words in a text can be considered as the nodes in a network, which manifests the structure of the text by showing the relationships (edges) between the words (nodes). A **weighted graph** is one where the edges may have different weights. [13, p. 34] The degree of node is the number of edges that connect to it. [13, p. 20] In a **directed graph** the edges have a specific direction between nodes. In an **undirected graph** the direction of the edges does not matter. The **indegree** of a directed graph is the number of incoming edges and the **outdegree** is the number of outgoing edges. The weighted degree is the sum of the weights of the incident edges. **Centrality** reflects how important a node is in a graph. [13, p. 74] There are several measures of centrality. The simplest is degree centrality. The **degree centrality** of a node in an undirected graph is equal to the number of neighbors of the node.

The method described in this paper uses a directed co-occurrence network constructed from adjacent terms (bigrams). The weights will be the term frequencies. The network is a directed graph because language has direction: the order of the words does matter. Once the word co-occurrence network is constructed, methods based on graph theory can be used to quantify the structure, including weight, centrality, indegree, and outdegree.

Figure 1. Word Cloud based on the text of the Heart Sūtra

The Nan Tien Institute Reader (NTI Reader) is an open source text reader for the Taishō Tripitaka with embedded Chinese-English dictionary and corpus analysis. [1] The NTI Reader was used as the source for text data, for word segmentation of the text, and computation of term frequencies. Gephi was used for network analysis and visualization. [2]

IV. Results

The *Heart Sūtra* has been analyzed according to the procedure described above. It was chosen to illustrate the methods with a small text before extending to the larger body of the *Taishō Tripitaka*.

A. Text visualization of the Heart Sūtra

The word cloud generated [1] for the *Heart Sūtra* is shown in Figure 1.

The raw term frequencies for the Heart Sūtra text are listed in Table I.

Table I shows that the term with the highest frequency in the *Heart Sūtra* text is 無 wú 'no.' The second most frequent term is 不 bù 'not,' and the third is 是 shì 'is.' Bigram and their frequencies were also computed. The bigrams starting with 無 wú 'no' are given in Table II.

The networks structure generated by Gephi[2] for

[1] Generated using https://www.jasondavies.com/wordcloud/

[2] https://gephi.org/

Table I
Unfiltered Term Frequencies in the Heart Sūtra text

Chinese	Equivalent	Frequency
無 wú	no	18
不 bù	not	9
是 shì	is	9
空 kōng	emptiness	7
故 gù	because	7
色 sè	form	6
咒 zhòu	mantra	6
般若波羅蜜多 bōrěbōluómìduō	prajñāpāramitā	5
揭帝 jiēdì	gate	4
行 xíng	volition	3
即 jí	exactly	3
想 xiǎng	perception	3
乃 nǎi	thus	2
受 shòu	sensation	2
大 dà	great	2
異 yì	different	2
眼 yǎn	eye	2
老 lǎo	old age	2
至 zhì	until	2
說 shuō	speak	2
一切 yīqiē	all	2

Table II
Bigrams starting with 無 wú in the Heart Sūtra

Term 1	Term 2	Frequency
無 wú 'no'	色 sè 'form'	2
無 wú 'no'	受 shòu 'sensation'	1
無 wú 'no'	眼 yǎn 'eye'	2
無 wú 'no'	意識 yìshí 'conscious'	1
無 wú 'no'	無明 wúmíng 'avidyā'	2
無 wú 'no'	老 lǎo 'old age'	2
無 wú 'no'	苦 kǔ 'suffering'	1
無 wú 'no'	智 zhì 'wisdom'	1
無 wú 'no'	得 dé 'obtain'	1
無 wú 'no'	所 suǒ 'that which'	1
無 wú 'no'	罣礙 guà ài 'hindrance'	2
無 wú 'no'	有 yǒu 'have'	1

the resulting co-occurrence network for the data in Table II is shown in Figure 2.

Compared to the word cloud, the word network illustrates the structure of the text in more detail by depicting the relationship between the high frequency terms and their related terms. Edges with higher word frequencies are shown with thicker lines.

The full set of bigrams and their frequencies for

Figure 2. Co-occurrence network for bigrams starting with 無 wú

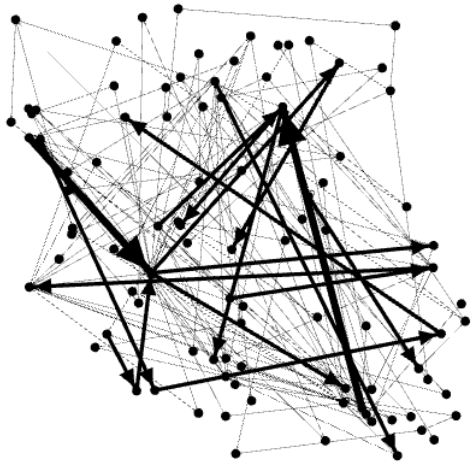


Figure 3. Co-occurrence network for the Heart Sūtra

whole text of the *Heart Sūtra* in a format similar to Table II was also loaded into Gephi. The co-occurrence network is shown in Figure 3.

The *Heart Sūtra* is only a short text and yet visualization of the co-occurrence network is very complex. An interactive tool like Gephi is more usable for exploring graphs than a static image like Figure 3. Eyeballing the figure, three concentrations of edges can roughly be made out. Interactive use of Gephi confirms that these are 無 wú , 不 bù , and 是 shì. This method of identification can be made more systematic by computing metrics for network centrality. The top ranking nodes by weighted degree were computed using Gephi and are listed in Table III. Weighted degree is a measure of centrality, which implies that this list should

be a candidate set of keywords. Note, however, that the terms 無 wú 'no,' 不 bù 'not,' 是 shì 'is,' 故 gù 'because,' and 亦 yì 'also,' should not be included as keywords because, as function words, they lack concrete meaning. This leaves 空 kōng 'emptiness,' 色 sè 'form,' 咒 zhòu 'mantra,' and 般若波羅蜜 bōrěbōluómìduō 'prajñāpāramitā' as the suggested keywords. The term 揭帝 jiēdì 'gate' is used for sound in the mantra, rather than meaning, so is not appropriate as a keyword.

Table III
Top terms in the Heart Sūtra ordered by weighted degree

Chinese	Equivalent	Weighted Degree
無 wú	no	31
不 bù	not	18
是 shì	is	16
空 kōng	emptiness	14
故 gù	because	14
色 sè	form	12
咒 zhòu	mantra	12
般若波羅蜜 bōrěbōluómìduō	prajñāpāramitā	10
亦 yì	also	8
揭帝 jiēdì	gate	8

B. Term frequency and keyword analysis of the Taishō Tripitaka

The NTI Reader automates computation of the term frequencies in the *Taishō Tripitaka*. This is done at three levels: (1) the entire canon, [14] (2) individual titles that are collections of scrolls (fascicles), such as the *Lotus Sūtra*, which consists of seven scrolls, [15] and (3) individual scrolls. [16] The top terms based on term frequency from this analysis for the entire scroll of *Heart Sūtra* (T 251) are shown in Table IV.

Table IV
Top terms in the Heart Sūtra ordered by term frequency

Chinese	Equivalent	Frequency
之 zhī	possessive particle	40
空 kōng	emptiness	28
者 zhě	nominalizing particle	21
相 xiāng	lakṣaṇa	17
不 bù	no	13

Note: Entire Scroll

Similarly to the discussion above, many of the top terms have a high frequency because they are common in general, especially function words like the possessive particle 之 zhī. The NTI Reader labels terms by domain area. The domain label 'Buddhism' was used to narrow the list to retain only Buddhist terms. The top Buddhist terms by term frequency are shown in Table V.

Table V
Top Buddhist Terms by Frequency of the Heart Sūtra

Chinese	Equivalent	Frequency
空 kōng	emptiness	28
相 xiāng	lakṣaṇa	17
佛 fó	Buddha	15
般若波羅蜜多 bōrěbōluómìduō	prajñāpāramitā	8
色 sè	form	8

Note: Entire Scroll

The results in Table V for the whole scroll of T 251 strongly overlap with the suggested keywords for the text of the *Heart Sūtra* only.

The NTI Reader does not provide co-occurrence networks. However, high frequency bigrams are provided as collocations for the entire corpus under the individual word entries. For example, the top collocations for 空 kōng are 空安靜 kōng ānjìng 'empty and tranquil,' 空靜 kōng jìng 'empty and tranquil,' and 飛空 fēi kōng 'flying in the sky.' [17] The collocations indicate the prodominant word sense used in the canon. The top collocations for 空 kōng are different senses from the sense used in the *Heart Sūtra*, which intends the Buddhist philosophy sense of śūnyatā.

V. Discussion

Discussion of results, challenges with the methods, and possible alternative methods are included in this section.

If only Buddhist terminology is considered then the term frequency and the weighted degree of the nodes in the network give similar results for the *Heart Sūtra*. Taking only the Buddhist terms from Table I, the predicted keywords are: (1) 空 kōng 'emptiness,' (2) 色

sè 'form ,' (3) 咒 zhòu 'mantra,' and (4) 般若波羅蜜 bōrěbōluómìduō 'prajñāpāramitā.' When the preface of T 251 is considered as well the list strongly overlaps.

A. Challenges with methodology

It may have made sense to filter the function words before performing the analysis since they did obscure the keywords. However, it was instructive to include the function words. In addition, the *Heart Sūtra* has a repetitive negating style of verse that depends heavily on function words.

One challenge is word sense disambiguation. That is, some words have multiple senses and sometimes this includes both Buddhist and secular senses, so it is hard for an algorithm to reliably determine if the intended word sense is Buddhist or not. The NTI Reader regards a word as Buddhist in the term frequency analysis if any of the senses are Buddhist. As just mentioned, the word 空 kōng may mean 'śūnyatā' in a Buddhist philosophical sense or 'empty' the ordinary sense. In the *Heart Sūtra* the Buddhist philosophical sense is intended.

An example term where this is problematic is the term 無 wú, which may mean the adverb 'no' or read as mó with a purely phonetic value in a Buddhist chant. In the *Heart Sūtra* the ordinary 'no' is intended. It has been eliminated from the Table V based on this interpretation but the term frequency algorithm did not detect this. Similarly, 有 yǒu has been removed on this basis as well.

A further challenge is word segmentation. Some terms occur both individually and also as part of multiword expressions. This can affect the values of the term frequencies. Word segmentation in the NTI Reader is based on an embedded Chinese-English dictionary, which includes many multiword expressions. For example, the term 色 sè 'form' is included within 無色 wúsè 'formless.' There are eight occurrences of 色 sè in T 251 including two occurrences in 無色 wúsè. Should we say that there are six or eight occurrences of 色 sè? With automated analysis, the terms embedded within multi-term expressions were not counted independently, so the answer will be six. While it is a subjective judgement, it may be reasonable to include all occurrences for the purpose of keyword analysis because the stylistic effect of repetition is important. The difficulty of terms included within other terms was overcome with manual checking of the term frequency data and adjusting where judged appropriate. Thus eight occurences of 色 sè are included in Table V.

Thirdly, the NTI Reader analysis includes the contents of the entire scroll of T 251, including the preface, which is longer than the text of the *Heart Sūtra* itself. In contrast, a human creating a word cloud may make a different choice, perhaps selecting only the section of text that is most interesting, typically the text of the sūtra only. It may be argued that this does not invalidate the results because the preface is discussing the sūtra text, so the term frequencies are still representative.

Finally, the NTI Reader has not implemented network analysis.

VI. Other Methods Considered

This section lists some other methods considered and attempted.

1) Controlled vocabulary: Using a controlled vocabulary has a number of advantages and also a number of difficulties. The advantages are that the words in the controlled vocabulary can be selected to be meaningful and broadly understood. Archaic variants are avoided in controlled vocabularies. The disadvantages are that the authors do not have a controlled vocabulary for Buddhism in Chinese and that an additional synonym analysis step is required. That is, words discovered in a text need to be compared against words in the controlled vocabulary to decide whether there is a synonym match. At present, the authors do not have such a list of synonyms to automate the task. Creating a controlled vocabulary and associated synonym list is a large task beyond the scope of the present study.

2) Term frequency-inverse document frequency: TF-IDF was discussed above in the literature review. The TF-IDF value for a term in a document is calculated as

$$tf - idf = tf \cdot idf$$

where *tf* is term frequency or word count in the document and *idf* is the inverse document frequency, which is computed as

$$idf = log10(N/df)$$

N is the total number of documents and *df* is document frequency.

The top terms from analysis based ranking by TF-IDF value and filtered for only Buddhist terms are shown in Table VI.

Table VI
Top Terms from TF-IDF Analysis of Heart Sūtra

Chinese	Equivalent	Term Frequency
界 jiè	realm	2
相 xiàng	lakṣaṇa	17
纏 chán	entangle	4
無罣礙 wú guà'ài	unimpeded	2
入滅 rù miè	enter Nirvāṇa	1

Note: Entire Scroll

The terms suggested by the TF-IDF analysis, subjectively, do not seem to be as relevant as the term frequency and the results of the network analysis.

3) Document classification: Document classification is a family of techniques that infers labels representing categories based on analysis of the text content. Possible methods of analysis include artificial neural networks [6, pp. 152-153] as well as traditional statistical approaches based on word and bigram frequency. Typically, however, the categories are relatively coarse, such as 'sports' or 'entertainment.' The methods used may be extended for keyword analysis but this depends a labelling system and would need a controlled vocabulary to enable more fine grained categorization. In addition, there is an important distinction between keyword and category. For example, the term *prajñāpāramitā* may be considered both a keyword and a category. However, the term 'emptiness' may be considered a keyword but not a category.

VII. Summary

The results showed that term frequencies filtered by terms labelled as Buddhist give a reasonable list of keywords. The simple nature of the method allowed exploration of data and reasoning within established linguistic concepts. Approaches that go beyond this simple approach are also available. Today deep neural networks are the focus of much state of the art research in natural language processing. [6, pp. 38-39] In the last few years deep neural networks have been used successfully for natural language processing tasks like translation and named entity recognition. However, artificial neural network models cannot easily refer to established linguistic concepts.

Network analysis is an additional area of recent research in natural language processing that does address keyword extraction and can be related to traditional linguistic theories. [13, p. 64] Possible future areas of application include identifying structure in texts, matching parallel texts, and user assistance in navigation and refinement of search results. The discussion in this paper indicates that the application of network analysis may be more promising in Buddhist textual studies where the goals are exploring and explaining whereas deep neural networks may be more promising in the areas of machine translation and text classification.

VIII. Abbreviations

LCSH Library of Congress Subject Headers
NTI Nan Tien Institute
SNA Social network analysis
SVM Support Vector Machine
T *Taishō shinshū daizōkyō* 大正新脩大藏經
TF-IDF Term frequency-inverse document frequency

References

[1] A. Amies, NTI Buddhist Text Reader Project, 2019, https://github.com/alexamies/buddhist-dictionary

[2] Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. InThird international AAAI conference on weblogs and social media 2009 Mar 19.

[3] R. Baeza-Yates, and B. Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology Behind Search, New York: Addison Wesley, 2011, p. 62.

[4] H. Bussmann, Routledge Dictionary of Language and Linguistics, Translated by Gregory Trauth and Kerstin Kazzazi, London: Routledge, 2006, p. 1186.

[5] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, "Devise: A deep visual-semantic embedding model. In Advances in neural information processing systems," 2013, pp. 2121-2129.

[6] T. Ganegedara, Natural Language Processing with TensorFlow: Teach Language to Machines Using Python's Deep Learning Library. Packt Publishing Ltd, e-book, 2018.

[7] C. Kay, "Issues for Historical and Regional Corpora: First Catch Your Word," In Archer, D (Ed.) What's in a Word-List?: Investigating Word Frequency and Keyword Extraction, Farnham: Ashgate Publishing, 2012, p. 77.

[8] S. Kenderdine, L. Lancaster, H. Lan, and T. Gremmler, "Omnidirectional 3D Visualization for the Analysis of Large-scale Textual Corpora: Tripitaka Koreana," In Second International Conference on Culture and Computing, 2011 Oct 20, pp. 27-32, IEEE, 2011.

[9] C.M. Lee, C.K. Huang, K.M. Tang, and K.H. Chen, "Iterative machine-learning Chinese term extraction. In International Conference on Asian Digital Libraries," Berlin, Heidelberg: Springer, 2012 November, pp. 309-312.

