2016 第七屆數位典藏與數位人文國際研討會

The 7th International Conference of Digital Archives and Digital Humanities 2016

數位學者，一個新興領域的產生？
Digital Scholars, an emerging profession？

日期：2016 年 12 月 01 日(四) - 12 月 03 日(六)
Date：December 1-3, 2016
地點：國立臺灣大學凝態科學暨物理學館國際會議廳
Venue: Center for Condensed Matter Sciences (CCMS), National Taiwan University, Taipei, Taiwan

主辦單位 Organizer：
國立臺灣大學數位人文研究中心
Research Center for Digital Humanities, National Taiwan University
協辦單位 Co-organizers：
中央研究院數位文化中心　The Academia Sinica Center for Digital Cultures
法鼓文理學院　Dharma Drum Institute of Liberal Arts
臺灣數位人文學會　Taiwanese Association for Digital Humanities

議　程

第一天12 月 1 日(四)	
8:40 ｜ 9:00	報　　到
9:00 ｜ 9:20	開 幕 式
9:20 ｜ 10:20	**專題演講(一)** **資料、檔案、中文佛典文獻學：談數位人文之挑戰** Christine L. Borgman/加州大學洛杉磯分校講座教授 主持人：趙飛鵬/國立臺灣大學中國文學系教授暨佛學中心主任
10:20 ｜ 12:00	**Panel A** **數位人文平台的未來構想** 主持人：項潔/國立臺灣大學
	• 〈從「機構導向資料庫」到「個人導向資料庫」：數位人文下一階段的可能發展〉翁稷安/國立臺灣大學 • 〈DocuSky：個人資料庫的建構與分析平台〉杜協昌/國立臺灣大學 • 〈三國演義人物說話關係之標註與呈現〉王景逸、黃家富/國立臺灣大學 • 〈以 DocuSky 為核心的工具開發與建置〉謝博宇/國立臺灣大學
12:00 ｜ 12:30	**海報發表宣傳** 主持人：蔡炯民/國立臺灣大學
12:30 ｜ 13:15	午　餐　時　間
13:15 ｜ 14:00	海　報　發　表
14:00 ｜ 15:30	**Panel B** **研究的變革：數位分析與文史學科的未來** 主持人：祝平次/國立清華大學
	• 〈數位分析與漢語方言研究〉楊秀芳、葉秋蘭/國立臺灣大學 • 〈一位臺灣古典詩研究者對數位人文的想像和運用〉施懿琳/國立成功大學 • 〈數位人文與歷史研究的互動：理論與實際〉薛化元/國立政治大學
15:30	茶　　敘

15:45	
15:45 | 17:00	**Paper Session 1** **脈絡中的文本：自動化取徑** 主持人：劉吉軒/國立政治大學
	• 〈資料化與地方歷史文獻的數位化、文本挖掘：以《中國地方歷史文獻資料庫》為例〉趙思淵/上海交通大學 • 〈邁向動態擴充的前現代中國文學數位圖書館〉德龍/哈佛大學 • 〈適用於中文史料文本之標記式主題模型分析方法研究〉陳奕安、江子揚、蔡銘峰、薛化元、劉吉軒/國立政治大學 • 〈使用 iAligner 進行語言內文本並列比對〉Tariq Yousef、Chiara Palladino、Gregory Crane/萊比錫大學

第二天12 月 2 日(五)	
8:20 | 8:40	報　　到
8:40 | 9:40	**專題演講(二)** **一個人文研究者對數位人文發展的幾點看法** 陳弱水/國立臺灣大學歷史學系教授暨文學院院長 主持人：王泰升/國立臺灣大學法律學院教授暨出版中心主任
9:40 | 11:10	**Panel C** **數位文本與文學研究** 主持人：蔡瑜/國立臺灣大學
	• 〈數位文本與文學研究〉羅珮瑄/中央研究院 • 〈群體傳記的數位分析：以葉德輝（1864-1927）出版《乾嘉詩壇點將錄》和《東林點將錄》為例〉謝薇娜/中央研究院 • 〈中國詩歌格律之重探與數位化研究：兼談「漢詩格律分析系統」的設計〉林偉盛、鄧賢瑛/國立臺灣大學、莊德明/中央研究院
11:10 | 12:25	**Paper Session 2** **新計算思維：以人文為本** 主持人：劉昭麟/國立政治大學
	• 〈資料計算於數位人文研究意涵的省思〉劉吉軒/國立政治大學 • 〈服務於中國歷史研究的網絡基礎設施〉王宏甦、徐力恆、包弼德/哈佛大學 • 〈數位人文學者的中介角色：連結物理學與戲劇學者〉Miguel Escobar Varela/新加坡國立大學

12:25 | 13:00	午　餐　時　間	
13:00 | 13:45	海　報　發　表	
	凝態館國際會議廳	楊金豹演講廳（104 室）
13:45 | 15:25	**Panel D** **宗教醫療數位平台之建置與應用** 主持人：釋惠敏/法鼓文理學院	**Panel E** **網路哲學** 主持人：蔡偉鼎/東海大學
	• 〈法的療癒資料庫研究與建置〉洪振洲、杜正民、黃舒鈴/法鼓文理學院 • 〈跟踪物質實踐：搜索情境化的亞洲醫藥知識與中國中古宗教文獻為例〉徐源/馬克斯・普郎克科學史研究所 • 〈身體與聖藥：藏密與道教的跨宗教對話〉梅靜軒/法鼓文理學院	• 〈數位年代中對物的重新追問〉洪世謙/國立中山大學 • 〈數位人文的現象學還原：從擬象到檔案〉高國魁/國立政治大學 • 〈數位時代的人文反思：以大數據為線索〉楊士奇/弘光科技大學 • 〈論大數據的知識論條件〉蔡偉鼎/東海大學
15:25 | 15:45	茶　　敘	茶　　敘
15:45 | 17:00	**Paper Session 3** **型式探求：文學與藝術** 主持人：鄭毓瑜/國立臺灣大學	**Paper Session 4** **語料庫語義：社會學應用** 主持人：賀安娟/國立臺灣師範大學
	• 〈概念模式與中國現代廣告社會〉白露/萊斯大學、陳靜/南京大學、鄧柯/北京清華大學 • 〈詩經的量化研究：發掘興體詩的隱藏節奏〉廖學盈/克萊蒙費朗學區 • 〈五代北宋山水畫的數位人文研究（二）：以「漁隱」主題為例〉王平/中國美術學院、鈕亮/中國計量大學、金觀濤/國立政治大學、劉青峰/香港中文大學、毛建	• 〈以語料庫分析取徑探究臺灣新聞中的跨性別：以聯合知識庫為例〉羅盤針、鄭碩、江安淇/國立臺灣大學、曾博揚/國立臺灣師範大學 • 〈大埔之歌：臺灣主流報紙中的「土地徵收」〉王章逸、闕河嘉/國立臺灣大學 • 〈臺灣獨立媒體中的基改食品〉郭柏傑、闕河嘉/國立臺灣大學

	波/中國美術學院	

第三天12 月 3 日(六)		
8:40 | 9:00	報　　到	
9:00 | 10:00	**專題演講(三)** **高等教育中的人文與數位人文** 包弼德/哈佛大學講座教授 主持人：黃寬重/中央研究院歷史語言研究所兼任研究員 暨長庚大學通識教育中心講座教授	
10:00 | 10:20	茶　　敘	
10:20 | 11:50	**Panel F** **學生培育：新世代人才的數位研究能力培育** 主持人：鄭文惠/國立政治大學	
	• 〈臺灣數位人文教育的困難與展望〉祝平次/國立清華大學 • 〈跨越範式：數位人文之人才培育及其多元挑戰〉邱偉雲/湖北經濟學院 • 〈重中之重：文化資產與數位素養〉Duncan Paterson/海德堡大學	
11:50 | 13:00	午　餐　時　間	臺灣數位人文學會年會
13:00 | 13:55	海　報　發　表	
13:55 | 15:25	**Panel G** **文本解讀的擴展、拆解和觀察** 主持人：唐牧群/國立臺灣大學	
	• 〈ADEPT：自動化資料豐富程序〉宋浩/藍星球資訊股份有限公司 • 〈《先秦諸子繫年》之數位設計與呈現〉林農堯、陳胤豪/國立臺灣大學 • 〈《春秋》三傳對讀系統〉趙叡、謝于琳/國立臺灣大學	
15:25 | 15:45	茶　　敘	

15:45 | 17:00	**Paper Session 5** **古今連結：文化資產** 主持人：陳淑君/中央研究院
	• 〈看似無關，實則關連：印度-太平洋南島航行與宗教網絡的數位人文時空地圖集〉卜道/國立政治大學 • 〈用數位工具挖掘 18 世紀德語歷史文獻〉王濤/南京大學 • 〈烏普薩拉大學 1602-1855 時期論文之數位化〉Anna Fredriksson/烏普薩拉大學
17:00 | 17:05	閉　幕　式

Conference Program

Day1: December 1(Thu.)	
8:40 | 9:00	Registration
9:00 | 9:20	Opening Ceremony
9:20 | 10:20	**Keynote Speech 1** **Data, Archives, and Chinese Buddhist Philology:** **Challenges for the Digital Humanities** Christine L. Borgman/Distinguished Professor & Presidential Chair in Information Studies & Director of the Center for Knowledge Infrastructures, University of California, Los Angeles Moderator：Fei-pang Chao/Professor of Department of Chinese Literature & Director of the Center for Buddhist Studies, National Taiwan University
10:20 | 12:00	**Panel A：A Future Framework of Digital Humanities Platform** Moderator：Jieh Hsiang/National Taiwan University
	• *From Institutional-Oriented Databases to Individual-Oriented Databases：The Next Step of Digital Humanities* (Chi-an Weng/National Taiwan University) • *DocuSky: A Platform for Constructing and Analyzing Personal Text Databases* (Hsieh-chang Tu/National Taiwan University) • *Tagging and Displaying Character's Conversational Relationship in the Romance of the Three Kingdoms* (Jia-fu Huang, Jing-yi Wang/National Taiwan University) • *Development and Deployment of Tools Based on DocuSky Platform* (Po-yu Hsieh/National Taiwan University)
12:00 | 12:30	One Minute Madness Moderator：Chiung-min Tsai/National Taiwan University
12:30 | 13:15	Lunch
13:15 | 14:00	Poster Session

	Panel B **Innovations in Research: Digital Platforms and the Future of the Humanities** Moderator：Ping-tzu Chu/National Tsing Hua University
14:00 ｜ 15:30	• *Digital Platforms and Chinese Dialect Studies* (Hsiu-fang Yang, Chui-lan Ye/National Taiwan University) • *The Possible Applications of Digital Humanities in the Study of Taiwanese Classical Poetry* (Yi-lin Shih/National Cheng Kung University) • *The Interaction between Digital Humanities and Historical Research: Theory and Practice* (Hua-yuan Hsueh/National Chengchi University)
15:30 ｜ 15:45	Coffee Break
	Paper Session 1 **Text, Context & Programming** Moderator：Jyi-shane Liu/National Chengchi University
15:45 ｜ 17:00	• *Digitization of Local Historical Archives, Creation of Metadata, and Datamining：The Example of the Chinese Historical Local Archives Database* (Si-yuan Zhao/Shanghai Jiao Tong University) • *Towards a Dynamic, Scalable Digital Library of Pre-modern Chinese* (Donald Sturgeon/Harvard University) • *An Enhanced Topic Model Based on Labeled LDA for Chinese Historical Corpora* (Yi-an Chen, Tzu-yang Chiang, Ming-feng Tsai, Hua-yuan Hsueh & Jyi-shane Liu/National Chengchi University) • *Intra-Language Text Alignment Using iAligner* (Tariq Yousef, Chiara Palladino & Gregory Crane/University of Leipzig)

Day2: December 2(Fri.)	
8:20 ｜ 8:40	Registration
8:40 ｜ 9:40	**Keynote Speech 2** **Remarks on the Future Development of Digital Humanities:** **A Humanities Researcher's Viewpoint** Jo-shui Chen/Distinguished Professor of Department of History & Dean of the Faculty of Arts, National Taiwan University Moderator：Tay-sheng Wang/Professor of College of Law, National Taiwan University & Director of National Taiwan University Press

9:40 | 11:10	**Panel C** **Digital Texts and Literature Studies** Moderator：Yu Tsai/National Taiwan University	
	• *Review on Digital Texts and Literature Studies* (Pei-hsuan Lo/Academia Sinica) • *A Digital Analysis of Group Biographies：A Research on Qian-Jia shi tan dianjiang lu and Donglin dianjiang lu Published by Ye Dehui (1864-1927)* (Severina Balabanova/Academia Sinica) • *Revisiting and Digitalizing of the Metrical Regulations of Chinese Pentasyllabic Poetry：Discussion of the Design of Metrical Regulation Analytic System of Chinese Poetry Project* (Wei-cheng Lin, Hsien-ying Teng/National Taiwan University、Der-ming Juang/Academia Sinica)	
11:10 | 12:25	**Paper Session 2** **Human-centered Computing** Moderator：Chao-lin Liu/National Chengchi University	
	• *Thinking on the Research Implications of Data Computation in Digital Humanities* (Jyi-shane Liu/ National Chengchi University) • *A Cyberinfrastructure for Historical China Studies* (Hong-su Wang, Lik-hang Tsui, Peter Bol/Harvard University) • *The DH Scholar as an Intermediary：Connecting Physics and Theatre Scholarship* (Miguel Escobar Varela/National University of Singapore)	
12:25 | 13:00	Lunch	
13:00 | 13:45	Poster Session	
	International Conference Hall	Yang Jinbao Lecture Hall (R104)
13:45 | 15:25	**Panel D** **Creating a Digital Platform for the Study of Religious Medical Traditions** Moderator：Huimin Bhikshu/Dharma Drum Institute of Liberal Arts	**Panel E** **The Philosophy of the Internet** Moderator：Wei-ding Tsai/Tunghai University
	• *Study and Building of a Dharma-Healing Database* (Jen-jou Hung, Aming Tu, Shu-ling Huang/Dharma Drum Institute of	• *The Question Concerning the Thing in the Digital Epoch* (Shih-chian Hung/National Sun Yat-sen University)

	Liberal Arts) • *Tracking Material Practice: Searching for Situated Knowledge of Asian Drugs in Medieval Chinese Religious Texts* (Michael Stanley-Baker/Max Planck Institute for the History of Science) • *Body and Sacred Medicine：A Dialogue between Tibetan Tantric Buddhism and Daoism* (Ching-hsuan Mei/Dharma Drum Institute of Liberal Arts)	• *The Phenomenological Reductions of Digital Humanity: From Simulacrum to Archive* (Kuo-kuei Kao/National Chengchi University) • *Humanistic Reflection in Digital Era：Based on Big Data Problems* (Shi-chi Yang/Hung Kung University) • *On Epistemological Conditions of Big Data* (Wei-ding Tsai/Tunghai University)
15:25 ｜ 15:45	Coffee Break	Coffee Break
15:45 ｜ 17:00	**Paper Session 3** **Pattern Recognizing:** **Art & Literature** Moderator：Yu-yu Cheng/National Taiwan University	**Paper Session 4** **Corpus Linguistics for Social Science** Moderator：Ann Heylen/National Taiwan Normal University
	• *Concept Modeling and Advertising Chinese Modern Society* (Tani Barlow/Rice University, Jing Chen/Nanjing University, Ke Deng/Tsinghua University) • *A Quantitative Research of the Book of Odes (Shijing)：The Discovery of the Underlying Rhythm in the Incentive Process* (Shueh-ying Liao/Académie de Clermont-Ferrand) • *The Digital Humanities Research of the Landscape Painting of the Five Dynasties and Northern Song Dynasty(2)：A Study of the "Fisherman-Hermit" Theme in Painting* (Ping Wang/China Academy of Art, Liang Niu/China Jiliang University, Guan-tao Jin/National Chengchi University, Qing-feng Liu/The Chinese	• *Applying Corpus Analysis to Explore Transgender in Taiwanese Newspapers* (Pan-chen Lo, Shuo Zheng, An-chi Chiang/National Taiwan University、Bo-yang Ceng/National Taiwan Normal University) • *An Effectiveness of Dapu Incident: A Corpus Content Analysis of Eminent Domain in Taiwan's Mainstream Newspaper* (Chang-yi Wang, Ho-chia Chueh/National Taiwan University) • *Independent News Media Coverage of the Genetically Modified Food in Taiwan* (Bo-jie Guo, Ho-chia Chueh/National Taiwan University)

	University of Hong Kong, Jian-bo Mao/China Academy of Art)	

Day3: December 3(Sat.)		
8:40 | 9:00	Registration	
9:00 | 10:00	**Keynote Speech 3** **The Humanities and the Digital Humanities in Higher Education** Peter K. Bol/Vice Provost for Advances in Learning and the Charles H. Carswell Professor of East Asian Languages and Civilizations, Harvard University Moderator：Kuan-chung Huang/Adjunct Research Fellow, Institute of History and Philology, Academia Sinica & Chair Professor, Department of Medical Humanities and Social Science, Chang Gung University	
10:00 | 10:20	Coffee Break	
10:20 | 11:50	**Panel F** **The Development of Research Students:** **The Cultivation of Digital Humanities Literacy in the New Age** Moderator：Wen-huei Cheng/National Chengchi University	
	• *Difficulties and Prospect of Digital Humanities Education in Taiwan* (Ping-tzu Chu/National Tsing Hua University) • *Cross-Paradigm: Talent Development and Its Multiple Challenges in the Digital Humanity Field* (Wei-yun Chiu/Hubei University of Economics) • *Matter Matters: Cultural Heritage Objects and Digital Literacy* (Duncan Paterson/Heidelberg University)	
11:50 | 13:00	Lunch	Annual Meeting of TADH
13:00 | 13:55	Poster Session	
13:55 | 15:25	**Panel G** **Restructuring and Visualizing Texts** Moderator：Muh-chyun Tang/National Taiwan University	

	• *ADEPT: Automated Data-Enrichment Processing Technologies* (Hao Sung / BluePlanet Data and Information Technology Inc.) • *Digitized Presentation of 'A Chronological Study of the Pre-Qin Philosophers' by Qian Mu* (Nung-yao Lin, Yin-hoe Tan/National Taiwan University) • *A Comparitive Reading System for the Three Commentaries of Chunqiu* (Jui Chao, Yu-lin Hsieh/National Taiwan University)
15:25 ｜ 15:45	Coffee Break
15:45 ｜ 17:00	**Paper Session 5** **Cultural Heritage: Connecting Past & Present** Moderator：Shu-jiun Chen/Academia Sinica
	• *Basic Cultural Elements, Seemingly Unrelated Yet Connected：Spatiotemporal Mapping Early Historical Religious Networks Points in Indo-Pacific Austronesia* (David Blundell/National Chengchi University) • *Text Mining Deutsche Textarchiv Using Digital Tools* (Tao Wang/Nanjing University) • *Dissertations from Uppsala University 1602-1855 at the Internet* (Anna Fredriksson/Uppsala University Library)
17:00 ｜ 17:05	Closing Ceremony

目錄

Contents

專題演講

Keynote Speech

Data, Archives, and Chinese Buddhist Philology： Challenges for the Digital Humanities

Christine L. Borgman*

Abstract

Scholars in the humanities are unaccustomed to viewing their sources of evidence as data or to sharing and releasing those data as part of the publication process. Meanwhile, as more archival materials are digitized and as more cultural information is created in digital form, humanities scholars have turned to computational tools for analysis and interpretation. In turn, as the digital humanities adopt data-intensive methods, they often become subject to open access policies that governments, funding agencies, and publishers impose on science. The transition is an uneasy one for many scholars. This presentation centers on a case study of a Chinese Buddhist philologist whose scholarship employs evidence from material objects and digital resources to study the communication of Buddhist texts ca. 3rd to 5th century C.E. He was an early adopter of CBETA, a digital counterpart of the Taisho edition of the Chinese Buddhist canon. As CBETA expanded in scope and features over the course of a decade, it grew in value as a data source. Because these new tools are integrated into the knowledge infrastructure that serves his community, his scholarly products have become more portable across platforms, increasing the likelihood they will endure. However, these infrastructures remain fragile as they depend on invisible work to curate disparate content and technologies. Chinese texts, both ancient and modern, are particularly difficult to digitize and encode for scholarly analysis. Problems of open access, data management, curation, preservation, and sustainability loom large for the digital humanities. The Chinese scholarship case study is set in the broader context of data in scholarly communication, drawn from the presenter's recent book, *Big Data, Little Data, No Data: Scholarship in the Networked World* (MIT Press, 2015).

* Distinguished Professor & Presidential Chair in Information Studies; Director, Center for Knowledge Infrastructures University of California, Los Angeles, Email: Christine.Borgman@ucla.edu.

資料、檔案、中文佛典文獻學：談數位人文之挑戰

Christine L. Borgman*

摘　要

人文學者重視推論的證據來源，但通常不會將其視為一般的科學性資料，也不須將這些原始資料當成著作發表的一部分，將其予以分享與發布。現在，隨著越來越多的檔案材料數位化，越來越多的文化資訊以數位化形式記錄，研究也開始轉向，人文學者嚐試利用各種資訊工程技術與工具，來進行分析與解釋。但是，當數位人文研究採用資料密集的研究方法時，在資料取得上，卻常常受制於政府、補助機構和出版商的資料開放與近用政策。面對這種研究環境的轉變，往往令許多學者感到不安。本演講主要以一個中文佛典文獻學研究者的個案為例，說明他如何利用實體物件與數位資源作為佐證，探討大約公元 3 至 5 世紀這段時期的佛典文本傳播。這位學者是最早一批採用「CBETA 電子佛典」中《大正新脩大藏經》等數位材料的研究者，隨著十多年來收錄範圍與功能不斷擴展，「CBETA 電子佛典」已經成為重要的研究資料來源。加上一些新工具的整合，更變成學術社群提供服務的知識基礎設施，這位學者的學術產出也更具有跨越平台的機動性，逐漸提高這些基礎設施支撐下去的可能性。然而，這些基礎設施仍是十分脆弱，必須倚賴許多默默付出的工作來支持，才能蒐集分散在各處的內容和技術。而在語言的處理上，不管是古代或是現代使用的中文，在文字的數位化或是提供學術分析的程式，都有很高的困難度。因此，從資料的開放、近用、管理、蒐藏、保存和永續等等問題，我們隱約可以看到數位人文研究的一些挑戰。這個中文佛典文獻學的案例，乃是設定在一個更廣大的脈絡情境，探討資料在學術傳播中的角色與定位，也就是我的近作《從巨量資料、小量資料、到沒有資料：網絡世界的學術研究》（麻省理工學院出版社，2015 年）所探討的主題。

* 美國加州大學洛杉磯分校資訊研究所特聘教授暨主任，知識基礎設施研究中心主任，Email: Christine.Borgman@ucla.edu。

一個人文研究者對數位人文發展的幾點看法

陳弱水*

摘　要

在臺灣，數位人文的發展大約開啟於三十年前，已經取得了重大的成果。臺灣發展數位人文的成果結合臺灣以外的各種數位建置（中國大陸、香港、日本等），為以漢文為核心的人文研究帶來根本性的變化。透過數位資料庫與數位工具，現在的研究者（特別是嫻熟數位工具的年輕研究者）所能掌握的資料量之大，所能研究的課題之多，是二、三十年不可想像的。

數位人文發展到今天，最大的成果是文獻資料的數位化，以及與此相應的檢索力量，本文所想特別考慮的是：接下來可以做什麼？如何超越文獻檢索？超越文獻檢索的數位人文工作其實已有不少，但成果似乎不如文獻資料庫顯著，本文想在這方面有所探討，希望能夠對超越文獻檢索的數位人文發展提出有益的建議。

此外，文獻檢索的成就是無庸置疑的，但目前的文獻資料庫還是可以有所改進，從而成為更便利、功效更大的研究工具，這也是本文想要討論的一點。

* 國立臺灣大學歷史學系特聘教授暨文學院院長，Email: joshuichen@ntu.edu.tw。

Remarks on the Future Development of Digital Humanities：A Humanities Researcher's Viewpoint

Jo-shui Chen*

Abstract

In Taiwan, what we know call digital humanities has been developing for about three decades. Great achievements were made during this period. The accomplishments from Taiwan together with those made in neighboring countries and regions such as China, Hong Kong and Japan have transformed fundamentally the ways researches are done in Sinology and parts of East Asian studies that are centered around premodern Chinese texts (*kanbun* in Japanese). On the basis of these achievements, present researchers, particularly the younger ones who are in general familiar with digital tools, have gained quick and deep access to source materials of an amount that is beyond the imagination of researchers only two decades ago. As the result of their ability to access a huge amount of materials and to penetrate them, the present researchers can also conceive and work on numerous topics that their counterparts in previous times did not dare to think about. In other words, the breadth and depth of humanities researches benefit immensely from the development of digital humanities.

It seems to me that the greatest achievement of digital humanities in Taiwan and neighboring regions is the digitalization of historical and literary texts, mainly in the form of full-text databases, and the accompanying search power. What I wish to consider particularly in this lecture is: What comes next? How do we go beyond the work of database construction and make even greater contributions to the study and culture of humanities? As a matter of fact, many works beyond the level of textual databases are already done. Yet their achievements are not as conspicuous as those databases, and they are probably underappreciated. My lecture aims to explore this aspect of digital humanities in the academic world that uses *kanbun* texts heavily. I wish to present useful suggestions in this regard.

In addition, although the achievement of full-text databases is unquestionable, there are still rooms for improvement as research aids and tools. My lecture will also touch upon this.

* Distinguished Professor of Department of History, Dean of the Faculty of Arts, National Taiwan University. Email: joshuichen@ntu.edu.tw.

The Humanities and the Digital Humanities in Higher Education

Peter K. Bol*

Abstract

Before addressing the digital humanities I should explain what I mean by the "humanities." They are two things at once. First, they are the works of literature and philosophy, art and music and history that have accumulated in diverse civilizations over time. Whether elite or popular, they are integral to the cultures in which people live. They provide us with means of communication and the stuff of shared memories; it grounds our debates over values and the creation of shared identities. Second, they are the disciplines that teach the critical appreciation of those works and inquire into the roles culture plays in our lives.

Here I am specifically concerned with the humanities both in higher education and during the many years of learning after college. Students in higher education today view their choices about what to learn through the lens of their career aspirations. This apparently holds for life-long learning as well -- at least the platforms that distribute massive open online courses (MOOCs) to many millions of learners across the globe promise learning opportunities that will directly benefit the learner's career. To be sure, there are professional careers to be had in literature, art, music, philosophy and history, but outside of teaching careers they are not many. In the US the number of teaching positions in the humanities has been declining, and there are ever fewer students who choose to major in the humanities.

The humanities are for the general education of all, but a means to a career for only few. Higher education is not the foremost producer of cultural goods, but it is our best means of showing students how to think critically about the culture that surrounds them, where they came from, how to express themselves, how to justify their views, and how to debate with others. Students are less inclined to use their 32 courses for studying works of literature and art, of music and philosophy, and gaining historical knowledge in their own language traditions, much less in cultures foreign to them. If

* Charles H. Carswell Professor of East Asian Languages and Civilizations, Vice Provost for Advances in Learning, Harvard University, Email: peter_bol@harvard.edu.

students do not take courses in the humanities, what then? When and how do the humanities reach them?

This brings me to the fast growing field of the digital humanities. What are they and how are they related to the humanities? There is a growing literature that addresses these questions. Digital technologies give unprecedented global reach, the ability to create easily shareable resources, new means of collaboration and new tools for the analysis of data. This is true for all fields of scholarship. However, reach, resources and tools require an ability to make use of a computational infrastructure of hardware and software. Humanists who make full use of digital media have had to learn new skills and new concepts. There are conceptual leaps they have had to make, from flat tables to relational databases, from searching texts to employing regular expressions and topic modeling, from maps to geographic information systems. Digital humanists are those who are able to make use of digital technologies. A few possess advanced skills, but most of us have learned that we can only succeed if we collaborate with those who have technical expertise. Those who do not code depend upon those who do.

There are humanists who make use of digital technologies for research in the humanities: to find trends in large text corpora, to map the spatial distributions of historical data and to uncover the networks in social connections, for example. There are writers, artists, historians, musicians and even philosophers who create works in digital media. And there are digital humanists who see the digital humanities as a field unto itself and call for a separation from the traditional humanities, not unlike how film studies separated itself from literary studies. Their writings make the case for digital humanities as an independent field that conducts research into how society uses digital media to communicate, to create meaning and memory, to debate values and to form identities, all issues of concern in the humanities. Those who advocate the independence of the digital humanities are understandably uneasy about being seen as merely using their skills in service of others, yet they are the ones who are most able to build the tools that will enable humanists to do research and teach in a digital environment.

Digital technologies originate in information technology, statistics and the social sciences. Digital humanities centers, of which there are now over 175, running from labs to well-staffed centers and degree-granting academic departments, have proven necessary to make these technologies useful in the humanities. There are three pressing needs in the humanities that require the help of scholars with advanced computational skills: tool integration, cyberinfrastructure, and online teaching and learning.

On tool integration. We now have numerous databases and utilities for online and offline data visualization, collaborative annotation, mapping, exhibitions, publishing and more. By and large these run independently, so that the output from a database query needs to be loaded into other software programs for further analysis and visualization. We need to tie such utilities together in an online environment. For example, a biographical database that can be queried online would ideally allow users in that same online setting to map their data, generate genealogical charts, visualize and measure social networks, and link these to online publications.

On cyberinfrastructure. A cyberinfrastructure is the system of connections between the layer of base technologies (computation, storage, and communication) and the layer of software, services, instruments, information and social practices applicable to specific projects and disciplines. One might think of the cyberinfrastructure as the network of software, data collections, personnel, best practices and standards independent of specific projects and disciplines, which facilitates the implementation of specific projects on general purpose base technologies.

The humanities and the less quantitative social sciences differ from the sciences in that they are necessarily embedded in language, and this creates challenges inherent in linguistic and literary interpretation, and all the more so for projects that require a cyberinfrastructure using non-alphabetic and non-roman script languages such as Chinese. A cyberinfrastructure in the humanities must take into account the language in which texts were written. It must also deal with two further impediments to communication. First, digital resources such as text databases are dispersed among many institutions and companies. Second, utilities such as dictionaries which could facilitate the online analysis of digital materials are either embedded in a particular resource or exist independently.

The goal of creating a cyberinfrastructure for the humanities cannot be accomplished by combining all searchable text corpora (or image or sound collections) into a single giant repository because the majority of databases are proprietary and access is subscription based. However, with the greater use of Application Programming Interfaces (APIs) it has become possible to create links between online databases and online text programs so that the functionalities of databases devoted to particular topics (dictionaries, places, people, government offices, religious sites) can be brought to bear on searchable text programs. Making it possible for public and proprietary databases to use such APIs to annotate their contents will greatly enhance their usefulness to many different research communities. The same functionality is now being developed for image collections – including art, maps, and scans of texts – that

adopt the IIIF standard. This allows the creation of a cyberinfrastructure to annotate and compare images while recognizing the institutionally dispersed and disparate nature of the digital resources.

On teaching and learning. Currently in the US 7 million students in higher education (1/3 of all) take at least one online course for credit, but two to three times as many post-graduates take online courses (principally MOOCs) without seeking credit. At Harvard 1.5 million unique learners have been actively involved in HarvardX MOOCs over the last three years. Irrespective of whether or not residential learning is more effective, there is an enormous global audience for continued learning. Five further findings from our experience: 80% of HarvardX learners say they are taking a course for life-long learning rather than career advancement (at least twice as many as the norm for MOOCs), 70% are post-graduates, the global median age is 28, 35% identify themselves as teachers, and one-third of Harvard's MOOCs are in the arts and humanities.

This presents humanists with an opportunity. If we think that the humanities matter, then we should consider going where the audience is: online. We may ask ourselves why our responsibility as teachers should end when students leave college. To teach the humanities online is to be a nascent digital humanist, but we need the help of the digital humanities to create the online utilities that will facilitate our students participation, engagement, and community that are integral to humanistic learning that MOOCs currently lack. Students may think of college as the beginning of a career and choose their courses accordingly, but if they want to continue to broaden their understanding of their society, their culture, their environment, and the course of their own lives then the humanities should be ready to help.

高等教育中的人文與數位人文

包弼德*

摘　要

開始談論數位人文前，我應解釋「人文」所指之意。此詞同時有兩個意思。首先，是不同文明中隨時間累積的文學和哲學、藝術與音樂以及歷史作品。無論精英或通俗，其與人類生活的文化不可分割。提供我們交流之法及共享記憶之事物；為我們辯論價值觀與創造共有認同奠定基礎。其次，則是教授上述作品的批判性鑑賞及探討文化在生活中所扮演角色的學科。

在此，我特別關注高等教育及大學後多年學習中的人文學科。現在高等教育的學生會透過他們的事業理想，檢視對學習內容的選擇。這顯然也適用於終身學習——至少提供大規模開放式線上課程（MOOCs）予全球數百萬學習者的平台，就讓他們獲得了直接惠及職業生涯的學習機會。可以肯定，文學、藝術、音樂、哲學及歷史中有專業的職業，但除教學外則不多。在美國，人文學科的教學職位日漸縮減，而選擇主修人文學科的學生也越來越少。

人文學科是所有人的通識教育，但只是少數人獲得職業的途徑。高等教育並非文化產品主要的產地，但卻是我們向學生展示如何批判性思考他們周遭的文化、他們來自何處、如何表達自己、如何證明自己的觀點，及如何與他人辯論時最好的途徑。在三十二門課程中，學生不太願意選擇學習自己語言傳統的文學、藝術、音樂和哲學作品及獲得歷史知識，對陌生文化的意願更低。若學生不選擇人文課程，怎麼辦？人文何時、如何才能接觸他們？

這使我轉向數位人文此一快速成長的領域。何謂數位人文、其與人文科學如何相關？針對上述問題的研究越來越多。數位科技提供了前所未有的全球可及性、創造易於共享資源之能力、協作的新途徑，以及資料分析的新工具。於學術各領域皆是如此。然而，可及性、資源及工具皆需使用軟硬體計算基礎設施的能力。要充分利用數位媒體的人文學者不得不學習新的技能與概念。他們必須在概念上取得跳躍性進展，從平坦表格到關聯式資料庫，從搜尋文字到使用正規表示式和主題模型，從地圖到地理資訊系統。數位人文學者是有能力利用數位科技之人。

* 美國哈佛大學東亞語言暨文明學系 Charles H. Carswell 講座教授，教學發展副教務長，Email: peter_bol@harvard.edu。

少數人擁有先進的技術，但我們中大多數人已認識到，只有和擁有技術專長的人合作才能成功。不會寫代碼的人依賴於會寫的人。

有人文學者在研究人文時利用數位科技：例如在大規模語料庫中尋找趨勢、繪出歷史資料的空間分布，及揭露社會聯繫中的網絡。數位媒體作品的創作者有作家、藝術家、歷史學家、音樂家、甚至哲學家。也有數位人文學者認為數位人文本身就是一個領域，並呼籲將其與傳統人文學科分開，這和電影研究如何從文學研究分離出來相類。他們的著作視數位人文為獨立領域，而於其中研究社會如何使用數位媒體溝通、創造意義和記憶、辯論價值觀和形成認同，即所有人文關注的議題。遭視作僅是使用技能為他人服務，數位人文學科獨立性的主張者對此不滿可以理解，但他們仍是最能為人文學者在數位環境中研究與教學建立工具的人。

數位科技起源於資訊科技、統計和社會科學。數位人文中心目前已有超過一百七十五家，從實驗室到人員齊全的中心和授予學位的學術部門，證明了在人文學科中運用科技之必要。人文學科中有三項迫切需求，必須具備進階計算技能的學者幫助完成：工具整合、網路基礎設施（cyberinfrastructure）以及線上教學與學習。

工具整合。我們現在有大量的資料庫和公用程式，可用於線上與離線資料視覺化、協作式註解、製圖、展覽、出版等等。總體來說，上述所言都是獨立運行的，因此資料庫查詢的輸出資料需要載入其他軟體程式，以進一步分析和視覺化。我們需要在網路環境中串連此類公用程式。例如，可線上查詢的傳記資料庫，理想中能讓使用者在同一網路環境中繪製資料、產生系譜圖表、視覺化和評估社會網絡，以及將其連結至線上出版品。

網路基礎設施。網路基礎設施為基礎技術（運算、儲存和通訊）層級和適用特殊項目與學科之軟體、服務、儀器、資訊及社會實踐層級間的連結系統。可能有人認為，網路基礎設施是獨立於具體項目與學科之軟體、資料收集、人員、最佳實踐及標準的網絡，而能促進基礎技術上具體項目之實施。

人文及量化較少的社會科學與科學的不同在其必然附於語言，而這造就了語言學與文學詮釋中的固有挑戰，對於網路基礎設施中需使用像是中文的非拼音和非羅馬字母語言之項目更是如此。人文學科的網路基礎設施必須考慮到文本所用的語言。並且還須面對另外兩種溝通障礙。首先，像是文本資料庫的數位資源分散於眾多機構和企業間。其次，可促進數位材料之線上分析的公用程式，像是字典，不是附屬於特定資源就是獨立存在。

為人文學科建立網路基礎設施的目標無法藉由將所有可搜尋的語料庫(或圖像、聲音收集）結合為單一巨大儲存庫來達成，因為大多數資料庫是專有的，必須訂閱使用。然而，隨著應用程式設計介面（API）獲得廣泛使用，線上資料庫和線上文本程式之間開始可以建立連結，如此特定主題（字典、地點、人物、政府機關、宗教場所）專門的資料庫功能便能支持可搜尋文本程式。讓公共和專屬資料庫可使用此類 API 來註釋其內容，其對許多不同研究團體的效用將大為提升。於採用 IIIF 標準之圖像收集，包括藝術、地圖和掃描文本上，也正在開發同種功能。承認數位資源之機構性分散與差別性質的同時，也能建立網路基礎設施來註釋並比較圖像。

教學與學習。目前美國有七百萬名高等教育學生（總數的三分之一）在修習至少一門線上課程以取得學分，但有二至三倍的研究生學習線上課程（主要為 MOOCs）並非為了取得學分。過去三年中，哈佛有不重複的一百五十萬名學習者在積極參與 HarvardX MOOCs 課程。無論在家學習是否更有效，繼續學習在全球有廣大的受眾。五項來自我們經驗中的進一步發現：HarvardX 學習者中有 80% 表示修課的目的是終身學習而非職業發展（MOOCs 中至少是平常的兩倍）、有 70% 是研究生、全球平均年齡是 28 歲、35% 表示自己是教師，以及哈佛大學 MOOCs 課程中有三分之一是藝術和人文科學。

這為人文學者帶來了機會。如果我們認為人文有重要意義，那麼就應考慮前往受眾所在之處：線上。我們當自問：為何學生離開大學後，我們做為教師的責任就應結束。進行線上人文教學即是成為新生的數位人文學者，但我們需要數位人文幫助建立線上工具，以促成人文學習中不可或缺的學生參與、投入和社群，而這是 MOOCs 目前缺乏的。學生或將大學作為職業生涯之始並相應選擇課程，但若他們想繼續拓展對其社會、文化、環境和自身生活過程的認識，那麼人文學科就應準備好提供幫助。

Panel A

數位人文平台的未來構想

A Future Framework of Digital Humanities Platform

Panel A

數位人文平台的未來構想

A Future Framework of Digital Humanities Platform

主持人	項潔（國立臺灣大學資訊工程學系特聘教授暨數位人文研究中心主任） Jieh Hsiang (Distinguished Professor of Department of Computer Science and Information Engineering, and Director of the Research Center for Digital Humanities, National Taiwan University)
發表人	翁稷安（國立臺灣大學歷史學研究所博士） Chi-an Weng (Ph.D. of Department of History, National Taiwan University)
題目	從「機構導向資料庫」到「個人導向資料庫」：數位人文下一階段的可能發展 From Institution-Oriented Databases to Individual-Oriented Databases：The Next Step of Digital Humanities
發表人	杜協昌（國立臺灣大學資訊工程系博士後研究員） Hsieh-chang Tu (Postdoctoral Fellow of Department of Computer Science and Information Engineering, National Taiwan University)
題目	DocuSky：個人資料庫的建構與分析平台 DocuSky：A Platform for Constructing and Analyzing Personal Text Databases
發表人	王景逸（國立臺灣大學資訊工程研究所碩士生） Jing-yi Wang (Master Student of Department of Computer Science and Information Engineering, National Taiwan University) 黃家富（國立臺灣大學資訊工程研究所碩士生） Jia-fu Huang (Master Student of Department of Computer Science and Information Engineering, National Taiwan University)
題目	三國演義人物說話關係之標註與呈現 Tagging and Displaying Character's Conversational Relationship in the *Romance of the Three Kingdoms*
發表人	謝博宇（國立臺灣大學數位人文研究中心研究助理） Po-yu Hsieh (Research Assistant of Research Center for Digital Humanities, National Taiwan University)
題目	以 DocuSky 為核心的工具開發與建置 Development and Deployment of Tools Based on DocuSky Platform

從「機構導向資料庫」到「個人導向資料庫」：數位人文下一階段的可能發展

翁稷安*

摘　要

隨著臺灣學界數位人文的蓬勃發展，積累大量的成果，或許已到了去思考下一階段如何發展的時刻。本文試圖由歷史回顧的角度出發，指出臺灣數位人文的生成和數位典藏之間密不可分的關聯，決定了數位人文的根本性格，「機構導向資料庫」就是這樣性格的具體呈現。「機構導向資料庫」是數位人文在學科建置的過程中，不可缺少的環節，然而數位人文還有另一自由、活潑的面向，「個人導向資料庫」的需求正體現了這一側面，希望藉由這樣轉向的提案，能替數位人文未來打造出更完整和豐富的學術風景。

關鍵字：數位人文、數位典藏、機構導向資料庫、個人導向資料庫

* 國立臺灣大學歷史學研究所博士，Email: giant.weng@gmail.com。

From Institution-Oriented Databases to Individual-Oriented Databases: The Next Step of Digital Humanities

Chi-an Weng*

Abstract

Digital humanities in Taiwan has seen rapid growth in the past few years. This paper reviews the hereditary relationship between digital archives and digital humanities in Taiwan, and shows how that has lead the dominance of institution-oriented databases that forms a pronounced characteristic of Taiwan's DH development. From a discipline development perspective, building institution-oriented databases is perhaps a necessary process. From a diversity angle, however, databases that cater to individual needs should be the important and unavoidable next step in enriching the landscape of digital humanities.

Keywords: digital humanities, digital archives, institution-oriented databases, individual-oriented databases

* Ph.D., Department of History, National Taiwan University. Email: giant.weng@gmail.com.

一、 前言：從「數位典藏」到「數位人文」

作為一新興的研究領域，「數位人文」的普及率在幾年之間於臺灣學界有了大幅度的成長，從僅少數人聽聞的嘗試，至 2016 年的今日已成為由科技部所推動的主題計畫之一，吸引了許多研究者的投入，並開始凝聚融合，形成己身獨特的學術性格和有些模糊的學術邊界。顧名思義，「數位人文」強調以「數位」與「人文」的結合，然而從既有的研究成果來看，「人文」一詞所指涉的，泰半以中文、歷史兩學門為主體，雖有人類學、文化研究、圖書資訊學等等學門的加入，仍屬少數。「人文」採取的是比較狹義或古典意義式的界定，擁有悠久量化傳統，並長期使用 SAS、 SPSS、Stata、R 語言、S 語言、LEM 等資訊軟體的「社會科學」或言「社會統計學」的研究取徑，並不包括其中，連帶所及，在歷史學門中視以數位技術進行計量為基礎訓練的經濟史，也很少在這波「數位人文」的浪潮中被提及。

是以，或可大膽推論，數位人文領域所偏重的，不是對「數字」的運算和統計，而是針對文史資料中「文字」或「圖像」，以數位技術去使用和呈現。由「數字」轉向「文字」，重視「圖像」以及圖形化，以「文」、「圖」為思考對象，是數位人文的新意所在，也呼應數個不同的學術趨勢，諸如歷史學近年對圖像或器物史料的重視，[1]又或者時下資訊領域熱門的「大數據」（big data）的分析。[2]倘若從實作的角度觀察，多數研究大抵依循著先建立一資料庫，再就庫上資料內容進行標注、分析、呈現的步驟，也因此除了那些熱門學術趨勢外，臺灣數位人文發展的淵源，或仍需溯及至「數位典藏」的概念，一定程度上甚至是數位典藏的某種延伸和進階。

以臺灣大學數位典藏團隊為例（即後今日臺灣大學數位人文研究中心前身），在本世紀初討論數典典藏時，反覆申說如下的觀念：

> 數位典藏是透過數位化技術，將珍貴的文化資產轉化成文字、影像、聲音或視訊等數位物件，數位典藏內容包括許多第一手資料……是對未來相關研究極為珍貴的研究素材及研究紀錄的來源。……數位資料便於檢索的與查閱的特性，使更多的研究者與學生樂意且容易取得第一手資料；更進一步來說，圖書館用數位方式保存資料，集合大量不同來源、類型、領域等豐富多樣的數位研究素材，藉由資訊技術所發展的各種工具，激發研究人員發現新的研究面向、產生新的研究方法，這將有助於提

[1] 相關研究甚多，並以取得豐碩的成果，不一一舉列，可參考黃克武主編，《畫中有話：近代中國的視覺表述與文化構圖》（臺北：中研院近史所，2003），以及 Peter Burke 著，楊豫譯，《圖像證史》（北京：北京大學，2008）。

[2] 「大數據」概念最完整而扼要的說明，見 Viktor Mayer-Schonberger 等著，林俊宏譯，《大數據》（臺北：天下文化，2013）。史學研究者對大數據硬用的想像，可參考黃銘崇，〈「巨量資料」概念下的史料收集與歷史書寫〉，「歷史學柑仔店」網站，2014.6.20（2016 年 10 月 8 日檢索）。

高學術研究從量變進而產生質變的可能性，造成新的學術研究典範。[3]

此外，洪一梅在回顧中央研究院史語所於數位典藏的成果時，從歷時性的角度，強調了數位典藏承先啟後的角色定位，從 1990 年代開始，依循著「學術研究活動的交互變動模式」，經由使用者導向的系統開發、人文學術研究環境的再造二階段的演進。[4]其中最關鍵的，不僅在於資料的保存，而是研究的應用或研究者的需求。綜合兩者，數位典藏並非只是單純將文字、圖像或器物的資料或藏品進行數位化的轉換，藉由電子全文化、詮釋資料的建置等步驟，給予現實之中被固定型式所綁定、限制的資料，於虛擬世界中的解放，帶來了研究上全新的想像。是以，數位典藏概念最主要的核心，是從傳統「靜態」的堆置、調閱，轉向網路時代「動態」的搜尋、應用。

由「數位典藏」為基礎變成「數位人文」，所依憑著邏輯，是依照文史學者進行研究的慣用順序，用最簡化的方式描繪，即在各個不同的圖書館或資料收藏地中尋找資料，再分門別類整理所得的資料，進行閱讀和筆記，然後開始自身的論述。數位典藏希望能替代圖書館角色，並逐漸延伸，由資料的集合成為研究環境的打造，這原本也是實體圖書館應有的功能之一。這延伸轉變的過程，也決定了數位人文的方式和規範，諸如機構典藏建置的思維，成為主導數位人文的重要因素之一。[5]也因此以各大教學、研究機構為主體，建置的大型資料庫成為早期數位人文研究前沿的主導者，「機構導向資料庫」是數位人文能取得豐富成果的關鍵，並在打造資料庫的基礎上，建立與不同學界、學者之間的合作模式，在可預見的未來，應該還是推動該領域持續前進的力量。然而，隨著技術的不斷變遷，以及人們對數位人文認識和接受的深入，「機構導向資料庫」模式也許已到了反思和改變的時刻；即使這種生成至「數位典藏」的模式，在實際運作上還將發揮很長時間的影響，但也應該有著不同的調整和偏重。

---

[3] 項潔、洪筱盈，〈大學圖書館數位保存與館藏發展策略：以臺大圖書館為例〉，《大學圖書館》，10 卷 2 期（2007 年 9 月），頁 11。

[4] 洪一梅，〈人文學術研究的數位新時代：史語所的思維與作為〉，《古今論衡》，第 20 期（2009 年 12 月），頁 133-154。

[5] 關於機構典藏的理念，可見項潔、洪筱盈，〈臺灣機構典藏發展芻論〉，《教育資料與圖書館學》，43 卷 2 期（2005 年 12 月），，頁 173-189。這篇論文是 2005 年由教育部委託臺灣大學圖書館的「建置臺灣學術研究資源中心（Taiwan Academic Research Electronic Library）運作架構、機制與執行執略計畫」的成果，雖還未有後來那麼明確的概念，但指出了將機構藏品數位化的好處和必要，並注意到「全文化」所能帶動的研究潛能。不只機構典藏，數位典藏對當前數位人文的影響還有其他的面向，加價運用是其中之一，這涉及了商業化的授權和營利，也因此除了學術、研究機構，大小不一的資料庫廠商也在其中扮演著重要的角色。隨著數位人文領域的成長，商業和學術之間如何取得平衡，諸如和廠商的合作、技術的轉型營利，以及文本數位化後的版權問題，都將成為令機構或個人使用者日後不得不面對的問題，該問題涉及層面甚廣，又部分將在下文討論，但並非本文關注的重點，需另為文討論。

二、 個人導向資料庫

「機構導向資料庫」面臨的最大挑戰，從一開始便存在於形成的探索中，不論何種形式的資料庫，在建構過程中，都強調以使用者（即研究者）的需求和習慣為參考的核心。隨著人們數位人文接受度的提高，體會到由各機構所開發大型資料庫在研究上的優勢，實際運用於研究後，很快就會延伸出新的需求：除了對大數據或巨量資料進行宏觀的分析，以及藉著檢索功能提取出有關資料外，更希望能建立起和自己研究連結，建立起個別或個人研究所需的中小型資料庫。這種「個人導向資料庫」的客製化需求，必將伴隨著數位人文概念的普及，呼聲越來越高。

此趨勢背裡涉及的，是文史學科實際操作的慣習，甚或方法論層次問題。對多數文史學者而言，在實作層次上，除去特定類別外，「大數據」難稱其為主流，甚或不是最直覺的思考或研究方式。若依循「大數據」最嚴苛的學術定義，文史學者所需的資料量，亦規模遠遠不足。文史之學是由文史學者以單數「人」的身份，去理解過去單數「人」或其集結，是以「人」為尺度，去構思和觀察進而開展推論，要達到「大數據」探勘窮盡人力所難以觀察的資料量，並勾勒「人」所不自覺的趨勢和規律，雖非不可能，但很難以視為人人可達成的目標。數位人文領域對「大數據」概念的宣稱，往往是常識性的理解和調整，對多數文史研究者而言，他們仍是在「大數據」時代，進行「小數據」研究的人們。

事實上為了呼應「個人導向資料庫」的需求，「機構導向資料庫」也逐漸發展成新的合作模式，當有此需求的使用者，苦於無法跨越技術門檻，和機構的合作成為常見的選項，發展出「機構-個人」或「機構-機構」的合作結盟。此合作結盟的出現，象徵著「機構導向資料庫」的自我演化和調整。然而，這樣的調整是否能呼應日益增加的個人客製化的需求呢？答案可能不太樂觀，涉及許多主客觀因素，最現實的，並非所有個人都有和專業機構互動的管道，即便在不動用人際網絡的前提下，直接尋找相關商業廠商，所需花費的金錢和溝通的時間，使多數使用者聞之卻步。

這並不表示沒有資源的文史研究者只能自外於數位人文的潮流外，倘若使用較廣寬的定義，僅視「數位人文」為某種研究的方法和工具，而非得一定要具有什麼明確規範的學術範疇；在僅需要工具或特定功能的協助，而不涉及領域前景的前提下，轉而使用許多非針對數位人文開發的軟體，利用單一軟體或不同軟體交相串聯下，即能獲得大致的滿足，土法煉鋼地勉強建立起屬於個別或個人需求的資料庫。

類似的軟體很多，此處無法一一列舉，僅試舉一套流程為例，說明這樣的可能性存在。一個資料庫的建立，可暫時粗歸納為下面幾個環節：（一）資料的輸入：如文字資

料的全文化或圖像資料的掃描。（二）資料的標注：即標示出全文資料中特定的詞彙，或給予特定的標籤。（三）資料的呈現：前兩步驟所處理好的資料給予圖表統計，或將結果給予視覺化的呈現。輸入、標注和呈現三者，包含了一資料庫最陽春的運作邏輯和原則，而只要能塞入流程之中，發揮一定功能的軟體，經由排列組合，得到可供說明和分析成果，即可視作運用數位方法於人文研究之中。試想下面情況，一位無任何資源的文史研究生，他立即的目標是完成論文，取得學業，考量到時限和可操作性，他所關心的議題和擁有的資料量多半僅有一定限度，除非十分幸運身處的科系有相關的機會，即便對數位人位十分嚮往，大概也無緣與「機構導向資料庫」有合作的機會，只能暫時旁觀，給予贊嘆和喝彩。然而，如果他能把握這三步驟的概念，在確定關注的問題意識，發揮想像力，即能利用週邊間接資源完成「個人導向資料庫」。

在輸入方面，圖像的部分難度不高，多數中高階影印機已具備影像掃描的可能，若數量龐大，在預算範圍之內，還可以請由外面的影印店代勞。若是物質史料，如雕像或器物等，以攝影或錄影的方式處理，亦不困難，所需的僅是時間。全文化的部分，目前許多 OCR 軟體，都可有不錯的成果，以 ABBYY FineReader 中的 ABBYY PDF Transformer+功能，在 PDF 轉成文字檔的效果十分成功，對中文印刷體的判讀十分準確，使用者僅需再行校對即可。標記的部分，只有上網尋找 annotating tool 就有很多可以使用的工具，大部分的筆記軟體也能發揮類似的作用，使用者須明白自己研究的限制，並非「大數據」的宏觀，要去發現未經察覺的史料線索，所以無需鉅細靡遺地標出所有的專有名詞，而是依自己的提問，把有用、需要觀察的詞彙標記出來。在那麼有限度的功能需求下，類似 Evernote 軟體所提供的 tagging 的功能，即能滿足；可以先運用手動的方式將已全文化的資料切割成適合的單位，再將每一單位存成個別筆記，針對部分關鍵進行標記，賦與標籤，最後運用這些標籤進行簡單的計數工作。在呈現上，則可以運用 Excel 強大的統計繪製功能，將計算的成果轉化成圖表；此外，如 google earth 和 Pajak 等軟體，也能快速的畫出簡易的地圖和人際網絡。

上述方式對比於「機構導向資料庫」所能提供的強大可能，彷彿是部東拼西湊的「拼裝車」，但只要問題意識明確，規模合宜，在有限的資源內，不用數位技術人員一旁協助，還是能克難而有效地將資訊技術和人文研究結合，打造屬於自己需要的「個人導向資料庫」，符合合廣義的數位人文界定。這樣的流程或許陽春，但卻不違反實務的經驗，從數位端的角度，這是建立巨型資料庫都必經的流程，不同者僅在於缺乏資訊科技人員的協助和互動，很多功能只能被捨去，並耗費個人而非集體的時間和精力，將建置的個人成本拉高。從文史研究的角度來看，這符合了在完全無電腦引進之前的手工作法，並獲得資訊技術這樣優秀「助理」的協助整理。無論如何，都滿足了以使用者為出發的預設。如同「拼裝車」的譬喻，也許一台拼裝車效能有限，還有污染、維修等隱藏成本的

耗費，但能以比一般車輛低廉的售價，達到駕車移動的目的，就可被視為大眾運輸的一環。

一旦我們將「數位人文」視為一種廣義的方法論，內化在文史研究的過程，那麼更個人、規模有限的研究需求，將變成資料庫發展的重要轉向，「個人導向資料庫」不見得必然取代「機構導向資料庫」，或發生如「『機構導向資料庫』的終結及其最後一人」的預言場景，但所帶來的轉變在所難免，大膽推論，或有機會成為無論廣狹義數位人文下一階段的基調，「機構導向資料庫」或數位人文的倡導者需要針對這樣的轉折做出回應和調整，唯有如此，才有可能帶動整體研究在方法或視角上的變化。[6]

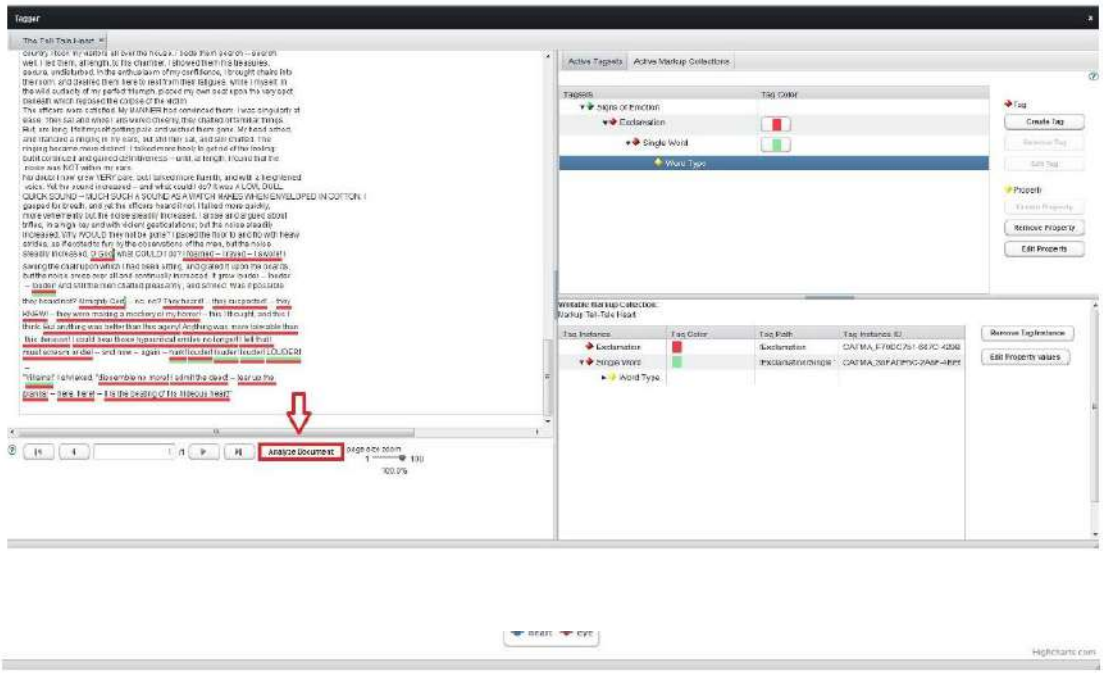
三、 面向個人需求的「機構導向資料庫」

「個人導向資料庫」的出現，象徵著這幾年數位人文在發展上的成功，吸引了更多的關注，讓更多人想要投身其中，同時也符合著數位人文一直以來以研究者需求為主要構思的原則。和「機構導向資料庫」最大的差異，在於數位典藏是種動態的運用，兼有「收藏」和「應用」兩個面向，個人導向的數位人文也許在收藏的「量」上無法媲美機構，但在應用的「質」上有著相同的需求，「機構導向資料庫」的主事者，除了要繼續發揮在資料量的優勢，讓更多的資料被數位化外，也必須回應個人在功能上的需求，提供未綁定典藏的純粹功能，提供獨立的提供工具。這也是為何 MARKU 這個仍持續開發中，作為計畫一部分的工具，能獲得人文學界那麼大的回響和重視，因為它讓文史學者看到自建資料庫的可能，即便只是很有限的希望火光。

作為數位人文當前最主要的推動者，可能在提供資料庫外，也需試著提供純功能的使用，即給予使用者隨意支配、具功能的「庫房」，或給使用者搭建簡易庫房的工具。類似的開發亦有許多，個別工具外，近來亦出現將這些功能整合於一的系統，此處僅舉德國漢堡大學語言、文學和媒體系（University of Hamburg, Department of Languages, Literature and Media）所開發的 CATMA（Computer Aided Textual Markup & Analysis）為例，該工具以文本分析的數位工具（Textual Analysis Computing Tools）為出發點，將 Tagger 和 Analyzer 的功能作為核心，製作出一套符合文史學者研究的流程，使用者可以上傳所要分析的文析，給多標注和分類，待完成之後即能畫出簡易的圖表工具，如下

[6] 需要聲明的，對如何界定「數位人文」的爭議，在西方近年也有很大的爭論，筆者本文所採取這樣個人、自由的論法，其實是受到批判的，見 M. K. Gold ed., Debates in the Digital Humanities. Minneapolis: University of Minnesota Press, 2012. 如本文中將反覆提及的，我不認為對數位人文的討論是非 A 則 B 的選擇題，一旦如此被抹去了該學門重要的學術特徵和性格，請見下文的申論。

圖所示。



這確實達成了部分使用者自建資料庫的需求，然而或許還有更本質的層次值得再深思，類似的建構工具承繼著「機構導向資料庫」很重要的特色，那就是自成一套首尾相銜的完整系統，即便可以輸出成不同存取的格式，核心的思考仍是封閉式的，希望藉由單一系統完成一站式的（one-stop services）開發，如何讓功能能不斷擴充，更形完備，打造一縱向——從頭至尾完成研究作業的宏大系統。從開發的角度來講，這樣的思考是必然的，亦無錯誤，「機構導向資料庫」決定性的優勢和賣點便在於「規模」，不是在典藏的「量」，或功能種類的「質」，務求無所不包的全能資料庫或系統。

必須再次重申，這樣的開發思維有其必要和正確，尤其在發展學術規範，不斷推進學術疆界的目的下，不應消失；但也不應是唯一的聲音。在筆者看來，數位人文從一開始，哪怕是在最原初的「動態」式數位典藏時，最核心或根本的性格，不是承襲自圖書館、博物館的收藏傳統，亦非近現代文史研究學門數百年所留下研究需求，而是來自數位端那自由、開放的活潑氣息，這才是數位人文學門最引人入勝之處。不在具體提供了什麼特別的自動化功能，而在於抽象宣示了：在數位時代來臨，事事物物都重新獲得新風貌之際，給人嚴謹印象的文史學門也能順應潮流，產生或重塑出新的可能。數位人文替傳統文史學界注入了一股全新的活力和空氣，賦與了呼應著新時代的契機。這種自由、創新的精神，成就了數位人文的風行，卻也成為數位人文領域在學科建制時最大挑戰，一方面要建立某種知識論或方法論的規範，另一方面則必須保持流動不居、挑戰既有邊界的想像力，兩股追求創造了數位人文內部的張力；「機構導向資料庫」和「個人導向

資料庫」，正是這兩股力量的體現。

前文所提及的「拼裝車」，雖然是不得不然的克難手段，但背後所顯現的，就是數位人文的開放精神，縱向、一站式的資料庫或系統是推動學科成型的安定力量，但往往是封閉的，造成各自為政、山頭林立的學術地貌。下一階段的數位人文發展或許就是要在這樣穩固的基礎上，創造橫向、節點串聯式的平台，以更開放的方式，納入更多人力或物力的資源。所謂的「縱向」或「橫向」，可視為兩種不同的「整合」，「縱向」的整合，是一種建立在巨型資料庫前提下，系統搭配指使用者研究的過程，整合資料庫內部的資料及功能。「橫向」的整合，則更偏向外部，整合的單位主體從資料庫變成了個人，在一定的標準規範下，最大化資料或工具的可擴充性和可連結性，打造能以接合各式資料或工具的平台。又或更開放的，只要有對接的數位窗口或孔道，直接「點對點」生成個人對個人的結合。「橫向」整合保有著拼裝的精神，但不用再生硬的硬湊各種不同的軟體，而是在一定規範和框架下，自由組合原初設計時即用來相互結合的原件。這看似有些抽象、過於理想的概念，並非空談，臺灣大學數位人文研究中心近年所推動 Docusky 既是對橫向整合的嘗試。

「機構導向資料庫」或「縱向」整合宛如蓋起一座座的大樓，「個人導向資料庫」或「橫向」整合則是由點串成線，進而交織成廣闊的平面，兩者兼有，才能建構出既廣且深的學術環境；兩者可以並行不悖，且理應並存。「數位人文」最理想的發展願景，就是必須在不斷建立學術規範的同時，又要同時有人去打破、挑戰規範。

「機構導向資料庫」除了要面對個人需求外，同時也要持續基礎工程，首先要發揮原本在「量」上的優勢，推動更多資料的數位化，唯有數位文本持續成長，以及作為基石的數位典藏不斷擴增，數位人文才能不斷成長；有一些工具的試驗和開發，也必須使用大量資料才能完成或發揮效果。只有在「量」上的發展，「大數據」於數位人文的運用，才能被落實，並啟發無數個人的使用者。其次，則是現實面的工作，一方面是建立具體可行的學術操作原則，譬如對數位文本的徵引，目前尚未有一致而統一的格式，結果多半還是依循紙本的規則，而無考慮系統更新、增添等因素，這對打造一虛擬研究環境，無論縱向或橫向都是十分不利的。又或者像是 metadata、全文的除錯或檢查機制，圖像、實物資料庫的搜尋或比對，圖文如何於資料庫中並存並用，甚或連橫向連結的基準，都還是必須仰賴機構才能推動。另一方面，機構要扮演起讓現實中個人連結的窗口，成為資訊技術人才和文史研究學員之間的媒合者，在促成合作外，並能找出對雙方學術成就都能有所肯定的衡量方式，在各種大型的計畫和標案之外，照顧到個人的需求，讓橫向連結的精神，由虛擬延伸到實存的世界。

四、 結語

Here's to the crazy one, the misfit, the rebel, the troublemaker, the round peg in the square hole, the one who saw things differently.

He's not fond of rules. And he had no respect for the status quo.

You can quote him, disagree with him, glorify or vilify him. About the only thing you can't do is ignore him. Because he changed things. He pushed the human race forward.

And while some may have seen him as the crazy one, we saw genius. Because the man who was crazy enough to think he could change the world, was the one who did.

這是 1997 年 Steve Jobs 在蘋果電腦廣告中的著名宣言，雖為廣告詞，卻成功捕捉了某種數位時代的探險精神，數位人文即便在學院之中，也唯有秉持這種精神才能不斷突破，推進研究的前沿。「機構導向資料庫」是在原本制度內尋求創新的可能，為發展立下根基，「個人導向資料庫」則希望能在那樣的基礎之，以更去中心、個人化的方式，開拓更多的可能。唯有召喚各式各樣不同的人投入，他可以是文史學者、可以是社會學者、可以資訊學者，甚或什麼身分都不是，從各自的角度和資源，去看待「數位人文」這新方法甚或新學科，「數位人文」所描繪的那些理想和願景才有實現的一天。

DocuSky：個人文字資料庫的建構與分析平台

杜協昌*

摘　要

隨著數位人文領域的開展，學術或大型機構所開發的傳統典藏資料庫，已經不再能滿足研究者的需求。這些典藏庫雖然能提供品質良好的文本，但內容的修訂擴增速度卻相當緩慢。此外，除了最基本的檢索功能，這些典藏系統幾乎都沒有提供進階的分析工具，能夠對使用者感興趣的文件進行彙整與統計。這些缺陷會阻礙使用者採行數位人文方法來對文本進行研究。

另一方面，文史研究者通常都在自己的電腦上存有感興趣的文本。若能提供一個系統，讓使用者能夠建構個人的文字資料庫 (text databases，簡稱文字庫)，支援全文檢索、後分類與詞頻分析等功能，並提供數位工具讓使用者能夠對建構的文字庫進行統計分析，將能從查找資料、內文比對、以及字詞相關統計等不同面向，增加使用者對文本內容的掌握。

DocuSky 就是在這些動機下所開發的系統。它允許使用者上傳全文、詮釋資料、以及經過標記的文本來建構個人文字庫。一旦文字庫建立完成，使用者就可以利用多種開放的數位工具來對它們進行存取與分析。在系統設計上，DocuSky 主張文本與工具必須分離、使用者介面必須可在瀏覽器上操作，並提供許多小元件來幫助工具開發者降低資料存取的障礙。我們認為，DocuSky 有潛力形成數位人文研究的平台，讓資訊工作人員與文史研究者在它的系統架構下共同合作，從而加速推展數位人文的應用。

關鍵字：DocuSky、個人文字資料庫、系統架構、後分類、詞頻分析、文本統計分析、數位人文研究平台

* 國立臺灣大學資訊工程系博士後研究員，Email: hsieh.chang@gmail.com。

DocuSky：A Platform for Constructing and Analyzing Personal Text Databases

Hsieh-chang Tu*

Abstract

The main, and usually the only, purpose of most traditional text databases (textbases for short) is to provide good contents with a retrieval system that helps one find desirable documents. This is often not sufficient for digital humanists who want to apply digital tools to explore properties in a fairly large subset of the textbases. On the other hand, humanists usually have interesting texts stored in their local disks. Finding and analyzing the statistical behavior of a dynamic subset of text files can be hard with the available search functions provided by personal computers. It is desirable to have a system that allows one to build textbases that support common retrieval functions and many other text analysis tools.

In this paper, I propose DocuSky that allows one to build personal textbases. This system supports fulltext retrieval, post-classification over a search result, as well as analysis on tagged terms. Fulltext retrieval is common for searching desirable documents in a database. For any search result, post-classification groups its metadata and shows the resulting distribution. Analysis on tagged terms, on the other hand, returns a list of tagged terms occurring in that search result. They are the three major functions offered by the well-known THDL (Taiwan History Digital Library) system. In addition to these elementary functions, it also provides various tools to help users analyze the contents in a textbase.

The advance of digital humanities requires closely cooperation of computer scientists/engineers and digital humanists. The system architecture of DocuSky encourages users and tool developers to re-think about the roles of text contents and analysis tools. In order to reduce the effort of tool development, DocuSky designs a

* Postdoctoral Fellow, Department of Computer Science and Information Engineering, National Taiwan University. Email: hsieh.chang@gmail.com.

set of APIs and provides some widgets to ease the access of personal textbases. Although DocuSky is still in its early development stage, it shows strong potential to become a platform that helps people work together.

Keywords**:** DocuSky, personal text databases, system architecture, post-classification, analysis on tagged terms, text analysis, platform for digital humanities

一、動機

在數位典藏 (digital archives) 盛行的年代，許多機構就曾開發有中大型的文字資料庫 (text databases，以下簡稱文字庫)。這些典藏機構所開發的文字庫，一般都提供品質良好的文本內容和詮釋資料 (metadata)，也通常會提供目錄或檢索機制來幫助使用者查找資料。即便有了這些品質良好的典藏庫，我認為在數位人文 (digital humanities) 的時代裡，並不該以此為滿足；還必須提供一個機制，讓研究者能夠建構自己的個人文字庫。

有兩項重要的理由。首先，為了提供優良的文本品質，機構必須投入相當多的時間與資源，來對文本內容與詮釋資料進行清理與校正。這使得在實務上，典藏機構的文字庫不僅在數量上有所限制，在內容更新上也相當緩慢。此外，雖然典藏機構所開發的文字庫一般具有不錯的品質，但它們多半都只提供檢索的功能，缺乏進階的標註與分析工具。這使得研究者僅能在文字庫中查找資料，無法對文本內容進行較深入的統計分析。在這些限制下，學者通常並不能僅依賴典藏庫來進行研究，而必須另尋其他文本資源來佐助。

另一項理由，則是換個方向，從人文研究者的角度來思考。人文學者為了研究需求，通常都會在個人電腦中儲存屬於自己的文本 (texts)，以方便隨時查找與參照。然而，當這些文本的大小和數量逐漸增加，現今個人電腦所提供的檔案儲存格式，以及資料搜尋的工具就顯得過於簡單，對研究者在進一步查找相關文件、或者對文本內容進行更細緻的統計分析構成障礙。如果能夠提供一個系統，讓使用者可上載文本建構成個人文字庫，並提供類似 THDL 系統[1]全文搜尋 (fulltext search)、後分類 (post-classification)[2] 以及標記詞彙頻率分析 (analysis on tagged terms，以下簡稱詞頻分析)[3] 的功能，就可在相當程度上，幫助研究者對這些文本內容擁有更好的掌控能力。

DocuSky 就是在上述動機下所設計開發的系統。注意到數位人文研究的開展，需要資訊與人文跨領域的合作（具資訊工程背景的技術人員負責工具的開發與維護，而人文學者則提供文本內容，並利用開發出來的工具對這些文本進行分析探討），我在系統架構的設計上納入工具與文本相互分離的原則，期盼這個系統能夠成為一個平台 (platform)，讓工具開發者與文本研究者都能在平台上發揮自己的專長，從而對數位人文的研究產生影響與貢獻。

---

[1] THDL (Taiwan History Digital Library, 網址為 http:// thdl.ntu.edu.tw) 為臺灣歷史數位圖書館的系統名稱。

[2] 對任意的搜尋結果，後分類可列出其詮釋資料的統計資訊。

[3] 對任意的搜尋結果，標記詞彙頻率分析可計算該結果的文本中，特定詞彙（一般就是經過標記後的詞彙）的出現統計。這些統計通常包含詞彙出現的文件數量，以及出現的總次數。

欲開發讓使用者自建文字庫的系統，必然得具備資料上載儲存、和數位典藏庫的相關技術。雖然現今已經有許多系統提供讓使用者免費（或以相對低廉的費用）將資料上載儲存的雲端服務 (cloud services)，而且數位典藏資料庫的開發技術也已經相當成熟，但在伺服器端允許使用者建構個人文字庫，仍是一項嶄新的嘗試與挑戰。我將在第二節介紹使用者在 DocuSky 建構文字庫的流程。DocuSky 所採取的建庫流程，與一般開發典藏資料庫的步驟不甚相同。尤其是，典藏的資料通常會採用機構所維護的一套共通文本與詮釋資料格式，但個別的使用者資料儲存格式卻不盡相同。DocuSky 解決這個問題的方式，是先要求使用者利用系統所提供的資料彙整或轉換工具，將文本與詮釋資料打包成系統所能辨識的格式；接著，再請使用者利用另一項工具將這份彙整後的檔案上載建庫。接著，我將在第三節討論 DocuSky 的系統架構與運作方式。尤其是，我將從軟體分工和工具文本分離的角度，說明這個系統的一些設計原則、以及可能帶來的影響。第四節介紹 DocuSky 的工具集。DocuSky 仰賴不同的工具來實作各種有趣的功能；這些工具能讓使用者對文本進行不同程度的操作與應用。論文的最後一節討論 DocuSky 近期的工作，以及我對這個系統的未來展望。

二、在 DocuSky 建構文字庫的流程

欲建構一份文字庫，使用者必須提供文本（texts）和相關的詮釋資料（metadata，又譯為後設資料或元資料，是描述這份文本的一些額外資訊）。建構文字庫所需面對的第一個問題，就是彙整這些文本與詮釋資料[4]，並將其存放系統所設計的特定儲存結構中。對於機構典藏資料庫而言，這通常並不構成困擾，因為每個機構對於自身典藏的文本與詮釋資料，都有維護一套適合機構本身使用的共通格式。另一方面，個人或不同機構存放資料的方式，就經常有相當高的變異性：這些資料可能被人們以特定的格式儲存在文字檔中，[5]也可能被使用者以不同類型的檔案分別存放。[6]要在 DocuSky 伺服器上，實作一份通用的程式來彙整各種類型的文本和詮釋資料格式，實務上顯然並不明智。

DocuSky 解決這項問題的方式，是將資料彙整的工作，交給使用者來負責處理： 1) 首先，要求使用者利用系統提供的工具，將欲建庫的文本和詮釋資料包裝成系統能夠識別的 XML 檔案， 2) 然後，使用者就可將這份 XML 檔案上載到 DocuSky 來建庫。

[4] 一般說來，需彙整的資料整除了文本和詮釋資料，還可能包含文本內容的標記資訊 (tagging information)。

[5] 例如 Kanseki Repository（網址為 https://www.kanripo.org）提供了大量的免費文本可供線上觀看或下載。下載後的每份文本，都被儲存成「同時包含詮釋資料與文本內容」的單一文字檔。

[6] 例如，使用者可能將文本儲存於一系列的文字檔（對這些文字檔的檔名進行編碼），而詮釋資料則彙整存放在 Excel 表單中（利用文字檔的檔名對該文本的內容進行連結）。

檔案上載後，系統可以選擇在適當的時機（例如，上載當刻或離線排程），從上載的內容中自動提取文本和詮釋資料的資訊、將這些資訊以適當的資料結構儲存起來、並對文本進行全文索引 (fulltext indexing)。當然，由於使用者資料格式的變異性高，有時 DocuSky 所提供的彙整工具並不足夠。這時，使用者可以調整手邊的資料格式，或者提出請求 (request)，尋求具資訊背景的技術人員開發新的彙整工具。一旦新的彙整工具被開發出來，它就可以被放到平台上，讓其他擁有類似格式資料的研究者使用。

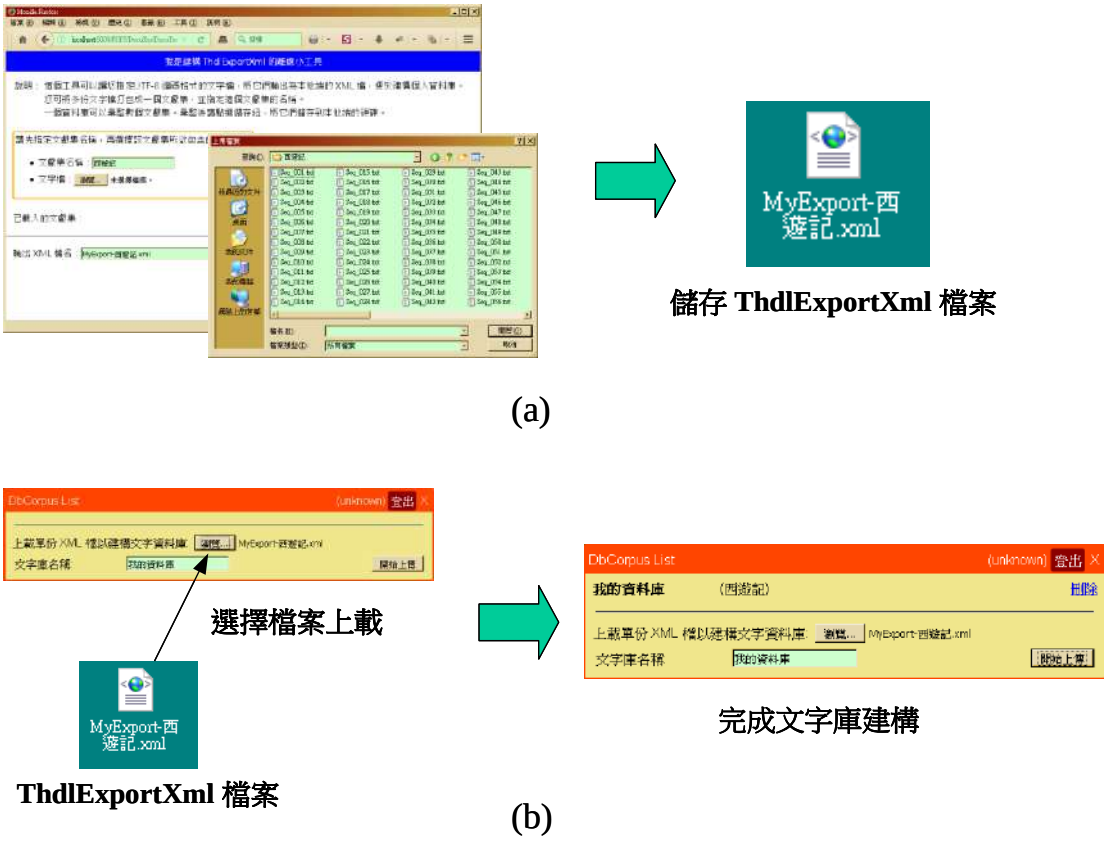
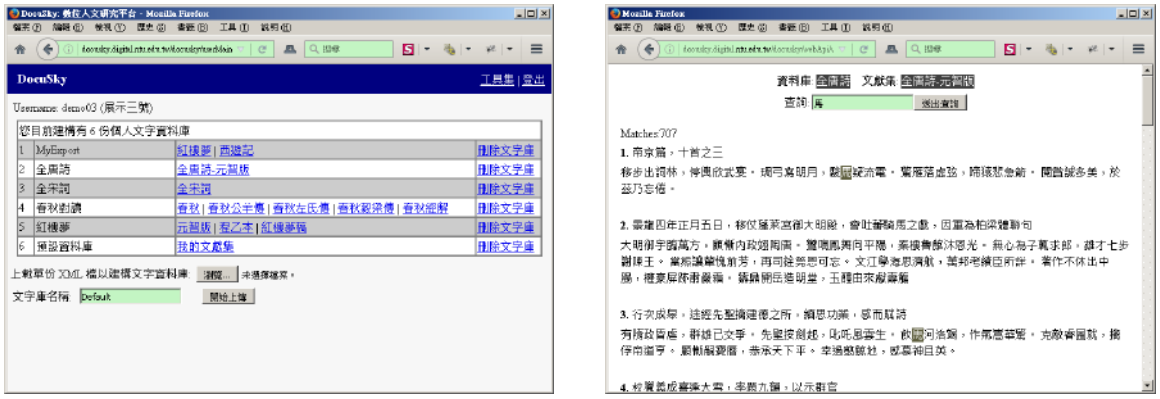


圖 1. 在 DocuSky 建構個人文字庫的流程被分為兩個步驟：首先，使用者需利用工具將文本兜組成 ThdlExportXml 檔案；接著，使用者就可將這個檔案上載到 DocuSky 來建庫。 (a) 左方截圖展示建構 ThdlExportXml 的工具頁面：使用者指定文獻集名稱（此例為「西遊記」），並準備從本地端硬碟載入欲建庫的文字檔。完成文字檔載入後，使用者可點選「點我儲存」，從而將工具產生的 ThdlExportXml 檔案儲存在本地端硬碟（右方截圖）。 (b) 左方截圖展示上載建庫的小工具，它讓使用者指定文字庫的名稱（此例為「我的資料庫」），並從本地端硬碟選擇欲上載的 ThdlExportXml 檔案。右方截圖顯示這份文字庫已建構完成。

最簡單的文字庫，只包含了純文字的內容，並沒有詮釋資料等額外的資訊。在這種情況下，我們假設使用者將文本儲存於多份文字檔。[7]參考圖 1(a)，它介紹如何運用 DocuSky 目前所提供的一項簡單工具，將文字檔彙整成可上載建庫的檔案。更仔細地

[7] 而且這些文字檔僅包含全文的內容，並沒有詮釋資料等額外的資訊。若文字檔中有利用特定的格式納入詮釋資料的資訊，我們可以開發另外的工具，將詮釋資料從文字檔中擷取出來。

說，這項工具要求使用者指定欲建構的文字庫和文獻集 (corpus) 名稱[8]，然後從本地端的磁碟選取欲建庫的文字檔。接著，這工具會將這些資訊通通包裝到 DocuSky 可辨識的 XML 格式檔案裡，讓使用者儲存在本地端的硬碟上。一旦完成這些資料的彙整，接下來使用者就只需上載這份 XML 檔案，然後等 DocuSky 將文字庫建構出來（參考圖 1(b)）。



(a) (b)

圖 2 (a) 使用者登入後，可以看到屬於自己的個人文字庫。一個文字庫可以包含多個文獻集（例如截圖中，文字庫「紅樓夢」可以包含「元智版、程乙本、紅樓夢稿」等三個文獻集），而每個文獻集可包含有多份文件。 (b) 點選文獻集名稱後，可對該文獻集進行全文檢索（截圖為在文獻集「全唐詩-元智版」下，查詢「馬」的檢索結果，系統找到 707 筆符合的文件）。

使用者登入系統後，可以在頁面上看到已建構完成的文字庫。點選文字庫的文獻集連結，就可看到文獻集的內容，並且對其進行全文查詢（參考圖 2）。

比較複雜的文本資料，除了全文之外還會包含詮釋資料（例如該文本的出處和作者資訊等），甚至可能包含使用者所加上的標記 (tagging) 或筆記 (notes) 資訊。現今已有許多利用瀏覽器介面幫助使用者對文本進行標記的工具。例如 MARKUS 就是一個專門為中文歷史文獻所設計的文本標記工具 (text annotation tool)。[9]使用者可利用 MARKUS 在文本的內容上標記日期、人名、地名、藥名等資訊，也可以在文本上加註一些個人的筆記。曹又霖在他的碩士論文（曹又霖，2016 年 8 月）中設計了一套轉換格式（稱為 STAML，Simple Text Annotation Markup Language），並利用 JavaScript 實作可在 MARKUS 輸出格式和 ThdlExportXml 格式互轉的開放元件。利用這份元件，我們在 DocuSky 提供一個工具，它可以讀入多份 MARKUS 的輸出檔，轉換並彙整成一份可在 DocuSky 建庫的 ThdlExportXml 檔案。換句話說，使用者可以先利用

[8] 一個文字庫可以包含多份文獻集，而一份文獻集可包含有多篇文件。例如，我們可以將古龍小說的每個章節視為一篇文件，將每套小說「絕代雙驕」、「三少爺的劍」、「陸小鳳傳奇」... 設為一個文獻集，從而建構出一個完整的「古龍小說」文字資料庫。

[9] http://dh.chinese-empires.eu/beta/

MARKUS 標註文本，然後透過格式轉換工具取得 ThdlExportXml 檔案，接著將該檔案上載到 DocuSky 來建構包含標註資訊的文字庫。

三、DocuSky 的系統架構

圖 3 是 DocuSky 的系統架構圖。圖左代表使用者端，圖右代表系統伺服器，而中間的垂直虛線表示網際網路。典型的操作流程如下：使用者先利用瀏覽器連上系統所提供的工具 (Web Tools)，由該工具負責使用者可操作的介面（例如上載本地端硬碟的文本檔案來建庫、或者從文字庫下載文件來進行分析）；若使用者欲進行的工作必須連上 DocuSky，就由該工具透過系統所提供的 Web API (Application Interface) 來存取所需的資訊。

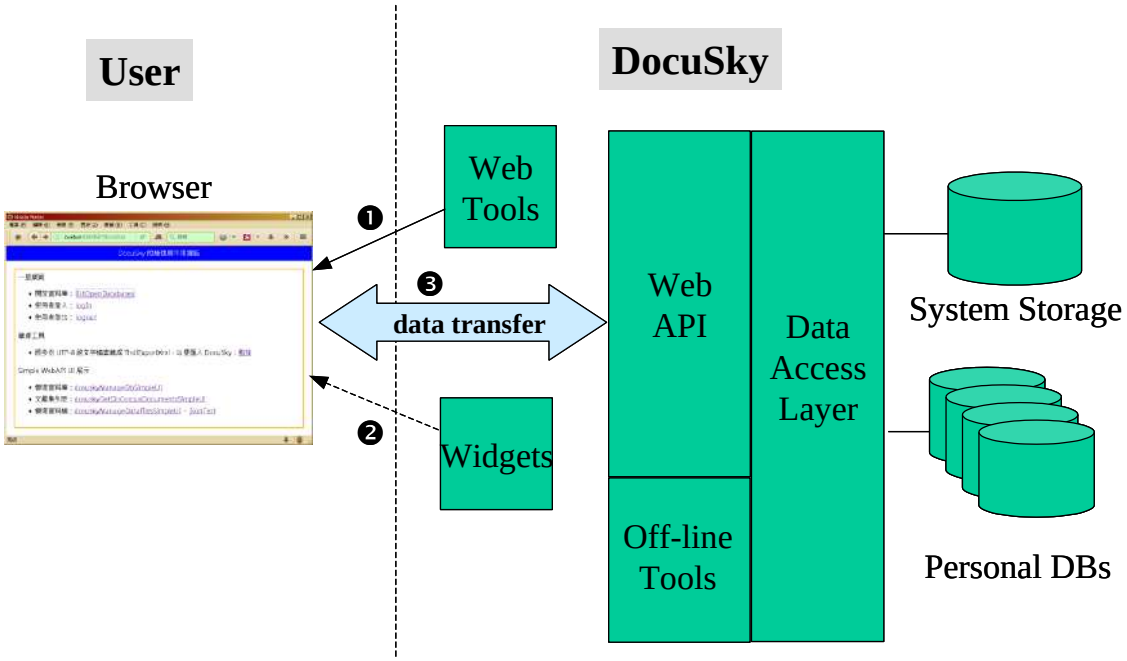


圖 3. DocuSky 的系統架構圖。典型的操作流程，是 ❶使用者利用瀏覽器連上系統所提供的工具 (Web Tools)，這份工具的作用可能是上載本地端硬碟文本檔案來建庫、對文字庫的內容進行標註、或者從文字庫下載文件來進行分析。 ❷工具可以引用套件小工具 (widgets) 來簡化開發的複雜度。 ❸瀏覽器和 DocuSky 之間的資料交換，都必須透過系統所提供的 Web API (Application Interface) 來進行存取。

由於 Web API 的運用較為複雜，若能將常用的一些功能包裝成 JavaScript 的小元件 (稱為 widgets)，就可以有效減低工具開發的複雜性。例如，工具開發者可以直接引用「取用文字庫文獻集」的小元件，很方便的取得文字庫的內容。這個小元件會檢查使用者是否已經登入（若尚未登入，會跳出一個要求使用者登入的對話方塊）。若已成功

登入，該元件會將該使用者的所有文字庫與文獻集列出，讓使用者選取欲進行分析的文獻集。使用者選取文獻集後，這個元件會從 DocuSky 取回這個文獻集的所有文件內容，放置於特定的 JavaScript 物件中。於是，工具開發者欲取得文本內容來進行加值應用，就不見得要瞭解 Web API 與 DocuSky 的溝通細節，而只需知道如何從這個物件取得該資料。例如，在（杜協昌，2015 年 11 月）我曾介紹過一個利用 JavaScript 所開發的詞夾子工具。透過引入 DocuSky widget，現在只需在原本的工具新增修改幾行程式碼，就可以取得個人文字庫的文件來進行分析處理（參考圖 4）。如此一來，工具開發者就可以節省不必要的成本，將心神專注在工具本身的功能開發上。

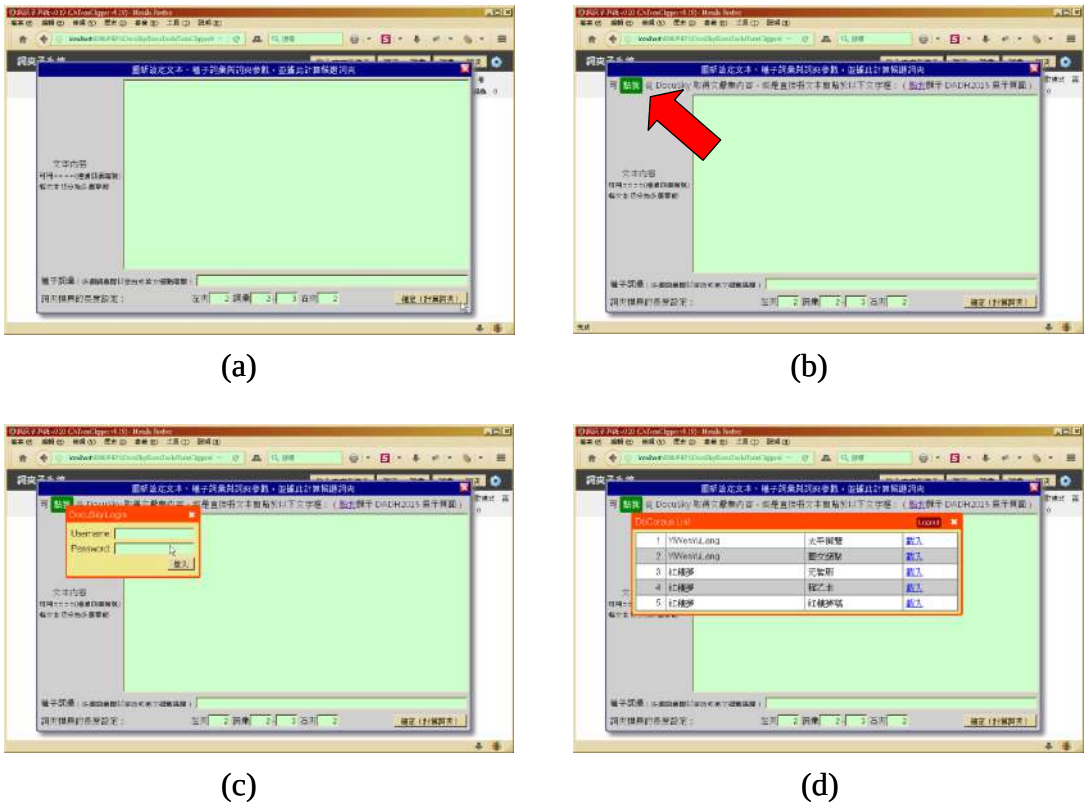


圖 4. DocuSky 的 widgets 具有簡化工具開發的優秀能力：圖 (a) 為（杜協昌，2015 年 11 月）論文所開發的詞夾子工具，它僅能利用剪貼的方式輸入文本。圖 (b) 顯示該工具引用 wedget 後，即可額外提供從 DocuSky 取得文本的功能。這個 widget 會負責「點我」按鈕（箭頭所指處）的所有行為。若使用者尚未登入，widget 會跳出要求登入的對話方塊，如圖 (c) 所示。圖 (d) 展示登入後所顯示的文字庫與文獻集列表。當使用者點選文獻集右方的「載入」連結後，widget 會負責取得該文獻集的內容，並放入特定物件以供工具取用。注意到，(b)(c)(d) 的登入與資料取得動作，都是由 widget 負責處理。

架構圖中的離線工具 (Off-line Tools) 一般並不開放給外界使用。這些工具是為了處理較大資料量的文字庫構建，或是耗時較長的的全文索引工作。這些工作通常需要大量的系統資源，並不適合在線上 (On-line) 處理。最後，在 Web API 與 Off-line Tools 與文字庫之間有一個資料存取層 (Data Access Layer)，這樣可以增加系統的彈性與可擴充性，減少日後維護的負擔。

圖 3 的架構圖雖然並不複雜，卻包含了幾項 DocuSky 在系統設計上的重要原則。首先，它展現系統對瀏覽器使用介面的喜好：使用者必須透過瀏覽器來操作工具（這表示至少在原則上，所有的人機介面的程式都必須透過 HTML5/CSS3 與 JavaScript 來實作）。[10] 此外，所有使用者端的資料存取都必須透過 Web API 來進行。不論工具的作用是在於上載建庫、查詢文件、文本標記、或者進行文本分析，它們都必須透過 Web API 來存取 DocuSky 的資料。這種必要性暗示原則上工具和文本資料應該是彼此分離獨立的（工具開發時，不必然假設運用在某份特定的文本上；文本上載建庫時，也不需假設使用特定的工具和用途），而這種獨立性可以對工具開發或應用帶來相當大的影響。

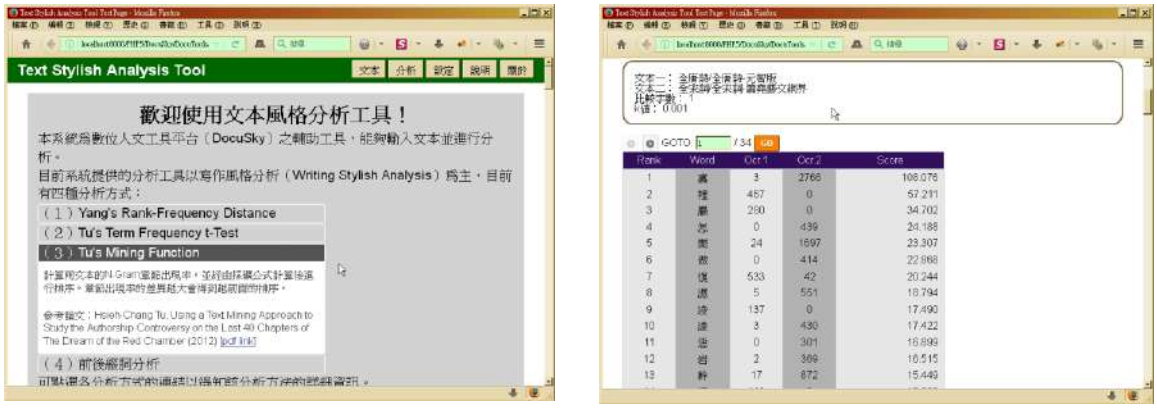


圖 5 實作論文〈利用文本採礦探討《紅樓夢》的後四十回作者爭議〉（杜協昌，2012 年 11 月）所提到的幾種文本風格分析演算法的 DocuSky 工具。左圖為該工具的使用頁面，右圖為運用其中 Tu's mining function (k=0.001) 分析「全唐詩-元智版」和「全宋詞-簫堯藝文網界」所得到的結果。從截圖中可看出，單字「巖」出現於「全唐詩-元智版」中的 280 篇文件，但是「全宋詞-簫堯藝文網界」卻一篇都沒有出現。另一方面，「怎、恁」則僅出現於後者。

其中一項值得提醒的，是這種獨立性可讓工具開發者和使用者，從各自的角度注意到，工具可以被套用到原本（工具設計之初）並未曾注意到的文本；因而可在相當程度上，鼓勵我們思考該工具的抽象性質與適用範圍。例如，在（杜協昌，2012 年 11 月）

[10] 這裡假設工具所欲執行的功能，都能透過瀏覽器執行 JavaScript 來達成，且使用者都是經由瀏覽器來連上 DocuSky。理論上會有例外的狀況：若有某項功能並不適合在瀏覽器上執行，我們需考慮開發在伺服器上執行、或者獨立下載執行的工具。此外，日後若欲開發行動裝置上的工具，它們很有可能是以非網頁形式的獨立 APP 來完成。

這篇論文中，我曾發展一個文本採礦（text mining，或稱為文本探勘）方法來比對《紅樓夢》前八十回與後四十回的文本，看看它們在常用的字詞頻率上有哪些顯著差異。將這個方法實作成 DocuSky 的工具後，由於工具和文本內容彼此獨立的性質，我們可以將欲比對的文本內容，從《紅樓夢》前八十回與後四十回，抽換成《全唐詩》與《全宋詞》。[11]在該論文中，這個採礦工具的主要目的是利用字詞頻率對《紅樓夢》前後進行寫作風格的比對，但由於它的抽象性質是比較任意兩份文本在字詞頻率上的差異，因此我們也可透過這個工具來發現一些在《全唐詩》經常出現、但在《全宋詞》卻鮮少出現的單字（例如巖、復、雲案、迴），並可找出許多後者經常出現、但前者極少見的單字（例如怎、恁）。[12]

四、DocuSky 的工具集

在 DocuSky 的系統中，所有可讓使用者操作的功能，都必須透過工具 (Web Tools) 來提供。除了系統登入後的檢索頁面（參考圖 6）外，DocuSky 預計在工具集提供幾種不同類型的工具：彙整資料的工具、建構與刪除文字庫的工具、文本標記與編輯工具、內容分析與文本觀察工具、以及詞夾子等其他的工具。

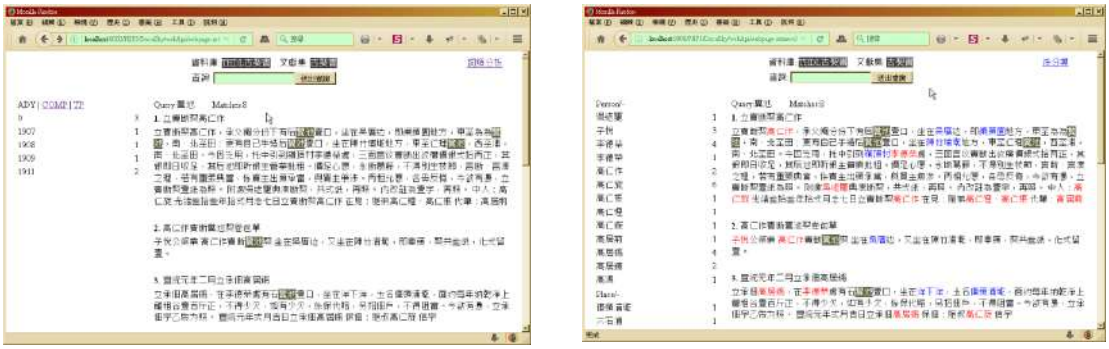


圖 6　DocuSky 支援後分類和詞頻分析的截圖。左圖的左方顯示查詢結果在西元年的後分類，而右圖的左方則顯示人名與地名詞彙的出現統計。由於 DocuSky 仍處於開發階段，它在許多地方的支援都相當陽春。例如，後分類和詞頻分析的欄位名稱，目前都僅以代碼顯示；此外，系統也尚未支援多面向搜尋 (faceted search) 等縮小查詢範圍的功能。這些都有待後續改善。

[11]《全唐詩》與《全宋詞》收納了多位作家橫跨數百年的作品，而詩和詞的文體也截然不同，因此對這兩份文本進行寫作風格的比對，顯然荒謬。然而，我們依然可以比較它們所使用的文字，看看這些文字在出現頻率上是否有顯著的差異，而這些差異在相當程度上也反應了作品所關注事物的轉變。

[12] 欲利用數位工具對文本進行分析，這些文本也必須數位化。《全唐詩》據稱共四萬餘首，在此實驗是採用元智大學 http://cls.hs.yzu.edu.tw/tang/ 的 47,957 首唐詩文本；另外，維基百科稱《全宋詞》共收錄 21,116 首，但本實驗的文本則是來自網路上收錄較完整「簫堯〖藝文〗網界」(http://www.xysa.com/) 的 18,932 首宋詞。利用 Tu's mining function，很容易發現「巖」字出現於 1,303 首唐詩，卻並沒有出現於宋詞；而「淡、閑、個、梅」等字在宋詞出現的頻率數倍於唐詩等。值得強調的是，要用人力發現文本中經常出現的字詞並不困難，但要確認某個字詞在文本中鮮少或完全沒有出現，通常就必須倚賴資訊工具。

彙整資料的工具在第二節就已介紹過，其目的是將使用者手頭的文本與詮釋資料，彙整成 DocuSky 可以辨識的格式。目前 DocuSky 僅支援 ThdlExportXml 格式。ThdlExportXml 是我將 THDL 系統的文本與詮釋資料匯出時所採用的格式，它基本上是以「文件」作為單位，每份文件包含了該文件所屬的文獻集、詮釋資料、以及內文和標記資訊。雖然 ThdlExportXml 的格式稍嫌老舊，但從多年來開發 THDL 相關系統的經驗，我相信這份格式具有足夠的彈性，可以滿足 DocuSky 在全文檢索、欄位檢索、後分類與詞頻分析的需求。

使用者將 ThdlExportXml 準備好了之後，就需要用到「建構與刪除文字庫的工具」來將這份檔案上載建庫。DocuSky 為此提供了一個 widget，將建庫和刪庫的主要工作都包裹在這個元件中。對於一個提供個人文字庫的系統來說，線上的文本標記與編輯工具相當重要，因為在許多場合下，使用者都需對資料進行標註與更正。可惜的是，限於人力不足，這部分的工具還未開發完成。

內容分析與文本觀察工具的目的，是讓使用者透過各種不同的分析工具，對文本內容有逐篇深讀之外的了解。例如，我們目前已經開發了數種文本字詞的統計工具，可讓使用者比對文獻集的字詞數量、計算每篇文件中特定詞彙的出現情形，也可讓使用者比對兩份文本在字詞頻率上的風格差異。本屆會議座談 (panel discussion) 的另一篇論文（謝博宇等，2016 年 11 月）中，將會對這些工具進行較為深入的介紹與應用。

五、近期工作與未來展望

目前 DocuSky 系統仍處於早期開發的階段，也僅提供有限的帳號讓具有高度意願者進行先期的試用。這篇論文中，我介紹了 DocuSky 的初步成果，但細心的讀者應可看出許多明顯從缺的重要工具、以及許多有待改進的功能和使用介面。例如，目前 DocuSky 雖已提供將 MARKUS 標記後的文本轉換成 ThdlExportXml 的工具（這份轉換後的檔案，可建構支援詞頻分析的文字庫），但尚缺乏能夠一併整合詮釋資料的工具。又如，從圖 6 可看出 DocuSky 在後分類和詞頻分析的顯示介面和功能都相當陽春，亟待後續改善。還有，上一節所提到的文本分析工具，基本上都僅利用全文的內容來進行統計分析。由於詮釋資料與標記後的文本能夠提供更多使用者關心的資訊，我們也應開發一些能夠對詮釋資料與文本標記進行統計分析的工具。

另外，若要讓 DocuSky 成為一個工具開發和文史研究者都能有所發揮的平台，這個系統目前顯然還欠缺一些重要的功能。例如，它至少需提供一個機制來讓這兩個領域

的專業人士溝通訊息（例如提供留言板的功能、或者利用 Facebook 等大型社群網站來進行聯繫）。透過訊息溝通的機制，研究者就有機會將自己需要的功能張貼出來，請有興趣的程式設計師幫忙開發合適的工具。當然，DocuSky 也必須提供一個方法，讓程式設計師將開發完成的工具上載給他人使用。當然，既然要稱之為平台，就應該提供資料分享的功能，因此也必須為此設計合適的資料權限控管機制。

除了以上所提到的一些近期工作，DocuSky 還有另一個重要的發展面向，那就是接受更多其他類型的文本資料。目前 DocuSky 僅支援 ThdlExportXml 格式，而這種格式僅適合處理一般具有詮釋資料和標記的文本。我們計畫讓 DocuSky 也能支援族譜類型的資料。在郭秀萍的研究（郭秀萍，2016 年）中，曾開發一份描述中文家譜的 JPML (JiaPu Markup Language) 格式。我們打算以此為基礎，制定一份 DocuZupuXml (DocuSky Zupu XML) 格式來支援族譜類型的資料。當然，若有充分的開發資源，我們也可擴充系統，讓使用者能夠建構個人的多媒體資料庫，對文本、圖片、影像、以及相關的詮釋資料進行更全面的彙整與應用。

我們也知道，當前世界上已有相當多機構提供開放的數位人文相關工具。例如，哈佛大學的 China Historical GIS[13] 提供了中國歷史各個時期的相關地圖資料；史丹福大學的 Paladdio[14] 提供視覺化工具 (visualization tools) 讓文史研究者能對複雜的資料進行分析；而 Voyant Tools[15] 也開發了各種可在瀏覽器上執行的視覺化工具，可以讓研究者對文本進行線上閱讀與分析。我們希望能夠在 DocuSky 上提供合適的工具和連結機制，讓使用者可以運用外界所開發的開放資源來對 DocuSky 進行加值，從而更有效地拓展數位人文的整合應用。

總結地說，雖然目前 DocuSky 系統仍處於開發的階段，但這篇論文已經展示許多它異於傳統典藏資料庫的優異能力。尤其是，DocuSky 可讓文史研究者建構自己的文字資料庫，並運用各類工具對這些文字庫進行分析與加值應用。另一方面，DocuSky 也提供許多小元件，它們能在相當程度上減少工具開發的負擔。我們樂觀地相信，未來有機會找到有意願合作的機構與個人，將這個系統擴充成一個數位人文的研究平台，並在這平台上產生許多有趣的應用。

[13] https://www.fas.harvard.edu/~chgis/
[14] http://hdlab.stanford.edu/palladio/
[15] https://voyant-tools.org/

參考文獻

杜協昌。2012 年 11 月。〈利用文本採礦探討《紅樓夢》的後四十回作者爭議〉。第四屆數位典藏與數位人文國際會議。

杜協昌。2015 年 11 月。〈半自動詞彙擷取〉：簡化的詞夾子方法以及其 JavaScript 元件的開發與應用。數位典藏 數位人文 DADH 2015 國際研討會論文集，台北。

郭秀萍。2016 年 1 月。〈中文家譜數位化研究〉。國立臺灣大學資訊網路與多媒體研究所碩士論文。

曹又霖。2016 年 8 月。〈文本標記格式的轉換與應用〉。國立臺灣大學資訊工程研究所碩士論文。

謝博宇。2016 年 12 月。〈以 DocuSky 為核心的工具開發與建置〉。DADH 2016 國際研討會論文集，台北。

三國演義人物說話關係之標註與呈現

王景逸*、黃家富**

摘　要

在小說這種文學體裁中，說話行為乃是最重要的角色活動，說話以及說話的參與者往往是促成小說情節發展的因素。因此對於說話內容以及參與者作分析，我們可以試著推斷出角色間關係的「質」與「量」。為此本研究建立了一套標註方法，並為此方法設計了一套標註系統，來協助使用者進行小說文本的標註，並試著對三國演義中的赤壁之戰內容做標註。在完成標註之後，本研究將資料應用在人物說話統計工具做標註結果的呈現，以及套用社會網路分析方法，這些方法包括了 Degree Centrality、Closeness Centrality、Betweenness Centrality、Modularity，並嘗試去解讀得到的結果。

關鍵字：三國演義、赤壁之戰、社會網路分析、角色關係、Docusky

* 國立臺灣大學資訊工程研究所碩士生，Email: kather0912888@gmail.com。

** 國立臺灣大學資訊工程研究所碩士生，Email: james30199@gmail.com。

Tagging and Displaying Character's Conversational Relationship in the *Romance of the Three Kingdoms*

Gia-fu Huang[*] , Jing-yi Wang[**]

Abstract

Conversation is an important part of novels. A story can be driven by the conversational contents and their participants. By examining the nature and the quantity of conversations between characters, we may be able to capture their relationship to each other. This research develops a novel tagging process and presents a system to help users tagging conversations in full texts. We demonstrate the tagging process by tagging Battle of Red Cliff, a famous story in the *Romance of the Three Kingdoms*. We also visualize the tagged results using various social network connectivity relations.

Keywords: Romance of the Three Kingdoms, battle of Red Cliff, social network analysis, role relationship, Docusky

[*] Master Student, Department of Computer Science & Information Engineering, National Taiwan University, Email: james30199@gmail.com
[**] Master Student, Department of Computer Science & Information Engineering, National Taiwan University, Email: kather0912888@gmail.com

一、 前言

在小說文本的探索當中，研究者們所關心的議題眾多，除了解析小說的時代價值，或是解讀作者筆下所嘗試表達的意義，角色關係也是相當有趣的議題。尤其近年來社會網路分析(Social Network Analysis)的熱潮興起，社會學者與文史學者開始利用社會網路分析相關的統計方法，來達到人物關係的量化與視覺化的呈現。社會網路分析將人際網路看成點和線組成的網路，網路圖上的節點代表的是個體，線則是個體與個體的關係。但是這些統計數字和複雜圖表代表的意義是什麼？我們必須回來探討更基礎的問題，也就是如何去界定兩個角色間的關係？這個問題有幾個層次：1、小說文本中什麼要素讓我們可以決定，小說文本中兩個角色之間是有關係的。2、這兩個角色的關係深淺，也就是關係的「量」是如何。3、這兩個角色的關係類型，也就是關係的「質」是如何。我們必須有辦法回答上述的三個問題，才能進一步探究社會網路分析中的統計數字和圖表，最後真正達到小說文本角色關係的解析。

二、 角色關係之判斷

如前述，角色關係是小說裡相當有趣的議題，而人的關係往往是很複雜的。基本上要判斷小說裡人物間的關係，主要就是依據該小說文本中的內容來做出相關的解讀，在此分類為下列四個要素，並分別予以介紹。

(一)說話內容：從角色說出來的話判斷出角色間的關係，例如：

> *言未畢，忽帳下一人出曰：「某自幼與周郎同窗交契，願憑三寸不爛之舌，往江東說此人來降。」曹操大喜，視之，乃九江人：姓蔣，名幹，字子翼，見為帳下幕賓。*

由說話內容可以得知，蔣幹與周瑜是小時候的同學關係。

(二)角色間的互動：角色 A 對角色 B 的行動。例如曹操和蔡瑁及張允有如下的橋段：

> *瑁曰：「軍尚未曾練熟，不可輕進。」操怒曰：「軍若練熟，吾首級獻於周郎矣！」蔡，張二人不知其意，驚慌不能回答，操喝武士推出斬之。*

曹操責罵了蔡瑁及張允，並使手下斬下兩人的頭，代表曹操不信任蔡瑁及張允，處於不信任的關係。

(三)角色心理想法：角色在該情境下的心理內容。例如蔣幹偷看周瑜書信之後，有這麼

一段：

幹思曰：「原來蔡瑁，張允，結連東吳！……」遂將書暗藏於衣內。再欲檢看他書時，床上周瑜翻身，幹急滅燈就寢。

蔣幹看了書信後，開始懷疑蔡瑁、張允，因此開始了不信任的關係。

(四)旁白：作者在小說中補充敘述人物時，有時會提到人物關係，例如第一回在介紹黃巾賊有這樣的內容：

時鉅鹿郡有兄弟三人：一名張角，一名張寶，一名張梁。

也就是這三個人是兄弟關係。

文本中角色關係之基礎是什麼？要處理小說中的角色關係，我們可以粗略地將它類比成現實生活中的人物關係，將小說中的角色視為實體，角色間的互動就是實體間的互動。在此我們討論的是直接的互動，假如沒有直接的互動，卻有血緣關係等等，這便不在我們定義的關係之內。現實生活中基礎關係之建立就是藉由互動來得到，例如交談、握手、擁抱等等，因此我們試著將關係之建立奠基於角色間實際之互動行為。實體間的行為會帶來相關的解讀，例如甲拍了乙一下，「拍」的行為到底是帶有調侃的意思，還是帶有一點責罵的意思，這就要從當時的情境來做判斷與解讀了。小說也是如此，三國演義中常常只會有一個「曰」字，這個曰代表責罵還是稱讚對方，就有賴讀者對於該情境的解讀。藉由這樣對角色間行為的詮釋，就能判斷出角色間的關係，例如友好、敵對關係。由於關係建立的基礎是小說角色間直接的互動，因此說話內容、旁白都不在互動關係定義的範圍之內。

角色間的直接行為建立了關係的基礎，而行為的多寡則決定了角色關係的深淺。而對角色間行為的詮釋，可以進一步瞭解到角色間的關係。如下範例：

次日，操酒醒，懊恨不已。馥子劉熙，告請父屍歸葬。操泣曰：「吾昨因醉誤傷汝父，悔之無及。可以三公厚禮葬之。」又撥軍士護送靈柩，即日回葬。

曹操對劉熙說了一句話，因此曹操和劉熙就建立起說話關係，而可從內文得知曹操哭著對劉熙說話，所以角色間行為詮釋就是哭訴，又一般人只會向親近的人哭訴，所以便能進一步得知兩人是友好關係。

藉由角色間直接的互動來決定角色關係的建立，由角色間行為詮釋來決定關係的「質」，再由角色間的直接行為發生次數來決定關係的「量」，有了這些瞭解之後，就能

進一步去處理文本，並將結果做後續社會網路分析的利用。

三、 方法回顧

研究小說人物關係的研究眾多，在此介紹和本研究比較直接相關的文獻，主要有 2010 年廖倆凡的〈中國古典白話小說中的社會網路關係：以《儒林外史》為例〉，以及 2015 年趙薇的〈社群網絡分析(SNA)在現代漢語歷史小說研究中的應用初探—以李劼人的《大波》三部曲為例〉。在廖倆凡的研究中使用的流程大略如下：抽取全部說話片段，提取每一節句的說話者、主要聽眾、次要聽眾等角色，最後將得出來的資料用社會網路分析方法來呈現網路圖並予以解釋；趙薇大致上沿用了廖倆凡的流程，但更細緻地去探討每一張網路圖中角色間的關聯，以及對該圖和劇情的呼應做了許多豐富且詳細的解釋。

但是就社會網路分析方法的適用性來說，這種標註方法是有可能導致錯誤解讀的。例如沒有對角色間互動的關係進行分類的話，很可能會將友善的互動和不友善的互動看成同一類，而導致將來在社會網路分析的應用與解釋上碰到困難。研究者們常關注的人際網路中群集問題，也就是所謂的團體，如果不對角色互動關係進行分類，只是將所有互動都看成同一類的話，則網路圖中的同一個群集代表的意義就會模糊不明。

標註系統的設計上則是參考了 Hamburg 大學團隊[1]所開發的 CATMA。CATMA 做為文本的標註系統，提供了基本的標註功能以及相關統計的呈現。但是由於不是針對小說這種體裁來做設計的，因此在針對小說類型文本的標註上會有它的限制與不便之處。

四、 文本標註方法概述

在小說這種文學體裁中，說話行為、說話內容、說話的參與者往往是推動小說情節發展的要素。根據初步統計，三國演義赤壁之戰中角色的互動行為中，說話行為就占了一半。因此本研究在人物關係的判讀與處理上，優先處理說話行為，暫不處理說話以外的行為。

人物說話行為是小說劇情開展的重要方式，因此試著以說話單位來保留住前述的資訊。一個說話單位包含了以下項目：說話內容、說話者、聽眾、說話類型、說話欄位置、是否接續前說話。說話內容保存的就是說話者之說話行為所呈現的內容；說話者即是說話行為的來源方，聽眾即是說話行為的接受方，由於考量到說話者和聽眾都有可能是大於一位，因此標註時說話者和聽眾都可以是複數；說話類型所保存的就是此說話行為的

[1] 詳細資訊請參考 http://www.catma.de/contact.html

詮釋，是威脅還是稱讚等等，因為一個說話類型也有可能是多個，因此說話類型數量上也可以是複數；說話欄紀錄的是此說話欄在文本中的位置。

在三國演義赤壁之戰說話單位的標註上，必須先決定要將說話單位類型分成哪幾類，在對於三國演義文本的說話類型進行了前二十個章回的觀察之後，歸納為下列 28 個類型：一般說話、報告、討論、心想、命令、宣布、評論、責罵、哀嘆、懷疑、拉攏、稱讚、請求、欺騙、詢問、抱怨、勸告、安慰、嘲笑、奉承、威脅、感謝、請求原諒、推辭、邀約、哭訴、自嘲、驚訝。在評估三國演義赤壁之戰該有的說話類型時，本研究盡量讓說話類型多且更細緻，因為之後要做角色關係的判斷時，越細緻化的分類對於關係之判斷會更有助益。

五、 文本標註系統介紹

為使標註文本更加方便快速，本研究設計出一套標註系統，能夠自動化提取小說中的說話片段，並協助研究者手動擷取說話者及聽話者，以及說話類型的標註等等。這部分放棄了自動化的方式，主要是因為自動化擷取說話者及聽話者和類型仍有相當的難度，因此先跳過自然語言處理的範疇。在整個文本標註完之後，必須對角色名稱進行整併，因為標註時可能會有角色別名的問題，例如劉備又稱玄德；關羽又稱關公、雲長，勢必要處理角色別名之狀況。最後，標註完的文本會得到一個相對應的文本 XML 檔，以作為後續的說話統計分析、社會網路分析之利用。

標註系統的使用流程如下：

1. 匯入文本及抓取說話內容的正規表達式，也可以直接匯入 Docusky 格式之 XML
2. 自動化抓取文本中的說話內容
3. 人工判斷是否有抓錯或缺漏的對話，可以使用標註系統裡的工具補足
4. 標註每個說話內容的 metadata
5. 切換下一個文本並繼續步驟 2
6. 整理人物名稱
7. 匯出 Docusky XML

圖 5-1 輸入頁面

此標註系統支援 txt 檔及符合 Docusky 格式的 XML 檔，輸入端支援多個檔案的輸入，並且支援同時輸入 txt 和 XML。這部分的設計是為了使用者的彈性考量，因為使用者可能一份編輯過的 XML 中，想要額外加新的內容去標註。預覽區會呈現匯入的所有檔案內容，使用者可以確認已經匯入正確的文檔後再來進行標註工作。

圖 5-2 標註頁面

首先提到版面配置，在畫面切分上將標註區和預覽區各分約 50%，預覽區 50%實際在使用上已經足夠使用者來判斷說話內容的 metadata。標註區提供了標註工具來協助使用者對文本進行標註，說明如下：

1. 說話內容：標註工具目前的對象，也就是特定的說話內容。為了方便使用者找出還沒進行標註的內容，尚未標註的會以黃色為底色，標註完的則是以藍色為底色。

2. 說話者：說話者名稱。

3. 聽話者： 受話者名稱。

4. 最近使用： 快取，暫存最近使用過的名稱，最多存放 5 個。這部分考量到使用者在標註時大部分時間都花在輸入說話者及受話者，，因此設計了快取(Cache)來改善，採用 LRU(Least Replacement Used)演算法。

5. 類型：說話類型。考量到不同使用者對不同文本會有不同的標註需求，因此 Type 的內容使用者可以自行新增。若使用者沒有這部分的需求，可以直接忽略這個功能，標註系統將會預設為一般對話。

6. 接續前對話：是否接續前面說話內容，使用者選擇是或否，也可以在旁邊的空白區輸入接續前面第幾個說話內容，標註完畢後系統會統一整理，讓該說話內容可以參照到特定的 ID。若使用者沒有這部分的需求，可以直接忽略這個功能，標註系統將會預設為否。

圖 5-3 角色名稱整理頁面

此頁面的主要功能為進行人名合併，例如劉備又叫玄德，因此可將劉備及玄德合

併成一個人，在最後輸出的 Docusky XML 中，兩人就會被視為同一個人而有相同 ID。

標註完成後，就可以使用下載功能，得到的檔案即是一份 Docusky XML。

圖 5-4 標註後的 XML 檔

圖 5-5 標註後的 XML 檔

將來可以再次匯入 XML 檔進行編輯，或是將此 XML 檔運用在其他的工具上，例如 Docusky 平台上支援的工具，或是 Gephi 等軟體。接下來會呈現將 XML 套用在角色說話統計工具，以及套用在 Gephi 軟體上得到的社會網路關係圖。

六、 方法與步驟概述

本研究以下列步驟來進行：

1. 使用標註系統對文本進行標註
2. 將文本匯入到角色說話統計工具
3. 先將 xml 轉檔成 gdf，再匯入 Gephi 軟體
4. 藉由 Gephi 製作文本的網路圖，並計算出相關的社會網路分析統計值

七、 人物說話統計圖表

標記完成的資料為了便於使用者觀察，為此而設計了對應的統計呈現工具方便讀者鳥瞰整份資料，讓使用者能從多個角度來檢視資料。

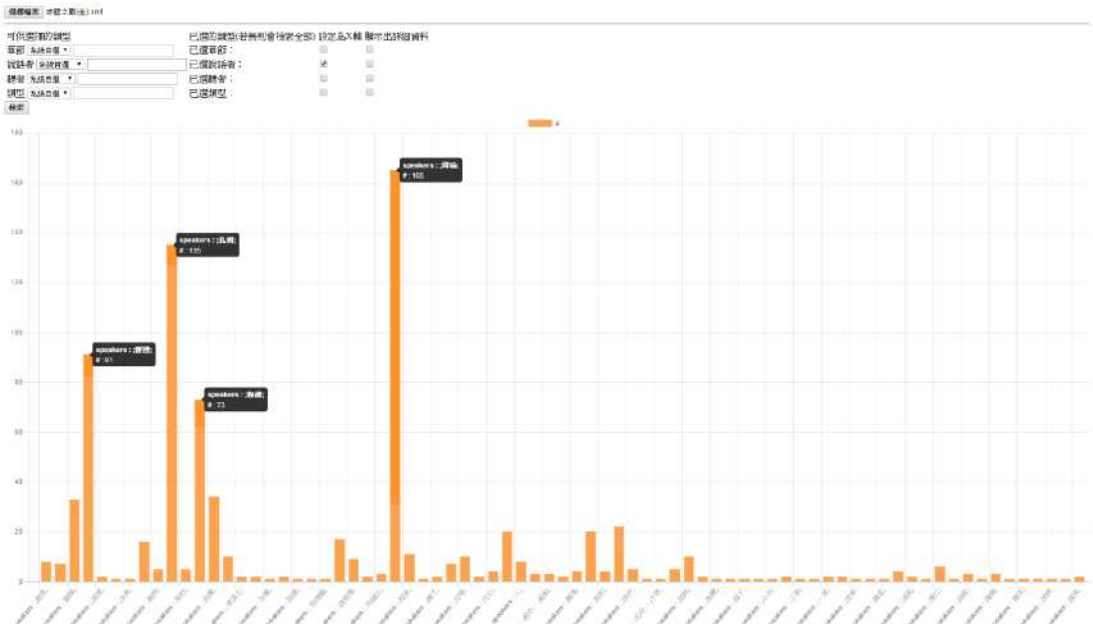


圖 7-1 角色說話頻率統計圖

上圖是以說話者為橫軸，所作出的統計圖表。可以見到，赤壁之戰有較多說話行為的角色為周瑜、孔明、曹操、魯肅，次數分別是 165(20.1%)、135(16.5%)、91(11.1%)、73(8.9%)。這四個人在整個赤壁之戰說話比例就佔了 56.7%。讓我們先粗略地假設他們在本章節內為主要角色，以他們為對象作更多類型的查詢。

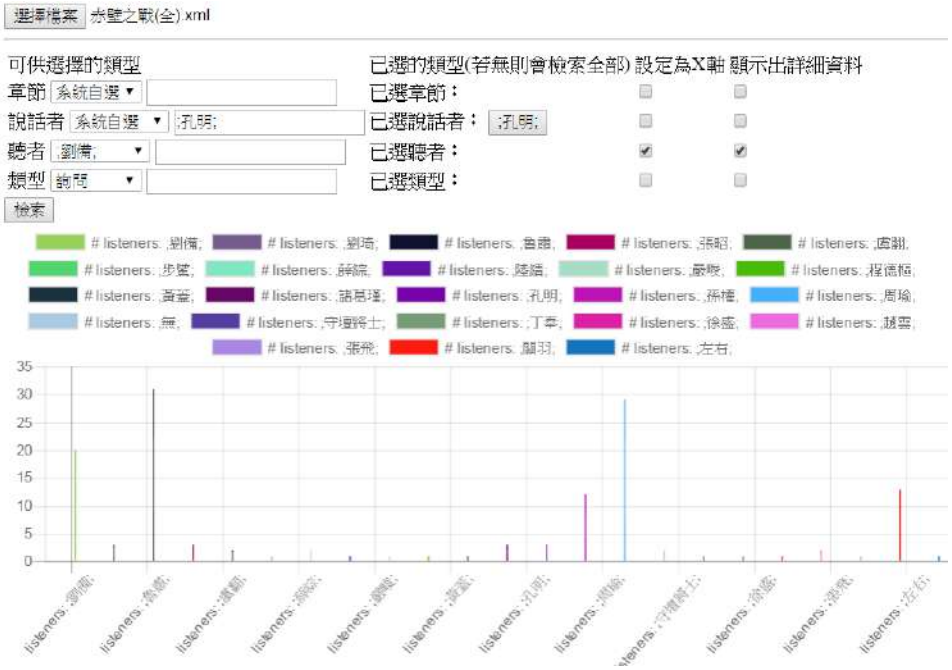


圖 7-2 孔明說話對象統計圖

上圖以周瑜為說話者，查詢聽者的分布，可以見到魯肅、孔明及蔣幹為最高，次數分別為 29(17.5%)、31(18.7%)、16(9.6%)。文章提到「魯肅與瑜最厚」，因此可以得知周瑜和魯肅關係最好，時常向魯肅訴說他的計策；周瑜利用孔明的智慧聯合打擊曹操，因此和孔明說了很多的話；而因為曹操派遣蔣幹來拉攏周瑜，於是周瑜和蔣幹也說了許多的話。接著想進一步知道他們說話類型的差異，繼續往下查詢。

圖 7-3 周瑜說話類型統計圖

以周瑜為說話者，並限制聽者為孔明、魯肅、蔣幹，顯示出說話的詳細類型以利分析。周瑜對魯肅的說話類型最多為一般對話(62%)，主要是因為周瑜會告訴魯肅他的計謀與策略；對孔明則是一般對話(32.2%)和詢問(29%)還有請求(19.3%)佔了大部分，因為周瑜時常向孔明商討與請教軍事策略，並尋求相關的協助；對蔣幹則是因為周瑜心知蔣幹為曹操說客，因此和蔣幹說了許多客套話，在分類上為一般對話(56.2%)。

八、 社會網路分析之應用

社會網路分析包含了許多的方法，以求出該網路中節點許多不同的統計值。這些統計值可以反映出該節點在整個網路中的特性，例如關鍵性、中心程度等等，協助研究者分析某個節點在網路中的角色。以下將以社會網路分析中常用的指標 Degree、Closeness Centrality、Betweenness Centrality、Modularity Class 來做為說明。

(一)Degree

Degree 代表是節點的連接數量，也就是一個點連出去的邊之總和。一個點連出去的邊數量越多，Degree 也會越大，代表該角色和越多角色說話。

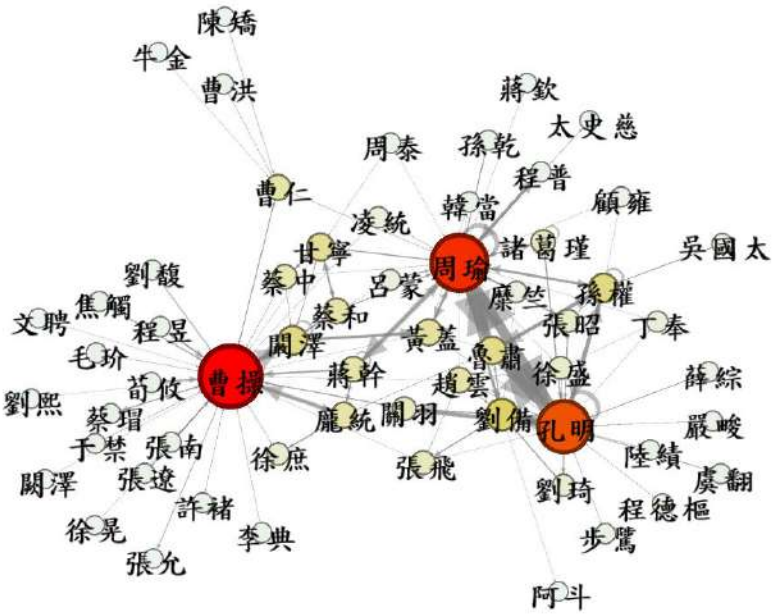


圖 8-1 赤壁之戰人物關係網路：Degree

圖 4 比較大及比較紅的點就是 Degree 較高的點。可以看出，孔明、周瑜及曹操是

Degree 最高的角色，分別是 38(10.6%)、46(12.8%)、49(13.6%)，參考附表 A-1。這三人就是在赤壁之戰中，和最多人說話的角色，占了總說話數的 37%，可以說他們是赤壁之戰主要的劇情推動者。

(二)Closeness Centrality

Closeness Centrality 是用來判斷一個節點的中心程度。中心程度越大代表越容易到達其他節點，也就是平均路徑更短。初步在角色說話關係形成的網路上來看，這個統計值的意義很模糊，因為說話關係包含了友善和非友善的互動，因此必須事先對說話類型進行篩選。本研究示範以友善說話關係進行篩選後，得出的 Closeness Centrality。友善的說話關係為 28 項說話類型中篩選出的 13 項，分別為報告、討論、哀嘆、詢問、稱讚、請求、抱怨、勸告、安慰、感謝、邀約、哭訴、自嘲。

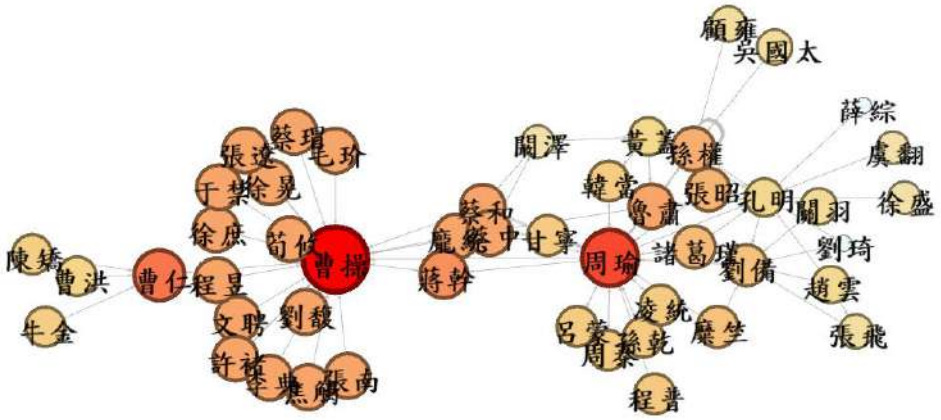


圖 8-2 赤壁之戰人物關係網路：Closeness Centrality

上圖即為角色友善說話關係的網路，比較大及比較紅的點是中心性程度較高的點。可以看出，曹操和周瑜在赤壁之戰的章節當中，是中心性最高的點，參考附錄 A-2。在友善的關係當中，兩人位於最核心的位置。這也是因為在赤壁之戰的章節當中，曹操和周瑜必須不斷地和手下將士有良好的互動，例如詢問計策或是激勵將士，才能在戰爭即將發生時做好準備。在所有的說話類型中詢問類型的部分，曹操就有 19 句(20.8%)，周瑜則是 42 句(25.6%)。

(三)Betweenness Centrality

Betweenness Centrality 是評估一個節點處於樞紐的程度。該值越大代表該點處於越

樞紐的位置，也就是說，如果將這個點拿掉，會有更多其他的點將會與整張圖失去連結，網路也會被破壞掉。本研究示範以友善說話關係進行篩選後得出的 Betweenness Centrality。友善說話關係為 28 項說話類型中篩選出 13 項，分別為報告、討論、哀嘆、詢問、稱讚、請求、抱怨、勸告、安慰、感謝、邀約、哭訴、自嘲。

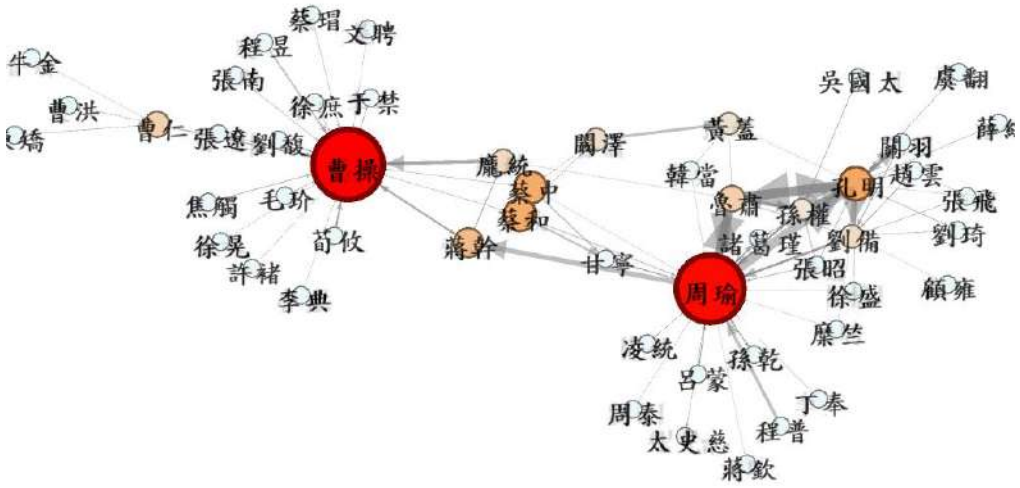


圖 8-3 赤壁之戰人物關係網路：Degree

上圖即為角色友善說話關係的網路，圖上比較大及比較紅的點代表就是 Betweenness Centrality 程度較高的點，也就是這個點的存在非常重要，少了它會讓整個圖嚴重地分崩離析。周瑜和曹操為 Betweenness Centrality 程度最高的角色，由此可以知道兩人是這場戰役真正不可或缺的關鍵。這點倒是蠻符合史實的。根據史實紀載，諸葛亮並沒有直接的參與，而只是與魯肅一起促成孫劉聯盟。

(四)Modularity Class

Gephi 中的 Modularity 演算法會將網路中關係較為緊密的點歸類為同一個群集。同一 Modularity Class 代表群集內的節點關係較緊密，而與網路中其餘的點較不緊密。本研究示範以友善說話關係進行篩選後，得出的 Modularity Class。友善的說話關係為 28 項說話類型中篩選出 13 項，分別為報告、討論、哀嘆、詢問、命令、稱讚、請求、抱怨、勸告、安慰、感謝、邀約、哭訴、自嘲。得出來的結果，同一群集內代表互相關係較密集且友善，可以將相同群集的概念理解為赤壁之戰中的同一陣營。

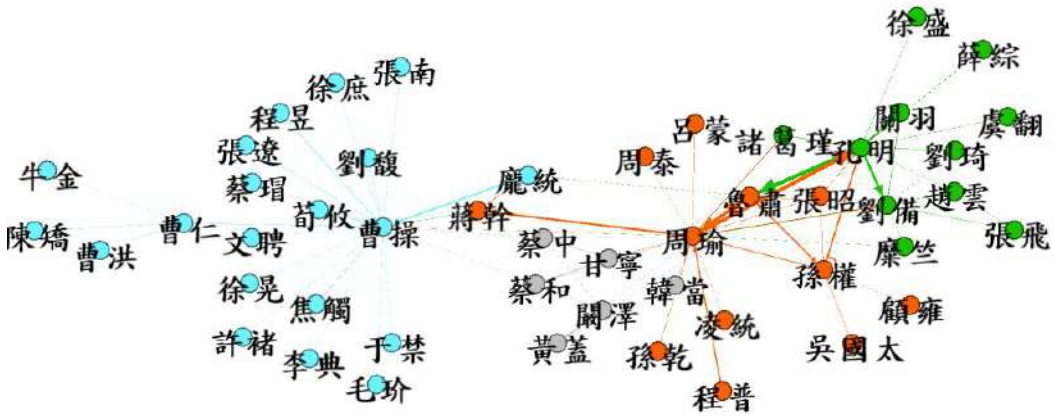


圖 8-4 赤壁之戰人物關係網路：Modularity

如上圖所示，依照演算法計算過後，可以分為 4 個群集，分別以紅藍綠灰色來區別。由於此時的圖形上的邊代表的是友善的互動，因此相同群集內代表友善互動較多，彼此之間關係更緊密。現在根據赤壁之戰內容，分別將藍、紅、綠、灰方分別命名為曹操陣營、周瑜陣營、劉備陣營、詐降陣營。值得一提的是詐降陣營，該陣營會被特別區分出來，是由於第四十七回周瑜手下黃蓋、闞澤以及甘寧要詐降曹操，期望藉由曹操的手下蔡中、蔡和的協助以投靠曹操。彼此之間有較為頻繁且深入的對話，因此這些人彼此的互動較為緊密，自成一個群集。

大致上 Modularity 的分類和赤壁之戰實際的陣營一致，除了部分的角色被歸類到錯誤陣營，例如蔣幹、龐統、徐盛、諸葛瑾、孫乾。被歸類錯誤的原因主要是由於該角色被主公派去拉攏或是誤導其他陣營的角色，因此和不同陣營友善說話數量較自家陣營友善說話數量為多，例如蔣幹為曹操派去拉攏周瑜的手下，和周瑜的互動較曹操為多；龐統是周瑜暗中使去獻給曹操連環計的將士，和曹操也有比較多的互動；諸葛瑾是周瑜遣去說服孔明來投靠的人；徐盛為周瑜派去請求孔明暫留幾天；孫乾是劉備派去恭賀周瑜勝利的角色。針對這個問題，如果將來標記的回數夠多，讓每個人說話的數量足夠多以反映出角色真正的互動網路，就能夠進一步減少這樣的錯誤。

參考文獻

Blondel, Vincent D, Guillaume, Jean-Loup, Lambiotte, Renaud and Lefebvre, Etienne. "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 , no. 10 (2008): P10008.

Brandes, Ulrik, Delling, Daniel, Gaertler, Marco, Goerke, Robert, Hoefer, Martin, Nikoloski,

Zoran and Wagner, Dorothea. "On Modularity Clustering." *IEEE Transactions on Knowledge and Data Engineering* 20 , no. 2 (2008): 172-188.

Ulrik Brandes . " A Faster Algorithm for Betweenness Centrality “. University of Konstanz Department of Computer & Information Science.2001.

Yannick Rochat . " Closeness Centrality Extended To Unconnected Graphs : The Harmonic Centrality Index “. Institute of Applied Mathematics University of Lausanne, Switzerland.2009.

廖傋凡。〈中國古典白話小說中的社會網路關係：以《儒林外史》為例〉。碩士論文，臺灣大學，2010。

趙薇，〈社群網絡分析(SNA)在現代漢語歷史小說研究中的應用初探─以李劼人的《大波》三部曲為例〉，第六屆數位典藏與數位人文國際研討會論文集，459-480，2015。

九、 附錄

表 A-1 角色 Degree 統計表

name	Indegree	outdegree	degree
趙雲	2	6	8
張飛	3	3	6
劉備	9	7	16
曹操	28	21	49
張遼	1	1	2
許褚	1	1	2
李典	1	1	2
關羽	3	3	6
劉琦	2	2	4
孔明	18	20	38
荀攸	1	1	2
魯肅	6	5	11
孫權	7	5	12
張昭	2	3	5
虞翻	1	1	2
步騭	1	1	2
薛綜	1	1	2
陸績	1	1	2

表 A-2 角色社會網路分析統計表

name	Modularity class	Closness centrality	Betweeness centrality
趙雲	6	0.275	0
張飛	6	0.251908	0
劉備	6	0.326733	102
曹操	2	0.407407	945
張遼	2	0.295652	0
許褚	2	0.295652	0
李典	2	0.295652	0
關羽	6	0.272727	0
劉琦	6	0	0
孔明	6	0.326733	283.4167
荀攸	2	0.292035	0
魯肅	6	0.4125	144.4167
孫權	6	0.347368	66.58333
張昭	6	0.346939	0
虞翻	6	0.251852	0
薛綜	6	0	0
黃蓋	20	0.34375	76

接表 A-1 角色 Degree 統計表			
嚴畯	1	1	2
程德樞	1	1	2
黃蓋	3	6	9
諸葛瑾	3	3	6
吳國太	1	1	2
周瑜	22	24	46
程普	1	1	2
呂蒙	1	2	3
甘寧	5	5	10
糜竺	1	2	3
蔡瑁	1	1	2
蔣幹	4	4	8
毛玠	1	1	2
于禁	1	1	2
蔡中	3	4	7
蔡和	4	4	8
闞澤	6	6	12
龐統	4	4	8
徐庶	2	2	4
劉馥	1	1	2
程昱	1	1	2
焦觸	1	1	2
張南	1	1	2
韓當	2	1	3
周泰	1	2	3
丁奉	2	1	3
徐盛	4	1	5
凌統	1	2	3
曹仁	5	3	8
孫乾	1	1	2
曹洪	1	1	2

接表 A-2 角色社會網路分析統計表			
諸葛瑾	14	0.340206	0
顧雍	6	0.263566	0
吳國太	6	0.263566	0
周瑜	14	0.464789	902.5
程普	14	0.32381	0
呂蒙	14	0.320388	0
甘寧	20	0.320388	0
糜竺	14	0.330097	0
蔡瑁	2	0.295652	0
蔣幹	14	0.294643	206.5833
毛玠	2	0.292035	0
于禁	2	0.292035	0
蔡中	20	0.478261	264
蔡和	20	0.478261	264
闞澤	20	0.266129	89
龐統	2	0.292035	105.5
徐庶	2	0.292035	0
劉馥	2	0.292035	0
程昱	2	0.295652	0
焦觸	2	0.295652	0
張南	2	0.295652	0
韓當	20	0.336735	11
周泰	14	0.32381	0
徐盛	6	0.251908	0
文聘	2	0.295652	0
凌統	14	0.320388	0
徐晃	2	0.295652	0
曹仁	2	0.297297	141
孫乾	14	0.320388	0
牛金	2	0.234483	0
陳矯	2	0.234483	0
曹洪	2	0.230769	0

以 DocuSky 為核心的工具開發與建置

謝博宇*

摘　要

DocuSky 為一協助人文學者進行數位人文研究的系統，其特點在於能夠讓學者以所持的電子化文本檔案自行建立專屬的個人資料庫，再利用 DocuSky API 的程式工具進行分析以進行研究。在 DocuSky 的架構下，這些工具主要透過網頁瀏覽器運作。此一特性為使用者（人文學者）與開發者（資訊技術人員）雙方帶來合作上的益處：一方面普遍有網頁使用經驗的使用者易於學習如何使用工具，另一方面開發者能憑借網頁與瀏覽器的特性，快速地實作跨平台的客製化工具。

本文所介紹的工具皆使用 DocuSky 提供的 widget 進行開發，藉由介紹這些工具的功能與應用範例，我們希望這些工具作為範例能彰顯 DocuSky 的特性，以及未來 DocuSky 在不同數位人文領域上的應用功能之可能。

關鍵字：DocuSky、數位人文平台、文本分析、格式轉換

* 國立臺灣大學數位人文研究中心研究助理，Email: pykenny@gmail.com。

Development and Deployment of Tools Based on DocuSky Platform

Po-yu Hsieh *

Abstract

DocuSky is a web platform designed for digital humanities research. It enables humanity scholars to build personal text-based databases, and to apply tools with analytical functions that may be useful for their research. This platform aims to clarify the roles of humanity researchers and analytical tool developers in order to facilitate cooperation. Under DocuSky's structure, most of the available work is performed on client-side browser, and this attribute brings benefits to both researchers and developers. For researchers, tools constructed in the form of web pages provide them with a relative familiar interface, thus reducing difficulties and troubles in using them. For developers, this attribute enables them to quickly develop customized cross-platform tools designed for specific purposes.

In this article, we introduce five tools making use of DocuSky's functionality. They use DocuSky's widgets to perform different tasks, and show efficiency and convenience on analysis of text data or constructing personal database. We hope the demonstration of these tools can shed light on DocuSky's main advantage, as well as stimulate thinking of further applications to humanities research.

Keywords: DocuSky, platform for digital humanities, text analysis, data transformation

* Research Assistant, Research Center for Digital Humanities, National Taiwan University. Email: pykenny@gmail.com.

一、 前言

一般而言，文本數位典藏(Digital Archive)的資料庫通常盡其可能提供人文學者研究上所需要的分析功能，如全文檢索、後分類(post-classification)、相關文件推薦、視覺化、搭配外部工具（如地理資訊系統、其他外部資料庫）等。然而人文學者隨著研究課題的相異或變遷，既有的功能開始不能滿足新的研究需求，而開始要求系統提供新的分析功能。系統的開發技術人員若選擇擴充系統功能滿足這些需求，一來系統會變得益加複雜，導致開發新功能的難度隨著時間增加，二來在也可能造成系統維護上的問題；若因維護考量選擇不擴充系統功能，則使用者可能會選擇放棄使用資料庫，回頭使用傳統的研究方式。本屆會議座談中所提出的 DocuSky 系統（杜協昌，2016 年 11 月）為這項兩難問題提出一項解決方案，即「工具與文本相互分離」原則。在這個原則之下，研究者對於資料庫的內容若有新的分析需求，可以選擇尋找提供相關功能且支援 DocuSky 的公開工具，或是委託第三方的資訊技術人員開發客製化的工具，以取代原本尋求系統的開發技術人員的途徑。

在上述原則下，DocuSky 選擇以網頁作為分析工具的主要介面。此種建置網頁工具為主的概念具有下列的優點：對於研究者而言，由於大部分研究者有透過瀏覽器使用網頁的經驗，因此操作網頁形式的分析工具能降低工具的學習難度，也因為只需要透過瀏覽器就能使用，研究者無需另行安裝程式就能快速地開始使用這些工具；對於資訊技術人員或工具開發者而言，由於大部分電腦都具備近期或最新版本的瀏覽器，這些瀏覽器不論種類，大多能支援主流的作業系統，以及涵蓋大部分瀏覽器端 JavaScript 規範[1]的功能，這些特性得以讓工具易於擁有跨平台(cross-platform)的性質，並且降低不同瀏覽器與平台間的維護難度。

DocuSky 提供工具開發者 Web API（Web Application Programming Interface，網路應用程式介面）以及數種 widget（套件工具）。開發者可以選擇直接在工具程式中使用 API，或是利用 widget 簡化一部份的程序。例如圖 1 是一個能夠閱讀與編輯 DocuSky 個人資料庫中文獻集內容的網頁，該網頁中使用的 widget 在確定使用者登入後會列出該使用者目前建庫的文獻集列表。開發者可以設置使用者在 widget 中選擇載入指定文獻集以後的動作，在此例中，當網頁從 DocuSky 載入使用者指定的文獻集後，會列出載入文獻集的文件內文，以及提供使用者編輯與將更動後的文件儲存回 DocuSky 的功能。

[1] 目前主要的規範為 ECMA Script，主流瀏覽器主要支援的為第五版。關於 ECMA 目前版本的規範，請參考網址：http://www.ecma-international.org/publications/standards/Ecma-262.htm

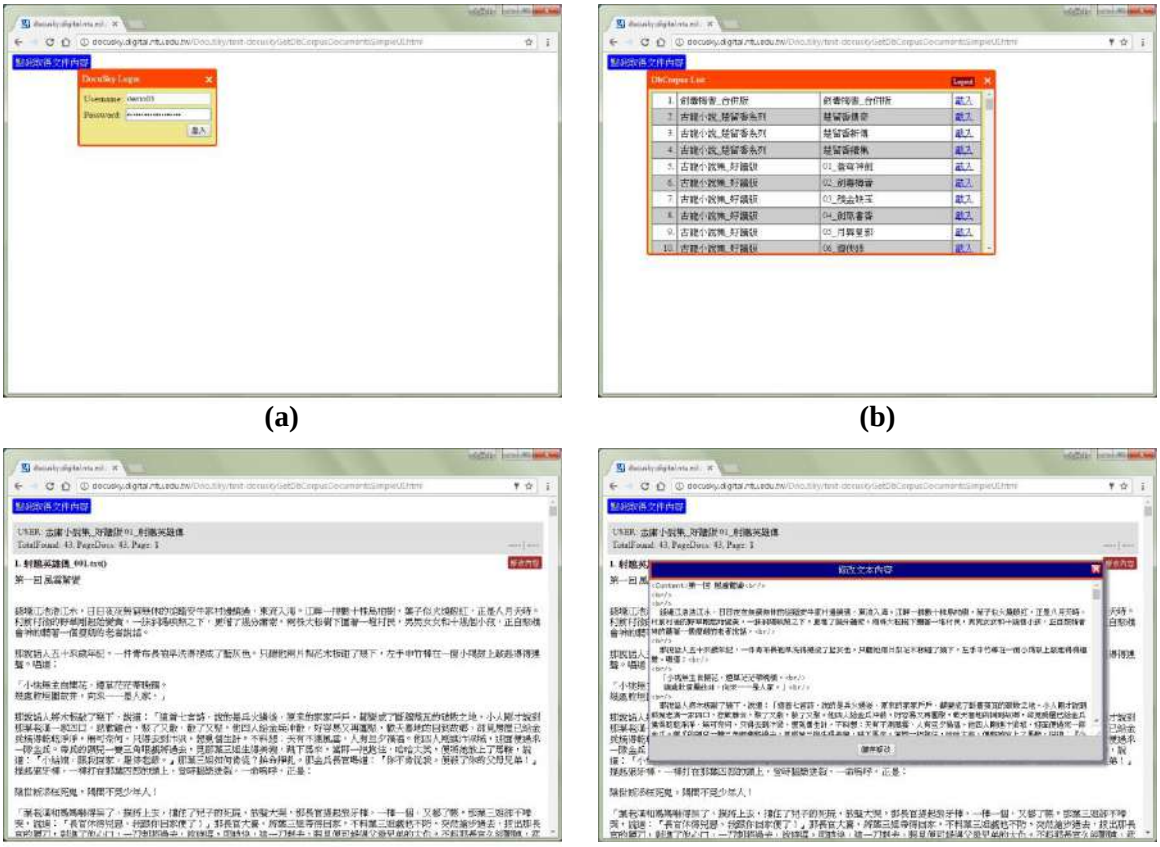


圖 1　DocuSky 所提供的 widget 使用範例網頁。(a)使用者在登入前 widget 會要求以 DocuSky 帳號登入。(b)在使用者登入後，widget 則會列出個人資料庫的文獻集列表。(c)使用者載入選擇的文獻集後，會依照 widget 中的設定，顯示載入的文件內文。(d)widget 亦提供更新文件的功能，讓工具可以修改並上傳文件。

本文將介紹數項使用 DocuSky 套件的網頁工具，這些工具擁有不同的功能與用途：「文獻集字頻統計工具」能夠提供使用者文獻集的基本資訊；「詞彙統計分析工具」、「度量衡轉換工具」、「文本風格分析工具」為研究者提供文獻集的分析功能；「STAML 格式轉換工具」讓研究者得以利用數項標記工具對文本進行不同類型資訊的標記，並生成能夠提供給 DocuSky 建庫的檔案格式。

二、　文獻集字頻統計工具

（一） 工具簡介

DocuSky 目前提供的離線(off-line)工具包含基本的全文檢索、文件後分類與分類詞彙統計，但是使用者對於新建立的資料庫可能有得知一些基本資訊的需求，如文件大小

與字數統計等。理論上能夠在使用者建庫時於伺服器端進行計算，並在使用者從 DocuSky 載入文件時獲取這些資訊，但是不同使用者對於基本資訊的需求可能相異，或是需要的資訊隨著時間增加，這會增加伺服器端維護上的困難。若是讓使用者將此類需求交由工具開發者，在工具中加上計算基本資訊的功能，一來更容易且快速針對使用者需求的變化，以客製提供此類資訊的功能，二來基本資訊通常不需要長久的時間計算來取得，而不會對原本的工具效能造成太大的影響。此一工具模仿文字編輯工具中的「字數統計」功能，在使用者從 DocuSky 載入一份文獻集後，工具會計算該文件集的總字數、使用字數、句長等統計資訊，並將所有已載入文獻集的資料顯示在表格當中。工具亦會計算文獻集中的高頻字提供使用者調閱。

圖 2　文獻集字頻統計工具畫面

（二） 範例：國小國文課本的字頻資訊比較

在這個範例中，我們將 103 年翰林版國小課本共十三冊的單冊內容建立個別的文獻集，並在統計工具中依序載入。所有載入的文獻集都會在「文件集列表」區塊中顯示該文獻集的基本統計資料，能夠藉此觀察各項統計數值的變動。載入所有文獻集後顯示的資訊如表 1 所示。從這些數據我們大致能夠得知一些訊息，例如「總字數」代表文獻集的文字總計（不包含標點符號、空白等），而在每冊課本的課數相異不大的情況下，我們能夠得知教材設計者預期學生能夠閱讀文章的長度的變動；「單字數」代表文獻集中不重複單字數的總計，而「累計單字數」則代表該文獻集和所有前面文獻集中不重複單字數的總計，因此我們也能從「累計單字數」的數值得知教材設計者對於不同年級預期的應習單字數。我們亦可在「高頻字比較」區塊中比對兩冊課本的全文字頻排名前 100 的高頻字，表 2 列舉各年級下學期課本中字頻排名前 15 的高頻字。

表 1　翰林版 103 年小學國文課本各冊文件基本資訊統計

冊次	課文篇數（文件數）	總長度	總字數	使用字數	累計使用字數
首冊	10	319	319	148	148
一上	8	399	325	105	210
一下	14	1969	1654	306	403
二上	16	3999	3409	595	709
二下	16	4229	3569	706	966
三上	16	7337	6263	969	1299
三下	16	7348	6313	1045	1574
四上	16	8562	7386	1193	1817
四下	16	10229	8896	1346	2078
五上	14	10471	9263	1324	2285
五下	14	11714	10125	1444	2450
六上	14	12308	10830	1518	2616
六下	12	10773	9536	1554	2779

表 2　翰林版 103 年小學國文課本下學期各冊高頻字表

序次	一下	二下	三下	四下	五下	六下
1	小	的	的	的	的	的
2	我	一	一	一	一	一
3	的	我	我	我	我	是
4	了	了	了	是	不	不
5	一	小	小	不	了	我
6	天	不	不	了	是	在
7	大	上	子	子	有	人
8	來	是	是	在	人	了
9	說	著	有	上	在	們
10	你	說	大	人	說	他
11	在	在	這	到	來	這
12	裡	有	來	有	們	有
13	好	來	上	著	到	生
14	不	好	著	來	著	和
15	子	媽	就	看	地	學

三、 詞彙頻率統計工具

（一） 工具簡介

此工具為臺灣大學數位人文中心與馬克斯普朗克科學史研究所(Max-Planck-Institut für Wissenschaftsgeschichte, MPIWG)的合作計畫中所開發的工具。這項工具能夠從本機端檔案或 DocuSky 建立要分析的文獻集與分類詞彙表，進而統計分類詞彙在各文獻集中的統計資訊，最後輸出成 CSV(Comma-Separated Values，逗號分隔值)格式檔案讓使用者做進一步的分析。有關於這項工具在上述合作計畫中的成果，請參考本屆研討會中徐源(Michael Stanley-Baker)博士的報告。

圖 3　詞彙頻率統計工具畫面

（二） 範例一：新聞文章的成語使用統計

在這個範例中，我們以教育部《成語典》[2]中的 5153 筆成語建立詞彙集，對 9997 篇臺灣媒體的新聞文章內文進行統計。在這些文章中，我們共找出 840 個成語。表 3 為全文詞頻(term frequency, tf)排名前 30 的成語列表，並列出詞彙的全文詞頻(term frequency, tf)[3]以及文件頻率(document frequency, df)[4]。

表 3　新聞文章中的成語統計結果

1-10				11-20				21-30			
名次	成語	tf	df	名次	成語	tf	df	名次	成語	tf	df
1	脫穎而出	29	28	11	捲土重來	17	16	21	雙管齊下	12	11

[2] http://dict.idioms.moe.edu.tw/cydic/index.htm
[3] 在一文件集中，某字詞的總出現次數。
[4] 在一文件集中，出現某字詞的文件總數。

2	層出不窮	25	24	12	責無旁貸	16	14	22	水到渠成	12	11
3	不約而同	22	22	13	雪上加霜	15	13	23	小心翼翼	12	11
4	拋磚引玉	20	20	14	不遺餘力	15	14	24	軒然大波	12	12
5	耳目一新	19	19	15	難能可貴	14	14	25	包羅萬象	11	10
6	不可思議	19	16	16	無獨有偶	14	12	26	避重就輕	11	11
7	當務之急	18	16	17	如出一轍	13	12	27	絡繹不絕	11	11
8	現身說法	18	17	18	舉足輕重	13	13	28	背道而馳	11	10
9	名列前茅	18	18	19	破天荒	13	12	29	信誓旦旦	11	11
10	無所適從	18	18	20	四通八達	12	12	30	不謀而合	11	11

（三） 範例二：金庸小說的中醫詞彙統計

在這個範例中，我們以中醫十二經與任脈、督脈的 365 個穴位名稱（包含部分穴位的別稱），以及衛生署出版《臺灣中藥典第二版》[5]一書所收錄 300 項中藥的名稱，對《倚天屠龍記》、《天龍八部》兩部金庸的武俠小說[6]進行詞彙統計。表 4 和表 5 列出兩部小說中部分章節中找出的詞彙名稱與統計資訊。

表 4　金庸《倚天屠龍記》一書部分章節出現的中藥名稱統計結果

回數	出現詞彙數	詞彙(詞頻)
第十二回	22	當歸(6), 獨活(5), 防風(4), 水蛭(3), 遠志(3), 乳香(1), 五靈脂(1), 人參(1), 大黃(1), 天麻(1), 柴胡(1), 沒藥(1), 牛膝(1), 牛黃(1), 白芷(1), 百合(1), 紅花(1), 羌活(1), 茯苓(1), 蘇木(1), 血竭(1), 附子(1)
第十三回	8	牛黃(7), 血竭(7), 獨活(2), 水蛭(1), 當歸(1), 知母(1), 遠志(1), 防風(1)
第三十八回	4	前胡(2), 三七(1), 獨活(1), 荷葉(1)

表 5　金庸《天龍八部》一書部分章節出現的穴位名稱統計結果

回數	出現詞彙數	詞彙(詞頻)
第五回	18	膻中(18), 氣海(9), 少商(7), 中極(2), 天突(2), 會陰(2), 關元(2), 三間(1), 不容(1), 中都(1), 天池(1), 廉泉(1), 承漿(1), 曲骨(1), 氣衝(1), 石門(1),

[5] 行政院衛生署臺灣中藥典編修小組，《臺灣中藥典（第二版）》（臺北市，衛生署中醫藥委員會，2013）。

[6] 此範例採用「好讀網」(http://www.haodoo.net/)的電子化版本。
天龍八部：http://www.haodoo.net/?M=book&P=65；倚天屠龍記：
http://www.haodoo.net/?M=book&P=57。

		神道(1), 神門(1)
第十二回	11	中都(2), 不容(1), 中渚(1), 大都(1), 太白(1), 崑崙(1), 液門(1), 石門(1), 肩貞(1), 陽池(1), 養老(1)
第二十回	9	人中(4), 大都(3), 不容(1), 二間(1), 前谷(1), 勞宮(1), 崑崙(1), 血海(1), 關門(1)
第三十五回	16	意舍(6), 不容(3), 大包(3), 天池(2), 玉枕(2), 胃倉(2), 陽池(2), 陽綱(2), 人中(1), 光明(1), 大都(1), 天井(1), 巨骨(1), 脾俞(1), 關衝(1), 陽谷(1)
第四十三回	14	不容(5), 廉泉(3), 血海(3), 陽白(3), 風府(3), 百會(2), 光明(1), 大椎(1), 太乙(1), 承泣(1), 梁門(1), 玉枕(1), 關元(1), 頰車(1)

四、 度量衡轉換工具

（一） 工具簡介

度量衡轉換工具的功能為截取文本中描述長度或面積等的計量結果的文字，將之轉換成其它度量衡系統下的量值。轉換功能原身為台灣歷史數位圖書館研究工具集的度量衡單位換算系統，預設有四種系統（清制、日制、公制、英制）下四種量（長度、面積、容量／體積、重量）的度量衡標準，提供使用者使用。

度量衡轉換工具在處理規模較大的文本中，能幫助使用者使用程式自動截取出符合系統已知度量衡系統、與數值系統的量值，並附上轉換的結果。未來預計將預設度量衡系統陸續增加，並加入自行上傳或輸入度量衡資料與轉換的機制、與將截取出的計量與轉換結果匯出成資料詮釋資料的機制，讓使用者能自行上傳文本中所使用的度量衡資料，再以本工具自動進行截取與換算並將成果加進資料庫中。期盼本工具逐漸完善後能協助使用者在制式的作業上節省大量的人力。

（二） 範例：古地契的度量衡資訊擷取

此處以兩筆 THDL 系統所收錄的地契文件為例，截取並顯示文中的度量衡資訊。兩份地契的內文與工具的分析結果（圖 4，圖 5）如下：

基隆三沙園 立盡根出稅地基字人盧江河，有承祖父應份得自己曠地壹所，址在奎籠三沙園。東至山坎腳為界；西至水退海底為界；南至王曾官墻壁為界；北至溝墘地面為界。闊伍丈，長二拾丈，四至明白為界。
今因乏銀別置，除問親堂不肯承受，外托中引就與陳順官稅出，地基過三年稅銀來佛銀參拾圓正，面約參年後每年地基銀拾圓，其銀即日仝中收訖，地基即付陳順官，起蓋築埕圍墻，任從番〔翻〕蓋，聽其所便，傳子孫及孫永為陳順官物業，不得異言生端，保此地基係是江河自己承祖父得物業，與叔姪兄弟無干，亦無重張典掛他人為碍，及內歷不明，為有不明江河出頭抵擋，不干銀主之事，此係兩

願，各無反悔，今欲有憑，立盡根出地基稅字壹紙為照。
再批明：即日仝中收銀參拾員。
再批明：後至劉九連厝前曠腳界。
作中人 陳細
代書人 吳玉 光緒參年參月 日 立盡根出稅地基字人 盧江河 知見人 母親 王氏 契尾 布字伍千參百號
光緒拾貳年肆月[7]

擷取	度量衡	制度	度量衡擷取轉換	
			制度	轉換結果
伍丈	長度	清制	清制	5丈
			日制	5丈2尺8寸
			公制	16公尺
			英制	17碼1英呎5.921517英吋
	長度	日制	清制	4丈1弓2尺3寸4分9釐10毫
			日制	5丈
			公制	15公尺15公分
			英制	16碼1英呎8.45705英吋
二拾丈	長度	清制	清制	2引
			日制	21丈1尺2寸
			公制	64公尺
			英制	3鏈3碼2英呎11.685587英吋
	長度	日制	清制	1引8丈1弓4尺3寸9分9釐10毫
			日制	20丈
			公制	60公尺60公分

圖 4　系統換算結果。輸入來源：國立台灣大學，《台灣歷史數位圖書館》，檔名：〈cca100003-od-ta_04239_000108-0001-u.xml〉

新竹縣東興庄二十張犁、北勢十張犁、舊社
大租戶 高福
立轉典庄業契人金聯安，有接典林朱氏等承劉朝珍原典王胡氏名下，割出東興庄業戶管收二十張犁、北勢十張犁、舊社等處。庄佃應納課租共穀捌百參拾陸石，仝業戶吳金桔輪流兩年值收壹年，應帶完正供穀壹百捌拾肆石柒斗肆升，又採買穀貳拾伍石，又錢粮廊餉耗羨銀貳拾貳元參角貳瓣伍周，如遇奉文加領兵米穀貳拾伍石。今因乏銀應用，先問原典主不能取贖，外托中引就高指一出首接典，即日仝中面議，照原價銀壹千陸百陸拾貳元交聯安收訖。其東興庄佃租凡遇值年盡付接典，銀主自立戶名，管收完課，將來聽原主王胡氏，自備典價取贖，聯安不得過問，保此庄業並無重張典掛他人違碍。如有來歷不明或來轉典以前有拖欠公私債項未清，聯安應支理與接手者無干。口恐無憑，合立轉典契壹紙，檢同上手契約參紙，共肆紙，付執為照。
即日，仝中收過典價番銀壹千陸百陸拾貳元足訖。

[7] 國立台灣大學，《台灣歷史數位圖書館》，檔名:〈cca100003-od-ta_04239_000108-0001-u.xml〉

陳且漣號
為中人 莊子榮號
張斗南號
代筆人 何長鑑號
道光拾柒年拾貳月 日 立轉典契人
契尾 高指一承典金聯安二十張犁等處大租用價銀壹千壹百四拾六兩七錢八分
納稅銀參拾肆兩肆錢零參厘四毫
布字捌百六拾壹號右給新竹縣業戶高指一 准此。
光緒拾五年拾月 日 [8]

擷取	度量衡	制度	度量衡擷取轉換	
			制度	轉換結果
捌百參拾陸石	容量	清制	清制	836石
			日制	479石8斗8升9合6勺116259.796237立方分
			公制	86565公升191.68毫升
			英制	297夸特4蒲式耳1加侖2夸脫1品脫3及耳4.234216液量盎司
	容量	日制	清制	1456石2斗9升8合4.059712勺
			日制	836石
			公制	150802公升361.6毫升
			英制	518夸特2蒲式耳1配克1加侖3夸脫1品脫3及耳4.276984液量盎司
壹百捌拾肆石柒斗肆升	容量	清制	清制	184石1斛2斗4升
			日制	106石4升6合4勺136607.41528立方分
			公制	19129公升250.6112毫升
			英制	65夸特5蒲式耳1坎寧1配克1加侖3夸脫3及耳2.801111液量盎司
	容量	日制	清制	321石1斛3斗1升4合0.759798勺
			日制	184石7斗4升
			公制	33324公升435.744毫升
			英制	114夸特4蒲式耳1配克1夸脫1品脫0.180897液量盎司
貳拾伍石	容量	清制	清制	25石
			日制	14石3斗5升7勺428407.39786立方分
			公制	2588公升672毫升
			英制	8夸特7蒲式耳1加侖1夸脫1品脫1及耳3.828774液量盎司
	容量	日制	清制	43石1斛4升9合5.9348勺
			日制	25石
			公制	4509公升640毫升
			英制	15夸特3蒲式耳1坎寧1配克1加侖3夸脫1品脫3及耳2.681728液量盎司
貳拾伍石	容量	清制	清制	25石
			日制	14石3斗5升7勺428407.39786立方分
			公制	2588公升672毫升
			英制	8夸特7蒲式耳1加侖1夸脫1品脫1及耳3.828774液量盎司
	容量	日制	清制	43石1斛4升9合5.9348勺
			日制	25石
			公制	4509公升640毫升
			英制	15夸特3蒲式耳1坎寧1配克1加侖3夸脫1品脫3及耳2.681728液量盎司
壹千壹百四拾六兩七錢八分	重量	清制	清制	71斤10兩7錢7分9釐10毫
			日制	71斤46匁9分4厘4.350587毛
			公制	42公斤776.04078公克
			英制	3夸特10磅4.85537盎司

[8] 國立臺灣大學，《台灣歷史數位圖書館》，檔名:〈cca100003-od-ta_04239_000108-0001-u.xml〉

圖 5　系統換算結果。輸入來源：國立台灣大學，《台灣歷史數位圖書館》，檔名：〈cca100003-od-ta_04239_000108-0001-u.xml〉。

第一筆契約為租約，其中描述租地的長寬：「伍丈」與「二拾丈」。工具集預設的度量衡系統中「丈」可能為清制系統與日制系統，因此，系統會將兩種可能都列出來，並進行換算，可以看出清制的「丈」與日制的「丈」長度略有不同，本章節中的計量屬於哪個度量衡系統，須交由使用者做判斷、選擇。

第二筆為典契，契約中列出米糧容量：「拾陸石」、「壹百捌拾肆石柒斗肆升」、「貳拾伍石」、「貳拾伍石」，與交易的銀重量值「壹千壹百四拾六兩七錢八分」。容量值皆被判定可能為清制系統與日制系統，重量值則判定為清制系統。

五、　文本風格分析工具

（一） 工具簡介

DocuSky 將網頁作為分析工具的界面帶來的其中一項益處，在於能夠讓其他學者易於取得分析工具，只要使用者將使用 DocuSky 的網頁分析工具架設在伺服器上，其他的 DocuSky 使用者便能將其個人資料庫的文獻集載入工具進行其他研究。

此工具實作第四屆數位典藏與數位人文國際研討會（DADH 2012）中，杜協昌博士於《利用文本採礦探討《紅樓夢》的後四十回作者爭議》（杜協昌，2012 年 11 月）一文中使用的數個分析方法，包含比較兩份文本中高頻字的兩樣本 t-檢定、文本採礦函數、前後綴詞分析、以及內文中提及楊智傑的 Rank-Frequency Distance(RFD)分析方法。我們將這些分析方法實作成網頁型態的 DocuSky 工具後，便能快速地將文本先透過 DocuSky 建庫，接著從工具載入後進行分析。

（二） 範例：《劍毒梅香》前後段寫作方式的比較

《劍毒梅香》為古龍早期之武俠小說作品。連載中途古龍便因故停止創作，只完成小說前四集共十四回的內容。而後續內容則由出版社託上官鼎代筆完成。《劍毒梅香》由上官鼎接手的續寫點曾經有所爭議，由於後續出版的版本的章節經過調整[9]，導致無法直接從後期的出版版本的章節判斷續寫點，只能從寫作風格和劇情的轉變進行猜測，而各方對於續寫點的看法也不一。《神君別傳》一書也因為初版後就絕版的關係難以取得。最後這些問題從最初版本（清華本）的取得、以及《神君別傳》手稿的公布與再版[10]而得以解決，由清華本的前四集內文與《神君別傳》開頭的前情提要內容，目前能夠

[9] 如南琪版的十五回、及現行風雲時代版的五十回。
[10] 陳曉林，＜古龍的遊戲之作：神君別傳＞，《劍毒梅香（下）》（臺北市，風雲時代，2009），頁 213-

確認《劍毒梅香》一書的續寫點。在這個範例中，我們希望能透過閱讀全文以外的方式，觀察《劍毒梅香》由不同作者所完成兩個部分的寫作差異。而以下會基於杜文中的分析過程來使用文本分析工具進行兩文本的比對。

我們使用的《劍毒梅香》電子化文本[11]的來源採用初版的章節編排，共四十回。以敘寫點的第十四回作為分界，將原來文本分為前十四回與後二十六回兩份文本。然後將前十四回（以下簡稱前半部）以(5, 5, 4)分為三個單元，後二十六回（以下簡稱後半部）以(5, 5, 5, 5, 6)分為五個單元。

首先利用這八個單元進行杜文中的 Yang's RFD 和 2-sample t-test。計算結果的矩陣如圖 6 與圖 7 所示。從 Yang's RFD 的分析結果以矩陣顯示進行觀察，會發現後半部的五個單元與前半部中三個單元的 RFD 普遍比與其他後半部單元之間的 RFD 高，大致能得到後半部各單元的寫作習慣相近的結論。但是比對的結果也顯示前半部第三個單元與全文其他七個單元的 RFD 皆偏高，這顯示該單元的用字習慣與全文其他部份皆有較大的差異。

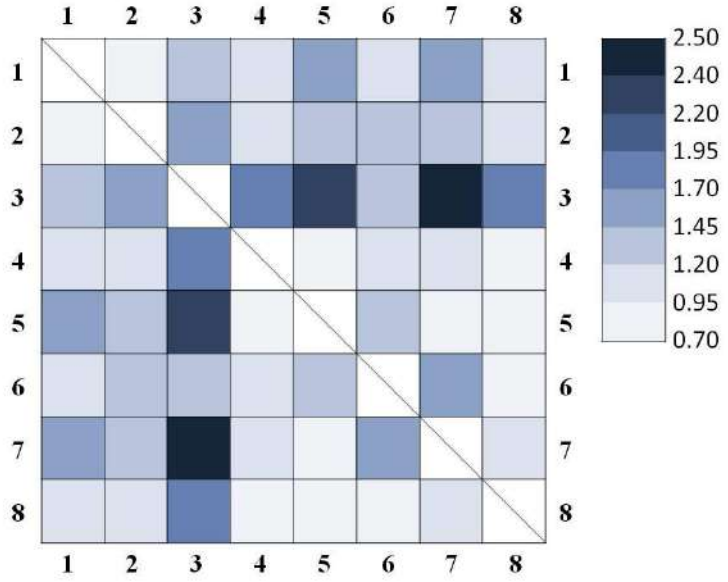


圖 6　計算《劍毒梅香》各單元間之 Yang's Rank-Frequency Distance 所形成的矩陣。

其中 1-3 為前半部之單元，4-8 為後半部之單元。

215。

[11] 此範例採用「好讀網」(http://www.haodoo.net/)的電子化版本。原文另外還有「尾聲」一章，由於內容過短，我們將該段內容與鄰近的第四十回合併。網址：http://www.haodoo.net/?M=Share&P=1031。

而從對全文字頻前 200 名的高頻字率[12]進行兩樣本 t-檢定的結果顯示，前半部各個單元和後半部單元之間字頻有顯著差異的單字數量比前半部其他單元之間相比的數量多，能夠得到前半部各單元寫作習慣相近的結論。但是在後半部單元的比較結果中，後半部的第四單元除了前半部的單元外，與後半部的第一、二兩單元的用字差異也偏高。

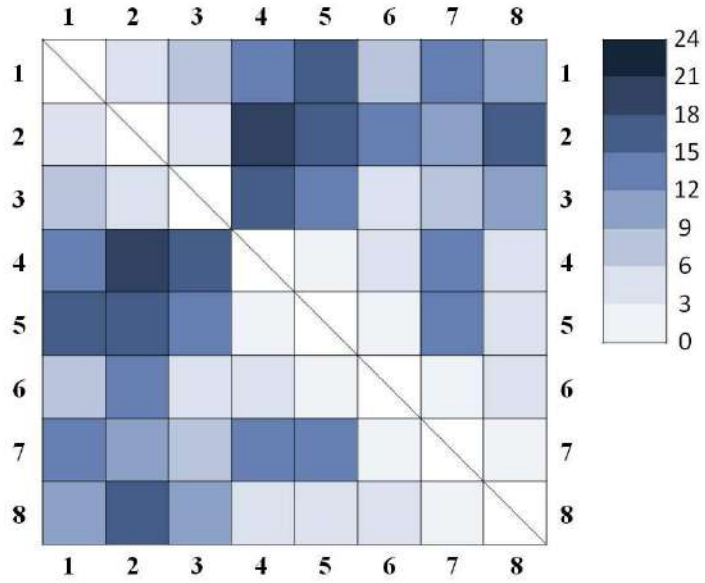


圖 7　計算《劍毒梅香》各單元間，字率平均值有顯著差異(p-value=0.01)的高頻字數量所形成的矩陣。其中 1-3 為前半部之單元，4-8 為後半部之單元。

使用杜文的採礦函數所得到的分數前 20 高的單字如表 6 所示。該表顯示「哩」、「啊」、「嘿」等語助詞，「嚓」、「叮」等狀聲詞，以及「跨」、「瞥」等動詞僅出現在後半部，而並未於前半部出現，由此就可以觀察到兩文本中用字模式上的不同。除了一些和劇情發展相關的單字[13]以外，從一些單字能夠觀察到前後半部在寫作習慣上的差異。以下試舉兩例：

一、「工」字的使用：「工」字僅出現在《劍毒梅香》的後半段。若利用綴詞工具計算「工」字的後綴詞（表 7），會發現「工夫」是該字在後半部最常出現的使用方式。而「工夫」一詞又具有兩種詞義，如下面的兩個篇段：

[12] 由於《劍毒梅香》各回之間的長度差異偏高，若採用原論文以詞頻（於單份文件中的出現次數）進行比較，則結果易受到文件長度影響。此處採用原論文中建議的替代計算方式，即「字率」。計算方式為：(文件中的出現次數 ÷ 文件總字數（不包含標點符號、空白等))。

[13] 如「詰」字的差異來自後半段劇情出現的招式「詰摩神步」、「元」字和後半部劇情出現的招式「歸元四象陣」與角色「金元伯」、「金元仲」有關。

表 6　《劍毒梅香》前後段的採礦函數分數前 20 名的單字列表（k = 0.02）

1-10					11-20				
名次	單字	前半段出現章節數	後半段出現章節數	分數	名次	單字	前半段出現章節數	後半段出現章節數	分數
1	哩	0	23	45.231	11	祥	0	13	26.000
2	啊	0	22	43.308	12	元	0	13	26.000
3	瞥	0	18	35.615	13	裂	0	13	26.000
4	嘿	0	17	33.692	14	溢	0	12	24.077
5	跨	0	16	31.769	15	叮	0	12	24.077
6	詰	0	14	27.923	16	崖	0	12	24.077
7	峰	0	14	27.923	17	拭	6	0	22.429
8	幌	0	14	27.923	18	抄	6	0	22.429
9	工	0	14	27.923	19	卅	0	11	22.154
10	嚓	0	13	26.000	20	暇	0	11	22.154

表 7　《劍毒梅香》後半部中計算「工」綴詞的結果（後綴字長度 1）

名次	單字	出現次數
1	工夫	7
2	工作	4
3	工整	2
4	工造	1
5	工打	1
6	工開	1
7	工心	1
8	工度	1

「……但那繪聲繪影的傳說到底也令他有點不安，不過他始終以為那多半是冒牌貨罷了，那知目下這個蒙面人那手振劍的工夫分明是七妙神君的特殊標誌……」（十九回）

「……海天雙煞陡然醒悟，他們已知中了對方的毒，由於不麻不癢的感覺，知道這毒性非淺，他們連檢驗毒傷的**工夫**都沒有……」（三十五回）

在這兩個篇段中，前者的詞義為「本領、用功致力的程度」或「武術」，而後者的詞義為「空閒時間」。而又由於「工夫」與「功夫」兩詞可通用，我們再觀察後半部中「功夫」一詞出現篇段中的詞義，會發現用法皆為前者的詞義，如：

「……只是連無恨生這等人物都未發覺船底被做了手腳，這些潛水夫的**功夫**可想而知了！」（十五回）

「平凡上人卻嘴帶笑容，一語不發。眾人雖不知這是什麼**功夫**，但都知這比金魯厄踩陷青磚又不知難了幾倍。」（二十九回）

這代表著在後半部中，有出現同義詞「工夫」和「功夫」兩字的併用，這有可能是作者筆誤，或是出版社校正上的問題。在前半部中並沒有出現「工夫」，但是有使用「功夫」一詞，觀察前半部中「功夫」出現的篇段，會發現除了其中一個的詞義為「空閒時間」之外，詞義皆為「本領、用功致力的程度」或「武術」：

「天絕劍、地絕劍不由大怒，那知那人根本不將他們放在眼裏，看了蘇映雪一會兒，臉孔一板，道：『你們還呆在這幹什麼，盧老頭子現在沒有**功夫**替你們醫病，你們快滾。』」（十三回）

此外，「功夫」一字在前半部中的總出現次數為 19 次，但是在後半部中的總出現次數則多達 153 次（「工夫」一詞則為 7 次）。使用頻率上的差異也是透過觀察這些分析結果所能發現兩者在寫作方式差異上的線索。

二、「拭」字的使用：「拭」字是僅在前半部出現的單字，若利用綴詞工具統計「拭」字的綴詞，並觀察這些綴詞出現的橋段（表 8），能發現該字常出現於「擦淚」情節的段落，如：

「侯二伸手**拭去**眼簾上的淚珠，強笑道：『故事講完了。』」（七回）

「他想抬起手來替她**拭去**頰上的淚珠，但是他覺得手臂竟全然失去知覺，像是已不屬於自己身體的一部分了。」（九回）

表 8　《劍毒梅香》前半部中計算「拭」綴詞的結果（後綴字長度 1）

名次	單字	出現次數	出現該詞的段落節錄
1	拭去	3	他想抬起手來替她**拭去**頰上的淚珠…
2	拭了	2	辛捷伸手**拭了**拭面上的雨水，又踱回簷下…
3	拭著	1	梅山民用手輕輕**拭著**頷下的微鬚，嘆道…
4	拭面	1	辛捷伸手拭了**拭面**上的雨水，又踱回簷下…
5	拭汗	1	天魔金欹悄悄伸手一**拭汗**，臉上現出痛苦的神色來…

而由觀察後半部中「淚」字出現的相關段落，則會發現提及擦淚的部份常使用「擦」字，如：

「『雲爺爺，你別傷心啦，你心中有事，說給風兒聽，風兒替你解憂。』雲爺爺悚然一驚，飲泣立止，雙袖**擦**淚。」（二十七回）

「她**擦**乾了淚，低聲問道：『大哥，你這大半年到了些什麼地方，伯父的仇報了嗎？』」（三十九回）

使用杜文的採礦函數所得的分數前 20 高的雙字詞如表 9 所示。同樣的，在排除與劇情相關的字詞[14]後，可見前後段寫作用詞上的差異。如「又說」、「非但」等連接詞只出現在前半部，而「一躍」、「退後」等動作相關詞只出現在後半部。此處討論列表中採礦函數分數最高的「他笑」一詞，作為透過雙字詞觀察寫作差異的範例。

「他笑」這個雙字詞只出現在前半部，使用綴詞工具找出該詞的後綴詞的結果如表 10 所示。從結果可以推測在後半部中，作者對於人物做出「笑道」、「笑著」這些動作的描寫可能更為習慣使用人名作為主詞，而非代名詞，如使用「辛捷笑著」而不是「他笑著」，或是習慣在主詞和動詞之間加上其他敘述詞語，如「他哈哈笑道」。若觀察後半部這些詞的出現段落，確實也能支持這樣的推論。

表 9　《劍毒梅香》前後段的採礦函數分數前 20 名的雙字詞列表（k = 0.02）

1-10					11-20				
名次	單字	前半部出現章節數	後半部出現章節數	分數	名次	單字	前半部出現章節數	後半部出現章節數	分數
1	他笑	12	0	43.857	11	風一	0	16	31.769
2	施出	0	20	39.462	12	神妙	0	16	31.769
3	吳凌	0	20	39.462	13	攻勢	0	15	29.846
4	喝一	0	19	37.538	14	時見	0	15	29.846
5	當下	0	18	35.615	15	透出	0	15	29.846
6	一躍	0	17	33.692	16	情他	0	15	29.846
7	神劍	0	17	33.692	17	又說	8	0	29.571
8	大衍	0	17	33.692	18	非但	8	0	29.571
9	退後	0	17	33.692	19	懷裏	8	0	29.571
10	十式	0	16	31.769	20	衍十	0	14	27.923

[14] 如「吳凌」、「風一」和後半段出現的角色「吳凌風」有關；「大衍」、「十式」、「衍十」和後半段的招式「大衍十式」有關。

表 10　《劍毒梅香》前半部中搜尋「他笑」綴詞的結果（後綴字長度 1）

名次	單字	出現次數	出現該詞的段落節錄
1	他笑道	4	**他笑道**：「想不到今日我也做了個摧花之客。」
2	他笑著	4	於是**他笑著**連連點頭道…
3	他笑聲	4	**他笑聲**未了，已是一聲驚呼…
4	他笑了	2	一種「後繼有人」的喜悅，使得**他笑了**。
5	他笑容	1	**他笑容**一斂，說道…

表 11　《劍毒梅香》前後段計算「笑道」綴詞，詞頻排名前 10 的詞彙列表（前綴字長度 1）

前半部				後半部			
名次	單字	全文詞頻	文件頻率	名次	單字	全文詞頻	文件頻率
1	冷笑道	10	7	1	一笑道	32	17
2	捷笑道	10	5	2	冷笑道	14	11
3	哈笑道	8	5	3	大笑道	11	10
4	聲笑道	4	3	4	哈笑道	9	7
5	他笑道	4	4	5	人笑道	7	4
6	手笑道	4	3	6	微笑道	6	5
7	飛笑道	4	2	7	捷笑道	5	5
8	嬌笑道	4	3	8	風笑道	4	4
9	娘笑道	3	1	9	口笑道	3	2
10	怪笑道	3	3	10	然笑道	3	2

表 11 為「笑道」綴詞的計算結果。可以從當中觀察到前半部「笑道」綴詞的詞頻前 10 高的詞彙當中，前綴詞和人名有關的有「捷笑道」（辛捷）、「飛笑道」（于一飛）、「娘笑道」（繆七娘），而後半部則有「捷笑道」和「風笑道」（吳凌風）。當中和主角相關的「捷笑道」（辛捷）、「風笑道」（吳凌風）在前半部中出現的總次數和相近（前半部出現 10 次，後半部共出現 5＋4＝9 次）。若再考慮前後半部的篇幅大小差異甚大（前半部共 14 回，後半部共 26 回），可以推論在使用「笑道」一詞上，後半部傾向於使用「（人名）－（敘述詞）－笑道」的文句結構，而較少使用「（人名）－笑道」的結構。

六、 STAML 格式轉換工具

（一） STAML 簡介

在與文本相關的數位人文研究當中，一部份的研究需要對文本進行標記(annotation)

以進行後續處理與分析，例如標記專有名詞後進行詞頻統計或文件分類，或是標記地名後與地理資訊系統(GIS)結合。此類標記工作或相關工具通常以特定格式儲存，例如：(1)針對文本資訊交流所設計的標記規範。如 TEI(Text Encoding Initiative)；(2)標記工具進行標記後產生的輸出格式，而此格式通常亦能被同一標記工具讀入，以利後續標記工作。如 MARKUS[15]所輸出的 HTML 格式檔案；(3)資料庫系統內部所使用的儲存格式，其標記內容能用於系統中的特定功能（例如顯示標記詞彙）。如臺灣歷史數位圖書館(Taiwan History Digital Library, THDL)中儲存文本的 XML 格式。這些標記格式的規格因目的不同而有所相異，然而研究者可能因研究內容與方法而需要使用多個標記工具，若是不同工具的標記格式之間不相容，則會造成處理上的障礙。一種基本的解決方法為開發兩種工具標記格式間的轉換工具，讓其他工具能夠使用自身工具標記後的文本。該項解決方法的主要問題在於，開發者的資源有限，只能在眾多標記格式中選擇一部分提供轉換功能。另外一種方法則是提供將格式轉為廣為公開的標記規範的功能，如一些工具提供轉換成前述(1)類型中提及的 TEI 格式之功能。然而 TEI 因為著重於保存文本資訊，並未含括標記工具除標記以外產生的額外資訊，如工具中的設定參數等。

為了解決此一問題。曹又霖的論文（曹又霖，2016 年 8 月）提供一項解決方案，即 STAML(Simple Text Annotation Markup Language)。該文以數位人文研究為重心，統整人文學者在處理文本資料時常用的標記資訊種類，以及標記工具產生的其他資訊，為不同 XML 格式的標記格式間提供一項中介格式。一旦一項工具建立了與 STAML 之間的轉換規則，則其他已經建立轉換規則的工具使用者就能經由 STAML 將該工具標註過的文件快速轉換成自身工具使用的標記格式。反之，該項工具的使用者也能利用其他工具標註過的文件產生的 STAML 格式檔案轉換為自身工具的標記格式。這也讓工具或格式的開發者只需要專注於與 STAML 的格式對應關係，而無須考慮與其他工具標記格式的對應關係（圖 8）。

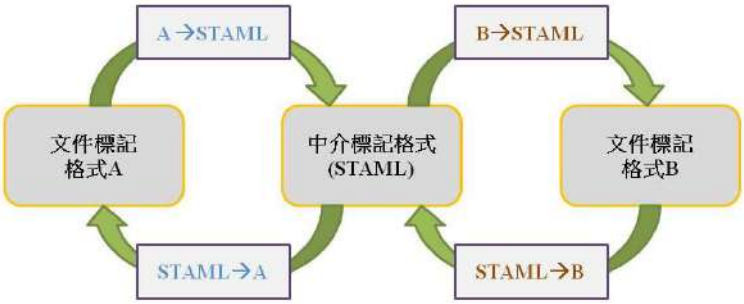


圖 8　STAML 的格式轉換示意圖

[15] http://dh.chinese-empires.eu/beta/

（二） 工具簡介

在曹文中亦提供標記格式與 STAML 格式間的轉換程式，該程式以 JavaScript 撰寫，只要給予 JSON(JavaScript Object Notation)格式的參數告知程式雙方標籤與屬性的對應關係，即可進行兩種格式間的轉換，該程式讓實作網頁格式的格式轉換工具變得更為容易。本工具即為曹文中範例所使用的轉換工具，能夠進行 MARKUS 標記工具產生的 HTML 檔案、THDL 所使用的 XML 格式、及 STAML 三種格式之間的轉換，前兩者格式與 STAML 之間直接透過曹文的程式進行對應轉換，而這兩種格式之間的轉換則透過先轉為 STAML 格式，再進行 STAML 與另一項格式的轉換來達成。下述的範例示範如何藉由該轉換工具使得利用數個標記格式不同的工具的標記工作變得可行。

（三） 範例：結合 MARKUS 與中西曆轉換工具的標記流程

在本例中，我們希望能將一份古地契檔案的文字內容使用 MARKUS 進行標記，然後利用接受 STAML 格式的中西曆轉換工具標注時間資訊，最後將這些檔案轉換為 DocuSky 可接受的上傳格式後利用 DocuSky 的功能進行建庫。在這項過程中除了 STAML 外，牽涉另外兩項格式，分別是 THDL 與 DocuSky 使用的 XML 格式(thdlExportXml)、以及 MARKUS 輸出的 HTML 格式。

首先，我們先從 THDL 取得一筆古地契[16]的資料，該資料的 thdlExportXml 格式內容如圖 9 所示。我們將這個檔案經由 STAML 轉換工具轉換成 MARKUS HTML 格式後，便能讓 MARKUS 讀入該檔案進行其他資訊的標記。在原本的 thdlExportXml 檔案當中，內文部份當中並未對人名、地名等資料進行任何標記，我們用 MARKUS 將人名和地名標出後，輸出成新的 MARKUS HTML（圖 10），而標記內容在 MARKUS 上的顯示如圖 11 所示。

[16] 國立台灣大學，《台灣歷史數位圖書館》，檔名:〈at1998_016_059.txt.xml〉。

```
<ThdlPrototypeExport>
    <documents>
        <document filename="at1998_016_059.txt">
            <corpus order="0">古契書</corpus>
            <compilation>高雄市立歷史博物館古文書</compilation>
            <compilation_name>高雄市立歷史博物館古文書</compilation_name>
            …
            <title>立承佃高居觀</title>
            <author order="0">高居觀</author>
            <author_name>高居觀</author_name>
            <topic order="0">開墾契</topic>
            …
            <doc_content>立承佃高居觀，今在嶺頂村李德榮處承出根面全民田壹號，坐產洋下洋，土名六石上坵即王珩，經丈壹畝弍分柒厘正，承來耕作，面約遞年認納乾淨上籠租穀肆百肆拾觔正，早六晚四送還理清，不敢色箊，不敢少欠，如有少欠，係保代賠，其田即听李家召佃別耕，且觀不得阻留之理。今欲有憑，立承佃字壹紙為照。 ↵
            宣統二年二月初三日立承佃字高居觀 ↵
            保佃：叔祖高體康  代筆：族叔高仁旋 ↵
            信字 ↵</doc_content>
            …
            </document >
    </documents>
</ThdlPrototypeExport>
```

圖 9　原始 THDL 的古地契 XML 檔案（局部內容）。來源：國立台灣大學，《台灣歷史數位圖書館》，檔名：〈at1998_016_059.txt.xml〉

```
<div class="doc" filename="at1998_016_059.txt" ···>
    <pre contenteditable="false">
        <span class="passage" type="passage" id="passage0">
            <span class="commentContainer" value="[]">···</span>
            立承佃<span class="markup manual unsolved fullName noCBDBID" type="fullName" cbdbid="">高居觀</span>，今在<span class="markup manual unsolved placeName" type="placeName">嶺頂村</span><span class="markup manual unsolved fullName noCBDBID" type="fullName" cbdbid="">李德榮</span>處承出根面全民田壹號，坐產<span class="markup manual unsolved placeName" type="placeName">洋下洋</span>，土名<span class="justAdd markup manual unsolved placeName" type="placeName">六石上坵</span>即<span class="markup manual unsolved fullName noCBDBID" type="fullName" cbdbid="">王珩</span>，經丈壹畝弍分柒厘正，承來耕作，面約遞年認納乾淨上籠租穀肆百肆拾觔正，早六晚四送還理清，不敢色篠，不敢少欠，如有少欠，係保代賠，其田即听李家召佃別耕，且觀不得阻留之理。今欲有憑，立承佃字壹紙為照。 ↵
            宣統二年二月初三日立承佃字<span class="markup manual unsolved fullName noCBDBID" type="fullName" cbdbid="">高居觀</span>↵
            保佃：叔祖<span class="markup manual unsolved fullName noCBDBID" type="fullName" cbdbid="">高體康</span> 代筆：族叔<span class="markup manual unsolved fullName noCBDBID" type="fullName" cbdbid="">高仁旋</span>↵
            信字↵
            <div id="metadataHidden" style="display: none">
                <corpus>古契書</corpus>
                <compilation>高雄市立歷史博物館古文書</compilation>
                <compilation_name>高雄市立歷史博物館古文書</compilation_name>
                ···
                <title>立承佃高居觀</title>
                <author>高居觀</author>
                <author_name>高居觀</author_name>
                <topic>開墾契</topic>
                ···
            </div>
        </span>
    </pre>
</div>
```

圖 10　經過 STAML 工具轉換，由 MARKUS 進行標記後的 MARKUS HTML（局部內容，人名與地名的標記以淺色顯示）。注意原檔案中的其他資訊被保留在”metadataHidden”這個標籤當中。

圖 11　圖 10 的 MARKUS HTML 檔案在 MARKUS 當中的顯示情形

雖然 MARKUS 提供了歷史人名、官職名稱、地名等資訊的自動標記功能，然而在時間資訊方面僅提供年號的自動標記功能。在標記流程中，我們希望能標記文件中的時間資訊，並且提供西曆的對應年份資訊。這邊我們提供了以 STAML 為輸入與輸出格式的中西曆轉換工具，該工具會比對出時間字串並進行標記，並且加注對應的西曆時間。我們可以將進行要標記時間的 MARKUS HTML 透過 STAML 轉換工具轉換成 STAML 格式，再將檔案輸入中西曆轉換工具進行標記（圖 12）。最後，使用 STAML 轉換工具將格式轉換回 MARKUS HTML 輸入 MARKUS，結果如圖 13。和圖 11 比較，可以看到除了原先的標記外，又多出了中西曆轉換工具所找出的時間標記。

```
<STAML>
    <metadata>
        <corpus>古契書</corpus>…
    <metadata>
    <article>
        <chapter><section>
            立承佃<person>高居觀</person>，今在<location>嶺頂村</location>…
            <datetime userdataRef="0">宣統二年二月初三日</datetime>立承佃字<person>高居觀</person>
            …
        </chapter></section>
    </article>
    <application><appdata>…</appdata></application>
    <userdata><data refID="0"><note>西元 1910 年</note></data></userdata>
</STAML>
```

圖 12　經過中西曆轉換工具標記後的 STAML 格式檔案。（局部內容，工具加上的標記以淺色顯示）

圖 13　(a)將圖 12 中的 STAML 檔案轉為 MARKUS HTML 後，經 MARKUS 讀入後的顯示情形。(b)中西曆轉換工具加入的西元年份資訊也在轉換格式後被 MARKUS 所顯示。

最後，將標記完成的文本輸出成 MARKSU HTML 格式檔案，由 STAML 轉換工具將檔案轉為 thdlExportXml 格式檔案，並上傳至 DocuSky 建庫。建庫後從 DocuSky 的個人資料庫網頁上觀看文件內容，可以看到透過 MARKUS 與中西曆轉換工具標記的詞彙會被標色顯示，左側的後分類欄位也顯示了人名與地名的類別，如圖 14 所示。

圖 14　將標記完的文件上傳至 DocuSky 建庫後，該項資料庫的網頁畫面。

七、 結論與展望

本文列舉了數項結合 DocuSky 功能的工具與應用實例，這些工具的功能有繁有簡，涵蓋協助研究者進行分析以及自行建庫兩個面向的功能。需注意本文列舉的工具並未涵蓋所有目前 DocuSky 所提供的功能，像是 DocuSky 的 Web API 和一部分 widget 提供的上傳與更新文件功能，和這些功能相關的工具有待後續開發。實務上也會需要整合數項 DocuSky 功能的工具，例如提供對純文字文本進行標記，再轉為 DocuSky 能接受的格式並上傳建庫如此一整套流程的工具。

DocuSky 尚處於草創階段，未來將會提供更完善的功能，如增加支援建庫的檔案格式、更完整的 Web API、與 widget 功能的改良等，以擴增 DocuSky 能夠支援開發的工具範疇和提昇工具的開發效率。而利用網頁容易公開使用的性質，亦能開拓讓外界對 DocuSky 工具進行改良或應用於其他研究的可能性。同時，STAML 也使得既有的文本標記工具與格式易於和 DocuSky 整合，協助研究者建立內含多項資訊的資料庫。我們希望以這些工具做為引子，能夠引發人文研究者與資訊技術者雙方思考 DocuSky 賦予數位人文研究的可能性，進而發掘更多可行的應用。

參考文獻

杜協昌（2016 年 12 月）。〈DocuSky – 個人文字資料庫的建構與分析平台〉。第七屆數位典藏與數位人文國際會議。

杜協昌（2012 年 11 月）。〈利用文本採礦探討《紅樓夢》的後四十回作者爭議〉。第四屆數位典藏與數位人文國際會議。

曹又霖（2016 年 8 月）。〈文本標記格式的轉換與應用〉。國立臺灣大學資訊工程研究所碩士論文。

Panel B

研究的變革：數位分析與文史學科的未來

Innovations in Research: Digital Platforms and the Future of the Humanities

Panel B

研究的變革：數位分析與文史學科的未來

主持人	祝平次（國立清華大學中文學系副教授） Ping-tzu Chu (Associate Professor of Department of Chinese Literature, National Tsing Hua University)
發表人	楊秀芳（國立臺灣大學中文學系退休教授） Hsiu-fang Yang (Emeritus Professor of Department of Chinese Literature, National Taiwan University) 葉秋蘭（國立臺灣大學中文學系計畫助理） Chiu-lan Ye (Project Assistant of Department of Chinese Literature, National Taiwan University)
題目	數位分析與漢語方言研究 Digital Platforms and Chinese Dialect Studies
發表人	施懿琳（國立成功大學臺灣文學系退休教授） Yi-lin Shih (Emeritus Professor of Department of Taiwanese Literature, National Cheng Kung University)
題目	一位台灣古典詩研究者對數位人文的想像和運用 The Possible Applications of Digital Humanities in the Study of Taiwanese Classical Poetry
發表人	薛化元（國立政治大學臺灣史研究所特聘教授） Hua-yuan Hsueh (Distinguished Professor of Graduate Institute of Taiwan History, National Chengchi University)
題目	數位人文與歷史研究的互動：理論與實際 The Interaction between Digital Humanity and Historical Research：Theory and Practice

Panel B

研究的變革：數位分析與文史學科的未來

台灣在 1984 年，距今 32 年，也就是個人電腦開始流行之際，中央研究院就已經開始製作全文檢索。當時隨著二十五史全文檢索的建置，還有共同研究平台。然後到 1990 代的末期，陳郁夫也建立了寒泉檢索系統、羅鳳珠也試著建立線上平台來進行詩詞研究和輔助教學，並嘗試了一些不同的路徑的研究方法。當然，還有其它各式各樣的資料的建置。之後，在 2002 年，國家啟動了數位典藏計畫，長達十年的時間，使得資料的建置更為多元、多樣；研究工具的開發也有了新的樣貌。整體來看，在這三十多年裏，台灣在數位人文研究的各個方面都有累積，但就文史學界的研究方法意識和具體的研究成果來講，則還有待今日之後的發展。到底數位人文研究方法和不同的文史學科未來的發展可以有什麼關係？過去比較隱形的數位研究方法將來是否可能變成各學科的正式訓練的一部分？未來我們可以期待什麼樣的數位人文研究成果？本場次的 Panel，將由楊秀芳、施懿琳、薛化元三位文史學者分別從小學研究、文學研究與史學研究，來討論以上問題。

Panel B
Research Innovations: Digital Platforms and the Future of the Humanities

At the time of personal computers being popular, the Academia Sinica began to develop a full-text search database in 1984, i.e. 32 years ago. With the construction of Scripta Sinica database, a collaborative research platform was also brought into reality. In the late 1990s, Professor Chen Yu-fu (陳郁夫) established the Han Quan (寒泉) search system, and Professor Lo Feng-ju (羅鳳珠) also attempted to create an online platform to conduct poetry studies, to assist teaching, and to try out different research approaches. In fact, there were a variety of databases also constructed. As the National Digital Archives Program launched in 2002, the databases have become more diverse and the research tools has also developed a new look during the 10-year span. Overall, the development of Digital humanities in Taiwan in the past 30 years has accomplished in many aspects; however, we are still seeking further development of research methods and solid results in literature and history studies. What relation could be developed between the research methods of digital humanities and the disciplines of literature and history? Could the digital research methods, which is less visible in the past, become a part of formal training of the disciplines? What research results of digital humanities could we expect to have in the future? In this panel, the above issues will be discussed from the perspectives of the traditional Chinese Linguistics and Philology, Literary Studies, and Historical Research by the panelists Hsiu-fang Yang (楊秀芳), Yi-lin Shih (施懿琳) and Hua-yuan Hsueh (薛化元) respectively.

數位分析與漢語方言研究
Digital Platforms and Chinese Dialect Studies

楊秀芳*、葉秋蘭**

摘　要

本報告以〈方言形態變化中的存古與創新〉為例，說明「漢字古今音資料庫」在研究過程中所提供的協助。

報告分三個部分：（一）本資料庫建置簡介，（二）數位分析與漢語方言研究，（三）本資料庫轉型再開發的可能性。

（一）本資料庫建置簡介

本資料庫最初是行政院國家科學委員會「漢學研究資料庫‧漢字古今音電腦檢索系統」研發計畫(1997-2000)下的成果，這個三年計畫結束之後，我們利用歷年的國科會計畫經費，繼續輸入語料並加強資料庫的檢索功能，至今 19 年。本資料庫所提供的可查詢漢字總共兩萬多字，主要根據宋代韻書《廣韻》收字。在這框架之下，輸入上古、中古、近代、現代四個階段的語音資料，並有日本、韓國和越南三地的域外譯音。上古階段包括先秦、兩漢音系；中古階段包括魏晉、南北朝、隋唐音系；近代音收錄元代《中原音韻》以及明代《洪武正韻》的音系；現代音部分則包括各大方言區的代表方言。目前總共輸入語音資料 1,211,610 筆。

建置本資料庫的最初構想，是希望能有一個可以交叉檢索的資料庫，以便於做古今漢語的分析研究之用。要達到這個目的，資料庫的檢索條件就必須盡可能細緻，因此從一開始，我們輸入的方言材料就是經過分析之後的聲母、韻母和聲調，使用者也還可以根據需要設計條件進一步分析，例如將韻母再分析出介音、主要元音、韻尾，以「音位」為單位來觀察；資料庫框架所提供的古音輸入項則包括各字所屬的調類、韻目、字母、清濁、等第、開合等，同樣盡可能切分細緻，以方便做不同條件的數位分析之用。

* 國立臺灣大學中文學系退休教授，Email: yanghf@ntu.edu.tw。
** 國立臺灣大學中文學系計畫助理。

（二）數位分析與漢語方言研究

本節以〈方言形態變化中的存古與創新〉為例，說明本資料庫的三種功能（1.快速獲取大量資料，2.掌握語音演變規律，3.判讀例外現象），同時也指出本資料庫在輸入內容和語言研究上有待開發的部分。

〈方言形態變化中的存古與創新〉探討「上」「下」的音義問題：古漢語「上」「下」是語義相反的一對指事詞，都有上、去兩讀。指示方位的「上」讀去聲，「下」讀上聲，它們都可作謂語，具有不及物動詞的性質。當不及物動詞出現在使動用法中，語義有了較大的改變，聲調也相應作了區隔：「上」從原來的去聲改讀為上聲，「下」從原來的上聲改讀為去聲，表現為「四聲別義」的構詞現象。「上」「下」這種平行發展的音義變化，見於《經典釋文》的紀錄，也程度不等的保留在現代方言中。

「上」「下」聲母俱屬全濁，今方言是否在語音上反映古漢語的形態區別，關鍵在於古全濁上與濁去今調是否合併。部分閩、粵、客方言能在聲調上區別古全濁上與濁去，這些方言的這兩類音讀有些反映了古漢語的語義用法區別，表現為一種「存古」現象；有些則與古漢語的語義用法不同，是「創新」演變的結果，而且各方言的「創新」變化又有不同。利用比較研究方法，比較這些不同的方言變體，可以重建古漢語的形態變化，補充《經典釋文》所未能完備的紀錄。從現代方言「上」「下」的形態變化看，古漢語形態變化是一種自然的語言現象，並非漢魏經師的人為讀書音。

進行這個研究，需要從眾多方言篩選出能夠區別古全濁上與濁去的方言，以便觀察它們有關「上」「下」的語義和用法。在研究過程中，我們得力於本資料庫的協助，簡述其中三種功能如下：

1.快速獲取大量資料

以粵語來說，目前本資料庫共收八片 79 個方言點的語料，我們在很短的時間內，從廣府片 32 個方言點中找出能區別古全濁上、濁去的 28 個方言點，其中有 24 個方言的「上」能區別古全濁上、濁去，但只有 7 個方言的「下」有全濁上、濁去之別，其餘方言的「下」幾乎都讀來自濁去的陽去調。這些能區分古全濁上、濁去的方言，成為本研究集中討論的對象。

2.掌握語音演變規律

在這個研究中，我們需要知道古全濁上聲字和濁母去聲字演變到現代各方

言讀為什麼聲調。利用本資料庫的「古音檢索條件」，可以立刻查檢出演變到現代各方言的聲調調類和調值。

3.判讀例外現象

本資料庫除了可以用來分析音變規律，對於有疑義的不規則音讀，可以設計條件，交叉檢索，判讀這些不規則音讀究竟是另外的音變規律，或者是真正的例外現象。

目前本資料庫輸入的都是個別的字音，因此詞彙、語句中可能出現的語音變化，以及與語義用法有關的語音區別，本資料庫都暫時無法呈現。因此在上述研究中，資料庫提供了很好的分析音變的功能，但對於語義用法的細緻區別，必須再查檢古籍文獻及方言田野調查報告，才能作深入的分析。

本資料庫名為「漢字古今音資料庫」，本來就是以提供語音資料為主，未來如果有機會，希望也能建置語義方面的資料，使它擁有更強大的功能。

（三）本資料庫轉型再開發的可能性

受人力資源及技術開發的限制，本資料庫目前只有靜態的資料呈現。我們期待能夠以此為基礎，不斷擴充改進，開發出新的格局，讓本資料庫成為具有人工智慧的好幫手。

Panel C

數位文本與文學研究

Digital Texts and Literature Studies

Panel C

數位文本與文學研究

主持人	蔡瑜（國立臺灣大學中國文學系教授） Yu Tsai (Professor of Department of Chinese Literature, National Taiwan University)
發表人	羅珮瑄（中央研究院中國文哲研究所計畫助理） Pei-hsuan Lo (Program Assistant of Institute of Chinese Literature and Philosophy, Academia Sinica)
題目	數位文本與文學研究 Review on Digital Texts and Literature Studies
發表人	謝薇娜（中央研究院中國文哲研究所博士後研究） Severina Balabanova (Postdoctoral Fellow of Institute of Chinese Literature and Philosophy, Academia Sinica)
題目	群體傳記的數位分析：以葉德輝(1864-1927)出版《乾嘉詩壇點將錄》和《東林點將錄》為例 A Digital Analysis of Group Biographies：A Research on *Qian-Jia shi tan dianjiang lu* and *Donglin dianjiang lu* Published by Ye Dehui (1864-1927)
發表人	林偉盛（國立臺灣大學中國文學系博士生） Wei-cheng Lin (Doctoral Student of Department of Chinese Literature, National Taiwan University) 鄧賢瑛（國立臺灣大學中國文學系計畫助理） Hsien-ying Teng (Program Assistant of Department of Chinese Literature, National Taiwan University) 莊德明（中央研究院數位文化中心研究助技師） Der-ming Juang (Assistant Research Specialist of Academia Sinica Center for Digital Cultures)
題目	中國詩歌格律之重探與數位化研究：兼談「漢詩格律分析系統」的設計 Revisiting and Digitalizing of the Metrical Regulations of Chinese Pentasyllabic Poetry：Discussion of the Design of *Metrical Regulation Analytic System of Chinese Poetry* Project
討論人	邱偉雲（湖北經濟學院新聞與傳播學院中文系副教授） Brian, Wei-yun Chiu (Associate Professor of Hubei University of Economics) 雷之波（中央研究院中國文哲研究所助研究員） Zeb Raft (Assistant Research Fellow of Institute of Chinese Literature and Philosophy, Academia Sinica)

Panel C

數位文本與文學研究

本組 panel 乃是由中央研究院中國文哲所劉苑如先生主持之「觀念·話語·行動：數位視野下中國/台灣多元現代性研究──葉德輝印刻書流通與人物往來的數位研究」與台灣大學中國文學系蔡瑜先生主持之「漢詩音韻分析系統」兩個科技部補助之數位人文專題研究計畫共同組織而成，兩者皆是從 2015 年 8 月開始執行，前者為二年期計畫，後者為三年期計畫，至今已具有階段性的研究成果，並於計畫執行期間發展出多項議題，涉及傳統人文學與數位人文學在問題意識、研究方法、知識技術、資源分配等各個面向的差異與互動。Panel 由三篇論文組成，除了發表階段性研究成果之外，亦有聚焦的核心關懷，從文學研究者的問題意識出發，數位人文學的成立與發展將如何建構一個新興的研究環境，包含數位文本的生產、主題資料庫的建立、資訊的加值服務，以及資料庫和系統設計背後所牽涉的理論內涵和技術問題。

首先〈數位文本與文學研究〉一文乃是站在文學研究者的角度，帶著高度的歷史意識和反省自覺，探討 2000 年以來數位技術與人文研究的互動歷程，其中特別針對數位文本的生產與再生產活動，如何改變研究者的閱讀方法、思維模式、以及問題意識的提出，同時也探索從傳統人文學研究過渡到當代的數位人文學，哪些東西不變、哪些東西改變、面臨跨學門跨領域的合作，雙方應該如何溝通與分工，又將面臨怎樣的挑戰，就知識結構與學術轉型的發展脈絡，提出觀察和建議。

其次〈群體傳記的數位分析：以葉德輝(1864-1927)出版《乾嘉詩壇點將錄》和《東林點將錄》為例〉一文乃是以葉德輝 1907 年出版的清代舒立《乾嘉詩壇點將錄》以及明代王紹徽《東林點將錄》兩個案例，探究數位文本所創造的知識系統。該文首先將原始的文本結構加以解構，透過資料庫串聯與時空座標改造成新的研究文本，藉以分析人際脈絡所產生的知識結構，並進一步利用數位工具如 Pajek 展開社群網絡分析。此一研究案例不僅僅探究明清時代對於文本的看法為何？清末民初的藏書家與出版家葉德輝又抱持怎樣的態度？他如何在晚清的人際關係和話語系統當中利用出版活動重新創造文本，而這種文本的生產與再生產歷程，與當代科技環境之下的數位文本相較，彼此的性質有何差異？傳統文本分析方法是否可以容納數位研究以擴展詮釋的角度？又或者說數位知識與技術能夠為傳統文本提供怎樣豐富的資源，以挖掘文本深層的脈絡，提供文學研究者更多更大的想像？而這樣的技術介入文學研究之後，又將形成怎樣的對話、衝擊和挑戰？

再者〈中國詩歌格律之重探與數位化研究：兼談「漢詩格律分析系統」的設計〉一文乃是從 20 世紀以來當代中國詩律學研究的學術課題出發，反省長期以來詩律研究所面臨的瓶頸，在於各種後見之明的格律條件往往過於繁雜瑣碎，而傳統研究方法又無從負擔巨量的音韻材料與文本檢驗，更遑論歷史語言學視野中語音演變和方言的介入，使得詩律學的研究進程遲滯不前，且難以取得共識。另一方面，隨著數位典藏的推動與積累，中央研究院「小學堂」資料庫奠基了豐富精確的語音文本，而各種類型的文獻資料

庫亦提供了古典詩歌的數位文本，在充分的條件匯聚之下，「漢詩格律分析系統」於 2015 年開始構思、撰寫程式、設計系統，以數位技術擅長處理巨量資料的優勢，嘗試回應傳統詩律學研究力有未逮的課題。本文除了說明「漢詩格律分析系統」的階段性成果之外，同時著重於設計過程當中，如何從文學研究的問題意識出發，假設和建立六朝時期五言詩的格律模型，經過巨量統計與分析之後，修正格律條件，再進一步考證文獻，呈現出南朝沈約四病、初唐元兢四病與元兢調聲三術等三種格律模型的發展歷程，並且以系統的設計和思維方式，整理廿世紀以來現代學術的詩律研究，歸納出可被規範的數種格律條件，建立近代格律模型。在這個過程裡，文學研究與數位技術之間有許多交鋒的機會，雙方各自有各自的思維邏輯，在互動之際型塑問題意識和研究方法，這樣的跨領域合作已不再僅以某一方馬首是瞻，而是各具主體、細緻分工，體現出作為方法的數位人文學，有其發展的前景和道路。

Panel C
Digital Texts and Literature Studies

This panel is jointly organized by the two research projects: "Ideas, Discourse, Motion: A Digital Perspective on China's/Taiwan's Multifaceted Modernity: A Digital Research of the Circulation of Ye Dehui's (1864-1927) Block Printing and of Social Networks" led by Prof. Liu Yuanju from Academia Sinica, and "Metrical Regulation Analytic System of Chinese Poetry" led by Prof. Tsai Yu from National Taiwan University, both projects starting in August 2015 on a two- and three-year term respectively. These projects already have research results, focusing on subject matter concerning problem awareness, research methods, technical knowledge, and distribution of resources. The panel presents three articles. The topics analyzed include how the development of the digital humanities can construct a new research environment, including the production of texts, the establishment of a topical database, the enhanced functions of data, and the theoretical and technical concerns when designing databases.

The article "Review on Digital Texts and Literature Studies" is written from the point of literary research, and with a high degree of historical awareness and introspection reviews the interaction between digital technology and humanities research, the production and reproduction of digital texts, how it affects the reading process, the way of thinking and the problem awareness, the changes involved in the transition from traditional research in humanities to digital research. It also reflects on the communication between the two and gives some suggestions about the development context of the structure and transformation of knowledge.

"A Digital Analysis of Group Biographies: A Research on *Qian-Jia shi tan dianjiang lu* and *Donglin dianjiang lu* Published by Ye Dehui (1864-1927)" examines the system of knowledge created by the pattern of a digital text based on Qing Dynasty Shu Li's *Qian-Jia shi tan dianjiang lu* and Ming Dynasty Wang Shaohui's *Donglin dianjiang lu,* both published by Ye Dehui in 1907. The analysis deconstructs the original text and creates a new one by linking different databases as well as by the spatial and temporal distribution of its components to examine knowledge structure created by the social networks using *Pajek* as a digital tool. The article seeks answers to questions such as the views on texts during the Ming-Qing period, how Ye Dehui created new texts in the context of social relationships and discourse as seen in his publishing activities, the differences of such texts from the digital text, the extent to which it is possible for traditional textual analysis to incorporate digital research in the interpretation process, the role of imagination in digging into the deepest levels of a traditional text through knowledge built by digital technologies, the dialogue, conflicts and challenges thus created.

"Revisiting and Digitalizing of the Metrical Regulations of Chinese Pentasyllabic Poetry: Discussion of the Design of *Metrical Regulation Analytic System of Chinese Poetry* Project"

takes as a starting point contemporary prosody studies in China and reflects on the difficulties facing this research field. The development of prosody studies has been hindered by the complexities of the conditions concerning the metrics, the impossibility to deal with a large amount of phonological material and the lack of enough experiments with texts, not to mention the speech evolution in historical linguistics and intervention of dialects. In the promotion of digital archives the *Xiaoxue tang* 小學堂 database has established a solid and precise sound of text, with a number of other types of textual databases providing digital version of ancient verse texts. In 2015 the project "Metrical Regulations of Analytic System of Chinese Poetry" started to design the respective programs in which it uses the advantages of the way digital technology handles big data in the attempt to research the issues that have not been addresses by traditional prosody. The article will discuss preliminary findings of the project, and will focus on how to create a metrical model of the five-syllable poems from the Six Dynasties when designing the database. It also shows how to correct the conditions for the verse after the examination of large amounts of data. Further textual analysis reveals the development of three types of verse: Shen Yue's "four defects" (441-513), Tang Dynasty Yuan Jie's "four defects" and "three methods adjusting tonal prosody". It systematizes the research on prosody during the 20th century from the point of system design, and generalizes the norms of the different verse types, thus establishing a modern verse model. Problem awareness and analytical methods emerge with the interaction between literary research and digital technology with their specific logic, where digital humanities describe a method with rich potential and perspectives.

群體傳記的數位分析：以葉德輝（1864-1927）出版《東林點將錄》和《乾嘉詩壇點將錄》為例

論文初稿，請勿引用

謝薇娜*

一、 前言

傳統文本分析往往將焦點放置在文字層面的詮釋，將文本視為其他文本所建構脈絡的一部分。在當代科技發展的環境之下，從數位人文研究的角度進行分析，原本只是佔據整體脈絡一部分的文字文本，還可以包括相關人名、地名的傳記資料，甚至於相關聯的圖片亦可蒐羅其中，理論上資料種類可以無限擴張，藉此重新賦予文本新的概念與意涵。傳記作為特殊的文體，在數位的時空分析中，如何創造一種新的文學經驗和典範？數位人文又如何改變文本所形成的知識系統？此二者是本文的重要論點。本文將以葉德輝 (1864-1927) 在 1907 年出版，由明代王紹徽所著的《東林點將錄》和清人舒立撰寫的《乾嘉詩壇點將錄》為案例，探究數位文本型態所創造的知識系統。第一個文本介紹乾嘉兩朝詩人事跡，第二個則以《水滸傳》的人物綽號與著名東林黨人相比擬。本文將在兩個層次上展開論述：一、介紹葉德輝出版點將錄的情景以及點將錄的文本特色；二、用數位人文分析呈現文本中的人物關係特徵。本文將原始文本進行解構，在透過空間分佈產生的新文本，分析人際脈絡所產生的知識系統，進一步探究數位人文工具的加入如何改變文本的概念。明清時代對文本的看法，是否與清末民初出版業者葉德輝對文本功能和意義的想法一致？這與當代科技環境之下的數位文本性質差異何在？傳統的文本分析方法，是否可以容納數位研究以擴展詮釋角度？本文試圖探討以上問題之際，更想談及數位人文研究者的挑戰：如何引入數位人文和數位工具到文史研究脈絡當中。

數位人文研究奠基於大量數據的量化分析。筆者認為，針對點將錄所代表的群體傳記進行量化研究，比起對個別傳主進行研究的好處在於：宏觀分析使文本

* 中央研究院中國文哲研究所博士後研究員。

語境化，這種語境讓我們更能了解不同文本出現的環境及其意義。我們可以將一個文本放在更廣泛的環境裡，探究其風格的發展特色、文本模式和主題的演變邏輯，以及不同文體在不同時代的盛行等面向。[1]

因此，本文擬從探索兩種點將錄文本的性質特色、這類文獻形式結構的意義、文本中羅列人物的篩選邏輯和標準，進而討論此類文體在中國文學脈絡中的價值和含意。

在傳統的閱讀和研究方法中，遠離那個時代的學者必須累積許多歷史知識，才能正確詮釋點將錄人物所代表的意義。然而，在數位人文時代，很多參考文獻可以透過建立資料庫和數位人文工具的分析，進而重建此類關聯，甚至呈現傳統文本無法表示的關係（特徵）。本文將以群體傳記的數位化為主要分析方法，探究數位工具對群體傳記帶來的新研究角度，展示以點將錄群體傳記為主的「遠讀法」(distant reading)[2]，如何促進對該時代歷史和文學特徵的宏觀分析。群體傳記的數位化和研究，已經擁有相當豐富的研究成果，例如：法鼓佛教學院對歷代高僧傳的資料加以數位化，並將相關成果應用於地理資訊系統，除能掌握其時空資訊外，更創造一個嶄新的研究資料系統。在群體傳記資料庫方面，哈佛大學、中央研究院歷史語言研究所與北京大學合作建置的《中國歷代人物傳記資料庫》(CBDB)，運用群體傳記學為理論基礎，對人物傳記（7-19 世紀）進行蒐集、整理與數位化，之後可以透過它所提供的分析工具對人際網絡進行分析。CBDB 的處理方式為「仿真陳述」(factoids)，即某一事實 (fact) 所涉及的材料對這一事實的說法。本文運用 CBDB 的優點在於，此系統能夠作為「群體傳記學」的分析工具，也可用於「社會網絡分析」(Social Network Analysis)。群體傳記數位研究這種研究方法，允許研究者從不同角度分析一個人的社會關係、了解他的影響力，或不同人物在一個時段內的屬性關係，以及他們各自的影響力。

二、葉德輝出版點將錄的動機以及點將錄的文本特色

葉德輝的出版動機必須放在時代脈絡中進行思考，他獲得文本的方式以及交換文本的現象，反應出時代的知識需求。一方面，藉此形成的人際網絡對於知識系統的生產相當重要。另一方面，文本的內容呼應清末民初的政治和社會環境等課題，因此探究葉

[1] 參看 Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History.* Urbana, Chicago, and Springfield: University of Illinois Press, 2013, p. 27.
[2] 參看 Moretti, Franco. *Graphs, Maps, and Trees: Abstract Models for A Literary History.* London-New York: Verso, 2005.

德輝出版的點將錄有助於分析清末民初對文本價值的概念。[3]

葉德輝出版過許多與傳記相關的書籍，如屬於地方志的《巴陵人物志》，以及《高力士外傳》、《楊太真外傳》、《安祿山事蹟》等屬於唐代小說一類的作品。[4]據葉德輝於《重刻足本詩壇點將錄敘》中所言，葉氏提及從小讀過袁枚 (1716-1797)《隨園詩話》、王昶(1724-1806)《湖海詩傳》等屬於詩文評點一類的書籍，更於文中強調對當時詩壇產生相當濃厚的興趣，萌生思欲出版詩集的想法。雖然部分詩集已經亡佚，但他找到《乾嘉詩壇點將錄》後，便下定決心要出版此部作品。在《重刻足本詩壇點將錄敘》一文裡，葉德輝於此呈現個人對於《東林點將錄》內容的看法，他認為「雖遊戲之作，能使讀者於百世之下，想像其生平，斯固月旦之公平，抑亦文苑之別傳？」[5]。

考察有關葉德輝出版點將錄的相關記載和材料，可以發現葉德輝在 1907 年八月下旬返回長沙，寓居洪家井，本月二十五日作〈重刊詩壇點將錄序〉一文。此外，〈點將錄附考〉未署年月，附印於《東林點將錄》後，推測或許為同一年所作。[6]

紀錄葉德輝出版活動的〈郎園出版書目〉[7]，收錄一則葉德輝考證《東林點將錄》作者的篇章，此文收錄在《乾嘉詩壇點將錄》所附的《點將錄附考》中。此書的出版時間，推測應該是在光緒三十三年。《乾嘉詩壇點將錄》於光緒三十三年初次刊印時，於書末附有《東林點將錄》，其後更於宣統三年(1911)與新刊《足本詩壇點將錄》合訂為一冊，又將《足本詩壇點將錄》插訂於《乾嘉詩壇點將錄》與《東林點將錄》兩書之間，版式字體全異。〈郎園出版書目〉還提到《乾嘉詩壇點將錄》的題面是由劉鈺所寫。

追溯《乾嘉詩壇點將錄》的來源時，葉德輝說：「光緒丁未，從長沙舊書攤購得同治已巳巾箱本，遂付梓人刊行。旋獲傳抄武進莊氏舊藏足本，較余本少訛字，諸人里貫仕跡，亦較余本稍詳，然所缺者猶多。余據吳鼎雯《詞垣考鏡》、李富孫《鶴征後錄》

---

[3] 在葉德輝與當時人士交換文本活動的問題上，參看沈俊平，〈清末民初版本目錄學家葉德輝與藏書家和版本目錄學家之交往活動〉，《書目季刊》，第 40 卷第 1 期，頁 13-27。

[4] 參見沈俊平：《葉德輝文獻學考論》，臺北：台灣學生書局，2012，頁 157。此外，《高力士外傳》、《楊太真外傳》、《安祿山事蹟》這三部小說有《唐人小說文種九卷》以及《唐開元小說文種九卷》的觀古堂刻本，後者為宣統三年 (1911) 刻本，詳見吳格，眭駿整理：《續修四庫全書總目提要．叢書部》，北京：國家圖書館出版社，2010，頁 123-4。

[5] 本文用舒位等原著; 楊楊編校，《三百年來詩壇人物評點小傳彙錄》，鄭州：中洲古籍出版社，1986 的版本，頁 56-7.

[6] 王逸明、李璞編著，《葉德輝年譜》，北京：學苑，2012，頁 147.

[7] 王逸明，孫有東，劉海翼等整理，〈郎園出版書目〉，收入：王逸明主編，《葉德輝集》，第一冊，北京：學苑，2007，頁 31。

及郡邑詩選、各省誌乘、詩人詩文本集、集中碑傳文字補之，而是書遂臻完善。」[8]

從以上證據可以得知，葉德輝於光緒、宣統年間刊印過兩種刊本。葉德輝刻書時會選擇內容完整的底本進行校勘，在針對文本內容進行考證，此處提及的兩種點將錄文本也不例外。葉氏於 1907 年刊行《乾嘉詩壇點將錄》後，復又得到足本抄本，在經過精心校勘之後，又於 1911 年重新刊刻。

此外，從葉德輝考證《東林點將錄》的文字內容可以得知，葉德輝也把《北略》及《遣愁集》進行過比較[9]，這些例子不只闡述清末獲得文本方式，更顯示出版者透過修改和重寫，對文本的介入狀況，呈現文本在不同時代的歷時變化。值得注意的是，因為政治的關係，《東林點將錄》一書的成書過程很複雜，傳說的作者也比較多，[10]說明葉德輝使用其他種類的文本對《東林點將錄》進行校勘時，《東林點將錄》已經存在許多不同的版本。[11]

《東林點將錄》顯然是在明代東林黨爭之下產生的作品，東林黨背後的歷史事實如何？1594 年顧憲成 (1550-1612) 因延推內閣的大臣時違背皇帝的意志，被削職歸里，與弟顧允成復修東林書院，攜高攀龍、錢一本、薛敖教等人在東林書院講學，議論朝政，得到許多士大夫的支持，學者、士紳聞聲響附，形成集團，被稱之爲「東林黨」。因他們主張開放言路，實行改良，遭到以東廠魏忠賢爲首的嫉視，用王紹徽造編《東林點將錄》等文件，試圖把東林黨人一網打盡。因此，最初的《東林點將錄》有著負面意義，是用《水滸傳》梁山起義者的形象比喻東林黨成員。然而，對梁山起義者的看法隨著時代而產生改變，開始呈現正面形象。於是，清代的《乾嘉詩壇點將錄》已經脫離最初《東林點將錄》的負面意涵，包含著對詩人及其作品的另一種正面肯定。葉德輝發現他擁有的《東林點將錄》文本與《北略》以及《遣愁集》所載者有不同之處，因此加以校對。值得注意的是，在葉德輝校對《東林點將錄》文本的當下，已經出現不同的版本，每次對文本進行修改時，文本都會隨著修改者個人的目的而有所改變。即使創造《東林點將錄》的原始目的在於呈現東林黨和魏忠賢之爭的一個側面，其記載格式仍繼續被後期作家運用。

點將錄有別於傳統的列傳，文本性質相當於評點小傳，內容不免簡略，省略許多人物描寫的細節，焦點放在人物的性格特徵，並用《水滸傳》的形象比喻詩人的風神。

---

8 舒位等原著; 楊楊編校，《三百年來詩壇人物評點小傳彙錄》，《重刻足本詩壇點將錄敘》，頁 56.

9 舒位等原著; 楊楊編校，《三百年來詩壇人物評點小傳彙錄》，頁 51-2.

10 參看（清）永瑢等撰，《四庫全書總目》，卷六二，北京：中華書局，1987，頁 559．

11 關於葉德輝的校勘看法，沈俊平，〈葉德輝對校讎學、目錄學、版本學三者關係的理解〉，《國立中央圖書館台灣分館館刊》，第六卷第六期，頁 28-35 以及文庭孝，劉曉英，〈葉德輝的校勘之功及校勘之法〉，《高校圖書館工作》，第 26 卷第 114 期，2006 年第 4 期，頁 13-15．

雖然《東林點將錄》出現在前，並且其中不少東林黨成員也寫了許多著作，然與《乾嘉詩壇點將錄》相比，前者沒有記載評點的部分，僅用《水滸傳》中 108 條好漢相應的外號，單純地與羅列出來的東林黨成員進行比附。兩個文本基本的構思則以不同人物描述當時政治或文學（詩壇）的某些情形。然而，這個情形與被篩選的人物，以及他們之間的關係有著密切的關聯。

據考察[12]，東林黨總共有309名成員，但只有108人被錄取在《東林點將錄》裡。點將錄發展脈絡自《東林點將錄》開始，中國文學歷史上出現詩壇「點將錄」（如本文的《乾嘉詩壇點將錄》）、詞壇「點將錄」（如：清末朱祖謀《清詞壇點將錄》）、小說「點將錄」等等。點將錄的文體特徵與九品制度密切相關。九品中正來源深遠，可追溯到魏晉時期官吏人才的判斷和甄選。「點將錄」不可或缺的部分是推舉派別領袖、羅列派別成員、甄別派人品第。用《乾嘉詩壇點將錄》的版本進行觀察，領袖是舊頭領托塔天王晁蓋所屬者；成員則是一百單八將；品第者則是天罡地煞之排名先後。在葉德輝的出版活動中，與人格及人才品鑑相關的文本，在其出版事業中佔有一定位置。他曾經輯佚並刊刻《山公啟事》，相關研究指出，此過程隱藏很多有關清代出版環境和文化的訊息，編輯本身「其實可視為一種再創作，也是一種重新的閱讀與詮釋，……作為批評當時的舊弊、新禍的依據，進而剖析這一類文人權衡於舊學與新知之間的時代感覺。」[13]雖然葉德輝認為讀者將點將錄看為一種「遊戲」，但是他以一定的態度和立場校對文本，此態度與其社會身分和時代的知識份子價值觀密切相關。

《東林點將錄》不是憑空出現的，其中所點一百零九人，大多數都是錢謙益《列朝詩集小傳》、朱彝尊《靜志居詩話》、陳田《明詩紀事》中提到者，然而在形式上點將錄與其他記載截然不同。[14]

再者，如果我們用以上的人物參考《東林黨籍考》相關傳記時，我們發現在楊漣的傳記提及左光斗、魏大中和魏忠賢，左光斗的傳記提及趙南星、高攀龍和楊漣，陳于廷傳記提及魏忠賢，或徐良彥的傳記提及魏忠賢和趙南星，皆使用簡略的方式敘述事件，

---

[12] 李棪，《東林黨籍考》一卷，收入：楊家駱主編，《中國學術名著》《史學名著》第四集第六冊，台北：世界書局，1961；張永剛，《東林黨議與晚明文學活動》，北京：中國社會科學出版社，2009，頁 273-274.

[13] 劉苑如，〈從品鑑到借鑑——葉德輝輯刻《山公啟事》與閱讀〉，《中國文哲研究集刊》，第三十八期，2011 年 3 月，頁 173.

[14] 筆者參考過幾個朱彝尊《靜志居詩話》的例子：丁元慶（卷 16，頁 367）、高攀龍（卷 16，頁 369-70）、孫慎行（卷 16，頁 376）、陳于廷（卷 16，頁 377）、徐良彥（卷 16，頁 379-80）、楊漣（卷 17，頁 386）、左光斗（卷 17，頁 386）、周順昌（卷 17，頁 392）。參看（清）朱彝尊，《靜志居詩話》，二十四卷，收入：《續修四庫全書》，第 1698 冊，上海：上海古籍出版社，2002.

或用很簡單的筆法描述他們之間的關係，因此，若僅依靠這種記載，讀者無法理解這些人物關係的複雜性。

從以上的例子我們可以看到，點將錄紀錄的個人細節相對於傳統傳記為少，點將錄的評點小傳顯然不是正式列傳的規整寫法，其在評論部分佔有一定篇幅，而其與《水滸傳》在名稱上的類比，說明中國文學歷史脈絡中不同文本和文體的互相關聯性。那麼，這兩種點將錄文本至少呈現三種文獻種類之間的互動關係：史書列傳、《水滸傳》和詩人類書（例如《靜志居詩話》）中詩人的傳記和評點。基本上點將錄省略許多歷史記載的細節，點將錄寫法的優點在於，文本層次上，人物之間不呈現任何關係，但這種人際關係仍舊存在，因此能以人物為主軸，將文本重新建構。在史書、小說、類書這樣性質不同的文本，彼此不斷地進行溝通，點將錄固定格式的出現，有其文學和歷史的重要性。更重要的是，用點將錄的寫法描述群體傳記，在篩選點將錄傳記的細節，可以看出作者各有其目的與動機。

作為群體傳記的特殊記載，「點將錄」的材料安排和文本結構很適合用數位人文工具進行分析，可以在一定的範圍內探索這部群體傳記的特色，並透過數位研究凸顯其功能、以及所顯示的社會關係屬性和文本框架流動的可能性。以上內容，將在下文的敘述中進行證明。

三、 數位人文對文本的分析

點將錄的結構模式延續魏晉以降九品中正的人物評議系統，如單純從傳統文本分析的角度觀察時，點降錄僅是人物名單的羅列而已，難以看出人物與人物之間的關係。然而，若借助數位工具來分析之，數位文本的視覺化的能夠突顯出文本中所隱含的人物關係，進而揭示人物條列新一類關係。點將錄為一種文學作品評論、且以傳記形式的文體書寫，其固定的體例格式很適合用於數位人文研究。從前述例子可見，很多傳統文獻（如史書列傳、地方誌、詩集、文集、書信等等）需要很長時間且需從許多材料中，才能夠看出某一人物與其他人物往來的細節，現代數位化的文獻一方面提供以不同問題意識為出發點所建立的資料庫，如點將錄的群體傳記資料庫。另一方面，社會網絡分析軟體的連結則有助於我們了解這些群體傳記之間人物關係的範圍與人物行為的特色等。本文將《東林點將錄》與《乾嘉詩壇點將錄》置在中國傳記文體的傳統之下，透過數位的社會網絡分析探究兩個文本中的人物影響力，藉此探討明代東林黨與清代詩人群體中重要人物之間的影響、中心性以及行為趨向。在兩種文本的數位分析基礎之上，本文將進一步討論數位人文工具如何對傳統文字文本提供另一種詮釋，進而探討這樣的詮釋異於一般

文本分析之處。

在具體的數位分析上，首先引介社會網絡分析的幾個關鍵概念——程度中心性（degree centrality）、親近中心性（closeness centrality）以及中介性（或也稱中間性）（betweenness centrality）——以突顯點將錄文體的社群特色。在社會網絡分析中，中心性分成三種形式：程度中心性、親近中心性、中介性。程度中心性用以衡量誰是一個團體中最重要或最有權力的中心人物。親近中心性是以距離為概念計算一個節點（node）[15]的中心程度，與別人愈近者則中心性愈高。如果以親近中心性的標準演算的數字高，說明這個節點更接近中心位置，也更具重要性。中介性則是衡量一個人作為媒介者的能力，是占據兩個人之間的重要位置的人，其重要性在訊息的交流和溝通，占據這樣位置愈多，這個人的中介性愈高。[16]上述三種概念有助於群體傳記的分析，故本文以三種中心性形式為本文的理論依據。

在實務操作上，本文以四個階段進行文本整理和初步分析：第一、建立點將錄兩個文本人物的基本資料庫；第二、在 CBDB 搜尋人物的個人編號後[17]，製作成 txt 格式的名單；第三、將名單匯入 CBDB「查詢社會關係網絡」的檢索系統裡，從既有的資料庫讓電腦依照不同的標準和特定的選項畫出人物之間的社會網絡關係；第四、運用 Pajek 的社會網絡分析工具對點將錄的人物關係進行進一步的分析。

1. 《東林點將錄》的社會網絡分析

為了瞭解點將錄人物的影響力，第一個我想測試的標準是中介性。首先，將《東林點將錄》人物編號的 txt 檔案輸入 CBDB 的系統，並按照搜尋的選項（親屬關係；朋友、家庭、宗教；師生關係、學術交流、主題相近、學術成員、文學藝術交往；政治所有選項；序跋、禮儀、傳記、論說；時間限：1550-1700）製作一個 Pajek 檔案，再輸入到 Pajek 的計算系統得到一個基本的網絡檔案。接著用這個初步的檔案讓電腦匯出一個人物中介性（betweenness centrality）的表單。在操作過程中可以發現，因 CBDB 製作的名單也包含了最近親屬關係，所以從原本的 95 個人物以中介性範疇的標準出現了 221 個節點。其中只有少數節點有一定的向量（按照中介性的標準與其他節點有連結）。選擇前 20 名展現中介性的程度的話，我們則可以看到東林黨網絡中最高中介性程度的人物（如 report 檔案）。

[15] 此外，在社會網絡分析中，「頂點」（vertex）是描述行動者的另一種術語，在本文兩者均可用。
[16] 參看，羅家德，《社會網分析講義》，北京：社會科學文獻出版社，2005，頁 150-162.
[17] 因在 CBDB 搜尋一個人物時出現幾個不同的傳記，筆者與《東林黨考籍》進行資料對比以免輸入不正確的傳記。

表 1: 《東林點將錄》中介性分析表

1. Betweenness centrality in N1 (221)

Dimension: 221
The lowest value: 0.0000
The highest value: 0.0158

Highest values:

Rank	Vertex	Value	Id
1	101	0.0158	王象春
2	125	0.0123	王象蒙
3	126	0.0023	劉學曾
4	31	0.0023	鄒元標
5	128	0.0023	甄淑
6	3	0.0017	黃尊素
7	37	0.0015	魏學濂
8	93	0.0014	王圖
9	34	0.0012	魏大中
10	127	0.0012	陳所學
11	84	0.0010	王象乾
12	119	0.0006	葉向高
13	120	0.0004	趙南星
14	6	0.0004	姚氏(黃尊素妻)
15	10	0.0004	黃宗會
16	38	0.0004	魏邦直
17	4	0.0004	黃曰中
18	25	0.0003	文震孟
19	24	0.0002	文元發
20	19	0.0001	周可賢

Sum (all values): 0.0461

表 2：《東林點將錄》親近中心性分析表

Viewing Vector --- 1. All closeness centrality in N1 (221)

1.	0.013575	錢謙益
2.	0.017163	劉應期
3.	0.038288	黃尊素
4.	0.024887	黃曰中
5.	0.017163	黃大綬
6.	0.026197	姚氏(黃尊素妻)
7.	0.017776	姚克俊
8.	0.024887	黃宗羲
9.	0.022624	黃宗炎
10.	0.024887	黃宗會
11.	0.022624	黃宗轅
12.	0.022624	黃宗彝
13.	0.009050	高攀龍
14.	0.009050	高世儒
15.	0.009050	錢曾
16.	0.009050	周茂蘭
17.	0.013575	周順昌
18.	0.009050	周冠
19.	0.013575	周可賢
20.	0.009050	孫慎行

表一顯示，王象春、鄒元標、黃尊素、魏大中、陳所學、葉向高、趙南星的中介性相當高，都在前二十名，表示這些人之間的關係比與其他人物的關係可能更密切。表二則是從原本的檔案將《東林點將錄》的人物按照親近中心性的標準排列，表二顯示按照名單節點排列的結果，而不是按照數值的高低。如果我們以錢謙益為例來看兩個表顯示

數字的意義，錢謙益在兩個表的排序中並非正相關。他不呈現在表一，但在表二按照親近性的標準，他有一定的數值。通過表二全表的分析可見，雖然在歷史文獻中錢謙益的關鍵性比其他東林成員要高，但他與其他成員（如：周順昌）親近性演算數值一樣，同時黃尊素親近性比錢謙益高。如同東林黨相關研究指出，[18]錢謙益無論在政治、學術、文學方面都是東林黨很重要的人物之一，但數位社會網絡分析顯示他的親近中心性不若現實影響力的狀況。由上述兩表數據可見一個人物在不同標準演算環境下，顯示的數值不一樣。這說明在人物關係方面，電腦的分析標準及提供的訊息比文本呈現的訊息具更多層次的意義，因而讓學者從更多方面詮釋相關數據。

為了更進一步了解兩個表有關錢謙益的數據以及他的影響，可以以一個節點製作 2-mode network，包含所有的叢集（clusters），展現一個人與其他人的關係。附錄圖 1 為錢謙益親近中心性的視覺化關係結構圖，圖中錢謙益的關係網絡看起來十分龐大，其原因在於與他相關聯的東林黨成員也呈現各自相關的親屬關係。2-mode network 這樣的分析讓我們更好地了解，在創造與一個人物或一個群體為中心的另一群體網絡時，無論在原本的網絡裡中心性率高或低的人物，在新的群體網絡中仍可以呈現其中心性的細節和特點，讓學者更進一步分析其影響力。

至於親近性中心的範圍（圖 2），我們發現東林黨人物的親近中心性主要與親屬有關。他們之間的互動相當低，甚至有些人的親近性中心指數為零，在圖上與其他人物沒有連接，例如公鼐、丁元薦、方震孺、王允成、汪文言、毛士龍、左光斗、沈應奎、李三才、周起元等等。

東林人物的程度中心性特色的視覺化如圖 3，人物標上加上了數字，並用方框的大小表示每一個人物的程度中心性。如圖所示，王象春、鄒元標、黃尊素、劉向高、陳所學、王圖的程度中心性相當高，或許由此可以顯示他們在東林黨人物網絡中的位置和影響力。圖 4 則是用 pajek 所有核心分割的功能表示不同叢集之間的複雜關係。

在親近中心性的標準社會網絡分析系統的結果中左光斗、李三才、周起元都偏低，然而在歷史中他們是相當積極的東林黨人物──左光斗是「東林黨六君子」之一，周起元是「東林黨七君子」之一，李三才則是東林黨很關鍵的人物，與被稱為「朝中重臣」的東林官員（如鄒元標等）曾一起做過官，關係十分密切。[19]其次，王象春、鄒元標、趙南星的中介性相當高，王象春與鄒元標的程度中心性也很高。既然李三才與鄒元標關係這麼密切，為何他們的中心性不一樣？王象春是山左三彥詩派的人，但屬於同樣詩派公鼐的親近中心性卻很低。黃尊素、魏大中分別為「東林黨七君子」以及「東林黨六君

[18] 張永剛，《東林黨議與晚明文學活動》，北京：中國社會科學出版社，2009.
[19] 朱文杰，《東林黨史話》，上海：華東師範大學，1989，頁 91.

子」的代表，卻他們的中介性比周起元和左光斗高。從這種初步的數據分析我們可以推測人物在不同詩派或政治團隊的影響力，在進一步考察這些人物與其他人物的互動如何影響到他們在團隊的形象、影響力的程度、他們之間互動的原因，進而推論有關當時歷史和文學發展的面向。

2. 《乾嘉詩壇點將錄》的社會網絡分析

首先，將《乾嘉詩壇點將錄》人物編號的.txt 檔案，並按照 CBDB 搜尋的選項（不包括財務、醫療、軍事、女性；時間限：1700-1850）（圖 5），輸入 Pajek 的計算系統，得到一個最初的網絡（如圖 6），再據以製作一個 2mode network (existing clusters only) 新的網絡呈現關係。第二種網絡好處在於可以看到人物的叢集安排。按照實際的叢集安排的圖，可以用不同的顏色標記最短距離的兩個頂點（closest vertices）（例如，王芑孫與顧光旭）（如圖 7）等等細節，並在本檔案出現的 1893 個重疊的部分（number of crossings），及哪些兩個頂點之間距離最短或最長（shortest/longest line）。這些功能可更進一步分析和了解人物之間關係的屬性和特色、交流的頻率、不同人物之間的關係距離大小，從而推測關係的密度。

在上述操作我們發現，從所有的叢集中（c0-c4）《乾嘉詩壇點將錄》所列的詩人以及點將錄作者本身大部分屬於叢集 c0：

表 3（顯示一部分的名單）

```
2-Mode Network: First set = Vertices, Second set = Clusters
================================================================================
 Time spent:  0:00:00

--------------------------------------------------------------------------------
Editing Network: 4. 2-Mode network [Existing Clusters only] obtained from C13 (183). Vertex:179
--------------------------------------------------------------------------------
        1:          44.179          val=1.00000             / 王夏.c0
        1:          49.179          val=1.00000             / 全祖望.c0
        1:          50.179          val=1.00000             / 杭世駿.c0
        1:          52.179          val=1.00000             / 胡天游.c0
        1:          55.179          val=1.00000             / 朱彭.c0
        1:          56.179          val=1.00000             / 王昶.c0
        1:          59.179          val=1.00000             / 彭兆蓀.c0
        1:          69.179          val=1.00000             / 沈清瑞.c0
        1:          77.179          val=1.00000             / 沈德潛.c0
        1:          80.179          val=1.00000             / 吳錫麒.c0
        1:          84.179          val=1.00000             / 王文治.c0
        1:          85.179          val=1.00000             / 顧光旭.c0
        1:          86.179          val=1.00000             / 曾燠.c0
        1:          90.179          val=1.00000             / 吳青梁.c0
        1:          91.179          val=1.00000             / 石韞玉.c0
        1:          93.179          val=1.00000             / 王芑孫.c0
        1:          94.179          val=1.00000             / 舒位.c0
        1:          97.179          val=1.00000             / 汪端光.c0
        1:          98.179          val=1.00000             / 謝啟昆.c0
        1:         101.179          val=1.00000             / 趙翼.c0
        1:         102.179          val=1.00000             / 孫士毅.c0
        1:         104.179          val=1.00000             / 王太岳.c0
        1:         109.179          val=1.00000             / 孫原湘.c0
        1:         111.179          val=1.00000             / 查禮.c0
        1:         112.179          val=1.00000             / 戴敦元.c0
        1:         113.179          val=1.00000             / 趙懷玉.c0
        1:         114.179          val=1.00000             / 齊召南.c0
        1:         115.179          val=1.00000             / 彭紹升.c0
        1:         116.179          val=1.00000             / 程晉芳.c0
        1:         117.179          val=1.00000             / 鄭燮.c0
```

有些詩人屬於其他叢集，例如詹應甲、翁方綱、畢沅、袁枚、許宗彥等屬於叢集

c1，孫星衍、錢大昕、阮元、陳文述屬於叢集 c2：

表 4（顯示全部名單）

```
--------------------------------------------------------------------------------
Editing Network: 4. 2-Mode network [Existing Clusters only] obtained from C13 (183). Vertex:181
--------------------------------------------------------------------------------
        1:           1.181         val=1.00000              / 阮元.c2
        1:           2.181         val=1.00000              / 阮玉堂.c2
        1:           3.181         val=1.00000              / 江承瑞.c2
        1:           4.181         val=1.00000              / 阮承信.c2
        1:           6.181         val=1.00000              / 江振箕.c2
        1:          11.181         val=1.00000              / 沈在廷.c2
        1:          12.181         val=1.00000              / 沈勤堉.c2
        1:          19.181         val=1.00000              / 陳遇清.c2
        1:          20.181         val=1.00000              / 陳基.c2
        1:          23.181         val=1.00000              / 陳廷慶.c2
        1:          28.181         val=1.00000              / 錢大昕.c2
        1:          30.181         val=1.00000              / 錢大昭.c2
        1:          33.181         val=1.00000              / 錢桂發.c2
        1:          51.181         val=1.00000              / 許乃穀.c2
        1:          62.181         val=1.00000              / 孫星衍.c2
        1:          64.181         val=1.00000              / 孫勳.c2
        1:          65.181         val=1.00000              / 孫星衡.c2
        1:          66.181         val=1.00000              / 孫馨.c2
        1:          83.181         val=1.00000              / 陳文述.c2
        1:          87.181         val=1.00000              / 許乃濟.c2
        1:          88.181         val=1.00000              / 許乃普.c2
        1:          89.181         val=1.00000              / 陳裴之.c2
        1:          92.181         val=1.00000              / 龔玉晨.c2
        1:          95.181         val=1.00000              / 張問榮.c2
        1:          96.181         val=1.00000              / 張問陶.c2
        1:         106.181         val=1.00000              / 盧見曾.c2
        1:         108.181         val=1.00000              / 戈源.c2
        1:         119.181         val=1.00000              / 戈濤.c2
        1:         124.181         val=1.00000              / 盧謙.c2
        1:         141.181         val=1.00000              / 張問安.c2
        1:         150.181         val=1.00000              / 錢東.c2
        1:         171.181         val=1.00000              / 盧荅文.c2
    Newline
```

錢枚、錢杜屬於叢集 c3，紀昀跟親屬關係的人則成立 c4 叢集。電腦分析的特色之一是從一個既有資料庫按照一定的標準取訊息並進行分析。從簡單格式的一個文本我們得到了相當複雜的結果，對研究清代文學歷史的學者來講，詮釋這些數據是更深入地了解這些叢集組成的邏輯的一個機會。每一叢集裡所篩選的人物的標準為何？這些規則是否符合一些實際的歷史過程或傳統文獻呈現不出來人物關係？不同群體之間存在的關係又如何？回答這些問題必須研究每一叢集裡的人物以及其關係特徵。換句話，電腦所進行的群組組合和排列，展現點將錄的另一種邏輯，文本按照這種邏輯重新組合起來，提供一個嶄新的詮釋、意義和價值。

運用 2mode network 的檔案再製作按照中心中介性、中心程度性及中心親近性的圖案，結果如圖 8（因顯示資訊最明顯，所以附錄只顯示 all degree centrality 的圖檔）。

表 5：《乾嘉詩壇點將錄》中介性分析表

1. Betweenness centrality in N1 (178)

Dimension: 178
The lowest value: 0.0000
The highest value: 0.0079

Highest values:

Rank	Vertex	Value	Id
1	45	0.0079	紀昀
2	28	0.0056	錢大昕
3	1	0.0020	阮元
4	31	0.0017	錢東壁
5	34	0.0017	錢東塾
6	33	0.0017	錢桂發
7	4	0.0014	阮承信
8	19	0.0013	陳遇清
9	106	0.0010	盧見曾
10	30	0.0009	錢大昭
11	38	0.0009	王爾達
12	13	0.0006	錢汝鼎
13	12	0.0006	沈勤堉
14	9	0.0006	張熙
15	110	0.0006	錢琦
16	60	0.0003	畢沅
17	6	0.0003	江振箕
18	152	0.0003	錢枚
19	71	0.0003	袁枚
20	2	0.0002	阮玉堂

Sum (all values): 0.0313

按照這三個標準以及從這些圖所展現的訊息我們可以看到，在前二十名中，錢大昕、阮元、畢沅、錢枚和袁枚中介性相當高，說明在《詩壇點將錄》人物的網絡中，在不同叢集人物的互動程度、中心位置或從中心相對的距離及其產生的影響力。

如附錄圖所示，《東林點將錄》與《詩壇點將錄》中的人物雖然在社會網絡分析中呈現出一定的程度中心性、親近中心性以及中介性的意義，但是也有許多人物與程度中心性的人物距離相當遠或者根本不產生任何關係。其次，我們可以觀察到《詩壇點將錄》的叢集組織比《東林點將錄》更明顯。東林黨和東林學院有很特殊的關係，[20]這種關係一部分是家庭的，一部分則是政治的。雖然有部分東林學院的成員同時也是東林黨黨員，但點將錄人物的社會網絡分析顯示《東林點將錄》所列舉的人物數據分析與實際的社會關係有差異──即人物的影響力與他們所參與的團隊沒有直接的關係。我們該如何解釋這樣的數據？這種安排以及缺乏人物之間高度密切關係的文本結構，呼應葉德輝在《詩壇點將錄》序所提及的話：「《詩壇點將錄》一書，乃以《水滸》一百八人配合頭領，或肖其性情，或擬其行止。」[21]點將錄的文學特色在於強調人格和行為，其次是呈現不同人物關係。依固定格式書寫的點將錄側重文學傳統脈絡的連續性和關聯性，這樣的關聯性比人物之間的關係更明顯。如果數位人文的研究呈現兩個點將錄的人物之間的關係與

[20] 張永剛，《東林黨議與晚明文學活動》，頁 22-25.
[21] 舒位等原著; 楊楊編校，《三百年來詩壇人物評點小傳彙錄》，頁 39.

現實存在的關係不同的結構，那麼可以說兩個文本的文學想像力比現實關係更重要。

若結合葉德輝的校對過程，並參考傳統文獻對點將錄的記載，我們可以發現《明史》、《清史稿》中沒有《東林點將錄》的相關記載，而《四庫全書總目・史部・傳記類存目四》卷六二著錄江蘇巡撫采進本《東林點將錄》一卷。[22]受此影響，此後各種目錄書大多將《東林點將錄》著錄於史部「傳記類」。這樣的書目歸類說明《東林點將錄》的性質仍與歷史有關，同時與小說脈絡的關係也很明顯。據葉德輝在附考裡所述，點將錄所錄的三十六天罡來源要追溯到宋代《宣和遺事》[23]，這樣比喻已經被根深蒂固在文學的想像力裡。如同本文的第一部分提到的《東林黨籍考》用描述性的敘事呈現人物關係，假如對這種關係缺乏量化的研究，讀者或研究者很難或無法了解宏觀的情境。像社會網絡分析和網絡視覺化的數量性技術（Quantitative techniques），有助於展現社會結構不同的面向和層次，從而輔助文學社會學的研究。[24]因此，本文以點將錄的例子示範在簡單格式的文本裡，讀者可以看到文學想像力與歷史事件之間的複雜互聯，說明連最單純的評點小傳背後存在著文學和歷史的多層次關係，而數位人文分析則可以幫助學者質問文學與歷史的交叉點，從其中的人物關係釐清時代的歷史脈絡、文學觀、甚至於重新思考這種關係如何促進文本的重新詮釋。

四、 結論

本文以點將錄作為彙錄特殊群體傳記的一個文本，並引用資料庫（CBDB）以及社會網路分析工具（Pajek）對文本進行數位人文分析，探討數位研究工具如何協助學者做出有關這些人物關係與時代背景之間的關聯，並藉由原來文本的解構和建構以及數位的改造，顯示詮釋明清歷史和文學的一個可能性，並呈說明在這個歷史脈絡中點將錄作為獨特文本的重要性。

本文證明一般文史研究以傳統的文本分析進行研究，此與數位人文研究所表示的文本中人物實際的關係（缺乏固定格式卻呈現內在的細節）存在這差別。透過本文分析，我們可以提出兩個看法：第一、假如傳統的文本提供文字方面的訊息，那麼數位工具能協助將文字重新加以整理，而呈現另一詮釋的可能性或展示傳統讀法無法顯示的細節。在文學歷史的語境中，這樣數位研究方法有助於學者對文本更全面的掌握。第二、作為群體傳記例子的點將錄分析證明，這些作者和與其相關的人物的中心性或邊緣性代表了

[22] （清）永瑢等撰，《四庫全書總目》，頁 559.

[23] 舒位等原著; 楊楊編校，《三百年來詩壇人物評點小傳彙錄》，頁 53.

[24] So, Richard Jean and Hoyt Long, “Network Analysis and the Sociology of Modernism”, *Boundary* 2, (40) 2, p. 155.

文學歷史脈絡中的某種意義。群體傳記其實是研究文學歷史中很重要的面向之一，且是文學歷史宏觀分析（macroanalysis）的不可缺乏部分。在「清代普遍的輯佚風氣，不僅名家屢出，一般士人、官員、藏書家和商人也都或多或少從事輯佚、校讎或考證工作」[25]的歷史情境之下，葉德輝對傳統文本的校對可視為一種符應時代的特殊知識創造，而當代研究環境之下的數位分析，則讓文本呈現多層次的內容。這過程中不只是文本的不同詮釋，更是研究工具轉變帶來的知識系統改變，而在這過程中數位人文的研究是不可代替的分析方法。*

關鍵詞：數位文本、群體傳記學、社會網絡分析（SNA）、清末民初、葉德輝

[25] 劉苑如，〈從品鑑到借鑑──葉德輝輯刻《山公啟事》與閱讀〉，頁 185.

* 本文附錄請於會後自會議網頁下載。

中國詩歌格律之重探與數位化研究：
兼談「漢詩格律分析系統」[1]的設計

林偉盛[*]、鄧賢瑛[**]、莊德明[***]

摘 要

奠基於漢語與漢字特殊質性上而形成的漢詩格律，是漢語詩學的重要內涵。過去對於漢詩格律的研究，由於技術工具的局限，往往只能透過人力處理相對有限的詩歌音韻資料，使得漢詩格律研究的觀念不得不趨於保守，而往往採用「以後律前」、「一以貫之」的方式，來處理動態發展的歷史現象，使得我們對於漢詩格律的理解往往受到侷限，甚而扭曲變形。近年來，由於研究工具大幅的改變，特別是數位技術的精進，使得巨量資料的統計與檢核，已可通過數位工具進行處理，既有的文獻資料庫和音韻資料庫，也使文本檢索與音韻分析的工作較諸傳統的方式更為便捷。惟目前的各種資料庫各有其不同的設計取向，彼此之間尚未能相互連結成一個因應詩學研究所需的數位工具。「漢詩格律分析系統」研發之目的，即在匯通語言學與詩學的研究資源，整合現有的數位內容，設計出適於漢詩格律研究需求的新工具，以從事傳統研究方法不易進行的研究課題，減少傳統研究繁瑣、重複的檢核與統計的流程，使研究者可將時間與心力轉向更高階的研究分析上，進而開拓漢詩格律研究的深度與廣度。

本文將以「漢詩格律分析系統」的設計為例，介紹系統研發的前因後果，探討數位技術如何回應文學研究之需求，共同合作研發出一個以文學研究為取向，回應文學研究關懷的新工具。「漢詩格律分析系統」的雛型已基本建置，提供五言四句詩的判讀分析。系統主要依據當前漢詩格律的研究成果，通過對於魏晉南北朝到初盛唐與聲律相關的文獻研究，擇取出南朝齊武帝（蕭賾，440－493 在位）永明年間以沈約（441－513）為代表的永明詩人的聲律論，與初唐時以元競

[1] 「漢詩格律分析系統」：http://xiaoxue.iis.sinica.edu.tw/hanshi。

[*] 國立臺灣大學中國文學系博士生。
[**] 國立臺灣大學中國文學系計畫助理。
[***] 中央研究院數位文化中心研究助技師。

（？－？）為代表的聲律論，建立起具規範性的格律規則，形成三種主要的格律模型：「沈約四病」、「元兢四病」和「元兢調聲三術」，作為系統分析判讀詩作格律的理論依據。系統的分析結果，也讓三種格律模型可以同時呈現，使用者可據此以進行不同時期的格律比對，進一步檢視漢詩格律由南朝永明時期到初盛唐的動態發展歷程。

「漢詩格律分析系統」的設計，是從文學研究的問題意識出發，結合數位工具在處理巨量音韻與文獻資料彙整上的優勢，以重探中國古典詩格律在不同的發展階段所呈現出的具體樣貌。由於思維邏輯上的差異，在系統設計、研發的過程中，文學研究與數位技術之間進行過多次的對話、溝通與切磋、琢磨，反映出各具主體、細緻分工的跨領域合作模式。文學研究也因著數位技術與工具的介入，使得研究者得以快速且準確地處理大量詩作的聲韻分析，以檢核通過文獻資料所擬定的格律規範，再依據結果反過來檢視所擬定的格律規範是否適宜，修正其中不恰當的格律條件。由於數位工具節省了原先所需耗費的時間與人力，使得文學研究者可以根據系統分析的結果進一步進行研究，從而發現原先受囿於詩作檢核樣本數不夠多、不夠全面，而難以進一步反省與討論的新議題。另一方面，藉由數位工具所進行的格律研究的成果，也可以讓研究者進一步反省過去傳統的研究方法，通過對於詩作材料的整理所歸納出的格律規範的合理性。本文的最後，反省了「漢詩格律分析系統」的設計過程中，文學研究與數位技術之間的互動，並探討數位技術作為一種文學研究的方法，相較於傳統研究方法所具有的優勢，和所可能存在的侷限。

關鍵字：五言詩、聲律論、詩學、小學堂、漢詩格律分析系統

Revisiting and Digitalizing of the Metrical Regulations of Chinese Pentasyllabic Poetry：Discussion of the Design of "Metrical Regulation Analytic System of Chinese Poetry"

Wei-cheng Lin*, Hsien-ying Teng**, Der-ming Juang***

Abstract

Metrical regulations of Chinese poetry rooted in the particular traits of Chinese language and Chinese characters are an important part of Chinese poetics. Due to the restrictions of technical research tools, previous research on Chinese metrical regulations has usually only dealt with limited amounts of phonetic data in poetry. In light of that, the concept of research on Chinese metrical regulations inevitably becomes relatively conservative. Scholars usually handle dynamic historical incidents by means of "applying rules in later generations to assume precedent" and "consistency," which limits or even distort our understandings of metrical regulations. Recently, thanks to enormous changes in research tools, especially improvements in digital technology, researchers are able to process statistic calculations and reviews of big data via digital tools. Existent databases of literature and phonological data also make textual searches and phonological analyses more convenient than traditional methods. However, various contemporary databases have different intensions and targets of design; they have not yet connected as a single digital tool that completely fulfills the needs of research of Chinese poetics. The purpose of developing "Metrical Regulations of Analytic System of Chinese Poetry" is to integrate research resources of linguistics and poetics, as well as existent databases, and to invent a new site that is tailored to the needs of research of metrical regulations of Chinese poetry. In that way, when one does research that is difficult to deal with through conventional methods, complicated and redundant static checking procedures can be largely decreased. Additionally, this could enable scholars to invest their time and energy into more advanced explorations and therefore increase

* Ph.D. Student, Department of Chinese Literature, National Taiwan University. Email:
** Program Assistant, Department of Chinese Literature, National Taiwan University. Email:
*** Assistant Research Specialist, Academia Sinica Center for Digital Cultures. Email:

the depth and breadth of research on Chinese poetics.

This paper takes “Metrical Regulations of Analytic System of Chinese Poetry” as an example to introduce the reason for its invention and to explore how digital technology responds to the demands in the field of literary research, as well as illustrate how scholars from different disciplines work in collaboration to design a new tool based on the concerns of researchers of literature. “Metrical Regulations of Analytic System of Chinese Poetry” already has a basic model which can analyze four-line pentasyllabic verses. Based on existing research findings on metrical regulations during the Wei, Jin, Southern and Northern Dynasties, the Early Tang and the High Tang, the system selects Shen Yue’s (441–513) metrical theory as representative of the Yongming era and Yuan Jing’s (?) theories as representative of the Early Tang in order to establish mandatory metrical rules and form three major metrical patterns: “four defects” proposed by Shen Yue, “four defects” and “three methods adjusting tonal prosody” proposed by Yuan Jing. These prosodic rules and patterns serve as criteria to examine and evaluate a poem; the system can also simultaneously display the analytic results of the three models. Users can compare metrical patterns from different eras and accordingly investigate the evolution of metrical regulations from the Yongming era to the Early Tang and the High Tang.

The invention of “Metrical Regulations of Analytic System of Chinese Poetry” initiates from problems in literary research and then incorporates the advantages that digital devices have in organizing enormous amounts of phonological and textual data in order to revisit different stages of Chinese metrical regulations. On account of differences in logic, scholars of literary research and digital technicians have been through numerous discussions and modifications in the process of design and invention. This process has revealed that the two research disciplines have their own subjectivities, and the interdisciplinary working pattern relies on neat divisions of work. On the one hand, for the sake of digital technology and tools, researchers of literature are able to quickly and accurately deal with analyses of metrical regulations of poems in large quantities in order to review the rules of metrical regulations synthesized from historical materials. On the contrary, based on the results, scholars can also reexamine whether the rules are appropriate and revise the criteria for checking. Thanks to digital tools, not only is much time and labor saved, but researchers of literature are empowered to do more advanced studies based on analytic results retrieved from the system. In light of

that, they can explore new topics that previously were subject to lack of sample and comprehensiveness. On the other hand, results analyzed by the system allow researchers to rethink traditional research methods and the generalization of rules of metrical regulation from poetic materials. Lastly, this paper reviews the interaction between research on literature and digital technology in the process of invention and discusses, as a method of research on literature, advantages and possible limitations of digital technology.

Keywords: pentasyllabic poem, metrical theory, poetics, xiaoxuetang, metrical regulations of analytic system of chinese poetry

一、 格律研究的現況與侷限

以「格律」為主要內容的新體詩歌的試煉成熟，係唐人在中國詩體發展上的一個顯著且重要的成就，奠定了中國詩歌主要的形式美典，表現在對於「對偶」與「聲律」的重視。唐人稱之為「近體」或「今體」，以與唐以前的詩歌作分別。所謂古、今「體」的差異，主要表現在「格律」的有無，因而唐人又稱「聲勢沿順、屬對穩切」者為「律詩」，[2]後人亦稱「近（今）體詩」為「格律詩」。

當今學界對於律詩格律的研究，主要依據清代學者所擬定之聲調譜或平仄論，如王士禎（1634－1711）《律詩定體》、趙執信（1662—1744）《聲調譜》、何世璂（1666－1729）《然燈記聞》和翁方綱（1733－1818）的《王文簡古詩平仄論》、《趙秋谷所傳聲調譜》、《五言詩平仄舉隅》、《七言詩平仄舉隅》，以及董文渙（1833—1877）的《聲調四譜圖說》等論著，並結合唐人詩作進行分析，以檢視唐人的詩體觀。近人王力（1900—1986）的《漢語詩律學》則將各家紛繁的格律規範予以系統化及分析說明，是相關著作中最具代表性的。[3]後來的學者如呂正惠、張夢機和許清雲等的格律論著，大抵亦不出王氏所擬構的格律規範而予以簡化。[4]

這樣一種研究取徑，蔡瑜先生稱之為「逆向的理解」；[5]其所擬構出的格律規範，則可稱為「近世格律」。由於唐代討論聲調格律的論著迄清已不見存於中土，清人乃係通過整理唐人詩作並予以分析歸納的方式擬構出所謂的近體格律。[6]將這樣一種逆向而得的格律規範落實於唐人詩作的具體分析上，討論唐人詩作合格與否的問題，在邏輯上不免存在著「以後律前」的窘況，如同蔡瑜先生所指出的：

> 以清人觀念為中心的聲律規程落實在唐人創作的聲律分析上，具現了極為繁瑣的分類系統，以及古律體互相滲透的問題，終至變例滿紙拗救充斥，難免有著

[2] 唐·元稹：〈敘詩寄樂天書〉，收入周相錄校注：《元稹集校注》（上海：上海古籍出版社，2011），卷五十六，頁 1361。

[3] 王力：《漢語詩律學》（上海：新知識出版社，2002）。

[4] 參見呂正惠：《詩詞曲格律淺說》（臺北：大安出版社，1991 年三刷）；張夢機：《古典詩的形式結構》（板橋：駱駝出版社，1997）；許清雲：《近體詩創作理論》（臺北：洪葉文化，1997）。

[5] 蔡瑜：〈唐詩律化的理論進程──以詩格為中心的探討〉，《唐詩學探索》（臺北：里仁書局，1998），頁 1。

[6] 清·仲是保〈聲調譜序〉：「唐詩聲調，迄元來微矣。明季寢失，古詩尤甚。吾虞馮氏始發其微，于時和之者，有錢牧齋及練川程孟陽。若後之婁東吳梅村，則又聞之于程氏者矣。顧解人難得，惟新城王阮亭司寇及見梅村，心領其說，方欲登斯世于風雅，執以律人，人咸自失。」收入《叢書集成新編》（臺北：新文豐，1985），第 78 冊，頁 253。案：仲是保為清初人，與趙執信（1662─1744）同時。可見相對於前朝，清人雖有較多探討唐代近體聲調的論著，但其時已難確切掌握唐詩聲調使用的實際情形。

界限模糊區劃困難的堪尬。[7]

另一方面，由於留存下來的唐人詩作數量匪鮮，僅依據清康熙（愛新覺羅玄燁，1661－1722 在位）時彭定求（1645－1719）等奉敕編校的《全唐詩》，即錄有「詩四萬八千九百餘首，凡二千二百餘人」，[8]更遑論唐以後的近體詩作數量。徒以人工的方式而欲以所擬定的格律規範檢核一首首的詩作，難免力有未逮。因此，在未經充分的詩作檢核以驗證所擬定的格律規範適當與否的情形下，往往又將所擬定的格律「一以貫之」，視為包含唐以後的近體格律的通則。在格律研究上普遍存在著「以後律前」和「一以貫之」的現象，是將格律假設為一靜態的完成體，而忽略了格律的生成與發展理應是動態演進的歷史現象。這樣的處理方式也使得我們對於格律的理解往往受到侷限，甚而扭曲變形。

另方面，隨著清末從日本傳回九世紀日本遣唐僧空海（774—835）所著《文鏡祕府論》一書中所載錄的唐人「詩格」論著，學者們也漸漸發現，近體格律實具有其動態發展的脈絡。唐人對於聲調的討論，也與後世整理所得出的格律規範不完全一致，最顯著的差異在於，後世整理的格律是字字定其平仄，但在唐人的「詩格」論著中，顯然更為彈性自由，往往只強調關鍵的音節點，以至於一些後世被認為「拗」而需「救」的作品，如依唐代「詩格」的規範來看，則實合格而無犯。[9]通過唐人「詩格」的論述來重新認識唐代近體詩聲律發展的軌跡，相對於前面一種「逆向」的方式，則可謂之「順向的理解」。這方面的研究雖也已經累積了一些成果，迄今卻仍是方興未艾的工作。[10]所面臨到的一個主要的困境在於，通過唐人「詩格」資料重建了當時的聲律理論後，欲將所得的理論應用在巨量詩作的分析，以檢核所擬定的格律，僅憑人工的方式仍不免力有未逮。

在過往的格律研究中，不管是採取「逆向」的抑或「順向」的理解徑路，欲以人工的方式將格律理論落實於巨量詩作的檢核上，難免力不從心，這是過往格律研究所面臨到的一個共同的困境。近年來，由於研究工具的大幅改變，特別是數位技術的精進，使得巨量資料的檢核已可通過數位工具來進行處理，因而也出現了一些將數位技術應用在格律研究上的數位系統，解決了過去以人工檢核巨量詩作所存在的困境。其中較具代表性的，是由北京大學數據分析研究中心和北京欣諾格科技有限公司聯合研發的「全唐詩

[7] 蔡瑜：〈唐詩律化的理論進程──以詩格為中心的探討〉，頁 2。
[8] 清·愛新覺羅玄燁：〈御製全唐詩序〉，收入清·彭定求等編：《全唐詩》（北京：中華書局，1979 年二刷），頁 5。
[9] 詳參蔡瑜：《唐詩學探索》，頁 23-100。
[10] 相關的論著有方瑜：《唐詩形成的研究》（臺北：牧童出版社，1975）、王夢鷗：《初唐詩學著述考》（臺北：臺灣商務印書館，1977）、張伯偉《全唐五代詩格彙考》（南京：鳳凰出版社，2002）和蔡瑜《唐詩學探索》等。

分析系統」與「全宋詩分析系統」，[11]系統的相關功能包含《全唐詩》和《全宋詩》的文獻、詩人傳記資料、格律資料、以及自定義區等。然而，對於系統所採用的語音材料來源為何，既不清楚，且過於簡略。至於其格律規範的設定，雖未加說明，推測應即「近世格律」，但缺乏規範說明的系統，顯然不能滿足嚴謹的學術研究需求。此外，各別的學者為了研究的需求，或亦自行建置格律分析系統作為自身研究之用，如杜曉勤所著《齊梁詩歌向盛唐詩歌的嬗變》一書，及充分利用了其與大陸首都師範大學電子文獻研究所所長尹小林共同研發的一套「詩歌聲律自動分析系統」，來處理研究過程中的詩律分析與統計。[12]系統所採用的格律規範，依據《齊梁詩歌向盛唐詩歌的嬗變》一書的內容，可推測應該也是採用「近世格律」。然此套系統並未公開，未能供研究者應用與檢核，顯然亦未能符合學術研究的需要。

由上述可知，不論是公開的抑或非公開，格律數位研究系統所採取的格律標準，主要仍屬逆向的「近世格律」，而「順向」的呈現從格律的發生與發展的系統，則仍付之闕如。可以說，目前近體格律的研究現況，不管是使用傳統的研究方法或者採取新近的數位技術，「逆向」的研究徑路仍是主流，這也使得我們今日對於格律發展的歷史圖像甚為單一扁平，更遑論具體掌握各時代的差異，和不同詩人的音韻風格。因而，建構一個採取「順向」的徑路，呈現出律詩格律動態發展樣貌的數位分析系統，既可與「逆向」的研究成果相互對話，也是現今格律研究的一個重要且迫切的突破口。

二、「漢詩格律分析系統」的發想與設計

國立臺灣大學中國文學系蔡瑜先生，有感於傳統格律研究方法的侷限性，與新興數位工具在處理巨量資料上的優勢，加之長期對於漢詩格律的興趣與關注，乃欲建置一個針對詩歌格律研究，能夠呈現漢詩格律發展脈絡的數位分析系統。「漢詩格律分析系統」，即是蔡瑜先生所申請的科技部補助專題研究計畫下，結合了詩學、語言學和數位技術等相關領域的專家，所共同設計研發的格律分析系統。

「漢詩音韻分析系統」的研發，係以漢詩格律研究之需求為主要取向，以匯通詩學與語言學的研究資源，整合現有的數位工具，以從事傳統研究方法不易進行的研究為目標。為滿足此一目的，需要整合現有的文獻資料庫和語言資料庫，以取得巨量的詩歌文本和每一個字音的音韻資料，進一步予以判讀、統計、分析和呈現。既有的文獻資料庫，

[11] 「全唐詩分析系統」：http://www.chinabooktrading.com/tang/；「全宋詩分析系統」：http://www.chinabooktrading.com/song/。

[12] 杜曉勤：《齊梁詩歌向盛唐詩歌的嬗變》（北京：北京大學出版社，2009），頁。

如中央研究院所建置的「漢籍電子文獻資料庫」，和由已故元智大學羅鳳珠教授所建置的「網路展書讀」等，將來都規劃作為「漢詩格律分析系統」的文獻資料來源，系統也將建構「總集」、「別集」、「域外漢籍」等項目，以對文獻資料庫所取得的詩歌文本進行分類，各項目下則依時代排列。在語音判斷的方面，則以臺灣大學楊秀芳教授所研發的「漢字古今音資料庫」作為基礎，該系統主要根據宋代韻書《廣韻》收字，而兼及其它字書、韻書和近代新增的少數字，與「漢詩格律分析系統」相關的先秦、兩漢、魏晉、南北朝和隋唐等各期的語音資料，該系統皆已建置，是目前最為完善的漢字語音資料庫。

在系統建置的規劃上，分為「學術研究」與「系統建構」兩個部分，主要由前者提供後者構擬的主要方向。現階段在學術研究方面，已完成從現存詩論文獻及唐人《詩格》著述中建立起理論的基點，如以沈約為代表的永明聲律規範，和以元兢《詩髓腦》為代表的初唐聲律規範。系統也據此而設計出「沈約四病」、「元兢四病」和「元兢調聲三術」這三種格律模型作為條件群組。由於「押韻」是詩歌成立的最基本條件，因此「押韻」亦被預設為必選的判讀條件。茲將「漢詩格律分析系統」依據學術研究方面的成果，在具體的「系統建構」上所呈現的條件群組說明如下：

（圖例說明：○代表單一漢字；●代表韻腳；◎代表可押韻亦可不押韻之處；🟡🔴代表遵守平仄律中不同聲的規範；🔵🟢代表遵守四聲律中不同聲的規範）

(一) 韻腳：兩句一聯，每聯的最後一個字押韻，首句可押可不押。

○○○○◎，○○○○●。○○○○○，○○○○●。

(二) 沈約四病：南朝齊永明時期的詩學理論以沈約的「四聲八病」為核心。「四聲」指平、上、去、入；「八病」則可分為兩組，分別為平頭、上尾、蜂腰、鶴膝，稱為「前四病」；大韻、小韻、正紐、旁紐，則為「後四病」。由於「後四病」在漢詩格律的發展上無法形成具體的規範，故系統從略，而著重於「前四病」的規則對五言詩格律的規範與發展。以下簡述「前四病」的格律條件：

1、 上尾：第十字不與第五字同聲，依此類推。即每句句尾不押韻字與押韻字不可同平、同上、同去、同入聲。

○○○○◎，○○○○●。○○○○🔵，○○○○●。

2、 鶴膝：第五字不與第十五字同聲，依此類推。即相鄰兩聯的上句不押韻之句尾字，不可同平、同上、同去、同入聲，須四聲迭代。

○○○○🟢，○○○○●。○○○○🔵，○○○○●。

3、 蜂腰：以五言一句為單位，每句第二、五字不可同平、同上、同去、同入聲。

○🔵○○🟢，○🔵○○●。

4、 平頭第一字：兩句一聯，每聯上、下句第一字不可同平、同上、同去、同入聲。

🔵○○○○，🟢○○○●。

5、 平頭第二字：兩句一聯，每聯上、下句第二字不可同平、同上、同去、同入聲。

○🔵○○○，○🟢○○●。

(三) 元兢四病：初唐元兢著有《詩髓腦》，繼沈約之後積極發展聲律理論。其「四病」的規範原則與沈約基本相同，惟放寬部分「同平聲」的限制。

1、 上尾：同沈約。

2、 鶴膝：同沈約。

3、 蜂腰：略同沈約。惟同平聲者為合格。

4、 平頭第一字：略同沈約。惟同平聲者為合格。

5、 平頭第二字：同沈約。

(四) 元兢調聲三術：元兢《詩髓腦》的「調聲三術」為「換頭」、「護腰」、「相承」，其中「換頭」又可細分為「雙換頭」與「拈二」，惟「拈二」為基本規範，分別規範「第二字」與「第一字」，故系統區分為二。依據文獻，「換頭」係針對第一、二字建立起平聲、去上入聲之二元結構與粘對交迭的規則，已顯露出過渡至平仄律的傾向。「相承」則因無法形成具體規範，故系統從略。

1、 拈二：每兩句一聯中，上、下句第二字平仄相對；每兩聯四句中，第二、三句第二字平仄相同。依此類推。

○🟡○○◎，○🔴○○●。○🔴○○○，○🟡○○●。

2、 換頭第一字：兩句一聯，每聯上、下句第一個字不可同仄（包括上、去、入聲），同平聲者為合格。

🟡○○○◎，🔴○○○●。🟡○○○○，🔴○○○●。

3、 護腰：兩句一聯，每聯上、下句第三字不可同上、同去、同入聲，同平聲者為合格。

○○🔵○◎，○○🟢○●。○○🔵○○，○○🟢○●。

「漢詩格律分析系統」目前雖尚處於研發階段，但業已初具規模，於民國一〇五年元月發表於中央研究院「小學堂」資料庫，如下圖：

系統主要分為「系統選單」、「檢索條件」、「聲調資料」和「格律分析結果」四個部分。使用者可輸入五言四句共廿字的漢詩文本，依據個人需求設定格律檢索條件。格律的檢索條件設計為評分函式（scoring function），依據對每條格律的重視程度來設計其分數。比如，「押韻」是一首詩成立的基本條件，在程式的設計上，即將此條件設定為能獲得最多分數者。亦即當某一種音韻組合不符合此項條件時，則無法被選出。經由評分設計，可以對各項格律的重要性進行區分。經由評分函式的運作，系統將自動計算出符合最多條件的分析結果，使用者亦可微調聲調資料，進行個人化的分析。

「系統選單」分為簡易查詢與進階查詢，簡易查詢節省「擷取聲韻」與「自動選音」兩個步驟，依據使用者所選擇的條件，直接呈現運算之後的結果；進階查詢則全面顯示系統運算的每一個步驟，供使用者檢驗，並可依使用者需求，調整細節，以達到個人化服務的效果。「檢索條件」包括詩文輸入欄、每句字數、韻書、格律條件等四個部分，目前已建置完成的部分包含五言、四句、廣韻、廣韻同用例、韻腳、沈約四病、元兢四病、元兢調聲三術等選項。「檢索結果」包含聲調資料和格律分析兩部分，上方為聲調資料顯示區，沿用漢字古今音系統的設計，針對詩歌文本的性質，預設值顯示詩文用字的聲調和韻目，使用者亦可依個人需求，在進階查詢部分檢閱、調整其他聲韻資料的細節；下方則為格律分析顯示區，除了韻腳分析為預設顯示項目外，其他的格律分析結果無論合格與否，均附上詳細說明，惟不合格者會以紅字標示，如下圖：

「漢詩格律分析系統」目前已建置完成五言四句、以《廣韻》為核心與三種格律規範的分析服務，未來將繼續就新增資料、開發技術、擴充研究等方面進行系統研發，並隨著語音與文獻資料的新增，調整使用者介面的操作與呈現方式。在格律規範部分，並不預設何種格律規範為唯一標準，而採用多種規範交叉檢索，以見典範的消長競爭，藉以提供變動的發展史觀。因此，接續將規劃增加盛唐以後「詩格」著作中的聲律理論，以及以王力為代表的「近世格律」模型，後者除代表著近體格律發展至當代學術所形成的基本共識，也可以作為與早期格律發展歷程比較的基礎，以凸顯漢詩在語音變化與格律形式之間的互動與流變。

三、「漢詩格律分析系統」設計過程中人文與數位的互動

「漢詩格律分析系統」的設計與建置，是欲通過數位技術與工具，以解決過往的「漢詩歌律」研究僅憑人力難以有效處理的困境，可視為發展迄今猶方興未艾的「數位人文學」的一個典型的成果。所謂「數位人文學」（Digital Humanities），在項潔、涂豐恩所著〈什麼是數位人文〉一文中曾予以簡要地說明：

> 它指的是那些唯有借助數位科技方能進行的人文研究。反過來講，數位人文的研究，即是企圖尋找在前數位時代中難以觀察的現象、無法想像的議題與無法進行的研究。[13]

[13] 項潔、涂豐恩：〈什麼是數位人文〉，收入項潔編：《從保存到創造：開啟數位人文研究》（臺北：臺

從這一段說明文字可以看到，「數位人文學」作為一門學問，係「以人文世界為核心關懷」。[14]與一般的人文學科不同的地方在於，「數位人文學」係以數位工具與技術「為人文研究提供了方法上的擴充」。[15]因此，所謂「數位人文」，顧名思義即是「『數位』和『人文』兩種不同領域的有機融合」，「是一個資訊技術與人文不斷互動的過程」。[16]也就是說，為共同解決一個人文研究的議題，人文學者與數位科技的研發者之間必須進行充分的對話與溝通，這是從事「數位人文學」研究的一個重要特徵。「漢詩格律分析系統」的設計與建置，正充分顯現這一特點。本文接續將探討「漢詩格律分析系統」在設計過程中人文與數位間的互動，希望以此為例，也可以提供其它欲從事或正在從事文學研究與數位技術相結合的研究者的參考。

重新審視「漢詩格律分析系統」的研發過程，大抵可區分為三個階段，各個階段也呈現出人文與數位之間的互動：

(一) 單一格律條件設定

系統設計之初，是從文學研究者欲節省時間與人力，使用數位工具來檢視巨量詩作是否滿足格律，以及滿足哪些格律條件的問題意識出發，乃將南齊永明（483—493）至初唐重要的格律規範，如押韻、平頭、上尾、蜂腰、鶴膝、第一字換頭、拈二和護腰等，逐項獨立設計，使系統進行各別格律條件的分析與判斷。然而，由於漢字具有「一字多形」的特性，如「東」和「东」是繁、簡體的區別，這對文學研究者完全不成問題，但就計算機的運作邏輯而言，若無法正確處理單個字形，便無法得到正確的聲韻資料，而影響分析的結果和品質。於是，在與數位研發者進行了充分的討論後，數位研發者為系統設計了一個異體字表，透過這個字表對異體字的定義，使得使用者在輸入各種異體字時，系統能順利地予以轉換成韻書的代表字。

除了「一字多形」的問題外，漢字也有「一字多音」的特性，對於計算機來說，可能導致運算出的音韻排列組合變得相當複雜。因此在系統設計之初，雖然能以資料庫查詢取代傳統人力檢索，但在格律分析的問題上，仍需仰賴人力挑選音韻，再以程式判斷是否合於格律，因此節省人力的程度仍相當有限。此外，僅設定單一格律條件讓使用者作選擇，其所反映的思維邏輯不免仍是「一以貫之」，而無法反映出各別格律條件的歷

大出版中心，2011），頁 11。

[14] 項潔、翁稷安：〈關於數位人文的思考：理論與方法〉，收入項潔編：《數位人文研究的新視野：基礎與想像》（臺北：臺大出版中心，2011），頁 13。

[15] 項潔、翁稷安：〈數位人文的變與不變〉，收入項潔編：《數位人文要義：尋找類型與軌跡》（臺北：臺大出版中心，2012），頁 14。

[16] 分見項潔、翁稷安：〈數位人文的變與不變〉，頁 15；項潔、涂豐恩：〈什麼是數位人文〉，頁 22。

史演變，如「蜂腰」和「平頭第一字」發展至初唐，同平聲的情況則被視為合格。

(二) 格律模型的設定與格律史的並呈

為呈現格律的動態發展樣貌，系統在格律條件的設計上，改採「沈約四病」、「元兢四病」和「元兢調聲三術」三種格律模型作為條件群組，提供使用者不同的觀察角度，也設計讓使用者可以選擇讓三種格律模型並呈，呈現出格律條件從南齊永明到初唐的歷時性演變脈絡。然而，在具體落實到系統的程式設計上，就面臨到一個文學研究者的思維與計算機思維的一個顯著的差異。對於文學研究者而言，在分析一首詩的格律時，格律是被視為一個整體來進行思考的。但對於計算機而言，則需要把具體的格律條件一一羅列，並決定其優先次序，程式才能為每種音韻組合是否合於格律進行評分進行分析判斷。這裡就反映出為了能呈現出文學研究者的需求，並解決前述一字多音所造成的格律分析的複雜性，進一步減輕人力在分析詩句的工作，數位研發者在程式設計上，採取以每個字為「節點」(node)，以字的語音為「分支」(edge)，將整首詩展開成一個樹狀結構，再配合樹狀結構的「走訪演算法」(tree traversal)，走訪每個葉節點。在此，一種走訪方式即代表著一種音韻組合。如此系統便可預選出一個最佳的音韻組合，再由使用者來決定是否要進行細部的調整。

格律條件方面，則設計為「評分函式」(scoring function)，讓系統可以根據對每條格律條件的重視程度來設計其分數。以五言詩為例，第二句和第四句的最末字為韻腳，必須押韻，這是一首詩成立的基本條件，在程式設計上即將此條件設定為能獲得最多分數者。當某一種音韻組合不符合此項條件時，則無法被選出。經由「評分函式」設計，便可以對各項格律的重要性進行區分。由此，也可以提供不同的格律組合，進行分析。

(三) 數位思維對於近世格律設計之影響

隨著南齊永明到初唐的代表性格律模型已基本建置，後續「漢詩格律分析系統」的開發也規劃納入以王力為代表的「近世格律」。蔡瑜先生組織其研究生助理們開了數次「近世通用格律」讀書會，旨在研討王力、呂正惠、許清雲和張夢機等學者的格律理論，並整理出哪些「近世格律」的規範是當今創作、教學和學術研究普遍通用的。

由於討論結果是為了「漢詩格律分析系統」中「近世格律」的建置，且此前為了如何讓系統具體呈現出符合文學研究者研究需求的格律分析結果，已與數位開發者有過數次討論，過程中對於數位思維也有一些基本認識。加之各家主張或強調的「近世格律」的規範甚為蕪雜，尤以王力的系統最為紛亂。而計算機思維的一個顯著特色在於，它必須獲得簡單而明確的指令才能順利進行。因此，在討論「近世格律」的格律規範時，便

帶著「如何將紛雜的格律規範化作明確的指令讓系統呈現」的問題意識進行思考與討論。「方法的變化，會帶來研究視野的變化。」[17]可說是人文與數位在經歷一段時間的對話與溝通後，自然會出現的現象。茲以「近世格律」中的「拗救」規範為例進行說明。

「近世格律」主要是由四個基本的平仄式組成，以五言為例，即：「仄仄平平仄」、「平平仄仄平」、「平平平仄仄」和「仄仄仄平平」。所謂「拗」，係指不合乎基本式中特定位置的字的平仄規定；所謂「拗救」，則是指出現「拗」時，用改變其它位置的字之平仄的方式以為補償。如「拗」而能「救」，便不算「病」。[18]經由比對上述諸家學者格律理論對於「拗救」的討論後，歸納整理出諸家共同強調的「拗救」原則有三：

1、 第一字拗，則當句第三字救（針對「平平仄仄平」）；

2、 第三字拗，則隔句第三字救（針對所有情況，但救後不可出現「下三平」）；

3、 第四字拗，則當句第三字救（針對「平平平仄仄」）

或隔句第三字救（針對「仄仄平平仄，平平仄仄平」）。

由這三點「拗救」原則可知，第一、三、四字發生「拗」的情況時，只要滿足上述三個原則中與拗字有關係的另一字的平仄與之相反，便有「救」的可能。反過來說，只有第二字和第五字一旦不合乎「近世格律」基本平仄式所規定的平仄，就一定是不合格律的。由此，便可以得到以下幾條關於「近世格律」的「拗救」規範：（綠底表示「拗」，黃底表示「救」）

1、 「仄平平仄平」合格律；

2、 「仄仄仄平仄，平平平仄平」合格律；

3、 「平平仄平仄」和「仄仄平仄仄，平平平仄平」合格律。

換言之，系統在分析一首詩是否符合「近世格律」時，首先判斷是否符合平仄基本式，接著再判斷有無符合「第二字粘對」、「無下三平」、「無拗」等規範，其中「拗」的部分只要滿足上述三種情況，便算「救」而非「病」。上述討論格律的取徑，已不盡同於傳統文學研究的思考，反映出文學研究者受到數位思維模式的影響來處理研究課題。

由上述「漢詩格律分析系統」研發的三個階段中所呈現出的人文與數位的互動可以

[17] 項潔、翁稷安：〈數位人文的變與不變〉，頁 14。
[18] 參見王力：〈關於「一三五不論」〉、〈拗救〉，收入氏著《漢語詩律學》，頁 83-91；91-100。

看到，雖然「數位系統應該要以人文研究作為思考的起點」，[19]但在系統設計的實際互動中，人文學者也可能受數位計算機思維邏輯的影響，而觸發新的思考方式。這或許不是人文學者從事數位人文研究必然會有的結果，卻是過程中可能碰撞出的特別的火花。

四、結論

本文以「漢詩格律分析系統」為例，討論數位技術如何應用於中國古典詩歌的格律研究。本文首先回顧了既有的格律研究成果，可以分為傳統方法與數位研究兩個部分。傳統的研究主要採取「逆向」的理解徑路，依據清人所擬定的聲調譜或平仄譜，來整理、歸納出格律規範。有些學者則依據唐人的「詩格」資料，嘗試以「順向」的理解徑路，依據唐人自身的論述還原唐代的格律觀。不管是採取「逆向」的抑「順向」的徑路，傳統的研究面臨到的一個共同的困境在於，欲以人力處理巨量詩作的格律分析，總是力有未逮。隨著數位科技的進展，前述的問題有了較好的解決工具。惟目前主要的數位格律分析系統，採用的皆是「近世格律」的規範，而未能反映出格律生成與發展的動態歷程。「近世格律」作為當今多數人所習用的格律規範，雖自成其理論體系，卻也不免存在著「以後律前」和「一以貫之」的問題，使中國詩歌格律的內涵變得平板和僵化。「漢詩格律分析系統」的設計與建置，其目的之一便是為要補足既有的格律研究所匱缺的面向，通過數位工具協助處理巨量的詩作分析，以協助重探中國詩歌格律發展的脈絡。

在本文的第二部分，介紹了「漢詩格律分析系統」從發想到設計的過程，以及目前已經建置完成的功能。第三部分，則重新反省了系統設計與建置的過程中，文學研究者與數位研發者之間的互動與溝通。系統的研發本係以文學研究的課題為核心關懷，因此在研發的過程中，文學研究者方面佔有相當的主導性。然而，在不斷溝通、對話的過程中，數位思維的邏輯與模式也漸漸對於文學研究者產生影響，從而開啟新的研究視野與徑路。這一點，具體反映在對於「近世格律」規範原則的整理工作上。

「數位人文學」作為一門尚處發展階段的學科，猶有許多的發展潛力。中國文學研究者在這一領域中的參與並不普遍，也未深入，是比較可惜的。雖然如此，既有的數位典藏與資料庫建置方面的成果，也已提供進一步深入研究，乃至於開發以研究為導向的數位研究系統的重要基礎。誠如項潔、翁稷安所指出的：「必須要有正確而完備的數位典藏和數位資料庫作為基礎，資訊和人文研究才能有更進一步合作的可能。」[20]而「漢

[19] 項潔、翁稷安：〈關於數位人文的思考：理論與方法〉，頁 16。
[20] 項潔、翁稷安：〈關於數位人文的思考：理論與方法〉，頁 14。

詩格律分析系統」正是奠基在既有文獻資料庫與語音資料庫的基礎上，進一步予以整合而開發出的以漢詩格律研究為取向的數位研究系統，屬於「數位人文學」從典藏到研究的一個嘗試，初步也有了一些成果。系統的雛型已經建置，且已公開於網路平臺。在接續的研發階段中，規劃整合既有的文獻資料庫，讓使用者可以直接在系統上點選詩作並進行分析，但過程中也遇到一些阻礙。其中一個現階段較難以克服的困難，在於目前對於「數位智財權」的認定仍未有共識，使得「漢詩格律分析系統」在欲進一步整合其它既有資料庫時，往往因「智財權」的認定不一，而無法順利取得部分資料庫的授權。「漢詩格律系統」在數位智財權方面遇到的問題，恐怕是許多數位系統建置時都會遭遇到的。此問題實非本文作者的能力足以給出答案的，謹盼對於數位智財權的議題能有更廣泛的討論，進而逐步凝聚共識，在不久的將來可以把相關規範明確地制定下來。

Panel D

宗教醫療數位平台之建置與應用

Creating a Digital Platform for the Study of Religious Medical Traditions

Panel D

宗教醫療數位平台之建置與應用

主持人	釋惠敏（法鼓文理學院校長、臺北藝術大學名譽教授、中華電子佛典協會主任委員） Huimin Bhikshu (Principal, Dharma Drum Institute of Liberal Arts; Emeritus Professor, Taipei National University of the Arts; Director, Chinese Buddhist Electronic Text Association)
發表人	洪振洲（法鼓文理學院佛教學系副教授） Jen-jou Hung (Associate Professor of Buddhist Studies, Dharma Drum Institute of Liberal Arts) 杜正民（法鼓文理學院佛教學系教授） Aming Tu (Professor of Buddhist Studies, Dharma Drum Institute of Liberal Arts) 黃舒鈴（法鼓文理學院專案研究助理） Shu-ling Huang (Project Research Assistant, Dharma Drum Institute of Liberal Arts)
題目	法的療癒資料庫研究與建置 Study and Building of a Dharma-Healing Database
發表人	徐源（德國馬克斯·普郎克科學史研究所博士後研究員） Michael Stanley-Baker (Postdoctoral Research Fellow of the Max Planck Institute for the History of Science, German)
題目	跟踪物質實踐：搜索情境化的亞洲醫藥知識與中國中古宗教文獻為例 Tracking Material Practice：Searching for Situated Knowledge of Asian Drugs in Medieval Chinese Religious Texts
發表人	梅靜軒（法鼓文理學院佛教學系助理教授） Ching-hsuan Mei (Assistant Professor of Buddhist Studies, Dharma Drum Institute of Liberal Arts)
題目	身體與聖藥：藏密與道教的跨宗教對話 Body and Sacred Medicine：A Dialogue between Tibetan Tantric Buddhism and Daoism

Panel D

宗教醫療數位平台之建置與應用

當代提及醫學這個語詞，主要涵攝起源自古希臘醫療框架，而後以近代科學為基礎而發展的西方醫學，以及根據東方傳統中國自然與人文哲學的中醫學。一般認知醫學屬於科學和生物技術領域，與宗教似乎有根本的差異，然而，宗教與醫療在幾千年的並存過程中持續相互影響與補充，乃是歷史事實。是自 19 世紀以來，近代醫學技術快速發展，使醫療型態轉變為高度專業分科化及大型機構化之後，才相對使得長久以來被信任的傳統醫療方式被慢慢地邊緣化，並被當成只是宗教行為或民俗醫療的一部份。因此，對於醫療與宗教兩者互涉關係的研究與檢討，無論對於了解歷史，或發掘傳統智慧以為今用，都是非常必要的。

道、佛、藏是在亞洲宗教醫療中很具代表性的三大傳統。早在先秦漢初的時候，處於形成過程中的中醫學就已深受道家思想影響，而啟發了未病先防，以及重視養心及治神的醫學思想基礎。兩漢時期傳入中國的佛教和東漢後期產生的道教，對中醫學的影響也無所不在。又佛教連同其醫藥於西元七世紀傳入西藏，雖當時期藏醫藥學已有相當程度的發展，在其既有基礎上融合了佛教醫藥，但在傳統上，藏族民眾幾乎都將其醫學體系歸溯到佛教中。

上述三大宗教醫療傳統各有其豐富的文獻資源，為了解亞洲宗教醫療的全貌，如果能用系統的方式建構或連結整合道教、漢地佛教和西藏佛教中，醫療相關的文獻資料庫，相信可有助於研究者更容易地取得跨領域的研究資料，以理解與探討亞洲宗教醫療相關論題。

Panel D
Creating a Digital Platform for the Study of Religious Medical Traditions

In present-day context, the term “Medicine” mainly affiliates with (a) Western medical tradition, which originates from ancient Greek medical schema and was further developed based on modern science, and (b) Chinese Medicine, which is based on traditional Chinese notions of philosophy of nature and the human being. Generally, medicine is perceived as a discipline on its own right placed in the domain of science and biotechnology. For this reason fundamental differences seem to exist between medicine and religion. However, the coexistence of both medicine and religion over the millennia has proven that they have sustained and complemented one another. Starting from the 19th century, the healthcare system became highly specialized and greatly institutionalized under the rapid development of modern medical technology, thus it gradually marginalized the long-trusted traditional medicine, confining it to being solely a part of religious practices or folk therapies. Studies and review on the interaction between medicine and religion is therefore pivotal, not only for historical appreciation, but also in order to uncover ancient wisdom for today’s applications.

Taoist, Buddhist in general and Tibetan medicine are the three main traditions which well represent Asian religious medical traditions. As early as the Qin to Han dynasty, the development of Chinese medicine was greatly influenced by Taoist thought and eventually constructed its fundamental medical epistemology as disease prevention as well as mind and spirit recuperation. The impact of Buddhism, which was introduced to China during the Han dynasty, and Taoism, which emerged in the late Han dynasty, on Chinese Medicine was ubiquitous. Buddhism, together with its medical practices, was introduced to Tibet in the 7th CE. Although the already well-developed indigenous medical tradition in Tibet integrated its fundamental practices with the introduction of Buddhist medicine, Tibetans traditionally attribute the origin of their medicinal practices to the introduction of Buddhism.

The three main religious medical traditions mentioned above are each rich in their resources. To better understand the macrocosm of Asian religious medical traditions, if sources from Taoism, Chinese Buddhism and Tibetan Buddhism could be systematically constructed as an integrated database, researchers might easily obtain interdisciplinary data for further studies and understanding of relevant topics within the scope of religious medicine..

法的療癒資料庫研究與建置

洪振洲*、杜正民**、黃舒鈴***

摘　要

佛陀以「老」、「病」、「死」三事出現於世，為眾生說所證法及調伏事，「醫方明」亦為研習佛法的五種學處之一。然在國內，以佛陀教理與佛教文獻，進行相關於「佛醫」（Dharma Healing）（包括治身的醫藥與治心的法藥）之研究，仍僅見零星討論，且大多僅針對單一經典或單一宗派思想進行研究。主因可能在於歷代佛教經典中，佛醫主題之相關內容散見於各經典，並未收錄於單一類別中，故研究者難以全方位收集相關資料進行大規模研究。

因此本計畫擬藉重數位人文技術，建置「法的療癒資料庫」。本資料庫的目的在於全面性的蒐集漢譯佛典中與佛醫相關的文獻內容，以協助研究者從巨觀的角度進行佛醫研究。本計畫擬蒐集之文獻，可依內容區分成兩大部分，包括：(1)以佛醫為主題之文獻(2)包含佛醫詞彙之文獻。後續並針對此兩部分文獻，進行內容分類、索引建置，並連結到佛教大藏經全文。

本計畫之成果，除可做為進行佛醫研究之優質資料來源之外，對於：瞭解與掌握佛醫的歷史演進與應用；學習在面對與處理病痛時，佛教提供的種種教法；以及瞭解把握生命乃至超越生死的佛典生命觀……等實際用途，亦期望能有所貢獻。

關鍵字：佛教、法的療癒、宗教醫療、大藏經、數位資料庫

* 法鼓文理學院佛教學系副教授，Email：jenjou.hung@dila.edu.tw。
** 法鼓文理學院佛教學系教授，Email：aming@dila.edu.tw。
*** 法鼓文理學院專案研究助理，Email：shuling.huang@dila.edu.tw。

Study and Building of a Dharma-Healing Database

Jen-jou Hung*, Aming Tu**, Shu-ling Huang***

Abstract

The Buddha appears in the world because of old age, sickness and death to preach the truth he has realized and expound that which should be quelled. Medicine is also one of the five fields of learning in Buddhism. However, in Taiwan there are only a handful of studies done on Dharma Healing (佛醫), which includes both physical and mental Buddhist therapy, based on Buddhist doctrines and texts. Most of these studies mainly focus on a single text, or are based on a particular school of thought. This may be due to the difficulty searching for data within such an enormous corpus of texts, as texts related to medicine are scattered throughout the canon. Dharma Healing texts have never been classified into a distinct collection throughout the history of texts' compilation.

Therefore, this project proposes to build a new database called 'Dharma-Healing Database' using the tools of the digital humanities. The purpose of this database is to gather all Dharma Healing related contents from Chinese Buddhist texts in the hope that it could aid in the effort of researchers to study Dharma Healing from a more comprehensive and macroscopic perspective. It intends to gather textual data from these two categories: (1) texts with Dharma Healing as their main theme, (2) texts containing terminology related to Dharma Healing. These texts will be stored in this database and further classified into smaller groups. Also indexes that link the terminology related to Dharma Healing and the full text pieces containing the terminology in the Buddhist Canon will be built.

This project hopes to produce a high quality data source for Dharma Healing research, but also enhance the understanding of the historical development and application of Dharma Healing, especially regarding the practical aspects of the Buddhist teachings, which the Buddha taught in order to face sickness, understand the meaning of this precious human life, and transcending life and death.

Keywords: Buddhism, Dharma Healing, Religious Medicine, Tripiṭaka, Database

---

* Associate Professor, Department of Buddhist Studies, Dharma Drum Institute of Liberal Arts. Email: jenjou.hung@dila.edu.tw.
** Professor, Department of Buddhist Studies, Dharma Drum Institute of Liberal Arts. Email: aming@dila.edu.tw.
*** Research Assistant, Dharma Drum Institute of Liberal Arts. Email: shuling.huang@dila.edu.tw.

緒論

佛陀以「老」、「病」、「死」三事出現於世，為眾生說所證法及調伏事，「醫方明」亦為研習佛法的五種學處之一。然在國內，以佛陀教理與佛教文獻，進行相關於「佛醫」（Dharma Healing）（包括治身的醫藥與治心的法藥）之研究，仍僅見零星討論，且大多僅針對單一經典或單一宗派思想進行研究[1]。主因可能在於歷代佛教經典中，佛醫主題之相關內容散見於各經典，並未收錄於單一類別中，故研究者難以全方位收集相關資料進行大規模研究。

本計畫擬關注此一佛教中與生命最切身相關的議題，藉重數位人文所發展之技術，建置稱為「法的療癒」之佛醫研究資料庫，以協助研究者從巨觀的角度進行佛醫研究。資料庫將以佛教大藏經為主要範圍，全面網羅佛醫研究第一手文獻，並提供能有效取得這些佛醫文獻之資訊檢索介面。本計畫成果，除可成為進行佛醫研究之優質資料來源之外，對於：(1)瞭解與掌握佛醫的歷史演進與應用；(2)學習在面對與處理病痛時，佛教提供的種種教法；以及(3)瞭解把握生命乃至超越生死的佛典生命觀……等實際用途，亦期望能有所貢獻。綜言之，期能追尋古德智慧，開展當代佛學研究內涵，有助於佛醫研究之推廣，並據以實踐「法的療癒」於日用。

本專案製作之「法的療癒」資料庫以佛醫研究為主題，將蒐集散諸於佛教藏經內之相關文獻，包括在原始佛教文獻（主要是阿含部與律部）中提及有關病苦、瞻病之記述，或診病、醫病的療法與藥方、還有在原始佛教後的三藏諸部類中，直接論及或部分涉及療病、醫藥以及醫喻等豐富文獻內容。由於，佛醫文獻跨越悠久的歷史年代，並且跨越漢、梵、巴、藏、英等多種佛典傳譯語言，本計畫預計先收集在漢譯佛典中與佛醫相關的文獻，再擴增到其他語言之佛典資料。初步完成後，視研究發展，可能跨越地域與學科，例如與印度、中亞等地的宗教醫療研究相互關涉；與現代醫藥和醫事、醫學人文、生命倫理、生死學等學科相互連結，進一步擴增本資料庫之可用性與參考性。

本計畫之重要性主要有三方面：一、為國內首見之以佛醫文獻為核心之主題式研究數位資料庫的建置計畫。過去的佛學數位文獻專案，大多以特定典籍為標的進行數位化；相較之下，主題式資料庫則以特定研究主題為核心，將相關的橫向與縱向資料加以收集與整合。以此方式建構之資料庫，更能符合特定領域之研究者之需求。二、可為國內推

[1] 查詢國內博碩士論文，可以發現處理佛醫議題的主要是碩士論文。年代較早的是黃文宏（2004）、陳麗彬（2005）和謝崧熙（2006）的碩士論文；沉寂近 10 年後，2013-6 年與佛醫議題直接相關的至少有 7 篇碩士論文（蘇蜂琪（2014）、李憶容（2015）、郭美純（2015）、廖彥博（2015）、黃子瑜（2016）、范明麗（2016）），近年來對此議題的關注似有增加的趨勢。相關論文都只處理特定範圍，如特定宗派、特定歷史期間或特定經典範圍等。

動佛醫研究建立重要基礎。本計畫為國內第一個旨在全面性及系統化長期推動佛醫研究的計畫。在本計畫建立佛醫經文全文之檢索與交互參照，並完成佛醫相關主題之結構化建構後，將期能成為於佛醫研究之基礎建設，進而促進佛醫研究的發展。三、透過與佛典數位化技術密切結合，本計畫的執行，有助於建置佛典主題式研究資料庫之經驗與能力之養成，而後可將此經驗與能力移轉應用於不同主題之佛典研究。

目標資料分析與資料庫建置

本計畫目標在於收集所有在漢譯佛典中與佛醫相關的文獻內容，細部將文獻內容區分成兩大部分，包括：(1)佛醫主題文獻，意指文獻內容與佛醫研究主題直接相關的文獻。(2)包含佛醫詞彙之文獻，意指文獻主旨並不一定直接與佛醫研究主題相關，但可能因為某些原因而在文字內容用了與佛醫研究主題相關的詞彙(如：藥材藥方、疾病與症狀、身體部分…等等)。在此兩種範圍定義之下的文獻集合，並非為兩個互斥集合，而是兩者間應具有一定程度重疊，如下圖 1 所示。但若以文件屬性來看，兩類文獻實具有相當程度的差異。就主題相關性來說，上述第一類的「佛醫主題文獻」較貼近資料庫的主題，可說是本資料庫的核心資料集，而第二類的文獻內容較類似於廣泛性的補足本資料庫的涵蓋範圍。而在文獻數量上，「佛醫主題文獻」的數量可能遠低於「包含佛醫詞彙之文獻」。所以在資料庫建置的策略上，為使有限的資源能發揮最大的效果，我們預計針對第一類的「佛醫主題文獻」，進行文獻研究，儘可能找出此類文獻的完整清單、據以建立數位目錄，並擷取相關內容建置數位全文，收錄於資料庫之內。此外，有些文獻內容具有其他多語文本之對應，針對此類多語對應資料，也將進一步進行多語文本的數位化處理，並提供多語對讀介面。而針對「包含佛醫詞彙之文獻」之部分，資料庫內將不收錄全文，而是改以紀錄相關詞彙的清單，與該詞彙出現於文章內之位置來取代。

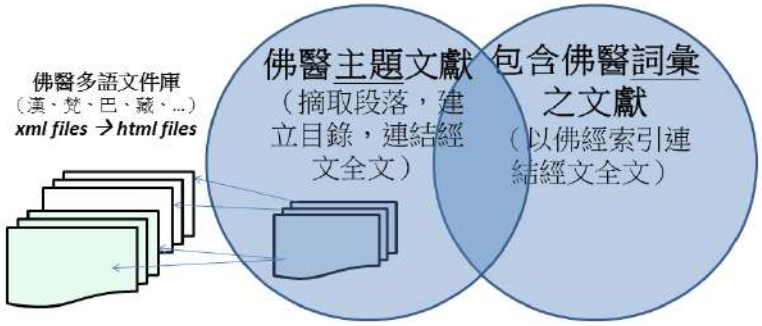


圖 1 計畫架構圖

文類 1: 佛醫主題文獻

最接近佛醫主題的文獻，當屬在三藏中直接論及療病、醫藥之內容。佛教典籍部帙龐大，要在茫茫經海中尋找佛醫主題文獻，並非易事。而目前佛醫主題的文獻的發表著

作方面，僅有《中國醫藥全書》[2]為較具規模的作品。此書全 100 冊，收錄論醫佛經、涉醫佛經、僧人醫著和居士醫著四部分內容，本計畫只參考第 1～23 冊的「論醫佛經」和「涉醫佛經」兩部分。此書雖然收錄大量所謂論醫和涉醫資料，但收錄原則不甚明確，也並未就佛醫論題提出參考性架構。雖然此書內容仍屬粗糙，但若先依此為出發點，就內容進行精讀、分類及整理，重新選錄出有關佛醫之段落，仍不失為一可行之道。而在實際進行資料處理過程中，考量到漢譯佛教文獻內容，也與佛教之思想發展以及在中國傳播的過程有關。因此，依照佛教發展脈絡的時間先後順序進行資料查找與精讀處理，再逐步往後開展，當為合理，故我們由代表原始佛教時期的《阿含經》和律部文獻著手進行。《阿含經》中記錄了豐富的佛陀有關「老」、「病」、「死」直接的身教、言教；在律部中，則有許多耆婆醫王建議比丘衛生保健事項的相關記載，以及療治比丘疾病的記錄。再往後發展的諸經論中，亦可找到許多佛醫主題文獻。

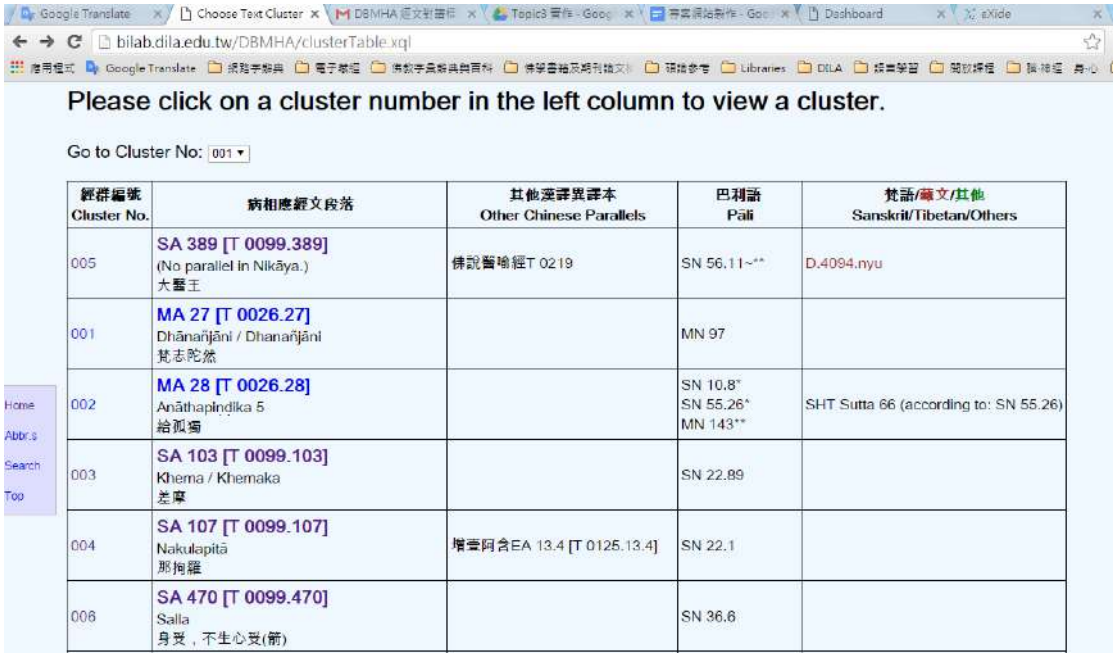
當我們尋找到確實相關於佛醫主題的文獻段落後，我們將該段落的基本資料(經名、經號、頁碼、起訖行號)等訊息，紀錄於我們的工作資料庫中。後續將利用此段資訊，由 CBETA2016 光碟中所附之原始檔案中，擷取相對應段落之 XML（eXtensible Markup Language）內容，加上適當的 TEI 框架後，後續儲存於本專案的 XML 資料庫之中[3]。

文類 2: 多語對應文本

在佛醫主題文獻的文獻集合中，如上節所述，漢譯《阿含經》是主要相關經典的來源之一。由於漢譯《阿含經》目前仍有不少多語對照的平行譯本，與豐碩的優質二手研究文獻，這些相關資料是研究漢譯《阿含經》內容時，不可或缺的重要參考資料。因此在呈現漢譯《阿含經》相關佛醫主題文獻經文的內容時，除直接顯示佛醫相關之段落外，我們將也參考《阿含經》多語對照相關研究成果，收集與其個別平行的多語文獻。接著透過網頁介面清晰地呈現漢譯《阿含經》佛醫主題文獻與其多語平行經間的對照關係，並提供平行經之間的對讀介面。這部分已有階段性成果如圖 2、圖 3。

[2] 釋永信、李良松（2011）。《中國佛教醫藥全書》。北京：中國書店出版社。

[3] CBETA 原始資料即為符合 TEI 規範之 XML 格式。



Please click on a cluster number in the left column to view a cluster.

Go to Cluster No: 001

經群編號 Cluster No.	病相應經文段落	其他漢譯異譯本 Other Chinese Parallels	巴利語 Pāli	梵語/藏文/其他 Sanskrit/Tibetan/Others
005	SA 389 [T 0099.389] (No parallel in Nikāya.) 大醫王	佛說譬喻經T 0219	SN 56.11~**	D.4094.nyu
001	MA 27 [T 0026.27] Dhānañjāni / Dhanañjāni 梵志陀然		MN 97	
002	MA 28 [T 0026.28] Anāthapiṇḍika 5 給孤獨		SN 10.8* SN 55.26^ MN 143**	SHT Sutta 66 (according to: SN 55.26)
003	SA 103 [T 0099.103] Khema / Khemaka 差摩		SN 22.89	
004	SA 107 [T 0099.107] Nakulapitā 那拘羅	增壹阿含EA 13.4 [T 0125.13.4]	SN 22.1	
006	SA 470 [T 0099.470] Salla 身受，不生心受(箭)		SN 36.6	

圖 2 漢譯《阿含經》佛醫主題文獻與其多語平行經對照表

SA 389 [T 0099.389] -- (No parallel in Nikāya.) 大醫王

Text-cluster:

SA389

T0219

圖 3 漢譯《阿含經》部分之佛醫主題文獻平行經對讀介面

文類 3：包含佛醫詞彙之文獻

而針對「包含佛醫詞彙之文獻」之部分，資料庫內將不收錄全文，而是改以紀錄相關詞彙的清單，與該詞彙出現於文章內之位置來取代。而這些相關的詞彙來源，將以《大藏經索引》（以下簡稱《索引》）內的資料為主要來源。《索引》由日本大藏經學術研究會邀請六所佛教大學負責編撰而成，是根據《大正新修大藏經》（以下簡稱《大正藏》）中之印度、中國、日本等三國撰述共 85 冊之內容，包括佛教名相與各種詞彙，依其在經文脈絡中的意義，進行系統性的分類，作成共 48 冊索引以利學者應用。其中本研究所關注的，是《大正藏》第 1-55 冊的印度和中國撰述，對應《索引》第 1-31 冊；以及

在《索引》所訂的 50 種分類項目中，選取包含較多佛醫相關詞彙的四項類別（「心理」、「生理・衛生」、「醫術・藥學」和「民族」），做為處理範圍。如同處理前述「佛醫主題文獻」之處理順序，針對詞彙執行建立連結經文全文的索引資料的工作時，我們也將先從阿含部再到律部及諸部等，逐步建置以佛經詞彙連結經文全文的功能。

數位化建置佛醫詞彙有三個主要目的：一、根據詞彙查找有意義的佛醫經文段落：可能可以依《索引》提供的佛醫詞彙，查找出有意義的佛醫經文段落。二、簡化並取代人工查找程序並提供線上經文查閱：《索引》雖提供索引資訊回查詞彙所在經文段落，但人工查閱是兩段式查閱，十分耗工費時。故擬將目前的紙本查找方式，建置成線上詞彙檢索系統，可直接由詞彙連結到經文段落查閱經文全文。三、可供未來進一步研究運用：在將詞彙數位化之後，或可便於針對佛醫詞彙進行時空分析或統計分析等，進一步發現有趣的研究議題或發展多元化的應用方案。佛醫詞彙建置已有階段性成果如圖 4。

圖 4 包含佛醫詞彙之文獻檢索

小結

本計畫目前仍在持續規劃與進行當中，未來仍有許多發展之可能性。在佛醫主題文獻方面，我們會繼續豐富佛醫文獻資料庫的內容，提供學習者與研究者使用。在佛醫詞彙處理方面，由於資料量大，處理過程繁瑣，將先以改善處理流程為目標，尋求更多數

位化分析模式或工具，以簡化資料處理過程並有效輸出質量穩定的數位化原始檔。

參考文獻

李憶容。2015。《佛教疾病療癒感應案例之生命轉化歷程探討》。南華大學生死學系碩士論文。未出版。

林秀砡。2013。《《藥師經》醫療觀之探析》。華梵大學東方人文思想研究所碩士論文。未出版。

范明麗。2016。《當代醫療方式轉向生命療癒之進路——以華嚴無盡網絡之圓融思想為導向》。華梵大學東方人文思想研究所碩士論文。未出版。

郭美純。2015。《宗教應用於身心療癒之歷程--佛教導向之個案研究》。樹德科技大學經營管理研究所碩士論文。未出版。

陳麗彬（釋見寰）。2005。《《雜阿含經》中佛陀對病苦的教示之研究》。華梵大學東方人文思想研究所碩士論文。未出版。

黃子瑜。2016。《漢魏兩晉南北朝佛教醫療觀的傳入與影響》。國立政治大學歷史學系碩士論文。未出版。

黃文宏。2004。《從神醫耆婆之醫療事蹟論其醫療方法及對佛教影響》。輔仁大學宗教學系碩士論文。未出版。

廖彥博。2015。《鬼魅侮身——道教和佛教實務處理之探討》。輔仁大學宗教學系碩士在職專班碩士論文。未出版。

謝崧熙。2006。《由整體療癒的觀點談疾病防治──以天台觀行與系統思考為進路》。南華大學哲學研究所碩士論文。未出版。

蘇蜂琪。2014。《佛教醫學思想之研究－以《阿含經》為論述核心》。南華大學宗教學研究所碩士論文。未出版。

釋永信、李良松。2011。《中國佛教醫藥全書》。北京：中國書店出版社。

身體與聖藥：藏密與道教的跨宗教對話

梅靜軒*

摘　要

本研究出發點是立基於前人所觀察到的，承認佛教與道教這兩個宗教傳統有諸多共同點的前提下展開。而切入的面向是從宗教醫療的角度分析所謂「轉凡成聖」的轉化機制是如何發生作用的。要特別說明的是，雖然佛教與道教所追求的目標—成就佛果抑或是成仙在本質上不同，筆者無意從佛教本位立場將道教神祈置於佛教的宇宙觀脈絡中。本研究擬將佛教與道教視為兩個平行發展的系統來分析其間的概念流動與交涉現象。一般而言，粗質身(coarse body)需要物質的滋養；而微細身(subtle body)的形成則有賴心念專注的訓練。如何點石成金地使肉身脫胎換骨轉化為非物質性的存在是十分有趣的一種宗教論述。研究初期將以涉及身心兩層面的法藥/聖藥作為比較研究的開端，期望法藥的生產與效用可使我們得以一窺另類的醫病關係。

關鍵字：身體、藏藥、聖藥、療癒、鍊金術

* 法鼓文理學院佛教學系助理教授，Email：chmei99@dila.edu.tw。

Body and Sacred Medicine：A Dialogue between Tibetan Tantric Buddhism and Daoism

Ching-hsuan Mei*

Abstract

This study is based on the observation of previous researches done by the pioneers of Tantric Buddhist and Daoist scholars. As they have already pointed out there are several common phenomenon in Tantric Buddhism and Daoism. I will follow up their observation to investigate the mechanism of transforming the physical body into sacred being from the perspective of medical-religion. Although Buddhism and Daoism have fundamental difference in terms of their ultimate goal, I have no intention to place Daoism under the cosmology of Buddhism but rather consider them as parallel religion system to analysis their interaction. General speaking, coarse body requires the nutrition of materials. The formation of subtle body is relied on the training of strict mental contemplation. Physical transformation is a very interesting religious discourse. On the preliminary stage of comparison, I will begin with the making of dharma medicine/sacred medicine which involves body and mind two aspects. Hopefully through this we will be able to unveil the undiscovered doctor-patient relationship.

Keywords: body, Tibetan medicine, sacred medicine, healing, alchemy

* Assistant Professor, Department of Buddhist Studies, Dharma Drum Institute of Liberal Arts. Email: chmei99@dila.edu.tw.

一、前言

中世紀印度的成就者運動提倡透過密續的修持、瑜伽、煉金術的鍛鍊，凡人可以晉級為成就者。此觀點認為，修行人圓滿自身，將可轉化為與理想的聖者一致的狀態。凡夫希望擁有一個不受死亡束縛的身體願望，最早可見於大約西元前 1200 年的《梨俱吠陀》(*Rig-veda*)的文獻中。而在《阿嚩婆吠陀》(*Atharva-veda*)中，學者發現大量關於使用咒語與藥方來恢復病人健康的醫療讚頌。後來的《阿育吠陀》(*Ayur-veda*)醫學中繼承了阿嚩婆吠陀傳統並發展出有名的回春的療法—rasayana。吠陀理論家從液態的汞與火性的硫磺起作用發生蛻變的觀察，歸納出水、火、風三要素的結合能使身體質變，產生轉化的結論。[1]

轉化肉身的觀念不只出現在南亞次大陸，東亞宗教同樣流傳著這樣的信念。Michel Strickmann (1981)在一會議論文集簡介中，回顧了藏學研究先驅石泰安(R.A.Stein, 1911-1999)的生平著作；其中特別引人注目的是他學術生涯晚年所熱衷的密續佛教與道教的比較研究。這兩個宗教傳統不論是在經典的書寫生產或是教法、修行人的特徵、儀軌表現等都有"驚人的"相似之處。[2] Michel Strickmann 自己也曾以中國為主要研究場域，探討了道教觀點的疾病與幾種具療癒功效的儀式類型。他指出療癒性儀式散見於中國密教-道教信仰脈絡中，例如符印的運用、驅邪儀式劇展等。

本研究出發點是立基於前人所觀察到的，承認佛教與道教這兩個宗教傳統有諸多共同點的前提下展開。而切入的面向是從宗教醫療的角度分析所謂「轉凡成聖」的轉化機制是如何發生作用的。要特別說明的是，雖然佛教與道教所追求的目標—成就佛果抑或是成仙在本質上不同，筆者無意從佛教本位立場將道教神祈置於佛教的宇宙觀脈絡中。本研究擬將佛教與道教視為兩個平行發展的系統來分析其間的概念流動與交涉現象。一般而言，粗質身(coarse body)需要物質的滋養；而微細身(subtle body)的形成則有賴心念專注的訓練。如何點石成金地使肉身脫胎換骨轉化為非物質性的存在是十分有趣的一種宗教論述。

二、身體與聖藥

在印度文獻中，最早有大量微細身概念的討論出現在奧義書中。特別是 *Taittiriya Upaniṣad* 中關於五種身或五種我的解釋，從需要食物來維繫的物質身到越來越微細、抽

[1] David Gordon White (2004), *The Alchemical Body: Siddha Traditions in Medieval India*, New Delhi: Munshiram Manoharlal.

[2] Micheh Strickmann (1981: VII-XV), *Tantric and Taoist Studies I: in Honour of R.A.Stein*, Bruxelles: Institut Belge Des Hautes Etudes Chinoises.

象的身體，諸如肉身(*anna-maya*)，氣息身(*prā&a-maya*)，意身(*mano-maya*)，識身(*vijñāna-maya*)以及自成樂身(*ananda-maya*)。此書還談到了微細身的內在剖析，其中有一個中脈、以及從中脈往四方擴展的各支脈。類似的概念在早期的印度奧義書類的文獻中、同時期的希臘文書中以及晚近出土的西元前一世紀的道教文獻中都也有類似的概念。[3] 而這些各類形式的身體的概念都是本研究所涉及的範圍。

身體的完美或缺陷乃至使人陷於痛苦狀態構成了「病」的概念；而減緩、逆轉疾病帶來的苦楚則是用藥的目的。由於此研究採取廣泛的身體定義，與之相對應的病與藥所涵蓋的範圍也隨之擴張。而這裡所謂的聖藥或法藥可簡要歸納為兩大類。一是指經儀式的過程聖化了原本是物質性的藥材，使之具有神聖的加持增添了藥效，因而得以療癒身心。另一個層次的聖藥專指意識轉化，幫助人從痛苦煩惱解脫的精神指導。以下將分別簡要說明密續佛教與道教視角下的身體以及這兩個重要的亞洲傳統宗教如何結合醫療知識以達到肉身轉化之理想。由於密續佛教在印度的發展是一個複雜龐大的主題，本研究的文獻將以西藏佛教所吸收內化的密續思想與教法為主要參考。

* 西藏密續佛教的身體

關於西藏佛教密續的身體觀，時輪金剛續文獻中有最詳盡的討論。Wallace (2009)曾指出密續佛教的論述足以證實身體除了是物質肉身之外，也涵蓋了非物質的意念狀態乃至純淨的無二本覺。[4] 這對於身體的討論提供了一個更開闊的空間。從時輪密續的角度來看，達到純淨佛果有諸多異名，如成就智慧身(*jñāna-kāya*)、俱生身(*sahaja-kāya*)、大樂身(*mahāsukha-kāya*)或清淨身(*viśuddha-kāya*)。而所謂的智慧身是已解脫無名煩惱的束縛與意識的層層障礙且不受物質限制而展現光明。可說是將大乘佛教的空性、無我概念延伸至無身與心二元組合，亦即從五蘊身心超脫的無質礙狀態。而正因為這種無質礙狀態使得智慧佛身得以遍佈一切時、一切處，如虛空般存在於各種物質世界。據此，圓滿智慧身的法藥可以廣義的包羅佛陀的教法。

回到凡俗的世界層面來說，身體的健康可歸納為身、心兩個面向。西藏醫學理論的根據主要來自於十二世紀的伏藏《四部醫典》。此部醫典有如藏醫史上最具權威的醫學百科全書。日後的醫者多半根據《四部醫典》理論再加以闡述詮釋。又因地理位置、氣候的差異，發展出南北醫學兩大學派。簡要來說，藏醫的基本理論是建立在隆(*rlung*)、

[3] Geoffrey Samuel (2013: 33-47), "The Subtle Body in India and Beyond" in *Religion and the Subtle Body in Asia and the West*, edited by G. Samuel and J. Johnston, London& New York: Routledge.
[4] Vesna Wallace (2009), "Why is the Bodiless (a$anga) Gnostic Body (jñāna-kāya) Considered a Body" in *Journal of Indian Philosophy*, 37:45-60.

赤巴(*mkhris pa*)、培根(*bad kan*)—風、火、地與水這三因的平衡。人體內的三因支配了七種物質(乳糜、血、肉、脂、骨、髓、精)與三種排泄物(糞、尿、汗)的運轉，若相互協調則得以維持健康，若失衡則百病叢生。不過 17 世紀左右，實證經驗開始挑戰西藏傳統對身體的認知。佛教密續理論所探討的氣、脈 (*cakra-nāḍi*)這些構成微細身的組成要素是肉眼所不能見的。該如何在神聖的佛語與經驗的觀察之間做出抉擇是藏醫們在西藏邁向現代化過程中所要面對的考驗。[5]

本研究的初期將以涉及身心兩層面的法藥/聖藥作為研究的開端。一般而言，「法藥成就術」(*sman sgrub*)指的是一系列準備與製作藥物的程序，透過儀式加持使藥物之療效倍增外，也對執行儀式者有益。修法者可獲得長壽、神通，乃至認識心性本質而得以轉穢為淨，並具有治療魔障病的能力。換句話說，「法藥成就術」除了轉化了外在藥材的特性，也因禪修之力轉化了修法者自身的狀態。Frances Garrett (2009) 指出「法藥成就術」是許多西藏的宗教活動之一，並不只限於醫學傳統的文獻。這個修法與西藏煉金術、植物學、瑜伽修行以及禪修都有密切關係。雖然最初源自於印度但西藏發展出自己的傳承與著作。不同的傳統各自有不同的法藥修練、藥材準備、生產製作等方法。[6] 研究文獻方面，將以蔣貢康楚 ('Jam mgon kong sprul Blo gros mtha' yas, 1813-1899) 所編輯的《大寶伏藏》(*Rin chen gter mdzod chen mo*) 作為核心文獻。上述的「法藥成就術」是流傳於寧瑪派八成就法中之一的「甘露功德」 (*bdud rtsi yon tan*)，這類的教法被收編在《大寶伏藏》的第四十五到第四十八冊中。

2007 年 TBRC[7]已經部分完成雪謙版《大寶伏藏》[8]的數位化，並於網路上提供全文電子文本(etext)。因此本研究擬結合資訊技術，發展相關電腦程式搭配《大寶伏藏》之電子文本，嘗試分析法藥的藥方類別、製作方法與效益。期望進而得以與道藏相關的資料庫連結，以更有效率的方式進行文獻的比較分析。希望透過不同資料庫的連結，生產傳統人文研究難以企及的量化，實踐更為全面性的跨宗教對話的理想。

---

[5]當藏醫們從實際的解剖經驗觀察身體的組成並進而發現了一些與經典記載相違的事實時，他們面臨了該如何順從傳統權威的難題。見 Janet Gyatso (2013) "Experience, empiricism, and the fortunes of authority" in *The Tibetan History Reader*, New York: Columbia University Press. 另外關於前現代的西藏醫學發展與佛教關係的探討，參考 Janet Gyatso (2015), *Being human in a Buddhist world: an intellectual history of medicine in early modern Tibet*, New York: Columbia University Press.

[6] Frances Garrett (2009), "The Alchemy of Accomplishing Medicine (*sman sgrub*): Situating the Yuthog Heart Essence (*G.yu thog snying thig*) in Literature and History" in *Journal of Indian Philosophy*, vol.37: 207-230.

[7] The Tibetan Buddhist Resource Center，http://www.tbrc.org/，2016/10/23.

[8] 這是由 Tsadra 基金會所支助完成的電腦輸入版《大寶伏藏》，由雪謙寺出版，以下稱之為雪謙版《大寶伏藏》。全集共七十冊，見 http://www.tbrc.org/#!rid=W1KG14/，2015/12/25。現有的 54 冊電子文本已在該網站上公開，供線上查詢了。我們也已取得該 54 冊資料之 XML 格式原始檔，用做為數位分析的來源資料。

Panel E

網路哲學

The Philosophy of the Internet

Panel E

網路哲學

主持人	蔡偉鼎（東海大學哲學系助理教授） Wei-ding Tsai (Assistant professor of Department of Philosophy, Tunghai University)
發表人	洪世謙（國立中山大學哲學所副教授） Shih-chian Hung (Associate professor of Institute of Philosophy, National Sun Yat-sen University)
題目	數位年代中對物的重新追問 The Question Concerning the Thing in the Digital Epoch
發表人	高國魁（國立政治大學社會學系助理教授） Kuo-kuei Kao (Assistant professor of Department of Sociology, National Chengchi University)
題目	數位人文的現象學還原：從擬象到檔案 The Phenomenological Reductions of Digital Humanity：From Simulacrum to Archive
發表人	楊士奇（弘光科技大學文化創意產業系助理教授） Shi-chi Yang (Assistant professor of Department of Cultural and Creative Industries, Hung Kung University)
題目	數位時代的人文反思：以大數據為線索 Humanistic Reflection in Digital Era：Based on Big Data Problems
發表人	蔡偉鼎（東海大學哲學系助理教授） Wei-ding Tsai (Assistant professor of Department of Philosophy, Tunghai University)
題目	論大數據的知識論條件 On Epistemological Conditions of Big Data

Panel E

網路哲學

本分組將「網路哲學」（The Philosophy of the Internet）設定為討論主題，以作為「數位人文」範疇下的一個子議題。所謂「網路哲學」，我們將之定義為：以一切**在**網際網路上、**隨著**網際網路、且／或**透過**網際網路（on, with and/or through the Internet）所發生的各種人文現象作為研究對象，並對之進行奠基性的哲學探討，以求深入考察今日諸般方興未艾的相關論題。本分組在本年度數位人文國際研討會中將聚焦討論以網路作為**生活世界**（Lebenswelt / life world）所促發的某些人文現象之特徵，以期對其進行理論性反思。

今日的網際網路幾乎已全面滲透到人類文明的各個層面，而成為一般人於日常生活中不可或缺的重要部份，且無疑已對吾人帶來諸多實質的影響。有鑒於電腦網路科技之發展日新月異、隨著數位化所帶來的生活急劇變化，各式各樣相關的新興問題遂應運而生，以致當今吾人所習以為常的日常生活與世界觀刻正急迫面臨著挑戰。尤其網路現已不再只被簡單看成是個將眾多電腦硬體連結在一起的資訊網絡，亦不再能被僅僅當作為一個可讓人暫時逃脫現實生活的虛擬場域。事實上，其毋寧更是一種不斷將虛擬世界聯結上現實世界而融合在一起的**擴增實境**（augmented reality）。在此，我們係**廣義地**理解「擴增實境」一詞，意指虛擬世界與現實世界交疊互動而混成一體——就此而言，其亦可被理解為「混合實境」（mixed reality）。也正因為今日虛擬世界與現實世界越來越緊密地交織在一起，迫使以往將這兩者理所當然地截然二分的傳統思維模式面臨挑戰，從而要求吾人對網路世界裡生發的事件進行更深入的哲學反思，以俾能一方面重新反思其就實踐哲學來看所蘊含的具體意義及實質效果，另一方面從新的視角來重構其存有學暨知識論上的理論基礎，並在最後通過審慎明辨箇中問題後對之提出適切的可能解釋與因應之道。因此之故，我們希望能透過組織「網路哲學」之分組專題討論，邀請已涉足此一領域的學者們提出其最新研究成果，並透過會議現場上的即席討論來刺激彼此思想之交流，增進彼此學習之機會，俾使哲學界同好對於相關議題有更多的密集討論。

考慮到與網路相關之具體議題範圍極廣，為求使現場討論能更為聚焦，以俾論文發表者與聽眾間均能充分且深入地進行爭論析辯，本分組於本年度特別設定以胡賽爾的「生活世界」（Lebenswelt）概念作為引導線索，試圖將擴增實境理解為數位時代的生活世界，以考察人類在數位網路科技影響下的實際生活中所面臨的結構性變化及其諸般問題。此一設定之背後有一前提，亦即：有鑑於作為擴增實境的網路世界無疑**即將**、甚至對某些人來說**早已**成為人類實際生存的**常態生活環境**，這不但意味著數位科技已在不知不覺中逐漸改變了吾人先前所習以為常的**世界觀**（Weltanschauung），更表示我們在某種程度上已能合理地將網路世界視作為人類的日常生活世界。在這樣一種已結合數位化的人文生活環境裡，表面上看似僅只在網路上呈現的「虛擬」行為終究最是會相應地在實際生活中產生某種程度的實效—不論其是正面的或是負面的效果。最難以被人忽

視、而且極具戲劇張力的例證，莫過於國際間自 2010 年起出現一連串透過社群媒體營造出來的政治活動（譬如突尼西亞的茉莉花革命、太陽花學運），其顯示社群媒體的使用者不再都是一群在虛擬空間裡說嘴取暖的逃離現實者，從而顛覆了先前網路文化觀察者們的諸多斷言—譬如其最常琅琅上口的「萬人響應，無人到場」。總之，諸般虛實交融的實例再再指出了今日的世界觀與生活世界皆已產生某種轉變，而這將迫使吾人醒悟到實有必要去探究其背後的結構及理論基礎。

今年度本分組一共提出四篇論文，分別探討網路世界裡的物質與非物質的關係、數位模控的社會系統如何展現、大數據在政治倫理及知識論面向上的問題。這四篇論文就論述內容而言，圍繞著網路現象之形上學、知識論、政治哲學、社會哲學等面向之意含來追問其本質性特徵。此外，我們還按照論文內容題材的相關度，將這四篇論文分成兩組來發表，並安排論文發表人兩兩互評，以期透過這種方式營造充分的對話空間，相信亦能讓與會者一同參與深入且實質的討論。

Panel E
The Philosophy of the Internet

The theme of this panel is "philosophy of the Internet" which could be regarded as a sub-topic under the category "digital humanities." The so-called "philosophy of the Internet" sees as object of its study all kinds of human phenomena occurred in the Internet, with the Internet, and/or through the Internet, and researches their philosophical foundations, in order to examine all relevant topics in depth. The discussion of this panel will focus on the characteristics of certain phenomena in the Internet, in order to do some theoretical reflections on them.

Today, the Internet as a life-world has almost penetrated into all levels of human civilization, and becomes an important part for everyday life. It is no doubt that the Internet has brought a lot of substantial impacts on us. In view of rapid development of computer network technology and drastic changes with digitalization of human life, many kinds of related problematic issues come into being, so that our everyday life and worldview nowadays is facing challenge urgently. In particular, the Internet is no longer simply seen as a connected network of computer hardware, and cannot be regarded only as a virtual field which let people escape from their real lives temporarily. In fact, it is more likely an augmented reality which constantly connects and integrates the virtual world with the real world. We understand the term "augmented reality" here in its broad sense and mean that the virtual world interacts and intermingles with the real world, so that both worlds can fuse into a unity – in this context, it can also be understood as "mixed reality." Since the virtual world and the real world intertwine with each other more and more closely now, the traditional way of thinking that both worlds naturally separate from each other, is facing challenges. We are required to rethink events happened in the Internet more deeply and philosophically, in order to be able to re-reflect on their concrete meanings and substantive results from the perspective of practical philosophy on the one hand, and to reconstruct their ontological-epistemological basis from a new perspective on the other hand. Finally, after deliberating on such problematic issues, we need to put forward some appropriate possible explanations or solutions to those problems. Because of the above-mentioned considerations, we organize a panel discussion on "Philosophy of the Internet" and invite those scholars who have been involved in this field to present their latest research results to the audience. By doing so, we want to stimulate more exchange of ideas between scholars, especially philosophers, so that they can learn from each other and focus more intensively on the philosophical topics about the Internet.

In order to make the discussion even more focused, our panel in the 7th Conference of DADH especially sets Husserl's concept of "life-world" (Lebenswelt) as a guide clue for the authors and the audience and base on such understanding to observe the structural changes and problems which the digital technology of the Internet brings to our daily life. In other words,

we try to understand augmented reality as our life-world in the digital age, since the online world as augmented reality is no doubt about to, or even already for some people, be the normal living environment for human's existence. This means not only that the digital technology has somehow gradually changed our worldview (Weltanschauung) we used to have, but also that we can now to some extent reasonably view the online world as world of human's daily life. In such a human living environment combined with digitization, the "virtual" behavior which seems to appear only in the Internet, will, in the end, have some concrete effects on real life - whether they are positive or not negative. In short, there are many cases of augmented reality which show us that the worldview and life-world in the digital age is much different from twenty years ago, and thus force us to explore the underlying structure and theoretical basis of human phenomena in the Internet.

There are four papers presented in this panel: Prof. Shih Chian Hung's article "Rethinking the Objects in the Digital Era" describes how the relationship between human beings and things has been changed in the digital world, so that we should re- define the meaning of thing and material. Prof. Hung tries to point out that both social relationship and personal relationship will be changed along with the new definition of thing and thus need a new philosophical anthropology to explain those new relationships. Prof. Kuo-kuei Kao's essay "The Phenomenological Reduction of Digital Humanities: From Simulacra to Archives" tries to answer the question: Whether the social systems of totally digitalized cybernetics will preserve some real traces of oral humanity or leave a few fictional marks of print humanity? Prof. Kao argues that behind the epistemological environment of digital humanities lies the political nature of technique and media. Though discussing with Heidegger, Derrida, Baudrillard and Groys, he indicates in the end a radical transformation of phenomenology. Prof. Shi-chi Yang's article "Humanistic Reflection in the Digital Era: based on Big Data Problems" deals with two problems of practical philosophy in the digital age, namely the moral status and privacy problem caused by data collection and reservation, and the ethical judgment problem caused by the calculation and application of big data. Prof. Yang points out that both problems arise from the development of digital technology and have serious impacts upon today's governance frameworks. He argue that the solution to both problem does not lie in another new technology, but in a philosophical idea, namely the ideal of good life. Prof. Wei-Ding Tsai's article "On Epistemological Conditions of Big Data" focuses on the epistemological status of big data and tries to analyze its outer and inner conditions. By way of comparing the difference between big data and small data, Prof. Tsai highlights an essential condition of big data that the quantitative change has produced a qualitative change, and indicates that the epistemological aspect of big data lies especially in the qualitative difference that big data concerns more correlation than causality. He concludes that the epistemological attitude of big data is actually pragmatic.

數位年代中對物的重新追問

洪世謙*

摘　要

隨著網路的出現，我們對於物質的想像便更加的貧瘠。例如網路交友、網路銀行和網路購物，這些看起來帶來便利性的網路活動，實際上卻是神秘化了許多實際的社會關係和生產關係。我們不再重視一些很直接的社會網絡，舉個十分常見的例子，我們對於一地有什麼樣的歷史建物或歷史事件不再感興趣，而是習慣於透過網路了解一地的文化與歷史，使得在查詢資料時，往往看見大量重複而了無新意的編年敘事方式，並且更多是過時的資訊（尤其是關於美食）。又例如網路買菜，傳統市場以往一直扮演的社區間資訊傳遞的空間，人們聚集在此，並不單純只是為了購得生活所需，還在此形成了社會關係，例如盤商之間的關係、人與土地的關係、人與社群鄰里的關係，然而，網路買菜的出現，看似帶來了便利，其實是神秘化了這些關係，切斷了這一切由物質和空間所展開的社會關係。

於是，關於數位人文的思考，重要的不是定義網路世界是什麼？或者描繪網路中的現象，重點在於，網路重構了我們生活的世界，我們活在一個以網路而構成的世界當中，網路不單單只是物或作為技術的存在，而是這個物展開或延伸了世界，世界不若以往是只有物理世界，而是它透過了終端機，延伸了以往的物理世界，形成了一種物質與非物質共在、共構的新世界，並因此影響了生存空間和生活方式。網路空間作為一種物質與非物質共構、共在的新世界，實際上可視為是在既有的現實空間之外，亦可視為是未到來的預存，相對於「現在」來說，它並非不在，而是「隱在」，然而只有當未到、當下與已逝三者的整體關係才共同構成了人的生活世界與生存意義。面對這個新延伸出來的世界，我們活在這樣的世界當中，要怎麼回應、因應這個世界所帶來的所有改變，並且思考及看待以往世界中的物以及關係。

進一步地說，對於網路科技的探索，面對的不是一個客觀的對象物或者某種科技發明，而是一種對於人及世界的探討，也因此我們可以說，這是一套新

* 國立中山大學哲學所副教授，Email: hungschian@gmail.com。

的哲學人類學(L'anthropologie philosophique)的思考。這裡的人類學並非指從考古或體質所界定的人類歷史，它不是將人視為觀察對象的某種自然科學的研究，而是對人所處的社會，由於生活空間的擴延、生存處境的改變，而重新思考人與世界的關係，將人重置於現實及歷史脈絡中，成為具體活生生的人，並且在這種具體的生存關係中重新討論人的意義。

根據上述，對於網路的思考，不能僅將其視為單純的物，不能僅討論網路的物質性，而是必須將其同時視為物質與精神，換言之，即便作為物質，網路亦同時帶來了實際上社會結構和精神意識的轉變。而此，又影響了我們對於生活和文化的想像與理解。是此，對於網路，不能僅將其視為單純的技術或物，而必須將它關聯到尤其所展開的新的世界，並因此重新討論和分析這個新世界所建構而成的政經結構、社會關係以及人與人之間的倫理關係等。

本論文的工作，便試圖從哲學的角度，尤其是以馬克思及其相關理論，重新說明物的意義。透過對於物的重探，賦予物質更多意義，而不是將物質僅視為簡單的原料或材料。眾所周知，馬克思的歷史唯物主義對物質與意識、物質與社會關係的論述，影響深遠，不論是後來的阿多諾或者當代對地理空間與文化有重大影響的 H.Lefebvre 和 D.Harvey。馬克思的名言：「物質生活的生產模式制約著整個社會生活、政治生活和精神生活的過程。不是人們的意識決定人們的存在，相反，是人們的社會存在決定人們的意識。」。換言之，物質不僅構成了我們的生活基礎、社會關係，也同樣的構成了我們的意識、文化、記憶、認同與歸屬。因此，若我們承認，我們正活在網路世界的年代，那麼網路就不單單只是物的存在，而是重新的構成了我們的社會關係，甚至影響了我們的人際關係（例如鄉民、網路霸凌，或者以網路通訊軟體作為主要交友與溝通，又或者因臉書發表個人言論，由於不同立場所造成的刪友潮）。

關鍵字：網路空間、物質、數位人文、馬克思、日常生活

The Question Concerning the Thing in the Digital Epoch

Shih-chian Hung*

Abstract

The thinking over digital humanities relies neither on the focus of defining what the cyborg world is nor on the description of the cyborg phenomenon. Instead, the internet reconstructs our living world; we are living in a world constituted by the internet and the digital. The internet is not merely the thing or the technological existence, but also the world extended and expanded by the thing. The world is not tantamount to the mere physical world of the past, but, through the computer terminal, it extends the physical world of the past, forming the coexistence between the material and the immaterial that exerts impact upon the living space and manners of living.

Hence, the thinking of the internet can no longer regard it as the pure thing or discusses the materiality of the internet. Rather, it should regard it as the matter and the spirit. The matter cannot only constitute our living foundation and social relations, but also constructs our own conscious, culture, memory, identity, and sense of belonging.

In the thinking of the digital epoch, we are living in the world newly extended, but how should we respond to and cope with the changes of the world, and how to think and approach the metamorphoses between the thing of the world of the past and its relations—these questions mentioned above will be examined from the structuralist perspective and the deconstructive perspective.

Keywords: Digital Epoch, thing, internet, coexistence, immaterial

* Associate professor, Institute of Philosophy, National Sun Yat-sen University. Email: hungschian@gmail.com.

隨著網路的出現，我們對於物質的想像便更加的貧瘠。例如網路交友、網路銀行和網路購物，這些看起來帶來便利性的網路活動，實際上卻是神秘化了許多實際的社會關係和生產關係。我們不再重視一些很直接的社會網絡，舉個十分常見的例子，我們對於一地有什麼樣的歷史建物或歷史事件不再感興趣，而是習慣於透過網路了解一地的文化與歷史，使得在查詢資料時，往往看見大量重複而了無新意的編年敘事方式，並且更多是過時的資訊（尤其是關於美食）。又例如網路買菜，傳統市場以往一直扮演的社區間資訊傳遞的空間，人們聚集在此，並不單純只是為了購得生活所需，還在此形成了社會關係，例如盤商之間的關係、人與土地的關係、人與社群鄰里的關係，然而，網路買菜的出現，看似帶來了便利，其實是神秘化了這些關係，切斷了這一切由物質和空間所展開的社會關係。

於是，關於數位人文的思考，重要的不是定義網路世界是什麼？或者描繪網路中的現象，重點在於，網路重構了我們生活的世界，我們活在一個以網路所構成、展開的世界當中，網路不單單只是物或作為技術的存在，而是這個物展開或延伸了世界，世界不若以往是只有物理世界，而是它透過了終端機，延伸了以往的物理世界，形成了一種物質與非物質共在、共構的新世界，並因此影響了生存空間和生活方式。換言之，我們活在一個較以往更加無邊界的世界，這個世界同時交錯了、涵蘊著實體的物理世界以及終端機中介後的網路世界，兩個世界同構了我們目前的生活世界。網路空間作為物質與非物質共構、共在的新世界，實際上可視為是在既有的現實空間之外，亦可視為是未到來的預存，相對於「現在」來說，它並非不在，而是「隱在」，然而只有當未到、當下與已逝三者的整體關係才共同構成了人的生活世界與生存意義。面對這個新延伸出來的世界，我們活在這樣的世界當中，要怎麼回應、因應這個世界所帶來的所有改變，並且思考及看待以往世界中的物以及關係。

根據上述，對於網路的思考，不能僅將其視為單純的物，不能僅討論網路的物質性，而是必須將其同時視為物質與精神，換言之，即便作為物質，網路亦同時帶來了實際上社會結構和精神意識的轉變。而此，又影響了我們對於生活和文化的想像與理解。是此，對於網路，不能僅將其視為單純的技術或物，而必須將它關聯到由其所展開的新的世界，並因此重新討論這個新世界所建構而成的政經結構、社會關係以及人與人之間的倫理關係等。

所以重點是思與物。但，我們究竟要思甚樣的物呢？物究竟要怎麼思呢？以及我們是否只能單獨的思考物？還是物，無法獨立思，只能是一種整體關係？

如同普羅米修斯的火，網路科技就像人類面對火一樣，是面對一項新科技，這個新科技，改變了人類的生活，也改變了人類的世界觀。與以往的新科技有很大的不同

在於，工業革命時的機器或者資訊革命時的（電視）媒體，這些新工具都與我們在同一個空間維度當中，亦即我們泛稱的現實生活當中。但網路，其存在於不同維度的空間中，我們稱之為虛擬空間，換句話說，網路空間不是以物理性的實體存在，我們的身體也並不真實存在於那個空間，但它卻真實存在，並且改變我們的生活和世界觀。因此，網路科技對我們造成了兩個具體的影響，一是改變了現實世界的邊界，其次是改變了真實與虛擬的界線。

傳統哲學中，人或事物皆關聯於基質（subjectum/ substratum），即 ousia，它是各種變化和偶性之下繼續持存而不變之物，變化只作為偶性而顯現於現象之中。笛卡兒的哲學以降，這種不可再倒退與懷疑的基質，轉變為我思主體，於是與主體相關的不再是不可動搖的基礎，而是事物的可見性，一種可計算的對象。人因此成了世界的主人，透過可見、可計算而將世界具體化為可控制的對象物，人在此情況中也成了被觀看的對象物。從海德格的話來說，這是一種對存有（Sein）的固置（gestellt），人以自己作為保證的基礎，實際上是人失去了其自身的基礎，人的可思與所思之物，都在人自己的範圍之內。是此，海德格認為我們應該要去思考那些未曾思之物，是這些未曾思才讓我們開始思考。於是他說：「思是通過存有(par l'Être)而為存有(pour l'Être)的任務」，正是思才讓存有透過人而說出了真理。（海德格，孫周興選編，1996，頁 359）

不同於笛卡兒的我思主體，海德格提出了「在世存有」以及「事件」（Ereignis），他說：「這個事件使人和存有在它們的本質的共在中彼此共屬」，透過這種回溯讓人重新的連結基質，亦即在人的活動之中有一種未被展開的潛在性，而此潛在性正是我們活動的根基，它是一切的基礎並藉由存在者的活動而顯現，於是人的活動（包括真理活動和藝術活動）都作為揭蔽，亦即將那些從不消失之物使其澄明（Lichtung），因此他說作為「此在」，人的本質就是存有的澄明（海德格，孫周興選編，1996，頁 370）。藉此，海德格將人的活動賦予存有學（ontologie）的意義。海德格在《存有與時間》訴諸一種生存論的基礎（fondation existentielle），也因此基礎存有學（ontologie fondamentale）成為建構一般存有學的先決任務。他指出對存有（Sein）的領悟是一種特殊性，一種優先性，亦是一種承擔（charge），存有是存在者的共同存在性（étantité）（Heidegger,1967: 12）。存有不是一般的存在者（étant / Seiendes），它不僅追問自身，也追問一般的存在者。對存有的遺忘與追憶，讓海德格將「在世存有」的意義表現為將存有帶到眼前或者使存有澄明，並使得存有作為由人的關係所展開的意義，且因此作為存在的意義。他認為真理的本質在於自由，而自由的可能性就是讓存在者存在，這意味著讓所有存在者都處於敞開的無蔽狀態，這同時

意味著人置身於存有狀態，是此，人才具有他得以生存的本質根據（海德格，孫周興選編，1996，頁 224）。換言之，存有作為人類活動所鋪展的整體關係與意義，這種將存在者置於存有的整體之下，也就將存有視為根基的本質，而「此在」是能夠向存有（Sein）敞開的唯一方式，海德格因此認為「此在」就是「在世存有」，他強調主體不是以認知的方式接觸世界，而是以體驗的方式，也就是一種將自身拋擲於世界而親臨世界，這是一種將己身投身於世界中的「在」。我們「在」世界中活動著，並在這些活動、關係中把握世界，因此我們無法以獨立於世界的意識活動來認識世界。

就海德格的觀點來看，世界是「此在」（存在者）整體的存在方式，它先行並與此在相關，規定著每一個具體的方式並使之可能，因此他說「存有的歷史承擔並規定著人類的任何條件與情勢（situation）」（海德格，孫周興選編，1996，頁 359）。世界必定與此在相關，世界是此在為此之故的基礎。「此在」作為「在世存有」，總已經是在世界的關聯之中的存在。世界作為世界化是給予性的，因為它給出意義，這意味著它敞開意義，而且世界和真理的關聯性就蘊含於其中。簡言之，對海德格而言，存有之經驗就是湧現（physis）、在場（Anwesenheit）、解蔽(aletheia)和發生(Ereignis)。只是這種經驗只能藉由人的活動，在關係中將其敞開、顯現。亦即存有需要人；反之，人只有處於存有的敞開中才成為人，二者彼此委身，相互依託（einander übereignet）。

其次，從解蔽與無蔽的觀點出發，海德格認為，現代技術就是解蔽，然而這個解蔽卻又被重新擺置以作為我們認識事物的方式。因此他說：「我們追問技術，是為了揭示我們與技術之本質的關係。現代技術之本質顯示於我們稱之為集置的東西之中。可是，僅僅指明這一點還絕不是對技術之問題的回答。」（海德格，孫周興譯，2011： 23）他認為以集置作為技術的本質並用以解蔽是危險的，因為這樣的去蔽並非從生產的意義，亦即不是從將尚未顯現之物展現出來（海德格，孫周興譯，2011：9），而僅僅只是技術作為計算而重新支配自然。笛卡爾哲學以降，人以主體的方式將世界化約為認識論對象，將現實簡化為由主體的意向性所能把握的範圍，甚至簡化為感官知覺，因此現實必須是主體可認知，可把握的。海德格認為這種從主體出發，加以現代技術所構成的世界圖像（Weltbild），就是我們現在認為的現實世界。然而他認為現代科技是一種「集置」（Gestell），它同時具有擺置（stellen）和聚集的特質，從而將事物以具有框架的方式呈現。於是這種由主體性和科技所構劃的世界圖像，看似是揭露了世界，卻反而使世界轉向、退隱。換言之，這種由建構所形成的世界圖像，讓世界成為了可思、可控制、可計算的世界，從而限制了我們認識世界的方式（Heidegger,1977:126-127；13-14）。人越技術化地理解自身與世界，越將自身提昇為

世界的主宰，也就使得人陷入了對技術的依賴，而忘了可從不同於製造的方式思考人的問題。

因此，海德格關於技術的沉思必須從此在的存有學，亦即它是一種在時間當中的活動關係而展開的，也因此它不同於笛卡爾以確定的計算為基礎的理性原則，而是它總是在尚未及曾是之間延綿（Er-streckung），這也同時讓海德格對科技的沉思帶往了對「在場」的批判探討，任何的在場其實都是已是（déjà-là），因此在場不可避免的具有並以時間性所構成，也因此不可能從當下或理性來理解，因為在可思之前，時間都已經在「在場」之中了。若將網路科技是為一種新技術時，我們便不該將網路簡單地視為是一項科技，其更接近於海德格稱之為「工具」（outil）。海德格認為工具是整體結構，工具顯現了關係及整體性，只有在這整體性關係之中，工具才因而實用（Heidegger,1967:68）。也因此，工具與人之間不可分，甚至，是因為與人的親近性，以及工具作為上手物 (*Zuhandenheit*)，也因此工具與人形成了生存關係。從這個角度看來，我們可以說網路科技就是另一種共在的它物，人同時跨越在不同維度空間的關係之中，人在網路空間中繼續一切活動，並與他人、他物繼續產生關係。因此，網路空間是與我們息息相關的共在（Mitsein），我與它之間共同活出了、展開了我們的「現實」，亦即它影響了、揭示了一種不同以往的生存空間和生存樣態。於是，網路空間，雖是虛擬，卻弔詭地擴延、增補了現實，二者共同的組成了我們的生活世界。從這個角度看來，面對網路科技，它作為與人共在的整體結構，應視為是人的新「上手物」，它幫助人類延伸了空間。當人的配備升級，人所涉入的世界範圍改變時，我們就更該重新省視人與己身生存的關係，人與世界的關係。我們更需面對網路所展開的新的生存空間與方式。也因此，我們必須重新探問世界是什麼？而當我們啟動這個問題時，伴隨而來的便是構成世界的是什麼？現實是什麼？以及生存是什麼？而這一切的問題，又都發生在時間性（temporalité）問題之中。

從上述觀點，我們可以說，對於網路科技的探索，面對的不是一個客觀的對象物或者某種科技發明，而是一種對於人及世界的探討，也因此我們可以說，這是一套新的哲學人類學（L’anthropologie philosophique）的思考。這裡的人類學並非指從考古或體質所界定的人類歷史，它不是將人視為觀察對象的某種自然科學的研究，而是對人的本質從存有論上重新討論，試圖將人從傳統形上學或人本主義的傳統解放，將人重置於現實及歷史脈絡中，對人所處的社會，由於生活空間的擴延、生存處境的改變，成為具體活生生的人，並且在這種具體的生存關係中重新討論人的意義。正如海德格所定義:「人類學並不是指某種關於人的自然科學的研究…它標誌著那種對人的哲學解釋，這種哲學解釋從人出發並且以人為指歸，來說明和評估存在者的整體。

(Heidegger,1977:133)」換言之，人不是某種具有本質性的存在，也不是某種被先行定義和給予的存在，而是在活動中展開其存在的意義與價值。也因此，當我們所面對網路這個新工具，新的人與世界的意義，就在於我們與這個工具之間的新關係中重新的開展出來。

綜合上述，我們進入了一種海德格指稱的人類疑難/困惑（L'énigme de l'humain），我們遇到的難題是，當網路科技成為一種新工具，新的載體，並且與人共同的形成了新的人的概念和新的世界時，網路所帶給我們的究竟是一種去歷史的年代還是一個「元」(L'archie)[1]世界？

一、世界存在於「不在」之中

首先，網路空間使我們重新思考「真實」的本體論，亦即所謂「真實世界」或「生活世界」，它已經無法像以往停留在具象的物理空間，或者，具象的物理空間已無法作為「現實」的唯一標準。因為網路空間，現在更大一部份也成為我們的生活世界，網路世界成為我們生存空間的延伸，即便它不具物理性，必須透過介面（interface）才能存在的空間，但它仍然構成了我們的生活世界，以及我們真實生活的一部分。換言之，網路空間並不完全外於我們的現實世界，而是滲透、交錯的構築了我們的現實生活，也因此，我們無法再以過去的基礎思考人的生存世界。

在傳統哲學中，所謂的現實通常指向我們所生存的世界，並將此現實世界視為真實，而現實世界指的就是人類的感知與活動發生的地方，亦是人類已知或被給予的狀態。然而，從海德格的角度看來，世界不在我們的生存之前，而在我們的生活關係之中展開，然而這樣的前後，不是時間序上的前後，而是一種內在性如何顯於外的問題，亦即這是一種藉由活動而方能展演的本體論問題。海德格指出，技術性越強，亦即時間測量的效果越精確、越詳細，就越沒有機會對時間的本真因素進行思考(Heidegger, 1967: 11)。因此他認為，在人們認為能夠測量、或談論時間之前，時間必

[1]這是德希達哲學中，非常難以漢語化的概念。某個意義上，「元」可以理解為海德格意義下的存有或者真理，亦可視為某種不以顯現方式而存在的崇高價值，德希達將這樣的概念形容為「存有的聲音」(voix de l'être)，只是這樣的聲音是沉默的、失語的(muette)(Derrida,1967 :36)，也因而它往往容易被某些可見的在場物(法律、制度、理論、知識典範等)替代。「元」不在傳統形上學二元對立的系統中，也不在一般的經驗中，亦非一般稱為意義的起源之物(Derrida,1967 :20)；同時，「元」遠非知識類型下的真理，亦非科學的研究對象，更無法以任何制度或概念將其化約(Derrida,1967 :83)。「元」描述一種無法預見的不可能價值（la valeur d'im-possible imprévisible），且這種價值總是相關於不可計算（incalculable）和經驗之外的他異性。換言之，它雖是個看似外在的新物，但這個外在之物其實是內在性，也就是這些交錯的內在關係所構成的複雜整體，顯現為一個全新的外在樣貌。換言之，雖然它外顯上是個新，然更徹底的說，它是內在元素的新的整體關係，只是在不同的年代當中，以各種不同的樣貌重新出現。

定有讓時間之為時間的根本性質，即「時間性」(Zeitlichkeit)。已是（逝）、將來與當下並非時間性的嬗遞或累積，而是作為綻出，它們於時間性中皆是同源。換言之，因為時間性能夠綻出已是、將來與當下等現象，因而能夠讓人們能夠將時間視作整體關係而加以闡釋 (Heidegger, 1967:329)。對 Heidegger 來說，關於時間性的探討，一方面能讓人們更本真地掌握時間，另一方面，這也是探討「存有」(Sein)的無可迴避的前提。

藉由海德格的概念，我們說明了在場與世界的關係，不論是在場還是世界，都應視為整體而不可分割的，每一個顯現的當下，其實都同時包含著「不顯」，他說：「當下作為原始時間性的樣式，始終包括在將來與曾在中」，且說：「所謂不在場並非一無所有，而不如說，不在場乃是那種恰恰首先躍居為有的在場現象，也就是隱蔽而豐富的曾在者和如此這般聚集起來的本質現身者的在場狀態….。這種不在本身就是一種尚未，也就是它不可窮盡的本質的隱蔽到達的尚未。（海德格，孫周興譯，2011：193-194）」。也因此「現實世界」實際上同時包含著「外-現實世界」，當下其實只是藉由不同的綻出而規定自身。

我們同樣可以從德希達以「在場」（圖像、聲音、符號…）所示與在場所是的差別，說明以往哲學將在場、能指視為真實，將缺席、所指視為虛擬的二分，其所出現的問題。換言之，所顯示的是否就是圖像所要表達的？長久以來，我們都將圖像視為一種確認，透過圖像紀錄一些已是/已逝的事物，並以詮釋的方式，說明它的背景和歷史脈落，並因此形成記憶，以及一系統性的歷史學科論述。

然而如同聲音一樣，當我們聽到聲音（voix）時，是聽到聲響（son）還是聽到音素（phonè）？還是實際上是聽到音響內部所帶有的意義，也就是說，當我們聽到一個聲音時，我們並不是因為某種音響或音位而理解事物，而是因為知道這個聲響背後的意義，甚至是這個聲響的整體脈落，於是乎我們知道目前發出這個聲音代表什麼意義。換言之，當我們聽懂一個聲響所表示的意義時，並不是聲響讓我們理解是聲音的意含，而是我們總已先在那個聲響所表現出來的意義賣落下，於是，並非聲音這個在場說明了意義，反而是那個鋪開意義的「沉默」說出了意義。亦即，聲音作為一種在場，其實只是說明了那個不在場，音響什麼都不代表，真正有意義的是那個不在場的沉默。換言之，聲音在此作為中介，一旦它是導向意義的意象活動，它不必然需要發出聲響，而可以是保持沉默的聲音，也正是這種沉默的聲音使意識成為在場。於是，德希達稱此沉默的聲音是現象學的聲音，它無關乎具體的、物理的聲響，而可以是意識自身。在此意義上，德希達可以說對聲音進行了一次現象學的還原，除了它並非只是具體的聲響之外，聲音並非外部的表述或對外部的「指示」（indication），聲音甚

至可以是先於一切固定意義指涉的音素，一種以觀念（idéal）方式的存在。將這樣的概念放到圖像，一張照片或影像之所以具有意義，並不在於圖像所示，這個圖像所示，往往並非所是，例如在圖像當中，我們永遠無法知道這個圖像的歷史或者內在心靈。換言之，和所有在場與缺席一樣，物有其差異性或說雙重性，並且是一種永不可能縫合的差異，在場（圖像、聲音、符號…）「所示」的同時，與其「所是」必然帶有差異性，他的示不盡然是它的是，它的「是」往往需要透過一種扭曲、轉譯或者組合。如同文化協會之中的辯士，透過他轉譯、扭曲圖像意義，而賦予了這種集會和影像本身政治意識（反帝、反殖民、反壓迫…）並因此可能出現政治行動和政治的歷史記憶。

也因此得知，物本身會有雙重性或者物本身帶有間性，也就是一直保持著可見與不可見、在場與缺席間的差異，正是這個差異維持著物的間性，這個空間可以讓其他事物，不論是已是或者未到來，都以此可以放到這個因為差異而展開的空間之中。

德希達認為「在場」僅是一種僭越(usurpation)，「在場」的僭越意謂著它將自己當成了事物的起源，而我們的歷史正是在這種「在場」的僭越中定義和思考本質(nature)及起源(origine) (Derrida, 1972a: 59)。而「不在場」其實只是遲到、扣留(rétention)，他認為若沒有這種扣留，我們便無法在結構中指出差異，是這種扣留成為一種痕跡(trace)[2]，顯現了差異，並且因為這樣的差異讓一切變化的自由成為可能(Derrida, 1972a: 68)。也就是說，在場與不在場是一種顯(可見)與不顯(不可見)間的差異，而正是這種差異，是一切的起源(Derrida, 1967a: 202-203)。

準此，德希達指出，我們所謂的世界是歧義的，一種是作為已在場的、構成性的當前「現實」，這樣的世界確定為純粹自身而封閉的現實。其作為世界觀念是預先給予的先驗被動性，並將無限可能性顯現為可能性，亦即將複雜性化約為簡單經驗，化為有系統而可用的形式。科學（尤其生物學）與人類學，便是以此預設著作為對象的動物世界或人的世界（Derrida,1987:76）。另一種世界則是作為可能經驗的無限界域（horizon infini），是一切判斷基礎的無限整體性(totalité infinie)，世界不作為被預定的確定性，而是複雜、無定限的（indéfini）。在此意義下，世界是無限的現實性，它由異於其自身的他物所孕育與支撐，界域和無限可能性不再有起源(Derrida,1990:187,191-192)。

[2]德希達的痕跡（trace）概念，它是既見證、對照而又否證，也因此，它往往指引了某一個面向，也同時指認了另一種可能的面向。而這種既承認又排除，也就讓事物之間產生意義上的差異，亦即產生了延異的效果，讓意義總是指向了他者而無法回歸到開端，而讓事物保持它的開放性，既懸置現實又展開另一種可能。對德希達來說，痕跡並非不在場亦非在場，痕跡是在場自我拆解、自我位移和自我參照（renvoi）的顯像，它並非一個確切的空間，而是刪除它所屬的結構。（Derrida, 1972: 25）

也因此「不在」(已逝或未到來）反而才是事物的本質（基礎），這也讓事物永遠存在一種不滿足和永遠在未來，也因此它依舊有無限的可完善性，正因為這個無限的可完善性，才讓事物成為了一種可變性。德希達認為，恰好是這個「不在」，逾越（transgression）了「現在」所劃定的範圍，使得「現在」的內部發生了變動，也因此這種越界在任何地方都不會作為既成事實而出現，也因此人既不會安於這種跨界，但也不會生活在他方（Derrida,1972b :21）。換言之，因為事物永遠具有不在場、不呈現的部分，這「不在」不斷的逾越事物的界線，反而使事物更顯現了完整性。

綜觀之，不論從海德格或德希達的觀點來看，世界一直是動態的，是在關係之中展開的，這也就讓「現實」留下了可能性。換言之，我們不該把現實就當成是眼前所是的樣態，而應該視為一種「能是」（Seinkönnen /pouvoir être）(Heidegger,1967 :201)，亦即現實亦包含著未到來的可能性（Möglichkeit）。在此觀點下，若我們將未到來的現實視為虛擬，那麼虛擬與現實是並存的，或者說，現實實際上包含了兩個面向，一個已在場的及一個尚未來臨的可能性和潛在性的在場。或者反過來說，虛擬是尚未實現化的現實，它並非模仿或再現而就是現實。現實不再僅是客觀的存在物，現實還包括了未到來的可能性，亦即傳統所界定的虛擬。

二、結語

根據前述，人的生存世界不是一個由科技所構劃的圖像，而是一整個與他者之間的關係活動所展開的世界，人不是精神與身體的二分，人的生存空間也不是虛擬與現實的二分，而是如德希達解構哲學所欲強調的不可化約並永遠留下未到來的可能性。而只有這種非主客二元的模糊性，才讓人和生存世界成為一種以親近性的共在所鋪展開的整體性，而非範疇或被給予物。我們可以將科技視為一種潛能（*potential*），這樣的潛能錯綜複雜的影響著我們的整體生活，而絕非單純從技術的觀點所能片面決定[3]。也就是技術可以帶來什麼改變，這件事不是目前所能構想到的，但重點是，我們要重探我們與技術之間的關係，使它不再作為眼前的技術，而是一個與我們生活環繞的相關物。因此，對於虛擬網路空間的省思，不該僅停留在對科技的理解，而應反思究竟虛擬網路空間是以何種方式重構或重新展示我們所生存的世界。尤其是當網路穿透了現實生活而成為現實並具有了真實性時，我們便更無法將虛擬與現實二分。

其次，數位年代中，彷彿「物」已消失，一切都可以化約為 0 與 1，並因此創造、拼貼、再製無數的聲音、符號和影像。然而，從在場與不在場的意義來看，這些聲

[3] Richard Sclove, Democracy and Technology, Guilford Press, 1995,p.7.

音、符號和影像，即便不作為具有物理空間的物，而僅是網路上數位化後的資料或產物，它同樣具有物的特質，也就是作為一種能指、在場，依舊作為一種所示。也因此我們可以說，數位年代下的數位產物，同樣具有所示與所是間的差距，同樣作為一種生產物，即便只是網路上生產文字或影像，它依舊可能被置入於符號意義的對應關係之中。是以，數位年代下對物的重新探問，不在於思考已在場之物，而在於，若物的資料或形式已改變，物不一定是具有物理性時，我們依舊可以看到符號、影像、聲音等具有物的雙重性，而我們於此年代的工作，便是繼續不斷在所示與所是之間的裂隙中，增補各種不同的意義，即便這樣的增補永遠不可能完成。

引用書目

海德格。1996。《海德格選集》（孫周興譯）。上海：三聯書店。

海德格。2011。《演講與論文集》（孫周興譯）。上海：三聯書店。

Derrida, Jacques. (1967). L’écriture et la différence. Paris: Seuil.

Derrida, Jacques. (1972a). Marges de la philosophie. Paris: Minuit.

Derrida, Jacques. (1972b). Positions. Paris: Minuit.

Derrida, Jacques. (1987). De l'esprit: Heidegger et la question. Paris: Minuit.

Derrida, Jacques. (1990). Le problème de la genèse dans la philosophie de Husserl. Paris: PUF.

Heidegger, Martin. (1967). Sein und Zeit, Tübingen.

Heidegger, Martin. (1977). The Age of the World Picture, in : The question concerning technology and Other Essays. New York: Harper & Row.

Heidegger, Martin. (2000). *Introduction to Metaphysics. Gregory Fried, Richard Polt(eds.). Yale University Press.*

Heidegger, Martin. (2010). *Qu 'appelle-t-on penser*? Aloys Becker, Gérard Granel(eds.). Paris: PUF.

數位人文的現象學還原：從擬象到檔案

高國魁*

摘　要

橫跨機械和電子複製時代，從存有學到解構學的思潮首先源出自然的退隱、接著轉入文化的延異，從而雖然已自存有的科技末日論朝向語言的技術救贖說進入了康復期，但是尚未探問「終極解決方案」執行下的世界狀態：全數位模控的社會系統是否將保存人性的實在蹤跡，或者餘留人文的虛構字跡？重建一脈現象學的系譜，本文主張，數位人文將有可能獲得從擬象到檔案的無條件還原。

本文將集中探討數位人文的技術本質和媒體現象。在這裡「數位人文」的所指將從知識論地圖上被拉出，轉而連接至現象學視域中考察。換言之，本文不再反思批判數位人文的學科結構，而要直觀理解數位人文的存在世界，及其歷史命運。本文假定，除非替換方法轉向真理，否則我們無法棄置技術媒體的人類學工具論，再生技術媒體的詩學政治觀。

總的來說，全文將分成四段論證。第一節「數位人文的知識環境」將要概觀數位人文研究的語言和制度環境，並且點明該知識學科化的結構限制，以及技術和媒體的功能解決。旋即擱下結構功能論的社會學批判，本文轉而探究技術和媒體的政治性本身。第二節「數位人文的技術本質」將從海德格的集置轉向觀論及布希亞的擬象流變說。第三節「數位人文的媒體現象」將從德希達的檔案病惡觀推導葛洛伊的檔案陰謀論。第四節「從擬像到檔案：藝術的救贖？」將比較布希亞和葛洛伊的交換理論，據此闡明當今數位媒體之於技術和藝術的曖昧關係。

就理論言，全文要試圖梳理一派獨特的現象學轉型，實質考察海德格存有學和德希達解構學已如何被布希亞和葛洛伊的現象學激進還原到空無如是的人文境地，並特別表達在後者對於前者有關技術集置和媒體檔案的原創回應上。以經驗說，文中也會適時援引今日世界的數位化現象闡明人文現象學，可能包括串媒介敘事和產消者網絡，以及現場攝影實錄和維基解密檔案等；同時，文末也將根據擬象和檔案的概念導引，逕行攝影和電視、書本和網路等技術媒體的類型比較。

關鍵字：技術、媒體、集置、擬象、檔案、靈光

* 國立政治大學社會學系助理教授，Email: pascalkk@nccu.edu.tw。

The Phenomenological Reductions of Digital Humanity：From Simulacrum to Archive

Kuo-kuei Kao*

Abstract

Spanning the ages of mechanical and digital reproduction, the intellectual trend has evolved from the ontological withdrawal of nature to the deconstructive différance of culture, and hence arrived at a period of convalescence in which the technological eschatology of Being could already aspire for the technical soteriology of languages. Nonetheless, it has yet to question the state of the world subject to an execution of "the final solution". The question is whether or not the social systems of totally digitalized cybernetics will preserve some real traces of oral humanity or leave a few fictional marks of print humanity? In the genealogical reconstruction of a phenomenological strand, this article proposes that digital humanities would be able to receive an unconditional reduction from simulacra to archives.

The essay will concentrate on uncovering the essence of technology and media phenomena. Herein the reference of "digital humanities" will be torn apart from the epistemological map to reconnect with the phenomenological horizon. In other words, the essay will no longer critically reflect on the disciplinary structure of digital humanities, but rather intuitively grasp the existing world of digital humanities and its historical fate. The essay assumes that unless methods are replaced with an orientation to truths, we cannot abandon anthropological instrumentalism while dealing with technical media and regenerate them with a poetic politics.

In general terms, the article will be divided into four sections. Section one entitled "Knowledge Environment of Digital Humanities" will give an overview of the current research environment of digital humanities by reference to language and institution. We will indicate the structural limitations that prevent this kind of research

* Assistant professor, Department of Sociology, National Chengchi University, Email: pascalkk@nccu.edu.tw.

knowledge from growing into a discipline, and the functional solutions offered by technique and media. But leaving aside a sociological critique of structural functionalism, the essay will turn to explore the political nature of technique and media in and of themselves. Section two entitled "Technical Essence of Digital Humanities" will talk about the turn of *Gestell* reflected by Martin Heidegger and the becoming of simulacra elaborated by Jean Baudrillard. Section three entitled "Media Phenomena of Digital Humanities" will debate on the evil sickness of archives envisioned by Jacques Derrida and the conspiracy theory of archives designed by Boris Groys. Section four entitled "From Simulacra to Archives: the Redemption of Art?" will compare the differences between two theories of exchange proposed by Baudrillard and Groys, and thereby explicate the ambiguous relationships of contemporary digital media with technique and art.

Theoretically, the essay seeks to establish the transformation of a unique phenomenological school. It essentially examines the ways in which Heidegger's ontology and Derrida's deconstruction undergo a radical reduction to a vacant state of humanity as such by the phenomenological attitudes of Baudrillard and Groys, which are particularly expressed in their original responses to technical *Gestell* and media archives. Empirically, the essay also appeals to digitalized phenomena in today's world as an illustration to the phenomenology of humanities. They might include transmedia storytelling and prosumer network, as well as live camera recording and wikileaks file release. Meanwhile, we can proceed to a typological comparison of technical media between photography and television or between book and internet according to the leading concepts of simulacra and archives.