[10] Library of Congress, "Library of Congress Subject Headings," accessed 27 May 2019, http://id.loc.gov/authorities/subjects.html.

[11] J.E. Mai, Looking for information: A survey of research on information seeking, needs, and behavior. Emerald Group Publishing, 2016, pp. 22-25.

[12] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Kindle. Cambridge ; New York: Cambridge University Press, 2008.

[13] R. Mihalcea, and D. Radev, Graph-based natural language processing and information retrieval. Cambridge: Cambridge University Press, 2011, pp. 163-165.

[14] NTI Reader, "Terminology Extraction and Vocabulary Analysis," viewed 26 May 2019, http://ntireader.org/analysis/corpus_analysis.html.

[15] NTI Reader, "Glossary and Vocabulary for The Lotus Sutra (Miaofa Lianhua Jing) 《妙法蓮華經》," viewed 26 May 2019, http://ntireader.org/analysis/taisho/t0262_analysis.html.

[16] NTI Reader, "Glossary and Vocabulary for Prajñāpāramitā Heart Sūtra 《般若波羅蜜多心經》," viewed 26 May 2019, http://ntireader.org/analysis/taisho/t0251_01_analysis.html.

[17] NTI Reader, "空 kōng kòng," viewed 31 May 2019, http://ntireader.org/words/3876.html.

[18] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," Text mining: applications and theory, 2010, pp. 1-20.

[19] P. Verdonk, Stylistics, Oxford: Oxford University Press, 2002, p. 41.

[20] J. Vinay, and J. Darbelnet, Comparative Stylistics of French and English: A Methodology for Translation, Amsterdam and Philadelphia: John Benjamins Publishing, 1995.

[21] K. Wales, A Dictionary of Stylistics. Routledge, 2014, pp. 244-246.

[22] I. Xie, and K. Matusiak, Discover digital libraries: Theory and practice, Elsevier; 2016 Jul 26, e-book, pp. 155-157.

[23] C. Zhai, and S. Massung, Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. Morgan & Claypool, 2016, p. 299.

[24] K.A. Zweig, Network analysis literacy: a practical approach to the analysis of networks, Springer Science & Business Media; 2016, pp. 27-28.

Combination of TEI and Python in Studies of Chinese Epigraphy in Singapore

Duoduo Xu
Department of Chinese Studies
National University of Singapore
Singapore
chsxd@nus.edu.sg

Francis Bond
Linguistics and Multilingual Studies
Nanyang Technological University
Singapore
bond@ieee.org

Kenneth Dean
Department of Chinese Studies
National University of Singapore
Singapore
chshead@nus.edu.sg

Abstract*—The present study describes a convergent digital humanities approach to Chinese Epigraphy in Singapore. The analyzed inscriptions are spread in temples, associations, and other sites and preserve a rich biographical resource for reconstructing the premodern history of Singapore, especially the related network and social activities. We present work to digitize these resources using TEI (Text Encoding Initiative) mark up. Chinese Epigraphy documents are transformed into TEI texts. A Python script is designed to extract the names of sponsors and related information, which are tagged with specific labels. This convergent approach provides an automatic way in managing large amount of data. It also reveals the interactions between philology-oriented encoding and digitization.

Keywords—Digital Humanities; TEI; Python; Chinese Epigraphy in Singapore

I. INTRODUCTION

Inscriptions on stones and many other mediums were considered a long-lasting method to record significant events in history and to make official important documents. These Chinese epigraphies preserved from the old times enrich the picture of cultural links among countries in the Sinosphere. The earliest relevant material discovered so far is a tombstone of 1264 A.D. in Brunei (Franke 1973).

Historical epigraphy in Singapore provides a rich resource for reconstructing the 'biography' of early Chinese clans. As shown in Figure 1, a typical stele epigraph is composed of a title, a preface stating the event, a panel displaying the major organizers, a long list of sponsors (usually referred as "Fangminglu 芳名錄"), and a dateline together with the corresponding persons erecting the stele. There are also other forms of epigraphic texts, e.g.: plaques and couplets. In these kinds of epigraphies, the main texts are compliments dedicated to the worshipped gods, while a dateline and donors are also included. These literal records depict the social network and interactions among community leaders, civilians, and organizations.

These epigraphic texts inscribed in stelae, wooden plaques, and metal boards spread, over time, among numerous temples, associations, and other old facilities. So far, scholars have compiled several collections of Chinese Epigraphy in Singapore, e.g.: Chen & Chen (1970), Lin *et al*. (1975), and Chng (1984). Among these catalogues, Dean & Hue (2017) is a recent edition with new field work data and bilingual interpretations.

Besides the systematically organized epigraphic texts, a digital version of the book, which can be connected to related databases (e.g.: the Singapore Biography Database, SBDB for short), is in preparation.

Funding agency: "Singapore Historical GIS Analysis: The Transformations of Chinese Institutions" (MOE2015-T2-2-135); "Chinese Epigraphy of Singapore: 1912-2019" (CACRG-1801).

Fig. 1. Sample of Chinese Epigraphy
(No. 40.05: Record of the repairs to this temple (Stone 1908) (74 x 45.5 cm), Dean & Hue 2017: 918).

In the project, the TEI (Text Encoding Initiative) standard is applied to the book by Dean & Hue (2017). This two-volume (1,422 page) collection of epigraphy in Singapore consists of preface, body of text part (63 chapters), and appendix. Each chapter documents one Chinese Temple in Singapore and follows a consistent framework. Therefore, one template can be highlighted for the TEI project, into which various contents can be filled. Another xml file including the description of the project (contributors, vernacular terms, and content list) wraps all the 64 xml files of each chapter and the appendix.

The preliminary task of the project was to tidy up the texts of the book, recognized from the pdf file output from InDesign by OCR (Optical Character Recognition), following TEI standard. Based on this text (made up between October and December 2018), further detailed mark-up of specific information is ongoing. The main target of this digitalization work is to get the whole list of sponsors who contributed and donated to the temples or associations, in order to analyse their social network.

Differently from the traditional catalogues, TEI mark-up implements metadata to the texts by adding various labels to the corresponding information. The prepared text in xml format can be transformed into various formats accordingly. Moreover, it is possible to extract information tagged by the same label through batch programming to be analysed for different research purposes. Python is chosen in this case.

II. Data Structure

A. Framework of the Target Text

According to the current template, each chapter of the book is divided into four sections: front pages, including the name and an overview of the Temple, the main parts describing the details of all sorts of epigraphic documents in the temple, notes of each chapter, and back pages including miscellaneous pictures of the temple. These four sections are at the same level in TEI language, encoded as , , , and , respectively. The term "div" is short for "division". Sections "notes" and "back" are optional according to the content.

The "front" part is a sort of cover page of each chapter. It includes the title of the temple/association (in Chinese and Romanized transcription of its dialectal pronunciation) and a general picture of the site. The title is marked by the label <head>. Its transcriptions in multiple languages are considered segments of the <head>, and each language is indicated by the corresponding code of ISO 639-3 (International Organization for Standardization). The label <figure> is applied to the sections of image(s). Each image, marked by the label <graphic>, is navigated by its local path. The bracket <pb n=""/> beside indicates the page number ("pb" is short for "page break").

```
<div type="front"><pb n="0001"/>
<head>
   <seg xml:lang="zh">粤海清庙</seg>
   <seg xml:lang="nan-Latn-x-Teo">Wak Hai Cheng Bio</seg>
</head>
<figure>
   <graphic url="images/vol1-ch1_01.jpg"/>
</figure>
</div>
```

Fig. 2. Sample of Marked-up "Front" Text (excerpted from Chapter 1).

Four hierarchical levels can be highlighted in the body of the text of each chapter. Therefore, three other layers of divisions are set under , numbered as "L2", "L3", and "L4". The text of Level 1 ( ) is the headline of the main text, transcribing the name of the temple in Mandarin Chinese, Chinese dialect, and English.

The text of Level 2 ( ) refers to secondary sections of the chapter. The main secondary sections include introduction, main god of the site, list of cultural artifacts, stelae, plaques, couplets, other cultural artifacts, and translation of the main text of the inscriptions. The headlines of these sections are bilingual, numbered in the format: chapter-number. section-number. The texts in this section are English dominant, with some Chinese translations. Some images in this section are with titles. These titles are labelled as <head> following <graphic> inside the label <figure>.

```
<div type="L2">
   <head>1.1.
      <seg xml:lang="en">1.1 Brief introduction to the Temple</seg>
      <seg xml:lang="zh">廟宇簡介</seg>
   </head>
   <figure>
      <graphic url="images/vol1-ch1_02.jpg"/>
      <head>
         <seg xml:lang="zh">粤海清廟</seg>
         <seg xml:lang="nan-Latn-x-Teo">Wak Hai Cheng Bio</seg>
      </head>
   </figure>
   <p xml:lang="en" xml:id="vol1-0001">The Wak Hai Cheng Bio Temple, located at 30B Philip Street,
is the oldest Teochew (Chaozhou) temple in Singapore, with an official founding date of 1826. ...</p>
   <p xml:lang="en" xml:id="vol1-0002">(For more information, see <bibl>Lim and Pang 2005: Vol.1,
199-202</bibl>; ...)</p>
   <p xml:lang="zh" corresp="#vol1-0001">粤海清廟，座落於30B號菲力街，是新加坡最古老的潮州廟宇。...</p>
   <p xml:lang="zh" corresp="#vol1-0002">（欲知更多詳情，請見林、馮2005：上冊，頁199-202；...）</p>
</div>
```

Fig. 3. Sample of Marked-up "L2" Text (excerpted from Chapter 1.1).

Level 3 texts (div type="L3") are general information of each epigraphic text. The Level consists of three blocks of encodings: 1) the figure of the object, title of the object (bilingual, if available); 2) label assigned to it in the book, and 3) physical data of the object, including: material, date, measurement, and reference. The third block is marked as <ab type="object">. The term <ab> is shortened for "anonymous block". This block includes four labels: <material>, <date>, <measure> (with the unit counted by centimetre), and <bibl>. The example below is a stele quoted from the 83rd artifact in the Lotus Mountain Double Grove Monastery (Chapter 45 of the book).

```
<div type="L3">
   <figure>
      <graphic url="images/vol2-ch45_39.jpg"/>
      <head>
         <seg xml:lang="zh">重修蓮山雙林禪寺碑記，天運庚申年</seg>
         <seg xml:lang="en">Stele record of the restoration of the Lotus Mountain Double Grove Monastery</seg>
      </head>
   </figure>
   <label>45.83</label>
   <ab type="object">
      <material>Stone</material>
      <date>1920</date>
      <measure unit="cm">220 x 113</measure>
      <bibl>Chen 1970: 156-157</bibl>
   </ab>
```

Fig. 4. Sample of Marked-up "L3" Text (excerpted from Chapter 45.5).

"L4" (div type="L4") is set for the transcriptions of epigraphic texts. In the initial framework, all kinds of transcriptions are processed as anonymous block. With the progress of marking-up, it is noticed that the epigraphic texts often include different sections, e.g.: head, dateline, preface, name lists of organizers and donors, expense of the project, signature, etc. These tags are not valid in <ab>. Therefore, it is necessary to extend the framework to the fourth level. Some short epigraphic texts are also marked-up within "L3". They don't affect the result of sponsors, since there are no donors' names in these epigraphic texts. Even though, in order to be consistent, they will be transformed. The next section are examples of epigraphic texts in "L4".

B. Types of Epigraphic Texts

The main types of epigraphic texts are stelae, plaques, and couplets. Other mediums with epigraphic texts are categorized as "other cultural artifacts" in the book, such as censors, bronze bells, shrines, beams, etc.

Stele text generally includes a headline (the title of the stele), a preface (the background and reason to build this stele), a name list of organizers, a name list of donors (and their donations), the total amount of donation, sometimes the itinerary of expenses, and the date of erection. Each is defined with <subtype> as "transcription", since the texts are transcriptions of the stele. Each transcription is assigned an <xml:id> number, which is composed by volume number and a four-digits serial number. The title of the stele is marked as <head>, and the preface <p> ("paragraph"). The name lists are encoded as <list> in the current template. Each donor and donation is an <item>. Inside <item>, individuals are tagged as <persName>, while organizations (association, store, boat, etc.) are tagged as <orgName>. The amount of donation is tagged as <measure>. Inside <measure> annotates the amount transcribed in Arabic numbers as <quantity> and the <unit> depending on the currency. The inventory of expenses follows the same format of <list>. The dateline of erection is marked as a paragraph (<p>), while the part of the string describing the date is tagged as <date>. In most cases, it is possible to 'translate' the lunar calendar dates into solar calendar days. Take the example below (No.50.06), the date "一九九二年歲次壬申仲冬吉日" means "an auspicious

day of the mid-winter in the year Ren Shen, 1992”. Mid-winter refers to the second month of winter in lunar calendar, i.e. the 11th month. Checking through the historical calendar, the 11th month in lunar calendar of the year Ren Shen (1992) is from November 24th to December 23rd. Therefore, the date can be furtherly specified by the tag <when-iso>, which presents the transformed date according to the solar calendar.

```
<div type="L4" subtype="transcription" xml:id="vol2-0695">
   <head>惠州會館籌建克力街新會所樂捐芳名錄</head>

   <p>本會新會所經籌委們多年之積極展開工作，以及獲得本外
   埠熱心同鄉之鼎力襄助下，終於如期建竣完工，並擇吉日於
   公元一九九二年十二月五日舉行隆重開幕典禮。
   此項歷史性之建舘任務，會舘同人對全體建委之努力不懈及
   熱心樂捐士女們作表揚功績和衷感謝外，謹致崇高敬意。
   茲將建委會及樂捐者之芳名列下留念。</p>

   <list>
      <item><persName>黃六輝先生</persName>  <measure quantity="20000" unit="yuan">貳萬元</measure></item>
      <item><persName>張金虎醫生</persName></item>
      <item><persName>何春泉先生</persName></item>
      <item><persName>阮志興先生</persName></item>
      <item><measure quantity="10000" unit="yuan">以上各壹萬元</measure></item>
      ...
   </list>
   <list>
      <head>舘會親鄉埠外</head>
      <item><orgName>武吉班讓雷厝公會</orgName><measure quantity="2000" unit="yuan">貳仟元</measure></item>
      <item><orgName>馬來西亞海陸會舘</orgName><measure quantity="632" unit="yuan">陸佰叁拾貳元</measure></item>
      ...
   </list>
   ...
   <list>
      <head>籌建委員會</head>
      <item><roleName>主　席</roleName>：<persName>張金虎</persName></item>
      <item><roleName>副主席</roleName>：<persName>廖基業</persName></item>
      ...
   </list>

   <p><date when-iso="1992-11-24/1992-12-23">一九九二年歲次壬申仲冬吉日</date>立</p>
</div>
```

Fig. 5. Sample of Marked-up “L4” Stele Text (excerpted from Chapter 50.4).

Plaque texts are generally composed of a short sentence stating the event, the main text, the donors, and the date. The four sections are displayed from the right to the left of the plaque. In the text of transcriptions, there are not titles for each plaque. While for the stele, the title is annotated below each picture of the stele in both Chinese and English. According to the “list of artefacts” of each chapter, the titles of plaques are the main text. The main text is bold among the other transcriptions. These main texts are labelled as <head>, being consistent with stele texts. The tag <dateline> is chosen for the sentence, which is located before the main text in TEI text. It is also a tag included to appear before <head>. The donors are encoded as <list> when the name list is long, and <p> when they are only one or two. The narrative text is marked as <p>, and the date in the end as <date> inside <p>, the same as the stele text.

Plaque texts have been translated into English in the book. The translations are processed as corresponding of their Chinese transcriptions: the subtypes are <translation>, and the serial number is encoded as <corresp="#">. These translations follow a consistent format: the main text in quotation, followed by a narrative description of the other parts of the plaque. Therefore, the main text is labelled as <head> in correspondence to its Chinese version, and the rest of the text is labelled as <p>. The example in figure 6 is the TEI text of a plaque from Wui Chiu Fui Kun (No.50.108).