Keywords: Technique, Media, *Gestell*, Simulacra, Archive, Aura

一、 數位人文的知識環境

假設在十九世紀中後期環繞著機械工程學（熱流力）和生理生物學（遺傳演化）的進步光環曾經誘人到促成實證社會科學知識的誕生，進而幫助了現代國家治理權力的壯大，那麼在二十世紀中後期延續著電子計算學（真空管）和基因生物學（人工試管）的創新能量似已深遠到引發人文學科跟著仿效社會科學的成功軌跡，開始讓人文學者也有意在正典文本的規範性評論和詮釋之餘，額外學習數據資料的經驗性描述和解釋。晚近十多年來，似乎有一個可泛稱作數位人文的學術社群正在成形。這個研究社群在文學、藝術和歷史，以及媒體、文化和政治的跨學科領域之間興起一股風潮，並且產出令人注目的研究成果。頓時之間，由近拉遠的閱讀方法竟能浮現另類的文學史圖像，而自小放大的網路分析亦可聚集顯著的傳播學訊息。與此同時，世界各地的大學更是陸續在文史科系所內或者院校級圖書館裡成立數位人文中心、建立電腦實驗室，最起碼也該提供電子化的授課教室。

並不若自然科學向前跟隨非人類物件和物質現象的「客觀發現」而同步轉移研究典範，人文學科始終返回人類語文和思維現象作為「原初經驗」，所以較容易忽略其研究對象也會隨著環境的遷移而出現本質的變化。如此看來，數位人文研究正可謂是在面對快速變異的當代世界之時應運而生的回答和反思。按此前理解，數位人文研究已不單純指涉人文學科怎樣功利性使用數位工具來強化知識宣稱的客觀效力，而是複雜隱含人文學科如何與數位環境進行象徵性交換，挑戰彼此固有的身份認同。只可惜數位人文研究的對象至今依舊不明，並且所在的環境仍然有待釐清。

文學和文化藝術雜誌《洛杉磯書評》（LARB）在 2015 年 3 月至 8 月間刊登了一項名稱為「人文學科中的數位性（The Digital in the Humanities）」系列專題。這項專題由 12 篇訪談稿所構成，內容為文學和媒體學者 Melissa Dinsman 針對 12 名數位人文研究的專家進行訪問。[1] 訪談者的初衷是想藉由教義問答形式辨明數位人文學科的共同旨趣，包括基本問題、方法和目標等，然而她旋即發現得到的回答大多圍繞在數位人文研究興起的制度和技術史，甚至還一度遭遇「數位人文是為無物」的直接否認（Moretti, 2015）。結果她只好把專題的定位調整為探問人文學科中的數位特性，終究無法確認數位人文的學科本性，更別說是回頭思索自然科學的文字屬性。

[1] 本節以下是從學門、性別和族裔分佈中擇取六位各具代表性學者的意見統整，具體指涉 Franco Moretti（Mar 2, 2015）、Alexander Galloway（Mar 27, 2015）、Laura Mandell（Apr 24, 2015）、Richard Jean So（Apr 28, 2015）和 Richard Grusin（Aug 18, 2015）共六篇訪談。系列訪談的全部內容可參見入口網頁 https://lareviewofbooks.org/contributor/melissa-dinsman。2016.09.30 瀏覽。

即便如此，我們還是得到了一些稱不上定義的基本認識。在訪談當中，學者們的不同說法可有下述概況。數位人文原來專指運用人文學科計算（humanities computing）方法產生的計算機批判（computational criticism），實質言即文體計量分析（stylometric analysis）和數位編輯工作等內容（Mandell, 2015; Moretti, 2015）。發展到後來，人文學科也能夠採納自然科學理性的思維，以數位表徵的經驗翻新解釋文學和藝術史觀，接著還可以結合制度和科技研究等方法，實地製作數位模式的媒體和藝術作品。持平地說，數位人文的夢想是藉由計算學分析進一步推展人文學科專長的詮釋學批判（Mandell, 2015）。在這理想下，人文學者被建議要提高數位識字率，但不會因而喪失批判理性和道德感性，反而能藉由該學習認識社會和文化原本就是個技術和象徵的共生系統，進一步認可電腦理性的運用確實有助學者觀察人類經驗，揭示新知（Galloway, 2015）。然而值得強調，數位人文學者絕非是在求諸電腦科學協助人文學科變得「有用途」，好似人文學科已與當今現實失去「關聯性」因而需要技術神的拯救。與此對反，數位人文學者是在試圖鼓勵人文學科善用電腦科學以提升對人類事物的巨觀理解，比方說在小說文學中和社群媒體上透過大量文本的資料彙整進而分類原型結構、梳理敘事情節，或是辨認訊息趨勢（Mandell, 2015）。照這正面形象，數位人文研究通常會往返在數理和語文的異質符號系統之間進行語言轉譯，簡單點說就是藉由比較工作產生知識。該類工作的基礎研究首先要重建數位特性在人類歷史長河當中發展的經驗敘述，其次需思辨數字和文字在文化和社會中是如何作用的理論闡釋，再次可站在數位人文的新視野上面對社會和政治的當代構成運用批判分析（Galloway, 2015）。檢視工作進程，數位人文考察全面揚棄了人文主義形而上學，因為人類的語言溝通已被設想成有如馬賽克鑲嵌細工的數位運算性質，而且社會的語言傳播也被認為具備有符號學的意義流變和系統論的機器操作等科學成分。

基於上述理解，我們可初步推論出數位人文研究的兩大知識環境，意即語言和制度。先從語言看，它顯然橫跨了自然人類和人造機器的兩種語言（Galloway, 2015; So, 2015）。它們的初級經驗模式是以語文詮釋和數理運算為質料，而次級展現型態卻是以印刷書本和電子螢幕為媒介。在平行的兩階段發展下，原先分屬於意義操演（聽說）和邏輯操作（算數）的力量後來反倒需要仰賴想像意識（讀寫象徵）和視覺感知（觀察幾何）的權能，結果則是產生具體和抽象經驗的雙重顛倒。這也許是為何現代人普遍認為閱讀書本相較於觀賞螢幕而言是更加費解的知識經驗，可若是把兩大媒介裡的質料回復，則前者反而成為較易理解的生活體驗。正因如此，數位人文研究傾向借助統計科學和視覺藝術的工具整合（Mandell, 2015）。換言之，只有當不可見的結構型態被轉碼以後重新顯示在可見的影像圖表上，我們方能展開剖面（cross-sectional）或鑲版（paneling）資料的實徵經驗分析。如同社會科學家處理量化資料，

數位人文學者也亟需學習圖檔管理員的資料收集技能，以及程式編撰員的資料解析能力，而兩者已相應地涉及藝術和科學的創新術（Moretti, 2015）。

再就制度說，數位人文學者可被懷疑與某種黑暗勢力共謀。直言之，這類新興研究似乎符合教育理性企業化（以評比階層分化大學和學者）和科學工作機器化（以設備去技能化勞力和智力）的新自由資本主義精神，從而在當前體制裡正處於優勢位置（Grusin, 2015; So, 2015）。但是，這種批判觀點是以結構認識為前提，而在這裡，數位人文的研究現況並無法玩弄生物政治的權力遊戲，因為學術社群裡的學者「個人」尚未從交換的「生產者」突變成競爭的「企業體」（Foucault, 2008: 118, 145）。就人事看，仍處在學術邊陲的數位人文研究還未沾染組織間層級化和組織內個體化的市場趨勢，反倒可說常見不只跨領域的團隊合作，還有獎助大資料庫建立而非專著書籍出版的反英雄式學術對話（Moretti, 2015, Mandell, 2015）。從事物說，科系所和圖書館爭相建立數位人文中心的確隱含資源積累的競爭態勢。然而，在這裡所謂的實驗室又可以是翻轉教室的有形譬喻，但現實中卻無形建立在學生個人筆電的串連網路裡（Galloway, 2015）。在新自由主義的膚淺印象下，數位人文的教育理念實際構築在知識份子消失的前提中，因此至終要承認公眾參與、俗民方言的數位消費文化，當中可見的是創作和出版、部落格和駭客，甚至惡搞、模因和同人誌等社群媒體現象（Galloway, 2015; Grusin, 2015）。

綜合而論，數位人文的語言和制度環境中隱含有人文知識科學化和工業化的雙重風險。這些風險可謂是數位人文學科自主化的限制，所以理應批判之。科學化的問題牽涉全自動化幻想，意即在數位人文學習中誇大讚揚資料庫存和數據採礦。工業化的問題關乎新自由化現實，也即在數位人文教育裡過份依賴國家獎勵和企業贊助（Grusin, 2015; So, 2015）。但跨越了批判視界，又可知數位人文學的最大特色絕不僅止於系統價值的符碼破解，而是還要在廢棄舊事物秩序以後建設新事物工程，意即使事物發揮積極作用，製作生活世界（Mandell, 2015）。吸收了量化世界觀，數位人文學不再強調必然否定依舊封閉的社會，而開始追求偶然偏離已然開放的社會，因為既身處在域外也棄置了例外，剩餘下局部相對的可比較性。值此無限相對論，數位人文學經常給人承諾的未來性，也總處在潛能的嬰兒期（Grusin, 2015）。然而在此，樂觀的期望感卻不有靠語言和制度的環境改善，而是仰賴技術和媒體的工具發明，好似數位人文學的成功機會維繫在新工具的強化效能，終究才可衝破環境限制。雖然數位人文學從不缺乏技術和媒體的理論研討，但是相比語言和制度的分析而言較難以達到批判要求，換言之更容易受到功能需求的誘惑，特別是當學者回到新奇的數位實習場域。然而，本文不再循線往下批判技術媒體端的工具解方是導致語言制度面的結構問

題被去政治化的元兇。這類目的論的社會學批判既使正確，但也無力。相反地，本文要追本溯源扣問技術和媒體自身的獨特政治性。

本文將集中探討數位人文的技術本質和媒體現象。在這裡「數位人文」的所指將從知識論地圖中被拉出，轉而放到現象學系譜上考察。換言之，本文不再反思批判數位人文的學科結構，而要直觀理解數位人文的存在世界，及其歷史命運。本文假定，除非替換方法轉向真理，否則我們無法棄置技術媒體的人類學工具論，再生技術媒體的詩學政治觀（poetic politics）。總的來說，以下主文將分成三段論證。第二節「數位人文的技術本質」將從海德格（Martin Heidegger）的集置轉向觀論及布希亞（Jean Baudrillard）的擬象流變說。第三節「數位人文的媒體現象」將從德希達（Jacques Derrida）的檔案惡病觀推導葛洛伊（Boris Groys）的檔案陰謀論。第四節「從擬像到檔案：藝術的救贖？」將比較布希亞和葛洛伊的交換理論，並據此闡明當今數位媒體之於技術和藝術的曖昧關係。就理論言，全文要試圖梳理一脈現象學的系譜，實質考察海德格存有學和德希達解構學已如何被布希亞和葛洛伊的現象學激進還原到空無如是的人文境地，並特別表達在針對技術集置和媒體檔案的原創回應上。以經驗說，全文會適時援引今日世界的數位化現象闡明人文現象學，並同時根據擬象和檔案的概念導引，逕行攝影和電視、書本和網路等技術媒體的類型比較。

二、 數位人文的技術本質

(一) 集置的轉向

由於全文以下將要運用現象學還原數位人文的世界命運，所以這裡似有必要從哲學史角度預先提出現象學概說，然後再帶出本小節的主題談論晚期海德格的技術存有學。胡塞爾（Edmund Husserl）的現象學旨在突破十九世紀末的實證主義和心理主義，本意是當面臨人事物現象時能夠獲得必然確證為絕對真實的理解（apodictic understanding），因而採取觀看（see, *eideo*）姿態描述現象的本質或理想結構。現象學可說是返回超驗意義（*sinn*）的意識哲學：它相信吾人若能從自然科學化的抽象現實脫離出來，則可以純粹直覺（eidetic intuition）面對現象明證（*evidenz*）。按這假定，現象學著重研究有關某事物的意識，亦即在主體的認識行為（*noesis* 能思）和認識對象（*noema* 所思）的內在連結中，事物立即被給出（given）或出現（appear）在意識經驗內的結構。作為認識論，現象學力圖超越自然—科學態度的限制，並重邏輯和語言的基礎，探問客觀認識的可能條件。有別於自然和科學態度，現象學態度強調經過嚴格的方法程序剝除思考的慣習和學習，接著方能返回事物本身發現新穎敞開的意識經驗或視域。超越認識論，胡塞爾的現象學後來又啟發了海德格的存有學，探問理想客體或說世界存有的意義和真理，以及心智思考或說心靈生活的本質和存在。

在胡塞爾和海德格之間有個存在現象學的橋樑。存在現象學結合了胡塞爾的現象學方法和瓦爾（Jean Wahl）稱作存在哲學的思潮。存在哲學不只反對實證哲學方法，也要反對觀念哲學系統，包括科學理性和歷史精神。它可以跨越一神教和無神論的分野，溯及謝林（Friedrich W. J. Schelling）和齊克果（Søren Kierkegaard）的自然哲學和神學，甚至歌德（Johann W. von Goethe）藝文和科學哲學。存在現象學意在彰顯存在的事實性（facticity）和情緒性（emotionality），又旨在宣稱存在先於本質或者存在即是本質（Wahl, 1969[1959]: 7, 15, 29）。回頭說，胡塞爾的現象學通常被理解成描述現象學（意向性的還原）、超驗現象學（互為主體的自我）和發生現象學（生活世界的危機）的三階段分期，而在這演化框架內又極易被認定殘留有觀念論和實在論的兩難困境亟待解決（Zahavi, 2003: 8-9; Hodge, 2008: 70, 73, 78）。所以，我們經常得知現象學運動的發展是先從描述心理學或心理人類學演進到超驗現象學，再從超驗現象學分岔出存在現象學。斷裂的發生在這裡泛指從認識論和人類學到存有學和存在論的理路轉向，接著沿新路向後人開啟在世存有、為己存有、無限它者和身體知覺的問題意識，同時揭示日常生活的存在結構，像是死亡、衰老、病痛和誕生的事實，以及焦慮、關懷、好奇、噁心、愛戀、憎恨、漠然、勞累和逃避的情緒。

進而修補斷裂，呂格爾（Paul Ricoeur）有個中肯說法證成超驗和存在現象學的融貫性。其一，現象學描述不僅只是經驗性反映事物，而是要將事物的出現（the appearing of things）本身給問題化。其二，存在現象學不是附帶在超驗現象學下的分支，而是把相同的現象學方法應用到存在這個不同對象上，意即把意識經驗再次還原為存在經驗。其三，前期胡塞爾談意向性之時便已隱含既是表意化言語也是直覺性知覺的雙重特性，從而充分理解到存有在世界裡的意識已化為語言及身體的存在（Ricoeur, 1967: 202-204）。與文化馬克思主義有別，從超驗到存在現象學已發現多種異質的意向性，不只橫跨認知與想像（虛構和現實），還能渲染情感（快樂和悲傷）、信念（希望和遺憾）、判斷（尊敬和鄙視）的色調，因此可包含但不執著在否定性主體裡（苦難、不幸、衝突和抵抗），提供了批判和實證立場以外的肯定性態度。這樣說來，既使像阿多諾（Theodore Adorno）已從馬克思（Karl Marx）回到黑格爾（Georg W. F. Hegel）反省左派政治哲學的限制，進而推展放棄同一概念、追求異質物體的美學政治，卻仍然慣用康德（Immanuel Kant）美哲學傳統倒退式閱讀胡塞爾，因此無法認同現象學的認識詩意，更別說存有詩學（Hodge, 2008）。[2]

[2] 基本上，現象學的認識對象跨越了實存物和幻想物、再現和物自身等區分，而認識能力也超出了認知的理解範疇，甚至超越認識論、倫理學、美學和神學的啟蒙理性分工，回歸到原初的多元意向性，因此胡塞爾的認識論不需預設康德的認知個體。此處理應強調胡塞爾的「意義」和「直覺」概念與康德哲學相異之處。關於意義，胡塞爾是在意義理論下提出超驗意義（*sinn*），並非認知理論下的經驗意義（*bedeutung*），因此是能夠包含現實內在性的一般內在性，某種明證的被給出性。至於直覺，胡塞爾

站在存在現象學的橋樑上，現象社會學能再對照出海德格存有學的基進性。存在現象學和現象社會學都想整合超驗現象學和實用主義哲學，故首先要把沒有被科學知識改造的特殊自然態度，意即生活常識給例外保留在還原程序中（Renn, 2009: 156-164）。依此「自然態度的存而不論」，海德格和舒茲（Alfred Schütz）紛紛提出共同存有或多重世界、行動計劃、工具分析、死亡焦慮等雷同概念（Renn, 2009: 153-156）。然而，師承胡塞爾的兩人在 1933 年前後的德國發生了交錯複雜的思想進展。

參考文獻

Althusser, L. (1965) *For Marx*. London: Verso.

Attell, K. (2015) *Giorgio Agamben: Beyond the Threshold of Deconstruction.* New York: Fordham University Press.

Baudrillard, J.(1975[1973])*The Mirror of Production.* St. Louis: Telos.

Baudrillard, J. (1981[1972]) *For a Critique of the Political Economy of the Sign*. St. Louis: Telos.

Baudrillard, J. (1987a[1977]) *Forget Foucault.* New York: Semiotext(e).

Baudrillard, J. (1987b[1984-1985]) *Forget Baudrillard: An Interview with Sylvère Lotringer*. New York: Semiotext(e).

Baudrillard, J. (1988[1987]) *The Ecstasy of Communication*. New York: Semiotext(e).

Baudrillard, J. (1990[1979]) *Seduction*. London: Macmillan.

Baudrillard, J. (1993a[1976]) *Symbolic Exchange and Death*. London: Sage.

Baudrillard, J. (1993b) *Baudrillard Live: Selected Interviews*, Mike Gane (ed.). London: Routledge.

Baudrillard, J. (1993c[1990]) *The Transparency of Evil.* London: Verso.

Baudrillard, J. (1994a[1981]) *Simulacra and Simulation*. The University of Michigan Press.

Baudrillard, J. (1994b[1992]) *The Illusion of the End*. Cambridge: Polity.

Baudrillard, J. (1996[1968]) *The System of Objects*. London: Verso.

Baudrillard, J. (1998a[1970]) *The Consumer Society: Myths and Structures.* London: Sage.

Baudrillard, J. (2001a[1999]) *Impossible Exchange*. London: Verso.

Baudrillard, J. (2001b) *The Uncollected Baudrillard*. Gary Genosko (ed.). London: Sage.

Baudrillard, J. (2003a) *Passwords*. London: Verso.

在《觀念 I》第 24 節中出現一段隱密的文字，其中德文的直觀（*anschauung*）一詞在文脈流動到下句時被默默替換成拉丁文的直覺（*intuition*）。按整段文意推論，胡塞爾也許是想從主體面把偏頗的直觀擱置成中性的直覺，並且從客體面把死沉環境啟動為活生現實，因此造成的總體效果是把能動性從主體移交給客體（Hodge, 2008: 76, 83-84）。

Baudrillard, J. (2005a[2004]) *The Intelligence of Evil, or the Lucidity Pact*. Berg Publishers.

Baudrillard, J. (2005b) *The Conspiracy of Art: Manifestos, Interviews, Essays.* Sylvère Lotringer(ed.). New York: Semiotext(e).

Baudrillard, J. (2008[1983]) *Fatal Strategies: Crystal Revenge*. London: Pluto.

Baudrillard, J. (2010[2008]) *Carnival and Cannibal.* London: Seagull Books.

Baudrillard, J. & Guillaume, M. (2008[1994]) *Radical Alterity*. New York.: Semiotext(e).

Bernstein, R. (1991) *The New Constellation: The Ethical-Political Horizons of Modernity/Postmodernity*. Cambridge: Polity Press.

Bourdieu, P. (1991) *The Political Ontology of Martin Heidegger.* Stanford: Stanford University Press.

Derrida, J. (1973[1967]) *Speech and Phenomena*. Evanston: Northwestern University Press.

Derrida, J. (1978[1962]) *Edmund Husserl's "Origin of Geometry": An Introduction.* Lincoln: University of Nebraska Press.

Derrida, J. (1982[1972]) *Margins of Philosophy*. Chicago: University of Chicago Press.

Derrida, J. (1992) "Before the Law", pp. 181-220 in *Acts of Literature*. Derek Attridge (ed.). New York: Routledge.

Derrida, J. (1993) *Aporias*. Stanford: Stanford University Press.

Derrida, J. (1995) *Archive Fever: A Freudian Impression*. Chicago: The university of Chicago Press.

Derrida, J. (1997[1967]) *Of Grammatology.* Baltimore: John Hopkins University Press.

Derrida, J. (2005) *Paper Machine*. Stanford: Stanford University Press.

Foucault, M. (2004[1977]) "Preface", in pp. xiii-xvi in *Anti-Oedipus: Capitalism and Schizophrenia* by Gilles Deleuze and Felix Guaratti, London: Continuum.

Foucault, M. (2008) *The Birth of Biopolitics: Lectures at the Collège de France 1978-1979.* Basingstoke: Palgrave Macmillan.

Gane, M. (2003) *French Social Theory*. London: Sage.

Groys, B. (2012a[2000]) *Under Suspicion: A Phenomenology of Media.* New York: Columbia University Press.

Groys, B. (2012b[2009]) *Introduction to Antiphilosophy.* New York: Verso.

Groys, B. (2013[2008]) *Art Power.* Cambridge: The MIT Press.

Groys, B. (2014[1992]) *On the New.* New York: Verso.

Groys, B. (2016) *In the Flow*. New York: Verso.

Hamauzu, S. (2009) “Schütz and Edmund Husserl: For Phenomenology of Intersubjectivity”, pp. 49-67 in *Schütz and His Intellectual Partners.* Constanz: UVK.

Heidegger, M. (1977) *The Question Concerning Technology and Other Essays*. New York: Harper & Row.

Heidegger, M. (2008) *Basic Writings: Key Selections from Being and Time to The Task* of *Thinking.* D. F. Krell (ed.). New York: Harper Collins.

Hodge, J. (2008) “Poetic Epistemology: Reading Husserl through Adorno and Heidegger”, pp. 64-86 in *Adorno and Heidegger*. I. Macdonald & K. Ziarek (eds.). Stanford University Press.

Pefanis, J. (1991) *Heterology and the Postmodern: Bataille, Baudrillard and Lyotard*. Duke University Press.

Rajan, T. (2002) *Deconstruction and the Remainders of Phenomenology: Sartre, Derrida, Foucault, Baudrillard*. Stanford University Press.

Renn, J. (2009) “Time and Tacit Knowledge: Schütz and Heidegger”, pp.151-176 in *Schütz and His Intellectual Partners.* Constanz: UVK.

Ricoeur, P. (1967) *Husserl: An Analysis of His Phenomenology*. Northwestern University Press.

Ricoeur, P. (1991) “Hegel and Husserl on Intersubjectivity”, pp. 227-245 in *From Text to Action*. Northwestern University Press.

Rockmore, T. (1995) *Heidegger and French Philosophy: Humanism, Antihumanism and Being.* London: Routeledge.

Sartre, J-P. (1997)《嘔吐》。台北：志文。

Wahl, J. (1969) *Philosophies of Existence: An Introduction to the Basic Thought of Kierkegaard, Jaspers, Marcel, Sartre.* London: Routledge.

Zahavi, D. (2003) *Husserl’s Phenomenology*. Stanford University Press.

Žižek, S. (1999) *The Ticklish Subject: The Absent Centre of Political Ontology*. New York: Verso.

Žižek, S. (2014) *Event*. New York: Penguin Random House.

數位時代的人文反思：以大數據為線索

楊士奇*

摘　要

本文以大數據為線索，梳理數位科技的發展所衍生的人性價值問題。其一，大數據的蒐集與記憶對人們所造成的道德地位與隱私、自由等問題；其二，大數據的演算與應用可能遭遇到的倫理判斷問題。本文指出，這些問題雖然是由於新興的數位科技發展而產生，對於現有的人類治理框架也有相當的衝擊，不過，由於數位科技的發展速度飛快，因應之道並非掛一漏萬地縫補治標，而是可由哲學基礎反思、建構人們理想的美善生活，並由此主導數位科技的發展方向。

關鍵字：大數據、演算、自由、平等、正義

* 弘光科技大學文化創意產業系助理教授，Email: simbaya47@gmail.com。

Humanistic Reflection in the Digital Era：Based on Big Data Problems

Shi-chi Yang*

Abstract

In this article, I deal with two human value issues arising from the big data problems based on the development of digital technology. First, people's moral status and privacy problems caused by the data collection and memories; Second, problems of ethical judgment encountered by the calculation and application of big data. I would like to point out that while these issues arise from the development of new digital technologies and have serious impacts on existing human governance frameworks, the solution is not to fix the symptoms but to reflect on the basis of the philosophical source. People can take the opportunity to build people's ideal of good life, and thus leading the development of digital technology.

Keywords: big data, algorithm, liberty, equality, justice

* Assistant Professor, Department of Cultural and Creative Industries, Hung Kung University, Email: simbaya47@gmail.com.

一、前言：由智慧生活說起

近年來，科技與網際網路（network）呈指數型發展的樣態與速度，頗有全面改變人類生活的企圖與趨勢，而以此為基礎所勾勒的新世代智慧生活情境，尤其引人入勝。例如，今時今日上市的新型汽車，已可搭配自動停車駕駛系統，而未來的設計方向，更是以全自動上路駕駛為開發目標。另外，以發展中的智慧家居為例，人們不僅能於千里之外遠端遙控家中諸如冷氣、電視、冰箱等各項電器用品，也能令各項電器自我偵測並回報如室溫冷熱、電視節目錄影、食物冷藏冷凍等狀態；未來當其他硬體條件（如記憶容量擴增、處理速度加快等）與環境（如廠商店家也進入網路連結等）能夠配合，冰箱還能自動判斷食物如雞蛋等的餘量狀況，透過網路連線向送貨到府的商店訂購添補，在主人下班返家的同時送到家門口，減少主人料理時缺東少西或購物往返的耗時不便。

對產業界而言，這僅是聊舉其一隅而已。在科技與網路的技術引領下，目前已進入日常生活中的智慧科技，涵蓋的範圍已然包括玄關、車庫、客廳、廚房、臥室、廁所、露台等空間，而照顧到的應用領域，則自生活民生資訊、居家安全維護、個人健康管理、多媒體娛樂、社區服務連結乃至於綠能環境監控等，發展內容可說包羅萬象，對人類的體貼則趨向無微不至。科技業界甚至誇下「沒有辦不到，只有沒想到」這樣的豪語，來形容新世代智慧生活無限寬廣的藍海發展空間，以及描繪人類即將體受之高端科技生活情境。人們樂於盡情揮灑創意空間、設想各種可能成真的便利先進生活，而過去科幻電影中看似虛妄的情節，如今都只等待科技的東風吹起，一切便可水到渠成地合理實現在我們的生活世界中。

這幅美好的生活圖像背後，人們需要準備什麼條件？或者，針對智慧科技對人們生活的各種體貼作為，人們需要付出什麼樣的代價？我們是否真的願意蛋商知道家中的餘蛋數量，或是得以藉此計算家人使用蛋的頻率，甚至因而掌握我們的行蹤與行程？當這套送貨到府的即時機制進入我們的生活之後，我們是否還擁有隨時隨意挑選蛋商、挑選雞蛋的自由？當全自動駕駛的智慧汽車實際上路時，人們所讓出的決策權，是否僅止於表面上的油門、煞車與方向盤而已？當智慧汽車遭遇意外狀況，它會如何判斷因應？判斷因應的原則依據又是什麼？我們是否能接受汽車（自動駕駛系統）的判斷所造成的後果？一旦肇事，相關的責任又該由何者擔負？是坐在駕駛座的人？汽車（自動駕駛系統）本身？還是汽車（自動駕駛系統的）製造商（設計者）？更進一步的問題是，汽車自動駕駛系統的運作內涵是什麼？我們如何能夠放心地相信、使用這套系統的運作與判斷，甚至是託付自己與他人的生命？我們是否有權針對自動駕駛系統的原理原則提供修訂意見、甚至自主地改變自動駕駛系統的判斷原則？

這些關於智慧生活情境的各項提問，背後針對、追索的，其實是人的存在處境問題：當這些智慧科技為人們處理生活「瑣事」、「代替」人們下判斷之後，生活於其間的人們所扮演的角色會是什麼？表面上，是智慧科技代行人們的決定（決策），執行人們的意志；但是在科技代行與人們的真實意志之間，似乎還有諸如人的存在地位、行動的判斷原則、人的權利損益等許多關於人性價值存續的環節，隱藏著晦暗不明的疑慮，需要進一步釐清。這些疑問可以進一步分三層次呈現：一，構成這些先進智慧生活的背後所需要的科技條件是什麼？它們在體貼、介入人們生活內容的同時，對於人們的存在處境或自由、平等、正義等諸般人性價值的影響又是什麼？二，這些智慧科技要能運作順暢，背後需要一套管理機制。那麼，管理這些智慧科技的原則或原理是什麼？管理者的正當性為何？又是由「誰」來管理？三，面對新世代數位科技與智慧生活難以抗拒的展開趨勢，人們能否、又是如何掌握或搭配這些科技，來完成人們想要的理想生活？在此之前，人們興許還要追問，人們想要的理想生活是什麼？我們如何得知與彼此確認？此間，第一個問題討論的是科技與人之間的關係，亦即，智慧科技（機器）如何彼此搭配、並「得知」人們的興趣與意志？這些體貼人的作為背後所依據的資料來源為何？第二與第三個問題，則無可避免地需要納入關於公共領域（當然也包括科技與私領域之間的關係與影響）的討論：第二個問題思考的是，「誰」能夠掌握這些資料？管理、運用這些資料的原理原則會是什麼？管理、運用所產生的結果又會是什麼？第三個問題則是反問人類自己：在人們追求數位科技所帶來之智慧生活的同時，這些科技是否回頭呼應了我們關於理想生活的要求？而人們關於生活的理想又是什麼？能公共地追求嗎？

本文將依序梳理這些問題。

二、數據中的人

前述所描繪的新世代智慧生活，主要奠基在以網際網路為基礎、物物相聯為基本概念的「物聯網」（Internet of Things），而有著蓬勃發展想像空間的智慧家居，僅是物聯網科技發展的其中一個環節而已。以目前網路科技的發達程度與應用展望，尚包括有所謂的智慧工業網、車聯網、行動網、社交網乃至於支付網等，正在串聯圍繞在人類身旁的各種物，自產業製造、運輸聯繫、隨身裝置、社交互動乃至於金融消費等，逐步以數位科技為基礎，構築著人類的新世代智慧生活網絡。根據國際研究暨顧問機構 Gartner 的研究分析，2016 年最新的前十大策略性科技趨勢分別為：（一）裝置網格、（二）環境使用體驗、（三）3D 列印材料、（四）萬物聯網資訊、（五）先進機器學習、（六）自動代理與智慧物件、（七）適應性資訊安全架構、（八）進階系

統架構、（九）網格應用程式與服務架構，以及（十）物聯網平台等。Gartner 的分析指出：「前三項趨勢針對的是實體與虛擬世界的整合，還有數位網格（digital mesh）的崛起。目前企業組織都把焦點放在數位商業上，但運算業務正在逐漸崛起。藉由運算，我們可以得知事件之間的關聯性與互連性，而這恰恰定義了未來商業。在運算業務當中，很多都是源於人們並非直接涉入的背景資訊，這樣的技術是拜智慧機器所賜才能實現…。最後四項則是 IT 領域為了支援數位及運算業務而產生的現有或新型架構及平台趨勢。」[1] 簡言之，這些影響多數企業組織的科技趨勢，正是以大數據（Big Data）、雲端運算（Cloud Computing）以及物聯網為主要發展概念核心。

物聯網、大數據、雲端運算這些數位科技的發展，為人類的生活帶來了可預測的根本性變化。這一切之所以成為可能，主要依賴三大基本科技：（一）能夠定義物體、並令物體彼此之間相互識別的辨識技術，（二）包括感測特殊氣體、溫濕度、加速度、光、超音波、壓力觸覺、色彩、磁場、語音互動等各式各樣的感測技術，以及（三）包括藍芽（Bluetooth）、ZigBee、WiFi 和行動通訊等各種長短範圍不一的無線網路通訊技術。這些技術讓各式各樣的物能（1）進行自我定義、與他物彼此識別，（2）對外在環境的刺激變化分別進行感測、分析、形成數據，並（3）透過無線網路通訊，將數據傳輸至後端運算平台（即目前所謂的雲端），進行數據資訊傳輸、互聯、運算與共用，既讓原本互不相屬、性質各異的各物彼此溝通交流，更將數據資訊的蒐集分析提供給（身處遠方的）人類作為「透明」管理、即時判斷的依據，乃至於提供明確的選擇建議，進一步實現了「人與物的對話」。簡言之，只要在各個物品上分別安裝具辨識、感測與無線通訊的晶片，那麼原則上各個物品之間就具備彼此聯繫的硬體基礎，而再搭配儲存硬體與雲端運算相關數據的程式軟體，將各個物品傳輸過來的數據經過運算後加以互聯運用，基本上一組物聯網也就這樣建立起來了。（陳儀等，2014）

要之，當代數位科技為人類所擘畫的新世界藍圖，主要透過感測與識別、紀錄與傳輸，以及數據與運算等三大科技佈樁奠基，藉以蒐集形成大量的資料數據作為背景土壤，透過數據的各種運算機制，以各種物聯網為具體建構呈現。就此而言，智慧生活的內容要能豐富而充分地展開，前提是必須要感測、蒐集、紀錄有足夠大量的數據，以及能夠讓各種物彼此銜接運作順暢的運算機制。其實，在電腦與網路科技逐步發展的同時，人類也就已經同步走入所謂的「大數據時代」了：不僅是對物的感測，包括人們的各項行動，也都被日新月異的數位科技所偵測、蒐集、儲存，快速累積成大量、龐雜的數據，供人們有目的、有方法、系統地進行挖掘、分析與利用，從中獲

[1] 參見〈2016 年 Gartner 十大策略性技術趨勢觀察〉，
http://iknow.stpi.narl.org.tw/post/Read.aspx?PostID=11665

取所需的情報與利益。誠然，由歷史的角度來看，人類的發展史幾乎可以是一部情資的蒐羅史，不過，與過去相當不同的是，當代數位科技在資料處理速度與記憶容量上的指數型發展，不僅連帶高速提升了人們在掌握資訊方面的深度與廣度，甚至這些仍在不斷膨脹成長的數據資料，即將（或已經）龐大到人們無法掌握處理的地步。也正是在此處，人們在新世代智慧生活中的存在處境，有了進一步思考的必要。

首先是人的道德地位問題。當人的各種行為、行動，因著數位科技快速、龐大的資訊處理能力而（全面地）被蒐集成為資料數據並被使用，人在智慧生活中的存在地位，是否和其他物一樣，由於被感測成數據資料而成為物的一環？在智慧生活中所謂的「人與物的對話」，究竟是物擁有了「智慧」而踏上人的位置與人對話，或是人被數據化了而以物的姿態與物對話？我們是否能夠、或者有必要梳理扭轉這個看似「物化人」的新型態機制？對於這個問題的進一步思考是：一個人是否可能透過各項相關數據的感測堆積、全面性地分析而被「重建」、「複製」？我們在數據資料上掌握到的個人特質，是否與實際上的個人特質全然相符？在數位科技的時代，我們是否能透過數據真實地認識人？

其次，從人們的行動被蒐集、儲存成資料數據的方式，乃至於數據被挖掘、分析、利用的方法，都有需要進一步考慮的自由、平等、正義等人性價值問題。例如，數位科技對於人們的行動所進行的各種甚至是全面的偵測、蒐集、儲存為資料數據，幾乎已經形同對人的監視。這是否已經侵犯了人們的隱私權或自由權？時至今日，面對這些等同監視的智慧科技近乎全面性的包覆，人們是否仍有能力拒絕這些偵測、蒐集，乃至於形成資料數據的紀錄？此外，在數據的利用方面，人們被偵測、紀錄而成的各種資料數據，是否能被平等地對待、評估？舉例而言，我們是否允許保險業者因為掌握了人們的致病相關數據而拒絕帶有高風險致病基因的人士投保健康或疾病保險？

其三，若是只針對各種物或人們的行動進行感測與紀錄，那僅是作為雜多的數據而已，尚非是具有意義的資料；要讓這些數據資料能夠「說話」、呈現出有用的價值，則需要進一步按目的而設計的運算法（algorithm），依所需加以挖掘、分析、評價。不過，在發展智慧科技方面，科學家們對於演算法發展前景的期許，顯然超出尚在藍海悠游的物聯網或智慧生活的要求甚多。目前，演算法的發展大致可歸納為五大演算學派，包括：將知識視為符號來組織結構的符號理論學派、由生物學的角度反推人類大腦學習模式的類神經網路學派、以物競天擇理論為知識演算基礎的演化論學派、主張知識學習的不確定與機率推理的貝式定理學派，以及以向量概念為演算預測基礎的類比推理學派等。這些學派根據不同的目的，各自發展出一套在特定範圍內能

有效解讀、利用數據的演算法，為人們開拓了新世代的智慧應用領域；而目前科學家最新的努力方向，並不滿足停留在為人們提供特定目的的數據解讀與應用，卻是試圖整合五大演算法、甚至是設計一個能全面性思考的「大演算」（the master algorithm），以開發能比擬人類思考的智慧機器為目標。（張正苓等，2016）換言之，智慧生活的具體情境目前尚未成熟，數位科技的發展卻已經在預想下一個能像人類那樣思考的智慧機器（人）的階段了。不過，無論是哪一種演算法學派，或者是能擬人思考的智慧機器，此處我們的問題是：在面對人世間的（倫理）兩難問題時，這些演算法是否能協助人們追求公平正義？例如，汽車產業大力發展的汽車自動駕駛系統，將會如何解決倫理學上著名的電車難題（the trolley problem）？我們是否能容許自動駕駛的智慧汽車，在面臨兩難狀態時選擇衝撞少數人、放棄少數人的生命權，甚至在危急時刻，容許汽車放棄車主的生命，以自我毀滅的方式，挽救更多人的生命？或者，相反地，無論遇到什麼樣的意外，都以保障車主的生命權利為優先，無視於其他可能在車禍中遭遇橫禍的無辜路人？我們好奇的是，自動駕駛系統會如何抉擇？而其抉擇的演算依據，背後的原則（倫理原則或是計算原則）依據又是什麼？顯然，這裡探問的是，智慧機器背後的演算法，是否有能力演算人們的倫理問題？若答案是肯定的，那麼這項能力是由於演算法本身就包含倫理原則，或者是依知識內涵演算而得出具倫理價值的結果？若答案是否定的，那麼人們是否有能力在數位科技為人們所建構的智慧生活情境中，介入或主導處理倫理問題？

這些問題，一方面是數據記錄所衍生的道德地位與隱私自由問題，另一方面則是關於數據利用上的原理原則問題。換言之，智慧生活的開展，除了新穎數位科技的運用之外，無可避免地必然要觸及到自由、平等、正義等重要人性價值的思考。問題是，在這一波數位科技的智慧生活革命，我們目前處在哪一個階段？我們現在還來得及介入嗎？

三、討論（代結語）：問題的整理與展望

(一) 監視：隱私、記憶與遺忘

數位科技發展至今，我們已經很難取消無所不在、無時不在的監視問題。然而，由隱私權、自由權的思路著手，似乎也只能說明數位科技對於現有人類治理架構的衝擊，並無法徹底解決數位科技對人們的窺探。以對於人權議題較為敏感的歐洲為例，歐盟最高法院曾於 2014 年 5 月做出一項有關「被遺忘權」（right to be forgotten）的判決，指出人們有權要求如 google 等網路搜尋引擎，移除有關個人敏感、無關、不當或過時的資訊，強調個人擁有個人資訊的最高處分權。儘管此項判決受到部分人士的歡

迎，不過，包括英國司法部、倫敦媒體監督機構（Index on Censorship, 言論審查指標）與 Google 等科技公司等，卻是消極地否定甚至是表明窒礙難行。除了司法部表態不支持此項判決之外，言論審查指標也指出，「被遺忘權」很有可能被那些想要漂白過去歷史的人加以反利用。[2]

當然，我們還可以進一步討論：過去犯錯的人是否無權「漂白」歷史？不過，也就是在安全防弊的思考上，以國家安全或者反恐為名的監視，其實正在透過各種人們想得到或想不到的管道，監視全球的各種人類活動。國際安全科技專家布魯斯・施耐爾（Bruce Schneier）便指出：「當我們在數位化生活中移動奔走，產生巨量資料時，政府和企業會蒐集並分析這些資料。我們通常不知道它們正在這麼做，而它們往往也沒有取得我們的同意。根據這些資料，它們會對我們每個人下出某種結論，而這個結論可能是我們所不同意或反對的，也可能會對我們的生活產生深遠的影響。我們可能不願意承認，但我們正活在大眾監控之下。」（韓沁林，2016：33）施耐爾此處指出的，不管是基於商業考量或是國家安全，重要的是，「我們通常不知道這些監控正在發生，而它們也不曾取得我們的同意」。此處，人的存在處境或道德地位再一次受到威脅：我們以為總有某些地方是可以讓我們保有隱私的，卻沒有想到在衛星、網路等科技的推波助瀾下，我們無所不在、無孔不入地被監視，赤裸裸地被監視。

由「被遺忘權」到全面被監視，數位科技帶來的，不僅是方便的智慧生活，人們被迫要讓出的人性價值，代價恐怕更多。更重要的是，當監視者從未過問我們的同意便進行監視，更在我們思考所未及的地方進行監視，那麼我們是否只能透過修改法律、制訂規範等司法途徑來進行近乎曠日廢時的防堵、修補，而不能有其他更積極的作為？更甚者，這些規範是否真能約束那些躲在暗處的監視？此外，一個較為正向的思考是，自古以來，由於各種可欲或不可欲的需求，人們從未停止對於他人的監視窺探——特別是秘密進行的、往往是非法的監視窺探。如今，數位科技的普及，一方面雖然使得監視窺探顯得更為方便，另一方面卻也讓監視這項議題無所遁形地展開在人群裡。一旦「所有人監視所有人」的時代來臨，那麼，或許也就是人們開始認真思考監視與人性價值之間的權衡問題了。

另一種監視問題，則是關於數據的運用問題。目前，人們偶而能感受察知的，大部分來自於商業上的應用。例如，企業透過人們在網路上的查詢、瀏覽紀錄，來推測消費者的喜好，以作為商品設計或行銷手段的依據；透過各種金融消費記錄，來察知人們的消費習慣、品味、乃至於購買意向，以提供消費者有興趣的相關廣告，刺激消費者的購買欲等等。這樣關於人們決策意向的監視，進一步發展出能依數據推算人類

[2] 楊芬瑩，〈「被遺忘權」，歐盟判決 Google 須遵守〉。http://www.storm.mg/lifestyle/31067

思考意向或行為意義的演算法：問題在於，一旦能擬人思考的智慧機器真的出現，又會對人們造成什麼影響？

(二) 權衡：演算的意義

有著百餘年歷史、具國際影響權威指標的英國標準協會（British Standards Institute, BSI），近日針對智慧機器的研發，發表了一套「機器人與機器系統的倫理設計與應用指南」（Guide to the ethical design and robots and robotics systems）。這是目前業界第一套關於機器人設計的倫理規範要求，旨在指導研究、設計與生產機器人的研究者與廠商，如何對機器人所能造成的影響做出道德評估，以確保人類生產出來的智慧機器，能夠融入人類既有的道德規範裡。這套指南開宗明義便提出：「不可設計出專門殺害或傷害人類的機器人；人類要為機器人的行為負責；任何機器人都必須有人負責，而且必須確保有人能為機器人的行為負責」[3] 等幾項明確將智慧機器的行為歸責於人類的要求，恰恰間接回應了本文前述關於智慧汽車自動駕駛系統的責任歸屬問題：亦即，智慧機器的行為判斷以及所造成的後果，必然要由人類來負責。按照這套倫理指南「免除」了自動駕駛系統的責任以及所瞄準的行為主體來看，顯然英國的倫理思考，會把較大的責任放在智慧機器的製造商身上。這不免就引發下一個問題：我們似乎在設計智慧機器之初就該把人類世界的道德規範「安裝」在機器裡，以便智慧機器能夠「聽話地思考」。只是，問題是否真有如此簡單？

就《指南》而言，英國的確將智慧機器「具有超越既有範圍之自學能力」視為一個機器製造商需要謹慎控制的危害因素之一。[4] 不過，這項要求一方面顯然與推展大演算理論的科學家們期待相悖，另一方面也確實為機器的「智慧」設下了限制：人們既然要為機器的行為負責，那麼理所當然地，機器的「思考」不能超出人們所授與的範圍。另一個問題是，即便機器的智慧受到設計上的限制，然而，機器的演算功能是否能完美地處理倫理問題？人類都未必能解決的倫理難題，智慧受到限制的機器能處理到什麼程度？

此處，我們只能暫時慶幸，按照《指南》的要求，科幻電影中的奇思妙想尚未有成真的可能。

(三) 小結

本文目前仍處於發展階段，相關問題的哲學思考，一如智慧科技的發展潛勢般，亦頗有泅游於藍海之感。不過，以自由、平等、正義等人性價值出發，我們至少可以

[3] 〈英國正式頒佈機器人道德標準：不許傷害、欺騙人類與令人成癮〉，參考自 http://www.inside.com.tw/2016/09/21/official%ADguidance%ADrobot%ADethics%ADbritish1/3

[4] 其他危害因素還包括了機器欺騙、令人成癮等。參考同上。

在幾個方向上努力：一是關於自由的權衡，二是平等的態度，都不應由於人們資訊的揭露，而有所差異對待。此外，智慧生活的相關數位科技，確實能協助人們過上更好的生活，不過，人們也應避免因貪圖過於方便的服務，而受制於機器的智慧演算。相反地，人們應可趁此高端科技發展的機會，重新思考人們的理想生活的內涵，追尋生命更為深層的意義。

參考資料

Floridi, Luciano.(2011). *The Philosophy of Information*. New York: Oxford University Press.

Freeman, Samuel. (2004). "Public Reason and Political Justifications," in *Fordham Law Reviews*, vol. 72, (5), pp.2021-2072.

Freeman, Samuel. (2007). *Justice and the Social Contract: Essays on Rawlsian Political Philosophy*. New York: Oxford University Press.

Gaus, Gerald. "The Rational, the Reasonable and Justification," in *The Journal of Political Philosophy* 3(3), pp.234-258.

Rawls, John. (1971a). *A Theory of Justice*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Rawls, John.（1999a）*A Theory of Justice*. Revised edition. Cambridge, Mass.: Belknap Press of Harvard University Press.

Rawls, John. (2001). *Justice as Fairness: A Restatement*. Cambridge, Mass.: Harvard University Press.

Rawls, John. (2005). *Political Liberalism*. N. Y.: Columbia University Press.

Reidy, David. (2007). "Reciprocity and Reasonable Disagreement: From Liberal to Democratic Legitimacy," in *Philosophical Studies* 132, pp.243-291.

Scanlon, T. M. (1998). *What We Ought to Each Other*. Cambridge, Mass.: Harvard University Press.

Smith, Adam. (2002). *The Theory of Moral Sentiments*. Knud Haakonssen (ed.), Cambridge: Cambridge University Press.

Smith, Steven. (2002). *Defending Justice as Reciprocity: An Essay on Social Policy and Political Philosophy*. Lewiston, N. Y.: Edwin Mellen Press.

布魯斯・施奈爾（Bruce Schneier）。《隱形帝國》（韓沁林譯）。台北：如果。

尼克・比爾頓（Nick Bilton）。2011。《一位數位移民的告白》（王惟芬等譯），台北：行人。

克里斯・安德森（Chris Anderson）。2013。《自造者時代：啟動人人製造的第三次工業革命》（連育德譯），台北：天下文化。
克雷・薛基（Clay Shirky）。2011。《鄉民都來了：無組織的組織力量》（李宇美譯），台北：貓頭鷹。
林宇玲。2014。〈網路與公共領域：從審議模式轉向多元公眾模式〉，《新聞學研究》，第 118 期，頁 55-85。
佩德羅・多明戈斯（Pedro Domingos）。2016。《大演算》（張正苓等譯），台北：三采文化。
哈伯瑪斯（Jurgen Habermas）。2002。《公共領域的結構轉型》（曹衛東等譯）。台北：聯經。
柯利・多克托羅（Cory Doctorow）。2015。《資訊分享，鎖得住？》（朱怡康譯），台北：行路。
馬丁・福特（Martin Ford）。2016。《被科技威脅的未來：人類沒有工作的那一天》（李芳齡譯），台北：天下。
麥爾荀伯格、庫基耶（Viktor Mayer-Schonberger、Kenneth Cukier）。2013。《大數據》（林俊宏譯），台北：天下文化。
麥爾荀伯格（Viktor Mayer-Schonberger）。2015。《大數據：隱私篇》（林俊宏譯），台北：天下文化。
傑瑞米・里夫金（Jeremy Rifkin）。2014。《物聯網革命》（陳儀等譯），台北：商周出版。
凱斯・桑思汀（Cass R. Sunstein）。2015。《剪裁歧見：訂作民主社會的共識》（堯嘉寧譯）。台北：衛城。
達娜・博依德（Danah Boyd）。2015。《鍵盤參與時代來了！》（陳重亨譯），台北：時報出版。

論大數據的知識論條件

蔡偉鼎*

摘　要

在數位時代裡，吾人所面對的大量資訊早已遠遠超出個別人力所能負荷分析的量，以致於有諸多問題不得不交由電腦來連結處理。這種透過電腦處理大量資訊的方式被稱為「大數據」（big data），其宣稱不必尋求因果性的解釋來保證兩事件間的關係是必然的，而只需滿足於其相關性（correlation）要求即可。也就是說，只要能確認兩事件之間具有強的相關性，就足以視為有效的科學證據，並充作進一步推論的前提。有鑑於大數據之應用範圍現已拓展到人文科學之領域，且成效卓越，以致有不少人視之為人類挖掘未知事物的一項最新研究利器，甚至樂觀地認為終可求助於大數據來取代人類理性思考的工作。面對這樣一種方興未艾的大數據研究趨勢，吾人實有必要對其進行哲學反思，以考察大數據作為理性思考的可能性條件。本文即試圖聚焦討論大數據的知識論特徵，並透過現象學觀察來提出幾點評論。

當吾人欲去考察大數據的成立條件時，首先可清楚看出其**外在條件**無疑係在於今日資訊儲存與處理技術的大幅精進，遂才令資訊的儲存量能快速地累積，並得以用大規模的方式來進行分析。然而，光是訴諸這些外在條件，並不足以凸顯出大數據之所以被接受為具有說服力的分析工具的核心原因。關鍵應是去接著剖析其**內在條件**，亦即其知識論條件，因為這才能進一步說明為何這裡存在著「量變產生質變」的效果。細究之，儘管大數據與以往小規模的資料分析方式──亦即所謂的「小數據」（small data）──同樣均是奠基於量化的思維模式，但這兩者間仍是有差異的。基本上，這裡出現研究態度上的三重改變：（1）由於能夠獲取及分析的資料量大為增加，不再需要依靠隨機抽樣的方式來分析數量較小的資料集；（2）不需再堅持一切資料都要數據精確，因其是以容忍測量上的誤差來取代抽樣所造成的誤差；（3）分析的重點在於從巨量資料中找到事物間的相關性來進行預測未來趨勢，而不必然需要先確定其彼此間是否存在因果關係。基於對上述的三重特徵差異的理解，筆者將先指明大數據所強調的「相關性優先於因果性」這種知識論態度可謂是一種實用主義

* 東海大學哲學系助理教授，Email: tsaiweiding.tw@thu.edu.tw。

（pragmatism）的態度，因為其只要求能達乎實效即可。此外，根據羅逖（Richard Rorty），實用主義式的知識論態度並不認為人類心靈能夠完全精確地反映實在。因此之故，大數據滿足於資訊不精確性的作法亦有其道理。接著，藉由重新反思休姆（David Hume）對因果關係的批判，吾人亦可瞭解大數據不特別考慮因果性的作法並不全然是毫無理性思辯的支持。最後，筆者試圖基於現象學理論指出，大數據強調事件相關性的作法其實還可以藉助現象學直觀方法來獲得另一層面的理論補充，因為現象學恰恰就是一種試圖在諸現象脈絡中洞察到其本質性關聯的研究方法。

關鍵字：大數據、因果性、相關性、實用主義知識論、現象學

On Epistemological Conditions of Big Data

Wei-ding Tsai*

Abstract

In the digital era, the amount of information which we faced is too large so that a lot of problems are far beyond human capacity to be analyzed and have to be handed over to computers to be processed. This way of dealing with a large amount of information by computers is called the "big data," and claims that it doesn't need to seek causal explanations to ensure the necessity of relationship between two events, but only satisfies with their correlation. In other words, as long as there is a strong correlation between two events, it is enough for us to regard it as a valid scientific evidence, and as a premise for further arguments. Since now the application of big date has been extended to the fields of humanities and already made remarkable achievements, a lot of people take big data as a new research tool to mine unknown correlations out of large information, or even optimistically to replace human's rational thinking. Faced with such an ascendant trend of big data, it is necessary to reflect on it philosophically and to study the conditions of possibility of the big data as rational thinking. This article tries to discuss on the epistemological characteristics of the big data and make some comments from the phenomenological perspective.

When we want to study the conditions of possibility of the big data, we can, first of all, clearly find out that its external conditions base beyond doubt on the fact that today's technology of information storage and processing develops so rapidly, that the information can be accumulated quickly and be analyzed in a massive way. However, resorting to these external conditions alone is not sufficient to explain why the big data can be accepted for a convincing tools to analyze information. Therefore, we turn to study its internal conditions, i.e. its epistemological conditions, in order to explain why the big data can achieve an unexpectable effect that the quantitative change can produce a qualitative change." Such an effect marks a status which differentiate the big data from the so-called "small data"—the way of dealing with a small amount of information

* Assistant Professor, Department of Philosophy, Tunghai University. Email: tsaiweiding.tw@thu.edu.tw

by computers—, although both all base on the quantitative patterns of thinking. Basically, there is a triple change of the research attitude in the big data: (1) the amount of available information has increased so radically that we no longer need to relies on the method of random sampling to analyze smaller set of data; (2) we no longer need to insist on the accuracy of all datum, because now we endure errors in measurement rather than errors in sampling; (3) the focus of the big data is to forecast trends in the future by analyzing the correlation between two things from huge volume of data, and it is not necessary to find out at first whether there exists a causal relationship between both things. Based on the understanding of three differences of attitude mentioned above, I will show that the epistemological attitude about the priority of correlation over causality in the big data is actually a kind of pragmatic attitude, because it only requires practical achievements. In addition, according to Richard Rorty, pragmatic theory of knowledge does not believe that human mind can reflect the reality totally and exactly. Thus, it could serve as a convincing reason to justify why the big data contents with the inaccuracy of information. practices are justified. Then, I will try to indicate through a new reflection on David Hume's critique of causality that the big data's epistemological attitude about neglect of causality is not entirely without any philosophical support.

Keywords: big data, correlation, causality, pragmatic epistemology, phenomenology

一、

我們已然生活在一個**數位化的世界**裡。這並不是說「我們在一個全然**虛擬的世界**裡過活」，而是說「我們生活在一個已被虛擬世界滲透的現實世界裡」。當我們使用平板電腦、智慧型手機來進行資料搜尋、線上購物、倉儲管理、行進導航、社群互動、遠距遙控等等事務時，我們均可說是即時地通過網路獲得資訊，。因為我們的生活世界這個世界是透過虛擬的數位化資訊，它逼迫我們去思考這個新的生活世界對我們的影響。在這個世界裡，我們的生活不斷被數位化的儀器。

在數位時代裡，吾人所面對的資訊總量不斷地大幅擴增，早已遠遠超乎單一個人分析能力所能負荷的量，以致於今日有越來越多的資料分析任務不得不交由電腦來連結處理。這種透過電腦處理巨量資料的技術可被稱為「大數據」（big data）——或被譯作「巨量資訊」。由於在茫茫的巨量資訊大海中搜尋有用的資料，就有如在暗無天日的地底裡挖礦一般，故此一過程遂被稱為「資料探勘」（data mining）。綜觀大數據作為資訊管理之技術，一開始多被應用於自然科學及商務金融的領域裡，近來更被運用到人文科學的研究上，後者即為所謂的「數位人文」（digital humanities）[1]。有鑑於大數據之應用範圍不斷向外拓展，且成效卓越，以致有不少人視之為人類藉以挖掘未知事物的一項最新研究利器，甚至樂觀地認為終將可以求助於大數據來取代人類理性思考的工作。譬如谷歌（google）所研發的圍棋對弈軟體「AlphaGo」在 2016 年 3 月間與韓國職業棋士李世乭對戰獲得四勝一負的佳績之後，越來越多人不再懷疑通過分析處理巨量資訊來自我學習的人工智能極有可能超越人類的理性思維能力。面對這樣一種方興未艾的大數據研究趨勢，吾人實有必要對其進行哲學反思，考察大數據作為理性思考的可能性條件，以避免在資料探勘過程中發生統計學上戲稱為「資料挖泥」（data dredging）──亦即發生錯誤歸因或不當關聯之結果──的情勢。

本文企圖對大數據進行哲學性的後設探討，以思索今日訴諸大數據來作為人文研究方法的哲學理論基礎何在。筆者在此係將這樣一種後設探討的工作稱作為「數位理性批判」（Kritik der digitale Vernunft）。須事先澄清的是，這裡所說的「數位理性」是意指那以大數據作為運作基礎的人工智能，因為後者確實能表現出一種類似於人類所具有的分析、整合、推理資料之高階理性認知能力──儘管其物質性構成基礎跟人類大不相同。[2] 再者，這裡所謂的「批判」，並不是指日常語言下的「批評」之義，而是挪用康德哲學意義下的「批判」概念，亦即去找出所研究對象之所以成立的可能性條件。簡而言之，本文是試圖去反思大數據之所以能被視為合理知識的可能性條件。

為達成前訂之目的，本文擬採取以下幾個步驟來進行處理：首先，筆者將考察資訊科學家們對大數據的解說，以剖析其對大數據之知識論條件的認知。有鑑於一般人對大數據的認知僅僅取決於電腦軟硬體設備上之差異，故對其本質的理解多係著眼於

[1] 有關數位人文之內涵及發展，請參考項潔與涂豐恩之簡介（項潔、涂豐恩，2011）。

[2] 參見：2010 年 6 月 23 日，德國《法蘭克福匯報》（FAZ）〈人工智能：數位理性批判〉（Künstliche Intelligenz Kritik der digitalen Vernunft）http://www.faz.net/aktuell/feuilleton/debatten/digitales-denken/kuenstliche-intelligenz-kritik-der-digitalen-vernunft-1997027.html?printPagedArticle=true#pageIndex_2

其某些**外在條件**，然而後者其實尚不足以證成大數據的知識論效果何以能不同於以往小規模的資料分析方式。為了說明這點，本文接下來將指出大數據之合理性毋寧是奠基於一種實用主義的知識論態度上。最後，筆者試圖基於現象學理論指出，大數據強調事件相關性的作法其實還可以藉助現象學直觀方法來獲得另一層面的理論補充，因為現象學恰恰就是一種試圖在諸現象脈絡中洞察到其本質性關聯的研究方法。

二、

若吾人欲去考察大數據的成立條件，自然應先確認「大數據」這個概念的實質內容，也就是說應先了解其定義，以確定「大數據」到底是什麼。可是，正誠如麥爾荀伯格（Viktor Mayer-Schönberger）與庫基耶（Kenneth Cukier）在 2013 年出版的名著《大數據》（*Big Data: A Revolution That Will Transform How We Live, Work, and Think*）中承認，「大數據」一詞迄今並未有一個公認的明確定義。儘管如此，他們仍將「大數據」的基本特徵描述如下：「資料量一定要達到相當規模才能做的事情（例如得到新觀點、創造新價值），沒有一定規模就無法實現，而且這些事將會改變現有市場、組織、公民與政府的關係。」（Mayer-Schönberger & Cukier: 2013, 14）簡言之，能夠被稱作為「大數據」的資料分析技術，至少應當滿足以下兩個條件，亦即：（1）資料量要很大，並且（2）要出現「量變造成質變」的效果。然而，這樣的描述其實並未**積極地**指出其實質內容，而是**消極地**限定其指涉範圍罷了。為了說明這點，以下分別對這兩個條件進行深入的考察。

首先，一般通俗對大數據的理解大多只限於前述之第一項條件。不過，我們可以發現第一項條件其實是含混的，因為這裡並沒有明確規定資料量要多大才能算是「大數據」。也正因為如此，人們往往只能以套套邏輯的方式來說「大數據就是數據規模很大」。然而，這並不能讓人對其有任何更多的理解。有鑑於此，有些人試圖給予更明確一點的規定，指其所處理的資料量之規模至少要達到 TB（即 1000GB）、PB（1000000GB）、甚至 EB（1000000000GB）才行。[3] 不過，如此之修訂依舊維持著某種含混性，致使其對於「大數據」的指涉範圍也只是游移於 TB、PB 與 EB 之間，而未敢乾脆斷然說其只須超過 TB 即可。顯然地，這樣的修訂並不是好的解決方案，畢竟此法還不足以給出一個精確的答案。另一個較多數人支持的修訂方案，則是延續美國 IT 產業分析師 Douglas Laney 在 2001 年提出的論點，主張「大數據」相較於以往的小規模資訊處理技術──即所謂的「小數據」（small data）──具有三項特徵，亦即：其在資料之容量（Volume）、速度（Velocity）以及多樣性（Variety）上均更為龐大。[4] 這個論點雖然也沒有明確規定資料量的規模要達到多大的範圍，不過它透過強調資料的類型要多、處理資料的速度要快，從而至少對「大數據」之指涉範圍做出了更進一步

[3] 參見：維基百科條目「大數據」說明。網址：https://zh.wikipedia.org/wiki/%E5%A4%A7%E6%95%B8%E6%93%9A#.E5.AE.9A.E7.BE.A9。擷取日期：2016/9/15。

[4] 同前註。

的限定。乍看之下，這似乎是個更好的提案，因為它為「大數據」補充說明了另外兩個不同的**質**（Quality）上的特徵，而不滿足於僅僅以資料之容量為唯一的關鍵。但只要再經細察，吾人即可發現這三個特徵基本上均仍是著眼於**量**（Quantity）上的差異，換言之，均是以**量化**的觀點來定義「大數據」的內容。縱使如此，其卻同樣未能明確限定量之大小，說明到底資料容量要多大、速度多快、種類多少才算是「大」數據。就此而言，這其實仍有犯循環定義之嫌，故並不算是真正解決了「大數據是什麼」這個問題。歸根究底，單憑訴諸量化的方式來定義大數據，至多只能說明其**外在條件**而已，尚不足以指明其關鍵之處。

無疑地，大數據的出現係緣於今日資訊儲存與處理的技術大幅精進，遂才令資訊的儲存量得以快速累積，並以大規模的方式來進行分析。然而，亦正因為科技發展迅速，以至於人們對數據規模的大小判定往往也會隨之相對化。昔日被視為巨大者，沒隔幾年就已淪為微小者。如此迅速的變化使得論者們對於大數據在量上的定義游移不決。甚至還讓他們意識到：光是訴諸這些可量化的外在條件，並不足以揭明大數據的核心意義。因此之故，無疑還有必要訴諸前述的第二個條件，以作為補充。

第二個條件是要求大規模的資訊處理分析能夠得出小規模資訊處理技術在先前所難以想像、也無法達成的結果，從而具有量變產生質變的效果。譬如谷歌公司（Google Inc.）曾經光靠比較分析美國民眾上網搜尋的關鍵字及搜尋頻率，就令人詫異地推算出有關美國流感傳播之時間與地區的即時資訊。（Mayer-Schönberger & Cukier: 2013, 8-10）如此這般找出事件關聯性的成果不但是要依靠運算了極大量的個別資料後才有可能達成，而且還遠遠超出了往昔認為需透過直接採集檢體或醫院疾病通報系統才能確定的認知模式。總之，這個條件是要求滿足一種**質上的差異**，以俾能與過去的資訊處理技術做出明確的區隔。乍看之下，這個補充條件更優於 Laney 的修訂方案，因為其不是僅只重視量上的差異而已。儘管如此，這個條件一樣也是含混的，因為它並沒有對**質上的內容**做出任何實質的斷言，從而也就未對資料分析對象的範圍給予明確的限定。事實顯示，它也不可能提出如此這般的斷言，否則的話，其就不再是「先前所難以想像」之事。由此看來，吾人若要進一步說明為何這裡存在著「量變產生質變」的效果，則不可滿足於僅從所分析的對象及其之間的關聯性來切入思考，因為如此一來除了只能說大數據會對一切可能對象的關聯性都保持開放之外，其他的甚麼也無法多說——這其實等於什麼也沒說。既然吾人無法去明確限定資料分析對象的範圍，那麼恰當的因應之道應是將關注的焦點從分析處理的**對象**轉移到分析處理的**過程**本身，以俾能指出大數據的**內在條件**。

簡言之，儘管大數據跟小數據同樣均奠基於量化的思維模式，唯這兩者間仍是有差異的。現在問題是，光訴諸量上的差異並不足以凸顯出大數據之所以被人接受為具有說服力的分析工具的核心原因所在。此外，透過研究對象之間的關聯性雖可標示出某種質上的差異，但其仍不足以充分說明大數據為何能有量變造成質變的效果。因此之故，造成質變效果的因素被人認知。麥爾荀伯格與庫基耶指出，這裡出現研究態度上的三重改變：（1）由於能夠獲取及分析的資料量大為增加，不再需要依靠隨機抽樣

的方式來分析數量較小的資料集；（2）不需再堅持一切資料都要數據精確，因其是以容忍測量上的誤差來取代抽樣所造成的誤差；（3）分析重點在於從巨量資料中找到事物間的相關性來預測未來趨勢，而不必然要先確定其彼此間是否存在因果關係。[5]

三、

基於對上述三重特徵差異的理解，筆者將先指明大數據所強調的「相關性優先於因果性」這種知識論態度可謂是一種實用主義（pragmatism）的態度，因為其只要求能達乎實效即可。此外，根據羅逖（Richard Rorty），實用主義式的知識論態度並不認為人類心靈能夠完全精確地反映實在。因此之故，大數據滿足於資訊不精確性的作法亦有其道理。接著，藉由重新反思休姆（David Hume）對因果關係的批判，吾人亦可瞭解大數據不特別考慮因果性的作法並不全然是毫無理性思辯的支持。最後，筆者試圖基於現象學理論指出，大數據強調事件相關性的作法其實還可以藉助現象學直觀方法來獲得另一層面的理論補充，因為現象學恰恰就是一種試圖在諸現象脈絡中洞察到其本質性關聯的研究方法。

參考書目

項潔、涂豐恩。2011。〈導論──什麼是數位人文〉，收錄於：項潔編輯，《從保存到創造：開啟數位人文研究》（9-28 頁），台北市：國立台灣大學出版中心。

Viktor Mayer-Schönberger 及 Kenneth Cukier 著。2013。《大數據》（*Big Data: A Revolution That Will Transform How We Live, Work, and Think*）（林俊宏譯）。台北：天下文化。

[5] 參見：Viktor Mayer-Schönberger 及 Kenneth Cukier 著，《大數據》（*Big Data: A Revolution That Will Transform How We Live, Work, and Think*）。

Panel F

學生培育：新世代人才的數位研究能力培育

The Development of Research Students: The Cultivation of Digital Humanities Literacy in the New Age

Panel F

學生培育：新世代人才的數位研究能力培育

主持人	鄭文惠（國立政治大學中國文學系教授） Wen-huei Cheng (Professor Department of Chinese Literature, National Chengchi University)
發表人	祝平次（國立清華大學中文學系副教授） Ping-tzu Chu (Associate Professor of Department of Chinese Literature, National Tsing Hua University)
題目	臺灣數位人文教育的困難與展望 Difficulties and Prospect of Digital Humanity Education in Taiwan
發表人	邱偉雲（湖北經濟學院新聞與傳播學院副教授） Wei-yun Chiu (Associate Professor of Hubei University of Economics)
題目	跨越範式：數位人文之人才培育及其多元挑戰 Cross-Paradigm：Talent Development and Its Multiple Challenges in the Digital Humanity Field
發表人	Duncan Paterson（德國海德堡全球情境亞歐卓越研究中心博士候選人） Duncan Paterson (PhD Candidate of the Cluster of Excellence Asia and Europe in a Global Context in Heidelberg, Germany)
題目	Matter Matters：Cultural Heritage Objects and Digital Literacy 重中之重：文化資產與數位素養

Panel F

學生培育：新世代人才的數位研究能力培育

本場 Panel 主要從教育場域、學生培育及數位人文學習與反饋機制等面向，凸顯新世代人才的數位人文素養的重要性與養成方式。Panel 計邀請三位學者講演，其一：清華大學中文系副教授祝平次〈台灣數位人文教育的困難與展望〉，以在台灣推廣「數位人文工作坊」為例，透過數位工具與文史研究實作，讓學員學習數位工具，如中國歷代人物傳記資訊庫、地理資訊系統和社會網絡分析軟體等，但由學習數位工具到有一定經驗及視野並以數位人文做為一種方法或手段，具體運用於實際的研究與教學上，仍亟須加強。此外，數位文本來源及資訊學者由程式分段運作到程式自動化之數位工具的建置以滿足研究與教學的需要，或大量文史學者的投入等，都是數位人文教育與研究發展上的整體性且重要性的建置。其二：湖北經濟學院新聞與傳播學院中文系副教授邱偉雲〈跨越範式：數位人文之人才培育及其多元挑戰〉，以政治大學金觀濤講座教授與鄭文惠教授所帶領的數位人文研究群為例，討論數位人才培育須正視的問題：一、人文與數位學員在提出與解決問題上所存在的差異性如何透過不斷的對話有所調和；二、因應學員數位背景知識不足，亟須透過教育增設質化與量化研究方法的基礎課程；三、因數位人文研究的探索性太強，須有高度熱情投入以求不斷的創發，且就業市場的缺乏，導致學員不易持續探索，這是人才培育上的現實困境。其三，海德堡大學博士候選人 Duncan Paterson〈重中之重：文化資產與數位素養〉，聚焦於數位人文的認識論、實踐及美學條件、教學效果等，以所任教的研究所提供計畫培訓課程、專題討論及年度研習會三種數位人文教育場合，突出在專門培訓課程中確立執行準則與標準程序，以促使學生能將方法應用於自己學習或研究的材料中，並積極與數位人文文獻結合，這在數位人文作為一門新興專業的教學，在方法與內容上缺乏廣泛認同的共識上，實質提供了一種數位人文教育的在地實踐。

Panel F
The Development of Research Students: The Cultivation of Digital Humanities Literacy in the New Age

This panel aims to highlight the importance of digital humanities literacy and its cultivation in terms of teaching venues, the development of research students, digital humanities learning and feedback mechanism. This panel invites three scholars to give presentations. In his paper ”Difficulties and Prospect of Digital Humanity Education in Taiwan”, Ping-Tzu Chu, Associate Professor of Department of Chinese Literature, National Tsing Hua University, presents his experience in promoting “workshop of digital humanities” in Taiwan. Through the workshop, the students are able to learn how to use digital tools such as China Biographical *Database* Project, geographic information system, software for social network analysis and so on. However, it is suggested that, regarding the application and implement of digital technology in humanities research and teaching, there is still plenty of room for improvement. In addition, it is indicated that the development of digital humanities education and research has achieved remarkable results as there are more humanities scholars involved in the field and various digital tools ranging from databases of texts, to program segmentation and program automation have also been established.

In his “Cross Paradigm: Talent Development and Its Multiple Challenges in the Digital Humanity Field”, Wei-Yun Chiu, Associate Professor of Department of Chinese at Hubei University of Economics, investigates the issues regarding the cultivation of future talents for digital humanities research by taking the digital research project led by Professor Guan-Tao Jin and Wen-Huei Cheng of National Cheng-Chi University for case study. This papers explores how to reconcile the differences regarding problem shooting between the students of humanities and digital technology through dialogue and communication. It also indicated that, in order to reinforce the knowledge of students on digital technology, it is important to develop and include the courses of quantitative and qualitative research methodology into syllibus. Finally, one of the difficulties in developing research talent is that, as the academic job market for digital humanities research is a small niche and digital humanities research demands constant dedication and passion, it is not easy for the research students to carry on their research projects or even interests.

Finally, in his paper” Matter Matters: Cultural Heritage Objects and Digital Literacy”, Duncan Paterson, Ph.D Candidate of University of Heidelberg, discusses the hidden assumptions about the epistemological, practical, and aesthetic conditions of digital humanities. In addition, he introduces the three venues for digital humanities teaching in his institute: project training sessions, research seminars, and annual. He indicated that by establishing the standards and procedures for work on these projects which take place in specialized training sessions, the students are able to apply methods and incorporate digital texts to their own

materials for research and study. Peterson gives a great example to illustrate the practice and implementation of digital humanities education as digital humanities teaching, as an emerging profession, is still devoid of a widely shared consensus about methods and contents.

台灣數位人文教育的困難與展望

Difficulties and Prospect of Digital Humanity Education in Taiwan

祝平次*

Ping-tzu Chu*

摘　要

數位人文工作坊 (www.tinyurl.com/dhintaiwan)設定的目標是讓學員學會課程中所介紹的數位工具。由於課程內容包含實作，所以現場氣氛較活潑。每一場工作坊也都安排有助教，可以現場馬上解決學員的問題，也能達成一定的學習效果。最重要的，課程也提供實做練習的資料，因此學員馬上可以了解數位工具的效果，。就此而言，工作坊的效果是具體可見的。然而，在工作坊結束之後，多少學員可以將課程中教授的數位工具運用到自己的研究或教學裡，則是另一回事。而且，光是工具的使用，很難讓學員進一步體會到數位人文方法做為一種研究手段的重要性。就此而言，一整學期的課程設計會較為理想，也較容易設定整體性的目標。

在《數位工具與文史研究》(tinyurl.com/dhtools/)這個整學期的課程裡，安排了兩位助教，教導不同的數位工具，從如何利用光學辨識製造數位文本、到利用微軟的套裝軟體 Word 和 Excel，再到利用文本分析、中國歷代人物傳記資訊庫、地理資訊系統和社會網絡分析軟體等工具。學期間，大部分同學的現場學習都能跟上進度，期末考試也是以測試工具操作為主。但學期課程結束後的追蹤卻顯示，這些研究所同學幾乎都沒有在自己的研究中應用這些數位工具。這突顯出來的是，數位工具的短暫學習無法配合長期研究目標的問題。這一方面是由於缺乏研究典範，所以這些工具能夠如何被具體運用於實際的研究計畫，可能還需要跨過一定的經驗及視野的門檻。另一方面，數位工具的使用，預設了資料量大的特性，這也使得初學者較為怯步。當然，資料量大也意味著必須在研究前取得較多的數位

* 國立清華大學中文學系副教授，Email: dh@ptc.cl.nthu.edu.tw。

* Associate Professor, Department of Chinese Literature, National Tsing Hua University. Email: dh@ptc.cl.nthu.edu.tw.

文本，這又突顯出數位文本來源的問題。數化高品質的文本，所需投入的時間精力都非常多，是單一研究者無法負擔的事。

總結而言，文史學科的同學學習數位工具的操作並沒有問題。是否要考慮進一步引進簡單的程式設計概念，則是另外一個問題。但就現階段而言，如果能組合好一套基本的數位工具來滿足研究的需要，程式設計的能力也非必要。就工具目標的設定，雖然過去有政治大學鄭文惠老師團隊的一些成果，但如何讓類似的成果能在不同的文史領域都可以出現，並且得到學界的適當評價，可能還有待研究者的努力。最後關於數位文本取得的問題，由於日本學者開發了 kanripo (www.kanripo.org)和法鼓山長期對於數位佛典的建置，應該在將來也可以獲得巨大的改善。數位人文研究的發展，需要大量文史學者的投入才可能成功，而這個過程，不只要仰賴於資訊學者的參與，更需要文史學界整體性環境的建置。

跨越範式：數位人文之人才培育及其多元挑戰
Cross-Paradigm：Talent Development and Its Multiple Challenges in the Digital Humanity Field

邱偉雲*

Wei-yun Chiu*

摘　要

本文以由政治大學中國文學系鄭文惠教授擔任總主持人，政治大學資訊科學系劉昭麟特聘教授，與政治大學統計學系余清祥教授擔任共同主持人之「觀念、事件、行動：中國近現代觀念形成與演變的數位人文研究」國科會整合型計畫（2013 年 10 月 1 日-2015 年 7 月 31 日）項下組成之「DHRG 政治大學數位人文研究群」為例，在第一個部分，揭示出研究群的課程設計、研究群成員之專業領域背景，研究群成員參與時的心得感想，以及參與研究群後的研究成果，也對研究群成員後續是否繼續進行數位人文研究的情況進行了追蹤。基於上述的基本資料，第二部分則分析了人文與資科、統計領域的研究群成員，在跨領域的課程學習過程中所遇到的困難，也提出一個在培育數位人文學新世代人才之際需聚焦商議的一個重點議題，即：究竟是需要一個兼通人文與數位方法的通才？還是培養人文與數位學者皆具備有與跨領域學者溝通的基本知識與對話能力？而在第三部分，則將討論聚焦在「DHRG 政治大學數位人文研究群」，在進行人才培育的過程中所發現的諸多問題。這些問題可以總結成三方面：一、人文與數位學人在問題提出與問題解決等兩個方面有觀念與方法邏輯上的差異，因此應當設法調和；二、就目前數位人文研究現狀來看，可以看見有些參與數位人文研究的年輕學人，時有熱情不足的現象，主要原因是背景知識不足，因此難以進入跨領域研究語境，建議政府能鼓勵大學在通識教育中為學生增設質化與量化研究方法的基礎課程，以增強跨領域對話與研究的基礎能力；三、目前在人才培育上，存在著後繼無力的現象，經過歸納後得出兩個主要原因：第一點是數位人文研究的探索性太強，致使參與人員容易回到自己熟悉的研究領域；第二點是就業市場的欠缺，導致參與人員無法完全投入在數位人文研究中。上述以「DHRG 政治大學數位人文研究群」為考察案例，進而提出的一些觀察與建議，或可提供未來有關單位在規劃數位人文學發展時作為參考。

* 湖北經濟學院新聞與傳播學院副教授，Email: brianacwu@163.com。

* Associate Professor, Hubei University of Economics. Email: brianacwu@163.com.

Matter Matters：
Cultural Heritage Objects and Digital Literacy

Duncan Paterson*

Abstract

Teaching an emerging profession means the absence of a widely shared consensus about methods and contents, which is substituted by a set of local practices. In response to these problems of a discipline in the making my home institute provides three venues for DH teaching: project training sessions, research seminars, and annual workshops. Because of our limited pool of programmers, our projects depend on the workforce of student assistants, postdocs, PhDs, etc. The instruction for work on these projects takes place in specialised training sessions. Research seminars, on the other hand, enable students to apply methods to their own materials and critically engage with DH literature. Our annual workshops week provides senior researchers and students from other departments with opportunities to engage with DH topics. I begin with three questions instructors need to answer when drafting their syllabi:

1. What objects / sources will I use?
2. What must students do to reach my teaching goals?
3. What makes some work better than others?

At the heart of these questions lie the hidden assumptions about the epistemological, practical, and aesthetic conditions of DH, in other words the transcendental conditions of DH and their effects on teaching.

What Objects / Sources Will I Use?

Comparing the selection of source materials in the Hachiman Digital Handscrolls Project (HDH) and Early Chinese Periodicals Online (ECPO) points towards a new role for cultural heritage institutions, with respect to digital projects. Despite repeated calls

* PhD Candidate, the Cluster of Excellence Asia and Europe in a Global Context in Heidelberg, Germany, Email: duncan.paterson@asia-europe.uni-heidelberg.de.

for a shift towards a database paradigm within the humanities, the document paradigm continues to thrive. Based on these projects I argue that the database paradigm's emphasis on distant reading style analysis fails to address the knowledge condition of many humanities questions. Humanists often deal with idiosyncratic sources unfit for quantitative analysis. The critical reflection on the epistemology of digital technologies, allows students to arrive at better ways of using the inherent scaleability and connectivity of digital artefacts. As a result we arrive at a document centred view of DH, that combines aspects of both paradigms.

What Must Students Do to Reach My Teaching Goals?

DH research requires practical skills for interacting with computers and software. I will discuss the role that markup languages in general, and X-technologies (xml, xslt, xquery) in particular, play in helping learners pose better research questions. Familiarity with prominent metadata standards is a core aspect of our DH teaching, and provides students with a skill-set that has direct research applications, for example in Modern Chinese Scientific Terminologies (WSC) or in their professional careers beyond academia. In the absence of widely accepted DH curricula and certification procedures, we need to define both the minimum level of digital literacy we can expect from humanities scholars, as well as the upper limits of where computer science (CS) departments fit in. Using examples from research seminars on book history and network science I will discuss administrative and practical hurdles that prevent the classrooms culture of learning-by-doing from being fully integrated in humanities study plans, such as lack of credits for acquiring programming skills, or discrimination of data publication as viable research result.

From special training sessions for student assistants to discussions over funding applications, successful communication between software engineers and traditional humanities scholars remains challenging. The collaborative workflows typical of DH projects are alien to traditional humanities teaching, which trains young scholars for the lonely task of writing manuscripts. To prepare students for participating in large DH projects we favour the use of collaborative workflows typical of software development. Besides the practical advantage of knowing how to use popular version-control software (e.g. GitHub, SVN, …), this also forces students to engage with aspects of software beyond the GUI, such as documentation manuals, the command line, and APIs. Tools for collaborative work thus help us to increase the digital literacy of our students, and to complete more challenging and interesting assignments as part of a team.

What Makes Some Work Better Than Others?

Lastly, both teachers and other members of the DH profession have to confront the question of what makes a work good or bad. In other words, we cannot escape aesthetic judgement, and its role for our understanding of the discipline. Aesthetics can differentiate a broad concept of DH from a narrow concept of computational humanities. In class and in project pitches, we often encounter ideas that are either highly interesting to humanists, or coders, but rarely to both. In my response to the third question, I argue that at the root of this situation are institutional biases of both humanities and (hard) science against art as a form of aesthetic expression. Even the approach of critical code studies (CCS) that applies literary categories , such as succinctness, voice, and creativity, to source code is too narrowly tied to textual metaphors.

New media offer students the ability to explore different forms of narratives. In my experience, it is beneficial to look at how non-textual narratives are treated in other disciplines, like musicology, art history, or performance studies. How we include design principles in our teaching, and which value we place on good design in our grading seems like the greatest challenge of DH teaching. On the other hand, it can also be one of the most exciting aspects for moving from STEM to STEAM values.

Based on student responses I will discuss some of the problems teachers and students encounter when using visual essays, video productions, and web-designs as study assignments. This problem continues to undermine DH teaching by devaluing the time and intellectual effort that goes into non-textual narratives. I suggest to use architects as a guiding metaphor for how DH might define itself, and adopt it’s teaching practices accordingly. Architects are trained to operate as part of a collaborative environment where engineering challenges, artisanal requirements, and artistic work come together. As a jack of all trades, and master of none, the architect’s main contribution lies in the conceptual thinking at the intersection of these various demands. In my view, the DH profession is in a similar position. The societal effects of modern computational tools require not only technical know-how, but also ethical and aesthetic deliberation. Successful training of students in DH produces researchers, curators, and designers capable to address such issues.

重中之重：文化資產與數位素養

Duncan Paterson*

摘　要

新興專業的教學意味著方法與內容上缺乏廣泛認同的共識，取而代之的是一套在地的實踐。為因應這門形成中的學科所生之問題，我們研究所提供了三種數位人文教育的場合：計畫培訓課程、專題討論及年度研習會。由於程式工程師有限，我們的計畫職必須依賴大量工讀生、博士後、博士等人力。而工作執行的準則與標準程序則在專門培訓課程中進行。另一方面，專題研討則讓學生能將方法應用於自己的材料中，並積極與數位人文文獻結合。而為期一週的年度研習會則提供其他系所的資深研究員和學生參與數位人文議題的機會。因此，在規劃教學大綱時需要考慮的三個問題：

1. 我要使用的物件/來源為何？

2. 學生要做什麼才能達成我的教學目標？

3. 是什麼讓有些作品比其他作品更好？

本報告將依序說明在這些問題中所潛藏的假設，這些假設關乎數位人文的認識論、實踐及美學條件，換言之，亦即數位人文的先驗條件及其教學上之效果。

* 德國海德堡全球情境亞歐卓越研究中心博士候選人，Email: duncan.paterson@asia-europe.uni-heidelberg.de。

Panel G

文本解讀的擴展、拆解和觀察

Restructuring and Visualizing Texts

Panel G

文本解讀的擴展、拆解和觀察

Restructuring and Visualizing Texts

主持人	唐牧群（國立臺灣大學圖書資訊學系教授） Muh-Chyun Tang (Professor of Department of Library and Information Science, National Taiwan University)
發表人	宋浩（藍星球資訊股份有限公司總經理特別助理） Hao Sung (Executive Assistant to General Manager of BluePlanet Data and Information Technology Inc.)
題目	ADEPT：自動化資料豐富程序 ADEPT：Automated Data-Enrichment Processing Technologies
發表人	林農堯（國立臺灣大學資訊網路與多媒體研究所博士候選人） Nung-yao Lin (Doctoral Candidate of Graduate Institute of Networking and Multimedia, National Taiwan University) 陳胤豪（國立臺灣大學歷史學研究所博士生） Yin-hoe Tan (Doctoral student of Department of History, National Taiwan University)
題目	《先秦諸子繫年》之數位設計與呈現 Digitized Presentation of ‘A Chronological Study of the Pre-Qin Philosophers’ by Qian Mu
發表人	趙叡（國立臺灣大學資訊工程學系碩士生） Jui Chao (Master Student of Department of Computer Science and Information Engineering, National Taiwan University) 謝于琳（國立臺灣大學資訊工程學系碩士生） Yu-lin Hsieh (Master Student of Department of Computer Science and Information Engineering, National Taiwan University)
題目	《春秋》三傳對讀系統 A Comparitive Reading System for the *Three Commentaries Of Chunqiu*

ADEPT：自動化資料豐富程序

宋浩*

摘　要

在數位典藏資料庫、數位圖書館、以及數位博物館的領域，詮釋資料的建立，經常是耗費最多人力時間成本的一項工作。同時，建立詮釋資料並不是一件簡單的工作，建立者需要對某個特定領域的知識有深入的了解，才能產出豐富、正確、精準的詮釋資料，進而詳實傳達數位資源的重要性。

正因為詮釋資料必須透過大量人力進行建置，因此在實務上經常採用「聯合目錄」的形式。亦即由原始資料典藏單位負責建立典藏物的詮釋資料，再提交至中央主管單位統一提供可整合檢索、瀏覽的介面。由原始資料典藏單位各別建立詮釋資料與數位化的過程稱為「分散建置」，而由中央整合並提供使用介面則稱為「集中管理」，這是綜合考量時間、人力、資源等因素後的平衡點。不過，聯合目錄形式資料庫所衍生的問題則是詮釋資料的填寫方式難以趨於一致，進而導致後續在瀏覽、檢索、與資料鏈結、交換上的困難。

本文提出一套資料前置處理的框架：ADEPT (Automated Data Enrichment Processing Technology)，目標是將符合都柏林核心集的輸入資料進行自動化的前置處理與豐富化。在 ADEPT 框架中包含了三個主要模組：驗證模組、正規化模組、專有名詞擷取模組。透過這些模組處理過的詮釋資料將趨向一致性、符合統一的格式，同時具備人、事、時、地、物等重要屬性資訊。除此之外，豐富化後的資料將更適合鏈結資料(linked data)，不但可與網際網路上的相關資料相互連結，更可讓詮釋資料進一步被加值利用，達到全民共享的目標。

關鍵字：ADEPT、數位典藏、數位人文、鏈結資料、資料正規化、專有名詞擷取

* 藍星球資訊股份有限公司總經理特別助理，Email: sung@blueplanet.com.tw。

ADEPT：Automated Data-Enrichment Processing Technologies

Hao Sung*

Abstract

Metadata, known as "data about data", is an important way to describe and utilize digital objects in digital archives, digital libraries, and digital museums. To present accurate, precise, and high-quality metadata is a critical task for the digital databases, and it requires not only a high cost of human resources, but also domain know-how.

Due to the labor-intensive nature of metadata construction, a model often employed in developing a large digital collection is to build different archives separately, then construct a central portal (such as a union catalog) for users to browse, search, and explore the entire collection. Although this model is effective in terms of time, manpower, and resources, it has some drawbacks. The main problem is inconsistency in the metadata constructed. This may be caused by misinterpretation of metadata attributes, different details when inputting data, or inadequate metadata format for interpreting specific data sets.

In this research, we propose ADEPT (Automated Data Enrichment Processing Technology), a framework for pre-processing data. ADEPT contains three primary modules: data verification, data normalization, and named-entity recognition. ADEPT aims to ensure data consistency and correctness, and increases data usability at the same time. Furthermore, the enriched metadata is more suitable for linked open data. By connecting related data, we can explore and share information and knowledge through the Web.

Keywords: ADEPT, digital archives, digital humanities, linked data, data normalization, terminology extraction

* Executive Assistant to General Manager, BluePlanet Data and Information Technology Inc. Email: sung@blueplanet.com.tw.

一、研究動機

在資訊化的時代中，數位資料庫系統具有不占空間、維護成本低、檢索迅速精準、增刪資料便捷等優點，無論是對於學者研究、或是一般民眾瀏覽皆是重要的資源。而數位資料庫系統的建置工作除了最基本的原始資料數位化、以及檔案永續保存外，還包括詮釋資料的建置，以及提供使用者瀏覽與檢索用的介面。然而，建立詮釋資料並不是一件簡單的工作，建立者需要對某個特定領域的知識有深入的了解，而維護者則需要了解如何讓眾多詮釋資料有著統一、一致的表達形式，同時還需要遵循詮釋資料的定義規範，考慮未來的互通性、交換性等建立原則，已經需要花費相當大的人力成本。因此，在建立詮釋資料初期再額外加上更多建置者所需要注意的規則與條件有實作上的難度。

目前在實務上詮釋資料的建立經常採用「聯合目錄」的形式，亦即由原始資料典藏單位對典藏物做詮釋資料的建立，再提交至中央整合單位統一提供一個整合檢索、瀏覽的介面。例如國家文化資料庫 00，便是由文化部轄下各機關單位、縣市文化單位負責建置各類型文化資料的數位資料庫，再將數位化成果提交至文化部中央整合，讓民眾可以透過這個整合性的資料庫一覽全國各地的文物典藏內容。由原始資料典藏單位分別建立詮釋資料與數位化的過程稱為「分散建置」，而由中央整合並提供使用介面則稱為「集中管理」，這個模式是時間、人力、資源等綜合因素下的平衡點。

為使分散建置的詮釋資料在集中管理的時候能批次匯入，在建立之初便需要先定義一個詮釋資料的格式，使各類詮釋資料得以互通，例如在數位圖書館中最被廣泛使用的「都柏林核心集」(Dublin-Core)0。但即便如此，在實際建立詮釋資料時，可能因為資料建置者對於物件本身的主觀認定不一致，像是顏色、形狀、觀察角度不同等原因，導致詮釋資料很難有著統一的填寫方式。進而使得詮釋資料在整合後，仍然看起來像多個分散的資料集、勉強放在同個資料庫中。統整過去多項研究 000000，分散建置的詮釋資料進而衍生的問題可歸納成：

- 資料重覆：同一筆詮釋資料可能在不同單位、不同資料集裡面重覆出現。
- 填寫格式不一致：因資料建立者的習慣不同，對於同樣一個事實的描述或有出入，如日期記為西元或民國紀年，地點則有縣市名、地標、或是詳細地址等差異。
- 缺少重要屬性：因資料建置者的主觀認定差異，可能有些重點屬性未被填寫，影響日後資料的可用性。
- 資料關聯性無法串聯：雖然各資料集的詮釋資料是由不同單位建立，但資料與資料間卻可能存在關聯性，分散建置難以依相關性做相互串聯。

本文中提出的資料處理流程框架稱為：自動化資料豐富處理技術 (Automated Data Enrichment Processing Technologies, ADEPT)，是為了解決分散建置的資料不一致、不完整的問題，並強化資料之間的關聯性、進而提高資料的可用性與可檢索性所設計。我將這個框架分成三個模組：

- 驗證模組：驗證模組的主要目的在於確認輸入的詮釋資料是有效的，其中不存在重覆的資料、不存在空白的資料、也不存在其他被定義為無效的資料。在流程的一開始若能確定資料的有效性，便能在接下來的模組中減少誤差的狀況。在資料驗證的技術上，有 LCS 演算法(Longest Common Subsequence)、正規表示法(Regular Expression)、文本雜湊法(Content Hashing)等現有的演算法或工具可以協助實作。
- 正規化模組：為了要讓不同人在不同地所分散建置的詮釋資料格式趨向一致，正規化模組中設計了時間正規化、空間正規化兩個子模組。基於時間維度與空間維度於檢索、瀏覽、排序、群組上的重要性，在正規化模組中將時間資訊一致化、地理空間資訊的標準化，以利後續模組中的再利用，及終端使用者的使用。
- 專有名詞擷取模組：考量終端使用者在使用詮釋資料時的使用特性，在專有名詞擷取模組中，透過自動化的程序將詮釋資料中的人、事、時、地、物等重要資訊取出並整理，以期每一筆詮釋資料中的專有名詞都能被取出，以強化資料的可用性與檢索性。

在經由前三個模組所驗證、正規化、專有名詞擷取後的詮釋資料，趨向一致性、符合統一的格式，同時具備人事時地物等重要資訊，比起原始資料更適合交換或開放，透過 Open API (Application Interface)、鏈結資料(Linked Data)的技術，可以讓詮釋資料得到更進一步的加值與利用，達到全民共享的目標。

二、研究方法

ADEPT 的輸入格式為 Dublin-Core 相容的 XML 資料。在具備正規化資料庫的情況下可以將資料正規化；在具備知識本體資料庫的情況下，能擷取出資料中的知識本體；輸出格式同為 Dublin-Core 相容的 XML 結構化資料，除原本輸入端資料不變之外，另附加上可被機器閱讀並具本體知識的輸出資料。

為確保 ADEPT 保有最大的彈性，在設計時我採用了模組化的設計。亦即每一段的資料處理工作可以視為單獨工作的處理程序。在處理流程的一開始，第一個模組首先要

確認輸入的資料是否符合 ADEPT 的輸入規則，文件需符合 Dublin-Core 相容的 XML 結構化資料，若經過格式驗證發現有誤，則流程停止，經過修正後再重新開始流程。若格式符合 ADEPT 的輸入規則，則進行必要欄位的驗證，依照必備欄位規則的設定，可以在這個流程中檢驗是否必備的資訊都齊全，若有缺少必須的資訊，則流程停止，經過修正後再重新開始流程。最後的驗證項目是偵測是否存在雜訊資料、重覆資料、無效的資料。藉由雜訊資料規則的比對，確認能進到後續流程的資料都是有效的資料。

第二個模組的工作是將資料正規化，目的是讓資料具備一致的格式，也確保資料的可被機器讀取性、資料間交換能力。第三個模組的工作是擷取出原始資料中的專有名詞。此處的專有名詞指的是人物名稱、團體名稱、其他關鍵字等。原始資料在建立時通常具有創作者、所屬團體／組織、創作、作品類型、收藏者、著作人等重要的資訊。利用知識本體資料庫的比對，以及詞夾子演算法的探索，可以從中得知原始資料中所提到的關聯人物、關聯團體、相關的事件背景、物件的分類等重要資訊。因此若能將這些資訊擷取出來，並補充到原始資料中，是 ADEPT 中的豐富化功能，提高資料的可利用性、可被檢索性、後分類、詞頻統計、脈絡分析等更進一步的使用與研究。

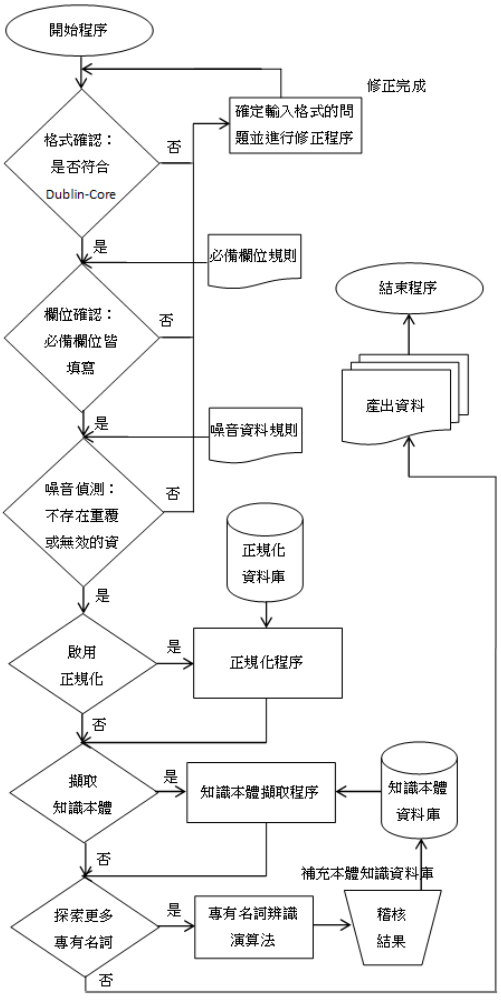


Figure 1. ADEPT 程序執行流程

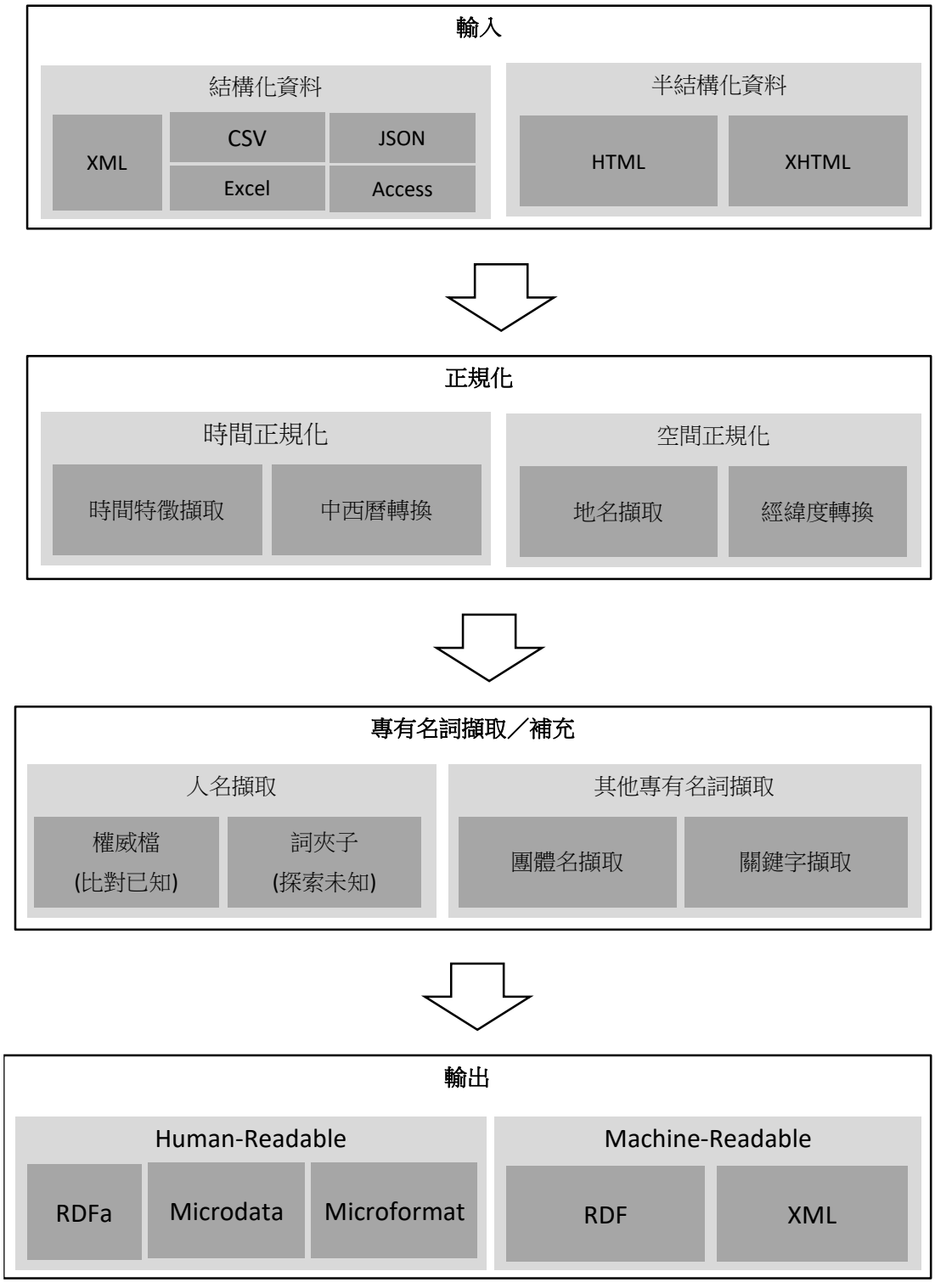


Figure 2. ADEPT 的處理模組

常見的數位化資料建置方式，是以結構化資料如 Microsoft Excel、XML(eXtensible Markup Language)、CSV(Comma-Separated Values)、JSON(JavaScript Object Notation)等

檔案格式或是關聯式資料庫相容格式如 SQL(Structured Query Language)、MDB(Microsoft DataBase file format for Access)。半結構化資料如 HTML(HyperText Markup Language)所建立。ADEPT 接受的輸入資料格式為 XML 格式的 Dublin-Core 及其延伸格式。在此將說明結構化資料和半結構化資料在轉換成 ADEPT 統一輸入格式時採取的策略。

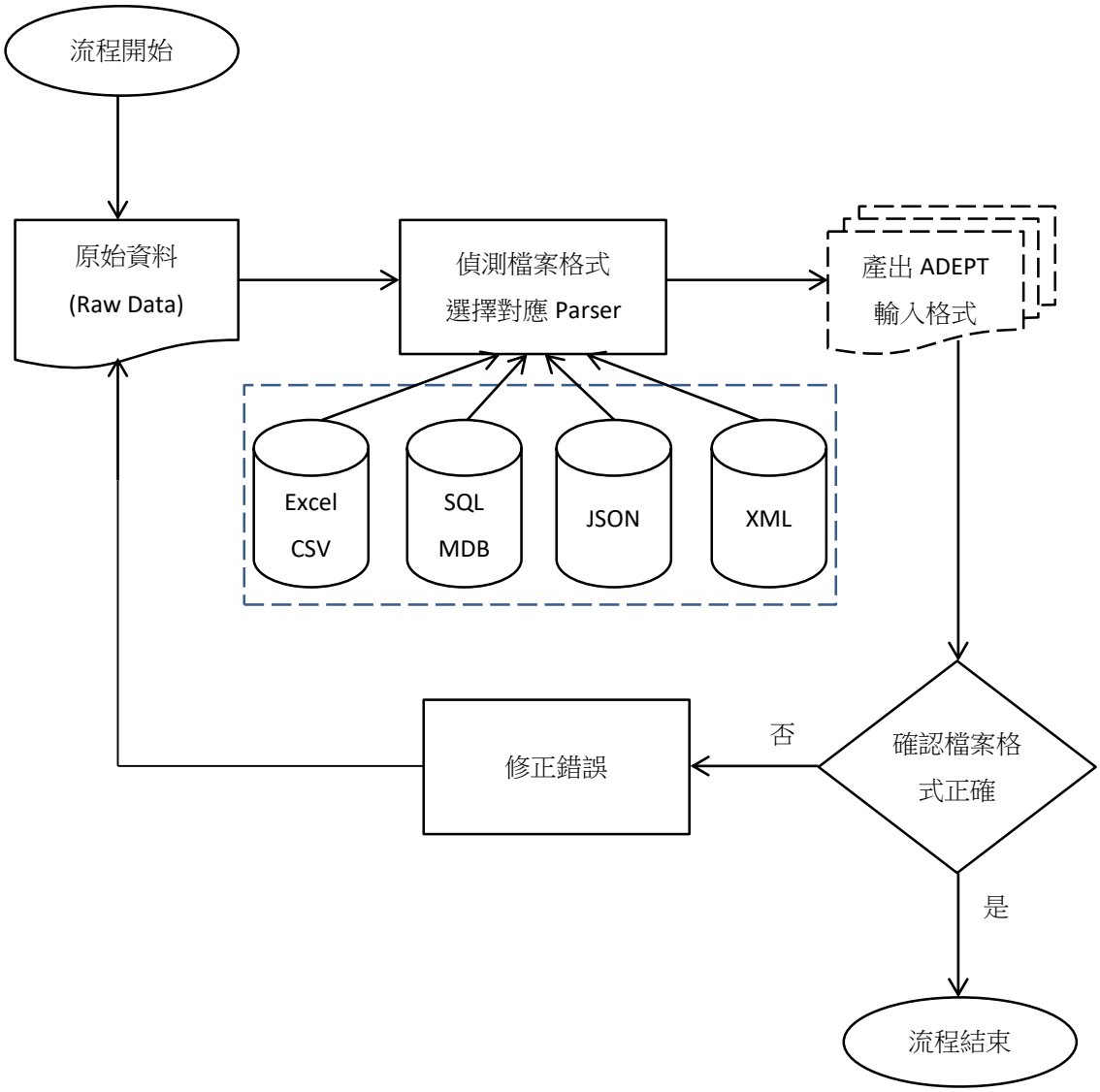


Figure 3. 結構化資料的處理流程

結構化資料(Structured Data)常見的視覺表達方式是二維的表格。X 軸註明了每個欄位(Field/Column)的定義，而 Y 軸則是依照這些定義所填寫成的資料。結構化資料因為本身已經具備欄位結構，因此在處理上只需要 Mapping 至 Dublin-Core 格式，並利用 Parser 轉換成統一的 XML，最後經由輸入格式檢查程式驗證即可。目前 ADEPT 已經實作了 Excel/CSV、JSON、SQL/MDB、XML 等四種最常見的結構化資料格式 Parser。依

照格式轉換所需，可以實作所需要的 Parser 於輸入模組中，以保留完全的彈性。

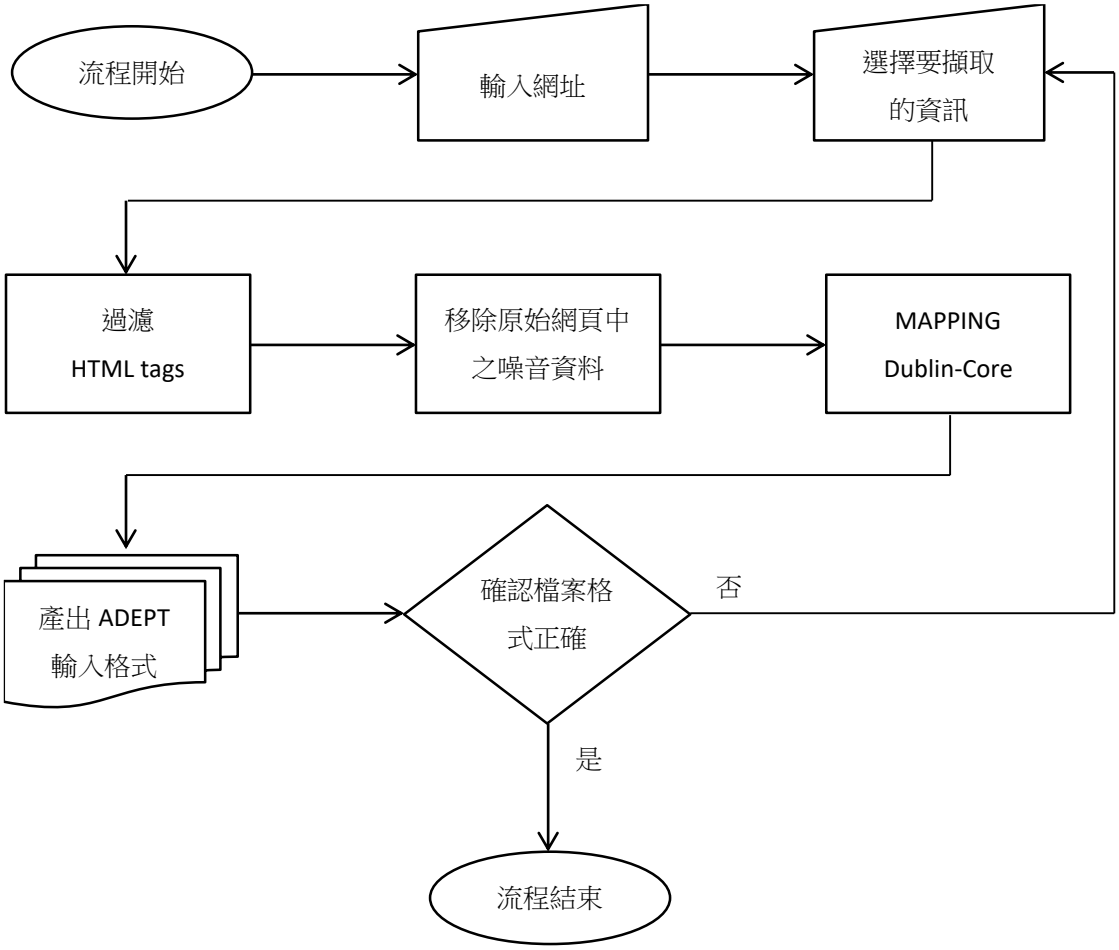


Figure 4. 半結構化資料的處理流程

半結構化資料(Semi-Structured Data)則是具有一部份詮釋資料、一部份全文資料的混合形式資料，最常見的半結構資料是 HTML 格式(HyperText Markup Language, 超文字標記語言)。半結構化資料不像結構化資料般的嚴謹，但仍然可從中抽取出許多重要的資訊。在 WWW 全面來臨的時代，有許多重要的數位資料是由 HTML 的格式所建立的，因此在 ADEPT 的系統架構中設計了專門為半結構化資料，亦即 HTML 的處理流程。目的是能自動的將半結構化資料所構成的網站也能夠分析出其中的結構化，並進一步的將資料正規化、自動擷取出人事時地物等關鍵資訊，讓網站形式的資料也可以受到完整的典藏與應用。

在 ADEPT 中的第二個步驟為正規化流程。此處的「正規化」指的是將資料的內容做某種統一化、一致化的處理。原因是數化位的資料通常由超過一位以上的建置者建立。可能是多個不同的單位在不同時間、不同地點、不同的計畫、不同的專案下所建立。因此，在不同時空環境下所填入的資料內容或有不同的觀點、不同的填寫方式、對欄位有不同的重視性等環境因素。這使得資料之間的表達方法有所落差，無論對使用者或是

系統技術本身，在檢索、排序、瀏覽或是後分類上來說都會是一個困擾。舉例來說，同一天的日期「1988 年 2 月 27 日」便可能有數種表達方式：「1988/2/27」、「民國七十七年二月二十七日」、「昭和 63 年 2 月 27 號」、「西元 1998 年 2 月 27」。

實作上利用正規表示法(Regular Expression)技術，可以找出在詮釋資料中是否存在時間 Pattern。若是存在時間 Pattern，則使用中西曆轉換資料庫 0 中提供的年代對應表做比對，轉換成一致的 YYYY-MM-DD (四位數年-兩位數月-兩位數日)。若原始資料為一個時間區間，如「乾隆八年」，則轉換成 YYYY-MM-DD(start) ~ YYYY-MM-DD(end)。

- 西式紀元
  - C.E. (Current Era, Common Era, Christian Era), 西元
  - B.C.E. (Before C.E.), 西元前
  - A.D. (Anno Domini), 拉丁文西元
  - B.C. (Before Christ), 西元前
- Patterns
  - a.d.,A.D.,c.e.,C.E.,ac,ce,AD,CE,西元,公元,西曆
  - b.c.,B.C.,b.c.e.,B.C.E.,bc,bce,BC,BCE,公元前,西元前
  - 00〇零1122334455667788 99

Figure 5. 西式紀元方式及正規表示法中允許的 Patterns 表達方式

- 中式紀元
  - 朝代、時期、皇帝名稱、年號、天干地支、節氣
- Patterns
  - 明,清,太平天國,日本,滿洲國(大滿洲帝國),中華民國,民國,民,日治,日治時期,日治時代,日據,日據時期,日據時代
  - 太祖,惠帝,成祖,仁宗,宣宗,英宗,景帝,憲宗,孝宗,武宗,世宗,穆宗,神宗,光宗,熹宗,思宗,福王,唐王,桂王,太宗,世祖,聖祖,高宗,文宗,德宗,末帝,昭宗,順帝
  - 洪武,建文,永樂,洪熙,宣德,正統,景泰,天順,成化,弘治,正德,嘉靖,隆慶,萬曆,泰昌,天啟,崇禎,隆武,永曆,天命,天聰,崇德,順治,康熙,雍正,[乾乾]隆,嘉慶,道光,咸[豐豐],祺祥,同治,光[緒緒],宣統,順天,天運,弘光,紹武,明治,大正,昭和,大同,康德,宣光,至正
  - 甲乙丙丁戊己庚辛壬癸、子丑寅卯辰巳午未申酉戌亥
  - 立春,雨水,驚蟄,春分,清明,穀雨,立夏,小滿,芒種,夏至,小暑,大暑,立秋,處暑,白露,秋分,寒露,霜降,立冬,小雪,大雪
  - 一壹壱弌元二貳弍三參叁參弐四肆五伍六陸七柒八捌九玖十拾廿卅〇零百佰00112233445566778899

Figure 6. 中式紀元方式及正規表示法中允許的 Patterns 表達方式

參照鄭鶴聲《近世中西史日對照表》0、郭廷以《太平天國曆法考訂》0 中的中曆年號與西元年對照，可以定義出資料庫的對照結構，如下圖 Figure 7、Figure 8 所示。

dynasty	emperor	empire	year	ganzhi	yearStart	yearEnd
清	世宗	雍正	2	甲辰	1724-01-26	1725-02-12
清	世宗	雍正	3	乙巳	1725-02-13	1726-02-01
清	世宗	雍正	4	丙午	1726-02-02	1727-01-21
清	世宗	雍正	5	丁未	1727-01-22	1728-02-09
清	世宗	雍正	6	戊申	1728-02-10	1729-01-28
清	世宗	雍正	7	己酉	1729-01-29	1730-02-16
清	世宗	雍正	8	庚戌	1730-02-17	1731-02-06
清	世宗	雍正	9	辛亥	1731-02-07	1732-01-26
清	世宗	雍正	10	壬子	1732-01-27	1733-02-13

Figure 7. 中曆朝代、帝號、年號與西曆轉換對應表 I

date 西曆日期	dynasty 朝代	emperor 帝號	empire 年號	empireYear 年	ganzhiYear 干支年	month 月	day 日	ganzhiDay 干支日	solar 節氣
17230605	清	世宗	雍正	1	癸卯	5	3	辛巳	
17230606	清	世宗	雍正	1	癸卯	5	4	壬午	芒種
17230607	清	世宗	雍正	1	癸卯	5	5	癸未	
17230608	清	世宗	雍正	1	癸卯	5	6	甲申	
17230609	清	世宗	雍正	1	癸卯	5	7	乙酉	
17230610	清	世宗	雍正	1	癸卯	5	8	丙戌	
17230611	清	世宗	雍正	1	癸卯	5	9	丁亥	
17230612	清	世宗	雍正	1	癸卯	5	10	戊子	
17230613	清	世宗	雍正	1	癸卯	5	11	己丑	
17230614	清	世宗	雍正	1	癸卯	5	12	庚寅	

Figure 8. 中曆朝代、帝號、年號與西曆轉換對應表 II

空間正規化的基本概念是透過地理座標系統的定位方式，將地標的經度與緯度擷取出來，將方便人類閱讀的地標名稱，轉換成機器可閱讀的座標資訊。空間正規化需要地標資料庫與經緯度轉換程式協助。這次實作上採用的地標資料(Position Of Interest, POI)來自於中華民國交通部運輸研究所所提供的公開資料，共計 22,651 筆 0。資料格式說明如下：

欄位名稱	說明
Landmark, 地標名	地標的名稱，例如觀光景點、學校、政府機關
Address, 地址	地址含鄉鎮縣市與道路名稱，部份具郵遞區號
Country, 國家	該地標所在的國家名稱
City, 城市	該地標所在的城市名稱
Lat, 緯度	地標的緯度，採度與小數制，正數為北緯；負數為南緯
Lon, 經度	地標的經度，採度與小數制，正數為東經；負數為西經
Acc, 精確度	地標的精確程度，1-9 共 9 個層級

Figure 9. 交通部運輸研究所 POI 的資料格式說明

精確度	說明
1	地標為國名層級，如日本、美國
2	地標為州、省等第一級的行政區域
3	地標為自治區、郡等第二級的行政區域
4	地標為鄉、鎮、縣、市等第三級的行政區域
5	地標為單一郵遞區號內的行政區域
6	地標為路名、街名層級
7	地標為兩條以上的街道構成的交叉路口
8	地標具有完整的地址
9	地標為明確的建築物、紀念館、購物中心、學校等單位名稱

Figure 10. 精確度的層級說明表

詮釋資料的內容中，往往包含了許多專有名詞在其中。舉凡人名、團體名、重大事件、特定領域的專有名詞等等。這些專有名詞不但經常是終端使用者檢索時的關鍵字，更是對資料做前分類、後分類、群組、聚集、分類的重要依據。因此在詮釋資料的處理過程中應當進行專有名詞的擷取與補充。然而，中文詞集是一個開放集合，在目前現階段的技術，並不存在任何一個可以將所有的中文專有名詞皆成功的辨識出、擷取出。ADEPT 採用準確率(precision)最高的詞庫為主，輔以準確率與回收率(recall)達到一個均衡的詞夾子演算法來探索更多新的專有名詞，成功達到專有名稱補充的目標。

Algorithm: 詞夾子演算法

1. 給定一個已知專有名詞集合 $P = \{P_1, P_2, P_3, \dots P_n\}$, $p \in P$

   給定一個已知非專有名詞集合 $\acute{P} = \{\acute{P}_1, \acute{P}_2, \acute{P}_3, \dots \acute{P}_n\}$

2. 輸出：P_{dic}為已知詞, $P_{candidate}$為候選詞，P_{clip} 為詞夾子

3. 對每個 Metadata 屬性 M 比對是否完全相符 $p \in P$

   若是，則將該名詞加入 P_{dic}

4. 對全文屬性 S 比對是否部份相符合 $p \in P$

   e.g. $(Clip_{prefix})\ p\ (Clip_{suffix})$

   若有，則將 $Clip_{prefix}$, $Clip_{suffix}$ 加入 P_{clip}

5. 以已知詞夾子 P_{clip} 對全文屬性 S 做比對

   若找到符合詞夾子的對象$(Clip_{prefix})P_{unknown}(Clip_{suffix})$

   且 $P_{unknown} \notin p$, $P_{unknown} \notin \acute{p}$

   則將 $P_{unknown}$ 加入 $P_{candidate}$

6. 驗證 $P_{candidate}$，確定為專有名詞則補充至 P 中
7. 是否尋找更多可能的夾子？
   若是，回到 4.
   若否，則專有名詞擷取停止

Figure 11. 詞夾子演算法

三、系統實作

ADEPT 使用者操作系統（ADEPT UI）採用 Web 技術實作，技術面在使用者端採用了 HTML5／CSS3／JavaScript／Bootstrap 等技術，在伺服器端則採用 WAMP 架構（Windows Server／Apache HTTP Server／MySQL Database Server／PHP Hypertext Preprocessor）。本研究以 Web 技術實作的優勢在於，可以讓多人在不同的時間、地點同時操作系統，以符合分散建置、集中管理的前置資料處理流程的系統使用方式。

Figure 12 展示了 ADEPT UI 的基本操作流程。不同單位的資料來源可由單一或多個不同使用者由 ADEPT UI 匯入，並選擇要執行的模組，包括正規化模組中的「時間正規化」與「空間正規化」，以及關鍵字擷取模組中的「人名關鍵字擷取」、「團體名關鍵字擷取」、「其他關鍵字擷取」等。若有其他自訂辭庫的話，也可以在最後選擇執行自訂辭庫內的關鍵字擷取。

待自動擷取程序完成後，系統將轉入驗證介面主控台，如圖 Figure 13，提供使用者檢視擷取結果。為方便使用者檢視，驗證介面將以後分類方式，搭配標註辭典中的擷取標籤和自動擷取出的擷取標籤，協助使用者篩選擷取結果的資料，並依照實際資料的狀況，選擇修正資料或保留擷取結果。

最後將完成驗證的擷取結果依照需求匯出成 XML、XML/RDFa、XML/Microdata 或是 XML/Microformats 等格式，做為匯入資料庫系統或 Linked Open Data 環境中的資料交換格式。

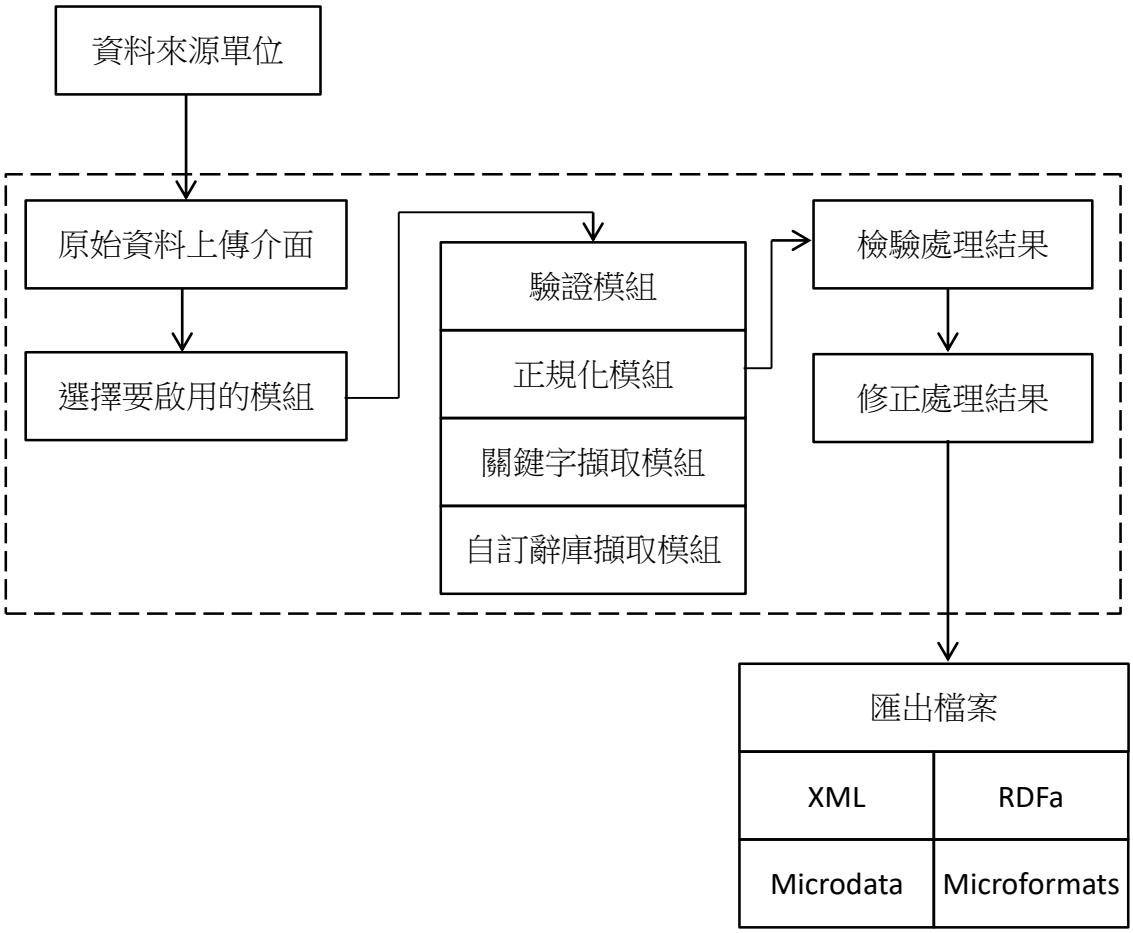


Figure 12. ADEPT UI 操作流程

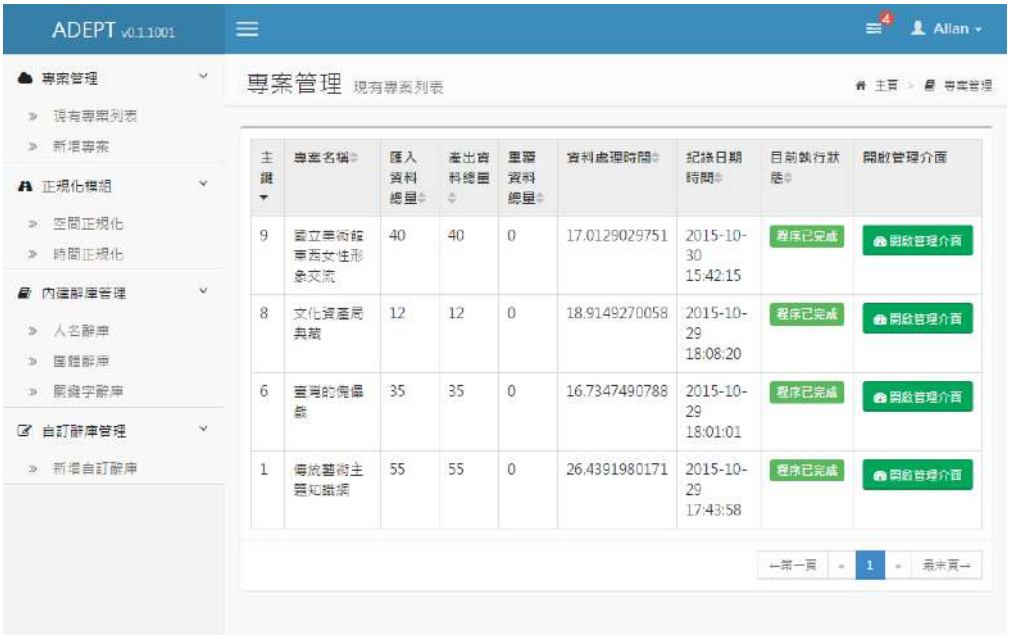


Figure 13. ADEPT UI 專案管理主控台畫面

圖

Figure 14 示範的資料為文化部於 2013 年起進行的文化資源庫計畫中來自國立傳統藝術中心製作的「傳統藝術主題知識網」的詮釋資料，其不但具備 Dublin-core 的 15 個核心欄位，包括 title、subject、type、creator 與 contributor 等欄位亦皆具備多重值型態。

圖 Figure 15 示範的另一個例子為文化部文化資產局典藏系統中的詮釋資料。此例中的詮釋資料較為精簡，僅有 Dublin-Core 中的五個欄位：identifier、date、title、description 與 subject，但卻具有重要的「時間資訊」1989 年、「空間資訊」宜蘭縣頭城鎮、「人物名資訊」林讚成、「團體名資訊」新福軒，以及「其他關鍵字資訊」開廟等。

```
<?xml version="1.0" encoding="utf-8"?>
<metadata>
    <identifier>M0000026</identifier>
    <subject>音樂</subject>
    <type>祭祀音樂</type>
    <date>2005/11/2</date>
    <title>巫師治病歌(一)</title>
    <title>misalisunalaliucikawasaymisalisin</title>
    <subject>阿美族</subject>
    <subject>儀式音樂</subject>
    <subject>巫師治病歌</subject>
    <description>儀式音樂臨時祭儀</description>
    <language>阿美語</language>
    <type>A原住民音樂</type>
    <type>A01阿美族音樂</type>
    <creator>台東馬蘭村阿美族</creator>
    <creator>吳榮順</creator>
    <contributor>郭英男</contributor>
    <contributor>林正鳳</contributor>
    <contributor>郭秀英</contributor>
    <contributor>吳瑞榮</contributor>
    <contributor>林正金</contributor>
    <contributor>郭秀珠</contributor>
    <description>巫師在阿美部落是很普遍存在的一種古老的制度。台東海岸一帶稱為
    cikawasay，卑南群尤其是馬蘭一帶都稱misalisin，阿美族宗教信仰（nokawasan）和祭
    祀儀禮（odmaknolioin）都需透過巫師的制度。在阿美族部落，巫師的職權可分三類
    ：（1）misalisin巫醫替人治病（2）mamaagag在一些祭儀中擔任祭司（3）piaraaw受戒
    養身。misalisin經過相當長時間的修練之後，就能替人醫治各種疾病。通常在治病時，
    巫師是有一個巫師群在老巫師備妥了地瓜葉、小米酒、行酒杯、酒甕等法器，就在他一
    面搖著竹枝施行法術下，飲唱起這首misalisinalaliu，其他巫師群起呼應，他們堅信除
    了這些法器之外，misalisin合得不順暢，病人是會病情加劇的。</description>
    <publisher>風潮有聲出版有限公司</publisher>
    <creator>吳榮順</creator>
    <relation>「台灣原住民音樂資料蒐集暨數數位化計畫」（第一期）</relation>
    <format>阿美族音樂簡介  PDF文件檔案  150KB巫師治病歌(一)  MP3聲音檔案  1.22MB
     1分20秒巫師治病歌(一)  WAV聲音檔案  40.2MB  3分59秒</format>
    <source>國立傳統藝術中心</source>
</metadata>
```

Figure 14. 一個國立傳統藝術中心 - 傳統藝術主題知識網中的詮釋資料範例

```xml
<?xml version="1.0" encoding="utf-8"?>
<metadata>
    <identifier>vidcat11</identifier>
    <date>1989</date>
    <title>新福軒頭城大溪開廟1(198912)</title>
    <description>宜蘭縣頭城鎮大溪海邊的這一座寺廟，在1992年春曾經有過一次翻修後神明入座的開廟門儀式，並聘請頭城新福軒傀儡戲團的林讚成擔任跳鍾馗開廟門及演出。</description>
    <subject>戲劇</subject>
    <subject>新福軒</subject>
    <subject>林讚成</subject>
    <subject>頭城</subject>
    <subject>開廟</subject>
</metadata>
```

Figure 15. 一個文化部文化資產局 – 文化資產局典藏系統中的詮釋資料範例

在 ADEPT 流程開始時，須給定一專案名稱，如「傳統藝術主題知識網」、「臺灣的傀儡戲」、「文化資產局典藏」、「國立美術館東西女性形象交流」等專案／專題名稱。再選擇原始詮釋資料檔案，原始詮釋資料來源可以接受 XML / XLS(Microsoft Excel) / CSV / SQL / MDB (Microsoft Access) / JSON 等格式，主要以 XML 做為資料交換格式。在各別模組的選擇上，可區分為「正規化模組」的啟用與選擇、「內建辭庫模組」的啟用與選擇，以及「自訂辭庫模組」中所啟用的各個自訂辭庫、需要進行自動擷取的欄位、支援多重欄位選擇等。將各別模組功能勾選完畢後，即會進行 ADEPT 處理程序。

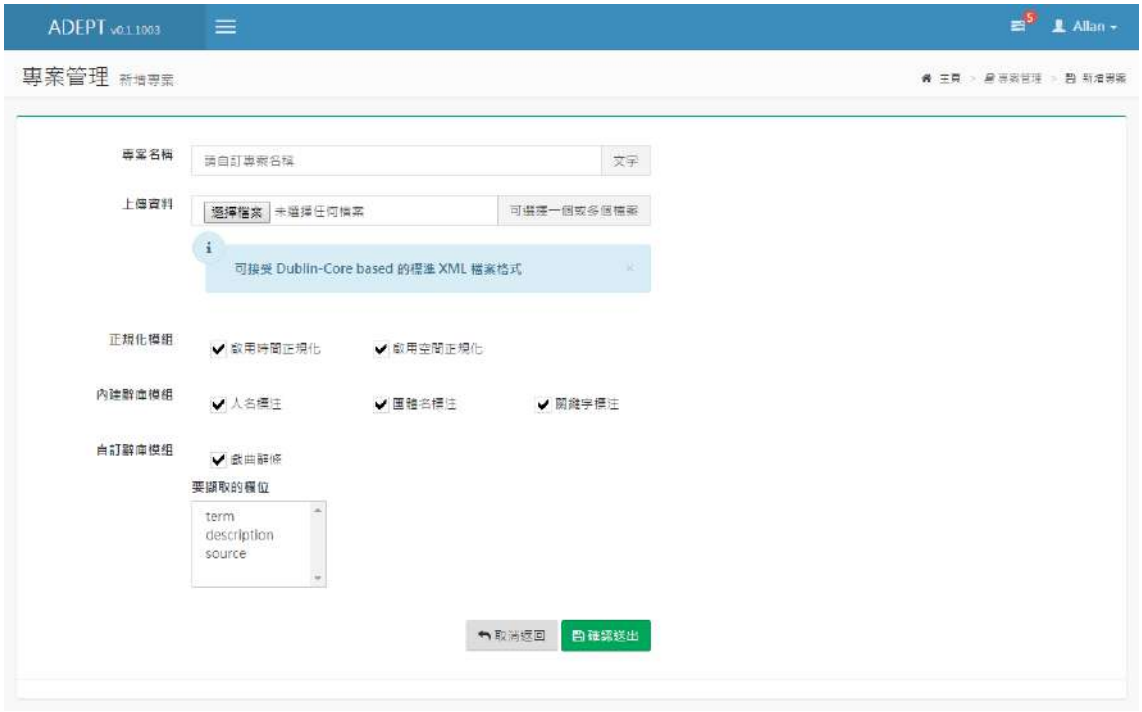


Figure 16. ADEPT UI 新增專案介面

處理程序完成後管理介面即可啟用，其目的是對每一批上傳的資料做統籌化的整理。該介面規劃成左右兩大區塊：左方區塊為匯入成果的後分類清單，並具備資料篩選功能；右方主畫面區塊上方規劃了資料匯入與產出的統計分析，以及匯出 XML 格式檔

案、匯出 RDFa 格式檔案、匯出 Microdata 格式檔案、匯出 Microformats 格式檔案四種方便於資料鏈結、資料交換的檔案格式；下方規劃成單一資料的管理，具有原始 XML 資料的瀏覽，以及各種專有名詞的擷取結果檢視與管理。

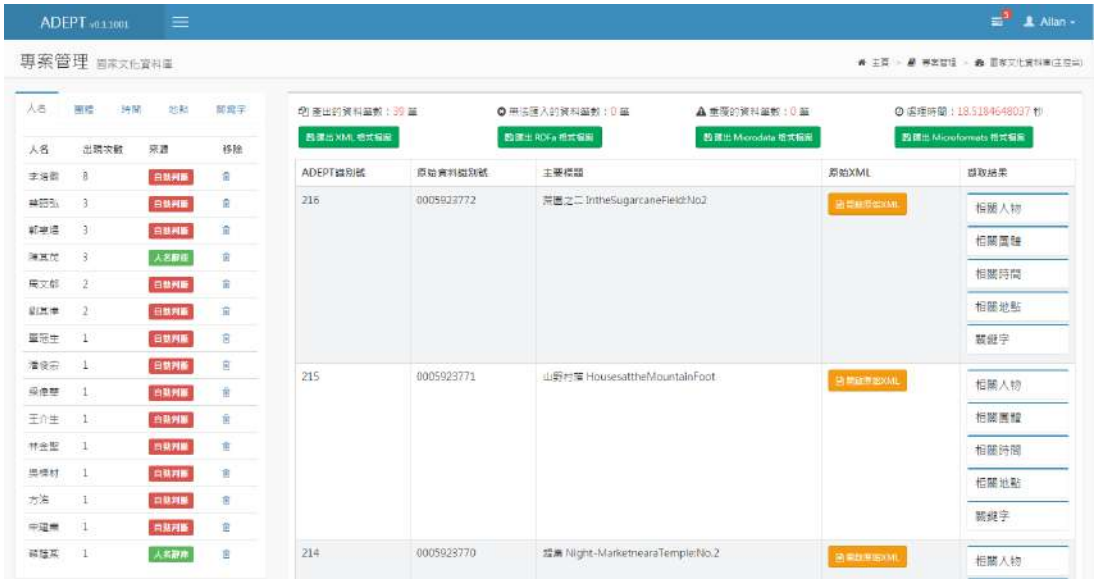


Figure 17. 管理介面主畫面

上圖 Figure 17 為資料驗證的管理介面主畫面。左方的後分類共計有人名、團體、時間、地點、關鍵字等五個頁籤，分別代表「人名擷取結果」、「團體名擷取結果」、「時間正規化結果」、「空間正規化結果」、「關鍵字擷取結果」。

在專有名詞擷取結果的頁籤中統計了該專有名詞出現的次數及來源，其中來源具有「來自辭典」或是由系統「自動判斷」兩種可能性。綠色的標籤「來自辭典」表示該專有名詞是從現有內建辭庫中，經子字串比對而擷取出，紅色的標籤「自動判斷」表示該專有名詞是符合詞夾子演算法的專有名詞 Pattern，判斷的專有名詞候選字詞。

Figure 18. 管理介面 – 驗證人名擷取之結果後分類

Figure 19. 管理介面 – 驗證團體名擷取之結果後分類

完成正規化的資料可後分類為時間與地點兩個頁籤。時間頁籤中的內容除了提供有正規化後的時間資訊，以西元（前）XXXX 年 XX 月 XX 日做為統一的表達格式之外，

並具備原始資料中所擷取出來的時間資訊及資料的數量；空間正規化後的結果呈現於地點頁籤中，具有地標名稱(POI, Position Of Interest)做為識別單元，包括十進位數字的經緯度（緯度,經度）、資料出現的次數，同時可以於數位化地圖中顯示，方便操作者檢視地點標示是否正確精準。

圖 Figure 20 呈現了時間正規化後的結果，即時間後分類；圖 Figure 21 中呈現了空間正規化後的結果，即地點後分類。另外在 Figure 22 為內建於空間後分類中的 Google MAP API 檢視器，以 Google Map 做為基底圖資呈現數位化地圖的服務，顯示其地標的名稱與經緯度資訊供操作者檢視。

Figure 20. 管理介面 – 驗證時間正規化之結果後分類

Figure 21. 管理介面 – 驗證空間正規化之結果後分類

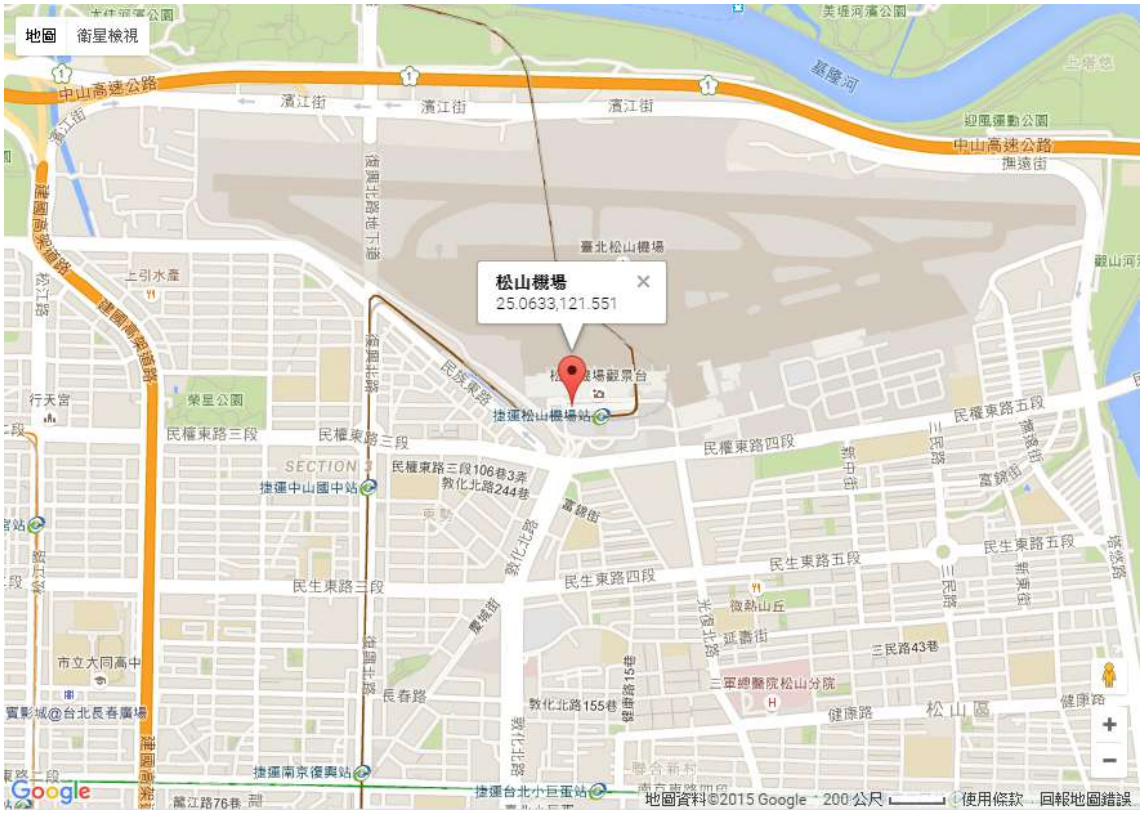


Figure 22. 內建於 ADEPT 中的 Google MAP API 檢視器

Figure 23. 管理介面 – 驗證關鍵字擷取之結果後分類

ADEPT 是一套將資料豐富化的流程，因此在經過格式驗證模組後，可確保資料格式正確、無雜訊資料及重覆性資料，當資料進入正規化模組的流程後，透過時間與空間

正規化的機制，使資料趨於一致性。而專有名詞的擷取流程則可將內建辭庫與自訂辭庫中的專有名詞擷取出來，並利用詞夾子與學習機制自動擷取其餘可能的專有名詞。最後經由操作者的驗證作業，刪去錯誤判斷的專有名詞後，資料即具備可輸出成容易使用、匯入、分享的檔案格式。在 ADEPT UI 的專案主控台所設計的資料匯出功能可依照需求的不同，匯出「XML 格式檔案」、「RDFa 格式檔案」、「Microdata 格式檔案」、「Microformats 格式檔案」等四個不同的檔案格式，以下以 XML 格式檔案為例說明。

圖 Figure 24 為完成資料匯出的範例：自第 25 行開始系統插入了一個<adept>的標籤，至第 46 行的</adept>結束。該例中具有一個不屬於人名辭典中的人名「吳榮順」在第 26 行被<people>標籤包含，但不具備 id 的元素屬性；具備一個存在於團體辭典中的團體名「國立傳統藝術中心」，因此被加上了 id=”1129” 的元素屬性，其中 1129 代表國立傳統藝術中心在團體名辭庫中的統一識別號；另外自 30 行～45 行具備了 16 個在關鍵字辭典中出現的關鍵字，皆分別加上了 id 的屬性使資料可與辭庫中的資料相互對應。

正規化資料的部份在第 28 行列出了時間正規化後的結果，原始資料中具備<date>2005/11/2</date>的時間資訊，在經過正規化後統一轉換成供機器閱讀的「+20051102」八位數字並包含西元前後紀年，並將較適合一般民眾使用的格式西元 2005 年 11 月 02 日於格式化輸入後列在標籤之中；空間正規化的資訊列在第 29 行，本例中具備一個被定義的地標「台東市南王村卑南族」，精準度為 5（精準層級為單一郵遞區號內的行政區域），緯度為 22.788903，經度為 121.11752。

```
1  <?xml version="1.0" encoding="utf-8" ?>
2  <metadata>
3      <record>
4          <identifier>M0000068</identifier>
5          <subject>音樂</subject>
6          <type>祭祀音樂</type>
7          <date>2005/11/2</date>
8          <title>獵隊凱旋歡迎歌</title>
9          <title>malikasau</title>
10         <subject>卑南族</subject>
11         <subject>南王村</subject>
12         <subject>儀式音樂</subject>
13         <subject>獵隊凱旋歡迎歌</subject>
14         <description>儀式音樂歲時祭儀</description>
15         <type>A原住民音樂</type>
16         <type>A06卑南族音樂</type>
17         <creator>台東市南王村卑南族</creator>
18         <creator>吳榮順</creator>
19         <contributor>台東市南王村卑南族婦女</contributor>
20         <description>這是一首獵人們從獵場回到部落後，婦女們在歡祝獵隊凱旋的聚會上所唱的不牽手而只歌唱的malikasau。歌
           使用男族人喜歡用的形式：不斷反覆一段特定旋律，而該段旋律前半段是用實詞演唱，後半段則改用無意義的虛詞母音演唱
           </description>
21         <publisher>風潮有聲出版有限公司</publisher>
22         <creator>吳榮順</creator>
23         <relation>「台灣原住民音樂資料蒐集暨數數位化計畫」（第一期）</relation>
24         <source>國立傳統藝術中心</source>
25         <adept>
26             <people>吳榮順</people>
27             <group id="1129">國立傳統藝術中心</group>
28             <datetime format="+20051102">西元2005年11月02日</datetime>
29             <location acc="5" lat="22.788903" lon="121.11482">台東市南王村卑南族</location>
30             <keyword>台東市南王村卑南族婦女</keyword>
31             <keyword id="5463">不牽手</keyword>
32             <keyword id="13836">台灣原住民</keyword>
33             <keyword id="24693">卑南族</keyword>
34             <keyword id="30939">南王村</keyword>
35             <keyword id="36061">原住民音樂</keyword>
36             <keyword id="45023">牽手</keyword>
37             <keyword id="45381">祭祀</keyword>
38             <keyword id="45400">祭儀</keyword>
39             <keyword id="45582">第一期</keyword>
40             <keyword id="46966">部落</keyword>
41             <keyword id="55959">歲時祭儀</keyword>
42             <keyword id="59705">歌唱</keyword>
43             <keyword id="69503">獵人</keyword>
44             <keyword id="69561">簡介</keyword>
45             <keyword id="72523">歡迎歌</keyword>
46         </adept>
47     </record>
48 </metadata>
```

Figure 24. ADETP 匯出格式 – XML with Dublin-Core/ADEPT

四、實驗結果

為檢驗 ADEPT 的資料處理效果，實驗資料以中華民國行政院文化建設委員會（現為文化部）於 2009-2010 的新版國家文化資料庫計畫，與中華民國行政院文化部於 2013-2014 的文化資源庫計畫所產出的數位化詮釋資料與數位物件成果做為實驗素材。國家文化資料庫中的詮釋資料由 175 個子計畫所建置，共約 100 萬筆資料；文化資源庫中的詮釋資料則由 187 個不同的網站或計畫分別建置，共約 160 萬筆資料。來自於如此多元的單位與計畫分散建置的資料，相當適合拿來做為 ADEPT 的實驗材料。

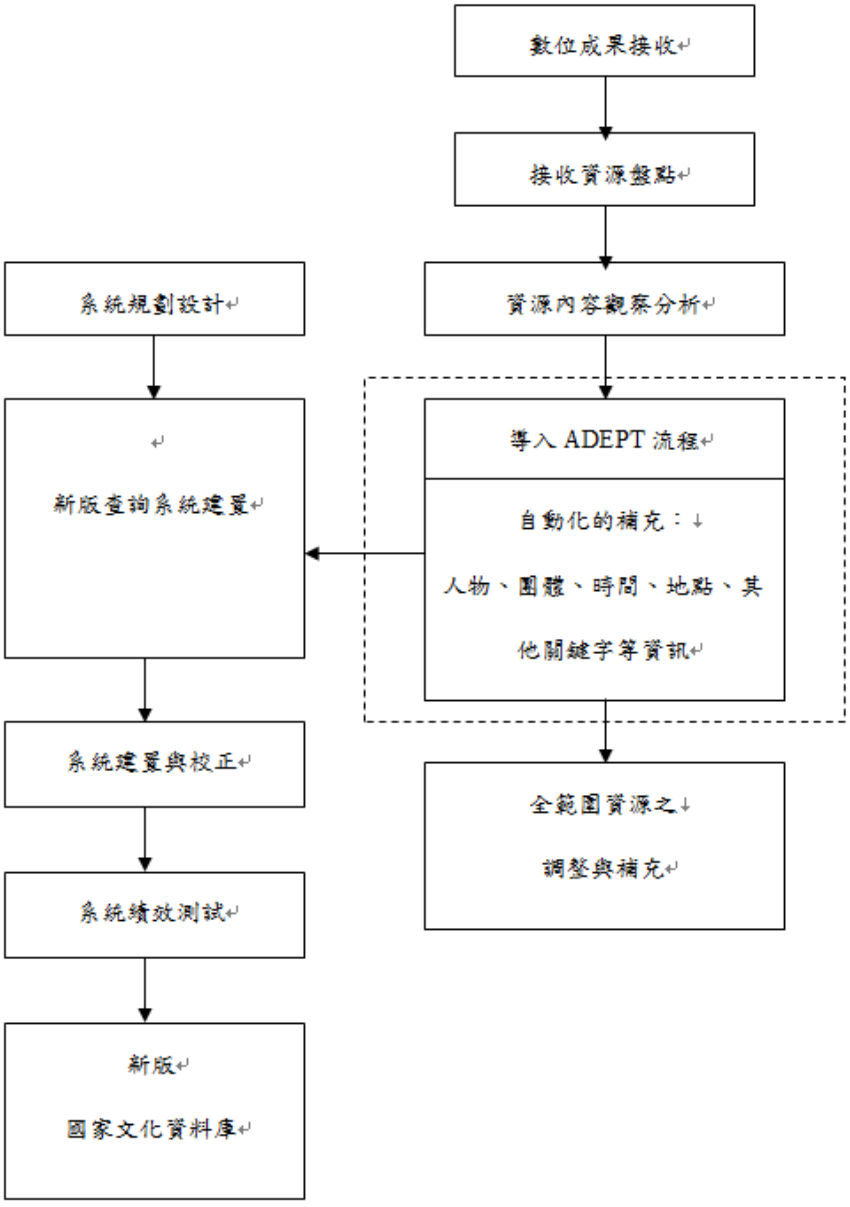


Figure 25. 新版國家文化資料庫計畫資料處理程序

在新版國家文化資料庫計畫中，ADEPT 處理的結果統計分析如下：時間正規化部份，依照時間 pattern 的偵測與正規表示法，成功轉換為 YYYYMMDD 的詮釋資料共 891,929 筆，約佔整體資料的 84% (891,929／1,064,213)；空間正規化部份，共有 674,175 筆詮釋資料擷取出具備地標資訊，約佔 63% (674,175／1,064,213)、具備超過一個以上地標的資料共 29,462 筆；人名專有名詞擷取部份，共計擷取出人名資料 655,388 筆，原人名權威已知資料 119,840 筆（共計 2,310 個人名，人名權威中共有 3,727 筆人名資料），以詞夾子演算法擷取出新人名 318,431 筆（29,584 個），共計有 438,271 筆詮釋資料具備人名，資料比例上約佔 41% (438,271／1,064,213)，而出現頻率最高的前五位人名資料為：中央社記者/組長陳永魁 (7,567)、作家柏楊 (7,286)、沙雕藝術家陳漢中 (6,605)、舞蹈家李天民 (6,958)與雕塑家楊英風 (3,556)，其中僅李天民為人名權威中已知人名；在團體名擷取部分，共計擷取出團體名資料 430,077 筆，原團體名權威已知資料 88,181 筆（共計 208 個團體，團體權威中共有 286 筆團體名資料），以詞夾子演算法擷取出新團體名 125,847 筆（3,450 個），共計有 214,028 筆詮釋資料具備團體名，資料比例上約佔 20%(214,028／1,064,213)，而出現頻率最高的前五組團體名資料為：國立傳統藝術中心(55,784)、臺灣總督府(9,340)、國立台灣美術館(5,343)、九歌出版社(3,531)、國家台灣文學館(2,550)；至於關鍵字擷取部份，共擷取出關鍵字資料 7,664,025 筆，共計有 913,352 筆詮釋資料具備關鍵字，資料比例上約佔 86%(913,352 / 1,064,213)，出現頻率最高的前五組關鍵字資料為：布袋戲(18,881)、北管(16,284)、版畫(9,929)、皮影戲(9,746)、崑曲(7,363)。

Figure 26. 新版國家文化資料庫的一般檢視模式

Figure 27. 新版國家文化資料庫的時間關係檢視模式

如前文所述，經由 ADEPT 流程處理過的資料會標示出正規化後的時間資訊、空間資訊、以及專有名詞的標註。針對上述的資料特性與豐富化後的成果，我設計了一個「三面向的檢索模型」，可將使用者所輸入的關鍵字檢索結果分成「一般條列式」、「時間軸關係」、「地理位置關係」等三個維度的呈現方式。使用者於檢索引擎中輸入搜尋關鍵字後，可將同一個關鍵字的檢索結果以一般的條列式方式檢視，如圖 Figure 26、或切換至時間軸關係檢視，如圖 Figure 27、或切換至地理位置關係檢視，如圖 Figure 28。

文化資源庫（http://nrch.culture.tw/）（由於文化部政策因素，網站實際名稱仍維持「國家文化資料庫」）自 2013 年 8 月上線至今僅三餘年，已有超過 250 萬瀏覽人次，較舊版上線五年僅 88 萬人次高出許多。文化資源庫以新版國家文化資料庫的首頁設計為基礎，即以搜尋引擎為主題，搭配近期一個月內的熱門搜尋關鍵字，如 Figure 269 所示。

Figure 28. 新版國家文化資料庫的地理位置檢視模式

Figure 29. 文化資源庫(文化部國家文化資料)首頁

本文中所介紹的實驗結果系統皆已上線，提供一般民眾更便利的檢索後結果分析，其總瀏覽人次與每月瀏覽次數充份說明新版國家文化資料庫與文化資源庫的可利用性較原先未經過 ADEPT 技術處理的版本擁有更高的瀏覽意願和重覆使用率。

五、結論與未來展望

分散建置、集中管理的環境已是目前數位化資料庫系統的主要建置模型方法之一。由於大量的詮釋資料是經由不同建置者、不同單位、不同計畫、不同專案所建立，因此在資料進入關聯式資料庫前，若未經過前置處理，將容易產生雜訊資料、格式不一致、重要欄位缺漏、資料難以關聯或共享等問題。因此 ADEPT 的目標是提供一套能自動化處理資料的前置處理程序，將正規化與專有名詞補充過的資料，結合三面向檢索模型，使詮釋資料與數位物件可在時間或空間為主軸的條件下，以不同面向、不同維度的方式進行檢視。

實驗結果顯示，ADEPT 流程前置處理過的詮釋資料、以三面向檢索模型建立而成的使用者介面，可以減少資料中的錯誤與雜訊、降低詮釋資料建置時期的人力成本、建立更多面向的檢索後分類、確保資料格式的一致性、補充缺漏的重要資訊，並以時間面向或空間面向為主軸檢視資料，將使用者操作介面的可使用性盡可能的提升，同時也便於將詮釋資料轉換成適合在網際網路上分享的鏈結資料格式。流量統計顯示，本文所提出的 ADEPT 對於兩套實作的系統成果，在瀏覽人次上都有顯著的提升。透過 ADEPT 程序，為數位典藏、數位人文或其他數位化資料庫系統的資料前置處理程序一致化、標準化。

ADEPT 的目標是設計給大量分散式建置的詮釋資料，做為資料前置處理的標準作業程序。在本次研究中我已經實作了 ADEPT 的操作系統，並將文建會與文化部兩個計畫中共計約 260 萬筆詮釋資料，經過 ADEPT 流程處理完成，且成功應用在已上線的資料庫系統網站運作。未來 ADEPT 的研究將以更跨領域、更全面性的詮釋資料與數位物件做為其能力的驗證。

參考文獻

DCMI Home: Dublin Core® Metadata Initiative (DCMI), http://dublincore.org/

文化部國家文化資料庫, http://nrch.culture.tw

交通部運輸研究所, http://www.iot.gov.tw/.

林信成、康珮熏。2005。 "建置以 OAI-PMH 為基礎的數位化新聞 Metadata 分散檢索系統", TANet2005。

洪淑芬、邱婉容。2005。 "國立臺灣大學圖書館數位典藏 Metadata 之設計與資料庫之建置彙整", 後設資料在數位典藏之研究發展：回顧與前瞻研討會論文集, pp.237-269。

郭廷以。1963。 太平天國曆法考訂。

張懷文。2004。 "詮釋資料與數位典藏長久保存取用－淺談 Preservation Metadata", 典藏國家型科技計畫電子通訊。

新版國家文化資料庫, http://newnrch.digital.ntu.edu.tw

陳昭珍。2002。"數位典藏計畫異質系統互通機制：以 OAI 建立聯合目錄的理論與實務", 國家圖書館館刊 91 年第 1 頁。

陳淑君、城菁汝、陳雪華。2013。"探索數位典藏的詮釋資料與索引典之多語化", 圖圖書資訊學刊第 5 卷第 2 期, pp.49-72。

國立臺灣大學資訊工程研究所數位典藏與自動推論實驗室, 國立臺灣大學數位人文研究中心, 中西曆轉換程式, http://140.112.30.230/datemap/.

鄭鶴聲。1994。 近世中西史日對照表。

賴忠勤。2007。"數位典藏建置規劃與管理", 佛教圖書館刊第 45 期。

《先秦諸子繫年》之數位設計與呈現

林農堯[*]、陳胤豪[**]

摘　要

《先秦諸子繫年》是中國歷史學家錢穆先生的重要著作之一，主要考證先秦諸子的生平事跡及其生卒年，藉以解決史料上的空闕及互歧問題。錢穆以古本《竹書紀年》為基礎訂正《史記．六國年表》之誤，初步整體性解決了戰國史編年缺乏可靠史料的問題。然而，《繫年》的研究成果卻長期以來未受學界重視。本文希望透過《繫年》文本的數位化處理，讓《繫年》的研究成果更便於利用。《繫年》之所以適用於數位化處理，是因為其論述的證據力乃建立在人物與事件在時空中緊密聯繫的關係上。交織的關係鏈索形成資訊相聯的文本脈絡，而數位人文的強處，正在於將這個複雜的文本脈絡加以分析及視覺化，使之更易於被利用。原書中的〈通表〉及〈索引〉，更為運用數位技術進行文本分析帶來極大的便利。數位化系統將初步嘗試整合網路資源，把《繫年》文本與線上的原始文獻作聯結，以方便學者參照上下文。進而，系統將利用書中的〈通表〉及〈索引〉建構《繫年》的文本脈絡，以突顯文本內部關鍵詞之間有意義的聯結，從而進行對比和分析。未來會把系統設置在數位人文平台 DocuSky 上，再整合開發大型可用以做比較及參照的工具，如：對照表、地理資訊系統、人物關係網絡等。藉由這些數位化工具的協助，建立可將先秦史料中之人、事、時、地、物與書，在檢索以後進一步分類與觀察的系統，為學者提供一個整合多種資源的簡易數位化環境，來進行文本分析的工作。

關鍵字：《繫年》、《紀年》、（史料的）可信度、（資訊的）相聯性、文本脈絡

[*] 國立臺灣大學資訊網路與多媒體研究所博士候選人，Email: nungyao@gmail.com。
[**] 國立臺灣大學歷史學研究所博士生，Email: astyh83@gmail.com。

Digitized Presentation of 'A Chronological Study of the Pre-Qin Philosophers' by Qian Mu

Nung-yao Lin*, Yin-hoe Tan**

Abstract

A Chronological Study of the Pre-Qin Philosophers, or *Xinian*, is an important work by the renowned Chinese historian Qian Mu. This monograph features an extensive and nearly exhaustive collection of information about Pre-Qin philosophers, starting from Confucius and ending with Li Si, spanning a period of three hundred years. Qian Mu meticulously and painstakingly separate facts from allegories of the biographic data of these philosophers – both prominent and obscure – that were extracted from the historic sources available at the time. His goal was to piece all the information together to form a seamless and uniform chronology. *Xinian* used the the *Bamboo Annals* (竹書紀年) to revise the chronology given in the *Shiji Chronology of the Six Warring States*(史記・六國年表). This was also an attempt to confirm the former book's authenticity, which was in doubt at that time. Despite its merits, however, *Xinian* has been underutilized in the academic circle. Through digitization, our work hopes to make Qian's efforts more accessible to scholars of early Chinese history and the general public. Since the validity of Qian's arguments rests mainly on the fact that all the information fit together as a whole, a digitized version of the *Xinian* should facilitate the visualization of the inherent informational connectivity that forms the crux of his study. We are aided in our work by the provision of comprehensive tables and indices that indicate the temporal, spatial, thematic, biographic and bibliographic web of relationships between the textual contents of the *argumentations* or *Kaobian* (考辨) - one hundred and sixty three in all - that form the main body of the study. The digitized system integrates network resources to facilitate the simultaneous display of the primary sources cited by Qian. The system would also establish connections between

* Doctoral candidate, Graduate Institute of Networking and Multimedia, National Taiwan University, Email: nungyao@gmail.com.
** Doctoral student, Department of History, National Taiwan University, Email: astyh83@gmail.com.

keywords found in Qian's argumentations with the help of the aforementioned tables and indices to map out the logical structure and contextual coherence of the work. The system will be hosted on the digital humanities platform DocuSky to integrate resources and provide a variety of digital tools for the comparison, referencing and analysis of texts. Users would eventually be able to sieve out vital information from historical texts through a search and present them in a comprehensive and meaningful way.

Keywords: *Xinian*, *Bamboo Annals*, reliability (of sources), inter-connectivity (of information), contextual coherence

《先秦諸子繫年》是中國歷史學家錢穆先生的重要作品之一，主要考證先秦諸子的生平事跡及其生卒年，藉以解決史料上的空闕及互歧性問題。錢穆以古本《竹書紀年》為基礎訂正《史記》之誤，初步整體性解決了戰國史編年缺乏可靠史料的問題。此書在方法上繼承了清儒的考據學，發前人所未發，具有重大的學術價值及意義，卻長期以來未受學界重視。《繫年》的文本特點在於〈考辨〉當中緊密相聯的人物與事件，以人物之間，以及事件，在時空中的緊密聯繫來證成史料的可信度。在這個意義上，《繫年》文本非常適用於數位化處理。經由關聯索鏈的視覺化來突顯個別關係之確立在整體中的意義。本文希望透過《繫年》文本的數位化處理，讓《繫年》的研究成果更便於利用，從而推進學術的整合與進步。

一、《繫年》的學術價值

中國第一部編年體歷史著作，是孔子據魯史改編的《春秋經》。《春秋經》始於魯隱公元年（前 722），止於哀公十四年（前 481），凡兩百四十二年。自此以往，每年發生的事，都有了歷史記錄。惟在《春秋》與《資治通鑑》（始周威烈王二十三年，前 403。）之間，尚有七八十年左右[1]的空闕。錢穆的《先秦諸子繫年》，正是為了審訂這七八十年的歷史而作。[2]

要填補《春秋》與《通鑑》之間，戰國紀年上的空闕，就必須利用《史記》的〈六國年表〉。該〈年表〉上承〈十二諸侯年表〉（前 841-前 477），依序記錄了周元王元年到秦二世三年（前 476-前 207）之間的史事。然而，由於六國史記為秦火所滅，〈年表〉只能根據《秦記》來書寫。問題是，《秦記》根據司馬遷的說法，既「不載日月，文略不具」；又因秦國自孝公以前，「僻在雍州，不與中國諸侯之會盟」，故忽於六國史事。〈年表序〉云：

> 秦既得意，燒天下詩書，諸侯史記尤甚，為其有所刺譏也。詩書所以復見者，多藏人家，而史記獨藏周室，以故滅。惜哉，惜哉！獨有秦記，又不載日月，其文略不具。然戰國之權變亦有可頗采者，何必上古？……余於是因秦記，踵春秋之後，起周元王，表六國時事，訖二世，凡二百七十年，著諸所聞興壞之端。後有君子，以覽觀焉。

因此，〈年表〉多疏漏，在所難免。《繫年》的學術價值，正在於利用《竹書紀年》以及諸子年世的相關資料等來糾正〈六國年表〉的錯誤，補充它的不足，從而建構出一個較

[1] 《左傳》的《春秋》經文止於哀公十六年（前 479）孔子卒歲，較《公羊》《穀梁》多兩年；《傳》文止於哀公二十七年（前 468），較《春秋經》多出十一至十三年。若依此計算則空白期為六十五年。
[2] 錢穆，《中国史学名著》（北京：生活・讀書・新知三聯书店，2000 年），頁 14。

可信的戰國編年史。

《竹書紀年》出土於西晉太康二年（281）。當時汲郡有人發魏襄王冢，得大批竹簡，整理者稱之為《竹書》。其中有《紀年》十三篇，按年依次記錄夏至戰國初期的史事。杜預〈左傳後序〉謂《紀年》：

> 起自夏殷周，皆三代王事，無諸國別。惟特記晉國，起自殤叔，次文侯、昭侯，以至曲沃莊伯。...晉國滅，獨記魏事。下至魏哀王之二十年（前 299）。蓋魏國之史記也。

自《紀年》出土以來，學者或用之以訂正古史。其中亦有用以校正《史記》者，如唐·司馬貞的《索隱》。然而，司馬貞並不信任《紀年》。〈燕世家索隱〉云：「紀年之書多譌謬，聊記異耳。」清代中期以來，學者就發現，《紀年》原書早已亡佚於兩宋之際。再加上《紀年》言三代事多與儒家舊說相違異，其記春秋時事又常化約《左傳》之文，與國史承告據實書者不同。因此，《紀年》的歷史價值，一直存有爭議。

道光年間，朱右曾（1838 年進士）因今本《紀年》「鼠璞溷淆，真贋錯」而將散見於古籍所引的《古文紀年》，掇拾成帙，輯成《汲冢紀年存真》一書，以為信據。民國初年，王國維又因朱書「尚未詳備，又諸書異同，亦未盡列，至其去取，亦不能無得失」而在《存真》的基礎上，更著《古本竹書紀年輯證》。又於民國六年（1917）仿惠棟《古文尚書考》之例，為《今本竹書紀年疏證》，將今本文句一一疏其出處。此書既出，在 20 年代「古史辨」疑古風潮的影響下，今本《紀年》為偽書之說遂在學界幾成定論。

《繫年》自 1923 年發意起草至 1935 年問世，正逢「古史辨」期間。錢穆先生基本上也認同今本《紀年》不可信的說法，卻有意證明古本《紀年》的史料價值。因為《紀年》是魏國史記，言戰國事，時代既近，自可信據。不應與三代及春秋時事之不可信據相提並論。再者，魏在戰國初年為東方霸主，握中國樞紐，《紀年》載秦孝公以前東方史實，亦當勝於〈六國年表〉，且可彌補〈年表〉所據《秦記》的不足。錢氏云：

> 夫《史記》之誤易見，捨《史記》而求是則難尋。《紀年》之佚文，散見於《集解》《索隱》諸家之注，以及《水經注》諸書者，其與《史記》異同，一一可按。然碎文單辭，知其異於《史》者，無以定其是。而《史》之異於《紀年》者，亦無以定其非。今〈六國表〉及諸〈世家〉，記事明備，一按可得。《紀年》遺佚散亂，荒晦難尋。學者既不以考年為重，好易惡難，習常疑怠，則亦誰為考覈詳定其是非者耶？夫判兩家之異同，貴乎參伍以為驗。求定《紀年》《史記》之得失，不得不參

伍以驗之於諸子。[3]

《繫年》利用古本《紀年》的材料來訂正〈六國年表〉的主張，其實是為了間接證明古本《紀年》的值價。然而，拿《紀年》與《史記》互校，會發生一個問題，即：「知其異於《史》者，無以定其是。而《史》之異於《紀年》者，亦無以定其非。」兩說之間的是非去取，究竟應如何把握？錢氏自創的新方法，就是「參伍以為驗」：透過先秦諸子年世的考訂來驗證史料的可信度。只要史料在諸多相關聯的證據當中，是有所支持的，最起碼是沒有牴牾的，就足以（姑且）信以為實，直到更有力的反面證據出現為止。具體來說，《繫年》的方法就是：「凡先秦學人，無不一一詳考」，從而做到「以諸子之年證成一子」，利用眾多資訊緊密結合的相互關聯性來證成個別資訊的可信度。

二、《繫年》研究成果之不好利用

錢穆先生對《繫年》的研究成果相當有自信。他說：「余為《先秦諸子繫年》,比論《史記》、《紀年》異同，自春秋以下，頗多考辯發明，為三百年來學者研治《紀年》所未逮。」[4]可見錢先生本人也是在《紀年》研究的脈絡下看待《繫年》研究成果的學術價值的。《繫年》問世之際，也普遍受到當時學者的好評。如顧頡剛說：「錢穆先生的《先秦諸子繫年》，雖名為先秦諸子的年代作考辨，而其中對古本《竹書紀年》的研究，于戰國史的貢獻特大。」[5]朱希祖亦云：

> 閱《先秦諸子系年》序。其書為北京大學史學系教授錢穆撰，統考戰國各國年代，頗多糾正《史記》謬誤，謂《竹書紀年》真為魏史，西周以前雖多臆測不可據，而戰國時事年紀實最正確，其論頗有見地。蓋以《史記》各本紀、世家紀年，多與諸子所記時事系年相抵牾，而以《竹書紀年》言之，則多密合，故不可以為偽書視之。他若〈蘇秦考〉謂《史記》、《戰國策》多本偽蘇秦、張儀之書，故蘇、張遊說各國之辭皆不足信，證據頗確實。[6]

《繫年》亦為陳寅恪所激賞。據朱自清的回憶：

> 晚（葉）公超宴客，座有寅恪。.....談錢賓四《諸子系年》稿，謂作教本最佳，其中前人諸說 皆經提要收入，而新見亦多。最重要者說明《史記·六國表》。但據《秦紀》，不可信。《竹書紀年》系魏史，與秦之不通於上國者不同。諸子與《紀年》合，

[3] 《繫年》，頁 42。
[4] 錢穆，〈略記清代研究竹書紀年諸家〉，《錢賓四先生全集》，冊 22，《中國學術思想史論叢》（八），頁 568 - 569 。
[5] 顧頡剛，《當代中國史學》，《當代中國學術叢書》（南京市：勝利，1947 年）。
[6] 朱希祖: 《朱希祖日記》（下冊），1939 年 2 月 12 日條，（北京：中華書局 2012 年），頁 1000。

而《史記》年代多誤。謂縱橫之說， 以為當較晚於《史記》所載，此一大發明。[7]

然而，錢穆先生對《繫年》的自信，再加上《繫年》問世以來所受到的好評及肯定，相對於該書在當今學界，尤其在《紀年》的研究領域上被重視及引用的程度，遠不成正比。例如，繼王國維的《古本竹書紀年輯證》而作的《古本竹書紀年輯校訂補》（范祥雍著，1957 年 9 月第一版。）以及《古本竹書紀年輯證（修訂本）》（方詩銘、王修齡著，1981 年 2 月第一版，2005 年 10 月修訂第一版。）都沒有引用到《繫年》的資料。80 年代以降，海內外學者再度關注《紀年》的歷史價值問題，由不同角度重新審視王國維的定論，且逐漸形成今本《紀年》雖有錯訛，卻非偽作的共識。然而，在這批海內外學者，諸如陳力、黃凡、張培瑜、邵東方、倪德衛、班大為及夏含夷等人的討論當中，幾乎都沒有引用到《繫年》的研究成果。[8]惟有程平山 2013 年出版的《竹書紀年考》提及錢穆及其《繫年》。[9]然而，也只不過散引《繫年・自序》中的見解，並對《繫年》針對《紀年》研究的相關內容作一簡略的介紹而已。並沒有針對其論證及得失進行深入的討論。

上述現象究竟應該如何解釋？尚有待研究。在此謹針對將《繫年》數位化的構想，提出一個假說。本文認為，《繫年》的研究成果之所以在學界未被充分利用，是因為《繫年》的論證構成了一個整體，而其論證的有效性是由整體來證成的。拿錢先生自己的話說，就是：「如常山之蛇，擊其首則尾應，擊其尾則首應，擊其中則首尾皆應。」這個相關聯的整體，固然增強了《繫年》考辨的可信度，卻造成學者採納其研究成果一定程度上的不便。因為，在採納《繫年》的部分結論以前，似乎有必要先對《繫年》有一個全盤的掌握，並考慮是否認同它含納所有細節在內的整體觀點。否則，就很有可能發生史觀或歷史事件上的矛盾而不自覺。在這個意義上，數位化的文本若能夠把那個「整體」分割成若干意義相近或相關的文本脈絡，在個別「考辨」之間建立關係鏈，讓人能夠快速掌握《繫年》所討論的不同問題之間的聯繫，對學者進一步利用《繫年》的研究成果進行有效的批評及運用，相信是有幫助的。

然而，一般在利用《繫年》的成果時，若只是想要檢索某位先秦諸子的生卒年，似乎沒有必要花大力氣去研究《繫年》以後，再來做決定。這裡還牽涉到另外一個問題：《繫年》基本上並不是一本工具書。但是，在基本上接受其整體結論的前提下，確實可

[7] 朱自清，《朱自清日記》，收錄《朱自清全集》，（南京市：江蘇教育出版社，1997 年），卷 10，頁 202。

[8] David N. Keightley, “The Bamboo Annals and Shang-Chou Chronology,” *Harvard Journal of Asiatic Studies* 38, no. 2 (1978): 423–38, doi:10.2307/2718906. David S. Nivison, “The Dates of Western Chou,” *Harvard Journal of Asiatic Studies* 43, no. 2 (Spring 1983): 481–580, doi:10.2307/2719108. 另見：邵東方、倪德衛主編，《今本竹書紀年論集》（台北：唐山出版社，2002）。該書收錄了 1983－2000 年之間，海內外研治《竹書紀年》專家學者的十五篇論文。

[9] 程平山，《竹书纪年考》，《竹書紀年與出土文獻研究 1》。（北京：北京中華，2013 年），頁 307-308。

以拿它當一種工具書來使用。惟世面上訂正〈六國年表〉的著作很多。就不完整統計，已有以下五至六種：

(一)、陳漢章，《史記六國年表新校正》，稿本，1920 年。[10]

(二)、武內義雄，〈六國年表訂誤〉，收錄《武內義雄全集》，卷六。（東京：株式會社角川書店，1945 年），頁 305-327。

(三)、華世出版社編訂，《中國歷史紀年表》。（臺北：華世出版社，1978 年）。

(四)、陳夢家，《六國紀年表；六國紀年表考證》。（臺北市：學海，1992 年）。

(五)、楊寬，《戰國史料編年輯證》。（臺北市：臺灣商務，2002 年）。

(六)、劉俊男，《《史記・六國年表》與史料編纂》，《古典文獻研究輯刊》，第 6 冊。（台北縣永和市：花木蘭文化，2010 年）。

其中華世出版社編訂的《中國歷史紀年表》，因為不載論述文字，可謂純屬於工具書用途。其中戰國的部分，也是利用《竹書紀年》來訂正〈六國年表〉。武內義雄與陳夢家的著作則為短篇論文。因為篇幅不長，較便於檢索之用。楊寬的《史料編年輯證》則是類似於《繫年》的大部頭論著。但楊氏主要著眼於國與國之間的關係以及政治事件上，與錢穆的關注點不同。劉俊男的《史料編纂》是一本碩士論文，主要分析司馬遷作〈六國表年〉的意圖，以及該〈年表〉的特色。內附「〈六國年表〉與相關本紀、世家記事對照表」、「秦國人物列傳記事編年表」、「六國人物列傳記事編年表」、「睡虎地秦墓竹簡《編年記》與《史記》相關記事對照表」等。嚴格說來，雖非正式訂正〈六國年表〉之作，卻可以看出〈年表〉內部的問題，可與《繫年》等論著的結論相互參照。以上幾種訂正〈六國年表〉的著作，其結論與《繫年》有何異同？這些異同，又是基於何種選擇而發生？其中是否容有優劣高下之分？在這些問題未解決以前，其實很難斷定，哪一家的編年比較可靠。在這個意義上，就算只想用作工具書，也無所適從。因此，如果有一個統一的數位平台，將所有戰國紀年相關的論著及史料匯整在一起，並利用數位檢索、對比的功能，使學者便於先參照各家說法，再做取捨，將可確信、及值得存疑的部分區別開來。如是集體協作下，或許就能真正突破《繫年》及其他相關論著原有的格局，共同建立起一個較完整可信的戰國紀年表。一方面可以針對個別論點有所對話；另一方面，也可以使這些著作原本隱涵的「工具書」功能，得以在真正「信實」的基礎上有所發揮。

三、《繫年》的文本特點

[10] 見引於程平山，《竹书纪年考》，《竹書紀年與出土文獻研究 ; 1》（北京：北京中華，2013 年），頁 1127。

《繫年》的內容以四卷〈考辨〉為主，共計一百六十三篇。稱「考」者，旨在考實；稱「辨」者，意在辨偽。〈考辨〉以孔子的生平事跡為首，以李斯為終，間及孔門弟子、墨子、吳起、老子、孟子等。四卷〈考辨〉將孔子以降，尤其是戰國年間的「世運之升降，史跡之轉換，人物之進退，學術之流變」鉤勒了出來。錢氏〈自序〉云：

> 晚周、先秦之際，三家分晉，田氏篡齊，為一變。徐州相王，五國繼之，為再變。齊、秦分帝，逮乎一統，為三變。此言夫其世局也。學術之盛衰，不能不歸於時君世主之提抑。魏文西河為一起，轉而之于齊威、宣稷下為再起，散而之于秦、趙，平原養賢，不韋招客為三起。此言夫其學風也。

復自介四卷〈考辨〉之內容云：

> 書分四卷，首卷盡于孔門，相宰之祿，懸為士志，故史之記，流為儒業，則先秦學術之萌茁期也。次卷當三家分晉，田氏篡齊，起墨子，終吳起。儒、墨已分，九流未判，養士之風初開，游談之習日起，魏文一朝主其樞紐，此先秦學術之醞釀期也。三卷起商君入秦，迄屈子沉湘。大樑之霸焰方熄，海濱之文運踵起。學者盛于齊、魏，祿勢握于遊仕。於是有白圭、惠施之相業，有淳于、田駢之優遊，有孟軻、宋鈃之曆駕，有張儀、犀首之縱橫，有許、陳之抗節，有莊周之高隱，風發云湧，得時而駕，乃先秦學術之磅礴期也。四卷始春申、平原，迄不韋、韓、李。稷下既散，公子養客，時君之祿，入於卿相之手，中原之化，遍於遠裔之邦。趙、秦崛起，楚、燕扶翼。然而爛漫之餘，漸歸老謝，紛披已甚，主於斬伐。荀卿為之倡，韓非為之應。在野有老聃之書，在朝有李斯之政。而鄒衍之頡頏，呂韋之收攬，皆有汗漫相容之勢，森羅並蓄之象，然猶不敵夫老、荀、非、斯之嚴毅而肅殺。此亦時運之為之，則先秦學術之歸宿期也。四卷之書，因事名題，因題成篇，自為起迄，各明一意。[11]

此外，錢先生還將〈考辨〉結論之梗概，繪製成四個與〈考辨〉起迄相應的〈通表〉。〈通表例言〉：

> 〈考辨〉百六十篇，因事命題，因題裁篇，各不相蒙。而〈通表〉則先後一貫。今〈考辨〉所詳，依次散注〈通表〉某年某事之下。〈通表〉為綱，而〈考辨〉為之目。〈通表〉如經，而〈考辨〉為之緯。亦有〈考辨〉所論，關涉廣泛，未可確歸某年某事者，亦隨宜附列。[12]

[11] 《繫年》，頁 47。
[12] 《繫年》，頁 590。

又，〈序言〉云：

> 前人為諸子論年，每多依據《史記・六國表》，而即以諸子年世事實系之。如據〈魏世家〉〈六國表〉魏文稱侯之年推子夏年壽，據〈宋世家〉及〈六國表〉〉偃稱王之年定孟子游宋，是也。然《史記》實多錯誤，未可盡據。余之此書，於先秦列國世系，多所考核。別為《通表》，明其先後。前史之誤，頗有糾正。而後諸子年世，亦若網在綱，條貫秩如矣。

可見〈通表〉其實就是〈六國年表〉的訂正，具體呈現了《竹書紀年》的史料價值所在。〈通表〉之後，還別附〈列國世次年數異同表〉、〈戰國初中晚三期列國國勢盛衰轉移表〉以及〈諸子生卒年世先後一覽表〉，「概括〈通表〉大體，用資參覽」。這些與〈考辨〉內容緊密相聯的〈表〉，皆有助於讀者快速直觀地把握《繫年》的總體內容及結論。

〈通表〉及其附〈表〉而外，《繫年》文本的另一大特點，就是錢穆自編的〈考辨索引〉及〈書名人名索引〉。〈考辨索引〉以人物、事件或議題的形式，將每篇考辨所涉及的內容總括起來。方便讀者依循其感興趣的人物、事件或議題找到相關的〈考辨〉。〈書名人名索引〉則將每篇〈考辨〉所提及的書目及人物標示出來。總的來說，這些〈索引〉具體標示了出每篇〈考辨〉之間的關聯性。因為每篇〈考辨〉在人物、事件及有關史料或史實的「問題」上，都與其他若干篇相關聯，故本書不並適合「線性閱讀」。這個特點非常符合數位化的功能需求。

上述《繫年》的文本特點，使《繫年》文本特別適用於數位化的處理。《繫年》的證據力，其實在於〈考辨〉當中緊密相聯的人物與事件。錢穆先生費盡心思繪製〈通表〉和編訂〈索引〉，正為便利讀者快速掌握人物，以及事件之間，環環相扣的關係。由此可以想見錢穆先生的用心。

〈索引〉當中現成的人物、事件及主題，為數位化以後，在〈考辨〉之間建立有意義的聯結，帶來極大的方便。〈通表〉則為〈考辨〉內容提供了直觀的時間與空間性訊息，將人物與事件，在時間前後的因果關係上，以及四維空間的遠近關係上聯結起來。構成一個人、事、時、地、物緊密相聯的文本脈絡。這樣的文本脈絡，非常適用於數位化的呈現。我們可以善加利用數位人文的強處，讓文本原具有的，在人物關係或社會網絡上，在時空關係及因果邏輯上等等的關聯性，更立體地呈在給讀者。

《繫年》文本的數位化處理，更有助於未來將《繫年》與其他相關的文本作聯結，對比文本內容乃至論點上之異同，讓學者更有效地利用《繫年》文本進行進一步的研究，從而更全面地評估《繫年》的學術價值。在《繫年》與其他文本的聯繫上，系統初步整

合網路資源，提供檢索線上原始文獻內容的便利。由於錢穆常透過二手研究的轉述來引用原始文獻，對基本史料不夠熟的人，很難一眼就看出錢氏的意思。以上功能可使讀者在不解之處聯結至原始文獻的上下文，以便利讀者跟上錢穆的思路。

〈考辨索引〉已經完成數位化，並建立數位的查詢及使用方法。〈通表〉由則尚未取的數位化文本，故未克處理。以下先介紹系統總體的構想及設計理念，再就設計的角度說明系統的實作方法及原則。

四、《繫年》的數位化：系統構想及設計理念

《繫年》的數位化系統，初步目標在於將《繫年》文本上網，以便於整合網路資源。完成此目標之後，再運用詞夾子工具與後分類的技術對《繫年》進行文本分析，進而逐步建立文本內部，以及《繫年》與其他關聯性文本的關鍵詞之間，有意義的聯結。例如：整合文本內部與人物、時間及空間相關的資訊，建立《繫年》文本的社會網絡圖像。抑或將《繫年》文本延伸出去，與其所引用到的原始文獻之線上資源做聯結，再比對兩者之間在意義相關的段落上，文字敘述的異同及其意義。

下一階段的研究與開發，由於數位人文平台的概念日漸成熟，再考量開發數位工具的通用性及廣度。預計將此系統轉移至數位人文平台 DocuSky 上，再整合開發大型可用以做比較及參照的工具，如年表、地理資訊系統、人物網絡關係等。在此平台的文本都能使用其工具，還可透過時間與空間資訊在 GIS 上的整合，具象性地對比出文本異同，進而發現有趣的問題。例如：在平台上有《繫年》及《史記》兩文本，可使用平台上的詞夾子工具，標註出人、時、地、物以後，利用對照表工具將兩個文本同時期的人物篩選出來，再觀察其空間位置。或將人在不同文本中，或是同一文本的不同段落中的時空資訊抽取出來，進行對讀，進而發現問題。學者可以利用統整在平台上的各式數位文本和分析工具（analytics tools）來做研究。相信透過數位平台的整合，能提供先秦歷史學者更好的數位人文研究環境。

五、《繫年》系統的設計：實作方法及原則

在經過一些缺字的處理後將《繫年》的文本使用 UTF-8 儲存。計算《繫年》不包含索引及通表總文字約 37 萬字，檔案大小約 1,100KB。以此資料量評估不需使用關聯式資料庫儲存，以降低運作與維護系統的成本。需要記錄資料則使用 JavaScript 或 JSON 儲存。如此系統主要由 HTML、TEXT 與 JavaScript 這三部分組成，《繫年》文本的修改則開啟純文字編輯；使用介面修改則開啟 HTML 檔案編輯；功能與詮釋資料則開啟對應的 JavaScript 編輯。依循此規則，便可以產生架構簡單與維護容易的系統。

(一)系統軟體架構──輕量級架構

系統主要用 JavaScript 寫成，使用純文字的文本資料。所需的詮譯資料也均存在 JavaScript 的獨立檔案中，以方便維護及擴充。《繫年》系統的網址在 http://pt-phil.appspot.com，建置在 Google App Engine 平台上。選擇雲端平台的好處在於 App Engine 會根據應用程式接收的流量自動調整應用程式設定，因此我們只針對我們需要使用的資源付費。只要上傳程式碼，Google 就會幫您管理應用程式的供應面。換句話說，我們不必建置或維護任何伺服器。建置可自動彈性調整資源配置的應用程式快速啟動，更快完成建置作業。除此之外負載平衡、健康狀態檢查和應用程式紀錄等內建服務能可以大幅提升部署網路和行動應用程式的速度。App Engine 也內建自動擴充功能，可讓您的應用程式依各種規模需求，從零至數百萬名使用者立即自動調整用量規模。目前開發的使用量極小，故不須任何費用。

(二)分散式雲端運

在雲端盛行的年代，如何依使用需求設計一個平均分散計算量的系統格外重要。在雲端計算量的分配上，將大部分系統計算資源移至每個連線的瀏覽器的客戶端上，讓伺服器簡單地只作提供檔案的運算。如此在頻寬足夠的條件下，即可應付大量的用戶使用。

(三)系統索引功能

將紙本的〈考辨索引〉設計成為主題式的標籤，標示在相關的〈考辨〉上。《繫年》共有 163 篇〈考辨〉，1417 條〈考辨索引〉。一條〈考辨索引〉對應至一至多篇〈考辨〉。讀者可藉由〈索引〉標籤探究某一主題的文本脈絡；若要對比事件的先後順序，則可以利用〈通表〉上的資訊；若要對照原始資料則可利用〈書名人名索引〉上的資訊。

(四)系統搜尋功能

在人文研究中盡可能找出所有線索是必要的，再者將最符合的結果依其重要性放排序。以這種思維下設計系統搜尋功能。使用者輸入關鍵字後，分析及計算關鍵字在文本中出現的位置及次數，作為排序優先權的依據。關鍵字在考辨條目中最重要；索引子主題次之；最後是該篇考證的文章中。如此可將所有關鍵字相關文章均找出且依重要性排序。

(五)人事時地物標註

將文本中的人、時、地、事及引用到的書，使用詞夾子工具找出且標註。其中事是指錢穆考辨索引條目。現今找出特定專有名詞作法大致上分為三種,一種是以人工撰寫

規則的 rule-based 方法，一種是以建置詞庫為主的 corpus-based 方法， 最後一種是利用學習方式 machine-learning 的方法。大部分的作法都是以詞庫為主，但是詞庫要建置完備並不容易。本系統的標註的是使用一個不建立詞庫的方法，來做專有名詞辨識及標註，此法張尚斌稱為詞夾子，在其論文中有說明：

> 詞夾子演算法來解決專有名詞辨識的處理，詞夾子是使用「前文」、「詞首」、「詞尾」、「後文」的組合。主要概念是利用文章寫作上的一些特定習性與字辭之間的耦合關係，來找出專有名詞。先給予樣本詞，然後找出和樣本詞相關的詞夾子，並利用這些詞夾子找出與樣本詞類似的候選詞出來，之後以迭代方式不斷的產生詞夾子和候選詞。[13]

(六)系統後分類功能

系統能將關鍵字檢索後的結果，依人、時、地、事（〈考辨索引〉主題標籤）及書籍作後分類計算，可以檢視查詢結果及分類的數量。方便將查詢目標以人、時、地、事及書的面向作為觀察。後分類的優勢在蕭屹灵的論文中指出：

> 為了使得典藏檢索系統能發揮檔案的價值，使用『屬性標籤』的資料結構整合詮釋資料,並引入了多維度的後分類導覽方式；更在此基礎下，進一步為系統加入檢索詞組控制介面，發展出擁有檢索詞組控制與後分類架構的整合式檢索介面系統之模型，以解決一般檢索系統之缺點,達成連貫檢索流程的目的。[14]

可見在現代的數位環境此功能是必備的研究工具之一。透過上述方法，建立《繫年》的數位化平台，以提供人文研究者數位化的使用環境。

六、《繫年》系統的使用方法

《繫年》系統的網址在 http://pt-phil.appspot.com 。《繫年》一書中的目次與考辨索引已數位化建立在系統的文本中，可以依目次或索引找到所要的考辨，當按下連結即會產生新的標籤頁呈現該篇考證。下圖 1 即是按目次載入第一篇考辨。

考辨索引共有 1417 條考辨子主題，亦可在系統中使用找尋相關資料， 如下圖 2。目次與索引皆是錢穆先生在此書中設計，讓讀者可以依線索查詢，而數位化後可以更快

---

[13] 張尚斌，《詞夾子演算法在專有名詞辨識上的應用──以歷史文件為例》。臺北：國立臺灣大學資訊工程學研究所碩士論文，2006 年。

[14] 蕭屹灵，《日治法院檔案系統及其後分類呈現》。臺北：國立臺灣大學資訊工程學研究所碩士論文，2008 年。

且整合兩者，提供此書的大致架構，例如「考辨與索引」功能。

圖 1 使用目次載入考辨

圖 2 使用考辨索引

在「考辨與索引」中可依考辨主題，展開其考辨索引之子主題。例如第三篇、〈孟懿子南宮敬叔學禮孔子考〉，展開後可以發現錢穆在其中設定許多子主題，瞭解這子主題後，才可一窺錢穆先生此篇考證的架構。同時亦可以發現這些子主題可能會與其它〈考辨〉相關也須一並瞭解。第三篇〈考辨〉按其子主題的所在考辨篇次可知，尚與 29 與 48 篇有相關，其中 48 篇是它篇最相關的，如圖 3。藉由以上的設計，可按錢穆先生的想法再延伸，以達讀者更易掌握其閱讀方向之目標。

圖 3 考辨與引索呈現考辨子主題

在每篇考辨的閱讀上，最上方兩個按鈕可以瀏覽鄰近前後的考辨篇次；之後是考辨主題及內文；最後是閱讀脈絡，如圖 4。在考辨內文中有許多超連結，可以連至網路上搜尋相關的內容，如果是書籍，會連到中國哲學書電子化計劃(http://ctext.org)中搜尋相關的全文，人名則是連結到維基百科(https://zh.wikipedia.org)，藉由外部資源的整合運用，建立整合的閱讀環境。

圖 4 考辨閱讀

閱讀脈絡則是結合考辨索引之子主題與考辨中的人、時間、地點、參讀書目及參讀考辨，如圖 5。參讀考辨篇次是指錢穆先生引用的其它考證篇次；參讀書目則是錢穆

先生引用書目。子主題則是整理考辨索引而產生的考辨架構。透過子主題的設計，可以瞭解此篇考辨涉及哪些重要議題？這些議題與其他考辨之間的無關為何，藉甲數位化的文本可以幫助研究者一目了然的看出及使用。這些功能即是整合運用錢穆的〈考辨索引〉，所以可以幫助我們快速找到我們所關心之議題在哪篇或哪幾篇〈考辨〉當中有涉及到，也是數位化文本的目的之一。

圖 5 四十八篇考辨脈絡

系統嘗試將本書大量以〈表〉及〈索引〉建立閱讀線索，重新設計且彼此連結參照內外部資源。將此書閱讀脈絡整合且隱藏於系統中，以資訊技術降低使用難度，將多元關聯相互連結。例如：當讀者閱讀考辨第二十一篇〈孔子過宋考〉，系統會呈現閱讀脈絡，讓讀者知道支持此考辨的子主題項目，而某些子主題若與其它考辨相關也會一併顯示，且可開啟閱讀。在考辨中錢穆亦提及有用到哪些考辨及書目，系統也會一併顯示及連結，外部書目會連至中國哲學書電子化計畫中搜尋。除此也將相關的人、時、地呈現。

全文搜尋功能可以針對感興趣的關鍵字搜尋所有的考辨。透過 JavaScript 實作檢索及後分類功能。以搜尋「孔子」為例，共找到 60 篇考證，按照相關程度排序。60 篇的檢索結果，按人、時、地、事及書作後分類，提供讀者觀察脈絡的不同面向，如圖 6。

圖 6 全文搜尋

選取後分類的不同維度，以觀察不同的角度。在搜尋孔子的 60 篇中，地點的視角最多的魯國與齊國，如圖 7。孔子是魯國人且曾在齊國作官，空間的角度很符合預期。

後分類的「事」則是使用錢穆先生考辨索引的子主題為事件的統計方式。例如孔子的 60 篇檢索結果，孔子自陳適蔡(22,19,31)、孔子厄於陳蔡之間(21,19,27)、子夏居西河 3(29,38,39)、南宮括(29,3,48)、子夏喪明(39,67,48)與孔子宰中都(12,6,4)。這五件事是與孔子的檢索結果最相關，如圖 8。

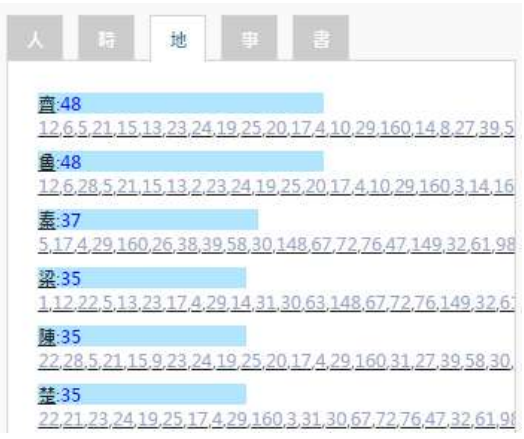


圖 7 以空間角度檢視查詢結果

圖 8 事之後分類結果

由「書目」的角度可以看見，錢穆先生在考辨時使用書目的狀況。在孔子的檢索結果中使用史記最多；使用自己考辨次之，之後左傳…等，如圖 9。

圖 9 書目後分類結果

圖 10 人的後分類結果

在孔子的檢索結果中，由「人」的角度看，孔子與孟子出現最多是正常的。但崔述居然出現在第三名，如圖 10。崔述為清代研究儒學，胡適稱崔述為「二千年來的一個了不得的疑古大家。」錢穆先生在考辨孔子的相關事跡引用許多崔述的研究。這也提供現在的研究者可藉由先人的成果繼續研究。

七、結語

目前系統可以閱讀及檢索《繫年》一書，將檢索結果依人、時、地、事以及書作分類。應該可以達到數位化文本基本的使用目的。〈通表〉與〈附表〉尚未數位化，無法有效整合時間與空間資訊，是最可惜之處。為方便將錢穆的研究成果與他所參考的原始以及相關的二手研究做對照，數位化以後的〈通表〉將用一種方便的呈現方式，與《史記．六國年表》、《竹書紀年》等原始文獻，以及楊寬的《戰國史料編年輯證》、陳夢家的《六國紀年》等戰國史編年研究的成果相互對照，讓學者得以在數位人文平台上進行資料的匯整，校讎及討論，從而得出一個真正將所有相關的研究成果匯整一處，更為信實的戰國編年史。

最後希望可以藉由本系統開發的階段性報告，尋得同好一起討論設計及技術上的問題，共享研究心得和成果。

引用書目

Keightley, David N. (1978). “ The Bamboo Annals and Shang-Chou Chronology.” *Harvard Journal of Asiatic Studies* 38, no. 2: 423–38. doi:10.2307/2718906.