Couplet texts are similar to those of plaques. The standard composition of a pair of couplets is: a line stating the event, two lines of main text (considered the title of the couplets), donors, and signature of calligrapher and date. The four sections are marked as: <dateline>, <head>, <p> or <list> (with specific <persName> or <orgName>), and <p> (with <persName> and <date> inside). Below is an example of couplets from Wui Chiu Fui Kun (No.50.26). Inscriptions on other artefacts are more or less in the same format of plaque epigraphs.

```
<div type="L4" subtype="transcription" xml:id="vol2-0699">
   <dateline>新加坡惠州會館
      新會所開幕暨成立一七一周年双慶誌禧</dateline>
   <head>精 誠 團 結</head>
   <list>
      <item><orgName>雪蘿莪豐湖聯誼社</orgName></item>
      <item><orgName>居鑒柔佛州惠州會館</orgName></item>
      <item><orgName>馬六甲惠州會館</orgName></item>
      <item><orgName>柴城惠州會館</orgName></item>
      <item><orgName>霹靂惠州會館</orgName></item>
      <item><orgName>太平惠州會館</orgName></item>
      <item><orgName>彭亨惠州會館</orgName></item>
      <item><orgName>東甲惠州會館</orgName></item>
   </list>
   <p>仝敬賀</p>
   <p><date when-iso="1992-12-05">一九九二年十二月五日</date></p>
</div>
<div type="L4" subtype="translation" corresp="#vol2-0699">
   <head>“With sincere spirit unite together”</head>
   <p>with congratulations on the double celebration of the new A
n Malaysian Wui Chiu Native-place associations</p>
   </div>
</div>
```

Fig. 6. Sample of Marked-up “L4” Plaque Text (excerpted from Chapter 50.5).

III. Python Programming

The target of the program is to extract the name lists of sponsors documented in the epigraphic texts in different Chinese temples and associations. In order to understand the context of the sponsors, the related information needs also to be elicited, including: the label, title, and date of the epigraphic text; origin of place, organization, title of the donor, donor’s name, amount of donation.

The preliminary module involved is “etree” of “lxml”, which is used to get the roots of xml documents. After getting the roots, the next step is to search tags used in the marked-up texts. The basic logic of the program is to find certain tags in the corresponding division (in this case, or), and then to elicit the texts connected with the tags. For example, to find the labels that are in , the code is:

```
for l in d.findall ('.//label', namespaces=ns):
    label = l.text
```

Since the TEI texts are separate xml files (one per chapter), the module “os” is imported for looping the program among the xml files. The filenames in the folder are segmented using the “format” function: format = xml.split ('.'). In this case, “format [-1]” refers to the file’s format, while “format [0]” refers to the filename before the dot. If the file is an xml document (if “format[-1] == 'xml'”:), the python module continues to work. In other words, the program loops around all the xml files (TEI files needing to be processed) in the folder.

The designed output is a spreadsheet. Therefore, the module “xlsxwriter” is imported to write the results. The output is named “donors”, and the code to build a new workbook is: book = xlsxwriter.Workbook ('donors.xlsx'). The result of each chapter is written in a separate worksheet, entitled as the filename of that xml document: sheet = book.add_worksheet (format[0]).

The labels of each column, from A1 to K1, are: Label, Epigraph, placeName, roleName, persName, orgName, quantity, quantity_value, size (cm) , Date, and Date-iso. “Epigraph” refers to the title of each epigraphic text, which is marked as <head>. For the plaques and couplets and other cultural artefacts there is not the amount of donation in the epigraphic texts. However, there is the size of the object, which is also marked by the label <measure>. It reflects the

amount of donation to some extent. For “Label”, located in “A1”, the code is: sheet.write ('A1','Label',bold). The element “bold” is an additional format. Since the labels occupy the first row of the sheet, the results starts from the second row (row = 1 col = 0).

The tags in the TEI texts to be read include:

<label>, <measure unit="cm"> from ;

<head>, <placeName>, <roleName>, <persName>, <orgName>, <measure>, <measure quantity="">, <date>, <date when-iso="">, from .

Most of the information to be written into the spreadsheet are texts of the TEI files, which are considered as “strings” in Python language. Some other information is annotated by the encoder inside the label, such as the transcription of <measure quantity="" unit=""> and <date when-iso="">, which are considered as “attributes” in Python language. The expression to elicit these two types of information are “.text” and “.get (' ')”. For example, the original date in epigraphic text is defined as date_1(column “Date”), and its corresponding date according to the solar calendar filled in <date when-iso=""> is defined as date_2 (column “Date-iso”), the expressions to elicit the two are: date_1=da.text; date_2=da.get ('when-iso'). Since there is more than one item after each label, the value of the cell needs to be set free before being defined.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Label	Epigraph	placeName	roleName	persName	orgName	quantity	quantity_	size(cm)	Date	Date-iso	
2	50.01	重建惠州會館碑			潮澄進		1000千元	1000	107 x 55	光緒二十年	1903	
3	50.01	重建惠州會館碑				永瑞隆	1000千元	1000	107 x 55	光緒二十年	1903	
4	50.01	重建惠州會館碑			王郁賢		400百元	400	107 x 55	光緒二十年	1903	
5	50.01	重建惠州會館碑				廣合源	200百元	200	107 x 55	光緒二十年	1903	
6	50.01	重建惠州會館碑			陳庚瑞		100百元	100	107 x 55	光緒二十年	1903	
7	50.01	重建惠州會館碑			吳光禹		100百元	100	107 x 55	光緒二十年	1903	
8	50.01	重建惠州會館碑				廣泰和	100百元	100	107 x 55	光緒二十年	1903	
9	50.01	重建惠州會館碑				南昌泰	100百元	100	107 x 55	光緒二十年	1903	
10	50.01	重建惠州會館碑			刘燦光		100百元	100	107 x 55	光緒二十年	1903	
11	50.01	重建惠州會館碑			張仲富		100百元	100	107 x 55	光緒二十年	1903	
12	50.01	重建惠州會館碑			黃茂昌		100百元	100	107 x 55	光緒二十年	1903	
13	50.01	重建惠州會館碑			黃禄勝		100百元	100	107 x 55	光緒二十年	1903	
14	50.01	重建惠州會館碑			林武家		100百元	100	107 x 55	光緒二十年	1903	
15	50.01	重建惠州會館碑			鄧連元		100百元	100	107 x 55	光緒二十年	1903	
16	50.01	重建惠州會館碑			刘秀合		100百元	100	107 x 55	光緒二十年	1903	
17	50.01	重建惠州會館碑			陳楨祥		100百元	100	107 x 55	光緒二十年	1903	
18	50.01	重建惠州會館碑				新就記	100百元	100	107 x 55	光緒二十年	1903	
19	50.01	重建惠州會館碑				新廣順	100百元	100	107 x 55	光緒二十年	1903	
20	50.01	重建惠州會館碑			陳庚先		＄〇十元	50	107 x 55	光緒二十年	1903	
21	50.01	重建惠州會館碑				廣萬和	＄〇十元	50	107 x 55	光緒二十年	1903	
22	50.01	重建惠州會館碑				祥安泰	＄〇十元	50	107 x 55	光緒二十年	1903	
23	50.01	重建惠州會館碑			陳在勳		＄〇十元	50	107 x 55	光緒二十年	1903	
24	50.01	重建惠州會館碑			余灶生		＄〇十元	50	107 x 55	光緒二十年	1903	
25	50.01	重建惠州會館碑				廣萬安	＄〇十元	50	107 x 55	光緒二十年	1903	
26	50.01	重建惠州會館碑			曾爵昇		＄〇十元	50	107 x 55	光緒二十年	1903	
27	50.01	重建惠州會館碑			邹鳳合		＄〇十元	50	107 x 55	光緒二十年	1903	
28	50.01	重建惠州會館碑			刘振泰		＄〇十元	50	107 x 55	光緒二十年	1903	
29	50.01	重建惠州會館碑			刘德合		＄〇十元	50	107 x 55	光緒二十年	1903	

02-vol2-ch50 | 02-vol2-ch51 | 02-vol2-ch52 | 02-vol2-ch53 | 02-vol2-ch54 | 02-vol2-ch55 | 02-vol2-ch56 | 02-vol. ...

Fig. 7. Sample of the output of the program.

IV. Interaction between TEI and Python

A. Data Structure Design

The template outlines that the data structure has undergone several major changes due to the epigraphic text types. In the first draft, epigraphic texts are marked-up within "L3" as anonymous blocks (<ab>). With the progress of mark-up, "L4" (div type="L4") becomes necessary for the epigraphic texts, especially for stelae including long name lists (the list of donors, organizers) and inventories of income and expenses. On the other hand, in order to extract specific information, distinct tags are required for different components of the text. For example, the title of the epigraphic text needs to be marked differently from the whole transcription, especially where it is not highlighted in the book.

TEI text has its restrictions on the labels, since the xml files need to be valid. Take the case of the title of an epigraphic text, for instance. If the transcription is marked as <ab>, labels like <head> cannot be inserted. The solution is to mark the block as a division. In many epigraphic texts, there is a line stating the event before the body text, which is considered the title. However, only a few labels can be placed before <head>, such as <dateline>, while other labels, such as <p> and <date>, will result as errors in the validation process of xml.

For another instance, the end of the division ( ) can only be followed by the end of the higher level of division, or the starting of another division . Apart from the stele, the basic information of the medium is in brackets after the English translation of that epigraphic text. If the details are marked as <ab>, they cannot be located after the of translation. Therefore, they are moved forward to be after the <label> and before of the transcription, which is consistent with the stele section. Moreover, the <label> located after the <head> of the image is also a sort of compromise with the original text, since <label> cannot be inside <figure> and <head> cannot appear independently after <figure>.

B. The Order of Tags

TEI encoding provides a comprehensive framework for all kinds of annotations of texts. Although the Guidelines contain examples, they allow also flexibility for different cases. The main target of TEI standard is to annotate a text in philological perspectives.

Take the person’s name, for instance. According to the TEI standard, the related information of a person should be included in the label <persName>. It is possible to further distinguish the surname and first name by the labels <surname> and <forename>. The person’s title marked as <roleName> is considered componential information of the person (Text Encoding Initiative Consortium 2019: 454).

On the other hand, in programming language, the text is read in a linear way. If there is a label inside another label, the string is interrupted by the internal label, and only the last level can be extracted. Take samples from the CES Project, for instances: “Mr. Lim Hng Kiang, Minister for National Development and Second Minister for Foreign Affairs (國家發展部部長兼外交部第二部長林勛強先生)” (No. 1.37) should be marked as <persName><roleName>國家發展部部長兼外交部第二部長</roleName>林勛強先生</persName>. However, if it is marked this way, only the text inside <roleName> can be extracted as the role name when searching for the person’s name. In other words, the internal label interrupts the label “embracing” it, and the string tagged by the external label cannot be recognized by the programming language. For this reason, <roleName> is marked separately outside <persName>. When the stage to assign “xml:id” to each person is reached, the details of a person can be wrapped in the general label <person> (Text Encoding Initiative Consortium 2019: 467).

In some special cases, such interruption is inevitable, e.g.: the title is inside the name. A sample of such name is “surname + title + given name”: “Tailor Chua Duan’in 蔡和裁傳引” should be marked as “<persName>蔡<roleName>和裁</roleName>傳引</persName>” (No. 47.02). However, it will result that only the character “Chua 蔡” can be found as the person’s name. The solution could be skipping the label <roleName>, leaving it to be analyzed in the output.

V. Conclusion

This paper described the digital humanities approach applied to the analysis of donors documented in Chinese epigraphy in Singapore. The Chinese epigraphic texts preserved in temples and associations provide a rich resource for biography studies on the pre-modern history of Singapore. Tens of thousands of names of individuals and organizations depict social activities in the past and provide information on events not recorded in historical texts.

One of the analyzed databases is the two-volume work, Dean & Hue (2017), a recent comprehensive and systematic compilation of Chinese epigraphic texts in Singapore from 1819 to 1911. In order to figure out the information of the sponsors' names and to make these data searchable for various purposes, TEI standard has been chosen to mark-up the texts. In the current data structure, there are four paralleled sections in each chapter: , , , and . And four hierarchical levels of texts in the body part of each chapter: , , , and . The basic information of epigraphy is in the third level, and the transcriptions of inscriptions are in the fourth level. Three main types of epigraphic texts (stele, plaque, and couplets) are presented to explain the different sets of labels used in marking the texts in . The whole book is divided into sixty-four xml documents: sixty-three chapters and one appendix.

Based on the marked-up texts, a Python script has been designed to extract the sponsors' names and the related information. The modules involved include "etree" of "lxml", "os", and "xlsxwriter". The first module is implemented to parse the xml files, the second is used from looping the Python program around the folder, and the third is utilized to write the results into an excel workbook. The primary logic of the program is to find certain labels and to read the texts tagged by them or the attributes annotated by the encoder.

The combination of TEI and Python is a mutual process. Although programming script is based on data structure, the features of programming language also influence the details in arranging the labels and the order of texts. TEI aims at a philological annotation of texts. It shows certain criteria that have to be respected, while keeping flexibility at the same time. Python provides a relatively accessible programming language to manage copious amounts of data in an automatic way. Around 5,5000 names in 64 files are involved in this research. This convergent approach requires a patient manual work in preparing TEI data and commitment in programming, which is challenging and promising at the same time.

References

[1] J. Chen & Y. Chen. Xinjiapo Huawen Beiming Jilu 新加坡華文碑銘集錄 [A Collection of Chinese Epigraphy in Singapore]. Hongkong: Chinese University of Hong Kong Press, 1970.

[2] D. Chng. "Xinjiapo Huawen Mingke Jilu Chubian 新加坡華文銘刻集錄初編 [A Preliminary Collection of Chinese Epigraphy in Singapore]," Proceedings of Humanities and Social Sciences 4, pp. 73-108, 1984.

[3] K. Dean & G. T. Hue. Chinese Epigraphy in Singapore, 1819-1911 新加坡華文銘刻彙編 1819-1911. Singapore: NUS Press, 2017.

[4] X. Lin *et al*. Shile Guji 石叻古跡 [Historic Sites of the Republic of Singapore]. Singapore: South Seas Society, 1975.

[5] O. S. Song. One Hundred Years' History of the Chinese in Singapore. London: John Murray, 1923.

[6] Text Encoding Initiative Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange, 2019. URL: www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

[7] W. Franke & T. Chen. "A Chinese Tomb Inscription of A.D. 1264, discovered recently in Brunei," Brunei Museum Journal 3:1, pp. 91-9, 1973.