Nivison, David S. (1983). “The Dates of Western Chou.” *Harvard Journal of Asiatic Studies* 43, no. 2: 481–580. doi:10.2307/2719108.

錢穆。2000。《中国史学名著》。北京：生活・讀書・新知三聯书店。

張尚斌。2006。詞夾子演算法在專有名詞辨識上的應用──以歷史文件為例。國立臺灣大學資訊工程學研究所碩士論文。

蕭屹灵。2008。《日治法院檔案系統及其後分類呈現》。國立臺灣大學資訊工程學研究所碩士論文。

《春秋》三傳對讀系統

趙叡*、謝于琳**

摘　要

《春秋》是中國歷史上最早的編年體史書之一，記載了上起魯隱公元年 (公元前 722 年)，下迄魯哀公十四年 (公元 481 年)，歷十二君，共二百四十二年的史事。所謂的編年體史書，就是“系日月而為次，列十歲以相讀”，以「年，時(季節)，月，日，記事」為體裁，記錄魯國與眾諸侯國的大事。

《春秋》三傳就是註釋《春秋》的史書，有左氏、公羊、穀梁三家，稱為「春秋三傳」。而三傳作者皆不同，各自所闡述的方式，對事情的看法不盡相同，也因此三傳彼此間對於某些史事的描述有所出入。東晉范寧評三傳時所說:「《左氏》豔而富，其失也巫（指多敘鬼神之事）。《穀梁》清而婉，其失也短。《公羊》辯而裁，其失也俗。」表明了不管在敘述還是在觀點上，三傳皆有不同的看法。

綜合各種不同的文本資料，加以比對觀察，推導出不同的見解或看法，是人文學者所熟悉的研究方式。也因此如何提供使用者合適的對讀系統，是數位人文時代的重要課題之一。臺大數位典藏與自動推論實驗室團隊過去曾開發針對共時性的文本，如清實錄、明實錄、朝鮮李朝實錄的對讀；以及內容本身互為說明，以關聯性為出發的乾隆朝會典與則例對照系統。在這些基礎上，試著以《春秋》經為對象，同時結合時間與內容的比對，並建立利於閱讀和分析的使用者介面，除了令該領域的專業研究能對《春秋》有更快的掌握，同時亦希望從工具的角度出發，讓這樣的閱讀和研究模式，能套用在不同的文本，成為未來對讀工具開展的基礎。

關鍵字：《春秋》、《春秋》三傳、對讀系統

* 國立臺灣大學資訊工程學系碩二學生，Email: ray20013247777@gmail.com。
** 國立臺灣大學資訊工程學系碩二學生，Email: jinnij11107@gmail.com。

A Comparitive Reading System for the *Three Commentaries Of Chunqiu*

Jui Chao*, Yu-lin Hsieh**

Abstract

Chunqiu, the history of Lu compiled by Confucius that spans from 722 BCE to 481 BCE, is one of the most important historical record in chronological form. To explain Confucius' very terse recording, three annotations, *Zuo Zhuan*, *Gongyang Zhuan* and *Guliang Zhuan*, were written. While each presents and interprets Chunqiu in its own way, together they are called the *Three Commentaries of Chunqiu*.

This paper presents an effort to develop a system that allows a reader to simultaneously read the three *Commentaries*. It utilizes the chronological nature of the records, treating the original writing in *Chunqiu* as a headline, and presents and compares the writings in the three *Commentaries*. Our approach is general enough to be applied to other chronologically recorded documents, such as diaries.

Keywords: Chunqiu, Three Commentaries of Chunqiu, simultaneous reading

* Master Student, Department of Computer Science and Information Engineering, National Taiwan University. Email: ray20013247777@gmail.com.
** Master Student, Department of Computer Science and Information Engineering, National Taiwan University. Email: jinnij11107@gmail.com.

一、動機

《春秋》是中國歷史上最早的編年體史書一，記載了上起魯隱公元年(公元前 722 年)，下迄魯哀公十四年(公元 481 年)，記錄期間內魯國與眾諸侯國的大事。《春秋》三傳則是解釋《春秋》經的史書，三傳分別為《左傳》、《穀梁傳》、《公羊傳》。其中三傳的作者皆不同，進而解釋《春秋》經的觀點也不相同，《左傳》會把經文的前因後果交代出來，形成一個完整的故事，《穀梁傳》、《公羊傳》則是去推測孔子為什麼會要這樣寫，也就是把《春秋》背後的評論給寫出來，然而兩本書可能會有截然不同的推論。

(一)研究價值

根據前面所述，我們可以將《春秋》經的經文看做是一個新聞標題，並未有很多解釋，敘述一個歷史事實以及暗蘊著一些作者自己的主觀看法。而《春秋》三傳則可以視為各自不同的媒體，依照著新聞標題做深度的解釋和探討。但也因為如此，《春秋》三傳作者在解釋經文時，也會帶入自己的觀點以及立場，甚至為《春秋》經的條目立下褒貶，《春秋》三傳間彼此的闡述也可能有所出入。正因為如此，《春秋》經與《春秋》三傳間的對照十分耐人尋味，也會是學者感興趣得研究目標。

(二)《春秋》文字簡練

《春秋》經一書文字簡練、記事簡略，但卻微言大義，暗含褒貶。文中雖然不直接闡述對人物以及事件的看法，但卻透過細節描寫、修辭手法，委婉的表達自己的主觀看法。因此若學者本身對於《春秋》經的了解沒有到一定的程度，直接研讀《春秋》是十分困難的，文行中少則幾字，如：

> 昭公三年：「夏，叔弓如滕。」

多則也不超過四十五字，如：

> 定公四年三月：「三月，公會劉子，晉侯，宋公，蔡侯，衛侯，陳子，鄭伯，許男，曹伯，莒子，邾子，頓子，胡子，滕子，薛伯，杞伯，小邾子，齊國夏，于召陵，侵楚。」

從這點來看，若能搭配上《春秋》三傳的注解，能夠讓學者更容易地對《春秋》經文有所了解，畢竟工欲善其事，必先利其器。而三傳各自有不同的特色，《左傳》具有文辭之美，而《公羊傳》、《穀梁傳》 則是以論說取勝，藉著《左傳》的事與《公羊傳》、《穀梁傳》兩傳的例，可以知道春秋時代的史事與聖人之意，使學者可以在三傳的輔助之下，對《春秋》經文有更深入的了解。

(三)翻看不易、對照困難

綜合以上觀點，在研讀《春秋》經時，《春秋》三傳可以說是十分重要的史料，經

由《春秋》三傳來了解《春秋》，是研讀《春秋》經的不二法門。而在數位典藏 (digital archives) 盛行的年代，已經有許多網站系統有《春秋》三傳的文本，如維基百科(wiki)、中國哲學書電子化計畫(ctext)等，不須使用笨重的實體書，也能輕鬆地從網路上取得這方面的資訊。

儘管如此，卻沒有一個系統可以”一次性的”參讀多本史料。我們設想一個情境，當學者想在同一個時間點之下，藉由三傳的輔助，品出孔子微言大義下所蘊含的意義，過程中必須頻繁的翻動書本，甚至不停地從三傳之中彼此切換，這對學者來說是很不體貼的。儘管有數位化的網站系統，學者還是需要開多個分頁，在彼此之間切換。這些動作依然會帶給學者很大的不方便，而繁瑣的切換也會造成研究以及對照上的失誤。

從這些角度為出發點，設計出一套可以讓學者方便使用的對讀系統，讓學者不需要去煩惱研究以外的事情，更專注於在《春秋》經與三傳之間的研究，讓” 翻動書本” 、 ”對應時間點”的工作讓系統代勞，增加學者的研究效率。

二、資料前處理

在這對讀系統中，我希望能做到《春秋》經與《春秋》三傳的對讀，不只是其君王年份的對應(如魯隱公元年)，我更希望能做到君王年份之下，各個條目之間的對應，例如我會希望系統能幫我在《春秋》經與《春秋》三傳之中，對應同樣年份中三月所發生的事情，藉此觀看三傳中不同的觀點。預做到這點，我必須將文本中每一條條目的時間點擷取出來標記，這樣才能達到我期望的目標。

《春秋》經與《春秋》三傳皆以編年體的方式編寫，又《春秋》三傳是解釋《春秋》經，因此體裁以《春秋》經為主，而《春秋》經以「年，時(季節)，月，日，記事」為體裁，結構如下：

年:魯國之君主、魯公在位紀年。

時:季節。四季之「春夏秋冬」。

月:「正月、二月、三月...」。

日:「甲子。乙丑、丙寅...」。

記事:短句構成。

以下以《春秋》經魯隱公二年為例：

年	時	月	日	記事
隱公二年	春			公會戎於潛
	夏	五月		莒人入向 (1)
				無駭帥師入極 (2)
	秋	八月	庚辰	公及戎盟於唐 (3)
		九月		紀裂繻來逆女
	冬	十月		伯姬歸於紀
				紀子帛莒子盟於密
		十有二月	乙卯	夫人子氏薨
				鄭人伐衛

可以發現在《春秋》經中，有大量的條目皆有記載其發生時間的時、月，而少數條目有記載發生的日。因此我可以以這個特性，以自動化的方式擷取出每一條條目的大致時間，以便我以後在對應上能有更多的資訊。至於少數沒有記載其發生時間的條目，我則以編年體的性質，”大致推斷出”其記載的時間點。以表格中(2)為例，「無駭帥師入極」並無任何與時間相關的資訊，以至於我沒辦法擷取出其確切的時間點。但以編年體的性質來看，記事的方式與時並進，雖然沒有時間的資訊，但我可以藉由其上下條目去推斷出大致的時間點，因此我以表格中(1)的五月與(3)的八月可得知(2)的時間點大致上是五到八月之間，由此給予標記。

三、使用者介面與系統工具介紹

既然將系統定位在能夠輕鬆地對讀多本史料，在網頁上的版面配置就是一大挑戰，既要讓多本史料一次呈現，也要提供方便使用者使用的輔助工具。

(一)使用者介面

圖(1)為系統介面，版面配置上《春秋》經的部分會比其他文本的空間都來的小，因《春秋》經記事簡略，相對於《春秋》三傳來說其條目字數較少，因此將比較多的空間分配給《春秋》三傳，使其較好閱讀。每一份文本皆有自己的滑軌，使用者可以自行滑動滑軌前往目標，亦可藉由年份書籤的功能來移動。年份順序皆是照著《春秋》經編年體的記事順序，由魯隱公一直到魯哀公。

圖(1). 對讀系統使用者介面，以《春秋》經以及《左傳》為例。

隨著使用者所要觀看的文本越來越多，每一分文本所分配到的版面也會越來越小，導致最後所有的字都擠在一起，擁擠不堪，可讀性直線下降。為了要解決此問題，此系統從《春秋》經與《春秋》三傳的性質下手。在論文的動機有提到，《春秋》經文字簡練、記事簡略，長則四十五字，短則一字，與《春秋》三傳大不相同。以新聞來比喻，可以把《春秋》經當作是新聞的標題，而《春秋》三傳則視為不同媒體對該標題的報導。

因此在版面配置上，將《春秋》經縮小在左上角，用下拉是選單的方式呈現目標時間點的年份以及條目，將比較大的版面讓給相對字數較多的《春秋》三傳（圖 2），讓整體的可讀性提高，多文本的情況下也不會讓版面上顯得凌亂。

圖(2). 再多文本的情況下，將《春秋》經縮小在左上角。

(二)系統工具介紹

接下來介紹系統中有一些能幫助使用者研究的工具，主要是以對照不同文本中條目的時間點為主：

1. 年份書籤
2. 條目錨點
3. 條目時間對應

(1) 年份書籤

可以看到圖(3)中標記的地方有一排記載了春秋所有君王名稱(e.g:隱公、桓公等)的書籤列，該書籤點開後可以看到該君王的統治年分(e.g:魯隱公元年…等)，而使用者點選後，會將頁面上有的文本滑動到該年分的區塊。此功能比較偏向是目錄，幫使用者直接滑動到欲研究的段落，減少使者自己尋找該年分的時間。

以魯閔公二年為例，若使用者想觀看春秋與左傳在魯閔公二年所記載的條目，使用書籤點擊及魯閔公二年後，下方文本會自動滑動到該年分的區塊。

圖(3). 點選魯閔公二年後，所以呈現的文本皆會滑動到魯閔公二年的位置。

(2) 條目錨點

除了年份書籤可以移動文本外，點選任何條目也可將滑軌得位置移動到不同文本但時間能互相對應的條目上，讓使用者不管在閱讀任何一份文本，也可迅速地找到其他對讀文本的相同段落。

以圖(4)、圖(5)為例，目標文本為《公羊傳》魯成公十二年，可以看到除了《公羊傳》在欲觀看的位置之外，《左傳》、《穀梁傳》皆在不同的位置。而現在我想以《公羊傳》魯成公十二年:「十有二年春…..自其私土而出也」為目標條目，觀看在《左傳》、

《穀梁傳》中相同時間點條目的闡述。

點擊目標條目後，可以觀看在圖(4)之中的藍框，系統會自動將版面上所有文本移動到與目標條目相對應的時間點(魯成公十二年春)。

圖(4). 圖中藍框部分為目標條文。

圖(5) 點擊目標條文後，會將不同文本皆移動至相同時間點。圖中可以看到在《左傳》以及《穀梁傳》皆移動到了與目標條目相同的時間點(魯成公十二年春)。

(3) 條目時間對應

此功能與**(2)條目錨點**功能相近，**條目錨點**是將現有文本皆滑動到點選條目的時間點，而**條目時間對應**則是能幫你標記出**其他文本同時間的條目。**

以圖(6)為例，呈現的文本為《左傳》、《公羊傳》、《穀梁傳》這三本書，而點選的條目為《公羊傳》魯僖公三年中的”三年春、王正月，不雨”，可以在《左傳》以及《穀梁傳》看到，系統會淡化不同時間點的條目，顯示出相同時間點的條目，而左上角縮小的春秋也會顯示對應的條目，簡單來說可以快速地呈現如下的對照:

《春秋》魯僖公三年:「三年春，王正月」。

《左傳》魯僖公三年:「三年春，不雨」。

《公羊傳》魯僖公三年:「三年春，王正月，不雨」。

《穀梁傳》魯僖公三年:「三年春，王正月，不雨。不雨者，勤雨也」。

這能讓使用者非常快速的拿到欲觀看的資訊，不需要一一去翻書也不用開很多網頁，減少過程中一些與研究無關的程序。

圖(6).以《公羊傳》魯僖公三年的目:「三年春，王正月，不雨」為例

四、近期工作與未來展望

本系統尚在開發以及改善的階段，亦有文本在時間標註上的一些錯誤，而我也會持續的更新，勘誤。也希望能有在《春秋》經與《春秋》三傳上研究的學者能提供我一些意見、回饋，讓這個系統能更容易地被使用以及推廣。

Paper Session 1

脈絡中的文本：自動化取徑

Text, Context & Programming

資料化與地方歷史文獻的數位化、文本挖掘：以《中國地方歷史文獻資料庫》為例

趙思淵*

摘　要

歷史文獻資料庫可區分為數位化（digitalization）、資料化（datalization）、文本挖掘（text mining）三種不同形態，迄今多數中文歷史文獻資料庫實現了數位化功能，部分地實現資料化功能，而能夠實現文本挖掘功能的則十分少見。數位化是將文獻的物理形態轉化為電子形態，資料化是將文獻轉化為可量化分析的資料，編制中繼資料（metadata）是主要方法。文本發掘是在此基礎上開發文本分析工具。《中國地方歷史文獻資料庫》以文獻學研究為基礎，建立特定的中繼資料結構，提供交叉導航、資料統計等多種功能，這些功能不僅可以幫助研究者找到自己的所需文獻，更可能幫助研究者發現新的研究議題。史學研究中，資料庫有必要被視作一種新的文獻形態，建立針對性的文獻學方法論。

關鍵字：地方歷史文獻、數字人文、文本挖掘、中繼資料

* 上海交通大學人文學院歷史系講師，Email: titaner@sjtu.edu.cn。

Digitization of Local Historical Archives, Creation of Metadata, and Datamining：The Example of the Chinese Historical Local Arcives Database

Si-yuan Zhao*

Abstract

The paper exam three concepts to define historical archives database that digitalization, datalization and text mining. Digitalization made the archives to be cyber texts, and datalization made the text quantized with metadata, text mining meant more analyze tools to be applied in the database. The database of Chinese Historical Local Archives was designed with the principles of datalization and text mining that constructed a modified metadata based on Dublin Core with an archive criticism. Cross search and statistics is also available in the database. The database is going to promote researchers to “analyze” but not “find” materials. It is also suggested that a new corresponding methodology on archives criticism should be applied to the database.

Keywords: historical local archives, digital humanities, text mining, metadata

* Lecturer, Department of History, Shanghai Jiao Tong University. Email: titaner@sjtu.edu.cn.

一、引言

數位化（digitalization）、資料化（datalization）、文本挖掘（text mining）是歷史文獻資料庫的 3 種不同形態。數位化是將文獻從物理形態轉化為電子形態，資料化是將電子形態進一步轉換為可識別的文本與可分析的資料，文本挖掘則是針對文本、資料做進一步的計量、相關性、GIS 分析。本文將嘗試提出並解釋 3 種資料庫形態分類的依據，並以《中國地方歷史文獻資料庫》為例說明如何實現資料化與文本挖掘。最後，本文將提出一個有待解決的問題，即史學研究中，資料庫是否已經有必要視作一種新的文獻形態，並建立針對性的文獻學方法論？

《中國地方歷史文獻資料庫》是 2012 年以來由上海交通大學出版社、圖書館、歷史系合作開發的資料庫。該資料庫由上海交通大學歷史系收集資料，並提出資料庫建設構想，於 2012-2013 年間由上海交大圖書館進行文獻整理與資料加工[1]，2013 年以來由上海交大出版社進行資料庫研發。該資料庫主要收錄上海交通大學 2009 年以來陸續收集的浙江、安徽、福建等地地方歷史文獻及 2007 年以來曹樹基教授收集、授權複製的《石倉契約》，總計近 35 萬件，目前已進入資料庫的有 10 萬餘件。

二、從資料化到文本挖掘：歷史文獻資料庫的演進

數位化並非一個新鮮概念，通常語境中，數位化是指將文獻的物理形態轉化為電子形態，或者說將類比資料轉換為二進位資料。但電子形態的文獻除了易於傳播外，並不能增強文獻的利用價值。如今天廣泛傳播於網路的書籍掃描電子檔，對讀者來說，只是將閱讀載體從紙本書變成了電腦螢幕，並未真正改變使用者利用文獻的方式。如果將“大資料時代”理解為書籍電子檔橫行的時代，則遠不能視為歷史學研究的重大變革。

真正能夠改變文獻利用方式的是資料化，也即將文獻轉化為可製表分析的量化形式。[2] 歷史文獻中包含的物價、產量、價格等資訊，可以被轉換為量化資料，其他描述性的資訊，也應通過某種形式轉換為可量化分析的資料，這是歷史文獻資料化的理想狀態。目前在針對歷史文獻的研究方法中，常用的是詞頻分析、GIS 以及關係網絡分析等。

資料化的意義是將利用文獻的方式從“讀”轉變為“分析”，其核心方法是重組文獻內容，置入使用者所建立的新的文本或資料結構中，也即文獻的結構化。歷史學研究中，這也並非新鮮事物。電腦出現之前，史學研究者已經在製作史料編年、人物關係表，經

[1] 李芳、陳進、王昕：《上海交通大學新藏地方歷史文獻的數位化建設規劃與實踐》，《大學圖書館學報》2015 年第 2 期，第 77-83 頁。

[2] （美）維克托·邁爾·舍恩伯格：《大數據時代：生活、工作與思維的大變革》，周濤譯，浙江人民出版社，第 104 頁。

濟史與社會史研究中也早已整理了各種資料序列。如何炳棣先生研究科舉與中國社會流動性的關係，在硃卷基礎上建立的龐大資料表並未借助電腦技術。電腦技術帶來的革新是使得研究者可以更高效率地建立並使用資料。如王業鍵先生主持建立的《清代糧價資料庫》[3]，該資料庫建成於 2008 年，最初依靠代碼表查詢資料，現在則已可利用下拉清單查詢時間、地域、糧別，是中國經濟史研究的基礎性資料。

給歷史學帶來真正深刻變革的是，電腦技術提供了分析資料化文獻的複雜工具。文本挖掘的理念，正是由此興起。從資料化到文本挖掘的演進，以"數位人文"（digital humanity）概念的興起為標誌。這一理念引導了資料庫建設、開發思路的轉變，人文學研究者不再是被動選擇既有的資料庫，而是參與資料庫建設過程，由其自身研究需要引導資料庫開發，資料庫開發過程也就成為其研究的一部分。

"數字人文"概念在 20 世紀 90 年代興起，逐漸取代 20 世紀 70 年代以來的"人文電算"（humanity computing）概念，成為一個增長迅速的交叉研究領域。項潔、王曉光等已經先後梳理了數位人文概念在西方學界的發展及其在中文人文研究中的適用性。[4] 我認為數位人文研究還可細分為文獻資料庫、線上博物館、網路文本（cyber born context）分析三個子領域。

總的來看，西方數位人文研究更多的力量集中于對網路文本的分析。[5] 互聯網出現以來所積累的各類型線上資料，數量已經十分龐大，自然成為研究者們關注的焦點。而且這類研究直接與互聯網經濟相關，很多不同學科的學者都有興趣介入。線上博物館所關心的則是如何將傳統藝術領域的"展示"轉變為線上的、視覺化的、互動的。

相較而言，歷史文獻雖然留存數量龐大，但已經電子化的規模仍遠遠少於網路文本，並且歷史文獻資料庫研究的收益回報也顯然低於網路文本研究。因此，針對歷史文獻的數位人文研究並沒有如網路文本那樣活躍。雖然如此，如前所述，不論在西方學界或中文學界，都已經有很多數位人文導向的歷史文獻數位化或資料分析研究。今後的歷史文獻數位化過程中，數位人文導向將是一個總的趨勢。

近年歐洲史研究中已經出現越來越多以文本挖掘為主要目的的資料庫或分析工具。如 Tara Andrews 開發的拜占庭文書校勘（critical editing）工具。[6] 此外，荷蘭、比利時等國家在 2013 年集中討論了"大資料"（big data）對歷史學研究的影響，他們所開發的

---

[3] 《清代糧價資料庫》，http://mhdb.mh.sinica.edu.tw/foodprice，發佈日期：2014，訪問日期：2016 年 7 月 26 日。

[4] 項潔、陳麗華：《數位人文——學科對話與融合的新領域》，項潔編：《數位人文研究與技藝》，臺灣大學出版中心，第 9-23 頁；王曉光：《"數位人文"的產生、發展與前沿》，《方法創新與哲學社會科學發展》，武漢大學出版社，2010 年，第 207-221 頁。

[5] David M. Berry ed. , *Understanding Digital Humanities*, Palgrave Macmillan, 2012, p4.

[6] Tara Andrews, *The third way: philology and critical edition in the data age*, working paper, in Lectio Round Table "Digital or critical/digital and critical?", Leuven, 2011.

Biland 資料庫以及 WAHSP 資料庫可以對 17-18 世紀歐洲的媒體資料進行詞頻分析、語言比較分析，為人文學者提供幫助。[7]

中國史研究中，21 世紀初時已有不少學者開始考慮如何使用資料庫便利文獻檢索與研究。[8] 這些討論中，多數學者的關注點是如何使用資料庫，而較少涉及如何開發針對性的資料庫，研究者參與資料庫設計、開發者更少。這一時期代表性的中國史史料資料庫是《文淵閣四庫全書》電子版與《中國基本古籍庫》。《文淵閣四庫全書》電子版由上海人民出版社與香港迪志文化公司、香港中文大學共同開發，於 1999 年投入市場。[9]《中國基本古籍庫》于 1998 年作為高校古委會專案立項，由北京大學領銜開發，完成於 2001 年，此後陸續投入市場。[10]

作為第一代中文史料資料庫，當時的主要技術難點是文字錄入與標準化，實際也就是資料化問題。《四庫全書》電子版在研發中曾與清華大學電腦系合作，開發"多特定人准規範手寫 OCR 引擎"，用於文字自動錄入。如何處理 Unicode 字元集之外的文字，以及如何利用 XML 語言建立文字標引，在當時都是有待解決的技術難題。[11] 傳統文獻學中的版本考辨，也是這類資料庫所面臨的困境，在當時的開發條件下並沒有很好地解決。[12] 此外，《四庫全書》電子版與《中國基本古籍庫》最初都使用光碟版發行，這是由當時的技術條件與網路速度決定的。

因此，以上資料庫所體現的設計理念是將其視作檢索、獲得文獻文本的儲存平臺。儘管當時的研究者已經認為"電子版不是紙版翻印"[13]，應當具有豐富的研究功能與工具，但是他們所指的研究功能主要還是檢索功能。

2007 年以來，歷史文獻數位化的範圍擴大到古籍以外。有越來越多學者討論民間歷史文獻、地方歷史文獻資料庫[14]，除歷史學者外，也有圖書館學學者基於各圖書館的館藏情況，提出特色文獻資料庫建設構想。但不論討論歷史文獻資料庫的使用或建設，多數研究者構想的主要是資料庫的資料儲存、文本檢索功能，而較少考慮如何使用資料

---

[7] Joris van Eijnatten, Toine Pieters, Jaap Verheul: "*Big Data for Global History: The Transformative Promise of Digital Humanities"*, Low Countries Historical Review, 2013, 128(4): pp. 55-77.

[8] 包偉民：《論當前電腦資訊技術對傳統歷史學的影響》，《杭州大學學報》1998 年第 2 期，第 1-8 頁；王文濤：《古籍數位資料應用與史學研究》，《史學月刊》2009 年第 1 期，第 119-125 頁；陳鵬：《新世紀以來的史料型資料庫建設與中國近代史研究》，《國家圖書館學刊》2013 年 6 期，第 33-38 頁。

[9] 程之：《香港推出〈文淵閣四庫全書電子版〉》，《出版參考》1999 年第 16 期，第 12 頁。

[10] 《中國基本古籍庫光碟工程基本完成》，《圖書館理論與實踐》2001 年第 02 期，第 74 頁。

[11] 張軸材：《〈四庫全書〉電子版工程與中文資訊技術》，《電子出版》1999 年第 03 期，第 3-6 頁；朱岩：《談古籍數位化》，澳門圖書館編：《"兩岸三地古籍與地方文獻"會議論文集》，澳門圖書館，2002，第 143-150 頁。

[12] 陳尚君：《〈中國基本古籍庫〉初感受》，《東方早報·上海書評》，2009 年 8 月 9 日。

[13] 朱岩：《〈四庫全書〉電子版問世的啟迪》，《中國圖書館學報》1999年第06期，第82-84頁。

[14] 趙思淵、湯萌：《上海交通大學新藏地方歷史文獻的分類法及其依據》，《上海交通大學學報（哲學社會科學版）》2014 年第 3 期，第 78 頁。

庫中說明研究者分析文本。

中文民間文書、地方文獻資料庫中，迄今文本分析、資料處理功能最為完備的是臺灣數字歷史圖書館（THDL），該資料庫由項潔教授領導的臺灣大學數位人文研究中心開發，主要收錄臺灣地區契約文書及臺灣總督府檔案。臺灣大學數位人文研究中心並不擁有這些資料的實體，而是以授權複製或者錄入為電子文本的形式建設資料庫內容。[15]

THDL 中提供詞頻分析、上下手契關聯分析、人物相關性分析等不同功能，還可以部分地實現契約空間分佈的展示。THDL 提出了資料庫建設的新理念，那就是資料庫的主要功能是為研究者提供研究環境並幫助研究者發現問題，而非僅僅是儲存與檢索。[16]

由上可見，迄今流行於網路中的中文歷史文獻電子資源中，數量最龐大的是掃描、錄文、影像等數位化資源，如大量的書籍掃描電子檔，以及部分全文檢索資料庫。此外借助電腦技術實現的文獻資料化成果正在逐步積累，如王業鍵先生主編的《清代糧價資料庫》。資料化基礎之上，文本挖掘的發展還比較有限，其代表是臺灣歷史數位圖書館。

數位人文導向，提供文本挖掘能力將是今後歷史文獻資料庫開發的大趨勢。但是，如何資料化？研發怎樣的工具能夠實現文本挖掘？中文史料數位化的進程中，以上問題還尚在探索之中，成熟的案例並不多。因此，我們在開發《中國地方歷史文獻資料庫》時，將以上問題作為我們的研究焦點。

三、基於文獻性質的資料庫結構與分析工具研發

我們在開發《中國地方歷史文獻資料庫》的過程中意識到，對文獻進行有效的資料化，並開發有效的分析工具，必須以對文獻性質的深入研究為基礎。資料庫開發中，我們主要面臨兩個問題，第一，如何針對地方歷史文獻的文獻性質，進行有效的資料化。資料化不僅僅是文字錄入，更重要的是為文獻設計中繼資料（metadata）。利用中繼資料標引並標準化文獻中的資訊，才有可能將文獻中的描述內容轉變為可分析的資料。

第二，如何從數位人文的理念出發，開發更多有助於研究者的分析工具。今天電腦技術能夠提供的分析功能非常多，但不同的軟體、分析工具，都對資料類型有特定的要求，因此需要考慮特定的文獻類型可以被處理成怎樣的資料形態，並據此做針對性的分析工具開發。為了解決這兩個問題，首先必須對地方歷史文獻的性質做一分析。

---

[15] 項潔、陳詩沛、杜協昌：《臺灣古契書全文資料庫的建置》，“第三屆臺灣古文書與歷史研究學術研討會”，逢甲大學歷史與文物管理研究所，2009 年 3 月 14 日，1-19 頁。

[16] 涂豐恩、杜協昌、陳詩沛、何浩洋、項潔：《當資訊科技遇到史料——臺灣歷史數位圖書館中的未解問題》，項潔編：《數位人文研究的新視野：基礎與想像》，臺灣大學出版中心，2011 年，第 21-44 頁；項潔、翁稷安：《數位人文和歷史研究》，項潔編：《數位人文在歷史學研究的應用》，臺灣大學出版中心，2011 年，第 11-20 頁。

本文所討論的地方歷史文獻，主要指兩類材料，一類文獻是留存於民間，產生於民間的日常生活，以手寫為主，未經過出版暨知識再整理的過程，也可稱之為民間歷史文獻或民間文書。[17] 這類文獻的具體內涵，我們曾另文敘述。[18] 另一類文獻是由地方政府形成的各種檔案。這裡所說的地方政府主要指作為“親民之官”的縣級或次縣級行政機構，對於明清時代來說，也可包含府（州、廳）級行政機構。如民國時期江津縣保留了 2 萬餘卷司法訴訟檔案，通過這些檔案，可對 20 世紀上半葉的江津地方社會做深入研究。

這類材料與一般意義上的“古籍”具有不同的文獻學特徵。古籍是經過有意識的書寫與知識再組織之後形成的，地方歷史文獻的文本形成後，沒有經過知識再組織的過程，而是在經歷了功能性使用的週期後，就被以其使用中的形態保存起來。這意味著，首先，這類文獻的每一件都是獨一無二的，幾乎沒有複本。進而，由於沒有複本且未經過知識再組織，這類文本並不形成版本，古籍則具有抽象概念的“書”與作為實體的“版本”之間的分離。[19] 這意味著整理地方歷史文獻時，版本整理、校勘不是主要難點。

地方歷史文獻與古籍的另一個差異是，地方歷史文獻更多情況是碎片化的，單個文本的字數少，古籍整理中所注重的文本內關係，如篇章順序、自校等，在地方歷史文獻中雖然也存在，但不是非常顯著。整理地方歷史文獻時更注重文獻之間的關係，以明清史學界整理過程最久的徽州文書為例，以下學者們所提出的徽州文書特性，或可啟發我們理解地方歷史文獻的特性。

表 1 徽州文書特性歸納

學者 / 提出特性	周紹泉	臼井佐知子	中島樂章	嚴桂夫 王國鍵	劉伯山
	具體性	史料豐富	原始性	系統完成	唯一性
	連續性		時間跨度大	年代跨度大	連續性
	啟發性		數量大	數量大	
	真實性		史料豐富		
	典型性				

（資料來源：周紹泉：《徽州文書與徽學》，《歷史研究》，2000 年第 1 期；臼井佐知子：《徽州文書と徽州研究》，載森正夫編：《明清時代史の基本問題》，汲古書院，1997；中島樂章著、郭萬平、高飛譯：《明代鄉村糾紛與秩序：以徽州文書為中心》，南京：江蘇人民出版社，2006；嚴桂夫、王國鍵：《徽州文書檔案的特點與價值》，《檔案學研究》，2001 年第 1 期；劉伯山：《徽州文書的遺存及特點》，《歷史檔案》，2004 年第 1 期。）

[17] 鄭振滿：《民間歷史文獻與民間文化傳承研究》，《東南學術》2004 年第 1 期，293-296 頁；梁勇、鄭振滿、鄭莉：《新史料與新史學——鄭振滿教授訪談》，《學術月刊》2012 年第 4 期，第 155-160 頁。
[18] 趙思淵、湯萌：《上海交通大學新藏地方歷史文獻的分類法及其依據》，第 76-87 頁。
[19] 喬秀岩：《古籍整理的理論與實踐》，《版本目錄學研究》第 1 輯，國家圖書館出版社，2009 年，第 7 頁。

周紹泉先生認為徽州文書具有真實性，因為徽州文書是從實際生活中直接形成的檔。他所說的典型性則是指利用徽州文書可以形成一個個具有代表性的個案研究。中島樂章所說的原始性，其含義接近與周紹泉先生所述的真實性，特別強調徽州文書來自實際生活。另外，中島樂章所說的豐富性是指："徽州學研究的最大優勢在於，以徽州文書為中心，大量地保存了長時期族譜等文獻史料和建築等非文獻史料。……有可能恢復包括民眾文化、日常生活在內的一個地方社會的全貌。"[20] 嚴桂夫和王國鍵所說的系統完整，與劉伯山所述的連續性具有相近含義，均強調徽州文書的來源是可追溯的，文書之間的內部聯繫是有機的，可以復原的。

以上各位代表性學者所提出的徽州文書特性，可以歸納為以下共同點：第一，所有學者都認為徽州文書存量之大，內容之豐富，是同時代其他文獻群難以匹敵的。第二，相對於傳世文獻，徽州文書的特別之處是其保持了原始記錄，同時具有完整的，有機的文獻內部聯繫。

地方歷史文獻的單件當然也具有研究價值。以契約文書為例，傅衣凌、章有義、楊國禎等前輩學者都曾依據一件件獨立的、經過選擇的契約解釋明清鄉村的地權結構。但隨著研究的深入，對單件文書的分析、考釋，常常不能滿足研究的需要，即使在傅衣凌先生開創契約文書研究的時期，當他對契約文書內容和類型進行了解釋和考釋之後，也轉入了以時間、地域等關係對多件契約做綜合分析的研究。可以說，地方歷史文獻碎片化的形態決定了其每一個單件的研究價值通常要置於一個整體中才能被發現，也即其研究應當以一個"文獻群"為單位展開。

以上差異決定了，地方歷史文獻不能使用既有的古籍資料化方法。多數古籍的資料編目，都可參照現代書籍標準。但在地方歷史文獻的文獻結構中，著作人、出版方、出版地點等都是不主要甚至是不存在的資訊。因此，必須設計針對性的中繼資料方案。

資料庫開發實踐中，我們參照圖書館界通行的都柏林原則（Dublin Core）設計了事主、題名、時間、地域、文獻類型等中繼資料項目。資料庫中中繼資料格式主要實現兩種功能。其一是識別每一件文獻，並說明文獻的性質，如文獻編號、資源類型。其二是對文獻內容的描述，地方歷史文獻所涉及的內容千差萬別，設計能夠適用於全部文獻的中繼資料是非常困難的。因此中繼資料的設計必須具有高度的彈性，能夠涵納多數文獻，如文獻名稱、涉及人名（事主）、文獻歸戶、日期等，幾乎所有文獻中都具備。但另一方面，針對存量特別多的文獻，也需要針對性設計。從目前粗略的統計看，契約、帳簿占到文獻收藏的 60%左右，因此也設計了如標的、金額等此類材料特有的元素。[21] 從資

[20] （日）中島樂章：《明代鄉村糾紛與秩序：以徽州文書為中心》，郭萬平、高飛譯，江蘇人民出版社，2010 年，第 43 頁。

[21] 張潔、李芳、湯萌：《契約文書描述性中繼資料規範設計與應用》，未刊稿。

料中提取中繼資料可以採用人工與半自動標記（semi-automate tag）甚至全自動的方式。上海交通大學目前採取的是人工編目的方式，但是社會學界及數字人文領域已有一些可應用于中文文獻的半自動標記工具[22]，可以預見，這將成為今後的一個趨勢。

四、歸戶：制度史源流、整理方法、中繼資料

以上中繼資料格式中，歸戶是我們首創的中繼資料項目。這個中繼資料項能夠幫助使用者感受到文獻本來的特性，也是進一步開發分析工具的基礎。“歸戶”中繼資料項體現了我們提出的基於對文獻性質的理解構建中繼資料結構的資料庫開發理念。

之所以提出這項設計，是因為我們面臨一個困境：地方歷史文獻與書籍存在文獻性質的差異，其研究價值必須以一個“文獻群”為單位，那麼，如何確定一個文獻群的範圍？如何在資料化中體現一個文獻群的內在聯繫？

一個具有研究價值的文獻群，應當是一組具有內在邏輯關係的文獻所組成的整體，特別是那些由生產自同一個來源的文獻所形成的整體，如出自同一個家族的全部文書，或同屬一個案卷（record）的全部檔案。凡是屬於同一個文獻群的文獻，即使是在研究者看來可能並無價值的殘件，整理時都應當全部收錄。《石倉契約》的整理與研究過程中，以上方法被歸納為“有機”的研究方法。[23]

進而我們發現，檔案學中的全宗原則、來源原則對如何界定一個文獻群有直接的借鑒意義。如果參照全宗原則與來源原則，來自明清賦役制度以及徽州文書的“歸戶”概念則是最有效界定文獻群的方法。

全宗原則和來源原則是 19、20 世紀之交檔案學逐漸發展出的檔案管理理論。16 至 18 世紀的歐洲國家，其檔案管理本來依據“事由原則”，即按照檔案內容對檔案進行分類保管。19 世紀之後，本來的王室檔案館與行政機關檔登記室逐漸轉變為國家檔案館，並且從封閉保密轉為開放查閱，檔案來源與檔案查閱需求也隨之多元化，因而，本來封閉的，依照邏輯進行主題分類的檔案管理辦法不再能滿足需要。有的檔案可以歸入多個分類，或者有的檔案不能按照現有分類歸檔，都給檔案管理造成困難。

1841 年，法國內政部第 14 號通令頒佈省檔案館條理，規定：“來源於一個團體、一個機構、一個家庭或者一個人的所有檔都要組成全宗；檔案管理人員不得把全宗拆散或將不同的全宗混在一起。”[24] 這一條例所規定提出了“尊重全宗原則”（the principle of

[22] 何浩洋（Hou Leong Ho）：《MARKUS：中文古籍半自動標記平臺》，www.academia.edu，發佈：2014-12，訪問：2015-11-27。

[23] 蔣勤：《清代石倉文書的“在地”與“有機”分析》，《上海交通大學學報（哲學社會科學版）》2014 年第 3 期，第 88-98 頁。

[24] 馮惠玲：《論檔案整理理論的演變與發展》，載吳寶康、丁永奎：《當代中國檔案學論》，中國檔案出

respect pour les fonds），成為“來源原則”、“全宗原則”之濫觴。

之後，1881 年德國國家檔案館發佈《國家機密檔案館檔案整理條例》，其中提出“國家機密檔案館內檔按其組成部分的來源進行整理”及“每一機關一旦開始移交檔，就要立即指定一部分庫房專放該機關的檔，在這部分庫房內，官方檔要保持它在有關機關活動過程中獲得的順序和標誌。”即“登記室原則”，此原則之後發展為“來源原則”。[25]

來源（provenance）在檔案學中指“向檔中心或檔案館移交檔之前，在事務活動過程中形成、保管和/或利用檔的組織或個人。”在此基礎上，來自一個組織或個人的全部檔案應當作為一個單獨的整體保存，不同來源的檔案不能混合，這就是現代檔案學中通行的“來源原則”。[26] 根據來源原則，檔案保管必須保持檔案的“來源聯繫”。[27] 也就是說，應當以文獻產生時的來源單位作為文獻保管的基本單位，從而避免打破文獻之間既有的有機聯繫。在整理文獻時，應當區別針對文獻實體的分類法和文獻內容的分類法[28]，通過兩套分類法的綜合編目，達到對文獻的整體使用。

地方歷史文獻中的每一個文獻群，正如同檔案學中所說的“來源”。近年民間文書整理中所提出的“歸戶”概念，與來源原則有相似之處。歸戶是一個來自明清賦役制度的概念，意指賦役過割至地權買入人戶，如清初陸隴其總結地方官的為政經驗，“受業之家”即地權買入方應當“割稅歸戶”，這裡的“歸戶”是一個動詞，為歸入買人戶之意。明清之際的賦役制度改革中，“歸戶”是一個總體性的原則。[29]

> 夫有田則有賦，頑猾抵官者，誠所當治，而善良樂輸者，要當與之覆議。其大要，則於移割宜加意焉。產去稅存，不可不察，民又以出業報者，便當關會受業之家，割稅歸戶，然後卻、與、除、退，庶幾無泛追、無濫罰、無推攤抵捱之弊。[30]

“歸戶”在明末演變為一個名詞，徽州文書中存在“歸戶親供冊”、“歸戶清冊”等賦役冊籍[31]，通常是一個納稅戶所有應納稅糧之土地的登記，與陸隴其所稱之“歸戶”涵義相通。根據目前學界對清代賦役制度的理解，這些納稅戶通常是一些虛擬戶名，其背後可

版社，1988，第 115-167 頁。

[25] 黃霄羽：《魂系歷史主義——西方檔案學支柱理論發展研究》，中國人民大學出版社，2006 年，第 35 頁。

[26] 黃霄羽：《魂系歷史主義——西方檔案學支柱理論發展研究》，第 31-32 頁。

[27] 馮惠玲、何嘉蓀：《全宗理論的實質——全宗理論新探之二》，《檔案學通訊》1988 年第 5 期，第 10-13 頁。

[28] 馮惠玲、李憲：《中國檔案分類法的理論與使用方法》，《山西檔案》1989 年第 2 期，第 44-46 頁。

[29] 劉志偉：《在國家與社會之間：明清廣東地區裡甲賦役制度與鄉村社會（修訂版）》，中國人民大學出版社，2010 年，第 201 頁。

[30] 陸隴其：《蒞政摘要》卷上第 12 頁，《官箴書集成》第 2 冊，黃山書社，1997 年，第 628 頁。

[31] 欒成顯：《萬曆九年清丈歸戶親供冊研究》，《中國歷史博物館館刊》1996 年第 2 期，第 79-93 頁。

以是個人、家庭、宗族、會社或其他社會團體。[32] 這些“戶”是納稅單位，同時也即經濟活動的單位，進而也是產生契約、帳簿等民間文書的基本單位。

整理、研究民間文書的學術史中，劉伯山較早將“歸戶”作為一項原則，認為徽州文書具有歸戶性。[33] 他在編輯《徽州文書》時將同屬一個家族的文書稱為歸戶文書。此後，越來越多學者將“歸戶性”視作民間文書的一項重要特性，研究者在整理清水江文書、太行山文書時，也開始重視歸戶整理的方法。[34]

正如檔案學對“來源”的理解越趨複雜，隨著文獻收集越來越豐富，作為文獻收集、整理基本單位的“戶”、“歸戶”也應當具有更豐富的內涵。事實上，早在 1962 年嚴中平先生已經提出一項針對收集工作的建議，希望能夠“完整地”收集徽州文書。[35] 我們認為嚴中平先生所說的“完整”已經包含了“歸戶”的整理原則。

正如檔案保管從事由分類轉向來源分類，保管、整理地方歷史文獻也應當以文獻群或“歸戶”作為基本單位，從而取代按照內容、年代等進行整理的原則。[36] 因為這些文獻本來是以文“戶”為單位產生的，以“戶”或文獻群為單位進行保管、分類，最能夠保持文獻內部的有機聯繫。同時，“戶”的所指也應更加豐富，舉凡家戶、家族、宗族、會社、寺廟等都可成為一“戶”。

因此在《中國地方歷史文獻資料庫》中，“歸戶”成為一個中繼資料項目，設計為“縣+姓氏”的形式，根據收集文獻時獲得的資訊，標注每件文獻所屬的縣份及姓氏，由此反映文獻與當地人群之間可能存在的關係。在徽州及浙南等文獻脈絡更清晰的地方，文獻的歸戶資訊還可細化到縣以下層級，也即其所屬的“都”、“圖”、村落等。但縣以下行政區劃層級幾乎每一縣均不相同，因此在按照“歸戶”資訊檢索的介面中，省去了縣以下層級，而在中繼資料中，則以文字形式保留了這些資訊。

為了彌補以上不足，中繼資料中又增加“批次”資訊，這是收錄於《中國地方歷史文獻資料庫》中每個文獻群的編號，此編號是根據各文獻群入藏時間製作的，文獻群中的每件文獻則在批次號的基礎上流水編號。批次號是對文獻群物理保存形態的反映。

---

[32] 劉志偉：《在國家與社會之間——明清廣東地區裡甲賦役制度與鄉村社會》，第 197-204 頁。

[33] 劉伯山：《徽州文書的遺存及特點》，《歷史檔案》2004 年第 1 期，第 125 頁。

[34] 張應強：《清水江文書的收集、整理與研究芻議》，《原生態民族文化學刊》2013 年第 3 期，第 33-38 頁；喬福錦：《歷史文獻學視域中的鄉村社會文獻整理》，《遼東學院學報（社會科學版）》2011 年第 3 期，第 103-112 頁。

[35] 嚴中平致中央檔案館函（1962 年 2 月 6 日），安徽省檔案館藏，轉引自嚴桂夫、王國鍵：《徽州文書檔案》，安徽人民出版社，2003 年，第 11 頁。

[36] 民間歷史文獻整理方法演進的學術史，參見楊培娜、申斌：《走向民間歷史文獻學——20 世紀民間文獻搜集整理方法的演進歷程》，《中山大學學報（社會科學版）》2014 年第 5 期，第 71-80 頁；張侃：《20 世紀以來民間文獻研究的學理述略》，“第七屆民間歷史文獻論壇”，廈門大學，2015。

五、文本挖掘：拓展資料庫應用的可能性

研發《中國地方歷史文獻資料庫》時，由於資金與技術的限制，並未設計嵌入資料庫的文本挖掘工具，但從數位人文的理念出發，設計了兩種檢索方式以及兩組檢得文獻分析工具。研究者利用這些工具，就有可能進行進一步的文本挖掘與研究。

資料庫提供的基本檢索方法是輸入任意詞在整個資料庫中檢索，或者以下拉清單方式在題名、事主、歸戶、事由、分類中用任意詞檢索，也就是通常文獻資料庫都具備的普通檢索與高級檢索。另一種檢索方式是多維分類導航，也即利用時間、地域、歸戶、分類法等方法交叉流覽、檢索，尋找文獻。對於檢索所得文獻，資料庫提供兩種分析工具，一種可以統計檢得文獻的地域分佈、年代排序、類型分佈，以及事主統計。另一種工具則可顯示檢得文獻的關聯文獻，如屬於同一批次、同一地域、同一歸戶或同一事主的文獻及其數量。

建立以上檢索與分析工具的意義是為研究者提供更好的研究環境。通過檢索找到資料庫中的資料，是研究者使用資料庫的最基本需求，但是，歷史學研究不僅需要找到資料，更重要的是發現資料間的關係。歷史學研究中，文本記錄中的時間、空間、人物無疑是最重要的三組關係，我們試圖在資料庫中提供相應的功能說明研究者揭示資料群在這三個方面的關聯性。依據時間檢索、檢得資料時間分佈統計正是為發現不同時間形成的資料間的關係而設計的。地域、歸戶等則是以不同形式分析、呈現資料的空間分佈。文本中所有的人物資訊則都被登記為事主。

以上功能設計還有可能進一步說明研究者發現新的資料或新的研究議題。以我們最近的一項研究為例，本來的研究計畫中，只是準備分析清代徽州契約中代筆人與買賣雙方的親屬關係，在利用事主相關功能檢索一批文書中代筆人的身份時，發現在一批契約中的代筆人江振玉同時還是當地編制歸戶冊的冊書，由此開始研究清代鄉村中同時擔任半職業代筆與稅收職役的人群。[37]

並且，編制中繼資料時提取了文書中的全部人物，也就有可能分析文書所反映的社會網路。以上述研究為例，根據該家族保存的 100 餘份契約，可立該家族清代、民國時期的土地交易記錄資料庫，圖 1 是根據資料庫繪製的當地土地交易社會網路，中可以觀察到 300 餘年間該家族有實力大宗購入土地的主要成員，及該家族購買土地時主要使用的戶名。利用同樣的方法，還可以分析當地土地出讓的主要趨勢，即圖 2。此外，利用分家書、家譜等其他資料，還可進一步分析圖 1、圖 2 中與該家族頻繁交易的人物身份，由此可瞭解當地社會關係網路對土地交易的影響。

[37] 趙思淵：《19 世紀徽州鄉村的土地市場、信用機制與關係網絡》，《近代史研究》2015 年第 3 期，第 96 頁。

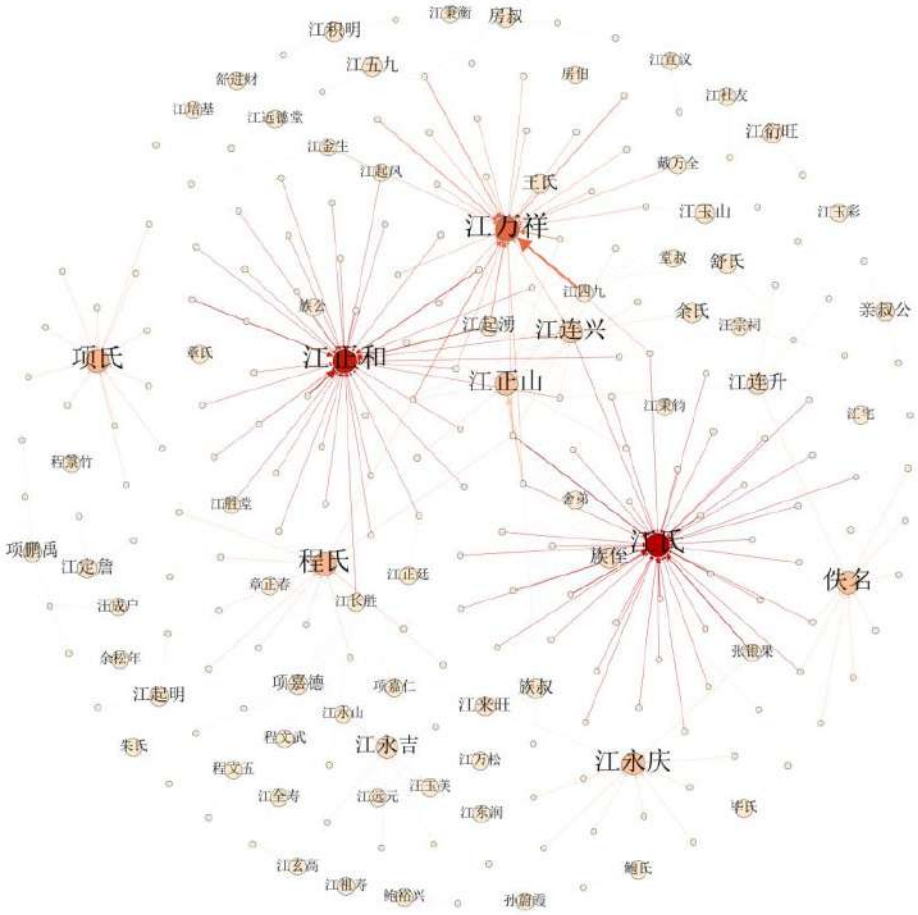


圖 1 歙縣棡橋江氏地權交易網路（獲得地權）

（資料來源：《中國地方歷史文獻資料庫》，http://www.datahistory.cn，批次號：0111120601。

說明：圖中以箭頭表示土地權利轉讓的方向，如圖中箭頭從江四九指向江萬祥，表示土地權利從江四九轉讓至江萬祥。圖中每個點的顏色表示其在交易中購入地權的次數，頻率越高，顏色越深。這裡所說的獲得地權，包括買入、典入、抵押等形式。）

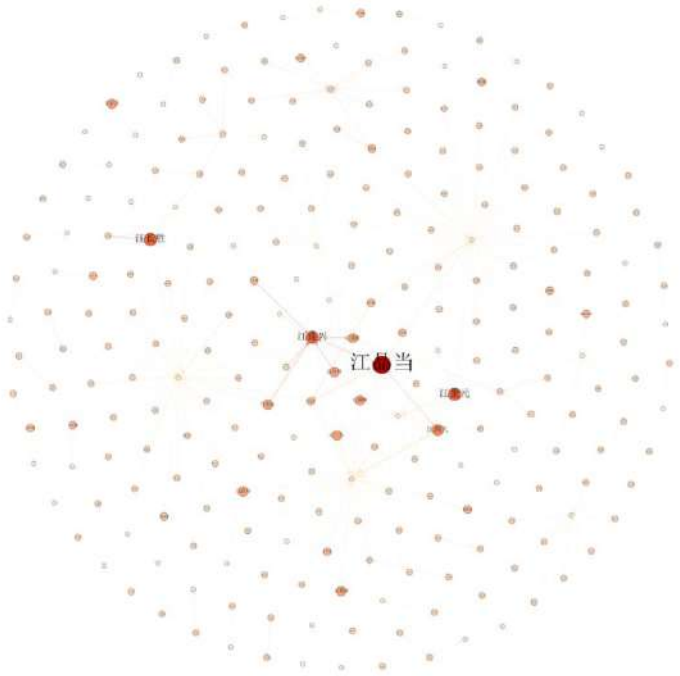


圖 2 歙縣棿橋江氏地權交易網路（出讓地權）

（資料來源：《中國地方歷史文獻資料庫》，http://www.datahistory.cn，批次號：0111120601。

說明：圖中每一個點的顏色、字型大小表示其地權出讓的次數，顏色越深、字型大小越大，表示地權出讓越頻繁。這裡所說的出讓地權，包括賣出、典出、抵押等。）

六、結論：建立史料資料庫的文獻學方法

數字人文是最近 20 年來新興的交叉學科概念，對歷史學研究而言，引入這一概念的意義是促進更多分析工具應用於文獻整理與解讀。與古籍不同，地方歷史文獻未經過知識重組，也相對碎片化，更注重文獻間的關聯性。《中國地方歷史文獻資料庫》針對文獻特性設計中繼資料結構，從而實現對文獻的多維度檢索。尤其是我們根據文獻特性所提出的“歸戶”資料項目，將幫助研究者發掘文獻的內在關聯。我們不僅期望這些功能設計可以便利研究者尋找史料，更期望以此幫助研究者發現新的研究議題。

中繼資料是可以被電腦識別的文獻描述，製作中繼資料，也就是將文獻資料轉化為可被檢索、分析的資料的過程，這是將物理形態的史料轉化為可分析的數字形態的關鍵。如何設計中繼資料結構，很大程度上決定了文獻資料可以被如何檢索、分析。另一方面，資料庫的使用者必須瞭解中繼資料結構，從而判斷哪些因素可能影響了分析結果。

因此，歷史文獻資料庫不僅是傳統史料的載體或“倉庫”，其本身也將日漸形成一種獨立的文獻形態。歷史文獻學對傳統史料已形成一套綿密、精細的處理方法，資料庫作為一新文獻形態也應當建立針對性的文獻學方法論。對中繼資料結構的考辨可能應當是此方法論的核心。史學理論對史實與史料關係的思考，也同樣適用於歷史文獻與資料庫。

Towards a Dynamic, Scalable Digital Library of Pre-modern Chinese

Donald Sturgeon*

Abstract

This paper contrasts two radically different approaches to full-text digital library design and implementation: firstly, the "static database approach", in which materials are firstly created, edited, and manually reviewed before being added to a generally static database system; secondly, dynamic approaches in which incompletely reviewed materials are imported into a dynamic system providing similar functionality, but within which significant further editing is intended to take place. To illustrate the technical challenges, benefits, and practical consequences of these two design approaches as reflected in a large-scale digital system, specific examples are drawn from the Chinese Text Project (http://ctext.org) digital library, which initially began as a primarily static database system, and has over time evolved into a primarily dynamic platform. This change has been motivated in particular by a desire to achieve a scalable, sustainable platform for the curation of textual data and metadata, to which new material can be easily added as well as improved over time, while requiring minimal administrative overhead. This paper argues that while there are technical challenges to a dynamic approach, the increase in scalability dynamic approaches offer can have significant advantages, including potential access to a "long tail" of data which might otherwise in practice be overlooked.

Keywords: chinese, digital library, database, wiki, API

* Postdoctoral Fellow of Harvard University. Email: djs@dsturgeon.net.

邁向動態擴充的前現代中國文學數位圖書館

德龍*

摘　要

本文比對兩種截然不同的全文數位圖書館之設計、實作研究取徑。將資料經建立、編輯與人工檢視等流程查核後，再於資料庫中更新資料，此為靜態資料庫取徑。動態取徑則允許於資料庫輸入未經完整流程檢視之資料，除提供類似資料庫的服務外，亦可繼續進一步編修資料內容。以中國哲學電子書計劃（Chinese Text Project，http://ctext.org）而言，此計畫於初期主要採用靜態資料庫建立，之後逐步轉變為動態平台。本文將以此計畫為例，說明兩種取徑於大型數位化系統建置過程中所反映之技術議題、優缺點及成果。此一取徑上的轉變，目標為希望以最少的管理成本，建立一個具可擴充性與持續性之平台，不僅可以用來匯集文本與其詮釋資料，也能輕易地隨時增補新的資料。本文認為，儘管動態取徑之技術要求較高，但在擴充能力上具備明顯優勢，尤其是對靜態取徑所忽略的長尾資料，更具有取用潛力。

關鍵字：中文、數位圖書館、資料庫、維基、API

* 美國哈佛大學博士後研究員，Email: djs@dsturgeon.net.

1. Introduction

The Chinese Text Project (http://ctext.org/) began life in 2005 as a web-based research tool for a handful of relatively short classical Chinese texts, implementing simple full-text search, cross-reference, and navigation functions. Since then the project has expanded significantly in scope and functionality to become one of the largest digital libraries of pre-modern Chinese. This expansion has been made possible by radical changes in both design and implementation, as well as strategies for curation and editing. Most significantly, it has necessitated a change from a traditional centrally edited database system to a version-controlled format editable by large numbers of geographically distributed contributors. These changes have been motivated in particular by a desire to achieve a scalable, sustainable platform for the curation of textual data and metadata, to which new material can be easily added as well as improved over time while requiring minimal administrative overhead. In an era in which statistical and "big data" studies of these materials are receiving increasing attention, aspects of these changes and how they have been implemented are likely to have relevance to the design of digital libraries and archives more generally.

In this paper, I contrast what I term the "static database approach" – in which materials are firstly created, edited, and manually reviewed before being added in their final form to a database-driven system whose contents are intended to be primarily static in nature – with dynamic approaches in which incompletely reviewed materials are imported into a system and made available immediately, with the expectation of further editing taking place from within this same platform. This latter class of system includes in particular widely used web-based systems such as content management systems and wikis, many of which are implemented using relational databases, but which provide editing functionality of some kind through their own distinct (not necessarily database-like) user-interface. I present aspects of the design and implementation of the Chinese Text Project as examples of some of the advantages and challenges of these two distinct design approaches as reflected in a large-scale digital library project.

2. Static Databases

"Static databases" in the sense used in this paper are database-driven systems which are static in that their contents are primarily intended to be fixed and unchanging. Such a database typically represents the culmination of a great deal of prior work: in the case of a database of

historical textual material, a long process of data selection, preparation, proofreading and formatting, as well as in some cases textual criticism and emendation by experts. The contents of the database may be updated over time, but this is performed centrally by a small team of individuals and editors affiliated with the project, and such updates are usually infrequent – perhaps on the order of once or twice a year. Well-known examples of this type of database project for historical Chinese texts include Academia Sinica's Scripta Sinica,[1] the Chinese University of Hong Kong's CHANT,[2] and a number of commercial database projects.

Static databases have many advantages which make them convenient to work with from both scholarly and technical points of view. Since their contents are not designed to change over time, typically a large amount of effort is put into ensuring the accuracy of their contents before they are published, in much the same way as would be the case with a printed scholarly reference work. This makes them attractive as "fixed targets" both in terms of scholarly citation, and in terms of requirements and expectations of how the software powering them should work: no technical allowance needs to be made for aspects of their contents changing, except during times of maintenance and upgrade. Textual references, links, and metadata within such databases are easily maintained technically, because the objects they relate are essentially fixed. Search indexes and other derived data supporting database functionality can be computed once and then left alone entirely until the next planned update for the same reason – and the length of time required to generate these is effectively irrelevant since they need only be modified on a very occasional basis.

3. Importance of Scalability

Despite the theoretical attractiveness of the static database, there are practical reasons why other approaches may also be worth considering, and these primarily relate to issues of scale. The Chinese Text Project is currently one of the largest textual databases of pre-modern Chinese writing, containing over 5 billion characters and 25 million pages of scanned source material. This naturally represents only a fraction of the entire corpus of pre-modern Chinese writing however, which includes both countless works which have not yet been digitized, as well as innumerable more distinct editions of the same works, all of which researchers would ideally like to have access to in digital, full-text form.

[1] http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm
[2] http://www.chant.org/

Creating a database of this material using traditional approaches requiring near-perfect transcriptions to be created before ingesting these into a static database tends to become cost prohibitive for very large volumes of data. Supposing, somewhat optimistically, that accurate input (or OCR with manual post-correction) of these historical texts could be completed at a cost of $1 per page, this would result in costs of $25m for initial transcription alone on this volume of data – and orders of magnitude higher for the full available corpus. While scholars may feel that investments such as these are justified and worthwhile, in practice funding is limited, and as a consequence the volume of data which can be handled by these techniques is necessarily curtailed.

One response to this type of limitation might be to simply incorporate available imperfect materials, such as unedited OCR (Optical Character Recognition) results, into a static database system and solicit corrections from users of the resource. Even with a static database approach in which content is not generally expected to change over time, improvements to publicly accessible material can still be proposed by users of an information system, for example through the reporting by e-mail or other means of corrections, emendations, and other suggested changes to a centralized group of curators who have access to a private editing interface of some kind.[3] Such an approach however introduces an obvious bottleneck which will become an increasing problem as the size of the dataset and user base increases, as every suggested correction must be first evaluated, and then, if appropriate, applied to the database by one of these curators. Additionally, since those proposing the changes have an expectation that a human being (most likely an expert) will be reviewing the submitted information, the content, format, style, and completeness of each such submission will often vary significantly.

A natural approach to address this type of issue is to streamline the mechanism by which changes are proposed, for example by the addition of a form with fixed fields which must be completed in a certain way to describe various aspects of a proposed modification. This potentially reduces the editing effort required per individual change, though the cost of evaluating the changes remains largely unchanged.

The reasonable expectation – particularly with respect to scholarly resources – that data should be as accurate as possible, when combined with the poor scalability of centrally edited databases with respect to the implementation of proposed corrections as well as limitations of

[3] Many static database systems do in fact incorporate this type of feature – an implicit acknowledgement that while their accuracy rates are high, they are not infallible.

available funding, thus place natural limits on the scope of material which can realistically be included in a static database project. Where large amounts of additional data are available in a form that is immediately useful but still requires significant amounts of editing – as is typically the case with data produced by imperfect procedures such as OCR or machine translation – compilers may be faced with a choice between including only a small fraction of the available material, or an inability to process the large volume of required corrections.

This leads to broader questions about the goals of scholarly information systems – for instance, is it preferable to have a small collection of highly curated, reliable data, or a much larger collection of potentially less reliable material? Is it possible to combine the two in a useful way? In principle, highly curated, reliable data is always preferable, but in practice the resources are not always available to rapidly produce such data in the quantities that may be needed for statistical studies or with sufficient coverage to satisfy the needs of researchers working with less commonly studied or less mainstream materials.

Dynamic platforms provide one approach to addressing this issue. Taking a more radical step, and a departure from the static model, the group of potential editors of the material is expanded beyond the core group of those affiliated with the resource, to include a larger, distributed group of people, not necessarily experts or people directly affiliated with the project, perhaps including anonymous or pseudonymous users, and potentially, as with projects such as Wikipedia or Wikisource, anyone at all. This step typically entails its own technical and administrative requirements, including in particular versioning – storing in complete detail the changing state of a resource, so that edits can be evaluated retrospectively after they have been applied and prior revisions restored if necessary. This in turn requires some type of user interface to present the content of revisions and differences between them, as well as to allow their reversion – something which may become more difficult for systems with complex ontologies. One relatively simple solution is to provide a serialization of suitably chosen units of data, for example as an XML document or fragment, which can then be edited and versioned – and differences identified and highlighted – as plain text.

By replacing the traditional approach in which data input and proofreading are performed outside of the user-facing part of the system, it becomes possible to allow for imperfect data to be made use of where it represents the best data currently available. While this may initially seem contrary to the requirement that data used for research purposes must be accurate, in reality imperfect data can still serve a useful role when there is no better alternative available

– it is important to remember that in many cases, the choice being made is not one between imperfect data and perfect data, but between imperfect data and no data at all. Provided that users of such a system are alerted to the limitations of the data and use it accordingly, what is being offered is often preferable to the next best alternative.

An example of the utility of such imperfect data (though itself still primarily a static database approach) is the Google Books project, in which full-text search of large amounts of scanned material is accomplished by means of uncorrected OCR results. These results inevitably contain errors (and in some cases such as historical works, large numbers of errors); yet they are at the same time widely used for the simple reason that they provide a far more efficient way of searching many of these materials than any available alternative method, such as visiting a library, locating the volume, and laboriously searching through it by hand. Such use may be problematic or unproblematic depending on how the resource is used: for example, in the very common case of a researcher wishing to identify a particular passage of text or the exact source of a quotation, there is effectively no scope for error due to imperfections in the OCR, since the user merely *locates* the passage using the OCR results, while confirming its contents by means of the scanned representation; by contrast, it would not be legitimate to use such data to confirm the *absence* of some piece of text within a volume using this type of data, since OCR errors could easily lead to a false negative. What the Google Books project has so far not attempted to do, however, is to co-opt users of the material to assist in identifying and directly correcting errors within it; while it makes use of imperfect data, it is still essentially a static database system, with no practical facility for users to contribute to the improvement of the material during their use of it.

4. From Static Database to Dynamic Platform

As the Chinese Text Project has grown from a small conventional database to a repository of texts many orders of magnitude larger in size, challenges such as those raised in the previous sections have arisen, and approaches to resolving them have been explored. Of these, the most significant break with the original static database design has been the development of a publicly editable wiki system for the curation of textual data. Unlike most mainstream wiki platforms such as MediaWiki, this imposes its own ontology in addition to the more common and general page-like structure employed by most wikis: all materials in this wiki consist of an “item”, which contains metadata about a text, and one or more “chapters”, which contain textual

content, and which are ordered as a single list within the item. This relatively simple restriction significantly simplifies many processing tasks compared with a more freely structured ontology, since the basic structure of any textual item can be inferred without requiring the parsing of any textual content – in addition to reducing processing overhead, loops and complex interdependencies of pages, which can easily arise in a simple page-based ontology in which pages can link freely to other pages, are not possible. All of these objects can be created, edited, and deleted by any user of the online system. Chapters themselves can contain various types of markup, for example describing the hierarchical structure of their own contents or their relation to a scanned representation of the same content in its source edition. Each chapter therefore has a serialization which can be edited directly or through special-purpose editing interfaces, which can be compared to highlight modifications between versions, and which can be parsed to extract textual and non-textual content as needed. Data requiring collation from many sources in real time can be extracted from the wiki serialization when changes are committed, and if necessary stored in derived database tables for rapid retrieval.

The first iteration of this public wiki system provided an editing interface similar to the traditional page-based editing interface of MediaWiki software (Figure 1): a user could elect to edit a chapter or text, correct it, and then submit the corrected version; a versioning system was also provided to record edits, show differences between edits, and revert edits. Users were asked to specify, where possible, the evidence for their emendations in the form of a link to the page of scanned material corresponding to the change. Modest numbers of contributions were received through this interface – on the order of tens or hundreds of corrections per month.

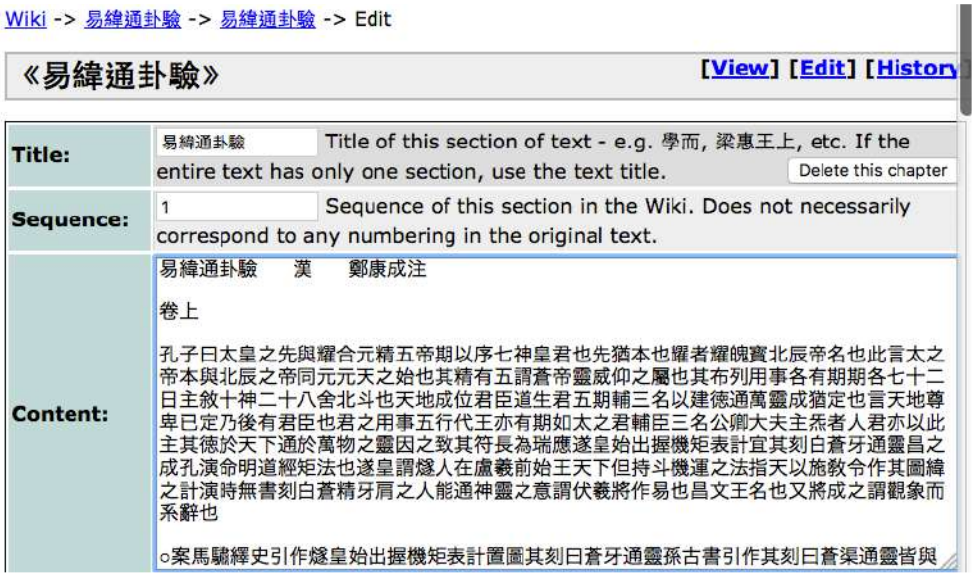


Figure 1: WikiMedia-like editing interface.

A subsequent addition to the project implemented a more ambitious and specialized editing interface, which allowed editing of the same underlying data but in a more intuitive way. This interface (Figure 2) consists of a side-by-side image and transcription view, which visualizes the underlying textual data as it corresponds with images of pages of the edition with which the textual transcription should agree. This view provides a natural way of comparing image and transcription, and also a more intuitive method of submitting corrections at the page level. The interface allows editing of the same underlying textual data used elsewhere in the database, presenting a simplified interface to the underlying XML serializations actually used to store the data. Crucially, this means that while the underlying representation retains complex structured data, knowledge of complex conventions – such as what constitutes valid XML – is not required to correct the material. While the original editing interface is retained to allow for more radical reorganization of textual data when needed, the side-by-side editing interface greatly simplifies the most common type of corrections.

Figure 2: Side-by-side transcription editing.

The introduction of this “side-by-side” editing interface had an immediate effect on the rate of user contributions: soon after its release, the number of edits submitted by users increased from tens per month to hundreds per day (Figure 3). At the same time, the additional interface provided a streamlined mechanism for evaluating corrections. The accuracy of edits submitted via a particular page of a text can be assessed by displaying that page, and further

simplified by highlighting the altered segment; the new interface makes this a simple matter of following links from the edit log.

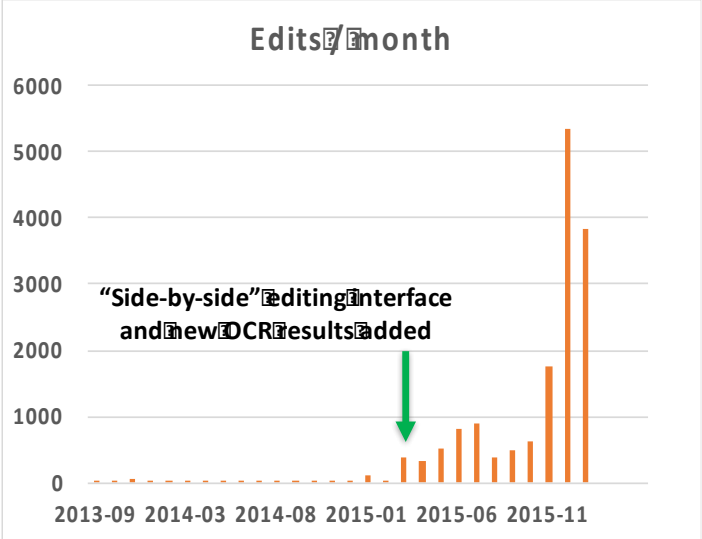


Figure 3: Crowd-sourced monthly edits to the Chinese Text Project Wiki section.

5. APIs and Integration with Diverse Scholarly Workflows

The primary motivation for employing crowd-sourcing in this project and enabling the inclusion of imperfect OCR results was to allow the inclusion of a larger amount of material than would otherwise have been feasible with a static approach. Part of the reason for doing this is the economy of scale involved in scalable digital systems: in contrast to the case of manual input or proofreading, in which doubling the volume of material can typically be expected to double the cost, after initial setup costs have been met the additional cost of performing automated procedures like OCR and database ingest on a similar volume of material is typically marginal. It makes sense to apply these techniques to as many materials of the same type as possible, because this maximizes the potential utility of the resource: while many users may be interested in a core subset of popular, commonly requested material, there is also a "long tail" of interest in less commonly studied material.

However, just as different individuals will be interested in different parts of the collection, so too they may wish to apply different tools and methodologies to their study of it. While a centralized system can provide direct access to commonly used functionality such as full-text search, it cannot hope to comprehensively account for all of the many possible scholarly use cases of these materials. One way to address this limitation is to create, in addition to a user

interface designed to be used directly by human users, technical means for integration with other software, such as Application Programming Interfaces (APIs). These provide fixed, documented mechanisms for the exchange of information among software components which may be created entirely independently of one another by different groups of people and without requiring close coordination of effort.

In the case of the Chinese Text Project, such a system aimed at facilitating decentralized collaboration and data exchange between projects has been implemented by means of three main components: an API allowing machine-readable access to textual data and metadata stored in the database, a "plugin" system, allowing individual users to define and install their own extensions into the site's user interface, and a system of textual identifiers used to uniquely refer to individual textual objects.[4] Plugins in this context are XML descriptions of a programmatic way of accessing an external resource while simultaneously providing the external resource with some sort of context-specific content. For example, a user looking up a term in the site's built-in dictionary may have a plugin for an external dictionary installed into their account, which will allow direct access to that same term in the external dictionary via a link from within the user interface.

The plugin system and API are also designed so that they can be used together to create more complex interactions. For example, a textual plugin may send a textual reference to an external site, identifying the particular section of some edition of a text currently open in the Chinese Text Project interface. The external site may then use this reference to request information about the resource from the API, such as its full-text content, metadata, availability of other editions, etc., and process this information using its own rules and logic. It might perform some statistical analysis, compare differences between editions, or do something else entirely – the key point is that this functionality can be created by others besides the developers of the library itself, yet immediately integrated into its interface in a natural way. Since plugins themselves are merely XML descriptions of how to access this external functionality, they are easily shared among users, and are designed to be installed by users who need not have any familiarity with XML – indeed, it is possible for an external resource to link directly to a page prompting a user to install their plugin, requiring no more work than clicking on a link labeled "Install".

[4] http://ctext.org/tools/api

This provision of API services facilitates a number of common use cases, but more importantly allows for adaptation to new and unforeseen uses, particularly when combined with other interactive functionality of the system. For example, one of the simplest applications of the API is implementation of a full-text export function, which in the most basic case allows download of full-text data as plain text. When combined with existing editing interfaces however, this aspect of the API makes possible the use of this database as a platform for, among other things, transcribing texts. While OCR results can be used as-is for full-text search and other purposes, individual scholars and research teams are also able to make use of the editing functionality to improve the quality of those texts which are important to them. A key advantage of this approach is that when this is done, not only is a transcription created for the use of that particular research project, but the transcription – together with the data linking all of its components to the original source images – immediately becomes part of a larger database system which can be expected to be maintained over the long term. In other words, while the system on the one hand provides an efficient mechanism for transcribing and proofreading texts, at the same time it also ensures that any effort put into such tasks will not be wasted, and that the corrected results can also be accessed by others in future.

A more sophisticated example of API and plugin use in practice is the MARKUS textual markup system for historical Chinese texts.[5] By means of the API and plugin system, it is possible for users to identify textual resources in the Chinese Text Project through its own interface, and then immediately import these into the MARKUS system with a single click, where they can be analyzed using MARKUS's own database of personal names, place names, temporal references, etc., and further manipulated and marked up using the latter's user interface which has been specifically developed for these tasks. In future, closer integration may also be possible – for example, automatically using metadata from the Chinese Text Project to determine how best to automatically mark up a given text (e.g. a Tang dynasty text cannot plausibly reference the name of a Ming dynasty individual), as well as to provide access to Chinese Text Project functionality (such as image search) from directly within the MARKUS interface.

6. APIs and Text Mining

[5] http://dh.chinese-empires.eu/beta/index.html

Since the data provided via API is produced in a single, consistent format, this also provides a convenient mechanism by which to extract larger amounts of data in bulk in order to apply statistical techniques such as text mining. The API itself is not tied to any particular programming language or environment, instead being accessible over the web just like an ordinary website – the only distinction being that the responses given are machine-readable rather than human-readable, and typically represented in JSON rather than HTML. Thus the API can be accessed from virtually any modern programming environment; additionally, for environments likely to be used by large numbers of people, it becomes practical to create lightweight "wrapper" modules tailored to particular programming languages, facilitating even more intuitive access to the data from within these environments. An example of this is the Python wrapper module,[6] which provides a simple set of functions returning data from the API in the form of standard Python data structures.

For this type of use case too, APIs offer advantages over alternative techniques such as simple database dumps or static copies of large volumes of data. Because textual data is assembled in real time, any necessary corrections to a text can be made prior to exporting it (or, if errors are encountered during processing, it can be corrected and the updated version immediately re-exported). This would likely not be possible with fixed bulk data downloads: given the billions of characters of text already stored in the database, and the constant stream of corrections and new additions, such inflexible methods of distributing content would inevitably take time to prepare, and thus lag behind the current state of the database itself.

APIs and appropriate client libraries also offer the possibility of obtaining the same underlying data in different forms depending on the desired application. When obtaining large amounts of data in bulk for text mining purposes, the structure of each text – how it is divided into paragraphs, chapters, and larger units – may be entirely unimportant, for example when building a statistical model of language use. For other types of analysis however, it may be necessary or advantageous to preserve aspects of the structure – for instance if comparing properties of different paragraphs, chapters or other textual units within or between texts, as might be done with topic modeling or in an authorship attribution study. Using an API and client library, functions can be provided to in the one case omit this unnecessary information – returning data simply as a string of text – while in the other case transform it into some convenient data structure.

[6] https://pypi.python.org/pypi/ctext

This aspect of the API and Python library has been of particular utility when using materials from the Chinese Text Project in teaching digital humanities methods to humanities students, for the simple reason that the convenience these mechanisms afford leads to simple yet clear and powerful programming constructs. This is illustrated by simple examples such as the program shown in Figure 4, which obtains the full text of the Dao-de-jing and prints out strings of characters matching a specified regular expression. This is a simple program containing only five lines of code, as might be used when teaching the principles of regular expressions, and this program can be expected to function reliably and predictably on any system which has the "ctext" Python module installed. Yet while the task this program performs is simple, it would be much more cumbersome in a teaching context to perform – as well as explain the relevant logic behind – even this same simple task without using this type of module. The natural way to do this would involve students downloading a file to their computer and having the program read the contents of this file into a variable. While a simple task for those with experience of programming, file input/output, and fully-qualified pathnames, this seemingly trivial task raises all sorts of unhelpful complications because it relies on an assortment of technical knowledge quite unrelated to the task at hand – to say nothing of further platform-specific complexities such as different pathname conventions on Mac OS and Linux versus Windows systems. Instead, the API and Python module allow a text to be loaded into a Python string variable using a single function, in a way that is far more intuitive to the beginning student. Moreover, given an awareness of the notion of Chinese Text Project textual identifiers – the "ctp:dao-de-jing" in this example – it is also entirely obvious how this example would be modified to perform the same task on a different textual object.

```
In [3]: from ctext import *

        laozi = gettextasstring("ctp:dao-de-jing")

        for match in re.finditer(r"足[^。，、；！？]", laozi):
            matched_text = match.group(0)
            print(matched_text)

        足者
        足見
        足聞
        足既
        足以
        足不
        足之
        足矣
        足以
        足下
        足者
        足以
```

Figure 4: Python program to apply a regular expression to a text.

Less trivial examples further indicate why this facility is a significant advantage in a teaching context. The program shown in Figure 5 obtains the contents of another text – the Analects – but this time instead of reading it into a single string, the data is read into a list variable with one paragraph in each list element. Again this could be accomplished using local files, but with considerably more work – and additional complexity distracting from the core logic being demonstrated (in this case, exhaustively listing social relationships of named individuals mentioned together in the same passage of the text for social network analysis).

```
from ctext import *

passages = gettextasparagraparray("ctp:analects")
people = ["有子", "子貢", "曾子", "子夏", "子游", "顏淵", "季路", "閔子騫", "冉伯牛",

print("graph {")

for passage in passages:
    for p1 in range(0, len(people)):
        for p2 in range(p1+1, len(people)):
            if people[p1] in passage and people[p2] in passage:
                print("  " + people[p1] + " -- " + people[p2])

print("}")
```

```
graph {
  樊遲 -- 孟懿子
  顏淵 -- 季路
  顏淵 -- 子路
  季路 -- 子路
  顏淵 -- 子路
  子貢 -- 子夏
  子貢 -- 子游
  子貢 -- 顏淵
```

Figure 5: Python program to output a network graph of relationships in the Analects.

Similar points apply to more complex examples: chapters of text can be read into variables without additional programming work, and for more complex tasks involving multiple texts, metadata can also be loaded automatically based on the same textual identifiers used to fetch textual data. This results in programs which can be easily modified and experimented with, performing their functions on any textual objects available in the library. A short Python program of about 40 lines of code allows experimentation with Principal Component Analysis on arbitrarily selected textual items and feature vectors, plotting the results on a graph with a legend indicating text titles and authors obtained via API. This level of flexibility decreases complexity to the point where there is no need to ask students to experiment with toy examples, nor to provide in advance a set of carefully prepared source material – working with any part of the corpus can be done with equal simplicity and through a consistent mechanism.

7. Challenges Implementing Complex Functionality

One practical challenge which has been faced with this particular digital library, and one often faced by digital systems migrating from a static to a dynamic platform, has been the integration of existing content from the static database format into the new dynamic wiki format. One approach (and perhaps the best long-term solution) is to re-implement existing functionality by serializing complex database content into a format which can be edited from within the wiki system itself, while simultaneously modifying query systems to operate directly upon data in this format. This comes however at a cost in complexity, both in terms of additional software development and the representations of data with which editors must interact. One of the most significant advantages of relational databases – and one which has contributed to their widespread use as "back ends" for systems with entirely distinct, non-database-like user interfaces – is the ease with which structured data can be represented, queried and further manipulated by software efficiently and in real time. Replicating this standard functionality of database systems is one challenge faced by large scale digital systems adopting an approach based upon a less rigid structure such as a wiki. At the opposite extreme from a relational database, content management systems and software like MediaWiki often initially impose very little structure on their contents at all beyond a simple page-based ontology. Querying this data in real time to organize it in useful ways therefore becomes more technically challenging. These issues may be compounded if non-experts are invited to take part in the editing process directly: depending on the platform, this may introduce the possibility of malformed data (e.g. ungrammatical XML, broken Wiki-code, etc.) being present in the dataset to be queried – something which would often not be possible at all with data stored in a relational database. Nevertheless, it may also be desirable to allow the possibility for editors to make radical changes to the contents and their organization, particularly when the goal is to allow editors to improve and reorganize preliminary data in need of editing. There is therefore a tension between how much structure is imposed on the data in order to simplify automated processing of it, and how much freedom is given to editors to simplify their tasks.

In the case of the Chinese Text Project, an alternative "hybrid" approach has been adopted to reduce the technical complexity of migrating from a static database with a complex ontology to a serializable wiki format. In this hybrid approach, existing data structures and user interfaces are retained for some or all of the existing static content, while new structures and interfaces for wiki content are developed in parallel. From a user perspective, both parts of the

system work in the same way – the only exceptions being that static content cannot be edited, and wiki content may not have certain specialized functionality available. This “hybrid” approach is made manageable by the presence of the API, which provides a single, unified interface to content which is in fact represented differently by the underlying system – implementation details which the API abstracts away. This approach has made it possible to vastly expand the range of content in this digital library, without the significant complexity of re-implementing all existing functionality from the static database system, and without removing functionality currently only implemented in the static format.

8. Conclusion

The transition from static database to dynamic platform has made possible an enormous expansion in the contents of the Chinese Text Project as well as greatly added to its utility. It has already hugely increased the level of user interaction, and facilitated tens of thousands of crowd-sourced contributions. While there are technical challenges to a dynamic approach, the increase in scalability makes possible new use cases, offers access to the “long tail” of data from texts which might otherwise be overlooked for full-text transcription, and provides a practical and open platform through which large amounts of material digitized in the future can be made immediately accessible while also being corrected over time by a group of geographically distributed editors.

適用於中文史料文本之標記式主題模型分析方法研究

陳奕安*、江子揚**、蔡銘峰***、薛化元****、劉吉軒*****

摘 要

本論文提出了一個適用於中文史料文本主題分析方法，主要是根據標記式隱含狄利克雷分布（Labeled Latent Dirichlet Allocation）演算法，使其可以透過人工標記的中文文本找出特定主題的相關詞彙。在我們的演算法中，我們額外地加上主題種子字詞（Seed Words）資訊，以增強 LDA 群聚過後的結果，使群聚過後的詞彙與主題的關聯度能夠獲得提昇。近年來，隨著網際網路的普及以及資訊檢索的蓬勃發展，同時由於數位典藏的資料成長，越來越多的實體書藉被編輯成數位版本並且加上後設資料（Metadata），在取得這些富有價值的歷史文本資料後，如何利用文本探勘技術（Text Mining）在這些資料上變成一項重要的研究議題。其中，如何從大量文本史料中辨識出文章主題更是許多學者感興趣的方向，而 LDA 主題模型則是在文字探勘領域中非常經典的方法。在此研究中我們發現傳統 LDA 對於群聚後的主題描述存在些許問題，包括主題類別的高隨機性以及個別主題的低易讀性，使得後續的解讀工作變得十分困難，因此我們採用了由 LDA 衍生出的標記式主題模型 Labeled LDA 演算法，限定能夠產生的主題類別以降低期隨機性，此外我們還加入了考量中文字詞的長度以及自定義的相關種子字詞等改進，使群聚出的主題詞彙能夠與主題更加相關，更加容易描述。

關鍵字：主題模型、隱含狄克雷分布、文字探勘

* 國立政治大學資訊科學系研究助理，Email: 102753031@nccu.edu.tw。
** 國立政治大學雷震研究中心助理研究員 ，Email: wallace0510@hotmail.com。
***國立政治大學資訊科學系助理教授，Email: mftsai@nccu.edu.tw。
**** 國立政治大學臺灣史研究所教授 ，Email: hy5595@nccu.edu.tw。
***** 國立政治大學資訊科學系特聘教授，Email: jsliu@cs.nccu.edu.tw。

An Enhanced Topic Model Based on Labeled LDA for Chinese Historical Corpora

Yi-an Chen[*], Tzu-yang Chiang[**], Ming-feng Tsai[***]
Hua-yuan Hsueh[****], Jyi-shane Liu[*****]

Abstract

This paper proposes an enhanced topic model based on Labeled Latent Dirichlet Allocation (L-LDA) for Chinese historical corpora to discover words related to specific topics. To enhance the traditional LDA performance and to increase the readability of its clustered words, we attempt to use the information of seed words and the Chinese word length into the traditional LDA algorithm. In this study, we find that the traditional LDA exists some problems about topic descriptions after clustering. We therefore apply the Labeled LDA algorithm derived from traditional LDA with the proposed improvements of considering the lengths of the words and related seed words.

Keywords: topic model, latent dirichlet allocation, text mining

[*] Research Assistant, Department of Computer Science, National Chengchi University. Email: 102753031@nccu.edu.tw.
[**] Assistant Researcher, Lei Chen Research Center, National Chengchi University. Email: wallace0510@hotmail.com.
[***] Assistant Professor, Department of Computer Science, National Chengchi University. Email: mftsai@nccu.edu.tw.
[****] Professor, Graduate Institute of Taiwan History, National Chengchi University. Email: hy5595@nccu.edu.tw.
[*****] Distinguished Professor, Department of Computer Science, National Chengchi University. Email: jsliu@cs.nccu.edu.tw.

一、 緒論

(一) 前言

隨著大數據時代的來臨，有越來越多的實體文本被轉為數位版本的形式儲存。在擁有大量的文本資料之後，人們便能透過資料科學（Data Science）等技術進行文本分析，並希望能夠從中得出真正有用的資訊，比如文章脈絡、趨勢，以及一些透過人工閱讀難以發現的特性，而主題模型（Topic Modeling）的出現對於這項工作則有顯著的幫助。在機器學習與自然語言處理等相關領域裡，主題模型是一種統計模型，被用於挖掘系列文本中所隱含的抽象主題，此概念最早始於 1998 年，但其真正被廣泛應用則是於 Latent Dirichlet Allocation (LDA) 演算法 [2] 被提出之後。

(二) 傳統主題模型與其限制

LDA 屬於典型的詞袋模型（Bag-of-words model），詞袋模型是資訊檢索領域中一種最基本的文件資料表示法。其將文件中出現的詞彙，想像是放在袋子裡零散而獨立的物件，如此一個袋子代表一篇文件[1]。LDA 將文章視為一組詞彙所集結而成的一個集合，詞彙與詞彙之間並無順序以及先後關連。 LDA 演算法可以將文章以多個主題的機率分布來表示，而每個主題又以詞彙的機率分布來表示，再藉由所群聚出的主題詞彙，人們便可以解釋這些主題，例如：演算法產生出主題詞彙「太陽能、核電廠、火力發電、污染」，可以解讀該主題可能為「能源與環境污染」概念之主題。

在 LDA 演算法出現之後，此演算法改變了傳統文章的表示形式，有許多的後續應用在之後被提出，然而，傳統 LDA 演算法仍有以下限制與缺點：

1. 無法適用於附含標記之文本

由於 LDA 演算法屬於無監督式學習（Unsupervised Learning)），無監督式學習是一種機器學習的方式，此學習方式在訓練模型時不需人力來輸入標籤，為監督式學習（Supervised Learning）和強化學習（Reinforcement Learning）等策略之外的一種選擇 [3]。因此，使用 LDA 時只需要將全文輸入便能得到其主題分布的結果，對於擁有文本但對其文本一無所知的使用者相當方便，但對於熟悉該文本的使用者，或者擁有附含主題標記之文本的使用者而言，他們已知部分文章隸屬於某些主題，對於演算法來說這是相當有用的資訊，但在傳統 LDA 演算法裡，這些資訊將無法納入考量；

2. 產生主題隨機性不一

LDA 演算法所產生的主題具有隨機性，每次聚出的主題種類與主題詞彙略有不

[1] http://terms.naer.edu.tw/detail/1679006/

同，也約略影響了每次群聚出的主題分布，將造成使用者對主題種類判別的困 難；

3. 主題詞彙難以解釋

由於 LDA 屬於詞袋模型（Bag-of-words model），其所產生的主題詞彙是零碎且不 連續的，對於不熟悉文本或非該文本領域專業人士而言，欲解讀這些詞彙之集合所代表的意義將十分的困難。

對於這些問題，近年來也陸續有學者提出改善的方法，Ramage 等人提出了 Labeled-LDA (2009) [6] 讓傳統 LDA 演算法可以加上已標記的資訊，Wang 與 Robert 等人先後提出了 Topical N-grams (TNG, 2007) [8] 與 Phrase-Discovering LDA (PDLDA, 2012) [5] 欲改善主題詞彙難以解釋之問題，詳細的介紹將於下一個章節做解釋。

(三) 研究目的

在本研究中，我們試圖將 LDA 演算法應用於一已附標記之中文文本，欲藉由其所群聚出的主題詞彙找出不同於已知主題，額外的表示詞彙。在套用於中文文本時，我們同樣面臨到上述問題，故本研究亦提出另一種混合方法來解決。我們由中文文本的斷詞問題發現一些特性，由於在中文領域裡一個「字」並不能完整的表達一個意思，一般認為「詞」才是最簡小有意義的一個單位，而其中長字詞更能夠明確的表達特定詞意，故本研究將詞的長度也納入生成主題詞彙之考量因素。此外，我們希望能夠透過已知知識更加地強化主題詞彙之群聚結果，將已知能夠表示某主題之詞彙視為種子字詞並將其納入演算法之考量，同時我們也思考詞性影響字詞易讀性之可能。總結來說，本研究之研究目的亦在於突破與改良上述傳統 LDA 演算法的的限制與缺點，並且將其套用於中文文本史料中，基於傳統的 Labeled-LDA 演算法，開發出專用於中文文本，同時考量長字詞、種子字詞以及詞性等混合方法的主題模型，並預期能夠找出較佳的詞彙表示該主題，使各主題分類更明確易讀。

二、 相關研究

多數 LDA 應用中文文本的相關研究屬於分類問題以及自動產生摘要等應用， 將 LDA 模型自動產生出來的詞彙作為分類問題中的特徵值，相對於傳統詞頻統計 以及 TF-IDF 的表示法，以主題以及主題所產生的詞彙表示一篇文章。這些研究皆 將 LDA 模型作為工具來使用，鮮少研究是直接針對中文文本對 LDA 模型之演算法進行改良。在前章節提到傳統 LDA 演算法有著以下一些問題：無法適用於附含標記之文本、產生主題隨機性不一，以及主題詞彙難以解釋等問題。非針對中文文本，以下幾類研究針對上述問題進行改良：

(一) 適用於已附標記之文本

傳統 LDA 演算法 [2] 屬於無監督式學習，所應用之文本並無需任何人工標記以及其訓練集，但倘若文本本身已附含標記之訊息，原始演算法沒辦法將這類資訊加入計算，為了解決 LDA 演算法不適用於已附標記文本的問題， Ramage 及 Hall 學者等人提出了 Labeled LDA 演算法 [6] ，他們透過每篇文章的標記資料，限制每篇文章的主題數量，每篇文章所表示的主題各有不同而不再是同樣的 k 個主題，此舉同時也限制了文章內的字詞所能分配的主題種類以及數量，降低了詞彙分布的隨機性。 LLDA 演算法成功將已附標記文本應用於原始 LDA 演算法之中，而且由於主題限制的原因，也一併降低了 LDA 演算法產生主題之隨機性，主題中的字詞也更加相關。

(二) 英文中的長字詞

除了產生主題之過度隨機性，主題描述困難亦是傳統 LDA 演算法的問題之一，由於群聚出的詞彙過於離散，若非對於文本有相關了解或是該主題類別的專業人士，僅憑藉單字與單字之間的關聯而欲描述該主題類別將是非常困難的工作，此類情況在中文文本上更為明顯。針對主題描述困難的議題，Wang 及 McCallum 等人（2007）提出了可能的解決辦法。他們在抽取主題詞彙的同時，用隨機的方式連接前字詞，使得主題字詞不再是一個一個獨立的 unigram，而是 bigram 或者 bigram 以上的 n-gram 字詞，讓一個主題可以擁有更多的線索去的描述 [8] ；而 Robert 與 William 等人（2012）更是將 Hierarchical Pitman-Yor Processes (HPYP) [7] 作為選定相連詞彙的辦法，藉以找出更符合詞意的片語字詞（Phrase）[5] 。

這些工作分別解決了 LDA 演算法無法將標記納入考量以及改善主題描述不易的問題，但對於中文的文本，中英文之間字（Charactor）與詞（Word）的意義並不相同，相較於英文的詞，中文的詞的概念更接近英文中相連字詞，為此，本研究也嘗試將相連字詞的概念轉換成中文的長字詞問題，希望藉由增加長字詞出現的比例，可以使得 LDA 演算法對於中文文本史料的主題描述更加容易。

三、 研究方法

本研究主要是基於標記式主題模型所做的修改，並且針對中文文本進行演算法的改良，在本章節的前半部，我們將會從傳統的主題模型開始介紹起；並於後半部說明與本研究演算法所改進的部份以及方法。

(一) 傳統主題模型簡介

在機器學習（Machine Learning）與自然語言處理（Natural Language Processing）等相關領域裡，主題模型（Topic Model）是被用來發現在一系列文檔中所隱含的主題分布。假設一篇文章在撰寫的時候代表某一種主題，我們可以推測在這篇文章之中的某些特定的詞彙出現之頻率將會大於或小於其他主題的文章，即不同主題所代表的詞彙將會有所不同，舉例來說：在討論狗的文章中，「骨頭」與「狗」等詞會可能的出現頻率較高；在討論貓的文章中，「貓砂」與「貓」等詞彙可能的出現頻率較高，而「於是」或者「關於」這類一般性詞彙出現的頻率則會大致相同。

如同於緒論中所提到，主題模型之概念最早於 1998 年由 Papadimitriou 及 Raghavan 等學者所提出，另一位學者 Thomas Hofmann 則是在 1999 年導入了機率的概念並且發表了 Probabilistic latent semantic indexing (PLSI) [4] ；然而最受歡迎也最具代表性的主題模型則是 LDA ，其是由 Blei 等人於 2003 年所提出並且實作，將文章所隱含主題由單一主題轉變為多重主題表示。

LDA 演算法將每個文章以多個主題的混合型態表示，並且可以產生相對應的詞彙集以及詞彙的機率來表示一個主題。透過每個主題所產生的主題詞彙，使用者便能夠解釋該主題（如：憲法、法律、政府皆與法律有所關聯，可以推測該主題為「法治相關」主題），同時也可以觀察到 LDA 屬於典型的詞袋模型（Bag-of-words model），其將文章視為多組詞彙所集結而成的一個集合，每組詞彙代表著一種主題，詞彙與詞彙之間並無順序以及先後關連。

(二) 隱含狄利克雷分布 LDA

接下來我們將從主題分布的原理到主題字詞的推論，更仔細的介紹 LDA 演算法。首先， LDA 遵從以下兩個步驟，將字詞從文件集中的每篇文章取出：

1. 隨機選擇一個主題的分布

2. 對文章中的每個字詞：

(a) 隨機從第一步驟產生的主題分布中挑選一個主題

(b) 隨機地從已知對應詞彙分布中挑選一個詞彙

這個機率模型很直覺的反映出了一篇文章表示著多重主題，每篇文章都有不同的主題分布（步驟一），且每篇文章的每個字詞皆是由其中一個主題中所挑選出來，而該主題的主題字詞又是由先前的文章分布中所聚集出來的（步驟二）。

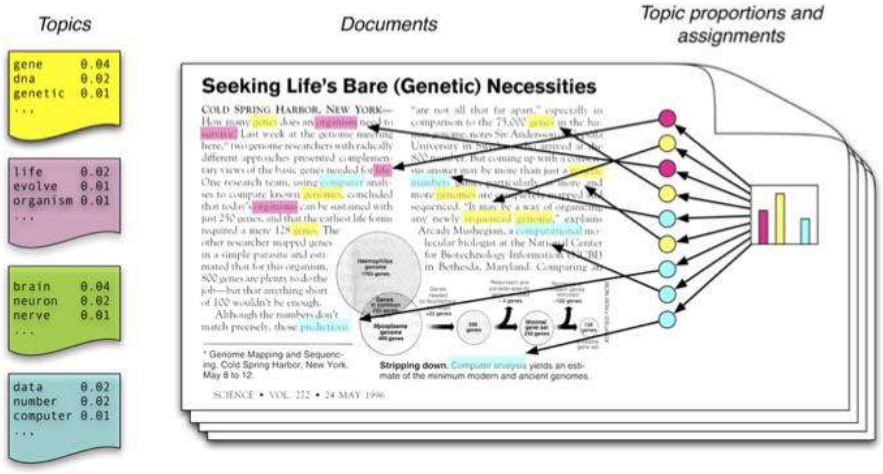


圖 1 主題於文章內文分布示意圖 [1]

利用以下符號，我們可以對整體過程做更正式的描述：

將所有的主題以 $\beta_{1:K}$ 表示，而 β_k 是在每個主題下各個字詞的機率分布（如圖一中的左半窗格）。 θ_d 表示第 d 個文章中各主題所佔比例（如圖一右側簡易直方圖所示）， $\theta_{d,k}$ 則是文章中主題 k 的所佔比例。 z_d 表示第 d 篇文章的主題分配，其中 $z_{d,n}$ 是文章 d 中的第 n 個字所分配到的主題（如圖一中央上色部分）。最後， w_d 表示從第 d 篇文章所觀察到的字（此為固定不變的元素），文章中的第 n 個字為 $w_{d,n}$ 。我們將上面所提到的符號用來描述 LDA 主題分布的生成過程，可以產生出以下式子：

這個式子說明了 LDA 的主題分布有著一連串的相依性，舉例來說，字詞所分配的

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d)\, p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right). \tag{3.1}$$

主題 $z_{d,n}$ 機率相依於該文章的主題分布 θ_d ；而所觀察到的主題字詞 $w_{d,n}$ 又與該詞所分配到的主題 $z_{d,n}$ 以及所有主題 $B_{1:K}$ 有所關連。這些相依性是在生成過程中的統計假設，在統計學裡稱作聯合機率分配，他們定義了 LDA 本體，並且能夠以圖 2 的圖形表示法來表現。

接著我們要進行的部分是 LDA 的計算問題，即在我們擁有（欲觀察的）文章後要

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \mid w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}. \tag{3.2}$$

如何計算出主題與文章之間的條件機率，也就是所謂的後驗機率計算。利用先前所標示的符號， LDA 的後驗機率為：

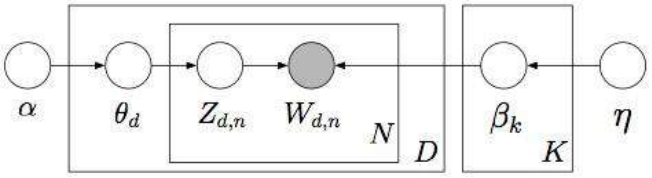


圖 2　LDA 模型圖形表示法（plate notation）

(三) 標記式 LDA

我們知道 LDA 可以適用大多數的文件集，但卻僅限於沒有標記資料的文件集，然而現今仍有非常多的資料集是擁有標記的，如網路新聞、部落格文章底下會有許多關鍵字分類，許多社群網路平台（如 Twitter 、 Facebook ）也有提供 hashtag 的標記功能， 我們假設這些分類的關鍵字及 hashtag 標記能夠代表該文章的主題，則這些資訊應該納 入 LDA 主題分布的考量之中，故某些學者提出監督式的 LDA ，使傳統 LDA 演算法能 夠在推論主題模型時將該資訊納入其中。

依照 Ramage 等人所提出的標記式隱含迪利克雷分布（Labeled LDA），我們將可以 簡單地用預先觀察到的文章主題（即人工標記的主題分類）去限制主題模型。下列是 Labeled LDA 演算法的生成過程，其中所使用的符號與前述傳統 LDA 演算法相同，而另外增加的符號是 Λ ， $\Lambda^{(d)}$ 表示該篇文章所標記的主題種類，其值是非 0 即 1 的整數陣列，從此文章與主題之間的分布 θ 關係不再只與 α 有關而是同時受限於 Λ 值。在演算法中，步驟一、二依序從每個主題 k 下依照迪利克雷分布的前參數（prior）η 抽出主題中字詞的分布 β ，接著傳統 LDA 演算法將從每篇文章 d 中依照另一迪利克雷分布的前參數 α 抽出各主題的分布 θ_d ，但在 Labeled LDA 中， θ_d 將被限制在只與該文章有關的標記 $\Lambda^{(d)}$ 中。如此一來，可以確保步驟九當中（同本文章節三的第二小節所提及） 所有文章字詞所分配到的主題 $z_{d,i}$ 皆被該文章既有之主題標記所限制。

為了實現此一目標，在Labeled LDA 中我們首先以主題標記的先驗機率Φ_k 產生出每篇文章的標記 $\Lambda^{(d)}$ ，接著定義 $\lambda^{(d)} = \{k \mid \Lambda^{(d)} \stackrel{k}{=} 1\}$ 為文章標記之向量。如此我們便能為每篇文章d 定義出一個大小為M_d 乘以K 的映射矩陣 $L^{(d)}$ ，M_d 即為該文章的標記主題

Algorithm 1 增強 LDA 之生成過程

1: **for all** topic $k \in \{1, ..., K\}$: **do**
2: Generate $\beta_k = (\beta_{k,1}, ..., \beta_{k,V})^T \sim \text{Dir}(\cdot \mid \eta)$
3: **end for**
4: **for all** document d: **do**
5: **for all** topic $k \in \{1, ..., K\}$: **do**
6: Generate $\Lambda_k^{(d)} \in \{0, 1\} \sim \text{Bernoulli}(\cdot \mid \Phi_k)$
7: **end for**
8: Generate $\alpha^{(d)} = L^{(d)}$
9: Generate $\theta^{(d)} = (\theta_{l_1}, ..., \theta_{l_{M_d}})^T \sim \text{Dir}(\cdot \mid \alpha^{(d)})$
10: **for all** i in $\{1, ..., N_d\}$: **do**
11: Generate $z_i \in \{\lambda_1^{(d)}, ..., \lambda_{M_d}^{(d)}\} \sim \text{Mult}(\cdot \mid \theta^{(d)})$
12: Generate $w_i \in \{1, ...V\} \sim \text{Mult}(\cdot \mid \beta_{z_i})$
13: **end for**
14: **end for**

$$L_{ij}^{(d)} = \begin{cases} 1 & \text{if } \lambda_i^{(d)} = j \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

數量（$M_d = \dot{\lambda}^{(d)}\dot{}$），各列$\mathbf{i} \in \{1, ..., M_d\}$，行$\mathbf{j} \in \{1, ..., K\}$：

換句話說，當此矩陣$L^{(d)}$的第i行中的第j列的元素為1時，即表示第i篇文章恰擁有主題 j 的主題標記，反之，元素為0 時，表示該文章所擁有的主題標記不為主題 j 在得到此矩陣 $L^{(d)}$ 之後，我們將它投影至狄利克雷主題前參數α 的參數矩陣上（其中，$\alpha = (\alpha_1, ..., \alpha_K)^T$ ）：

$$\alpha^{(d)} = L^{(d)} \times \alpha = (\alpha_{\lambda_1}^{(d)}, ..., \alpha_{\lambda_{M_d}}^{(d)})^T \tag{3.4}$$

我們可以明顯的看出此映射矩陣的維度與該文章所標記的主題數量成正比。舉例來說，我們假設K = 4 ，而某篇文章d 的標記$\Lambda^{(d)} = \{1, 0, 1, 0\}$ ，這代表$\lambda^{(d)} = 1, 3$ ， 如此一來$L^{(d)}$ 將會變成：

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

接著，演算法將根據迪利克雷分布的前參數 $\alpha^{(d)}$ 來抽出主題分布 $\theta^{(d)}$ ，而此時的 $\alpha^{(d)} = L^{(d)} \times \alpha = (\alpha 1, \alpha 3)^T$ ，如此便將能將主題分布限制在文章所標記的主題一以及主題

三之中。

透過標記式 LDA 演算法，我們已經有效的將文章的主題分布限制在文章的標記之中，並且透過實驗可以發現群聚出的主題詞彙相較於傳統 LDA 演算法更加的集中，意義更加的明顯。接著我們將繼續以適用中文文本為宗旨進行其他的改良。

(四) 適用中文文本之改良

由於傳統 LDA 演算法最早是適用於英文文本，在將 LDA 套用至中文文本時會遇到一些問題，其中一個主要的問題源自於中英文之間對於「字」與「詞」所描述的意義 並不相同。在英文中，每個單字（Word）是由字元（Character）組成，而中文則是以 詞作為最小的的單位元；在英文中的一個句子中可以將每個單字獨立來看，每個字皆 能表達完整意思，但中文則否，例如：

「今天天氣晴朗」

若將每個單字切開來看：

「今 / 天 / 天 / 氣 / 晴 / 朗」

雖然每個字都有其意義，但卻不能表達完整意思。因此在套用 LDA 演算法至中文文本之前，斷詞的問題是必須優先解決的。

1. 斷詞問題

在斷詞的問題中，中文斷詞相對於英文斷詞困難許多，延續上一個例子，原句為：「今天天氣晴朗」要如何正確的斷句成「今天 / 天氣 / 晴朗」，而非「今 / 天天 / 氣 / 晴 朗」或「今 / 天 / 天 / 氣 / 晴 / 朗」便是重要的課題之一，假使在英文中：「Today is a sunny day」只要以空格作為分界便可以輕易斷句成「Today / is / a / sunny / day」，但中 文裡並沒有此類將詞語特意獨立之特性。故要達成中文斷詞必須要先有一個字典，並且該字典之詞庫越大越好，但是現實的問題是並沒有一個字典包含所有的詞彙，根據不同類型的文本，所欲切分之詞彙也有所相異，因此在進行中文斷詞工作的時候，還需要依據文本特性的自定義字典，一旦擁有此自定義字典，斷詞的效果方能獲得顯著提升。

隨著詞庫字典以及個人的定義字典越來越大，詞彙量大到一定的程度後，字典的查詢工作將會消耗大量的時間，為了節省此一部分所消耗的時間，在進行查詢工作之前，我們可以將整本字典建立成一個字典樹（Trie 樹），此一字典樹保存著所有單字詞，以及該單字所接續下一個字的機率值。建立完字典樹之後，在斷詞的時候即可將原句子逐一比對，判斷原句中是否含有符合字典樹裡所包含的字詞，從而得到原句裡

所有的切分組合，再將所有的組合以一個有向無環圖（Directed acyclic graph, DAG）表示，最後再計算出最佳的切分方式，得到初步的斷詞結果。

然而，對於未知的詞彙（即不存在於字典中的詞彙），在剛才的結果中會以單字詞斷開，因此還需要透過 HMM Viterbi 演算法，計算出單字詞與單字詞之間可以合成新詞彙的機率值，藉此我們才能獲得最終的結果。

2. 長字詞優先

在中文的詞彙裡有一種現象，字詞越長所能表現的意義愈顯明確，例如「大」字可以形容體積、容量、數量、強度上超過一般，或是超過比較的對象，「大」字也可以是名詞，指相對於「小」的字詞，或者年紀較大的人；相對於「大」，「大學」的意義較明確一些，但仍然可以表示成國家最高學府，或者是四書《論語》、《孟子》、《大 學》、《中庸》中的「大學」；而持續增加字詞的長度，四字詞的「政治大學」所能表現的意義更加的明確，指的是位於臺北市木柵區，台灣的國立大學之一。

由於這樣的一個特性，在中文斷詞的處理中，除了考量字詞出現的頻率，也會一併考量字詞的長度，因此，在增強 LDA 的演算法裡，我們也將字詞的長度當做考量之一，讓 LDA 在抽取主題詞彙的時候能夠以長字詞為優先。

同樣的，為了實現此一目標，我們首先給定每個字詞一個相對應的權重 w_i，並且定義權重向量$W = \{w_1, w_2, \ldots, w_V\}$：

$$w_i = base^{|V_i|} \tag{3.5}$$

我們將 W 標準化至總和為 1，接著我們定義一個乘法 @ ：

$$A@B = \begin{pmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} a_1 b1 & a_2 b_2 & a_3 b_3 \end{pmatrix}$$

接下來我們即可重新給定新的 η 值 ηˆ ，其中：

$$\hat{\eta} = W@\eta \tag{3.6}$$

接著演算法將依循狄利克雷分布從前參數 ηˆ 抽出主題中的詞彙分布 β ，而此時的 ηˆ 已 經包含了字詞的長度資訊作為權重。

3. 種子字詞

由於 LLDA 演算法已經將傳統 LDA 由無監督式演算法進一步修改為半監督式演

算 法，使用者在使用的同時對該文本或欲觀測的主題已有初步的了解，因此我們假設文 本的已知主題下已經有一些既定的主題詞彙，而這些詞彙是由使用者認定能夠真實表 現主題內容，我們將這些詞彙稱作種子字詞，我們希望藉由先前提及的權重概念，將 種子字詞納入考量，間接影響其他的字詞，提昇演算法所群聚出的詞彙質量。因此， 我們將權重向量 W 定義為：

$$w_i = \begin{cases} 1 & \text{if } V_i \text{ is seedword} \\ w_i & \text{otherwise.} \end{cases} \tag{3.7}$$

如此將提高種子字詞的權重，並同時考量了其他長字詞的權重。

如上所述，種子字詞是依照使用者（文本史料擁有者）所認定能夠真實表示主題意義的詞彙，因此，在種子字詞的提取方面亦是採取人工標記的方式，由使用者在事前標記好各個主題所可能隱含的種子字詞，並由以上的權重向量公式將其納入演算法之中。本研究中所使用的種子字詞則是有多個版本，最初的版本是由薛化元教授團隊所提供，由團隊史學家以其經驗統整出該主題所應包含的詞彙，同時，隨著實驗的進行依照主題群聚的結果產生，團隊史學家們再根據所產生的主題詞彙新增或刪減種子字詞，而後再將新標記的種子字詞加入演算之中，周而復始。

四、 實驗結果與討論

(一)實驗設定

1. 資料集以及資料前處理

在實驗的史料文本部分，我們所使用的資料集為《自由中國》雜誌，其為一半月刊的雜誌，由胡適以及雷震等人所創辦，並於中華民國發行，以擴展民主自由空間為宗旨的政治刊物，其收錄的範圍是 1949 年至 1960 年間共 23 卷又 5 期，內容包含言論自由、地方自治、司法獨立、反對問題等相關議題以及其主張，為眾多史學專家在研究台灣民主運動所用。

本研究所使用之《自由中國》文本是由政治大學雷震研究中心[2]薛化元教授團隊所提供，原始文章共 3651 篇，約一千六百多萬字，其中，附含主題標記之文章數有 1973 篇，統計後約七百多萬字（經斷詞處理後為四百多萬詞），而由於一字詞在中文裡不太能夠表達完整意義，故實驗中我們除了移除標點符號外，同時也將一字詞移除，處理過後的文本總詞數剩下兩百多萬個。

[2] http://leichen.nccu.edu.tw/leichen/

事前標記的部分包含文章主題以及種子字詞，在《自由中國》文本中共含有 23 個主題，並且各個主題皆有與主題相關的種子字詞（主題二十三：文藝類除外）。我們額外在所有的文章都加上主題零，定義為一般類的主題，目的在篩選出各篇文章皆有出現的通用詞彙，使其他主題所群聚出的詞彙獨特性能夠更加地顯著，同時也可以確保文本中僅含一種主題標記的文章，所群聚出的詞彙仍保有「分配」的空間。

表 1 原始文本與附標記文本之統計資料

	原始文本	附標記文本
文章篇數	3,651	1,973
文本總字數	16,171,034	7,569,758
去除標點符號及英文	13,700,366	4,399,346
處理後總詞數（二字詞以上）	4,917,782	2,696,892
處理後相異詞	258,658	178,289

2. 斷詞工具

在斷詞處理部分，我們使用開源斷詞工具－結巴中文分詞（jieba3）幫助我們進行 斷詞的工作，由於結巴最初是以簡體中文開發，在斷詞表現簡體中文較優於繁體中文， 但是近期結巴發佈了針對繁體中文的的字典，在使用該字典後繁體中文的表現已獲得 有效提昇。另一方面，我們也透過結巴提供的自定義詞典功能建立了使用者字典，此 字典亦由薛化元教授團隊所提供，該團隊擁有足夠的歷史相關知識，以及對《自由中 國》文本的了解，所提供的字典對於斷詞有相當的幫助。

3. 量化評估標準

我們利用資訊檢索中對檢索系統的評量方法，作為增強 LDA 所群聚出字詞的量化評量。我們將群聚出的結果以人工標記出正確答案，評斷的基準是由史學專家以經驗初步判斷是否與該主題相關，並且與實際文本（另外建立文本的檢索系統）交叉比對後所得到的結果。有了各主題的正確答案，並能透過以下方法來評量結果：

Precision（準確度 / 查準率）

Precision 指的是檢索結果的準確度，意即在檢索結果中含有與檢索內容相關文章

$$Precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \qquad (4.1)$$

[3] https://github.com/fxsjy/jieba

的比例，其公式如下：

從公式中可以看到準確度是取所有檢索到的結果裡，與欲查詢的問題（query）有相關的文章數，所有檢索中若出現較多的相關文章，則準確度值越高，其值越高越好， 最大值為 1。

Recall（召回率 / 查全率）

相較於準確度希望檢索結果之中相關文章的比例，召回率注重的是所有相關文章中被檢索到的比例，其公式如下：

$$Recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{revelant\ documents\}|} \tag{4.2}$$

與準確率相同，召回率的值介於 1~0 之間，其值越高越好。

Average precision（平均準確度）

在每個主題下我們所標記為正確答案的字詞數量不一，而由於準確度與總相關文章數成正比，若直接使用準確度來作為評量，對於正確答案較少的主題較難看出其成效，故我們使用平均準確度以及召回率作為最終的評量標準，其公式如下：

其中， P 即為準確度，而 rel(k) 表示該文章是否相關，若相關則是 1，反之為 0。

$$AveP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{number\ of\ relevant\ documents} \tag{4.3}$$

此外，平均準確度也可以設定僅計算檢索排序前 n 個結果：

$$AveP@n = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{min(m,n)} \tag{4.4}$$

其中，m 指的是所有相關文章的數量（即正確字詞數量），計算的時候分母取m 與n 的最小值，假設所有相關文章的數量為12 ，而n = 10 ，此時則僅計算10篇文章的平均準確度。

Mean average precision（平均準確度均值）

平均準確度均質為平均準確度的延伸，主要用來評量整個檢索系統，將每次 query 所算出的平均準確度再取平均：

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{4.5}$$

公式中， Q 即是 query ，而本實驗中所算的 MAP 則是將主題類別當成是 query ，依每種主題所計算出之平均準確度作為最後的評測。

表 2 主題標記之名稱及主題詞彙整理

	主題名稱	標記文章數	原始種子數	群聚後種子數	新增種子數
主題一	刊物的立場與反省	21	2	8	7
主題二	自由民主的基本概念	160	6	57	53
主題三	法治	59	5	31	30
主題四	表現自由／出版法問題	163	3	50	48
主題五	其他基本權問題	83	3	61	58
主題六	責任閣制／責任政治	33	6	36	32
主題七	行政中立—— 國民黨退出軍警特	34	7	4	4
主題八	司法	61	10	66	61
主題九	立法院問題	44	4	35	32
主題十	監察	23	8	39	31
主題十一	考試	13	8	37	31
主題十二	在野黨	102	4	40	36
主題十三	國民黨體質改造／國民黨問題	44	1	5	4
主題十四	地方自治	42	5	42	38
主題十五	地方選舉問題	80	1	49	48
主題十六	軍隊	65	6	60	57
主題十七	教育／救國團	201	7	37	34
主題十八	外交／聯合國問題	156	8	57	55
主題十九	總統三連任問題／修憲／國大	69	9	35	28
主題二十	反共救國會議	35	5	9	7
主題二十一	反共	188	2	40	39
主題二十二	經濟／財政	241	8	50	47

(二) 實驗結果分析與討論

此一小節我們將呈現增強 LDA 對中文字詞的改進成果，並且依續對質化以及量化的結果進行分析。我們以原始 LLDA 的結果作為基準，而後與長字詞優先以及考量種子字詞的方法比較。

1. 長字詞優先

首先我們看到中文裡長字詞優先部份的改良，根據公式，我們將長字詞權重調整為 base = 10 、 β = 0.123 ，其結果如下表 3。在表 3 當中我們列舉了四種主題，首 先可以看到的是各主題中 LLDA 的表現結果，在有限制條件的演算法之下，各個主題所群聚出的字詞已經有一定程度的相關性，像是主題二裡面「自由」、「民主」、「平等」等詞，或是主題十六的「軍人」、「軍官」、「軍隊」等詞，這些字詞所表現的主題意義已十分相近。在增強 LDA 的部份我們可以看到僅考量長字詞的情況之下，群聚出的字詞已經改變許多。在新的群聚結果裡有非常多的新字詞出現，而其中又以長字詞居多，像是主題二裡面出現了「自由經濟」、「資本主義」等詞彙，主題三則有「司法行政部」、「懲治貪汙條例」等超過四字詞的詞彙出現，這些詞彙如同我們所預期的，符合中文裡長字詞更能明確表現詞意的特性。原演算法所獲得的結果會出現「政府」、「國家」、「社會」等較普遍的詞彙，對於一般使用者來說，這些普遍性詞彙較通俗易理解，而改良過後的增強 LDA 演算法則能夠得到像是「司法行政部」、「圓山飯店」以及「懲治貪汙條例」等意義明確的詞彙（甚至是專有名詞、人名等等），對於已熟知文本的史學家來說，該類字詞將是較佳的觀察目標。

2. 考量種子字詞

接著我們繼續看到考量種子字詞的結果，此部份我們根據公式 3.7 將各主題標記之 種子字詞權重調整為 1，剩餘的詞彙權重則是延續公式 3.5 以及公式 3.6 所生成之 wi ， 將長字詞一併納入考量。在實驗中我們所擁有的原始文本標記共有 23 個主題，其中第 23 個主題（文藝類）未含有種子字詞之標記，故評量的時候不予採計。下表 4 以及表 5 中我們列舉剩餘 22 個主題排序前五十的平均準確度（AP@50）並且計算其均 值（MAP），並且以 LLDA 演算法作為實驗的基線（baseline）。

在實驗中，我們演算法所使用的種子字詞有好幾組，最初所設定的種子字詞是團隊史學家所認定在該主題分類中應該出現的詞彙。而後，隨著實驗的進行，我們將增強 LDA 演算法的群聚結果交由團隊史學家進行分析，進而從中抽取出更新後的種子字詞，我們再將新的種子字詞納入增強 LDA 演算法之中，週而復始，最後得到最終的結果。在表 4 中，我們可以看到在各個主題底下增強 LDA 的平均精準度皆優於傳統 LLDA，在平均準確度均值的表現，增強 LDA 為 0.4752 ，更是倍數於 傳統 LLDA 的 0.2040 ，同時在僅考量長字詞以及套用最初版本種子字詞的增強 LDA 也有平均準確度均值 0.3633 以及 0.4106 的表現。

表 3 主題詞彙結果比較

主題二		主題三	
LLDA	僅考量長度	LLDA	僅考量長度
自由、民主、國家、政治、人民、思想、社會、個人、反共、政府、他們、組織、中國、平等、個人自由	自由、個人自由、自由主義、平等、個人、個體、群體、權威、君主、民權、中山先生、容忍、自由經濟、資本主義、愛民	法律、法治、人民、憲法、政府、總統、國家、行政、民主、機關、美國、規定、守法、議會、命令	蔡金塗、司法行政部、監所、羅氏、治外法權、陪審員、圓山飯店、附徵、白鵬、升等考試、主席台、懲治貪污條例、行仁、搗毀、高等法院

主題十二		主題十六	
LLDA	僅考量長度	LLDA	僅考量長度
反對黨、政黨、民主、人民、政治、民主政治、國家、國民黨、新黨、政府、執政黨、在野黨、組織、政權、問題	反對黨、政黨、新黨、在野黨、執政黨、民主政治、在朝黨、組黨、強大、在野、黨員、政黨政治、執政、領袖、知識分子	軍人、軍事、軍官、官兵、政府、軍隊、待遇、生活、我們、士兵、訓練、軍中、人員、海軍、國軍	軍官、官兵、軍隊、軍人、軍中、士兵、海軍、退除役、新軍、國防會議、勞軍、部隊、待遇、國軍、空軍

此表擷取主題二、三、十二、十六，分別代表「自由民主的基本概念」、「法治」、「在野黨」以及「軍隊」，每個主題顯示排序前 15 的主題詞彙， 其中，粗體字表示更改演算法後新出現三字詞（含）以上的詞彙。

3. 因應主題隨機性之處理

由演算法得知，標記式主題模型限制了主題詞彙的分配，並在某種程度上降低主題群聚詞彙的隨機性，但在上面的實驗中，這樣的隨機性對實驗結果仍會產生些許影響，為了降低主題隨機性對實驗結果所造成的影響，在 LLDA 以及增強 LDA 我們 都採用以下辦法：首先，將所有的實驗執行數次（表 5 中的數據為每組執行二十次之結果），接著擷取出每組實驗排序前 100 個主題詞彙，並對每組詞彙所對應的機率值加總平均，最後再依據重新計算過後的機率值進行排序。同樣的，在表 4.5 中列舉 22 個主題排序前五十的平均準確度（AP@50）並計算其均值（MAP）。

在表 5 中可以發現，主題詞彙經過重新排序後，各組實驗的 MAP 均有所提昇，其中，傳統 LLDA 演算法由 0.2040 提昇至 0.3413 ，而增強 LDA 則是由原先的 0.4752 提昇至 0.5272 ，此結果顯示主題隨機性對實驗中各個演算法之表現有著實質上的影

響。另外，由表 4 以及表 5 中可以看到主題一以及主題七在對照組中表現普遍低落，即使加上主題隨機性的處理仍不見起色，此兩組主題分別表示「刊物的立場與反省」及「行政中立 ── 國民黨退出軍警特」，在標記的文章篇數稍嫌不足（參考表 2 ），經討論後亦認為主題一屬於意義較不明確之標記，可能為整體表現低落的原因之一。

除了平均準確度，我們還可以看到增強 LDA 在種子召回率的表現，透過圖 3a 我們可以看到在不同演算法底下各個主題的平均召回率（@K），在 K=100 的情形底下，僅考量長度的增強 LDA 演算法的平均召回率有 0.6771 ，而加入種子字詞後則是達到 0.9110 ，平均召回率提昇了 34.5% ，對比傳統 LLDA 演算法的 0.5153 更是提昇了 76.8% 。另外，從平均精準度均值的圖表（圖 3b）中可以看到，在 K=50 之後其值逐漸上升，其原因在於各個主題中所標記為正確答案的主題詞彙平均僅有 38.5 個，根據公式 4.4 中平均準確度的算法，分母為 K 值與總相關字詞數量之最小值，故分母不變的情況之下分子逐漸增加，均值自然逐漸提昇。

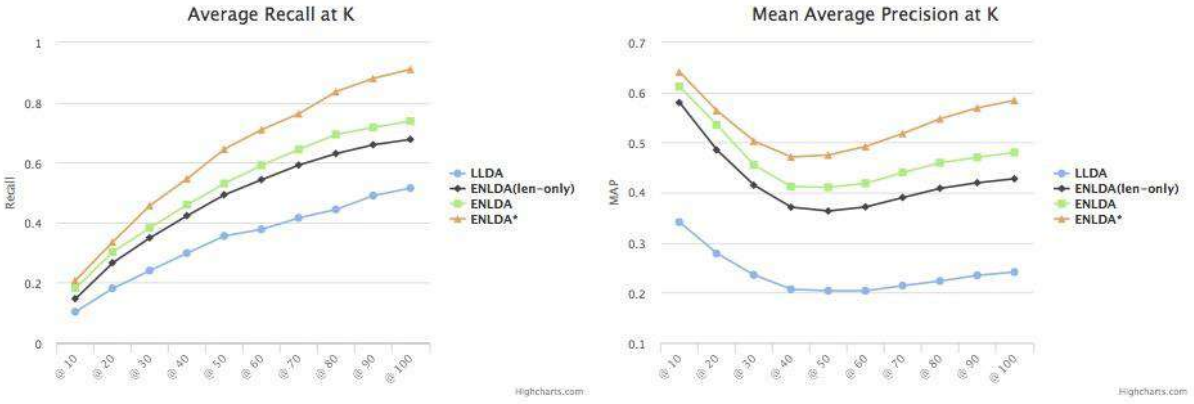
4. 與傳統關鍵字提取方法比較

在此組實驗中，將主題模型所產生之主題詞彙與傳統關鍵詞提取技術（TF-IDF）所產生之關鍵詞進行比較，實驗目的有二：一、觀察兩種方法所提取出的字詞有無差異；二、觀察加入種子字詞是否能增近新的隱含詞彙之生成。在 TF-IDF 的方法中，將 1494 篇文章（去除文藝類）重新分類成 22 篇長文（22 類主題），將相同主題標記的文章視為同一篇文章，若該文章有兩種以上的主題標記，則以相同內容重複在不同篇長文內的方式處理，最後擷取各新文章中 TF-IDF 值排序前 100 之詞彙作比較評估。

表 4 主題模型對個主題之平均準確度（Average Precision at 50）

	LLDA	增強 LDA (僅考量長字詞)	增強 LDA (使用初始種子)	增強 LDA
主題一	0.0417	0.0283	0.0000	0.2282
主題二	0.1080	0.4844	0.5258	0.6225
主題三	0.0181	0.1504	0.2871	0.5552
主題四	0.1811	0.4596	0.4629	0.4583
主題五	0.0699	0.5485	0.5024	0.5784
主題六	0.1623	0.0239	0.3339	0.5840
主題七	0.0078	0.0000	0.0000	0.2619

主題八	0.3732	0.5981	0.5963	0.5935
主題九	0.2819	0.3751	0.3867	0.3801
主題十	0.1819	0.4818	0.5094	0.5035
主題十一	0.1477	0.3014	0.3005	0.3634
主題十二	0.1574	0.4109	0.4932	0.5567
主題十三	0.0044	0.0042	0.2400	0.4333
主題十四	0.2295	0.5077	0.5681	0.5403
主題十五	0.4720	0.6239	0.6216	0.6212
主題十六	0.4116	0.5938	0.6028	0.6668
主題十七	0.4180	0.5351	0.5348	0.5339
主題十八	0.3075	0.5291	0.5361	0.5333
主題十九	0.2301	0.3652	0.3785	0.3684
主題二十	0.0419	0.1808	0.3735	0.2920
主題二十一	0.3322	0.4188	0.4046	0.4056
主題二十二	0.3103	0.3716	0.3745	0.3744
MAP	0.2040	0.3633	0.4106	**0.4752**



平均召回率@K　　(b) 平均準確度均值 @ K

圖 3　評量結果比較圖

表 5 主題模型對個主題之平均準確度之二（Average Precision at 50）

	LLDA	增強 LDA (僅考量長字詞)	增強 LDA (使用初始種子)	增強 LDA
主題一	0.0000	0.0229	0.0216	0.3894
主題二	0.5223	0.4242	0.4330	0.6706
主題三	0.0904	0.2925	0.3310	0.5353
主題四	0.4767	0.4383	0.4414	0.4916
主題五	0.2624	0.5760	0.5277	0.6948
主題六	0.3402	0.2560	0.3354	0.6124
主題七	0.0000	0.2955	0.2917	0.5250
主題八	0.5028	0.5970	0.5971	0.5975
主題九	0.2996	0.3680	0.3677	0.3817
主題十	0.3141	0.4715	0.4866	0.5433
主題十一	0.2386	0.2581	0.2777	0.4049
主題十二	0.4767	0.4557	0.4326	0.5590
主題十三	0.1240	0.2174	0.2233	0.4810
主題十四	0.4061	0.4981	0.4799	0.6023
主題十五	0.5536	0.6074	0.6221	0.6216
主題十六	0.6039	0.6366	0.6553	0.7578
主題十七	0.4768	0.5483	0.5337	0.5508
主題十八	0.4886	0.5325	0.5180	0.5325
主題十九	0.2904	0.3500	0.3383	0.3690
主題二十	0.2916	0.3189	0.3760	0.4877
主題二十一	0.4195	0.3974	0.4072	0.4069
主題二十二	0.3309	0.3724	0.3724	0.3835
MAP	0.3413	0.4061	0.4123	**0.5272**

在評估方式方面我們仍然使用平均準確度以及平均準確度均值作為評估的標準，但在正確答案的部份有著些許的差異，為了能夠觀察出加入種子字詞是否能夠群聚其他與主題相關的隱含詞彙，在此部份的增強 LDA 之中，我們所採用的種子字詞為原始版本的種子字詞（參照表 2 第四欄），而最終拿來計算成績的正確字詞則是將最終標記的種子字詞減去原始版本種子字詞的「新增字詞」（參照表 2 第六欄），實驗結果如表 6 所示。

由結果可以看到，傳統 TF-IDF 所提取之結果與主題模型所群聚之結果的關聯性不高，以 MAP 來看準確度僅有 0.1553 ，表示大部分由主題模型找出的隱含詞彙，透過 TF-IDF 並無法找到。同時，我們也看到在個別主題當中，絕大部分的主題加入種子字詞後，準確度皆有一定程度的提升，而某些主題底下更有高於兩倍的提升，由此可見種子字詞對於群聚隱含詞彙的影響力。另一方面我們也看到主題十以及主題二十二（「監察」以及「經濟財政」）底下，TF-IDF 的表現約略優於主題模型。

表 6 傳統關鍵字提取方法與主題模型之平均

準確度比較（AP@50）

	TF-IDF	LLDA	增強 LDA
主題一	0.0130	0.0000	0.0247
主題二	0.1948	0.3668	0.3477
主題三	0.0434	0.0443	0.2501
主題四	0.0344	0.3687	0.3288
主題五	0.2079	0.2520	0.4787
主題六	0.0939	0.2493	0.2086
主題七	0.2738	0.0000	0.2917
主題八	0.3014	0.3381	0.4125
主題九	0.2318	0.1904	0.2467
主題十	0.2677	0.1332	0.2549
主題十一	0.1321	0.1789	0.1752
主題十二	0.0530	0.3464	0.2948
主題十三	0.0000	0.1250	0.2609
主題十四	0.1158	0.3766	0.4201
主題十五	0.2110	0.5651	0.6350
主題十六	0.2347	0.4875	0.5556
主題十七	0.1291	0.4003	0.4381
主題十八	0.2409	0.4088	0.4338
主題十九	0.0998	0.2301	0.2743
主題二十	0.1429	0.1152	0.1578
主題二十一	0.0923	0.3770	0.3701
主題二十二	0.3031	0.2514	0.2870
MAP	0.1553	0.2639	0.3249

表 7 各組實驗與演算法之平均準度均值（Mean Average Precision）

	TF-IDF	LLDA	增強 LDA	增強 LDA
實驗一	-	0.2040	0.3633	**0.4752**
實驗二	-	0.3413	0.4061	**0.5272**
實驗三	0.1553	0.2639	-	**0.3249**

5. 主題詞彙分析

除了以最終人工標記之解答所進行的量化分析，我們也將各主題所聚集的主題詞彙交由團隊內史學專家進行一些質化研究，其中某些主題底下隱含大量的人名詞彙引起團隊的注意，部分人名整理如下表 8：可以發現許多人名與該主題有相當程度的關聯，例如在「自由民主的基本概念」之主題下，胡適，即胡適之，為近代中國著名的自由主義者，同時也是《自由中國》雜誌創辦人雷震之自由民主思想的啟迪與推促者；而杜威則是胡適的老師，曾對中國教育界以及思想界有著重大的影響，其反對傳統的灌輸和機械訓練、強調從實踐中學習的教育主張，對中國著名的教育家、思想家都有一定影響。其他如中山先生、洛克、盧梭等人皆為著名的政治、哲學、教育思想家，對於近代自由民主思想、人權運動的啟蒙皆有相當程度的貢獻，而馬克斯、希特勒則分別代表共產主義及法西斯主義，可說是自由主義的反面論述。

表 8 主題群聚之人名詞彙整理

自由民主的基本概念	胡適、中山先生、馬克斯、盧梭、希特勒、杜威、洛克
法治	蔣公、白鵬、張君勱、黃啟瑞、張釋之、李璜、劉金華、黃黎洲
其他基本權問題	蔡金塗、盛世才、程維賢、李葆初、李唐、孫秋源、馮丹白
司法	谷鳳翔、延憲諒、孫秋源、李國楨
地方自治	李福春、鳳姐、黃啟瑞、李賜卿、周氏、黃千里、王國柱
反共救國會議	張君勱、董時進、唐筠卿、鄭所南

再來我們看到「司法」主題所群聚的人名詞彙，可以發現其結果與民國 47 年所發生的「奉命不上訴」之司法案件相關，民國 47 年，台中地檢處檢察官黃向間起訴南投縣長李國禎貪瀆，一審判決無罪；黃向堅不服，準備上訴，不料卻被當時的首席檢察官延憲諒批示「奉命不上訴」，而當時的司法行政部部長即為谷鳳翔。而孫秋源為《自治雜志》之經營人，其在 1961 年曾和經商的廖啓川因涉嫌反國民黨以及主張「台獨」，分別於居處遭台灣「警備總部」逮捕。

在其他的主題底下，同樣可以發現相關的人名出現，像是「地方自治」裡的黃啟瑞、李賜卿；「反共救國會議」底下所出現的張君勱、董時進，在各主題之中團隊史學家皆找到許多代表性的人物，而透過群聚人名詞彙之質化分析後，我們得以證實中文文本在經由強化標記式主題模型演算法所得之結果與其相對主題標記的關聯度。

(三)小結

在實驗階段，我們首先對資料集進行斷詞、去除標點符號以及一字詞等前處理，接著將主題標記以及種子字詞標記加入至演算法之中。實驗結果發現，在考量長字詞作為權重之後，各主題之間所群聚出之主題詞彙，其表現意義更加明確，並且有更多的人名及專有名詞出現在前一百個群聚結果之中，此結果將有利於專業人士進行文本分析。此外，我們將事後標記的種子字詞加入其中，並且利用召回率（Recall）以及平均準確度均值（Mean Average Precision）做為驗證，由結果可知本論文中提出的增強型 LLDA 演算法於中文文本的主題聚類的表現上皆優於傳統的 LLDA 演算法，最後我們再與傳統關鍵詞提取的 TF-IDF 方法比較，驗證了主題模型所群聚出詞彙的獨立性。

五、 結論

本研究基於傳統的標記式主題模型（Labeled Latent Dirichlet Allocation），開發了適用中文史料文本的分析方法。我們企圖利用中文長字詞意明確的特性，加上自行定義的主題詞彙（稱之為種子字詞），讓主題模型所群聚的隱含主題詞彙更加明確易讀且切合主題意義。在實驗中將改良之演算法與傳統標記式主題模型（LLDA）與傳統關鍵字提取方法（TF-IDF）進行比較，結果發現長字詞與種子字詞的加入確實能讓演算法群聚出更多有意義之隱含詞彙，並讓文本擁有者或專業領域人員作進一步的分析，使其對該文本有更進一步的了解。主題模型的概念已出現約十餘年，相關研究大多聚焦在其後續應用之開發；而本研究主要關注於群聚詞彙之易讀性的改良，並嘗試由中文詞語特性切入，對於後續主題模型應用於中文文本分析應有些許貢獻。在未來研究方向則可以對語言特性做更深入的分析，如詞性對詞語意義的影響，此外，也能將改良後的模型套用至以往常見的應用，例如：文件分類、資訊檢索、摘要提取等應用。然而，本研究最大的限制在於文本的收集，文本擁有者必須對相關領域有一定了解，才能對個別主題進行標記，或者提供自行定義的詞彙，應用門檻相對提高。不過現今的社群網路平台多有提供標記功能，相信未來要取得已附標記之文本的難度應會有一定程度的降低，本研究的應用亦會更加廣泛。

致謝

本篇論文得以完成首先必須感謝數位人文計畫團隊，在合作的期間團隊密集的開

會、溝通以及相互幫助。感謝雷震中心薛老師的團隊，你們所提供的文本以及完整後設資料的標記，是本研究幕後最大的功臣。再來要感謝蔡銘峰老師，不論在演算法的改進方向、實驗設計以及論文的撰寫與編修，都非常感謝老師的指導以及幫助。在本次投稿的論文撰寫部分，也要特別感謝子揚以及團隊，有你們在質化分析上的幫助，此論文才能夠的完成也順利被評審接受。最後，感謝薛老師以及劉老師在每次開會的討論中給予的寶貴意見。

參考文獻

D. M. Blei. (2012) Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

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

G. E. Hinton and T. J. Sejnowski. (1999). *Unsupervised learning: foundations of neural computa- tion*.

T. Hofmann. Probabilistic latent semantic indexing. (1999). In *Proceedings of the 22nd annual in- ternational ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

R. V. Lindsey, W. P. Headden III, and M. J. Stipicevic. (2012) A phrase-discovering topic model using hierarchical pitman-yor processes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 214–222. Association for Computational Linguistics.

D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics.

Y. W. Teh. (2006). A hierarchical bayesian language model based on pitman-yor processes. In *Pro- ceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992. Associ- ation for Computational Linguistics.

X. Wang, A. McCallum, and X. (2007). Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 697–702. IEEE Computer Society.

Intra-language Text Alignment Using *iAligner*

Tariq Yousef[*], Chiara Palladino[**], Gregory Crane[***]

Abstract

The purpose of the paper is to introduce an in-development tool for intra-language text alignment. The tool uses syntax-based dynamic programming methods to compute the optimal alignment of two or more parallel texts in the same language. The paper will introduce the subject of intra-language text alignment, report on the chosen algorithm and its modifications to perform optimal alignment, then focus on a variety of use cases from the field of historical languages, with particular regard to Ancient Greek and Latin.

Keywords**:** Intra-language alignment, natural language processing, alignment algorithm, historical documents, ancient Greek, Latin.

[*] PhD Candidate, University of Leipzig. Email: Tariq.Yousef@uni-leipzig.de.
[**] PhD Candidate, University of Leipzig. Email: chiarapalladino1@gmail.com.
[***] Alexander von Humboldt Professor of Digital Humanities, University of Leipzig; Adjunct Professor, Classics, Department of Computer Science, Tufts University. Email: Gregory.Crane@tufts.edu.

使用 *iAligner* 進行語言內文本並列比對

Tariq Yousef[*], Chiara Palladino[**], Gregory Crane[***]

摘　要

本論文目的為介紹一個發展中的語言內文本並列比對工具。本工具使用的方法為以語句結構為基礎的動態程式，計算同種語言的二或多個平行文本間之最佳並列比對。本文將介紹語言內文本並列比對的主題、說明所選擇的演算法及用以進行最佳並列比對之修正方案，然後專注於來自古老語言領域的各種使用案例，特別是古希臘語及拉丁語。

關鍵字：語言內並列比對、自然語言處理、並列比對演算法、歷史文件、古希臘語、拉丁語

[*] 萊比錫大學博士候選人，Email: Tariq.Yousef@uni-leipzig.de。

[**] 萊比錫大學博士候選人，Email: chiarapalladino1@gmail.com。

[***] 萊比錫大學亞歷山大·馮·洪堡德數位人文講座教授、塔弗茲大學電腦科學系院暨古典文學系兼任教授，Email: Gregory.Crane@tufts.edu。

1. Introduction

Text alignment is one of the most important tasks in Natural Language Processing, and an important supporting task for statistical machine translation methods [1]. In general, text alignment tries to define the correspondence between two or more parallel texts: when the texts are in two different languages, it is called cross-language alignment; if they are in the same language, it is called intra-language alignment.

Automated alignment methods on parallel texts have much potential in the field of Textual Criticism: they provide high support for the individuation of variants across different versions of the same text. Therefore, they are particularly useful in the philological process of collation and the consequent reconstruction of the manuscript transmission [2]. In the field of Western Modern Literature, computational methods for automated textual comparison have been particularly successful, especially in the case of alive texts, *i.e.* where the writing process is documented by multiple authorial versions [3]. However, historical sources, such as ancient Greek or Latin works, provide insights into a different situation: ancient texts as we read them are the result of a reconstruction, depending on a sometimes nebulous transmission, where the degree of authorial contribution can only be derived from circumstantial evidence. It is, therefore, highly important to document the editorial process of the selection of variants used to reconstruct the text printed in the edition: this task is traditionally assigned to the so-called *apparatus criticus*, which, however, is in itself a choice and eliminates the context of each editorial decision [4]. Digital Humanities methods offer the opportunity of overcoming these limitations: not only they facilitate the mechanical task of comparing texts and individuating the divergences, but they can also help to document the process and visualize it in its completeness.

2. Related Work

In the field of Digital Humanities, there has been a strong interest in computer-aided comparison of multiple text versions since the early 2000s. Various projects have been developed in this field: Juxta Commons (http://wiki.tei-c.org/index.php/JuxtaCommons) is a family of softwares that enables users to see how texts change between various editions and revisions up to 20 versions. The software automatically compares various versions, showing them on an interface where the user can compare the differences. Juxta provides different types of visualization: the texts can be explored as overlapping versions, with highlights for areas of

variation between witnesses; changes can be displayed side by side, where connecting lines in the corresponding shades indicate areas of divergence; histograms with bars of varying lengths to feature the amount of divergence.

Juxta Commons has been integrated with the Versioning Machine as an alternative visualization. The Versioning Machine (http://v-machine.org/) provides a comparison set based on panels: it displays multiple versions of texts encoded in a basic TEI XML format, allowing comparison of various types of diplomatic transcriptions or editions. The visualization provides also a user space for annotation, reading in-line textual notes and bibliographical references, and an image viewer. The comparison panel displays the full versions of the compared texts, highlighting divergences when the cursor moves to single areas.

The CollateX software follows the Gothenburg model, resulting from a 2009 fusion with Juxta (http://collatex.net/), and whose primary concern was to separate various computer-supported tasks of collation which were independently developed by the two projects. CollateX is a software for expressing textual variance by using a graph-based data model. It offers a simple tokenizer, which splits plain text or evaluates a customizable XPath expression or an XML document with a list of node values. In the preprocessing stage, the only optional configuration is normalization/regularization, mainly dealing with punctuation and orthographic differences through word stemming. The alignment itself is then realized by finding a set of matching tokens and aligning them in the visualization data model, which follows the tabular representation of collation results by sequence alignment, but also models textual variance internally with Variant Graphs, where each edge contains label tokens indicating the compared versions and their relations of equivalence. Variant Graphs have also been implemented by the interactive interface of Stemmaweb (https://stemmaweb.net/) to allow user annotation and modification of the structure. Both tools provide plain horizontal layouts without highlighting or distinguishing the essential features of the graphs: the visualization has been recently implemented by Stefan Jänicke with the Sentence Alignment Flows [5].

The most recent project of critical textual comparison focused on ancient documents, developed in Leipzig by the team of Charlotte Schubert, is called eComparatio [6] and allows the comparison of an arbitrary number of texts which are shown in different panels with the variations highlighted in the visualization interface. The output of the application is aimed at

the creation of an apparatus of variants for born-digital editions of ancient texts. It is, therefore, especially aimed at critical and philological work.

The majority of the mentioned softwares does not offer complete text alignment visualization, as the type of comparison performed focuses on the individuation and measurement of divergences, without including spaces of similarity. *iAligner* offers the complete range of alignment, providing a way to visualize similarities, gaps and divergences across multiple texts, and also including various refinement options available to the user, according to the possibly different purposes for the textual comparison.

3. Methodology

In general, we differentiate between two types of alignment, according to the number of parallel texts: pairwise alignment for two texts, and multiple alignment for more than two texts. The former detects similar patterns between two parallel sentences: insertions and deletions are used to transform one sentence into the other, in order to produce the highest similarity score.

Three main types of algorithm are used to produce pairwise alignment: dot-matrix methods, dynamic programming, and word methods [7]. *iAligner* applies a dynamic programming method to produce pairwise alignment by using a modified version of the Needleman-Wunsch algorithm, which is used in Bioinformatics to perform optimal alignment of DNA sequences [8]. The method used for *iAligner* does not use any lexical or syntactic information (POS, lemma, etc.) to perform the alignment: instead, it takes into account the order of the words in a sentence, and the order of the characters in a word.

3.1. The alignment Algorithm

Given two sentences S1 and S2 with length n and m respectively, we put the in the form of two-dimensional array (similarity matrix M, and fill it with values according to a scoring function. Then, we find the alignment score at M[m][n]: we start from this point and follow the arrows to the original cell M[0][0] to get the optimal alignment.

The score function is used to measure the similarity between the two sentences. It takes three cases into account: matching (*m*), mismatching (*mis*) and insertion/deletion (*indel*). The last two are assigned as penalties, and *indel* produces a gap in the aligned output. The values assigned to each case are determined experimentally: if we consider the mismatching as the

worst case, then we assign it a higher penalty than the gap, while if the gap is considered worse than mismatching, we assign a higher penalty to the former [**Table 1**].

$$M[0][j] = j * indel \quad where \quad j \in [0, n],$$
$$M[i][0] = i * indel \quad where \quad i \in [0, m]$$

$$M[i][j] = max \begin{cases} M[i-1][j-1] + m \searrow & Matching \\ M[i-1][j-1] + mis \searrow & Mismatching \\ M[i][j-1] + indel \downarrow & Gap\ in\ S1 \\ M[i-1][j] + indel \rightarrow & Gap\ in\ S2 \end{cases}$$
$$Where \qquad 0 < i <= m \ , \ 0 < j <= n$$

Table 1. The score function of the alignment algorithm

3.2. The modification to the Needleman-Wunsch Algorithm

Given two sentences S1 and S2 with length *n* and *m* respectively, the Needleman-Wunsch algorithm in its basic form compares each token of S1 with each token of S2, and produces a search space = *n* * *m*. We optimized this algorithm by reducing the search space, since we do not need to compare each token for each sentence: our algorithm compares a token W at the position *i* in S1 with a range of tokens [*i*-*k*, *i*+*k*] in S2 with length of 2*k*+1. The search space is therefore reduced from *n* * *m* to 2*k* * *m*, where *k* < *n*/2; the value of *k* differs according to the two compared texts: if the texts have approximately the same length, the best value of *k* is set at 5. In the case of two texts with a large gap in length, the value of *k* can be changed to *n*/4 [**Figure 1**].

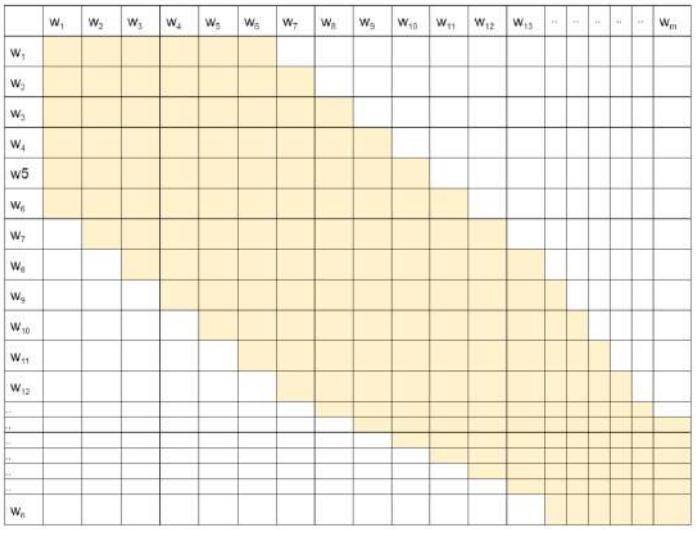


Figure 1. Reduction of the search space in the optimized Needleman-Wunsch algorithm. The white cells represent the search space of the algorithm in its normal form, the yellow cells the optimized search space.

We also modified the score schema in accordance with various refinement options made available to the end user (see below).

4. Workflow

The web service currently allows to perform the alignment in two ways: by uploading files in tabular or plain format, or by directly pasting the text on the editor provided in the website. As a third option, it is also possible to run the Python code independently: the code can be accessed via the Github repository of the project (https://github.com/OpenGreekAndLatin/ILA_python).

The input text is parsed and splitted into a list of text groups, each group consists of multiple parallel sentences. During the preprocessing stage, the splitted sentences are passed through a simple tokenizer, which takes white spaces and punctuation marks as delimiters to split the sentence in a vector of single tokens. Once the tokenization is completed, the algorithm is applied to produce the output. According to the number of parallel sentences, *iAligner* selects which algorithm to use: if the group consists of only two parallel texts, it uses the pairwise alignment method to produce the output; if the groups consists of multiple texts, it uses progressive multiple alignment methods.

Various additional refinement criteria are directly available to the user to provide further options, depending on the specific purpose of the comparison:

- *Ignore non-alphabetical*: ignores symbols such as punctuation and numbers, anything that is not an alphabetical character.
- *Case sensitive*: detects variation between words according to the case.
- *Ignore diacritics*: ignores any type of diacritical character, including punctuation.
- *Levensthein distance*: it is a revised version of the Levensthein algorithm [9], used to increase the tolerance threshold on the alignment of similar words. The threshold is currently 66%, but in the future it will be possible to adjust it according to the users' needs.

5. Alignment Output and Color-key

The aligned sentences are displayed in different shades of color according to the degree of match and to the chosen refinement criteria. When the sentences are not completely aligned, the longest common substring is also displayed at the bottom. In general, perfectly matching tokens are highlighted in green, non aligned tokens are in red.

The color-key currently differs between pairwise alignment and multiple alignment, but it is going to be harmonized in a second stage of the work. In the case of pairwise alignment, the main visualization option displays the parallel texts on two lines; tokens aligned according to the case and ignored nonalphabeticals are displayed in light green; tokens aligned according to the Levensthein distance are in blue-green. Gaps are also individuated and highlighted in yellow [**Figure 2**].

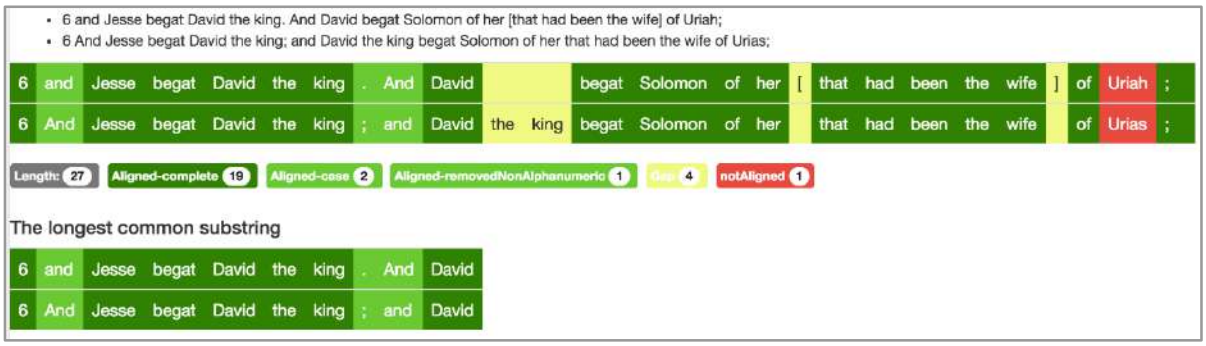
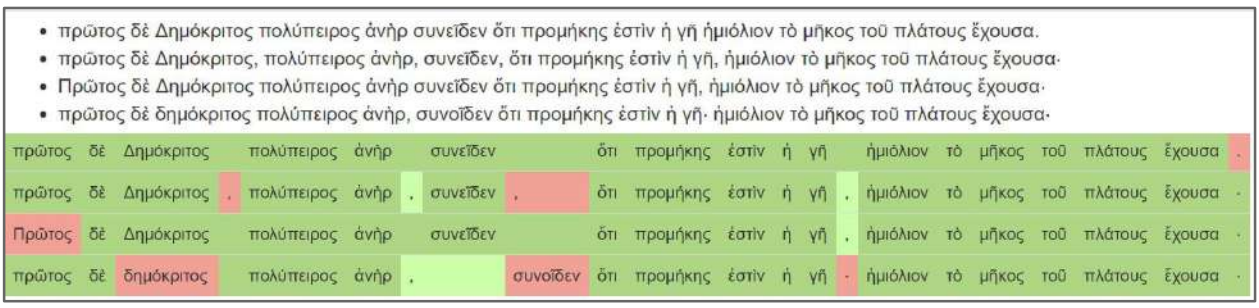


Figure 2. An example of pairwise alignment visualization

In multiple alignment, the visualization is currently simplified to the basic version to facilitate the individuation of variants, but it will be harmonized with the previous one in future developments. Similar tokens across individual witnesses are highlighted in light green,



isolated variants are, as always, in red [**Figure 3**].

Figure 3. An example of multiple alignment visualization

6. Evaluation

In order to evaluate our results, we have used the F1 score, which measures the accuracy of an experiment by measuring the recall *r* , which reflects the sensitivity of the algorithm, and precision *p*; *p* is the rate of correct positive results related to all positive results, and *r* is the rate of positive results to the number of all positive results that should have been returned [**Table 2**]. The value of the F1-score ranges from 0 (i.e., least accurate) to 1 (i.e., most accurate). The F1-score is computed by the following formula:

$$F1 = 2 (p \ r) / (p + r)$$

We performed the evaluation on three samples of Ancient Greek, Latin and Arabic. The evaluation on Greek and Latin was measured on two long sample texts from the *Patrologia Graeca* and the *Patrologia Latina* respectively: in both cases we aligned two good OCR outputs of two samples from Vol. 72 and 9 respectively (see below), and measured the degree of match among tokens. The arabic texts used to evaluate the algorithm are three different translation of the bible: we used the Simplified Arabic Translation (SAT), the Good News Arabic Bible (GNA) and the New Arabic Version (NAV). In this case, we selected some chapters an aligned them: the recall and F1-score appear to be slightly lower than in the two other cases, as the conjunction "and" (و) in Classical Arabic is connected to the following word: this case typically tends to produce errors in the alignment performance.

	Pairwise Alignment							
	Tokens	Aligned correctly (TP)	Not Aligned correctly (TN)	Aligned incorrectly (FP)	Not Aligned incorrectly (FN)	*p*	*r*	F1
Greek	**1117**	**678**	**439**	**0**	**0**	**1**	**1**	**1**
Latin	**1083**	**827**	**211**	**0**	**0**	**1**	**1**	**1**
Arabic	**1250**	**868**	**345**	**0**	**37**	**1**	**0.96**	**0.98**

Table 2. Evaluation on pairwise alignment

7. Case Studies

7.1. Manuscript Alignment

Manuscript collation is a typical case study for philological analysis, and automatic textual alignment can offer an important aid in this process. We experimented with the transcriptions of three manuscripts of Plato's *Crito*, the Clarkianus 39 (B, saec. IX), known as one of the most important witnesses, the Tübingensis Mb 14, (C, saec. XI), whose importance as an independent witness has only recently been recognized, and the Parisinus Graecus 1808 (saec. XII), probably an apograph of the Marc. App. Class. IV 1 (T, saec. X).

The study of Plato's transmission is still at its beginning, due to the long-lasting neglect of two manuscript families, hastily removed from the stemma to the advantage of the sole Clarkianus [10]. The independence of C from B has been convincingly demonstrated in the case of *Phaedo* [11], and it calls for a re-examination of the manuscripts of the *Crito* as well [12].

We processed the three transcriptions directly from the command line, and displayed the results on a dedicated section of the iAligner website (www.i-alignment.com/crito/). The preprocessing stage required a limited amount of human intervention in the choice of the limits of the sentences to align, as a rough textual division was already provided by the captions of the dialogues' *personae*.

A general analysis of the output has shown that C tends to differ from B especially in variants of transposition, but there are also some instances where it curiously diverges both from B and from the Parisinus. None of these variants is significant to the interpretation of the work, but they are of some relevance as disjunctive errors, as they cannot be easily motivated as scribal distractions or omissions. Examination of the evidence reinforces the idea that the Tübingensis belongs to an independent family, and it deserves further attention. Here we show the most interesting instances.

The texts of the transcription are shown in the following order: 1. Clarkianus 39 (B); 2. Parisinus 1808; 3. Tübingen Mb 14 (C) .

Figure 4 shows a section of the comparison of *Crito*, 47c-d. This piece is particularly interesting in the alignment, as it shows some instances of typical divergences both in the Parisinus and the Tübingensis. The transposition of the sequence ἡμῖν ἐστιν after νῦν and before ἡ βουλὴ is specific to this manuscript; on the other hand, the most substantial difference is in this case the pronoun at the end of the sentence, where B has the usually accepted lesson αὐτὴν (referred to δόξα above), against ταυτὴν in the Parisinus and αὐτοὺς in the Tübingensis.

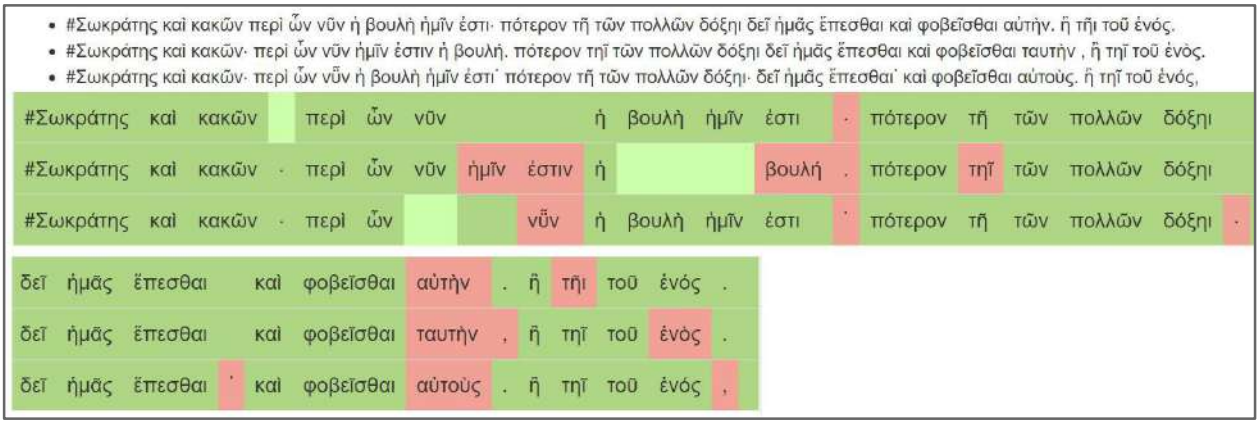


Figure 4. *Crito*, 47c-d. αὐτὴν B vs ταυτὴν Paris. vs αὐτοὺς C.

Figure 5 shows a comparison of section 52a of the *Crito*. The article of the vocative ὦ Σώκρατες is missing from the main manuscript, where in this case the main edition by Burnett follows the right lesson ὦ Σώκρατες (although it is not stated in the *apparatus* which witness was followed in particular in this case). This is a typical instance where it is difficult to hypothesize scribal distraction, as small sections of words such as articles or adpositions are not likely to be arbitrarily added by copyists, rather omitted. Moreover, the alignment offers another instance of transposition: where B has καὶ σὲ σώκρατες, essentially followed by the Parisinus which has καὶ σὲ ὦ σώκρατες, C transposes to ὦ σώκρατες καὶ σὲ.

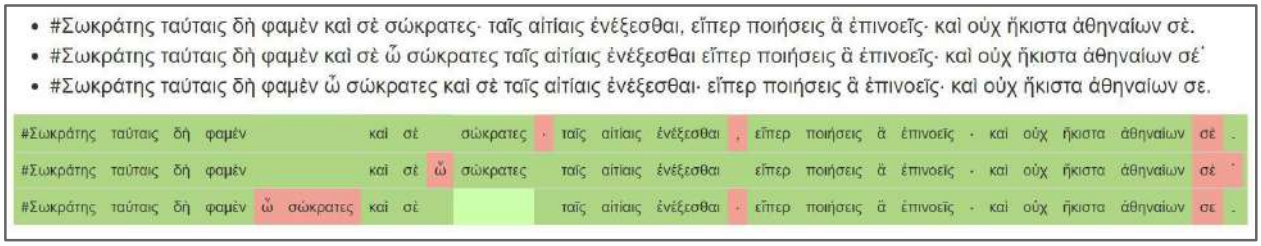


Figure 5. *Crito*, 52a.

Figure 6 shows section 52b aligned. This case is again significant for the divergence in the variant οὐδαμῶς σε C vs οὐδαμόσε of B and Parisinus.

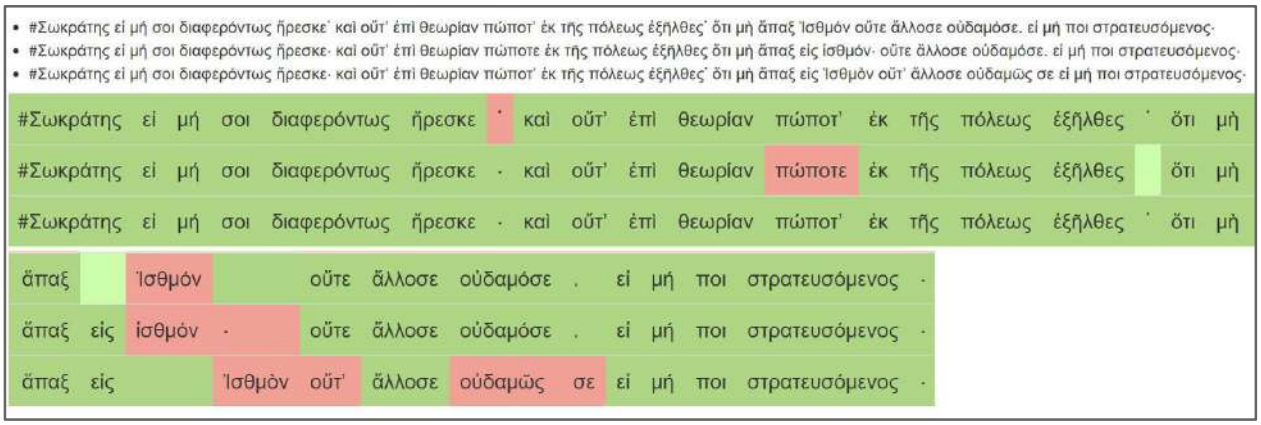


Figure 6. *Crito*, 52b. οὐδαμῶς σε C vs οὐδαμόσε of B and Parisinus.

7.2. OCR Output Alignment

Dynamic programming for the detection of textual variants is not limited to critical textual work, but can be also fruitfully used in order to reduce the amount of human intervention on long correction tasks that require the comparison of very large chunks of texts.

An example of this situation is offered by the stage of post-correction of OCR outputs. To improve the correctness of OCR-generated texts more outputs are generated and then compared. Therefore, we used the tool in order to detect residual errors in corrected Data Entry and to improve the accuracy in the workflow. We aligned two OCR outputs of the Patrologia Graeca and created an environment for manual correction, where users can choose between two alternatives recognised as variants by the tool and provide additional data to improve the performance of the OCR engine.

We tested the workflow on two extracts from vol. 72 of the *Patrologia Graeca*, the large collection of Christian authors writing in Greek, consisting of 161 volumes and put together by Jacques Paul Migne between 1856 and 1866. The Greek text of the scanned volume was processed through the same OCR pipeline to obtain two slightly varied text outputs, which were then aligned directly from the command line (www.i-alignment.com/ocr/), and used for the evaluation on Greek (see above). We also created an environment for the choice of the correct variant [**Figure 7**], which can be used in the context of collective manual post-correction, frequently used in the field of Citizen Science. This is going to provide a suitable framework for digitization efforts on large corpora that need nevertheless the human eye to be perfected in their conversion to the new format.

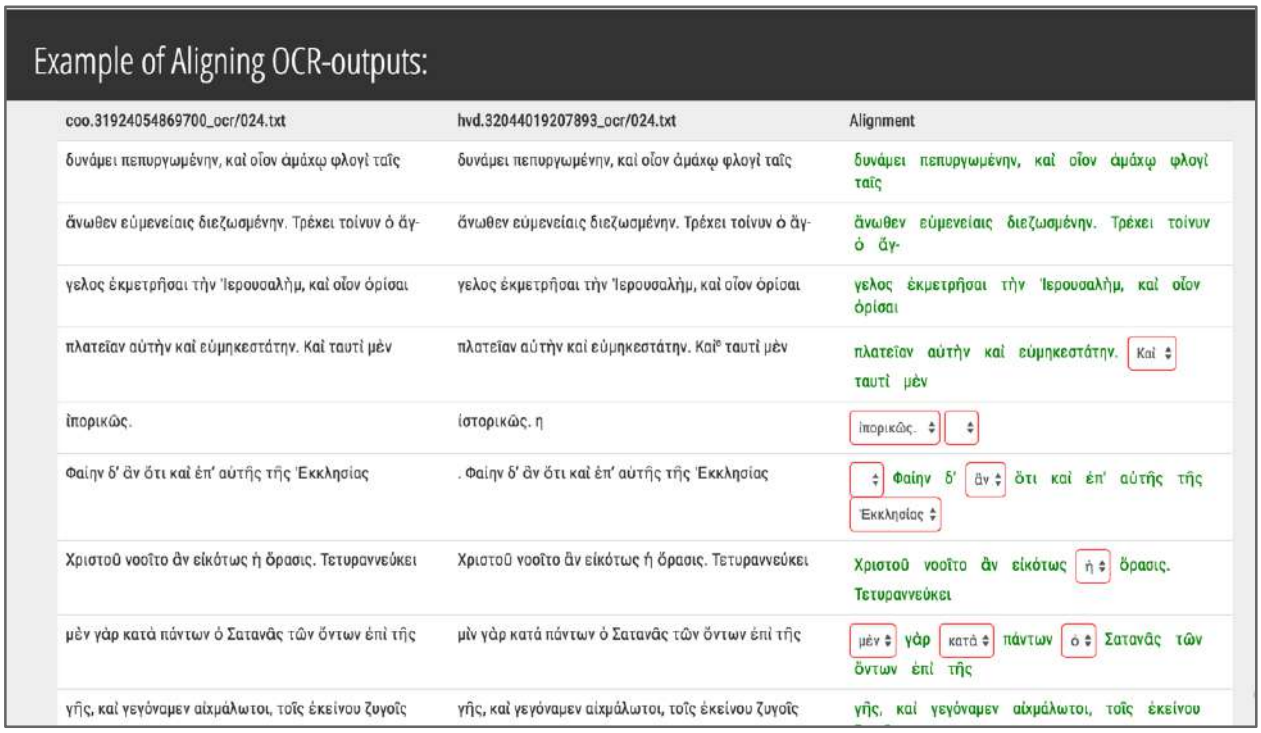


Figure 7: alignment of two OCR outputs from the *Patrologia Graeca*. The third column shows the overlapping sections and offers the user the choice between two variants where the two texts diverge.

It is also possible to compare the resulting TEI XML file to another source, whose correctness is higher, which may be another XML transcription, a plain text or HTML file, or raw OCR-generated text in hOCR. The output is a version of the TEI XML source file that has <corr> or <sig> tags added to capture possible variants. The aim of this particular alignment is essentially to test OCR performance against a perfectly correct version of the same collection. We tested this approach on two texts extracted from Vol. 9 of the *Patrologia Latina*, the second big effort by Migne, consisting of 221 volumes and currently involved in a progressive

collective digitization effort within the Open Greek and Latin project[1]. One of the texts was extracted from a private collection and was a perfected XML version, the other was a rough OCR output [**Figure 8**]. The results of the comparison were very encouraging and we examined and detected recurring errors manually, which will then be corrected partly automatically and partly by users.

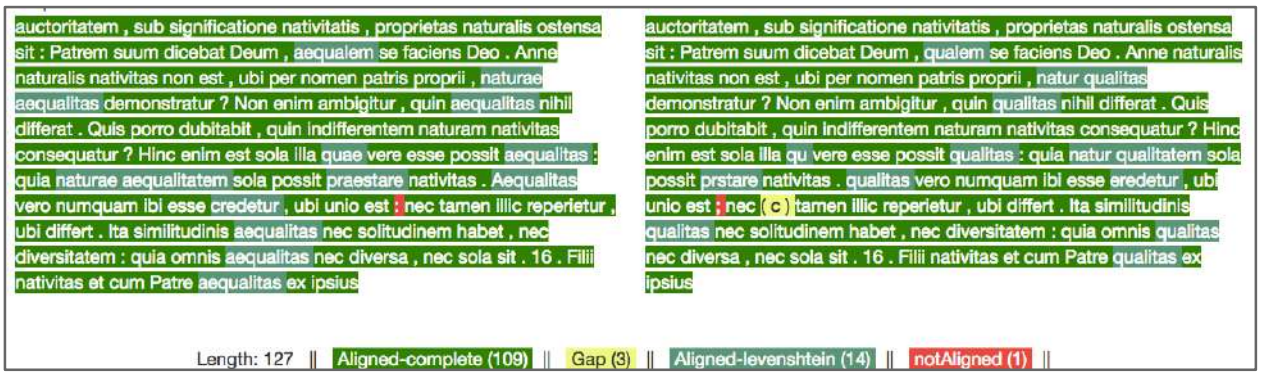


Figure 8: *Patrologia Latina*: OCR output vs. correct version: www.i-alignment.com/pl/

7.3. Edition alignment

From the scholarly point of view, the comparison of editions of contested texts is increasingly important, either for the history of the transmission of a text, or for the history of Classical Scholarship in itself.

We tested this approach by running two tests on two different editions of Aeschylus' *Suppliant Maidens* [13][14] [**Figure 9**]. This tragedy is highly problematic in terms of tradition, dating and *constitutio textus*: being transmitted by one witness only, the so-called Codex Mediceus, Laurentianus 32, 9 of the 9th or early 10th century, traditionally indicated with the letter M, which contains all Aeschylus' seven tragedies [15]. The text offered by this manuscript is not only heavily corrupted, but was also subject to alterations before the Mediceus was copied; the difficulty of the text adds to the controversial problem of the date of the play and its place into the not surviving tetralogy of the *Danaids* [16]: despite being traditionally considered the oldest tragic play surviving in European culture, the recent discovery of the P. Oxy 2256, containing a *hypotyposis* to the *Suppliants*, has changed the situation once more, suggesting different hypotheses of datation [17].

1 The project is maintained by the Universities of Leipzig and Tufts (http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/). The repository is publicly available on GitHub: http://opengreekandlatin.github.io/patrologia_latina-dev/.

In this situation, to be able to retrace the history of the critical editions of the text across time can be helpful not only to inspect how the modern tradition changed according to hypotheses of datation and philological reconstructions, but also to recollect the divergences in interpretation and conjectures advanced by editors across time. In a future stage of the work, when the download option in XML format will be made available, it will be possible to create whole collections of conjectures as they can be extracted from diverging editions.

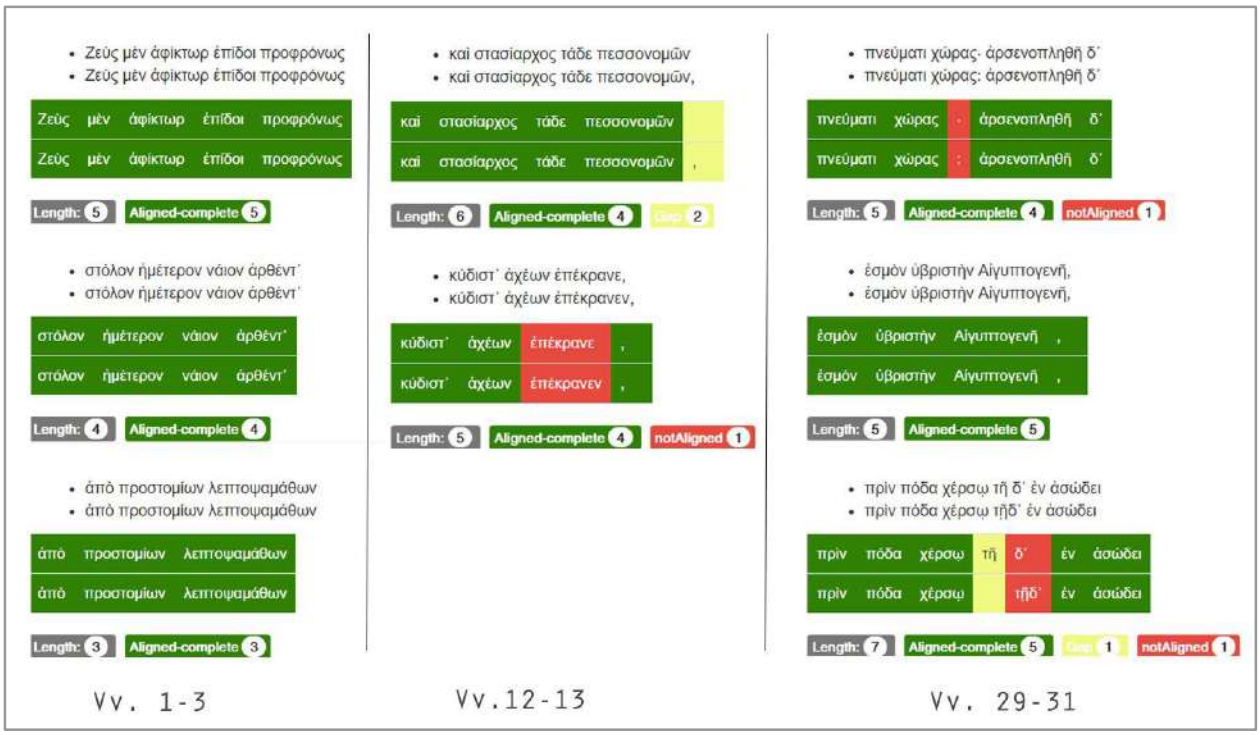


Figure 9: three excerpted sections of *Supplices* aligned. www.i-alignment.com/Aeschylus

8. Conclusion and Future Work

Current results encourage both applications on scholarly editorial practice and on larger efforts for the detection of a high amount of variants.

Future efforts will be addressed at providing a variety of download options of the aligned texts, in order to make the variants available in a workable format such as XML and CSV. We are currently improving the performance of the tool on Classical Arabic, and have recently introduced some language-dependent options, designed in the code to overcome existing ambiguities which are peculiar to specific languages, such as Latin, Greek and Arabic. These options will be refined to allow an even more precise and language-focused alignment performance.

Current limits in the alignment output, such as the minor capacity in handling crossing instances, due to the use of a syntax-based algorithm, will also be addressed.

Acknowledgements

Our thanks to the Humboldt Chair of Digital Humanities of the University of Leipzig for supporting the project. The transcriptions of the manuscripts of the Crito *were generously provided by Ana-Katerina Gorički and Giuseppe Celano.*

References

Berti, E. 1976. I manoscritti del Critone di Platone e la prima famiglia dei mss.. *Hermes*, 104,2, 129-140.

Boschetti, F. 2007. Methods to extend Greek and Latin corpora with variants and conjectures: mapping critical apparatuses onto reference text. *Proceedings of the Corpus Linguistics Conference* (Birmingham, July 27-30).

Carlini, A. 1972. *Studi sulla tradizione antica e medievale del Fedone*, Edizioni dell'Ateneo, Roma.

Dekker, R.H., van Hulle, D., et al. 2014. Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Digital Scholarship in the Humanities*, Oxford, Mar 2014.

Jänicke, S., Büchler, M., et al., 2014. Improving the Layout for Text Variant Graphs, *VisLR: Visualization as Added Value in Development, Use and Evaluation of Language Resources, Proceedings*, 41-48.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10,8, 707–710.

Lobel, E. 1952. *Oxhyrhynchus Papyri* XX, n. 2256 fr. 9a.

Makedon, F., Owen, M., et al. 1998. HEAR HOMER: A multimedia-data access remote prototype for ancient texts. *Proceedings of ED-MEDIA'98, World Conf. on Educational Multimedia and Hypermedia (June 20-25)*, Freiburg.

Mount, DM. 2004 (2nd ed.). *Bioinformatics: Sequence and Genome Analysis.* Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.

Needleman, S.B., and Wunsch, C.D. 1970. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology,* 48,3, 443–453.

Pasquali, G. 1959 (2nd ed.). *Storia della tradizione e critica del testo*, Le Monnier, Firenze.

Salanitro, G. 1968. La data e il significato politico delle "Supplici" di Eschilo, *Helikon* 8, 311-340.

Schubert, C., Meins, F., et al., 2016. eComparatio - Editionsvergleich, in *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, 882-883 (http://dh2016.adho.org/abstracts/93).

Sidgwick, A. 1902. *Aeschyli Tragoediae. Cum fabularum deperditarum fragmentis poetae vita et operum catalogo*. Clarendon Press, Oxford.

Smyth, H.W. 1922. *Aeschylus,* Suppliant Maidens, *with an English translation*. Vol. I, W. Heinemann, London.

Turyn, A. 1943. *The manuscript tradition of the tragedies of Aeschylus,* Polish Institute of Arts and Sciences in America, New York City.

West, M.L. 1973. *Textual criticism and editorial technique,* De Gruyter, Stuttgart.

Paper Session 2

新計算思維：以人文為本

Human-centered Computing

資料計算於數位人文研究意涵的省思

劉吉軒*

摘　要

數位人文根基於數位科技與人文議題的交會與融合。本文基於跨領域研究經驗的累積，以數位框架的觀點，嘗試歸納資料計算對於人文研究的意涵，並以部分國內研究案例具體情境，對應抽象意涵的實際展現，提供數位人文共同成長的參照與省思。本文提出資料計算型塑數位人文研究典範的本質特徵，包括：規模、維度、智能、視覺呈現、脈絡連結、獨立檢驗、跨越框架基礎等，協助年輕世代學者建立跨領域合作的路徑參照，也與相關學術社群共同歸納省思。數位人文跨領域研究可以從互動合作的物理變化到交融協作的化學變化，期待有志之士齊心協力、共同開創。

關鍵字：數位人文本質、資料計算、研究意涵特徵

* 國立政治大學資訊科學系特聘教授，Email：liujs@nccu.edu.tw。

Thinking on the Research Implications of Data Computation in Digital Humanities

Jyi-shane Liu*

Abstract

Digital humanities has been rooted on the intersection and fusion of digital technology and humanity study. Based on accumulated experiences in interdisciplinary research, this paper adopts a viewpoint of digital framework and attempts to characterize research implications of data computation on humanity study. The abstract implications are manifested by actual research contexts of several domestic case studies, so as to provide reference and contemplation for the collective growth of digital humanities. Essential characteristics of digital humanities research paradigm inferred by the data computation model include: scale, dimension, intelligence, visualization, contextualization, independent validation, and common grounding. These characterization may help establish transdisciplinary pathways for younger generation scholars and provide further consideration with the research community. Transdisciplinary digital humanities research entails both physical changes of collaborative interaction and chemical changes of creative fusion. It is hoped that more will join the endeavor and begin a great journey of new adventure.

Keywords: characteristics of digital humanities, data computation, research implications

* Distinguished Professor, Department of Computer Science, National Chengchi University, Taipei, Taiwan. Email: liujs@nccu.edu.tw.