Appendix: Python Script

```
from lxml import etree

# loop
import os

# write into spreadsheet
import xlsxwriter
book = xlsxwriter.Workbook('donors.xlsx')

# add formats
bold = book.add_format({'bold': True})

listofxml = os.listdir("./")
for xml in listofxml:
   format = xml.split('.')
   if format[-1] == 'xml':

      doc = etree.parse(xml)

      ns = {None:'http://www.tei-c.org/ns/1.0',
        'xml':'http://www.w3.org/XML/1998/namespace'}
      root = doc.getroot()

      sheet = book.add_worksheet(format[0])

      # cols =
['Label','Epigraph','roleName','persName','orgName','quantity','quantity_val
ue','size','Date']
      sheet.write('A1','Label',bold)
      sheet.write('B1','Epigraph',bold)
      sheet.write('C1','placeName',bold)
      sheet.write('D1','roleName',bold)
      sheet.write('E1','persName',bold)
      sheet.write('F1','orgName',bold)
      sheet.write('G1','quantity',bold)
      sheet.write('H1','quantity_value',bold)
      sheet.write('I1','size(cm)',bold)
      sheet.write('J1','Date',bold)
      sheet.write('K1','Date-iso',bold)

      # start from second row
      row = 1
      col = 0
      for d in root.findall('.//div', namespaces=ns):
         if d.attrib['type'] == 'L3':
            epigraph = ' '
            size = ' '
            for l in d.findall('.//label', namespaces=ns):
               label = l.text
               for s in d.findall('.//seg', namespaces=ns):
                  if s.attrib['{http://www.w3.org/XML/1998/namespace}lang']
== 'zh':
                     epigraph = s.text
               for a in d.findall('.//ab', namespaces=ns):
                  for me in a.findall('.//measure', namespaces=ns):
                     size = me.text

            date =' '
            for da in d.findall('.//date', namespaces=ns):
               date_1=da.text
               date_2=da.get('when-iso')

            for i in d.findall('.//item', namespaces=ns):
               place = ' '
               role = ' '
               name = ' '
```

```
    org = ' '
    quantity = ' '
    quantity_value = ' '
    for pl in i.findall('.//placeName', namespaces=ns):
        place = pl.text
    for r in i.findall('.//roleName', namespaces=ns):
        role = r.text
    for p in i.findall('.//persName', namespaces=ns):
        name = p.text
    for o in i.findall('.//orgName', namespaces=ns):
        org = o.text
    for m in i.findall('.//measure', namespaces=ns):
        quantity = m.text
        quantity_value = m.get('quantity')

    #print(row,label,role,name,org,quantity,date)
    sheet.write(row, 0, label)
    sheet.write(row, 1, epigraph)
    sheet.write(row, 2, place)
    sheet.write(row, 3, role)
    sheet.write(row, 4, name)
    sheet.write(row, 5, org)
    sheet.write(row, 6, quantity)
    sheet.write(row, 7, quantity_value)
    sheet.write(row, 8, size)
    sheet.write(row, 9, date_1)
    sheet.write(row, 10, date_2)
    row+=1

#print(epigraphic text marked as <ab> in TEI)
place = ' '
role = ' '
name = ' '
org =' '
quantity = ''
for a in d.findall('.//ab', namespaces=ns):
    for pl in a.findall('.//placeName', namespaces=ns):
        place = pl.text
    for r in a.findall('.//roleName', namespaces=ns):
        role = r.text
    for p in a.findall('.//persName', namespaces=ns):
        name = p.text
    for o in a.findall('.//orgName', namespaces=ns):
        org = o.text

if role or name or org or quantity:
    sheet.write(row, 0, label)
    sheet.write(row, 1, epigraph)
    sheet.write(row, 2, place)
    sheet.write(row, 3, role)
    sheet.write(row, 4, name)
    sheet.write(row, 5, org)
    sheet.write(row, 6, quantity)
    sheet.write(row, 8, size)
    sheet.write(row, 9, date_1)
    sheet.write(row, 10, date_2)
    row+=1

# save
book.close()
```

Community Level Old Place Names in the Northeast of Thailand for a Historical Digital Gazetteer

Yoshikatsu Nagata
Graduate School of Engineering
Osaka City University
Osaka, Japan
nagatay@osaka-cu.ac.jp

***Abstract*— To identify locations of old place names is important to studies on community history. However, as place names are not always stable for long, information on relationship between old place name and its corresponding current place name is indispensable. Present-day place names in digital form are readily utilized in various information services. But, practical cases of integration of old place names precisely in a wide region in digital form are not much ready. The major target of this article is community level place names in past one century in the Northeast of Thailand. Current achievement of linking place names between different age, and further tasks to improve the content of a digital gazetteer is described.**

Keywords—old place names, digital gazetteer, Northeast Thailand

I. INTRODUCTION

To identify corresponding locations of old place names found in old materials and heard in interviews is important to studies on community history. However, as place names are not always stable for long years, information on relationship between old place name and its corresponding current place name is required to identify the location of old place name. Locations of old place names should be in geographic coordinates on current standard geodetic system to be integrated into a geographic information system, GIS, for a spatial analysis.

Place names show geographical features and natural environments in general. Therefore, old place names which are different from their current names may suggest such features and environments in past which cannot be easily understood by current names. Many cases are observed that old place names in the local major language of the community were changed to names in the standard language of the nation, thus, old place names may include suggestive information on history of migration and development of rural community.

Present-day place names in digital form are readily utilized in various information services. For old place names, one model case is open for public which integrates place names of Japan in late 19th century and thereafter [1][2]; however, practical cases of integration of old place names precisely in a wide region in digital form are not much ready.

The author have several experiences of field surveys on development history of rural communities in the Northeast of Thailand. In interviews to informants of surveyed communities, many place names of village level of the birthplace of their selves or their ancestors have been mentioned. Many of such place names are memorized in old names but not always same with present-day names. Thus, sometimes it is not easy to identify their corresponding places. Their memory of ancestors' migration path is by place name at the time but not by place name at present day. Fortunately, a set map which covers the whole Thailand was completed in early 20th century or the day of several generations before, and place names of village level are drawn on the map. These place names on the map are indispensable to identify the location of a place in old name. But there are many difficulties to link old place names and present-day place names.

In this article, major causes of difficulty in linking old name and present-day name are described. Then, the degree of difficulty observed in the latest geographical analysis is shown. The further plan of integrating such old place names into a historical digital gazetteer is explained.

II. IDENTIFICATION OF OLD PLACE NAMES

In this article, old place names of major target to be identified their corresponding present-day names are names used in early 20th century in the Northeast of Thailand. Some parts of Lao P.D.R., the adjoining area of the Northeast of Thailand beyond the Mekong River, are also referred because many place names are mentioned as ancestors' original places in interviews in the Northeast of Thailand.

A. Materials

Place names have been collected using following materials of set maps and printed gazetteers.

(a) Set map of scaled 1 to 200,000 prepared by the Royal Survey Department, Thailand, and published around 1920 and later. Hereinafter, these are referred to as RSD maps.

(b) Set of the L708 series topographic maps scaled 1 to 50,000 prepared by the Royal Thai Survey Department in cooperation with the U.S. Army Map Service. They were published around 1960 and later. Hereinafter, these are referred to as L708 maps.

(c) Set of the L7017 series topographic maps scaled 1 to 50,000 prepared by the Royal Thai Survey Department and published around 1980 and later. Hereinafter, these are referred to as L7017 maps.

(d) Printed gazetteer in the period of the World War II, "Gazetteer to Maps of Thailand", published by the U.S. Army Map Service in 1944. Hereinafter, this is referred to as MOT.

(e) Printed gazetteer "Thailand: Official Standard Names approved by the United States Board on Geographic

This article is a part of the research supported by JSPS KAKENHI Grant Number JP24500312, JP26300010 and JP19K12700.

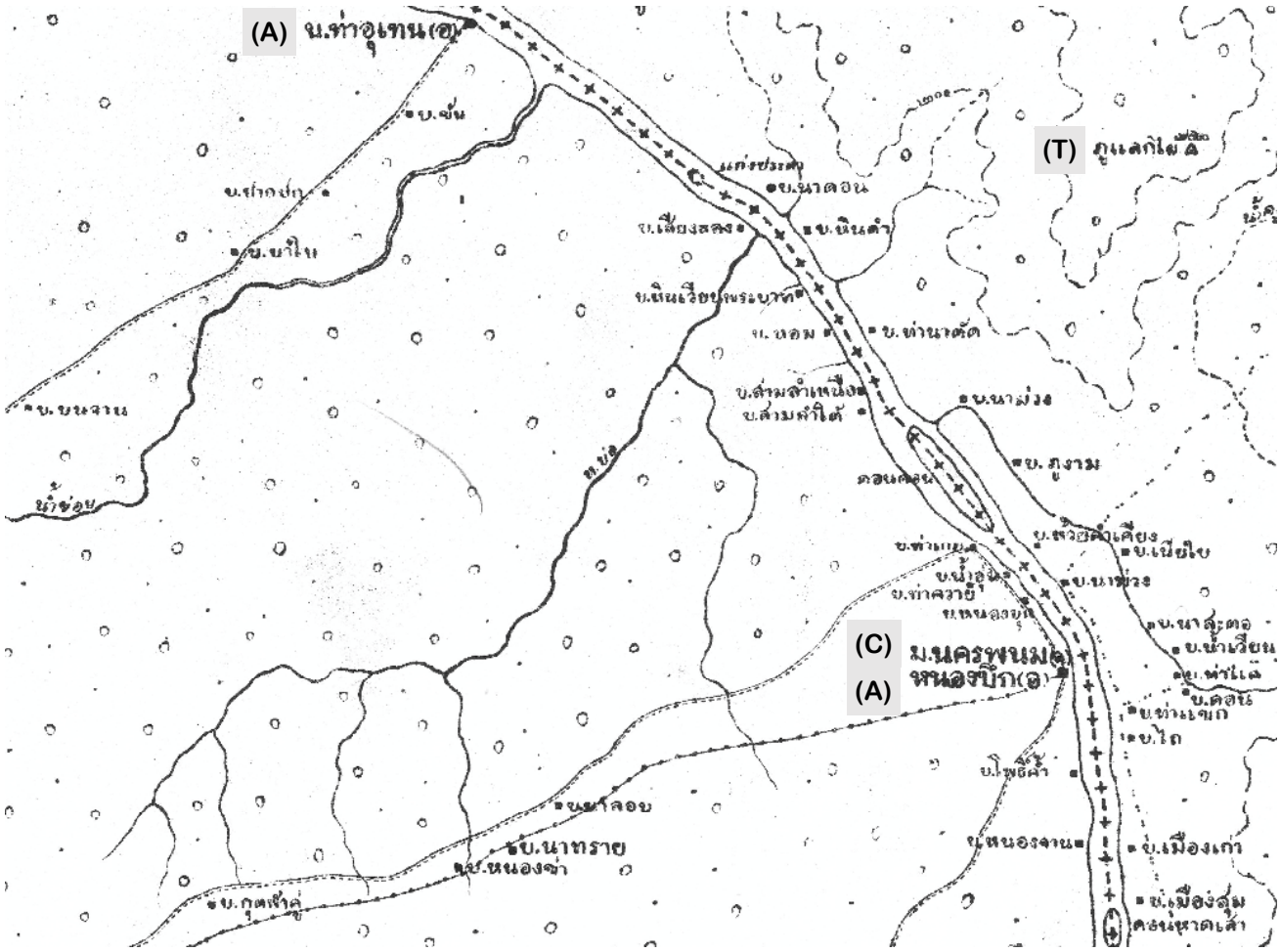


Fig. 1. Place names in the early 20th century around Nakhon Phanom (source: 'นครพนม' RSD)

Names" published in 1966. Hereinafter, this is referred to as OSN.

The RSD set map is the most important source material in this article since it records village level place names of about a century ago in Thai script. Fig. 1 is a part of the RSD map with index title "Nakhon Phanom". As in Fig. 1, village level names on both sides of the Mekong River are shown in Thai script. One province level name, marked as '(C)' in Fig. 1, and two district level name, marked as '(A)', are also shown.

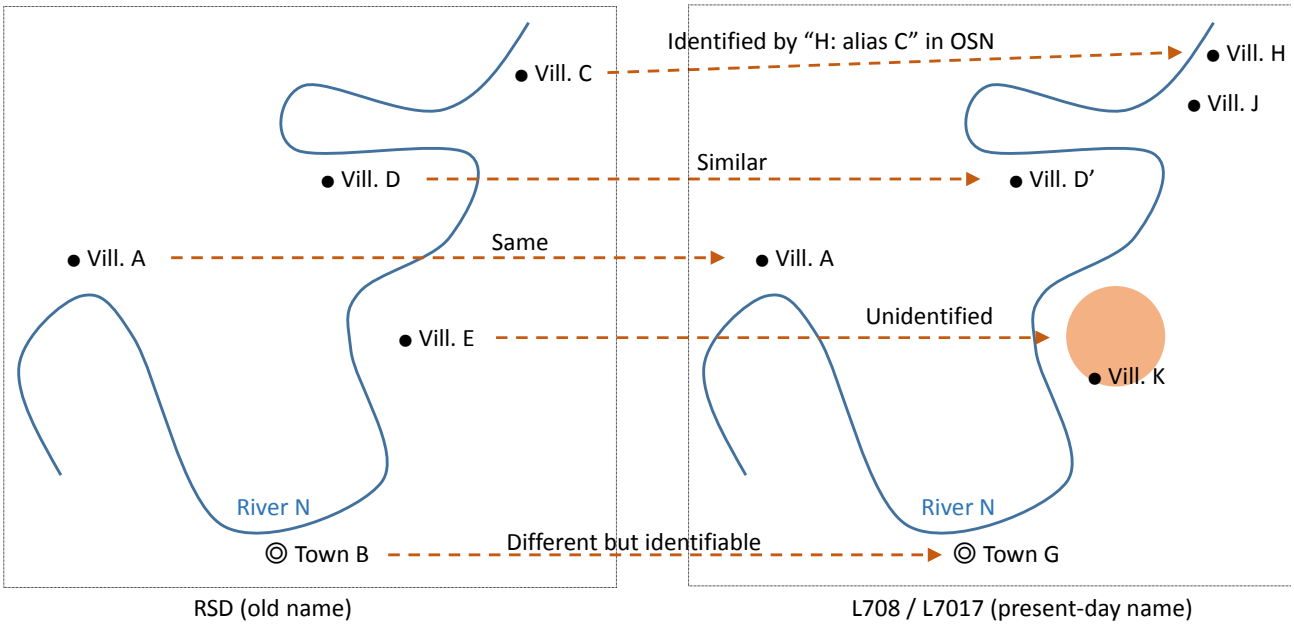


Fig. 2. Identification of place names

Source	Place name	Location
RSD	ผาหนาว (Pha Nao)	103° 33′ E 17° 26′ N
OSN	Pha Nao / Phan Na	103° 34′ E 17° 26′ N
L708	พันนา (Phan Na)	103° 34′ E 17° 26′ N

Fig. 3. Different names among sources

Triangulation point is also drawn, but only one point is found in Fig. 1, as marked '(T)'.

The MOT compiles many place names collected from various source materials at the time including the RSD set map; however, the MOT lists Romanized place names but not in Thai script. Therefore, homophone names in Thai cannot be distinguished by Romanized script.

B. Identification

Place names on the RSD maps have been tried to be identified their corresponding present-day place names or locations by the author. In this process, other materials mentioned above have been frequently referred. Fig. 2 is a conceptual diagram of this process. If a place name on the RSD is same or similar to a name on the L708 or the L7017 and their relative position to a stable landmark, such as the bending point of the river, major city or town, and peak of the mountain, the old place name can be identified to be the present-day name. In Fig. 2, 'Vill. A' is a case of same name, and 'Vill. D' is a case of similar name, then they can be easy to be identified. The case of 'Town B', it can be identified to be 'Town G' since they are located at the same place and only one large community in the area though names are different. If a name around the estimated corresponding old place name differs from the old name like a case of 'Vill. E' in Fig. 2, further information must be required to determine that those two names show the same place or not. The case of 'Vill. C'

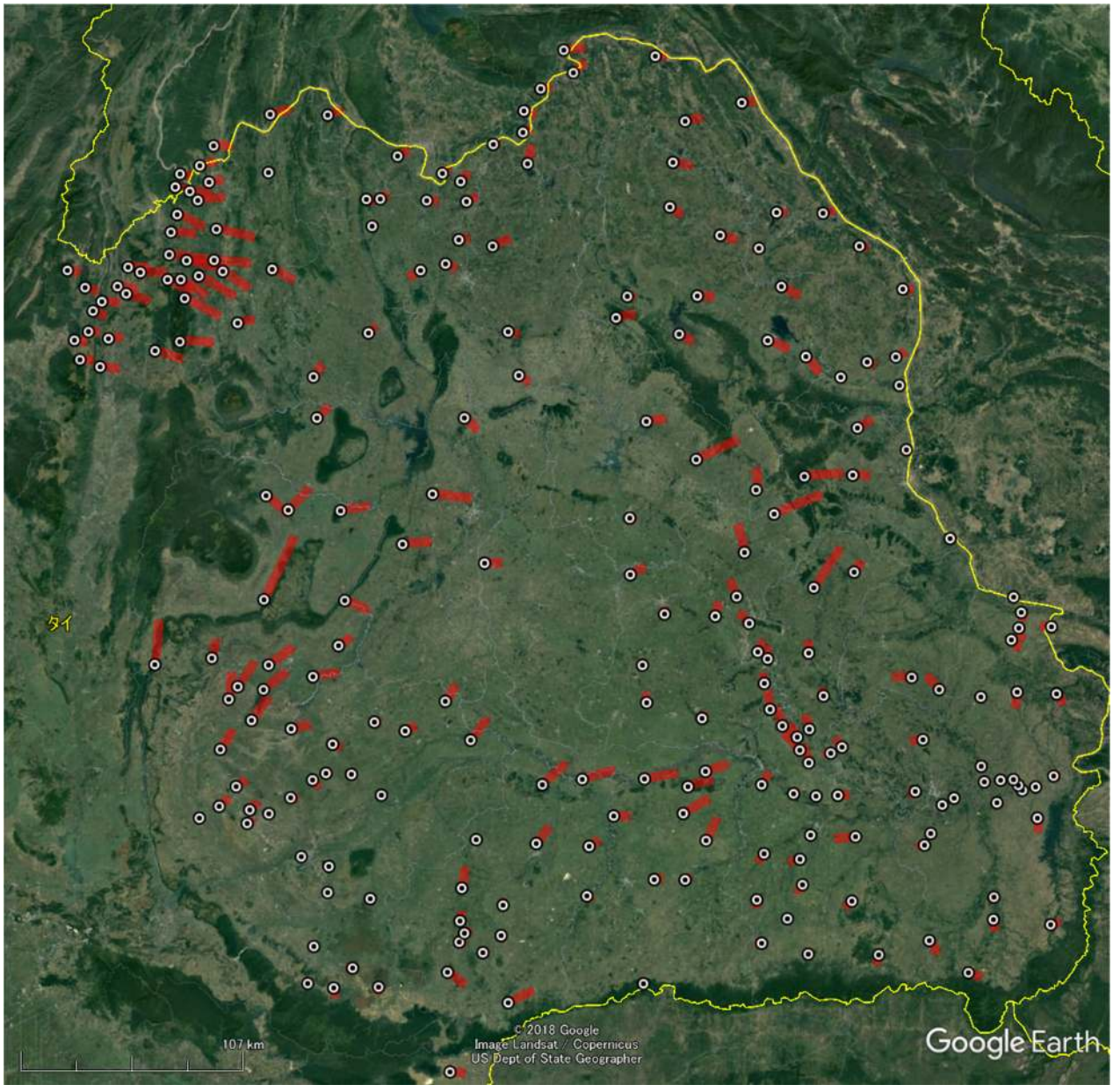


Fig. 4. Locational Difference of Places in the RSD

explains that it can be identified as 'Vill. H' since the OSN gives 'Vill. H' as alias of 'Vill. C'.

Among the source materials above, thus the OSN lists many alias names of a place which are proved indispensable to link different looking place names between the present and the past. Fig. 3 shows an example of a place where the place name in the RSD differs from that in the L708. Those two names 'Pha Nao' and 'Phan Na' look roughly similar, but two names cannot be sure to designate a same location. But for the entry of 'Phan Na' in the OSN, 'Pha Nao' is also listed as the alias of 'Phan Na'. Therefore, it is possible to identify those two names designate the same location.

In the RSD maps of scaled 1 to 200,000 of one century ago, about 3,500 place names were read, and more than half of them have been identified their corresponding place names of present-day. Fig. 4 shows the tendency of locational difference between the location designated in the RSD and the location presented in the L708 of the same place. In Fig. 4, the black dot shows the location designated in the RSD, and the red bar shows the distance to the correct location. It can be easily observed that such locational difference is not negligible when identifying corresponding place of a village level place name. At the worst case, the difference is more than 20km.

Fig. 4 also shows that the direction and the distance to the correct location differs place to place. In some areas close to the triangulation stations at the day, the difference is small, but due to small numbers of triangulation stations installed at the day, in many areas far from the triangulation stations the locational difference shows large. Though the longitude and latitude values of the edge of a map are drawn in the RSD maps, these geographical values should be treated as just for reference, and more accurate values should be gained after the relationship of old name and present-day name is confirmed.

TABLE I. SIMILARITY IN NAME

Similarity	Place names	Ratio
Same	991	28.6%
Similar	425	12.3%
Local Language	16	0.5%
Same Meaning	16	0.5%
Slightly Different	188	5.4%
Different	332	9.6%
Unidentified	1,497	43.2%

In the RSD maps, place names are shown in Thai script but rather described to show the pronunciation of major language of a community. But, in later years, such place names in local words have been gradually converted to words in national standard language. Many cases can be observed that changes in name from Lao word to Thai word, but these changes are so far easily traceable. Not a few cases are also observed that changes in name from Khmer words. Other typical change in place name is that from a name composed by words designating natural environment to a name with words which show conceptual goals such as friendship and development. In addition to changes in name, abandonment of a village or relocation of a village had been common in rural areas. Thus, there are many factors to make difficult to

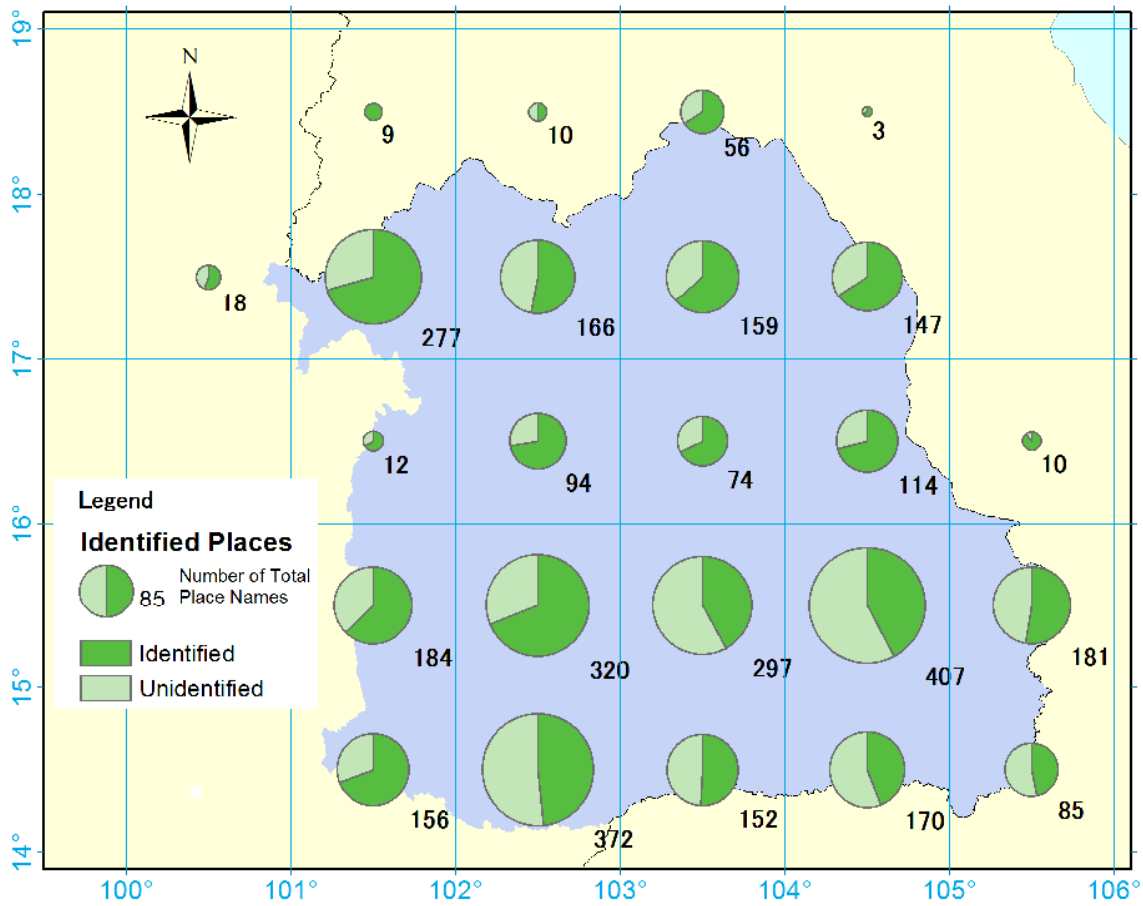


Fig. 5. Ratio of Identified Places

identify the exact location of old place names even they are shown on the RSD maps.

Table 1 shows number of places classified by similarity in place name between name in the RSD and present-day name. Places in the RSD which cannot be identified their present-day names are classified as 'unidentified'. As in Table 1, more than 40 percent of places use same or similar names for long and their location could be identified. About 15 percent of places could be identified even their old names differs from present-day names. The rest more than 40 percent of places are not identified at present, and further supplemental information must be explored.

The ratio of unidentified places differs from area to area. In Fig. 5, the ratio by each map sheet of the RSD, or rectangular area of one degree of latitude and one degree of longitude, is shown.

Areas located south from 16 degree North latitude show relatively high ratio of unidentified places. In these areas, there is a wide basin of the Mun River where many villages located in lowland tend to be relocated or abandoned in decades later. Ethnic background of these areas is another factor of showing insufficient identification. Since the adjoining region is Cambodia, many Khmer speaking communities were named in their language in past..

Fig. 6 brings close observation of similarity in place names. Fig. 6 covers Nakhon Phanom province, one of nineteen provinces in the Northeast of Thailand, showing similarity by place, or village level community. In many communities of this area, senior people explain that their ancestors of several generations before migrated from the opposite side of the Mekong River, Lao P.D.R., and many of such communities still keep more or less some relationships between their original communities in Lao by family or by community.

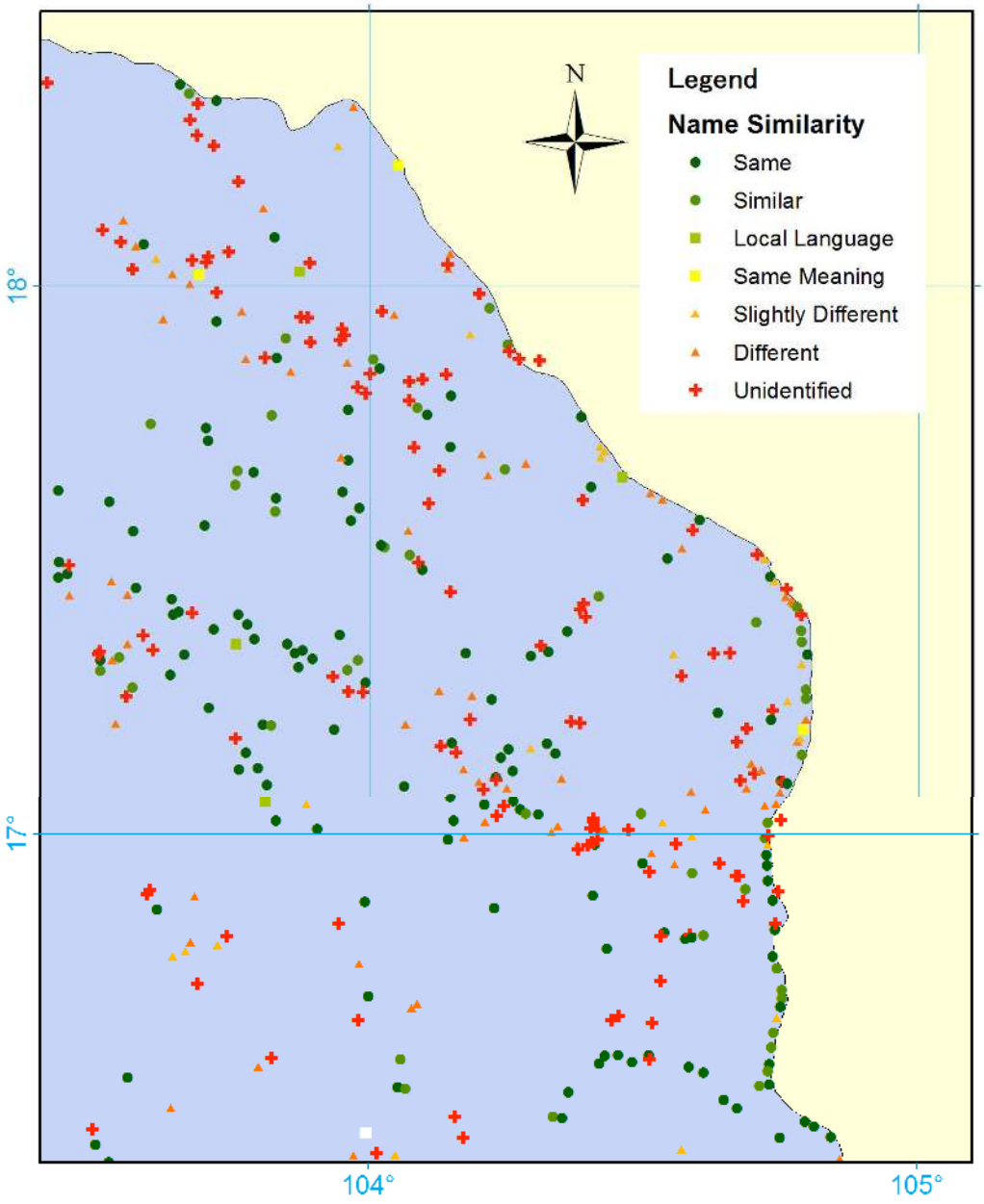


Fig. 6. Similarity in Names of Old Place Names at Nakhon Phanom Province

III. FURTHER TASKS

It must be a big loss for academic activities if place names in people's memory or historical record of a community cannot be mapped on the present-day materials. As the case described in this article deals old place names of merely several generations before, there must be positive possibility to obtain further information of such place names. One emergent activity is to collect and record personal memories of senior people who well know the history of the community and neighboring area. By conducting interviews to those people, some information on missing links of unidentified old place names can be expected available before such information will disappear forever.

Field surveys to interview to senior villagers are planned in Nakhon Phanom Province by the author with cooperation of local researchers in coming three years as a feasibility study to explore unpublished records or memories of changes and aliases of local place names.

Though absolute geographical coordinates of places drawn on the RSD maps are not reliable, relative location to nearby places can be utilized. For the field surveys, it is better to estimate corresponding places of unidentified name than to rely on the location originally drawn in the RSD. In fact, unidentified place names shown in Fig. 6 are placed in the figure with their estimated geographic coordinates calculated by nearby identified places.

Some part of progress in changes in place names in the Northeast of Thailand is already integrated into a digital gazetteer by the author. And further progress will be continuously integrated to make it richer. This historical digital gazetteer may contribute as one of academic basic infrastructures.

REFERENCES

[1] Oketani, I., 2009, The Database of Topographical Maps and Place Names, IPSJ SIG Technical Report, Vol.2009-CH-83 No.3, 1-8, 2009

[2] Yotsui, K., Sekino, I., Hara, S., Oketani, I., and Shibayama M., 2010, Construction of an Historical Gazetteer Based on 1:50,000 Maps from the Meiji and Taisho Eras, IPSJ Symposium Series, Vol.2010 No.15, 211-216

The native and cultivated *Phalaenopsis* species from biogeography to horticulture

Xiao-Lei Jin
Department of Biological Sciences
National Sun Yat-Sen University
Kaohsiung, Taiwan
shelleyjin25@gmail.com

Chi-Chu Tsai
Kaohsiung District Agricultural Research and Extension Station
Pingtung, Taiwan
tsaicc@mail.kdais.gov.tw

Yu-Chung Chiang*
Department of Biological Sciences
National Sun Yat-Sen University
Kaohsiung, Taiwan
yuchung@mail.nsysu.edu

Abstract**—***Phalaenopsis* is one of the important commercial orchids in the world. Members of the *P. amabilis* species complex represent invaluable germplasm for the breeding program. However, the phylogeny of the *P. amabilis* species complex is still uncertain. The *P. amabilis* species complex (Orchidaceae) consists of subspecies *amabilis, moluccana,* and *rosenstromii* of *P. amabilis,* as well as *P. aphrodite* ssp. *aphrodite*, *P. ap*. ssp. *formosana*, and *P. sanderiana*. The moth orchid (*Phalaenopsis* species) is an ornamental crop that is highly commercialized worldwide. Molecular identification based on microsatellite loci is an important technology to improve the commercial breeding of the moth orchid. There are more than 30,000 cultivars have been enrolled at the Royal Horticultural Society (RHS). This study were to reconstruct the phylogeny and biogeographical patterns of the species complex using Neighbor Joining (NJ), Maximum Parsimony (MP), Bayesian Evolutionary Analysis Sampling Trees (BEAST) and Reconstruct Ancestral State in Phylogenies (RASP) analyses based on sequences of internal transcribed spacers 1 and 2 from the nuclear ribosomal DNA and the trnH-psbA spacer from the plastid DNA. Ggenomic microsatellite primer sets were developed from *Phalaenopsis aphrodite* subsp. *formosana* to further examine the transferability of across 21 *Phalaenopsis* species.