一、前言

數位人文根基於數位科技與人文議題的交會與融合。早期的人文計算（humanities computing）以人文研究為主體，以當時功能有限的電腦為輔助工具，協助文本資料的初階分析。隨著數位科技大幅進步，數位科技於人文研究所扮演的角色開始提升至夥伴關係，數位人文宣言 2.0（Schnapp, et al., 2009）認為計算科技已成為探討現代人文議題的充分條件，不僅發展出混合式研究方法，更進而創新學科範式。隨後，Berry（2011）提出計算轉向（computational turn）概念，討論計算技術如何在傳統知識框架之外，轉變我們組織與使用資訊的能力，主張必須以計算為關鍵議題，才能更深入理解數位世界所帶來的各項人文衝擊。Hayles（2011）則思考數位科技是否帶來人文研究本質上的轉變，並認為數位素養（digital literacy）將是數位人文研究者必須具備的基礎能力。因此，人文學者與資訊學者應攜手建構數位型態研究特質的共同認知，掌握資料計算能耐概念，以驅動數位人文的持續開創與進展。本文基於跨領域研究經驗的累積，以數位框架的觀點，嘗試歸納資料計算對人文研究的意涵，並以部分國內研究案例具體情境，對應抽象意涵的實際展現，提供數位人文共同成長的參照與省思。

二、意涵特徵

資料計算是電腦運作的核心，所有的電腦應用都是各類資料與演算法的結合孕育而生。當人文活動及社會情境開始以數位型態表示與中介，這些數位摺疊的樣貌成為可供資料計算的對象，也產生了利用電腦機器龐大計算效能的機會。以數位人文研究中的文本分析而言，人類的文字語言成為計算資料，透過文本探勘、自然語言處理及計算語言學等演算分析技術，展開以資料為核心的實證研究，從議題發想到實驗假設與驗證，以各種資料計算結果提供的客觀證據，進行人文屬性的解讀與發現。本文認為資料計算帶來至少以下幾種轉變特徵，進而形塑數位人文的新研究範式。

(一) 規模

數位資料能被納入研究的範圍，僅受限於電腦機器的記憶容量及運算能力，以目前的技術標準而言，數百萬本、甚至千萬本書籍的文本規模已經是可實現的目標。因此，文本分析資料計算的規模可以百倍千倍的大幅超越研究者個人窮盡一生的閱讀認知能量，以遠距閱讀（distant reading）模式（Moretti, 2013），透過局部文字語言特徵

的搜尋比對統計分析，快速的取得聚焦資訊，並建立宏觀輪廓。國內中文數位文本已開始達到億字以上的規模，如中華電子佛典協會的歷代藏經（CBETA, 2016）、中國近現代思想及文學史專業數據庫等（鄭文惠，2014）。特定主題來源的文本，如《自由中國》政論期刊從 1949 年至 1960 年出版 23 卷，內容共計超過一千萬詞，《雷震日記》從 1948 年至 1977 年內容共計約一百四十五萬詞。研究資料的規模龐大帶來的重要改變包括研究問題的情境與內容，打開了過去受到資料面限制的思考空間，而能以更大的框架或更細微的觀察，進行議題的探索與假設的驗證。

(二) 維度

文本具有許多資料屬性，能被歸屬分類，而在維度空間上區隔分布，成為比較分析觀察發現的基礎。數位資料計算讓文本在多元維度的設定及維度空間的配置更容易也更快速，因而有助於更彈性的研究設計與更廣泛的議題探討。例如，二二八事件本地新聞史料彙編的文本中，可以設定事件爆發一週內，官方報紙與民間報紙的維度，進行類別文本之間的資料計算，比較報導立場與言論態度上的差異；也可以聚焦於單一官方報紙，以時間軸觀察議題或氛圍的變化，而和歷史研究對事件發展階段的定位相互參照驗證。另外，文本資料屬性有外顯與內隱之分，外顯屬性為原先具備而容易使用，但內隱的屬性則是內含於文本內容，人工辨識區分成本較高，必須依賴資料計算才能有效的使用。例如，雷震日記文本分析可以採用外顯屬性，即年度或人生階段（政論十年、入獄十年，筆耕十年），內隱屬性在資料計算的支援下，可以多元定義，如親屬家人或人事時地物等。因此，文本分析的維度空間更加寬廣，而提供更多的研究議題想像。

(三) 智能

數位資料從最基礎的 0 與 1 位元到抽象概念符號，有許多的表示層次及對應的資訊單元，讓各種分析應用目的的計算技術有非常寬廣的發展空間。以文本資料而言，人類的語言文字轉化為數位資料後，可以被拆解為一個字元，甚或一個音節的資料單位，而從音節、筆畫、部首、字詞、詞組、句子、段落、篇章、特定主題文稿集或無所不包的語料庫，都可以被設定為計算對象。過去持續爭論的人文議題，開始能被以計算方式，從大量資料中萃取出充分的證據資訊，而得到可供客觀邏輯演繹的結論。例如，杜協昌（2014）將《紅樓夢》前 80 回與後 40 回的內容，以計算方式比較詞彙特徵的顯著差異，從文本中發掘新事證，支持前後兩部為不同作者的結論。資料計算最重要的意涵是讓人文議題與機器智能接軌，許多舊議題能以新方法新角度論證，許

多新議題能被想像探索開發，人文研究開始進入結合人類智能與機器智能的新境界，未來的豐富研究圖像令人期待。

(四) 視覺呈現

資料的數位性質在各式計算方法的操作下，可以產生高度彈性的資料觀察能力，在各種分析工具的參數設定下，資料的維度與面向可以被輕易的選擇、調控與投射，而可以多視角呈現資料內容的豐富樣貌。當資料量遠超過一般人有限的資料解讀與資訊認知能力時，從資料中擷取出有用資訊的彈性呈現能力，可以提供宏觀的全貌描繪，快速取得統整性資訊或浮現出顯著資訊，也可以聚焦到特定對象或區塊，進行細微的爬梳檢視。資料計算的豐富資訊呈現方式再搭配資料視覺化技術，以多元圖像呈現資訊意涵，更直接訴諸人類的視覺訊息接收能力，展現出更有力的資訊傳遞效果，開發了人文研究的新型溝通媒介。例如，將中華民國政府遷台後至 2014 年間行政院內閣團隊之上司與下屬資料匯集統整，圖 1 呈現出行政權演化的宏觀結構，從早期的單軌延續到近期兩次政黨輪替而產生的三角形態，其中也部份揭露特定官員的特殊角色與位置，提供後續研究的參照資訊。

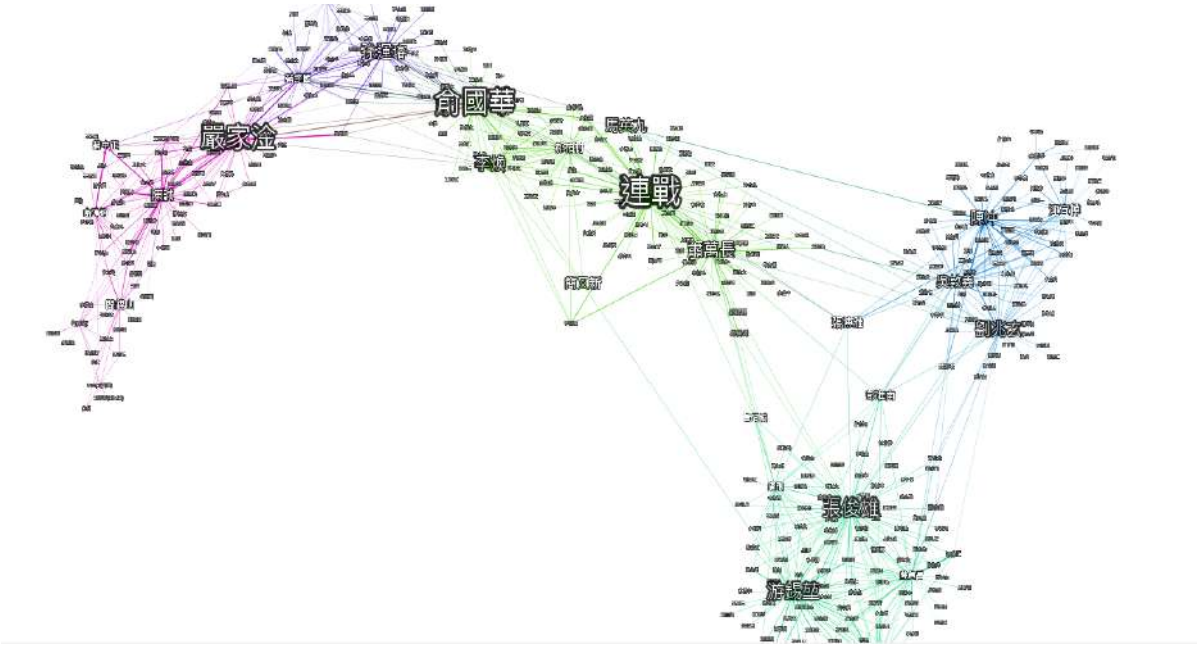


圖 1　中華民國遷台後行政院內閣演化結構(資料來源：作者)

(五) 脈絡連結

隨著人文資料的大幅成長，各種資料之間也將產生在主題、對象或概念上的連結，而能提供更完整的意義與樣貌。資料之間可能存在顯而易見的關聯，也可能隱藏著微妙的關聯，無論是主觀的認定，或是客觀的檢驗，都可以透過資料的計算分析過程，進行主導性的建立或是探索性的辨識，而取得資訊的連結與整合，協助建構出更完整的脈絡資訊與知識體系。例如，以中文約一萬四千個同義詞及兩萬七千六百餘個同義關係的同義詞典，可以快速建立同義詞彙關聯脈絡，並可聚焦特定概念，觀察局部或全貌資訊。圖 2 展示和含有「苦」字的詞有同義關係的局部詞彙網絡，圖 3 則為其整體詞彙網絡，發現帶有「苦」概念的同義詞彙共有 220 個，可分為 18 個獨立子群，27 個不同的次級概念，如「勞苦、勞碌……」、「勤苦、辛苦……」、「刻苦、吃苦……」、「儉樸、節儉……」等。

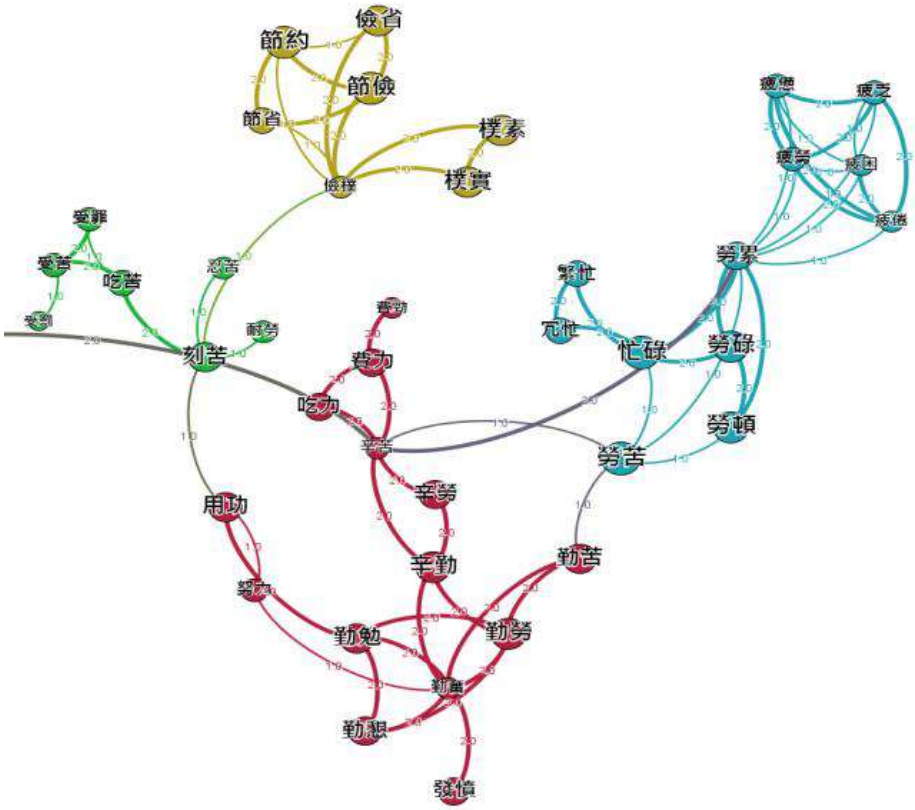


圖 2　以「苦」為例之同義詞脈絡呈現：局部網絡 (資料來源：作者)

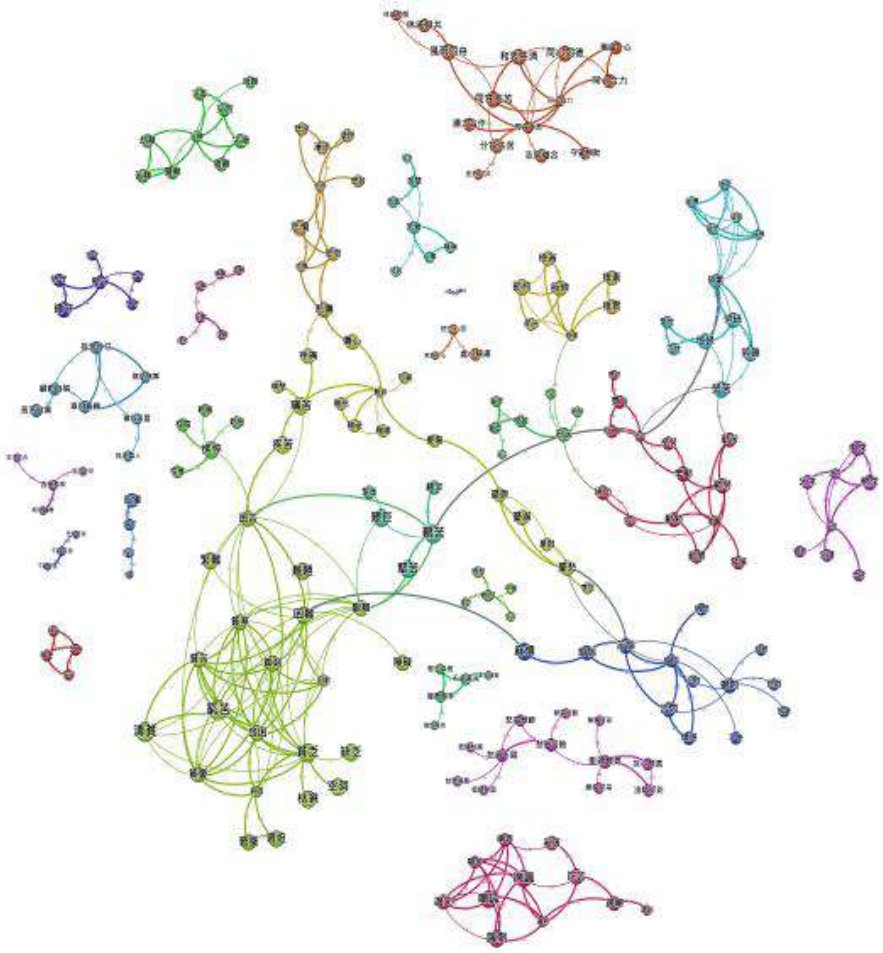


圖 3　以「苦」為例之同義詞脈絡呈現：整體網絡(資料來源：作者)

(六) 獨立檢驗

資料的數位型態讓資料便於公布流通，人文研究不再受限於研究素材的寡佔問題。同一份資料、同一研究議題、同一分析方法，不同研究團隊可以對各自的實作分析結果，獨立觀察解讀，而相互比較或前後驗證；也可以用不同的方法，比較結果的差異，測試方法的優劣；或是探討不同的議題，嘗試不同的發現；或是累積相關資料的分析成果，整合為更完整的研究論述。資料就如同實驗室中的材料，可以反覆操作萃取其中的重要資訊，實驗過程與結果都可以被複製檢驗，而提升研究成果的客觀性與正確性。人文研究將因獨立檢驗能力的提升，而帶來社會影響的擴散，並促進研究社群的共同成長進步。例如，亙古彌新的《紅樓夢》研究，在數位文本公開流通的有利條件下，可以展開許多新議題、新方法、新事證的計算挖掘研究，從主角性格、作

者辨別、角色關係網絡，到文學評析、歷史情境、語言行為等更廣泛的研究議題面向，都可以帶來豐碩的新發現。

(七) 跨域框架基礎

計算邏輯的操作有很大一部分是以資料、資訊與知識為主要概念元素，並以建立概念元素的系統性連結為目的，打造從資料到資訊、知識的萃取提煉過程，而達成去蕪存菁、甚至轉化生成的效能。計算邏輯的設計與目的的設定，需要資料領域專業的參與引導，並提供專業知識概念脈絡，以利計算邏輯實質作用的實現。因此，資料計算即是一個跨域框架的本體，數位人文則是在此基礎框架上，對人文資料與人文議題的不同專業領域，如文學、歷史、語言、政治、經濟、傳播等，嘗試展開垂直與水平的各種跨域組合與交融。例如，鄭文惠等（2015）提出以構詞概念為基礎之詞彙擷取邏輯，結合文學分析的思維角度，考察中唐詩歌中的白色字詞的使用，快速建構出白色構詞的搭配系統，從而推展出豐富的隱喻與文化象徵，並引申至情感現象與色彩政治的論述。Liu, et al.（2016）則運用語言學的情態表示概念，於重大歷史事件新聞文本中，分析情態詞使用的分布與變化，再依據不同情態類別的意涵，觀察推論官方對重大社會事件的言論立場與態度。

三、結語

由於學術領域的分化獨立發展與專精訓練養成，人文領域與資訊領域在學術目標、研究方法、專業語言、甚至價值體系等方面，都存在不可忽視的差異。數位人文代表著一個新研究典範開創的機會，但必須以相關學術社群的共同參與投入與相互認知理解為基礎。資料計算突顯了人文研究客觀證據的份量，在議題想定、方法設計、結果解讀、發現論證等層面，打開了更豐富的樣貌，同時，也刺激學術社群的理念激盪與學術對話。資料計算更是大數據分析的核心基礎，也是人文社會研究的重要新思維與新面向（劉吉軒，2016）。本文提出資料計算型塑數位人文研究典範的本質特徵，協助年輕世代學者建立跨領域合作的路徑參照，也與相關學術社群共同歸納省思。數位人文跨領域研究可以從互動合作的物理變化到交融協作的化學變化，期待有志之士齊心協力、共同開創。

致謝

本文研究成果來自於科技部數位人文研究計畫「開放性數位工具平台發展-以台灣自由人權系列文本為例之社群參與及數位人文研究」(MOST 104-2420-H-004-032-MY3)之經費支持。同時，也感謝政大研究團隊台史所薛化元、資科系蔡銘峰、心理系顏乃欣及英文系賴惠玲等教授之學術互動與知識激盪。

參考文獻

杜協昌。2014。〈利用文本採礦探討《紅樓夢》的後四十回作者爭議〉。項潔編。《數位人文研究與技藝》。臺北：臺大出版中心。

鄭文惠。2014。〈從人文到數位人文：知識微縮革命與人文研究範式的轉向〉，人文與社會科學簡訊，15 卷 4 期，頁次：169-175。

鄭文惠、劉昭麟、邱偉雲、許筑婷。2015。〈情感現象學與色彩政治學：中唐詩歌白色抒情系譜的數位人文研究〉,「東亞聚焦：第六屆數位典藏數位人文國際研討會論文集」，頁次：481-522。台灣大學數位人文研究中心。

劉吉軒。2016。〈大數據分析與人文社會科學跨領域研究應用〉，傳播文化，15 期。

Berry, D. M. (2011). The computational turn: Thinking about the digital humanities. *Culture Machine*, Vol. 12.

CBETA–中華電子佛典協會. Retrieved June, 28, 2016, from http://www.cbeta.org/

Hayles, N. K. (2012). How we think: Transforming power and digital technologies. In *Understanding digital humanities* (pp. 42-66). Palgrave Macmillan UK.

Liu, J. S., Lee, C. Y., & Ning, K. C. (2016). Evaluating Modal Use in News Corpus for Constructing Rhetorical Context of Historical Event, *Digital Humanities 2016 Conference Abstracts*, pp. 262-266, Krakow, Poland.

Moretti, F. (2013). *Distant reading*. Verso Books.

Schnapp, J., Presner, T., & Lunenfeld, P. (2009). The digital humanities manifesto 2.0. Retrieved July, 2, 2016.

服務於中國歷史研究的網絡基礎設施

王宏甦[*]、徐力恆[**]、包弼德[***]

摘　要

數據庫、研究項目數量和參與中國文史數位研究的人員大幅增加，使得為中國歷史研究建立相應的網路基礎設施變得必要。網絡基礎設施可以起的作用在於連接對一個學科有用的電腦軟件、數據集、人才、實務做法、標準和合作模式，促進研究的進步。本文將具體論述為何要營建中國歷史研究的網絡基礎設施，以及如何從資源的共享和成員的交流兩方面實現這個目標。

關鍵字：網絡基礎設施、中國歷史、數位人文

[*] 美國哈佛大學「中國歷代人物傳記資料庫」項目經理，Email: hongsuwang@fas.harvard.edu。
[**] 美國哈佛大學「中國歷代人物傳記資料庫」博士後研究員，Email: tsui01@fas.harvard.edu。
[***] 美國哈佛大學副教務長、東亞語言與文明系查理斯·卡威爾（Charles H. Carswell）講座教授，Email: peter_bol@harvard.edu。

A Cyberinfrastructure for Historical China Studies

Hong-su Wang[*], Lik-hang Tsui[**], Peter K. Bol[***]

Abstract

The proliferation of databases for the study of Chinese history and the increasing numbers of researchers taking part in their development calls for a cyberinfrastructure. A cyberinfrastructure can be conceived as a network of discipline-specific software applications and data collections and also of the personnel and the set of best practices, standards, and collaborative methods they establish. This paper discusses how participants in such a cyberinfrastructure for historical China studies can share their resources and how their communication can be facilitated by various technologies and mechanisms.

Keywords: cyberinfrastructure, chinese history, digital humanities

[*] Project Manager, CBDB, Harvard University. Email: hongsuwang@fas.harvard.edu.
[**] Postdoctoral Fellow, CBDB, Harvard University. Email: tsui01@fas.harvard.edu.
[***] Vice Provost for Advances in Learning and the Charles H. Carswell Professor of East Asian Languages and Civilizations, Harvard University. Email: peter_bol@harvard.edu.

一、引言：中文數位人文網絡基礎設施的實現

美國學術團體協會（ACLS）在 2005 年發佈的研究報告《我們的文化共同體》（Our Cultural Commonwealth）提出人文、社會科學應該像自然科學研究一般，有自己的網絡基礎設施。[1] 網絡基礎設施（cyberinfrastructure）的層次介於基礎科技和具體用於某研究項目、某學科和實踐的特定科技之間，可以說處於中層。它可以起的獨特作用，在於連接對一個學科有用的電腦軟件、數據集、人才、實務做法、標準和合作模式。[2] 我們希望在本文處理的議題如下——我們為何要在中國歷史研究領域營建這種網絡基礎設施，以及如何實現這個目標。

和自然科學相比，人文學科和部分社會科學學科（尤其是其中量化特點不明顯的學術領域）深深浸淫在語言之中，也深受語言的特點影響。就以主題模型（topic modeling）為例，當這方法用於中國文史研究時會面對頗多挑戰。一般使用這種研究方法時，認定每個詞之間的空格就代表分詞的區隔，但中文文本的情況並非如此。例如，當機器閱讀「中華民國」這四個中文字時，它可以認定那是四個詞，也可以是兩個各為兩個字的詞，也可以是一個四字詞。怎麼讓機器獲得判斷的能力，需要人文學者利用他們的知識介入。

因此，這意味著我們必須深入調查、瞭解語言的特點（例如是古代漢語、現代漢語），才能建立專門用於中國歷史研究的數位人文工具。如果不正視這一點，數位研究計劃之間的溝通和連接將很難進行，數據的分享也會面對很多障礙，導致閉門造車的弊病。不過，近年學界對我們會議的主題——數位人文的興趣越來越濃，用於統計、社會網絡分析、地理空間分析和地圖繪製、文本標註和挖掘、製作主題模型，還有建立關係型數據庫（relational databases）或物件導向式數據庫（object-oriented databases）的電腦軟件如雨後春筍。過去很難用，或者很難獲得的軟件，現在已經是隨手可得，其操作也簡便許多了。項目數量和參與數位研究的人員大幅增加，也為中國研究數位人文建立相應的網路基礎設施開始有了成熟的條件。就如其他學科和區域研究的專家一樣，我們領域裡不少學者都感到應該開展這種工作。因此有了本文的寫作。

網絡基礎設施的建設涉及很多問題，本文只打算討論資源共享和成員交流兩個主要的方面。首先，是資源分享。資源分享對數位人文研究有巨大的幫助——它使資源更容易被整合，人文學者可以透過被大量連接著的邊界對象（boundary objects）發現

[1] 參閱：http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf。
[2] 參閱：https://www.nsf.gov/cise/sci/reports/atkins.pdf。

和自己研究有關的其他領域的知識。[3] 數據庫項目亦可藉此避免重複勞動和浪費資源，並透過網絡基礎設施的網絡交流，以自己的品牌推進數位人文領域的發展。

重要的是，資源不只是數據。這裡的所謂資源還包括工具、研究方法、技術等。這些內容的分享，並不只是把共享資料分發出去給不同成員。如果分發出去的資料沒有必要的說明，或如果分享不是通過規範的方式，這些資料就很難被潛在的使用者利用。[4] 除分享的方式外，版權和非公開資源也是需要討論的問題。在網絡基礎設施的合作中，非常歡迎商業公司參與。對於商業公司來說，他們會樂於分享那些對其商業運作有幫助的資料，譬如他們的數據庫的收錄書目和試用資源。因此需要將分享方式和分享權限一併考慮，來設計中文數位人文網絡基礎設施的資源分享如何實現，其細節又應該如何設計。關於有哪些核心的數位人文工具需要進行開發，促進網絡基礎設施的成熟，也是本文會探討的問題。

第二，成員交流。通常數位人文項目，無論是學術機構還是商業機構的項目，都會有自己的合作者。有些合作是與數據、軟件公司一起處理技術上的問題，有些是尋找有專門知識的學者來解決學術問題，不一而足。這種合作方式可以稱為面向具體計劃的合作。學者 Marina 將其歸為小組（small teams）、本地組織（local organizations）、分散組織（distributed organizations）幾類。除此之外，還有一種是組織鬆散的社區（loosely organized communities）[5]——這種社區，正是建立中文數位人文網絡基礎設施中應該規劃建立的。一方面，這種鬆散組織的社區可以為中文數位人文項目各自獨立的小圈子引入更多潛在的合作者。另一方面，透過組織討論和會議，合作者可以接觸自己圈子和領域外的群體，接觸新理念、新材料、新方法。而這些成員應該包含各種參與者，不應只限於單一種類。大學機構、研究團隊、數據庫計劃、圖書館、出版社、商業數據庫公司、基金會等等，都應該涵括在內，才適應當前中國文史數位研究的確切狀況。

數字人文的研究計劃往往是集體協作，也經常採取國際合作的形式。一些服務中國歷史研究的學術合作已經存在十年以上，例如筆者幾位共同參與運營的「中國歷代人物傳記資料庫」（CBDB）就是一個已經運作超過十年的國際合作項目，由哈佛大學、中央研究院和北京大學共同開發。這數據庫目標在於系統地收錄中國歷史上所有

[3] Jirotka, Marina, Charlotte Lee, and P. Olson, “Supporting Scientific Collaboration: Methods, Tools and Concepts,” *Computer Supported Cooperative Work* 22, no. 4 (2013): 687-688.
[4] Barkhuus, Louise, and Brown, Barry. "The Sociality of Fieldwork: Designing for Social Science Research Practice and Collaboration." Proceedings Of The 17th Acm International Conference On Supporting Group Work, 2012, 41.
[5] Jirotka, Marina, Charlotte Lee, and P. Olson, “Supporting Scientific Collaboration: Methods, Tools and Concepts,” *Computer Supported Cooperative Work* 22, no. 4 (2013): 702.

重要的傳記資料，並將資料開放供學術研究之用。本文的分享，代表著這種合作摸索出來的經驗和展望。

二、資源的共享方式

我們認為，中文數位人文網絡基礎設施的資源共享方式主要可以採取兩種方式：一種是 API 分享，另一種是文件分享。

(一) API 分享

首先討論 API 分享。API （application programming interfaces） 的中文翻譯是「應用程序接口」[6]，它是一種非常快捷、有效的數據分享方式，是允許數據庫之間互相溝通的接口。在理想狀態下，透過 API，每個數據庫都可以取用其他數據庫的信息，來補充自身比較欠缺的資料, 而不必在自己的數據庫中重新輸入這些數據。

在資源分享中，用戶或者其他項目可以直接引用開發者的 API，甚至數據所有者無需專門投入任何精力，便可以達成資源分享。例如，CBDB 就有為其他數據庫提供 API 服務，讓任何數據庫都可以取用 CBDB 人物傳記資料，並以自己的數據庫方式呈現出來。目前，CBDB API 支援兩種查詢方式：用人物的 ID 查詢 ，或用人名查詢（漢字或拼音）。[7]

因為有這樣的功能，所以其他數據庫或數位平台就可以加以利用。譬如，MARKUS（瑪庫斯）平台利用 CTEXT（「中國哲學書電子化計劃」）的 API 來導入古籍文本，又利用 CBDB 項目的 API 來查詢全文文本中的人名，以便進行標註。[8] 這些文本經過標註以後，可以方便文史學者解讀文獻，甚至是引入其他方式對文本信息進行數位分析。這大概是目前領域內利用 API 達致資源分享，為學者的研究實現更多可能性的最佳例子。

[6] 由於本文受眾為從事中國文史研究的各種機構和人員，海峽兩岸的情形都有涉及，所以本文用語不全用臺灣術語和譯名，一般採用較普遍用法。

[7] 更詳細的情況請參閱：http://projects.iq.harvard.edu/chinesecbdb/cbdb-api。

[8] MARKUS 是萊頓大學的何浩洋（Brent Hou Ieong Ho）和魏希德（Hilde De Weerdt）開發的中文文本標記系統，網址是：http://dh.chinese-empires.eu/beta/ 。CTEXT 是一個開放的線上電子圖書館，為學者提供中國古籍文獻文本，網址是 http://ctext.org/。下文對這兩者還會有更多介紹。

1762

CBDB ID: 1762
索引/中文/英文名稱: /王安石/Wang Anshi
指數年 (index year): 1080
生年: 北宋元祐5年 (1021)
卒年: 北宋元祐1年 (1086)
享年: 66
郡望: 太原
出處: 宋人傳記資料索引(電子版), 頁1536
註: Wang(2) Anshi [1762] Yi(3)'s [7082] son, Guan(1)'s [1841] grandnephew, Anguo's [7076], Anli's [1760], and Anshang's [1761] brother, Fang's [1803] uncle, Jue(1)'s [1796] great grandfather, Zhu(1) Mingzhi's [526] and Shen(2) Jichang's [1445] brother-in-law, Cai(1) Bian's [8131] father-in-law and uncle-in-law, and Wu(3) Anchi's [1957] father-in-law. Anshi was Zhang(1) Jifu's [195], Xu(3) Xi's [753] and Yang(2) Wei's [2032] patron, Li(2) Ding(1)'s [1097] patron and teacher, Gong(2) Yuan' s [941] teacher, and the teacher of Wang(2) Pin's [7381] uncle, Boqi [3970]. He was the grandnephew of Wang(2) Guanzhi [3965], the father-in-law of Zhou(1) Jiazheng's [480] son, Yanxian [3250]. His mother was Wu(3) Shi and his wife was Wu(3) Shi. He once arranged for a marriage between his wife's younger sister and Wang(2) Ling [3967] whose scholarship he admired. The sister's, and therefore Wang Anshi's wife's, grandfather, Wu(3) Min [4017], had the same generational name and address as Anshi's maternal grandfather, Wu(3) Tian [7396], so that it can be assumed that they were brothers or cousins. When Wang Anshi composed the funerary inscription for the two women's father, Wu(3) Ben [4020], in 1054, neither daughter was married. Wang(2) Ling died in 1059 at the age of 28. Wang Anshi was 32 in 1054. Wang(2) Ling was Wu(3) Yue's [1992] maternal grandfather. The mother of Yan(5) Shu's [2073] grandnephew, Fang [4118], was the younger sister of Wang(2) Anshi's wife, Wu(2) Shi. Songshi yi, 5.5b discusses the coalitional significance of the fact that Lu(9) Jiawen's [1294] son married the daughter of Wang Anshi's son, Pang(a) [3968], and that Cai(1) Bian [8131] married Pang's elder sister. Anshi was a friend of Wang(2) Ping's [1856] son, Hui [3958] and his older and younger brothers were tongnianyou with the sons of Chen(1) Jiansu [7101]. XCB, 177.2a, 179.3a, 184.15a, 188.8a, 189.3a, 18a, 191.9a-10a, 192.4a-4b, 7a; XCBSB, 4.4b, 15.17a, 16.14b; Shen Gou, WJ, 2.48b; SHY:ZG, 5.1a; Zeng Gong, WJ, 45.4b; Du Dagui, 'xia,' 14.1a; Wang Anshi, WJ, 97.998, 98.1012; Wang Ling, WJ, 'fu,' 13b. CBD, 1, 277-281.From Hartwell's ACTIVITY table:1075: Working on the railroad all the time all the time all the time

別名: 字介甫，室名、別號半山老人、謚號文，小字獾郎。

地理資訊: **籍貫(基本地址)**: 宋朝 / 江南西路 / 撫州 / 臨川
出處: , 頁
註:

任官:
- **參知政事**: 地點: **宋朝**。起始年: 1069。 赴任。
  出處:宋人傳記資料索引(電子版)(頁1536)。
  註:Hartwell defined the office as Canzhi zhengshi (參知政事)
- **三司度支判官**: 地點: **宋朝**。起始年: 1058。 赴任。
  出處:宋人傳記資料索引(電子版)(頁1536)。
  註:Hartwell defined the office as Duzhi panguan (度支判官)
- 正授 **群牧判官**: 地點: **宋朝**。起始年: 1054。 終止年: 1055。 赴任。
  出處:未知。
  註:Hartwell defined the office as Qunmu panguan (群牧判官)
- **同中書門下平章事**: 地點: **宋朝**。起始年: 1071。 赴任。
  出處:宋人傳記資料索引(電子版)(頁1536)。
  註:Hartwell defined the office as Tong Zhongshu Menxia Pingzhang shi (同中書門下平章事)

圖 1：CBDB API 的輸出格式。這裡僅截取了一個人物的其中一部分資料。

（圖片出處：http://projects.iq.harvard.edu/cbdb/cbdb-api）

第二，使用大數據的研究方法，常常需要批量抓取特定資料，來進行統計運算或可視化分析。這種情況下，具有一定格式並欄位名清晰的 API 輸出有利大大簡化數據的清理工作。[9] 另外，學術用途的數據抓取一向是商業數據庫公司不太瞭解，或諱莫如深的行為。但在處理大數據的研究方法下，數據抓取絕不是數據盜取，而是因為研究需要而必須進行的研究步驟。英國已經通過新的規定，指明只要這種做法不是商業行為，就會受到版權法例保障，不會違反法規。[10] 對於商業公司來說，與其抵制這種學術上的做法，不如透過 API 的設置來開放，並規範這種數據抓取的行為。第三，API 可以鼓勵大家對數據所有者的數據進行二次開發，反而增加用戶對數據庫的使用需求，突顯原數據擁有者的貢獻。例如，德國馬普科學史研究所（Max Planck Institute for the

[9] 熟悉數據研究工作的同仁都會瞭解，清理數據的工作往往佔據研究者的大量時間、精力。
[10] 詳見 Ross Mounce, “The right to read is the right to mine: Text and data mining copyright exceptions introduced in the UK,” LSE Impact of Social Sciences Blog, June 4, 2014, http://blogs.lse.ac.uk/impactofsocialsciences/2014/06/04/the-right-to-read-is-the-right-to-mine-tdm/.

History of Science）對現成方志數據的二次開發引起越來越多的學者對方志的興趣，肯定會刺激更多人購買相關的電子資料。[11]

中文數位人文網絡基礎設施的 API 在資料流動方面可以分成三種類型：

一種是**輸出分享數據**。大多數數據庫 API 屬於這一類型。在合作中，如果合作機構已經有比較完善的 API，徵得允許後，就可將其直接收入中文數位人文網絡基礎設施 API 網站（詳見下文），還有 GitHub 之類的平台。如果合作機構有完善的 API，但並不完全向外公開，征得同意後，收入時標註其使用必須獲得授權。如果合作機構願意分享其資料，但是暫時沒有開發 API 系統的資源，我們可以建議對方使用 API 自動生成工具。

API 自動生成工具是由 CBDB 項目提出的一個快速生成簡單 API 的工具。此工具要求用戶將要分享的數據製作成一份，或若干份第一行為欄位名的 CSV 表格，放到指定的資料夾下。之後，由程式自動讀取該資料夾下所有文檔，將資料的檔名作為表名存入 MySQL 數據庫，並生成可以輸出 JSON 文檔的 API 系統。此 API 系統將 CSV 表格前三個欄位的欄位名當做查詢變數，第三個欄位可以設定為允許進行「大於」和「小於」的運算。該工具有兩個版本，其中一個是本地版。如果合作機構的服務器是 PHP 和 MySQL 結合，並且希望自己來維護自己的 API，我們可以協助他們搭建本地 API。如果對方沒有資源提供 API 服務，可以將本地服務器上指定的資料夾告訴我們，網絡基礎設施的程式會定期抓取資料，在基礎設施的主機上代為提供 API 服務。

第二種是**在線工具的功能分享**。它的資料流向是輸入。譬如 MARKUS 提供 API 可供使用者導入自己的文本進行標記，CartoDB 和 PLATIN 等工具提供 API 允許用戶提交特定格式的數據，為數據實現可視化，部分滿足研究者的學術需求。[12] 另外，我們也呼籲相同或相關領域的在線工具使用相同的參數名稱來接收輸入的資料，以利連接。

第三種是為數據庫之間提供**數據之間的關聯**（database ID cross link）。它的功能是為不同數據庫的同一實體提供鏈接，使得系統之間的相互操作（system interoperability）成為可能。[13] 不同中文數據庫中的人名、地名、官名、書名等常有重

[11] 參閱：https://www.mpiwg-berlin.mpg.de/en/research/projects/departmentSchaefer_SPC_MS_LocalGazetteers

[12] CartoDB 是一個建雲端計算服務平台，提供地理空間資訊的展示，使用者可以上傳空間數據，或連接網上資料，即可發佈地圖服務，操作比一般 GIS 軟件簡單。PLATIN（Place and Time Navigator）是一個類似的工具，對時間數據的支持比較強。PLATIN 源代碼可以參閱：https://github.com/skruse/PLATIN。

[13] 可以參考 DH2012 會議以下小組的論文：「用於中國文史的群體傳記學、文本挖掘、地理信息系統和系統間的相互操作」（Prosopographical Databases, Text-Mining, GIS and System Interoperability for Chinese History and Literature），參閱：http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/prosopographical-databases-text-mining-gis-and-system-interoperability-for-chinese-history-and-literature.1.html

合。在數據搜集過程中，搜集者常常會利用各種資料，對不同數據庫的資料來進行消歧工作（disambiguation）。[14] 在此過程中，很容易發現不同數據庫中的相同實例。因此，我們希望建立一個可以進行授權讀取和寫入的數據庫，中文數位人文網絡基礎設施的成員可以使用它來建立和查詢不同數據庫之間 ID 的關聯。[15]

接下來，我們希望所有提供 API 的機構為 API 提供一個 META 元素。這個 META 元素用於系統地描述當前資料的基本資訊，即元數據（metadata）。在理想情況下，它的子元素（child elements）如下：topic, period, type, URL, description, version, date, call_freq_limit。以下對它們逐一解釋：

- topic 元素。此元素描述該 API 的內容主題。其內容（values）可以包括人名（person）、地名（address）、機構（institution）、職業（career）、教育（education）、物質（material）、書目（bibliography）等。
- period 元素。此元素描述該 API 內容的時代。其子元素為 century_begin（西元世紀始），century_end（西元世紀終），dynasty（朝代）。西元世紀值為數值，如 1 表示公元一世紀，10 表示公元十世紀，-1 表示公元前一世紀。朝代用詞規則可參考 CBDB 的朝代表格代碼表（關於代碼表，請見下文）。允許同時記錄朝代和公元世紀將便於區分同名朝代，以及同時期但不同朝（如宋、遼、金、元）的狀況。同時也允許用戶只填寫朝代或年號，或對沒有時間資訊或時間資訊不明確的資料留空。
- type 元素。此元素描述該 API 的文件類型。其內容（value）包括 JSON 代碼、文本（text）、圖片（image）、視頻（movie）、聲音（sound）、文檔（file）等等。
- URL 元素。當 type 元素為圖片、視頻、聲音、文件時，我們會建議分享者使用 URL 元素的值來定義這些資源的路徑，而不是直接將其編碼存放在 JSON 結構裡面。
- description 元素。此元素允許開發者使用一段文字來描述該 API。
- version 元素。此元素值（value）用以描述該 API 的版本。如果未設置，初始值為 0.01，每次更新 API 資料的時候自動加 0.01，方便記錄、追蹤。

[14] Chao-Lin Liu, Guan-Tao Jin, Hongsu Wang, et al., "Textual Analysis for Studying Chinese Historical Documents and Literary Novels," *Proceedings of the ASE Big Data & Social Informatics* (2015).
[15] 我們意識到臺灣的大型數位典藏計劃在連接多個數據庫的工作上積累了不少經驗，是我們希望可以借鑒的。相關平台網址為：http://digitalarchives.tw/。

- date 元素。此元素描述該 API 更新的時間。格式為 YYYYMMDD。
- call_freq_limit 元素。此元素的值描述該 API 建議請求資料（request data）的時間頻率。其描述方法為反斜線（slash）右邊表示毫秒數（millisecond），左邊表示在右邊的毫秒數下建議抓取資料的次數上限。譬如 1/100 表示最多允許每一百毫秒抓取一次。1/0 或者留空，則表示無限制。數據所有者應當注意的是，如果 API 被其他平台（platform）使用，或者二次開發，抓取頻率將會由第三方使用者決定，或者多個 IP 的分散需求將會被合併到一個 IP（二次開發的伺服器）上，建議抓取頻率在這種狀況下通常沒有辦法被執行。

最後，中文數位人文網絡基礎設施 API 將會分類公佈在網絡基礎設施網站上。同時分享到 GitHub。對於無法訪問 GitHub 的中國大陸用戶，API 信息同時也應發佈在「碼雲」（GIT@OSC）和 CSDN 的 CODE 上。[16] 如此，才能保證不同成員都能使用相關程序。

(二)文檔分享

除了 API 分享外，另一種簡易快捷的分享方式是文檔分享。分享的內容最常見的是數據庫的離線版本，以及在線版本的 Datadump 分享。另外還有程式碼、PDF 文檔、圖片、音頻、視頻、從數據庫中抽取出來的部分資料、製作數據庫的原始數據或中間數據產物、實用的詞彙集（dictionaries）等。

文檔分享有三個問題需要討論：文檔的持久性（sustainability）、文檔的版本管理（version control）、文檔的描述（description）等。

首先，文檔的持久性。數位項目常常遇到的風險是，當計劃結束後，項目資料和工具是否可以長期維持有效。因此我們建議將需要分享的文檔除了放在項目或者學校的網站上，還要分享到 Dataverse 和 ICPSR 等平台上。[17] 這些有固定機構支持的平台可以保障資料長期被妥善保存。除北美之外，Dataverse 在歐洲、亞洲的大學和研究所中有許多分支版本和類似項目，例如是北京大學的的開放研究數據平台、香港科技大學的 DataSpace 等。[18] 因此，通過和 Dataverse 生態圈的各系統進行合作，可以保障分享的資料在世界各地的下載速度。

---

[16] 「碼雲」網址為：https://git.oschina.net/。CSDN 的 CODE 網址為：http://code.csdn.net/。

[17] Dataverse 計劃是一個開源的研究數據存儲軟件，由哈佛大學的量化社會科學研究所（Institute for Quantitative Social Science）開發，全世界目前有 20 個收藏庫。平台網址見：http://dataverse.org/。ICPSR 是美國校際社會科學數據共用聯盟（Inter-University Consortium for Political and Social Research），是目前全球最大的社會科學資料中心。網址見：http://www.icpsr.umich.edu/icpsrweb/index.jsp/ 。

[18] 前者網址是：http://opendata.pku.edu.cn/；後者網址是：https://dataspace.ust.hk/ds/service/about。

第二，文檔的版本管理。電子數據和紙質文獻的重要不同之一是：紙質文獻出版之後，即使是再版，舊版並不會因版本迭代而消失。但電子數據發佈之後，常常會被迭代修改，舊的版本再也不能被找到。這種修改對於資料分享和引用會產生一定風險。譬如，一個數據庫用自己的人物 ID 連接到另一個數據庫的人物 ID 上，但是在資料更新的時候，對方數據庫的這個人物被刪除了，那麼連接就會意外中斷。如果沒有版本聲明，那麼這在邏輯上就成為一個錯誤的資訊 ——一個人物被連接到另一個不存在的 ID 上。而如果有版本聲明，在邏輯上便是正確的——當前數據庫的一個人物被連接到另一個數據庫某個版本的某個 ID 上。CBDB 項目把歷年離線數據庫保存在項目網站上，供使用者下載，所以一旦發生上述問題，用戶就可以下載特定版本的數據庫來確認需要的資料。另外，在 Dataverse 和 ICPSR 上各有版本管理機制。

第三，文檔的描述。我們希望每份資料的分享都擁有必要的說明內容。Dataverse 和 ICPSR 設計有描述性欄位供上傳者描述自己的資料，這對資源的分享和試用特別有幫助。也正因為這一點，我們不建議中文數位人文網絡基礎設施成員在分享資料時只使用 Dropbox、Microsoft OneDrive、Google Drive、百度雲等雲端儲存工具。原因在於，這些工具並不鼓勵用戶對分享的資源進行描述。資源和描述的分離對於資料的持久使用和正確使用有潛在危害。

鑒於在 Dataverse 和 ICPSR 分享數據會稍微增加數據庫運行的人力成本，對於不希望自己投放資源維護 Dataverse 和 ICPSR 更新的機構，我們提出基於 Dataverse 和 ICPSR API 自動上傳和分享的工具。和上文提到的 API 自動生成工具類似，此工具也會有本地維護和遠端維護兩種版本。本地維護版本的運行環境最好是 Python 3.x，此版本可以將特定資料夾中的資料和同名 JSON 檔傳給 Dataverse 和 ICPSR，JSON 檔的內容為 Dataverse 和 ICPSR 需要的描述性資料。我們可以提供一個網頁界面，讀取和生成 Dataverse 和 ICPSR 描述性資訊，讓使用者透過這個界面生成用來描述資料的 JSON 檔（並且允許用戶上傳修改已經生成的 JSON）。

對於遠端版本，如果使用者希望資料保存在自己的服務器上，使用者可以提交給我們用來抓取資料的 URL 跟檔名，要分享的資料需要與同名 JSON 一併放在此 URL 對應的服務器本地目錄下。程式會定期檢查這些目錄下的資料是否有更新，如果有更新，便自動更新 Dataverse 和 ICPSR 的資料。如果使用者並沒有自己的服務器，或不希望由自己的服務器保存這些文檔和 JSON，可以將文檔和 JSON 一併傳給中文數位人文網絡基礎設施的維護小組，由他們代為上傳。

三、分享權限

數據的獲取和開放程度是中國數字人文面臨的另一大挑戰。對於數據所有者來說，數據分享可以讓更多人使用自己的數據，讓自己的數據在更多的研究中發揮作用，減少重複性工作。如果項目之間有足夠的協調，就不用再浪費資源在重複的工作之上，例如多個機構為版本完全相同的書進行電子化。但另一方面，對數據被惡意抓取，以及對服務器可能帶來巨大訪問量壓力的擔心，使得數據所有者，特別是商業數據庫所有者對數據分享持相對保守態度。單以中國古代典籍為例，電子化材料的分享權限程度仍不夠透明。不過需要注意的是，數位人文網絡基礎設施的構建不是為了把所有數據都全部合併到一起。其任務更多在於推動數據分享方式的規範化和開放風氣之推廣。以下一部分要討論的是正是分享權限問題。

(一)分享授權的形式

分享權限就授權形式劃分，可分三種：完全公開分享、部分公開分享、授權分享。

首先討論**完全公開分享**。完全公開分享是非盈利公開數據庫項目常用的分享形式，其授權方式絕大多數可以納入 Creative Commons 架構下。Creative Commons 4.0 共提供六種權限，以授權嚴格由寬鬆到嚴格排序如下：

- 署名（Attribution）
- 署名（Attribution）、相同方式共享（Share Alike）
- 署名（Attribution）、禁止演繹（No Derivative Works）
- 署名（Attribution）、非商業性使用（Noncommercial）
- 署名（Attribution）、非商業性使用（Noncommercial）、相同方式共享（Share Alike）
- 署名（Attribution）、非商業性使用（Noncommercial）、禁止演繹（No Derivative Works）

以 CBDB 項目為例，我們的分享方式是署名、非商業性使用、相同方式共享。它表達的意思如下：公開使用時，必須要提及我們項目的名字。在非商業使用的狀況下，可以自由的使用。如需商業使用，則必須聯絡我們，得到批准。如在非商業狀況下使用我們資料，資料分享的協議跟我們必須保持一致。

接下來討論**部分公開分享**。這裡討論的數據庫是部分或者全部資料不向公眾公開，因此不包括商業資料。這種數據庫通常只提供給局域網內，或認證用戶使用。不公開的原因通常是因為預設用戶群體非常小、系統無法支撐全面公開帶來的訪問量、數據庫尚未開發成熟、版權問題的限制、資料涉密等，甚至一些數據庫是因為研究資金已經用完，所以沒有人員或機構持續維護，無法繼續保持上線。對於非公開的數據持有者來說，其資料可以分為兩種：一種是可以公開的資料，另一種是不公開的資料。我

們應當鼓勵其找出並開放他們資料中可以向公眾開放的部分，並為其提供一定的技術支援。對於不便分享的資料，也必須充分尊重其意願。

最後討論的是**授權分享**。授權分享的例子中最常見的是商業數據庫。授權的方式較常見的有 IP 限制和用戶登錄限制。此外，基於數位人文的研究方法越來越偏向透過大數據得出結論，而通常商業數據庫為了限制資料被盜取（基於傳統的數據庫查詢需求），會設計一些防止抓取的機制。有鑒於此，我們會建議資料所有者面向大數據分析的需求，設計出授權的方法，譬如數據挖掘授權（data mining authorization）和 API 令牌（token）。如果這樣做，就意味著授權分享對商業數據庫來說不只是銷售部門的工作，跟技術部門同樣息息相關。另外，我們也會建議商業數據庫所有者可以向聯合機構出售使用權限。譬如地方志在每個學校裡的需求量可能不大，尤其是中國研究者數量有限的歐美大學。對學校圖書館來說，往往不值得為很小的使用需求，花一大筆經費購買全面的地方志數據庫。但是如果能夠面向多個機構的聯盟出售使用權限，譬如若干有相同研究興趣的學者建議各自的學校聯合購買全面的地方志數據庫，為該聯盟獲得訂閱權限（alliance license），那麼每個用戶或學校需要分攤的價格就會低得多。[19] 這樣做一方面可以讓研究者和用戶受益，另一方面又能讓數據庫開發商賣出更大量的數據庫授權，達到共贏。

(二)分享授權的數據持有者

數據分享的數據持有者可以被劃分成四類：圖書館類、公開非盈利數據庫類、商業數據庫類、非公開數據庫類。這一部分將討論這四類分享者可能分享的內容，以及我們構想的理想分享方式。

首先是**圖書館類**。圖書館擁有完備的書名和元數據資訊，並且圖書館的檢索通常是面向公眾開放的，因此這已經是圖書館類機構必然對外共享的資源。我們會建議圖書館基於這些資料建立 API 系統，允許用戶以元數據的欄位，對數據庫進行多條件的查詢，並且以 JSON 或 XML 的格式來返回元數據的內容。在建立 API 系統的時候，也建議允許用戶用索書號的一部分作為分類檢索條件。

另外值得注意的是，當下有不少圖書館已經開發了自己的 API 系統，但其操作往往相對複雜。我們會建議圖書館在搭建 API 平台的時候，易用性是需要特別考慮的，與其把複雜的檢索狀況納入檢索環節，不如把複雜的條件查詢交給用戶在抓取資料之後自己去做，把更多的精力投入在提高查全率（recall rate）和相關度排序上。有時用戶把資料抓取之後放入 Excel 軟件，利用其中一些基本功能就能完成他們需要做的工作。

[19] 德國一些機構對東亞研究數據庫的資源管理有類似做法，可以參閱：http://crossasia.org/en/service/xasia/projekte/virtueller-campus.html。

而如果要設計數據庫滿足這些需求，往往需要花費數十倍的精力和代碼。懂得自己寫程式抓取資料的用戶，通常都具備進行清理資料的基本技術。當然，我們知道具備這種技能的人文學者仍屬非常少數。

除了元數據，現在許多公共圖書館和大學圖書館都開始了圖書數字化計劃。大多數圖書電子化資源會允許用戶公開閱讀或下載。這也是圖書館類成員在分享數據方面得天獨厚的資源。因為圖書館類機構的運行相對穩定，所以我們不主動向圖書館項目建議將自己的數據分享至 Dataverse 或者 ICPSR 上。當然，如果圖書館項目願意做這樣的備份分享，是應該鼓勵的。此外，我們會建議圖書館的 API 可以限定只檢索公開的電子文檔資料，或者根據公開的數字化書目來建立獨立 API 元數據（metadata）查詢系統。對於非本校用戶來說，公開的電子文檔是他們更頻繁查詢的內容。

由此，值得順帶一提的是，我們可以預見大學圖書館以後在中國文史的數位人文研究中會扮演非常重要的角色。[20] 它們既有管理甚至是創建數字資源的經驗，又需要面對讀者、用戶和研究者，具備擔當橋樑角色的條件。它們和科研人員的合作會是推動數字學術（digital scholarship）發展的關鍵，也自然是網絡基礎設施必然要涉及的建設者。也是因為這一層原因，不少中外圖書館方面越來越注意數位人文的發展如何推進對其中文館藏的管理和研究。

第二是**公開的非盈利數據庫**。公開的非盈利數據庫通常願意自己的資料被更多人使用。他們可以分享的內容非常豐富：

i. 離線版數據庫和在線版數據庫的數據轉儲（data dump）。如果公開非盈利數據庫有離線版本，通常會提供給用戶免費下載。我們建議數據庫所有者在發佈資料的時候注意版本管理（version control）問題。使用者引用數據庫資料的時候，可以引用版本資訊，如此就不會導致因為版本更迭，導致無法重現查詢結果的的狀況發生。此外，建議在發佈資料的同時也分享到 Dataverse 和 ICPSR，確保項目結束或者沒有資金維護的時候，用戶仍能下載到數據庫的歷史版本。至於共享方式，上文已有討論。

ii. 子數據集（sub-datasets）。有三種子數據集對用戶會非常有幫助：有價值的中間產物（包括初始的錄入記錄）、以特定查詢條件查詢出來的結果、有工具屬性的代碼表（code tables）。

[20] 數位人文的發展對圖書館本來就有很大的影響，可以參閱 *Digital Humanities in the Library: Challenges and Opportunities for Subject Specialists*, ed. Arianne Hartsell-Gundy, Laura Braunstein, and Liorah Golomb (Chicago: Association of College and Research Libraries, 2015).

首先，有價值的中間產物（intermediate）。從原始數據到成為數據庫的資料，數據會被按照數據庫模型以及項目中工作人員的知識和判斷被抽象、修改。然而從文本開始的研究不只是在數據庫模型架構下的研究。隨著數據離完成越近，數據對於在數據庫模型架構下做的研究越有用處，越加方便。另一方面，對於在架構外做研究，則會越來越不方便。因此在沒有版權問題的條件下，我們建議數據庫項目分享有價值的中間產物。

第二，以特定查詢條件查詢出來的結果。在我們幫助學者用 CBDB 做研究的時候，會發現學者們常常會重覆檢索一些特定內容。如果以 CBDB 為例，例子包括各朝代的進士、歷代在某省曾任政府官職的人名等。以最經常被查詢的結果反向猜測用戶最經常問的問題，在這個基礎上抽取出數據建立子數據集，對於改良數據庫效用、增強用戶體驗和優化查詢速度都非常有幫助。

第三，有工具屬性的代碼表。做數據挖掘（text mining）的學者常常需要人名表、作品表、官名表、地名表、社會機構表等作為字典（dictionaries），或所謂分類（taxonomy）中的分類項來挖掘數據。[21] 分享數據庫的代碼表，非常有助滿足數據挖掘的需求。

iii. 內部工具分享。每個數據庫都會有一些面向特定需求的程序。譬如朝代年號轉換、特定格式的轉換、字符集識別、自動編碼（coding）、自動校對、自動消歧（disambiguation）等。[22] 這些用於臨時工作的程式通常不會被大家分享到 GitHub 上，因為這些程式要處理的內容太特殊，面對的任務太專門。但是在中文數位人文資料處理的領域裡，不同研究者難免會做一些相似的工作。因此，這些程式碼可能複用率不高，但有很大可能性可以為其他開發者、用戶提供思路和基本程序架構。因此，我們建議網絡基礎設施的參與者分享內部工具程式碼，即便是一些規模很小的，或用途特殊的程式。分享的網站建議是 GitHub、「碼雲」、 CSDN 等。並且向我們提交這些的維護項目。我們會將其分類，並發佈在中文數位人文網絡基礎設施網站上。

[21] 一個例子是在地方志裡挖掘歷史人物的數據，介紹參閱 C. L. Liu, C. K. Huang, H. Wang and P. K. Bol, "Mining local gazetteers of literary Chinese with CRF and pattern based methods for biographical information in Chinese history," *2015 IEEE International Conference on Big Data*, Santa Clara, CA, 2015, 1631-1632.

[22] 比如，CBDB 為了內部的工作需要，制訂了不少實用小程序：譬如清除圖片上專名和書名符號的程序；將一頁中多欄排版的圖片切割拼合成一欄的程序；基於地名不同時間斷面將上層地址和下層地址自動整理出歸屬關係的程序；將固定格式的 XML 格式轉換成 Markus 格式的程序；將有方向的成對資料自動減半成無方向無重複資料的程序等等。

第三是**商業數據庫**。商業數據庫可以分享的資料有兩種，一種是無需授權的公開資料，另一種是授權開放的資料。

無需授權的公開資料中，最常見的是數據庫或數位典藏收錄書籍書目。商業數據庫項目通常會公佈下載自己收錄的書目和資料的元數據作為向訂閱者推銷的材料，我們非常希望數據庫所有者可以將書目和其元數據製作成 API，供用戶查詢。用戶可以直接查詢多個數據庫的 API，快速找到自己需要的書籍被哪一個數據庫收錄。除收書書目以外，商業數據庫還會常常開放部分資料，以供試用。譬如中央研究院的「漢籍全文資料庫」的免費部分[23]、愛如生的「搜神」[24]，以及「中國知網」（CNKI）的篇目查詢、篇目的元數據以及引文目錄。[25] 這些資料如果可以製成 API，對於用戶進行大數據分析會有莫大幫助，實現許多在數位人文時代才可以完成的研究。

對於開放授權的資料，上文已經討論過在大數據時代下，研究者有了對數據批量抓取，然後使用統計、地理分析系統（GIS）等軟體對其進行分析的新需求。因此，我們希望數據庫公司可以為自己的數據開發 API 系統，並向購買方提供數據挖掘（text mining）權限的購買選項。據我們所知，哈佛大學圖書館已和其訂閱的一些商業數據庫達成協議，為圖書館用戶、學者、研究項目獲得數據挖掘的權限，以便對數據進行研究。CBDB 的部分數據挖掘工作就是基於這種安排達成的。

第四是**非公開的數據庫**。上文已經討論過對非公開數據庫不公開數據的原因。對於這些非公開的數據庫，我們可以建議其開放兩種資料：

i. 公開的研究成果目錄和數據庫的基本信息。對於小範圍學術性共享的數據庫以及因為系統尚未成熟至可以開放的數據庫來說，也許不容易做到數據公開。但是，我們可以建議數據所有者公開根據這些數據進行研究並已經公開發表的論文、專著書目和元數據。我們可以幫助其基於這些資料創建 API 查詢系統。如果允許，我們希望這些機構能將數據庫的基本狀況，包括數據庫名、參與者、簡介、開始時間、使用材料的目錄等也一併製作入 API 查詢系統。數據庫的元數據和包含哪些數據表，如果它們能公佈，也可以在不公佈數據的情況下，讓學界對其獲得最大程度之瞭解。

ii. 對於不涉及版權以及不涉密的數據庫，我們會希望這些數據所有者可以將資料以壓縮包的形式存檔至 Dataverse 和 ICPSR。如果是專題性數據集，除了傳到這

[23] 網址為：http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm。
[24] 網址為：http://m.soshen.cn/soshen/。
[25] 網址為：http://www.cnki.net/。

些大型平台以外，也可以同時上傳到特定科目的同類平台，例如是 LingHub。[26] 對於項目本身來說，版本管理（version control）是個必須處理的問題，其中重要的原因是為了保持數據庫的可持續性（sustainability）。在數位領域內，我們都見證不少數據庫在項目結束，或者資金短缺的情況下被迫停止對數據庫的開發。隨著項目網站的關閉或失效，原數據庫的資料再也無法被訪問。[27]因此，我們建議即便是受眾狹窄的、還不成熟的、無法承受大流量訪問的數據庫，也將自己的數據庫（包括 Datadump）。如果同意，將程式碼一併壓縮打包會更好）分享至 Dataverse 和 ICPSR。

四、跨項目電子化工具

對於中文數位人文項目的發展，有兩件任務是最核心的：一是創建更多的電子化資料，二是使用電子化資料做出更多的研究。這一部分將介紹一些已經完成和尚在概念階段的中文數位人文工具和方案，它們或有助於電子化更多資料，或有助於推進數位人文研究，並且絕大多數是免費向公眾開放的。我們認為，這些是構建用於中國歷史的網絡基礎設施的重要手段。以下將從五個方面說明。

(一)跨庫書目檢索系統

在查閱電子典籍的內容時，我們常常希望知道某部書或者一部書的某個版本是否已被電子化，而如果有的話，究竟是可以被檢索和挖掘的文本資料，還是掃描圖像檔。這時我們通常只能在各個數據庫逐一檢索，或者逐一下載各數據庫的目錄來進行檢索。CBDB 項目在日常工作中為了解決這個問題，開發了一個跨庫檢索的原型系統。只要輸入書名，就可以對多種數據庫和大型叢書的書目內容進行檢索，迅速查出要在什麼地方找到某部書的電子版。

[26] 網址為：http://linghub.lider-project.eu/。

[27] 又或是項目本來以某個版本的程式開發，之後網站管理員對程式進行更新，在新程式版本中某些語法被改動，導致網站的一些功能失效。

圖片 2：跨庫檢索原型系統的界面和檢索結果舉例。

目前我們更新資料的辦法是從各數據庫網站上下載書目文件，將其手動加入跨庫查詢系統。然而，這種資料收集方式還不理想——它對數據的更新非常不利，一旦數據庫增加了新的資源，本系統無法進行自動更新。我們建議數據持有者將數據庫書目以 API 的方式共享，這樣跨庫查詢系統可以直接調用 API。一旦 API 被更新，查詢結果就可以在短時間內更新。數據格式——也就是書目資料的編排方式，也應該劃一處理。德龍博士未來會接手該系統的後續開發，亦會收錄更多數據庫資源，並提供更強大的檢索功能。預期該系統會和 CTEXT 有更好的整合，並向公眾開放。

開發這樣的書目查詢 API 不僅對用戶有幫助，對於各種數據持有者來說也是一個有效地宣傳他們數據的途徑。如果數據庫持有者暫時沒有計劃和人力來做開發，也可以利用上文提到的 API 自動生成工具，自動將表格轉化成 API。

(二)OCR 技術與中文文本資源的開放

CTEXT 和 CBDB 項目在最近幾年從不同的角度在探索高效中文 OCR （Optical Character Recognition，光學字符識別）的方案。它們在這方面的努力預期會加強學界對中文材料的辨識能力。由於包含中文資料的全文數據庫已非常普遍，業界內已有多種多樣，因而對所有這些數據庫都有幫助的 OCR 技術對於網絡基礎設施的構建無疑可以有重要作用。

圖 3：CTEXT 對古籍進行 OCR 的結果。

（圖片出處：http://dsturgeon.net/harvard-yenching-ocr/）

在 CTEXT 項目中，哈佛大學費正清中國研究中心博士後研究員德龍（Donald Sturgeon）博士探索了使用機器學習（machine learning）的方法快速將中國古代刻本轉成可進行全文檢索的資料。如今，他正為哈佛-燕京圖書館（Harvard-Yenching Library）將該館掃描的中國古代文獻全部轉成可檢索的文本資料。[28] 值得強調的是，CTEXT 項目正向互聯網免費公開沒有版權問題的、用戶可以參與校對的繁體電子文本資源，而且當前的校對用戶貢獻數量頗為可觀。由此，CTEXT 項目已經成為進行中文數位分析的一個重大數據來源。

至於 CBDB 項目，則提出了整套面向現代印刷文本的文本化方案。這種電子化的流程利用高效率的掃描器和專門編寫的電腦程式，改良電子化和數據擷取的工作。此流程強調擇取格式規整、經過排版的文獻版本，然後用電腦算法，處理書頁上的格式和標點符號，把文字辨識的錯誤率減到最低。通過長期搜集使用 OCR 軟件時的積累的校對記錄，建立起一套可以自動偵錯的校對平台，使得人工校對工作量得以減輕。此

[28] 有關消息參閱：http://dsturgeon.net/harvard-yenching-ocr/。對這種 OCR 方法特點的介紹，參閱：http://ctext.org/instructions/ocr。還可以參閱：https://cpianalysis.org/2016/06/08/crowdsourcing-apis-and-a-digital-library-of-chinese/。

工作流程能較有效率對文獻進行電子化，並轉化成結構化數據，放入關係型數據庫中供學者使用，為大批收錄、處理中國古代資料提供一種較有效率、較可靠的方式。[29]

以上介紹兩個項目的 OCR 方法無疑會改變電腦處理中國歷史文獻的方法，是數位人文研究在 OCR 技術上值得注意的發展。將更多資料電子化是中文數位人文數據庫的重要努力方向之一，在這個目標下，有些項目提出了以技術為主導的思路，有些項目提出了以人工為主導的思路。這兩種思路在本質上沒有衝突，可以並駕齊驅。以技術為主導的思路不會忽視人工檢查的重要作用，以人工為主導的思路如果有技術的輔助，也可以大大提升人工作業的效率。如何組織人力，以及最大化減低工作成本、增加工作準確度，同樣是專門的技術，值得一起探索。因此。我們希望中文數位人文網絡基礎設施可以成為分享和推進電子化文本技術的渠道。對於商業數據庫來說，這方面技術也預期會帶來頗大的商業收益。

(三)標記與可視化工具

MARKUS 是萊頓大學的何浩洋（Hou Ieong Brent Ho）博士和魏希德（Hilde De Weerdt）教授開發的中文文本標記系統。此系統免費向公眾開放，用戶可以使用這個系統自動標記人名、別名、時間、地名、官名等信息。系統可以自動識別以上各種類型的數據並且自動標記，並且允許使用者利用正則表達式（Regular Expressions）原理以及「詞夾子」算法（keywords clipper）進行快速批量標記和關鍵字的擷取。[30] 自動和批量標記完成之後，可以人工對標記結果進行核實。當前 CBDB 項目正在使用 MARKUS 來對《宋會要》和唐代墓誌進行文本挖掘，從中提取人物傳記資訊。MARKUS 為這些工作提供了極大的便利。[31]

另外，隨著近年數據科學（data science）的興起，當前有許多開放易用的可視化的工具。不少在網上設立平台，允許用戶導入自己搜集、處理過的數據，製成圖表、圖形、地圖等等。坊間已經有不少開放的工具，能讓數據庫具備各種可視化功能。由於類似工具眾多，這裡不一一贅述。[32] 因此，在數據庫開放的過程中，我們沒有必要重

[29] 參閱徐力恆、王宏甦：《中國古人傳記資料的數碼化方式芻議：從歷史文本到結構化數據的半自動化流程》，發表於香港公開大學 2016 年度數碼文化與人文學科研究所會議，2016 年 7 月 5 日。

[30] 參閱何浩洋：《MARKUS：中文古籍文本半自動標記平台》，https://www.academia.edu/11078612/MARKUS_%E4%B8%AD%E6%96%87%E5%8F%A4%E7%B1%8D%E6%96%87%E6%9C%AC%E5%8D%8A%E8%87%AA%E5%8B%95%E6%A8%99%E8%A8%98%E5%B9%B3%E5%8F%B0。關於詞夾子，參閱謝育平，《同位詞夾子主題式分類詞庫萃取演算法》，載《數位人文研究的新視野：基礎與想像》，臺北：國立臺灣大學出版中心， 133-162 頁。

[31]參閱 Lik Hang Tsui & Hongsu Wang, “Creative Uses of MARKUS in the China Biographical Database Project,” MARKUS Research Blogs, Oct. 2016, http://dh.chinese-empires.eu/forum/topic/5/creative-uses-of-markus-in-china-biographical-database-project/

[32] 例如，地理信息可視化的工具除了上述已經提到的，還有以下多個：WorldMap（http://worldmap.harvard.edu/）、Palladio（http://hdlab.stanford.edu/palladio/）、Leaflet

新花費大量資源重新開發類似工具。選擇使用這些工具，首要目的不在於展示的美觀，而是為了有助文史研究者更好地組織資料、提出並回答研究問題。

(四)代碼表

無論是關係型數據庫或數據集，使用者都會用到代碼表（code tables）。這些代碼表是一些通用數據的字典。目前適合用於中國歷史研究，又是免費、公開的代碼表有很多。僅舉出來自少數幾個項目的例子，譬如 CBDB 項目的人名表、地名表、官名表、入仕方式表、親屬關係表、社會關係表、社會機構表等等；[33] CHGIS 的歷史地名表、地名坐標、地名層級關係等；[34]「明清婦女著作」（Ming Qing Women's Writings）項目的女性作家人名表、作品表；[35] 法鼓佛教學院「佛學規範資料庫」的佛教時間、人名、地名表等等。[36]

這些開放且免費的代碼表對數據庫的製作、連接，以及學者進行數位人文的研究提供了許多的便利。我們也希望參加中文數位人文網絡基礎設施的成員在製作數據庫時，能製作並且分享更多代碼表，使得共享數據更加豐富，亦有助數據的規範化和除錯。

(五)API 和數據分享工具

API 分享和數據分享工具可以在幾乎是零成本的基礎上，為數據庫項目創建簡單的 API 系統，以及將數據分享至可靠的、可以進行版本管理的存儲空間上。這方面內容已經在上文詳細說明，此處不再贅述。

五、成員交流

成員交流的方式有兩種，一種是基於互聯網的**非常規交流**，另一種是**常規交流**。以下將討論如何營造最有利於交流網絡基礎設施相關問題的具體環境。我們應該盡量鼓勵合作機構有多於一個成員參與交流。在交流中能得到管理者和技術人員兩方面參與討論，對於共同解決問題非常有幫助。另外，人事改動是不同項目常會碰到的事情，也是對交流的挑戰之一。如果一個項目中參與交流的人員始終是同一個成員，那麼一旦經歷人事改動，就會帶來與合作機構之間的聯絡中斷的危險。因此，保證項目中有多於一位成員參與對外交流工作，並在交流中保持翔實的文獻記錄，對項目之間的長期合作也是很重要的。

（http://leafletjs.com/）等。這些工具大大簡化了地理信息可視化的開放和共享。社會網絡關係的可視化計劃則有：D3js（d3js.org）、D3PLUS（d3plus.org）、Processing（processing.org）等等。

[33] 參閱：http://projects.iq.harvard.edu/cbdb/download-cbdb-standalone-database。

[34] http://maps.cga.harvard.edu/tgaz/。

[35] 參閱：http://digital.library.mcgill.ca/mingqing/chinese/download.php。

[36] 參閱：http://authority.ddbc.edu.tw/docs/open_content/download.php。

另外，這種交流必須是沒有預設的跨學科溝通。文科學者在網絡基礎設施構建中面臨的一大問題是怎麼可以跟電腦，甚至是美術設計、統計等學科的學者交流。在數位人文的研究範式下，人文學者未必要掌握其他學科的專門知識，但應該有具備跟其他學科交流的能力，在溝通過程中能從對方的角度考慮問題，不能過於在意學科之間的藩籬。這種能力的培養還是要從實踐中摸索，這屬於數字人文發展的一個主要挑戰。

(一)基於網路的信息溝通方式

基於網路的信息溝通方式大致有三種，網站、郵件組和社交網絡小組。接下來對這三種溝通方式的用途逐一進行討論。

首先是**網站**。網站為所有中文數位人文網絡基礎設施參與者提供了一個公開無交流障礙的交流平台。我們可以使用 WordPress 平台快速搭建便於使用的網站。在這個網站上我們可以設置如下幾種內容：

第一，定期採訪成員或者專家學者來撰寫採訪稿。這樣做可以為項目成員或者數位人文領域的專家提供可以充分表達自己觀點的平台。另外，由有經驗的數位人文專家進行採訪，可以相互碰撞出有趣的話題，也是對各個項目的推廣。

第二，為不同機構提供發佈新聞和專題稿的賬號。這樣這個網站可以成為參與成員的宣傳渠道。

第三，我們可以開發抓取項目新聞的程式，從參與方的網站上定期自動採集新聞資料，例如是某數據庫開發了新功能，或發佈了剛剛電子化的古籍。然後，發佈到網絡基礎設施的網站上。如此一來，不再需要各個項目都有人專責在某處發佈消息，減輕參與者的工作負擔。

第四，鼓勵參與者在網站上發佈一些日常的工作經驗和發現。這些像工作筆記的經驗分享雖可能是面對非常具體的問題而寫，但是解決問題的思路或部分代碼可能會對其他學者和專家在碰到相似問題時有幫助，不需要在相似的問題上再花費更多時間。而且這些經驗的分享會促進成員之間的交流，甚至引出一些意想不到的話題。這對網站的活躍程度，以及使網絡基礎設施的交流變得活躍大有幫助。[37]

第五，科普性文字。我們可以定期發佈一些科普性的文章，並且向一些媒體投稿，將網站鏈接附在投稿的文章下面。[38] 這對於讓更多的人了解什麼是數位人文，甚至用

[37] 比如 CBDB 項目在最近設立了博客，網址為：http://projects.iq.harvard.edu/cbdb/092016。
[38] 類似嘗試，可以參考近來成立的數位人文微信公眾號「01 Lab」。

數位人文的工具解決一些小型的學術問題（譬如本科學生的課程論文，或者 Quora、「知乎」之類網站上有趣的問題等）。

第六，資源的分類和匯總。分享的內容上文已有討論。

同類網站在其他服務研究的網絡基礎設施已有成立，值得參考，例如是歐洲數位研究基礎設施 DARIAH（Digital Research Infrastructure for the Arts and Humanities）。[39] 除此之外，我們還可以根據參與者使用通訊工具的情況，在各大社交媒體註冊中文數位人文網絡基礎設施的公共賬號，把網站的內容推送給用戶和大眾。

第二是**郵件組**（listserv）。在數位時代，郵件組已經算是古老的交流工具，但仍是一些學術社群進行交流的關鍵工具。即使是數位人文這個領域，也有同類郵件組。[40] 相對網站來說，它的優點是能第一時間把郵件推送給所有參與成員。幾乎所有人都有工作或私人郵箱，所以我們可以藉助郵件組來發佈新聞，以及使之成為中文數位人文網絡基礎設施成員向其他成員請求答疑解惑的平台。此外，由於郵件組的封閉性，如果需要在內部分享不能全面公開的資料，郵件組也是不二的選擇。因此，我們建議把郵件組當做一對多廣播的主要平台。

第三是**基於社交賬號的虛擬社區**。社交賬號的最大問題是不同地域的人習慣使用不同的社交賬號，譬如 Facebook、微信、Line、Whatsapp 等。因此，它們不適合一對多的消息廣播。另外，由於虛擬社區的封閉性，它同樣也不像網站是適合發佈面向公眾信息的平台。社交賬號的優點是方便實時討論具體問題。通過多人在短時間內的互動問答和陳述，可以快速清楚地描述問題、獲得反饋，以及在交流互動中產生進一步思考。因此社交賬號可以成為進行主題討論的工具。最後，因為社交賬號虛擬社區的非正式性，一些在正式、嚴肅的場合不適合討論的問題，也可以在社交賬號的虛擬社區中討論。

對於中文數位人文網絡基礎設施的組織者來說，建議註冊所有以上的賬號，並且為網絡基礎設施建立上述所有賬號的討論社區。

(一)成員的常規溝通

中文數位人文網絡基礎設施成員的溝通方式可以有如下幾種：大會、定期會晤、主題討論、訪談和項目合作。提出這種建議的用意是結合線上線下不同形式，促進對話。參與者得包括範圍可以根據不同議題、領域而定，有時全體參與交流，有時和特定合作方集中交流，以此保持靈活度和高的針對性。

[39] 參閱：http://www.dariah.eu/。
[40] 例如是 Humanist Discussion Group，網址為：http://dhhumanist.org/。

首先是召開一次由中國數位人文核心參與者出席的**大會**，討論諸如本文提出的計劃，探索其可行性，並搜集其他意見，以謀網絡基礎設施的構建。這個大會應該有各個重要數位研究計劃、典藏、工具的參與者出席。目前，三位筆者已經擬定相關的會議方案，正聯同 CTEXT 的主管德龍博士聯絡各方專家，希望啟動這個計劃。

其次是**定期會晤**。我們建議中文數位人文網絡基礎設施組織定期會晤，譬如舉辦年會。在年會上，成員做正式的研究報告或者報告其一年的工作。會晤以學術報告的形式進行，其目的首先是促進數據開發者和研究者交流——讓數據開發者更好的從需求出發進行數據的開發。同樣對於研究者來說，可以藉此了解更多的數字資源。其次，是促進數字人文學者交流，分享研究成果。還有，可以促進數據開發者之間進行技術上的交流，甚至協調數據工作的分工。

再其次是**主題討論**。每年裡面數位人文領域會產生很多主題新聞。大到新書出版，技術的重大突破等等；小到論文發表，數據庫發佈，新項目啟動等。這些主題新聞中會有不少可以成為中文數位人文網絡基礎設施討論的主題，甚至有一些主題新聞就是就來自成員本身。對於這些主題新聞，我們可以通過網站和社交賬號組織討論。對於有價值的討論內容，可以整理成稿件，發佈在網站和社交賬號平台上，以廣流傳。

第四是**訪談**。我們可以每隔一段時間逐一採訪中文數位人文網絡基礎設施的成員。我們可以將訪談內容整理成文發佈到網站上，甚至對訪談進行錄音或錄像。根據 CBDB 過去舉辦活動的經驗，訪談主持人建議嚴格確定，不應隨意尋找。這是因為主持人的知識背景、如何設計問題和引導討論的技術對訪談內容會有決定性的作用。至於訪談方式，不用只局限於面對面對話，Skype 視頻也是合適的途徑。

第五是**項目合作**。我們鼓勵中文數位人文網絡基礎設施成員之間進行項目合作。在前三種交流中，都可以促進成員之間的相互了解。而一旦成員之間發現相互的資源互補時，一起申請項目是非常好的合作方式。在同一項目框架下也會更促進成員之間的交流和溝通。

總而言之，共同建立一個連接不同項目和專家的基礎設施，對從事數字人文工作的所有成員都是有利的。這是中國文史研究界面對當下研究狀況應該考慮的問題。即使有限度地進行這種數位方面的合作，仍是對各方都有利的。我們懇切希望這樣的對話能儘快開始——當研究者開始商量怎麼合作時，用於中國歷史研究的數位人文作為一個領域才算正式出現，相應的學術共同體也會因而逐步形成。我們在學界必先達成共識，認同必須開發網絡基礎設施，各政府和私人機構才會相應地作出配合。以上構想是我們倡議的內容，還望各位向我們提出建議，一同參與這種數位人文網絡基礎設施的營建。

The DH Scholar as an Intermediary： Connecting Physics and Theatre Scholarship

Miguel Escobar Varela*

Abstract

This paper describes three collaborations between theatre scholars and physicists. It focuses on three interdisciplinary studies of Javanese theatre: the application of signal processing to the analysis of theatre video recordings, an analysis of network structures in the fictional universe of wayang kulit, and a biomechanichal study of movement in Javanese dance. These collaborations are both challenging and fascinating, and show a real potential for increasing knowledge in different, previously unrelated fields. Collaborations of this nature require scholars to become intermediaries who constantly translate concepts, methods and expectations between different fields of knowledge. This paper argues that the essential skills of a DH Scholar are the capacity to be an intermediary and to collaborate with other intermediaries.

Keywords: networks, biomechanics, signal processing, theatre, Indonesia

* Assistant Professor, National University of Singapore. Email: m.escobar@nus.edu.sg.

數位人文學者的中介角色：連結物理學與戲劇學者

Miguel Escobar Varela*

摘　要

本文描述戲劇學者與物理學者之間的合作案例，主要內容為三個關於爪哇戲劇的跨學科合作研究，包括以訊號處理技術分析戲劇影像紀錄、哇揚皮影戲（wayang kulit）中虛構世界的網路結構分析、以及爪哇舞蹈動作的生物力學研究。這些研究除了成果令人振奮與著迷之外，也展現出以往被視為不相關之學科在合作時發掘新知識的潛力。跨學科合作需要學者扮演中介角色，持續在不同知識領域之間傳遞概念、方法與可能性。本研究認為數位人文學者需足以扮演各學科之間的中介角色，或與其它中介者合作，此為數位人文學者之重要必備能力。

關鍵字：網路、生物力學、訊號處理、戲劇、印尼

* 新加坡國立大學助理教授，Email: m.escobar@nus.edu.sg。

1. Introduction

This paper describes three collaborations between theatre scholars and physicists. It focuses on three interdisciplinary studies of Javanese theatre: the application of signal processing to the analysis of theatre video recordings, an analysis of network structures in the fictional universe of *wayang kulit,* and a biomechanichal study of movement in Javanese dance.

2. Signal Processing and the Study of Theatre Video Recordings

The area of the digital humanities where quantitative methods have made their greatest impact is the study of literature. Well known approaches include “distant reading” (Moretti, 2013), “macro-analysis” (Jockers, 2014) and stylometry (Holmes, 1998). Researchers routinely apply a methods from statistics, machine learning and graph theory to study literary objects. Some of these methods have been used to study artistic production which is not primarily linguistic, such as the visual and performing arts. Little attention has been devoted to the quantitative analysis of visual data in a way comparable to the quantitative analysis dedicated to linguistic features. Image processing has been used for the study of paintings (Jacobsen and Nielsen, 2013; Mureika and Taylor, 2013), but this work is developed solely from an engineering perspective and not from an inter-disciplinary digital humanities perspective.

However, we believe that the study of areas such as the performing arts has much to gain from the application of quantitative tools; for example, the usage of image processing techniques for the study of performance recordings. This is particularly important since the amount of available recordings of theatre and performance is growing, as digital archives become more prominent (Sant, 2017; Leonhardt, 2017). Eventually we hope that 'distant' or 'macro' analyses of these recordings will become widespread. However, we believe that a necessary precursor for this is the validation of image processing methods as reliable tools for the close analysis of performance recordings. Our study (Escobar Varela and Parikesit, forthcoming) aims to contribute to this development by using an example from the Javanese tradition of *wayang kulit* (leather puppets). In the study of theatre in general, and puppetry in particular, the analysis of motion is a key area of concern. In the article, we described how we use image processing tools to measure the average motion speed of each scene in a *wayang kulit* show and suggest that there is a significant connection between the average speed of the different scenes and their narrative function. There is a progressive increase in the average speed of each scene until the performance reaches the *gara-gara* scene, a light-hearted comical interlude. This interlude, which is a common section in *wayang* shows looks like a dip in the graph, before speed picks up again in the concluding scene of the sow. The video we analyzed in this article was a recording of Catur Benyek Kuncoro’s *Wayang Mitologi* [Mythology

Wayang, 2012], which is part of the Contemporary Wayang Archive (cwa-web.org).

3. Fictional Networks in *Wayang* Mythological World

Theatre depends on relationships. A theatrical production is impossible without the collaboration and co-presence of different people. Even the most minimal productions require at least a performer and one spectator. And dramatic texts often also explore connections between characters, even if only a single character is portrayed on stage. Network analysis is one way in which these relationships – both on and offstage – can be analyzed. However, the usage of network theory has rarely been used to study theatrical relationships. By network theory, we refer to the mathematical study of networks and not to methodological approaches such as ANT (Actor-Network Theory) which have indeed been often used by theatre scholars.

The lack of applications of network theory to study theatre is surprising since a famous theatre scholar – Jacob Levy Moreno – was instrumental to the development of network analysis. In theatre studies, Moreno is more commonly remembered as the father of sociodrama. But he was also the inventor of sociograms: visual representations of connections between people. He was one of the first people to realize that networks could depict social relationships. However, until recently, the study of networks often focused exclusively on the mathematical properties of networks. This trend has been challenged by some digital humanities scholars - such as Moretti, Elson and Finn - that work at the intersection of quantification and interpretive analysis in the study of literary networks. Following this direction, the objective of a collaboration between Andy Schauf and Miguel Escobar Varela is to combine the quantitative analysis of theatre networks with a detailed overview of the context in which those networks emerged. We focus specifically on the networks of the *wayang kulit* mythology and analyze the complex ways in which communities (protagonists, antagonists, gods, ogres and clowns) are interconnected in this fictional universe. When we remove different communities and measure the topological properties of the resulting network we realize that the *carangan* [side story] characters have a very important structural role. Their presence in the network has the effect of increasing both the algebraic connectivity and the diameter of the network. In other words, the effect is counterintuitive. Removing this groups increases the average “tightness” of the connections (density and algebraic connectivity), while also pushing some of the elements of the network farther apart from each other (increasing the diameter). Although the presence of the *carangan* and characters dampens the density and connectivity of the network, they also bring distant characters closer together. In other words, the *carangan* characters are not “outsiders”, as scholarly tradition suggests, but essential to the topology of this fictional universe. Like many great stories, the Mahabharata is a tale of warring factions. Scholarly

commentary often focuses on the struggle between these competing sides. But what traditional scholarship misses is brought into sharp focus by network analysis: the story does not just involve two sides, but several smaller subgroups that are caught in conflict where identity fault lines are not easily discerned. The quantitative tools of network analysis can thus aid scholars of culture to move away from reductive binary oppositions into a complex tapestry of conflicting allegiances.

4. Biomechanics of Javanese Dance

In early 2016, Luis Barraza and Miguel Escobar Varela conducted a study where they attached motion and electromyographics sensor to dancers in order to measure differences between dance styles in a quantitative way. Dance is one of the most important aspects of Javanese intangible heritage. Its academic study has a long history but it has been limited to qualitative descriptions. We believe that a more systematic, quantitative analysis of dance is needed in order to better understand the evolution of dance forms, create more effective cultural policies and even train future generations of dancers. For this purpose, we investigated the kinetic and kinematic differences between Javanese dance styles at the gait analysis laboratory of a local Singaporean university. One professional Javanese dancer was recruited and instructed to perform movements that correspond to different character types in the *Sendratari* dance-drama: vigorous (*gagah*), ogre (*raksasa*) and refined (*lanyap*). The dancer stood up from a kneeling position in the way that befits each character type. A motion capture system and force plates were used to measure the kinematics and kinetics of each standing movement. One factor ANOVA was used to compare the movements. Our results showed that the only character type which had a significant difference was the vigorous character. This result is surprising since a qualitative analysis of the characters would describe the vigorous and refined characters as being more closely related. However, our results show that the underlying structures of the movements of refined characters and ogres are actually very similar. This insight into Javanese dance can only be obtained through quantitative analysis. Thus, the results illustrate the way a scientific understanding of the biomechanical signatures of the different character types can help dance scholarship.

5. Conclusion

The collaborations described above are both challenging and fascinating, and show a real potential for increasing knowledge in different, previously unrelated fields. Collaborations of this nature require scholars to become intermediaries who constantly translate concepts, methods and expectations between different fields of knowledge. This paper argues that the

essential skills of a DH Scholar are the capacity to be an intermediary and to collaborate with other intermediaries.

Bibliography

Escobar Varela, Miguel and Gea Oswah Fatah Parikesit (2016, forthcoming), 'A Quantitative Close Reading of a Theatre Video Recording, *Digital Scholarship in the Humanities.*

Elson, David, Nicholas Dames and Kathleen R. McKeown (2010), 'Extracting Social Networks from Literary Fiction', *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 138-147.

Finn, Ed (2013). 'Revenge of the Nerd: Junot Díaz and the Networks of American Literary Imagination', *DHQ: Digital Humanities Quarterly* 7.1.
Holmes, David (1998). 'The Evolution of Stylometry in Humanities Scholarship', *Literary and Linguistic Computing 13(3*): 111-117.

Jacobsen C.R and M. Nielsen (2013). 'Stylometry of paintings using hidden Markov modelling of contourlet transforms', *Signal Processing 93*: 579–591.Moreno, Jacob Levy (1934) *Who Shall Survive?* (New York: Beacon House).

Jockers, Matthew L. (2013) *Macroanalysis: Digital Methods and Literary History* (Baltimore: Urbana University of Illinois Press).

Kaplan, Debra (2015). 'Notes from the Frontier: Digital Scholarship and the Future of Theatre Studies', *Theatre Journal 67(2)*:347-359.

Leonhardt, Nic (ed) (2016). *The Routledge Companion to Digital Humanities in Theatre Research* (Routledge, forthcoming).

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

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

Mureika, J.R. and R.P. Taylor (2013). 'The Abstract Expressionists and Les Automatistes: A sharedmulti-fractal depth?' *Signal Processing 93*: 573–578.

Sant, Toni (ed) (2016). *Documenting Performance: The Context and Processes of Digital Curation and Archiving* (Bloomsbury, forthcoming).

Paper Session 3

型式探求：文學與藝術

Pattern Recognizing: Art & Literature

Concept Modeling and Advertising Chinese Modern Society

Tani Barlow[*], Jing Chen[**], Ke Deng[***]

Abstract

In the late 19th century, thousands of industrially produced consumer items flooded into extraterritorially governed, internationally regulated, Chinese, treaty port cities. Foreign commodities were products, and formed the backbone of new, urban, popular consumer culture. Consequently, the advertising industry infiltrated commodity brands and branding techniques into everyday life making commodity images a paramount symbol of civilized urban life. Ads became a powerful vehicle for circulating desirable commodities in a modernist commercial culture. For us the significant point is that black and white newspaper ads are "[m]inor transient documents of everyday life," ephemera, which we see as a graphic epistemology. The ads or "graphs" forward axiological ideas like "modern things are clean" or "people are mammals," for instance, but they do it wordlessly. (Drucker) Advertising ephemera thus provides researchers with the conditions for thinking about modernity par excellent since it breaks data free of its origins to demonstrate how concepts embedded in ads ingratiate all consumer cultures. Social theorists and historians in the first third of the 20th century agreed commercial ephemera were modern and had value.

To force advertising to speak clearly, we launched the Chinese Commercial Advertisements Archive ("CCAA") and "metadated" (Lev Manovich's term) more than ten thousand high quality images from microfilm copies of five, major, commercial, Chinese newspapers, in the period of 1880 to 1940.[i] CCAA applies customized metadata schema based on the structural standard, Dublin Core, to each digital image of advertisement, entering all relevant information e.g., brand icon, word texts, street names and business titles. Our metadata include: descriptive content, contextual

[*] Professor, History Department, Rice University, USA. Email: barlow.tani@gmail.com.
[**] Associate Professor, Art Institute, Nanjing University. Email: cjchen@nju.edu.cn.
[***] Associate Professor, Statistical Research Center, Tsinghua University. Email: kdeng@tsinghua.edu.cn.

information, bibliographical, technical and image sources of location, copyright status, and owning institution. Our objectives are (a) to make useful advertising data available to cultural critics and historians and (b) to understand how commercial cultural life emerged in China in the late 19th and 20th centuries and saturated the historical unconscious.

In order to explore the mechanisms of images, texts, commercial products, advertising industry, global capital and modern disciplinary order, we propose a concept model rooted in content analysis and designed to reveal connections among modernist fields where visual culture (text/image), commercial power (commercial products, advertising industry and global capital) and social theory (modern disciplinary order) combine and the everyday of ordinary people as transient ephemera conveying philosophical ideas. Our aim is to produce more knowledge about advertisements and the advertising industry, and to provide a big picture of advertising (aka distant reading) while challenging traditional research methods. Before going through, we have overcome many of these roadblocks using statistical methods for Chinese text mining and knowledge discovery. Text mining allows us to: 1) discover potential associations among features and terms extracted from advertisements; 2) build links among these and ideological trends in the treaty port urban areas of China during our period by developing Deng Ke's statistical text mining method to establish indices of technical terms (TT) and metadated association patterns among technical terms (APTT).[ii]

With the case of Dr. Williams, we would be able to see better that the Dr. Williams campaign might be a jumping off point for understanding why elites in modern China bought quack meds like Pink Pills, Pinkettes and She-Ko and how Dr. Williams, an ambitious global company, play the global marketing game with a clear, consistent strategy, strong financial support and mature executive decision making skill.

Keywords**:** advertisements, ephemera, archive, text mining

概念模式與中國現代廣告社會

白露*、陳靜**、鄧柯***

摘　　要

19 世紀晚期，數以千計的日用消費品洪水般湧入為治外法權所管轄、國際規約控制的中國口岸城市。外國商品成為了全新的都市流行文化的主要載體。與之相應地，廣告業巧妙地將商品品牌及品牌技巧滲入到老百姓的日常生活之中，使得商品圖像成為了一種極為重要的、都市文明生活的象征符號。廣告成為了一種強有力的方式，在現代主義的消費文化中推動着充滿了欲望的商品流通。其重要意義在於，這種黑白兩色的報紙廣告像蜉蝣一般，是"日常生活的轉瞬即止的記錄。"我們將之視為一種圖像性的認知論。廣告或者"圖形"用一種無言地方式傳達了諸如"現代生活是整潔的"或者"人是哺乳動物"這樣的價值觀。廣告蜉蝣也因此為研究者提供了條件去思考關於現代性的優越性，特別是在廣告從其原初狀態解放出來成為數據後，就能夠為我們所用去證明嵌入在廣告中的概念是如何迎合那些商業文化了。

為了進一步發掘廣告的意義，我們啟動了"中國商業廣告數據庫項目"（CCAA）並且開始"元數據化"上萬張的高質量圖片。這些圖片是從 1880 到 1940 年間的五份主要的商業中文報紙的縮微膠卷轉換而來。在 Dublin Core 的標準上，CCAA 使用了定制化的元數據框架，對每一張外國商品廣告的數字圖片進行標註，輸入所有可能的信息，比如品牌名、文本和交道名和公司名等等。我們的元數據包括描述性內容、語境信息、文獻信息、技術和圖片信息等等。我們的目的是：1）使得廣告數據對文化批評家和歷史學家可用；2）對商業文化生活在 19 世紀的晚期和 20 世紀在中國的出現及其對歷史無意識的滲透得以理解。

為了進一步探索圖像、文本、商品、廣告、全球資本和現代規訓秩序的內在機制，我們提出了一種基於內容分析的概念模式，以期能解釋現代主義領域中的內在關聯性。在這個現代主義的領域中，視覺文化（文本/圖像）、商業力量（商

* 美國萊斯大學歷史系教授，barlow.tani@gmail.com。
**南京大學藝術研究院副教授，Email: cjchen@nju.edu.cn。
*** 北京清華大學統計學研究中心副教授，Email: kdeng@tsinghua.edu.cn。

品、廣告和全球資本）以及社會理論（現代規訓秩序）結合起來，使得日常生活中的人們接受蜉蝣產物所傳達的哲學觀念。我們的目的是發掘關於廣告及廣告業更多的知識，並提供一種方式，得以從遠距離審視在更大範圍內的廣告業。我們使用統計學方法幫助我們克服障礙，並期望文本挖掘來幫我們發現：1） 發掘由廣告內提取出的詞語及特征所具有的內在關系；2）簡歷這些詞語及內在關系與中國通商口岸內的意識形態趨勢之間的關系。我們主要依賴的方法是鄧柯博士提出的文本發掘方法，通過建立技術術語索引和技術術語內部的關聯模式。在韋廉士醫生的個案中，借助統計數據和文本發掘的初步結果，我們得以更清楚地理解韋廉士在現代中國中的位置，並能理解韋廉士醫生，作為野心勃勃的全球公司，是如何以一種清晰的、持續的策略、強大的經濟支持和成熟的決策制定技巧來進行全球市場推廣的。

關鍵字：廣告、蜉蝣、檔案庫、文本挖掘

Introduction

DH projects are time consuming and costly. What do we actually get out of them? What does DH provide intellectually that we cannot achieve using traditional scholarly means? Our team is inventing ways to exploit digitized, metadated 20th century, black and white, newsprint, and cartoon style advertisements to extract unprecedented historical truths. We consist of a historian, DH cultural studies scholar and a statistician, working together to develop a new resource, processing text/image resources innovatively, and statistically querying new DH data in novel ways. Our objectives are (a) to make useful advertising data available to cultural critics and historians and (b) to understand how commercial cultural life emerged in China in the late 19th and 20th centuries and saturated the historical unconscious.

Why a Chinese Commercial Advertising Archive (CCAA)?

Tani Barlow has shown that commercial ads are historically embedded text/images that began appearing in Chinese lithographic print news media in the late 19th century. Commercial centers, or "treaty ports," participated in a new, world-wide, ad industry that developed slogans, sophisticated cartoon drawings, innovative fonts and syntax to sell machine-made, branded, commercially exchanged large and small commodities. Ads became a powerful vehicle for circulating desirable commodities in a modernist commercial culture. For us the significant point is that black and white newspaper ads are "[m]inor transient documents of everyday life," ephemera, which we see as a graphic epistemology. The ads or "graphs" forward axiological ideas like "modern things are clean" or "people are mammals," for instance, but they do it wordlessly. (Drucker) Social theorists and historians in the first third of the 20th century agreed commercial ephemera were modern and had value. Barlow has argued that ads are part of the modern disciplinary order consisting of psychology, sociology, political science, etc.. Supporting generic ads there eventually emerged an entire commercial industry devoted to buying and selling advertising space. Since people determine social value our potential scholarly users can exploit our ad archive to discover not only how 1920s and 1930s media space was sociologically quantified, bought and sold, but also how brilliant impresarios like Morishita Hiroshi, C.P. Ling and Carl Crow creatively pioneered complex ad campaigns that were pedagogical, linguistically innovate and philosophically unprecedented.

There are several reasons to invest time and intellectual creativity into the CCAA. First, the black and white drawn ad is specific to a 60-year period that began and ended. The mature ads made a debut in 1919 and, when Japan declared total war against China and publishing houses moved inland, the ad industry went into hiatus. Thus we can periodize modern ads; they have a beginning and an end. Second, archived advertisements clearly show the modern Chinese people's imaginative experiences. Along with more conventional historical evidence, like maps, documents, literature, and social science disciplines, ads help historians puzzle out commodity-human relationship at a specific moment in time. Cultural life in new cities, filtered through a novel commercial culture, changed how people creatively altered their ways of life. The ad archive can pictorially show modern citizens enjoying their new commodity-object world. While these are not mirror images of real people the picture element of sophisticated advertisements shows an idealized world where modern commodities, medicines, social activities (drive a car, going to a dance, cleaning or sewing, fertilizing a field) are normative. Also, in the shift to lithography, the black and white advertising image was unprecedented. Ads combined pictorial elements from Dianshizhai huabao, Republican iconography, and the modern arts. Never before the late 19th century had Chinese commercial advertising circulated graphs idealizing, selling machine produced modern commodities to a mass buyer public.

Why CCAA is a Digital Humanities Project:

Our project is singular because we began from a traditional historian's monograph and repurposed Barlow's evidence. The CCAA seeks to raise and resolve questions that individual or even collaborative groups of traditional scholars cannot address unless we provide them *data*. Data and evidence are generically distinct. When Chen, a DH scholar, looks at traditional research evidence she sees a grid that does not summarize commonality but rather selects information implicit in the traditional archival materials, such as image description, transcription of texts, geospatially traceable addresses and street names, commercial information of commodities, and linguistic neological sociology of ads, and she applies systematic metadata schema to convert the information into structured data. What she creates is a database that can count, compare, associate, and generalize in response to new questions. Knowing how many images a product advertiser sold, knowing how many square centimeters an advertiser bought to showcase a new ad, finding out what the ratio of space to price was in 1933 mass media; all of this feeds back into queries that establish a commercial nexus for the emerging popular commodity culture. That, in turn, means

we can establish how norms changed with some certainty on the levels of graphesis, image and rhetoric. For instance, even the small data already supports arguments about probable advertising strategies. How? The data reveal conscious decisions about marketing in specific media, at definitive dates, and describes a new rationality. Advertising is a creative genre but it is never random because it always seeks profit. This reality allows Chen to pursue implicit motives of capitalists and capitalism. While "capitalism" is broadly transformative in the widest sense, it is not abstract nor is it an act of god. Individuals, groups, companies, industries contingently decide about investing, producing and selling specific commodities. They aggressively initiate new marketing and distribution systems. This deepens the hypothesis that industrially produced and branded commodities transform not just the visual landscape but also how consumers understood themselves, their everyday lives, and their capacities.

Quantitatively and qualitatively a digital archive is distinctive because it makes entry into the mechanics of capitalism possible but also, it makes data unlike a conventional library collection. First, data in a digital archive is finite. CCAA focuses exclusively on foreign, transnationally iconic ads for international corporations like General Electric and Jintan. We have excluded national brands for two reasons. First, metadating and digitizing all commercial advertising would be exponentially costly and difficult. Second, the relative selling power of national and international brands is an unknown; scholars have not yet determined what a "national brand" is in relation to a "transnational brand." Sherman Cochran, Howard Cox, Karl Gerth, William Kirby, and Madeleine Zelin, all of them traditional historians, present a general understanding. But they do not agree about what is a national versus an international firm. Second, because writing the history of business or capitalism is so difficult we have structured CCAA to help scholars approach questions indirectly as well as directly. If, for instance, in 1935 we see a relative shift in which there are more cosmetics and hygiene product ads than medicinal tonic ads, we can concentrate on specific corporate histories. Why was Brunner Mond Corporation investing in ad space? Do we see a corresponding rise, in production, new product lines, distribution networks, Nationalist government purchasing and did the ad campaign yield more profit? There is no direct route here and the data does not, itself, prove any of these assumptions to be true. However, we do know that ad information helped business operatives to write better understand strategic, targeted, decision making, in profit seeking activity. Third, advertising is a special genre because it's entire *raison d'etre* is profit. Ads are pedagogic and culturally rich, but unlike a poem or a short story, an ad sells a product or service. Finally, the CCAA helps scholars to figure out how the news industry profited individual

owners. John Major (*Shenbao* and *Dagong bao (L'Impartial*), Nakajima Shin (中岛真雄, *Shengjing shibao*), and other foreign owned news media – most Chinese news outlets had foreign founders – bought and sold advertising space in major markets. They were not philanthropists, they were profit seekers and investors. A question that CCAA opens is how the business end of the news industry worked.[1]

CCAA is intended to be a flexible resource. We provide data that will support many queries and projects. In this paper we are presenting three basic concerns. First, the paper shows how Chinese neologisms and advertising images are connected. Second, we present core statistical information about advertising and publication. Third, in response to a traditional humanities question, "what kind of sub-dialects do we see in 20th century newspapers," we are also developing a thesis about the rise of a new Chinese selling language. Exploiting singular algorithms Professor Deng Ke is developing we text mine to clarify the history of modern Chinese commercial language and the new commodity culture.

Thesis One: Graphesis and Society

An immediate question is how modern Chinese language and visual media are modern. We focus on neologisms or calques (new words) to illustrate the integration of social theory and advertising culture. For instance, one of the most important of all the modern words in 20th century Chinese is "society." "Society" is not a descriptor but a category of experience. It has a long career and its novelty has been established using traditional history methods. But the word is also a part of everyday life in 20th century commercial culture. This advertisement shows a group of people heading into the Gate of Happiness. "The gate of happiness, the gate to pass through for happiness and the road of good fortune. Happy New Year to the gentleman who loves smoking Chienmen. Happiness and manifold fortune," the copy reads. The crowd walking toward the product is neither a family group, nor gender segregated club, nor a social class or an aristocracy or a so-called identity like "modern girls;" it is a crowd

[1] Paul French, *Carl Crow: A Tough Old China Hand: The Life, Times, and Adventures of an American in Shanghai*, Hong Kong: Hong Kong University Press, 20 ...173

in a society composed of crowds. The graph communicates the new social formation defined, in part by desire for artfully packaged new commodities and

In the second advertising we see a similar kind of graphesis, visual coding that communicates modernism and the modern age.

綠陰深處空氣新
車馳道上速且靈
欲求平穩無他術
銀殼汽油最稱心

Air is so fresh in the deep green space [of the boulevard]
The new gas auto goes down the road quickly and lithely
Without Shell brand APC gas
Finding a smooth ride is not possible

The graphic element shows a public park in a Chinese city, no doubt in the treaty port a foreign concession because it has wide boulevards and a big space for driving your car. The gate advertises Silver Shell APC gas. The implied meaning (the car has already passed through the gate) is that buying this branded gas and motor oil takes you into modern China. Orderly, fresh, unpolluted, spacious and comfortable, the society of the future is just around the corner. CCAA is replete with images just like these. Under the category or metadata of "society" we can calculate the frequency and the breadth of ads that either use neologisms like "society" or depict crowds of people engaged in social activities in the landscape of the new urban society.

Thesis Two: Metadating

Our second problem is how to archive ad images. This thesis does not only address the technical and practical workflow, discussed below, and which involves digitization, annotation and visualization. It touches directly on the core theoretical issue which is the visibility of information. Visibilizing information is the foundational sine qua not for data analysis and DH's potentially revolutionizing interpretation models. This is the problem Professor Lev Manovich has elaborated the concept of "metadating," in his key publication, *Metadata, Mon Amour*. Manovich proposed that, "metadating the image" transforms an old paradigm where

scholars did "studies" or "worked on" images, into a new approach. According to Manovich the new approach 1) invents new systems of image description and categorization; 2) creates new interfaces to image collections, ways that users can exploit and interpret data; 3) offers new kinds of images on a "super-human" scale of visuality. What Manovich terms the struggle between human subject and the larger visuality of images resituates both the relation of visual images to creators and, probably more importantly, how scholars approach a large scale and sometimes unanticipated visual reality. We confront the question: how do computers help us to sort and present graphesis, i.e., the meaning that massifed advertising images generate? In this battle, establishing metadata is most crucial, as many successful projects regarding visual images have established[2].

Manovich's theory of "metadata" opens possibilities at two levels for CCAA. Firstly, metadating emphasizes a relationship between metadata and objective images, in this case. These real or objectively existing images are a foundation. Collectively they distinguish a process of extracting information. The images are no longer just an "archive" in the sense of library science, or simply the target of what in library science is called "annotation." Moreover, archive, noun or verb, is linked to traditional preservation methods, which include ways of isolating groups of materials on the basis that these are original, authentic and consequently should be preserved. This set of associations about originality and value do not work when it comes to ephemera.

CCAA is an archive of ephemera. The ads are neither original nor authentic; they are composite and infinitely repeatable. Moreover, while digital media can appreciate the ephemerality of our media world millions of these materials appears everyday and archiving them has proven impossible. Given the limitations that scholars of contemporary culture confront, they have begun repurposing annotation to mean a technical process. They focus on image metadata annotation, but they leave out context information. The result is that the relation of image to the production end is not considered or included. Consequently many projects that deal exclusively with contemporary data (e.g., Twitter, FB, Snapchat, etc.) are not open to historical analysis. Metadating is more flexible in that practical sense. Metadating provides the spaces where scholarly use can be built into coding. Secondly, metadating conceptually is more about image but also interface image, which will involve the interaction between human and machine. From the perspective of digital humanities, this concept connects

[2] Manovich, Lev. Metadata, 2002. Mon Amour, http://manovich.net/index.php/projects/metadata-mon-amour

the different levels of hierarchy structure of digital project, from database to interface, then to the user reaction. Metadating images is a crucial part of CCAA. As an image-object and a research-oriented online archive, CCAA is obliged to record all relevant information such as ad illustration, brand icon, textual words and syntax, and mapping information like street names and business titles. After a long process of learning and testing, building into the work the authority, generality, consubstantially and consistency of specific metadata, we were able to begin applying customized metadata schema based on the structural standard, Dublin Core, to each digital image of advertisement. The metadata includes descriptive data of image content, contextual information data about the commodity and advertising industry related to the images and bibliographical data of newspaper, technical data about the digital file as well as information about the source of image, its present location, its copyright status and the owning institution. The latter is necessary because commercial firms are copyrighting images and restricting their common use.

The descriptive metadata mainly describes the textual and iconic contents of images. These include the title, keywords, textual content and a description of the entire image which helps researchers who are seeking basic visual and textual data/information to enable access and the ability to search the whole database quickly and efficiently before they click on and download an images. These descriptive data also make it possible to make the image available for text mining and new kinds of analysis because we transform visual icons into the texts and transcribe all word texts included in each images.

We consider these images to be both visual and cultural products, economic and social commodities. The social and economic information of each commodity in the ad is coded under the following categories: brand/name, commodity category, company, agency, addresses of company and agency. Some of them are extracted from the content of images but some are from external resources. For example, the brand/name of commodity is always very easy to recognize but the company or agency is not. We use the codebook to resolve this problem and to insure that the data is consistent. But at same time, we realized that we couldn't just follow a codebook because information changed over time and in the different decades and newspapers.

The contextual information data about the commodity and the advertising industry that produced and disseminated images provide core information about these advertisements.

The bibliographical data of newspaper covers all publishing information including "Volume Number," "Issue Number," "Coverage Spatial," "Page Number," "Newspaper Publisher," "Issued Date," "Printer," and "Editor." These data provide background information and help researchers to find the source newspaper and process statistical analysis for information about frequency of a specific ad published in one or several newspapers in a year or over many years. "Technical data" describes the technical processes of digitization and archiving.

Subject
Ad Title
Ad Full-text
Ad Description
Ad Agency

Descriptive data

Commodity Brand
Commodity Category
Company
Agency
Company Nationality
Company/Agency Address

Contextual info. data

Volume Number
Issue Number
Coverage Spatial
Page Number
Publisher
Issued Date
Press
Chief Editor

Bibliographical data

ID
Source
Copyright
Collection

Technical data

CCAA has not yet developed a successful readable-pattern that will allow machines to recognize all visual information. Technically this capacity is not yet generally available although we anticipate it will arrive eventually. To compensate we have adapted the more traditional method in which we reduce the image description to one or a few verbal labels (called "keywords"). In other words, we use natural languages to serve as our meta-language for describing images. For instance, keywords describe our categories: 1) "iconic figures" includes animal, plant, human and sub-categories like female/male, old/young/middle age people, foreigner/Chinese, child/infant; 2) in natural language we refer to "design style" allowing us to categorize things like package illustration, context illustration and text; 3) under the category we call context information we rank class, nationality of company and geographical locations (headquartered in Delaware, USA, produced in Shanghai, distributed in

the Jiangnan region, etc). Extracting and summarizing keywords in subject also supports the searching function. That means CCAA users have easy access to full-package information and a clue about how to grab the visual elements designed into the ads. They can directly search for human girls or non-Chinese appearing figures. Because large size image file are heavy, some users will have limited access because their Internet speed and operating environments are deficient. Describing categories fully helps these users to gain targeted access at demand. This is our first, necessary step toward developing future Image Annotation Software. Our position is that relations between image and context information is as important as the content description, so will be developing ways of extending concept networks to link the CCAA to dimensions outside image world.

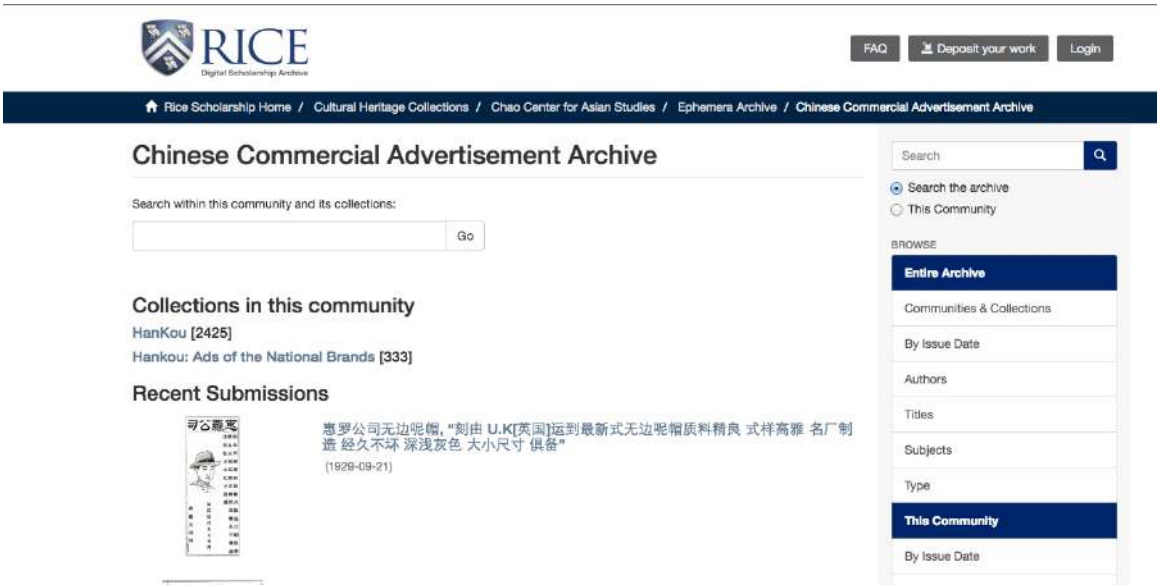
Eventually CCAA will archive more than 192,484 High-quality (>300 dpi) digital images of ads for international brands published in five Chinese newspapers, Shanghai, Shen-pao (申报), Tianjin's Ta Kung Pao (大公报), Shenyang's Sheng-ching Shih-pao (盛京时报), Hankou's Hankow Times (汉口中西报) and the Yuehua Bao (越华报) in Guangzhou, along with research metadata for each image.

NEWSPAPER	NUMBER	Resource	Original Format
Hankow Times	12,961	Shanghai Library; Peking University Library	Microfilm
Yuehua Bao	4,719	Sun Yat-Sen Library of Guangdong Province	Microfilm
Shengjing Shibao	54,804	The National Library of China	Microfilm
Dagong Bao	60,000	Library of University of Washington, Seattle	Microfilm
Shen Bao	60,000	Shanghai Library	Microfilm

Currently we have digitized and annotated about 69,061 ad images selected from three newspapers, including 15,261 advertisements in Hankou Times (1906-1937)， 4,419 advertisements in Yuehua Bao (1931-1938) and 49,381 advertisements in Shengjing Shibao (1906-1938). All images with metadata will be available gradually across two sites: the archiving site (scholarship.rice.edu) hosted and maintained by Fondren Library, Rice University and a dedicated website at the Amazon Elastic Compute Cloud (Amazon EC2), are hosted and maintained by the Luce Foundation funded Ephemera Project. In order to harvest and relocate the metadata from DSpace at Rice to the back-end system of Omeka, programmer

Jin Ying developed a DSpace REST API Harvester to bridge two systems and keep the consistency of metadata across the platforms extra cost of re-typing all image metadata into Omeka.

Because Chen decided to apply the standard metadata schema to images, CCAA is available to install and launch in different systems. For instance, with the help of Prof. Jieh Hsiang, we have also installed a portion of the images and their data set to the Taiwan History Digital Library. This data set can be analyzed using tools developed in the Taiwan History Digital Library.3



Rice University Location under the title "Chinese Commercial Advertisement Archive," scholarship.rice.edu

[3] Prof. Jieh Hsiang, "Taiwan History Digital Library," National Taiwan University tested our data and the results are available at address below :http://thdl.csie.org/HankouTimes_YuahuaBao/RetrieveDocs.php

Prof. Jieh Hsiang, "Taiwan History Digital Library," National Taiwan University tested our data and the results are available at address below

Thesis Three: Concept Modeling: From Metadata to Knowledge

In order to explore the mechanisms of images, texts, commercial products, advertising industry, global capital and modern disciplinary order, we propose a concept model rooted in content analysis and designed to reveal connections among modernist fields where visual culture (text/image), commercial power (commercial products, advertising industry and global capital) and social theory (modern disciplinary order) combine and the everyday of ordinary people as transient ephemera conveying philosophical ideas. We focus primarily on content analysis in order to exploit advertising's structured data. Our aim is to produce more knowledge about advertisements and the advertising industry, and to provide a big picture of advertising (aka distant reading) while challenging traditional research methods.

Our first step was using an R program to help calculate the frequency of some combinations of variables from the data. We listed several combinations of factors, and aimed to build the frequency table for each combination. The table is carefully analyzed. Particularly, we used moving average method to analyze combinations related to time. The combinations of factors concerned here include seven single variables, 14 types of combinations consisting of two variables, and 6 types of combinations consisting of three variables. The data consists

of 69, 062 advertisements.[4] From this here, we get some simple and obvious phenomena characterizing our ads, such as, which years had the most ads in newspaper, which countries had the most ads in each newspaper, which companies had the most ads, which commodities had the most ads, which company had the most ads in the different categories of commodities and which country had the most ads in the different categories of commodities, etc.

Although a complete set of newspapers does not exist, and thus our graphics are incomplete, we still get a general picture of the relationship among the commodity, advertisements, categories, and years. Immediately this small sample yields obvious but not expected outcomes. For example, in these three major commercial newspapers the majority of commodity ads are for Japanese products and the next highest national brand is Canadian. Six of the top 10 commodities measured by regular appearance of advertising in Hankou Times and Yuehua Bao, are from the same company, Dr. Williams brand. Dr. Williams brand products also take up the majority position in Shenjing Shibao. The result of analyzing combination of commodities and category, shows another Dr. Williams product, She-ko (如意膏), appears in the top of ranking in the category of medicine ads. Accordingly, among the 14 countries that advertised in the news media, Canada bought more ad space in the category of medical advertising than any other nation. This result shows how important Dr. Williams brand was historically speaking. In tight focus, although we know from traditional research methods that the bulk of early advertising in Chinese language newspapers consisted of medicines and tonics, Dr. Williams has not been the focus of investigation.

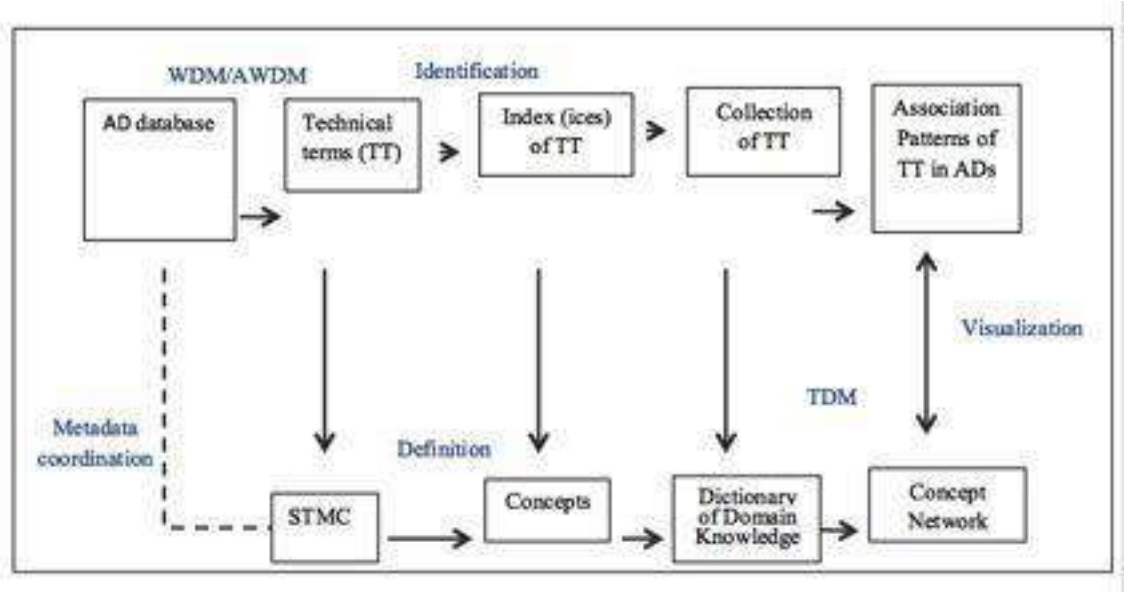
Inspired by the newly discovered importance of Dr. Williams brand, we took a further step to explore texts, and particularly terms used in Dr. Williams ads to promote the commodity line. Prof. Deng Ke's Unsupervised Chinese Text Mining via a Statistical Word Dictionary Model applied statistical methods developed for Chinese text mining and knowledge discovery.5 Deng's approach resolves a number of obstacles in Chinese text mining. For example, word boundaries in Chinese are invisible; worse, ad slogans are not punctuated. Transcribed raw data from archived advertisements are just a sequence of unsegmented Chinese characters, which makes mining in Chinese comparatively tough. Moreover, most

[4] Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.

[5] Deng, K., Geng, Z. and Liu, J. S. (2014), Association pattern discovery via theme dictionary models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76: 319–347. doi: 10.1111/rssb.12032

Chinese text mining methods depend on high quality training data, and will fail if the target texts are remarkably different from the training data. However, considering that the advertisements that interest us are from regional newspapers published over a long period of time, even ad-writing styles are uncertain due to local linguistic differences. We cannot rely on current training data employing modern Chinese to establish models for mining 1920s syntax, vocabulary, punctuation (or lack of it) word use, semantic references, ideograph variation for 100-year-old print media. The third obstacle is that these ad texts contain lots of unstable, idiosyncratic technical terms, like company names written in different ways, differently transliterated brand and product names and so only visible as an effect of text mining processes. This makes it difficult to distinguish technical terms, our true interest, from noise or background words.

Our method is based on a statistical model termed the "word dictionary model" (WDM). Although the WDM is not new, effective and scalable methods for analyzing Chinese texts based on it have not been done. This is likely because of two key challenges: the initiation of the unknown dictionary and the final selection of the inferred words. The Word Dictionary Model (WDM) and Advanced Word Dictionary Model (AWMD) are ideal tools for word discovery, text segmentation and named entity recognition of Chinese texts when training data are not available. WDM can be extended into an AWDM to achieve automatic recognition of TT (i.e., distinguishing technical terms from background words/phrases). The Theme Dictionary Model (TDM) is efficient for detecting association patterns, and the Concept Network (CN) is powerful for knowledge presentation and discovery.



In our case, we applied this method to CCAA's full-texts, 612, 479 characters, 2, 841 unique items of each style of 19,681 items of ads and we found 53, 015 segmented words in 20 minutes. SWDM uses a statistical model selection strategy to score each inferred word, giving rise to a natural ranking and the final selection of the words, so it helps us to select the "meaningful" words and then to make a dictionary. The extension from WD to AWD hasn't been automatically realized so the recognition of technical terms mainly depends on human vision and handwork. However this process is necessary for establishing a TT collection and making the TT index. We selected 37,207 terms including noun, adjective, verb, adverb and quantifier firstly from 53, 015 segmented words to establish the indices of TT for ads. Further, in order to know what kind of terms ads used to promote the product to users in the branding companion and which group of words generally used in the specific product advertising, we will need to dig into the full texts of ads and find out the association patterns among technical terms (APTT) of ads. We are currently working on three newspapers and one company case, Dr. Williams.

The Case of Dr. Williams Brand Products Campaigns

Obviously it is now possible to do statistically driven case studies. One important starting point is to gather data. Canadian Dr. William Frederick Jackson created a tonic that he sold to fellow Canadian, Senator George Taylor Fulford, a chemist and politician. Fulford was by all accounts an advertising genius. He developed the Dr. Williams brand product campaign. The Dr. Williams brand entered the scene in 1866 when Fulford Trading Company began advertising its product; a tonic against tuberculosis. Fulford patented it for general use following the 1891-2 influenza pandemic and marketed his quack medicine in eighty-two countries around the world. This much is either known or suspected. But historical concern about why Dr. Williams advertising is significant or noteworthy immediately runs into interpretive question. As Paul Pickowizc, Kuiyi Shen and Yingjing Zhang have noted in their study, Liangyou, Kaleidoscopic Modernity and the Shanghai Global Metropolis, 1926-1945 (Brill, 2013), the Dr. Williams campaign might be a jumping off point for understanding why elites in modern China bought quack meds like Pink Pills, Pinkettes and She-Ko. The Young Companion ad campaign may, they argue, represent elite attitudes toward the modern body. There are, we suggest other routes to grasping why Dr. Willams brand sold well and is historically significant. Although our routes also contribute to historical understanding they

rely less on traditional interpretative methods because metadating (not presumptions about perception or "identity") can ask different questions, address different historical strata and reconstruct another picture, altogether.6

This question opens our investigation. How much money the Dr. William's Corporation spend advertising its product line and which markets did it dominate? Since Japan colonized Northeast China ("Manchuria"), why did Dr. Williams buy so much ad space across the full spectrum of the news media including Northeast China, with its Japanese dominated media? We know that United States based transnationals also advertised in Northeast China until Pearl Harbor attack in 1941. But the question remains open. In a case like corporate imperialismin China, were so-called multinationals like Dr Williams, Goodyear Tires and so on, also transnational advertisers? To what degree did national transnational corporations compete with one another? What impact on global or transnational corporations do patterns of advertising disclose? Here many plausible reasons can be considered: George Taylor Fulford was just a better ad entrepreneur than Nakayama Ichiro; blanket campaigning proved themselves so profitable Dr. Williams invested heavily in them; Dr. William's singularly attaractive, story-image format ad proved particularly enticing; Dr. Williams' sold many products under its brand line; maybe (not likely) the product sactually worked against tuberculosis and eczema; relatively low prices meant consumers chose a less effective medicine over national or other more expensive national internationals; or Dr. Williams skillfully targeted small domestic commodity users in all new sub-marketing systems.

What does our data do that conventional methods cannot? First, rather than focusing in on one magazine or newspaper, CCAA shows that advertising campaigns were directed across regional media and they involved planning to accomplish media saturation. Data mining opens the capacity for comparative study of a journal like Young Companion and its tiny targeted readership. Very likely the same readers also consumed mass newspapers, but with the larger DB we can let go of reception questions and generalize about the broad scope of all opinion and life-style publications. Mass marketing is constitutive in the sense that it assimilates ideas, words, concepts and categories and disseminates these across vast cultural

---

[6] According to Wellcome researcher Julia Nurse (http://blog.wellcomelibrary.org/2015/03/dr-williams-pink-pills-for-pale-people/) the branded medicine consisted of iron oxide, magnesium sulfate or iron and liquorish and sugar. For Paul Pickowizc et al eds, book study see http://ebookcentral.proquest.com.ezproxy.rice.edu/lib/rice/reader.action?docID=4003978#

spaces. Our three markets (soon to become five) are distinctive but each shows excessive branding of specific international brands. Vernacular sociology is the term for advertising language and it transcends specific journals or newspapers. Niche markets and mass markets can be brought into a common project because CCAA assesses, reconstructs, recognizes and tracks mass markets, mass commodification, and the new commercial mass languages.

Second, specific campaigns and Dr. Williams is the example here, are ideologically complex. As argued shortly, concept mining helps exploit ideologies at the linguistic or subconscious level. But much can be gleaned comparatively when the metadated cartoon images are analyzed. In this capacity, questions of sexual difference, social role, social reproduction, health and illness, nationalism, scienticity and many other factors can be assessed. Dr. Williams' advertisings are charming because each has a story and a picture of a sufferer. This format resembles the social survey in contemporary sociology (see Barlow, Event) and is one way to exploit the information is to remind historians that generic mediation separates the reader or audience from products and allegations about what commodities meant to people. The ability to count is decisive. Writing an anatomy of the campaigns over many decades suggests how sociological categories diffused into mass readerships. Because tuberculosis was a common and deadly disease during these years the structure of the campaigns also carry the possibility for studying attitudes towards modern medicine and eugenic health.

Third, Dr. Williams' is only one of hundreds, perhaps thousands of external multinational corporate brands. Each has distinctive advertising. All of them have corporate histories and have left data behind that make possible measuring profit margins, if not public attitudes toward modern commodity use. Amassing many case studies will open up generalizations about mass mediation and modernity that until DH became possible lay far out our reach as scholars. Here we are not interested in how people received the media (reception theory) or what we can glean about their possible libidinal introjection of new norms (theory of emotions). We are not using the metadata to speculate about perception. Rather we can open a line of questioning about the mechanics of advertising and consequently the focused effort of corporate capitalists and national international corporations to shape the modern world.

So, according to the frequency of some combinations of variables from ads in HKTS and YHB, we find out the following interesting results about Dr. Williams Medicine Company and its commodity. Among all medical companies, Dr. Williams had the most ads, 3, 476 items and She-Ko, the skin cure commodity, most ads, 962, among all 7 Dr. Williams commodities

and not the one best-recognized and assumed to be the premier commodity, Pink Pills For Pale People (韦廉士大医生红色补丸). Taking a close look at the data from 1934 in HKTS & YHB, the most complete data, we found 2, 502 medicine ads, 1,029 ads for Dr. Williams Pink Pills, and 4,08 ads for She-Ko, which means She-Ko had a ratio of 16.3% and 39.65% to general medical and Dr. Williams ads respectively. Meanwhile, another result indicates that Dr. Williams had a clear strategy for advertising since all its commodities occupied the lower-right position of the seventh page of every newspaper. Searched with a combination of "Nationality" and "Page" shows similar results. In the SCSP collection, Dr. Willams' Pink Pills For Pale People (韦廉士大医生红色补丸) has more medicine ads than any other branded product and also appeared regularly on the eighth page's middle right, the lower right and the lower middle. Japanese companies bought space for 19, 545 medicine ads, which is far more in terms of national branding than Canada (6, 245), but none of these 80 Japanese companies competed with the mega-company, Canada's only true success story, Dr. Williams Medicine Company. All of these simple statistical results suggest an image of Dr. Williams Medicine Company, as an ambitious global company with a clear, consistent marketing strategy, strong financial support and mature executive decision making skill. That said, the question is how Dr. Williams Medicine Company packaged its commodities and how it developed branding ideas and an attractive vocabulary to promote their products and ideas to ordinary people? How can we develop this question sufficiently to extrapolate answers to it it?

What we have done is dig into the full-texts of 6 commodities of Dr. Williams to see how the company used words, phrases and ideas. As mentioned earlier, we selected 53, 015 terms including noun, adjective, verb, adverb and quantifier from the segmented words of texts of three newspapers to establish the indices of TT for ads. After the 1st round of result, the noise became obvious mostly because specific terms, like company names, address name and prices clotted the data. We narrowed the number of terms down to 13, 665 and reapplied this TT index to the full texts of 6 selected Dr. Williams products. We found some result related to the APTT of Dr. Williams with the setting (length<5, frequency>0.01). For example, there are 347 groups of the combination of 2-4 words that have been found out, which have the high score of

frequency among 1,832 groups of words. These groups include 121 words that could be assumed as the core technical terms of ads of Dr. Williams.

35	補血	9	濕疹	5	小女	3	療治
32	便秘	9	瘋濕骨痛	5	康健	3	皮膚諸恙
31	補血健腦	9	皮膚	5	氣管	3	肝經失調
27	消化	9	虛弱	5	灼傷	2	不調
24	馳名	9	風濕骨痛	5	衛生小書	2	價銀
23	大便	8	便閉	4	世界馳名	2	兒女
20	天下馳名	8	傷風	4	口臭	2	呼吸器
18	濕骨痛	8	大便秘結	4	微利	2	失眠
18	血虧	8	服用	4	性和	2	寒熱
15	治愈	8	血液	4	恢復	2	山嵐瘴癘
15	腸胃	7	不取	4	止痛	2	強壯
15	血薄氣衰	7	早老	4	痰厥	2	感冒
15	試服	7	未老先衰	4	皮膚諸	2	感激
14	健康	7	芬芳	4	絞痛	2	房事無能
14	各症	7	血虧腦弱	4	肚痛	2	數劑
14	小書	6	內臟	4	諸虛百損	2	潤腸
14	胃不消化	6	大便閉結	3	來書	2	無力
14	頭痛	6	強健	3	傷風痰厥	2	爽適
13	婦科各症	6	操勞過度	3	全愈	2	牙痛
13	治療	6	疼痛	3	口氣穢濁	2	病菌
12	蛔蟲	6	腦筋衰殘	3	味美	2	痛苦
12	體力	6	腹瀉	3	喉痛	2	發生
11	出牙	6	跌損	3	山林瘴氣	2	神經
11	發熱	6	軟弱	3	康強	2	耐心
11	統治	6	面疹	3	敝局	2	肝火
11	郵奉	5	中國各處	3	新血	2	胃呆
10	復原	5	之功	3	有力	2	自述
10	泄瀉	5	口氣	3	有序	2	舌苔
10	消化不良	5	嘔吐	3	活潑	2	體虛
10	購買	5	嬰兒	3	男女老少		
10	食積			3	痛瘋		

At that point we returned to the full text of ads to see if what the Associate Pattern of Technical Terms had found was generalizable. For example, the combination of “妄語” and “不智” appears many times in the same sentence : “以身體健康之重要而欲節省數毫洋殊為不智故凡遇經售家以影射紅丸兜售稱其價錢而效相同者妄語也慎勿置信.” So we could assume this group is an associate pattern of technical terms and that discovery could help us to find fixed phrases (expression pattern or writing style) in other ad texts. As we refine the APTT and bring into the data considered all of the five newspapers we expect to reveal unconscious associations and new language forms. While these conclusions will not prove anything about elite attitudes towards the modern body, they do imply that arguments about body image are relatively difficult to evidence. These are speculative ideas about presumed opinion and as

with many reception theory based arguments, they are neither demonstrably true nor demonstrably untrue. They represent the speculative opinions of highly educated and motivated historians who hail from an era a century after all the tiny decisions about advertising images and languages were made.

Preliminary Conclusions

CCAA works from a singular archive consisting of metadated, profit-oriented, internationally design mature, transnationalized advertisings extracted out of five commercial treaty port newspapers in Chinese cities during the 1920s and 1930s. We are aware that historians and cultural studies scholars are beginning to consider how these advertisements can be interpreted, what they mean. Our case study and our general argument are only one small report in what perforce will be a multi-year project. But this report does promise access to a sub-conscious language and a vast unknown reservoir of visual and literary associations. The associations and the metadated images are overwhelming evidence of change. Modernity is a language and it is a way of life; it is a visual order and internationalized nationalist visions. That is what our project can do, Now that we are celebrating the publication of our first data sets we ask you to join us in developing ways of approaching data with fresh research questions.

---

i Manovich, Lev. Metadata, 2002. Mon Amour, http://manovich.net/index.php/projects/metadata-mon-amour and Marie Claire Bergere, *The Golden Age of the Chinese Bourgeois, 1911-1937* (Cambridge: Cambridge University Press, 1989)

ii Deng, K., Geng, Z. and Liu, J. S. (2014), Association pattern discovery via theme dictionary models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76: 319–347. doi: 10.1111/rssb.12032

詩經的量化研究：發掘興體詩的隱藏節奏

廖學盈*

摘　要

本研究以量化分析重新探勘詩經的語言風格，作為詩經興體詩的實驗性釋讀基礎。分析文本採用阮本《毛詩正義》全數經文及其章句分段，作為釋讀方法的參照母本。興體通常由兩組詩句並排而成，詩句之間意義關聯模糊。第一組「起句」多引景物，第二組「應句」多題人事。以毛詩正義而言，毛公所傳經文，獨舉「興也」於起應之間；除卻毛傳，難辨興體。為了確立句組的意義關聯，歷史注釋常倚賴類比、因果、時序等途徑進行宏觀的解說。近代始有人注意押韻、句法、套語、虛詞、疊音等構成的微觀結構。興體的判準歷來眾說紛紜，卻一直被奉為中國文學抒情傳統的核心。本研究即以數位工具進行計量分析，單就全本經文內部產生的數據，捕捉選詞用字的節奏型式，重新描述興體詩的語言特徵。

關鍵字：數位人文、計量分析、節奏分析、詩經、興體

* 法國克萊蒙費朗學區漢語教授，Email: shuehyingliao@gmail.com。

A Quantitative Research of the *Book of Odes* (*Shijing*)： the Discovery of the Underlying Rhythm in the Incentive Process

Shueh-ying Liao*

Abstract

In the *Book of Odes*, how to describe the rhythmic pattern has been a question. In order to answer the question, this study presents an algorithm in order to extract the rhythm from the text at the line (*ju* 句) level and interpret the incentive process (*xing* 興)：1. Note down the frequency of each Chinese character in each position in each line, 2. For each line, the variation of frequency of each Chinese character can be interpreted as literary rhythm, 3. Compare each line by the defined literary rhythm which can be identical, complementary or symmetrical, 4. Such comparison is expected to explain the feeling of a reader about a textual unity, especially between lines produced with the incentive process. This work proposes a new way of reading and allows evaluation of style distance between poetic texts.

Keywords: digital humanities, quantitative analysis, rhythmic analysis, *Book of Odes* (*Shijing*), incentive process (*xing* 興)

* Professor of Chinese Language Académie de Clermont-Ferrand, France. Email: shuehyingliao@gmail.com.

一、研究概述

本研究以量化分析重新探勘詩經的語言風格，歸結一套「節律修辭 (rhythmic figure)」，作為詩經興體詩的實驗性釋讀基礎[1]。分析文本採用《毛詩正義》完整經文及其章句分段，作為實驗材料[2]。

《毛詩》是西漢毛公一派傳授的三百零五首周朝詩歌，即本研究使用的「經文」，原為地方性傳本[3]。毛公總結前人意見、講解詩句的紀錄稱作「傳文」，亦即《毛傳》。東漢鄭玄循「傳」釋「經」作「箋」(猶如筆記備忘便條)，是謂《鄭箋》。唐代孔穎達疏通歷代經解，以《鄭箋》為宗，完成「注疏」，稱作《毛詩正義》。朝廷將之視為正統，立於學官，教習背誦。

「興體」則是一種文學形式，通常由兩組詩句並排而成，詩句之間意義關聯模糊。第一組「起句」多引景物，第二組「應句」多題人事[4]。「毛傳」標示興體，以「興也」兩字點明之後，開始講解「起句(景物)」，啟發學生聯想「應句(人事)」的線索。今日，經、傳、箋、疏合一的通行文本，已經直接將「毛傳」對「興體」的解釋附在「起句」之後。文本上來看，「毛傳」是在「起句」和「應句」的接合處，辨識出「興體」：一種需要倚賴感覺和想像才能完成的間接類比。然而，除卻「毛傳」，單憑經文，難辨「興體」何在。宋代朱熹倡導諷誦涵泳，純粹讀經，不假傳箋注疏，認為讀者自然能掌握文意。其結果顯示，朱熹「詩經集傳」標註了 112 處「興體」，比對「毛詩正義」所錄的 116 處「興體」，只有 75 處相符。這說明「興體」的「起句」和「應句」之間，意義是否隔閡，關聯是否直接，因人而異。傳統注疏常倚賴類比、因果、時序等途徑對讀者進行「起句」和「應句」之間的想像引導。近代始有人注意押韻、句法、套語、虛詞、疊音等構成的微觀結構[5]，「起句」和「應句」似乎能藉由音響節律串連起來。

---

[1] 本研究處理「興體詩」遺留在文本上的具體結構方式(process)，因此不從意象(image)或動機(motive)較為抽象的層次來分析「起句」和「應句」之間的關聯。

[2] 李學勤，2001，《十三經注疏整理本：毛詩正義》，台北：台灣古籍出版社。

[3] 鄭玄，《詩譜》：「魯人大毛公為訓詁，傳於其家，河間獻王得而獻之，以小毛公為博士。」

[4] 「興體」在境外漢學著作中常見的翻譯和理解如下：[法] Marcel GRANET (1884-1940)直接翻作「比喻(comparaison)」(見« *Fêtes et Chansons Anciennes de la Chine* »，23 頁)；[法] Jean-Pierre DIENY (1927-2014)在古詩十九首的譯本中曾認為，這種結構旨在營造一種「象徵的意義 (sens symbolique)」(見« *Les Dix-Neuf Poèmes Anciens* »，62 頁)；[法] François JULLIEN, (1951-)則強調這是起心動念(motivation)間「物我自然交感激發」(incitation spontanée)的一系列過程(« *La Valeur Allusive* » ，73 頁)；[法] François MARTIN (1948-2015)逝世前最後以「啟發性的文學手法(procédé incitatif)」來理解興體(筆者博士論文法文標題中的「興體」即接受評委 François MARTIN 要求定為 « *L'usage de la figure rythmique dans l'analyse du procédé incitatif* »，2015 年)；最後，[美] Stephan OWEN (1946-)則以「情感意象「(affective image)」強調「起句」景物的抒情特質。

[5] 例如王靖獻著名的論文《鐘與鼓 ： 詩經的套語及其創作方式》(葉珊和謝謙翻譯) (1990。成都：四川人民出版社。)或德國漢學家 Ulrike MIDDENDORF 的文章〈詩經之微指——以心理語言學理論分析《木瓜》、《東門》〉(陳致主編。2010。《跨學科視野下的詩經研究》。上海：上海古籍出版社。212－235

「興體詩」因而「意味深長」卻「意思模糊」，一直以來被奉為中國文學抒情傳統的核心[6]。本研究即以數位工具進行文本探勘，倚賴電腦對每個詩句進行「組合分析」，列出所有隱藏的句式，隨後反查經文中所有符合隱藏句式的詩句，進行「頻率計算」。這些步驟企圖模擬讀者「諷誦涵泳」所有詩句後，對全部經文「隱藏句式組合」和「字詞分布頻率」的終極認識。實驗初步發現，我們可以用「詩句內的常用字詞與位置」歸納出「高頻用詞模式」，進而定義所謂的「節奏」。最後，將此「節奏」回用於分析興體詩句，發掘出「詩句的強調點」、「起句、應句間的節奏關係」以及「新的意義解讀層次」。

二、節奏的來源 ： 潛藏的句式

本文主張，興體可以藉由「句(line)」層次選詞用字的排列節奏來探勘。以全本經文作為「對照組(母本)」，計算每個字符(音或字)在每個句中每個位置「個別」和「共同」出現的頻率後，每個字符在每種組合方式下可以被標註為「較常見(+)」或「較罕見(-)」。這種二元變換的組合型式，在此稱作「文學性的節奏」。全面分析「對照組」之後，設定每首興體詩為「實驗組（子本）」，根據閱讀所及範圍，逐句、逐段或逐章配對相同(identical)、互補(complementary)或對稱(symmetrical)的型式，建立閱讀範圍內句和句之間的節奏關聯。文本處理步驟主要如下：

(一)組合分析和頻率計算

計算對照組(母本)全數經文中每個字符(音或字)在每個句中每個位置「個別」和「共同」出現的頻率後，每個字符在每種組合方式下被標註為「較常見(+)」或「較罕見(−) 」。以詩經首篇關雎首章次句「在河之洲」為例：

表 1. “在河之洲” 潛藏的句式組合分析和頻率計算

階層	型式	標籤	數量	頻率	附註
0	在河之洲	+ + + +	1	0.01%	1. 「〇」 表示選擇字符選擇軸可能的位置。在本研究中被考慮為主題成分的可能來源。 2. 「字符」表示組合軸的主要構成
1	〇河之洲	− + + +	1	0.01%	
	在〇之洲	+ − + +	1	0.01%	

頁。)。此外，池昌海的《先秦儒家修辭要論》(2012。北京：中華書局。)也羅列了更早期顧頡剛的「協韻說」(1982。《古史辨 III》。上海：上海古籍出版社。672-690 頁。)和朱自清的「套語說」(2004。《詩言志辨》。廣西：廣西師範大學出版社。39 頁。)。

[6] 徐復觀。1974。〈釋詩的比興:重新奠定中國詩的欣賞基礎〉。1996。《中國文學論集》。台北：學生書局。91-117頁。

	在河〇洲	++−+	1	0.01%	成分。在本研究中被考慮為閱讀時節奏成分的可能來源。 3. 同樣階層的組合條件下(亦即字符選擇軸「〇」的數量)，一個相對高頻率穩固的潛藏句式將被考慮為該階層代表性的節奏型式 (表中灰階表示的部份)。 4. 這裡提出的是一種理想模型，也就是文人諷誦涵泳後理論上的終極結果。實際上，每個讀者對文本熟悉程度不同，節奏型式的選擇也會出現差異。 5. 「〇〇〇〇」 即詩經所有四字句。
	在河之〇	+++−	4	0.05%	
2	〇〇之洲	−−++	1	0.01%	
	〇河〇洲	−+−+	1	0.01%	
	在〇〇洲	+−− +	1	0.01%	
	〇河之〇	− ++−	7	0.10%	
	在〇之〇	+−+−	18	0.25%	
	在河〇〇	++−−	4	0.05%	
3	〇〇〇洲	−−−+	2	0.03%	
	〇〇之〇	−−+−	353	4.84%	
	〇河〇〇	−+−−	8	0.11%	
	在〇〇〇	+−−−	48	0.66%	
4	〇〇〇〇	−−−−	6575	90.20%	

(二) 選擇每個階層下代表性的節奏型式

表 2. “在河之洲”潛藏的代表性節奏型式

階層	型式	標籤	數量	頻率	可能組成主題的元件
0	在河之洲	+ + + +	1	0,01%	ø
1	在河之〇	+++−	4	0,05%	洲
2	在〇之〇	+−+−	18	0,25%	河, 洲
3	〇〇之〇	−−+−	353	4,84%	在, 河, 洲
4	〇〇〇〇	−−−−	6575	90,2%	在, 河, 之, 洲

普通詞素和特殊詞素二元變換的組合型式，即本文所謂文學性節奏的來源。例如，讀者遇到「在河之洲」時，若在第一階層辨識出「在河之〇」為常見的結構 (詩經中出現 4 次)，句末「〇」位置上的「洲」將顯得突出而有特殊意義。常見的結構容易誦讀，可形成主導的節奏。但是，隨著讀者辨識選擇軸數量的變化，閱讀節奏也會變化。例如，察覺「在河之洲」有兩個選擇軸的讀者，理論上會將第二階層最常見的「在〇之〇」選作節奏型式。此時，位在第二個和第四個位置上的「河」和「洲」將會顯得特殊，可作為描述詩作主題的積極成分。

(三) 聯結各階層可辨識的高頻節奏型式

取任何興體詩為實驗組(子本)，限定反復誦讀的範圍，逐句、逐段或逐章配對限定範圍內的節奏型式。以下為逐句配對同章節同階層相同型式的節奏所形成的聯結範例：

表 3-1. 聯結相同的節奏型態範例

編號	標題	起句		興	應句		階層
272	黍苗	芃芃黍苗	陰雨膏之		悠悠南行	召伯勞之	2
		芃芃〇〇	陰雨〇〇		悠悠〇〇	召伯〇〇	

表 3-2. 聯結互補的節奏型態範例

編號	標題	起句		興	應句		階層
200	巷伯	萋兮斐兮	成是貝錦		彼譖人者	亦已大甚	2
		〇兮〇兮	成是〇〇		彼〇人〇	〇〇大甚	

表 3-3. 聯結對稱的節奏型態範例 (對稱+互補)

編號	標題	起句		興	應句				階層
181	鴻雁	鴻雁于飛	肅肅其羽		之子于征	劬勞于野	爰及矜人	哀此鰥寡	2
		〇〇于飛	肅肅〇〇		之子〇〇	〇〇于野	爰及〇〇	哀此〇〇	

(四) 文本的統一性

聯結高頻節奏型式的手法，試圖探源讀者隱約感覺的文本統一性。上述範例在平行結構(parallelism)的研究中司空見慣，但是，本研究強調，每一種結構的識別和聯結都需要考量文本自身的語言慣性。根據定義的探勘條件，計量分析提供完整的文本觀察記錄，藉由各種要素的出現頻率和分布位置來描述語言風格的特徵。透過反查，我們還可以知道實驗組的興體詩中，各種高頻組合型式，在對照組的全本經文中喚起的一系列意義網絡。例如詩經中所有「在〇之〇」型式的句子，第二個字若不是地點就是河流 (「周」、「坰」、「浚」是地點；「水」、「河」、「洽」 、「渭」 是河流。) ；第四個字則必然指出該地點或該河流的某個區域(「下」、「將」、「庭」、「洲」、「涘」、 「湄」、 「滸」、 「潯」、「郊」、 「都」、 「野」、 「陽」都是表示位置的詞彙)。這樣的調查容許我們假設，在某個詩歌創作傳統底下，「人」和「事」的詞彙與「在〇之〇」的型式不相容。也就是說，一個即興創作的樂師，一旦選用了「在〇之〇」發展詩句，就會順勢唱出地方或水域。而一個熟習詩經的讀者，一旦選擇了「在〇之〇」頌讀詩句，則有機會「自然」想起地方或水域：

表 4. 以「在〇之〇」節奏型式閱讀「在河之洲」時喚起的一系列意義網絡

句	在	河	之	洲	數量	出處	興體詩
型式	在	〇	之	〇	18	詩經	9/18
系列	在	坰	之	野	4	297 駉	否
		浚		都	2	053 干旄	是
		渭		涘	1	236 大明	否
		河		滸	1	071 葛藟	是
		水		涘	1	129 蒹葭	是
		河		洲	1	001 關雎	是
		浚		郊	1	053 干旄	否
		河		涘	1	071 葛藟	是
		水		湄	1	129 蒹葭	是
		浚		下	1	032 凱風	是
		洽		陽	1	236 大明	否
		河		滸	1	071 葛藟	是
		周		庭	1	280 有瞽	否
		渭		將	1	241 皇矣	否

(五) 可視化潛藏型式的聯結

興體詩「起句」和「應句」之間模糊的聯結，可以藉由上述步驟辨識出來。它可以是簡單的節奏型式聯結，也可以是複雜的意義網絡聯結。這些分析並非定義什麼是「興體」，而是說明「興體」在閱讀中如何成為可能。未來透過分析結果和傳統注疏的對勘，我們可以進一步考察該文體被建構的過程，也能重新審視過去的詮釋。例如詩經首篇「關雎」，首章次句「窈窕淑女」，在「毛傳-鄭箋-孔疏」的詮釋傳統下，其閱讀的節奏型式可能是「窈〇〇〇」(4 次)或「〇窕〇〇」(4 次)，致使解經側重「淑女」。然而，一些讀者在反覆諷誦涵泳中也許看出不一樣的風景：高頻出現的「〇〇淑〇」(7 次)和「〇〇〇女」(35 次)讓處在第 1、2 位置的「窈窕」特別引人注目[7]。

二、節奏和主題的互補關係：詩化的過程

[7] 馬瑞臨，《文獻通考》，卷一百七十八，經籍考五：「夫關雎、鵲巢，閨門之事，后妃夫人之詩也。」

文本探勘過程中發現，高頻節奏型式，其選擇軸「〇」內的字，通常是描述主題的積極成分。詩經詩化語言的程序，也許是透過這種節奏和內容互補的後設語言活動(metalinguistic activities)來實現。已知詩經詩作標題，常取自詩中某句或某句中數個字。為了觀察節奏和內容的互補情況，我們測試，是否組合分析得出的高頻節奏型式，其選擇軸「〇」內的字即為標題。下表羅列所有情況，以一個羅馬字母代表一個漢字，不同羅馬字母代表不同漢字：

表 5. 詩作標題和來源詩句的比對

情況	來源詩句類型	高頻節奏型式	選擇軸「〇」內的字	對應的詩作標題比例	
	階層 4				
1	ABCD	〇〇〇〇	ABCD	22.95%	
	階層 3				
2	ABCD	〇B〇〇, 〇〇C〇,…	CD, AB, … (@*)	18.03%	56.72%
3	AACD, ABCC,…	〇〇C〇, 〇B〇〇,...	AD, AC, … (&*)	1.97%	
	階層 2				
4	ABCD	A〇C〇, AB〇〇,…	BD, CD, …	25.9%	
5	ABCD	？	？	10.82%	
6	AACD, ABCC,…	AA〇〇, 〇〇CC,...	CD, AB,…(&*)	8.85%	
	階層 1, 階層 0				
7	ABCD	！	！	11.48%	

「AA」,「 BB」,「 CC」, 或 「DD」 表示句中出現疊音。
「?」 所有組合型式頻率相同無法比較。
「！」標題選擇不只牽涉來源詩句。
「@」= 加入補充篩選條件，組合分析後，標題會出現在選擇軸較多的一側。
「&」= 人工檢查，標題大多避免使用疊音。

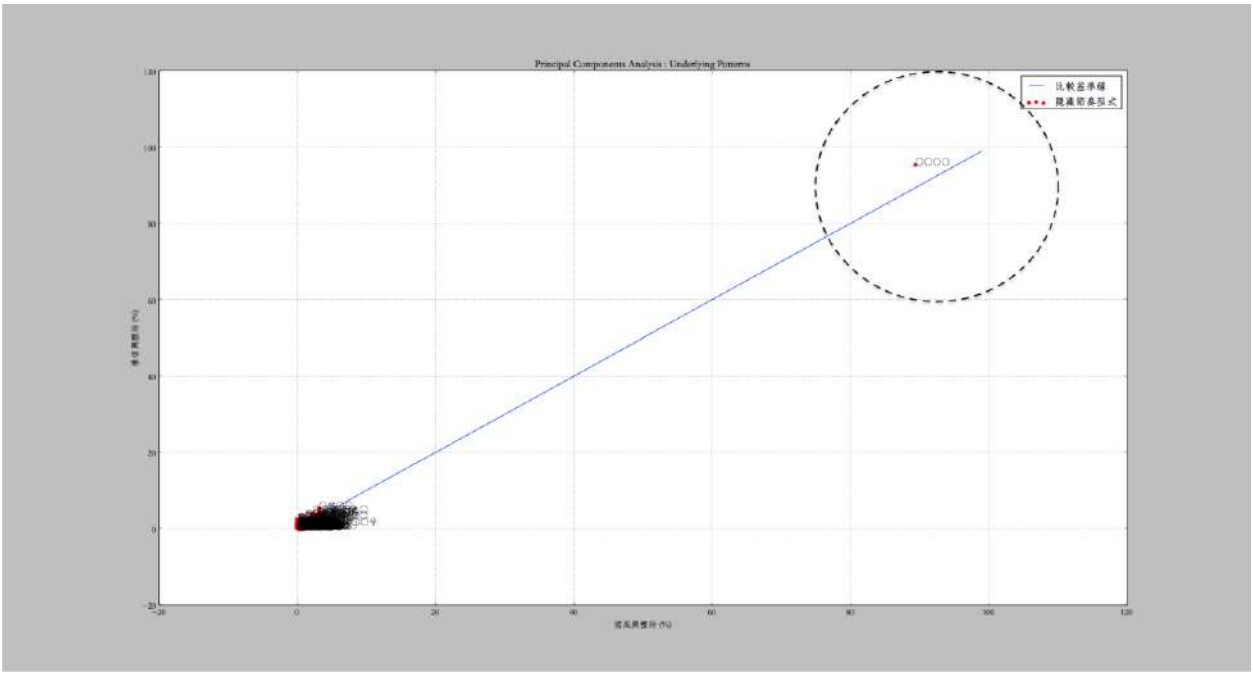
結果說明，作為預測標題而言，組合分析準確度不高(25.9%，見情況 4)。人工輔助或針對文本特性加入補充篩選條件，可以在情況 2、3、4、5 和情況 6 中辨識出正確的關鍵字(比例分別是 56.72%和 8.85%)。情況 1 是指完整詩句被取作標題，每一個位置都是選擇軸，每一個字都被取用。情況 7 是指，標題的選擇方式不明(11.48%)。本測試只考慮字型組合，尚未結合已經得出的字音頻率、字音長短、長音短音在每個句中每個位置出現的頻率[8]。然而，我們或可假設，在句的層次上，節奏字符與主題字符排列的位置多半呈現互補的傾向。

[8] Jacques Vergne 曾做過英、法、德、義語虛詞位置盲測的實驗，在沒有提供任何外部資訊的情況下(例如，沒有事先定義一個已知的虛詞列表)，單憑文本內部的探勘結果，若某詞彙對比前後相鄰的詞彙

三、節奏要素分析 (Principal Components Analysis, PCA)：國風和雅頌的興體詩

某文本高頻率使用的節奏型式(rhythmic pattern)群組，即該文本的節律修辭要素，可以作為語言特徵的客觀判準，運用 PCA 的基本技巧對照文本之間的風格差異。例如，將毛詩正義中 116 首「興體詩」分作國風和雅頌兩組來進行比較。圖例 1 顯示，四字句「〇〇〇〇」雖為兩組共有的高頻節奏要素，但它在雅頌的「興體詩」中更為常見：

圖例 1. 興體詩節奏型式要素比較：四字句型式

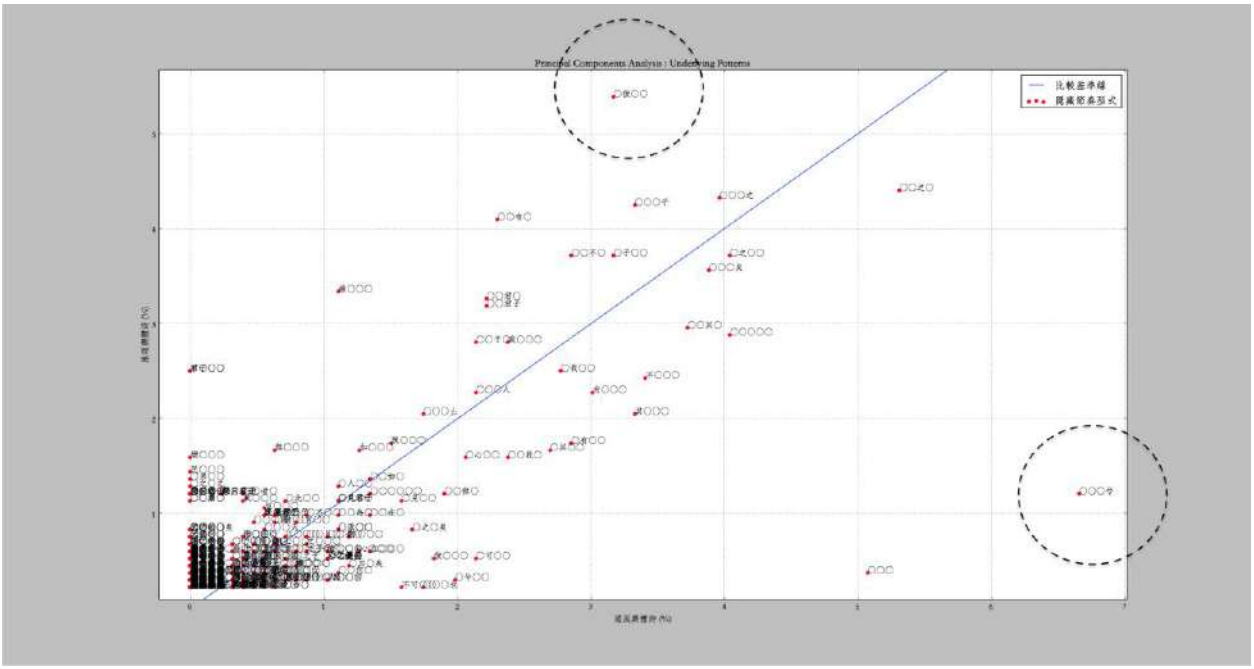


*每個紅點代表一個節奏型式。Y 軸對應該型式在雅頌所佔比例，X 軸對應同樣型式在國風所佔比例。

為了瞭解詳情，我們放大圖例 1 中左下角群聚的節奏型式。局部放大結果，如下面圖例 2 所示：

「詞頻較高」且「音節較短」，該詞彙很可能就是虛詞（詳細計算公式請參閱：〈 Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource 〉 ，收錄在《 *JADT 2004 : 7es Journées internationales d'Analyse statistique des Données Textuelles* 》，第 1157-1164 頁。）。詩經文字節奏的產生，也常憑藉虛詞或各種高頻實詞的「虛用」來達成填補音節、平衡節奏或延續音響。這是筆者探索詩經節律的主要線索。

圖例 2.　興體詩節奏型式要素比較：特色節奏型式



我們發現兩組都使用「〇〇〇兮」型式，但它在國風興體詩中特別突出(近 X 軸最大值，鄰 Y 軸最小值，見圖例右側下方)。或者，國風興體詩常用「〇彼〇〇」型式，只不過，以它來描述雅頌興體詩的風格會更為妥當(近 Y 軸的最大值，居 X 軸的中間值，見圖例中間上方)。

要素分析可以作為文本風格辨異的客觀判準。美國漢學家 RUSK Bruce (2012)曾就詩經接受史的角度，透過「詩學術語」在漢語典藉中的變異和分佈，來描述「世俗詩歌」和「正統經典」的交互作用[9]。全書總共探討五個交互作用的層面，如下：

表 6. 五個「世俗詩歌」和「正統經典」的交互作用層面

章節	標題	內容	頁次	頁數
1	*Poems* and Poems	「詩」作為「作品」的稱呼	15-56	41
2	Re-collections	「詩」作為「專輯」的標題	57-94	38
3	In the image of Classic	「詩」作為「意象」的範本	95-114	20
4	Circulation des Troposphere	「詩」作為「修辭」的判準	115-158	44
5	Inventious Discovery	「詩」作為「偽作」的權威	149-200	52

其中最值得注意的情況，就是第三章作者在分析「詩歌語言」特質時，篇幅明顯較少。這是因為「詩學術語」標籤的變異和分佈不能良好表達「詩歌語言」結構的變異和分佈。類似的問題也出現在池昌海(2012)針對先秦文學所進行的「語音修辭」研究[10]。詩經的部份，研究調查一開始就先預設了一系列「襯音」、「疊音」[11]和「虛

[9] RUSK Bruce, *Critics and Commentators: The Book of Poems as Classic and Literature (Harvard-Yenching Institute Monograph Series)*, Harvard University Asia Center, 2012.
[10] 池昌海，2012 年，先秦儒家修辭要論，北京：中華書局，209-211 頁。

字」的列表，接著根據書卷、篇章、句組、地域、階級、體裁等類別，計算這些「裝飾音響」的數量、分佈和相關性。這項研究調查了「裝飾音響」的使用情況，卻沒有提供相應的詮釋方法。例如，作者分類「疊音」的作用有「襯音」(補足音節)、「摹聲」(模仿聲響)、「肖形」(事物畫面)、「繪色」(顏色字眼)、「狀態」(人物心理)、「品性」(人物特質)六種。前兩種的判別相對容易，沒有太多的問題；但是，後四種作者最終只能屈就傳統注釋，並未在其量化研究的基礎上提供進一步的解釋。為什麼「灼灼」是「狀桃花之鮮」？為什麼「依依」盡「楊柳之貌」？為什麼「杲杲」為「出日之容」？「瀌瀌」擬「雨雪之狀」？這些「疊音」在非物質性文本的傳遞過程，並沒有部件或部首作為意義的提示，僅按照傳統注釋說「灼灼」是「狀桃花之鮮」，不能說明「灼灼」的音響與「桃花的樣貌」有何淵源[12]。

上述兩個研究都試圖描述詩歌語言的風格。或從注釋標籤的使用來觀察，或從音響元件的分佈來估量，皆沒有充分考慮文字在句中的擺放順序和相對位置。然而，這些正是使詩歌有「節」能「奏」的中樞、令文人咬「文」嚼「字」的氣骨。

四、實驗性的釋讀方法：階層式的節奏型式

單純的數量分析很難滿足文學家的需求。這裡試擬一個「量」轉「質」的分析方案。我們對詩句進行組合分析後，可將詩句「恆定」的元件標記為「+」(常見而普遍的元件)，將「可變」的元件標記為「−」(少見而特殊的元件)。例如「在河之洲」在階層 2 的視角下(讀者辨識到兩個可能的選擇軸)，「在〇之〇」為其最穩固的結構，我們可以將「在」和「之」標註為「+」。處在選擇軸「〇」上的特殊元件「河」和「洲」則標註為「−」。整句標註結果為：「+ − + −」。

實驗性的釋讀原則如下：所有恆定的組件是詩作普遍的成分，帶有節奏的特性；所有可變的組件是詩作特殊的成分，帶有主題的特性。

如果將一首「興體詩」中所有的詩句都如此標註，把相同類型的標籤按群組分類，最後依各群組被歸類的先後次序分層，我們觀看作品的方式將產生層次感。以詩經第 257 首「桑柔」第一章的「興體」為例，直讀：「菀彼桑柔，其下侯旬。捋采其劉，瘼此下民。不殄心憂，倉兄填兮。倬彼昊天，寧不我矜。」如果進行組合分析，甚至將組合範圍擴大，兩句一簇的情況下，「菀彼桑柔其下侯旬」潛藏的其中一種恆定型式是「〇彼〇〇其〇〇〇」(有 6 個選擇軸，因此是階層 6 的閱讀角度)，應標記為「− + − − + − − −」。同理，所有的詩句都兩兩一簇進行同階層的組合分析，並選出每簇最高頻率的恆定組合。

---

[11] 「襯音」和「疊音」即一般所謂「襯字」和「疊字」。池昌海認為這些元素很多沒有明確意義，建議純粹以音響來看待。

[12] 今人構擬的古音也只是一種理論上的推測。在 Baxter-Sagart 的古音重建工程中，自 1.0 版(2011)到 1.1 版(2014)均未見「灼」、「杲」、「瀌」等字的擬音。詳見：BAXTER William H. & SAGART Laurent, *Old Chinese: A New Reconstruction*, Oxford University Press, New York, 2014.

最後，根據標籤異同對句子進行群組分類：順著直讀的排序，由首句起，第 1、4、7 句最先配對，納入第一層；第 2 句尋無匹配，孤立起來；其次第 3、8 句完成配對，向右納入第二層；最後，第 5、6 句配對，再向右納入底層。所謂分層，即是考慮同樣型式的節奏，可藉由反復閱讀，在詩章內形成隱約的復沓感。於是，「桑柔」首章離析出三個節奏復沓層和一個孤立節奏層：

表 7. 興體詩「桑柔」的實驗性釋讀：復沓層次分類群組

詩句順序	孤立的節奏層次	第一復沓層次	第二復沓層次	第三復沓層次
1		菀彼桑柔 (−+−−)		
2	其下侯旬 (+−−−)			
3			捋采其劉 (−−+−)	
4		瘼此下民 (−+−−)		
興				
5				不殄心憂 (−−−+)
6				倉兄填兮 (−−−+)
7		倬彼昊天 (−+−−)		
8			寧不我矜 (−−+−)	

此時，孤立的「其下侯旬 (+ − − −)」顯然突出；節奏相對的「不殄心憂 (− − − +)」和「倉兄填兮 (− − − +)」遠在底層。前者追憶一度的安逸，後者復沓無盡的悲憤。任何讀者會察覺如此微妙的舖設嗎？高本漢的英文譯本似乎反應了這個深層結構。譯文道出：第二句是一個意義非凡的時刻 (even 帶有 kairos 含特殊意味的時間)，而五、六兩句卻是持續惡化的分分秒秒 (unceasing 和 long-continued 則帶有 chronos 無差別意味的時間)：[13]

表 8. 興體詩「桑柔」的實驗性釋讀：高本漢英文譯本對應析離出的復沓層次

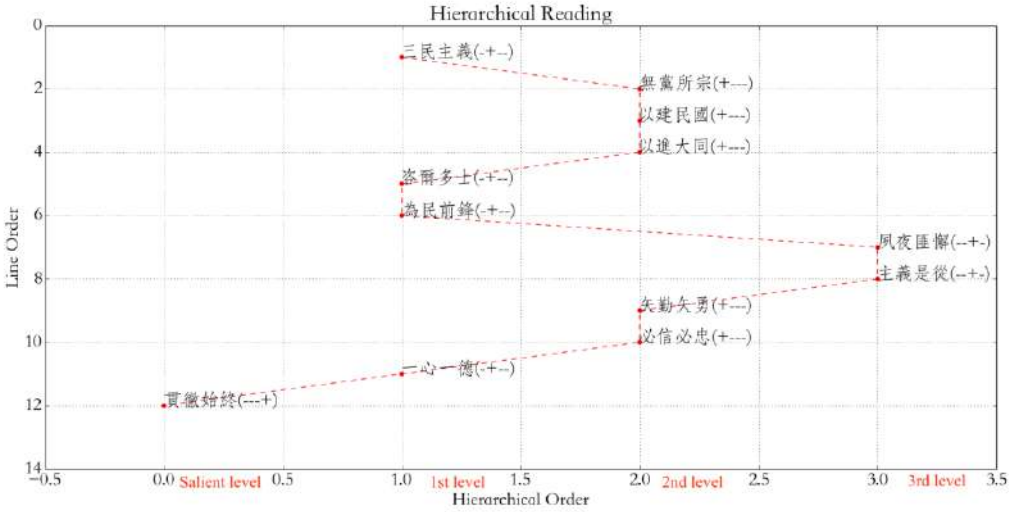
句	孤立的節奏層次	第一復沓層次	第二復沓層次	第三復沓層次
1		*Luxuriant is that softness of the mulberry tree,*		
2	*Beneath it, it is <u>even</u> (shade everywhere);*			
3			*But if one plucks, it will be destroyed;*	
4		*Suffering is this lower people.*		
	興			
5				*<u>Unceasing</u> is the grief of the heart;*
6				*The affliction and distress are <u>long-continued</u>;*

[13] Bernhard KARLGREN, *The Book of Odes: Chinese text, transcription and translation*, The Museum of Far Eastern Antiquities, Stockholm, 1974 年, 第 220 頁。

7		*Grand is that great Heaven,*
8		*Why does it not have pity on us?*

作為想像的實驗，我們還可以拿「三民主義歌」的歌詞來作測試。一般認為，這首歌的歌詞模仿了古雅的四言詩體製，且使用了詩經常見的「咨爾(女)」這樣一個遠古的祈使句型。取「詩經」作為對照組(母本)，以「三民主義」作為實驗組(子本)，定「階層 3」為閱讀角度，判定每個「句」的代表節奏型式[14]，最後進行階層式的節奏型式分析，可得出下圖結果：

圖例 3. 「三民主義歌」的實驗性釋讀：復沓層次分類群組



比對現行的樂曲，除了首句和末句，分類群組的變換與樂句的變換同步。演唱習慣上，樂句最終高昂的結尾，與分析結果中孤立而突出的節奏層次相應：「貫徹始終(- - - +)」。

上述實驗性的手段無法斷言任一節奏型式為必然，也無法預測作者或讀者選定的選擇軸數量和位置。但是，事先取得作者或讀者熟習的文本作為對照組，有助推論隨機文本的幾種可能寫法或讀法，以及連結句子的節奏型式。

五、結論和未來研究方向

探勘文本內部的語言規律，本研究開放一個實驗性的釋讀方法。其間生成的字音字型分佈資料庫能成為「可查詢」和「可再生」的音韻學研究資源。現階段結論如下：「興體」只是詩句間意義或節奏的模糊聯結，過去注釋家隱約察覺卻無法言明的結構，我們現在可以使用數位方法，參考確切數據，在不同層次上辨識出來。本研究提出的「節律修辭」仍可商榷，但所有探勘出的潛藏句式足以微調詩經文體風格描述

[14] 亦即設想「三民主義歌」的每個四字句都有三個選擇軸。例如，「三民主義」將被拆解成「三〇〇〇」、「〇民〇〇」、「〇〇主〇」和「〇〇〇義」等四種型式，隨後計算這四種型式在詩經中出現的頻率，最後得出最高頻率的「〇民〇〇」為代表節奏型式，標記為「- + - -」。

的準確度。詩經十五國風、大雅、小雅、周頌、魯頌和商頌的文風，也可以依照本研究建議的步驟進行歷時性和共時性的比對。我們承認，詩經常見的四言型式「○○○○」已經可以支撐聲律的感覺；大量的單音詞彙，也方便調整句式之間的節奏聯結。只是，本研究最終的目的，是要在詩經音樂失傳且古音可能沒有四聲的條件下，考察詩句之間是否存在任何型式的「對位(counterpoint)」。例如本文末段嘗試將「高頻用詞模式」定義下的「節奏型式(字/○)」轉換為「二元樣板型式(+/−)」，進行句組間的對位，形成循環節奏，作為理論上的復沓。最後，伴隨各種語言規律的發掘，我們或可作一個平凡的經典讀者，與傳統注疏家平行，自由閱讀，自由對議。

參考文獻

Baxter, W. H. (2013). *Old Chinese: A new reconstruct*ion. Oxford University Press.

Diény, J.-P. (1963). *Les dix-neuf poèmes anciens*. Paris: Presses Universitaires de France.

Granet, M. (1929). *Fêtes et chansons anciennes de la Chine*. Paris: E. Leroux.

Jullien, F. (1985). *La valeur allusive des catégories originales de l'interprétation poétique dans la tradition chinoise: Contribution à une réflexion sur l'altérité interculturelle.* Paris: Ecole française d'Extrême-Orient.

Karlgren, B. (1950). *The book of odes*. Stockholm: Museum of Far Eastern Antiquities.

Martin, F. (1995). *Le Shijing, de la citation à l'allusion : la disponibilité du sens*. Paris : Extrême-Orient, Extrême-Occident.

Owen, S. (1992). *Readings in Chinese literary thought.* Cambridge, Mass: Council on East Asian Studies. Harvard University.

Rusk, B. (2012). *Critics and commentators: The Book of poems as classic and literature*. Cambridge, Mass: Harvard University Asia Center.

Vergne, J. (2004). *Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource.* Actes des JADT 2004.

陳致編。2010。《跨學科視野下的詩經研究》。上海：上海古籍出版社。

池昌海。2012。《先秦儒家修辭要論》。北京：中華書局。

馬瑞臨。1983。《文獻通考》。台北：臺灣商務印刷館。

毛亨，鄭玄，孔穎達。(2001)。《十三經注疏整理本：毛詩正義》（李學勤主編）。台北：臺灣古籍出版社。

王靖獻 (WANG C.H)。1990。《鐘與鼓 ： 詩經的套語及其創作方式》（The Bell And The Drum : Shih Ching As Formulaic Poetry In An Oral Tradition）（葉珊、謝謙譯）。成都：四川人民出版社。（原作 1974 年出版）

徐復觀。 2001。《中國文學論集》。台北：臺灣學生書局。

朱熹。1983。《詩經集傳（八卷）》。台北：臺灣商務印書館。

五代北宋山水畫的數位人文研究（二）：以「漁隱」主題為例

王平*、鈕亮**、金觀濤***、劉青峰****、毛建波*****

摘要

據畫史畫論文獻著錄，五代北宋時期山水畫的立軸橫卷類有 1200 餘幅作品，本研究只選取有實體流傳於今的 120 餘幅山水畫圖像為實驗素材。（另附加南宋山水畫 240 餘幅作品為輔證。）這些山水畫的圖像所涉主題約有 30 餘種，如「雪景」、「寒林」、「秋山」、「溪山」等。本研究僅以畫作中出現「漁舟」、「漁人」這兩種相對容易辨識的圖像做為擷取數據對象，共約 22 幅（著錄 132 幅）。其步驟是，首先對漁舟、漁人圖像造型流變、人物身份在不同時期的變化做出一般描述性統計分析，然後通過文本挖掘（包括畫題、題跋、相關畫論），提取該山水畫全幅的題材物類、畫題語義和情境設定等內容的特徵詞，針對這些特徵詞構建相關的語義網絡。在圖像分析和相應文本挖掘互相參照的基礎上，將山水畫的物理圖像特徵作為圖像數據，將表達山水畫內容的題材物類、畫題語義和情境設定等文本資訊作為屬性數據來構建圖像資訊數據模型。

借助這樣的圖像資訊模型可實現三個平行方向的研究路徑。1）以時代變遷為時間序列，通過非監督機器學習中的文本挖掘手段觀察畫作的流派風格的演變特點。2）以選定的能夠代表風格變化的局部圖像為分析對象，利用其特徵值，通過圖像統計手段觀察這些圖像自身的相似或變異特點。3）將基於文本挖掘的和圖像統計的分析兩相對照形成對畫作風格變遷的一個實證判斷。

最终籍由山水畫「漁隱」圖式及其語義的微观案例研究，我們實驗了圖文數據模型和數位人文研究方法在中國古代繪畫研究中運用的有效性。

關鍵字：數位人文、山水畫、圖像、漁隱、五代北宋、儒學

* 中國美術學院藝術人文學院博士研究生；貴州黔南民族師範學院副教授，Email: wping9@163.com。
** 中國計量大學經濟與管理學院講師，機器學習和文獻計量，Email: niutyut@126.com。
*** 國立政治大學講座教授；中國美術學院南山講座教授，Email: gtqf1908@gmail.com。
**** 香港中文大學中國文化研究所榮譽研究員，Email: 2869961913@qq.com。
***** 中國美術學院藝術與人文學院教授，Email: artmjb@163.com。

The Digital Humanities Research of the Landscape Painting of the Five Dynasties and Northern Song Dynasty（2）：A Study of the 「Fisherman-Hermit」 Theme in Painting

Ping Wang[*]、Liang Niu[**]、Guan-tao Jin[***]

Qing-feng Liu[****]、Jian-bo Mao[*****]

Abstract

In total, there are over 1200 hanging and hand scroll landscape paintings from five dynasties period to North Song dynasty which have been recorded in various painting histories, theories, and historical documentaries. In this study, we selected more than 120 paintings extant as experimental samples and another 240 paintings from South Song dynasty as supporting materials. The selected paintings represent over 30 categories of motifs, such as Snowy Scene, Wintry Forests, Dwellings in Autumnal Mountains, The Mountain and Stream and etc. The image capture technology is used on paintings with two easily recognizable motifs 'fish-boat' and 'fisherman', of which, 22 out of 132 document recorded paintings were selected. The experiment was carried out in several steps: the first step is to establish a descriptive and statistical model to analyze the formative evolution and character identification of fish-boat and fisherman in different time periods; the second step will construct a semantic network with feature words extracted from the motif information, thematic content and scenario settings of the paintings; finally, based on the cross-references of the image analysis and text mining, it attempts to establish a Full Image Information Retrieval Database for landscape paintings and related descriptions.

[*] Ph.D. Student, China Academy of Art; Associate professor, Qiannan Normal College for Nationalities. Email:wping9@163.com.
[**] Lecturer, College of Economics and Management, China Jiliang University. Email:niutyut@126.com.
[***] Chair Professor, National Chengchi University; Chair Professor, China Academy of Art. Email: gtqf1908@gmail.com.
[****] Honor researcher, The Chinese University of Hong Kong. Email: gtqf1908@gmail.com.
[*****] Professor, Advanced School of Art and Humanities, China Academy of Art. Email: artmjb@163.com.

The Image Information Model makes three collateral studies possible: 1. Through text mining of unsupervised machine learning method, it observes the stylistic evolution of paintings in a time series of dynasty changes; 2. By using image statistical tools, it observes the similarity and variation characteristics of the analytical objects, in this case, the selected local images which can represent stylistic changes in paintings; 3. By comparing the results of text mining and image statistical analysis, it gives an empirical judgment of stylistic changes of the paintings.

At last, By studying “Fisherman-Hermit” Theme pattern in painting and its semantics，We test a validity of method of picture-text data model and digital humanity used in chinese classic painting.

Keywords: digital humanities, landscape painting, image analysis, fisherman image, Northern Song Dynasty, Neo-Confucianism。

一、前言

在中國繪畫史領域，關於山水畫為何於五代北宋時期崛起，其根本的原因是什麼？同時，由於畫作流傳的大量散佚以及真偽混雜等因素的存在，五代北宋時期的山水畫究竟該是何種面貌？這兩個問題自從 20 世紀初宋畫研究興起以來即成為學界反覆討論的核心問題。可以看出，前一個是繪畫發展的後設層面的疏證問題，後一個是本體層面的繪畫原型的考證問題。就目前學界的研究態勢而言，脫離繪畫作品從文學和著錄史料探究山水畫之風格，或是僅從繪畫作品之圖像內容探究山水畫之生成與發展是學術之兩大主流。與此同時，美術史考古與鑒定領域的交流互通亦顯困難，這使得藝術史中的諸多問題都被懸置起來。

我們的期望是通過數位人文技術的介入，從而貫通藝術史和圖像以及鑒定三個研究領域。在 2014 年的研討會上，我們提交的以「求真」為主題的研究議題即是初步的探索。近兩年的進一步研究，我們圍繞這一研究思路以及資料庫的建設，對五代宋代山水畫進行了主題和題材分類、圖像的提取與標注、著錄和關聯詩文的文本匹配，並在有限範圍裡探索山水畫圖像的自動識別的可能性。本文僅以「漁隱」主題為考察範例。因從圖像標注資訊的統計中，「舟船」和「漁夫」題材的數量異常之多，且其形象屬類之變化與山水畫發展的風格變遷脈絡亦呈現出對應關係。故而，我們認為對此主題的研究既頗學術價值，亦是對數位人文研究方法在繪畫史領域運用可行性探討的較佳課題。

二、圖文數據庫的建構

(一) 圖像數據庫的建構

開展數位人文研究的前提和基礎是全文檢索數據庫的建立，圖像的數位研究自不能外。但在當前的數位技術條件下，要如文本資料一樣實現中國古代繪畫山水畫圖像的自動識別、提取與標注是不可能的。所以，我們只能選擇以人工的方式截取圖像，再進行分類和標注。（但是，從圖像自動識別的技術角度來看，對圖像的分類截取，是 AI 進行圖像識別深度學習的訓練庫建設之必須。）

在具體工作中，我們以《宋畫全集》（浙江大學出版社，2008）和《故宮藏畫大系》（國立故宮博物院，1993）為基礎圖像來源，在軟體中對掃描圖像中的物類進行截取存檔。而後，參照國立故宮博物院「典藏資料檢索系統」的主題標註和五代北宋時期繪畫理論中的主題詞，整理出標註詞典，對截取的圖像存檔做標註分類。以此建立的基礎圖像檢索系統可進行圖像的屬類檢索。在標註的過程中，我們對山水畫圖像的作者資訊、典藏資訊和學界的鑒定結論做了匹配。（見圖 1）

圖 1 山水畫圖像檢索系統結構

(二) 文本數據庫的建構與關聯

山水畫於魏晉興起之時即與山水文學相呼應，在唐代則出現「以詩入畫」和「題畫詩」之互動，至宋代則有蘇軾之「詩畫本一律」、張舜民之「詩是無形畫，畫是有形詩。」等主張，即見「詩畫」關聯之深。同時，檢閱有關古代詩詞「主題」的研究著述，亦可發現詩詞與繪畫在主題上的統一性，如「漁隱」在唐宋詩詞中即是一顯題。那麼，在數據庫中，以「主題詞」為仲介將繪畫圖像與詩文文本相對應，構建一互通之系統，對於繪畫的研究未嘗不是一裨益。圖 2 即是依此思路設計的交互性圖文數據庫模型。

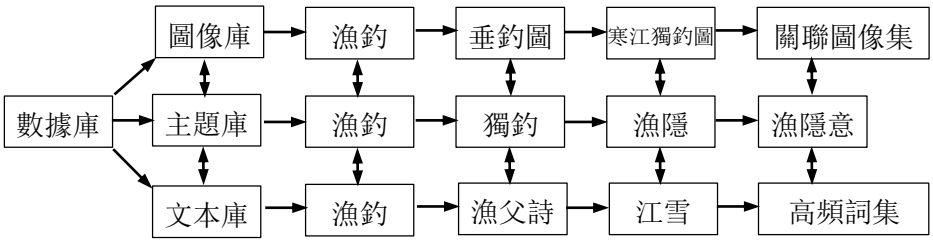


圖 2 圖文數據庫架構示意圖

此數據庫架構中「主題詞」的構成，一部分源於書畫著錄文獻中題名與題跋所整理的「主題詞集」，另一部分源於《全唐詩》、《全宋詞》、《歷代詩詞庫》和《典故大全》中整理的「主題詞集」。圖像庫中的圖像標註與分類皆引用此「主題詞集」。文本庫的運行亦是以「主題詞集」為詞典，進而採用 LDA 主題模型對文檔詞頻矩陣進行運算，以計算出語料文檔中的重要文本及高頻詞序列。此中，對文本最優主題之確定，學界目前

有貝葉斯統計中的標準方法[1]， KL 距離法[2]，餘弦相似度方法[3]， JS 距離方法[4]。緣於貝葉斯標準統計方法的簡潔和計算的效率，且已被大量研究課題所使用，本實驗亦採用之。

對於文本挖掘，我們的基本思路是一“降維”的過程，即文本被表示為一些特徵詞彙組合。通過特徵詞彙的計算以實現對文本隱含信息的揭示。鑒於可計算性是基於詞項之上，而詞項間的存在價值僅是概率價值，因此“作者的個性特點”在這種概率式的“詞袋模式”下即無區別了。（亦即“詞袋模式”階段的文本處理是去作者化的）詞袋“語料”對於歷史觀念研究之價值，即在於它去主體的“時間之維度”和“空間之維度”。同時，在對文本的處理上因介入了計量的視角，解釋者主體的思辨被分析所取代，它提供的是一種整體論（holistic）式的研究，這種整體論概念具有多義性、語境關聯性和滲透性三個特點。[5]由此，文本挖掘完成了兩重去主體化：一是去除文獻作者的主體身份，一是通過計量模式去除解釋者先入為主的身份，從而使文本提取的語料集呈現一客觀之立場。此時，我們作為解釋者介入，通過文本高頻詞、核心詞溯源，于時間層面觀察詞項的觀念變遷，于空間層面則可揭示觀念的地理轉移。這一階段聚類計量與文本解讀的結論既可佐證人文研究者的「理想類型（韋伯語）」，也可能開出一些新的啓發式的研究向度。

三、圖像中的「漁隱」

(一)五代北宋山水畫中的「漁舟」主題

首先，我們在圖像庫中以「舟船」為第一條件檢索，在五代北宋時期的山水畫中，「舟船」這一題材在 52 幅畫作[6]中出現，有「舟船」圖像 295 幅。（此中有部分遠景中極小的成組舟船、主要形體掩藏於其它物象之後無法做單體拆分的舟船組都做 1 幅計）如果剔除明顯為偽作的畫作，在「傳為」和「真跡」的 57 幅作品中則有 26 幅出現「舟船」圖像。以舟船的「形類」為條件進行檢索，可以看到有：扁舟（無蓬）、蓬舟、帆

[1] Griffiths T L, Steyvers M. Finding Scientific Topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(1):5228-5235.
[2] Rajkumar Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations[C] // Advances in Knowledge Discovery and Data Mining, Mohammed J. Zaki, Jeffrey Xu Yu, Balaraman Ravindran and Vikram Pudi (eds.). Springer Berlin Heidelberg, 2010:391–402.
[3] Cao Juan, Xia Tian, Li Jintao, Zhang Yongdong, and Tang Sheng. 2009. A density-based method for adaptive LDA model selection [C] . //Neurocomputing — 16th European Symposium on Artificial Neural Networks.2008,72,7-9: 1775-1781.
[4] Romain Deveaud, Éric SanJuan, and Patrice Bellot. Accurate and effective latent concept modeling for ad hoc information retrieval[J].Document numérique, 2014,17(1): 61-84.
[5] 布洛克曼：《結構主義》，北京：中國人民大學出版社，2010 年：第 1-2 頁。
[6] 本文所涉及的五代北宋山水畫包括《宋畫全集》與《故宮藏畫大系》所收錄，及其它圖冊中遼金作品 3 幅，李唐作品 4 幅，共計 117 幅。

船、並舟四大類，蓬舟下又有短蓬、長蓬與布蓬三類，帆船下有掛帆與落帆兩類。以舟船的「用途」為條件檢索，則有泊舟（無用）、漁舟、渡舟、遊船、貨船、勞作之船共 6 類。（見圖 3）以「位置」為檢索條件，則可見近、中、遠三個景層之中皆有舟船圖像。以「組合」為條件，在同幅作品中則有單隻、兩隻和多隻的分類，在多隻中則有單一分散、多隻組合和多樣組合，多隻組合中又有同形類、同用途和多類等組合。單就「舟船」中之「漁舟」而言，漁舟的判別是依賴於舟船上人物之行為。檢索舟船上人物之「行為」則有垂釣、網捕、扳罾、行船、休憩、擺渡、遊覽、勞作和其它 9 類。「漁事」之舟船，有僅容一人之扁舟、有可兩人之舟、還有將兩艘窄長之舟連結為一之「罾船」。

圖 3 北宋王希孟《千里江山圖卷》中「舟船」的用途

由以上圖像資訊聚類可知，五代北宋山水畫中「舟船」圖像的出現頻率頗高，「舟船」在畫面中之形類、用途、組合方式亦多。從檢索的截取圖像反溯作品及其創作之年代，則可以看到，北宋前期的畫作中舟船之屬類較少，且「舟船」之用途以「渡舟」和「漁舟」為多；從幅式角度來看，立軸類作品中「舟船」屬類較少，而橫卷類作品中「舟船」屬類較較繁多，最少者如《匡廬圖》僅 1 艘「短蓬舟」，最多者如《千里江山圖卷》有 122 艘、各「形類」各「用途」屬類皆備；從作者身份來看，畫院畫家作品中「舟船」屬類較多，而隱逸與士人類畫家的作品中「舟船」屬類較少。

(二) 五代北宋山水畫中的「漁隱」主題

山水畫中漁舟之判別依賴於「漁夫」之先行判定，但是，畫面中亦常見不乘舟之漁夫。檢索「漁夫」，可見有「有舟」與「無舟」兩大類；「無舟」即指漁夫身在台、岸、漁棚等處，尤其「扳罾」一類漁夫多在岸上；「有舟」則實包括「離舟」與「在舟」兩類，「離舟」多為漁夫歸泊後登岸之圖景；以「漁捕」為條件檢索，漁夫從事漁事的狀態則有「垂釣」、「未釣」、「扳罾」、「搬網」、「撒網」、「養殖」六類。人物之「組合」則

有單人、一船夫與一漁夫，多人 3 類。依漁夫服裝款式和著色、舟船上之器物等資訊，判別漁夫身份有「平民」、「士夫」、「孩童」和「不辨」四類。可見五代北宋時期山水畫中的「漁事」非僅垂釣與網捕兩類，而是含括了客觀生活中「漁業」活動的主要類別。

以往的繪畫史研究中，研究者多不具體區分「漁事」之類型和「漁夫」之身份，常以「漁樂」和「漁父」稱之。這固然與畫中點景人物尺度較小有一定關係，但若將「漁事」圖像聚類比對，即可發現，越是寥寥草筆的圖像往往越是契合某一個圖像程式。尤其五代北宋時期山水畫中，即便是程式化最突出的「垂釣」圖像，其中的差異也頗多，且值得深入研究探討。（如表 1 所示）

表 1 五代北宋山水畫中的漁夫「垂釣」圖像

作者	作品名稱	真偽	繪製年代	形態 1	形態 2	身份
董源	瀟湘圖卷	傳	元以前	垂釣	群釣	不辨
董源	龍宿郊民圖	傳	明以前	獨釣	獨釣	不辨
董源	夏景山口待渡圖卷	傳	元以前	垂釣	群釣	不辨
巨然	富春山居圖	偽	明	垂釣	獨釣	士人
李成	寒江釣艇圖	傳	元	垂釣	獨釣	士人
燕肅	春山圖卷	傳	金後	垂釣	獨釣	不辨
許道寧	雪溪漁父圖	傳	北宋後	垂釣	未釣	平民
許道寧	秋江漁艇圖卷	真	北宋	垂釣	對飲	平民
許道寧	雪景	偽	明	垂釣	賣魚	平民
惠崇	溪山春曉圖卷	傳	南宋	垂釣	對釣	士人
郭熙	雪景	偽	明	垂釣	獨釣	平民
王詵	漁村小雪圖卷	真	北宋	垂釣	對坐	士人
王詵	江山疊翠圖卷	偽	明	垂釣	對釣	士人
李公麟	溪橋散步圖	偽	清初	垂釣	對釣	士人
李唐	清溪魚隱圖卷	真	南宋	垂釣	獨釣	不辨
太古遺民	江山行旅圖卷	真	遼	垂釣	獨釣	不辨

由表 2 中可見，五代北宋時期山水畫中出現「垂釣」圖像的作品共計 16 幅。這些作品中，「漁夫」垂釣不僅僅是獨身垂釣一種，數艘舟船一起垂釣，或者在垂釣之時「對談」、「對飲」的圖式也較普遍。其中，確為五代北宋山水畫真跡的僅《秋江魚艇圖卷》

和《漁村小雪圖卷》2 幅，其「漁夫」、「舟船」皆以組合圖式出現。李唐(1066─1150)繪製於南宋的《清溪漁隱圖卷》在題名中明確使用「漁隱」一詞，但此題名題簽「清溪魚隱」者為明王顯之，李唐隱藏在圖卷末樹幹上的落款和拖尾南宋人題跋中皆未提及此圖最初之題名。如此，則「漁隱」題名在山水畫中使用的起點只能懸疑擱置。但從圖像角度看，在「舟船」圖像集中卻可以聚類的方式提出「漁隱」圖像的基本範式。（見圖 4）

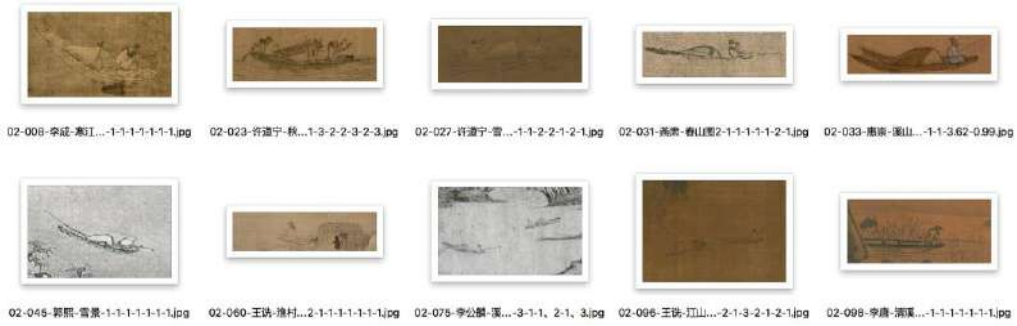


圖 4 「漁隱」主題的基本型

「**在雪天或蕭寒的氣候中，一中老年漁夫，戴斗笠披蓑衣，獨坐短蓬舟尾垂釣。**」（見圖 5）此即是「漁隱」主題的基本形，術語稱為「圖式」或「圖像典範」。（本文用「圖式」）將此「漁隱」圖式拓展到南宋繪畫圖像集中考察，亦可見在近 100 幅以團扇和冊頁為主體的作品中，有 17 幅使用且完全遵從了這一「漁隱圖式」。另有 10 餘幅「獨釣」圖式中，漁夫不戴穿斗笠蓑衣，而是做儒生便服形象，手中或坐側常有書籍、竹笛、酒具等，此可視為「漁隱」圖式的衍生形。再進一步擴展到元明清時期的山水畫中考察，更可見但凡畫作中出現「漁釣」形象，皆遵從「漁隱圖式」，且造型刻劃越是簡略則越是近似典範圖式。

於此，我們或可做一結論，即「漁隱圖式」是在五代北宋時期橫卷類山水畫中發展起來，至南宋形成了兩個「圖像定式」，在元明清時期的山水畫中「漁隱圖式」普遍存在。與此同時，我們又不禁追問，山水畫中的「漁隱」主題為何是這樣一種圖式？又是何種價值觀促成了這一圖式的形成？又是何原因使得南宋之後山水畫中「漁夫」形象皆遵從這一圖式而少例外？顯然，這三個問題並不是繪畫圖像和藝術史內部所能回答的，所以與山水畫、畫家關係密切的詩文語料拓展和關聯考察就成為必要。


圖 5 漁隱圖式

四、「漁隱」的寄寓

（一）詩文中的「漁父」主題

在「主題庫」中，與「漁」有關的有「漁父」、「漁翁」、「漁隱」、「漁樂」、「漁家」等，其中「漁父」和「漁翁」頻度最高，元代以前共有 1052 條文本。對這些詩文做詞頻矩陣運算，生成主題集 18 個，出現頻率 20 次以上的語詞 47 個。（見表 2）

表 2「漁父」、「漁翁」詩文高頻詞統計

滄浪	何處	蘆花	扁舟	不知	萬裡	一葉	不見	桃花	歸去	江湖	煙波
53	48	48	47	44	40	39	36	34	32	30	30
歸來	蓑衣	人間	明月	悠悠	月明	江上	孤舟	青山	江頭	無人	不可
29	29	28	27	27	27	26	25	25	24	24	23
平生	生涯	江山	風雨	鷗鷺	萬頃	春風	千古	夕陽	瀟湘	一聲	
23	23	22	21	21	21	20	20	20	20	20	

對表 2 中的高頻詞做語義和文本溯源疏證，即可發現，如「滄浪」一詞，其語源直指先秦之《孺子歌》:「滄浪之水清兮，可以濯我纓。滄浪之水濁兮，可以濯我足。」此詩歌又見於《孟子·離婁上》和《楚辭·漁父》; 又如「蘆花」、「明月」「萬裡」、「桃花」、「青山」、「桃花」、「歸去」、「江湖」、「蓑衣」、「風雨」、「鷗鷺」等詞，皆指向唐代張志和及其所作《漁父》詞；又如「夕陽」、「瀟湘」、「一聲」三詞則明確指向唐柳宗元「漁父」詩。綜合聚類統計后，可看到，這些高頻詞基本出於七則典故，如下表 3 所示：

表 3 「漁父」典故出處及主題詞一覽表

典故出處	**主題詞**
莊子・漁父	漁父、孔子、天下、澤畔、聖人、道、真、無累、葦間、江湖…
孟子・離婁上	仁、孺子歌、滄浪、清、濯纓、濁、濯足、安危、自取…
楚辭・漁父	屈子、漁父、澤畔、皆濁、獨清、見放、聖人、獨醒、皆醉、滄浪、濯纓、濯足…
吳越春秋	伍子胥、漁父、千尋之津、麥飯、魚羹、蘆中人、百金之劍、富貴、莫相忘、自沉…
後漢書・遺民傳	嚴子陵、帝、賢、羊裘、釣澤、士、有志、偃臥、富春山、釣台、特徵、嚴陵瀨…
張志和漁父詞	白鷺、桃花、箬笠，蓑衣、春江、細雨、江上、雪、月圓、醉宿、蘆花、風波、玄真、南溪、垂釣、秋山、野艇、倚檻、漁竿、是非、斜暉、醉宿、釣台、歸去、白浪、荷衣、窮、巴陵、仙…
柳宗元漁父詩	漁翁、西巖、欸乃、一聲、中流、相逐、清湘、楚竹、千山、人蹤滅（無人）、孤舟、蓑笠翁、獨釣、江雪、…

細說此六則典故詩文分別是：(1)《莊子・漁父》中虛構孔子遊蔡坐講時，漁父不邀而至，勸教孔子「謹脩而身，謹守其真，還以物與人，則無所累矣。」[7]之理，孔子敬漁父為「有道之賢」。(2)《孟子・離婁上》載孺子歌「滄浪」，而孔孟反說「不仁者不可与言」和「人當自立」之理，此中雖無漁父，但卻與「漁父」主題牽涉較多。(3)《楚辭・漁父》中，屈子貶放沅湘而遊時，漁父專候於澤畔說屈子以「聖人不凝滯於物，而能與世推移」之理，屈子否，漁父乃笑歌《孺子歌》而去。(4)《吳越春秋》中，伍子胥逃難，漁父助其渡河，為隱伍子胥行跡自沉於江，此漁父不同其他乃捨生取義者。(5)《後漢書・遺民傳》所載為光武帝同學嚴子陵不慕權貴，漁隱不受徵召之事。(6)唐人張志和（732—774）亦是徵召不受，棄官棄家，漁隱江湖，作《漁父》詞六首而名起唐宋。7）柳宗元（773—819）貶放永州時（805─815）作《漁翁》和《江雪》等詩，因其文章彪炳，亦廣受關注。

在相關研究著述中，有不少研究者將「漁父」主題的來源歸因為上古傳說中的「隱逸」之風，此雖不可說為無理據，但從文本「關鍵詞」疏證的角度來看，「漁父」傳統的生成則不外於上述七則典故文本。而由「漁父」生出直接「漁隱」一詞及其主題，則始見於北宋陳克（1081─1137）《奉題董端明漁父醉鄉燒香圖十六首・漁父七首其五》，[8]而作為正題廣泛使用是在南宋，如胡仔(1110─1170)著《苕溪漁隱叢話》一書，又有《題苕溪漁隱圖》詩三首，皆用「漁隱」題。(另以「漁隱」為關鍵詞，在《中國基本古籍數據庫》中檢索校證，結果亦相同。) 從詞義所指判別，「漁父」與「漁隱」在漢唐至北宋時期詩文中實為一義，在南宋則分別使用，其意圖當是對隱為「漁人」這一行為的價值予以區分，並且這一趨勢與山水畫圖像中「漁隱圖式」的形成在時間上基本對應。那麼，當時人為何要從「漁父」中另立出「漁隱」一詞呢？這一問題是值得繼續探討的。

（二）「漁父」和「漁隱」的價值判定

所謂「隱逸」即是對「入世有為，肯定禮制教化」持否定態度者，此種態度的形成需分作兩種客觀環境區別對待：一為戰亂動蕩時代，「隱居」對士人而言實是無奈自保之困局。一為太平盛世時期，「隱居」則是因志願不能適時申達而選擇的退避。既然「有為出仕」和「經濟天下」對儒生士夫來說是人生價值所在。那麼，儒生士夫對「隱退」和前代「隱士」之事跡做何種價值評判就必須要做分析辨解。對於「漁父」相關之莊子、屈子、嚴子陵、張志和及柳宗元五人的「隱退」事跡，唐宋士人多有評判。總括如下：

[7]莊周《莊子》(南華眞經卷第十)，四部叢刊景明世德堂刊本。中國基本古籍數據庫，2006 年。
[8]其詩為：「志和漁隱古仙真，霅水風流見後身。蓑笠何須訪圖畫，貂蟬凜凜在麒麟。」陳思《兩宋名賢小集》卷一百三十六《陳子高遺稿》，清文淵閣四庫全書本。中國基本古籍數據庫。

（1）唐時《莊子》因道教而興，宋代儒者亦好《莊子》。但《莊子》立論多對孔子極盡批判貶斥，故士人對《莊子》多主持「會通觀」，進而發展為「以儒釋莊」、「援莊入儒」。如王安石認為《莊子》「矯天下之弊，用其心亦二聖人之徒」[9]，蘇軾則認為對《莊子》當「實予而文不予，陽擠而陰助之」(《莊子祠堂記》)。這與唐宋詩詞中只說「莊子」而不說莊文中「漁父」的現象相應。此外，《莊子》中除《漁父》外，《秋水》、《外物》、《田子方》、《刻意》等篇亦有垂釣之「隱者」，在後世唐宋詩詞中皆無提及。

（2）屈子作《漁父》其意非是讚頌「漁父」，而是抒發自己即便身受放逐亦抱負不渝之志。對於屈子，賈誼即有《吊屈原賦》一文，論士人即便身遭放逐不得志時，當選擇隱逸而不可以自殺洩忿。唐宋詩詞雖多以「文儒」讚頌屈子，但對由屈子引出的「忠與怨」、「隱與死」議題討論也多，如韓愈、司馬光、晁補之、朱熹等。[10]但就「漁父」的典故而言，《楚辭．漁父》是七則中唯一的反面形象，故唐宋九十餘首引用此典的詩詞中也只說屈子而不說此「漁父」。

（3）嚴子陵是士夫作為「漁父」形象出現的第一人。唐代吟誦其事跡的詩就有 20 多首，宋代則有 40 多首。董弅纂《嚴陵集》，而最著名的莫過於范仲淹《桐廬郡嚴先生祠堂記》一文，其評價嚴子陵「…惟先生以節高之。既而動星象，歸江湖，得聖人之清。」[11]可做普遍觀念看。而由「先生成光武之大，光武成先生之高」引發出「君臣合德」之旨，遂使「漁父」具有了極正面又極光輝的儒家價值。

（4）五代兩宋時期，尤以張志和《漁父》詞影響廣大。通過題名（詞調）、題注、核心詞的檢索統計，五代兩宋時期就有 781 首以「漁父」為題或論及「漁父」的詩詞。其中，與張志和遙和的詩詞有 130 餘首，用及張志和《漁父》六首主題詞者有 420 餘首，乃一大系統。宋人認為張詞寫「漁家之樂，其樂無風波之患。對面已有不能自已者，已隱然躍於言外……」[12]而對張志和的「隱逸」事跡則多不討論。如此「漁父」及其「隱逸的生活」則化為一主題集，成為士人不合而隱、辭職罷免生活中托志抒懷的載體。

（5）柳宗元在兩宋的影響主要籍由蘇軾在黃州貶居時之發明，以及劉克莊、嚴羽的再次推崇。柳宗元一生基本在放逐中渡過，卻于放逐中不墮其志，發揚儒學，關心時政，教化一方，成就政務與文章。由於兩宋時期黨爭不斷，士人常有貶謫與自放，柳宗元的詩文事跡便成為「古士人之模範」與默契。[13]於是，「漁翁」、「江雪」二詩中「獨釣」

[9]（宋）王安石：《臨川先生文集》卷第六十八《莊周上》，四部叢刊景明嘉靖本。中國基本古籍數據庫。參見簡光明《宋代「援莊入儒」綜論》(《嘉大中文學報》2009 年 9 月第二期，第 121- 150 頁。)

[10]參見林姍：《宋代屈原批评研究》，福建師範大學博士學位論文，2011 年。

[11]（宋）范仲淹：《范文正公文集》卷第七《桐廬郡嚴先生祠堂記》，四部叢刊景明翻元刊本。

[12]黃蘇：《蓼園詞評》，唐圭璋：《詞話叢編》，北京：中華書局，1986 年，第 3023 頁。

[13]參見衣若芬：《瀟湘文學與圖繪中的柳宗元》，《零陵學院學報》，2002 年 9 月，第 23 卷第 1 期。及赵东丽：《宋代柳宗元诗歌接受研究》，河南大學碩士學位論文，2006 年。

與「釋懷」的語義在兩宋之際化成為「漁父」主題的個體精神狀態，（漁父的生活內容主題集被削解）并最終生成了「漁隱」這一更為具體的文化符號。

（6）《孺子歌》在唐宋詩詞中出現時，其核心詞「滄浪」、「濯纓」、「濯足」的關聯詞集皆指向《楚辭》中「漁父歌」和屈子事跡，未見有用《孟子・離婁上》中論「仁」之義者。《吳越春秋》中的「漁父」引用頻次也是極低，關注者較少。

由上述「漁父」到「漁隱」的語義轉變及時人價值評判中，我們可以看到：「漁父」在先秦文獻中，是以虛構的儒家文化的逆反形象而出現，《莊子》中的「漁夫」討論的是「入世有為」問題，《楚辭》中的「漁父」討論的是「忠恕」問題，《孟子》「孺子歌」章節討論的是「德性自立」問題，《吳越春秋》中的「漁父」討論的是「義利」問題，嚴子陵隱為「漁父」本牽涉的是「名禮」問題唐宋則引申為「君臣合德」問題，張志和隱為「漁父」引發的是「窮達進退」問題，柳宗元「漁父」引出的則是「退而成德」問題。可以說，縱觀古代文藝之主題集，似乎少有如「漁父」主題這般牽涉儒家文化系統中如此核心命題者。由「漁父」向「漁隱」的蛻變既體現出儒學對道家之學的消解與轉化，也體現著宋明理學於「成德」與「修身」（即「內聖」）命題在儒學系統中建構之功。「漁隱」主題的典範化傳達的「窮居成德」之價值立場，是理學家（尤其是程朱理學家）所崇尚的，也是漢唐時期儒學價值中缺乏的。[14]

五、結論

「隱士」作為政治文化的逆反者，在中西方文化中皆存在。但「漁父」和「漁隱」作為文化符號，卻只存在於東亞儒家文化圈中。作為文藝主題，「漁父」和「漁隱」是中國古代詩詞、繪畫和音樂（古琴《欸乃曲》）所共有。雖然「漁父」主題的語源可追溯至先秦道家的道家傳統，但是，隨著儒學成為中國文化的主體，並在發展中逐步消解融合道家學說，「漁父」主題的意指也逐步蛻變，最終形成了「漁隱」這一指寓「具有堅定志願又不掛礙名利」的文化符號，其在南宋之後詩詞與山水畫中的盛行，則是對理學「內圣」之道的圖像呼應與呈現。

籍由山水畫「漁隱」圖式及其語義的研究，我們實驗了圖文數據庫和數位人文研究方法在中國古代繪畫研究中運用的有效性。就數位繪畫研究方法可歸納以下幾點說明：

[14]參見金觀濤、劉青峰《中國思想史十講》，北京：法律出版社，2015：第五講之「《近思錄》：理學修身結構之呈現」。與余英時《朱熹的歷史世界》，北京：三聯書店，2011：第八章。

1. 圖像數據庫的建設，在當前條件下尚無法實現早期古代中國畫圖像的 AI 自動識別與標註，人工截圖與標註在技術上看似落後，但卻是今後實現 AI 自動識別與標註的必要基礎。

2. 通過文本沉澱和圖像資訊校正獲得的「主題詞庫」，對於圖文數據庫的建設和運行具有至關重要的作用，它是圖像標註、圖像與文本數據子庫的中介。

3. 單純的圖像數據庫對於繪畫史研究或繪畫風格的研究，和單純的文本數據庫對於繪畫的圖像的研究都是藝術史研究領域的一邊，通過觀念史和圖像聚類等數為人文方法的介入，對於圖像研究和藝術史研究或許可以打開一道貫通之門。

由於五代北宋時期的山水畫數量較少，且真偽問題極其突出。我們的工作即是以此最複雜的圖像與文本材料為實驗對象，以最大的難題來考驗圖文數據庫的可行性與有效性，所以我們期待著同行研究者的關注與批評。我們希冀著圖像識別等新技術對美術史研究帶來的技術革新。

Paper Session 4

語料庫語義：社會學應用

Corpus Linguistics for Social Science

以語料庫分析取徑探究臺灣新聞中的跨性別：以聯合知識庫為例

羅盤針[*]、鄭碩[**]、江安淇[***]、曾博揚[****]

摘　要

本文關注臺灣媒體對跨性別者及跨性別議題的報導與再現。以聯合知識庫中 1951 年至 2016 年與跨性別有關的新聞為例，透過詞頻、共現、關鍵字檢索等語料庫分析取徑，觀察報導常用的詞彙及特定詞彙的使用狀況，再透過詞彙間的共現關係建立詞彙網絡，以引入網絡分析，觀察新聞中與**跨性別**相關的詞彙及**跨性別**與**變性**、**男扮女裝**、**人妖**等詞彙間的關係。本文發現新聞提到跨性別、變性時再現的形象相對男扮女裝、人妖來得正面；跨性別一詞與性別平權運動密切關聯，而變性牽涉的面向更廣。本文也嘗試以歷史研究的視角將變性相關報導分期，透過觀察顯著詞，發現 1990 年代的報導主題自醫療面向轉往社會面向。而這些結果也顯示語料庫取徑可協助論述分析的進行。

關鍵字：語料庫分析、詞彙網絡分析、跨性別、變性、聯合知識庫

[*] 國立臺灣大學歷史學系大學部學生，Email: b02103015@ntu.edu.tw。
[**] 國立臺灣大學歷史學系大學部學生。
[***] 國立臺灣大學社會學系大學部學生。
[****] 國立臺灣師範大學歷史學系大學部學生。

Applying Corpus Analysis to Explore Transgender in Taiwanese Newspapers

Pan-chen Lo[*], Shuo Zheng[**], An-chi Chiang[***], Bo-yang Ceng[****]

Abstract

The study focuses on how transgender people and related issues were reported and represented by Taiwanese newspapers in the late 20th century. We applied corpus analysis techniques, including frequency analysis, co-word analysis, and KWIC, to analyze transgender-related news text collected from Taiwanese newspaper database Udndata during 1951-2016. Our objective is to find out what kinds of knowledge or image in regard to transgender the readers could receive. We inspected what are the most frequently used words and how specific words were applied by the news. The result suggested that kua-xing-bie (transgender) is often mentioned along with LGBT issue. However, instead of transgender issues, LGBT issues occupies the center of narrative attention. It was also discovered that western nations like the U.S and the U.K. have a lot to do with the term kua-xing-bie in Taiwanese newspapers.

Network analysis was also introduced through co-word network created based on co-word relationship between words to investigate the relationship and distinction between transgender-related words like kua-xing-bie, bian-xing (transsexual), nan-ban-nu-zhuang (a man dresses up as a woman), and ren-yao (kathoey). It was discovered that the former two were represented in a more positively way than the latter two in news, and kua-xing-bie was strongly related to gender rights movements while bian-xing involving more extensive aspects.

From a historical perspective, we divided news about bian-xing into two periods: before and after 1990s and practiced keyness analysis. The result showed that the subject of news has shifted from medical matters to social issues in the 1990s. Additionally, these results indicated that corpus analysis could assist the process of discourse analysis.

Keywords: corpus analysis, co-word network analysis, transgender, transsexual

[*] Undergraduate Student, Department of History, National Taiwan University. Email: b02103015@ntu.edu.tw.
[**] Undergraduate Student, Department of History, National Taiwan University.
[***] Undergraduate Student, Department of Sociology, National Taiwan University.
[****] Undergraduate Student, Department of History, National Taiwan Normal University.

一、 前言

在今日台灣的性別平權議題討論中，跨性別已非陌生的詞彙，一般指稱內在認同、外在表現異於自己出生的性別或社會所規範性別之人，不過這個詞涵蓋的意義範圍邊界為何，至今仍眾說紛紜。台灣第一本跨性別研究專書《跨性別》的編者何春蕤寫道：「研究跨性別的學者都苦於定義的問題」，並將跨性別視為包含多種概念的大傘術語（umbrella term）[1]。除了學者正在琢磨，學院外的社會亦開始認識跨性別。新聞媒體扮演其中重要的角色，在報導中使用跨性別一詞，不僅將這個詞語帶到作為讀者的社會大眾眼前，隨之而來的描述和觀點也可能形塑讀者對跨性別的印象。這個過程是媒體對於跨性別的再現（representation）。

20 世紀人文社會學科經歷所謂「語言學轉向」（linguistic turn），其中一重要主張是認為不存在先於語言的概念及意義，概念必須置於具體的語言和字詞組合才能被理解。本文受此啟發，認為變性雖在學術概念上可納入跨性別大傘下，但由於新聞使用兩個詞的情境不盡相同，傳遞訊息時賦予兩者的意義範疇也會有些差別，簡而言，報導會使這兩個詞給人的印象及聯想到的事物有所差異。除變性，人妖、扮裝皇后、第三性、陰陽人、男扮女裝、女扮男裝…這些被何春蕤等學者收入跨性別大傘下的詞也應存在類似狀況。透過新聞內容的文字，跨性別一詞與其他詞彙形塑了讀者的跨性別知識圖景，其中跨性別未必能如大傘般地包覆這些詞彙，而是呈現彼此意義範疇部分交疊、部分分離的狀態。

對於此一知識圖景，本文嘗試以語料庫分析方法來勾勒及觀察，並加入時間性、歷史的考察。關於這方面，獨立研究者陳薇真《臺灣跨性別前史》即以新聞作為史料，是台灣第一本欲研究歷史中的跨性別之專書，其中也為台灣的跨性別史作了一些分析；[2]不過就字詞而言，陳薇真所關注的正是跨性別一詞出現之前變性、人妖相關的報導，因此對於跨性別一詞在新聞的浮現及之後內涵如何發展較少著墨，本文透過語料庫分析方法，觀察量化資料，不同於林、陳兩人缺乏數據的質化研究，或許能與之相互對話或補充，描繪當代歷史尚未受到太多矚目的一個側面。

本文將會做出些許解釋和定義，然而探索和嘗試方法的過程本身亦是重要的結果。因此也提供視覺化分析成果如詞頻分析資料表、詞彙網絡圖等等，期望如此一來無論字詞本身的意義範疇或字詞之間的關係網絡，皆是可再探勘與詮釋的開放空間。

[1] 何春蕤主編，《跨性別》（桃園：中央大學性／別研究室，2003）。
[2] 陳薇真，《台灣跨性別前史：醫療、風俗誌與亞際遭逢》（臺北：跨性別倡議站，2016）。

二、 相關研究回顧

本研究欲以語料庫、網絡分析等方法研究新聞報導中的跨性別，而英國學者 Paul Baker 就曾以英媒關於跨性別的新聞為例，展示詞頻、詞彙共現（co-occurrence）、關鍵詞檢索（KWIC）等語料庫分析方法（corpus analysis）對進行批判論述分析（critical discourse analysis）可能的貢獻[3]。該文分析 *transgender*、*transsexual* 與 *tranny* 等詞彙的總詞頻及其共現詞的使用，呈現指稱跨性別的詞語在英國報導中被使用的特性，例如 *transgender* 一詞涉及的形象較為正面、*tranny* 則明顯較負面。

該文旨在示範研究方法，並嘗試與已有的跨性別研究對話，例如其分析成果支持英國跨性別女性主義者 Jane Fae 認為英媒將跨性別者做為其他少數族群代罪羔羊的觀點。Paul Baker 利用的語料庫規模較小，僅包括 *The Times, The Telegraph, The Sun* 等英媒於 2012 年發行與跨性別有關的報導，共 902 篇新聞、661189 字。

本研究基本構想與 Paul Baker 相似，但方法的運用則稍有不同，除了由詞頻分析進行觀察，亦嘗試結合網絡分析。另外，雖然英文的 *transgender*、*transsexual*、*tranny* 未必能輕易地對應為臺灣所使用的跨性別、變性等詞彙，但詞彙間有褒貶意義落差的現象可能是跨文化、跨語言存在的。一般認為跨性別是由 *transgender* 翻譯而來，也常見將兩者並置、互相參照的寫法，本研究無法處理翻譯的議題，但希望透過以近似的方法分析比較，可以看出兩個詞彙另一層面的關係。

至於臺灣媒體對跨性別的再現，前言提及林佳緣的碩士論文亦曾以批判論述分析解讀跨性別相關報導。[4]採用中國時報、民生報、聯合報三家報紙自 1994 年到 2004 年間對本地跨性別者的報導作為素材，同樣以 Norman Fairclough 的批判論述分析為主要的研究方法，討論新聞如何塑造跨性別者的形象與經驗，並揭露新聞背後所隱藏的文化意涵。同時，也採用深度訪談法理解跨性別者在閱讀新聞報導後的感受，以及自身經驗與報導異同之處，討論跨性別者如何被媒體再現。

林佳緣建議往後的研究進一步分析不同時期掌握話語權的人，以及價值觀、權力機制如何在不同時期中改變，並彰顯跨性別者的差異性，例如社經地位高低的影響。本研究關注有關跨性別的不同詞彙在新聞中呈現的差異，或許也能呼應是身分不同的跨性別

[3] Paul Baker, Bad wigs and screaming Mimis: “using corpus-assisted techniques to carry out critical discourse analysis of the representation of trans people in the British press”, in Christopher Hart, Piotr Cap （Eds.）, *Contemporary critical discourse studies* （London: Bloomsbury. 2014）

[4] 林佳緣，〈跨性別媒體再現與主體解讀之分析研究〉，世新大學性別研究所碩士學位論文，出版年份：2006。

者被媒體再現，暗示跨性別內部的分歧。

在方法上，林佳緣僅以跨性、變性、女變男、男變女為新聞搜尋關鍵字，討論集中在變性；跨性別則被其認定為是倫理上用來取代變性的「更中性的稱呼」，跨性別作為新聞用詞的一面並未被詳細討論。另外，其雖然列出使用新聞的時間範圍，但並未以歷史視角觀察變化，而這是本研究欲嘗試的方向之一；在這部分本研究也將嘗試以量化成果與陳薇真的跨性別歷史研究對話，在此暫不詳述。

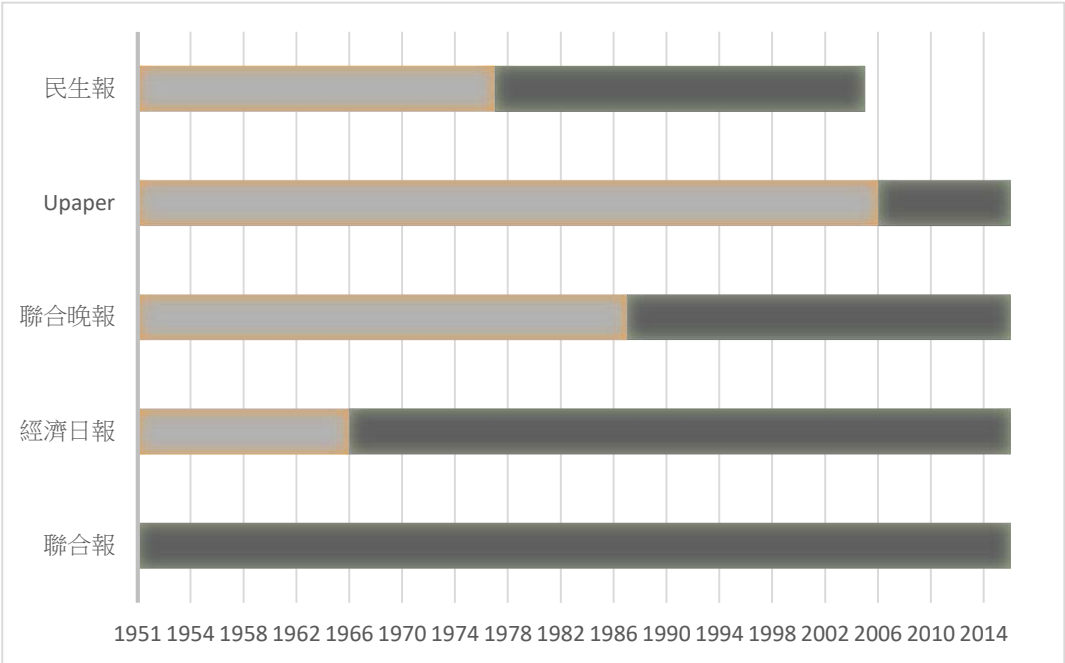
以下由分析直接提及跨性別的新聞（簡稱跨性別新聞）開啟本研究的探索。

三、 跨性別新聞的語料庫分析

（一） 語料庫來源：聯合知識庫

本研究的新聞資料來源為聯合知識庫，使用報紙包括聯合報、聯合晚報、民生報、Upaper 以及經濟日報，發行日期涵蓋 1951 至今，各別發行時間見下頁表。聯合知識庫時間涵蓋範圍較大，並提供全文。相較之下，同為新聞知識庫的慧科新聞知識庫雖然收錄較多報系，但其僅提供 1998 年至今的新聞，無法觀察進行時間跨度相對較大的觀察。

在聯合知識庫以跨性別為關鍵字搜尋得到的報導共 351 篇，約十五萬字左右，自 1993 年起直到 2016 年 8 月，時間跨度近 23 年。



表一　新聞發行時間

（二） 總詞頻分析

首先進行斷詞及詞頻分析，得到全部詞彙的總詞頻及文件詞頻。這些報導中，除了作為搜尋條件的跨性別，性別、女、男無論是總詞頻或文件詞頻的值都明顯突出，即新聞提到跨性別時幾乎都會在敘述中使用三者(參考表二)。總詞頻很高的還有同志一詞，顯示新聞提及跨性別時傾向表示多元性別的意涵，接近英文 transgender 意義，而非指男女兩性間的互動（cross-gender）：

> 所有「驕傲華堡」販售所得將捐給漢堡王麥克拉摩基金會，作為授予明年高中畢業的 LGBT（男女同志、雙性戀與跨性別者）學生獎學金。[5]

從引文中可以見到報導將跨性別與男女同志、雙性戀並提以說明 LGBT 的內涵，另一個總詞頻很高的詞彙同性戀也存在相似的狀況。雖然本研究最初以跨性別為關鍵字搜尋，但閱讀新聞內容後發現，將跨性別與同性戀、雙性戀並提的報導中，往往僅有此一句敘述帶到跨性別，並未多加著墨。

而自己則可能強調跨性別與個人想法有關，與自己共現關係較顯著的詞還有認同，應能支持此推測。例如：

> 透過變性手術從男性變為女性的許百欣說，自己身為男性但從小一直認同自己為女性，但在社會觀念封閉下，只能選擇結婚生子。[6]

而從其他詞彙如人權看來，可能有不少新聞提及臺灣跨性別平權運動的論述；兒廁所、婚姻等詞彙則呈現較具體的主題，可能是與運動有關的訴求。

（三） 特定詞彙分析

1. 話語權：說、指出

除直接觀察詞頻較高的辭彙，也能從新聞特性出發，透過特定辭彙進行語料庫分析。郭文平認為透過說、表示、指出等詞與其他詞彙的共現關係可以掌握新聞訪問對象；[7]此方法也能用於觀察跨性別新聞中話語權問題。說的總詞頻（575）與文件詞頻（194）都不低。進行共現詞分析並配合關鍵詞檢索掌握脈絡，可發現說的主詞多是受訪講述自身經驗的跨性別者，如婚姻註記遭內政部撤銷的吳伊婷、吳芷儀或者印尼跨性別者特迪：

---

[5] 劉利貞，〈漢堡王挺同志 推彩虹套餐〉，《經濟日報》，2014 年 7 月 3 日，第 A8 版（國際企業）。
[6] 陳麗婷，〈性別變更困難 在英他算男生 來台他變女生〉，《聯合晚報》，2013 年 11 月 21 日，第 A11 版（健康）。
[7] 郭文平，〈字彙實踐與媒介再現〉，《新聞學研究》125 期（2015：臺北），頁 118-119。

現年廿七歲的吳伊婷說，直到青春期，才懷疑自己為何身為男生，高三時向家人出櫃，卻被送到精神科治療。[8]

指出的主詞有聯盟、研究、協會、團體，透過關鍵字觀察脈絡，仍能掌握辭彙所指為何。如聯盟為「高雄同志遊行聯盟」、「花蓮同志遊行聯盟」；協會為「性別不明關懷協會」、「台灣性別平等教育協會」、「性別人權協會」；團體為「同志團體」、「性別團體」、「人權團體」、「婦女團體」、「跨性別團體」；研究則為「性別研究」、「性／別研究」等：

花蓮同志遊行聯盟指出，每個人都能學著愛人、也有被愛的權利，因此不管異性戀、同性戀、雙性戀、跨性別，甚至是具有更多元性別特質的人們，在社會中，都應該享有平等的公民權利。[9]

2. 區域：美國、亞洲、台北

從地名、國名來觀察，美國的總詞頻很高，英國、國際也不低（見下表），可見不少報導提及外國——又以歐美居多。報導國外事例以美國的跨性別運動、法案推行最多，其次為英國等歐洲地區的性別運動之報導。美國總統歐巴馬公開表態支持同志的新聞即是一例，這樣的報導可能為讀者提供「美國支持同志、多元性別認同」的認知：

美國總統歐巴馬八日呼籲，不要再透過心理治療，試圖改變男男同性戀和跨性別未成年人的性傾向。[10]

至於運動者與學者多是受邀來台參與國際研討會、講座。值得注意的是，何春蕤主持的中央大學性／別研究室[11]常出現在相關報導中：

中央大學幸運獲得英國官方贈送來自英國、剛上市的性別研究專書……何春蕤指出，今年五月十一日到十六日，英國婦女及性別研究學術訪問團來台訪問時，包括約克大學女性研究中心主任潔克遜、李滋大學跨領域性別研究中心主任羅絲妮、薩塞克斯大學女性研究中心主任艾賀恩等人皆是「性／別研究」專家[12]。

---

[8] 鄭宏斌，〈爭婚姻平權 她們大方站出來〉，《聯合報》，2013 年 7 月 12 日，第 A8 版（生活）。

[9] 范振和、徐庭揚，〈認知、情感、心胸 都多元 花蓮彩虹遊行 上街呼喊幸福〉，《聯合報》，2014 年 9 月 28 日，第 B1 版（宜花・運動）。

[10] 田思怡編譯，〈愛她？害他！科學證據：轉化治療造成傷害 歐巴馬：別再強迫治療性向〉，《聯合報》，2015 年 4 月 10 日，A19 版（國際）。

[11] 中央大學「性／別研究室」成立於一九九五年十月，研究室以階級族群、年齡、性別（gender） 等「社會差異」或「別」（differences）為原點，再結合同性戀等「性」（sexualities） 議題，從批判的人文社會理論出發，以同性戀、女性主義、性教育、性學、性／別文化、性／別文學、性 ／別倫理與哲學為研究焦點。

[12] 陳大鵬，〈性議題 跨領域研究 新視野 中央大學昨獲贈英國新書 何春蕤：每一本都令人流口水〉，《民生報》，1999 年 8 月 11 日，第 06 版（文化與藝術）。

其他來臺的外國人士尚有費雷思（Leslie Feinberg）、葛傑密（Jamison Green）等。

另外，許多提及歐美的新聞與非學術書籍、電影有關，顯示這種文本形式是台灣民眾認識國外跨性別訊息的管道。例如 2000 年得到奧斯卡大獎的美國電影《男孩別哭》，相關報導中包含跨性別者處境的討論；或者 2015 年的《丹麥女孩》。書籍則有《寂寞之井》、《藍調石牆 T》等。

除了觀察歐美資訊的跨國輸入，從地名出發的詞頻分析也呈現臺灣內部區域間的差異。台北的總詞頻很高，可能與在臺北進行的倡議活動較頻繁有關，例如：

> 台北街頭昨天「性別大解放！」[13]

不過，這也可能反映出台北相對其他區域有較多資源製作新聞、發布量較高，與台北有關的事件容易曝光。

表二　跨性別新聞詞頻分析結果：按總詞頻排序

詞彙	總詞頻	文件頻率
性別	795	200
同志	772	133
女	714	200
跨性別	633	351
男	595	180
廁所	460	57
自己	349	150
同性戀	312	120
婚姻	269	72
同學	260	80
人權	222	103

表三　區域詞彙詞頻分析結果

詞彙（同類詞）	總詞頻	文件頻率
全球（世界、國際）	204	100
美國	190	98
台北（臺北）	167	127
英國	50	27
丹麥	28	6
法國	22	10
歐洲	11	20

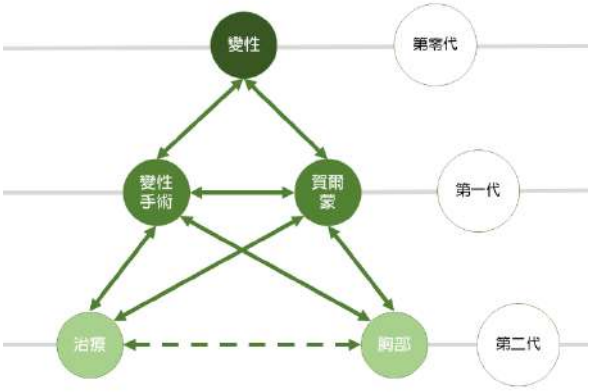
（四） 詞彙網絡分析

如上節所示，詞頻分析提供從各個詞彙分別再延伸討論的可能；然而詞彙間在報導內容中的關係卻無從得知。因此以下將對跨性別新聞進行辭彙網絡分析，尋找與跨性別共同出現的詞彙，並歸納出跨性別新聞常見的報導方向。

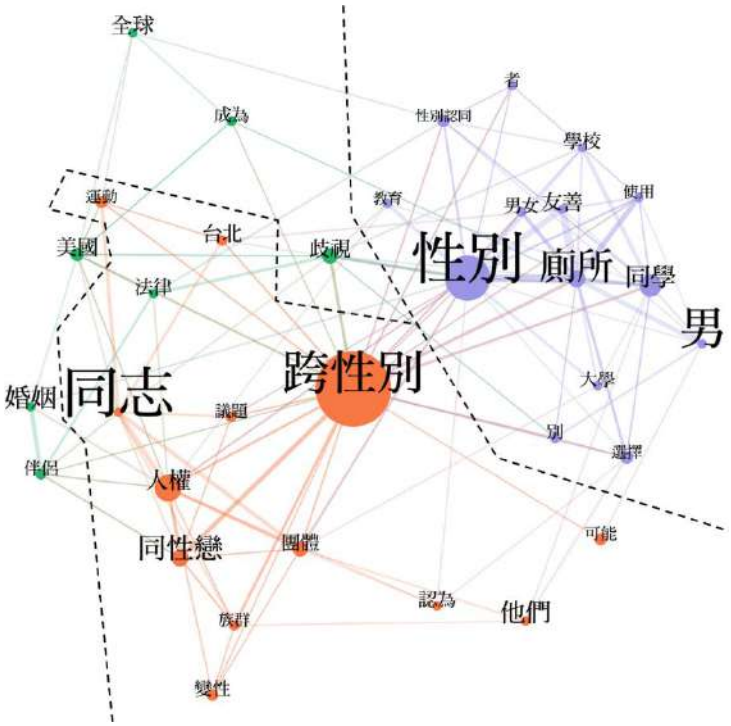
[13] 鍾蓮芳，〈同志遊行 台北街頭性別大解放 變性人、雙性戀者、SM 團體全出籠 要平等、要小孩 還有公民權〉，《民生報》，2004 年 11 月 7 日，第 A3 版（生活新聞）。

1. 研究方法簡述

詞彙網絡分析是本研究嘗試結合語料庫分析和社會網絡分析的新方法，故於此稍作說明。首先將跨性別作為第零代關鍵字，找出上述詞頻分析成果中總詞頻大於 220、t-score 大於 5 的詞彙，作為第一代關鍵字；接著以第一代關鍵字經由同樣條件找出第二代關鍵字。為將網絡規模維持在可進行細部解讀的大小，並考量詞彙與第零代關鍵詞的相關程度，只觀察至第二代關鍵詞為止。以變性為例（見示意圖），第一代有變性手術、賀爾蒙等關鍵字，變性手術再衍生出第二代治療等、賀爾蒙再衍生出第二代胸部等。另外，也可能因該第二代詞彙在其他路徑為第一代詞彙，而可見兩個第二代詞彙間關係的狀況。



圖一　共現關係示意圖



Filter	Appearance
Node (34): Freq. > 70 Edge (183): T-score > 5	Node: — Color: Modularity Class — Size: Betweenness Centrality — Text Size: Freq. Edge: — Weight: T-score

圖二　跨性別新聞詞彙網絡分析成果

2. 分析成果

成果如（圖二）所示，透過計算網絡關係分出三群，[14]可以看到總詞頻最高的兩個詞彙性別與同志分別與周圍詞彙在跨性別的兩側形成詞群。性別在右側紫色群，與廁所、學校、友善、同學等詞共現關係顯著，此外性別認同、教育也值得注意。這些詞彙組成類似以下的內容：

> 世新大學昨天啟用全國大專院校第一個性別友善廁所，在校內傳播大廈和舍我樓四樓建置二處，提供跨性別同學舒適的如廁環境。[15]

同志則在橘色群，雖然字體極大，表示該點 Betweenness Centrality 的點大小卻相對偏小，人權、同性戀的點則較大，代表「同志」並非連結其他詞彙的重要橋樑。

比較起來，紫色群偏向針對跨性別者校園生活的討論，顯示校園是觀察臺灣跨性別權益運動的一重要空間。橘色群涉及的新聞內容則多將跨性別納入同志、LGBT 的定義底下，不少報導將跨性別與同性戀並提、或將變性人與同性戀並提，作為對同志族群分支的補充說明。至於左上方的綠色群包含美國、全球、婚姻等，可能與上一節提到媒體大量報導美國的跨性別議題有關。這三個詞彙群大概呈現跨性別新聞常見的用語及主題。

整體而論，由這些詞彙組合成的主題包含了爭取權益、追求平等、推廣教育、反歧視及創造友善環境，是倡議者的語言，似乎顯示跨性別在新聞中出現時較為正面，不過這也可能反映新聞書寫與倡議團體提供的新聞稿接近。

---

[14] 分群即社會網絡分析中的 Modularity。參考 Vincent. D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in *Journal of Statistical Mechanics: Theory and Experiment*, 2008 （10）, P1000.

[15] 鄭語謙，〈學生反應兩極 世新設友善廁所 男女共用〉，《聯合報》，2011 年 10 月 27 日，第 A6 版（生活）。

3. 小結

跨性別自 1993 年開始出現在新聞中，[16]多被描述為同志族群的其中一種，與同性戀、雙性戀並提，但採用此種敘述提的新聞中跨性別往往是附屬的、出現在補充說明的括弧內。另外，分析結果反映美國及一些歐洲國家是臺灣媒體吸收、傳遞跨性別資訊的重要來源，又可分為間接報導社會運動、法案政策、文藝作品或外國人士受邀來台；至於臺灣本地跨性別的消息多與台北有關，較具體的議題則包含校園、廁所等面向。

總的來說，跨性別在新聞中具有高度的倡議色彩，報導對詞彙所描述對象的態度友善。同時這個詞多與知識分子有關，可能具有一定的社會地位屬性。

四、跨性別、變性、人妖、男扮女裝的詞彙網絡分析

如前所述，在學術概念上可被納入跨性別大傘下的變性等詞彙，在新聞報導行文中與跨性別存在某種程度的落差，因此，本研究接著另外再挑選幾個關鍵字搜尋報導，並將所有新聞（包含跨性別新聞）同時進行詞彙網絡分析，觀察使用這些詞彙的新聞與提到跨性別的新聞存在什麼關係和差異。同時本研究也期待以網絡分析成果呈現臺灣新聞超過半個世紀編織而成的跨性別知識圖景。此章分析暫不考慮報導或跨性別本身隨著時間可能的變化。

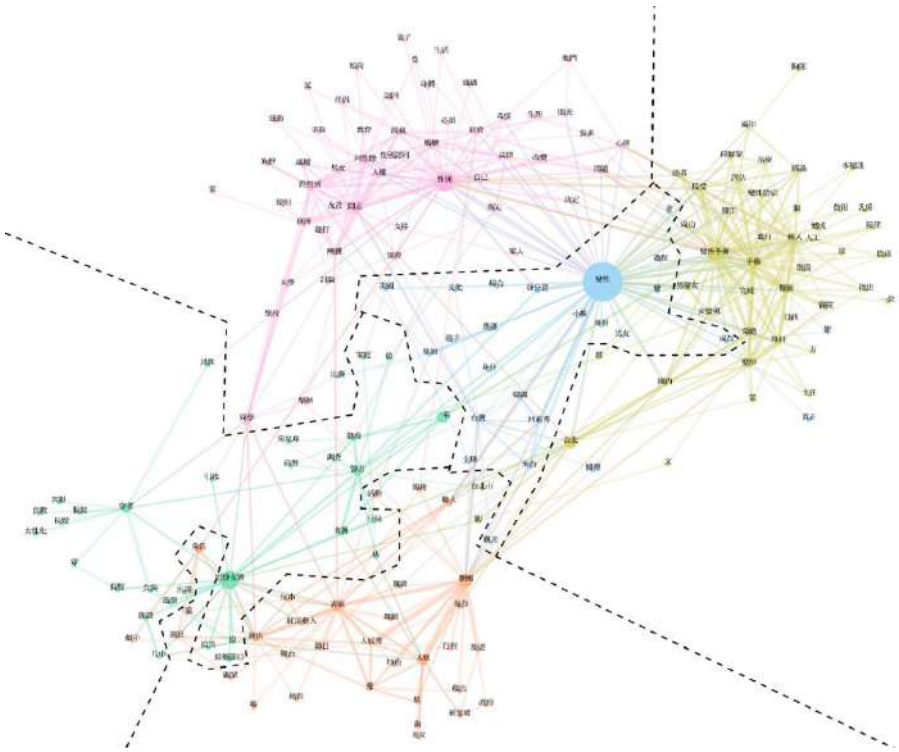
本研究首先選取不男不女、陰陽人、人妖、第三性、男扮女裝、變性、女扮男裝六個詞彙。這些都是坊間常見、學者在說明跨性別概念時會列舉的詞。其中變性指透過手術或施打賀爾蒙改變身體的性別；陰陽人指的是生來擁有兩種器官的人，也稱雙性人；人妖一詞最早出現在晚明文獻，指不符社會性別規範者，多用以稱呼從東南亞來台的變性或扮裝表演者；男扮女裝、女扮男裝常出現在表演藝術或影視娛樂相關的新聞，描述重點看似在於外在裝扮，實則隱含一套「衣服—本體」、「外在—內心」之二元對立預設；第三性、不男不女則如同字面上的意義，指稱男女兩性之外的他者，是性別二分框架的產物。為剔除與主題無關的新聞、使搜尋條件更符合需求，將變性改為變性人 or 變性男 or 變性女 or 變性手術作為實際搜尋資料庫時所用的關鍵字。[17]

由於初步搜尋後，以不男不女、陰陽人、第三性得到的新聞數量較少，而女扮男裝在 2006 年後不再出現在新聞中，因此本研究選擇變性、人妖及男扮女裝做為與跨性別比較的詞彙。一共得到 5077 篇新聞，約二百二十萬字左右。時間範圍橫跨 65 年，從

[16] 不過事實上該篇新聞提到的是「跨性別穿著」，這也顯示跨性別的意思其實仍有一定的彈性。
[17] 若以「變性」搜尋，會將「可變性」、「多變性」等與主題無關的詞彙列入，故改以「變性人 or 變性男 or 變性女 or 變性手術」搜尋。

1951 年直至 2016 年。取得文本建置語料庫後，如前章所述進行斷詞及詞頻分析。接著將跨性別、變性、人妖、男扮女裝作為第零代進行詞彙網絡分析，成果見（圖三）。

粗略觀察該網絡分析成果圖，可發現大致分成五個詞彙群，上方粉色群是跨性別為第零代的衍生詞，以性別為中心，即上一章考察的跨性別新聞；右上方黃綠色群以手術為中心，是變性為第零代的衍生詞，但變性本身屬於中央的藍色詞彙群，同時位居整個網絡的中心；左下角青綠色群以男扮女裝為中心，下方是人妖、泰國、表演為中心的紅色群。為方便閱讀，以下以局部圖分別呈現五群，可以與網絡整體互相參照。點大小則為 Betweenness Centrality，代表一個點出現在所有點到點最短路徑的次數，高 Betweenness Centrality 代表該點為各點間重要的橋樑 。如圖所示，變性即為在全文詞彙網絡中連結各群的點。

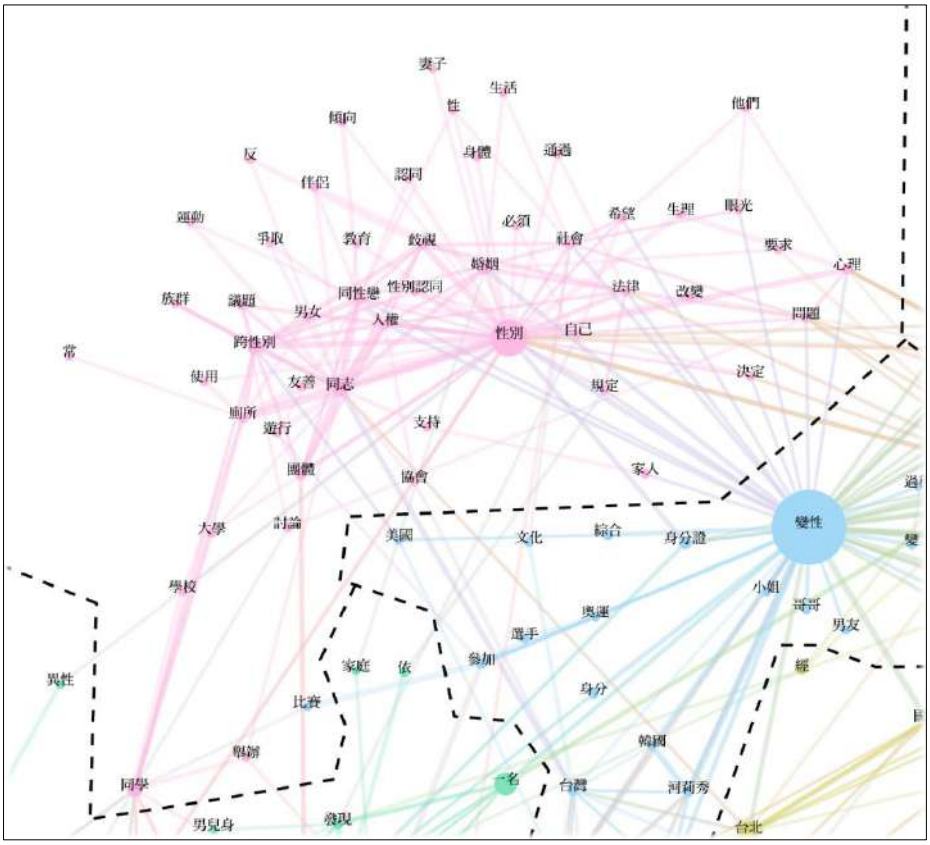


Filter	**Appearance**
Node (184): Freq. > 220 Edge (814): T-score > 5	Node: — Color: Modularity Class — Size: Betweenness Centrality Edge: — Weight: T-score

圖三　跨性別、變性、人妖、男扮女裝新聞詞彙網絡分析成果

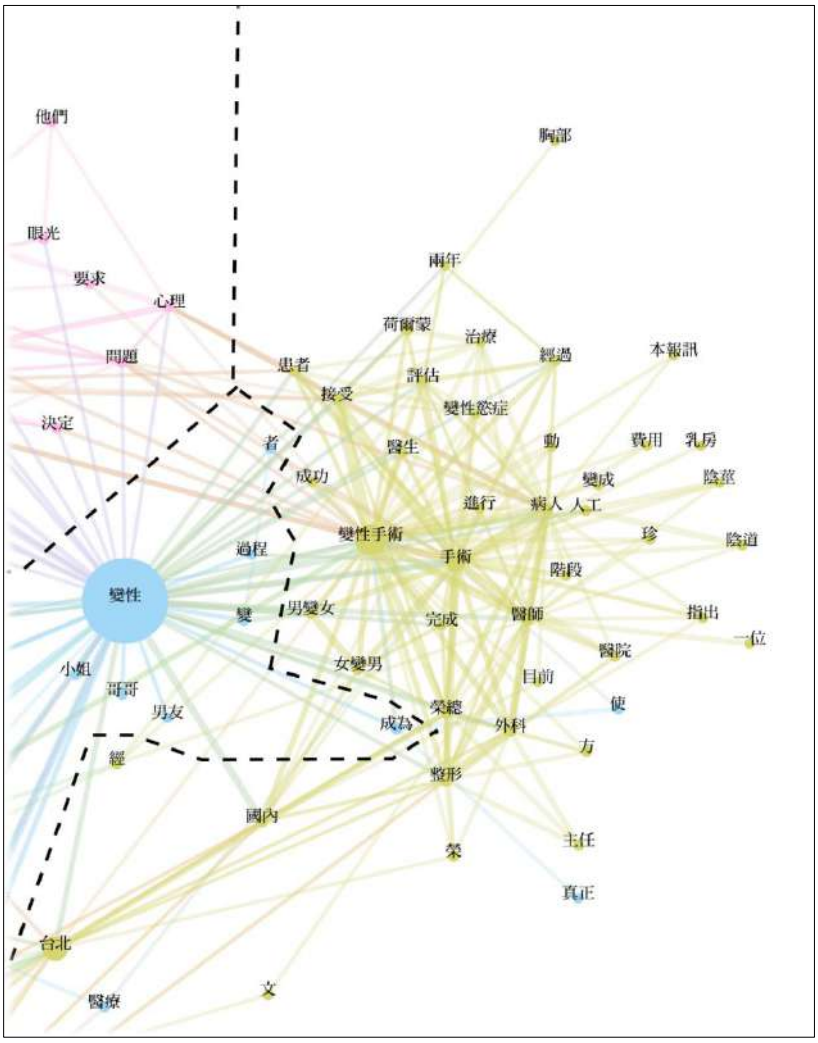
（圖四）是以跨性別為第零代衍生的詞彙群。跨性別與友善、同志、同性戀、人權、性別認同、歧視、婚姻、廁所之間的關係較緊密，呼應前一章提到跨性別是倡議色彩較高、對描述對象較友善的詞彙。

值得注意的是其中 Betweenness Centrality 最高的詞彙是性別，就詞彙群內部看來，意味提及跨性別的報導幾乎很難未使用性別一詞；然而，性別在整體網絡僅作為此一詞彙群內的橋樑，顯示新聞在敘述男扮女裝、人妖的主題時可能較少強調「性別」此抽象的概念。



圖四　詞彙網絡局部圖：跨性別群

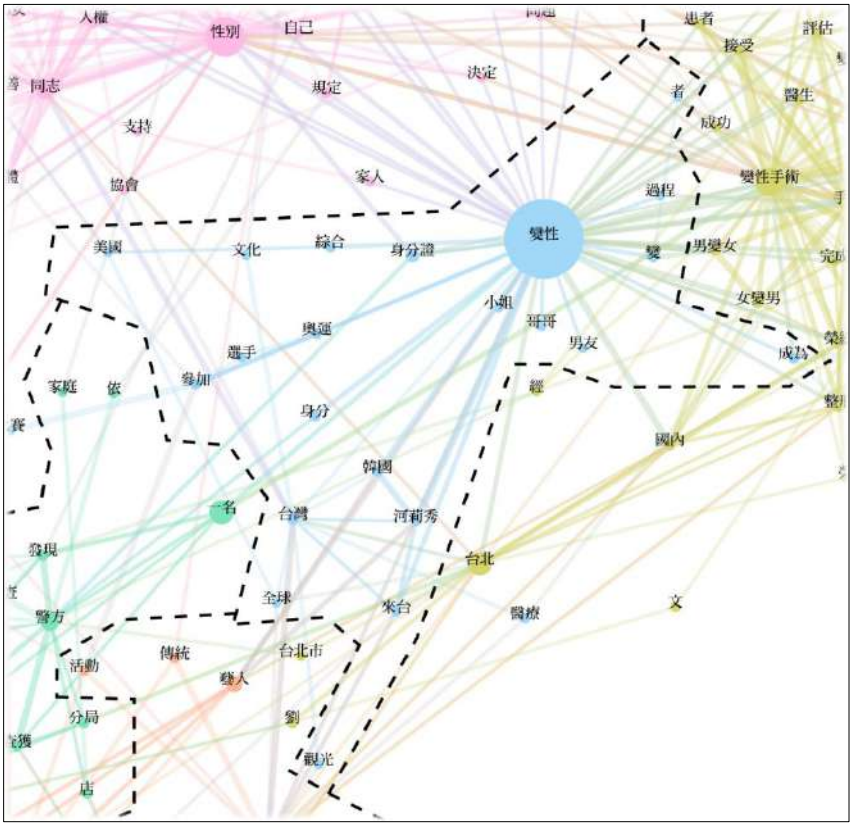
至於該詞彙群的邊界，自己、法律、改變、問題、決定、心理、生理、眼光位於右方，代表在提及變性、手術的新聞也常出現，與變性衍生的詞彙群關係較緊密，同樣的，這些詞彙組成的敘述顯然也較少在男扮女裝、人妖的新聞出現。另外同學、學校、大學位於左下方，就整體網絡觀察可見應是由於同學和穿著、男扮女裝、表演的共現關係較強所導致，即新聞報導與校園相關的跨性別議題時除了廁所也常提到扮裝表演和穿著，而這類報導與變性的關聯則較低。



圖五　詞彙網絡局部圖：變性手術群

（圖五）是以變性為第零代衍生的詞彙群。變性手術、手術、醫生、醫師、接受、患者、病人、變性慾症、評估、荷爾蒙、治療彼此關聯緊密，頻繁出現互為第一、二代的情況；變性手術、手術的 Betweenness Centrality 較高。該詞彙群代表討論變性相關醫學資訊的報導，不過顯然只是變性新聞的其中一個面向。

至於邊界，台北、整形因為與泰國的共現關係也不低而位於該詞彙群下方、左下方；詞彙群正左方與變性連結的詞彙則有男變女、女變男、成為、過程等。



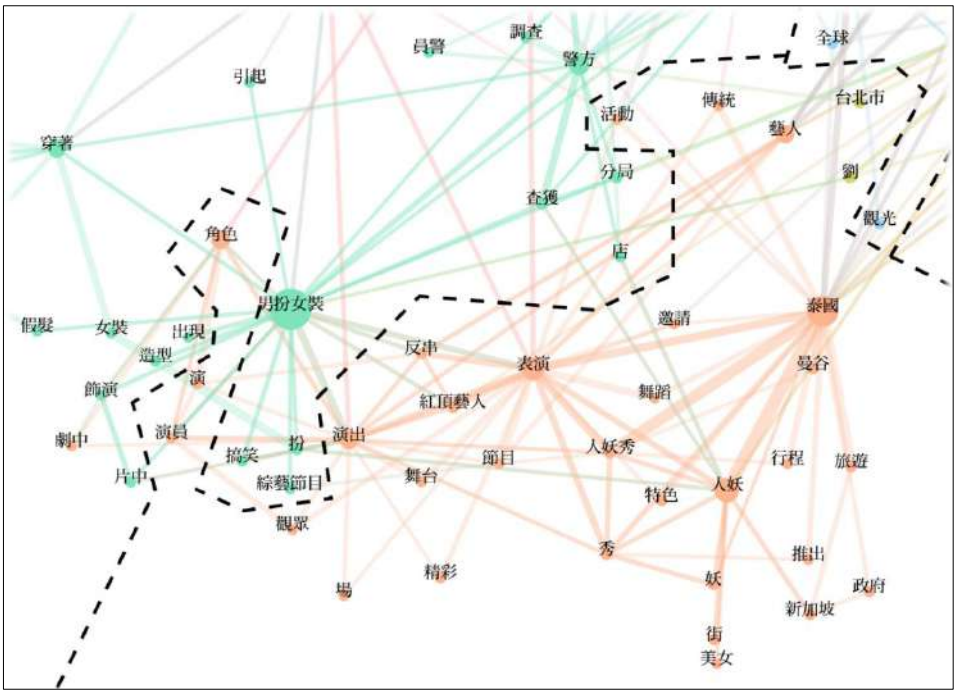
圖六　詞彙網絡局部圖：變性群

（圖六）是位於整體網絡中央的詞彙群，包含變性及其他關鍵字衍生的詞彙，不像其他詞彙群有關聯緊密的幾組詞彙。

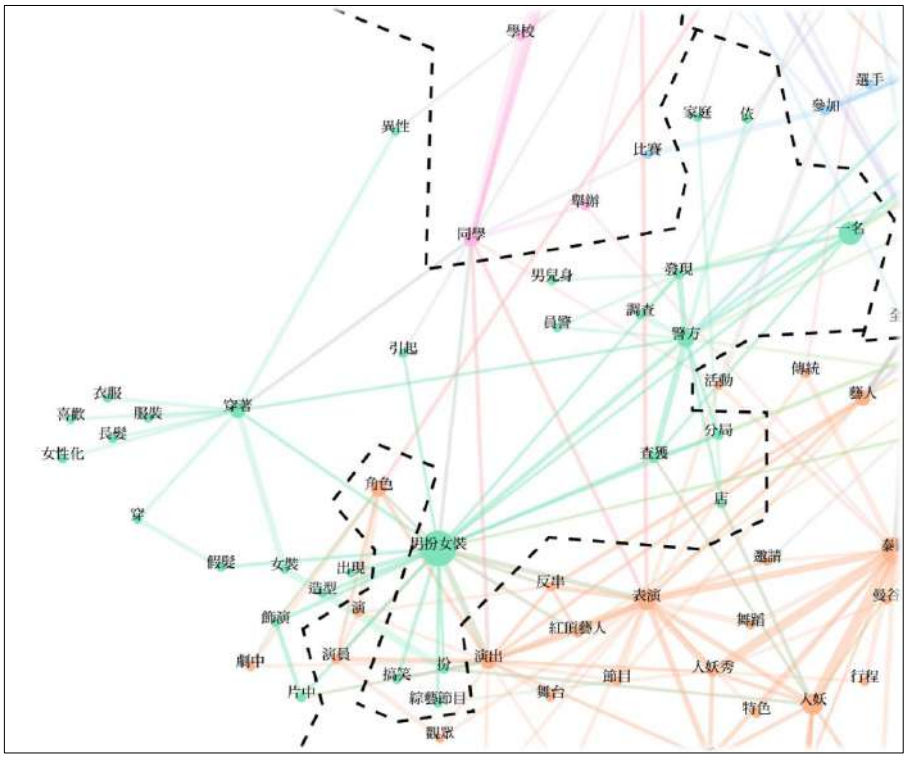
變性的點相當大，Betweenness Centrality 明顯是整體網絡中最高的，這代表與變性一起出現在報導中的詞彙較廣泛，與提及跨性別、人妖、男扮女裝的新聞常使用的詞彙範疇皆有重疊。本研究也嘗試刪去變性，發現整個網絡就變得相當鬆散、分群明顯。

該詞彙群由變性衍生的詞彙包含參加、選手、奧運、身分證、身分、小姐、男友、河莉秀（韓國變性藝人）等等，（圖五）比較起來，這些詞彙展現變性在醫療之外的面向，包含完成手術後的變性人欲參加奧運、更換身分證或作為演藝人員來台等在社會上的種種情況。

另外，位居整體網絡正中心的臺灣、全球 Betweenness Centrality 也不低，是連結各詞彙群的橋樑。



圖七　詞彙網絡局部圖：人妖群



圖八　詞彙網絡局部圖：男扮女裝群

（圖七）、（圖八）分別是以人妖、男扮女裝為第零代衍生的詞彙群，是整體網絡中關係最緊密的兩個詞彙群，故一起說明。

（圖七）當中 Betweenness Centrality 較高的是人妖、表演、泰國，右方的詞彙包含泰國、旅遊、政府、推出、行程，明顯與泰國觀光發展有關；至於左方的表演、反串、紅頂藝人、演出、舞台、節目、觀眾、演員等詞彙則與臺灣的娛樂表演之文化產業有關，也是與（圖八）詞彙群的邊界。

（圖八）當中 Betweenness Centrality 較高的是男扮女裝，相當靠近與（圖七）詞彙群的交界，與其關聯緊密的詞彙如扮、搞笑、綜藝節目、飾演、片中、劇中、造型等顯然與臺灣影視產業的新聞有關。左方則有一組及中描述男扮女裝本身的詞彙：穿著、服裝、衣服、長髮、喜歡、女性化；右上方的詞彙接近整體網絡的中心，大致看起來可能與犯罪、社會事件的報導有關，包含警方、男兒身、發現、調查、查獲、員警、分局、店等。

對於整體詞彙網絡尚有許多可深入詮釋的方向，在此僅再提出兩點討論：

其一，跨性別、變性在新聞涉及的範疇較接近，男扮女裝、人妖的關聯也很緊密。從報導大致看來，前兩者強調主體自我認同及對理想身體的追求，後兩者則偏向他人眼光下的形象；這個分野多少呼應陳薇真提出的白天與黑夜之比喻。提及跨性別、變性的新聞多談論作為公民的權益，是主體發聲、能見光的白天；不過跨性別者能否在報導中真正說出自己的話又是另一回事。而男扮女裝、人妖相關新聞多與表演有關，是混雜不可說、不願說之事的黑夜，表演者是被觀看的，其吸引力正來自社會窺視「異常」的獵奇心理，媒體報導多少與此有關。

其二，變性本身作為橋樑，其作為第零代的衍生詞被分散至各群，可能因為除了與醫療、手術有關的報導，新聞也常敘述變性的其他的面向。這並不是在說變性比起跨性別在臺灣的新聞中更像一把大傘，但提示研究者進一步考察變性。

五、　引入歷史視角：變性的顯著詞分析

上一章分析高達五千篇的報導，呈現跨性別相關新聞的多樣性及詞彙間的網絡關係，卻無法呈現超過半個世紀以來臺灣的新聞、社會可能發生的變化。而新聞文本往往具有反映當下的特性，尤其本研究觀察的報紙每日更新，時間標記清楚，既值得也十分便於運用以進行語料庫分析。因此這一章嘗試以歷史研究視角出發，結合語料庫取徑的顯著詞分析（Keyness Analysis）[18]，期待能發揮新聞材料於時間面向的價值。

如前所述，幾乎近世紀末臺灣媒體才逐漸採用跨性別一詞，這與美國同時期跨性別

[18] Scott, M. & Tribble, C., 2006, Textual Patterns: keyword and corpus analysis in language education, Amsterdam: Benjamins

運動的開展應有緊密關係，在此先將目光移回臺灣本地。那麼，跨性別出現前，新聞如何稱呼及再現與今日所謂「跨性別」有關的人群？陳薇真《臺灣跨性別前史》以「前史」為題，正是意識到不能僅關注跨性別一詞，而將焦點置於早於跨性別存在、社會用以描繪挑戰性別二分、身心一致等框架之人的詞彙；書的三個主題「變性」、「第三性公關（男扮女裝陪侍）」、「人妖」，恰對應本文上一章挑選與跨性別對照的關鍵字。其中，變性主題無論在陳薇真書中或資料庫中涉及的時間跨度都較大，且本研究實際操作語料庫分析後發現詞彙顯著現象也較明顯，故以變性為例開展進一步討論。

搜尋變性（變性人 or 變性男 or 變性女 or 變性手術）共得到 1499 篇新聞。為了進行顯著詞分析，本研究以 1993 年為界，分出前、後期，前期有 372 篇；1993 年以後則有 1127 篇。而「前期顯著」，是指相較於 1993 年以後，該詞在 1993 年前更容易出現，反之亦然。顯著詞分析結果見（表四）。

本研究以跨性別首次見於報導的 1993 年為分界線，並非在暗示跨性別一詞或相關概念的出現對於媒體如何再現變性存在影響；恰好相反，是期望透過觀察變性新聞在此時間點前後的差異，推測臺灣語境中的跨性別在什麼樣時機浮現，即在提問：世紀末以前，台灣媒體、乃至整個台灣社會對跨性別相關的主題（如變性）已經積累了哪些討論？

表四　變性前後期顯著詞分析

顯著詞列表（局部）	
前期顯著	後期顯著
珍	河莉秀
謝尖順	利菁
病人	變性手術
變性慾症	同志
醫師	泰國
治療	韓國
施行	南韓
台大醫院	龍唐
正常	林國華
精神科醫師	
醫院	
醫界	
榮總	
法律	

首先觀察顯著詞出現的人名，前期的謝尖順、珍以及後期的河莉秀、利菁、龍唐、林國華為新聞人物。謝尖順為 1953 年臺灣第一個接受變性手術的案例，珍代表的則是 1980 年代起刊載於《聯合報》，由變性人珍口述的口述記錄；河莉秀於 2002 年首次出現於新聞中，通常被稱為「韓國變性藝人河莉秀」，因此韓國也成為後期顯著詞之一；利菁於 2004 年首次見於新聞中，被稱為「購物台天后」，變性手術則是在述及其過往才會提及；龍唐是泰國著名的變性拳王，也是模特兒、演員；林國華則與後期顯著的其他三人不同，是家境貧苦、不幸失業但曾獲社會贊助至泰國變性的市井小民。

前期有高曝光度的兩人以接受變性手術的患者形象出現，而後期有高曝光度的三人，則是出現於影視娛樂版的藝人。乍看之下，此變化似符合變性人從進行手術到能以術後的身體及身分開展新生活的歷程，暗示變性人在技術條件上可能越來越容易走到手術完成後的下一步。然而走下手術台的變性人未必受社會接納，仍可能遭到歧視、壓迫，或者發現變性並非自己所想的美夢成真：在後期顯著的另一個人名，是即使完成變性手術、成功換發女性身分證，仍因生活壓力而選擇自殺結束生命的跨性別者林國華。

另外，前期顯著的詞彙都跟醫學方面有關，如病人、變性慾症、醫師、治療、施行、台大醫院、精神科醫師、醫院、醫界、榮總。在後期這類有關醫療的詞彙相對不顯著，僅有變性手術一詞。也就是說，1990 年代之前，臺灣的新聞曾對變性的醫療層面——進行手術——有許多討論。透過關鍵字檢索，發現新聞內容包含手術可行性和評估流程、變性手術是否符合道德等等。醫療人員常受訪發言提供專業建議，並指出法律爭議，因此醫師、精神科醫師為前期顯著詞，例如：

> 最實際的一個難題是屬於法律上的，由於動這種手術必須先割除乳房及子宮，而我國法律規定，除非是為挽救病人生命，醫生不得隨意割除病人的器官，否則要負「傷害人體」的刑責。[19]

> 李聖隆說：只要手術不出問題，病人不提出告訴，醫師為病人動變性手術後，不會造成嚴重刑事問題。但是，這種手術會引起許多民事糾紛，包括戶口及身分的性別，各種證件上性別與權益的關係，娶嫁時的特殊身分，及是否妨害兵役法等。…如果手術失敗，病人可據以控告醫師，請求法院以重傷害罪處理。過去，國內一位著名婦產科醫師曾在為病人做了類似的手術後，即因手術失敗，遭到病人控告。[20]

報導提及變性牽扯許多問題，在法律規範還沒完全完善之前容易引起糾紛。這讓醫界普遍對於是否進行手術感到為難。而具有指標性的大醫院如榮總、台大醫院在 1980 年代

[19] 賴淑姬，〈新聞網外 第三性的祈望〉，《聯合報》，1975 年 9 月 11 日，第 12 版（聯合副刊）。
[20] 本報訊，〈法律問題比什麼都複雜〉，《民生報》，1981 年 3 月 17 日，第 4 版（文化風信）。

末率先開啟嘗試，獲得媒體高度關注，因此成為前期顯著詞，例如：

> 榮總這一陣子先後做了幾次變性手術，結果都相當成功，燃起不少變性慾症病人的希望。事實上，變性手術已經不能算是特殊的尖端手術，在技術上也沒有那麼困難，只是過去囿於國情和國內的醫療制度，許多變性慾症病人因而吃了不少苦。
>
> 過去，國內教學醫院的整形外科幾乎不為病人動這項手術，使許多病人不得不轉而求助於私人醫院，但私人醫院的主刀醫師又多未受過完整的整形外科訓練，手術結果可想而知。……後來，少數教學醫院的整形外科醫師，或許因為實在無法推辭病人的懇求，曾實施過這項手術，但在態度上仍然不敢公開。對於變性手術，醫界仍有一股反對的力量。
>
> 醫界既已承認變性慾症這種疾病，又發展出變性手術與術前的評估標準，應已沒有再反對它的理由。對變性慾症病人而言，手術既是唯一解決病痛的措施，又為什麼不能享有合法而且合格的治療呢？……衛生署正式同意這項手術，榮總挺身實施，相信對病人絕對是有益無害的。[21]

除了手術本身，變性慾症、病人顯示早期醫界尚未將變性去病理化。事實上，今日變性慾症仍然是醫界在使用的術語，但報導多已不用這個詞彙，這或許暗示媒體已經意識到這些詞彙具有病理化的負面意義，但更有說服力的解釋應是新聞已經較少對此有多著墨。即在 1990 年代臺灣的醫界及媒體大量進行對變性手術的討論，確立了「變性手術」的概念，使之成為後期顯著的詞彙。而後變性手術的過程及科技已經「不是新聞了」。不過，臺灣的變性手術真的盡善盡美嗎？這又是另一個話題。

還有一個後期顯著詞是同志，從內容來說可分成兩類，一種是強調某些同志性別認同與生理性別不符，故進行變性手術，另一種則是在談論多元性別議題時將變性人是為同志族群內部的一支。這與第三章跨性別新聞分析提到的狀況類似，由於變性被視為跨性別的其中一員，在今日同志運動的論述中跨性別相對變性更常被與同性戀並提，提及跨性別的年度新聞數量也在 2010 年之後超越變性。

整體而言，以 1990 年代為分界來觀察變性相關的新聞，在此之前醫界完成了手術相關的討論，報導也經歷從醫療轉向演藝、社會接納的過程。而這個轉變多少說明了為何在上一章的整體網絡中變性衍生出兩個詞彙群，以及與變性共現關係強的詞彙較分散、與其他關鍵詞衍生的詞彙群皆有疊合，乃是由於變性相關的報導涉及面相較廣泛的緣故。

---

[21] 李師鄭，〈一分鐘短評 讓病人享有合法又合格的醫療〉，《民生報》，1989 年 1 月 6 日，第 23 版（醫藥新聞）。

本研究在此方面的發現多可呼應陳薇真的質性研究，其指出 1988 年後臺灣變性的標準流程逐漸確立，2000 年後開始出現變性社群，與新興同志運動結合，相關新聞論述轉向國際人權的討論等；透過顯著詞分析所發現的跡象都可支持以上的論點，揭示語料庫分析取徑對於研究跨性別的可能貢獻。

六、 結語

本研究以臺灣新聞中的跨性別為題，同時也是嘗試以語料庫分析取徑、網絡分析來觀察報導以幫助批判論述分析的嘗試。根據初步進行詞頻分析、詞彙網絡分析的結果，跨性別此一詞彙從 1993 年起開始出現在臺灣的報導中，通常具有較正面的意涵，與人權、同志運動、社會議題有關，是倡議者的用語，同時相當受歐美地區的影響。

若僅以跨性別為關鍵字搜尋材料來理解台灣的跨性別議題，可能遺漏其他豐富面向，如「變性」就是理解臺灣跨性別相關新聞與論述的重要主題。而在 1990 年代前後，無論質性研究或顯著詞分析都顯示變性的討論重心逐漸從原本的醫療轉向其他主題。

不過，詞彙網絡分析顯示，媒體再現的跨性別和變性並未完全涵蓋何春蕤等學者定義的「跨性別」族群。跨性別與變性在新聞的意義範疇較為接近，是屬於公民權利的討論；人妖和男扮女裝在新聞中被呈現的情況較接近，兩者相對變性、跨性別仍有一段距離、暗示這些詞彙具有不同的階級位置或屬性，須深入檢視。影視與娛樂或許為「人妖」及「男扮女裝」創造了較不受非議、無關道德的空間，但仍屬於「非日常」的範疇。甚至可能如陳薇真所說，成為變性、跨性別者白天爭取權益時遺落在黑夜的姊妹。

以上都只是對於跨性別議題非常粗略的推論，且主要侷限在男跨女、男變女、男扮女的人群。至於取徑，本研究認為 Paul Baker 對英媒進行的研究方法同樣能適用於觀察臺灣的報導，即以詞頻、關鍵字檢索、共現等語料庫分析方法，協助進行新聞的批判論述分析。同時，本研究還加入以共現為基礎的詞彙網絡分析，除了視覺化地呈現詞彙間的共現關係、打開探索詞彙意義範疇的新地圖，也藉此引入網絡分析方法。儘管分析五千篇報導地詞彙網絡將 65 年的新聞置於同一個平面，缺乏時間變量，透過將不同時期的新聞分別繪製詞彙網絡、比較不同時期的詞彙網絡應能改善這點。此外，本文也試圖以顯著詞分析引入歷史研究的視角，目前僅以 1993 年為分期來比較前後期顯著詞，若透過其他方式分期，應能達到更細緻、更豐富的觀察。

參考資料

Mike Scott & Christopher Tribble, Textual Patterns: keyword and corpus analysis in language education. Amsterdam: Benjamins, 2006.

Paul Baker, Bad wigs and screaming Mimis: “using corpus-assisted techniques to carry out critical discourse analysis of the representation of trans people in the British press”, in Christopher Hart, Piotr Cap (Eds.), *Contemporary critical discourse studies*. London: Bloomsbury, 2014.

Ulrik Brandes, A Faster Algorithm for Betweenness Centrality, in *Journal of Mathematical Sociology* 25(2) (2001), 163-177.

Vincent. D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10), p.1000.

田思怡編譯,〈愛她？害他！科學證據：轉化治療造成傷害 歐巴馬：別再強迫治療性向〉,《聯合報》, 2015 年 4 月 10 日，第 A19 版（國際）。

本報訊,〈法律問題比什麼都複雜〉,《民生報》, 1981 年 3 月 17 日, 第 4 版（文化風信）。

何春蕤主編,《跨性別》。桃園：中央大學性／別研究室，2003。

李師鄭,〈一分鐘短評 讓病人享有合法又合格的醫療〉,《民生報》, 1989 年 1 月 6 日，第 23 版（醫藥新聞）。

林佳緣,〈跨性別媒體再現與主體解讀之分析研究〉。世新大學性別研究所碩士學位論文。2006。

范振和、徐庭揚,〈認知、情感、心胸 都多元 花蓮彩虹遊行 上街呼喊幸福〉,《聯合報》, 2014 年 9 月 28 日，第 B1 版（宜花・運動）。

郭文平,〈字彙實踐與媒介再現〉,《新聞學研究》125 期。2015：臺北。

陳大鵬,〈性議題 跨領域研究 新視野 中央大學昨獲贈英國新書 何春蕤：每一本都令人流口水〉,《民生報》, 1999 年 8 月 11 日，第 06 版（文化與藝術）。

陳薇真,《台灣跨性別前史：醫療、風俗誌與亞際遭逢》。臺北：跨性別倡議站，2016。

陳麗婷,〈性別變更困難 在英他算男生 來台他變女生〉,《聯合晚報》, 2013 年 11 月 21 日，第 A11 版（健康）。

鄭宏斌,〈爭婚姻平權 她們大方站出來〉,《聯合報》, 2013 年 7 月 12 日，第 A8 版（生活）。

鄭語謙,〈學生反應兩極 世新設友善廁所 男女共用〉,《聯合報》, 2011 年 10 月 27 日，第 A6 版（生活）。

劉利貞,〈漢堡王挺同志 推彩虹套餐〉,《經濟日報》, 2014 年 7 月 3 日，第 A8 版（國際企業）。

賴淑姬,〈新聞網外 第三性的祈望〉,《聯合報》, 1975 年 9 月 11 日，第 12 版（聯合副刊）。

鍾蓮芳,〈同志遊行 台北街頭性別大解放 變性人、雙性戀者、SM 團體全出籠 要平等、要小孩還有公民權〉,《民生報》, 2004 年 11 月 7 日，第 A3 版（生活新聞）

聯合知識庫 http://udndata.com/（檢索日期：2016 年 8 月）

大埔之歌：臺灣主流報紙中的「土地徵收」

王章逸*、闕河嘉**

摘　要

本研究藉由分析臺灣主流報紙關於「土地徵收」報導，探討新聞媒體在土地徵收事件中所扮演的角色。土地徵收在台灣已行之有年，大抵以經濟發展為論述基調。人權團體往往會批判臺灣主流報紙在土地徵收議題報導上，淪為政府宣傳的工具。然而，在 2010 年苗栗縣政府強行徵收大埔農地引發抗爭事件之後，新聞媒體報導開始重視土地徵收之人權議題討論，包括土地正義、居住正義。主流新聞媒體似乎不再獨尊經濟發展為地方開發的論點。

土地徵收與抗爭是一個具有貫時性特徵，並且帶有衝突價值的議題。因為主流新聞媒體對社會大眾認知的影響力，實有必要檢視主流新聞媒體在此議題報導的價值取向。再者，主流新聞媒體在 2010 年大埔事件後，改變土地徵收議題報導的固有偏好，主流新聞媒體的報導內容與價值的轉變，在臺灣近年土地徵收爭議事件、鄉村社會發展論述的關係為何，乃是本研究的旨趣。

本研究採用語料庫為基礎的研究取徑，以台灣四大報紙（聯合報、自由時報、中國時報、蘋果日報）在 2008 至 2015 年間，對土地徵收的報導編輯成為語料庫，進行批判分析討論。研究焦點為：（1）2008 年至 2015 年之間，台灣四大報紙在土地徵收事件的報導趨勢為何？（2）以大埔事件為界，主流報紙報導土地徵收事件時有什麼轉變？（3）大埔事件如何影響報紙媒體的報導？

研究結果顯示四大紙的土地徵收報導數量與該年土地徵收爭議事件密切相關。大埔事件確實影響四大報對土地徵收報導方式；大埔事件發生之後，土地徵收的報導的引述對象和論述框架趨於多元。最後，雖然四大報紙報導因大埔事件開始的社會運動使得立場趨於多元且重視人權討論，但是報社對於後續土地徵收議題的報導的仍顯示其特定的價值立場，特別是以結合政黨等政治語言混淆土地徵收的合法性討論。

土地徵收的得與失並非單一論述可以詮釋，本研究以語料庫的方式檢視臺灣四大主流報紙報導，梳理其在臺灣紛擾的土地徵收現象所扮演的角色。

關鍵字：土地徵收、土地抗爭、臺灣主流報紙、語料庫研究、大埔事件

* 國立臺灣大學生物產業傳播暨發展學系研究生。

** 國立臺灣大學生物產業傳播暨發展學系副教授，通訊作者，Email: hchueh@ntu.edu.tw。

An Effectiveness of Dapu Incident：A Corpus Content Analysis of Eminent Domain in Taiwan's Mainstream Newspaper

Chang-Yi Wang*, Ho-chia Chueh**

Abstract

This research examines changes of mainstream newspapers coverage of eminent domain after the Dapu incident in 2010. Taiwanese government's power of eminent domain has long been used to acquire property for public use, justified by community development discourse. Mainstream media often acts as a state apparatus in reporting eminent domain issues, and result in critiques from human rights group. Yet, it seems that mainstream media has changed its stance since the Dapu incident in 2010, i.e. more attention on debates of land justice and housing justice issues, rather than mere community development concern.

We combine methods of corpus-based analysis and critical discourse analysis in analyzing coverages of four mainstream Taiwan newspapers (United Daily News, Liberty Times, China Times, Apple Daily) between 2008 and 2015. Our research questions are 1) what is the trend of these newspapers coverage on eminent domain? (2) what are characteristics of these coverages before and after the Dapu incident? (3) how does the Dapu incident affect the mainstream media report?

The results show that the number of eminent domain news each year corresponds to eminent domain dispute in that year. The Dapu incident has affected the ways in which mainstream news media represent eminent domain; that is, after the Dapu incident, the news of eminent domain started to cover more sources of information, adopt more frames, and pay more attention on human rights issue. Despite of this change, eminent domain represented in mainstream news tends to complicate it with political party rivalry, and obscure discussion of legitimacy of eminent domain. This research contributes to understanding of how Taiwan's mainstream newspaper represent eminent domain and whether social movements such as the Dapu incident has affected the stance of mainstream newspaper.

Keywords: eminent domain, land conflict, mainstream newspaper, corpus-based method, Dapu incident

* Master Student, Department of Bio-Industry Communication and Development, National Taiwan University.
** Associated Professor, Department of Bio-Industry Communication and Development, National Taiwan University. (Corresponding Author). Email: hchueh@ntu.edu.tw.

一、研究背景

本研究藉由分析臺灣主流報紙關於「土地徵收事件」的報導，探討新聞媒體在土地徵收事件論述的改變。土地徵收在台灣已行之有年，自台灣的經濟發展從農業轉向工業、高科技產業後，政府主導之工業區、科技園區與「發展」劃上等號，因此，近年來地土地徵收報導之中，大多會出現計畫園區的開發，新聞媒體的土地徵收報導也大致以經濟發展為論述基調。然而近年來在許多反徵收團體、台灣農業陣線相繼成立後，土地徵收事件的報導似乎有不同的聲音出現。自 2010 年苗栗縣政府強行徵收大埔農地、引發抗爭事件開始，報導中開始出現土地徵收之人權議題的討論，包括土地正義、居住正義。主流媒體似乎不再以經濟發展為單一之價值論點。本研究以這樣的社會環境作為基礎，探討主流媒體在土地徵收事件的報導的轉變。

台灣的土地徵收由來已久，直至近年來才設立專法規定。在 2000 年前土地徵收相關的法源分散於都市計畫法、平均地權條例、產業升級條例等，直到制定「土地徵收條例」後才有土地徵收的專法。其立法宗旨雖為「為實施土地徵收，促進土地利用，增進公共利益，保障私人財產，特制定本條例」，但有需多土地徵收之爭議被提出。首先，「公共利益」定義的擅用，影響地主權益（徐世榮、廖麗敏 2011）；第二，原本土徵條例補償的辦法「公告土地現值加成補償」有失公平，被徵收戶往往不能獲得相對應的補償（陳瑩真，2004），最後都市計畫法的執行以以及社會對地方政府開發計畫園區的憧憬，讓土地徵收事件的數量更盛（蔡偉銑，2014）。針對上述的質疑，政府也在 2012 年公布新的土地徵收條例辦法，改以市價徵收作為補償的原則。

各地自救會以及台灣農村陣線藉由抗爭，在新聞媒體報導中有更多土地徵收論述的空間，因此，「土地正義」的價值思考也越發被大眾所知道。「土地正義」包含多樣意義，在公共利益的大旗下，也應檢視資源分配合理性，考量土地資源獨特性；除此之外，也針對經濟發展、政府與人民權力對等之決策程序等進行反思（蔡培慧，2011）。在這樣土地意識發展下，關乎土地徵收的社會運動越發蓬勃及受到大眾所重視，尤其苗栗 2010 年 6 月 9 號怪手毀田、2013 年 7 月 18 日強拆張藥房兩次事件更獲得許多報導與關注。

土地徵收鑲嵌了政治性、經濟考量與人民權力等多種價值，報紙媒體則是這些立場價值之戰場之一。非政府組織如台灣農業陣線、大埔自救會、竹北璞玉自救會等常是土地徵收事件的消息來源，而政府亦會對其立場在報紙上刊登政令之宣導，例如 2013 年 7 月 11 日四大報之頭版廣告皆由苗栗縣政府所刊登，由此可見報紙媒體仍為議題爭鋒相對之場域。報紙雖然在其他媒體平台如電視、網頁、行動裝置的影響下，閱報率逐年

地下降，然而報紙除了仍有一定數量的訂閱者，報社也積極數位化，發行免費閱讀的網路報紙，以及可以使用行動裝置閱讀的系統。雖然閱讀報紙的人數下降，但報紙經由數位化，點閱率卻相對的上升，報紙實際上仍有一定的閱讀量，實仍具備研究之價值。本研究為探討臺灣主流報紙對土地徵收報導之轉變，以《聯合報》、《自由時報》、《中國時報》、《蘋果日報》作為主要研究對象，收集四大報 2005 年至 2015 年「土地徵收事件」相關的報導，以語料庫研究方法中的共現詞、顯著詞、關鍵詞檢索分析文本，並且以批判論述分析檢視報導中的權力關係的運作，以此回答大埔事件在台灣土地徵收的發展中帶來什麼影響？而主流媒體又如何因此改變、或者沒有改變？

二、文獻回顧

(一)台灣近代農地徵收之發展

台灣近代的經濟發展由農業轉向工業為主，因此農地釋出給予工業成長的空間。由政策與法令觀之，於 1960 年代頒布的《獎勵投資條例》是因工業而進行土地徵收的濫觴，而後分別有 1965 年的《加工出口區設置條例》、1979 年《科學工業園區設置管理條例》、1990 年《獎勵投資條例》，最近的則是 2010 年改版的《產業創新條例》（徐世榮，2016）。這些工業、高科技產業園區的畫設，大多是以農地作為釋出來源，而這樣的產業轉型則是以經濟效益作為主要考量，經濟價值主宰產業發展之主軸（鄭欽龍，1988、劉泰英，1988）。在這樣的意識下，2000 年修正的《農業發展條例》鬆綁了農地所有權的限制，農業一直以來的「農地農有」堅持轉變為「農地農用」，甚至出現農地的管制阻礙農業發展的評論（黃樹仁，2002），除此之外，雖然特定農業區仍有不得隨意開發的限制，然而《農業主管機關同意農業用地變更使用審查作業要點》仍大開農地開發的門戶，竹南大埔科學園區用地的徵收法源便是來自於此。

在上述政府的規劃與條例改革之下，土地徵收在學界有相當多的討論。研究的取向大致可分為三個類別，一為法學界對土地徵收條例之適用性、法理學之討論（陳立夫，2008；陳文貴，2003；廖學能，2014）；二為地政學者針對土地徵收之賠償、執行等做討論（周信燉，2004；戴秀雄、李立達，2007），最後則是社會政治經濟、土地正義的討論（黃樹仁，2002；李素蘭，2010；徐旭，2014；廖本全，2014；徐世榮，2015）。其中，本研究欲探討之土地正義相關概念於 2010 年後才有較多討論，而土地正義的論述，大抵是在討論土地徵收之公益性、必要性、補償制度的失衡與不公平、抨擊開發與經濟發展至上，資本浮濫凌駕於農地保護及居住正義等土地的多元價值之上等（鍾麗娜、徐世榮，2013）。

土地正義的論述主要來自於徐世榮（2010）、蔡培慧（2011）、詹順貴（2011）等人。

這些學者大多為近年來反對土地徵收的 NGOs 的參與者，當媒體報導土地徵收相關事件時，他們成為報導中訊息的來源之一，土地正義與人權的論述也因而可以與土地徵收連結，賦予土地徵收不一樣的詮釋。對應近年來之土地徵收爭議事件，大埔事件最為受到關注，然而觀察大埔案受到媒體報導的狀況，2010 年 6 月 9 日大埔發生怪手毀田事件，公民記者大暴龍於 13 日刊發的報導《當怪手開進稻田中》於網路引起軒然大波。相較而言，主流報紙媒體僅有聯合報與自由時報在地方版進行報導，其他主流媒體包括電視則毫無聲響，彷彿這件事情沒有發生過一樣（莊豐家，2011）。在怪手毀田事件之後，土地徵收的社會輿論達到高點，同年度的七月十七號更是發起「凱道農民守夜行動」。以下為近年來較有爭議之農地浮濫徵收案例整理：

表一、台灣近年較具爭議之徵收案件

地點	開發案
新北淡海	淡海新市鎮特定區第二期
桃園大園、蘆竹	桃園航空城特定區計畫案
新竹竹東	變更新竹科學工業園區特定區主要計畫
新竹竹北、芎林	台灣知識經濟旗艦園區（璞玉計畫）
苗栗竹南大埔	變更新竹科學園區竹南基地暨周邊地區特定區
苗栗後龍灣寶	後龍科技園區
台中后里	台中科學園區三期
台中烏日	烏日溪南產業特定區計畫案
台中大雅	中部科學工業園區台中基地附近特定區計畫
彰化二林	台中科學園區第四期二林基地
彰化田中	台灣高鐵彰化站

資料來源：台灣農村陣線

媒體在一連串的篩選之後所做的報導，才能將社會運動成為「事件」，媒體是社會運動傳播思想的重要管道（Gamson & Modigliani, 1989; Pride, 1995）。而在台灣近年來的土地徵收報導中，大埔事件似乎是一個轉捩點，土地徵收這個議題開始因此成為主流媒體關注的焦點，台灣長久以來的土地徵收歷史似乎有了新的篇章，以此作為基點，本研究首先想要探討的是在 2008 年至 2015 年間，土地徵收報導的趨勢為何？藉此了解台灣近期的土地徵收報導狀況。

(二)社會運動與主流新聞報導

社會運動必須藉由大眾媒體的報導以增加運動訴求的散播。Lipsky（1968）認為社會運動包含了運動者、傳播媒介、大眾、抗爭對象等四種元素。此外，Gamson（1975）

亦認為非政府組織、弱勢團體等需藉由媒體獲得社會支持、引發大眾討論，並且藉此團結內部意見與成員。朱慕涵（2007）針對「保留樂生療養院」社會運動，分析媒體於倡議活動之功能在於可進一步形成民意壓力，來影響政府對樂生療養院之政策。媒體對社會運動而言是相當重要的資源，透過媒體傳達訴求後，才能進一步匯集民意、影響政策。

社會運動與媒體之間的關係，可見王嵩音（1997）探討原住民還我土地運動在報紙媒體報導與再現之情形，其研究結果顯示不同報紙所使用的消息來源有很大的差異，且各家報紙大多在抗爭發生後才會有報導產生；翁秀琪（1994）對照婦女運動之社會真實，以及報紙對婦女運動報導所建立之媒介真實兩者之間的差異，顯示民營報紙媒體常有「衝突化」之刻板印象化過程；由此可知，抗爭的場面是媒體在報導社會運動時偏好的畫面與因子（孫秀蕙，1994），除了抗爭之外，組織化、有利的政黨關係、有社運團體支持且迫切推動的政策四個元素為影響報導的因素（Amenta et al., 2009）。近年來的台灣農民運動，則可見林如森（2014）探討「一一二三　與農共生」此抗爭，原本在大眾傳媒中，農會、信用部被畫上黑金、派系等符號，然在社運組織的資源、專業程度、協調能力的匯集下，取得媒體的近用權，進一步改變社會大眾對農會的認知。

雖當代公民新聞崛起與社群網站之風行，讓社會運動得以用另一種管道散布資訊與聚集聲音（莊豐嘉，2011；陳佳君，2015），然我們亦不能否認傳統主流媒體對型塑社會大眾意見，及與社會運動團體雙向影響之角色。以蔡培慧（2010）針對台灣農村陣線行動與組織之反思，文中提及主流媒體如何受到社運團體如農陣，以及一連串事件的影響，而改變其論述方式以「圈地」這樣古老卻反映社會政商壟斷的辭彙，描述土地徵收及產創條例的運作：「真實是一場社會行動。從論述而起的爭戰，建基於對社會現實的分析，也要回到社會現實的變革與否加以檢驗…運動團體已介入論述的壕溝，然而，我們知道這一場爭戰的近距拉扯，來回折衝仍在持續，面對情勢、冷靜分析、有機集結、持續介入，才有趨近真實的可能」（引自蔡培慧，2010）。蔡於文末針對農陣行動的反思，足見社運組織作為一異議團體，可透過其行動影響媒體的表述，成為一重要節點與樞紐。

媒體與社會運動是一相互依存之關係，過去亦有許多研究分析媒體如何報導社會運動（Gamson & Wolfsfeld, 1993、鄭婉婷，2013、張讚國、劉娜，2015）。記者需要素材，而社會運動需要被看見，兩方隨著社會環境、輿論、價值觀改變而有高度的互動。媒體是社會運動極力拉攏的夥伴，然而主流媒體卻常被詬病為「政府的傳聲筒」（胡元輝，2007），成為政府傳聲筒的原因，除了可能是服膺於權力，再加上記者常迫於截稿的時間壓力，依賴特定消息來源如記者會的公關稿，因而缺乏多方的查證（Dunwoody, 1997），造成官方的觀點與說法，最常充斥載報紙或者電視畫面上（臧國仁、鍾蔚文，2000）。除此之外，劉華真（2008）的研究也指出，媒體需要依賴企業的廣告收入，這樣的關係

則顯示媒體某個程度受控於財團、企業等第三方的利害關係之下。主流媒體於這樣政治、經濟環境中，既不能得罪政府，財源也需要依賴資本方。

觀察近年來土地徵收的論述中，「土地正義」在大埔事件後成為社運團體論述的核心之一，然而，這樣的訴求是否亦能在主流報紙上傳達？由以上文獻的回顧與觀察出發，本研究第二個問題想要探討的是主流報紙在報導土地徵收事件的論述中，蘊含了甚麼價值取向？以了解大埔事件以及自救會的訴求，對於土地徵收報導有什麼影響。

(三)社會運動的議題建構

社會運動透過媒體傳達其訴求，然而，媒體又是如何影響群眾？本研究從議題建構觀點出發，探討媒體報導的建構效果。議題建構理論主要探討媒介與社會政治權力間的關係，Lang and Lang（1991）討論水門事件中「消息來源」與「新聞室」的關係，分析消息來源如何建構議題，讓新聞的報導與消息來源所希望的方向相符。由此可知議題建構理論問題核心在探討「媒介與社會政治權力之間的關係」（Semetko et al. , 2013）。而社會運動報導中議題建構的社會角力，或可從報導中的消息來源中看出端倪。林怡瑩（2004）整理1980年國內傳媒再現社會運動的特徵，發現媒體不僅傾向醜化社會運動，官方的消息來源亦始終是媒介論域中最強勢的發言者，而異議團體的聲音則極為邊緣化。簡曉娟（2011）針對高捷泰勞事件的研究也指出，該事件報導主要消息來源大多是官方、資方（勞委會、高捷與其他民間企業）和仲介業者，而抗爭主體的泰勞或者非政府組織的勞工團體觀點則較為邊緣化，此外，報社「商業導向」的經營方式會排擠外勞議題的再現空間，使得記者「特定衝突化情節」的報導偏好，造成報導「事件化」的呈現。

除了議題建構之外，框架亦為新聞研究中常用來探討媒體意識形態的概念。媒體在報導事件時，會以一個價值架構來模塑觀眾的認知，而這種方式常具有社會與政治的效果。媒體是社會真實的一部分，透過將各種權力的論述接合，形塑大眾所認知的社會真實，並影響社會真實發展的方向(Gurevitch, M. et al., 1982)。關於社會運動的框架，Snow& Benford（1992）提出社會運動的目標，是提出一套重新認知世界的參考框架，讓任何群體雖然沒有直接經驗被壓迫的事實，但能夠指認壓迫並視之為不義，進一步採取批判與反抗的態度，最後喚起參與者的熱情與信念。

將社會運動與媒體框架連結，Gamson（1984、1988、1995）指出，媒體中對於事件的詮釋框架是相互競爭後的結果，對於社會運動而言，成功的框架建構在於社運組織能否將組織框架成功轉化為公共定義，吸引民眾與新聞媒介接受，因此，行動者往往會打造「不公義框架」，目的在於促使群眾產生一種道德憤慨(moral indignation)的政治意識。在台灣相關的研究方面，彭慧蕙（1998）、劉于禎（1999）對勵馨反雛妓運動的研究中

也發現選擇性的以「兒童人權」的訴求框架，使反雛妓運動取得一定的社會發言地位，而林常富（2009）以「830嗆馬大遊行」做電視新聞報導的框架分析，分析結果發現因自身對於四權的角色認知，在報導中新聞事件時常將責任歸屬於政府，並常以個人化的事件將新聞擬人化，增加新聞的戲劇性。

大埔事件作為台灣土地徵收的重大事件是土地徵收報導的轉捩點。Pride（1995）「事件中心社會運動理論」(event-centered social movement theory)的觀點指出，社會運動的「關鍵事件」，常具有吸引大眾注意的強烈效果，並會因此加速社會對問題的集體定義過程。以議題建構跟框架的概念，我們檢視隨著大埔事件興起的土地正義與居住正義等人權討論，是否影響主流報紙的土地徵收報導？而大埔事件所引起之漣漪到底如何擴散？

三、研究方法

為檢視近年來土地徵收事件報導的轉變，本研究以「語料庫文本分析」與「批判論述分析」作為研究方法。論述分析一直以來都被詬病有刻意挑選支持論點資料的缺點（Widdowson, 1995），然而，電腦科技及語言處理進步所產生的語料庫文本分析方法，則讓批判論述分析得以使用大量的資料，並從中汲取特定的語言模式來進行分析。過去語料庫文本分析常被使用在語言學的領域，但語料庫近年來也常被用來進行不同主題的論述分析，例如男同志的公共論述（Baker, 2006a）、英國報紙中的難民論述分析（Baker, et al. 2008）、美國報紙中的北韓（Kim, 2014）。上述的研究都是針對特定族群的媒體再現進行分析，找出新聞媒體之中的論述模型，藉此觀察媒體中對該族群特定的論述。

除此之外，本研究也以批判論述分析來解析土地徵收事件中的權力關係。批判論述分析是與論述分析最大的不同在於，批判論述分析主要的目標是將文本之中隱藏的權力關係及意識型態揭露（Wodak & Meyer, 2009; Baker, et al, 2008），批判論述分析亦將論述視為一種社會實踐，因此，檢視論述中的權力關係可以進一步分析社會建構的過程，並且找出其中的失衡與不平等。雖然批判論述分析有其堅定的立場，但通常會因只進行少數文本的探索、僅由研究者主觀來挑選文章進行詮釋而備受批評（Mautner, 2009）。相較於批判論述分析，語料庫文本分析方法最大的優點是可以藉由電腦的運用，避免研究者的偏見（Baker, 2006b），除此之外，他也能讓我們觀察論述中的增值效果(incremental effect)，亦即找出語言隨著時間的演進，在論述建構中的轉變（Baker, 2006b）。

本研究以 2008 至 2015 年間關於土地徵收事件的報紙報導為分析文本，為符合構成「土地徵收事件」之要求，我們以「土地徵收」及「抗爭與各大相關事件（例如大埔、航空城）」作為關鍵字，使用「慧科大中活新聞網」、「聯合知識庫」、「知識贏家」蒐集台灣四大主流報紙《自由時報》、《中國時報》、《蘋果日報》、《聯合報》的報導，最後結

果共有 3,055 篇報導。

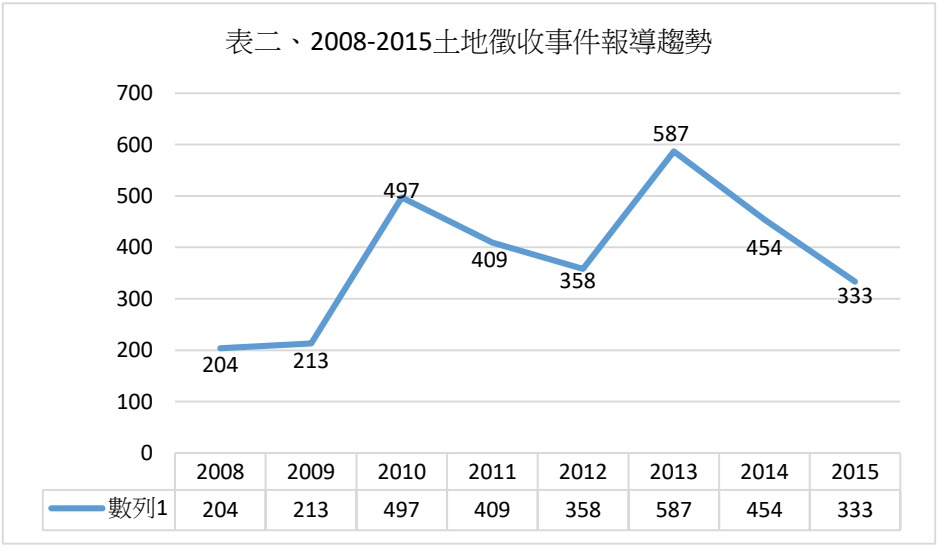
在語料庫分析軟體方面，本研究以「庫博中文語料庫分析工具」進行分析。不同於以往中文的語料庫研究，必須先以中研院的中文斷詞系統斷詞後，再以其他軟體進行分析，庫博整合中文斷詞與語料庫分析的功能，並且新增停用詞、自建辭典、同類詞等功能，讓中文語料庫的分析更為貼切與靈活。而在分析方面，本研究以 corpus-based 為主，經由媒體報導的觀察以及文獻的探討，以大埔事件作為一重要線索。本研究主要使用「共現詞」、「顯著詞」、「KWIC」這些分析方式進行。首先，我們觀察台灣土地徵收事件報導的趨勢，且為探求大埔事件對於土地徵收事件報導的影響，我們以 2010 年 6 月 9 號怪手毀田事件為界，將語料庫分為前半與後半，以顯著詞找到兩組語料庫的差別之後，再以共現詞與 KWIC 檢視報導面向的不同。除此之外，我們也將語料庫中所有提及「大埔」的報導提取出來，並以 2008-2015 的語料庫作為參照語料庫，找到大埔語料庫的顯著詞，藉由這些顯著詞，分析大埔事件對於土地徵收報導的影響。

四、研究結果

(一)報紙中的土地徵收

土地徵收要成為事件，其中一個因素是要透過媒體的報導，媒體是構築大眾認識抗爭的重要媒介。為了檢視本研究的主要探討對象「土地徵收事件」，我們以土地徵收事件相關的語彙建立語料庫。觀察台灣四大報 2008 年至 2015 年的土地徵收事件報導數量，大致的報導趨勢如下：

表二、2008-2015土地徵收事件報導趨勢

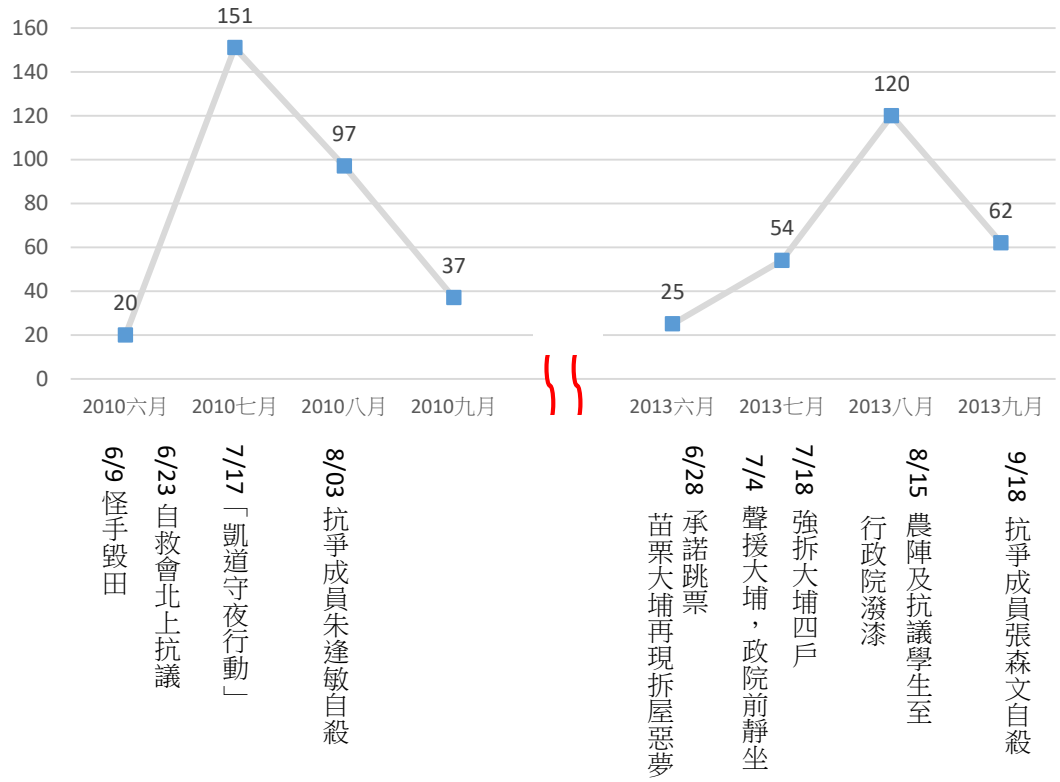


自 2008 至 2015 年，相關的報導數量共有 3055 則，在報導的趨勢圖中，可以發現 2010 年及 2013 年為報導的高峰。2010 年六月至七月為苗栗大埔事件爆發，自 6 月 9 號

怪手毀田事件發生、6 月 23 號大埔農民至總統府及監察院抗議、7 月 17 號「台灣人民挺農村 717 凱道守夜行動」、8 月 3 號大埔朱逢敏阿嬤喝農藥自殺，大埔的徵收事件在台灣造成轟然巨響，2010 年中 6 月 10 號至八月底的土地徵收報導數量就有 268 則，三個月的報導數量就佔了該年度的五成。

2013 年的高峰，前半年多為「桃園航空城」計畫的報導，另一重大事件則為 7 月 18 日「苗栗縣政府強拆大埔四戶」，再加上 9 月 18 日大埔四戶中張藥房的老闆張森文自殺，也為土地徵收事件投下一顆震撼彈。從七月的強拆民房，到九月因政府拆房的行動，導致主要抗爭者自殺，大埔事件的爭議性與話題性再度成為報紙報導的焦點。

表三、2010與2013大埔事件報導數量與事件對照表



報導數量急遽上升的原因，各地自救會的發聲以及台灣農村陣線帶起具規模的抗爭，符合新聞報導抗爭所需的「衝突性」，加上怪手毀田、農民夜宿凱道等具新聞性的畫面，還有前後都有人因抗爭自殺的戲劇性發展，新聞媒體因而緊追大埔事件的發展進行報導，除此之外，報紙具有公共討論平台的功能，因此成為支持或反對徵收意見論戰的匯集之處，造成在大埔事件後報導數量急劇增加。

除了大埔事件造成的報導高峰，從報導數量中也可以發現，近幾年土地徵收事件的

報導數量整體呈現上升的趨勢。觀察這些年份報導的事件，較為零星的徵收事件必然帶有抗爭的元素，而這些徵收仍以「事件式」的報導為主，亦即媒體僅陳述徵收事件發生的經過，卻沒有進一步追蹤後續發展或者結果。除此之外，有許多土地徵收重大的事件在近五年發生，例如桃園航空城的推動、台灣知識經濟旗艦園區、中科的二林園區、南鐵東移案等，而這也是導致後半段報導數量增加的原因。但以規模而言，這些爭議事件的開發規模都不比大埔事件小，且都有自救會及 NGO 組織抗爭行動，激起的報導卻有那麼大的落差？就新聞性而言，大埔事件誇張的毀田、拆屋，再加上有兩位當地的居民因犧牲，是導致大埔事件大量曝光的原因。

報紙媒體的報導大抵上會反映社會事件的發展。從報導數的趨勢中，我們得知近年來重大土地徵收事件帶動了媒體對土地徵收的報導數量，藉由檢視媒體報導數量增減，發現大埔事件是土地徵收報導大量出現的濫觴，而媒體的報導大致上也跟隨社會的徵收動態呈現。

(二)報導中的變與不變

1. 大埔事件前語後的顯著詞比較

土地徵收事件的報導數量在大埔案的影響下有顯著的差異，進一步，我們以大埔事件發生（2010 年 6 月 9 號大埔毀田事件）作為分界，分為前後兩個語料庫，並以中研院製作的平衡語料庫做為參照語料庫，檢視兩者的顯著詞彙。

表四、大埔事件前顯著詞列表[1]

序位	顯著詞	序位	顯著詞	序位	顯著詞	序位	顯著詞	序位	顯著詞
1	縣府	10	公頃	19	抗爭	28	工程	37	縣長
2	地主	11	農業	20	住戶	29	**拆遷**	38	報導
3	**徵收**	12	**用地**	21	二林	30	**說明會**	39	**地價**
4	**土地**	13	**補償**	22	航空城	31	政府	40	**公告現值**
5	發展	14	特定區	23	**土地徵收**	32	白布條	41	**拓寬**
6	園區	15	高鐵	24	都市計畫	33	**重劃**	42	開發案
7	**區段徵收**	16	居民	25	自救會	34	市府	43	辦理
8	抗議	17	道路	26	計畫	35	興建	44	縣議員
9	中科	18	陳情	27	**徵收土地**	36	民眾	45	中科四

圖示	利益相關人詞彙	發展與計畫詞彙	抗爭相關詞彙	補償相關詞彙	徵地相關詞彙

1、2 大埔事件前後之顯著詞 Keyness 皆大於 700，p-value 皆小於 0.05

表五、大埔事件後顯著詞列表[2]

序位	顯著詞	序位	顯著詞	序位	顯著詞	序位	顯著詞	序位	顯著詞
1	徵收	10	農民	19	園區	28	補償	37	環評
2	土地	11	自救會	20	用地	29	劉政鴻	38	陳情
3	發展	12	市府	21	抗爭	30	道路	39	住宅
4	航空城	13	農地	22	桃園	31	政府	40	工程
5	地主	14	內政部	23	戶	32	捷運	41	苗栗縣
6	縣府	15	抗議	24	苗栗	33	徵收土地	42	案
7	大埔	16	土地徵收	25	都市計畫	34	民眾	43	開發案
8	區段徵收	17	居民	26	拆	35	市價	44	淡海
9	政府	18	計畫	27	徵地	36	拆遷	45	吳志揚

圖示	利益相關人詞彙	發展與計畫詞彙	抗爭相關詞彙	補償相關詞彙	徵地相關詞彙

表四與表五為大埔事件前後兩個語料庫前 45 個顯著詞列表。我們將顯著詞依照主題分類，共有「利益相關人詞彙、發展與計畫詞彙、抗爭相關詞彙、補償相關詞彙、徵地相關詞彙」六組。首先，利益相關人的詞彙，以政府方、被徵收方、自救會三方為主。就政府方而言，前後的差別在於後期的顯著詞出現中央部門「內政部」(14)」以及地方首長的名字「劉政鴻(29)、吳志揚(45)」，這顯示了報導中政府的層級由地方提升到中央政府，並且，重大徵收地區的地方政府首長較常被提及或者發言較受引用。在被徵收者方面，最多使用的詞彙皆為「地主(前排序 2，後排序 5)」，然而，在事件後的顯著詞中，則有「農民(10)」一詞出現，顯示在後期的報導農民成為較受重視的主體。最後，自救會一詞前後兩者的差別則在於序位，大埔事件前後自救會分別排名 25 以及 11，後期的排序明顯往前，代表在事件後的報導中自救會較受重視。

發展相關的分類，無論是事件前或後，「發展」一詞的排序都非常前面(排名第五與第三)，這代表報導徵收土地事件時，通常伴隨著發展相關的討論，而後續相關的詞彙則以不同時期的開發案為主。在此，我們雖可辨認不同時間的重點開發案，然而，報導中關發展討論的差異卻無從得知，後續我們將以共現詞與 KWIC 做更詳細的分析。

補償相關詞彙中，無論前後都出現「補償(前 13、後 28)」以及價格相關的詞彙，如事件前的「地價(39)、公告現值(40)」，事件後的「市價(35)」，這樣的轉變是因為土地徵收條例於 2011 年的修改，地主的補償從公告現值地價改為市價判定。最後，徵地相關詞彙的部分，事件前與後都有出現徵收、區段徵收等報導土地徵收時常出現的字詞。然

而，兩者的差別在於，事件發生後的被徵收物出現了「農地(13)」，這使我們聯想到利益相關人分類中出現的「農民」。在兩時期顯著詞的對比之下，農業的元素更為彰顯。

從兩時期的顯著詞中，可得知不同時間報導內容的轉變，這些顯著詞大抵不脫土地徵收報導中常見的用詞。然而從台灣產業以及農地政策的發展來看，1993 年的「農地釋出方案」及 2000 年通過的「農業發展條例修正」，顯示政策正逐漸棄守「農地農有農用」的原則（楊閎仁，2002）。產業發展則可見 1960 年代在「獎勵投資條例」下設置的加工出口區，以及 1991 年為推動新興工業而設置的「促進產業升級條例」等「國家重大計畫」，政府對工業的鼓勵及對農業的棄守在政策的對比下格外明顯。台灣輕農重工的社會環境，進一步影響土地的使用。

土地使用的決策常以個體效用最大化為原則，土地通常會用於產生最大經濟效益的用途之上（Irwin and Geoghegan, 2001、Nelson et.al, 2001）。然而，從關鍵詞的分析當中，儘管發展仍是報導的基調，被埋沒在計畫園區聲浪中的農民卻在大埔事件後成為報導當中的主體之一，在以工商業發展為主的台灣社會中，這樣的轉變格外引人注目。

2. 不同時期的「發展」

土地徵收是犧牲人民權益來促進公共利益，而在台灣的產業觀念中，所謂公共利益即是開發具經濟價值的產業特區，因此，發展的概念貫穿報導，而發展也是政府說服地方的主要說法。大埔案造成全台對土地徵收的注目，而後除了有一連串的倡議，甚至有法條的修改，而這些改變是否會影響報導中對於經濟的詮釋呢？在前述發展與計畫的顯著詞分析中，無法透過顯著詞辨別兩個時期的差異，因此，我們以全部語料庫的詞頻做為依據，找出「開發(3,664)、發展(1,998)、產業(1,060)、經濟(840)」四個詞彙做為發展的同類詞詞組，再進行此詞組的共現詞分析。分析結果如下：

表六、發展詞組的共現詞分析

分期	共現詞詞彙
大埔事件前 (2008/01/01 至 2010/06/08)	發展、地方、園區、帶動、專區、公頃、未來、土地、方式、整體、區段徵收、許可、精密、計畫、機械、取得、聯合、專用區、大埔美、科技、引進、觀光、就業、基地、航空城、工業區、繁榮、進駐、特定區、地區、模式、後續、加速、新市鎮、都市、實質、協會、農業、進行、時程
大埔事件後 (2010/06/09 至 2015/12/31)	發展、園區、計畫、台灣、整體、帶動、地方、新市鎮、工業區、專區、面積、都市、淡海、土地、未來、航空城、周邊、許可、加速、專用區、二期、聯合、進駐、永續、國家、促進、地區、區段徵收、引進、成長、區域、農業、繁榮、重要、招商、機場、投資、觀光、科學園區、工業、就業、二林、建設、需求、潛力

註：共現詞依 t-score 分數排列，且 t-score 皆大於 1.645 以上

觀察大埔事件前後發展詞組的共現詞，可發現報紙在講述發展時，多以各種計畫連結：「園區、專區、專用區、航空城、大埔美、工業區、新市鎮、淡海、二期科學園區、工業」，並且出現與之相關的詞語如「促進、地方、未來、繁榮、就業、加速、帶動、建設、成長」。這些詞彙皆傳達較為正向的意涵，其原因為在報導中大多會引述政府關於發展計畫的意見，因而有這樣共現詞的結果呈現。我們以 KWIC 回到內文做觀察，可發現所謂繁榮發展的概念，的確都是以各種計畫作為主軸，且發言人大多都是政府：

(此報導引述前民航局局長李龍文言論)國外**航空城**成功案例，大部分先由機場本身改造做起，藉由機場經營效率提高，帶動機場周邊區域的人流、土地及**產業**活動，進而進行機場周邊區域**開發**，帶動機場周邊城市的**繁榮發展**。（中國時報， 2008/09/25）

張國棟(彰化二林鎮鎮長)說，為了促進**地方繁榮**與**國家發展**，公所當初努力協調土地徵收以及「南投—彰林高壓電塔」的反對聲浪，只為了成全大局，且馬英九總統 98 年主持**二林園區**動土，**地方**熱情迎接，如今，政府一句話就說不做了，「我們好像被甩了一巴掌，情何以堪?政府威信蕩然無存，**經濟產業發展**究竟走向哪裡?」（聯合報，2012/03/22）

報導中的「發展」，多是計畫園區帶來廠商，廠商帶來就業機會，進而帶動地方的繁榮發展，就發展建設與繁榮的連結而言，兩個時期並無太大差別。然而，我們再將共現詞中出現的「未來」(在兩組內分別排名第七與第十五)加入 KWIC 的條件，探知兩個時期對於發展期待的差異。檢索的結果發現，兩個時期對於未來似乎有不同的想像。在大埔事件之前，政府官員對未來的詮釋，大多傾向於計畫園區帶來的經濟發展來討論：

縣府並發表聲明說，這項土地區段徵收一切合法，群創**未來**將興建 7.5 代或 8.5 代面板廠，創造兆元以上產值，**未來**將形成北群創、中友達、南奇美三大光電面板製造基地，帶動中下游產業發展，這是對苗栗縣非常有利的投資案，不做才對不起鄉親。(自由時報，2009/12/19)

在 2010 年之後，政府官員口中的未來，雖主要仍是在講述計畫園區的優點，但在報導中則出現不一樣的敘述方向。「未來」似乎不再只是被計畫園區綁架，在園區帶來的經濟發展之餘，也有其他面向的考量：

內政部長江宜樺昨天表示，現行土地開發制度值得檢討，**未來**土地開發涉及土地徵收，若徵收範圍內有住宅聚落，將要求申請人提出**安置方案**，並提出**耕種權益**的保障措施。(聯合報，2010/10/07)

彰化縣府建設處長陳文慶表示，「彰南產業園區」希望廠商進駐後能吸引年輕人返鄉就業……縣府**未來**會嚴格執行**空氣污染防制、水污染防治、生態環境維護**等環境保護對策，也會要求進駐廠商異味防制效率需達80%以上，並需規畫20到30公尺寬度之隔離綠帶或設施，維護環境品質。（聯合報， 2014/07/11）

由上述兩個報導可知，雖然同樣在講述土地的開發，政府在增加就業的說法之餘，開始注重包括環境、居住、耕種的權力。這樣的轉變，我們可以從關鍵詞分析中的利益關係人進行分析。在關鍵詞分析中，前後兩組皆以「縣府、地主」為首，其他政府相關的語彙則無太大不同，然而，在民眾與NGOs的部分則有明顯不同。表五中，「農民(排序10)」成為顯著的詞彙，而大埔事件前的被徵收人則是地主、居民、住戶，除此之外，事件後顯著詞的被徵收物中亦出現「農地(排序 13)」，由此可知，在大埔事件後的報導中，對於農業有較多討論，報導中才因此出現「耕種權益」。除此之外，NGOs（自救會）在兩組語料庫都有出現，然而大埔事件前自救會在關鍵詞排名第二十五，事件後則上升到第十一，報導中自救會的顯著程度明顯增加。自救會的立場與論述常與政府對立，報導中自救會比重的增加，代表在爭議事件中，人民組織的媒體近用權較受重視，除此之外，組織化除了讓爭議事件更容易見於新聞報導中（林如森，2014），連帶的也讓反對政府的論述被看見。因此，政府必須為此做出回應，不同於經濟發展的說法亦因此增加。

3. 土地徵收事件報導中的「農業」

地徵收事件大多是政府為了執行產業計畫而進行，而這些計畫為尋求土地來源且兼顧成本，較便宜且面積較大的農業區常成為計畫園區的用地。在關鍵詞中，兩個時期被徵收者最大的不同在於後期「農民」出現，除此之外，也出現被徵收物「農地」，由此可見在事件後，農業的元素似乎較常被提及。有了這樣的觀察之後，為探知兩時期對於農業態度的差異，我們以KWIC回到文本檢視。以全部報導的詞頻分析做為依據，找出農業的概念詞彙「農民(2375)、農地(1797) 、農業(647)、農村(434)」成為農業詞組，並以該詞組進行共現詞的分析，結果如下：

表七、農業詞組的共現詞分析

分期	共現詞詞彙
大埔事件前(2008/01/01至2010/06/08)	徵收、重劃、耕作、台糖、港務局、觀光、再生、耕種、抗議、聚落、工廠、財團、買、社區、附近、罰、農保、變更、農用、森林、保護帶、不滿、建地、湖山水庫、輔導、100、污泥、現在、住、維生、變更為、公告現值、工業區、祖先、公頃、一七、高速公路、價格、環中東路、市價、交會處、林務局、橫越、權益
大埔事件後(2010/06/09	徵收、大埔、凱道、耕作、夜宿、政府、優良、抗爭、苗栗、引發、團體、集中、事件、農委會、糧食、強徵、建地、相思寮、

至 2015/12/31)	生產、保護、土地徵收條例、農、農用、灣寶、再生、休耕、怪手、務農、總統府、台灣、重返、政策、保留、工業、耕種、凱達格蘭、農舍、苗栗縣、陣線、關心、犧牲、各地、吳敦義、農保

註：共現詞依 t-score 分數排列，且 t-score 皆大於 1.645 以上

無論事件前後「徵收」都是跟農業詞彙中貫串的主題，然而，兩個語料庫的共現詞分析結果則有很大的差異。2008 年至 2010 年在土地徵收事件中跟農業共現的詞彙，除了當時的事件或開發項目例如「湖山水庫、工業區、環中東路、工廠、堤防、財團」，也有因土地徵收而影響農民福利的「農保」。然而，表七中跟土地徵收較有關的共現詞，可以看到「公告現值、價格、市價」等詞彙，藉此，以 KWIC 檢視文本，我們得以推知大埔事件前農地徵收的抗爭，大多的訴求皆與價格有關：

> 目前**農地公告現值**每坪僅兩萬一千元，與市價每坪十四萬元價差七倍，他抨擊：「這跟用搶的有何差別？」（蘋果日報，2008/04/30）

> 蚵農林先生也說，中科沒有與當地居民溝通協調就強行徵收，且**徵收價格**還比以前**農民**承購的**價格**低，令人無法接受。（自由時報，2009/11/13）

回到當年的社會概況，本研究認為農地徵收常出現價格議題的原因，是因為在 2011 年修改土地徵收條例之前，土地的徵收價格都以「公告地價加四成」為主，也因此，當時與農業相關的土地徵收事件中，抗爭多是因為價格不如農民預期所導致。

在大埔事件發生後，對於農田的徵收，出現「強徵、犧牲、圈地」這樣較強烈徵收方式的敘述，同時也有「保護、保留」等呼籲式動詞的出現。共現詞中的「怪手、凱道、夜宿、苗栗、重返、總統府」等則多與大埔事件後，接續的抗爭有關。然而，我們從共現詞中也可以看到許多在 2010 年之前沒有過的討論，例如共現詞「糧食」，是農業價值中「糧食安全、糧食危機」的討論：

> 坐在凱道，抗議政府粗暴徵收**農地**，時隔一年仍無作為，要求立即修改「土地徵收條例」惡法。此外，農民也要求**農業**水資源**分配正義**及**糧食安全**的訴求。（聯合報，2011/07/17）

> 「台知園區」九十九%為特定**農業區**，現有**農地**達三百公頃，竹北周邊產業用地也早已供過於求，目前正值全國**糧食危機**之際，政府居然要毀壞良田，重現去年「大埔事件」。（中國時報，2011/10/22）

比較大埔事件前後與農業相關的土地徵收報導共現詞中，可以發現在事件後，農地的土地利用價值不再只有價格這樣單一的衡量標準，而是進一步延伸至糧食安全、糧食

危機，此時的價值並不僅只用金錢衡量，而是考慮農業無形的、意義上的價值。

第二個研究分析中，我們針對土地徵收報導的變與不變，以農業及發展的概念字彙進行分析。研究結果顯示，對經濟發展的詮釋仍是由科技園區、計畫區主宰，然在大埔事件後的報導中，我們可以看到不一樣的價值取向出現。討論土地徵收時，不再只有對未來美好的憧憬，而是出現對權益、環境的關心。而報導中的改變，我們從農業相關詞組的分析中也發現，前期以補償價格為抗爭導向的農地徵收，到近期不僅在徵收的事件中農業角色的重要性上升，也出現以農業價值的訴求，土地的使用不再獨尊經濟利益。

針對這樣的轉變，本研究認為大埔事件是為相當重要的轉捩點。因此，最後我們將針對跟大埔事件有關的報導做詳細分析與檢視，以了解大埔事件如何影響台灣的土地徵收事件報導，此外，也深入解析大埔事件後報導中，是否真的會訴諸人權相關的討論。

(三)大埔事件的再現與框架

第一個研究分析，顯示大埔事件在土地徵收報導數量的突出，而第二個研究分析則針對報導內容做探討，結果亦表示大埔事件後，報導面向已有所改變。我們從前面兩個分析可以得知大埔事件對於土地徵收報導而言有其重要性，然而，這樣具指標性的事件媒體又如何報導呢？

1. 新聞媒體中的大埔事件

本研究從 2008-2015 年的土地徵收報導的語料庫中，將提及「大埔」的報導提取出來，成為另一子語料庫，共有 633 篇報導。接著將 2008-2015 年全部的報導作為參照語料庫，找出大埔語料庫前七十五個關鍵詞並依照其特性分組，以識別當報導提及大埔事件時的顯著詞彙。

表八、大埔語料庫的關鍵詞分析

大埔語料庫關鍵詞(各分類按照 keyness 的分數排列)	
人	農民、劉政鴻、馬英九、吳敦義、吳揆、人民、四戶、行政院長、李鴻源、張森文、**徐世榮**、**學者**、江宜樺、**詹順貴**、**蔡培慧**、彭秀春、陳文彬、阿嬤、發言人、**教授**
地區	大埔、苗栗、竹南、凱道、相思寮、田寮洋
物	農地、怪手、農村、農業、事件、張藥房、農田、工業區、鞋、稻田、農業區
組織	政府、**自救會**、內政部、總統府、**台灣農村陣線**、農委會
行動	**修法**、土地徵收、拆、**圈地**、夜宿、徵收、聲援、保留、**毀**、**剝奪**
權益	**權益**、**必要性**、**土地正義**、**人權**、**公益性**、**利益**

其他	**土地徵收條例**、南鐵、政治、農、台灣、市價、爭議、徵收案、社會、利益、力量

註：顯著詞 keyness 皆大於 30，p-value 皆小於 0.05

扣除與大埔事件直接相關人物(例如苗栗縣長劉政鴻,當時的中央政府官員馬英九、吳敦義、李鴻源等人以及大埔事件被徵收人張森文、彭秀春等)，學者的角色成為報導的顯著詞，例如徐世榮、詹順貴、蔡培慧等。隨著高知識份子的加入，帶動土地徵收事件中非政府組織的話語權及詮釋深度，連帶在組織與其他分類中，「自救會、台灣農村陣線」跟「權益、必要性、土地正義、人權、公益性」也分別成為關鍵詞。

除了人物的差別之外，藉由行動分類的關鍵詞，更可以看出報導中對於土地徵收的態度。修法所言之「法」，是「土地徵收條例」以及「農村再生條例」的修改，報導中抗爭的訴求提升至法律的討論：應要求苗栗縣政府歸還侵奪自農民的大埔農地，並儘速完成「土地徵收條例」修法，保障人民財產權及生存權。(中國時報，2011-07-16)，在這樣的論述下，政府一直以來土地徵收爭議中「依法行政」的說法也遭到質疑：土地開發案依法行政 面臨挑戰（聯合報，2013/07/19)。

除了修法之外，在動作分類中「圈地、毀、剝奪」也帶有其他不同的意涵，在報導中，中世紀的「圈地」被拿形容政府徵收土地的行為，例如狀告神農五穀大帝 苗栗農民今夜宿凱道 抗議「圈地」(中國時報，2010/ 07/ 17)；「毀」則是因怪手毀田事件而被大量的出現，最後剝奪則常與權益連結，例如：台灣農村陣線發言人蔡培慧反問:「行政無法負荷，就能**剝奪**人民**權益**嗎?」(聯合報，2013/08/23)。

利用修法、圈地、剝奪等動詞的檢視，我們發現大埔事件報導跟所有的報導比較起來，似乎更注重權力的訴求，因此也出現相對應之「土地徵收條例、權益、必要性、土地正義、爭議、人權、公益性」等語彙，這些都顯示土地徵收報導取向的差異。其中，土地正義為 2010 年農民於凱道抗爭的口號，而自此之後土地正義便成為土地徵收事件中常被提及的概念，甚至影響政府對於土地徵收向來以補償為主的論述。土地正義向來多為 NGOs 的訴求，然而，爭議事件越來越多的情況下，政府也開始逐漸引用此概念：

> 內政部長李鴻源說，土地徵收是推動建設必要的手段，不可能廢除。但是，只要「長官」同意，他就能提出一套「**土地正義**」與「經濟開發」的土地徵收政策（聯合報，2013/09/23）

> 房地合一稅明年上路，工業區住宅和農地農用改革也持續在推動，國民黨主席朱立倫昨(15)日表示，**土地正義**以及社會公平正義的追求是國民黨創黨精神（中國時報，2015/07/16)。

2. 「大埔」與「南鐵」

我們在其他的分類中，發現「南鐵」竟排名於前端，因此進一步檢視大埔事件與南鐵東移地徵收案的關聯。土地徵收本就是一連串政治作用後的結果，在政府與地方派系利益交換的「恩庇侍從」之下，才得以進行土地徵收的計畫（徐世榮，2016），然而，大埔與南鐵的兩個事件的比擬，讓我們得以一瞥土地徵收中的政治元素。大埔與南鐵的兩案有較多的連結始於 2013 年 8 月 21 日，當時的內政部長李鴻源將兩事件相互比擬「李鴻源：大埔、南鐵類似的區段徵收」（聯合報，2013/08/12），自此之後，在南鐵案的爭議中便有更多政治相關的討論。我們整理語料庫中同時提及大埔以及南鐵案的報導，發現共有三十七篇，其中聯合報佔二十六篇、自由時報七篇、中國時報與蘋果日報各兩篇。台灣的主流報紙的政治立場向來明顯，從篇數來看，親近國民黨的聯合報佔據最多的報導數，在大埔事件已成為社會大眾公認之違反土地正義的 2013 年，將當時具爭議且是由民進黨執政區推動的南鐵案與大埔事件連結，其政治目的可見一斑。

除了報導數量的差異，從報導內容更可以看到聯合報連結兩事件的意圖：「台南正在上演大埔案翻版」，台南鐵路地下化路線東移須拆遷四百零七戶，反台南鐵路東移自救會認為違反居住正義（聯合報，2013/08/18）。除了以「大埔案翻版」作為比擬，聯合報甚且出現題目為「南方的大埔」的社論，該篇文章更是直接的將以民進黨作為攻擊對象，將政治元素融入土地徵收事件當中：

> 大埔案，九百餘戶同意，四戶有異議；在苗栗現場，不論大埔當地及苗栗全境，對縣政府幾乎一片頌揚之聲，反而是民進黨及農陣等外人緊咬不放，甚至升高至「今日拆大埔(四戶？)，明日拆(中央)政府」的地步。相對而言，南鐵案則有精華地段四百餘戶未能擺平，街市現場掛滿言詞辛辣的抗議標語，但農陣對南鐵案卻虛應故事，民進黨亦稱相關爭議只是「技術問題」（聯合報，2013/08/24）

相對而言，立場偏向民進黨的自由時報，提及兩爭議事件的方式卻有很大的差異：

> 民進黨發言人鄭運鵬表示，南鐵東移案與苗栗大埔案**差很多，不應拿來不當類比**，台南市政府針對南鐵案，持續和民眾溝通，甚至挨家挨戶拜訪，也做了很多安置措施與政策，希望南鐵團體能與市府做更多溝通。（自由時報，2015/05/14）

當政治性成為土地徵收報導的主軸，而大埔做為土地徵收中「惡行」的框架，我們得以藉此觀察政治爭鬥的過程。同日相關的報導中，聯合報以三篇報導詳細解釋大埔與南鐵案的差異，除了政府的說法之外，皆會引述自救會的說法：

> 『反南鐵東移』自救會長蘇俊文說，南鐵案與大埔案根本就是一樣，都是攸關人

權和土地徵收的問題。（聯合報，2013/08/22）

然而，自由時報卻以學者的敘述美化賴清德的徵收：

大埔案讓類似的徵收案都存在『不好的氛圍』，導致執行單位進退失據，事情也越來越難處理。台南市長賴清德如能更有耐心，與居民做更深入的溝通，將有機會樹立土地徵收的典範。（自由時報，2013/08/22）

當土地徵受事件的報導加入政治因素，為追求「平衡報導」的確會引述不同立場團體的說法，但比較兩報引用的方式以及選擇，立場似乎已先於平衡。

最後，本研究也發現大埔事件影響了許多土地徵收事件的報導。藉由南鐵以大埔作為比擬的報導方式出發，我們以「翻版、第二」等詞彙檢所，發現大埔事件已成為媒體報導土地徵收事件的模板：

和美開發案地主：縣府強徵民地 大埔案翻版（自由時報，2015/03/14）

地被規劃公園 地主抗議「大埔翻版」（聯合報，2013/08/29）

央北區段徵收 農民抗議大埔翻版（聯合報，2012/09/14）

同時也出現「大埔第二、下一個大埔」的比喻方式：

綠議員：徵地開發小心 別變大埔第二（聯合報，2014/08/13）

淡海新市鎮 下一個大埔？（聯合報，2013/09/23）

最後，我們回到大埔徵收案的發展檢視其轉變。大埔自救會於 2009 年開始就竹南科學園區的徵收案抗爭，當時並未受到社會大眾矚目，各家報紙也僅就當天的抗爭有單則的報導，然而在經過 2010 怪手毀田、農民夜宿凱道，2011 農民重返凱道、2013 強拆張藥房等事件後，新聞媒體大量且深入的報導使「大埔」成為「土地徵收疑慮」的代名詞，而這樣的現象，可以以 Fiske(1987)所提出的「水平互文性(horizontal intertextuality)」解釋，水平互文性意旨閱聽人會因為文本的相似性而產生疊加的狀況，因而在土地徵收事件的報導中，大埔做為重大的爭議事件，橫跨在不同的事件中產生的連結的作用。大埔的爭議眾所皆知，本研究透過語料庫及文本分析，顯現大埔對於後續土地徵收案報導的影響，大埔已成為土地徵收中揮之不去的陰影，但同時他也帶起土地正義這道曙光。

五、結論與討論

本研究旨在探討近年來土地徵收事件報導論述的改變。藉由語料庫文本分析的方式，

讓本研究得以用較大的文本基數，探討土地徵收事件的報導在大埔事件發生後的改變。研究分析的結果，第一，因政府的作為與抗爭過程的新聞性與戲劇性，大埔事件受到媒體大量的報導，儘管其他大型爭議徵收案的徵收規模遠大過於大埔，但大埔事件之衝突性及因此激起的討論，是造成大埔案在短時間報導量遠大於其他事件的原因。第二，經由關鍵詞與共現詞、KWIC 的分析，發現大埔事件前與後報導的取向稍有差異，整體而言雖仍是以產業發展為主，但大埔事件後有較多農業的討論，對於發展也會同時提及永續、權益的概念。第三，大埔案成為媒體在報導土地徵收時的框架，例如報導中常出現「大埔第二」的用法。並且深入檢視南鐵案與大埔案在媒體上的呈現時，發現政治力於此有明顯的展現。

過去有許多社會運動的媒體研究，多是探討社會運動如何較常出現於新聞媒體、社會運動哪些因子較常出現在報導中（Andrew & Caren, 2010; Amenta et al., 2009; Caren et al., 2011），例如議題、組織、行動方式、地區等變數。然而，本研究透過語料庫，與其他社會運動新聞研究不同之處在於，我們主要針對報導「內容」的發展深入探討，而非社會運動在媒體上的呈現，亦即，本研究著重的是土地徵收抗爭導致媒體報導內容轉變的作用。Gamson 及 Wolfsfeld 於 1993 年提出社會運動與媒體是一交互作用系統，社會運動影響媒體更關注運動的持續性、報導框架等。本研究的分析驗證了台灣的土地徵收抗爭的確影響了媒體的報導，發生大埔這樣關鍵的事件發生後，再加上自救會及農陣發起的數次大型的抗爭，使得媒體上土地徵收事件的報導增加，並且報導的取向也有所轉變。除此之外，本研究亦呈現台灣新聞媒體中土地利用的轉向。土地的使用常以最大經濟效益做為考量（Irwin and Geoghegan, 2001），然而，我們發現主流報紙在報導時，亦會提及土地權及農業重要性等意識，這樣的聲音也打破過去以價格考量土地的思考方式。

議題建構與框架的部分，大埔事件是近年來最具爭議之土地徵收事件，對於其他的徵收事件報導亦有很大的影響。過去的研究指出，當媒體報導社會運動時，政府是最常被引用的來源，亦即對於該事件媒體上的議題建構，主要是以政府作為發展軸心（林怡瑩，2004；夏曉鵑，2011）。然而本研究發現，相較於整體的土地徵收事件報導，在大埔事件相關的報導中，農民成為報導中最為顯著的主體，並且學者與 NGOs 也較常出現。過去的研究並無法支持我們的發現，但這卻說明了大埔事件的特殊性。在大埔事件的相關的報導中，以議題建構的觀點來看，媒體不再只以政府的資訊為尊，而是在報導中加入更多元的聲音。在框架部分，社會運動組織在媒體的報導中，建立框架以利訴求並不少見，然而我們發現大埔做為土地徵收事件的「關鍵事件」，除了加速社會對於該問題定義的過程（Pride, 1995），進一步成為代表「土地徵收惡行」的框架，這樣的現象在其他社會運動並不常見。

本研究以語料庫文本分析與批判論述分析進行土地徵收事件的研究，其研究限制，首先，受限於資料庫，文本資料的收集並不算完整，因而對於研究的分析或結果有一定的影響。再者本研究以主流四大報做為探討對象，然而，獨立媒體與社群媒體在當代社會亦具備相當的重要性，此為本研究無法顧及之處。因此，未來的研究方向或可從更多元的媒體管道進行分析，讓研究的結果更為完整。

本研究針對土地徵收事件報導的內容做了許多討論，並且解析大埔事件的重要性，最後，我們仍須正視的是土地徵收報導中各種權力的關係。前述我們已經分析南鐵徵收案報社的立場，而撰寫研究分析與結果時，正好發生高雄市果菜市場徵收的事件，儘管抗爭成員以絕食這樣激烈的方式抗議，但該事件在各大報紙媒體並無太多的報導。到底是什麼影響徵收事件的報導？是因為蔡英文政府上台嗎？抑或是因為南北的地區差異？在經過 2016 年藍綠兩黨政權轉移後，土地徵收的爭議並未因此止歇，儘管本研究的結果，發現報導的內容的確與過去有所不同，但土地徵收的取捨與執行仍主要由政府決定。「大埔事件」已過，台灣政府與人民是否能記取教訓，在期望經濟持續發展之餘，仍可以不忘大埔帶來的傷痛與損失。土地爭議的未來如何發展，仍值得我們細細探究與觀察。

參考文獻

Amenta, E., et al. (2009). All the movements fit to print: Who, what, when, where, and why SMO families appeared in the New York Times in the twentieth century. *American Sociological Review*, 74(4): 636-656.

Andrews, K. T. and N. Caren (2010). Making the news movement organizations, media attention, and the public agenda. *American Sociological Review,* 75(6): 841-866.

Baker, P. (2006a). *Public discourse of gay men*. London: Continuum.

Baker, P. (2006b). *Using corpora in discourse analysis*. London: Continuum.

Baker, P., et al. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society,* 19(3): 273-306.

Caren, N., et al. (2011). A social movement generation cohort and period trends in protest attendance and petition signing. *American Sociological Review,* 76(1): 125-151.

Dunwoody, S. (1997). Science writers at work. *Social meanings of news: A text–reader*. Thousand Oaks, Calif.: Sage: 155-167.

Gamson, W. A. (1975). *The strategy of social protest*. Homewood, IL: Dorsey Press.

Gamson, W. A. (1984). *What's News: A game simulation of TV news*. New York: The Free Press.

Gamson, W. A. (1988). Political discourse and collective action. *International social movement*

research, 1(2): 219-244.
Gamson, W. A. (1995). Constructing social protest. *Social movements and culture,* 4: 85-106
Gamson, W. A. and A. Modigliani (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology*: 1-37.
Gamson, W. A. and G. Wolfsfeld (1993). Movements and media as interacting systems. *The Annals of the American Academy of Political and Social Science*: 114-125.
Gurevitch, M., et al. (1982). Culture, society and the media. London: Routledge.
Kim, K. H. (2014). Examining US news media discourses about North Korea: A corpus-based critical discourse analysis. *Discourse & Society,* 25(2): 221–244.
Lang, G. E. and K. Lang (1991). Watergate: An exploration of the agenda-building process. *Agenda setting. Readings on media, public opinion and policymaking*, 277-289.
Lipsky, M. (1968). Protest as a political resource. *American Political Science Review,* 62(04): 1144-1158.
Mautner, G. (2009). "Corpora and critical discourse analysis." Contemporary corpus linguistics: 32-46.
Pride, R. A. (1995). How activists and media frame social problems: Critical events versus performance trends for schools. *Political Communication* 12(1): 5-26.
Semetko, H. A., et al. (2013). The formation of campaign agendas: A comparative analysis of party and media roles in recent American and British elections, Routledge.
Snow, D. A. and R. D. Benford (1992). Master frames and cycles of protest. *Frontiers in social movement theory*: 133-155.
Widdowson, H. G. (1995). Discourse analysis: a critical view. *Language and literature* 4(3): 157-172.
Wodak, R. and M. Meyer (2009). Methods for critical discourse analysis, Sage.
王嵩音。1997。*台灣原住民還我土地運動之媒體再現*。中華傳播學會 1997 年年會論文。
朱慕涵。2007。學生社會運動議題倡議策略之研究-以青年樂生聯盟推動[保留樂生療養院]議題為例（未出版之碩士論文）。國立臺灣師範大學，台北市。
李素蘭。2010。台灣區段徵收制度之政經分析（未出版之碩士論文）。國立臺灣大學，台北市。
周信燉。2004。土地徵收程序中被徵收人之陳述意見機會。*土地問題研究季刊 3*（2）：32-45。
林如森。2004。公共傳播與農民運動--以[一一二三與農共生]為例（未出版之碩士論文）。國立臺灣大學，台北市。
林怡瑩。2004。環境風險，環境運動與媒體：以台灣焚化爐政策爭議的媒體再現為例。國立政治大學，台北市。
林常富。2009。電視新聞框架研究-以電視新聞報導集會遊行事件為例（未出版之碩士論文）。國立政治大學，台北市

胡元輝。2007。*媒體與改造*。台北市：商周。
孫秀蕙。1994。環保團體的公共關係策略之初探。*廣告學研究*（3）：159-185。
徐世榮。2010。違背土地正義的浮濫徵收—土地淪為政府發展經濟的金雞母。*當代*（242）：102-105。
徐世榮。2015。浮濫徵收與土地正義。*臺灣社會福利學刊 12*（2）：1-13。
徐世榮。2016。土地正義：從土地改革到土地徵收，一段被掩蓋，一再上演的歷史。新北市：遠足文化。
徐世榮、廖麗敏。2011。建構民主人權的土地政策。*台灣社會研究季刊* 84：403-429。
徐旭。2014。社區動員如何可能？－以灣寶與大埔反土地徵收抗爭為例的比較研究（未出版之碩士論文）。國立臺灣大學，台北市。
翁秀琪。1994。我國婦女運動的媒介真實和〔社會真實〕。*新聞學研究 48*：193-236。
張讚國、劉娜。2016）。從定調到解釋性界限：占中運動，商業報紙與獨立媒體。*傳播研究與實踐 6*（1）：45-77。
莊豐嘉。2011。台灣公民新聞崛起對公共政策之衝擊--從樂生，大埔到反國光石化事件之比較分析（未出版之碩士論文）。國裡臺灣大學，台北市。
陳文貴。2003。土地徵收處分廢止概念之釐清。*法令月刊 54*（9）：40-45。
陳立夫。2008。析論我國土地徵收法制上之爭議問題。*臺灣土地研究 11*（1）：1-35。
陳佳君。2015。網路時代社會運動組織的傳播策略—以[文林苑]都市更新抵抗運動為例（未出版之碩士論文）。國立臺灣大學，台北市。
陳瑩真。2004。土地徵收補償中的估價問題（未出版之碩士論文）。國立台北大學，新北市。
彭慧蕙。1997。大眾傳播媒介與社會運動—以[反雛妓]社會運動為例（未出版之碩士論文）。中國文化大學，台北市。
黃樹仁。2002。心牢：農地農用意識形態與臺灣城鄉發展。台北市：巨流。
詹順貴。2011。市價徵收，不是土地正義。*司法改革雜誌*（86）：73-73。
廖本全。2014。歧視與暴力下的土地掠奪：台灣環境正義與人權的凝視。*台灣人權學刊 2*（4）：137-150。
廖學能。2014。土地徵收之法律問題探討-以農地為論述中心（未出版之碩士論文）。國立臺灣大學，台北市。
臧國仁、鍾蔚文。2001。災難事件與媒體報導：相關研究簡述。*新聞學研究*（62）：143-151。
劉于禎。1999。非營利組織行動暨訊息研究--以勵馨社會福利事業基金會的〔反雛妓社會運動〕為例（未出版之碩士論文）。國立中正大學，嘉義縣。
劉泰英。1988。現階段臺灣農地問題之探討。*臺灣農地政策研討論*，9-18。
蔡偉銑。2014。新竹科學園區政策過程的重新檢視。*人文及社會科學集刊 26*（3）：427-481。
蔡培慧。2010。真實是一場社會行動反思台灣農村陣線的行動與組織。*台灣社會研究季*

刊（79）：319-339。
蔡培慧。2011。[土地正義的堅持與實踐：大埔事件一年過後]專題引言。*台灣社會研究季刊*（84）：397-401。
蔡培慧。2011。土地正義的理由。*生態臺灣*（33）：10-13。
鄭欽龍。1988。紓緩土地取得的困難。*經濟前瞻*（11）：13-16。
戴秀雄、李立達。2007。論土地徵收條例第 51 條第 2 項未於規定期限內繳還徵收價額者不發還土地之適法性分析。*土地問題研究季刊 6*（2）：16-34。
鍾麗娜、徐世榮。2013。科技數字至上的迷失-市價徵收與土地正義間之恐怖平衡。*土地問題研究季刊 12*（4）：51-64。
簡曉娟。2011。暴動？抗暴？論移工團體與新聞媒體對[高捷泰勞事件]的意義建構與互動分析（未出版之碩士論文）。國立臺灣大學，台北市。

台灣獨立媒體中的基改食品

郭柏傑*、闕河嘉**

摘　要

本研究以台灣獨立新聞媒體所發行之電子報為研究對象，試圖探究獨立新聞媒體如何呈現「基改食品」議題，並提供哪些潛藏的論述框架。基改食品之爭議擴及了人類健康、環境衝擊、經濟發展以及宗教倫理等，其中媒體更是該風險溝通中很重要的一環。近年來台灣食品安全風波不斷，因而極具爭議的「基改食品」也成為大家關注的焦點。大眾媒體往往被視為是風險傳播的關鍵媒介，其風險宣傳的內容及溝通效果影響消費者對於基改食品的認知。

獨立媒體被視為與主流媒體相對抗，其近年來在讀者數量上有明顯的增長。然而，獨立媒體有時也會複製主流的價值框架、靠進優勢的權力體系，或淪為某些異議團體的附庸，失去其自主性。故分析獨立媒體如何對消費者形塑「基改食品」之形象仍有其必要性，藉由台灣獨立新聞媒體對「基改食品」所提供之觀點，揭示其所擁護的立場。本研究以六個獨立新聞媒體作為研究對象。其中包含台灣前五大獨立新聞媒體：「風傳媒」、「關鍵評論網」、「新頭殼」、「民報」、「上下游新聞市集」（創市際雙周刊，2016），以及台灣最大的科學網站社群「泛科學」。

本研究的三個具體研究問題為（1）探究六大獨立新聞媒體在「基改食品」之相關報導中，涵蓋了哪些「行動者」？（2）指出六大獨立新聞媒體之「基改食品」報導凸顯了誰的觀點？不同利害關係者對於「基改食品」所抱持的立場又為何？（3）檢視六大獨立新聞媒體「基改食品」報導之本質偏向「風險」傳播或是「危機」傳播？研究方法以語料庫導向（corpus driven）及文本分析方法為基礎。分析結果發現，獨立新聞媒體中的基改食品報導主要由七個「行動者」所組成，分別為「基改」、「農業」、「食品安全」、「利害關係人」、「政策行動」、「地理區位」以及「爭議討論」。此外，政府部門以及專業人士是報導當中的主要消息引述來源，而消費者與農民雖然在報導中被提及多次，但是兩者的言論幾乎是

* 國立台灣大學生物產業傳播暨發展學系碩士生。

** 國立台灣大學生物產業傳播暨發展學系副教授，通訊作者，Email: hchueh@ntu.edu.tw。

被獨立新聞媒體所忽視。而在基改食品的報導論述方面，獨立新聞媒體則涵蓋了「風險」以及「危機」傳播。

雖然基改食品之媒體再現已有不少相關研究，然而多數研究僅聚焦於主流媒體。隨著國內獨立媒體之閱讀人數的攀升，我們試圖以獨立媒體為主體，提供呈現「基改食品」的另一種聲音。

關鍵字：基因改造、獨立新聞媒體、媒體論述、風險/危機傳播、語料庫分析

Independent News Media Coverage of the Genetically Modified Food in Taiwan

Bo-jie Guo[1*], Ho-chia Chueh[**]

Abstract

This article examines independent news media coverage of genetically modified (GM) food in Taiwan. Several studies have suggested that news media coverage of GMOs has strong impact upon public understanding on GM foods. Independent media is often regarded as free of influence of government nor corporate interests, and thus has minimal bias. We chose Taiwan's 6 main independent news media (the *Storm, The News Lens*, *Newtalk, Taiwan People News, News & Market* and *PanSci*) to examine their coverage on GM food. We aim to 1) identify 'actors' that played in the coverages of GM foods; 2) examine the preference 'resource' these media employ and they ways in which they represent GM food debates; 3) consider the nature of GM food coverage as a form of risk communication or crisis communication. This research is conducted based upon a corpus-driven approach. We find that seven actors are included in the coverage: GM, agriculture, food security, stakeholders, policy, geographical location and controversy issues. Government officials and expert (research scientists and university lectures) are the most frequently quoted sources. The nature of independent media coverage of GM food is complex; there are mixed forms of risk and crisis communication at different periods. This research contribute to media analysis of GM food coverage in the ways that alternative media represent GM food and that may affect the public's understanding.

Keywords: gm foods, alternative media, media discourse, corpus-driven, Taiwan

[*] Master Student, Department of Bio-Industry Communication and Development, National Taiwan University.
[**] Associated Professor, Department of Bio-Industry Communication and Development, National Taiwan University. (Corresponding Author) Email: hchueh@ntu.edu.tw.

一、緒論

近年來台灣在食品安全風波不斷的情況下，「如何選擇安心食材」以及「如何吃得健康」等議題成為社會大眾關注的焦點；因此除了「有機食品」的議題之外，極具科技風險爭議的「基因改造食品」也成為國人所關注的焦點。在政策推動方面，2014 年台灣的新《食安法》當中，加強了基改食品之管理，要求業者應建立基改食品原料供應來源及流向的追溯系統（劉家瑜，2014）。此外，衛福部更於 2015 年公告「基因改造食品標示新制」，強化基改食品標示資訊之揭露，除了將「食品添加物」以及「散裝食品」納入實施範圍之外，也將原本基改原料摻入量大於 5%即視為基改食品的標準降低至 3%，比過往標準更為嚴格。而部分民間團體對於基改食品議題的積極投入，更直接影響了相關法條的制定。主婦聯盟環境保護基金會與綠色陣線協會、台大農藝系種子研究室於 2008 年組成「台灣無基改農區推動聯盟」；其中更於 2014 年發起「校園午餐搞非基」運動。最終，立法院於 2015 年 12 月三讀通過「學校衛生管理法」之修正，明定學校供應膳食者禁止使用基因改造生鮮食材及其初級加工品，基改食品全面退出中小學營養午餐。

全球對於基改食品的爭議至今仍爭辯不休，爭議面向主要擴及了人類健康、環境衝擊、經濟發展以及宗教倫理（ethics）。根據侯新龍（2007）指出，基改食品對「人類健康」之疑慮分別有抗性基因蛋白、過敏反應等問題；然而，部分科學家卻認為基改食品能夠創造更健康的作物（healthier crops）以及改善食品本身的營養價值（Lore & Imungi & Mubuu, 2013）。「環境衝擊」有超及雜草（super weed）的產生、「基因外流」導致生態系統破壞等潛在風險；持反向意見者則認為發展基改食品反而能減少化肥、農藥使用、減少對環境的化學汙染（科技新報，2014）。

反觀「經濟發展」層面，台灣政府組織抱持明顯的正向態度。台灣經濟研究院於 2013 年《基因改造產業發展與趨勢報告》中指出：「面臨未來全球人口爆炸性成長、可耕地減少、氣候變遷加劇等因素，為滿足全體人類對糧食作物之需求，使用快速有效益的作物育種技術，（基改）必將成為未來農業上的主流趨勢」，當中更提及基改技術具有增加產量、降低作物生產成本、增加農業管理彈性等助益。但郭華仁教授（2014）曾公開表示，糧食議題涉及層面錯綜複雜，單一化地將基改食品視為解決辦法，是一種「化約論」的極致（引自郭華仁個人 facebook 社群平台）。最後，在「宗教倫理」當中，由於生物科技得以將不同物種間之基因進行轉殖，有可能會產生植物基因組中含有動物基因（轉自台灣法律網）；也有人認為基因改造等同於人類扮演造物者的角色（playing God），強行違反大自然法則（Morse, 2016）。Dibden 等人（2013）指出，未來農業政策勢必朝向兩種典範發展，分別為農業生物科技（例如：基因改造、動物無性繁殖）以及生態性農業（例如：有機農法、環境友善耕作），然而各國在這兩者間的偏好選擇至今仍無定論。

過往國內研究指出，民眾接收基改食品相關訊息的主要來源分別是：「報紙、雜誌或書籍」（64.5%）以及「電視或廣播」（63.0%）（孫智麗，許嘉伊，劉翠玲。2007），顯示出「媒體」成為建構社會大眾「基改食品」之概念很重要的一環。根據創市際雙周刊第五十九期（2016）針對「新聞資訊網站調查與使用者蓋況」的分析顯示，「獨立新聞媒體」在人數使用量上有「突出」的成長。獨立媒體一般被視為與主流媒體相對抗（管中祥，2011），能夠藉由被忽視觀點之成現，形塑反抗文化或進行反壓迫（郭文良，2010）；然而獨立媒體有時候也會複製主流價值、或淪為特定團體的附庸（管中祥，2014）。因此，台灣獨立新聞媒體如何呈現基改食品之相關議題，對消費者進行基改食品之概念型塑有其深入探究的必要性；並藉由報導中消息來源之引述，檢視台灣獨立新聞媒體對「基改食品」所提供之觀點，以揭示其所擁護的立場。

二、文獻回顧

(一)「基改食品」之新聞框架

基改作物的商業化發展最早始於 1996 年在中國種植能抵抗毒素病的基改菸草，其後在 1990 年代中期美國等地也開始同意並發展基改農業。然而，基因改造技術雖然已發展 20 年之久，各界對基改食品的各項疑慮仍未達成共識。這些「各說各話」的現象以及相互對立的聲稱，影響了報紙媒體對於基改作物的報導，並轉向影響公眾對基改作物的消費選擇（Morse, 2016）。此外，根據國際農業生物技術應用服務組織（ISAAA）在 2014 年所做的統計，從 1996 年至 2014 年間全球種植基改作物面積成長了約 6.5 倍，達到 18 億公頃；這似乎能假定「基因改造食品」已成為媒體不容忽視的議題。

關於基改食品在媒體報導中的框架分析，國內外已有一些相關研究。在台灣研究部分，謝君蔚和徐美苓（2011），分析《聯合報》與《中國時報》自 1994 年 1 月至 2006 年 12 月的基改食品相關新聞。研究發現，歸類到「進步包裹」（科學萬能、經濟掛帥）的新聞占最多，雖然有逐年減少的趨勢；屬於「危害包裹」（禍延子孫、健康疑慮等）與「關切包裹」（妥善管理）的新聞總比例則不相上下，前者的新聞數量變化逐年增加，後者則在醞釀期（1994-1999）上升最多，之後數量略微減少。

在國外研究部分，Nisbet 和 Lewenstein（2002）檢視美國兩大媒體《紐約時報》（New York Times）與《新聞週刊》（Newsweek）自 1970 年至 1999 年與生物科技相關（biotechnology-related）的新聞（基改新聞即被研究者視為科物科技底下的一環）。研究結果指出，「進步」幾乎是報導中的主要論述，並強調生物科技與「科學進步」、「經濟

願景」框架的連結；此外，科學家以及政府官員主導了報導中的發言權。但自 1995 年開始，因複製羊、基因療法致死等案例出現，「倫理」（ethics）框架首次在報導中出現，並呈現比以往負面的爭議論述。此外，Price（2006）分析《時代雜誌》（Time）與《新聞週刊》（Newsweek）於 1994 至 2003 年間封面故事有關「食物」的框架，探究與食物相關的個別議題（例如：肥胖、基改食品）。研究結果顯示，「基改食品」新聞採用的論述框架，有凸顯科技超越自然的優勢趨勢。

由上述文獻得知，在台灣以及美國的報紙媒體當中，對於基改食品的「正向」論述是居多的，但是涉及爭議議題的報導也有逐年攀升之趨勢。然而，這些研究樣本距今已有十年以上之久，國內獨立新聞媒體對於基改食品之新聞議題框架偏好也尚未有相關研究，因此值得我們進一步檢視。

(二)獨立新聞媒體之於主流新聞媒體

獨立新聞媒體又可以稱之為另類新聞媒體、小眾新聞媒體、草根新聞媒體（張傳佳，2013）；獨立新聞媒體被視為與主流新聞媒體相對抗（管中祥，2011），然而這樣的「抵抗」不僅僅表現在內容之上，含包括了整個新聞媒體的運作以及產製方式（成露茜，2009）。值得注意的是，獨立新聞媒體與主流新聞媒體並非是一分為二的，賀照緹（1993）指出，小眾新聞媒體會不斷地在原有的權利立場中，策略性地改變它的抗爭位置，甚至在經過一段時間之後，反倒成為主流。

台灣獨立新聞媒體的發展可以追溯至異議媒體的崛起，並且與政治反對運動相關，將此另類媒體視為對抗國民黨統治的工具（管中祥，2009）。創立於 1997 年的《生命力新聞》可以被稱作是台灣獨立新聞媒體發展的指標，而幾乎同時期的《苦勞網》也被視為先鋒（莊豐嘉，2011）。伴隨著網路發展以及影音串流技術提升，獨立新聞媒體有了更多元的發聲管道；其中以 2007 年成立的《PeoPo》公民新聞平台最具代表性，該平台強調「開放」、「行動」、「分享」，提倡公民自主機制（管中祥，2009）。隨後，國內開始誕生越來多綜合性新聞的獨立新聞媒體平台，例如：《新頭殼》、《關鍵評論網》等。

郭文良（2010）以 Downing（2001）對獨立新聞媒體之解釋為基礎，整理出獨立新聞媒體的四個主要精神，分別為：（1）在主流文化中直接或間接表達反對立場、（2）最積極的主動閱聽人與媒體使用者、（3）反抗文化與抵抗、（4）寄存於社會運動，與意識覺醒相關聯（conscious-raising）。因此，獨立新聞媒體其內部組織除了遠離資本主義式的思考方式之外，在內容呈現上也運用了有別於主流媒體的視角與框架，以非商業模式提供平等的媒介進用、捍衛弱勢（或少數）團體的主體性（張傳佳，2013）。「理想中」的獨立新聞媒體呈如上述所定義，被視為是對主流媒體或主流意識形態的抗議或補充

（管中祥、張時健，2004）；然而相反的，獨立新聞媒體也可能靠近優勢的權力體系、或未必站在弱勢者的位置，即使在社會中發出異見，也不具獨立新聞媒體應有的主體性（管中祥，2014）。

伴隨國內民眾閱讀獨立新聞媒體的人數明顯增長，國內獨立新聞媒體以何種報導方式呈現「基改食品」之議題？又如何處理「基改食品」不同面向之爭議？其背後的潛藏意識形態有待深入揭示。

(三)研究問題

綜觀文獻當中對於基改食品的爭議討論以及國內獨立新聞媒體閱讀率的攀升，本研究試圖透過台灣具代表性的六大獨立新聞媒體，檢視它們如何呈現「基改食品」議題，並提供哪些潛藏的論述框架。以下發展出三個具體研究問題：

1. 探究六大獨立新聞媒體（風傳媒、關建評論網、新頭殼、民報、上下游新聞市集、泛科學）在「基改食品」之議題報導中，涵蓋了哪些「行動者」？
2. 指出六大獨立新聞媒體之「基改食品」報導凸顯了誰的觀點？不同利害關係者對於「基改食品」所抱持的立場又為何？
3. 檢視六大獨立新聞媒體「基改食品」報導之本質偏向「風險」傳播或「危機」傳播？

三、研究設計

根據創市際雙周刊第五十九期的調查結果顯示，台灣前五大訪客數最多的獨立新聞媒體分別為「風傳媒」、「關鍵評論網」、「新頭殼」、「民報」以及「上下游新聞市集」；此外，「泛科學」為台灣最大的科學網站及社群，且經營模式符合獨立媒體之內涵，故本研究乃以此六大獨立新聞媒體作為研究對象。

由於獨立新聞媒體成立的時間差異甚大、並皆不超過十年，因此樣本蒐集區段以各大獨立新聞媒體的成立時間為始，直至 2016 年 9 月 31 日為止。在樣本來源方面，「關鍵評論網」、「新頭殼」、「民報」以及「上下游新聞市集」以慧科大中華新聞資料庫進行樣本蒐集，而「風傳媒」及「泛科學」則以各自網頁中的歷史新聞作為資料庫來源。搜尋關鍵字為「基改」、「基因改造」；扣除文章重複連結以及不符合研究旨趣者（例如：基改蚊、基改胚胎等），分別搜集到「風傳媒」41 篇、「關鍵評論網」74 篇、「新頭殼」51 篇、「民報」59 篇、「上下游新聞市集」149 篇、「泛科學」32 篇，共 406 篇

研究設計以語料庫導向（corpus driven）作為分析依據，並輔以文本分析方法進行研究問題探究。語料庫分析方法能夠降低研究者潛在意識的偏見，並以「資料」作為研究的出發點，讓整體的語言模式以及趨勢被徹底顯現。除此之外，透過量化統計的數據，也能夠揭發主流論述以及各種看世界的方式（林意璇，2015）。而藉由語料庫導向的分析模式，能夠擺脫「描述性分析」窘境，深入挖掘語料庫文本結構與模式、輸出文本當中隱含的意義（郭文平，2014）。本研究所使用的語料庫分析元素分別有：詞頻（frequency）、文檔詞頻（document frequency）、搭配詞（collocation）、關鍵詞脈絡檢索（KWIC）、詞叢（cluster）。而論述（discourse）被賦予具備各人偏好系統的特性（Peter, 2014），文本分析注重文本間的結構關連性與互動意涵並強調意義的詮釋，並通常只針對一種社會製成品進行分析，例如：新聞報導、文學作品、電影等（遊美惠，2000）；因此輔以文本分析，進一步凸顯語料庫分析後的結果差異，將其予以解讀。

本研究採用目前功能最為完善的中文語料庫分析軟體—「庫博中文語料庫分析工具」（Corpro）進行研究分析。該軟體除了內建台灣「中央研究院中文斷詞系統」之外，也設有「自建詞典」功能，研究者能依照自己的研究問題進行重新斷詞。「同類詞編輯」功能則能夠集結相同概念之詞彙，進行概念與概念（或整體）之間的比較。輔以「停用詞」功能，移出過多不必要的資訊，降低研究者資料判讀之干擾。

四、研究發現

(一) 基改食品報導中的「行動者」

基改食品報導中涵蓋了許多錯綜複雜的元素，例如：政策、食安議題、農業發展等「非人類」面向、以及利害關係人等「人類」面向。本小節試圖打破一分為二的框架分類方式，選擇以行動者網絡（ANT）之概念，將獨立媒體中的高頻詞彙視為是共同建構基改食品報導的「行動者」，彼此之間相互交織、影響。進一步，我們將不同類型之行動者予以分類，深入了解行動者涉及了基改食品中的那些面向。

由於六大獨立媒體（「風傳媒」、「關建評論網」、「新頭殼」、「民報」、「上下游新聞市集」、「泛科學」）個別樣本數懸殊差異很大（最多為《上下游新聞市集》149 則，最少為《泛科學》32 則），若將所有獨立媒體集結為一語料庫進行詞頻分析，會產生樣本數較少的媒體來源其高頻詞彙被忽略之情形。因此本研究分別建置六大獨立媒體之語料庫，將個別語料庫中前 100 個高頻詞彙進行交叉比對（一個詞彙同時出現於兩個語料庫以上即納入表格進行整理），統整出 131 個獨立媒體在基改食品報導中偏好使用詞彙（表一）。

表一 六大獨立新聞媒體高頻詞彙統整（共 131 個）

N	Word	N	Word	N	Word	N	Word
1	基改	**34**	全球	**67**	玉米	**100**	是否
2	食品	**35**	社會	**68**	許多	**101**	無法
3	消費者	**36**	基因	**69**	食藥署	**102**	人體
4	產品	**37**	市場	**70**	孩子	**103**	學生
5	台灣	**38**	種	**71**	國際	**104**	候選人
6	問題	**39**	如果	**72**	拜耳	**105**	能夠
7	美國	**40**	需要	**73**	科學	**106**	鮭魚
8	作物	**41**	瘦肉精	**74**	農	**107**	學校
9	基因改造	**42**	因此	**75**	希望	**108**	午餐
10	使用	**43**	相關	**76**	土地	**109**	校園
11	政府	**44**	農產品	**77**	廠商	**110**	食材
12	食物	**45**	基因改造食品	**78**	成為	**111**	基改食品
13	可能	**46**	推動	**79**	進口	**112**	市長
14	認為	**47**	產業	**80**	增加	**113**	指出
15	安全	**48**	大豆	**81**	不會	**114**	市府
16	生產	**49**	孟山都	**82**	結果	**115**	教育
17	健康	**50**	食安	**83**	對於	**116**	進行
18	影響	**51**	不同	**84**	都是	**117**	檢驗
19	風險	**52**	技術	**85**	必須	**118**	蔬菜
20	農藥	**53**	業者	**86**	科技	**119**	未來
21	有機	**54**	公司	**87**	只是	**120**	民眾
22	可以	**55**	標示	**88**	國家	**121**	主婦聯盟
23	沒有	**56**	表示	**89**	目前	**122**	添加物
24	種子	**57**	經濟	**90**	造成	**123**	要求
25	農民	**58**	不是	**91**	在地	**124**	食用
26	因為	**59**	方式	**92**	甚至	**125**	現在
27	農業	**60**	管理	**93**	非基改	**126**	轉基因
28	環境	**61**	吃	**94**	提供	**127**	發展
29	黃豆	**62**	開始	**95**	政策	**128**	嘉磷塞
30	自己	**63**	人類	**96**	糧食	**129**	我們
31	研究	**64**	營養午餐	**97**	科學家	**130**	印度

32	原料	**65**	透過	**98**	生物	**131**	重要
33	種植	**66**	其實	**99**	實驗		

根據這前 131 個高頻詞彙，我們進行「行動者」字面意義的語義分類（semantic categories），以利探究獨立媒體在報導基改食品相關新聞時，可能涉及的對象、事件或議題。「語義分類」主要根據（一）詞彙特性、或（二）字彙背後可能連結的新聞主題做為歸類方式（郭文平，2014）。由於「基改」（基改、基因改造、基因改造食品、非基改、基改食品、轉基因）為主要行動者，意即行動者們所共同關心的議題，因此不納入表中。此外，「爭議討論」主要依據文獻回顧中基改食品之爭論面向（debates）進行詞彙選取，歸納結果如表二所示。

表二　基改食品報導中「行動者」類目表

行動者	**子類目**	**使用詞彙**
農業		作物、農藥、有機、種子、農業、種植、種、黃豆、農產品、大豆、玉米、農、土地、在地、糧食、農藥、蔬菜、嘉磷塞
食品安全	食材	食品、食物、原料、食材
	飲食	吃、食用
	食安議題	瘦肉精、食安、營養午餐、鮭魚、午餐、添加物
政策行動		推動、標示、管理、政策、檢驗
利害關係人	政府部門	政府、食藥署、市長、市府、候選人、學校
	專業人士	科學（家）
	NGO 組織	主婦聯盟
	消費者	消費者、孩子、學生、民眾
	業者	業者、孟山都、公司、拜耳、廠商
	農民	農民
地理區位		台灣、美國、全球、國際、印度
爭議討論		健康、風險、環境、社會、產業、經濟、人類、人體

藉由表二可以得知，獨立媒體在基改食品報導中，主要由七個行動者所組成；分別為「基改」、「農業」、「食品安全」、「利害關係人」、「政策行動」、「地理區位」以及「爭議討論」。為了凸顯不同行動者在基改食品報導中所涉及的面向，我們先將行動者進行同類詞編輯，並運用搭配詞分析（collocation）予以探究（表三），以下說明之。

表三　行動者之顯著搭配詞彙（T 值排序前 30 個具體運用詞彙）

行動者	**搭配詞彙**
農業	非基改、殘留、生產、農民、進口、基改、**使用**、作物、棉花、認證、**供應**、契作、本土、基因改造、自給率、台灣、除草劑、吉園圃、**推廣**、農場、栽培、噴灑、友

	善、蔬果
食品安全	校園、安全、非基改、標示、基因改造食品、包裝、在地、散裝、學校、基改、學童、管理、搞非基、供應、安心、衛生、承諾書、遵行、加工、採用、問題、來源、推動、管理法
政策行動	食品、基改食品、原料、基因改造、基因改造食品、自主、基改、強制、散裝、非基改、校園、規定、業者、源頭、食安、遵行、法規、落實、教育、規範、產品
利害關係人	要求、午餐、基金會、媒體、**環境保護**、衛生法、輔導、保障、把關、提供、**營養午餐**、自主、權益、呼籲、種子、**收購**
地理區位	基改、進口、孟山都、農業、國家、FDA、鮭魚、產業、糧食、豬肉、棉花、標準、市場、核准、黃豆、中國、貿易、藥物管裡局、農業部、基改食品、加拿大
爭議討論	影響、造成、危害、發展、安全、生態、飲食、友善、食物、有害、正義、文化、教育、大眾、永續、作物、致癌、基改食品、科技、消費者

「農業」為獨立媒體討論基改議題不可或缺的一部分。我們選用「使用」、「供應」、「推廣」三個動詞進行詞叢分析（cluster），以釐清獨立媒體對於基改食品之偏好為何。檢視「使用」、「供應」、「推廣」其後相連接的受詞可以發現，三者詞彙所連接的最高詞頻之受詞分別為：「非基改」（67 次）、「有機」（12 次）、「非基改」（10 次），因此在基改食品報導中，「非基改」以及「有機」反而有較多的論述。進一步探究被視為與農業生物科技相抗衡的「有機」，發現「有機」一詞顯著搭配的詞彙，除了蔬菜、認證、標章之外，「非基改」與它也具有強烈的搭配關係（T：4.110）。針對「有機」及「非基改」兩個詞彙進行關鍵詞檢索（KWIC），得到「有機」及「非基改」在媒體報導當中多數呈現正面形象，成為取代基改食材的選擇，例如：「鼓勵學校多採購品質良好安全食材，包括部分**有機**農產品及**非基改**黃豆食用，提昇營養食安（民報，2015.09.02）」、「建議素食者，除了在選購黃豆及其製品時，多選擇**有機**、**非基改**的產品，也要記得均衡攝取各種豆類（上下游，2016.06.23）」。

基改食品也被獨立媒體視為是食安問題的一環。縱使「基改食品」的益處與危機尚無明確定論，但是在獨立媒體報導中多半將基改食品視為是「問題食材」的潛在對象，例如：「大多數民眾對於進口的 230 萬噸黃豆中九成以上為**基改**豆的事實，毫無所悉；長期大量食用基改黃豆的族群也並不了解，可能將自身與家人暴露在未知的**食安**風險中（上下游，2014.04.17）」。此外，從「校園」、「學校」、「搞非基」、「承諾書」等詞彙得知，由民間所發起的「校園午餐搞非基」運動受到獨立媒體熱切關注，該活動即聚焦於基改食品對人類健康影響的討論，例如：「**食安**議題仍是候選人及選民關注的焦點。台灣無**基改**推動聯盟日前發起『校園午餐搞非基』運動（風傳媒，2014.11.19）」。「校園午餐搞非基運動」同樣影響到基改食品的相關「政策行動」，並持續至「學校衛生法」修正通過。我們將「搞非基」以及「學校衛生法」進行文檔詞頻分析（document frequency），

並藉由原始文檔檢視以「學校衛生法」之修訂時間為基準（2015 年 12 月 14 日），檢視兩者之間的報導數量差異。分析結果得到，「搞非基」相關報導數量共有 41 篇、而「學校衛生法」相關報導數量僅有 19 篇。因此，獨立媒體在基改食品的報導中，並非僅僅擔任政府政策「傳聲筒」的角色，對於民間發起的相關運動，也具高度重視。

基改食品不同面向的議題所擴及到的利害關係人也有所不同。我們選取了由搭配詞結果所呈現出來的主要涉及議題（營養午餐、環境保護、收購），進行議題與利害關係人的搭配檢視。在「營養午餐」議題中，「政府部門」、「消費者」為媒體報導中的主要提及的利害關係人；「環境保護」議題則有「消費者」、「專家人士」以及「農民」；「收購」議題則有「業者」與「農民」。然而，光是從字面上並無法進一步得知利害關係人彼此之間在獨立媒體中所佔居的要角，這有待於至第二個研究問題做持續的深入分析。

台灣目前尚未開放基改作物之栽培，大宗穀物同時長期仰賴美國供應（每年約進口 500 萬噸玉米及 250 萬噸黃豆，位居全球第三，僅次於歐盟、日本），以美國當地栽種面積推估，台灣黃豆市場有七成以上屬於基改黃豆（潘若琳、顏良恭、吳德美，2009），因此基改原料、食品的「進口」議題也是獨立媒體討論的一環。其中，美國的跨國農業生物技術公司「孟山都」的一舉一動更是受到矚目，推估其主要原因為孟山都壟斷了全球超過九成以上的基因改造種子。而美國食品藥物管理局（FDA）核准「基改鮭魚」上市，當中「不必特別標示」之規定也引發各界擔憂，甚至將它稱之為「科學怪魚」，例如：「就連美國當地消費者團體也抗議頻頻，對於基改鮭魚更冠上『科學怪魚』的封號（民報，2015.11.21）」；縱使台灣目前暫不開放基改鮭魚輸入，但其在獨立媒體中也成為重要話題。

在爭議討論當中，獨立媒體對於基改食品的爭議主要可以劃分為「健康」（飲食、致癌）以及「環境」（生態、永續）兩大面向。藉由「健康」的搭配詞分析可以得到，其與「環境」也經常伴隨出現（t-score：5.533），因此基改食品報導中的「健康」及「環境」議題，之於獨立媒體來說是同等被受關切的。針對「健康」及「環境」之搭配詞進行觀察，「健康」顯著搭配的詞彙有「風險」、「人體」、「安全」、「危害」；而「環境」有「破壞」、「危害」、「風險」、「汙染」。故基改食品與兩者爭議面向之間的關係，在獨立媒體當中都是傾向於負面論述的，以「健康」面向來說，例如：「除草劑「嘉磷塞」（Glyphosate，台灣產品名『年年春』）的致癌風險，從『可能致癌（Group 2B）』提高到『很可能致癌（Group 2A）。』（上下游，2015.12.11）」；「環境」面向例如：「基改作物可能有造成環境大災難的風險。此類作物，違反上帝造物法則，破壞了植物的生命科學（民報，2014.06.02）」。

從上述基改食品報導中的「行動者」分析得知，每個行動者並非各自一分為二，而是彼此之間相互牽連、影響（例如：「食品安全」涉及了「政策行動」等）。然而，從字面上的行動者語意分類並無法深入探究獨立媒體對於基改食品之立場，接下來本研究將鎖定獨立媒體在基改食品報導中的「利害關係人」，進一步挖掘獨立媒體如何建構基改食品報導。

(二)基改食品報導凸顯了「誰」的觀點？

為了凸顯「誰」在獨立媒體中主導發言權，我們藉由表二從高頻詞彙所歸納出來的「利害關係人」作為類目建構依據，分別為：政府部門、專業人士、NGO 組織、消費者、業者、農民。然而，若僅將利害關係人鎖定在高頻詞彙所列出的對象上，可能會造成部分利害關係人被遺漏的問題。故我們同樣從個別獨立媒體的語料庫中，以詞頻分析列表蒐集出現在報導中的利害關係人，整理結果如表四所列。

表四　基改食品報導中被提及的利害關係人

類目	涵蓋詞彙
政府部門	政府、食藥署、市長、市府、候選人、學校、農委會、縣府、衛福部、立委、教育局、衛生局、教育部、立法院、農業局、農業部、農會、經濟部
專業人士	科學（家）、教授、專家、學者、營養師、郭華仁、研究所、研究、博士、學術、科學界、醫界、林杰樑、毒物科、中研院
NGO 組織	主婦聯盟、基金會、綠色和平、里仁、慈心、黃嘉琳、賴曉芬
消費者	消費者、孩子、學生、民眾、學童、人民、市民、婦女、年輕人
業者	業者、孟山都、公司、拜耳、廠商、企業、高志明、義美、供應商、食品業
農民	農民、農友、農夫、農家、農人

運用同類詞編輯功能，可以得到在六大獨立媒體中「政府部門」被提及的次數最高（2820 次），依序為「業者」（1954 次）、「消費者」（1638 次）、「專業人士」（1532 次）、「農民」（838 次）以及「NGO 組織」（357 次）。「政府部門」以政府、學校出現的次數最為突出，市長、食藥署、候選人則不相上下，故能推測基改議題也成為近年選舉話題的一部分。「業者」又以孟山都最受媒體廣泛討論，顯示出該跨國企業對於全球基改農業的影響力。「消費者」以民眾、孩子、學生、學童為主要論及對象，基改食品對孩童的健康影響為報導中經常討論的面向。「專業人士」當中，科學以及學術領域比起醫學領域更受獨立媒體青睞。最後，主婦聯盟出現的次數遠遠超越其他組織團體，成為獨立媒體中「NGO 組織」的代表。而通常「農民」的指稱比起其他類目其對象較為單一、明確，故無需進一步細分探究。

從詞頻分析來看，僅能得知利害關係者在基改食品報導中的出現情形，並不能肯定獨立媒體是否引述其發言。因此，我們藉由表一中的高頻詞彙，蒐集「引述動詞」詞彙，

並將其與「利害關係人」進行詞叢分析（以「引述動詞」為中心，羅列出「前一詞彙」含蓋哪些利害關係人），輔以關鍵詞脈絡索引，窺探基改食品報導中有哪些主要的引述來源。「引述動詞」包含了：「表示」、「認為」、「希望」、「指出」，分析結果如表五。

表五 引述動詞與利害關係人之詞叢分析

引述動詞	利害關係人
表示	政府部門（41）、專業人士（15）、業者（13）、NGO 組織（6）
認為	專業人士（22）、政府部門（8）、NGO 組織（5）、業者（4）、消費者（4）、農民（1）
希望	業者（5）、政府部門（4）、專業人士（3）、消費者（1）、農民（1）
指出	專業人士（32）、政府部門（12）、NGO 組織（3）、業者（3）

（單位：次數）

從表五可以發現，「政府部門」以及「專業人士」是獨立媒體在基改食品報導中主要的引述來源，而「消費者」與「農夫」雖然在報導中被提及多次（分別為 1638 次、838 次），但是兩者對於基改食品的「言論」在獨立媒體當中幾乎被消音的。透過關鍵詞脈絡索引分析，「政府部門」引述部分主要可以分為兩個面向，其一為基改食品相關政策之頒布內容，以及政府針對市面上基改食品抽驗之調查結果，其二則為農業相關主管單位補助、輔導地方農民轉作非基改作物之宣傳報導，例如：

> **食藥署**長葉明功今（9）日表示，若立院通過食管法新增規範，強制業者將產品 原料、成品等送驗，衛福部最快在明年 6 月可公告相關細項（新頭殼，2013.12.09）

> 食品藥物管理署（簡稱**食藥署**）於 11 日公告，年底前提供豆漿、豆腐、豆花、 豆干、豆皮、大豆素肉等食品，只要使用到基因改造原料，都需要在店內掛牌標示（上下游，2015.08.16）

> **食藥署**表示，歷年抽驗常見基因改造食品標示結果，不合格率已由最高的 16% 降至每年平均 7%，顯示業者已願意並積極配合基因改造食品原料的標示要求（民報，2015.05.19）

> **農業局**表示，「非基因改造大豆契作計畫」以契約收購方式，輔導海線地區農民於二期作轉作非基因改造大豆（民報，2016.07.18）

在「專業人士」的引述上，可以劃分為學術領域以及科學領域兩大引述來源。在學術領域中，以台大農藝系「郭華仁教授」的發言最常被媒體採用；此外，學術領域的發言絕大多數持「反基改」立場。除了強調基改食品對於人體健康的潛在風險、並批判基改科技能化解糧食問題之理想，同時也針對基改食品之相關政策提出質疑。反觀科學領域，獨立新聞媒體多數引用國外相關研究報告，而對於基改食品的益處或危機也存在正、

反兩種論述之引用。持正向論述的言論表示，基改技術能解決糧食危機、以及改善孩童營養不良的問題，並認為基改食品並無證據對人體有害。持反面論述則強調基改食品在「動物實驗」中的負面影響。然而，科學領域對於基改食品的反面論述並不像學術領域來的肯定、絕對，而是偏好使用「尚無明確定論」、「風險」等較為保守的字眼。

學術領域之引述例如：

台大農藝系**教授郭華仁**表示，素食人口、過敏患者及學童，是受到基改食品影響最大的 3 大族群，因此基改食品「絕對不適合發育中的孩子」（風傳媒，2014.11.19）

如**郭華仁**老師所言，「基改科技是化約論的極致，妄想將複雜的糧食議題 單一化成只要透過基改科技就能解決。」（上下游，2015.09.03）

台大農藝系**教授郭華仁**批評，政府提出政策前，應明確指出發展基改的市場、投入資金、環境風險等，台灣根本沒有基改競爭力（上下游，2015.02.10）

科學領域之引述例如：

許多**科學家**和支持者確信黃金稻米可解決開發中國家，特別是撒哈拉沙漠以南和東南亞地區因缺乏維他命 A 導致的死亡及兒童失明問題（風傳媒，2016.07.03）

馬丁（Cathie Martin）是歐洲植物**研究**機構約翰・英尼斯中心(John Innes Centre)的高級**研究**員，長年致力基因改造作物研究。她表示，過去 25 年裡超過 500 個獨立**研究**機構對基因改造農產品的安全性進行了大量研究，至今沒有任何**科學**證據顯示它們在環境影響和健康安全方面比傳統作物和食品風險更高（風傳媒，2016.05.28）

世界上已有多篇**研究**指出 GMO 食品具有危害人體健康的風險，如 Carman et al.（2013）在長期以基改作物餵食豬的實驗中得到會引起子宮腫大及嚴重胃炎 的結論、Kessler（2013）根據 Sralini 科學家研究，指出被餵食基改玉米的大鼠有產生惡性腫瘤的現象（民報，2015.05.07）

縱使「業者」並非是獨立媒體在基改食品報導中的主要引述對象，但值得一提的是，義美食品公司總經理高志明的發言幾乎占據「業者」引述來源的一半；其餘的業者引述部分，則多為學校團膳供應商針對「基改食品退出校園」之規定，提出相關成本需調漲的問題。我們將「義美」以及「高志明」編輯成一組同類詞，並檢視其在「表示」、「認為」、「希望」、「指出」四個引述動詞中，在「業者」類目裡頭分別占了多少比例。分析結果為，「表示」38.46%、「認為」50%、「希望」20%、「指出」100%；整體來看，「義美」以及「高志明」，總共占了「業者」當中的 44%（其餘無特定指稱之業者占了 56%）。

高志明的發言呈現強烈的「反基改」態度，認為推行基改作物即為「禍害農地」，同時表示對美國「基改鮭魚免標示」的憂心，例如：

> 行政院科技會報傳出，政府將擬定《基因改造科技管理條例》並聚焦「基因作物種植、動物養殖」的貿易輸出，鼓勵國內基改動、植物研究與生產，搶攻國際千億元市場。食品大廠**義美**總經理**高志明**為此炮打政府欠缺「永續經營台灣」的錯誤思維。（風傳媒，2015.02.09）

> **高志明**也舉自己親身經驗表示，過去有不少人來請他支持基改產品，自己也知道這些人背後的壓力，但只要問他如果是你自己，你會買基改或者非基改食品給自己妻子、小孩吃？答案卻都是 100%是非基改（民報，2016.06.06）

> **義美**食品總經理**高志明** 21 日在臉書提出 5 點疑問呼籲總統、立委參選人，及早因應，否則市面上很快就可以買到「免標示的美國基改鮭魚」了（新頭殼，2015.11.21）

最後，在「NGO 組織」的消息引述來源中，以主婦聯盟環境保護基金會的秘書長賴曉芬掌握了話語權；極少部分則為國際綠色和平組織的發言。媒體引述的 NGO 組織言論，對於基改食品皆抱持「反面」立場，主婦聯盟的言論聚焦在基改食品與食安問題的連結，同時也扮演監督政府施政的角色。值得注意的是，「上下游新聞市集」與其他獨立新聞媒體相較，在 NGO 組織中明顯偏好使用主婦聯盟的言論，例如：

> **主婦聯盟**秘書長**賴曉芬**表示，近幾年大家非常重視食安，台灣地狹人稠，很難完全隔離基改污染，「而且我也不相信政府能做好管制」，一旦基改入侵台灣農地，對於環境、生態、農業都會有很大的危害（上下游，2015.02.10）

> **主婦聯盟**環境保護基金會秘書長**賴曉芬**批評，婆婆媽媽就是去市場買散裝的東西，食藥署之前也說過，散裝標示基改很難管，「但現在都說源頭要把關，追蹤追溯也都建立了，為什麼最後要放棄？」（上下游，2014.11.04）

經由上述對「利害關係人」之分析，我們可以歸納出在基改食品的報導中，「政府部門」以及「專業人士」之發言為獨立媒體主要凸顯的對象；藉由「利害關係人」的關鍵詞脈絡索引，大致窺探出不同利害關係人對於基改食品所抱持的立場。進一步，為了探究獨立媒體對於基改食品之報導語意為何，我們使用了「國立台灣大學情緒詞典」作為「正向」及「負向」詞彙語意之依據，並以「基改」概念（包含了：基改、基因改造、基因改造食品、非基改、基改食品、轉基因）進行搭配詞分析。分析結果顯示，「基改」與「正向」（T：23.194）、「負向」（24.791）語意詞彙皆有高度的顯著關係，而「負向」語意詞彙則略為突出。因此我們可以暫時認為，獨立媒體報導基改食品時，會盡量平衡

正、負語意的詞彙使用，以確保客觀立場。

(三)風險／危機論述傳播

進一步分析獨立媒體屬於何種論述傳播形態（風險/危機），我們先將獨立媒體之基改食品報導數量進行分年圖繪製，從中觀查報導數量的變化（圖二）。由於獨立媒體成立時間皆不超過十年（最早為「新頭殼」2009 年 9 月），因此我們以月份作為區分單位，並以最晚成立的「民報」作為計算起始基準（2014 年 9 月），至 2016 年 9 月 31 日為止。

圖二 六大獨立媒體之基改食品報導數量分年圖（灰底為報導數大幅攀升區段）

從圖二的趨勢圖來看，基改食品之報導數量呈現起伏不定的現象，並無穩定攀升或下降之趨勢。但是仍然可以找出五個較為明顯的報導數量高峰，分別為 2015 年 2 月、2015 年 5 月、2015 年 12 月、2016 年 5 月以及 2016 年 8 月。我們即各別針對這五個高峰之「攀升區段」，返回文本探究其報導內容，以判定獨立媒體在基改食品報導中的論述傳播形態。這五個高峰攀升區段分別為 2014 年 6 月至 2015 年 2 月（第一區段共 122 篇）、2015 年 3 月至 2015 年 5 月（第二區段共 53 篇）、2015 年 10 月至 2015 年 12 月（第三區段共 45 篇）、2016 年 3 月 2016 年 5 月（第四區段共 30 篇）、2016 年 6 月至 2016 年 8 月（第五區段共 29 篇）。接著，以「基改」概念詞彙進行搭配詞輔以關鍵詞脈絡索引分析，試圖了解各攀升區段的基改食品報導之報導面向為何；表六即為各區段的搭配詞分析結果。

表六 「基改」概念詞彙之搭配詞分析（T 值排序前 30 個具體運用詞彙）

區段	搭配詞彙
第一區段	作物、標示、校園、食品、午餐、原料、台灣、食材、使用、玉米、種植、黃豆、推動、安全、風險、食物、採用、產品、成份、**承諾書**
第二區段	作物、標示、食材、午餐、使用、木瓜、食品、校園、黃豆、原料、玉米、全面、台灣、產品、推動、種子、政策、種植、要求、**業者**、美國、進口、風險

第三區段	**鮭魚**、**美國**、台灣、表示、玉米、食材、**校園**、表示、使用、作物、食物、食用、公司、**初級**、安全、核准、技術、禁止、黃豆、上市、人類、**加工品**、成份
第四區段	作物、玉米、種植、標示、食品、農產品、台灣、技術、**歐洲**、黃豆、**研究**、禁止、進口、大豆、食物、原料、安全、種植、環境、傳統、**報告**、**歐盟**、產品、**美國**、風險
第五區段	作物、大豆、技術、安全、生物、俄羅斯、種植、反對、生產、玉米、表示、**契作**、管理、食品、中國、產品、推動、種子、認為、農業、國家、基因、標示、禁止、**綠色和平**、農民

透過關鍵索引得知，第一區段的報導多數聚焦於「校園午餐搞非基」的活動。而該活動也搭上了台灣 2014 年 11 月的九合一選舉（地方公職人員選舉）話題，要求候選人以簽署「基改作物退出校園午餐承諾書」的方式換取支持選票：「台灣無基改推動聯盟今早在立法院高喊『你推校園搞非基，家長選票投給 你』，要求候選人簽署**承諾書**，讓基因改造黃豆退出校園營養午餐（上下游，2014.10.02）」。因此在此階段的報導，除了藉由食安問題將基改食品的安全性予以延伸討論之外（尤其是食用級與飼料級黃豆之爭議），地方公職人員候選人簽署承諾書與否也成為媒體追逐的焦點，例如：

> 無黨籍市長候選人柯文哲至南港舊莊里進行座談，志工們聞訊相招前往，一位家長在台下拿著麥克風向柯文哲和在座民眾發言闡述要求安心的無基改校園午餐，抱著幼兒的年輕媽媽則走上台把自行印出的**承諾書**遞上給柯 P 等候選人，成功拿到首善之區重量級候選人簽字承諾（上下游，2014.11.27）

第二區段的報導依然以「校園午餐搞非基」之活動為主軸，並持續追蹤九合一選舉過後，已簽署承諾書的當選者是否遵循承諾，正視基改食品議題（「木瓜」一詞雖然與「基改」具有顯著的搭配關係，但從文檔詞頻得知「木瓜」在該區段僅出現三篇報導，故不列入討論）。同時「業者」的言論在此區段有明顯的增加趨勢，其主要是針對「校園營養午餐若全面禁止使用基改原料，將導致成本調漲」之問題提出討論及回應，例如：

> 目前有越來越多的人，了解到基改食品的潛藏風險，因此，聽到台北市即將推動「**非基改營養午餐**」而感到振奮。但，也有不少人私下熱議，像雲林這樣的 「窮縣」，有「資格」推動「非基改營養午餐」嗎？（關鍵評論，2015.04.30）

> 台北市餐盒食品**商業**同業公會理事長陳明信表示，非基改食材價格大約高出 30～35%，主要是豆製品，目前每校每週大約有四道豆製品，若要全面使用非基改，每校需調漲 5 元（上下游，2015.03.17）

第三區段主要有兩個報導議題，其一為立法院三讀通過《學校衛生法》部分條文之修正，明定學校供應膳食者禁止使用含基改生鮮食材以及初級加工品。其二為美國聯邦

食品藥物管理局（FDA）核准基因改造鮭魚上市供人類食用，以及各界對於「販售時不必特別標示」之規定所提出的疑慮，例如：

> 立法院今（14）日上午三讀通過**《學校衛生法》**部分條文修正案，明訂全國學校供應營養午餐、大學餐廳，將全面禁用基因改造食材及初級加工品，也就是說，包括基改的黃豆、玉米等食品都將在校園禁用（民報，2015.12.14）

> **美國**食品藥物管理局（FDA）日前宣布被批評者稱為「科學怪魚」的**基改鮭魚**為「安全可食」，並准予上市，且販售時還可以不須加註基改標籤。對此不僅美 當地消費團體抗議，義美食品總經理高志明今（21）日也臉書提出 5 點疑問呼籲政府對基改鮭魚應即早因應（民報，2015.11.21）

第四區段的報導較為多元，除了有基改食品殘留農藥（嘉磷塞、萊克多巴胺）的風險報導之外，也聚焦於台灣加入跨太平洋夥伴協定（簡稱 TPP）對基改食品進口管制之爭議。除此之外，也關注於國外關於基改食品之風險研究的相關報告（美國國家學院、英國皇家學會），以及該地區利害關係者對於此報告提出的回應，例如：

> 台灣校園午餐禁用基改食品的規定剛上路，立刻遭美國貿易代表署將此一禁令列為「技術性貿易障礙」。對於對 **TPP** 不存一絲懷疑的善良台灣人而言，這該是個嚴重的警訊（關鍵評論，2016.04.14）

> 地位崇高的**美國國家科學、工程與醫學學院**（National Academies of Science，Engineering and Medicine）17 日發表一項重要的研究報告，結論認為：基改作物與傳統育種（conventional breeding）作物相比較，兩者對人體健康、環境的影響並無差異（風傳媒，2016.05.18）

> **英國自然科學研究學會**「皇家學會」（Royal Society）會長、2009 年諾貝爾化學獎得主拉瑪克里希南（Venki Ramakrishnan）呼籲歐洲各國重新評估對基因改造 作物的禁令。他認為，基因改造技術本身並不危險，歐洲國家不應對「整個科技」加諸禁令（風傳媒，2016.05.24）

第五區段的報導依然聚焦於不同的基改食品議題（「俄羅斯」與「中國」皆只有一篇的報導；前者報導俄羅斯境內全面禁止基改食品之規定，後者報導中國全力推動基改技術之計劃）。國內部分關注於農業局提出的「非基因改造大豆契作計畫」之實施；國外部分則以諾貝爾得主呼籲「綠色和平組織」停止抵制基改食品之言論，延伸討論基改食品的潛在風險，並出現「基改迷思」的相關報導，例如：

活化農地、改善海線二期再生稻品質參差不齊等問題，台中市政府農業局去年輔導143位農民，將100公頃二期作再生稻轉作非基因改造大豆，今年將續推此**契作**計畫，目標擴大到200公頃（風傳媒，2016.06.21）

110位諾貝爾獎得者聯名發出一封公開信，敦促國際環保組織**綠色和平**（Greenpeace）停止反對基因改造農作物。對此，**綠色和平**則回應，這項聲明是「公關手段」，目的是為了影響國會對於強制標示基改食品的立法。（新頭殼，2016.07.01）

無論支持與反對，我們都必須了解：「基改」是跨領域的議題，包含生命科學、農業科學、與社會科學。身為消費者除了認識基改技術之外，了解基改技術對人類的影響也是很重要的。以下列出常見基改農業**迷思**…（泛科學，2016.08.04）

從上述五區段的報導內容分析，得到獨立新聞媒體之基改食品報導同時存在「風險」以及「危機」論述。國內民間組織將基改食品搭上台灣食安議題的風潮，同時結合選舉的話題，引起媒體追隨，將基改食品視為食安危機底下的一環。此外，國、內外基改食品的政策頒布（由其與台灣貿易關係緊密的美國）也是促使報導數量增加的關鍵因素。而藉由國外對於基改食品的相關研究報告，獨立新聞媒體也企圖梳理基改食品衍伸的爭議，向讀者進行基改食品的風險傳播。

五、結論

在獨立新聞媒體之基改食品報導中，主要由七個行動者所構成，分別為「基改」、「農業」、「食品安全」、「利害關係人」、「政策行動」、「地理區位」及「爭議討論」。在利害關係人中，「政府部門」、「業者」及「消費者」是經常被提及的對象；然而實際掌握發言權的僅有「政府部門」及「專業人士」。在報導數量趨勢上，獨立媒體對基改食品之報導並無明顯向上或向下攀升之趨勢，而是起伏不定。藉由報導數量「攀升區段」之內容探究，得到基改食品報導同時存在「危機」及「風險」兩種傳播論述型態。

從過往相關研究中得知，在基改食品報導的議題框架中，多數為「人類健康」的討論，「食品安全」框架並不突出。然而「健康」與「食安」的概念雖然頗為類似，但仍然存在差異；「健康」關注對人體的益處、營養，「食安」則是強調人對食物的恐懼、懷疑（Cahill & Morley, 2010）。因此，「食品安全」成為基改食品報導中行動者一員，與台灣近年來食安問題風波不斷之社會情境脈絡緊密相關，「基改食品」被部分利害關係人視為是問題食材。此外，「有機」為基改食品報導中主要的反面論述，並且絕大多數呈現「正面」形象，「有機」食材成為替代基改食材的最佳選擇。

在基改食品報導中，雖然基改食品與「消費者」、「農民」具有直接影響關係，但是在獨立新聞媒體當中兩者的言論是被忽視的。而這樣的研究發現與謝君蔚和徐美苓（2011）分析國內「主流媒體」的結果一致，具權威性或專業性的消息來源在基改相關報導中仍被視為較具參考價值。對於「政府部門」之引述以政策頒布及市場抽樣調查之結果報告為主；「專業人士」則偏好引述學術領域以及科學領域之見解。「學術領域」對基改食品採取堅決反對的態度，其中又以郭華仁教授之發言最受獨立新聞媒體青睞；「科學領域」則涵蓋了正、反立場，然而即使持反對立場，也傾向使用較保守的用詞。值得一提的是，義美食品公司總經理「高志明」在「業者」的引述中幾乎佔了一半，其強烈「反基改」的言論成為獨立媒體中業者的主要觀點，並壓縮了其他異議業者的聲音。

破除過往媒體偏好擔任政府「傳聲筒」之想像，由民間團體所發起的「校園午餐搞非基」活動更是引起了獨立新聞媒體對於基改食品議題的大幅報導，獨立新聞媒體不再單純追隨政府政策的腳步，甚至成為監督政府施政效能的角色。在報導論述本質上，危機論述有其明顯的「在地性」，故在報導中承襲了台灣近年來一系列的食安問題，將基改食品視為是國內食安危機中的一員。此外，國外基改食品的相關研究報告也會促使獨立新聞媒體追隨關注，並藉此重新向讀者梳理基改食品可能造成的潛在風險。

在未來研究發展上，本研究將持續蒐集獨立新聞媒體之基改食品報導；並進一步各別分析此六大獨立新聞媒體，探究不同獨立新聞媒體是否對基改食品也抱持著相異的意識形態。此外，研究者目前找到國內關於主流媒體在現基改食品之最新研究為謝君蔚和徐美苓（2011）的期刊論文；然而該論文至今也相隔五年之久，主流媒體在這期間所強調的議題框架或許有所轉變、甚至可能有新的議題框架產生，故本研究也計畫建置台灣主流新聞媒體之基改食品報導的語料庫，進行獨立與主流新聞媒體之間，在基改食品報導呈現上之比較。

參考文獻

賀照緹（1993）。小眾媒體・運動文化・權力──綠色小組的運動形式及生產條件分析。輔仁大學大眾傳播研究所碩士論文，未出版，台北。

游美惠（2000）。內容分析、文本分析與論述分析在社會研究的運用。調查研究期刊，8: 7-42。

侯新龍（2007）。基因改造作物的發展現況及爭議。教師之友，48: 5-9。

孫智麗、許嘉伊、劉翠玲（2007）。我國消費者對基因改造食品認知程度與接受度之調查分析。農業生技產業季刊，食品生技，11: 80-85。

成露茜（2009）。另類的媒體實踐。《批判的媒體識讀》，頁 371-387。臺北：正中。

管中祥（2009）。光影游擊最前線：台灣另類媒體 2007-2008。新聞學研究，99: 201-220。
郭良文（2010）。蘭嶼的另類媒體與發聲：以核廢料與國家公園反對運動為例。中華傳播學會，17: 43-74。
莊豐嘉（2011）。台灣公民新聞崛起對公共政策之衝擊──從樂生、大埔到反國光石化事件之比較分析。台灣大學政治學研究所碩士論文，未出版，台北。
管中祥（2011）。弱勢發聲、告別汙名：台灣另類「媒體」與文化行動。傳播研究與實踐，1（1）: 105-135。
謝君蔚、徐美苓（2011）。媒體再現科技發展與風險的框架與演變：以基因改造食品新聞為例。中華傳播學刊，20: 143-179。
張傳佳（2012）。獨立/主流媒體的環境報導──以國光石化開發案為例。台灣大學新聞所碩士論文，未出版，台北。
劉家瑜（2014）。新舊時安法比一比廠商停看聽。貿易雜誌，276: 50-53。
科技新報（2014）。基因改造作物先誰先受害？是環境還是人？科技新報，生物科技。民國 103 年 7 月 19 日，取自：http://technews.tw/2014/07/19/gm-crop/。
台灣經濟研究院生物科技產業研究中心（2014）。基因改造產業發展與趨勢報告（2013），台灣經濟研究院生物科技產業研究中心。
管中祥（2014）。另類媒體的生存。天下雜誌，獨立評論。民國 103 年 6 月 26 日，取自：http://opinion.cw.com.tw/blog/profile/47/article/1565。
林意璇（2015）。台灣報紙再現同性婚姻的語料庫與論述分析。政治大學新聞研究所碩士論文，未出版，台北。
郭文平（2015）。字彙實踐及媒介再現：語料庫分析方法在總體經濟新聞文本分析運用研究。新聞學研究，125: 95-142。
創世紀雙周刊（2016）。新聞資訊網站調查與使用概況。創世紀雙周刊，59: 1-29。
Cahill, S., & Morley, k. (2010). Coverage of organic agriculture in North American newspapers. *British Food Journal*, 112(7): 710-722.
Dibden, J., & Gibbs, D., & Cocklin, C. (2013). Framing GM crops as a food security solution. *Journal of Rural Studies*, 29: 59-70.
Lore,T. A., & Imungi,J. K., & Mubuu. K (2013). A Framing Analysis of Newspaper Coverage of Genetically Modified Crops in Kenya. *Agriculture & Food Information*, 14: 132-150.
Morse, S. (2016). They Can Read All About It: An Analysis of Global Newspaper of Genetically Modified Crop Varieties Between 1996 and 2013. *Agriculture,* 45(1): 7-17.
Nisbet, M. C., & Lewenstein, B. V. (2002). Biotechnology and the American media: The policy process and the elite press, 1970 to 1999. *Science Communication*, 23, 359-390.
Price, J. (2006). *Burgers and broccoli: The framing of food on U.S. news magazine covers, 1994-2003*. Paper presented at the annual meeting of the International Communication Association, Dresden, Germany.