Keywords— Biogeography, Demographic dynamics, Phylogeny, Species complex, Vicariance, *Phalaenopsis*, Microsatellites, Polymorphism, Transferability, horticulture

Introduction

Orchidaceae is regarded the world's biggest family of flowering crops with approximately 35,000 species. [18]. Vandaceous orchids include *Vanda, Ascocenda, Phalaenopsis, Renanthera, Rhynchostylis and Aerides*. The genus *Phalaenopsis* is often referred as moth orchid and comprises about 66 species (They consist of approximately 66 native species worldwide, 56 of which are extant) [13]. *Phalaenopsis* species is widespread in the Himalayas of northern India, South India, Sri Lanka, Southeast China, Taiwan, Indonesia, Thailand, Myanmar, Malaysia, the Philippines, Papua New Guinea and northeastern Australia (Species of *Phalaenopsis* been found throughout tropical Asia and the larger islands of the Pacific Ocean.) (Chen and [8,13]. *Phalaenopsis* divided into five subgenera according to the pollinia numbers [13] and molecular evidence [58]: the four pollinia clades of subgenera *Proboscidioides, Aphyllae, and Parishianae* and the two pollinium clades of subgenera *Polychilos* and *Phalaenopsis*. The Polychilos and *Phalaenopsis* were each subdivided into four sections among these subgenera, *Polychilos, Fuscatae, Amboinenses, Zebrinae* and *Phalaenopsis, Deliciosae, Esmeralda, Stauroglottis,* respectively [13,17]. Recently, the plastid genome of *Phalaenopsis aphrodite* been wholly sequenced [7], and molecular phylogenies of *Phalaenopsis* species also have been conducted based on the internal transcribed spacer (ITS) of the ribosomal DNA (rDNA) and plastid DNA [58, 65, 66, 67]. The internal transcribed spacer (ITS) region of nuclear ribosomal DNA (nrDNA) has provided valuable information for determining phylogenetic relationships of *Phalaenopsis* at intra-generic levels [66] and at the species complex level [67]. Plastid DNA has also extensively applied to evolutionary and phylogenetic research (Palmer 1987). Taberlet et al. (1991) developed a series of universal primers for several non-coding plastid regions. The trnL (UAA) intron, the trnL-F (GAA) spacer, and the atpB-rbcL spacer have been successfully used for the phylogenetic study of *Phalaenopsis* at an intra-generic level [58] and to reveal the natural hybridization of *Phalaenopsis* [66]. Some studies have used different markers to clarify the phylogenetics and biogeography of *Phalaenopsis* [68]. The aim of this study is to use different DNA fragments to explore the biogeographic relationship of native *Phalaenopsis* species, to use different sequence fragments to develop molecular marker in order to detect and trace back the related intergeneric, interspecific hybrid *Phalaenopsis* species and to develop molecular biotechnology for rapid identification of *Phalaenopsis* species.

I. Materials and Methods

Phylogenetic reconstruction and historical biogeography inference of the *Phalaenopsis amabilis* species complex

Thirty-nine accessions of the *P. amabilis* complex were obtained from 13 different populations, and three species of the *P. schilleriana* complex were used as outgroups. Using the cetyltrimethylammonium bromide (CTAB) method [59], total DNA was extracted from fresh etiolated leaves. To reconstruct the phylogeny and biogeographical patterns of the *P. amabilis* species complex using Neighbor Joining (NJ), Maxinum Parsimony (MP), Bayesian Evolutionary Analysis

Sampling Trees (BEAST) and Reconstruct Ancestral State in Phylogenies (RASP) analyses based on sequences of internal transcribed spacers 1 and 2 from the nuclear ribosomal DNA and the *trn*H-*psb*A spacer from the plastid DNA.

Identification of plastid genome types of *Phalaenopsis* hybrids

The plastid *trn*L intron sequence was determined for 54 native *Phalaenopsis* species. The inheritance of the plastid genome of three interspecific hybridizations of *Phalaenopsis* species including *P.* Yungho Gelb Canary, *P.* Timonthy Christopher, *P.* Rainbow Chip were determined based on inspection of the *trn*L intron sequence. The native *trn*L sequences were used to identify the plastid genome type of various *Phalaenopsis* hybrids. PCR-amplified DNA sequencing was used to determine the *trn*L intron genotypes of 54 *Phalaenopsis* species, representing over 95% of the living species diversity within this genus, and these sequences were submitted to GenBank (accession numbers: AY265742–48, AY265750–61, AY265763–87, AY265793, AY265795–800, DQ194981–82, DQ195040).

Genomic in situ hybridization (GISH), PCR-RFLP and RFLP analyses of intergeneric hybrid species

The various techniques including GISH, PCR-RFLP and RFLP analyses of nrDNA. are used to evaluate the genetic inheritance of these hybrids. The plant materials used in this study consisted of F1 intergeneric hybrid seedlings derived from an *Ascocenda* John De Biase 'Blue' female parent and a *P*. Chih Shang's Stripes male parent. Young root tips were collected from the two parents and the putative hybrids. These root tips were then treated for chromosomes slides. Non-overlapping chromosomes were selected for the GISH experiments. Genomic DNA was isolated from fresh leaves using the CTAB method. External transcribed spacer (ETS)-specific primers for PCR amplification were designed based on reference sequences from the GenBank database. The PCR products were separately digested with HaeIII and electrophoresed.

Screening, sequencing microsatellite loci, and primer designation across Genus *Phalaenopsis*

Sequencing, de novo assembly and develop EST-SSR loci

For Illumina transcriptome deep sequencing, total RNA was extracted from the fresh leaves of *P. aphrodite* subsp. *formosana* using the RNeasy Plant Mini Kit (Qiagen, Germany) according to the manufacturer's protocol. The library was sequenced using Illumina HiSeq™ 2000 (Illumina Inc., San Diego, CA, USA) according to the manufacturer's instructions. De novo transcriptome assembly for the high-quality reads (Q < 20) was performed using Trinity software. SSR loci were isolated in all the unigenes from *P. aphrodite* subsp. *formosana* with SciRoKo 3.4 software. The specific primers for each of EST-SSR loci were separately designed using BatchPrimer3 developed by You et al. To validate the polymorphism and transferability of the EST-SSR markers derived from transcriptome deep sequencing, 22 native *Phalaenopsis* species and 12 commercialized cultivars were the plant materials respectively examined. One hundred milligrams of fresh leaves were ground in liquid nitrogen, and genomic DNA was extracted using the CTAB method. PCR conditions and IRDye label procedure were referenced from Tsai et al [65].

Develop transferable microsatellite markers from *P. aphrodite* subsp. *formosana* using the modified magnetic bead enrichment method.

The DNA sample from *P. aphrodite* subsp. *formosana* was screened for microsatellites using the modified magnetic bead enrichment method. Total DNA was extracted from tissue culture seedlings or young leaves following the procedure by a Plant Genomic DNA Extraction Kit (RBC Bioscience, Taipei, Taiwan). Sequences containing microsatellites were detected using Tandem Repeats Finder version 4.09, and primer pairs were designed for microsatellite loci with suitable flanking regions to amplify using FastPCR software version 6.5.94 [34]. To verify the effectiveness and polymorphisms of microsatellite loci, all primer pairs designed for amplifying these microsatellites were tested using the *P. aphrodite* subsp. *formosana* DNA samples together with the other 20 *Phalaenopsis* species. The Bayesian clustering method was used to estimate genotyping group information and genetic components for 21 *Phalaenopsis* taxa with the assistance of STRUCTURE ver. 2.3.4 [19,47].

II. Results and Disscussion

Biogeography of the *Phalaenopsis amabilis* species complex inferred from nuclear and plastid DNAs

Our research was to reconstruct the phylogeny and biogeographical patterns of the *P. amabilis* species complex using Neighbor Joining (NJ), Maximum Parsimony (MP), Bayesian Evolutionary Analysis Sampling Trees (BEAST) and Reconstruct Ancestral State in Phylogenies (RASP) analyzes based on the sequences of internal transcribed spacers 1 and 2 from the nuclear ribosomal DNA and the trnH-psbA spacer from the plastid DNA. The *P. amabilis* species complex spans a wide geographic range covering southeastern Taiwan, the Philippines, Borneo, Sumatra (Mentawai Is.), Java, Sulawesi, Molucca Is., New Guinea, and northeastern Australia (Queensland) (Fig. 1), which has a complicated biogeography as these regions border two palaeocontinents [69]. These regions have drawn the attention of biogeographers for a long time and there have been several phylogeographic breaks (Fig. 2) [44,46,71]. Species or subspecies of the species complex shared most morphological characters, except for lip-midlobes and calli. The distributional patterns coincided with the geographic isolations via historical phylogeographic breaks between Borneo + Palawan and the Philippines (Huxley's Line), between Borneo and Sulawesi (Wallace's Line), between Sulawesi and Molucca Is. (Weber's Line), and between Sulawesi and New Guinea + Australia (Lydekker's Line) [40,44,46,71]. This species complex is an excellent tool for studying biogeography due to the distribution of this *P. amabilis* complex across the region between Southeast Asia and Australia, where several phylogeographic break lines have been identified. Thirty-nine accessions of the *P. amabilis* complex were obtained from 13 different populations, and three species of the *P. schilleriana* complex were used as outgroups. A pattern of vicariance, dispersal,

and vicariance + dispersal among disjunctly distributed taxa was uncovered based on RASP analysis. Although two subspecies of P. aphrodite could not be differentiated from each other in dispersal state, they were distinct from P. amabilis and *P. sanderiana*. Within *P. amabilis*, three subspecies were separated phylogenetically, in agreement with the vicariance or vicariance + dispersal scenario, with geographic subdivision along Huxley's, Wallace's and Lydekker's Lines. Molecular dating revealed such subdivisions among taxa of *P. amabilis* complex dating back to the late Pleistocene. Population-dynamic analyses using a Bayesian skyline plot suggested that the species complex experienced an in situ range expansion and population concentration during the late Last Glacial Maximum (LGM). The reduction of suitable habitats resulted in geographic fragmentation of the remaining taxa.

Plastid *trn*L intron polymorphisms among *Phalaenopsis* species used for identifying the plastid genome type of *Phalaenopsis* hybrids

Horticultural breeding by hybridization remixes floral characters such as the colours, shapes, and sizes, to generate diverse cultivars and varieties. Based on the elevated breeding and cultivation methods for light and feeding regulation and growth in the breeding and polyploidy of interspecific and intergeneric hybrids, improving the long-lasting quality of floral characteristics made *Phalaenopsis* one of the essential orchid sources for cut-flower crops. The *Phalaenopsis* genus species are the most common epiphytic monopodial orchids with unique composition for their distinctive and diverse flowers (with massive interspecific hybrids [59]. There are two indigenous *Phalaenopsis* species native to Taiwan, the *P. aphrodite* subsp. *formosana* and *P. equestris* [9], both of which have been classified as the Phalaenopsis section [13]. *Phalaenopsis aphrodite* subsp. *formosana* is well known as the Taiwan moth orchid, has been widely used as an essential breeding hybrids parent and is one of the most critical progenitors for the traits of modern extensive and white of floral organs commercial hybrids breeding [54]. *Phalaenopsis* equestris is another critical breeding parent for the small type of multi-flowers and artificial hybrids with white petals and sepals and a red lip [42,55]. *P*. Chih Shang's Stripes is an interspecific hybrid, which the genetic background includes species of *P. amboinensis, P. lueddemanniana, P. amabilis, P. aphrodite, P. equestris, P. schilleriana, P. sanderiana*, and *P. stuartiana* [43]. Although most intergeneric hybridizations in orchids fail because of irregular meiosis resulting from weak parental genomic homology [35,52], intergeneric hybridization among vandaceous plants is possible [29]. For example, Ascocenda John De Biase 'Blue' is an intergeneric hybrid between Ascocentrum and Vanda species [43]. The cultivar is known for its blue flowers and has been propagated and commercialized. Intergeneric hybrids have recently been created between *Phalaenopsis* and *Ascocenda* cultivars to bring orange colour into hybrid cultivars [39]. The Royal Horticultural Society (RHS) orchid database has registered more than 45,000 *Phalaenopsis* cultivars to date [49].

However, complicated phenotype and lengthy juvenile phase make it hard and time-consuming to identify *Phalaenopsis* plant varieties and cultivars. Moreover, by incorporating morphology, physiological development, and environmental variables as well as their complicated interactions, the traditional horticultural breeding method for fresh *Phalaenopsis* cultivars makes the breeding effect unpredictable and uncertain. Molecular markers could provide sensitive and accurate tools for identifying species and cultivars. In addition, molecular data was applied to determine the inheritance of the natural hybrid, *Phalaenopsis* x intermedia, showing *P. aphrodite* was the maternal parent, and *Phalaenopsis equestris* was the paternal parent [66]. The development of an extremely reliable, quick and inexpensive method to differentiate and identify *Phalaenopsis* species and cultivars seedlings is therefore essential and helpful for improving breeding effectiveness. In addition, the development of molecular markers could be applied to paternity assessment, phylogenetic reconstruction and long-standing problems relating to the reproduction of *Phalaenopsis*. Microsatellite markers with elevated polymorphism, codominant heritage and reproducibility features [48] are helpful instruments for plant genetics and crop breeding, including fruit tree [10,12,36,57] and orchid [63,64].

Universal primers for the trnL intron and trnL-trnF spacer were developed by Taberlet et al. [53] and have been used successfully to identify DNA sequences that are useful for phylogenetic markers at the intrageneric level, such as within Miscanthus, Saccharum [Poaceae; 27], Moraea [Iridaceae; 23], and Allium [Liliaceae; 71]. Furthermore, because organellar genomes are often uniparentally inherited, plastid and mitochondrial DNA polymorphisms have become molecular markers for investigating sex-biased dispersal and the directionality of introgression [72]. In our study, the plastid trnL intron sequence was determined for 54 native *Phalaenopsis* species. The inheritance of the plastid genome of three interspecific hybridizations of *Phalaenopsis* species was determined based on inspection of the trnL intron sequence. In addition, the native trnL sequences were used to identify the plastid genome type of various *Phalaenopsis* hybrids. PCR-amplified DNA sequencing was used to determine the trnL intron genotypes of 54 *Phalaenopsis* species, representing over 95% of the living species diversity within this genus, and these sequences were submitted to GenBank (accession numbers: AY265742–48, AY265750–61, AY265763–87, AY265793, AY265795–800, DQ194981–82, DQ195040). Almost all *Phalaenopsis* species were found to bear unique trnL intron sequences resulting from mutations, insertions/deletions, or both (Fig. 3&4). Molecular evidence has demonstrated that maternal inheritance of the plastid genome occurs during interspecific hybridization of *Phalaenopsis* species. Therefore, the plastid genome type of *Phalaenopsis* hybrids can be determined by comparing the trnL intron sequences of the hybrids to the GenBank database. The plastid genome type of the hybrids that are revealed through this analysis can be used to re-evaluate their genealogies because plastid DNA is maternally inherited.