Paper Session 5

古今連結：文化資產

Cultural Heritage: Connecting Past & Present

Basic Cultural Elements, Seemingly Unrelated Yet Connected：Spatiotemporal Mapping Early Historical Religious Networks Points in Indo-Pacific Austronesia

David Blundell*

Abstract

Developments in digital humanities have matured crossing boundaries of geography and history to include a wider range of disciplines. Advances in geographic information systems (GIS) computing and information infrastructures offer researchers the possibility of reconsidering the entire strategy of analysis and dissemination of information. It enables humanities scholars to discover relationships of memory, artifact, and experience that exist in a particular place and across time. We are challenged to imagine new methods for doing research and making results available to broader user communities. Can we find meaning and innovation (through) digital humanities beyond what has been traditionally part of scholarly efforts?

This paper is about integrating an atlas of historical data through our mutual trans-disciplinary synergies. It is a collaborative project with the Electronic Cultural Atlas Initiative (ECAI) Austronesia Team and Atlas of Maritime Buddhism (http://ecai.org/projects/MaritimeBuddhism.html), University of California, Berkeley. We comprehensively search for spatiotemporal points to where religious networks were transmitted through sea voyaging. Our research indicates Austronesian-speaking navigators were plying the Indian Ocean in the first millennia BCE circulating in Southern Asia carrying Indic merchants and *dharma* monks across tropical seas to mainland Southeast Asia far-flung islands (e.g., Indonesia).

Keywords: Indo-Pacific Austronesia, Monsoon Asia, GIS spatiotemporal mapping, religious networks, archaeology, early history

* Asia-Pacific SpatioTemporal Institute (ApSTi) Top University Project in Digital Humanities Research and Innovation-Incubation Center, and International Doctoral Program in Asia-Pacific Studies National Chengchi University, Anthropology / Language Editor Electronic Cultural Atlas Initiative (ECAI), UC Berkeley. Email: pacific@berkeley.edu.

看似無關，實則連結：印度-太平洋南島航行與宗教網絡的數位人文時空地圖集

卜 道*

摘 要

數位人文學的發展已臻成熟，跨越地理學、歷史學，涵括了範圍更為廣大的學科領域。地理資訊系統在電算能力與信息基礎結構的進步與發展，提供研究者重新思考整體的分析策略與資訊傳播的可能性。同時，也讓人文學者能夠發現長時期存在於一特定地的記憶、文物與經驗，彼此之間的關聯。學者面對的挑戰來自於想出做研究的新方法，並能夠將研究成果提供給更多的社群參考使用。在傳統的學術方法與努力之餘，我們還能透過數位人文學找到意義與創新嗎？

本篇論文闡述藉由跨學科的多元合作，繪製整合歷史資訊地圖。這是加州柏克萊大學電子文化地圖協會之南島研究團隊與海上絲路地圖集的共同計劃。透過全面而整合地尋找因為海上航行而傳播擴散的宗教網絡在某一特定的時空據點，我們的研究顯示南島語族的航海者，在西元前1000年的時候，就已經載著從印度文化圈而來商旅與行遊僧人，航行在印度洋上，逡巡於南亞地區，跨越了熱帶海洋，也到達東南亞半島，往返在南海的諸多島嶼之間。

關鍵字：印度/太平洋南島語族、季風亞洲、地理資訊系統時空製圖、宗教網絡、考古學、早期歷史

* 國立政治大學研究暨創新育成總中心，頂尖大學數位人文計畫亞太時空資訊研究室，政大社科院亞太研究碩博士學程，加州柏克萊大學電子文化地圖協會人類學 / 語言編輯，Email: pacific@berkeley.edu。

1. Introduction

Developments in digital humanities have matured crossing boundaries of geography and history to include a wider range of disciplines. Advances in geographic information systems (GIS) computing and information infrastructures offer researchers the possibility of reconsidering the entire strategy of analysis and dissemination of information (Gregory 2014). It "enables humanities scholars to discover relationships of memory, artifact, and experience that exist in a particular place and across time" (Bodenhamer 2010 *et. al.*). We are challenged to imagine new methods for doing research and making results available to broader user communities. Can we find meaning and innovation digital humanities beyond what has been traditionally part of scholarly efforts?

Histories are always multiple and incomplete. For any aspect of the past, there may be many narratives or none. It is useful to distinguish between the past, what happened, thus history, accounts of the past; and heritage, which is those parts of the past that affect us in the present. To be more precise as a student of history, it depends on the documentation of the past. That is to say the events that have transpired are no longer directly knowable. The past is knowable only indirectly through histories in various forms such as descriptions and narratives of what happened (Buckland 2004).

This paper is about integrating an atlas of historical data through our mutual trans-disciplinary synergies. It is a collaborative project with the Electronic Cultural Atlas Initiative (ECAI) Austronesia Team and Atlas of Maritime Buddhism (http://ecai.org/projects/MaritimeBuddhism.html), University of California, Berkeley (see Blundell 2014). We comprehensively search for spatiotemporal points to where religious networks were transmitted through sea voyaging. Our research indicates Austronesian-speaking navigators were plying the Indian Ocean in the first millennia BCE circulating in Southern Asia carrying Indic merchants and *dharma* monks across tropical seas to mainland Southeast Asia far-flung islands (e.g., Indonesia).

Our purpose is to explore information on ocean transport networks of religions from ports of India and Sri Lanka across Monsoon Asia. These original field-research findings[1] are based on documentation of pilgrims and their routes, early ship technology, navigation, and archaeology. It is to study the interplay of ancient cultural pursuits in the archaeological and

[1] Research conducted in 2016 funded by a Taiwan Ministry of Science and Technology (MoST) grant.

textual records (see Lammerts 2015, Munoz 2016).

Methodological questions were created on issues of research design and strategy as an empirical science. The research examines what extent did international religious systems, such as beliefs in the *dharma*, beginning about 2,300 years ago, spread into ocean island areas of Monsoon Asia facilitated by Austronesian navigation?

The aim is to recount narratives from historical records of religious transmissions, aesthetics, and partnerships implicit as conceptual underpinning of advanced hermeneutics research in our qualitative tradition, critical, and able to enrich and deepen perspectives *based on cultural elements seemingly unrelated, yet connected*.

Our spatiotemporal interfaces provide new methods of integrating primary source materials into crosswalks of interactive visualizations (Blundell and Zerneke 2014). Utilizing GIS we are able to chart the extent and dynamics of specific traits of cultural information creating layered maps. These elements are transmitted and based in places through languages and belief systems across Indo-Pacific Austronesian island systems.

We are reexamining through digital and spatial humanities the extent of religious transmission and its influence on humanity in the present creating a time-enabled Web-based, animated, visual, colorful, anthropological cultural atlas serving as *a local community bulletin board and for scholarly exchange*.

2. Research Applications in Spatial Humanities

The first important factor is synergy among the researchers in *Old Maps, New Views* at our Asia-Pacific SpatioTemporal Institute (ApSTi), Top University Project in Digital Humanities, National Chengchi University. We research independent projects, yet collaborate and integrate the project with GIS shared tools to create a unified spatiotemporal map. The challenge accepted by our projects is to break new ground, developing new knowledge using digital tools to produce results that could not be achieved through traditional research in any single discipline. It is leveraging data from disparate databases to create integrated systems and customizable visualizations. The shared infrastructure enables a collaborative method, linking scholars and their work from around the Pacific and beyond. The projects address the issues of geographic and temporal representation that often makes use of maps in history challenging. Best practices are being developed collaboratively with our global community of

scholars to document data sources and types of uncertainty and ambiguity to enable and encourage re-use of the data by future scholars. Collaborating with Academia Sinica open source archives supports sustainability and open access for project results.

Our mutual integrated research mapping strategies and methodologies are based on cultural and environmental attributes. These attributes are compiled in a GIS gazetteer style spread sheet for generating spatiotemporal mapping data sets. As with the other projects members, there is a creation of an annotated interface across the Individual projects strengthened by the synergy of collaboration.

The first important factor is synergy among the other researchers of this proposal. We will share a commonality based on a platform of mutual research strategies and methodologies based on cultural and environmental attributes. These attributes are to be listed in a GIS gazetteer style spread sheet for generating spatiotemporal mapping data sets. As with the other project members, we are creating an annotation interface across the sub-projects, strengthening the synergy model of the propose collaboration. The geo-referenced cultural entities could be stored in a Mutually Agree GIS-enhanced relational database such as PostgreSQL with PostGIS, allowing for automatic data input and processing in combination with a multiple user and multiple device annotation and processing, guaranteeing at any time the referential integrity of the data through the technologies of locking, triggers, constraints, and typed data. The relational database is the back-end for storing, processing and mapping the different data sources. It also connects to different front-ends for data input and data management, e.g., Web-pages, command-line interface tools for data analysis, such as ArcGIS, GRASS and R. Satellite images, digital maps, shapefiles and raster data are created and managed in a distributed revision control system such as Git. After being modified or created by individual researchers, they are checked in and become equally available on the file server for Web-services and analysis tools. This is a technical contribution from Integrated Project member Oliver Streiter's tomb research methodology.

Data analysis also includes infrared scanning and high definition 3D modeling (Chandler 2005 *et. al.*) provided by Department of Land Economics, National Chengchi University, for *Old Maps, New Views* projects. For example, Borobudur, Java, Indonesia, the outrigger ship '*oru*', 9th-century CE, will be scanned for analysis of unseen attribute components (see Fig. 6).

Project management derives from the mandate of our integrated projects based on "GIS Outreach" supported by our Core team member Jihn-Fa Jan, Department of Land Economics,

National Chengchi University, is working with his students on experimentation on 3D viewing using digital camera images. They are using Pano2VR software to stitch many overlapping photographs to produce a panoramic view that can be shown on a regular browser. Viewing a file just requires dragging it to an Internet browser. Adobe Shockwave Player is downloaded for viewing the files. Using Pano2VR can produce high-quality panoramic view, which can be integrated with our integrated projects Website (www.apsti.nccu.edu.tw).

Therefore our individual research projects are not separate and will contribute to lists of data and ways of managing the data sets. For specific requests, research specialists are volunteering findings from archaeological sites, libraries, museums, private collections, and written materials, related to:

religious networks
attributes and motifs
multilingual scripts
landscapes (forms of land, sea to mountain)
materials: stone, wood, earthenware, tiles, etc.
water systems and flows with depths and mud flats
sites with locational ‘sense of place’ (localities), geomancy, sacred sites
mapped points, and point clusters (spatiotemporal points)
migrations and movements (pathways)

Collaborators trained on the common reference set can handle the collaborative annotation interface and standard analysis tools to work on the data, in the annotation and the analysis, irrespective of the individual projects of *Old Maps, New Views*, reducing the amount of training needed and increasing the flexibility of the team when tackling specific questions, such as the spread of a specific features across the region of Monsoon Asia with different environments and features (e.g., temples, tombs, monuments, block impressions, motifs, written sculpture, earthenware, and landscapes) with time layers.

Fig. 1. Monsoon Asia. Spruneri composite map of India and Southeast Asia in ancient times, including Ptolemy's *Geography* (c. 150 CE) in lower left corner, 1855.

3. Project of Intersections, Points and Connecting Lines, Not Boundaries

This project is about integrating a spatiotemporal atlas of historical Indo-Pacific regions through our mutual interdisciplinary synergies to comprehensively search for points to where cultural networks were transmitted through sea voyaging. We integrate information of historical Monsoon Asia radiating from prehistory with Taiwan as a point of reference for Austronesian navigation networks visualized in dynamic maps. Our research is based on geographic information systems (GIS) and digital and spatial humanities to find '*elements seemingly unrelated, yet connected*' as a holistic spatiotemporal atlas.

Across the region of Monsoon Asia, there is documented evidence of Austronesian navigation out of Taiwan dating back as to about 2000 BCE (Bellwood 1995; Bellwood and Dizon 2013). By the first millennium BCE, outrigger '*oru*' voyaging spread to the Malay-Indonesian side of the Indian Ocean, and voyaged regularly to South Asia and East Africa by first millennium CE (Blench 2010).

About 500 BCE, the Megalithic cultures of Southern Asian were transforming into a literary religious system known as the *dharma*. By this time, Austronesian navigation was widespread. My hypothesis is that when the *dharma* spread abroad from about 300 BCE, Austronesian navigators transported it.

The project Atlas will be an array of points with enriched GIS data including the attributes in text and spatiotemporal visualization record. Data collections are being sourced across the region on Monsoon Asia from my personal research as data collected over the years, and new sources being discovered.

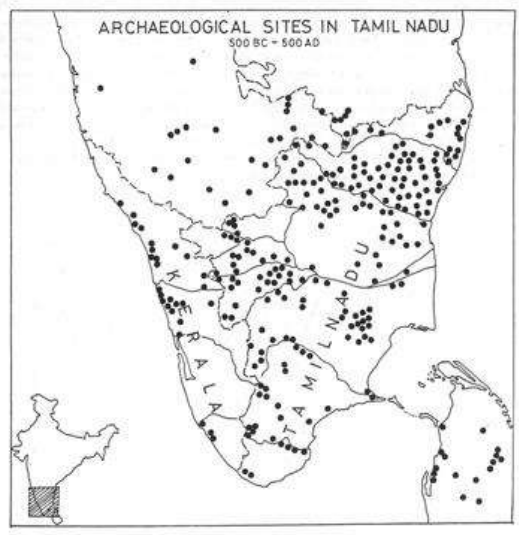


Fig. 2. South India, Sri Lanka, 500 BCE to 500 CE Megalithic sites transforming to *dharma* sites. See Himanshu P. Ray, *The Winds of Change: Buddhism and the Maritime Links of South Asia*. Oxford: Oxford University Press, 1994, p. 15. (data from N. R. Banerjee, *The Iron Age in India*, New Delhi, 1965, pp. 41-43).

Our supposition is the *dharma* as a literary belief system was carried as far as writing could be traced on palm leaves, metal, and stone. Our further hypothesis is that the *dharma* moved out by sea travel onboard ships with seasoned mariners who I suppose were Austronesian navigators as mentioned in Southern Asian literature and in stone relief imagery. Yet, there are gaps in the record. So to remedy this, I am taking my lifetime pursuit of knowledge, and academic network to further trace the extent of seemingly unrelated cultures intersected, and its periphery.

Can we find meaning and innovation digital humanities beyond what has been traditionally part of scholarly efforts? Our spatiotemporal interfaces provide new methods of integrating primary source materials into crosswalks of interactive visualizations. Utilizing geographic information systems (GIS) we are able to chart the extent and dynamics of specific traits of cultural information creating layered maps. These elements are transmitted and based in places through languages and belief systems across Malay and Indonesian island systems.

A far-reaching goal of the project is to further standards in cartographic strategies through the utility of digitalization and animation of old maps content format giving new possibilities in the hands of local and international collaborators. In our research, we have found that sea ports are orientated with a mountain peak serving as a navigational point (see below). 3D mapping for the project could provide new guidance for developing best practice

standards applied to databases giving interactive multimedia utility aspects. This allows uniting the context of environmental landscapes with cultural data for making new enhanced possibilities in spatial humanities of scholarly results.

Fig. 3. Bujang Valley, Kedah, Malaysia. *Dharma* related archaeological sites, 2nd-13th-century CE. Composite topo-map of sites Merbok River estuary (left) (Wheatley 1961, Fig. 44) with picture of Mount Jerai (right), a navigational maker to the Bujang port.

When the *dharma* was being transmitted physically, culturally, and linguistically through channels created by Southern Asian pilgrims travelling abroad, what other forms did the *dharma* take, what did it become, and why it spread elsewhere? As a philosophical term, the concept *dharma* is perhaps most commonly used to refer to a scope of mental and physical laws of nature—and related back to what people experience, and perceive as truths, or phenomena. There are innumerable systems employed for splicing phenomena into different categories. For instance, the individual can be viewed as a conglomeration of aggregates—the body, feelings, perceptions, mental formations, and consciousness—all of which are *dharma*.

The concept *dharma* predates the Buddhist tradition, appearing in the texts of the Vedic hymns. Buddhism is a continuum of Southern Asian beliefs, and adopted the term referring to (1.) the collection of the Buddha's teachings, which are recorded in various compilations of *sutra*, and (2.) phenomena of the belief in nature in general. The collection of the Buddhist teachings varies in scope depending on the region. For example, various indigenous texts have come to be regarded as canonical throughout the history of Chinese Buddhism, thus Chinese, Tibetan, and Thai Buddhists all refer to something different when they say "Buddha-Dharma."

However, in principle there are core texts, practices, and beliefs that connect Buddhist traditions together. The essence of the dharma should be the same everywhere, yet there is debate as to what constitutes this essence. That is, until an individual confirms the truth of the *dharma* through practice.

For the Buddhist practitioner working on understanding the nature of reality, there are Three Jewels to help. They are the *dharma* (cosmic law), the essential form of the Buddha (awakened), and the *sangha* (community). With these three attributes to embrace, the believer grasps onto something that gives a sense of worth and continuity with the world at large.

We use *dharma* as a concept to describe the early South Asian notion of maintaining cosmic order by spiritual awareness, devoted to a path, and assuring human continuity with respect for life. It originated in literature from Vedic times (1500 to 500 BCE), and grew with Buddhism from the 3rd century BCE. In the 1st millennium it spread as a concept or notion ushering Southern Asian beliefs carried through derivatives of the Sanskrit language, and local derivatives. These beliefs were considered Buddhist, Jain, and Hindu—all modern glosses for historical and complex sets of beliefs deriving from sources predating the present day organized religions. Yet, there is an underlying aesthetic sense of a flourishing *dharma*, or cosmic flow, from the Neolithic circulating across the Indian Ocean. Buddhism, a complexity of pre-Vedic derived beliefs, and pre-existing notions mixed to produce a *raj* (kingdom)—the authority ordering civilization in place and time. I see this as key to understanding the cohesive weave of concepts found in *rasa*: a holistic aesthetic value system of fundamental ideas for life that are pervasive in the region. It is an aesthetic system dealing with the value of perception, taste, and related experiences. It identifies traits and clues within the workings of a culture to create understandings and judgments. This aesthetic experience forms intrinsic attitudes vis-à-vis things specifically or generally recognized as worthy of attention in a society (see Blundell 1996).

Early metropolitan Southern Asia ascended through cultural exchange from a prehistoric common dominator shared across the Monsoon region. Urbanization was seeded from a spirit that proliferated cultural diversity with a social structure of theocratic civil administration. The vital energy embodied in an ancient standing stone was refined to immaculate form: a stone stele depicting priests backed against the monistic divine. The transition from the Neolithic to metal tools and civilization transformed development processes with continuity in Southern Asia.

The *dharma* circulated by sea, yet my question concerns the navigators who operated the ships. Going back in time, to early history to the Vedic days about 3,500 years ago, and before writing to the Neolithic cultural stratum, across Southern Asia, coexistence with the solar cycle (wheel), air (ether, wind, or lofty), soil (earth layer, organic base), and water (conveying hydraulics from lakes, rivers to oceans) meant the basis of substance. It was a prime for basic life, and for humans to respect, and give their observance, at first in the oral traditions, later as text.

We believe the seafaring Neolithic cultures of Monsoon Asia expanded and continued on demand for the trade opportunities. Today known as Austronesian speakers, this language family navigated the seas along the East Asian coast, to Taiwan, the archipelagos of the Philippines, Indonesia, and Malaysia. These navigators sailed directly from Borneo to East Africa, landing on Madagascar. The voyages are traced through artifacts and languages from the first millennia BCE. The Austronesian speaking navigation diaspora provided transportation to merchants and monks plying to Southeast Asia. What they shared in common were animist origins, a belief in nature—the stars, flora of life, sacred rocks, flowing water, gales of prevailing winds. At ports, like the ones found in Southern Sri Lanka, Malaysia, and Indonesia, where mariners and merchants (Devendra 2013) boarded ships with outrigger to traverse Monsoon ocean expanses.

Around the 3rd century BCE to the tenth century CE from South Asian ports launched an unfettered dharma across to new lands carried by heavy cargo laden ships. Through using evidence from texts, inscriptions, ethnography, and archaeology, it is traced though regional, yet interconnected networks across Southern Asia, origins of the state, aesthetic systems, trade routes, and theocracies.

In India Buddhist shrines flourished across the land and many were later re-molded by Hindu practitioners. A multitude of beliefs prevailed from a megalithic common denominator coalesced as the overarching literary dharma, then continued in transformation through clashes supported by political and military powers. These processes continue up through the present day.

Our purpose is to explore information and research on transport networks of religions in South and Southeast Asia, documentation of pilgrims and their routes, ethnology of ship technology, navigation, and historic climate variations. ECAI Austronesia Team and *Atlas of Maritime Buddhism* supports mapping sites for constructing a method to integrate data into an

interactive map interface. The research includes to explore the physical feasibilities of the stitched and lashed sea craft ‘*oru*’ and its variations across Monsoon Asia: a navigation network of staged voyaging across the Indian Ocean to Southeast Asia and the South China Sea. Destinations were to seats of kingdoms and trade centers where the word of the *dharma* and its faith developed in a healthy or vigorous way, especially as the result of a particularly congenial environment of the region. We trace the earliest evidence of trans-ocean sailing craft across Monsoon Asia.

Many factors influence what histories are, or can be written. As heritage is legacy from the past that we compose life with in the present and give to the future as reference for local identity, it’s also a marker for universal human appreciation.

4. What Are We Looking For?

The project is not about making collections for a digital archive. It is about locating rare data, few and far between. Data in the time frame is important, yet not systematically connected before. The data represents artifacts related to the spread of religious beliefs attributed to the South Asian *dharma* dating to 3rd century BCE to an opened ended time frame (part of the research question, e.g. 13th century CE). The *dharma* from South Asia provided primary teachings based on *ahimsa* (non-violence), natural forces, and deity systems attributed to early Hindu and Buddhist practitioners. Physical manifestations are found in:

1. Port facilities (seaside and upriver)
2. Landmarks (mountains and monuments)
3. Architectural styles
4. Geometric designs, symbols, motifs, e.g., auspicious plants and animals
5. Statutes and figurines
6. Landscaping and water systems
7. Trade beads (used initially as prayer beads) and amulets
8. Writings inscribed in stone, pottery, and palm leafs.

These are mostly fragments, broken and scattered pieces. They are found in the custody of museums, libraries, and private collections, and on land terrain (as mountain peaks), or sea terrain (as shipwrecks). How could they be sourced?

Fig. 4. Studied religious network points (Blundell 1975, 2003), registered in a GIS gazetteer by Jeanette Zerneke.

We are able to chart the extent of specific traits of cultural information on a map through GIS gazetteer style spreadsheets for generating data sets. Before the data was expressed in one dimension on maps with boundary lines and drawn-in dashes. Our project system is based on 'GIS points' of enriched data information. These are charted and visualized through computational analysis creating an innovative digital 'malleable structure'. This gives the researchers an expanse of data in layers of time depth across space.

Using methodologies in the digital humanities offers new ways of finding seemingly hidden evidence and fresh ways of connecting large arrays of scattered attributes. Archaeology is also periodically revealing new findings of artifacts in larger quantities. Conventional research methods in the social sciences are my case studies. These case studies contributions serve as a data points. An example of a new source finding for the project is stone relief panel, 9th-century CE, from Borobudar, Java, Indonesia, discovered at Bujang Valley, Malaysia, coinciding with Javanese occupation (see below). This one example connects Bujang Valley and Borobudar through artifacts. The project will collect such data to plot points on a GIS map showing connecting lines between the points, and clusters of points to visualize nodes.

Fig. 5. Buddhist stone relief panel, 9th-century CE, from Borobudar, Java, Indonesia, found at Bujang Valley, Kedah, early trading port in Malaysia, giving evidence of linkage to Java.

5. Some Research Attributes

'Mariners, merchants, monks' (Devendra 2013) provides research attributes for the GIS mapping and historical narrative. For mariners, the objective is locating where navigators voyaged, from what ports, and to trace routes. For example tracing Indian Ocean navigation: (1.) a type of ship, outrigger '*oru*' of Austronesian design found on the stone relief panel of Borobudar, Java, Indonesia. For merchants, what trade goods are found, such as pottery, metal, and other items to be identified? For example: (2.) a bronze bowl est. 2nd-centry BCE of South Asian design found at Khao Sam Kaeo, Muang District, Chumphon Province, peninsular Thailand. And for monks, there are religious *dharma* artifacts, temples, and other evidence of the beliefs found in iconography, symbols, and motifs.

For example: (3.) inscription of Buddha-Guptha found at Seberang Peri, with *dharma* teachings and *stupa* image, containing Sanskrit in Pallava-Grantha script: giving thanks for a safe passage from Raktampittika, 5th-century CE, Bujang Valley, Kedah, Malaysia.

What are the attributes?

1. Contextual location if finding (provenience)
2. Present location (collection)
3. Date of origin (approx.)
4. Size (dimensions)
5. Weight (if applicable)
6. Material content
7. Inscription (if applicable)
8. Art motifs.

The art motifs include geometric designs, symbols, motifs, e.g., auspicious plants and animals. These attribute classifications are useful for the other *Old Maps, New Views* projects for the analysis of motifs that appear on tombs, woodblock impressions, or in literary forms. The project GIS gazetteers will generate maps to include detailed information for sharing in the group of projects. This comprehensive cultural history of Southeast Asia from prehistory to Neolithic and Bronze-Iron age times through to the major Hindu and Buddhist civilizations, to around 1,300 CE. Southeast Asia has recently attracted archaeological attention for the first recorded sea crossings; as the region of origin for the Austronesian population dispersal across the Pacific from Neolithic times; as an arena for the development of archaeologically rich Neolithic materials (Glover and Bellina, 2011).

Ship

Fig. 6. Outrigger ship '*oru*' found on the stone relief panel, 9th-century CE, Borobudur, Java, Indonesia

.**Bowl**

Fig. 7. Bronze bowl, woman and man, lotus bottom, 'Austronesian' circular 'stamped style' impressions, est. 2nd-century BCE, Khao Sam Kaeo, Muang District, Chumphon Province, peninsular Thailand (see Glover and Bellina, 2011).

Inscription

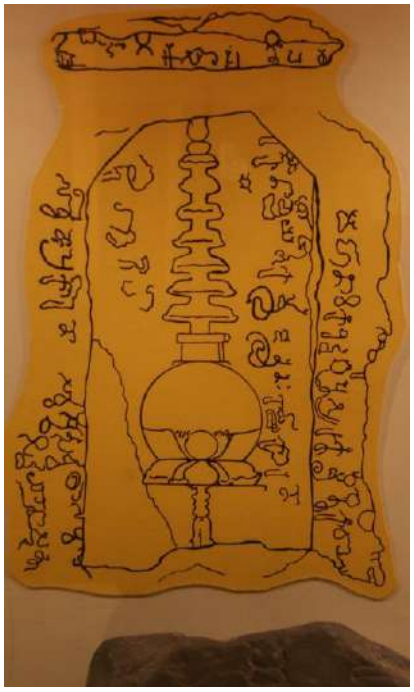


Fig. 8. Buddha-Gupta inscription on stone, from Seberang Peri, with *dharma* teachings and *stupa* image, containing Sanskrit in Pallava-Grantha script: giving thanks for a safe passage from Raktampittika, 5th-century CE, Bujang Valley, Kedah, Malaysia (Blundell 2015).

6. Complex Project, No It's Simple

To the reader, it could seem this project is vast and complex. The project is about finding intersections between Austronesian navigators and practitioners of the *dharma*. The simple question is where do they occupy the same location and time? What seems to be vast and complex is the data collection, yet not so. Our network has helped over the years, and now through spatial humanities mapping, the enthusiasm has increased for *looking at old data with new ways of viewing it*. The power of digital humanities to sort out seemingly unrelated arrays of information into comprehensible and meaningful understandings is the key to the project. We are soliciting old materials and new findings for computational analysis. Again, the research is to study the interplay of ancient cultural pursuits in the archaeological record. Methodological questions were created on issues of research design and strategy as an empirical science.

To summarize, the project is about mapping with GIS points indicating *dharma* intersecting with Austronesian evidence. The data is scattered and seemingly unrelated. Yet, the project has enlisted numerous volunteers interested in registering their collected historical data in spatiotemporal coordinates. This in turn is entered into gazetteer style spread sheets generating dynamic maps.

7. Digital Humanities Significance

The challenge accepted by this project is to break new ground, developing new knowledge using digital tools to produce results that could not be achieved through traditional research in any single discipline. It is leveraging data from disparate databases to create integrated systems and customizable visualizations. The project infrastructure enables a collaborative method, linking scholars and their work from around the Pacific and beyond. By using the methodologies developed by ECAI and it's affiliates for the *Atlas of Maritime Buddhism*, the project is addressing the issues of geographic and temporal representation that often makes use of maps in history challenging. Best practices are being developed collaboratively with ECAI and our global community of scholars to document data sources and types of uncertainty and ambiguity to enable and encourage re-use of the data by future scholars. Collaborating with Academia Sinica and ECAI open source archives supports sustainability and open access for project results.

To Sum Up, We are:

- Integrating with open source archives to support sustainability.
- Using digital tools to create something that couldn't be done before.
- Using collaborative methods to link scholars and their work.
- Using best practices of documenting sources and levels of uncertainty and ambiguity to enable re-use of the data by future scholars.

References

Banerjee, N. R. 1965. *The Iron Age in India*. New Delhi.

Bellwood, Peter. 1995. Austronesian prehistory in Southeast Asia: Homeland, expansion and transformation. *The Austronesians: Historical and Comparative Perspectives.* P. Bellwood, J. Fox, and D. Tryon, eds. Canberra: Australian National University, Department of Anthropology of the Research School of Pacific and Asian Studies, comparative Austronesian project publication. Pp. 96-111,

Bellwood, Peter, and Eusebio Dizon. 2013. The archaeology of the Batanes Islands, northern Philippines: 4000 years of migration and cultural exchange. *Terra Australis* 40.

Blench, Roger. 2010. Evidence for the Austronesian voyages in the Indian Ocean**.** In *The Global Origins of Seafaring*. Atholl Anderson, J.H. Barrett & K.V. Boyle eds. Cambridge: McDonald Institute. Pp. 239-248.

Blundell, David. 1975. *Metropolitan Ascent of Southern Asia*. Ms. based on of early Indo-European, Dravidian, and Chinese literature for the understanding of the growth and dynamics of ancient cities and trade routes in Southern to East Asia and across the Indo-Pacific oceans.

Blundell, David. 1996. Aesthetic ethos. *Bulletin of the Department of Anthropology*, National Taiwan University 51: 43-58.

Blundell, David. 2003. Metropolitan ascent of Southern Asia (partial version). *Proceedings for the Symposium on Indian Religions, Art and Culture*. Taipei: National Museum of History. Pp. 103-136.

Blundell, David. 2014. Dharma civilization and stitched outrigger navigation: Contributions to ECAI Project on Maritime Buddhism. *Buddhist Culture and Technology: New Strategies for Study*. International Council for 11th United Nations Day of Vesak. Buddhist Perspective towards achieving the UN Millennium Goals. Vietnam. Pp. 41-63.

Blundell, David. 2015. Bujang valley—The seat of all felicities. *Eastern Horizon*. May. Pp. 17-21.

Blundell, David, and Jeanette Zerneke. 2014. Early Austronesian historical voyaging in Monsoon Asia: Heritage and knowledge for museum displays utilizing texts, archaeology, digital interactive components, and GIS approaches. *International Journal of Humanities and Arts Computing*. 8: 237-252.

Bodenhamer, David J., John Corrigan, and Trevor M. Harris. 2010. *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Bloomington & Indianapolis: Indiana University Press.

Buckland, Michael. 2004. Histories, heritages, and the past: The case of Emanuel Goldberg. *The History and Heritage of Scientific and Technical Information Systems.* W. B. Rayward and M. E. Bowden, eds. Medford, NJ: Information Today. Pp. 39-45.

Chandler, J. H., J. G. Fryer and H. T. Kniest. 2005. Non-invasive three-dimensional recording of aboriginal rock art using cost-effective digital photogrammetry. *Rock Art Research* 22(2): 119-30.

Devendra, Somasiri. 2013. Mariners, merchants, monks: Sri Lanka and the eastern seas. In Satish Chandra and Himanshu Prabha Ray, eds. *The Sea, Identity and History: From the Bay of Bengal to the South China Sea*. New Delhi: Society for Indian Ocean Studies, 2013. Pp. 169-220.

Glover, C. Ian, and Bérénice Bellina, 2011. Ban Don Ta Phet and Khao Sam Kaeo: The earliest Indian contacts re-assessed. In Manguin, P.-Y. and Mani eds. *Early Interactions between South and Southeast Asia: Reflections on Cross-Cultural Exchange*. Singapore, New Delhi: ISEAS, Manohar. Pp. 17-46.

Gregory, Ian N. and Alistair Geddes. 2014. *Toward Spatial Humanities: Historical GIS and Spatial History.* Bloomington: Indiana University Press.

Lammerts, D. Christian, ed. 2015. *Buddhist Dynamics in Premodern and Early Modern Southeast Asia*. Singapore: Institute of Southeast Asian Studies.

Munoz, Paul Michel. 2016. *Early Kingdoms: Indonesian Archipelago & the Malay Peninsula*. Singapore: Editions Didier Millet.

Ray, Himanshu P. 1994. *The Winds of Change: Buddhism and the Maritime Links of South Asia*. Oxford: Oxford University Press.

用數位工具挖掘 18 世紀德語歷史文獻

王濤*

摘　要

主題模型是新近開發出來的研究方法，對於拓展數位人文的研究路徑非常有價值。LDA 是主題模型演算法之一，將它運用到“德語文獻檔案”收錄的 1700-1800 年間的文獻，在歸納、分析文本的主題後，對主題模型方法的有效性進行評判。主題模型的演算結果讓我們對 18 世紀德意志精神世界有了更加立體的認知：18 世紀的作者具有強烈的歷史意識，對知識體系的構建異常積極，小說受追捧與公共領域的興起密切相關，宗教啟蒙是時代主題。這些結果表明，啟蒙運動具備多重面相。在歷史研究中需要將以主題模型為代表的遠距離閱讀與細讀有機結合起來，才能夠得到更具說服力的研究成果。主題模型作為一種文本挖掘的方法，仍然存在改進的空間，而這種進步需要人文學者與計算專家的通力合作。這也是數字人文繼續發展的必由之路。

關鍵字：數字史學、主題模型、德意志、啟蒙運動、遠距離閱讀

* 南京大學歷史學院副教授，Email: t.wang@nju.edu.cn。

Text Mining *Deutsche Textarchiv* Using Digital Tools

Tao Wang *

Abstract

Topic modeling is a newly invented tools to do text mining by digital humanities. LDA is one of the algorithm be used in the processing of *Deutsche Textarchiv* during 1700-1800. The results will be evaluated according to the historical context. The application of topic modeling in the research work of 18th century will shed new light on the world enlightenment. As we can see from the different topics, the authors in 18th century had deep Historical consciousness, novels were very popular reading materials and religious enlightenment is important phenomenon deserving close observation. The results turn out to emphasize the multifaceted of enlightenment. It also to be noticed that topic modeling need improvement to meet the need of historical research and distance reading can only be used as an auxiliary means.

Keywords: digital history, topic modeling, German, Enlightenment, distance reading

* Associate Professor, School of History, Nanjing University. Email: t.wang@nju.edu.cn.

數位史學（digital history）在西方學界方興未艾，國內學者近年來也開始涉足。除了必要的理論探討外，[1] 史料型資料庫建設是主要的成果呈現形態，而有歷史特質的個案研究基本上以量化歷史的面目出現，用資料庫方法梳理觀念史的研究以對關鍵字頻的統計為依據。[2] 數位史學當然不能止步於資料庫的建設，量化歷史或者詞頻統計的方法也不是數字史學的全貌。某種意義上說，歷史研究的史料除了容易量化的資料外，更多是無法量化的文本，因此對資料庫進行有效的資訊提取與視覺化呈現，才是數位史學的核心價值。先行一步的西方學者已經在使用主題模型（Topic Modeling）的方法對大規模文獻進行資料採擷，[3] 拓展了數位人文（Digital Humanities）的研究路徑，在史學研究領域，也有值得期待的可能性。本文將在關於德意志啟蒙運動的研究實踐中使用這種工具，並結合具體案例對其有效性進行評判。

一、主題模型的基本概念

近 700 份文獻，字元數在 3000 萬左右，要用什麼方法在最短時間內瞭解文獻的整體面貌，並對文獻內容進行整理？傳統方法是讓不同的人同時閱讀，做讀書筆記，然後分享閱讀成果，最終整合成一份讀書報告。這種合作閱讀（collaborative reading）的方

[1] 早年間已經有國內學者注意到了“數字史學”這個概念，從史學史的角度發佈了一些介紹性文章，參見王旭東：《數字世界史：有關前提、範式及適用性的思考》，《安徽大學學報》2006 年第 6 期，第 96-101 頁；周兵：《歷史學與新媒體：數字史學芻議》，《甘肅社會科學》2013 年第 5 期，第 63-67 頁；牟振宇：《數字歷史的興起：西方史學中的書寫新趨勢》，《史學理論研究》2015 年第 3 期，第 74-81 頁；以及王濤：《挑戰與機遇：數字史學與歷史研究》，《全球史評論》2015 年第 8 輯，第 184-201 頁。《史學月刊》在 2015 年第 1 期組織了“電腦技術與史學研究形態筆談”，2015 年 12 月 4-7 日上海大學主辦了“傳承與開啟：大資料時代下的歷史研究”主題研討會，呈現了中文語境中“數字人文”研究的最新進展。

[2] 具有代表性的資料庫包括“中國基本古籍庫”、“晚清民國期刊全文資料庫”等；量化歷史的研究成果參見王躍生：《民國時期婚姻行為研究》，《近代史研究》2006 年第 2 期，第 26-44 頁；梁晨、李中清：《無聲的革命：北京大學與蘇州大學學生社會來源研究》，《中國社會科學》2012 年第 1 期，第 98-118 頁；梁晨、李中清：《大資料、新史實與理論演進》，《清華大學學報》（哲學社會科學版）2014 年第 5 期，第 104-113 頁；梁晨、董浩、李中清：《量化資料庫與歷史研究》，《歷史研究》2015 年第 2 期，第 113-128 頁。值得一提的還包括陳志武主導的北京大學經濟學院量化歷史研究所。觀念史的研究參見金觀濤、劉青峰：《中國近現代觀念起源研究和資料庫方法》，《史學月刊》2005 年第 5 期，第 89-101 頁；金觀濤、劉青峰：《歷史的真實性：試論資料庫新方法在歷史研究的應用》，《清史研究》2008 年第 1 期，第 90-108 頁。

[3] 人文學科領域的研究成果包括 David Newman, Sharon Block,“Probabilistic topic decomposition of an eighteenth century American newspaper,” *Journal of the American Society for Information Science and Technology*, vol. 57, no. 6, 2006, pp. 753-767; Sharon Block, David Newman,“WHAT, WHERE, WHEN, AND SOMETIMES WHY: Data Mining Two Decades of Women's History Abstracts,” *Journal of Women's History*, vol. 23, no. 1, 2011, pp. 81-109; David Mimno,“Computational historiography: Data mining in a century of classics journals,” *Journal on Computing and Cultural Heritage*, vol. 5, no. 1, 2012, pp. 1-19.另有里士滿大學（University of Richmond）的尼爾森（Robert K. Nelson）對 1860-1865 年間出版的《每日快訊》（*Daily Dispatch*）的資料採擷，見 http://dsl.richmond.edu/dispatch/ pages/home

式，通常被用來處理龐雜的文獻資料。它能夠提升搜集資訊的效率，[4] 但也有明顯劣勢：它基於多人協作，處理資訊的標準因人而異，讓內容整合的客觀性大打折扣。

更重要的是，這種傳統的方式是一種直接的（direct reading）、近距離的（close reading）的閱讀，處理資訊的容量非常有限。正如克雷恩（Gregory Crane）在 2006 年提出的那樣，“你怎麼處理 100 萬冊的圖書？”[5] 在資訊爆炸的網路時代，更有大量有效資訊淹沒在無關文獻的海洋，人力的局限性在這裡暴露無餘。為此，文藝理論家莫萊蒂（Franco Moretti）曾經提出“遠距離閱讀”（distant reading）的概念，[6] 其初衷實則沿襲了合作閱讀的方式。專注機器學習與自然語言處理的專家，設計出“主題模型”的演算法，能夠在無須人工參與的前提下發現和歸納文本的主題內容。這種統計模型工具用機器閱讀的形式兌現了遠距離閱讀的理念，為解決文獻增量超出人類理解極限的狀況找到了出路。

主題模型的工作原理立足於人類的寫作習慣。寫作者在創作文本時，都會預設若干主題。為了凸顯某個主題，作者會在遣詞造句時調用具有相關聯的詞彙，在主題模型的術語中，這些具有相關性的詞彙被稱為“詞群”（bag of words）。舉個例子，歌德在構思《少年維特之煩惱》（*Die Leiden des jungen Werthers*）時，[7] 會設計不同主題，並用不同的文字展現出來。作為一部愛情小說，“愛情”（Liebe）一定是絕對的主題，但歌德也不會排斥對其他主題的描述，否則小說的可讀性降低，對社會的描述也會非常扁平化。因此“自然”（Natur），“藝術”（Kunst）以及“社會”（Gesellschaft）等，也是可能的主題內容。為了描繪這些主題，歌德在寫作中會調動相應的詞群，例如，在描繪維特令人心碎的愛情時，一定會出現高頻率地出現“Liebe”（愛情）、“Hertz”（心）等，也會有“umarmen”（擁抱）、“küssen”（吻）等，或者頻率較低的“ewig”（永恆）、“morgen”（明天）等詞彙。其他主題也有類似的詞群以及頻率。基於這樣的創作習慣，如果我們能夠統計詞群，就能把握與之對應的主題，進而瞭解整部文獻內容。

[4] 合作閱讀的方法在文學研究領域使用較多，相關研究包括，Larry Isaac, “Movements, Aesthetics, and Markets in Literary Change: Making the American Labor Problem Novel,” *American Sociological Review*, vol. 74, no. 6, 2009, pp. 938-65.

[5] Gregory Crane, “What do you do with a Million Books?” *D-Lib Magazine*, vol. 12, no. 3, 2006. 克雷恩是古典學教授，“珀耳修斯數碼圖書館”（Perseus Digital Library）的專案主持人。

[6] 莫萊蒂最初在 2000 年的一篇論文中提到了“遠距離閱讀”的概念，參見 Franco Moretti, “Conjectures on World Literature,” *New Left Review*, no. 1, 2000, p. 54-68.

[7] 本例參見 Matt Erlin, ed., *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*, New York: Camden House, 2014, p. 59.

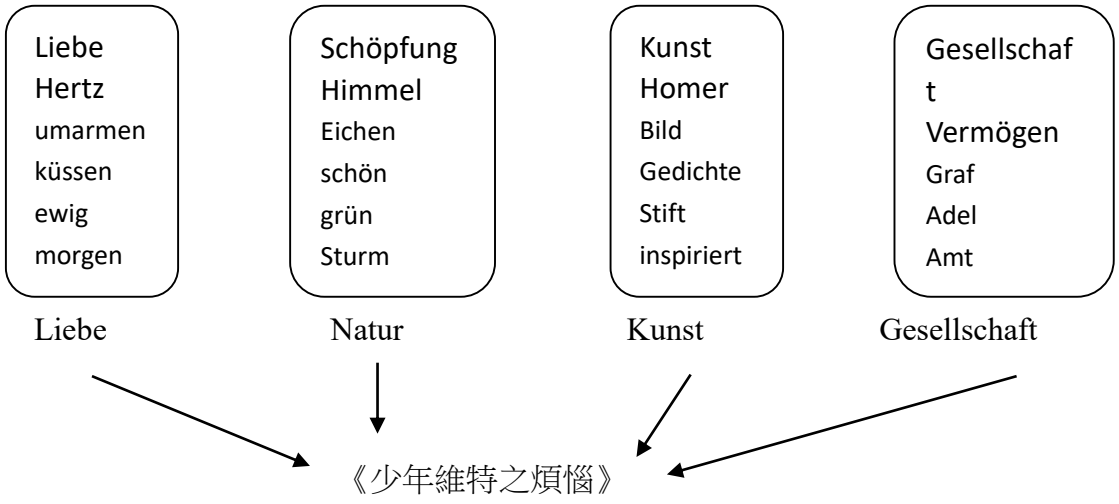


圖 1：歌德之維特的主題創作

在上述思路的指引下，佈雷（David Blei），吳恩達和喬丹（Michael Jordan）於 2003 年提出了"隱含狄利克雷分佈"（Latent Dirichlet allocation，簡稱 LDA），[8] 成為主題模型最常用的演算法。LDA 通過特定公式計算詞彙出現的頻率，並將相互關聯的詞彙作為結果輸出。這種模型是一種無監督學習的演算法，具有剛性的客觀性，即事先不需要研究者對文獻內容有任何瞭解，也不需要進行人工標注、設置關鍵字等主觀處理，而完全由電腦程式自動完成對文獻主題的歸納。主題模型試圖用數學框架來解釋文檔內容，這種做法看似同人文學科的習慣並不相容。但是 LDA 輸出的結果是一組有意義的詞群，而非純粹的統計資料，人文學者能夠使用這些詞彙進行定性分析，證實或者證偽一些猜測，[9] 將定量統計的客觀與定性描述的開放充分結合起來，所以此方法在人文學科領域極具應用的前景，特別是對動輒數以萬計的文獻，主題模型的計算能力非常誘人。[10]

基於 LDA 的理念，電腦專家邁克卡倫（Andrew McCallum）寫出軟體 MALLET，讓歸納整理文獻主題變成簡單的命令錄入，開始被人文學者廣泛使用；[11] 特別是在紐曼（David Newman）和同事用 JAVA 開發出圖像介面的主題模型工具套件（Topic Modeling Tools，TMT）之後，使用者甚至不需要瞭解繁瑣的命令符，進一步

[8] David Blei, Andrew Ng, Michael Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4–5, 2003, pp. 993–1022.
[9] David Blei, "Topic Modeling and Digital Humanities," *Journal of Digital Humanities*, vol. 2, no. 1, 2012, pp. 10-11.
[10] David Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, 2012, p. 77. 關於"主題模型"從概念、應用到工具的梳理，請參見《數位人文雜誌》在 2012 年的專刊，Daniel Cohen, ed., *Journal of Digital Humanities*, vol. 2, no. 1, 2012.
[11] 適合歷史學家瞭解 MALLET 的使用指南，參見 Shawn Graham , Scott Weingart and Ian Milligan, "Getting Started with Topic Modeling and MALLET," *Programming Historian* (02 September 2012), http://programminghistorian.org/lessons/topic-modeling-and-mallet.html。另外有 Ted Underwood, Scott Weingart, Miriam Posner 等學者關於主題模型的博文，亦可參考。

降低了應用門檻，讓主題模型成為人人能夠上手的工具。

二、“德語文獻檔案”簡介

主題模型的優勢是能夠對海量文獻進行高效率的分析。這裡涉及到兩個問題。

首先，“海量”是多少？Paper Machines是另一款可以進行主題模型分析的工具，其使用手冊上注明，成功進行主題模型的下限是50份文獻。[12]毫無疑問，過少的文獻，我們完全可以直接閱讀，獲取有效資訊的準確率一定高於機器識別。50份文獻也是一個略指，並沒有對每份文獻的具體字數進行說明：實際上，將文獻段落劃分為不同文檔，會影響主題模型輸出的結果（雖然可能僅僅是某些詞彙的改變）。

其次，什麼樣的文獻能夠進行主題模型分析？由於主題模型需要電腦對文字進行識別，所以需要把紙質文獻轉化為數字文檔，即要對文字資料的影像檔進行識別處理（即所謂光學符號識別，Optical Character Recognition，簡稱OCR）。但我們知道，OCR的錯誤率是無法回避的問題，特別是對歷史文獻而言，OCR的輸出結果總是差強人意。我們在本文使用的文獻集中在18世紀，都是用花體字（Fraktur）印刷，轉換出來的純文字更是錯誤頻出。對OCR文檔進行清理，必要時用規則運算式（regular expression）提高工作效率，也是我們進行主題模型分析的準備步驟。

實際上，這兩個問題都指向了文獻數位化的狀況。可以毫不誇張地說，文獻的數位化，是開展數位人文研究的前提。作為史學研究者，我們或許更能體會何謂巧婦難為無米之炊，史料就是我們研究的依據；沒有經過數碼化處理的史料，等同於史學研究無米下鍋。在這個意義上，建立史料的電子資料庫，是一項基礎設施建設。雖然它在客觀上加劇了文獻爆炸的事實，導致信息量太多以至於無法消化（too much to know），[13]但卻是“數字史學”研究展開的第一步。

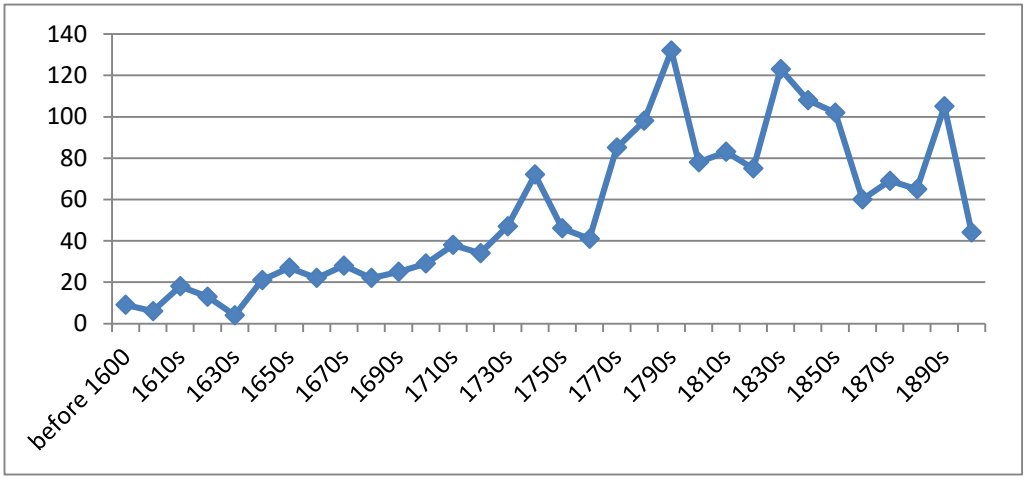
西方學界很早就意識到這點。本研究使用的數位文獻，就受益於數位化基礎設施建設的先期成果。主體文獻來自“德語文獻檔案”（Deutsche Textarchiv，簡稱DTA），是一個涵蓋了從15世紀到20世紀初跨度達500年的德語文獻資料庫，當前收錄的文獻近1800件，文獻類型包括書籍、報紙等，並在不斷擴充。[14]“德語文獻檔案”其實是歐盟範圍內CLARIN的一個子項目。CLARIN的全稱是“通用語言庫與技術基礎設施”（Common Language Resources and Technology Infrastructure），其宗旨是對人文社會科學領域的語言材料進行歸檔與數碼處理，實現資料共用，推進學術研究；各個歐盟成員國都有相應

[12] http://www.papermachines.org/wiki/page/Basic_Troubleshooting
[13] 這裡借用了哈佛大學歷史系安·布雷爾（Ann Blair）教授近著的標題，Ann Blair, *Too much to know: managing scholarly information before the modern age*, New Haven: Yale University Press, 2010.
[14] 資料庫的網址為 http://www.deutschestextarchiv.de

機構負責搭建各自語種的文獻資料庫，德國的成果之一就是DTA。[15]

本文集中分析“德語文獻檔案”收錄的1700-1800年間共計644件文獻，字元數總量近3000萬。這個時間段的劃分，是由“德語文獻檔案”資料庫的特性決定的。“德語文獻檔案”收錄的德語文獻有多個來源，[16] 其原則不是為了窮盡某個年份的文獻，而是要兼顧學科的全面與版本的首創。雖然資料庫收錄文獻跨度達500年之久，但從表1可以看出，文獻數量的年代差異非常明顯。1700年之前的文獻相對較少，1800年之後的文獻明顯增多。根據主題模型原理，過少或者過多的文獻都會左右結果的輸出，影響我們的分析；縱觀整個18世紀的文獻，既有康得、席勒、洪堡等重要歷史人物的作品，也有被歷史湮沒的小人物的文字，甚至匿名者，雖然收錄的文獻僅僅是這個時代所有文獻的很小一部分，但它們極具代表性，能讓我們比較全面地探尋時代面貌。另外，選擇相對較小的文本容量，主要是考慮到能與人工閱讀對照分析，方便我們對主題模型的有效性進行評判。

表1：文獻的年代分佈



600多份文獻達到了運用主題模型工具的標準。這些文獻的長短參差不齊，既有阿諾德（Gottfried Arnold）涉及教會史的大部頭，[17]單篇就有10萬字之巨；也有僅僅隻言片語的宣傳單。[18]需要指出的是，文獻的統計單位以其原始形態為依據，即一部書記為一份，多卷本的書每卷單獨計數，至於下文提到的報紙，以合訂的一期為一份。在我們的

[15] CLARIN 的網址為 https://www.clarin.eu/。關於 CLARIN 整體狀況，可以參見 Martin Wynne, “The Role of CLARIN in Digital Transformations in the Humanities,” *International Journal of Humanities and Arts Computing*, vol. 7, 2013, p. 89-104. 涉及德國專案的技術指標，工作流程，請參見 Christian Thomas, “Making great work even better. Appraisal and digital curation of widely dispersed electronic textual resources in CLARIN-D,” in Jost Gippert, ed., *Historical Corpora. Challenges and Perspectives*, Tübingen: Narr Verlag, 2015, pp. 181-196.

[16] http://www.deutschestextarchiv.de/doku/textquellen

[17] Arnold, Gottfried: *Unpartheyische Kirchen- und Ketzer-Historie*. Bd. 2 (T. 3/4). Frankfurt (Main), 1700. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/arnold_ ketzerhistorie02_1700.

[18] Wahrhaffter Abriß, Deß Wunder-Geschicht, so sich Anno 1702. den 29. Aprill in [...] Wienn [...] zugetragen. [s. l.], 1702. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/nn _abriss_1702.

分析中，每份文檔被視為最小的研究單位。對於內容龐雜的單個文獻之所以沒有按照章節繼續劃分，是因為進一步的切割會破壞專著的語義完整，在返回原文進行細讀分析時發生錯位。

三、主題模型的運用與分析

LDA主題模型是十多年前提出的概念，其間不斷有新工具被開發出來。本文主要採用具有圖像介面的MALLET。實際上，現在被人文學者用來進行主題模型分析的Paper Machines及Tethne等工具，都內置MALLET的內核，其後臺的演算法基本相同。

主題模型的工作原理雖然不要求使用者事先對文獻內容進行瞭解，但為了讓輸出的結果為人類理解，並被用作進行定性分析的材料，需要設置一些參數。其中一個重要參數是想讓機器演算出多少主題，並用多少關鍵字進行表達。[19]考慮到“德語文獻檔案”文獻類型的多樣性，以及文獻大小的巨大差異，我們將主題的數量確定為40個，每個主題用20個詞進行表達。將全部644份文檔導入程式之後，我們得到了一個完整列表。本質上說，擬合出來的主題復原了18世紀德意志的歷史畫面與精神世界，涉及了廣闊的內容。它是沒有任何人為因素參與的場景重建，呈現出來的形態令人矚目。[20]

(一) “德語文獻檔案”的整體狀況

從某種意義上說，主題模型就是將總量達幾千萬字元的文獻，濃縮到用800個主題詞去理解。仔細觀察全部40個主題的詞群，我們發現有一些詞彙在不同主題頻繁存在。這或許是一個從整體上理解“德語文獻檔案”的指標。我們可以用文字雲的工具，統計主題詞的頻率，得到視覺化的結果（圖2）。

圖2：主題詞的文字雲

[19] https://code.google.com/p/topic-modeling-tool/wiki/TopicModelingTool

[20] 由於篇幅原因，我們無法展現全部 40 個主題的詞群結果。在後文的分析中，我們會重點介紹一些主題詞群。感興趣的讀者可以與作者聯繫索要原始資料。

文字雲透露了一些資訊。在全部主題中，諸如Menschen（人）、Wasser（水）、Art（藝術）、Lieb（愛）等詞彙高頻率出現。我們可能會認為，這些關鍵字大概反映了18世紀的某種時代風貌，即對自然與人文的關注。這個判斷也與既有的研究成果不謀而合。許多傳統研究者提出，德意志文化存在"自然崇拜"的主題，自然景觀被賦予了崇高的意味，而其中流露出宗教虔敬的特質則為早期浪漫主義的出現提供了養分。[21] 當然，僅憑幾個關鍵字就引申出整個18世紀的時代精神，這種推斷在邏輯上可以存疑；另外，這些詞本來就是德語中常用的詞彙，如果沒有上下文的語境，它們並不能提供更多的所指。對於文本分析而言，關鍵字的文字雲功能有限，正如有學者強調的那樣，在人文學科的研究中，文字雲只不過提供了漂亮的裝飾而已，[22]對於研究者開展有營養的文本分析遠遠不夠。為此，我們需要對各種主題進行更加精細的解讀。

首先，我們可對這40個主題再進行分類。通過梳理不同主題，發現有些主題雖由不同詞群構成，但在講訴具有相關性故事。照此思路，將40個主題劃分成12個大類：

表2：主題的內容標籤

類型	主題	類型	主題
經濟類	1, 8, 14, 19, 25	哲學類	7, 30, 39
歷史類	2, 9, 12, 15, 17, 24	政治類	11, 28
家　庭	3，27	宗教類	20, 23, 26, 29, 33, 34
自然科學	4, 13, 16, 18, 21, 22, 40	法律類	38
情　感	5, 10, 37	旅　行	31, 36
醫學類	6, 35	技　術	32

這個主題標籤要比文字雲更能說明問題，至少有兩個非常明顯的特徵。

首先，自然科學與宗教類兩個看似對立的領域都表現得異常活躍。這讓我們意識到，啟蒙時代是一個科學與宗教並存的時代。我們會在後面繼續討論這個話題。

其次，在12大標籤中，除了家庭與情感類有較強的感性色彩之外，其他10大標籤都偏重理性的知識體系。實際上，如果我們深入挖掘的話，與"家庭"相關的主題，有相當一部分文獻涉及與園藝、烹飪、衛生等生活常識相關的內容，那麼整個"德語文獻檔案"所具有的"百科全書式"的氣質就更加明顯了。換句話說，這份文獻是一部格調極高的書

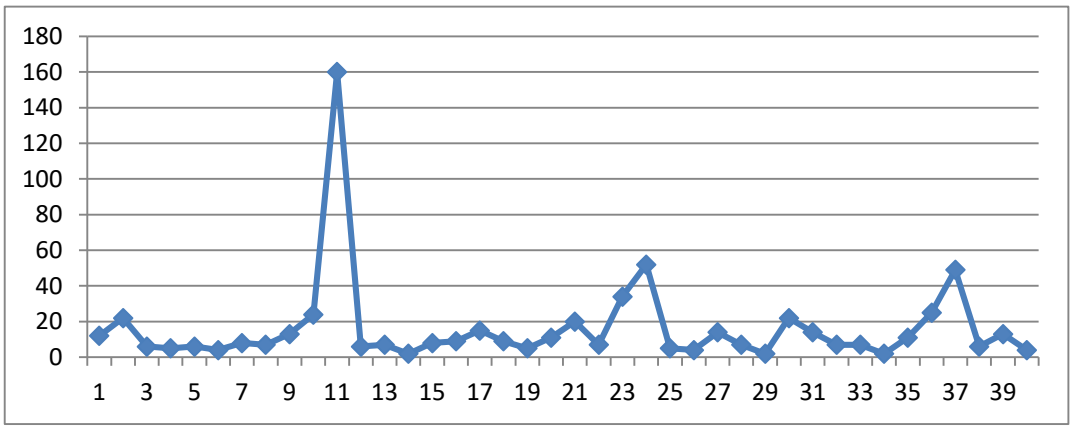
[21] 康得對"優美感和崇高感"的論述，無疑是"自然崇拜"的寫照；活躍於 18 世紀晚期的瓦肯羅德（Wilhelm Heinrich Wackenroder）是早期浪漫主義的奠基人之一，他的代表作《一個熱愛藝術的修士的內心傾述》亦收錄於"德語文獻檔案"，見 Wackenroder, Wilhelm Heinrich; Tieck, Ludwig: Herzensergießungen eines kunstliebenden Klosterbruders. Berlin, 1797. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/ wackenroder_herzensergiessungen_1797.

[22] 參見加州大學馬傑維斯基（Majewski）對運用數字史學研究美國鐵路的書評，見 John Majewski, "Review of The Iron Way," *The Journal of Southern History*, vol. LXXIX, No. 3, 2013, p. 714.

單，它從另外一個維度證實了啟蒙學者建構“知識樹”的努力，[23]在整個18世紀，不僅有知識的提供者，也有知識的消費者，“啟蒙運動的生意”因此得以如火如荼地進行。表2讓我們直觀地看到，法國的狄德羅《百科全書》式的生意經，只是整個18世紀知識經濟的冰山一角，我們在這裡所分析的近700份文獻，也僅僅展示了“圖書產業”（book industry）的一個側面。

根據LDA的邏輯，一個文檔可以包含若干主題；反過來，同一個主題可以在不同文檔中呈現。因此，我們通過統計某個主題對應文獻的數量，就能夠瞭解不同主題在整個“德語文獻檔案”中的強度。

表3：主題在文獻中的分佈



結果出人意料。主題11一枝獨秀，它包含的詞群為：

11. koenig stadt herr general herzog koenigl armee kaiser schweden sten majestaet fuersten grafen graf reichs kam koenigs erhalten hof neue（國王 城市 統治者 將軍 大公 軍隊 皇帝 瑞典 選侯 伯爵 朝廷 等）

雖然存在一些干擾詞彙，比如Koenig，Koenigl，Koenigs應該被視為同一個詞（它們應該是OCR不完善帶來的缺陷），同理還有Grafen與Graf；neue作為形容詞，不應該有太多所指；sten或為德語序數詞尾碼的誤判，但是我們仍然能夠確定主題11跟政治與戰爭相關。當然，我們無法確定它同歷史相關還是更多與時事相關，這需要我們返回對主題11的形成做出貢獻的文本。LDA使用百分比來描述文檔與某個主題的關聯度，我們調查發現，有將近130個文檔與主題11的相關度超過40%。這一批文獻的內容或許很能說明一些問題。為此，我們需要對它們進行細讀，大致梳理其內容。

這部分文獻中，有涉及時政的報紙，也有歷史體裁的書籍。在貢獻度最高（50%以上）的幾個文檔中，全部來自報紙。“德語文獻檔案”所收錄報紙的來源比較單一，主要

[23] 羅伯特·達恩頓：《屠貓記》，呂建忠譯，北京：新星出版社，2006 年，第 202-203 頁。

是《漢堡通訊》（*Staats- und Gelehrte Zeitung des Hamburgischen unpartheyischen Correspondenten*），它是1712-1851年間在漢堡出版的第一份跨區域的報紙，具有廣泛的讀者群。[24]從這個角度看，主題11涉及更多時政性內容。儘管目前"德語文獻檔案"收錄《漢堡通訊》的年份有限，1771、1789以及1790這三個年份還是在主題11的表述中凸顯出來。1789年的特殊性不言而喻，7月13日的報導，就涉及到了巴黎風起雲湧的局勢。[25]從隨後《漢堡通訊》的跟進報導可以看出，德意志的讀者對正在法國上演的重大事件高度關注，並持續到1790年，這些材料能夠成為分析法國大革命對德意志時局影響的切入點。[26]1771年的報紙只有7月2日到8月7日一個多月的時間。這個時段最重要的事件是仍在進行的第五次俄土戰爭，《漢堡通訊》也有跟蹤報導。[27]從《漢堡通訊》追蹤熱點問題的各種嘗試可看出，德意志人有強烈的全域觀念，視野早已超越本土。這並不令人意外，因為報紙的作者名錄，包括萊辛、赫爾德、利希滕貝格（Lichtenberg）等重要啟蒙學者。[28]從這個意義上說，德意志啟蒙學者宣佈做"世界公民"的理念，並沒有停留在泛泛而談的層面，還試圖在普通民眾的日常閱讀中進行推廣。

惟一一個與主題11高度關聯（關聯度46%）的非報紙文獻是席勒的《三十年戰爭》，[29]它為我們提供了該主題的歷史維度。這是一個並不令我們感到意外的文獻，三十年戰爭本來就是政治史上的大事件，大量出現Armee（軍隊）、Koenig（國王）、Hof（朝廷）等詞彙是必然的事情。Schweden（瑞典）的出現也並不意外，因為瑞典是三十年戰爭的重要參與者；然而，當我們注意到《漢堡通訊》中也有涉及瑞典的報導時，[30] 就無法區分被MALLET挑選出來的這個"瑞典"是歷史的指向，還是時政的指向。這或許是主題模型作為一種演算法的缺陷。

我們通過對主題11的深度分析，得到如下幾個結論：

首先，主題模型對文獻有較好的歸納能力，它能夠將報紙這種文獻類型劃歸到一個

[24] Susanne Haaf and Matthias Schulz, "Historical Newspapers & Journals for the DTA," in *LRT4HDA*, 26 - 30 May 2014, Reykjawik, Iceland.
[25] *Staats- und Gelehrte Zeitung des Hamburgischen unpartheyischen Correspondenten*. Nr. 115, Hamburg, 21. Juli 1789. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/hc_1152107 _1789/1.
[26] Rolf Reichardt,"Deutsche Volksbewegungen im Zeichen des Pariser Bastillesturms. Ein Beitrag zum sozio-kulturellen Transfer der Französischen Revolution," *Geschichte und Gesellschaft*. Sonderheft, vol. 12, 1988, S. 10-27. 另外可以參見王濤：《入侵與解放背景下的革命：美因茨共和國的歷史解讀》，《世界歷史》2015 年第 4 期，第 47-58 頁。
[27] *Staats- und Gelehrte Zeitung Des Hamburgischen unpartheyischen Correspondenten*. Nr. 105, Hamburg, 2. Julii 1771. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/hc_1050207 _1771.
[28] Holger Böning, "Hamburgischer Correspondent: Journal der Epoche, " *Zeit Online*, 6. Juni 2012.
[29] Schiller, Friedrich: Geschichte des dreyßigjährigen Kriegs. Frankfurt u. a., 1792. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/schiller_krieg_1792.
[30] 比如在 1790 年的報導中，Staats- und Gelehrte Zeitung des Hamburgischen unpartheyischen Correspondenten. Nr. 67, Hamburg, 27. April 1790. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/hc_ 672704_1790.

主題之下，說明LDA的演算法對報紙內容的挖掘具備合理性。這也提醒我們，有必要對文獻類型進行劃分，分別開展主題模型的梳理，或許能夠得到更加精細的結果。我們將在第三部分繼續這個話題。

其次，組成主題11的詞彙以描述社會等級地位的術語為主，直觀地描繪出一種歷史畫面：在啟蒙時代，等級觀念仍然是社會生活的主流。這是一個非常合理的判斷。在《漢堡通訊》的大量時政報導中，有許多與皇親國戚活動相關的報導，折射出社會上層的活躍度，如何解釋這種現象？我們認為，這恰恰是啟蒙運動在神聖羅馬帝國展開的一種方式。德意志的啟蒙作為後起之秀，其迅猛的發展要得益于君王與貴族的支持；換句話說，"自上而下"的傳統是德意志文化的特質，在推廣啟蒙理念的事務上同樣如此。開明專制被用來描述這個時代德意志的政治結構，而其中的代表者正是普魯士國王腓特烈（Frederick the Great，1712-86），他對啟蒙思想家的資助有目共睹，以至於自視啟蒙運動的領軍人物。他曾經大言不慚地對伏爾泰說："我的主要職責是同傲慢與偏見鬥爭……啟蒙心智，扶植道義，讓民眾追隨天性獲得幸福。"[31]其他的統治者，比如巴登的大公弗裡德里希（Karl Friedrich），在其長達73年的政治生涯中，也是啟蒙運動的重要資助者。主題11將表達社會等級的詞彙凸顯出來，佐證腓特烈的自誇並非空穴來風，說明主題模型的結果投射了18世紀德意志的社會現實。

另外幾個被大量談及的主題包括主題24與37。我們先來看看它們各自的詞群：

24. immer ganz zeit ganzen unsre nie nichts wenigstens vielleicht recht geschichte einmal macht lange gesellschaft endlich genug buch lassen wenig（總是 全部 時間 從不 至少 右的 歷史 社會 終於 書籍 至少 等）

37. herr vater liebe mutter nichts frau hand kam einmal immer recht herz ganz gut tage fort lieber machte kind liess（統治者 父親 愛 母親 虛無 妻子 手 永恆 心 善 日子 小孩 閱讀 等）

從類型上看，主題37涉及的內容非常明確，它與人類複雜的感情相關，而主題24的情況要複雜一些，我們先行討論主題37，稍後對主題24再做解析。

主題37的詞群讓我們聯想到了家庭、感情。我們返回查看文獻，證實了這個假設。對該主題做出貢獻的48份文獻中，絕大部分可以歸在"小說"的文獻門類之下，其中就包括了歌德的名著《少年維特之煩惱》，以及據稱第一位德意志女性作家拉洛施(Sophie von La Roche)的傷感文學代表作《施特恩海姆小姐的故事》（*Geschichte des Fräuleins von*

[31] Giles MacDonogh, *Frederick the Great: A Life in Deed and Letters*, New York: St. Martin's Press, 2000, p. 341.

Sternheim）。[32]

小說貢獻出這個極具感性的主題，可以從幾個層面解讀。首先，它與學者們已經觀察到的“閱讀社會”（Lesegeschelschaft）的興起直接相關。德意志民眾的閱讀習慣在18世紀開始出現重大轉變，恩格辛（Rolf Engelsing）用“閱讀革命”（Leserevolution）進行概括，經歷了從“精讀”向“泛讀”的轉換。[33]雖然恩格辛的結論有失偏頗，但它確實說明關乎人性體驗的文學體裁大受歡迎。實際上，許多圖書館的館藏記錄就是明證：德語寫作的書籍遠遠超過拉丁語書籍，小說是書單上的絕對主力。[34]在真實的閱讀實踐中，精讀與泛讀往往結合在一起，許多讀者會對打動內心的小說反復研讀，“維特熱”所帶來的社會問題，或許是一個極端案例，但它充分顯示了文學讀物受民眾追捧的程度。

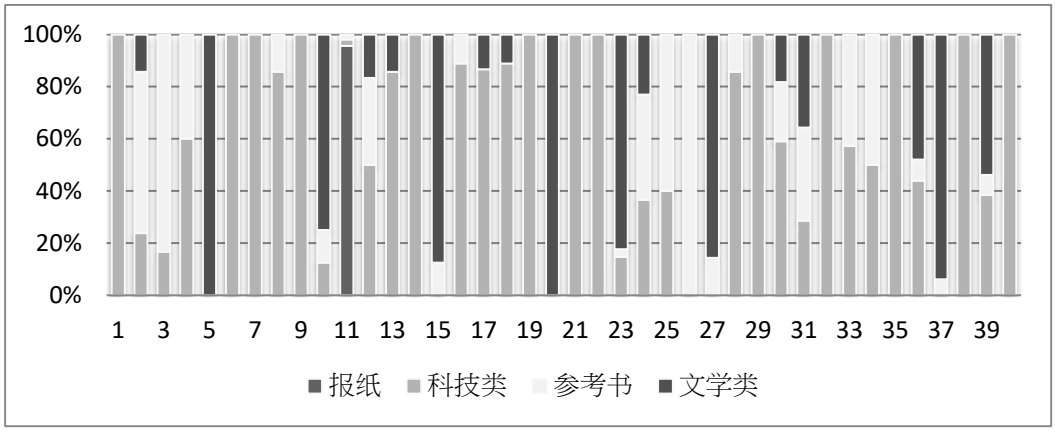
其次，這也跟整個啟蒙時代的策略相關。1791年一位弗萊堡的觀察者曾經總結到：“一般而言，存在學院的啟蒙以及民眾的啟蒙兩種類型。前者是後者的引路人，它高舉火炬。一種理念可能首先在大學的講堂經過二十甚至三十多年的闡釋，才能夠被大眾接納，並得到推廣。”[35] 18世紀的德意志世界存在“閱讀熱”（Lesewut）的社會現象，當時的人們形成了一種共識，試圖通過閱讀來修煉自己的啟蒙氣質，獲得提升自身素質的力量；許多小說作者也會在曲折的故事情節中夾帶私貨，從而讓文學讀物具備了傳播科學知識的功能，[36] 成為推廣啟蒙價值理念的工具：在小說的創作與消費之間，18世紀的“公共領域”以一種媒介傳播的方式得到建構。

主題11與主題37存在的這種狀況，讓我們意識到有必要對“德語文獻檔案”進行類型劃分。我們將全部644份文獻歸納為四類：報紙、文學類、科技類以及參考書。這是非常粗線條的劃分，實則在一個類型下，還可以有更多細類，比如科技類文獻其實涵蓋了人文與自然科學，歷史、政治學與物理學、生物都被囊括到同一類型下。但這種大類的劃分也具有合理性。我們通過檢視不同文獻類型在各種主題上的表現，發現文獻類型與主題的對應度非常明顯（表4）。最極端的例子是全部報紙只同主題11掛鉤，而上文分析過的主題37基本由文學類文獻構成，此類型還貢獻了主題10、20、23以及27。參考書由於總體數量較少，主題呈現度偏低，但仍然同主題2、3與26有強烈的依存度。科技類文獻相對最多，貢獻了更多獨立的主題，包括1、9、17、19、21、22、32、35、38以及40。

---

[32] [La Roche, Sophie von]: Geschichte des Fräuleins von Sternheim. Hrsg. v. Christoph Martin Wieland. Leipzig, 1771. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/laroche_ geschichte01_1771.
[33] 恩格辛將節點設置在 1750 年，精讀是對少量書籍的反復閱讀，而泛讀是對大量書籍的快速閱讀，參見 Rolf Engelsing,“Die Perioden der Lesergeschichte in der Neuzeit. Das statische Ausmass und die soziokulturelle Bedeutung der Lektüre,” *Archiv für Geschichte des Buchwesens*, vol. 10, 1969, S. 977–983.
[34] Dorinda Outram, *Panorama of the Enlightenment*, London: Thames & Hudson, 2006, p. 69.
[35] Notker Hammerstein, *Aufklärung und katholisches Reich*, Berlin: Duncker & Humblot, 1977, S. 12.
[36] Alan Kors, ed., *Anticipations of the Enlightenment in England, France, and Germany*, Philadelphia: University of Pennsylvania Press, 1987, pp. 171-177.

表4：文獻類型與主題的對應



上文提到主題24的詞群比較模糊，其中包含了諸如Geschichte（歷史）、Gesellschaft（社會）等術語，更多的則是與時間相關的形容詞與副詞。我們從直覺上判斷，它或許貼近歷史和社會問題。經過梳理發現，為該主題做出貢獻的文獻總計52件，我們流覽了它們的內容，確實存在歷史體裁的文獻，比如席勒的《什麼是普世歷史及其學習的意義》（*Was heißt und zu welchem Ende studiert man Universalgeschichte*?），以及莫澤爾（Justus Möser）的《奧斯納布呂克的歷史》（*Osnabrückische Geschichte*），但也有歷史社會題材的小說。[37] 這意味著，主題24所呈現的跨界傾向非常明顯，即兩種以上的文獻類型都有出現，類似的主題還包括30、31、36、39。仔細考察這些主題的內容，大多數跟哲學、地理等相關，而這些方向本來就具有跨學科的特質。

對於主題24這種情況，我們還可以將它與其他幾個主題結合起來考察。從表2的歸納可以看到，主題2，9，12，15，17與主題24同屬一種類型。這一個大類都與歷史問題相關。換句話說，啟蒙時代對歷史問題的關注度極高，不論在專業研究領域，還是在文學創作中，寫作者都具有強烈的歷史意識，民眾也對歷史懷有極大興趣，這是18世紀德意志的一大特色，以至於"歷史主義的興起"也需要從啟蒙運動那裡找源頭。[38]

(二) 主題的演變趨勢

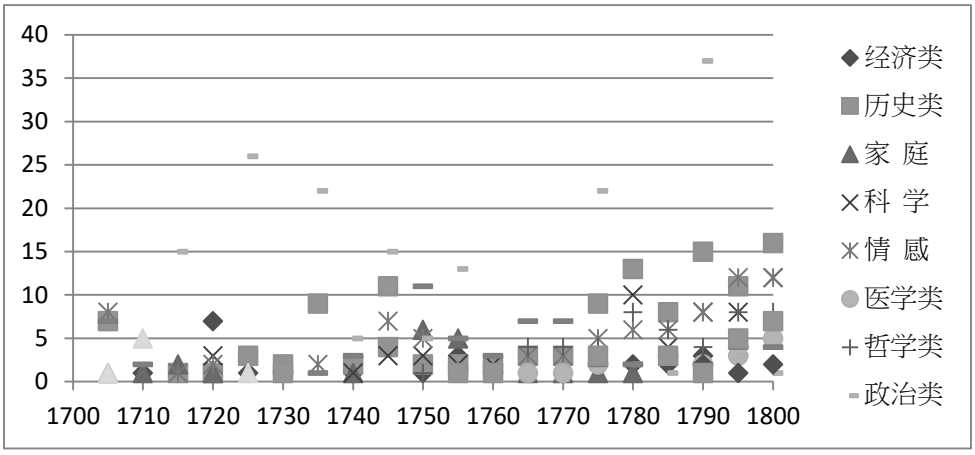
在進行主題模型分析之初，我們將644份文獻按照年代排列，並且在MALLET的演算法中，加入了保持排序的參數，從而讓主題模型能夠反映時間的變遷狀態。這對於我

[37] 例如 Hippel, Theodor Gottlieb von: Lebensläufe nach Aufsteigender Linie. Berlin, 1778. in: Deutsches Textarchiv <http://www.deutschestextarchiv.de/hippel_lebenslaeufe01_1778>

[38] Rudolf Vierhaus,"Geshichtsschreibung als Literatur im 18. Jahrhundert," in Karl Hammer, Hrsg., *Historische Forschung im 18. Jahrhundert: Organisation, Zielsetzung und Ergebnisse*, Bonn: Röhrscheid, 1976, S. 416-31. 德國歷史學家梅尼克的代表作《歷史主義的興起》是對這個問題最好的總結，參見弗裡德里希·梅尼克：《歷史主義的興起》，陸月宏譯，南京：譯林出版社，2010 年。

們分析趨勢演變十分有利。

表5：主題類型的縱向變化



我們發現，大部分主題類型都保持著年度的穩定性，但有幾點值得注意。首先，政治類題材由於時政性，它在增量上的節點，往往能夠與重大歷史事件的節點找到對應關係。對法國的即時關注最為明顯。其次，情感類、歷史類與宗教類在整個18世紀都是非常重要的主題類型，但歷史類主題在1780年代後期有一個明顯的增加趨勢。最後，科學類主題似乎在世紀末才有增長過程，可能的解釋在於，“德意志文獻檔案”所收錄的科學類文章，以合集為主，往往都在18世紀的下半葉出版。

(三)　類型細分下的主題模型

如前所述，“德語文獻檔案”的文獻存在四種類型，如果我們對每個類型進行主題模型的分析，會有特別的發現嗎？由於文獻類型劃分後文本容量減少，我們將主題數量設置為25個，構成主題的詞群仍為20。一些隱匿的資訊在匯出的結果中確實被揭示出來。

報紙的主題詞再次印證了這種媒介的特殊性。在它的詞群中，大量出現與時間相關的詞彙，並有許多地名，包括倫敦（London），法國（Frankreich），斯德哥爾摩（Stockholm），君士坦丁堡（Constantinople）等。這些資訊透露，《漢堡通訊》的時政意味濃厚，而且胸懷天下，沒有把眼光局限在漢堡或者德意志一隅。

參考書類文獻體現出來的主題也極具特色，有一些與生活相關的知識，涉及烹飪（主題3、21），園藝（主題12），狩獵（主題15），藝術（主題5、19），以及旅行（主題18），基督教也是一個重要內容（主題4、7、17）。參考書類型的文獻本來就是實用常識的彙編，在德意志的語境中，類似現在的生活指南，這個背景跟18世紀興起的“民眾啟蒙”（Volksaufklärung）緊密相關。尤其是“德語文獻檔案”收錄了一些可以被標識為“家政文

學”（Hausväterliteratur）的文獻，[39] 是民眾啟蒙的重要讀物。民眾啟蒙被視為德意志啟蒙的獨特面，從主題模型挖掘出的詞群可看到，這個運動很好地符合了它所追求的方向。

在文學類文獻中，毫無意外地，我們發現了大量與人、愛情、生活、美相關的詞彙。最為突出的是，這個文獻類型下出現了與死亡的主題（3、14、25），這在其他類型中並不存在，甚至Tod（死亡）這個詞都沒有出現在詞群中。生存還是死亡，這是一個問題。莎士比亞讓這個難題成為了文學創作的永恆主題，在18世紀德意志的文學作品中也不例外。當然，我們如果考察與Tod成對出現的詞群，也能夠將主題賦予基督教的含義。

科技類文獻彙集了人文與自然學科，主題模型的演算法在某些主題上對它們進行了區分，例如主題11是純粹的語言藝術類，主題24講述了政治問題，主題3是生物學內容，而主題23與醫學相關。

比較令人意外的是，在科技類文獻中，基督教仍然具有存在感。這是由於我們將人文與自然科學文獻混為一談的結果嗎？為此，我們將歷史、神學、經濟學、政治學等學科類型剔除，把純理科的文獻單列出來，再次進行了主題模型的分析。我們仍然得到了如下的詞群：

8. himmel leben welt liebe menschen geister mensch erde engel hoelle geist dinge guten wissen gott kommt wort sehen glauben geistlichen（天堂 生活 俗世 愛 人類 修士 世界 天使 地獄 精神 善 智慧 上帝 等）

顯然，它仍然可以被歸類于基督教主題。從文獻來源上追溯，這個主題指向了斯威登堡（Emanuel Swedenborg）的選集。斯威登堡是瑞典科學家，主要從事自然科學與工程學研究。他在年輕時遊學歐洲，還曾就教於牛頓，但中年以後突然開始對神秘主義發生興趣，最終在神學方面取得極大成就，其代表作是關於來世、天堂以及地獄的研究。斯威登堡習慣用拉丁語寫作，“德語文獻檔案”收錄的是翻譯成德語的選集，[40] 由於斯威登堡的科學家身份，把他的選集納入神學，或者自然科學分類，似乎都說得過去，當然這會影響主題模型的結果輸出。當我們把斯威登堡的作品刪除之後，像主題8那樣明顯具有宗教意味的主題確實消失了，但諸如上帝（Gott）、火（Feuer）等能讓人聯想到基督教的詞彙仍然閃現。

斯威登堡的跨界身份給我們的分析帶來了一點麻煩，但這段小插曲卻恰恰體現了18世紀的獨特性。斯威登堡在1741年出現的精神危急雖然極具個人色彩，但嚴肅學者對宗

[39] Holger Böning, Reinhart Siegert, Hrsg., *Volksaufklärung: eine praktische Reformbewegung des 18. und 19. Jahrhunderts*, Bremen: Edition Lumière, 2007, S. 92-93.
[40] Swedenborg, Emanuel: Auserlesene Schriften. Bd. 2. Frankfurt (Main), 1776. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/swedenborg_schriften02_1776.

教信仰的熱衷在18世紀德意志並非個例。結合前面各種文獻類型中都能夠宗教主題掛鉤的事實，我們可以判斷，所謂18世紀的理性時代，其實也是一個無法回避基督教的時代。一個可能的解釋是，啟蒙時代是多維度的存在，各種文體都在談論基督教，恰好說明宗教問題的普遍性。批判啟示與信仰，批評基督教的文獻當然存在，但它們往往是遵循思維邏輯的討論，而不是非理性的斥責；實際上，存在批判基督教言論的同時，也有大量教導人們如何做一名更合格基督徒的讀物。[41]從這個意義上說，啟蒙時代對基督教的態度要比我們的想像複雜得多。嚴格地講，啟蒙運動具備多重面相，宗教的啟蒙也是時代主題之一，[42]這種概括應該會拓寬我們對18世紀的認知。

通過對純理科文獻的分析，我們發現了一個有趣的內容。在主題10中，多次出現"日本"（Japan）這個關鍵字。這個主題的詞群揭示了同政治和地理的相關性。通過查閱文獻，我們找到對這個主題做出貢獻的文檔恰好是兩部關於日本地理的科技作品。[43] 換句話說，主題模型的演算法精准地從近100份文獻中找到了一個特別的內容，體現出這種分析工具的高效率與準確度。

四、對主題模型的反思

我們在研究中驗證了主題模型的有效性，並從"德語文獻檔案"中發現了一些不被學者重視的歷史現象。雖然這些內容算不上顛覆性的成果，但是與傳統研究相比，主題模型在效率上有傳統路徑無法比擬的優勢。不要忘記，我們對18世紀德意志的認知，經歷了好幾代學者的研究積累，而主題閱讀的工具在沒有任何人工干預和先入為主的前提下，在短期內完成了對啟蒙時代的畫像，這本身就是一個成就。當然，我們在本文僅僅對幾個具有代表性的主題模型結果進行了分析，還有很多內容值得深入挖掘；其次，我們處理的德語文獻並非新史料，想要獲得全新的發現比較有難度。可以肯定的是，如果我們的研究物件足夠合理，主題模型不失一款有效的歷史研究工具。

不過，謹慎的學者會傾向於認為，主題模型"產生的問題和帶來的啟示或許一樣多"。[44] 毫無疑問，主題模型輸出的結果如果不經過學者的解釋，就是一堆片語而已；而要想

[41] 這是參考書類文獻的一個獨特類型，比如 Modestinus, Theophilus: Freymüthige Doch Bescheidene Unterredungen Von Kirchen- Religions- Politischen- und Natur-Sachen. Frankfurt (Main) u. a., 1737. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/modestinus_unterredungen_1737.
[42] David Sorkin, *The Religious Enlightenment*, New Jersey: Princeton University Press, 2008, pp. 3-5
[43] 兩部書對日本的歷史地理狀況，政治結構，宗教體系進行了細緻描述，見 Kaempfer, Engelbert: Geschichte und Beschreibung von Japan. Hrsg. v. Christian Wilhelm von Dohm. Bd. 1. Lemgo, 1777. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/kaempfer_japan01_1777；Thunberg, Carl Peter: Reisen durch einen Theil von Europa, Afrika und Asien [...] in den Jahren 1770 bis 1779. Bd. 1. Übers. v. Christian Heinrich Groskurd. Berlin, 1792. in: Deutsches Textarchiv, http://www.deutschestextarchiv.de/thunberg_reisen01_1792.
[44] Benjamin Schmidt,"Words alone: dismantling topic models in the Humanities," *Journal of Digital Humanities*, vol. 2, no. 1, 2012, p. 50.

讓分析符合歷史學科的規律，仍然需要研究者對文獻形成的歷史背景、所處的社會環境等有一定的把握。所以，在數字史學研究的名義下展開的合作，歷史學家永遠在場：精於機器學習的計算專家提供智慧化的工具，歷史學家貢獻專業化的分析。從上文的論述可以看出，LDA的演算法勝在對大資料的歸納能力，以及挖掘隱含資訊的效率。如果在數字人文研究中將遠距離閱讀與細讀有機結合起來，而不是相互對立，[45] 應該能夠得到更具說服力的研究成果。

另外，主題模型僅僅是一種研究工具，我們對"主題模型"的應用前景審慎樂觀。LDA對大資料的解析能力令人鼓舞，研究者會傾向於研究宏大主題，使用動輒上萬的文獻，這在方法論上固然沒有什麼問題，但無法結合細讀的結果輸出，其合理性是值得懷疑的，甚至是危險的。[46] 換句話說，用主題模型的演算法獲取詞群僅僅是研究開始的第一步，要想透過有限的主題詞挖掘合理解釋，歷史學家的定性分析功力不可埋沒。同時，LDA的演算法還會拋棄那些由於樣本過少而被程式視為無法構成主題、但對歷史研究可能仍然具有意義的內容。這種省略是否合理，在不同的結果輸出中如何取捨，類似的問題都需要結合具體文獻、具體的研究項目進行討論。

有一點可以肯定，史學研究中出現更多數碼工具的介入，將是不可避免的趨勢。毫不意外地，2015年8月在中國濟南召開的第22屆國際歷史科學大會，專門設置了數位史學的討論單元，"歷史學的數位化轉向"乃是大勢所趨。[47] 它將在宏觀層面影響歷史學的整體面貌，在微觀層面改變個體史學研究者的工作方式。當然，主題模型作為一種文本挖掘的方法，仍然存在改進的空間，而這種進步需要人文學者與計算專家更緊密的通力合作。這也是數字人文繼續發展的必由之路。

致謝

本文寫作得到了哈佛大學CBDB中心王宏甦、徐力恒兩位學友的幫助。南京大學歷史學院舒小昀對本文提出了建設性的修改意見。

[45] David Mimno,"Computational historiography: Data mining in a century of classics journals," p. 18.
[46] Maurizio Ascari,"The Dangers of Distant Reading: Reassessing Moretti's Approach to Literary Genres," *Genre*, vol. 47, no. 1, 2014, pp. 1-19.
[47] 王育濟編：《中國歷史評論》2016 年第 11 輯，上海：上海文化出版社，2016 年，第 152-176 頁。另見瑪麗亞塔·希耶塔拉：《歷史學的數位化轉向》，《世界歷史》2016 年第 1 期，第 29-32 頁。

Dissertations from Uppsala University 1602-1855 at the Internet

Anna Fredriksson*

Abstract

At Uppsala University Library a long term project is ongoing, which aims at making the dissertations, that is theses, submitted at the University of Uppsala 1602-1855 easy to find and read via the Internet. Our vision is to make it possible in the future to search through the entire text mass of Uppsala dissertations and theses, from 1602 till today, in one search session.

The work includes metadata production, scanning and OCR processing as well as publication of images of the dissertations in full text searchable pdf files. So far approximately 2300 dissertations have been digitized and made accessible at the Internet via the DiVA portal, Uppsala University's repository for research publications. In this first stage of digital publication the emphasis is on the period 1778-1855. All in all there are about 14 000 dissertations of 20 pages each in average to be scanned – that is, approximately 280 000 pages or images.

My conference paper will describe more closely the arguments for scanning these dissertations, the use of them within research, the practical work with the project and the new possibilities for research it creates. Finally I will mention some issues central to us as a digitizing institution.

Keywords: digitization, old print, accessibility, full text databases, Uppsala University

* Assistant Keeper of Manuscripts of Uppsala University Library, Uppsala, Sweden. Email: anna.fredriksson@ub.uu.se.

烏普薩拉大學 1602-1855 時期論文之數位化

Anna Fredriksson*

摘　要

烏普薩拉大學圖書館為使 1602 年至 1855 年之間遞交給校方之論文可更容易由網際網路取得，正持續進行一項長期計畫。此計畫之願景，在於希望未來可經由單次檢索取得所有 1602 年以降之烏普薩拉大學論文之全文檢索結果。

此一數位化工作之內容，除將論文可供全文檢索的 PDF 檔案以數位影像方式出版外，亦包含詮釋資料生產、掃瞄與光學字元識別等流程。1778 年至 1855 年間之資料為目前第一階段之數位化重點，迄今完成約 2,300 件論文之數位化，成果可由烏普薩拉大學於網際網路上提供之研究出版品典藏庫 DiVA 平台加以取用。整體而言，約有 14,000 篇論文需進行處理，平均一篇論文約 20 頁，掃瞄作業量約為 280,000 頁的影像。

本會議論文將論證此論文數位化作業之必要性與用途，以及此計劃之實務面向議題與產生新研究之可能性。最後將以烏普薩拉大學為例，論述數位化機構之部分核心議題。

關鍵字：數位化、舊籍、進用性、全文資料庫、烏普薩拉大學

* 瑞典烏普薩拉大學圖書館館員， Email: anna.fredriksson@ub.uu.se。

What is a university without its dissertations? In them everything that the University stands for is fused together. They are education of a researcher and execution of research at the same time. They are the melting-pot of what we knew and what we will know, what is old and what is new, the experience of the supervisor and the curiosity of the student, in which tradition is confirmed at the same time as it, in the young scholar's or scientist's interpretation, takes a step away from itself in constant change. In the dissertation, or theses, all this is at display for us to see and judge. As creations of society they reflect what is discussed at the universities, which in their turn mirrors what is going on in the state and in the intellectual world on the whole. This makes dissertations valuable documents for understanding society, and historically, continuity and change in that society. In fact, studying them is an easy and excellent way of seizing the spirit of the time.

At Uppsala University Library a long term project is ongoing, which aims at making printed dissertations, submitted at the University of Uppsala from the very first appearance of them in 1602 till the year 1855 easy to find and read via the Internet.[1] Our vision is that in the future, it should be possible to search through the entire text mass of Uppsala dissertations and theses, from 1602 till today, in one search session.

The work includes metadata production, scanning and OCR processing as well as publication of images of the dissertations in full text searchable pdf files. So far approximately 2350 early modern dissertations have been digitized and made accessible at the Internet via the DiVA portal, Uppsala University's repository for research publications.[2] In this first stage of digital publication the emphasis is on the period 1778-1855. All in all there are about 14 000 dissertations of 20 pages each in average to be scanned – that is, approximately 280 000 pages or images.

1. Uppsala University and Its Role in the Development of the Swedish State

[1] The project is described briefly at Uppsala University Library's homepage www.ub.uu.se. : http://www.ub.uu.se/finding-your-way-in-the-collections/early-printed-books/early-dissertations-in-full-text-on-the-Internet/

[2] http://www.ub.uu.se/publish/about-diva/
http://www.diva-portal.org/smash/search.jsf?dswid=-4401&searchType=SIMPLE&faces-redirect=true&query=&af=%5B%5D&aq=%5B%5B%5D%5D&aqe=%5B%5D&aq2=%5B%5B%5D%5D

Uppsala University was founded in 1477, a time which, in Swedish history, is still counted as the Middle Ages.[3] However, warfare, poverty, and lack of political interest resulted in that the university was inactive almost the whole of the 16th century. In the beginning of the 17th century, new political ideals and resources resulted in a renovation and expansion of the University along with Sweden's educational system as a whole. In fact, there was a renovation of the whole *social system* and the state as such. This age is generally considered as the time when the foundations of modern Sweden were laid.

In the building of the new state the universities played an important part, and Uppsala University had a special position.[4] Close to the capital and government of Sweden, not only geographically, but in many other respects as well, here the development would take place that would make Sweden evolve into a modern state and an integral part of Europe. Not least, Uppsala University was the school that would train and morally mould young men destined for leading positions in the state and for service in the newly founded state institutions. To the universities of Uppsala and later on other Swedish universities, young men from all over the country went for higher studies, to become judges, priests, physicians, teachers and administrators in the service of the state. Understandably, the history of Uppsala University is thus an important part of the history of the Swedish State and even of Western history itself.

All along from the new start, Uppsala University produced dissertations, as did all other universities in Europe, as part of education, but with many other functions in addition.

2. Dissertations and Disputations[5]

In early modern Europe disputations were generally required to obtain a degree, which in the case of Uppsala gave the graduate license take part in public debate and to teach privately after the consent of the dean and the faculty. In that way, the University, as part of the State, had some control of what was being taught and by whom. In the beginning of the 17th century

[3] The most initiated works on the history of Uppsala University are Claes Annerstedt, *Upsala universitets historia* Part 1-3, Uppsala 1877–1914 and Sten Lindroth, *Svensk lärdomshistoria*, P. 1-4, Stockholm 1975-1981.
[4] Sten Lindroth, *Svensk lärdomshistoria. 2, Stormaktstiden*, Stockholm 1975.
[5] Introduction to the subject of early modern dissertations and disputations in the western tradition is given in two articles in *Historisches Wörterbuch der Rhetorik* Bd. 2 (Hg. Gert Ueding, Tübingen 1994) by Hanspeter Marti: "Disputation", coll. 866-880 and "Dissertation", coll. 880-884. Regarding Scandinavian dissertations, see Bo Lindberg, "Om dissertationer", in *Bevara för framtiden. Texter från en seminarieserie om specialsamlingar.* Uppsala 2016, s. 13-39. Short introductions in English are found in f ex Krister Ö stlund, *Johan Ihre on the origins and history of the runes: three Latin dissertations from the mid 18th century*, Uppsala 2000, p. 14–19 and in Peter Sjökvist, *The music theory of Harald Vallerius: three dissertations from 17th-century Sweden*, Uppsala 2012, p. 11-13.

and all the way to the 19th century, dissertations at Uppsala University were composed as a departure point for an oral examination, the *disputatio*, in which the student was to show his skills in argumentation in Latin on the points of discussion displayed in the dissertation. The dissertations were printed in many copies and distributed or sold both before and after the disputation. The supervising professor, the *praeses*, presided at the examination and was responsible for whatever was printed in the dissertation. He also generally defended the contents at the disputation whenever the student failed to do so.

The prerequisites for authoring or, I would rather say, in many cases composing a dissertation were a bit different than they are in most countries today. First of all, it is seldom all together clear who should be considered the author: The supervising professor, or the student. In some cases it has been possible to establish this relationship, but generally we have to be content by stating that it is a product of collaboration between the two. If there was not a third party involved, that is. Probably the main authorship was known to the persons involved through some tacit understanding. Not until 1852 it was stated in the University statutes, that the author of the dissertation should be the respondent. But even today you will find professors who claim they wrote the greater part of their student's dissertation, as well as you will find students who claim that they came up with the idea that forms basis to their professor's brilliant new book. However, generally, this question is, and was, solved by peaceful means.[6]

The question of authorship traditionally has been topical in the overall discussion about the value of dissertations as a scholarly product. In today's research however, this question is considered to be of lesser importance, as the dissertations are now more often seen as products of a certain period of time, a certain discourse, or a certain academic culture rather than of a specific person.

3. Why Did We Prioritize Dissertations and How Are the Dissertations Used Today?

In the great mass of historical documents, on may ask why we prioritized to digitize dissertations? First of all, due to their merit of being open windows into a certain period of

[6] Cfr Ku-ming (Kevin) Chang, "Collaborative production and experimental labor: two models of dissertation authorship in the eighteenth century", in *Studies in History and Philosophy of Biological and Biomedical Sciences* 41 (2010), p. 347-355.

time, early modern dissertations are valued research material. Therefore the physical items of the dissertations were frequently on loan already before the digitizing of them started. In research today, we see that the material is frequently consulted in all fields of history: They provide scholars in the field of *History of Ideas* and *History of Science* with insight to the status of a certain subject matter in Sweden in various periods of time, often in relation to the contemporary discussion at the European continent. The same goes for studies in *history of literature* and *history of religion*. Many of the dissertations treat subjects that remain part of the public debate today, and are therefore of interest for scholars in the *political and social sciences*. The languages of the dissertations are studied by scholars of *Semitic, Classical and Scandinavian languages*, and the dissertations often contain the very first editions and translations of certain old manuscripts in for example *Arabic and Runic script*. In one period of time a popular topic of the dissertations was the description of regions in Sweden, most often the student respondent's home province. Often this description was the first one ever made, and thus, these dissertations can provide evidence of *antiquities* and other features of the landscape now lost or worn down. Studying which dissertations were allowed for a disputation and which were not also gives insight in the area of *freedom of speech and censorship*. There is also a *social dimension* of the dissertations worthy of attention, as dedications and gratulatory poems in the dissertations mirror social networks in the educated stratum of Sweden in various periods of time. *Illustrations* in the dissertations were often made by local artists or the students themselves, an "industry" worthy of studying. The great mass of gratulatory poems mirrors a less well-known side of poetry in early modern Sweden. Dissertations are also witnesses to how various areas of society are interlinked: A dissertation from the 17th century, for example, discussing which organ in the body has the preeminence, the brain, the heart, or the liver, of course mirror the development of our knowledge about how the organs interact in the body and which are their functions[7]. But it also pictures a time when the body was considered a miniature universe, analogue to the great universe outside the body. A world, in which the king generally was considered the head controlling every limb all the way down to the toes.[8] And vice versa, a world in which the limbs should serve the head. In that connection, findings regarding the question of which organ was the most important one in the body was also something for the professors of philosophy of government

[7] Johan Franck (praes.), Joel Jonæ Kylander, *De trium partium principum, cordis, cerebri et hepatis principatu. disputatio medico-philosophica*, Uppsala 1634.

[8] In this specific dissertation these three organs are said to be analogue to the Emperor, the King and the Leader of the State, which three, according to the author, must cooperate in harmony in order for the body, or world, to be healthy.

to consider. Further, as Jesus Christ is quoted in the dissertation regarding God being the Head, this dissertation most probably also got the attention of the Theologians of the University. Every little dissertation was written in a bigger context, and it says a lot about that context.

From a methodological point of view, the great mass of dissertations and their uniformity make them especially suitable for comparative and longitudinal studies. The multitude of them also gives good chances for scholars to find material previously little used or not used at all in previous research. As a bonus, in the clever dissertations, you will get a rich fixed- in-time bibliography to the subject, made up by the bulk of references to the relevant literature of the day. Because of the above, the library, in dialogue with scholars of the university, considered this material to have great potential for future research.

Secondly, the library found it comparatively easy to digitize the dissertations. They are quite easy to handle and of a standard format, and not particularly sensitive. Also, there was a largescale project going on at another Swedish university, Södertörn University College outside of Stockholm, cataloguing the collection of old Swedish dissertations in their care. In this newly founded university college, this collection of old dissertations donated to them by chance was regarded a treasure and perhaps also worked as a way of strengthening their identity as a research institution. Thanks to that, we had the advantage of that an enormous amount of metadata describing Uppsala dissertations already existed. We just had to add to it. In return, Södertörn University College can now use our digitizations of the same titles. But already before that, the collections of Swedish dissertations were described in literature, as Swedish collectors published bibliographies of our dissertations already 250 years ago.[9] In contrast to many other European libraries, which considered dissertations as a material of lesser importance, the physical items of our dissertations are also organized, bound and easily accessible at our library. All of this means that this year, the cataloguing of the ca 14 000 Uppsala dissertations according to moderns standards, in MARC-format, will be completed, which make them searchable according to subject, year of publication and title word, which

[9] Johan H. Lidén, *Catalogus disputationum in academiis et gymnasiis Sveciae* […] *quotquot huc usque reperiri potuerunt* […], *sectio I. Disputationes upsalienses 7450* (Uppsala, 1778); Gabriel Marklin, *Catalogus disputationum in academiis Scandinaviae et Finlandiae lidenianus continuatus* […], *sect. I. Disputationes Upsalienses 3034 annis 1778–1819* (Uppsala1820); id., *Ad catalogum disputationum* […] *Lidenianum supplementa* […] (Uppsala, 1820); id., *Catalogus disputationum* […] *Lidenianus iterum continuatus* […], *sectio I. Disputationes upsalienses 3089 annis 1820–1855* (Uppsala, 1856).

was not possible before. These updates, together with the fact that they are becoming available in digital form have resulted in that the interest in them is greater now than ever.

A third reason why we prioritized dissertation is a practical one, namely, we had a place where we could store and display the digitizations, the DiVA database. This is the database in which researchers, scholars and students of Uppsala University today, and other Swedish universities, too, register their publications with the option to publish them digitally. This fits the older dissertations perfectly, because no doubt they are a product of Uppsala University. Therefore, we could start publishing digital versions of cultural heritage material although a platform for that specific purpose was not yet developed.

4. Other Digitization Projects in and Outside Sweden Involving These and Dissertations

Dissertations and theses of today are very often, besides of being printed, published electronically. Many universities are also digitizing their dissertations retrospectically and make them available open access. In Sweden such activities are going on, or was going on, at for example the universities of Umeå, Gothenburg and Stockholm. These are quite young universities, founded in the 20th century or even in the 1950's and 60's, which means that their collections are not that large, and that the material allows them to use a robot scanner without problems. But they have other issues, such as copyright. Stockholm University started scanning their dissertations without consulting the authors first. A few of the authors reacted, the discussion was picked up by media, there were great protest – and the whole project is now paused and the digitized dissertations are currently not available online. Gothenburg solved this by offering all dissertations authors a digitization and publication of their dissertations for free on request. At Umeå University, the digitizers contact the authors or their ancestors personally, with a letter, and a contract, to make sure they approve of the online publication. Needless to say, the overwhelming majority approves. An author wants to be read, and there is almost no money to make on a dissertation, at least not 30 years later. Similar procedures are carried out at other young universities in the Nordic countries, such as the northernmost university of the world, Tromsö in Norway. Our fellow university in Lund founded in 1666 so far refrained from digitizing their printed dissertations.

I need not mention more examples here outside Scandinavia: They can easily be tracked down via the internet.

I should also mention that libraries often collaborate with commercial actors. For example, the offsite repository of the USA's Digital Dissertations Library is commercial. This can work as a solution in the case of minor universities, which do not have the infrastructure for scanning the dissertations in-house. However, a commercial solution generally means that you can't use your own library's material freely, and calculations and empirical tests have shown that in-house scanning, for those who can, is not much more expensive than outsourcing.[10]

The Scandinavian projects mentioned above and other projects described on the net almost all have in common that they do not include dissertations older than from the end of the 19th or from the 20th century. Although I have tried, I have not so far found very many examples of institutions digitizing dissertations older than that. One predecessor of ours is however the University of Helsinki in our neighbouring country Finland. Initiative taken and money granted by an enthusiastic and very rich book collector, the University Library digitized a selection of dissertations from former Å bo Academy.[11] At the European continent The Max Planck Institute for European Legal History catalogued and scanned about 8,000 title pages of Legal dissertations of the 16th -18th centuries from the universities of the Holy Roman Empire.[12]

5. How the Dissertations Are Used

Who uses our dissertations, then, printed or digitized? As regards the digital version, we actually don't know exactly. We are very curious and we hope to get means to make user-based tests. We do have statistics, which I will touch upon below. In real life, in the reading room and via the mailbox or phone, we meet, not surprisingly, foremost scholars of the University, primarily our own University, from our neighboring country Finland and from the Baltic States, which were for some time within the Swedish realm. Among our dissertation

[10] Such calculations were generally carried out internally without publication of the results. An exception is Piorun M, Palmer LA. Digitizing Dissertations for an Institutional Repository: A Process and Cost Analysis. *Journal of the Medical Library Association : JMLA*. 2008;96(3):223-229. doi:10.3163/1536-5050.96.3.008.

[11] http://blogs.helsinki.fi/natlibfi-bulletin/?page_id=245

[12] http://www.rg.mpg.de/library/dissertations, see also: Amedick, Sigrid, Juristische Dissertationen des 16. bis 18. Jahrhunderts : Erschließung und Digitalisierung von Schlüsselseiten. In: Digitale Bausteine für die geisteswissenschaftliche Forschung hrsg. von Manfred Thaller, Göttingen 2003, S. 86-101 (Fundus: Forum für Geschichte und ihre Quellen; 5) ;Haben, Doris, Ende des Dornröschenschlafes. Moderne Erschließung juristischer Dissertationen des 16. bis 18. Jahrhunderts aus dem Gebiet des Alten Reichs In: B. I. T. online. Zeitschrift für Bibliothek, Information und Technologie mit aktueller Internet-Präsenz 5, 2002, Heft 1, S. 35-40.

users, there is also quite a number of scholars from other parts of the world. One could ask, why any historian outside Sweden would want to study Swedish dissertations? Sweden did not produce the great philosophers or theologians, nor is it the place of origin of new movements or lines of thought within science and society. In the 18th century, it is true, Sweden is regarded to have been in the forefront within the natural sciences, with names such as Linnaeus within botany, Scheele in chemistry and Celsius in physics.[13] Still, in the big perspective, historically, Sweden is one of those countries, which in the area of science and learning more or less shared the values, objects and methods of the Western world as a whole. Thus, to study Swedish science and scholarship is to study an important part of Western science and scholarship. Also, as indicated above, as compared to other European universities' dissertations, they are quite easily accessible.

Many projects are going on right now which include our dissertations as research material or which have them as their primary source material. I will mention some of international interest:

As part of the Erfurt /Halle University project *International law between natural law and a code of civility: discourses of international morality in early modern northern Europe*, Dr. Pärtel Pirimäe (Tartu) studies the discourses of international law and morality in early modern Europe.[14] He then focuses on the formation of a Eurocentric conception of 'the law of civilized nations', which replaced the universalist, natural law–based aspirations of seventeenth-century international law scholars. He also looks at the reception and development of these ideas in Northern Europe.

With its basis in Berlin, the project *Apotheosis of the North –the glorification of Sweden and Finland in the Baroque studies of the Antiquities*[15] is going on, engaging scholars and students in Finland and Germany. The aim of the project is to trace the impact of Olof Rudbeck the Elder's Goticism on the following generations. Rudbeck in the end of the 17th century published the monumental *Atlantica*, in which he, primarily using linguistic evidence, identified both Sweden and Finland as the "origin of all nations", and Scandinavia as being

[13] For an account in English of early Swedish science, see Colin A. Russell, 'Science on the fringe of Europe: Eighteenth-century Sweden' in *The Rise of Scientific Europe, 1500–1800*, ed. D. Goodman and C.A. Russell (London, 1991), ch. 12, 305–32.
[14] https://www.uni-erfurt.de/projekte/natural-law-project/projects/research-projects/international-law-between-natural-law-and-a-code-of-civility-discourses-of-international-morality-in-early-modern-northern-europe/
[15] https://blogs.kent.ac.uk/ewto-news/2014/11/24/workshop-apotheosis-of-the-north-berlin-finnland-institut-16-17-december-2014/

Plato's Atlantis. The project aims to study the impact of these theories and how adherence to them could affect academic careers in the generations to come.

Within the framework of the project *Encounters with the Orient*, Prof. Outi Merisalo (Jyväskylä) and Prof. Bernd Roling (Berlin) in *Biblical encounters: The Linneans and the Bible*[16] studies the relationship between new methods within science, linguistic studies and the interpretations of the texts of Bible. Two questions explored are "To what extent did findings of early modern natural sciences make an impact on the exegesis of the Bible?" and "How did this influence the image and representation of the Biblical world (not to say: the orient) itself?"

The object of a project of Dr. Meelis Friedenthal (Tartu) is to analyze the intellectual tradition of the Baltic Sea region during the period of the Swedish Empire (1611–1721).[17] The approach is to examine university disputations and to investigate the presence and reception of new philosophical ideas. He will then use quantitative text analysis methods on digitized disputations in order to detect changing patterns of thinking. Previously Friedenthal studied the concept of tolerance in early modern German university disputations.

In Sweden, too there is a big project now in its final stages, *Academic culture in the Baltic Sea region in the Early modern period*, which involved eleven scholars from three countries under the head of Prof. Bo Lindberg (Gothenburg) and Prof. Erland Sellberg (Stockholm).[18] The empirical material has, to a large extent, consisted of dissertations, orations and lectures. The research have dealt with official rules and academic practice, the use of Latin and vernacular languages, peregrinations and disputations *extra patriam*, the relationship between scientific and rhetorical Latin.

Within the area of History of Medicine and History of ideas, within the project *Medicine at the Borders of Life: Foetal Research and the Emergence of Ethical Controversy in Sweden*[19] Dr. Maja Bondestam (Uppsala) includes dissertations in the study of extraordinary births and their value in Swedish medicine 1660–1830. Among other things the project examines the emergence of unexpectedly shaped human fetuses as objects of medical inquiry.

[16] https://www.kent.ac.uk/ewto/projects/Biblical%20Encounters/Linneans.html
[17] http://www.swedishcollegium.se/test/subfolders/Fellows/Profutura/Meelis_Friedenthal.html
[18] http://ostersjostiftelsen.se/projekt/453-the-academic-culture-in-the-baltic-sea-region-during-the-early-modern-period
A conference proceedings volume with the preliminary title Early modern academic culture is forthcoming, editors being Prof. Bo Lindberg and Prof. Emer. Erland Sellberg.
[19] http://www.idehist.uu.se/research/research-projects/medicine-at-the-borders-of-life/

Needless to say, there is a vibrant activity at the European continent with regard to dissertations, especially in the German speaking area: The last decade's publications of rich anthologies on the subject bear witness to this fact.[20] The research conducted in Northern Europe however often focus on the scholarly traditions and discourses of one specific university or region. Therefore research looking at the broader picture and asking fundamental questions regarding the role of the dissertations or theses within the society of knowledge are welcome contributions to the discourse.

Many ask us about the language. As all other countries of Europe and North America, Latin was the vehicle for academic discussion in the early modern age. Therefore, the great majority of the dissertations are written in Latin. In the first half of the 19th century, Swedish became more common in the dissertations. Among the ones digitized and published so far, a great deal are in Swedish. Now, it has been asked by critics, why we should bother about this old literature, since people don't know Latin anymore. This may be partly true as regards Swedish scholars: Outside Sweden the situation is another. In many western countries Latin is introduced early on in the curriculum, and for those trained to be scholars of history, it is part of their education. To these researchers, the Latin dissertation actually would be easier to read than the ones in Swedish – and thus they will conduct research concerning our country that would otherwise probably not be conducted. But even if Swedish students of today did not take Latin in school, as scholars of history, they through their everyday work obviously do learn enough Latin to identify which documents are important to them and to recognize if a passage treats the topic of their interest. They can also extract the most important information from it. If the document appears to be central, it is possible to hire a translator.

But we believe that we also reach out to the so called "ordinary people". The older dissertations treat every thinkable subject and they offer pleasant reading even for non-specialists. They can also be surprisingly topical. For example, two issues discussed almost daily in Swedish newspapers today, "how to accept foreigners in the country" and "begging" were discussed some hundreds of years ago, too. The dissertations on these topics display interesting perspectives and deliberations which no doubt shaped the way we think about

[20] See Schwinges, Rainer Christoph & Schöpfer Pfaffen, Marie-Claude (red.), *Examen, Titel, Promotionen: Akademisches und staatliches Qualifikationswesen vom 13. bis zum 21. Jahrhundert*, Basel 2007; Gindhart, Marion. & Kundert, Ursula. (red.), *Disputatio, 1200-1800: Form, Funktion und Wirkung eines Leitmediums universitärer Wissenskultur*, Berlin 2010; Sdzuj, Reimund, Seidel, Robert & Zegowitz, Bernd (red.), *Dichtung, Gelehrsamkeit, Disputationskultur: Festschrift für Hanspeter Marti zum 65. Geburtstag*, Wien 2012; Gindhart, Marion., Marti, Hanspeter., Marti-Weissenbach, Karin. & Seidel, Robert. (red.), *Frühneuzeitliche Disputationen: polyvalente Produktionsapparate gelehrten Wissens*, Köln 2016.

those matters in Sweden today. There are also often references to both philosophical and theological dilemmas and standpoints which are almost totally missing in the debate of today. Whatever the subject, the cases of an internet visitor by chance stumbling into one of our dissertations when searching for a certain word or subject, or an ancestor, would be many. Not often does a Google search result in texts like these, taking a wide range of arguments into consideration! Perhaps we also this way contribute to these people's education and pleasure. That could explain why in the first six months after the digital publication of the first 2300 dissertations there had been over 23800 downloads of these old texts, and over 81 200 visitors on their unique web pages. A press release in May and June 2015 resulted in over 20 000 downloads these months only. And then you have to consider that there are only 9 million inhabitants in Sweden all in all, in contrast to Taiwan's 23.5 million. It is interesting to note that all these visits and downloads occurred even if we don't – or perhaps *thanks to* that we don't – neither offer nor demand advanced technologies for the use of these dissertations.

In fact, advanced technologies do not seem to be central to an increased use of digitized material. The experience of a colleague of mine at a library in the United States which has been working with DH within the humanities for a long period of time, was that a very small group of users would ask for advanced technologies for to be able to conduct their research, whereas the overwhelming majority needs guidance for to be able to use the most common and simplest programs, which will enhance their research results a lot. One of our meetings the other day with a group of researchers using our digitizations confirmed that: On our question about what was most important for them when using our database, they answered: 1) That the material they want is there 2) That it is easy to find.

A study within library science have shown that a reason for not citing a certain work is that it is hard to obtain,[21] and yet others have shown that scholars tend to cite works which are already in their room more than other works. This is not very surprising to anyone. It's also well known to librarians that professors tend to hoard library books in their own rooms. So, if there is a work claiming the same thing slightly more convincingly, but which is not in the room, this work will most probably not be cited, unless it is one of those that you "must" cite. Consequently, a good strategy for a work to be read and cited is to be in the room. To be

[21] See f ex M.D. White and P. Wang, 'A qualitative study of citing behaviour: contributions, criteria, and metalevel documentation concerns', *The Library Quarterly*, 67/2 (1997), 122–54 (esp. 149–52), in which 'hard to obtain and read' is given by scientists as a reason for not citing a specific item.

in a person's computer is definitely to be in the room. Now that our dissertations are on the internet, we believe that they will be studied more than ever before. I have already cited the results for the first six months of our digitizations on the internet. Recent statistics for the whole period from the first publication in spring 2015 record all in all over 380.000 visits on their unique web pages and over 42.000 downloads! And then there are just 2.300 of them out there so far. We are surprised, but the increased use is obviously true about younger dissertations, too. The digitization of 300 medical printed dissertations alone at the Lamar Soutter Library of the University of Massachusetts Medical School meant that the collection was used 17,555 times in ca 1,5 years as compared to 723 times in the 5 years before electronic publication.[22] Now we have to consider that the grounds for calculation differ. The gap would probably not be so enormous, and the numbers for prints would be higher, it we counted all the times those printed volumes were shut and re-opened by the same reader in one research session.

6. The Digital Publication and New Possibilities for Research

The database in which our digitized dissertations are stored and presented is, as said above, Uppsala University publications database, i e, the same database in which researchers, scholars and students of today register and digitally publish their publications. It's a Fedora based repository which was developed and is maintained by Uppsala University Library. DiVA is managed as a consortium including all in all 40 member institutions, and it accepts members both within and outside Sweden. DiVA started to work in 1995 and has been stuffed with contents by its own users since then. At Uppsala University it is obligatory for researchers and other employees to register their publications in DiVA. All doctoral students must post their doctoral theses electronically in DiVA.

According to Swedish Law ('lagen om pliktexemplar'), everything published electronically in Sweden should be downloaded and stored by our National Library. That means that all e-publications, including our newly made digitizations of old dissertations, end up in the National Research Libraries database LIBRIS. Having all universities' publications in one database of course facilitates for researchers to find texts on their topic, and in this case you can conduct searches in both LIBRIS and DiVA. Most importantly it makes it

[22] Piorun M, Palmer LA. Digitizing Dissertations for an Institutional Repository: A Process and Cost Analysis. *Journal of the Medical Library Association : JMLA*. 2008;96(3):223-229. doi:10.3163/1536-5050.96.3.008.

possible to make word searches which could result in set of first-hand-material from over a 400 years period of time in the same topic or on the same issue. A great deal of medical terms of diseases and body parts, chemical designations, and, of course juridical and botanical terms are Latin and the same as were used 400 years ago, and can thus be used for localizing text passages on these topics. This could prove to be interesting even to natural scientists of today who could get a full overview of experiments conducted in their field and theories previously dismissed or forgotten. I don't need to remind of that the Nobel Prize in Medicine last year was given to Youyou Tu for her discoveries based upon information in old medical texts.[23] Our botanical dissertations are already highly frequented, perhaps by botanists getting information of now forgotten habitats of their plant. But the form of the text can be studied, too: Linguists would find it useful to make quantitative studies of the use of certain words or expressions, or just finding the words of interest for further studies. The usefulness of full text databases are all known to us. But often you as a user get *either* a well working search system *or* a great mass of important texts, and seldom both. This problem is addressed here by the interconnection between the publication database DiVA and the previously mentioned Swedish National Research Library System LIBRIS. The combination makes it possible to take the advantage of the Library systems professionally "handmade" and controlled metadata within an advanced search system, thus reducing the internet problem with too many irrelevant hits. Scholars are all too well acquainted with the problem with internet searches based on metadata extracted automatically from certain parts of the text. Apart from the enormous number of irrelevant hits, searches for old prints often give quite misleading results, due to the problem with OCR of older texts. Also, the subject classification of the library catalogue proves very helpful in the case of old texts, as obsolete and variant spelling forms, sometimes not known to us, occur in the dissertations and make word searches less successful.

7. "Alvin" – a Digital Platform and Repository for Cultural Heritage Material

As part of our digitizing work, our library also developed a combined repository, platform and interface for other kinds of digitized cultural heritage material, such as

[23] Nobel prize winner Youyou Tu studied old herbal recipies, which resulted in a new way to treat malaria. See Youyou Tu's autobiography at Nobelprize.org or https://www.nobelprize.org/nobel_prizes/medicine/laureates/2015/tu-bio.pdf p. 275 sqq.

manuscripts, maps, pictures, music, coins and museum specimens. This platform, which is called “Alvin” and launched just a couple of years ago, allows for adding more information about each item, and also link related objects and literature. The Alvin system was, just as DiVA, organized as a consortium, so that members both contribute to its contents and finance its support, upgrading and development. Our most dedicated partners are the universities of Gothenburg and Lund, but so far 20 larger and smaller institutions, even public authorities are part of the Alvin group.

Both systems DiVA and Alvin offer stable search and publication systems – linked to certified contents, for free. These simple tools give our users the possibility to quickly learn and use the system, experiment and suggest further developments. For example, one of our users used the possibility to order digitizations of dissertations of his choice “on demand” – we offer the service called EOD, “E-books on demand” for our material printed before 1855 – for getting digitized dissertations published in Alvin. All EOD digitizations namely end up in Alvin, (although a copy of every digitized dissertation is also preserved in DiVA). In this repository, it is possible to interconnect documents and files to visualize the context of a document. This user, a scholar of numismatic, started interlinking documents both within Alvin and to Diva, in order to connect them to the special internet page he created. Another group of users used Alvin as a tool for digitizing, presenting, and linking documents central for an upcoming conference. The participants could then prepare themselves well for the conference by studying both the set of primary sources and the secondary literature connected to them. The fine thing is, that the contents and the links will still be there and useable for others after the conference.

8. The Practical Work within the Project and Related Issues

The practical work with the scanning is carried out within our own library by our image technicians. This is quite convenient both when it comes to logistics and control of the process and the handling of the documents. It also makes communication easier. In scanning, various techniques are used depending on the format, binding and age of the dissertation. The younger items can be scanned using a scan robot, whereas older items and those including illustrations must be scanned manually. In this mass digitization, with certain demands with regard to speed and due to our use of a robot scanner, it has proved necessary to accept images of a somewhat lower, however of course fully readable, quality. The images of the

text pages are OCR-processed in order to create searchable full text pdf files. The OCR process, too, gives various results depending on the age and the language of the text. OCR-processing dissertations in Swedish and Latin from ca 1800 onwards result in an OCR text of a high degree of accuracy, that is, around 90%, whereas older dissertations in Latin and in languages written in other alphabets will contain some inaccuracies. On this point we are not satisfied. Almost perfect result when it comes to the OCR-read text is a basic requirement for the full use and potential of this material. However, in this respect, we are dependent upon the technology which is available at the market, as commercial actors so far provided the best and safest product. These products were not developed for to handle printing types of various sorts and sizes from the 17th and 18th century, and the development of these techniques, except when it comes to "fraktur", is slow or non-existing. This however does not seem to be regarded a great obstacle by our users. They still express gratitude about the fact that the dissertations are there at the net *at all* and that they are OCR:d *at all*. We may however assume that demands will grow exponentially.

9. How to Use the Digitizations

If you want to pursue further studies on the documents, you can to download the documents for free and save them on your own computer. We have considered the fact that not everyone has a great computer and therefore we made them as "light" as possible. There is free software on the internet and instructions[24] that help you merge several documents of your choice to one document, in order for you to be able to search through a certain mass of text. If you are searching for something very particular, you could of course also make a word search in Google. One of our wishes for the future is to make it possible for our users to search in several dissertations of their specific choice at one time directly in DiVA, without them having to download the documents to their computer.

If you want to correct the OCR text, Wikisource can provide a structure for that, however, you will have no control over your document and other users can interfere with your correcting. On the other hand, if you are lucky, someone else will engage and do the job for you. A safe way is however to keep it on your own computer: At the image file, you just mark the whole text mass, copy and paste it into a new word-document. Then you will see the

[24] http://www.wikihow.com/Merge-PDF-Files
A search google on "merge pdf" will lead to several free programs for merging files.

OCR text which was hidden behind the image and you will also see all its flaws. Put the image page side by side with the word document and start correcting.

It would be wonderful to be able to collect all the work that our users carry and get it back into the database as additional information to the document. As for now, we do not have the administrative tools for to do that, but we are working on it.

Finally: Most important for us today within the dissertation project:

1) Further scanning to add more dissertations to the database
2) Better OCR for older texts
3) Easier ways to search in a large text mass of your own choice.
4) Help users to find use existing free resources for the expanded use of the texts

10. What We Produce and Its Place in the Society of Knowledge and the Area of DH.

As you have understood, what we have here is not rocket science. We simply put together, interconnect, and present our resources in new ways. Among our most valuable resources we count professionally collected, authorized metadata and documents from every period of time treating every thinkable subject in the world. We are greatful to have well kept, functional library systems, and institutions of the State, the University Libraries and the National Library, appointed to care for its long term maintenance and update. Also, we have persons working with those things who understand the material and who are open for discussion with the users.

Even so, we would love to see further development of digital techniques for the enhanced use of these texts. Therefore we aim to increase our collaboration with researchers who wants to explore new methods for to make more out of them. Doing that however, we always have to take into account that there are special demands from society when it comes to the work we, as a institute of the state, are conducting – in contrast to the work conducted by f ex Google books or research projects with temporary funding. For example, we are expected to produce both images and metadata of a reasonably high quality- a product that the university can ‘stand for’. What we produce should have a lasting value – and ideally possible to use for centuries to come.

What we produce should also be compatible with other existing retrieval systems and library systems within Sweden and in the world. Important, in my opinion, is reliability – that researchers and community feel confident about the system and contents – and citability – that the contents are regarded reliable and stable to a degree that researchers dare cite them in their publications. A great problem with research on digitally born material is, in my opinion, that the material itself constantly changes, both with respect to their contents and the place where to find them. This puts the fundamental principle of modern science, the possibility to control results, out of the running. This is a challenge for DH which, considering the pace its development has today, I am sure will be solved in the near future.

Reality is *so far* not moving in the same speed as our imagination, and it is my conviction, said in my profession as a scholar of Language rather than in my profession as a librarian, that the great digital revolution *so far* lies in the fact that a great mass of people can access a great amount of literature easier than ever before, and make word searches in the full text documents of this literature. In my opinion, which I am open to discuss, a great deal of development of technique within the area of DH *so far* regards finding documents easier, finding new ways of squeezing new information out of these texts, and to present the results in new, even more pedagogical ways. It is about development of research methods and development of pedagogical ways to work – but eventually it will probably also change both how we conduct research at all, and the topics of our research, and it will create new data to explore.

Such changes the academic world has witnessed more than once in the course of history. And, to get back to where I started, some day we will be able to track those changes by studying the theses of today's universities.

It is a dream of our Head Librarian that he will one day be able to present the Vice-chancellor of the University with one USB containing all dissertations ever published by our University. Just the thought of it, 400 years of labor and invention, which had such impact on our society, in a space smaller than a box of matches. It's a captivating thought.

唐代交通圖與 GIS（地理資訊系統）的運用

朱開宇*

摘　要

本論文研究構想源自嚴耕望先生《唐代交通圖考》(以下簡稱《圖考》)一書，由於仔細閱讀並讀懂該書，且將嚴耕望先生《唐代交通圖考》書中嚴謹精密的考證內容與譚其驤先生《中國歷史地圖集》(以下簡稱《歷史地圖集》)第五冊唐代部分的圖資訊息結合在一起，儘可能力求嚴謹與無所疏漏地落實於當代中國地圖學史上具里程碑的申報館《中華民國新地圖》(以下簡稱《新地圖》)上，進行 GIS 數位化，因而衍生出一些關於唐帝國軍政佈置與相關空間意涵等問題意識。蓋《圖考》一書細究其內涵實包含了專精（考證精詳）與廣博（唐代的軍政佈局與如何利用交通控制、統整帝國）兩部分，而交通又是空間發展的首要條件，因此，極適宜於運用 GIS 完成具有高度價值的數位化地圖，因為 GIS 本為處理空間資訊的有利工具。有其專精，故能體會其運用歷史地理學之方法，且將其考證一一落實於數位化之底圖；有其廣博，故而最後完成之成果能顯示出唐朝這樣強盛且地域廣泛的帝國在空間上的運作與構成，尤其是邊防區的軍政佈局。這樣的數位化的地圖能完全彰顯《圖考》一書的價值，更易使人明瞭，又能避去其艱澀的缺點。

因此，本研究關涉到幾個主題，其一是理解嚴耕望先生《圖考》所運用的歷史地理學之方法，這也涉及為何嚴耕望先生會運用當代地圖來復原唐代交通，同時也可以使我們理解歷史地理學中如何運用文獻、考古、地圖與當代交通、地理生態環境之報導等來考證唐代之交通與地理行政點。其二是對申報館《新地圖》優點之論述，為何可以《新地圖》作為數位化〈唐代交通地理資訊系統〉的底圖，又為何《新地圖》上的交通路徑極為適宜作為復原唐代交通的參考道路。其三是本〈唐代交通地理資訊系統〉乃結合《圖考》、《歷史地圖集》與《新地圖》三者優點的成果，而譚其驤先生《歷史地圖集》有著遠較《圖考》更為多的地理·行政點，將其精準地歸位於《新地圖》上，能夠更為全面地觀覽唐朝的軍政佈置，也能更全面地理解這些為數眾多的地理·行政點與申報館《新地圖》路徑、河流、山脈等的關係；此外，有許多譚其驤的地理·行政點並不在嚴耕望先生所考證的交通

* 開南大學通識教育中心專任助理教授，Email：chuki235@gate.sinica.edu.tw。

道路上，卻位於《新地圖》路徑上，彼此串聯著，可以將這些聯結著譚其驤先生地理·行政點的《新地圖》路徑，視為唐代交通道路的復原。可以說，搭配譚其驤先生《歷史地圖集》與申報館《新地圖》，更能全面地顯示與復原唐代的交通（許多《新地圖》路徑經過譚其驤的地理·行政點）與軍政佈局，且有行政區劃可為參考。這樣重新繪製的〈唐代交通地理資訊系統〉，將嚴耕望與譚其驤兩位先生的優點結合，同時亦能使《圖考》一書的價值更被彰顯、突出，其書中所蘊含的廣博面（軍政佈局與帝國控制）更能為人所體會，在閱覽此數位化之地圖時，空間上的意涵能充分顯現並被充分體會。最後，筆者以為《圖考》不僅是集歷史地理學之大成的著作，也復原了唐代較為重要且可考的交通道路，更為重要的是嚴耕望先生嘔心瀝血準備大半生所欲完成的傳世鉅著，是有其大問題與大關懷的，亦即唐朝如何運用交通和配置在道路或道路附近而能以制高點控扼道路的軍、鎮、戍、守捉與城堡等鎮戍體系，以及州（郡）縣等行政體系、通道上的驛傳體系來控制帝國、駕馭邊區、防禦敵人，大唐帝國盛時的遼闊版圖是如何構成與維持，在整個唐代興盛與中衰的歷程中面對突厥、吐蕃、回紇與南詔連番興起與挑釁，唐朝又是如何在邊防上應對的。蓋大唐帝國的強盛與版圖之遼闊，與其帝國的開放性格與政治、軍事的宏偉遠略自是息息相關，其中，將高明的軍政配置與對交通的高度重視結合起來，運用地形與交通作成環環相扣的軍政佈置，由兩京輻射出的全國性控制網，與對邊防區精密的戰略佈置，更是其能成就其帝國的重要原因。《圖考》一書所要論述的核心主旨不外乎此。因此，嚴耕望先生所重視者，凡州、縣、軍、鎮、關、戍、館、驛，有可考者，皆表而出之，以明交通之正確路線。可以說，交通道路系統是帝國的動脈，州、縣、軍、鎮、關、戍、館、驛正是佈置在大動脈與小動脈上的控制點，或使這些動脈的運行能保持順暢。

嚴耕望先生的鉅著《唐代交通圖考》嚴謹詳實且廣博宏偉，可謂致廣大而盡精微，已堪為傳世之作。譚其驤先生《中國歷史地圖集》亦為中國學術界關於歷史地理的重要參考圖籍。而關於唐代交通與歷史地理等議題，由於唐代特殊的歷史性格，委實是唐史一重要課題。且嚴先生的《圖考》一書，不僅精微考證而已，其撰書意旨隱含著欲以交通為經，唐代史事為緯，而欲成一家之唐史，書中實含有唐朝如何成為一偉大帝國的主旨。從《圖考》中可窺視唐朝的政軍格局與邊防佈置，這些都反映了唐朝的特殊性格與造成遠大帝國的原因。另一個與此相關的是，唐帝國的交通系統對於帝國的確立，有著密不可分的關聯。正因唐帝國對於交通的重視，並運用交通來控制整個帝國，所以對唐代交通與相關行政、軍事建制的文章相當的多，《圖考》也是在關注這重要議題下產生的鉅著。正如先生於《唐代交通圖考·序言》中開宗明義所闡明的，「交通為空間發展之首要條件，蓋

無論政令推行，政情溝通，軍事進退，經濟開發，物資流通，與夫文化宗教之傳播，民族感情之融合，國際關係之親睦，皆受交通暢阻之影響，故交通發展為一切政治經濟文化發展之基礎，交通建設亦居諸般建設之首位。」蓋中國疆域遼闊，交通建設尤為要務，而唐朝又極為重視交通建設，《圖考》所涵蓋的，舉凡軍政佈局及與異族之軍事進退等，藉由交通視角切入，實際上可以以空間作為深化理解歷史的一種思維。地理資訊系統（Geographic Information System；GIS）本為處理空間分析的一種有力工具，而理解歷史不僅於文字與文獻一途，還可以以空間作為一種理解歷史的思維，如同圖像、考古或藝術史作為理解歷史的一種思維般，仍有其意義與價值。何況，《圖考》一書本就文獻精密考證與作者厚積薄發的文史涵養而成。如何運用 GIS 的有力工具，以空間作為理解歷史的視角，而不僅只限於作為一種示意的圖，《圖考》可以成為一有力的運作與示範。蓋《圖考》考證精詳，且涉及唐帝國的整體軍政佈局，可以說唐帝國正是以空間的思維來運作、擴展與維持其帝國的發展。也正因為唐朝重視交通與空間作為帝國發展之憑藉，唐代的歷史地理學方有高度的發展，當代相關研究成果，諸如西北史地、域外交流、漕運與交通、邊防與軍政佈置等，才如是受重視且重要。而這些涉及唐代歷史地理的文獻與學術成果，不僅高度涉及空間思維，相關成果對於地理與地圖的運用都很重視，復由於此一領域之議題對於唐史之重要性，正可使空間作為一種理解唐帝國歷史的思維與憑藉，成為有意義的學術課題。因此，筆者歷時六年的時間，以 1930 年代刊載、近代中國地圖學上具里程碑式的申報館《新地圖》為底圖，詳細閱讀《圖考》與相關歷史地理研究，將《圖考》與《歷史地圖集》中的相關資訊以極為嚴謹、務求精準的態度，落實於申報館《新地圖》上，且《新地圖》繪有當時中國的道路系統，這些道路系統反映著現代科技較為深遠地影響、改變地貌前，符合自然地形與人文歷史演變下的道路，且多具有「大道」的性質，實可作為本研究數位化〈唐代交通地理資訊系統〉的參考道路。正因對交通路徑、地理·行政點、河川湖泊與山脈的整體關聯、相對位置等的每一環節都細心注意，仔細查對嚴耕望先生《唐代交通圖考》與譚其驤先生《中國歷史地圖集》之相關資訊，將這些資訊嚴謹地落實於《新地圖》之〈地形圖〉與〈人文圖〉中，真正將相關空間物件之定位精準化，儘可能無所遺漏疏忽地找出可能錯誤，才能真正顯示有意義的唐代交通圖，一個不僅只是示意的交通圖，而能落實於精準地圖上的地理·行政點、村落城鎮、精準方位、里距，呈現出真正空間意義的交通路線圖。正因為能準確反映交通路線及其與河流、山脈等之地理形勢，更能看出嚴耕望考證與譚其驤圖繪的地理．行政點，所反映出的軍政佈置，尤其是軍、鎮、戍、城堡等所處之重要地理條件；亦即，在精準的定位後，空間的歷史意義得以呈現。

可以說，筆者是將《圖考》的精密考證與《歷史地圖集》的周全相結合，以申報館《新地圖》為底圖，運用 GIS 的新技術，作成精密地數位化與繪圖，並期待結合後的成果，可深化為一種比較與討論，作為對學術的一份貢獻。諸如：我們可藉由嚴耕望考證的唐代道路及其性質分類與《新地圖》路徑作對比，來討論道路變遷的社會史意義。舉例言之，在交通路線方面，屬於紅色線段的唐代路線，由於被分為三級，其中屬於主要驛道的粗紅線段可以說絕大部分均與《新地圖》路徑疊合，那麼沒有與《新地圖》路徑疊合的，在唐代時為何會有如此重要的交通功能，當代又為何沒有路徑通過；而屬於普通道路的紅色虛線段，絕大部分都沒有與《新地圖》路徑疊合，它們在唐代又扮演什麼功能而顯諸史冊呢？這些探討實足以去闡明歷史的變遷。又如地理·行政點的比較方面，我們雖有嚴耕望精密的考證論述，卻沒有譚其驤的考證文字，而僅有其所繪之圖。不過，由嚴耕望所引之史料與其論證之方法、過程，亦可來檢視譚其驤所繪之圖是否合理，至少可以試圖對比出兩者思路的不同。甚至，藉由 GIS（地理資訊系統）的運用，我們甚至可以補充或核對嚴耕望考證之內容。我們除可觀覽有哪些地理·行政點是嚴耕望所有而譚其驤所無，又有哪些是嚴耕望所無而譚其驤所有，藉此窺視兩人著重視角之不同處。亦可藉由兩人對同一地理·行政點不同之結果，而以嚴耕望先生《圖考》一書之考論與徵引史料，以作為探討兩家差異形成之所由，乃至判定孰是孰非的可能依據。我們還可透過當代相關的歷史地理論著之考證，來佐證嚴、譚兩家之是非。同時，這些地理·行政點所位於《新地圖》之交通條件，諸如交通路徑、河谷地帶等，亦可作為觀察這些地理·行政點的形勢條件。

最後，若總結嚴耕望先生《圖考》，以及筆者數位化〈唐代交通地理資訊系統〉的成果，筆者試著提出一個極有可能的觀點，那即是中國的交通道路系統，發展到唐時期，已經到了完備成熟的階段，宋以後的交通道路系統，基本上承襲著唐代所已發展出的規模。嚴耕望先生便曾指出唐代是中國道路系統的「真正大定」時期，筆者更由數位化時對嚴耕望與譚其驤的地理·行政點的精密定位來進一步確認，由於嚴謹地將嚴耕望與譚其驤的地理·行政點落實於《新地圖》中後，可以發現《新地圖》路徑上貫穿了唐代的地理·行政點，即可為證明。筆者在數位化此〈唐代交通地理資訊系統〉後，可以發現無論是嚴耕望或譚其驤的地理·行政點，基本上有相當高比例都與《新地圖》路徑或河流有相當關聯，不是位於《新地圖》路徑上（或鄰近處），便是位於河流沿岸。雖然，還是有些地理·行政點並不在《新地圖》路徑上，這中間既蘊含了可能已然改道，但考慮《新地圖》路徑多為溝通大城市的大道、區域間的主要連繫道路，以及行經受天然地形限制下長期作為通道的諸項性質，似仍可推斷在唐代尚有其他較為細小的道路存在，亦即，在唐代

有比《新地圖》路徑還要更為深入一般村鎮或較為深僻地區的道路存在。正因為到唐時期，中國道路系統的完備基本完成，所以遠程貿易開始興盛。吾人若細究《圖考》竭澤而魚所引的材料，一言以蔽之，政治（含軍事）實籠罩全局，蓋此本為中國近世前之社會性質所致，商業經濟的活絡尚未到達一高度階段，政治仍為推動社會之主要力量，而交通道路之開展，因政治、軍事所需，實憑藉政治力而成，然正因中古完成了此一政治力打通全身經脈的任務，故長程貿易開始萌生，商業經濟得以大幅前進，在其他條件相輔相成下，得以邁入近世社會的階段。

關鍵字：嚴耕望、《唐代交通圖考》、譚其驤、《中國歷史地圖集》、申報館《中華民國新地圖》、GIS（地理資訊系統）、唐帝國、交通、軍政佈局

Digitization of the Tang Dynasty Transportation Map with Geographic Information System (GIS)

Kai-yeu Chu*

Abstract

Tang dai jiao tong tu kao 唐代交通圖考 (briefed as Tu kao) compiled by Yan Gengwang 嚴耕望 is widely accepted by the scholarly community as a detailed and masterful piece of textual documentation. With potentials of GIS (Geographic Information System) today, it may be worthy of further effort to expand readers' perception of the landscape from the original *Tu kao* into a spatial dimension aided by the historic change of geography and administrative offices. A breakthrough in this technical aspect would be especially relevant given the fact that, according to *Tu kao*, the Tang Dynasty took spatial networks as a communication means to run its civilian/military governance throughout its entire empire. Methodologically, this researcher reviewed in details the contents in the *Tu kao* including the communication routes, regional centers of administration, and terrains and waterways. Cross-references are made with the assistance from the fifth volume of *Historical Atlas of China* (*Zhong guo li shi di tu ji*) 中國歷史地圖集 complied by Tan Qixiang 譚其驤 and from the *New Map of the Republic of China* 中華民國新地圖 issued by *The Shen Bao* 申報. The results of this academic effort prove substantial. The revised digital map seeks to position the historical sites more accurately, represent spatial meaning of historic events, and more importantly, it lays bare issues related to change of communication networks seen from the social perspective.

Keywords: GIS (Geographic Information System), Yan Gengwang, *Tang dai jiao tong tu kao*, Tan Qixiang, *Historical Atlas of China*, Tang Dynasty, *The Shen Bao, New Map of the Republic of China*, communications, civilian/military governance

* Assistant Professor, Center for General Education, Kainan University. Email: chuki235@gate.sinica.edu.tw.

南京濱水景觀的轉型（1920s-2000s）：基於歷史地理資訊系統的製圖探討

徐振*、韓凌雲**

摘　要

很多中國城市的起源與河流有著密切的關聯，人們在利用河流建設城市的同時也對河濱乃至河流進行著改造。自從早期工業化和現代化以來，在很多中國城市的河岸帶發生了劇烈的變化。通過追溯濱水區風貌的變化可以理解城市和水的關係，以及其中反映出來的不同代理人塑造城市景觀中的態度和作用。這些對審思當今城市與水的關係頗有價值，在實踐上亦可以為景觀管理提供參考。

南京城毗鄰長江，為重要的歷史文化名城，也長期居於中國經濟文化最發達的長江中下游地區，具有大約 2480 年建城史和幾百年建都史，歷代城池的形成和轉變都與河流利用和改造密切相關。奠定南京老城現在格局的為明代都城，明代建都時都城範圍達 41 平方公里，有 13 座城門（其中 2 座水關），系世界上最大的城牆。明代都城城牆修建時結合前代遺存、功能分區、自然地形尤其是水系，所形成的骨架迄今仍為南京城市風貌精華所在。

基於歷史地圖、檔案和圖畫，本文以歷史地理的視角追溯南京濱河地區 1920s 至 2000s 之間的景觀演變軌跡。作者在梳理了南京城市及水系演變概略的基礎上，對城市沿河風貌的研究包括這幾個方面：

（1）以 GIS 為平臺，通過對不同時期精確歷史地圖的校正和疊加，對河流及其他水體的形態和功能變化進行製圖；（2）結合繪畫作品、歷史圖片分析沿河景觀的演變，包括歷史風貌的變化分析，視廊、眺望點和沿河視域變化；（3）典型地段如秦淮河周邊的城鎮平面變化、土地利用和建築肌理，以及對規劃決策的回顧。

作者利用地理資訊系統進行地圖的校正、疊加分析，發現其在研究景觀過程中具有四個方面潛力：

* 南京林業大學風景園林學院副教授，Email: xuzhen@njfu.edu.cn。
** 江蘇第二師範學院城市與資源環境學院講師，Email: landscaping@163.com。

（1）校正不同比例和方向的地圖，便於研究者發現變化。如今數位地籍和遙感圖像更易獲得，可以為歷史地圖研究的拓展資訊管道和空間範圍。

（2）通過屬性查詢便於發現平面和統計指標。而且通過對城市形態學中的平面單元、建築肌理和土地利用的分析，可以自動或者半自動地甄別出景觀分區，適合於對大尺度或初步的景觀風貌的分析。

（3）河流廊道涉及公眾福利、自然保護和利用等多個方面，應該以綜合的視角來進行研究。GIS 可以作為將形態研究與生態、社會經濟因素關聯的平臺。志願者提供的移動設備定位資料，則有望提供更多的研究機遇。

（4）在景觀轉變過程，雖然政府主導的規劃很少得以徹底的施行，但是政府仍是景觀演變的主導者。借助 GIS 平臺，通過疊置規劃與現實因素，研究者可以評價規劃的績效。這種評價回饋對於研究者和決策者都頗有價值。

關鍵字：濱水區、歷史地理資訊系統、製圖、景觀史、南京

Historical GIS Approach for Mapping Transformation of the Waterfront Cityscape in Nanjing (from 1920s to 2000s)

Zhen Xu *, Ling-yun Han**

Abstract

Originated from riverside settlements in the ancient, many Chinese cities took shape in virtue of rivers and simultaneously shaped the waterfronts. Drastic transformations of riparian areas have taken place since the early modernization and industrialization of China, which have direct influence of the present townscape. Tracing the riparian landscape changes will provide insights to understand the correlation of rivers and cities, and decipher attitudes and roles of different agents in urban physiognomy, which are valuable for landscape management.

Located in the south-east part of China along the lower reaches of the Yangtze River, Nanjing is noted for prosperous culture and economy, 2480 years of urban history and hundreds years of being the capital. The shaping of urban areas are closely relevant to the rivers especially Qinhuai River which connects to the Yangtze River. Established in 1368, Ming Dynasty Capital city laid the framework of the present Nanjing city. With the largest ancient city wall, the Ming Capital city has 13 gates (including two water gates) and covered the urban area of 41 km^2. The city is featured by its harmonious combination of urban landscape with historical relics and natural elements including the rivers.

From historical-geographical perspective, the authors briefly portray the river changes of Nanjing with maps, archives and paintings etc., in the span of 1920s to 2000s.Taking the city and rivers history as studying background, the authors reconstruct the riparian landscape transformation of Nanjing in the 20th century from

* Associate Professor, College of Landscape Architecture, Nanjing Forestry University. email: xuzhen@njfu.edu.cn.
** Lecturer, School of City, Resource and Environment, Jiangsu Second Normal University. Email: landscaping@163.com.

the following aspects: (1) rectifying and overlaying the historical maps with GIS platform to trace the river morphology and function changes; (2) demonstrating the landscape changes along the main rivers by paints and photographs, analyzing the view corridors and viewsheds; (3) urban morphological analysis (town plan, land utilization and building fabric) of typical areas, and retrospection of relevant planning decisions.

The authors adopt GIS to rectify historical map and conduct overlay analysis, and find its potential to conduct landscape process research as follows:

(1) Researchers can easily rectify maps and overlay them to find the changes despite their different scale and orientation. Since digital cadastre and remote sense image are more available, we could broaden the sources of information channels and conduct large scale analysis for GIS are compatible with these data.

(2) It's convenient to identify elements by location or attribute, and demonstrate them graphically. After all, based on the plan unit, building fabric and land utilization, landscape regionalization can be conducted automatically or semi-automatically, which is especially suitable for large scale and preliminary analysis.

(3) Involving public welfare, nature preservation and utilization, river corridors should be researched in comprehensive way. GIS can be served as open platform for ecological, socioeconomic research that correlates with morphological analysis. With voluntary LBS data, more research opportunities will emerge.

(4) For landscape transformation, the government plays more important role than other agents meanwhile few of the government's plans implemented thoroughly. We can evaluate the plans' performance and the positive and negative factors by overlaying and comparing the plan and the reality in GIS. The evaluation feedbacks will provide valuable knowledge to researchers and decision-makers as well.

Keywords: waterfront, historical GIS, mapping, landscape history, Nanjing

數位人文與地方宗教研究：
以北港武德宮分靈的 GIS 時空分析與社會網絡為例

Digital Humanities and the Study of Local Communal Religion: Centering on the GIS Spatio-Temporal Analysis and Social Network Analysis of Branches of *Wude* Temple of Beigang

林敬智*

Ching-chih Lin*

摘　要

本文將以數位人文方法與技術如何能夠輔助地方宗教研究為核心問題意識，聚焦於地理資訊系統（Geographical Information System, GIS）和社會網絡分析（Social Network Analysis）兩種方法，以實際案例展示利用數位人文工具可以在既有傳統典範研究方法的基礎上，發揮大數據與視覺化的優勢，互補不足。本文將以北港武德宮五路財神信仰在台灣 40 多年的發展歷史為例，觀察其數千分靈在台灣、甚至海外的擴展，將其置於 GIS 時空地圖上，檢視在不同歷史發展階段中分靈傳播的空間特性，與其時之政治、經濟、社會、信仰的脈絡進行相互參照，進一步理解地方宗教信仰如何與社會經濟發展和更大的時代背景交互影響的過程。

GIS 時空地圖可以將武德宮各地分靈進行空間分佈的視覺化呈現，並比較不同時期的空間分佈之間的差異（參見圖 1 與圖 2），觀察其擴展的趨勢，顯與雲林嘉義地區人口外移至台灣各主要都會地區為主，在東部宜蘭、花蓮、台東等地則並非主要移入地區，另外從地圖上亦呈現出武德宮分靈初期並未進入桃竹苗等客家地區，應與族群、語言等因素有關，值得進一步探索。而圖 3 與圖 4 則呈現出 2014 與 2015 兩年中大型活動的交陪關係，可以觀察其向全省著名宮廟的交陪

* 政治大學宗教研究所助理教授、亞太時空資訊研究室秘書，Email: cclin52@gmail.com。

* Assistant Professor, Graduate Institute of Religious Studies/Asia-Pacific Spatio Temporal Institute, National Chengchi University. Email: cclin52@gmail.com.

關係亦逐步擴大，透過建醮與南巡活動建立全國性的知名度。

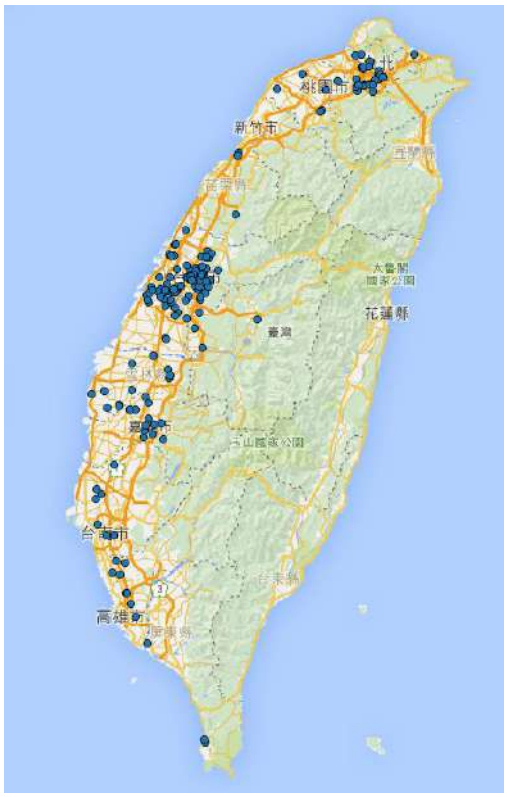


圖 1 北港武德宮早期分靈

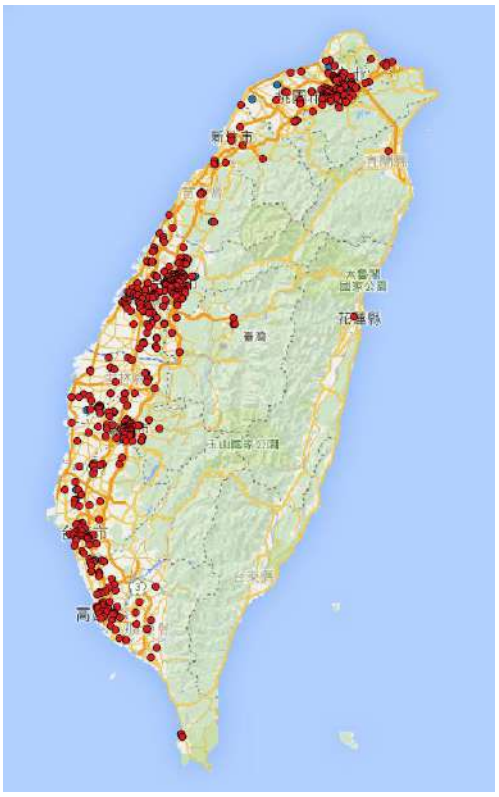


圖 2 北港武德宮近期擴張

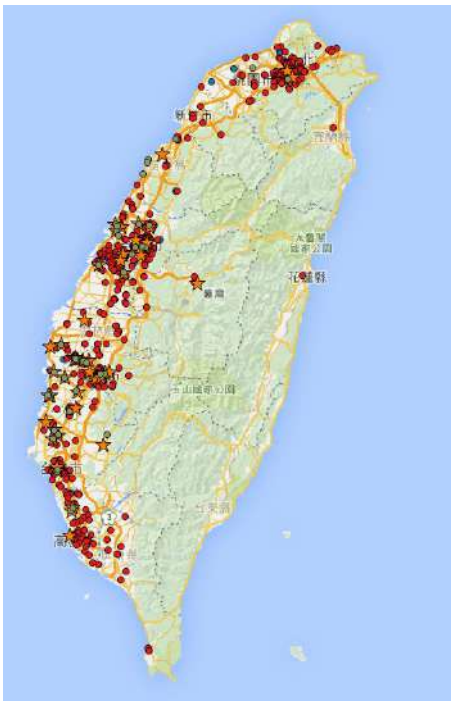


圖 3 武德宮分靈與 2014 建醮交陪

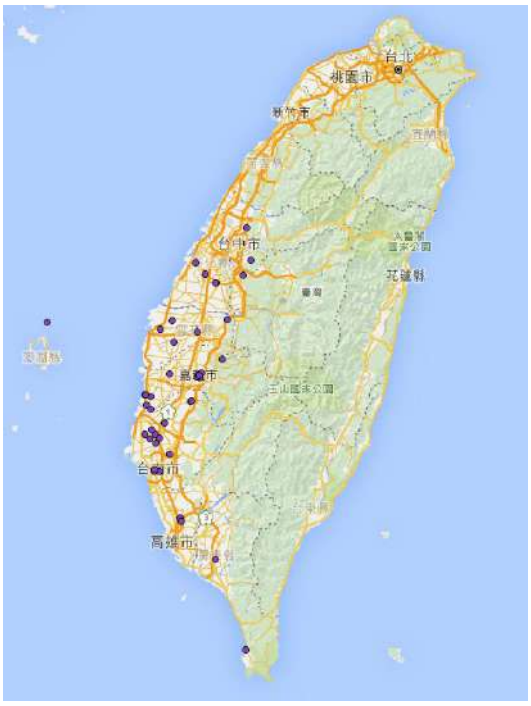
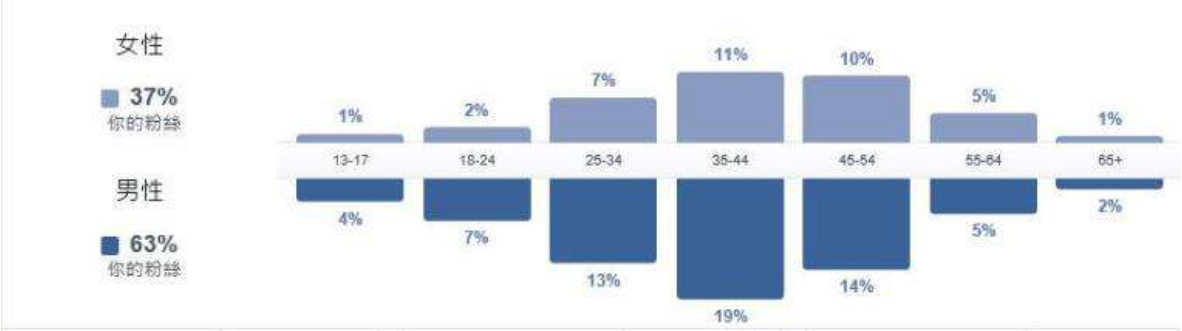


圖 4 武德宮 2015 南巡交陪

此外，武德宮於 2014 年「百年大醮」後，其各地分靈宮壇與信徒和主醮的吳政憲道長延陵道壇之間，建立的新社會網絡，特別是在網路上的社群媒體 Facebook 上，值得觀察當代道教與地方信仰之間如何透過網路與數位科技開拓虛擬與實際的社會空間與網絡，透過臉書粉絲資料之背景分析，將有助於瞭解信徒之性別、地域、國籍等背景，並利用社會網絡分析工具辨析粉絲之間所形成之次網絡結構（圖 5、圖 6）。截自 2016 年 10 月 14 日為止，北港武德宮的臉書網頁已有 300,600 人按讚，到訪人次到達 271,478 人；而其內部不公開的臉書社團「天官武財神信仰聯誼會」已有 4624 位成員，透過臉書上的社群網絡串聯信徒間的活動與資訊交流，並分享個人的靈驗經驗。從圖 3、圖 4 的粉絲分析可以觀察到信眾在台灣以外，以東南亞國家為大宗，包括馬來西亞、新加坡、香港、印尼、越南、泰國等地，而臉書所未包括的中國大陸地區，近期也開闢了微博網頁，並已開始拓展籌備在廣東、江西、南京、上海等地建立分靈，其未來的成長趨勢亦值得觀察。



國家/地區	你的粉絲	城市	你的粉絲	語言	你的粉絲
台灣	222,514	T	66,791	繁體中文（台灣）	221,862
馬來西亞	8,332	New Taipei City Xinbei	32,614	簡體中文（中國）	6,239
新加坡	1,455	Kaohsiung Kaohsiung	30,481	美式英語	4,592
香港	908	Taipei Taipei	16,568	繁體中文（香港）	2,219
泰國	659	Tainan Tainan	16,094	英語（英國）	1,162

New Taipei City Xinbei

圖 5 北港武德宮臉書粉絲背景分析 2015 年 12 月 20 日

圖 6 北港武德宮臉書粉絲背景分析 2016 年 7 月 20 日

北港武德宮草創於 1970 年代，四十多年來的發展已有數千分靈四散於全台各地，甚至擴展至東南亞、澳洲、美國、歐洲等地，財神信仰的傳佈與台灣經濟起飛有著十分密切的關聯，並與城鄉之間和跨國企業的經濟移民之流動具有高度的重疊性。晚近武德宮自身亦善於利用臉書與微博等網路社群媒體串聯四散各地的分靈信徒，還進一步吸引來自海外的信徒前來朝聖，全球化的發展過程值得關注。本研究亦將藉由 GIS 時空分析與社會網絡分析方法，剖析此一全台最大之天官武財神信仰如何貼緊時代脈動，利用臉書與微博等社群媒體推動其全球化與在地化的擴張，藉此觀察華人民間信仰的擴散現象與近代台灣經濟發展、城鄉和跨國移民之間的密切關係。

本研究還利用 GPS 全球衛星定位系統輔助紀錄神明遶境的路線，在 GIS 地圖上清楚呈現行經路線的城鄉差異，以及沿途交陪之宮廟與神明會等組織（圖 7）。2014 年的建醮活動與 2015 年的南巡活動更進一步促進武德宮在各地的分靈擴散，和並與各地重要宮廟之間建立聯盟與交陪關係。本文將透過 GIS 與社會網絡分析工具進一步梳理地方宗教交織的社會網絡關係，以及祖廟與分靈宮廟之間的

網絡、譜系、與階層化關係，進而與既有之相關研究典範進行對話，比對實證資料與既有理論之間的異同（圖 8）。

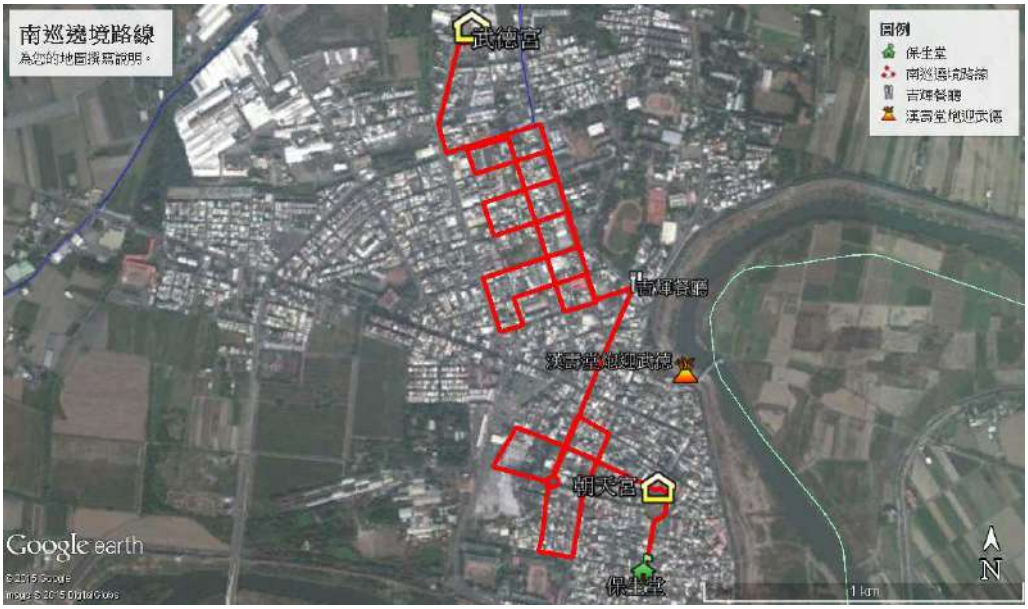


圖 7 武德宮 2014 年建醮南巡邊境北港鎮之路線圖

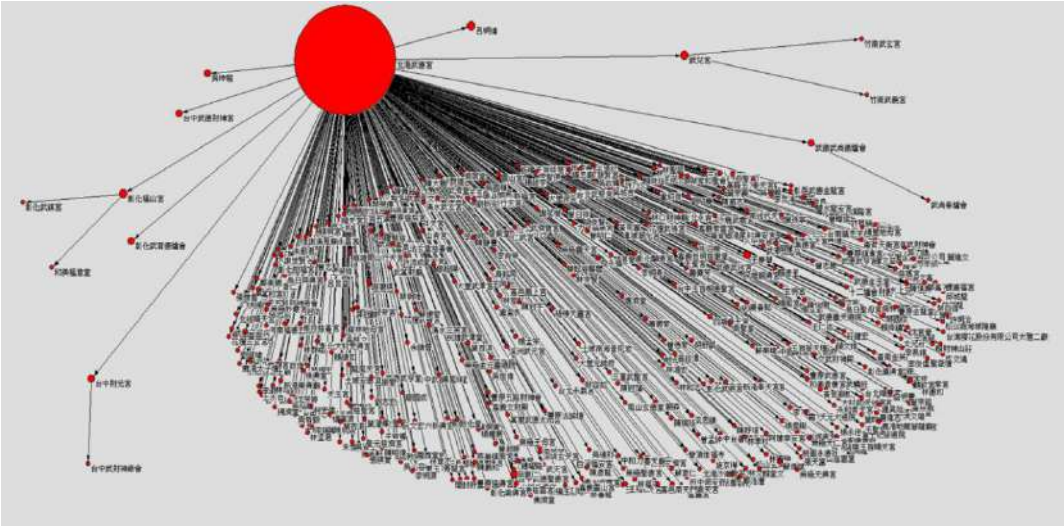


圖 8 武德宮分靈社會網絡三代

綜而言之，本文將以北港武德宮之分靈網絡和臉書粉絲網絡為例，進行 GIS 時空分析、社會網絡分析，利用數位人文的實證資料與視覺化輔助，進而與既有之研究成果進從比較與對話，一方面反思傳統研究方法之侷限，因一方面也檢討數位人文應用於實際案例所可能遭遇之限制，探索數位人文與傳統研究途徑之間如何相輔相成。

關鍵字：GIS 時空分析、社會網絡分析、華人民間信仰、祖廟與分靈、武德宮天官武財神

Simultaneous Invention or Propagation of Cultural Practices? Using Time-Distance Correlations for the Identification of Centers and Peripheries in the Transformation of Cultural Landscapes

Oliver Streiter*

Abstract

In this paper we are going to present a theory, according to which the *tanghao*[1] on the tombstones of Penghu (澎湖Pénghú) was invented on Xiyu (西嶼Xīyǔ) in order to replace the expression of loyalty to the Qing (皇清) as the tombstone topic (橫題) directly after the occupation of Penghu by the Japanese Imperial Army in 1895. From Xiyu, the *tanghao* spread to those islands of Penghu that also had used the loyalty expression to the Qing. On islands that had used under the Qing the *jiguan* (籍貫), a place name referring to a hometown in China, e.g. Wang'an, the *tanghao* was introduced later, when local carvers gave up their business and were replaced by more globally operating carvers. Finally, through the emigration of carvers and their students from Baisha (白沙Baíshǎ), Huxi and Makong to Taiwan in the Japaneses colonial period and in the early Republican period, the tradition of carving the *tanghao* arrived in Taiwan, where, promoted through the KMT, it became at the end of the 20th century the most frequent form of a focus on tombstones. Monte Carlo Sampling is used to

* Associate Professor, Department of Western Languages and Literature, National University of Kaohsiung; Zhengzhi Daxue Asia-Pacific Spatio Temporal Institute, French Center for Research on Contemporary China. Email: ostreiter@nuk.edu.tw.

[1] 堂號, a place name that servers as an identifier of a clan. On Taiwan, the most common tanghao is the primary tanghao, also referred to as Datanghao (大堂號) or Baijiaxing (百家姓 Baíjiāxìng) tanghao. The primary
tanghao is the place where the bearer of a Chinese surname has first been mentioned in literature. For the process
of sinicisation and the attachment of border communities through their new Chinese surnames to the cradle of the Chinese culture as an imagined homeland, the tanghao was traded in the canonical book Baijiaxing (百家姓),
which is usually translated as "The Hundred Family Surnames". Through this reading primer, people acquired literacy over the last eight centuries and learned, en passant, their tanghao (Taiwan Provincial Government, 1979;
楊緒賢 Yáng Xùxián (Yang Shi-hsien), 1979; Streiter & Goudin, 2013).

verify the origin of the *tanghao* in Penghu and Taiwan. The evaluation of randomly generated patterns of propagation through time-distance correlations confirms for Penghu and Taiwan the Xiyu Island as the only point of origin of this new carving practice. Trying to identify in addition intermediate centers from which the *tanghao* was propagated, the outcomes of purely statistical evaluation methods don't match the location of tombstone carvers who are known to have promoted the *tanghao* on Penghu and Taiwan. Enriching the algorithm with features that characterize carving tradition or carving schools will be necessary in the future to obtain patterns of propagation that can be reliably interpreted.

Keywords: Xiyu, Penghu, Taiwan, tombstones, tanghao, propagation, Monte Carlo sampling, time-distance correlation

「堂號」是同時發明或文化傳播的實踐？以時間距離的相關性識別中心與外緣在文化景觀中的轉變

Oliver Streiter*

摘　要

在本篇論文中我們將呈現一個理論，根據這個理論，進而探討澎湖西嶼發現的墓碑堂號。自 1895 年被日本皇軍佔領澎湖後，墓碑上的橫題轉變，取代原本表達對清國忠誠的「皇清」使用。過去從西嶼開始在澎湖小島間傳播的堂號使用，即是在日本殖民時期表達對清國的忠誠一種方式。在清國時期，墓碑的橫題最初是使用籍貫，表示對應到一個中國家鄉的地名，例如：萬安。而堂號是較晚才引進的使用方式，原先澎湖本地的墓碑雕刻師放棄了他們的職業並且被較有專業技術的墓碑雕刻師取代。並且在日本殖民時期與國民政府早期，透過雕刻師與其學徒們自白沙、湖西和馬公移居到台灣本島的遷移，雕刻堂號的傳統被傳入。加上後來國民黨提倡使用堂號，使得堂號變成 20 世紀末最常被使用在墓碑上的橫題。本論文以蒙地卡羅取樣法驗證堂號在澎湖與台灣的起源。以隨機排序的模式，從時間距離的相關性證實在澎湖列島和台灣本島中，西嶼是使用堂號作為雕刻實踐的唯一起點。當本研究試圖去識別堂號傳播的中心，發現純粹在統計的估算結果上，不吻合一般已知的當地墓碑雕刻師提倡堂號在澎湖列島與台灣本島的傳播路徑。本研究期許運用不同的變因且多樣的算法，呈現墓碑雕刻的傳統與雕刻師徒制下一個可信的傳播路徑模式。

關鍵字：西嶼、澎湖、台灣、墓碑、堂號、傳播、蒙地卡羅取樣法、時間距離相關性

* 國立高雄大學西洋語文學系副教授，Email: ostreiter@nuk.edu.tw。

1 Introduction

1.1The ThakBong Project

The research presented in this paper is part of a larger project, named *ThakBong*[2], a project in the Digital Humanities, which was started in 2007 with the intention to document and research funerary and epigraphic practices on Taiwan Streiter, Goudin, and Huang (2011). Throughout the nine years of its existence however, this research has extended its scope beyond Taiwan and reached Penghu and Jinmen, and, for comparison, Japan, China, and places where migrants from China have moved to. Our research equally develops techniques and interfaces for the storage, annotation, transcription and distribution of data and media for scientific and educational purposes. In addition, our research explores and develops analyses and visualization techniques to make data interpretable for a broader public.

The entire archive includes currently 238,994 images taken on 602 days of fieldwork on 814 graveyards, recording 62,768 tombs. On Taiwan only, 561 gravesites[3] have been documented, capturing and annotating 174,490 photos of 44,614 tombs.

1.2 From Taiwan to Penghu

Doing fieldwork on Taiwan and trying to understand the obtained data, we soon realized that an explicit spatial framing limited our ability to interpret the phenomena we could observe. We speculated that data on funerary and epigraphic practices from outside Taiwan might hold the key for the understanding of the observed variation of practices on Taiwan. In the process of this contextualization, Penghu was studied intensively. With the discovery of Qing period and Japanese period tombs, it became apparent that data from Penghu held potentially a key for the study of tombs on Taiwan. The interpretation of the Penghu data was facilitated by the natural fragmentation of Penghu into relatively homogeneous and autonomous islands. Exchange between these islands is restricted to specific channels, which, once identified, allow to understand some of the factors that transform cultural practices.

---

[2] The name ThakBong is derived from Taiwanese, where 讀墓 (thak-bong) means "*to study tombs*".

[3] We prefer for Taiwan and Penghu the term *burial ground*, meaning a landscape that hosts tombs, over the term *graveyard*, which might wrongly suggest an organized ensemble of tombs. Traditional burial ground in Taiwan are referred to as 亂髒 (luànzàng, unsystematic burial), 墓地 (mùdì, burial ground) or 墓阿埔 (bong-a-po) which all oppose the so-called normal graveyards 示範公墓 (shìfàn gōngmù). Only gravesites that have been set up in the last decades by local authorities or church communities follow a model of regularly spaced and regularly oriented tombs. Normal graveyards have been first introduced in Taiwan by the Japanese colonial authorities, but after 1945, most of these graveyards have lost again their regular structure.

Table 1: The number of tombstones collected in so far in the ThakBong project.

	Taiwan, Penghu and Jinmen	Penghu	Xiyu
Wanli -- 萬曆	2		
Chongzhen -- 崇禎	4		
Yongli -- 永曆	1		
Zheng-- 鄭氏王朝	59	2	1
Kangxi -- 康熙	6		
Yongzheng -- 雍正	1	1	
Qianlong -- 乾隆	55	12	5
Jiaqing -- 嘉慶	47	6	1
Daoguang -- 道光	147	27	6
Xianfeng -- 咸豐	59	9	2
Tongzhi -- 同治	81	14	4
Guangxu -- 光緒	172	98	59
Meiji -- 明治	342	28	15
Taishō -- 大正	663	51	38
Shōwa -- 昭和	2,518	249	118
Republican -- 民國	39,343	1,470	295

As a result, on Penghu, we undertook between 2010 and 2016 15 field-trips documenting 65 burial grounds on eleven islands through 14,871 photos of 3,192 tombs. Although the numbers of Penghu seem to compare disfavorably to those of Taiwan, we estimate to have covered more than 60% of all existing tombs, which is a much larger percentage than on Taiwan. In addition, Penghu has a much larger proportion of older tombs. Tombs of the Shōwa period for example are relatively seen four times more frequent on Penghu than on Taiwan, see Table 1.

2. The Development of Tombstone Inscriptions on Penghu

2.1 Penghu under the Ming

After the last evacuation of Penghu under the military ban in the 15th century, fishing communities were re-settled under the Ming towards the end of the 16th century and military forces were stationed on Penghu from 1603 on. In 1622 the Dutch Vereenigde Oost-Indische Compagnie (VOC) (United Dutch East India Company) occupied Penghu but eventually had to leave in 1924 and moved to Taiwan. The VOC was ousted from Taiwan in 1661 by the Ming-loyalist Koxinga (鄭成功 Zheng Chenggong), triggering in China another maritime ban under the Qing and subsequently a migration wave to Penghu between 1662 and 1664. After the battle of Penghu in 1683, the Qing conquered Penghu and shortly later also Taiwan.

Having experienced maritime ban, relocation, foreign occupation, migration, the loss of the emperor, local people on Penghu were aware that their presence on the islands would depend on the government and its ability to stay in power. Especially during the period of the Zheng regime, the time most known Ming tombstones fall into, a considerable part of the population were Ming loyal soldiers, prepared to face the Qing that were gathering their strength at the Chinese coast. Loyalty and support of the current government, whatever it was, might have been the tactics of most inhabitants to avoid future uncertainties. This support found its reflection, on Penghu and Taiwan, through a focus position on tombstones that expressed the loyalty to the Ming as 皇明 (huángmíng), or, shorter, as 明 (míng). Two known Ming tombstones that we can associate with an island of the Penghu archipelago are on Xiyu, where the West fort had been set up under Koxinga and where potentially larger amounts of soldiers were located. A third known Ming tombstone currently cannot be associated with a specific island.