Characterization of Genomic Inheritance of Intergeneric Hybrids between *Ascocenda* and *Phalaenopsis* Cultivars by GISH, PCR-RFLP and RFLP

Random amplified polymorphic DNA (RAPD) also has been conducted to reveal the phylogenetic relationship of *Phalaenopsis* species [3,19]. Chuang [14] reviewed several

accessions of *Phalaenopsis aphrodite* subsp. *formosana* and several associated Philippine *Phalaenopsis* species based on RAPD and inter-simple sequence repeat (ISSR) molecular markers. The findings showed that these two molecular techniques could provide informative markers to distinguish those from closely associated samples. In our study, in order to confirm the inheritance of putative hybrids obtained from the artificial hybridization of A. John De Biase 'Blue' (female parent) and P. Chih Shang's Stripes (male parent), the various methods including GISH, PCR-RFLP and RFLP analyses of nrDNA are used to evaluate the genetic inheritance of these hybrids. Genomic in situ hybridization (GISH) and restriction fragment length polymorphism (RFLP) analyses are strong molecular analytical methods. GISH was created by Schwarzacher et al. [51] to detect various chromosomes and big sections of DNA in interspecific or intergeneric hybrids and to evaluate chromosome pairing activity, translocation breakpoints and the genomic composition of polyploid crops. [47,51]. Botstein et al. [2] created the RFLP analysis that was commonly used in the development of genetic markers and linkage maps [5]. In both interspecific Phalaenopsis hybrids, RFLP was also used to demonstrate the maternal inheritance of cpDNAs [6]. Furthermore, a modified form of RFLP with the benefits of both RFLP and PCR, PCR-RFLP, was used to create high-fidelity, high-efficiency molecular markers requiring a reduced DNA concentration [41]. Recently, PCR-RFLP analysis of nrDNA has also been used for cultivar identification in intergeneric and interspecific hybrids [16,28,33,37,45,60,62]. GISH analysis showed the presence of both maternal and paternal chromosomes in the cells of the putative hybrids indicating that the putative hybrid seedlings were intergeneric hybrids derived from the hybridization of A. John De Biase 'Blue' (female parent) and P. Chih Shang's Stripes (male parent). In addition, GISH analysis showed no chromosome recombination in the hybrids (Fig. 5). Furthermore, twenty-seven putative hybrids were randomly selected for DNA analysis, and the external transcribed spacer (ETS) regions of nrDNA were analyzed using polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) and RFLP analyses to identify the putative hybrids. RFLP analysis showed that the examined seedlings were intergeneric hybrids of the two parents. However, PCR-RFLP analysis showed bias to maternal genotype. Both GISH and RFLP analyses are effective detection technology to identify the intergeneric hybridization status of putative hybrids. Furthermore, the use of PCR-RFLP analysis to identify the inheritance of putative hybrids should be carefully evaluated.

RNA-Seq SSRs of Moth Orchid and Screening for Molecular Markers across Genus *Phalaenopsis* (Orchidaceae)

SSRs have become common markers for population genetics [31,32,65], hybrid detection [38], mapping of linkages, genetic fingerprints [11,56], developmental history [20,21], and taxonomy [25,26]. EST-SSRs individually developed from the *Phalaenopsis* ESTs database have got 42 [24] and 261 EST-SSR loci [74]. Nine-hundred-fifty potential SSRs in *Phalaenopsis equestris* were found by largescale BAC end sequencing [30]. Deep sequencing techniques give the ability to generate countless SSR markers much quicker and at a reduced price than library-based techniques [1,4,15,50,73]. In our study, we performed de novo deep transcriptome sequencing of *P. aphrodite* subsp. *formosana* to analyze EST-SSR, develops molecular markers, and test the transferability among most *Phalaenopsis* members that are used as parents for moth orchid breeding. On the other hand, we extend to use more microsatellite loci as well as more extensive species testing in the next study to enhance the discriminatory power between the Phalaenopsis genus. We used Illumina sequencing technology to evaluate *Phalaenopsis aphrodite* subsp. *formosana*, a significant breeding species, and to create expressed sequence tag (EST)-simple sequence repeat (SSR) markers. After de novo assembly, the obtained sequence covered 29.1 Mb, about 2.2% of the *P. aphrodite* subsp. *formosana* genome (1,300 Mb), and a total of 1,439 EST-SSR loci were found. In order to validate the established EST-SSR loci and assess the transferability to the *Phalaenopsis* genus, thirty EST-SSR loci trinucleotide motifs were randomly chosen to design EST-SSR primers and assess the polymorphism and transferability across 22 indigenous *Phalaenopsis* species (Fig. 6), which are generally used as relatives for the reproduction of moth orchids. Of the 30 EST-SSR loci, ten polymorphic and transferable SSR loci across the 22 native taxa can be obtained. The validated EST-SSR markers were further proven to discriminate 12 closely related *Phalaenopsis* cultivars (Fig. 7).

Screening transferable microsatellite markers across genus *Phalaenopsis* (Orchidaceae)

In this study, the DNA sample from *P. aphrodite* subsp. *formosana* was screened for microsatellites using the modified magnetic bead enrichment method. Sequences containing microsatellites were identified using Tandem Repeats Finder version 4.09, and primer pairs were intended for microsatellite loci with suitable flanking regions to amplify using FastPCR software version 6.5.94 [34]. To verify the effectiveness and polymorphisms of microsatellite loci, all primer pairs designed for amplifying these microsatellites were tested using the P. aphrodite subsp. formosana DNA samples together with remaining 20 *Phalaenopsis* species. In total, 146 repeatable amplicons with length variation were screened from 28 microsatellite primer pairs in 21 species. The Bayesian clustering method was used to estimate genotyping group information and genetic components for 21 Phalaenopsis taxa with the assistance of STRUCTURE ver. 2.3.4 [48]. The allelic polymorphism information content (PIC) values reflect the extent of allele diversity among the species. The PIC value in our study is more significant than previous studies on *Phalaenopsis*. Due to the high transferability to species of the subgenus *Phalaenopsis*, these newly developed microsatellite primers can apply to establish a standard molecular identification operating system in Phalaenopsis. For genetically delimiting 21 species of the genus *Phalaenopsis*, a model-based Bayesian clustering algorithm was performed in STRUCTURE 2.3.4. The result showed that the first two best clustering numbers are K = 2 and K = 4. The assignment test by Bayesian clustering analysis reveals a similar result with molecular phylogeny patterns described by Tsai et al. [61]. The Bayesian clustering analysis based on EST-SSR loci could not get high resolution between either subgenus or sections within subgenus [63]. Compare to EST-SSR results published by Tsai et al. [63]; these newly developed genomic

microsatellite loci have higher resolution than EST-SSR loci when a study on native moth orchids.

III. Conclusions

Moth orchids (*Phalaenopsis* spp.) are among the most graceful and favourite plants which are highly commercialized worldwide. The *Phalaenopsis* species are valuable genetic resources for the breeding of hybrids in the horticultural market. The molecular identification markers are an essential technology for the breeder to improve the commercial cultivars. In our study, due to the DNA fragments of the internal transcribed spacer (ITS) of nuclear ribosomal DNA and the trnH-psbA intergenic spacer of plastid DNA reconstructs *P. amabilis* complex kinship and biogeographic relationship exploring the dynamic development of its ethnic history. On the other hand, there are breakthroughs in the genetic investigation and retrospective techniques of horticultural crossbreeding lines by using different molecular marker techniques and sequence fragments, the nucleotides of hybrids are rapidly clarified. Our results showed that it is not difficult to obtain universal SSR markers by in *Phalaenopsis* species. We developed many primer sets for the polymorphic microsatellite loci of *Phalaenopsis*, which are highly transferable among related species of the genus *Phalaenopsis*. Based on these transferable markers, delimitations between subgenera and between sections inferred by the Bayesian clustering analysis indicate that these SSR markers reveal high taxonomic resolution for paternity and hybridization application among genus *Phalaenopsis*. We clarified the kinship and biogeographic relationship of the *P. amabilis* complex and provided useful and cheap DNA barcoding markers for molecular breeding.

Figures and Tables

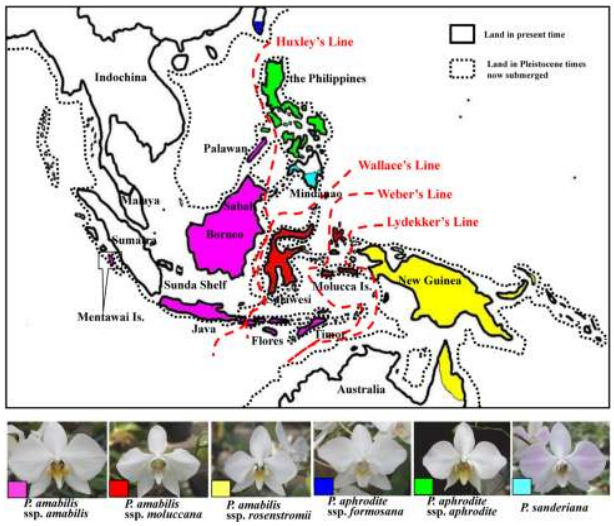


Fig. 1 Geographical distribution of six species/subspecies of the *Phalaenopsis amabilis* species complex and Southeast Asia landmasses between the Pleistocene and the present. In Pleistocene times, Indochina, Malaya, Sumatra, Java, Borneo, and the Philippines were interconnected and were separated from Sulawesi by the Makassar Strait. Four phylogeographic break lines were shown in red dashed lines (modified from [6]) and distribution region of six species/subspecies of the *Phalaenopsis amabilis* species complex drawed by different color in the map. Images for six species/subspecies of the *Phalaenopsis amabilis* species complex were photographed by Tsai, C.C. (the second author)

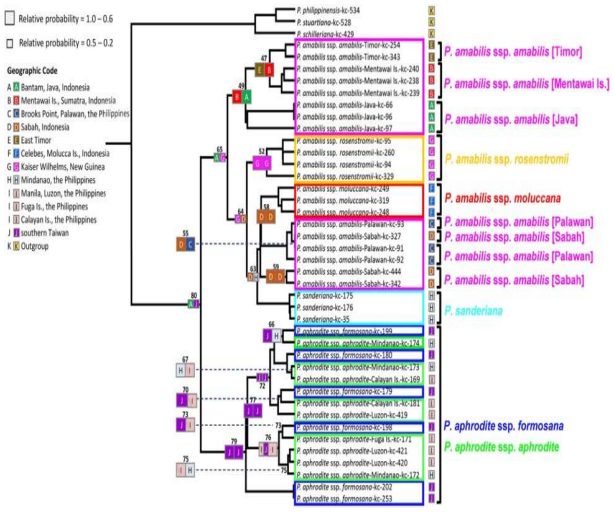


Fig. 2 Historical biogeographical reconstruction using Lagrange on the *P. amabilis* species complex topology. Coloured squares indicate reconstructed ancestral ranges and the square size is proportional to the probability of the reconstructions (see Table 7). The geographic ranges of species are displayed at right. [left | right]: 'left' and 'right' are the ranges inherited by each descendant branch (in the printed tree, 'left' is the upper branch, and 'right' the lower branch). The distribution areas of extant accessions of *P. amabilis* species are marked in capitals A–J (A: Bantam, Java, Indonesia; B: Mentawai Is., Sumatra, Indonesia; C: Brooks Point, Palawan, the Philippines; D: Sabah, Indonesia; E: East Timor; F: Celebes, Molucca Is., Indonesia; G: Kaiser Wilhelms, New Guinea; H: Mindanao, the Philippines; I: Manila, Luzon, Fuga Is., and Calayan Is. in the Philippines; J: southern Taiwan), respectively

Fig. 3 The polymorphic sites identified in the multiple sequence alignment for 12 species of the subgenus Phalaenopsis. Dots (···) indicate identical nucleotides, and dashes (– – –) indicate insertions or deletions.

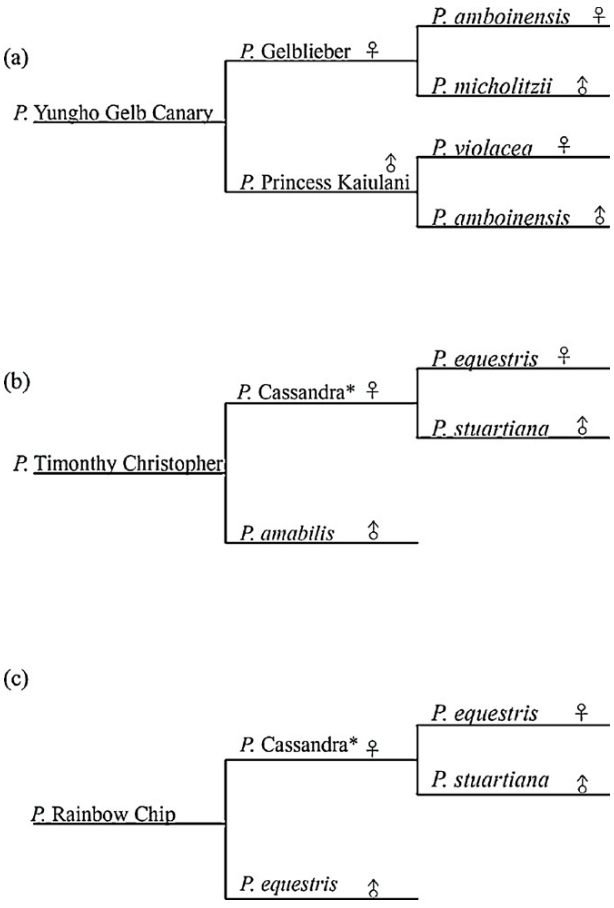


Fig. 4 The genealogies of *Phalaenopsis* Yungho Gelb Canary (A), *P.* Timonthy Christopher (B), and P. Rainbow Chip (C). These genealogies were redrawn based on the information from the Wildcatt Database. The asterisk represents the hybrid with the wrong genealogy resulting from the inverted submission between female parent and male parent.

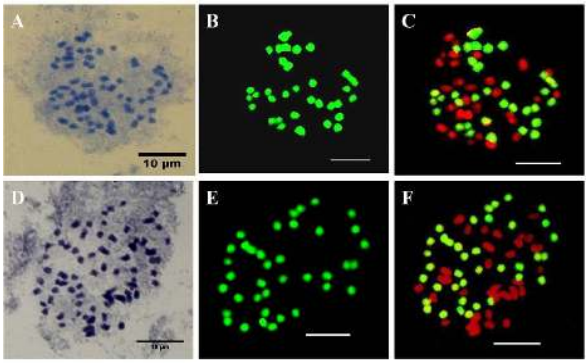


Fig. 5 GISH of somatic metaphase chromosomes from intergeneric hybrids between Ascocenda John De Biase 'Blue' (♀) and *Phalaenopsis* Chih Shang's Stripes (♂). Giemsa-stained chromosomes before GISH (A, D) and after GISH using total DNA from *Phalaenopsis* Chih Shang's Stripes, as revealed with FITC (B, E). A composite image of chromosomes counterstained with propidium iodide (red) and showing FITC signals (green) (C, F). Scale bar: 10 μm.

Fig. 6 Images of 22 *Phalaenopsis* species. Images (a)-(v).

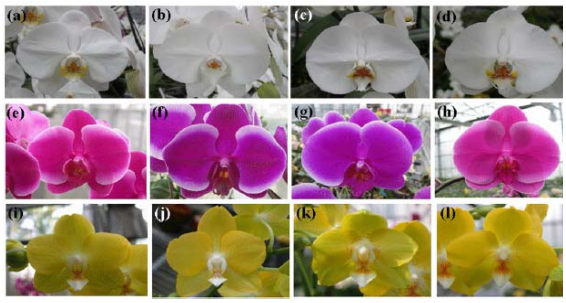


Fig 7. Images of 12 *Phalaenopsis* varieties. Images (a)-(l).

REFERENCES

[1] Abdelkrim J, Robertson BC, Stanton JAL, Gemmell NJ (2009) Fast, cost-efective development of spe- cies-specific microsatellite markers by genomic sequencing. Biotechniques 46: 185–192. doi: 10. 2144/000113084 PMID: 19317661

[2] Botstein D, White R, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32:314–331. PMID: 6247908

[3] Cai J, Liu X, Vanneste K, Proost S, Tsai WC, Liu KW, et al. (2015) The genome sequence of the orchid Phalaenopsis equestris. Nat Genet 47: 65–72. doi: 10.1038/ng.3149 PMID: 25420146

[4] Cavagnaro PF, Senalik DA, Yang L, Simon PW, Harkins TT, Kodira CD, et al. (2010) Genome-wide characterization of simple sequence repeats in cucumber (Cucumis sativus L.). BMC Genomics 11: 569–586. doi: 10.1186/1471-2164-11-569 PMID: 20950470

[5] Chang C, Bowman JL, DeJohn AW, Lander ES, Meyerowitz EM (1988) Restriction fragment length polymorphism linkage map for Arabidopsis thaliana. Proc. Natl. Acad. Sci. USA 85:6856–6860. PMID: 2901107

[6] Chang SB, Chen WH, Chen HH, Fu YM, Lin YS (2000) RFLP and inheritance paterns of chloroplast DNA in intergeneric hybrids of Phalaenopsis and Doritis. Bot. Ball. Acad. Sin. 41:219–223.

[7] Chang, C.C., Lin, H.C., Lin, I.P., Chow, T.Y., Chen, H.H., Chen, W.H., Cheng, C.H., Lin, C.Y., Liu, S.M., Chang, C.C., Chao, S.M., 2006. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol. Biol. Evol. 23, 279–291.

[8] Chen WH, Chen HH (2011) Orchid biotechnology II. World scientific Publishing Co. Ltd, Singapore

[9] Chen WH, Wang YT (1996) Phalaenopsis orchid culture. Taiwan Sugar 43:11–16

[10] Chiang YC, Tsai CM, Chen YKH, Lee SR, Chen CH, Lin YS, Tsai CC (2012) Development and characterization of 20 new polymorphic microsatellite markers from *Mangifera indica* L. (Anacardiaceae). Am J Bot 99(5):e117–e119

[11] Chiou CY, Chiang YC, Chen CH, Yen CR, Lee SR, Lin YS, et al. (2012) Development and characteriza- tion of 38 polymorphic microsatellite markers from an economical fruit tree, the Indian jujube. Am J Bot 99: e199–e202. doi: 10.3732/ajb.1100500 PMID: 22539510

[12] Chiou CY, Chiang YC, Chen CH, Yen CR, Lee SR, Lin YS, Tsai CC (2012) Development and characterization of 38 polymorphic microsatellite markers from an economical fruit tree, the Indian jujube. Am J Bot 99(5):e199–e202

[13] Christenson, E.A., (2001) *Phalaenopsis*. Timber Press, Portland, OR, USA, p. 330. Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19, 11–15.

[14] Chuang HT (2002) Identification of some species in the genus Phalaenopsis to Taiwan and Philippine by using RAPD and ISSR molecular markers. Mater Thesis, Graduate Institute of Agriculture, National Chiayi University, Chiayi, Taiwan. (Chinese with English abstract)

[15] Csencsics D, Brodbeck S, Holderegger R (2010) Cost-eVective, species-specific microsatellite devel- opment for the endangered dwarf bulrush (Typha minima) using next-generation sequencing technol- ogy. J. Hered 101: 789–793. doi: 10.1093/jhered/esq069 PMID: 20562212

[16] Devos N, Raspé O, Oh SH, Tyteca D, Jacquemart AL (2006) The evolution of Dactylorhiza (Orchida- ceae) allotetraploid complex: Insights from nrDNA sequences and cpDNA PCR-RFLP data. Mol. Phylo- gen. Evol. 38:767–778.

[17] Dressler RL (1993) Phylogeny and classi cation of the orchid family. Cambridge University Press, Cambridge

[18] Dressler RL (1993) Phylogeny and Classification of the Orchid Family. Dioscorides Press, Portland, Oregon, U.S.A.

[19] Fu YM, Chen WH, Tsai WT, Lin YS, Chou MS, Chen YH (1997) Phylogenetic studies of taxonomy and evolution among wild species of Phalaenopsis by random amplified polymorphic DNA markers. Rept Taiwan Sugar Res Inst 157: 27–42. (Chinese with English abstract)

[20] Ge XJ, Hsu TW, Hung KH, Lin CJ, Huang CC, Huang CC, et al. (2012) Inferring mutiple refugia and phylogeographical patterns in Pinus massoniana based on nucleotide sequence variation and finger-printing. Plos One 7: e43717. doi: 10.1371/journal.pone.0043717 PMID: 22952747

[21] Ge XJ, Hung KH, Ko YZ, Hsu TW, Gong X, Chiang TY, et al. (2015) Genetic divergence and bio- geographical patterns in Amentotaxus argotaenia species complex. Plant Mol Biol Rep 33: 264–280.

[22] Goh MWK, Kumar PP, Lim SH, Tan HTW (2005) Random amplified polymorphic DNA analysis of the moth orchids, Phalaenopsis (Epidendroideae: Orchidaceae). Euphytica 141: 11–22.

[23] Goldblatt, P., Savolainen, V., Porteous, O., Sostaric, I., Powell, M., Reeves, G., Manning, J.C., Barraclough, T.G., Chase, M.W., (20020 Radiation in the Cape flora and the phylogeny of Racock irises Moraea (Iridaceae) based on four plastid DNA regions. Mol. Phylogenet. Evol. 25, 341–360.

[24] Han SY (2005) Molecular cloning and characterization of cDNA-SSRs in Phalaenopsis. Master Thesis, Department of Biology, National Cheng Kung University, Tainan, Taiwan. (Chinese with English abstract)

[25] Hite JM, Eckert KA, Cheng KC (1996) Factors affecting fidelity of DNA synthesis during PCR amplifica- tion of d(C-A) n-d(G-T)n microsatellite repeats. Nucleic Acids Res 24: 2429–2434. PMID: 8710517

[26] Ho CS, Shih HC, Liu HY, Chiu ST, Chen MH, Ju LP, et al. (2014) Development and characterization of 16 polymorphic microsatellite markers from Taiwan cow-tail fir, Keteleeria davidiana var. formosana (Pinaceae) and cross-species amplification in other Keteleeria taxa. BMC Res Notes 7:255. doi: 10. 1186/1756-0500-7-255 PMID: 24755442

[27] Hodkinson, T.R., Chase, M.W., Lledo, M.D., Salamin, N., Renvoize, S.A., (2002) Phylogenetics of Miscanthus, Saccharum and related genera (Saccharinate, Andro- pogoneae, Poaceae) based on the DNA sequences from ITS nuclear ribosomal DNA and plastid trnL intron and trnL-F intergenic spacer. J. Plant Res. 115, 381–392.

[28] Holderegger R, Angelone S, Brodbeck S, Csencsics D, Gugerli F, Hoebee SE, et al. (2005) Application of genetic markers to the discrimination of European Black Poplar (Populus nigra) from American Black Poplar (P. deltoides) and Hybrid Poplars (P. x canadensis) in Switzerland. Trees 19:742–747.

[29] Holttum RE (1952) Hybridisation in Sarcanthinae: Its significance and limitations. Orchid J. 1:58–61.

[30] Hsu CC, Chung YL, Chen TC, Lee YL, Kuo YT, Tsai WC, et al. (2011) An overview of the Phalaenopsis orchid genome through BAC end sequence analysis. BMC Plant Biol 11: 3–13. doi: 10.1186/1471- 2229-11-3 PMID: 21208460

[31] Hsu TW, Shih HC, Kuo CC, Chiang TY, and Chiang YC (2013) Characterization of 42 microsatellite markers from poison ivy, Toxicodendron radicans (Anacardiaceae). Int J Mol Sci 14: 20414–20426. doi: 10.3390/ijms141020414 PMID: 24129176

[32] Huang CL, Ho CW, Chiang YC, Shigemoto Y, Hsu TW, Hwang CC, et al. (2014) Adaptive divergence with gene flow in incipient speciation of Miscanthus floridulus/sinensis complex (Poaceae). Plant J 80: 834–847. doi: 10.1111/tpj.12676 PMID: 25237766

[33] Jian L, Zhu LG (2013) Genetic diversity of Cymbidium kanran detected by polymerase chain reaction– restriction fragment length polymorphism (PCR-RFLP) markers. J. Plant Breed. Crop Sci. 5:158–163.

[34] Kalendar R, Lee D, Schulman AH (2009) FastPCR software for PCR primer and probe design and repeat search. Genes Genomes Genom 3(1):1–14

[35] Kamemoto H, Shindo K (1964) Meiosis in interspecific and intergeneric hybrids of Vanda. Botan. Gaz. 125, 132–138.

[36] Lai JM, Tsai CC, Yen CR, Ko YZ, Chen SR, Weng IS, Lin YS, Chiang YC (2015) Molecular characterization of twenty polymorphic microsatellite markers in the polyploid fruit tree species *Syzygium samarangense* (Myrtaceae). Genet Mol Res 14(4):13013–13021

[37] Li D, Qi X, Li X, Li L, Zhong C, Huang H (2013) Maternal inheritance of mitochondrial genomes and complex inheritance of chloroplast genomes in Actinidia Lind.: evidences from interspecific crosses. Mol Genet Genomics 288, 101–110. doi: 10.1007/s00438-012-0732-6 PMID: 23337924

[38] Liao PC, Tsai CC, Chou CH, Chiang YC (2012) Introgression between cultivars and wild populations of Momordica charantia L. (Cucurbitaceae) in Taiwan. Int J Mol Sci 13: 6469–6491. doi: 10.3390/ijms13056469 PMID: 22754378

[39] Liu YC, Lin BY, Lin JY, Wu WL, Chang CC (2016) Evaluation of chloroplast DNA markers for intraspeci c identi cation of *Phalaenopsis equestris* cultivars. Sci Hortic 203:86–94

[40] Lydekker R. (1896) A Geographical History of Mammals. Cambridge University Press, United Kingdom.

[41] Manhart L, McCourt RM (1992) Molecular data and species concepts in the algae. J. Phycol. 28:730– 737.

[42] Men S, Ming X, Wang Y, Liu R, Wei C, Li Y (2003) Genetic transformation of two species of orchid by biolistic bombardment. Plant Cell Rep 21(6):592–598

[43] Moir WWG (1995) In: An orchid database. Wildcatt Database Co., USA.

[44] Moss SJ, Wilson MEJ. (1998) Biogeographic implications of the Tertiary palaeogeographic evolution of Sulawesi and Borneo. In: Hall R, Holloway JD, editors. Biogeography and Geological Evolution of SE Asia. Leiden: Backhuys.

[45] Nwakanma DC, Pillay M, Okoli BE, Tenkouano A (2003) PCR-RFLP of the ribosomal DNA internal tran- scribed spacers (ITS) provides markers for the A and B genomes in Musa L. Theor. Appl. Genet. 108:154–159. PMID: 12955208

[46] Pianka ER. (1994) Biogeography and Historical Constraints. In: Evolutionary Ecology. 5th ed. New York: HarperCollins College Publisher, p. 15–40.

[47] Piperidis N (2014) GISH: resolving interspecific and intergeneric hybrids. Methods Mol. Biol. 1115:325–336. doi: 10.1007/978-1-62703-767-9_16 PMID: 24415482

[48] Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends Plant Sci 1(7):215–222

[49] Royal Horticultural Society, 2015: Available: http://www.rhs.org.uk/Plants/RHS-Publications/Orchid- hybrid-lists

[50] Santana Q, Coetzee M, Steenkamp E, Mlonyeni O, Hammond G, Wingfield M, et al. (2009) Microsatel- lite discovery by deep sequencing of enriched genomic libraries. Biotechniques 46: 217–223. doi: 10. 2144/000113085 PMID: 19317665

[51] Schwarzacher T, Leitch AR, Bennett MD, Heslop-Harrison JS (1989) In situ localization of parental genomes in a wide hybrid. Ann. Bot. 64:315–324.

[52] Stine M, Sears BB, Keathley DE (1989) Inheritance of plastids in interspecific hybrids of blue spruce and white spruce. Theor. Appl. Genet. 78: 768–774. doi: 10.1007/BF00266656 PMID: 24226004

[53] Taberlet, P., Gielly, L., Pautou, G., Bouvet, J., (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Mol. Biol. 17, 1105–1109.

[54] Tanaka Y, Katsumoto Y, Brugliera F, Mason J (2005) Genetic engineering in oriculture. Plant Cell Tissue Organ Cult 80(1):1–24

[55] Tang CY, Chen WH (2007) Breeding and development of new varieties in *Phalaenopsis*. In: Chen WH, Chen HH (eds) Orchids biotechnology. World Scienti c Publishing Co. Pvt. Ltd, Singapore, p 1–22

[56] Tsai CC, Chen YKH, Chen CH, Weng IS, Tsai CM, Lee SR, et al. (2013) Cultivar identification and genetic relationship of mango (Mangifera indica) in Taiwan using 37 SSR markers. Sci Hort 164:196– 201

[57] Tsai CC, Chen YKH, Chen CH, Weng IS, Tsai CM, Lee SR, Lin YS, Chiang YC (2013) Cultivar identi cation and genetic relationship of mango (*Mangifera indica*) in Taiwan using 37 SSR markers. Sci Hortic 164:196–201

[58] Tsai CC, Chiang YC, Huang SC, Chen CH, Chou CH (2010) Molecular phylogeny of *Phalaenopsis* Blume (Orchidaceae) on the basis of plastid and nuclear DNA. Plant Syst Evol 288(1–2):77–98

[59] Tsai CC, Chiang YC, Lin YS, Liu WL, Chou CH (2012) Plastid trnL intron polymorphisms among Phalae- nopsis species used for identifying the plastid genome type of Phalaenopsis hybrids. Sci. Hortic. 142:84–91.

[60] Tsai CC, Chiang YC, Liu WL, Huang SC, Chou CC. (2009) Intergeneric hybridization, embryo rescue and molecular detection for intergeneric hybrids between Ascocenda and Phalaenopsis. Acta Horticul- turae 829:413–416.

[61] Tsai CC, Huang SC, Chou CH (2005) Molecular phylogeny of *Phalaenopsis* Blume (Orchidaceae) based on the internal transcribed spacer of the nuclear ribosomal DNA. Plant Syst Evol 256:1–16

[62] Tsai CC, Huang SC, Huang PL, Chen FY, Su YT, Chou CH (2006) Molecular evidences of a natural hybrid origin of *Phalaenopsis* × *intermedia* Lindl. J. Hort. Sci. Biotech. 81:691–699.

[63] Tsai CC, Shih HC, Wang HV, Lin YS, Chang CH, Chiang YC, Chou CH (2015) RNA-seq SSRs of moth orchid and screening for molecular markers across genus *Phalaenopsis* (Orchidaceae). PLoS ONE 10(11): e0141761

[64] Tsai CC, Wu PY, Kuo CC, Huang MC, Yu SK, Hsu TW, Chiang TY, Chiang YC (2014) Analysis of microsatellites in the vulnerable orchid *Gastrodia avilabella*: the development of microsatellite markers and cross-species ampli ca- tion in *Gastrodia*. Bot Stud 55:72

[65] Tsai, C.C., Huang, S.C., Chou, C.H., (2006) Molecular phylogeny of *Phalaenopsis Blume* (Orchidaceae) based on the internal transcribed spacer of the nuclear ribosomal DNA. Plant Syst. Evol. 256, 1–16.

[66] Tsai, C.C., Huang, S.C., Chou, C.H., (2009) Phylogenetics, biogeography, and evolutionary trends of the *Phalaenopsis sumatrana* complex inferred from nuclear DNA and chloroplast DNA. Biochem. Syst. Ecol. 37, 633–639.

[67] Tsai, C.C., Sheue, C.R., Chen, C.H., Chou, C.H., (2010) Phylogenetics and biogeography of the *Phalaenopsis violacea* (Orchidaceae) species complex based on nuclear and plastid DNA. J. Plant Biol. 53, 453–460.

[68] Turner H, Hovenkamp P, van Welzen PC. (2001) Biogeography of Southeast Asia and the West Pacific. J Biogeogr; 28:217–30.

[69] van Oosterzee P. (1997) Where Worlds Collide: The Wallace Line. Ithaca, Cornell University Press, New York.

[70] Van Raamsdonk, L.M., Ensink, W., Van Heusden, A.W., Vrielink-Van Ginkel, M., Kik, C., (2003) Biodiversity assessment based on cpDNA and crossability analysis in selected species of Allium subgenus Rhizirideum. Theor. Appl. Genet. 107, 1048–1058.

[71] Wills, D.M., Hester, M.L., Burke, J.M., (2005) Chloroplast SSR polymorphisms in the Compositae and the mode of organellar inheritance in Helianthus annuus. Theor. Appl. Genet. 110, 941–947.

[72] Zalapa JE, Simon PW, Hummer KE, Bassil NV, Senalik DA, Zhu H, et al. (2012) Mining and validation of pyrosequenced simple sequence repeats (SSRs) from American cranberry (Vaccinium macrocarpon Ait.). Theor Appl Genet 124: 87–96. doi: 10.1007/s00122-011-1689-2 PMID: 21904845

[73] Zhang SM, Chen C, Chen FF, Wang T (2013) Analysis on Genetic Diversity of 16 Phalaenopsis Culti- vars Using EST-SSR Markers. J Plant Genet Resour 14: 560–564. (Chinese with English abstract)

Author Index