The tombs under the Zheng regime, however still have another, more subtle feature. Many Ming period tombstones show a date inscribed on the tombstone, expressed as Chinese lunar year, an unchanged circle of sixty metaphysical names. Most tombstones under the Zheng regime however don't use a second component of a date, the era name (年號, niánhào), which is the name the emperor gave to his reign, obligatory to mark the distinction between e.g. the year 1681, 1741 or 1801. The historical context of this omission under the Zheng regime is well known: The last Southern Ming emperor Zhu Youlang (朱由榔, Zhū Yoúláng), who had given the name *Yongli* (永歷, Yǒnglì) to his era, had been executed by the Qing in 1662, the same year that the Ming loyalists under Koxinga took over Taiwan from the Dutch. After 1962, the Era Yongli (永歷) continued to be used on official documents and inscriptions of the Zheng era. More than 95% of all known Ming tombstones however don't mention the era, causing much debate among historians about the dating of these tombstones. Nevertheless, it became a common understanding that a tombstone with a loyalty expression referring to the Ming and a lunar year without an era that could fall into the period from 1662 to 1683 is to be considered as a Ming tombstone of the Zheng era, see 朱鋒 Zhu Feng (Zhu Feng) (1953; 林衡道 Lín Héngdǎo (Lin Hengdao) 1969; 石萬壽 ShíWànshoù (Shi Wanshou) 1975).

2.2 Penghu under the Qing

Throughout all dynasties, the omission of the era name in combination with a Chinese lunar year is a systematic reaction of local people to the change of a dynasty. Best documented

is this omission where we have rich data, e.g. in the early Japanese colonial period. This omission, very likely an implicit expression of loyalty to the Qing government, was common in the Meiji (明治) era. Inscriptions of the Japanese era were more systematic in the Taishō (大正) era and were barely absent in the Shōwa (昭和) era. After WWII, we can also observe the omission of the republican era name (民國), although less systematically then during the Japanese colonial period. The shift from Ming to Qing is documented only with a few tombs, but it is very well possible, that people under the early Qing might have also omitted a Qing era name for some time on Taiwan. This is suggested at least by one of the earliest Qing tombs on Taiwan, the grave of the Zhongzhou Chen (中州陳, Zhōngzhoū Chén) in Beimen, Tainan.

The tomb of the wife of 陳興桂 (Chén Xíngguì), 陳氏鄭 (Chén Shì Zhèng), is located in the south of the Beimen village in Tainan, Taiwan. Mrs Chen had died under the Ming in 1661, but her tomb was set up in 1698, seventeen years after her death, under the Qing. The tombstone has two remarkable features. First, the era name has been omitted, no longer, as under the Zheng regime, in the absence of a living emperor, but, most likely, as an implicit expression of loyalty with the preceding Ming dynasty. Second, the explicit loyalty expression on top of the tombstone has been replaced by the place name 浯江. This is one of the first placenames to be found on Qing tombstones on Penghu and Taiwan. The placename, as a choice of focus, thus originated possibly out of a discomfort with the new colonial government. Placename, which offer a perfect excuse for not using a loyalty expression, developed into the most common focus in Taiwan, outnumbering the expression of loyalty to the Qing by the factor 15.

This replacement on Taiwan of the loyalty expression by a place name referring to China, the so-called *jiguan* (籍貫 jíguàn) contrasts with the shift that took place on some islands of Penghu at the same time. Here, the expression of loyalty, once common during the Ming era, was continued in a modified form as 皇清 (huángqīng). Although the dissatisfaction with the Qing government and the abuse by Qing soldiers might have been similar in Tainan and on Penghu, the degree of anxiety might have been different. While in Tainan, people could escape from the influence of the Qing authorities by moving north, south or east, on Penghu there was no such escape.

Unfortunately, we cannot reconstruct the transition of tombstones on Penghu from Ming to Qing in detail, as the earliest preserved Qing tombstone dates from 1733, which is 50 years after the establishment of the Qing on Penghu. This tombstone can be found on the burial site of Dong'an on Wang'an, a few hundred meters from the harbor, with a loyalty expression to

the Qing carved into granite. Granite tombstones can be usually found where boats where unloaded and thus indicate a direct shipping link to China. The majority of loyalty expressions however can be found not on Wang'an, but on Xiyu, where the Qing took over the West Fort from Koxinga and added the East Fort in 1883. Shipping links to China, as does the vicinity of forts, underline the opportunistic nature of the loyalty expressions. It's main function might have been to guarantee good relations with the authorities. But as the shift from granite to local materials for tombstones suggests, these shipping links might have become less important on Wang'an after 1735.

With the maritime connections between Fujian and Wang'an weakening and the center of the Qing government on Penghu far away, Wang'an finally adopted the *jiguan* as its principal focus on tombstones about 1750, while the loyalty expression continued to be used on Xiyu. Makong acquired the central position it has today towards the end of the Qing period, due to the outstanding protection its harbor offered to larger vessels which could not be pulled ashore. This rise of power of Makong under the Qing reflects on tombstones through a shift from the *jiguan* to the loyalty expression about 1820, a period in which the Qing started to invest in Makong. Smaller islands without a carver were served from nearby islands. As a consequence, the development of the focus position on Tongpan, close to Xiyu, resembles that Xiyu. Likewise, the development on Dongji, closer to Wang'an than to Xiyu, resembles partially that of Wang'an.

2.3 Penghu under the Japanese

When the Japanese arrived on Penghu in 1895, two distinct epigraphic practices were thus in usage, the loyalty expression and the *jiguan*, depending on the vicinity to Qing executive forces. However, those carvers and families that had traditionally inscribed an expression devoted to the Qing must have experienced a crisis, as this kind of inscription was no longer opportune. Having witnessed the Japanese occupation of Penghu by force and being abused regularly by the Japanese forces, local people understood the risk of continuing this practice.

In Makong carvers could reanimate the practice of the *jiguan*, which had fallen out of usage in Makong only 70 years earlier. In addition, through the quality of the stone and the relative protection from winds, tombstones in Makong retained their readability for more than 80 years, so that even forgotten *jiguan* could be recovered.

On Xiyu, in contrast, the solution to the question of how to carve the focus of the tombstone could not follow any historical example as wind and weather had eroded the visibility into the

past. On Xiyu, however, the loyalty expression had been used almost exclusively for 150 years, eroding the memory of the *jiguan* in those families that in 1895 had no *zupu*[4]. The alternative, to continue with the same kind of focus position, e.g.an opportunistic expression of loyalty to the Japanese emperor (皇日, huángrì), might not have been a strategic option as Penghu might have returned some years later to the Qing, which after their return would not have been amused by such a welcoming attitude towards the Japanese. It was thus in Xiyu where, within less than two years, a very important epigraphic invention took place.

The vacant position left by the loyalty expression was finally filled by professional carvers in Xiyu with a new element, the *tanghao*, more specifically the *Baijiaxing tanghao*, which through the ambiguity and flexibility of this notion as well as the increasing market share of those carvers who applied the *tanghao* gradually spread over the islands of Penghu.

After the WWII, the practice of the *tanghao* expanded to all corners of the archipelago, even to the small island of Jiangjun'ao, where before the arrival of the *tanghao*, tombstones had no inscription at all. On some islands the *tanghao* thus arrived for the first time in the 1960s (Jiangjun'ao), 1970s (Niaoyu) or 1980s (Hujing). This massive application of the *tanghao* is largely the result to the import of tombstones from two carvers, one in Baisha and one in Makong, who especially towards the end of 2000 assumed an absolute monopoly on the archipelago. What unites these carvers is that the carver in Baisha and the father of the carver in Makong have learned with a Master from Penghu in Yanshui, Tainan, returning then back to Penghu. This Penghu carver working in Yanshui however had additional students which might have continued the carving tradition that originated on Xiyu on Taiwan.

3. From Penghu to Taiwan

3.1 Migrating Carvers

As the result of interviews held with tombstone carvers on Penghu and Taiwan we could reconstruct fragments of the migration of carvers and their disciples from Penghu to Taiwan in the Japanese colonial period and early Republican period. Through these accounts we can identify Penghu tombstone carvers with an almost identical conception of how stones are to be

[4] 族譜, a genealogy of the male line, usually written by professional writers for a clan. *Zupu* have been frequently copied, e.g. when the quality of the paper started to degrade or when a clan split up and on these occasions probably rectified and embellished.

measured and carved in three regions of Taiwan: 1)Yanshui, Tainan, Japanese colonial period 2) Gangshan, Kaohsiung, early Republican period and 3) Fangliao (Pingdong) from 1970 on. The carver in Fangliao is still practicing and his carving can be unambiguously identified with the practices found on Penghu.

Future research will thus attempt to enrich our knowledge on the migration of tombstone carvers, their carving techniques, their apprentices and where these move to. At the same time these initial data on carvers' migration from Penghu suggest that the origin and the spreading of the *tanghao* on Taiwan might be a result of this migration of carvers and the continuation of the carving practice that had developed on Xiyu under very specific conditions. If support for this theory can be found, this might be particularly important as currently there is no alternative hypothesis of how the *tanghao* might have sprung up on Taiwan independently from Penghu at about the same time. Aa there is no conceptual relation between the *tanghao* and the political period of its emergence, it is not the period as such that might explain the propagation of the *tanghao*. Instead we believe that it is the encounter of tombstone carvers from Penghu and their strong belief in the *tanghao* with the rising number of sinicized Indigenous and Han who had lost their zupu or forgotten their *jiguan* that facilitated the acceptance of the *tanghao* on Taiwan.

3.2 Modeling the Spreading of the Tanghao

To test the hypothesis of a single origin, we intend to apply an algorithm to our data that models the propagation of the *tanghao* with potentially more than one points of origin. Using Monte Carlo sampling, we randomly generate patterns of how the *tanghao* might have spread from one burial ground to the next Silva et al. (2015). Using a non-deterministic approach to the establishment of intermediate hypothesis, we seek to reduce the number of assumptions we have to make that might influence the final outcome. Yet we consider the following four assumptions to be necessary in order to produce an interpretable model: a) Each burial ground is connected through an incoming and/or outgoing edge to another burial ground, b) a burial ground that we consider a center of the promotion and propagation of the *tanghao* must have more than one outgoing edge, c) a burial ground may only connect through an outgoing edge to another burial ground if the first occurrence of a *tanghao* on the second burial ground falls between the first and the last occurrence of a *tanghao* on the first burial ground and d) only tombs before 1980 are considered for this sampling, as after 1979, through the promotion of the *tanghao* through the KMT, the *tanghao* no longer spread locally Streiter and Goudin (2013) but through the print media. If *n* is the number of burial grounds, the intermediate hypotheses

generated through sampling have between 1 and *n*/3 points of origin, i.e. burial grounds without an incoming edge. The larger the number of burial grounds, thus, the smaller the percentage of intermediate hypotheses created that have only one point of origin.

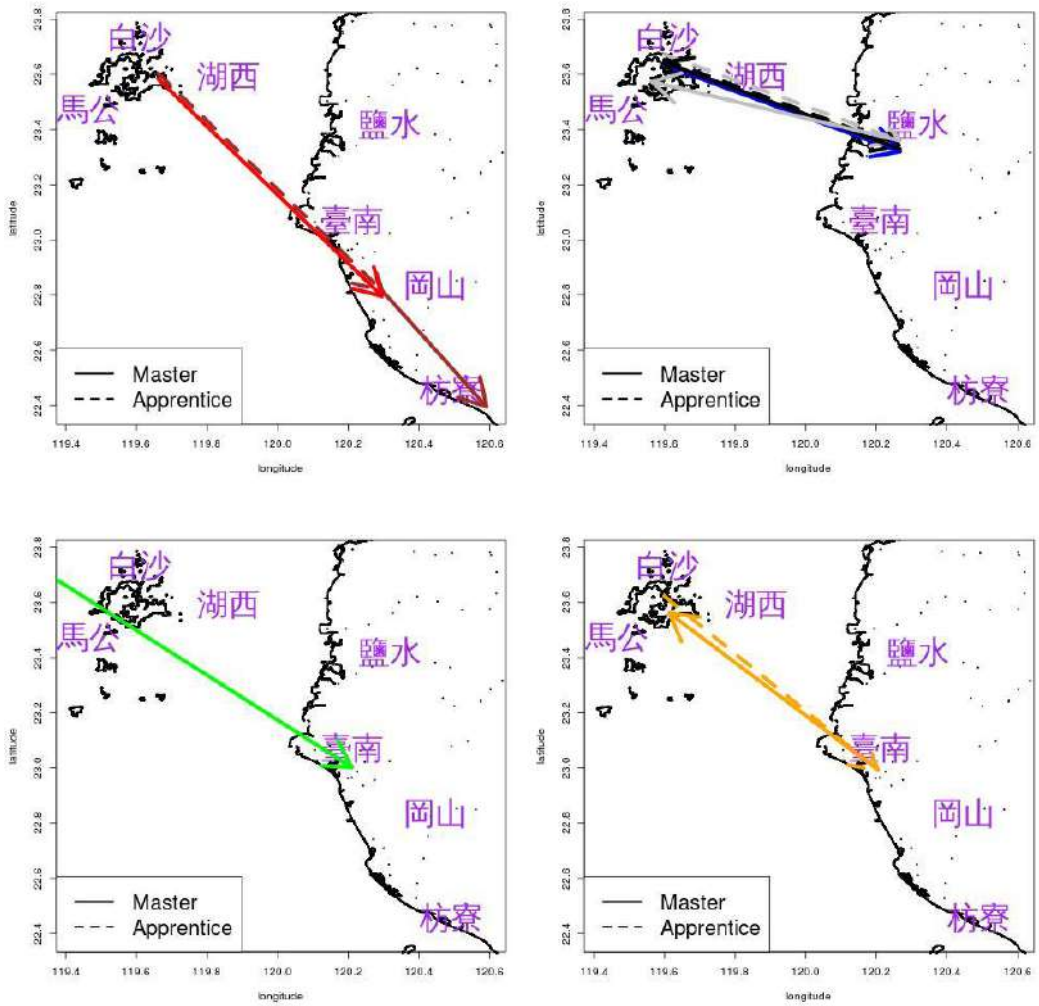


Figure 1: Four schools of carvers on Taiwan, two of which, in Yanshui, Gangshan and Fangliao originate from Penghu. Top left: Student from Penghu learns in Tainan and returns to Penghu. Bottom left: A carver moves from Huxi to Ganshan and takes a student from Huxi, who later works in Fangliao. Top right: A carver migrates from Baisha to Yanshui. Two of his student from Penghu open later shops in Baisha and Magong. The shop in Magong is now in its second generation. Bottom left: A student from Penghu learns with an autodidact Master in Tainan and returns to Magong.

One edge thus connects the geographic references of two burial grounds and the temporal indications of the first occurrence of a *tanghao* on each of these burial grounds. From this are derived a geographic distance and a temporal distance between the burial grounds. On the basis of a time-distance correlation over the entire set of edges, we can identify the points of origin

in all data. This is exemplified through Model 1, which selects the hypothesis with the smallest p-value of the correlation of the geographic and temporal distance.

As Figure 2 shows, this model identifies, as expected through our fieldwork, Xiyu as the point of origin of the *tanghao* on Penghu. This model also identifies reasonable intermediate centers. However, this model creates unwanted particularly long edges, as the property of a correlation is fulfilled equally with long and short distances. However, long distances over long time periods with no intermediate centers are either unlikely to occur or difficult to explain.

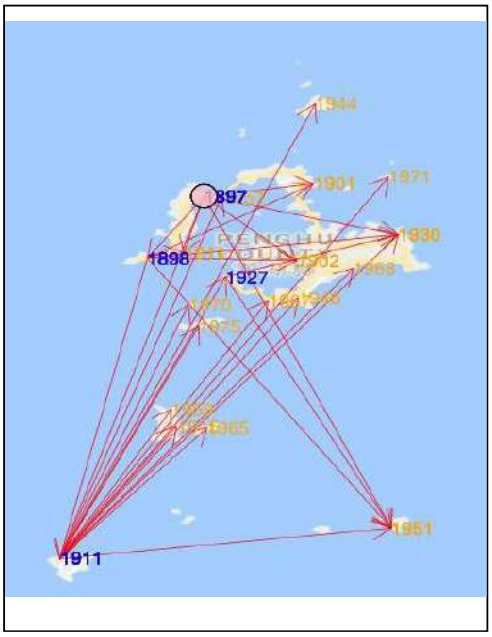


Figure 2: Model 1 for Penghu: Minimizing the p-value of the correlation.

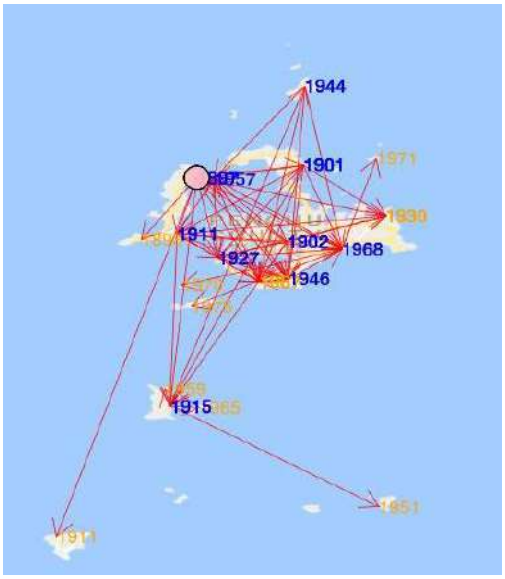


Figure 3: Model 2 for Penghu: Limiting the p-value of the correlation to 0.05, and minimizing the average spatial distance of burial grounds.

To remedy the weakness of Model 1, we apply the following changes. Of all hypotheses generated, we retain only hypotheses with a correlation p-value below the significance level of 0.05 and select the hypothesis that has a minimal average spatial distance.

Figure 3 shows that Model 2 confirms the unique point of origin in Xiyu and shows a reasonable development of the *tanghao* from a center around Xiyu, Baisha and Magong (馬公 Mǎgōng)[5] to the periphery. However the smooth development of the *tanghao* is achieved at the cost of multiplying the number of intermediate centers, which are probably not associated with

[5] 媽公, capital and main embarkation point of Penghu County. Erected in the 1880s, it became an important base for the Japanese imperial navy in WWII. Before 1920, it was known as 媽宮澳 (Magong'ao), then the Japanese authorities changed its name to 媽公 (Magong).

a carver or a carver tradition. As Model 2 has no qualitative data about carving traditions, beyond the carving of a *tanghao,* we see no point in augmenting the model at this stage statistically, before adding qualitative constraints, requiring a certain consistency in the carving tradition on both sides of an edge.

4.Testing and Results

The algorithm of the spreading of a feature on tombstones has been developed and tested with the Penghu data. This had the advantage that the research area is relatively limited and meaningless outcomes of the algorithm can be easily identified, given our knowledge of the field. In a second step the algorithm was in an unmodified form applied to the subset of the data that includes Penghu and Taiwan. The two main questions we would like to be answered by this algorithm are whether it would identify one or more points of origin and, second, whether the algorithm would identify as intermediate centers regions that we associated with Penghu carvers on Taiwan. The outcome of this application of this algorithm to the data of Penghu and Taiwan is shown in Figure 4. It confirms Xiyu and Penghu as the source of the *tanghao* on tombstones on Taiwan. However, the intermediate centers that are generated between Penghu and the points of latest arrival do not match the geo-references that we have associated with Penghu carvers that have promoted the *tanghao* on Taiwan. We hypothesize again that providing the algorithm with qualitative data on carving tradition, such as character variants, tombstone width and number of characters, will provide a refined model of how the *tanghao* spread from Penghu to Taiwan.

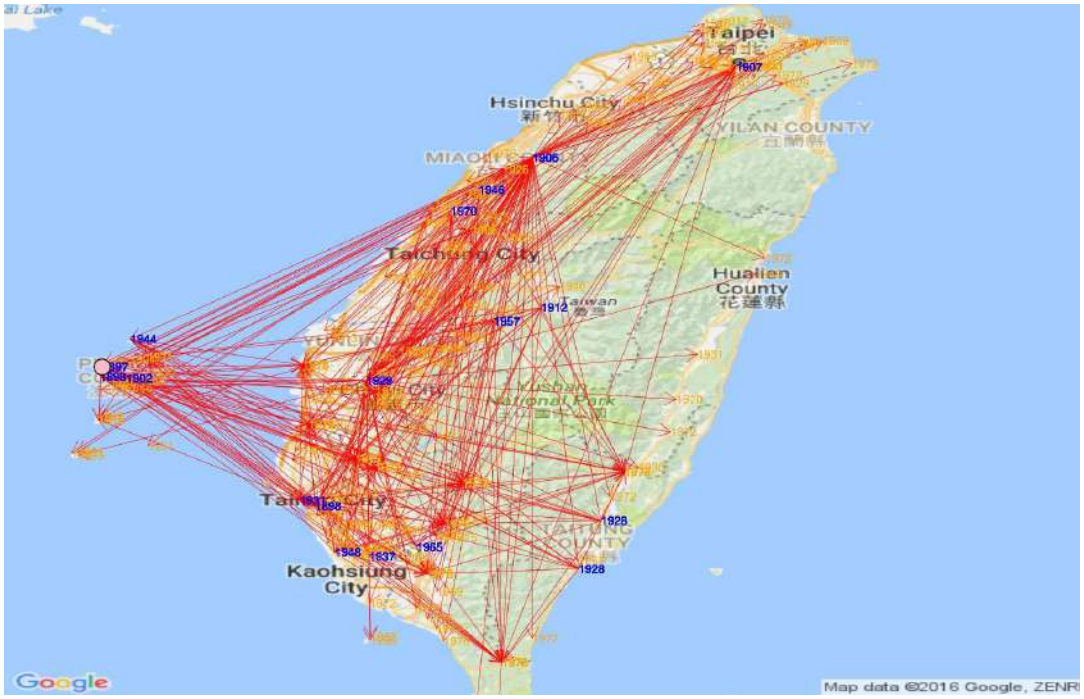


Figure 4: Model 2 for Penghu and Taiwan: Limiting the p-value of the correlation identifies a point of origin to 0.05, and minimizing the average spatial distance of burial grounds.

5.Conclusion

In this research we have shown that the *tanghao* as focus on tombstones has been invented most probably by a tombstone carver on Xiyu, Penghu under three very specific conditions. These are 1) the extensive usage of expression of loyalty to the Qing 皇清 (huang qing) on tombstones under the strong military presence of the Qing on Xiyu, 2) the erosion of memories of the jiguan, the ancestral place in Fujian, in those families that a) did not have a zupu and b) hadn't used the jiguan on tombstones for more than a 150 years and 3) the arrival of the Japanese forces on Penghu, an event which forced local people to replace the loyalty expression 皇清 by something less offensive to the Japanese authorities and more integrative with respect to the construction of a new in-group identity under a colonial administration. The implicit reference in the *tanghao* to the *Baijiaxing* and behind it, the division of China under the Song and its later unification, might have served as shared narrative for an in-group identity.

Through the migration of carvers from Penghu to Taiwan, we hypothesize, the *tanghao* as epigraphic practice arrived on Taiwan, where it filled a gap for those families that never had a *jiguan* or that had through the centuries forgotten their *jiguan*. Our modelling of the propagation of the *tanghao* confirms Xiyu as the point of origin for Penghu and Penghu as the

point of origin for Taiwan. However, the purely statistical nature of the modelling, without the ability of the algorithm to access the quality and similarity of carving traditions cannot provide a convincing model of how precisely the *tanghao* propagated from its point of origin to its periphery, mostly on the east coast and the very south of Taiwan.

In the followup of this research we will identify the dimensions along which carving traditions can be compared, fill existing gaps in the annotation and experiment with similarity metrics that can be included into the model in a robust way, so that incomplete data, such as partially unreadable transcriptions, can be included as much as possible in the simulation experiment.

Having determined in a first step where and when the *tanghao* emerged, the second step must be how it propagated through time and space. Only then we will be in state to provide more subtle hypotheses of why this development took place.

References

Silva, Fabio, Chris J. Stevens, Alison Weisskopf, Cristina Castillo, Ling Qin, Andrew Bevan, and Dorian Q. Fuller. (2015). "Modelling the Geographical Origin of Rice Cultivation in Asia Using the Rice Archaeological Database." *PLoS ONE* 10 (0).

Streiter, Oliver, and Yoann Goudin. (2013). "The Tanghao on Taiwan's Tombstones. The Recuperation of Tactics for a National Space." *Archivi Orientalni* 81 (3). John Benjamins Publishing Company: 459–94.

Streiter, Oliver, Yoann Goudin, and Chun (Jimmy) Huang. (2011). "ThakBong, Digitizing Taiwan's Tombstones for Teaching, Research and Documentation." In *TELDAP 2010 - The International Conference on Taiwan e-Learning and Digital Archives Program*, 146–57. Taipei, Taiwan: TELDAP agency.

朱鋒 Zhu Feng (Zhu Feng)。1953。〈臺灣的明墓雜考 Taíwān de míngmù zákǎo (On Ming tombs in Taiwan)〉，*臺灣文化* 3（2）： 480–90。

林衡道 Lín Héngdǎo (Lin Hengdao)。1969。〈臺灣現存的名墓 Taíwān xiàn cún de míng mù (Famous tombs preserved in Taiwan) 〉，*臺灣文化* 19（3-4）： 54–55。

石萬壽 Shí Wànshoù (Shi Wanshou)。1975。〈記新出土的明墓碑 Jì xīn chūtǔ de míng mùbeī (On recently unearthed Ming period tombstones〉， *Taiwan Historica* 26（1）： 37–47。

文本自動標註與事件擷取技術於漢籍全文資料之時空資訊加值應用

白璧玲[*]、賴郁婷[**]、黃惠敏[***]、吳承翰[****]、蔡宗翰[*****]、范毅軍[******]

摘　要

本研究整合文本自動標註、事件分群、與地理資訊等技術，以編年體之《明實錄》為分析標的，實踐以衛所事件為中心之數位人文研究。本研究的主要目標為：探索衛所相關的事件發展，包含將段落依事件類型分類，並依照各段落相關之人物、地點、時間，對各事件進行抽絲剝繭的分析。本系統利用事件分群技術，辨識出《明實錄》中各衛所相關段落所描述之事件類型，例如朝貢、升遷、軍事衝突等；透過自然語言處理技術，辨識出各段落中所提之時間、人名、地名、衛所名稱、官名。上述兩項自動辨識的少量錯誤均可透過編輯介面進行人工修正。此外，本系統亦結合其他服務，提供數位人文研究所需之必要分析工具，包含：(1)結合 CCTS-API 地圖服務來定位衛所位置，結合文本與空間；(2)結合時間軸元件將事件呈現於縮放年表上，以供觀察事件的發展趨勢；(3)結合統計圖表元件，展現各朝個別事件類型之發生頻率。最後，本研究設定江西諸衛所為標的，以本工具分析衛所設置與運糧等事件相關數據，使該等事件的脈絡，有更客觀具體的呈現。

關鍵字：漢籍電子文獻資料庫、時空資訊系統、自然語言處理、文本自動標註、事件擷取

[*] 中央研究院人文社會科學研究中心地理資訊科學研究專題中心博士後研究員，電話: (02)27857108#103 email: lingpai@gate.sinica.edu.tw。

[**] 國立中央大學資訊工程研究所碩士研究生。

[***] 中央研究院歷史語言研究所計畫助理。

[****] 國立中央大學資訊工程研究所助理。

[*****] 國立中央大學資訊工程研究所教授、中央研究院人文社會科學研究中心地理資訊科學研究專題中心副研究員，電話: (03)4227151#35203 email: thtsai@csie.ncu.edu.tw。

[******] 中央研究院歷史語言研究所研究員、人文社會科學研究中心地理資訊科學研究專題中心執行長，電話: 27829555#184 email: mhfanbbc@ccvax.sinica.edu.tw。

Automatic Text Markup and Event Extraction Technology Integrated with Spatial-Temporal Information Infrastructure for the Value-added Application of Scripta Sinica Database

Pi-ling Pai[*], Yu-ting Lai[**], Hui-ming Huang[***],

Cheng-han Wu[****], Richard Tzong-han Tsai[*****], I-chun Fan[******]

Abstract

This paper proposes a digital humanity research method targeting wei-so events by exploiting automatic text labeling and event clustering to analyze *Ming Shilu* and linking wei-sos and locations to the map. The main goal of this paper is to investigate the evolvement of events by (1) classifying paragraphs into dozens of event types and (2) studying who were involved in the event and when/where the event happened.

To accomplish the first goal, for each paragraph, we employs event clustering to recognize event types mentioned in that paragraph, such as tribute, promotion, or military conflict. For the second goal, we use natural language processing to recognize named entities including temporal expressions as well as names of persons, locations, and wei-sos. Recognition errors can be corrected manually via the edit interface. In addition, we offer a web-based analytics tool by integrating our system with three web services/components. (1) By querying the CCTS-API service, wei-so mentions can be located and the texts are linked to the map. (2) By utilizing the timeline component, events can be displayed chronically for observing the

[*] Postdoctoral Fellow, Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica. TEL: (02)27857108#103 email: lingpai@gate.sinica.edu.tw.

[**] Graduate Student, Department of Computer Science & Information Engineering, National Central University.

[***] Assistant, Institute of History and Philology, Academia Sinica.

[****] Assistant, Department of Computer Science & Information Engineering, National Central University.

[*****] Professor, Department of Computer Science & Information Engineering, National Central University; Associate Research Fellow, Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica. TEL: (03)4227151#35203 email: thtsai@csie.ncu.edu.tw.

[******] Research Fellow, Institute of History and Philology, Academia Sinica; Executive Director, Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica. TEL: 27829555#184 email: mhfanbbc@ccvax.sinica.edu.tw.

evolvement of the specified event types. (3) By using bar and pie charts, the count of every specified event type in every emperor’s reign period can be visualized in an organized fashion.

Finally, we conduct a case study targeting the Chiang-hsi guards by using the proposed tool to obtain statistics regarding establishment and grain transportation event types, making that the context of that event types present in a more objective and concrete way.

Keywords: Scripta Sinica database, spatial-temporal information infrastructure, natural language processing, automatic text markup, event extraction

數位人文研究著眼於透過資訊技術來處理數位化的資料，藉以改變人文研究方法，不僅能夠對大量資料進行有效率、有系統地分析，更在於拓展人文研究視野、轉變知識傳遞方式。以史語所長期以來維運的漢籍電子文獻資料庫而言，緣起於史籍自動化的理念，持續收錄中國傳統人文研究相關史料，進行數位化處理，並透過網頁介面提供全文檢索服務，篩選關鍵字所在的文本章節段落，便於從中查閱所需內容，然隨著新技術與應用理念的拓展，進一步對漢籍電子文獻資料庫進行加值應用、朝數位人文理念發展之可能性增加。

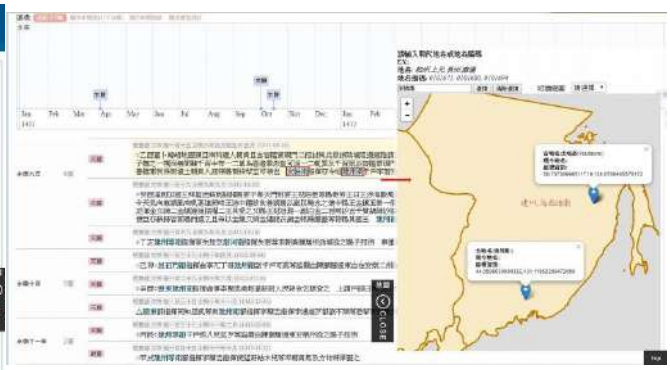
本研究嘗試運用文本自動標註與事件擷取方法，結合時空資訊技術，對漢籍電子文獻進行實作，並以實際歷史研究課題的發掘為導向，來探討此應用模式於數位人文研究的可行性。初步設定以《明實錄》之編年體史料為實驗文本，並以明代衛所作為地名識別之類型來測試。衛所制度是明代特有的軍事制度，於各地設立都司衛所，作為地方控制、邊區防守的單位，其興革與轄屬基本上涉及政治軍事形勢的發展，因此，透過對文本所述都司衛所名稱進行標註，就事件類別的時間排比，並且與衛所分布地圖作一整合，可一覽政軍局勢變動下的衛所任務特性，亦可探討區域內衛所組織的運作方式，以及空間結構的特性。此構想奠基於標註參照的「中華文明之時空基礎架構」（Chinese Civilization in Time and Space，CCTS）衛所沿革資料庫與 CCTS-API 之地名時空對位機制的建置，並預期藉由文本標註與事件類型識別方法的測試，以及圖文整合檢索系統的設計，逐步加以實現。

本研究建立的衛所名稱標註系統，主要分為兩個主要層面，首先在於衛所候選詞的辨識。鑑於《明實錄》中的衛所名詞構詞有一定的規則，我們採用半自動生成之構詞模板來擷取文本中的衛所候選詞。此方法類似規則式（Regular Expression）方法，但字串與模板匹配時允許插入、刪除與取代，故具有較大的彈性；此方法對於提升自動分詞的準確率，並有效擷取候選詞的可行性，已見於相關研究（Chang et al., 2015）。對於辨識出的衛所候選詞，我們計算其與 CCTS 衛所沿革表中各衛所正規名稱的最小編輯距離（Minimum Edit Distance，MED），選擇 MED 最小的衛所正規名稱，又經由衛所名稱模板分析，得知後綴詞（衛、所、府或司）為鏈結衛所的必要條件，因而再結合後綴詞擷取結果來對應。此外，衛所隨時間而多有變革，則須考量該衛所名稱出現在《明實錄》文本段落之紀日資料，經由中西曆轉換與時間區間計算後，將衛所沿革表中符合條件的編碼附加至標註結果。

針對上述標註詞彙之辨識效果，我們抽選《明實錄》衛所標註結果中 10%的資料量進行人工檢驗，評估其準確率與召回率，分別為 84.78%與 91.34%。為區別《明實錄》章節段落敘事之性質，進一步就文本之衛所標註結果，以非監督分群與分類方法來進行

事件分析，初步可自動辨識為 60 個事件類型。對於衛所自動標註與事件分類結果，我們建立了「明實錄衛所事件檢索系統」，提供研究者在此網頁介面上進行事件類型編輯，並輸入欲查詢之衛所與事件類型，搜尋相關文本段落內容。藉由網頁檢索功能設計，不僅可篩選衛所事件相關文本段落進行排比與解讀，並可運用時間軸與統計圖表展示功能，概覽事件類型之歷時分布特性（見圖一），同時，衛所標註所賦予的位置編碼，透過 CCTS-API、結合地圖介面來呈現事件所述相關衛所的空間位置分布，達到文本可視化的效果（見圖二）。凡此皆利於探索衛所相關研究議題。

圖一 事件檢索網頁功能範例



圖二 事件檢索結果與地圖整合範例

本研究成果實際運用於歷史研究課題的效用，鑑於中央研究院歷史語言研究所于志嘉先生對於明代江西地區衛所有諸多深入研究，首先即以江西衛所作為範例來探討。明代江西衛所設置時間，大部分集中在洪武年間（于志嘉，1995），除了直隸於前軍都督府的九江衛（洪武 22 年(1389)調京軍而設）之外，萬曆《大明會典》江西都司所轄 3 衛 11 所，在《明實錄》明確記載設置時間者，為永新所、建昌衛（洪武 2 年(1369)改為千戶所）、廣信所、南昌左衛（永樂元年(1403)改為南昌護衛）、南昌前衛、贛州衛、袁州衛等，對於南昌衛則僅記載洪武 13 年(1380)裁撤之變動；以上事例都可由本系統「置衛設員」此一事件類型檢索得知。不過，南昌衛雖然在洪武 13 年(1380)已裁撤，但在本系統洪武 35 年(1402)與嘉靖元年(1522)的「置衛設員」類型中，仍可以檢索到南昌衛的事例，此外，永樂元年(1403)已改為南昌護衛的南昌左衛，也屢見出現。由此可知，南昌衛、南昌左衛復設的事例，無法全由「置衛設員」類型中得知，仍須探索其他事件類型來追溯。

由研究資料所見，萬曆《南昌府志》曾記載正德 16 年(1521)併南昌前、左衛為南昌衛一事（于志嘉，1995:1006），而對照《明實錄》記載南昌前衛與南昌左衛之相關事例出現時間，亦截至正德 16 年(1521)3 月；前此，南昌護衛之事例僅截至景泰 7 年(1456)10

月，且之後再度出現南昌左衛之事例，此呼應明代衛所沿革資料所見天順元年(1457)廢南昌護衛、復置南昌左衛之考證結果。因此，文本中不同時期出現的衛所名稱，其所代表意義不盡相同，若景泰 7 年(1456)10 月至正德 16 年(1521)3 月之間，《明實錄》所載「南昌等衛」事例，即須解讀為南昌前衛與南昌左衛。事實上，《明實錄》記載正德 16 年(1521)3 月南昌衛再度設置，乃「併江西南昌前右二衛為南昌衛」[1]，但南昌右衛的設置始末在一般研究中未見探討，無法進一步論述。無論如何，依據本研究的事件類型分析，將上述衛所變動的事例擷取出來，有助於進一步就相關問題作一了解，而南昌衛於正德 16(1521)年 3 月設置，就《明實錄》撰述內容已說明與「宸濠之變」有關，則衛所設置實與軍事防禦相關，亦為檢索時需要相互對照、檢視的事件類型（見圖三）。

圖三 江西衛設置相關之事件類型檢索比對

本研究藉由事件分群所歸納之衛所事件類型，可以反映衛所之功能與演變。以明初江西諸衛所的任務而言，在《明實錄》記載內容所見主要為操練征守、平定各地亂事，其中，贛州衛在平定南方相鄰的廣東山區亂事上，扮演著重要的角色，袁州衛則著重平定省境亂事與屯守；由「以袁州等衛官軍之戍贛者代領漕運」之記述，可見袁州衛具有漕運功能，而袁州衛偏重漕運的任務，又與「袁州等衛在吉安下流便於漕運」的地理位置因素有關[2]（見圖四）。若針對單一衛所來分析，即以袁州衛為例，相關的事件類型與時間分布，可以迅速透過網頁檢索來呈現；集中在洪武年間的事例，以平亂與設置為要，

[1] 《明實錄》卷 197，正德十六年三月，頁 3679。引自中央研究院歷史語言研究所「漢籍電子文獻資料庫」。

[2] 《明實錄》卷 11，弘治元年二月，頁 250-251。引自中央研究院歷史語言研究所「漢籍電子文獻資料庫」。

其後散見各時期者，除了人事命令相關事件類型外，其任務包括修城、屯田、賑濟、運糧、造船、屯戍、守備等，其中運糧與造船事例反映其漕運功能，主要出現在弘治、嘉靖年間（見圖五）。據考，永樂 2 年(1404)以後江西大量衛軍投入軍屯，在《明實錄》中的記載，主要見於調撥屯糧或受災蠲免所提及事宜，其後又以衛軍河運漕糧，使得江西大部分衛所逐漸兼有運糧或造船任務；然屯軍兼運、運軍又差役繁重的結果，導致逃亡日多，至萬曆年間出現對漕運軍役的改革措施（于志嘉，2001）。因此，進一步檢索江西漕運相關衛所，包括南昌、吉安、袁州、九江等，觀察其事例分布；結果顯示嘉靖、萬曆年間漕運相關事例居多（見圖六），且嘉靖 43 年(1564)對於造船物料加以規範、漕船團造、避免遲誤漕糧等記述，也呼應當時漕糧軍役繁重、積弊漸多的狀況：

> ……嚴催江西安福所新造船料使廉能衛官領之久任責成不許他用其袁州五衛船廠改於吉安南昌衛改於九江各就產木近地團造以後不許再更一有司遲誤漕粮及一應輕齎船料運軍口粮之類行撫按監充官從重參劾……[1]

事實上，屯軍在轉任漕運以後，轉由餘丁承擔屯田，待嘉靖年間江西條鞭法出現，逐步規範均徭銀的徵收，限餘丁及屯田，又有依戶等高下派分正役的規定，而衛官佔役包納之弊也明文禁止（于志嘉，1997）。對此，透過系統檢索漕運相關事例所見萬曆 14 年(1587)12 月「丙戌更定江西南昌衛所軍餘丁差徵銀募役不許衛所官占役包納從按臣陳有年請也」條，可探知其梗概。

圖四 袁州衛地理位置之地圖參照

圖五 袁州衛事件類型檢索

[1]《明實錄》卷 540，嘉靖四十三年十一月，頁 8743。引自中央研究院歷史語言研究所「漢籍電子文獻資料庫」。

圖六 江西漕運相關衛所與事件類型檢索

本研究運用文本自動標註技術與地理資訊科學方法，建立基於明代衛所名稱的地名自動標註與事件分類研究架構，並透過包括衛所名稱在內的地名時空座標對位機制及地圖介接功能設計，賦予漢籍全文資料庫空間屬性，並形成圖文整合之加值應用模式。本研究以明代衛所來實作《明實錄》之命名實體識別方法，乃考量衛所具有職官與地名之雙重特性，且各衛所相關事件之時空分布往往反映衛所功能特性與變動，在尺度上亦得以由微知著，拓展研究視角；同時，本研究透過實作結果與既有文史研究成果相互對照檢驗，探討用於實際研究的可能性，亦評估相關問題之解決方法。後續將在此基礎上進一步結合相關研究課題來推演，逐步探索本應用模式於數位人文領域的研究運用價值。

參考文獻

于志嘉。1995。〈明代江西兵制的演變〉，*中央研究院歷史語言研究所集刊*，66（4），頁 995-1074。

于志嘉。1997。〈明代江西衛所軍役的演變〉，*中央研究院歷史語言研究所集刊*，68（1），頁 1-53。

于志嘉。2001。〈明代江西衛所屯田與漕運的關係〉，*中央研究院歷史語言研究所集刊*，72（2），頁 301-338。

漢籍電子文獻資料庫，台北：中央研究院歷史語言研究所。

Chang, Yung-Chun, Chen, Cen-Chieh, Hsieh,Yu-Lun, Chen, Chien-Chin & Hsu, Wen-Lian (2015, July). *Linguistic Template Extraction for Recognizing Reader-Emotion and Emotional Resonance Writing Assistance*. Paper presented at the meeting of ACL-IJCNLP, Beijing, China.

基於文脈的古漢文引述偵測研究：以佛經為例

唐國銘*、黃乾綱**

摘　要

引述偵測 (quotation detection) 研究是從大量的文字資料中，辨識出引述句，以便進行進一步的引述分析 (quotation analysis)。現有的引述偵測研究，是以定義引述句法 (quotation syntax) 偵測引述句法中的引述線索為主 要方法。引述句法分成引述來源 (source) 、引述線索 (cue) 及引述文 (quotation content) 等三個部分[1, 2]。根據引述文和引述來源的關係，引述句又可分為直接引述 (direct quotation) 和間接引述 (indirect quotation) 兩種[2]。

然而，現有的引述句法的定義，無法含蓋古漢文中所有的引述情況。由於古漢文中有許多專門為某經典所做的「論」、「注」或「疏」等類型的著作，因此在上述基本的引述句法的定義範圍外，古漢文中還可觀察到只有引述文沒有引述線索及引述來源的引述句，以及為了描述經典中文字位置的引述句。古漢文的引述句抽取及辨識，其形式較為複雜，大致分為三種類型 ：

(1)完整引文的引述句：有完整引述對應字串，可能有 cue 的字詞。

(2)不完整引文引述句：無完整的引述對應字串或字串太短，可能有 cue 的字詞。

(3)描述位置的引述句：使用描述位置的結構及指向性的文字。

由於古漢語的引述句無法僅用條列「來源、線索、引述」的引述句法來描述。因此，本論文不單靠引述線索來偵測基本的引述句法，而是用樣式探勘 (pattern mining)等方法，直接偵測語料庫中的引述文[3]。取得引述文後，再分析其前後文，以找出可以較全面描述古漢語引述句的引述結構 (quotation pattern)。經過觀察發現佛教文獻的引述句，常使用描述位置的結構及指向性的文字，來表達引述。因此，取出引用文獻中的高頻詞為結構及指向的候選詞，從中去除在經文中同為高頻的詞後，再以出現於數詞的前後為條件，便可篩選出結構、指向等詞，並以

* 國立臺灣大學工程科學暨海洋工程學系博士生，Email: d965251013@ntu.edu.tw。
** 國立臺灣大學工程科學暨海洋工程學系副教授，Email: ckhuang@ntu.edu.tw。

此為基礎再取得描述位置的引述句。

本文研究的方法係設法從出處文本(Attribution Texts)與引用文本(Citation Texts)中，找出引述句及引述樣式。分為以下三層：(1)引述偵測層(Detection Phase)(2)抽取精煉層(Refinement Phase)(3)整合彙集層(Integration Phase)。在引述偵測層，共有四種引述偵測(Quotation Detection)方式，第一種是引述內容匹配(content matching), 找出出處文本與引用文本中匹配的字串, 做為候選引述句。第二種是線索詞搜尋(Cue word search)，根據可能的線索詞，找到候選引述句。第三種是引述規則搜尋(Rule based search)，根據人工經驗提供的引述規則，找到候選引述句。第四種是引述位置探勘(Location text mining)，根據可能的引述規則及引述位置，分別找到候選引述句。在抽取精煉層, 第一種的引述內容匹配方式找到的候選引述句。需再運用過濾器(filter)或線索詞探勘 (Cue word mining)將引述句粹取出來。第二種的線索詞搜尋方式，係直接需運用線索詞探勘所產生的線索詞庫(Cue word library)來找出引述句。第三種及第四種方式則視需要運用部分排序文本比對(Partial order text alignment)找到並確認引述句及其內容。找到的引述內容及線索，需進行精度評估(Precision Evaluation)，以確定演算模型的正確性。最後，再經過整合彙集層，將四種引述偵測方式加以整合將，以提高引述偵測的整體精確度。引述句抽取的架構詳如圖1.。

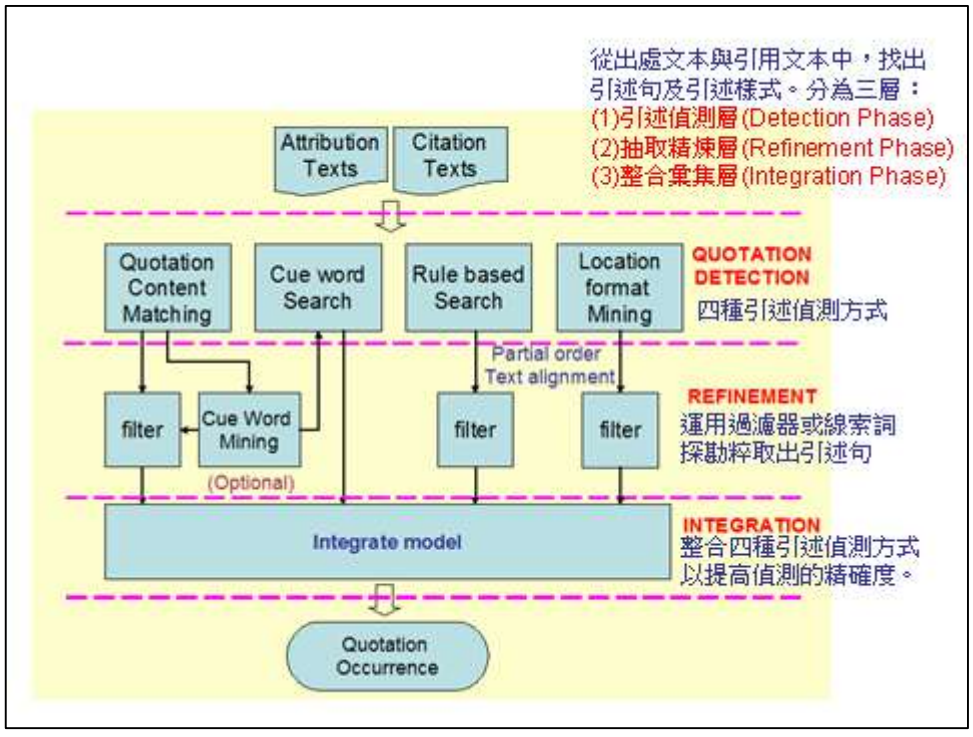


圖 1.本研究的引述句抽取架構

佛經是古漢文中重要且數量龐大的一類文獻，且佛經中也含有古漢文引述的所有情況。因此，本論文以漢文「大方廣佛華嚴經八十卷」及其註疏「新華嚴經論」為對象，進行古漢文引述自動偵測的實驗[5]，詳如表1.。

表1.實驗語料庫統計資料

attribution texts (經)	大方廣佛華嚴經（八十卷） CBETA 經號 T10n0279，80卷。共72萬餘字。[5]
citation texts (論)	新華嚴經論 CBETA 經號 T36n1739，40卷。共36萬餘字。[5]

以「新華嚴經論」23卷至29卷為例，分別計算Accurecy = (tp + fn) / (tp + fp + tn +fn)，Recall = tp / (tp + tn)，Precision = tp / (tp + fp)

F-score = 2 / (1/precsion + 1/recall) 。目前的實驗結果 Accurecy 84.1 ; Recall 77.2 ; Precision 56.6 ; F-score 65.4 。詳如表2.。將持續調整過濾器及整合模型，以獲得更佳的實驗效果。

表2.實驗結果數據

Text no.	tp	tn	fp	fn	Acc.	Recall	Prec.	Fscore
T36n1739_023	152	28	79	968	**91.3**	**84.4**	65.8	**73.9**
T36n1739_024	156	44	102	883	87.7	78.0	60.5	68.1
T36n1739_025	158	56	182	682	77.9	73.8	46.5	57.1
T36n1739_026	172	35	138	814	85.1	83.1	55.5	66.5
T36n1739_027	200	79	151	588	77.4	71.7	57.0	63.5
T36n1739_028	231	58	108	671	84.5	79.9	**68.1**	73.5
T36n1739_029	108	46	139	834	83.6	70.1	53.8	53.8
	1177	346	899	5440	**84.1**	**77.2**	**56.6**	**65.4**

未來，將逐步確認大量佛教文獻的引述內容，並從索引了解文獻相互引述的狀態，以建立佛經引述的脈絡及關係網。本研究所提出的方法，適用於所用古漢

文及其它現代文獻(如法條判例文獻)，可正確地找到引述內容，以及出處。本論文尚未完成古漢語引述結構的完整探勘分析。完成古漢語的引述結構分析，並證明其一致性及適用範圍，將是本研究下一階段的主要目標。另外，為發展本研究的可能應用，已完成「華嚴經引述檢索系統」服務實驗平台(http://service.quotation.cklab.org/)。

關鍵字：引述偵測、古漢文、古漢文引述句、文字探勘、大方廣佛華嚴經

參考文獻

Bonami, O. and D. Godard, *On the syntax of direct quotation in French.* Proceedings of the 15th International Conference on Head-driven Phrase Structure Grammar, 2008: p. 358-377.

O'Keefe, T., et al., *A Sequence Labelling Approach to Quote Attribution.* Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012: p. 790-799.

Pareti, S., et al., *Automatically Detecting and Attributing Indirect Quotations.* Acl, 2013: p. 989-999.

Smith–Waterman algorithm, in *Wikipedia*. 2016.

CBETA 中華電子佛典協會. 2010, Chinese Buddhist Electronic Text Association (CBETA).

Context-based Quotation Detection for Ancient Chinese Texts：A Case Study in Chinese Buddhist Sutras

Kuo-ming Tang*, Chien-kang Huang**

Abstract

Quotation detection research is the identification of quotation sentences from huge volumes of texts, in order to enable the further conducting of quotation analysis. The main methodology of current quotation detection research is through the definition of quotation syntax to detect quotation cues within. Quotation syntax can be divided into the three main components of source, cue and quotation content [1, 2]. According to the relationship between quoted text and quotation source, quotation sentences can be classified into the two types of direct quotation and indirect quotation.

However, the definition of quotation syntax in fashion cannot encompass all situations of quotations in ancient Chinese. This is because ancient Chinese corpus has many compositions specially for a certain scriptural book, in genres termed 「論」(Abhidharma), 「注」(Notes) or 「疏」(Commentaries). Hence aside from the basic definition of quotation syntax as above, one may see quotation contents without the cue and the source , as well as instances of quotation that describe the location of the quotation in Scripture. The extraction and identification of quotation in ancient Chinese is more complex. Broadly speaking there are three main types:

(1). Quotations quoting completely, with complete corresponding quoted text string, may have cue;

(2). Quotations quoting incompletely, without complete corresponding quoted text string or it may be too short, may have cue;

* Ph.D. Student, Department of Engineering Science, National Taiwan University. Email: d965251013@ntu.edu.tw.

** Associate Professor, Department of Engineering Science, National Taiwan University. Email: ckhuang@ntu.edu.tw.

(3). Quotation describing the quoted text's location in Scripture, using words indicating the structure and location of the quoted text.

Since we cannot just use source, cue and quotation content as above to describe quotation sentences in ancient Chinese, this essay would henceforth not just depend on cues to detect the basic contours of quotation syntax. Rather methods like pattern mining will be used to directly detect quotation sentences in the language corpus [3]. After we have obtained the quotation sentences, we would further analyze the context, so that we may fuller describe the quotation pattern in ancient Chinese. We could observe that quotation sentences in the Tripitaka often make use of words indicating the structure and location of the quoted text, conveying an instance of quotation. Thus, we would pick out frequently occurring words in the quoting text as candidate words so indicating structure and location. Those words (characters) also occurring frequently in the quoted scriptural text will be eliminated. Then after considering a few characters before and after the remaining ones as selection condition, we would have sieved out characters indicating structure, location, etc.. This will be the basis to obtain quotation sentences indicating where they are quoting from.

This research attempts to discover quotation sentences and quotation type from attribution texts and citation texts. Our process has three phases: (1). Detection Phase; (2). Refinement Phase; (3) Integration Phase. Four methods are employed in the detection phase. The first is content matching, discovering text strings which match in both the quoting text and the quoted text. They will become candidates of quotation sentences. The second is cue word search, picking up quotation candidates through possible cue words. The third and fourth are rule based search and location text mining respectively. Quotation candidates are discovered by possible quotation rules and locations. Coming to the refinement phase, candidate quotations through content matching will be treated with filters or cue word mining to distill out quotation sentences. Other candidates we found in the detection phase, depending on our needs, will be subjected to partial order text alignment, to find out and confirm quotation sentences and their contents. Quotation contents and cues found need to pass through precision evaluation to ascertain the accuracy of our algorithm. Finally, the integration phase pull together results from all four detection methods to raise the overall accuracy of quotation detection. The whole process can be seen in Fig. 1.

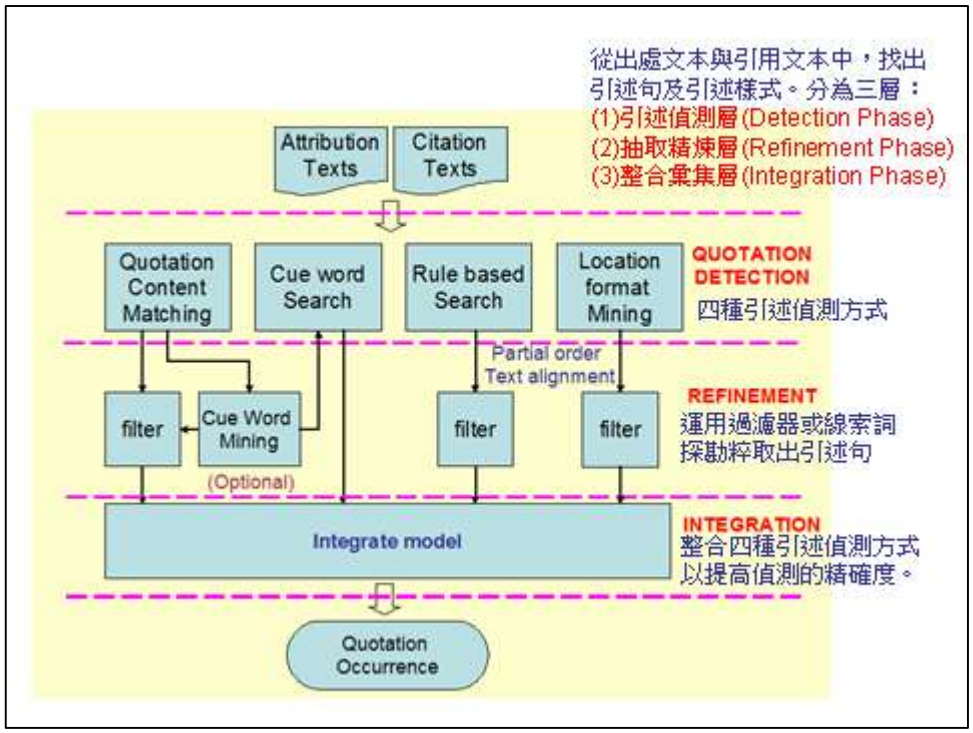


Figure 1. Proposed structure of the process of this study

The Buddhist Tripitaka is an important and huge collection of ancient Chinese texts. Furthermore within the Tripitaka there are all situations of quotation in ancient Chinese. Thus this essay makes use of the Chinese version of "Avatamsaka Sutra (80 rolls version)" and "Contemporary Treatise on the Avatamsaka Sutra" as objects for experiments on automatic detection of quotation in ancient Chinese [5]. Details of the results are shown in Table 1.

Table 1: Statistical Overview on the Experiment's Language Corpus

Attribution texts (Sutra)	Avatamsaka Sutra (80 rolls version) (CBETA,T10no.279), 80 rolls, totally over 720,000 characters [5]
Citation texts (Abhidharma)	Contemporary Treatise on the Avatamsaka Sutra (CBETA,T36no.1739), 40 rolls, totally over 360,000 characters [5]

From vol. 23 to vol. 29 in Contemporary Treatise on the Avatamsaka Sutra. Our current experimental estimate Accurecy = (tp + fn) / (tp + fp + tn +fn), Recall = tp / (tp

+ tn), Precision = tp / (tp + fp), F-score = 2 / (1/precsion + 1/recall). results show scores in accuracy, recall, precision and f-score to be 84.1, 77.2, 56.6 and 65.4 respectively, details data of the experiment result are shown in Table 2. We are continually fine tuning our filters and the model in the integration phase, towards even better experimental results.

In the future, we will progressively identify quotations in ever more Buddhist texts, indexing them to understand the state of mutual quotation, so as to construct the network and relationship web of quotations in the Tripitaka. The method proposed in this research can be applied to all ancient Chinese texts, as well as other contemporary corpus (e.g., court judgment cases), enabling correctly discovering quotations and their original locations. This essay has not exhausted the complete mining and analysis of quotation structure in ancient Chinese. Its completion and further proving of its consistency and range of applicability will be the main target of the next stage. Please also note that the Avatamsaka Sutra Quotation Search System experimental service platform (http://service.quotation.cklab.org/) is now complete.

Table 2: the data of Experiment result

Text no.	tp	tn	fp	fn	Acc.	Recall	Prec.	Fscore
T36n1739_023	152	28	79	968	**91.3**	**84.4**	65.8	**73.9**
T36n1739_024	156	44	102	883	87.7	78.0	60.5	68.1
T36n1739_025	158	56	182	682	77.9	73.8	46.5	57.1
T36n1739_026	172	35	138	814	85.1	83.1	55.5	66.5
T36n1739_027	200	79	151	588	77.4	71.7	57.0	63.5
T36n1739_028	231	58	108	671	84.5	79.9	**68.1**	73.5
T36n1739_029	108	46	139	834	83.6	70.1	53.8	53.8
	1177	346	899	5440	**84.1**	**77.2**	**56.6**	**65.4**

Keywords: Quotation detection, ancient Chinese, Quotations in ancient Chinese, text mining, Avatamsaka Sutra

文字探勘於中文資料前置處理技術之研究：以博物館數位典藏為例

賴鼎陞*

摘　要

本文探討中文文字探勘的相關議題，提出一個完整的中文資料前置處理流程，其架構分為: 微觀、中觀、宏觀等三個層次，並以一個實際的博物館數位典藏資料集為例，進行三個階段的實驗評估。此架構具有開放性，在未來進行數位典藏相關文字探勘研究時，可做為實驗設計與評估的系統化方法。

關鍵字：文字探勘、資料前置處理、詞頻分析、數位典藏

* 國立故宮博物院教育展資處助理研究員，Email: sam@npm.gov.tw。

Research on Chinese Data Pre-Processing Techniques for Text Mining: A Case of Museum Digital Archives

Ting-sheng Lai*

Abstract

This paper explores the related issues in Chinese text mining, and proposes a complete processes for Chinese data pre-processing. The architecture is divided into three stages: Micro-level, Meso-level and Macro-level. To evaluate the three stages of the experiments, a practical dataset of a museum's digital archives is used as an example. This architecture of the proposed processes is open and can be used as a systematic method for experimental design and evaluation in the future related research.

Keywords: texting mining, data pre-processing, term frequency analysis, digital archives

* Assistant Researcher, Division of Education, Exhibition, and Information Services, National Palace Museum. E-Mail: sam@npm.gov.tw.

一、前言

近年來，國際間越來越多的學術或研究單位，積極整合跨單位資源，成立數位人文相關研究中心，並積極推動各項跨域合作計畫。

在數位人文研究的領域中，博物館、典藏單位經常扮演內容提供者的角色，尤其國內各館所從過去參與數位相關計畫以來，持續產出眾多的數位影像、後設資料等檔案資料，更是繼續深化發展數位人文重要機會。

然而，各館所目前的數位典藏檔案資料，除了影像授權、文化創意產品設計等應用外；後設資料（Metadata）目前僅能透過傳統的網站資料庫形式，提供公眾或內部檢索、查閱，在國外相關學術、實務的領域，也無明確的創新運用。推究其原因，各館所缺乏專門技術人力，以及研究工具; 而館外的研究人士，通常也不容易理解館所的典藏文本（Archival Context），是極待克服的關鍵問題。

因此，本研究以博物館文本分析為發展方向，目標是將文字探勘（Text Mining）方法，應用於博物館數位典藏之後設資料分析，適當地與最新的數據分析（Data Analytics）相關技術接軌，促進博物館數位人文研究的發展。

目前越來越多國際研究學者專注研究資料前置處理（Data Pre-Processing）相關議題，並探討其在文字探勘所扮演的關鍵性角色（Munková et al., 2013; Vijayarani et al., 2015），然而目前對於中文文字探勘領域，目前對於資料前置處理的特定議題，仍缺乏專門性研究。

本文將針對目前文字探勘（Text mining）應用於中文文本分析的限制，以及缺乏典藏專業詞庫等相關議題進行分析，進而訂定適當的研究方法和實驗，以期透過良好資料整備（Data Preparation），提高研究效能。

二、研究方法

本研究採用實驗研究方式，文本資料以國立故宮博物院之器物類數位典藏為主。實驗資料的取得，是以一般研究者為中心，直接由故宮網站資料庫擷取文物之後設資料，經編整後計約一萬八千餘筆數位檔案（Document）。

為了進行連續性的實驗操作，資料需要運用適當的演算法逐步處理，以期推算出不同資料量、資料格式和內容的資料。為實驗研究需要，本文定義為：來源資料（Source Data）、語料庫（Corpus）、樣本資料集（Sample Dataset）、目標資料集（Target Dataset）等。

資料整備方面，在資料擷取（Scraping）、過濾（Filtering）等程序，採用Python[1]、SQL語言相關工具，以轉換處理來源資料和語料庫，進而輸出適當的樣本資料集。而在集料庫（Datasets）的整備，以及後續的主要實驗操作，都是以R語言[2] 進行處理。

本研究針對資料前置處理的各種可能的項目，歸納並設計出三項主要實驗，於本文定義為：微觀層次（Micro-Level）、中觀層次（Meso-Level）、宏觀層次（Macro-Level）。為了能有效評估實驗效果，本研究將過濾出「銅器類」的「題名」，計約四千六百餘筆檔案，做為樣本資料集。

三、實驗研究

第一個實驗，是在微觀層次，主要先進行「資料清理（Data Cleaning）」。清理的項目，除了標點符號等非文本相關符碼之外； 一些存在於題名中的強勢 （domimant）詞彙，例如: 周、西周、戰國等年代相關字詞 （註：青銅禮器多數產製於先秦），但在文本分析並不需要，可先移除。

而數量眾多的「銅印」，上面的「印文」也登錄在題名中，故需將標點符號「和」的內容自原題名抽離，並註記至下一個欄位，避免文本分析時的干擾。

其次，是基於文本內關鍵詞（KWIC，Keyword-in-Context）的概念 （Kowalski & Maybury, 2002），將目標資料進行索引。索引的方式，是採「中文斷詞」操作，因為目前的斷詞系統，皆未包含華夏文物相關語料，故作者自行建置典藏專業之「自定詞庫」。

在不同中文斷詞方面，本研究選擇以「人工斷詞」、「Jieba 套件[3]」、「Rwordseg 套件[4]」等三種方法，配合作者自訂的自定詞庫，針對銅器題名進行斷詞，並比較三種方法的差異。

在成效評估方面，則以「詞頻 （Word Frequency）」統計分析，結果發現，而三種方法皆達合理分詞效果 （表一）。再以「詞雲 （Word Cloud）」方式進行視覺化呈現 （圖一），三種方式皆可適切表現樣本資料集的主要詞彙與相對重要性，故可繼續進行下個層次的實驗。

[1] https://www.python.org/
[2] https://www.r-project.org/
[3] https://github.com/qinwf/jiebaR
[4] https://r-forge.r-project.org/R/?group_id=1054

表一 三種分詞方法的詞頻比較（僅列出最高的 10 個詞）

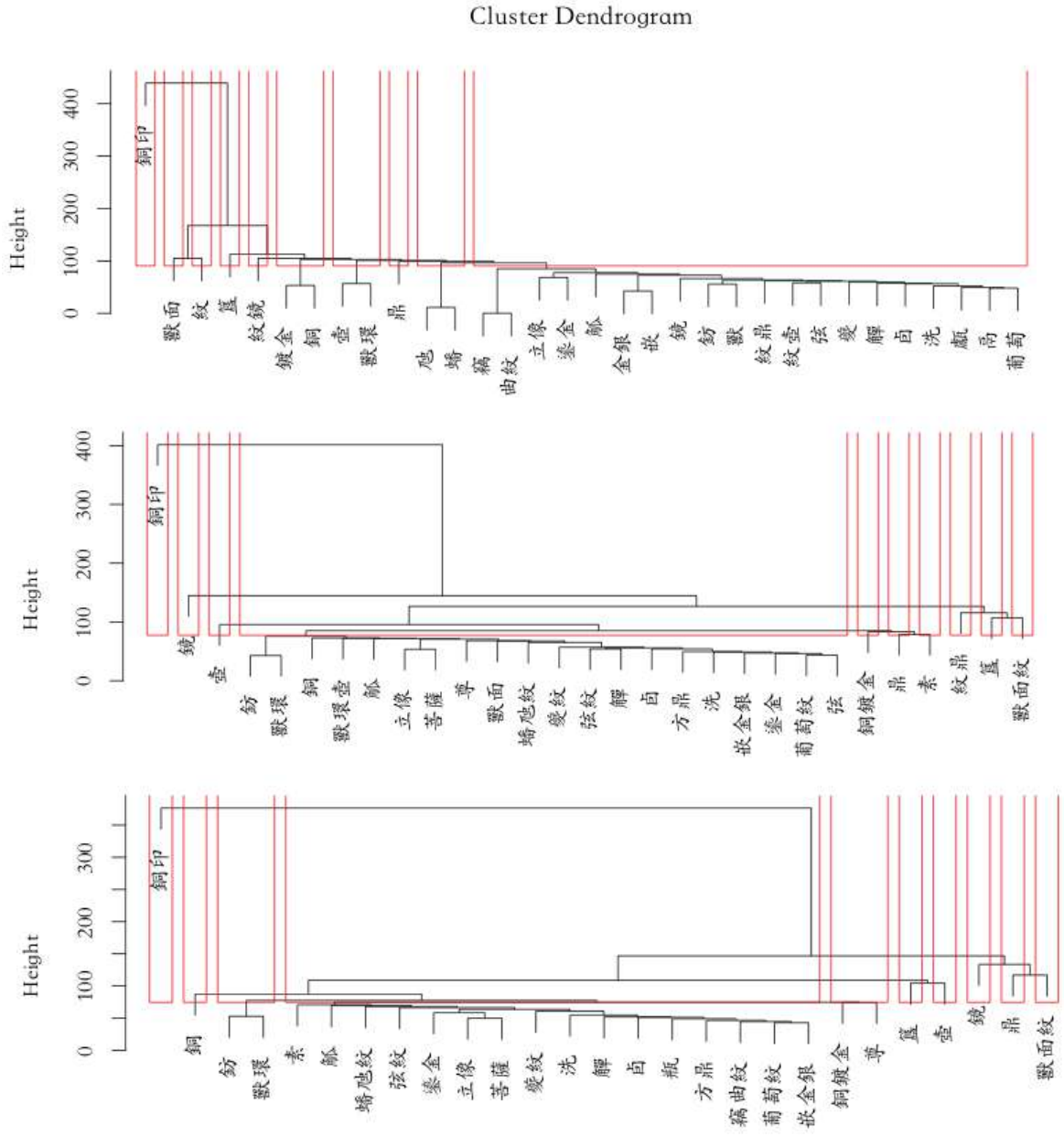
Jieba 套件		Rwordseg 套件		人工斷詞	
銅印	1604	銅印	1604	銅印	1603
紋	554	鏡	293	鼎	348
獸面	365	獸面紋	271	獸面紋	348
鼎	216	紋鼎	235	鏡	318
銅	207	簋	200	壺	247
簋	204	壺	151	簋	201
獸環	158	觚	112	銅	165
紋鏡	152	素	107	獸環	154
觚	122	鼎	104	觚	116
鎏金	118	銅鍍金	100	鎏金	115

圖一 三種分詞方法的詞雲比較

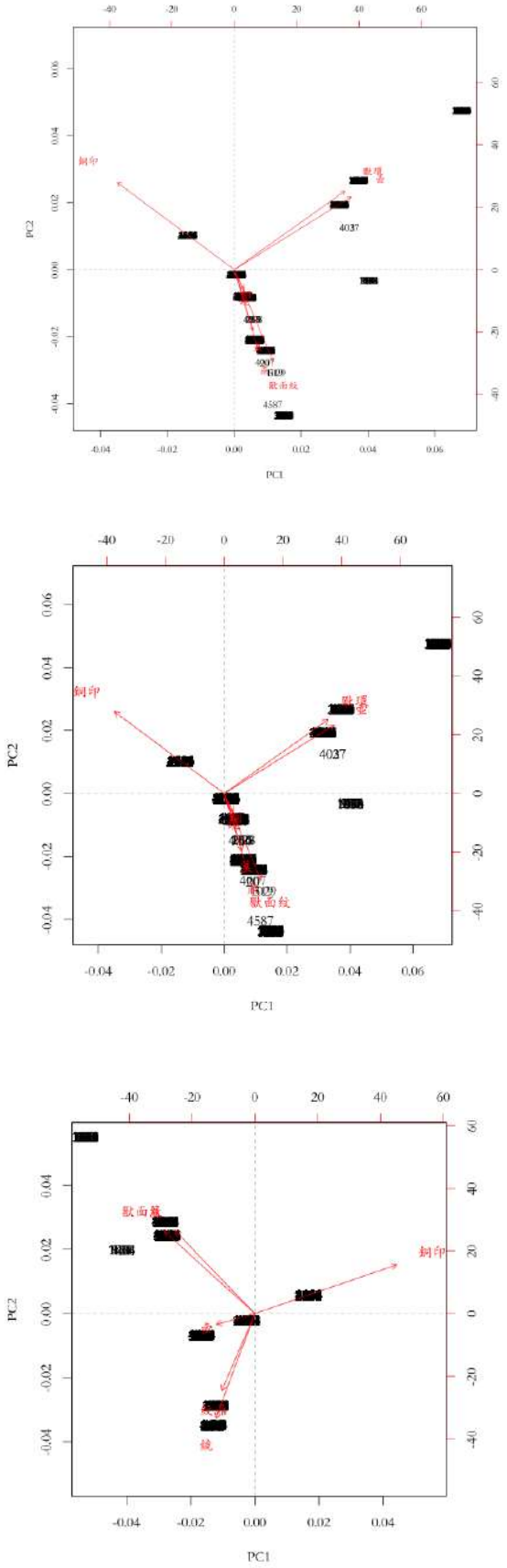
而第二個實驗，是在中觀層次，主要目的是自動處理關於「非初始的、亦非最終」的中介資料。此階段主要先建立「文檔—詞條矩陣 （DTM，Document-Term-Matrix）」，並依矩陣的稀疏率（Sparsity）計算，排除部分發生率低的詞條，以增加高頻率詞條的能見度。

其次，是運用機器學習（Machine Learning）相關的方法試算，以確保第一階段目標資料可被自動運算，並順利移轉至下一階段繼續處理。本實驗採用「 階層式集群分析 （Hierarchal Clustering）」（Manning et al., 2008）、「主成份分析 （Principal

Component Analysis）」二種機器學習方法，運用上述第一個實驗的三種已分詞的輸出資料，進行運算、繪圖。結果發現，此二種機器學習方法皆可達合理分群效果 （圖二、圖三），故可繼續進行下個層次的實驗。



圖二 比較三種分詞方法的階層式集群分析



圖三 比較三種分詞方法的主成份分析

最後，第三個實驗，是在宏觀層次。此階段主要是進行「分類（Categorization）」操作 （Feldman & Sanger, 2007），基於文本外關鍵詞（KWOC，Keyword-Out-of-Context）的概念（Kowalski & Maybury, 2002），為目標資料進行索引。

有些檔案資料的屬性與其他資料差異較大，故相關詞條會影響整體資料集的文本分析，所以運用分類方法將檔案適當區隔，以改善各分類的文本分析交果，如：「銅印」、「銅鏡」的藏品數量相對很多，故可用索引方式註記為不同類別。

分類的演算法由筆者設計，而預先定義分類項目，可將各檔案以階層式索引註記。結果發現，最終目標資料可被有效分類，以提高整體分析結果。

四、結語

綜上所述，依據三個實驗的結果與步驟，本研究可總結出一個資料前置處理的流程，在資料層次，可分為：微觀、中觀、宏觀等三個層次的架構。而此架構具有開放性，除了博物館數位典藏目錄之外，亦可適用於具有相似屬性的資料集，例如: 圖書館、美術館、檔案館等。

本研究所提出的中文資料前置處理技術，可提供一個系統化的資料整備流程，對未來進行不同資料集間的文本分析比較，或是跨資料集的文本統合分析，可做為實驗設計與評估之用，例如「帶扣」在銅器、瓷器、玉器都有；而「琺瑯」也有銅器和瓷器，有助於博物館持續發展數位人文研究的深度和廣度。

參考文獻

Feldman, R., Sanger, J. (2007). Chapter 4: Categorization. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge: Cambridge University Press, p.64,81

Kowalski, G. J., Maybury, M. T. (2002). Chapter 6: Document and Term Clustering. Information Storage and Retrieval Systems: Theory and Implementation, 2nd Edition. New York: Kluwer Academic, p.71-104.

Manning, C. D., Raghavan, P., Schütze, H. (2008). Chapter 17: Hierarchical Clustering. Introduction to Information Retrieval. Cambridge: Cambridge University Press, p.377-401.

Munková, D. Munk, M., Vozár, M. (2013). Data Pre-Processing Evaluation for Text Mining: Transaction/Sequence Model. Procedia Computer Science. 18. p.1198-1207.

Vijayarani, S., Ilamathi, J., Nithya. (2015). Preprocessing Techniques for Text Mining - An Overview. International Journal of Computer Science & Communication Networks. 5 (1), p7-16.

從數位人文觀點看語藝在運算轉向的機會與挑戰

劉大華*、黃鈴媚**、林頌堅 ***、曹開明 ****

摘　要

一、　研究緣起

近年來，隨著鉅量資料（Big Data）普及應用與資料分析技術與工具的演進，數位人文學者高呼「向運算轉」（computational turn），吸引人文、藝術、社會等不同學科的研究者意識到運算取向的重要性，並跨越了學門的界線，在數位人文（Digital Humanities）的大傘下，藉由運算技術的輔助，合作探求新的研究範式，補充與修正既有理論概念的基礎（Berry, 2011）。

語藝（rhetoric）源於古希臘時期，是一具有長久歷史與濃厚人文特質的學科，關注人們如何使用語言符號與修辭策略，建構認同和達成說服的目的（Bizzell & Herzberg, 2000）。然而隨著人們溝通形式的轉變，當網路逐漸成為人類溝通的重要場域時，Lanham 在 1989 年率先提出「數位語藝」（digital rhetoric）的概念，意指在電腦中介傳播情境下，對網路上所繁衍的數位文本進行語藝探究（Eyman, 2015）。Zappan（2005）指出數位語藝研究主要聚焦在（一）探究與分析數位文本中所使用的語藝策略、（二）確認新媒體的語藝特徵、作用與限制、（三）探究數位認同的形塑、（四）探究建構語藝社群的效用。Zappan 強調，除了將傳統語藝研究轉移到數位空間，同時也要從重視語藝理論的核心概念與批評方法上的擴展與整合。

二、　問題意識

然而當前數位語藝研究的一大挑戰在於研究方法上的受限，網路時代的社群溝通所產製的文本數量龐大、累積迅速，傳統質性的取徑已難以掌握語藝情境的廣度，更難捉準語藝的時機性（Kairos），恐導致文本分析時缺乏足夠論證依據。

* 世新大學傳播博士學位學程博士生，Email: agogonono@gmail.com。
** 世新大學口語傳播學系教授，Email: lmhuang@mail.shu.edu.tw。
*** 世新大學資訊傳播學系助理教授，Email: scl@cc.shu.edu.tw。
**** 國防大學政治作戰學院新聞學系助理教授，Email: tommy.intw@msa.hinet.net。

為此，本研究選擇社群媒體的數位文本為研究個案，試圖在「理論驅使」(theory-driven)與資料趨使(data-driven)，以及「近讀」與「遠讀」(Brummett, 2010; Moretti, 2005）這兩條軸線上取得平衡，尋找契合語藝研究的混合取徑。

在理論方面，本研究以 Bormann 的「符號融合理論」概念為引導，其理論主要預設在於強調：(一）人們藉由溝通創造真實；(二）個人不僅採取符號創造真實，個體藉由符號融合、創造出社群共享的意義。而由於人們使用符號互動與詮釋的過程，猶如是種戲劇形式，透過描述「人物」、「行動」及「場景」的主題來形成「幻想主題」(Fantasy theme)，並透過「覆誦」作用形塑出「語藝視野」(rhetoric vision)，也就是共享的意義(Bormann,1972, 1980a,1982b, 1982c, 1983,1985a, 1985b, 1990; Shields & Preston,1985;Cragon & Shield,1992)。

三、 研究方法

在研究文本與分析的方法上，本研究結合自然語言處理（Natural Language Processing）與計算語言學（computational linguistics）的文字探勘技術，其文本分析步驟為：

（一）選擇個案：2014 年 3 月 1 日至 4 月 30 日期間，國內 Facebook 上 11 個反核粉絲專頁。

（二）搜集文本：本研究透過網路爬文，以「反核」、「核能」、「核四」為關鍵字，蒐集核能相關之公開貼文與回文，共得貼文共 3424 則，回文 6443 則。

（三）斷詞（word segmentation）

1. 初步斷詞：首先 R 軟體中已普遍用來進行中文文字探勘的 Jiebar 套件，僅以內建的詞庫進行第一次的斷詞。

2. 未知詞偵測(Unknown Word Detection)與複合詞擷取(Extraction of Compound Words)：本研究使用 bigrams(兩個連續單詞)的方法，考量其不用詞庫的優點，將初始斷詞後所得的語料庫，計算任意兩個前後緊連詞語的共現次數，接著排序檢視所有 bigrams 的組合，以長詞優先（Maximum Matching）為原則找出未知詞和較具語意的長詞，將其新增到自建辭典

3. 重新斷詞：如此反覆上述斷詞與自建新詞的程序，直到新增的新詞趨於飽合。

4. 清理資料：移除英文字、數字與停止詞（Stopword）後，依據最終得到的語料庫，篩選其中詞頻大於 150 次，詞長 2 個以上，以及詞語出現的文章數大於 10

篇及小於 500 篇，過濾得到貼文 366 個詞與回文 420 個詞，列為代表文本意義的重要關鍵字。

（四）詞語分群

以傑卡德相似係數(Jaccard similarity coefficient)計算重要關鍵字矩陣（Gomaa & Fahmy, 2013)，計算文章中詞語的共現關係。接著找出詞語的叢聚情形來代表用來形塑語藝視野的幻想主題。這裡採用近鄰傳播(Affinity Propagation)做為分群的演算法（Frey & Dueck, 2007），其優點是不需要先指定分群數目的特性，根據資料點(data point)之間的相似度(similarity)分群，同時考慮各點為潛在的群中心(exemplar)，可以降低錯誤率並節省大量的時間。最後在貼文與回文各得 31 與 16 個詞群，然後以研究者對文本的瞭解，進行主題（詞群）的辨識、命名與編碼。

四、 結論

本研究的結果顯示，數位語藝結合電腦運算的文本分析工具，可大幅提昇語藝研究的效率。另一方面也驗證，傳統的語藝理論與批評方法只要加以調適，仍能在數位語藝研究的實踐中扮演理論指引的重要功能。

本研究同時體認到，斷詞是中文文本探勘的成敗關鍵，本研究認為在初始斷詞後藉由 Bigram 的計算來增加斷詞的正確性有幾個優點，一是其原理容易理解，適合初學者使用。其次是效果顯著，在操作的過程中，研究者可適時介入來提升未知詞的偵測與複合詞的組合，有利於三種情況：1. 還原未收錄於辭典中的人名、地名、行動標語、口號等新詞，例如「林義」、「義雄」（人名）、「非核」、「家園」（標語）等的。2. 結合有助語義判斷的複合詞，如「依賴」接合「進口」。3. 區分不同複合詞在語義上的歧異，例如「輻射汙染」與「輻射劑量」。

作為語藝研究與電腦運算工具的初探性研究，建議未來應繼續擴展應用的層面。近讀與遠讀是研究者看待事物的焦距調整，而非完全對立，研究者應回到問題本身，以解決問題為目的，然而研究品質的良窳需有賴研究者兼顧理論深度，以及對運算工具與原理的理解。

關鍵字：數位語藝、符號融合理論、斷詞

The Computational Turn : Exploring the Opportunities and Challenges of Rhetoric Research from the Perspective of Digital Humanities

Ta-hua Liu[*], Lin-mei Huang[**], Sung-chien Lin[***], Kai-ming Tsao[****]

Abstract

In recent years, with the popularization of Big Data and the progress of data analysis techniques and tools, several humanists have point out the slogan of "computational turn", which get the recognition of scholars from the humanities, art and society.

Rhetoric originated from ancient Greece. It is a discipline with long history and strong humanistic characteristics. It focuses on how people use language symbol and rhetoric strategies to construct identity and to persuade others (Bizzell & Herzberg, 2000).

With the change of people's communication practice, Lanham first proposed "digital rhetoric" concept in 1989. However, the challenge of current digital rhetoric research is the limitation of methods.

This study chooses the digital text of facebook as a case study, and attempts to explore the relationship between "theory-driven" and "data-driven" , and balance "close-reading" and "distant reading" (Brummett, 2010; Moretti, 2005)

The results of this study show that the combination of digital rhetoric and computer operations, can greatly enhance the efficiency of the study. However, the quality of research depends on both the theoretical depth and the understanding of computing tools by researchers.

Keywords: digital rhetoric, symbolic convergence theory, word segmentation

[*] Ph.D. student, Ph.D. Program in Communication Studies, Shih-Hsin University. Email: agogonono@gmail.com.
[**] Professor, Department of Speech Communication, Shih-Hsin University. Email: lmhuang@mail.shu.edu.tw.
[***] Assistant Professor, Department of Information and Communications, Shih-Hsin University. Email: scl@cc.shu.edu.tw.
[****] Assistant Professor, Department of Journalism, Fu Hsing Kang College. Email: tommy.intw@msa.hinet.net.

Employing Digital Tools to Re-examine Distributional Hypothesis: A Case of *Kèjiā* 'Hakka' in News Corpora

Huei-ling Lai [*] , Yi-lun Weng [**], Chao-lin Liu [***]

Abstract

The present study aims to re-examine the distributional theory and compositionality with a semantic vagueness term by using digital tools. Distributional hypothesis claims that the semantic distribution of words is determined by how often the words occur next to one another. For instance, the occurrence of the word *kèjiā* in a news item could refer to anything associated with the concept Hakka, including culture, language or people. Employing text-analysis tools, the present study retrieves *kèjiā* and its collocate from the four major newspapers in Taiwan and four aspects of collocation strength are measured: (1) frequency, (2) mean and variance of the distance between *kèjiā* and its collocates, (3) Pearson's chi-squared test and (4) pointwise mutual information. Then the top 15 verb collocates are chosen for coding the referential domain of *kèjiā*. Our computation gives rise to finer-grained distinction of various collocates, and resolves referential vagueness that challenges the claim of distributional semantics and compositionality. The findings provide implications of a strong tendency toward the most prevalently discussed and concerned topics regarding Hakka issues in Taiwan news.

Keywords: distributional hypothesis, compositionality, *kejia* 'Hakka', humanity digital tools, semantic vagueness

[*] Distinguished professor, Department of English, National Chengchi University. Email: hllai@nccu.edu.tw.
[**] Research assistant, Department of English, National Chengchi University. Email: winnie8289@gmail.com.
[***] Distinguished professor, Department of Computer Science, National Chengchi University.
Email: chaolin@nccu.edu.tw.

以數位人文工具重新檢視詞彙共現分布理論：以新聞語料庫中的「客家」詞彙為例

賴惠玲[*]、翁翊倫[**]、劉昭麟[***]

摘　要

本研究旨在運用數位人文工具計算語意模稜的詞彙共現分布，藉以強化語意分布理論及語意組合性。詞彙分布假設認為詞彙的語意分布取決於詞彙與其相鄰詞的共現頻率。舉例而言，新聞媒體中詞彙「客家」的出現可指涉各種與客家相關的概念，包含指其文化、語言、及客家族群。運用文本分析工具，本研究以「客家」為關鍵詞，從臺灣四大報紙的語料中檢索並擷取其搭配詞，並從四個面向測量其共現強度：（1）共現頻率（2）「客家」與共現詞彙之平均距離及標準差變異 （3）皮爾森卡方檢定（4）逐點間之互信息。計算結果之前 15 名動詞再針對「客家」的指涉範疇做第二次的人工解碼。本研究更精密準確的計算共現詞彙，同時解決語意模稜的詞彙對語意分布理論及語意組合性的。研究發現也意涵由臺灣媒體最關切並頻繁報導的客家議題，可看出語言使用、媒體及社會之密切關係。

關鍵字：詞彙共現分布理論、語意組合性、客家、數位人文工具、語意模稜

[*] 國立政治大學英國語文學系特聘教授，聯絡信箱：hllai@nccu.edu.tw。
[**] 國立政治大學英國語文學系專案助理，聯絡信箱：winnie8289@gmail.com。
[***] 國立政治大學資訊科學系特聘教授，聯絡信箱：chaolin@nccu.edu.tw。

1. Introduction

Distributional hypothesis claims that the semantic distribution of words is determined by how often the words occur next to one another. However, since sentences are formed by individual words, how can the meaning of a sentence be derived when some of the lexical items are vague? Semantic vagueness might cause the difficulty of parsing. For instance, the word *kèjiā* (Hakka, 客家) can refer to anything associated with the concept Hakka such as culture, language or people even if they all have the same distributional environment. Employing a structural approach and statistics tool can help computer programs better understand the relationship and distribution of words.

2. Distributional Semantics, Compositionality and Semantic Vagueness

Distributional semantics assumes that semantic representations can be built in the form of high dimensional vector spaces through a statistical analysis of the contexts in which words occur. The basic idea of distributional semantics is that "words which are similar in meaning occur in similar contexts" (Rubenstein & Goodenough, 1965). A correlation between distributional similarity and meaning similarity exists, allowing us to utilize the former to estimate the latter. For instance, if two linguistic forms *w1* and *w2* tend to have similar distributional properties, and they do not occur with the other entity *w3*, then we may infer that *w1* and *w2* belong to the same linguistic class.

Compositionality is the mechanism in which the meaning of a sentence is formed by the composition of the meanings of its components. The principle is based on the assumption that the hierarchical syntactic structures classify sentences into subparts, sub-subparts, and sub-sub-subparts, and ultimately into individual words. Hence, semantic representations can be derived compositionally due to a one-to-one correspondence between their syntactic structures. However, compositionality runs into several problems, in particular when word meanings are ambiguous or vague (Pelletier 2004). Semantic vagueness in which a lexeme is underspecified with certain information gives rise to a lower level of semantic specificity (Tuggy, 1993). For instance, words of occupation are often underspecified for gender. *The doctor just examined the patient* can refer to a female or a male doctor. Another kind of indeterminacy has to do with referential identity. To resolve such a lack of information exhibited by a semantic vagueness term requires increasing its specificity (cf. Cruse

1986). One way is to add syntagmatic modifiers as syntagmatic arrangement of two or more lexical items is widely used for modulating information for specificity. The word *kèjiā* (Hakka,客家) is semantically vague as it can refer to anything associated with the concept Hakka such as culture, language or people. In this study, employing digital tools, we try to resolve the semantic vagueness of *kèjiā*, presenting association measures that quantify the strength and the reliability.

3. Methods

3.1. Materials

The corpora come from four major newspapers in Taiwan — *Knowledge Management Winner* (KMW), *Udndata* (Udn), *Liberty Times Net* (LTN), and *Apple Daily* (AD). All news articles are extracted from January 1, 2005 until December 31, 2015. Articles which contain the keyword *kèjiā* are extracted to establish the database. In total 168,116 tokens are found.

3.2. Procedure

The text-analysis tools are employed to search the word *kèjiā* in the database. Chinese sentences are segmented into words to find collocates of *kèjiā*. The statistical evaluations between *kèjiā* and its collocate are calculated with four methods: (1) Frequency, (2) Mean distance and Standard deviation, (3) Pearson's chi-squared test, and (4) Pointwise mutual information.

3.3. Segmenting the text

Text segmentation is considered an important step for Chinese natural language processing tasks since Chinese words can be composed of multiple characters without space in-between. Sentences and segment words need to be divided correctly, so as to acquire correct word frequency and find collocations. To this end, Patricia tree (PAT Tree) is used to segment texts and retrieve collocations, allowing us to collect a corpus, construct a PAT tree, extract significant lexical patterns, and do text segmentation on other documents. With the text segmented, we can then tackle collocations, finding *kèjiā*'s collocates and detecting their collocation strength.

3.4. Measures of Collocation Strength

3.4.1. Frequency

In order to find the content collocates of *kèjiā*, we remove all the function words such as conjunctions, articles, auxiliary verbs, and pronouns and words with frequency less than 100. We then choose the most frequent 15 noun content words (including *kèjiā* and its collocate).

3.4.2. Mean distance and Standard deviation

Although frequency-based search works well for fixed phrases, many collocations consist of two words that stand in a more flexible relationship to one another. If the distance between two words is not constant, the fixed phrase approach would not work. One way of discovering the relationship between two words is to compute the mean and variance of the distances between two words. This information can be used to discover collocations by looking for pairs with low deviation. Consider the formula in (1).

(1) $$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$$

In the formula, n is the number of times the two words co-occur, σ^2 and μ are the variance and the mean of the co-occurrence. If the value of d_i is the same in all cases, then the variance is zero. If the offsets are randomly distributed, then the variance would be high. Variance-based collocation discovery is an appropriate method to find word combinations that are in a looser relationship than fixed phrases and that are variable with respect to intervening material and relative positions. A low deviation means that the two words usually occur at about the same distance. Zero deviation means that the two words always occur at exactly the same distance.

3.4.3. Pearson's chi-squared test

High frequency and low variance can be accidental. If two words co-occur a lot just by chance, they do not form a collocation. Hence, a further aim is to compare the observed frequency in the table with the expected frequency under the independence hypothesis. A chi-square (χ^2) test can be applied to categorical data to evaluate how likely it is for any observed difference between the sets to occur by chance. The χ^2 statistic is calculated by

summing the squares of the differences between observed and expected values divided by the expected values, as shown in (2)

(2)
$$x^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where O_{ij} and E_{ij} respectively are the observed frequencies and the expected frequencies for cell (i, j), i and j separately represent the i-th rows and the j-th column of a cross table. If the χ^2 value exceeds the critical value, then we can reject the independence hypothesis, which indicates that two words co-occur by chance.

3.4.4. Pointwise mutual information

Pointwise mutual information (PMI) is a measure of association used in information theory and statistics (Church et al., 1991; Church and Hanks, 1989; Hindle, 1990). This type of mutual information is roughly a measure of how much one word tells us about the other, that is the association strength between the two items. In our study, PMI expresses the direction of the association (attraction vs. repulsion) and the strength of the association on a logarithmic scale. Mutual information is a log likelihood ratio of the probability of the bigram *P(x, y)* and the product of the probabilities of the individual words *P(x)P(y)*:

(3)
$$\mathrm{PMI}(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$

When two words only occur together, $i_n(x, y) = 1$; if they are distributed as expected under independence, $i_n(x, y) = 0$ as the numerator is 0. Finally, when two words occur separately but not together, we define $i_n(x, y)$ to be −1, as it approaches this value when $p(x, y)$ approaches 0 and $p(x)$, $p(y)$ are fixed.

4. Collocation coding

After finding and computing collocations, we selected 15 most frequent verb collocates. The results showed *jǔbàn* ‘hold’, *zhǎnxiàn* ‘show’, *chénglì* ‘establish’, and *tǐyàn* ‘experience’ score relatively higher than other cases in terms of the four calculations, indicating that they carry stronger collocation strength with *kèjiā*. Particularly, *zhǎnshì* ‘display’ tops the other collocates in terms of PMI value, indicating the strongest meaning relation with *kèjiā*. However, this word does not show high frequency, nor does it show close distance

with *kèjiā*. The VP structure which contains *kèjiā* alone cannot be determined due to the vagueness of *kèjiā*. Our next step is to categorize all verb collocations, trying to resolve the ambiguity. Three items are considered: the predicate of verb collocations, the referential domain of *kèjiā* and the news topic they belong to. The coding was conducted by four research assistants who have linguistics training for at least 2 years. In order to make sure the credibility of coding results, they exchanged the data for the second round of examination.

5. **Discussion**

Across the 15 predicates, the most frequently reported topic in the news is art and culture, followed by life and style and politics. Fourteen verb collocates appear in the art and culture topic, and the verb *tǐyàn* ‘experience’ exists in the art and culture topic most. This shows the strong correlation between *kèjiā*, its collocates and the news topics. Table 1 shows what Hakka matters are associated with each verb: for instance, *guīhuà* ‘to plan’ is mostly related to Hakka cultural parks, or Hakka earthen buildings; *zhīchí* ‘support’ is all related to presidential candidates; *tuīguǎng* ‘promote’ is related to Hakka culture, followed by Hakka language and music.

Table 1. The predicate of the verb collocates

Rankingguīhuàtuīguǎngtuīdòngzhǎnxiànzhǎnshìbǎocúnzhēngqǔchénglìngxiànchénglìdǎzàozhīchífāzhǎnchuánchéngjǔbàntǐyàn
1 文化園區客家文化客家事務客家精神客家文物客家文化客家家票客家家風情客家後援會客家桃花源蔡英文社區客家精神客家歌比賽客家家風情
‘culture park’‘Hakka culture’‘Hakka affairs’‘Hakka spirit’‘Hakka cultural’‘Hakka culture’‘vote’‘Hakka custom’‘Hakkaporting fan’‘Hakka a Utopia’‘Tsai Ing-’‘comunity’‘Hakka spirit’‘Hakka a music contest’‘Hakka custom

					relic ’						we n’					
2	客家圓樓	客家話	客家文化	客家風情	客家粄仔	客家聚落	經費	精神	客家委員會	臺灣	馬英九	客家產業	客家戲曲	音樂會	客家擂茶	
	‘Hakka earthen building s’	‘Hakka language’	‘Hakka culture’	‘Hakka custom	‘Hakka rice noodles’	‘Hakka village’	‘found ing’	‘spirit’	‘Hakka Affairs Council’	‘Taiwan’	‘Ma Ying-jeou’	‘Hakka industry’	‘Hakka music’	‘concert’	‘Hakka tea’	
3	客家土樓	客家歌謠	客家語	客家特色	客家米食	客家文物	補助	特色	客家電台	客家圓樓	朱立倫	客家族群	客家母語	嘉年華	客家文化	
	‘Hakka earthen building s’	‘Hakka music’	‘Hakka language’	‘Hakka character’	‘Hakka food’	‘Hakka cultural relic’	‘subsidy’	‘character’	‘Hakka radio’	‘Hakka earthen building-ings’	‘Chu, Li-luan’	‘Hakka group’	‘Hakka language’	‘carnaval’	‘Hakka culture’	

The syntagmatic co-occurring collocates increase semantic specificity to the vague term *kèjiā*, giving more concrete specification of the meaning domain expressed by the concept Hakka. For instance, the two cases *zhīchí kèjiā zǐdì* ‘support the Hakka candidate’ and *chéngl ì dì-yí-ge kèjiā diàntái* ‘established the first Hakka radio’ and *zhǎnxiàn kèjiārén de*

rénwén jīngshén ‘showed the Humanism of Hakka people’ more specifically highlight which referential domain *kèjiā* denotes.

6. **Implications**

This study, employing digital tools, provides more precise computation regarding words whose referential meanings are undetermined. The research outcome contributes to two aspects. On the one hand, the analysis remedies the weakness of distributional hypothesis and compositionality for constructing meanings of large chunks from the meanings of their components. The procedure demonstrates how digital more rigorous methods can help resolve referential uncertainty due to a semantically vague lexical expression. On the other hand, the findings show which aspects of Hakka matters are most prevalently reported in Taiwan media. Planning Hakka-image parks or promoting Hakka music or food are the top two news topics in the news media, showing the governmental efforts in these aspects regarding Hakka issues. Specifically, the results strongly reflect media framing effects whereby news media convey government efforts to appeal to Hakka people’s anxiety from the loss of their cultural tradition and their language. The high frequent straight news report related to *kèjiā* ‘Hakka’ seems to provide certain communication effects to the Taiwan society in general and to the Hakka ethnic groups in particular--Hakka issues are now being emphasized by the government. However, such a media framing effect can give rise to stereotype effects toward anything that can be referred as *kèjiā* Hakka’. The media can serve as a two-edged sword. The consequences of such effects that come from the intertwining relations of media, society and language can be further investigated.

參考文獻

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Church, K., Gale, W., Hanks, P., & Kindle, D. (1991). Using statistics in lexical analysis. In U. Zemik (Ed.), *Lexical acquisition: exploiting on-line resources to build a lexicon* (pp. 115-164). Psychology Press.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1989). Parsing, word associations and typical predicate-argument relations. In *Proceedings of the workshop on Speech and Natural Language* (pp. 75-81). Association for Computational Linguistics.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics* (pp. 268-275). Association for Computational Linguistics.

Kilgarriff, Adam. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, *1*(2), 263-276.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Pelletier, F. J. (2004). The principle of semantic compositionality. In S. Davis & B. S. Gillon (Eds.), *Semantics: A Reader* (pp. 133-156). Oxford: Oxford University Press.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8*(10), 627-633.

Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, *4*(3), 273-290.

基於單類別支持向量機之文風辨識：以金庸武俠小說為例

林峻虓*、施朵舲**、徐碩***、高憶瑄****、陳光華*****

摘　要

文風辨識一直以來都是個重要問題，傳統文風辨識常採用的方法包含考證文本以外資訊、檢測敘事內容、分析文本思想等，多需以人力閱讀，經大量統計及手動分析後才能獲得少量的資訊，不但容易有閱讀者主觀意識和推論的涉入，更重要的是曠日費時，卻不保證獲得足夠有力的證據和資訊。隨著數位工具的進展，部份依賴於統計的分析方法開始有了一線曙光。如果文本數量夠多，屬於該文風群集的可信度也足夠高，那麼便有機會引入機器學習的技術來幫助獲得需要的資訊。本研究採用「單詞」而非「單字」的頻率作為一個文本的文風特徵，並以機器學習的單類別支持向量機（one-class support vector machine）演算法訓練出屬於該類型文風的文風分類器（classifier），可用於辨認一個輸入文本是否是該類別文風的作品。

實驗是以「金庸的武俠小說」作為目標文風群集，從金庸超過四十萬字的長篇武俠小說文本，先挑出出現頻率最高的前 300 個詞彙，再人工篩選出其中和文意較無關的「虛詞」和與敘事風格較相關的「敘事慣用詞」作為金庸武俠小說的文風關鍵詞，共計 245 個詞彙，依據獲得的關鍵詞抽取出各文本的特徵向量，用以訓練文風分類器。為了評估分類器的效能，我們以金庸自己的武俠小說作為正向文本 (屬於金庸武俠小說的文本)，以數本「其他作家寫的武俠小說」和「非武俠小說」兩種文本研究為負向文本 (非屬於「金庸武俠小說」的文本)，將兩者拿來交給分類器進行分類。結果顯示，本研究提供的方法效果非常顯著，不管是正向還是負向測試資料，評估正確性都非常高。在本實驗最後採用的模型，正向資料的分類正確性為 82.81%（318/384），負向資料的分類正確性為 95.65%（901/942）。而模型預測錯誤的《白馬嘯西風》在經過詞頻分析與人工閱讀後證實與金庸其他武俠小說確實有相當大的不同，因而模型之預測實屬合理，更進一步彰顯了本實驗的模型在快速辨別文風差異、探索研究方向上的潛力。

關鍵字：分類、單類別支持向量機、文風

* 國立臺灣大學資訊工程學系學生，Email: b01902064@ntu.edu.tw。
** 國立東華大學國際企業學系學生，Email: jiang19930903@gmail.com。
*** 國立臺灣師範大學東亞學系學生，Email: hsus@ntnu.edu.tw。
**** 國立臺灣師範大學國文學系學生，Email: ks072882@gmail.com。
***** 國立臺灣大學圖書資訊學系教授，Email: khchen@ntu.edu.tw。

The Identification of Writing Style based on One-Class SVM: A Case Study for Chinese Martial Arts Novels of Jin Yong

Chun-huang Lin [*], To-ling Shih [**], Shuo Hsu[***]

Yi-hsuan Kao[****], Kuang-hua Chen[*****]

Abstract

The writing style is an important characteristic of texts and authors, which can convey not only narrative technique but even philosophy behind the words, and so that has been widely used as one of criteria in author identification, text similarity, and content-based classification. A classic method for identification of the writing style is to select some context-free words as keywords, and to carry out statistical analysis on the frequency of keywords. However, traditional statistical analysis tools are too simple to detect complex frequency patterns, and the performance does strongly rely on keyword selection. Given enough texts belonging to the target style, this work uses machine learning as a new analysis tool, and extends keyword selection criteria to include words which are non-context-free but related to the target style. The frequencies of keywords are used to compose feature vectors and one-class SVM (support vector machine) is applied to training a binary classification model based on feature vectors. The trained model could predict whether a text belongs to the target writing style. Experiments done on Chinese martial arts novels of Jin Yong show that the performance is significant. The model achieves 82.81% (318/384) accuracy on positive data (texts of Jin Yong's martial arts novels), and 95.65% (901/942) accuracy on negative data. This work also investigates false negative texts manually, and finds that the style of those texts do show obvious differences from other Jin Yong's works, which indicates that the proposed method can rapidly detect outliers for advanced researches.

Keywords: classification, one-class SVM, writing style

[*] Undergraduate Student, Department of Computer Science and Information Engineering, National Taiwan University. Email: b01902064@ntu.edu.tw.

[**] Undergraduate Student, Department of International Business, National Dong Hwa University. Email: jiang19930903@gmail.com.

[***] Undergraduate Student, Department of East Asian Studies, National Taiwan Normal University. Email: hsus@ntnu.edu.tw.

[****] Undergraduate Student, Department of Chinese, National Taiwan Normal University. Email: ks072882@gmail.com.

[*****] Professor, Department and Graduate Institute of Library and Information Science, National Taiwan University. Email: khchen@ntu.edu.tw.

一、問題與回顧

文風辨識一直以來都是個重要的問題，除可應用在作者分析、文章相似性評估等小範圍的群集分類問題外，還可用來分析文章類別、時代文風、族群文風等對象範圍較廣的群集分類問題。傳統文風辨識常採用的方法包含考證文本以外資訊、檢測敘事內容、分析文本思想等方式，多需以人力閱讀的方式，經大量統計及手動分析後才能獲得少量的資訊，不但容易有閱讀者主觀意識和推論的涉入，更重要的是曠日費時，卻不保證獲得足夠有力的資訊和證據。

隨著數位工具的快速進展，部份依賴於統計的分析方法開始有了一線曙光。以《紅樓夢》的作者考證問題為例，如果從文風的角度切入《紅樓夢》的作者考證問題，可以把問題轉化成探討《紅樓夢》前八十回和後四十回是否屬於同一個文風群集（曹雪芹的文風）。事實上，這個視角從高本漢開始，陸陸續續有許多人使用統計方法進行了一系列的研究，其中許多人使用了虛字字頻分佈作為重要的文風特徵；近年來杜協昌（2012）在《利用文本採礦探討紅樓夢的後四十回作者爭議》中，除了嘗試借助了統計工具和虛字字頻作為特徵外，還把「虛字使用習慣」拓展成「用詞習慣」的概念，進一步引入了資料探勘（data mining）的技術，透過資料探勘找出頻率有顯著差異的單字詞（unigram）和雙字詞（bigram）進行詞頻統計，成功地給出了許多前人難以獲得的有力證據支持後四十回作者另有其人。由於資料探勘需要大規模的詞頻比對，這樣巨大的運算量若沒有數位工具是非常難以觸及的。

文風辨識問題的類型非常多，為了更具體地進行研究，本研究針對的問題為「特定文風辨識」，即會先鎖定某個特定的文風群集，以及一部分已知為該文風的文本資料，目標是給出一個具體而有系統的方法來辨別一份未知文風的文本是否屬於該文風群集。《紅樓夢》的作者考證問題事實上也可以視為這類問題，即令特定文風為「曹雪芹版本的《紅樓夢》文風」，已知的文本為《紅樓夢》的前八十回，而研究目的則是判斷《紅樓夢》後四十回是否屬於「曹雪芹版本的《紅樓夢》文風」群集。

在《紅樓夢》的考證問題中，由於已經知道「至少前八十回是由同一人所著」這個重要的背景知識，並且只對「後四十回是否也都為前八十回的作者所著」有興趣，因此會有許多有利於處理這個議題的假設：

1. 前八十回的用詞習慣相當接近，因而統計出的詞頻穩定性很高。
2. 可以將《紅樓夢》的一百二十回文本視為兩份文本，前八十回一份，後四十回一份，因為前八十回已經確定是同一作者，後四十回是不是同一人所作並非研究的重點，縱使後四十回有些仍是曹雪芹親作，只要不是大部分皆由曹雪芹所作就不影響分析的結論。
3. 兩份文本的內容長度都相當充足，在內文中統計詞頻不用擔心因為文長不足導致詞

頻分佈歧異性過大的問題。

上述的 3 個假設不僅讓統計方法的可信度大幅提昇，也讓研究時只需要專注於兩者間的不同之處是否合理，在設定策略和統計標的上都更加明確。然而，在更一般化的特定文風辨識問題中，這 3 個假設都不復存在，且問題的目標也不侷限在「辨別某特定文本是否為特定群集」，而是「辨別任意文本是否為特定群集」。如此一來，上面提及的研究採用的策略皆會因為「屬於特定群集的文本間詞頻的歧異性」和「非特定群集的文本的不明確性」而失去其效力，無法作為一般化問題的解決方案。

會遇到這樣的困境，追根究底，是因為和原先的《紅樓夢》作者考證問題相比，一般化的特定文風辨識問題要區分的兩個類別（屬於該特定文風群集和不屬於該文風群集）的集合都包含大量的文本，而原先的考證問題中，兩個類別的集合內都只有一份文本。然而，也正因「大量文本」的特性，讓需要大量資料進行學習的機器學習（machine learning）有機會挺身而出。幸運的是，上述的兩個問題，在機器學習裡面都有答案。

本研究將採用杜協昌（2012）以用詞習慣作為文風代表的概念，利用用詞習慣作為文本的特徵，並用機器學習的單類別支持向量機（one-class support vector machine, one-class SVM）作為學習的演算法（algorithm），基於部份給定的目標文風的文本，建立一個自動生成的文風分類器，可用於快速地辨別任意未知文風的文本是否屬於目標文風群集。

本文分為前後兩大部份，第一部份為分析問題的抽象層面與提出解決方案，第二部份則為實驗細節和進一步的探討。第一部份由第二節到第四節構成，第二節將描述拓展杜協昌的文風代表概念的動機與方法、第三節分析先前的研究方法在本研究目標問題上會遇到的瓶頸，並說明引入機器學習的動機，接著第四節介紹機器學習的基礎知識與本研究採用的機器學習演算法。第二部份則由第五節到第七節構成：第五節將介紹建構文風分類器的具體細節，並說明本研究中實驗的各項細節設定；第六節則以金庸的武俠小說為例，建構出能夠辨認文本「是否為金庸所作的武俠小說」的分類器，實驗本研究提出的方法的實際效果；最後在第七節中揭露模型預測結果能帶來的相似度訊息，並以簡單的詞頻分析證明分類器的預測無誤，藉此展現本研究提出的分類器的實用性以及在提供研究方向上所能帶來的巨大價值。第八節則是簡短的結論。

二、文風定義與代表

傳統上「文風」作為一個抽象的概念，其實甚少有非常嚴謹的定義。舉例而言，唐朝的山水田園派、社會寫實派等是寫作題材上的不同；社會學中的左派、右派思想是立場風格上的不同；就連一樣是道家思想家，老子、莊子表達思想的敘述手法也有很大的

差異。上述三例雖然是截然不同的切入點，但文字上都可以說是「文風不同」。學術研究上，也確實從這些切入點都可以看得出文本的文風是否類似，因此可以說上述的三例都是文風的重要特徵，可以作為文風的「代表」，卻沒有一者可以完全表述文風的全體面貌。

既然文風難以一言以蔽之，那麼進行文風分析時，切入面向的選擇便相當的重要，需要依據研究方法和研究標的的不同，選擇以何種面向作為文風的代表。然而，如果我們想要以統計方法和機器學習這類工具輔助我們分析文風，那麼文風的「具體性」便影響重大。如果我們選擇以思想、立場之類的抽象表徵作為文風的定義，那麼除非研究者對文本有非常充分的背景知識，否則難以用簡單的方式來表示「思想」的數值特徵；反觀杜協昌在《利用文本採礦探討紅樓夢的後四十回作者爭議》選擇以「用詞習慣」作為文風代表的策略就顯得相當合適，不僅可以直接以用詞頻率作為客觀的數值特徵，更可以在近乎不了解《紅樓夢》的抽象內容的情況下，區分出思想和題材皆接近相同的兩個可能不同的作者完成的作品，對於「區分作者」這樣的研究標的而言可謂強而有力的最佳選擇。

若進一步探討以「用詞習慣」作為文風代表的特性，我們還能夠發現，用詞習慣事實上也能部份地包含其它文風代表的意義。舉例而言，在山水田園派的詩詞中，比較可能出現的則是「林」、「桃花」、「琴」、「靜」等更像是自然景觀或休閒相關的詞彙；而在社會寫實派的詩詞中，比較容易看到「朱門」、「病」、「烽火」等與亂世相關的詞彙。又比如在左派的主張裡面，比較容易看到「弱勢」、「公平」、「福利」等字眼，在右派的主張中則比較容易看到「自由」、「市場」和「效率」等字眼。儘管用詞習慣距離完整地表現題材內容或者思想立場仍有相當的距離，但在區分完全不相似的兩份文本上仍能夠展現明顯的差距。

基於上述的優點，若要選擇一個一般化的文風代表來表達大多數的文本特徵，用詞習慣確實為一個客觀、易得又不失效果的指標。因此，本研究採用了杜協昌以用字習慣作為文風代表的看法，並捨棄僅以單字詞和雙字詞作為統計標的的方式，直接以分詞軟體將文本全數分詞，以分詞的結果進行統計。

在統計完文本各個詞彙的頻率後，為了能夠讓機器學習的演算法使用詞頻作為文本的特徵以進行學習，必須把選出的關鍵詞頻的資料轉換成以數值表示的向量，作為一份文本的數值特徵。第五節會更詳細地說明如何獲得文本的詞頻與關鍵詞，並且將關鍵詞頻作為文本的特徵向量。

三、傳統方法的瓶頸與改善方案

第一節曾經提及過去的研究策略在「屬於特定群集的文本間詞頻的歧異性」和「非特定群集的文本的不明確性」兩者的影響下無法發揮其效力，以下將詳細分析為何會有這樣的現象。

首先必須回顧前人研究採用的策略：趙岡、陳鍾毅（1980）透過抽樣分析用「兒」、「在」、「了」、「的」、「著」五個虛字在前後兩部份的出現頻率，並以 t 檢定值（t-test）計算出前後文同屬一個分佈的機率極低；Chan（1986）透過隨機抽樣詞彙進行前後兩部份相關係數的統計認為前後皆為一人所著；余清祥（1998）除了組合多個虛字進行線性判別分析外，還進一步利用了詩詞數量進行統計檢驗以及變動點觀測，判斷前後不同作者；何光國（2002）以「的」、「地」、「得」三個助詞的字頻分析認為前後同作者；最後是杜協昌（2012）透過資料探勘找到頻率相差極大的用詞，分析其使用習慣後，判斷前後不同作者。

統整上面各家的作法，可以歸納出：

1. 選擇關鍵字詞的方法有三種：

(1) 研究者自行選定。

(2) 隨機抽樣。

(3) 透過比較兩種群集的文本使用資料探勘獲得。

2. 獲得關鍵字詞後的分析方法有五種：

(4) 針對各個關鍵詞進行統計分佈檢驗。

(5) 針對關鍵詞在文本中的使用習慣進行人工分析。

(6) 針對多個關鍵詞進行線性判別分析。

(7) 變動點觀測。

(8) 計算前後關鍵詞相關係數。

下文將討論「屬於特定群集的文本間詞頻的歧異性」和「非特定群集的文本的不明確性」會帶來的問題，並列舉各個問題會影響的研究方法：

1. 屬於特定群集的文本詞頻的歧異性

如果特定群集裡面有多份文本，而文本之間的關鍵詞並不全然相同，那麼在選擇關鍵詞上將會有「哪個詞具有代表性」的問題，進行頻率分析時如何定義「基礎詞頻」也會是困難的問題。舉例而言，若「特定群集」定義為「法國大革命相關作品」，則「自由」與「平等」皆是可想而知的出現頻率極高的關鍵詞。然而，若有一書專門討論自由，一書專門討論平等，那麼兩書皆應屬於特定群集的範疇，但兩者間「自由」和「平等」的

詞頻卻很可能有天壤之別。在這種狀況下，如果兩詞都不選取，顯然喪失了兩個極為重要的關鍵詞；只選取其中一者作為關鍵詞，另一者類型的書籍很容易會被誤判；兩者都選取，則所有基於單一關鍵詞均勻分佈假設的統計方法皆難以適用，如上述的(4)、(6)、(7)、(8)。而兩者都不選或者只選其中一者帶來的巨大負面影響也讓(2)變得容易不實用。

2. 非特定群集的文本的不明確性

當非特定群集的文本含有複數以上文本（事實上是所有不屬於特定群集的文本都屬於此類）時，研究者就很難基於非特定群集文本的特性選擇關鍵詞，因為此類型的文本繁多而雜亂（以本研究中的實驗為例，所有不是金庸的武俠小說的文本皆為此類），難以得知此類文本的關鍵特性，從而設計特別能夠區分兩者的關鍵詞。這影響了(3)方法的適用性，另外也讓(5)方法的成本過高（所有的可疑文本皆須人工檢驗一次）。

至此，除了(1)方法仍然未被明確否定外，(2)到(5)等方法在面對如此一般性的問題皆遇到了直接的挑戰，而研究者選定關鍵詞則會因為缺乏能夠抵抗詞頻歧異性的後端分析方法而顯得左右為難。究其原因，是因為傳統的統計方法往往基於許多變數獨立性以及分佈狀態的假設，且難以表達多個變數間的複雜關係。在特定群集為「法國大革命相關作品」的假設題目，一個直觀的評斷方式為計算「自由」與「平等」兩者頻率相加的結果，但這在普通的統計模型中是難以被表達的，而要研究者自行處理這類問題，則不但考驗研究者的觀察能力，更難以推廣到多變數之間的關係。

另一方面，利用傳統的統計方法，往往也讓研究者在選詞上承受極大壓力。研究者面對數量極為龐大的文本，難以像面對《紅樓夢》時一樣透過閱讀來獲得相關的前備知識，只能針對既有的統計結果或自己的猜想選擇關鍵詞；然而當統計顯示未知類別的文本和已知類別的文本有關鍵詞頻上的差異時，又不容易判斷究竟是選詞策略錯誤，還是真的兩者為不同群集。

最後，基於要人工檢驗的分析方式雖然在研究者具有背景知識的情況下能夠彌補工具上的缺失，卻沒辦法面對大量的數據以及其他背景知識不足的資料。

基於上述的考量，一個解決方法是更換獲得關鍵詞後的分析方法。新的分析方法必須滿足：

- 同時能考量多個不同的關鍵詞，並能夠表達較為複雜的詞與詞之間的關係。
- 若考量的關鍵詞中有無意義的雜訊（noise），能夠不太大地影響辨識的能力。
- 有足夠可信的判斷機制，使得自動判斷文本群集的正確性足夠高到不需要人力檢驗。

幸運的是，在資料足夠充足的情況下，機器學習正好可以解決上述的所有問題。在機器學習中，演算法處理的輸入正是高維度的向量，因此如同第二節結尾提及將所有選取的關鍵詞頻一起整合成向量，交由演算法學習向量各個維度間的複雜關係；若向量中

有特定維度對最終的預測結果貢獻不大，演算法會自動降低該維度的影響力，因此研究者便可以更無畏地選擇潛在的關鍵詞；演算法內部定義的函數通常足夠複雜，且可以簡單地檢測模型（model）的精確度，因此研究者若能夠透過機器學習獲得足夠好的模型，便能夠放心地以該模型進行實際的預測而不須另外親自檢驗。簡言之，機器學習透過學習向量維度間的複雜關係的能力以及大數據的優勢，解決了文本詞頻歧異性與選詞雜訊的問題，更由於其可計算性大幅降低了對人力的需求。

第四節將更詳細地說明機器學習的抽象內涵。值得一提的是，「非特定群集的文本的不明確性」在機器學習中將帶來不同問題，這將在第四節第二小節的討論中呈現並獲得解決。

四、機器學習、單類別分類與單類別支持向量機

第二節已經介紹了以用詞習慣（詞頻）作為文本特徵的方法，因此在把文本轉換為數值上的特徵向量後，一份文本在抽象概念上便可以不再是許多文字的有序組合，而是一個高維度的數值向量，向量的數值則由文本的詞頻決定（為了方便表達，下稱此向量為該文本的「詞頻向量」）。由於這樣的表達形式正是機器學習中演算法看待資料的方式，因此在本節會以詞頻向量作為文本的代表。

下文將會簡單地介紹機器學習、單類別分類與單類別支持向量的的基本概念和直觀意義，以便在獲得機器學習的結果後，能夠解讀模型預測結果的物理意義。

(一) 機器學習

機器學習是近年來興盛的計算機科學的一門領域，可以視為人工智慧（artificial intelligence，簡稱 AI）的一種實現方式。在機器學習裡面，使用者必須先提供一些資料給計算機，透過機器學習的演算法，電腦會學習出一個模型，使用者再利用此一模型來對其他資料進行分類（classification）或者數值上的預測。因此，大致上機器學習包含下面四個基本的步驟：

1. 將原始資料預先處理成對應的數值向量。
2. 將訓練機器用的資料（訓練資料，training data）提供給機器學習的演算法進行學習，機器在學習後會傳回學習到的模型。如果學習出的模型目標是用於預測資料的分類，則此模型又常被稱為分類器（classifier）。
3. 檢測機器學習到的模型其預測效果是否合乎預期，方法通常是將一些已知結果且不在訓練資料內的的資料交給模型預測，再檢視預測的結果是否和實際結果相合。此步驟通常稱為驗證（validation），驗證用的資料則稱為驗證資料 (validation data)。

4. 若模型的預測效果良好，則用於原先要做的實際需求。

(二) 單類別分類

本研究的目標是判斷某文本的文風是否屬於某個文風群集，因此可以看成是在「屬於目標文風群集」和「不屬於目標文風群集」兩種結果之間進行分類的二元分類問題（binary classification）。

一般在解決二元分類問題時，使用者會提供兩種類別的資料給演算法學習，演算法主要在於學習如何「區別」兩種不同的類型，輸出的模型可以看成是一條界線，用來劃清兩個群集。然而，對於我們的目標問題而言，這樣的學習方式卻會遇到資料失衡（unbalanced data）的問題──「屬於目標文風群集」的文本量，相較於「不屬於目標文風群集」的文本量猶如滄海一粟。以我們接下來會用來實驗的金庸武俠小說為例，金庸的武俠小說雖然數量不少，有高達十五部作品，但除了該十五部作品外全世界所有的文本皆屬於「非金庸武俠小說」的範疇。想要將全世界的文本都列入訓練資料不僅是不可能任務，就算完成了也很容易因為兩種類別資料量的巨大差異而影響模型的效能。

既然「非金庸武俠小說」類的文本不能盡收囊中，如果想要使用一般的二元分類演算法，就難以避免要在該類別中抽樣數個文本作為代表。但新的問題是，「非金庸武俠小說」的類別太過龐大以至於同屬該類別的的文本歧異性也非常巨大，就算要抽樣某個子集合來進行訓練，該子集合所需的文本數量仍然多到難以負荷；與此同時，由於母集合太過廣泛難以取得，如何均勻地獲得子集合也是困難的問題。

以上的分析說明了「非目標文風群集」的資料雖然俯拾即是，但由於數量過大和歧異性過高，反而失去了實用性，如同當只有「目標文風群集」的資料可以使用，這正是前面提及的「非特定群集的文本的不明確性」問題在機器學習過程的表現方式。所幸，機器學習對這類型的問題也已經有些研究，稱為「單類別分類問題（one-class classification）」。

在現實生活中，有些狀況下雖然是二元分類問題，卻難以取得其中一種分類的資料。本研究的目標問題是其中一個例子；另一種例子則是如工廠的自動警報系統。如果工廠希望設計一個系統可以透過檢測各個監控器的情況來自動判斷是否有緊急危難，那麼很容易遇到的問題是大多數時間獲得的資料都是「沒有緊急危難」時的資料，因為大部分時間工廠都是處於正常運作的狀態。透過人工模擬來獲得「緊急危難」時的資料，不僅成本非常高昂，也會因為人類思考與經驗的侷限性而產生設計上的盲點，遺漏一些可能會發生的緊急危難。此時機器學習的策略必須要改變，不再是在給定的兩個分類的資料之間學習界線，而是在僅給定其中一種分類、或者給定的兩個分類終有一個分類的資料數量極少的情況下，進行學習。事實上，這仍是可能的，一如人類縱使只有看過蘋果，

看到榴槤時仍能非常肯定那不是蘋果。事實上，機器學習的研究也已經提供了一些單類別分類演算法來解決這樣的問題。

(三) 單類別支持向量機

單類別支持向量機便是其中一種知名的單類別學習演算法，運用機器學習常用的支持向量機（support vector machine，簡稱 SVM）的技術來解決單類別分類問題。支持向量機的細節內容由於和本研究較不相關且包含大量的數學演算，在此省略，但支持向量機的本質卻是具有物理的意義。有關 SVM 較為詳細的討論，可參考 Schölkopf, Platt, Shawe-Taylor, Smola, and Williamson（1999）。

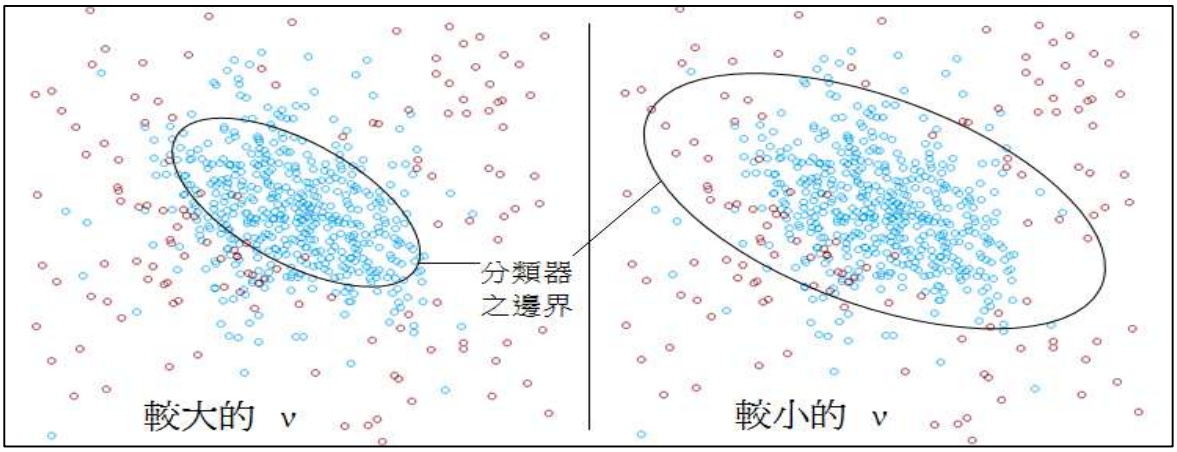
由於所有資料皆以數值向量的形式輸入、運算與輸出），因此所有的資料都對應到一個高維度的向量。若將向量視為平面中的點座標，那麼一個二元分類器對應的便是高維度空間中的某個封閉子空間，在子空間內的資料點屬於一個分類（為了方便敘述，設此為類別一的資料），在子空間外的資料點屬於另一個分類（類似地設此為類別二的資料）。一個直觀的想法是，給定「應當在分類子空間中的點集合」，即訓練資料後，若能夠訓練出一個合理的封閉子空間包含盡量多的訓練資料點，則這樣的一個子空間便能夠作為一個分類器，空間內者為與訓練資料相同的類別，空間外則為另一個類別。

然而，若僅將「包含點數」作為評斷分類器的優劣標準，最好的分類器理所當然應當包含整個母空間，如此便可以包含所有的資料點，這顯然會使得分類器變得毫無作用──只要將任意輸入都判斷為同一種分類即可。因此，單類別支持向量機用了一個恰好相反的想法來建構分類器：給定訓練資料和訓練資料錯誤率的上界 ν，生成一個盡量小且錯誤率小於 ν 的子空間。儘管不全然正確，但直觀上較大的 ν 允許較多錯誤資料點，從而允許子空間體積較小；較小的 ν 則為了能夠符合較嚴格的錯誤率限制，子空間可能不得不更大些。由此可見，機器學習縱使能夠自動產出模型，還是需要有經驗的使用者透過觀察模型的運行狀況以調整參數，才能產出能夠符合實際需求的模型。

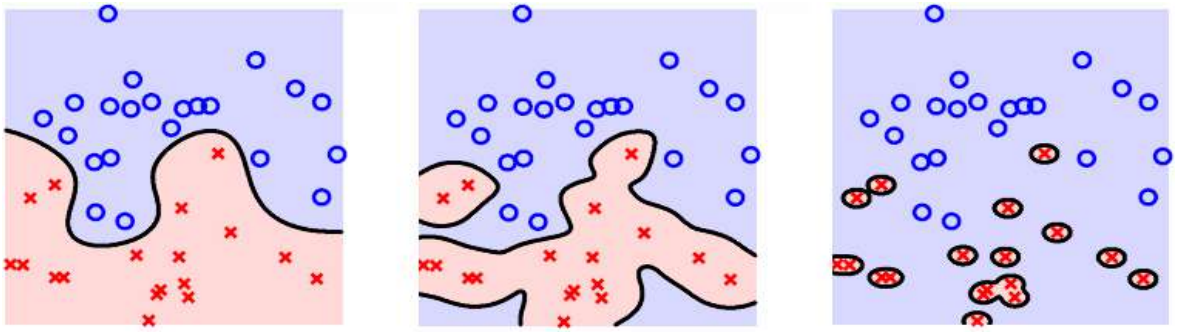
圖一是 ν 參數意義的示意圖。為了方便圖示，假設向量的維度恰為二，因此所有的向量恰好對應到平面上的其中一點，封閉空間則退化為一封閉的平面，在圖中為了簡單以橢圓示意。圖中藍色點為類別一的資料點、紅色點為假想的類別二的資料點。不難看出 ν 參數的設定其實是「錯判類別一」和「錯判類別二」之間的權衡。

此外，單類別支持向量機本質上仍是支持向量機，因此也有支持向量機部份的參數 γ 需要調整。γ 參數主要用來決定模型邊界函數的複雜程度，越高的 γ 允許越高的模型複雜度，雖然可以有效的讓模型在訓練資料上的預測結果更加準確，卻也有可能過度符合 (over-fitting) 訓練資料導致生成的模型不符合實際使用。圖二以一般的二元分類為例，比對三種不同數值的 γ 訓練出來的支持向量機模型的結果。ν 和 γ 對單類別支持向

量機而言都是極為重要的參數，第五節將會說明調整參數的方針。



圖一：單類別支持向量機 ν 參數意義示意圖 (資料來源：臺灣大學林軒田教授之機器學習課程講義)



γ = 1, 10, 100

圖二：單類別支持向量機 γ 參數意義示意圖(資料來源：臺灣大學林軒田教授之機器學習課程講義)

五、建構文風分類器

第四節已經說明機器學習的四個步驟，以下將依序說明這四個步驟在本研究中的實作方式，並註明各個步驟使用的第三方的軟體。本實驗採用的文本為金庸新版之武俠小說全集，以回目為文本的基本單位，總共 15 本、376 回。

(一) 文本資料的預處理

由於本研究選擇以「用詞習慣」作為文本的特徵，因此需要對所有文本分詞(斷詞)，將句子拆解成數個單詞的組合（如「今日天氣很好」應被分為「今日」、「天氣」、「很」、「好」四個詞），以便進行接下來的統計。實驗使用 Python3 程式語言的 jieba 中文分

詞套件[1]進行分詞。此分詞套件除了有品質優良的繁體詞彙資料庫外，還具有自動學習新詞的功能，因此作品中的專有名詞（如人名、物名）或者常見的新詞彙也有許多能被自動學習，增加分詞的準確度。

分完詞後，由於詞彙種類眾多，且並非所有詞彙都能作為文風的代表，因此需要設計一套準則，以挑選具有文風代表性的關鍵詞彙，作為文本的文風特徵。這部份對於最後模型的效果影響重大，是研究者自己需要運用背景知識和經驗的關鍵步驟。然而第三節的討論已經提過，機器學習模型的引入允許可選的關鍵詞數量大幅增加，也使得不適用的關鍵詞帶來的壞處明顯降低，因此研究者在關鍵詞的選擇上可以有更大的自由。比起確切要選哪些關鍵詞，研究者可以花更多心思在思考要選哪「類」關鍵詞。

本研究挑選關鍵詞的準則，考慮了下列三點特性：

特性 1. 關鍵詞的出現次數必須夠多，否則出現的頻率不容易穩定，難以訓練出有價值的模型。

特性 2. 本次實驗的標的為小說，並且係以「回目」為文本的單位，因此小說的劇情走向很容易影響詞彙使用的頻率。為此，選擇的詞彙應當盡量避開可能會因為劇情進度而改變出現頻率的詞彙。傳統上，這樣的詞彙多為虛詞。

特性 3. 除了「小說」外，本次的標的另外還強調了「武俠」的特性。除了虛詞外，有些詞彙雖然不是虛詞，卻是金庸寫武俠小說時的敘事慣用詞（如：劍、武林、左手、喝道、性命等），因而仍能作為金庸武俠小說作品的關鍵特徵。

因此確認最後關鍵詞的步驟如下：

(1) 以金庸超過 40 萬字的長篇小說文本為基礎，選出前 300 個出現頻率最高的詞彙。以本實驗的第 300 個常用詞彙在 270 回文本中出現了 1169 次。

(2) 人工從 300 個詞中篩去既不是金庸武俠小說的敘事慣用詞、也不是和劇情無關的虛詞，最終留下 245 個詞[2]。

特性 3 正是本次實驗中需要仰賴於研究者背景知識的部份，選擇時涉及許多主觀的認知。類似地，一開始選擇的詞彙數量（300）也是研究者自己根據經驗主觀地做出來的結果。最終留下的關鍵詞越多，演算法最終能產出的模型就越複雜，雖然區別與辨識的能力較強，但也因此需要更多的訓練資料，否則會導致模型不夠符合資料 (under-fitting)。這兩個變因都是研究者在使用機器學習時需要取捨的部份，可以透過機器學習的驗證步驟來測試出模型當前的瓶頸，並以之進行對應的修正。

[1] https://github.com/fxsjy/jieba

[2] 此 245 個詞彙列表可參見 https://github.com/b821213/Chinese-Author-Text-Style-Distinguisher/blob/master/feature_list

選擇完關鍵詞後，對於每個文本都計算出每個關鍵詞在該文本中的出現頻率（即：該關鍵詞的出現次數除以該篇文本的總詞彙數），將 245 個關鍵詞的每一個關鍵詞的出現頻率視為一維，從而產生一 245 維之向量，即為該篇文本的特徵向量。

(二) 使用單類別支持向量機訓練模型

單類別支持向量機演算法是使用 Python3 程式語言的 sklearn 套件內部的 svm 模組[3]。對於一般的二元分類問題而言，參數調整的方針是最大化最後的驗證階段的準確率 (accuracy)，然而在單類別學習問題，由於手上擁有的資料只有其中一種類別（或者只有非常少許的另一種類別），因此若以最大化驗證準確率為目標，僅須將參數 ν 設為 1 即可，然這樣的結果卻不切實際。一個權衡的方法是把方針改為令驗證階段的準確率處於一個合理的區間，如此一來既不會讓誤認目標文風的情況太嚴重，也能預期誤認非目標文風的狀況相較之下較合理。而何謂「合理」則再次仰賴於研究者的背景知識和經驗。在本實驗中，最後我們選擇的合理區間為 85%-90%。

另外，最後的實驗並沒有使用《連城訣》和《白馬嘯西風》作為訓練資料，原因是後續的實驗指出這兩者很可能並非傳統的金庸武俠小說，不能算是合理的訓練資料。

(三) 驗證模型之效能

驗證階段需要以訓練出的模型預測獨立於訓練資料的驗證資料，以評估模型在實際應用上的表現。在二元分類的問題中，當然希望驗證資料同時包含兩種分類的資料；然而在單類別學習問題中，僅有一種分類的資料，因而只能驗證模型面對該分類的準確率。

機器學習通常會從全部的資料中預先保留一部分作為驗證資料，再將剩下的資料作為訓練資料，以確保驗證資料獨立於訓練資料。然而有時原始資料可能數量已經不多，若要保留具有足夠代表性的驗證資料，就可能面臨訓練資料不足的問題；若驗證資料數量不足，則可能沒辦法精準地評估模型的效能。

為了能夠充分地利用為數不多的原始資料，機器學習中有一種驗證策略稱為交叉驗證 (cross validation)。其內容為將原始訓練資料分成若干個子集合，每次都依序取其中一個子集合作為驗證資料，剩餘的子集合作為訓練資料用來訓練模型，再用模型來預測方才保留的驗證資料。待所有子集合都已經被作為驗證資料過後，便能算出一個近似於使用所有原始訓練資料進行訓練後的模型的準確率。

本實驗以書目為集合分割的單位，在訓練資料上進行交叉驗證，以之獲得模型在「屬於金庸武俠小說」的文本上的準確率。如上所述，由於《連城訣》和《白馬嘯西風》並

[3] 事實上目前 sklearn 內部的 SVM 模型底層皆為另一公開的多語言套件 LIBSVM。

未被列入訓練資料，因此也不會在驗證階段進行驗證。

(四) 實際預測資料

實際運用階段不會再有「不屬於目標文風」缺乏代表性文本的問題，可以使用模型測試任意的文本。在本實驗中，「屬於目標文風」的部份由於文本數量不多，且已經盡數作為訓練資料，就不再另外測試這類的文本，而直接以交叉驗證的結果作為模型在「屬於目標文風」上效能的展示。除此之外，實際選擇了下列幾份文本進行測試：

(1) 被排除在訓練過程外的《連城訣》、新版《白馬嘯西風》。

(2) 原版《白馬嘯西風》。

(3) 同屬武俠小說大家的梁羽生武俠小說作品 16 部[4]。

(4) 其他武俠小說作品 3 部[5]。

(5) 非武俠小說作品 2 部[6]。

選擇大量武俠小說的原因在於，當預測「不屬於金庸武俠小說」的文本，最容易被誤判的應為用詞、劇情都更接近的其他作者之武俠小說；武俠小說中絕大部分為梁羽生的作品則是由於在實驗中無意間發現梁羽生的部份作品被模型預測和金庸之武俠小說有相當高的相似性，因而更大規模地測試梁羽生的其他作品。

六、「金庸之武俠小說」文風分類器之實驗結果

以下皆是以金庸武俠小說全集除《連城訣》和《白馬嘯西風》外之文本組成的訓練資料在 $\gamma = 0.01$、$\nu = 0.5$ 的設定訓練模型之結果：

(一) 訓練資料之交叉驗證

表一是基於訓練資料交叉驗證之結果。令人驚訝的是其中多數的準確率非常高，尤其在超過 40 回的超長篇小說，除《笑傲江湖》和《鹿鼎記》外準確率都高達 90% 以上，其中《射鵰英雄傳》和《神鵰俠侶》高達 97.5%，《倚天屠龍記》更高達 100%，足見此模型在預測目標群集文本上有相當高的準確性和穩定性。

(二) 《連城訣》與《白馬嘯西風》

[4] 《龍虎鬥京華》、《草莽龍蛇傳》、《塞外奇俠傳》、《七劍下天山》、《江湖三女俠》、《白髮魔女傳》、《萍蹤俠影錄》、《冰川天女傳》、《還劍奇情錄》、《散花女俠》、《女帝奇英傳》、《狂俠天驕魔女》、《風雲雷電》、《牧野流星》、《彈指驚雷》、《武林天驕》

[5] 慕容美《七星劍》、平江不肖生《江湖奇俠傳》、古龍《絕代雙驕》

[6] 曹雪芹《紅樓夢》、沈從文《邊城》

表二是對《連城訣》、《白馬嘯西風》（新版）與原版《白馬嘯西風》的預測結果。相較於其他金庸之武俠作品，《連城訣》預測的正確率明顯地較原先低了許多，而《白馬嘯西風》無論是經過大幅修訂的新版，還是最早的連載原版，竟都被模型認為完全非金庸之武俠小說作品，第七節將對此進一步的說明與討論。

(三) 其他武俠小說

表三是對梁羽生作品在內的所有其它武俠小說的預測結果，其中《龍虎鬥京華》至《女帝奇英傳》乃是依照梁羽生寫作時間排序的連續作品。注意到這裡的預測正確代表模型預測為非金庸之武俠小說。由資料可見模型對於非金庸武俠小說的預測也相當出色，絕大部分武俠小說都被徹底地否定為金庸之武俠小說作品，唯有梁羽生的《白髮魔女傳》中竟有高達 50% 的回目被預測為金庸所作之武俠小說。若進一步觀察《白髮魔女傳》附近成書的作品，會發現從《七劍下天山》以後，預測錯誤的比例逐漸上升，到《白髮魔女傳》為最高峰，接著在《萍蹤俠影錄》也不低的錯誤率後，接下來的作品又回到接近 100% 的正確率。這背後也許意味著梁羽生在某段時間的寫作風格有接近金庸之可能，然而還需要更多的後續研究來支持這種觀點。

(四) 非武俠小說

如同表四所列的數據，所有非武俠小說都被正確地預測為非金庸武俠小說。由於相較於其他武俠小說，非武俠小說被預測正確較為預料之內，因此沒有挑選太多的此類小說進行辨識實驗。

表一：金庸武俠小說訓練資料之交叉驗證結果

書名	預測正確回數	總回數	正確率
《書劍恩仇錄》	18	20	0.900
《碧血劍》	18	20	0.900
《射鵰英雄傳》	39	40	0.975
《雪山飛狐》	10	10	1.000
《神鵰俠侶》	39	40	0.975
《飛狐外傳》	19	20	0.950
《鴛鴦刀》	1	1	1.000
《倚天屠龍記》	40	40	1.000
《天龍八部》	44	50	0.880
《俠客行》	17	21	0.810
《笑傲江湖》	30	40	0.750
《鹿鼎記》	37	50	0.740
《越女劍》	1	1	1.000
總計	313	353	0.887

表二：《連城訣》與《白馬嘯西風》預測結果

書名	預測正確回數	總回數	正確率
《連城訣》	5	12	0.417
《白馬嘯西風》	0	11	0.000
（舊版）《白馬嘯西風》	0	9	0.000

七、淺探《白馬嘯西風》之預測結果

在第六節第二部份我們看到《連城訣》與《白馬嘯西風》在所有其它作品都預測極為正確的情況下，在正確率上有明顯的落差，尤其《白馬嘯西風》無論是原版還是新版皆被預測為 0%，難以將錯誤歸咎於編修的過程。底下將針對《白馬嘯西風》逐層檢視，根據此一模型的物理意義上，找出《白馬嘯西風》被預測錯誤的原因。

(一) 透過不同參數的模型檢測對應資料點與分類子空間之距離

第四節對單類別支持向量機的說明曾經簡單介紹學習出的模型在幾何中的直觀意義，以及各個參數對最後模型的影響。事實上透過觀察資料點在不同參數產生的模型中的預測結果，可以給出資料點在高維空間中的可能位置。舉例而言，如果《白馬嘯西風》的文本其實和金庸的其他武俠小說文本很類似，只是「不夠類似」，那麼在高維空間中這些點的位置應該是在模型分類器邊界的邊緣。如此一來，當調整 ν 和 γ 的值，讓模型對應的分類子空間在形狀、大小或者方向上有些改變時，應該至少會有幾回文本會因為這些改變而進到分類子空間的內部。換言之，資料點在不同參數下預測的結果，能讓使用者推斷出資料點和目標群集的距離，抽象的意義即為資料點和目標群集的相似度。

表三：其他武俠小說預測結果

書名	作者	預測正確回數	總回數	正確率
《龍虎鬥京華》	梁羽生	13	13	1.000
《草莽龍蛇傳》	梁羽生	12	12	1.000
《塞外奇俠傳》	梁羽生	28	28	1.000
《七劍下天山》	梁羽生	27	30	**0.900**
《江湖三女俠》	梁羽生	36	48	**0.750**
《白髮魔女傳》	梁羽生	16	32	**0.500**
《萍蹤俠影錄》	梁羽生	23	32	**0.719**
《冰川天女傳》	梁羽生	40	40	1.000
《還劍奇情錄》	梁羽生	14	14	1.000
《散花女俠》	梁羽生	35	36	0.972

《女帝奇英傳》	梁羽生	32	32	1.000
《狂俠天驕魔女》	梁羽生	120	120	1.000
《風雲雷電》	梁羽生	66	66	1.000
《牧野流星》	梁羽生	64	64	1.000
《彈指驚雷》	梁羽生	20	20	1.000
《武林天驕》	梁羽生	17	17	1.000
《七星劍》	慕容美	28	28	1.000
《江湖奇俠傳》	平江不肖生	43	43	1.000
《絕代雙驕》	古龍	126	126	1.000

表四：非武俠小說

書名	作者	預測正確回數	總回數	正確率
《紅樓夢》	曹雪芹	120	120	1.000
《邊城》	沈從文	21	21	1.000

為此我們以 ν= {0.1, 0.2, …, 1.0} 和 γ = {0.0001, 0.001, …, 1.0} 的所有參數組合（共 50 種）訓練模型，其中僅有 3 種在驗證階段準確率超過 80%；而這 3 種參數組合所形成的模型中，《白馬嘯西風》的兩種版本皆仍是 0% 被判為金庸之武俠小說作品，與此同時《白髮魔女傳》在這 3 個模型中最低也有 46.875% 被預測為金庸之武俠小說作品。由此可知，在空間中確實《白馬嘯西風》的各回都離模型對應的子空間非常遙遠，反而《白髮魔女傳》有將近半冊幾乎穩定在模型的子空間中。

(二) 以詞頻檢視《白馬嘯西風》和其他金庸武俠文本之差異

由於模型是基於金庸武俠小說全集除《連城訣》與《白馬嘯西風》以外之文本作為訓練資料，與選出的 245 個關鍵詞訓練而成，表五羅列此 245 個關鍵詞在《白馬嘯西風》與訓練文本中的頻率變化率，並將變化率為正和變化率為負的詞彙分別列出變動最大的前 50 名。一個詞彙 w 的頻率變化率的定義為

$$(p_1(w) - p_2(w)) \Big/ p_2(w)$$

其中 $p_1(w)$ 為 w 在《白馬嘯西風》中的頻率，$p_2(w)$ 為 w 在訓練文本中的頻率。

表五：《白馬嘯西風》與訓練文本平均關鍵詞詞頻差異

#	詞彙	頻率變化率	#	詞彙	頻率變化率	#	詞彙	頻率變化率	#	詞彙	頻率變化率
1	沒有	2.48	26	這是	0.72	1	沖	-0.96	26	使	-0.58
2	很	2.41	27	不會	0.70	2	派	-0.95	27	心下	-0.57
3	這樣	1.81	28	可是	0.70	3	均	-0.94	28	更	-0.57
4	聽到	1.49	29	知道	0.70	4	天下	-0.93	29	對方	-0.57
5	這裡	1.35	30	的	0.65	5	兒	-0.93	30	自	-0.56
6	之中	1.17	31	之後	0.65	6	罷	-0.92	31	當下	-0.56
7	兩個	1.14	32	心中	0.64	7	與	-0.92	32	甚	-0.53
8	忽然	1.09	33	便是	0.62	8	即	-0.88	33	說話	-0.53
9	突然	1.06	34	身上	0.61	9	這位	-0.79	34	知	-0.53
10	死	1.06	35	下來	0.59	10	如此	-0.76	35	倘若	-0.52
11	聲音	1.05	36	十分	0.56	11	功夫	-0.75	36	二人	-0.52
12	地	1.03	37	用	0.55	12	既	-0.75	37	如	-0.52
13	這時	0.97	38	過去	0.51	13	請	-0.71	38	女子	-0.51
14	著	0.89	39	卻是	0.48	14	眾	-0.70	39	罵	-0.50
15	誰	0.89	40	起來	0.48	15	以	-0.69	40	事	-0.49
16	走	0.89	41	在	0.48	16	字	-0.68	41	無	-0.48
17	地下	0.87	42	還是	0.48	17	於	-0.68	42	做	-0.48
18	之間	0.81	43	之下	0.46	18	老	-0.67	43	忙	-0.47
19	會	0.79	44	這個	0.46	19	劍	-0.65	44	待	-0.46
20	於是	0.79	45	大家	0.45	20	當即	-0.64	45	不知	-0.44
21	敵人	0.78	46	一陣	0.43	21	出	-0.63	46	不由得	-0.43
22	一會	0.77	47	一驚	0.43	22	今日	-0.62	47	才	-0.42
23	一個	0.76	48	一聲	0.41	23	如何	-0.62	48	所	-0.42
24	沒	0.75	49	了	0.41	24	之	-0.61	49	兵刃	-0.41
25	見到	0.73	50	跟	0.41	25	笑	-0.60	50	之時	-0.41

仔細觀察變化率為正（出現頻率變高）的詞彙和變化率為負的（出現頻率變低）的詞彙，可以發現變化率為正的多半為日常生活中會用到的白話詞彙（如：沒有、很、這樣、聽到、這樣、還是等），而變化率為負的則多為寫作甚至古文中較長使用的文言詞彙（如：均、即、既、甚、之時等），且越文言下降幅度越大。若進一步將具有替代性的詞彙列出（表六），更能明確地看到用字風格明顯偏向更白話的說法。

表六：具有替代性詞彙之變化率消長

詞彙一	頻率變化率	詞彙二	頻率變化率
沒有	2.48	無	-0.48
很	2.41	甚	-0.53
這裡	1.35	此	-0.38
這樣	1.81	如此	-0.76
這時	0.97	此時	-0.32
心中	0.64	心下	-0.57

(三) 人工分析文本背景

為了更精確了解文本的背景和金庸其他作品有何不同，我們親自閱讀了《白馬嘯西風》的內容。相較於金庸的其他武俠小說，《白馬嘯西風》有下列幾點特色：

1. 故事背景在新疆，角色也多為新疆人，不同於其他作品多為中原人。
2. 主角不再是有奇遇或者武功高強的男子，而是始終受人保護的少女。
3. 主軸情感由江湖義氣轉為個人愛情與類親情之情感。
4. 敘事之旁白更貼近觀察者之第一人稱視角，全知的第三人稱視角頻率較低。

雖然未有進一步的量化分析，但閱讀上確實能感受到和其他金庸武俠作品的巨大差異。至此，將《白馬嘯西風》全文皆判斷為「非金庸之武俠小說文本」實屬合理之舉──《白馬嘯西風》雖然是金庸所作，但已經不全然是一部金派武俠小說了。

八、結論

本研究試圖透過引入機器學習的工具來拓展既有以用詞習慣來辨識文風的方法，在擁有足夠多目標文風群集文本的前提下，大幅簡化關鍵詞的選擇難度、提昇可選關鍵詞的數量，並能描述遠比傳統統計模型假設更為複雜的用詞習慣，建立了自動化文風辨識的先決條件。在以「金庸的武俠小說文本」為目標群集進行的實驗，本研究提出的方法獲得了巨大的成功，並成功地預測出其中與眾不同的文本(《連城訣》與《白馬嘯西風》)，提供重要的線索以供後續著手研究。

在進一步研究《白馬嘯西風》的過程中，還看到本研究提出的方法持續提供抽象概念上的相似度提示；基於詞頻抽取特徵的特性也讓所有轉換過的資料仍保有原先文本的直觀屬性，讓研究者能夠在得知預測結果後反過來從詞頻中尋找文本的特別之處，更彰顯了此方法具有明確物理意義所能帶來的優點。

本研究並無意取代傳統文風辨識的方法，事實上在僅須針對特定文本進行研究，或

者研究者擁有的資料數量不足，又或者目標文風難以應用「用詞習慣的代表」的情況下，傳統文風辨識的方法相較於本研究提出的方法仍有極大的優勢。本研究的核心貢獻在於能夠短時間內在大量的資料上，僅基於少量客觀的事實，自動提供粗糙但足夠可信的分類，藉以提示研究者可能值得研究的目標以及部份資料的獨特之處。希望越來越精準且快速的數位化工具的使用，能夠讓極為珍貴的研究人力與時間投資在更有意義和發展性的研究議題上，加速整體人文學界的進步、開拓人文領域的新可能。

誌謝

本文為國立臺灣大學夏季學院「認識數位人文」課程的研究成果，該課程的授課老師為項潔教授、岳修平教授、陳光華教授、闕河嘉教授、唐牧群教授，感謝各位教授的熱心指導與建議。

參考文獻

Chan, B.-C. (1986). *The Authorship of the Dream of the Red Chamber based on a Computerized Statistical Study of Its Vocabulary*. Hong Kong: Joint Publishing Co Ltd.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., and Williamson, R.C. (1999). *Estimating the Support of a High-Dimensional Distribution*. Technical Report, Microsoft Research, MSR-TR-99-87. Retrieved Oct. 20, 2016 from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-99-87.pdf.

何光國。2002。從漢語白話文虛字「的、地、得」的運用論作者寫作個性。傳統中國文學電子報，131。

余清祥。1998。統計在《紅樓夢》的應用。政大學報，76，303-327。

杜協昌。2012。利用文本採礦探討《紅樓夢》的後四十回作者爭議。第四屆數位典藏與數位人文國際研討會。

趙岡、陳鍾毅。1980。紅樓夢研究新編。臺北市：聯經出版社。

報紙之「移工」相關詞彙選用變化：以《自由時報》為例[1]

高竹瑩[*]、葉奕辰[**]、林荷鎧[***]、黃志揚[****]

摘　要

在臺灣，經常可見報紙使用「外勞」、「外傭」等詞彙，報導與外籍勞工相關新聞，甚至將其國籍與性別、工作等結合，形成印尼籍女看護、菲傭等專用術語。近年來，台灣的東南亞籍移工呼籲以「移工」替代「外勞」等歧視性字眼，報紙也在此浪潮下新增「移工」、「外籍勞工」等指稱詞。然而，報導用詞的選擇改變並不能等同於移工去污名化的成功，因此本研究以《自由時報》為研究對象，研究報紙如何呈現東南亞籍移工議題。

本研究以語料庫導向取徑（corpus driven approach）為基礎，結合庫博中文語料庫分析工具（CORPRO pre-alpha）與批判性文本分析作為研究方法。本研究的兩個研究問題為 (1)在 2001 年到 2016 年 8 月 13 日這段時間內，「外勞」及其替代詞彙在報紙上逐年使用的趨勢為何？ (2) 在 2001 年到 2016 年 8 月 13 日這段時間內，報導內容是否隨著指稱詞彙的不同而有所轉變？

研究成果發現，臺灣媒體在 2001 年到 2016 年這十六年之間，雖然仍以「外勞」一詞最常被使用，但「移工」、「外籍移工」等詞彙的使用頻率有增加的趨勢。進一步檢視文本內容後發現，「外勞」在僱傭關係、政策、犯罪等面向上被使用的情形較其他替代詞彙普遍。而雖然亦有越來越多的報導透過記錄、理解這群移工來形塑他們在台灣的形象，但是舊有的移工呈現依然未隨新指稱詞的納入而轉變。

近年來已有許多探討東南亞籍移工權益的研究，然而多數聚焦在政策面向。本研究則試圖以《自由時報》的移工再現為重點，透過分析「外勞」及「移工」等詞彙各自的搭配詞揭示報紙潛在態度，並能檢視移工去污名化的訴求是否得到適當回應。

關鍵字：外勞、移工、外籍勞工、外籍移工、國際移工

---

[1]本文作者感謝北二區的通識教育課程「認識數位人文」的教授們，項潔、闕河嘉、岳修平、陳光華、唐牧群對於本研究的啟發；特別感謝闕河嘉指導本篇論文的發展。

[*] 國立臺灣大學圖書資訊學系學生，Email: b02106045@ntu.edu.tw。
[**] 國立臺灣大學圖書資訊學系學生，Email: b03106045@ntu.edu.tw。
[***] 國立臺灣大學外國語文學系學生，Email: b04102030@ntu.edu.tw。
[****] 臺北市立大學資訊科學系學生，Email:U10416001@go.utaipei.edu.tw。

A Corpus Analysis of Representation of South-Asia Migrant Workers in Taiwan Newspaper[1]

Chu-ying Kao[*], I-chen Yeh[**], Ho-shiuan Lin[***], Chih-yang Huang [****]

Abstract

This research examines the word-choice of representing Southeast-Asia migrant workers in the *Liberty Times*, one of Taiwan's mainstream newspapers. South-Asia migrant workers have called for a change in how media refers to them, especially in the way they are mentioned. They urge the public to use politically-correct nouns such as "migrant workers" (移工) to substitute discriminating names such as "foreign labor" (外勞). We chose the *Liberty Times* as our targeted press for its self-claimed "Taiwan first, Liberal First" newspaper aim. This research first analyzed the trend, between 2001 and 2016, of different word used when referring to Taiwan's Southeast-Asia migrant workers, and further critically examined the coverage when different word choices are employed. The result shows that *Liberty Times* has increased the use of suggested politically-correct words. However, these Southeast-Asia workers have not been continually reported with certain ideology, despite its attempt of adding documentary reports. This research provides understanding of the ways in which Taiwan mainstream newspaper responds to the call of unfair representation of migrant workers in Taiwan.

Keyword: migrant worker, Liberty Times, newspaper, representation, human right

[1] We appreciate professors who have guided us thorough the summer course 'Introduction to Digital Humanity', Jieh Hsiang, Ho-chia Chueh, Siou-ping Yue, Kuang-hua Chen and Muh-chyun Tang. Special thanks goes to Ho-chia Chueh for her help in the development of this research

[*] Undergraduate Student, Department and Graduate Institute of Library and Information Science, National Taiwan University. Email: b02106045@ntu.edu.tw.

[**] Undergraduate Student, Department and Graduate Institute of Library and Information Science, National Taiwan University. Eail: b03106045@ntu.edu.tw.

[***] Undergraduate Student, Department of Foreign Languages and Literatures, National Taiwan University. Email: b04102030@ntu.edu.tw.

[****] Undergraduate Student, Department of Computer Science, University of Taipei. Email:U10416001@go.utaipei.edu.tw.

從改革開放到富國強民：
1980 年到 2010 年人民日報風格變化研究

梁家安[*]、余清祥[**]、何立行[***]

1978 年中共第十一屆三中全會確立了「對內改革、對外開放」的戰略決策，自此開啟了中國的「改革開放」時代。於是，從上世紀 80 年代到本世紀第一個十年，自晚清以來備受內憂外患困擾的中國終於得到一段較長時期的安定，並在這三十年中快速成長為經濟大國。有些人開始以「盛世」、「強國」形容現在的中國，但也有人注意到中國的經濟發展並未像其他許多新興國家一樣帶動政治的民主化。中國的社會顯然因為改革開放而更多元，但政府的意識形態和思維有什麼樣的變化？或許我們可以從中共機關報《人民日報》的風格變化窺見。

本研究從生物多樣性的角度切入，以 1980 至 2010 年中國《人民日報》的文章報導為素材，藉由判讀各年度的常用字彙及關鍵詞，探究改革開放與思想觀念變化間的關連。

比較《人民日報》最常出現的單字、雙字詞，發現兩者的型態非常不一樣，兩兩年度間的重複程度差異頗大。以各年度出現次數最多的前十名單字、雙字詞為例，在此僅列出 1993～2000 年的結果，常見單字的年度重複程度非常高（表 1），代表常見單字的變化不大，然而前十大雙字詞的結果很不一樣（表 2），相鄰兩個年度的前十大雙字詞為未必有較高的重複率，而且重複率似乎並不存在絕對遞增或絕對遞減的函數關係。白話文中的重要觀念多以雙字詞（而非單字）呈現，雙字詞重複率不高代表觀念變化相對頻繁，以常見雙字詞為研究標的，可探究哪些想法始終如一、哪些想法隨時間逐漸發生或消逝。

[*] 國立政治大學統計系研究生，Email: 104354031@nccu.edu.tw。
[**] 國立政治大學統計系教授，Email: csyue@nccu.edu.tw。
[***] 國立清華大學中文系專案經理暨助理教授，Email: lillianlhho@gmail.com。

表 1、《人民日報》前十大單字兩兩年度的重複字數（對角線以下省略不表示）

年度	1994	1995	1996	1997	1998	1999	2000
1993	9	9	9	9	8	8	9
1994	---	10	9	10	9	9	10
1995	---	---	9	10	9	9	10
1996	---	---	---	9	10	9	9
1997	---	---	---	---	9	9	10
1998	---	---	---	---	---	9	9
1999	---	---	---	---	---	---	9

表 2、《人民日報》前十大雙字詞字兩兩年度的重複字數（對角線以下省略不表示）

年度	1994	1995	1996	1997	1998	1999	2000
1993	7	8	9	6	8	7	9
1994	---	7	9	7	8	6	8
1995	---	---	8	5	8	6	8
1996	---	---	---	7	9	7	9
1997	---	---	---	---	6	4	6
1998	---	---	---	---	---	6	9
1999	---	---	---	---	---	---	7

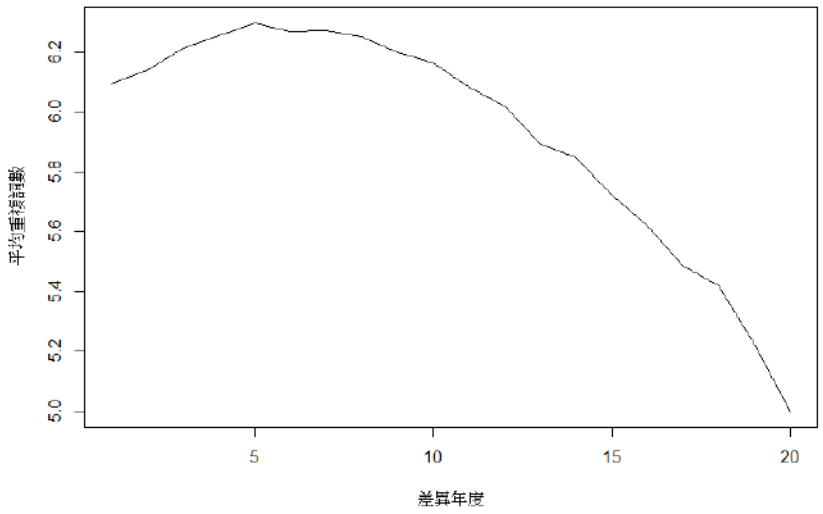


圖 1、《人民日報》年度間前十大雙字詞的平均重複詞數

這種雙字詞隨時間產生、消滅的現象，可以解釋為思想及觀念的變遷，和生物物種隨環境變遷而演化成更適合生存的想法類似，因此本研究將引進生物多樣性的想法，藉由辨識具有代表性的關鍵物種(Keystone Species)，探索《人

民日報》各年度的生態系及其特徵。雙字詞有不少特性與物種(Species)接近，以上述兩兩年度間的前十大雙字詞為例，間隔年數與的前十大雙字詞的重複字數雖然不是絕對遞增或遞減，但間隔時間自 5 年開始，呈現逐年線性遞減的現象（圖 1），從間隔 5 年平均字數約 6.3 字降至間隔 20 年的 5.0 字。

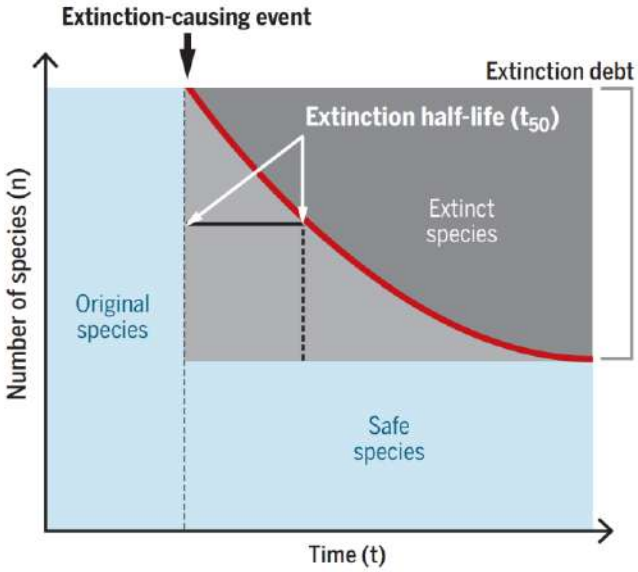


圖 2、物種絕種的速度（來源：2016 年七月 Science 期刊）

這個現象和生物界的物種減少（或滅絕）類似。美國知名的《科學》雜誌在 2016 年 7 月有篇研究觀察 1990 年代至 2015 年的植物，發現約 39 萬種植物中有 5 萬種植物可能已經滅絕，其滅絕趨勢如圖 2。小型哺乳類動的物種滅絕趨勢也類似，只有下降曲線形狀為線性或指數有差異，及下降速度稍快一些。

初步分析發現幾個有趣的結果。首先，有些文革色彩較重的詞彙如「階級」、「革命」僅在 80 年代初期上榜，之後完全消失。「我們」在 90 年之後不再上榜。有的詞彙晚出，但 90 年代後幾乎年年上榜，如「發展」。類似的還有「經濟」，1991 年首次上榜後，只在香港回歸 1997 年、北約轟炸南斯拉夫中國使館的 1999 年、法輪功成員除夕夜在天安門自焚的 2001 年、上海合作組織憲章發表的 2002 年、及中國主辦奧運和殘障奧運的 2008 年被擠出榜外。可以說這些詞彙在 90 年代後已成為《人民日報》生態系中的基石物種。有些雙字詞，如「國家」、「人民」、「主義」、「問題」等，在 30 年間詞頻一直相當高，這些詞彙類似基石物種的特徵尤其可以在它們偶然落榜時展現出來：它們落榜的年份通常可以發現較多具新聞時效性的專有名詞上榜，顯示外來物種在一定時間內對生態系造成短暫衝擊。但「發展」首次出現的 1988 年較為反常，「人民」、「主義」兩個基

石物種雙雙落榜，卻沒有較明顯具新聞時效性的詞彙上榜。此外，「人權」、「美國」也是《人民日報》的基石物種，其意涵、行文脈絡與話語功能則有待檢視。還有一個值得注意的現象是：「憲法」、「權利」在 1982 年為常見十大詞彙之內，之後卻不再進榜。

再看根據單字出現頻率計算所得熵(Entropy)值的變化（圖 3），熵與生物多樣性有關，數值愈大代表愈多元（或開放？）。《人民日報》的熵值從 1980 年開始逐年上升，至 1989 年達到高峰而後急降，猜測是 1989 年六四天安門事件造成中共內部緊縮，直至 1997 年香港回歸之後方有較明顯的回升。

圖 3、歷年《人民日報》的熵值

綜合以上觀察，可以看到改革開放後的三十年間，《人民日報》的風格變化有穩定的走向，但也有一些斷裂和曲折需要進一步探究。例如，今天當大家提起改革開放，大多數人自然而然想到的便只是經濟的快速成長，英文也直譯為 Chinese economic reform。然而，嚴格排除政治改革是否真為改革開放的初衷？以數位人文研究方法探究三十年間《人民日報》風格的變化，至少「憲法」、「權利」的消失和「發展」、「經濟」晚出，及熵值的起落，顯然都見證著經濟起飛的背後仍有其他截斷的伏流。接下來將延續上述研究，使用前三十大（或前五十大）常見的雙字詞，引進空間統計（Spatial Statistics）、核修勻法（Kernel Estimation）、社群網絡（Social Network）等統計方法，連結雙字詞間的關連，判斷哪些雙字詞為關鍵物種，並建構《人民日報》在各年度的「生態結構」，探討發展經濟如何從為改革開放的手段轉變為目的，多元發展的可能性又是如何被富國強民的單一目標所取代。

關鍵字：文字採礦、改革開放、人民日報、生物多樣性、關鍵物種

Style Change of *People's Daily* in 1980-2010 after Chinese Economic Reform

Ka-on Leong[*], Ching-syang Jack Yue[**], Li-hsing Ho[***]

Abstract

The Chinese Economic Reform that started from 1978 not only has brought the world a new super power but also changed the Chinese society drastically. As the mirror of the society, the language used in People's Republic China also reflects the change and experienced an on-going process of transformation. In this study, we propose to use the new digital humanity techniques to study the language change of contemporary China. In particular, we adapt the concept in species diversity to explore the style change in language. The specific topics of interest include the extinct species, the emerging species, and the connection between species. First, we treat the top 100 phrases appearing the most often every year as the dominated species and use them as the study objects. The top phrases off the list can be treated as the extinct species and they can provide useful information about, such as, the attributes of style change in language. On the other hand, the information of emerging phrases (or new phrases) can be used to further verify the change of style.

Keywords: text mining, species diversity, keystone species, economic reform, People's Daily

[*] Master Student, Department of Statistics, National Chengchi University. Email: 104354031@nccu.edu.tw.

[**] Professor, Department of Statistics, National Chengchi University. Email: csyue@nccu.edu.tw.

[***] Project Manager & Adjunct Assistant Professor, Department of Chinese Literature, National Tsing Hua University. Email: lillianlhho@gmail.com.

從國家至上到民之所欲：
中華民國總統直選前後就職演說風格變化研究

黃于珊[*]、何立行[**]、余清祥[***]

摘　要

中華民國從 1948 年的第一任總統，今年（2014 年）產生了第十四任總統，早期透過國民大會代表間接選出總統，1996 年後轉變為由公民直接投票選出。總統選舉制度的改變（總統直選），背後當然代表臺灣整體社會氛圍的變化，不止在於政治，對於臺灣的國際與歷史定位、政府與民眾的關係，過去這二十多年來有不少本質上的演變。總統直選對總統候選人及主政者的規範更為直接，如同臺灣的民意代表必須面對選區選民，社會氛圍改變了總統面對群眾的態度，從和民眾有距離的國家領導人、到傾聽民意的全民總統，總統的態度和角色由歷屆總統的當選演講稿裡，不難看出其中的微妙變化。

本研究透過中華民國十四屆總統當選人的就職演講，從演講稿的風格變化中，探究那些重要觀念隨時間轉變（消逝、產生），哪位總統提出開創性的想法，具有非常不同的典範轉移及新思維。十四屆就職演講稿共有約 46,000 字、將近 2,000 個不同字彙，六位總統的風格由總字數、字彙數、TTR （Type/Token Ratio，平均每個字出現新字彙比例；或可譯為標記類型的比例）判斷，風格差異頗大。不過，因為中文白話文的特色大多透過雙字詞描述重要觀念，所有總統的常用單字相去不遠，因此本研究將以雙字詞為分析的基本單位。

將雙字詞視為基本分析單元，首先需判定哪些雙字詞具有代表性，這個步驟可透過基本統計（出現頻率最高者）配合人工挑選（專家意見），由於全文字數不超過五萬字，預期所需時間不會太長。另一種可能是藉由統計工具挑選有意義的雙字詞，以數量方法決定哪些詞彙與統計觀念一致，可減少主觀認定的疑慮，避免使用者影響分析結果。例如：結合齊夫法則(Zipf's Law)及離群值，先以出現

[*] 政治大學金融系研究生，Email: 104352027@nccu.edu.tw。
[**] 清華大學中文系博士後研究員，Email: lillianlhho@gmail.com。
[***] 政治大學統計系教授，Email: csyue@nccu.edu.tw。

頻率多寡找出前三十名（或前五十名）雙字詞，預期這些雙字詞的出現次數與齊夫法則有關，圖 1 為蔡英文總統演講稿的常見雙字詞，大約前一百個詞彙的出現頻率與排序(Rank)倒數成正比，可藉此標示不符合倒數關係的雙字詞，並進一步判定是否可作為代表蔡英文總統演講稿的指標變數。

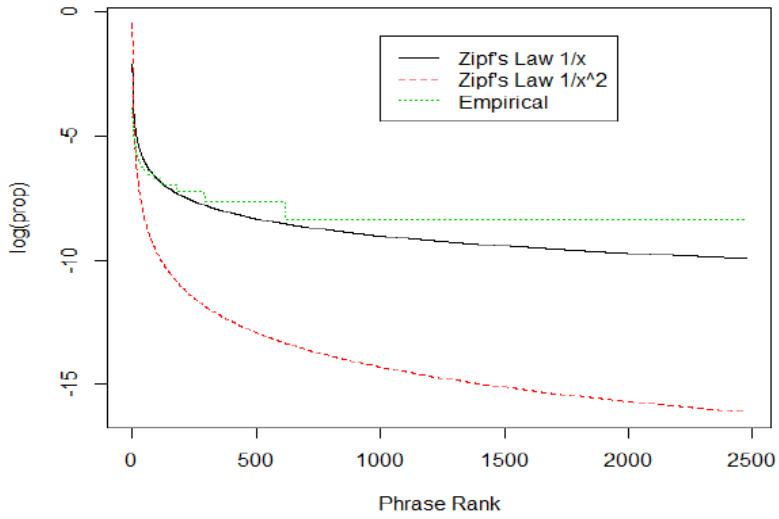


圖 1、蔡英文總統演講稿最常出現詞彙與齊夫法則

確認雙字詞之後，將先以關連性(Association)分析為主，包括群聚分析(Cluster Analysis)及社群網絡(Social Network)等，判定每位總統的詞彙間之關連性，作為描述演講稿的特徵。除了上述常見的關連性分析方法，本研究也將套用估計密度函數的方法，以類似空間分析(Spatial Analysis)的想法，將雙字詞之間的「距離」列入考量，作為雙字詞間關連性的測量依據。例如：套用核估計法(Kernel Estimation)，以每個詞彙為中心，測量 50 個字之內其他詞彙的出現次數，並以核函數為加權依據，得出的數值愈高、代表兩個詞彙的關連性愈高。這些結果可判斷某些詞彙歷屆總統都會使用，但其意義是否保持不變，像是「我們」、「民主」等雙字詞。

初步分析發現幾個有趣的結果。可以看到歷屆總統間某些觀念的消長，例如：蔣介石總統的「大陸」及「反共」等之後都未再出現，「三民主義」在蔣經國總統之後也消失了，「臺灣」一詞從李登輝總統之後開始都會提到，「歷史」及「萬歲」只有陳水扁總統使用。以典範轉移的角度思考，李登輝總統的開創性最強，除了大家熟知的「經營大臺灣、建立新中原」外，幾個常見詞彙後續陳水扁、馬英九、蔡英文三位總統都持續使用。

關鍵字：文字採礦、總統演講稿、社群網路、估計密度函數、關連性分析

Style Change of the Presidential Inaugural Address of the Republic of China before and after the Direct Presidential Election

Yu-shan Huang[*], Li-hsing Ho[**], Ching-syang Jack Yue[***]

Abstract

Up to 2016, there are 14 elected presidents of the Republic of China since 1948 and the presidents were elected via the National Assembly (i.e., indirect election) before 1996. The transform of presidential election indicates the change of international role for Taiwan, as well as the relationship between Taiwan's government and people, for the past 20 years. The direct presidential election implies that the presidential candidates need to meet the people and listen to their needs in person. It would be interesting to see whether the inaugural addresses can reflect the change of presidents' role and attitude before and after the direct election. We study the style change of the inaugural addresses of all 14 presidents via the perish and emerge of key phrases from the speeches. In addition to the vocabularies, TTR (Type/Token Ratio), and Zipf's law, we also adapt the notion of species diversity for data analysis. In particular, we apply the techniques of social network and density estimation, to detect the relationship between key phrases.

Keywords: text mining, presidential inaugural address, Zipf's Law, density estimation, social network

[*] Master Student, Department of Money and Banking, National Chengchi University. Email: 104354031@nccu.edu.tw
[**] Project Manager & Adjunct Assistant Professor, Department of Chinese Literature, National Tsing Hua University. Email: lillianlhho@gmail.com.
[***] Professor, Department of Statistics, National Chengchi University. Email: csyue@nccu.edu.tw.

「中國大陸研究」期刊主題分析：自動化方法之應用

邵軒磊*、曾元顯**

摘　要

中國大陸做為社會科學學術研究對象，隨學科變遷以及兩岸情勢發展，近年知識界已逐漸有「反思中國研究」之呼聲。本文應用文字探勘於「中國大陸研究」，以探索此一學科各種研究主題與脈絡的可能性。本文主要的研究問題意識有二：其一、能否使用「共現字分析」（Co-word Analysis），並以程式自動化大量歸納社會科學形式之論文主題與共同關鍵字？其二、若前述問題可以解決，那麼近年來中國大陸研究相關的「關鍵字」有什麼？

「共現字分析」乃計算任意兩文件的相似度，從而將相似的文件歸類（cluster）。此一研究法在科學計量學（Scientometrics）、文本分析（Content Analysis）或文字探勘（Text Mining）中皆有相關應用。其可便於瞭解待分析文件中包含的各種主題概念，從而對各個主題與各篇文章的欄位屬性（如：作者、機構、國家、出處、年代等）進行交叉分析，用以透露文獻內含的研究主題、各主題的影響主力（個人、機構、國家）、研究社群之聚落分佈、主題趨勢的變化。[1]以往類似的研究，是以人力方式使用數量比率的方法進行研究。[2]

較諸傳統以人工建構領域知識或分類架構的方法，本文使用之概念，在於提供由下而上（bottom-up）、資料驅動（data driven）、證據為主（evidence-based）的研究線索，能補充甚至對照專家之主觀體驗，從而得到啟示。本研究使用CATAR（Content Analysis Toolkit for Academic Research）程式。[3]對文章的「篇名

* 國立臺灣師範大學東亞系助理教授，Email: Hlshao@ntnu.edu.tw。

** 通訊作者，國立臺灣師範大學圖書資訊學研究所教授，Email: samtseng@ntnu.edu.tw。

[1] 曾元顯，「文獻內容探勘工具-CATAR 之發展和應用」，*圖書館學與資訊科學*，第 37 卷第 1 期（2011 年 4 月），頁 31-49。.

[2] 邵軒磊，「中國研究議程之系譜--以日本國際政治學會誌為例」，*問題與研究*，第 51 卷第 1 期 （2012 年 09 月），頁 23-54。

[3] Yuen-Hsien Tseng and Ming-Yueh Tsay, "Journal clustering of Library and Information Science for subfield delineation using the bibliometric analysis toolkit: CATAR", *Scientometrics*, Vol. 95, No. 2, pp. 503-528, May 2013..

與摘要」讓 CATAR 進行「共現字分群」(Co-word clustering)，以自動歸納出待分析文章的主題類別。

承此，本文使用主題分析工具（CATAR），對《中國大陸研究》期刊於 1998~2015 年刊載之論文，透過論文的篇名與摘要文字，從事主題群聚(clustering)分析，藉以辨識顯著的研究主題，及其關鍵字，並以此觀察各主題發展趨勢。結果呈現出《中國大陸研究》之 473 篇文章，可歸類為七大主題，每一主題各有關鍵字。

表 1 各分群與重要關鍵字

分群	關鍵字
經貿關係 C1	兩岸、經貿、海歸、交流、經濟、政府、台灣
軍事外交 C2	安全、國際、軍事、全球化、戰略
內部經濟 C3	結構、華裔、土地、人民幣、財政、農民
政治制度 C4	類型、組織、政治、主義、極權
中共黨政 C5	政黨、民主、社會、態度、制度、決策、政治、黨國
文化意識 C6	斷裂、文化、民族、血緣、近代
社會議題 C7	人口、城市、香港、環境保護、辦學

其次，從每個主題的發表量（包括「發表數量」、「發表數量百分率」）之變化，可看出歷年期刊（或研究者）偏好主題之演進。

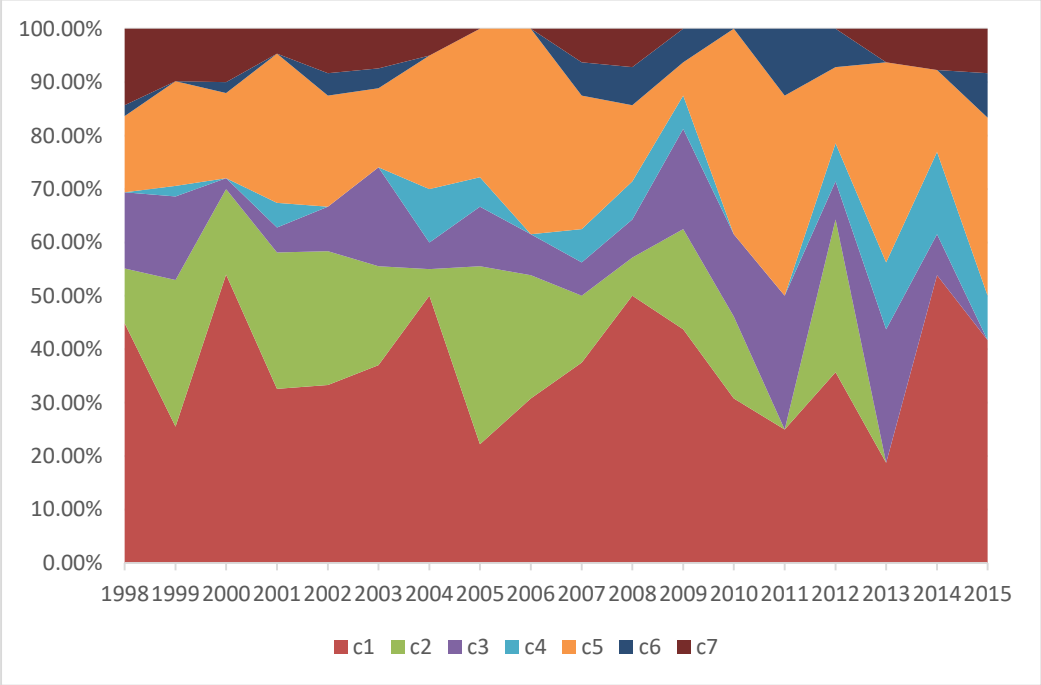


圖 1 各年各主題發表比率表

總上述圖表看來，C1 為《中國大陸期刊》之主流，比例始終排名第一，為穩定發展之研究主題，特別是近年隨中國在世界經濟角色之增強，發表比率也有微幅增加的趨勢。而次一穩定發展之研究主題為 C3，每年也均有穩定成長。C5 之研究，雖然較少，但每年也有固定的發表數量。上述三個分類較為穩定，且總體而言持續發展。相對稀缺之主題為 C6，除 2011 年外，比率均為個位數，是比較少人研究的主題。在 2006-2012 年較為風行，但近年較少研究。C2 在前半期一直是《中國大陸研究》期刊之研究主力，但近幾年逐漸減少，由一些關鍵字觀察，可發現其研究多與 C1 主題相近，可能是主題轉移之原因。C7 主題之變化較大，大略呈現週期性變化。初期、中期、末期均有發表量，但過程中有萎縮，特別是在 2008-2012 期間，幾乎沒有。因此是否該主題在 2008 年前後有變化？或是期刊之性質改變？則是可以深入思考之議題。C4 這一主題一直有少量發表，近年亦有漸漸變多傾向。

從趨勢可以看出，在「中國大陸研究」之發表傾向中，存在兩個主流議題，其他主題的年度篇數則變化較大。本研究貢獻為：1. 呈現了台灣「中國研究」逐年文章主題與關鍵詞彙的演變趨勢。2. 將所有「個別」的研究找到共通分群之可能性，並且試做了少量樣本的分群，其分群結果與常識相符，顯示此自動化工具應用到本領域也有一定的成效。3. 此主題分類自動化技術可篩選出各個分群關鍵字，提供研究者方便查閱過去的研究趨向。4. 根據各主題逐年的趨勢演變，提供後續研究的方向指引。

關鍵字：中國大陸、主題分析、自動化歸類、知識探勘、研究趨勢

Text Mining based Topic Analysis: A Case Study of Journal “Mainland China Studies”

Hsuan-lei Shao*, Yuen-hsien Tseng**

Abstract

With the rapid evolution of cross-strait situation, “Mainland China” as a subject of social science study has evoked the voice of "Rethinking China Study" among intelligentsia recently. This essay tried to apply an automatic content analysis tool (CATAR) to the journal "China Mainland China Studies"(1998-2015) in order to observe the research trends based on the clustering of text from the title and abstract of each paper in the journal. The results showed that the 473 articles published by the journal were clustered into seven salient topics. From the publication number of each topic over time (including "volume of publications", "percentage of publications"), there are two major topics of this journal while other topics varied over time widely. The contribution of this study includes: 1) We could group each “independent" study into a meaningful topic, as a small scale experiment verified that this topic clustering is feasible. 2) This essay reveals the salient research topics and their trends for the Taiwan journal "Mainland China Studies". 3) Various topical keywords were identified, providing easy access to the past study. 4) The yearly trends of the identified topics could be viewed as signature of future research directions.

Keywords: Mainland China studies, topic analysis, automatic clustering, research trends, knowledge mining

* Assistant Professor, Department of East Asian Studies, National Taiwan Normal University. Email: Hlshao@ntnu.edu.tw.
** Corresponding Author, Professor, Graduate Institute of Library and Information Studies, National Taiwan Normal University. Email: samtseng@ntnu.edu.tw.

《中庸》注疏思想史的數位人文研究：宋明理學的形成和演變

邱偉雲*、金觀濤**、劉青峰***、鄭文惠****

摘　要

本文研究目的有二，在數位技術方面，是以建構文言文之準詞彙擷取技術為主要目標，在人文研究方面，是希望將政治大學歷史與思想數位人文實驗室研究團隊（Digital Lab for History and Thoughts），過去用在近現代觀念史研究中的各種數位方法，推廣到經學研究領域中。

本文是以《中庸》為研究焦點，研究範圍包含了《禮記正義》中的「中庸」章原典，以及自漢代以降的注、疏文本。依照我們預設的研究框架，希望能從歷代《中庸》章句注疏中，觀察宋明理學的形成與演變軌跡。因應上述題目，我們將會依照時間先後，於漢、唐、宋、元、明五朝，各自尋找具有代表性的《中庸》注疏，作為考察理學形成與演變的基礎。上述考察脈絡的合理性，來自於前人多指出，理學家奉《中庸》一書為圭臬，因為理學的主軸為「心性論」，而《中庸》正是儒家經典中最能體現心性論的著作，因此本文以《中庸》注疏發展為考察焦點，當能適當地勾勒出理學形成與演變的軌跡。以下依照上面問題意識，詳細說明研究範圍如下。

（一）在理學形成階段，我們預計考察《禮記正義》中的「中庸」原典、漢代鄭玄（127-200）注、唐代孔穎達（574-648）疏，以及宋代朱熹（1130-1200）《中庸章句》，透過對不同時代的中庸注疏考察，掌握歷代對於《中庸》詮釋重點的樣態，並進一步歷時性地考察《中庸》注疏在漢、唐、宋之間的發展，以此作為理學形成階段考察的主要內容。（二）在理學演變階段，我們預計在上述理學形成階段的研究基礎上，加上元代許衡的（1209-1281）《中庸直解》，以及明代

* 湖北經濟學院新聞與傳播學院中文系副教授，Email: brianacwu@163.com。
** 國立政治大學中國文學系講座教授，Email: gtqf1908@gmail.com。
*** 香港中文大學中國文化研究所榮譽研究員，Email: 2869961913@qq.com。
**** 國立政治大學中國文學系教授，Email: wenhuei_cheng@yahoo.com.tw。

周汝登（1547-1629）的《四書宗旨》中的《中庸》注疏文字，还有智旭 (1599-1655)的《中庸直指補注》。在前者部分，元代許衡為朱子學傳人，透過《中庸直解》與朱熹《中庸章句》的比較，我們可以看見元代朱子學發展中，對於宋代朱子詮解《中庸》思想詮釋上的連續性繼承與非連續性轉化；而在後者部分，明代周汝登為著名的陽明後學，從他《四書宗旨》對中庸的注疏文字中，當可看見明代心學家對於中庸的詮釋發展；而與宋元朱子學的中庸詮釋比較中，更可見與宋元兩代朱子理學思想在宋明理學共性上的連續性，以及在「性即理」與「心即理」兩種不同的思想框架模式上所呈現出的非連續性。而在智旭（1599-1655）的《中庸直指》部分，透過與其他《中庸》註疏文字相比較，我們則可看見智旭聯結《中庸》與佛學思想的軌跡。以上即是本文在研究材料上的範圍以及研究步驟上的說明。

而在研究方法部分，根據以上所列的問題意識與研究範圍，我們預計使用過去已在近現代觀念史研究中不斷使用，且證明有效的數位人文研究法，嘗試運用於古典文獻的研究之上，這些方法包括：擷取文本關鍵詞彙的斷詞技術、關鍵詞叢研究法、概念群的權重計算。

在第一個 N-gram 技術部分，中文斷詞問題一直以來都是資科學界重要的研究議題之一，在過去我們已透過 N-gram 方法尋找白話文本中的準詞彙，發現對於尋找白話文本中的關鍵詞工作很有用，但 N-gram 的技術是否可以以及如何推廣於文言文本研究，目前仍有待試驗與解決。我們認為假使能夠在過去研究基礎上進行調整，將 N-gram 技術推廣至文言文本研究中，將能使數位人文學研究的視野更為開闊，得兼處理傳統與當代文本，因此本文第一個在方法上的研究重點，正是試圖在白話文斷詞技術的基礎上推進，進而尋找到能處理文言文本準詞彙以及擷取關鍵詞的方法。

而在第二個關鍵詞叢研究法部分，本文將以過去在白話文本上的研究經驗與技術為基礎，考察移植到文言文本中的可能性。本文將使用關鍵詞叢研究法，分析《禮記正義》中的「中庸」原典，與後代各種注疏。進行下列兩點工作：第一，擷取各個注疏文本中的關鍵詞叢及其詞叢結構；第二，比較歷代中庸注疏的關鍵詞叢，進行分析並歸納出中庸詮釋發展中的「連續性繼承」與「非連續性創發」等兩種現象。由於中文詞彙在不同語境下，意義會隨之改變，此種詞彙多義性問題需要設法解決。本文將從目前數據庫中所蒐集到的歷代《中庸》註疏文本中，找到大家用的最多的一個關鍵詞彙，然後以該詞彙為觀察對象，勾勒並考察其在各文本中的共現詞彙，借此共現詞彙的比對與分析，勾勒出不同作者、不同語境，

對同一詞彙所產生的影響，藉此分辨出同一詞彙，具有不同的詮釋現象，其背後觀念框架的差異，這樣的研究取徑能解決詞彙多義性的問題。

在第三個研究方法部分，我們想要解決的是《中庸》是否確實為儒家經典中，最能體現心性論的這個問題。為此，我們將先參考前人研究，客觀且公正的標定一批能夠體現心性論的詞彙；接著我們會再以儒家重要典籍為研究對象，如十三經，從中計算這些儒家經典中，佔有心性論相關詞彙的權重，藉此研究，我們可以客觀的說明，在儒家經典中，《中庸》是否為傳世文獻中最能體現心性論的著作？或是有其他經典更能體現心性論這個問題。我們將藉由以上三點工作，考察宋明理學形成與演變軌跡。以上是本文之研究方法與步驟說明。

在本文研究結果部分，我們會先區別出《中庸》原典與註疏文本的連貫性詞彙，以及原典中原來沒有的非連續性詞彙，藉此考察《中庸》原典與注疏思想間的新陳代謝現象。我們考察《中庸》注疏中的詞彙，發現某些特定關鍵詞彙的出現，是伴隨著理學思想史的發展，進而才產生詞彙的出現/消亡現象的，王汎森稱此為「詞彙的新陳代謝」。舉例如《中庸》原典中原先只有「未發」一詞，而其相對詞「已發」，在漢代鄭玄、唐代孔穎達、宋代朱熹、元代許衡的中庸注疏本文中皆未見，要到明代周汝登的注疏中才出現，形成「未發/已發」的相對詞組，其中特別有趣的是，朱熹早已曾在其他非《中庸》註疏文本的書信中談到「未發/已發」的概念，但為何就不在《中庸》註疏中提到，這其中的原因也是值得深入考察之處。根據中國思想史發展脈絡上來看，「未發/已發」相對詞組之成形，基本上是受到陽明心學一體觀出現，與朱子體用論並列後，帶起一體無別/體用先後之問題意識下所產生，正是在上述思想史發展脈絡下，才會出現「已發/未發」此組對立詞組。其他我們還觀察到例如「本體」、「氣先」、「幽明」、「心本」...等等詞彙，乃是宋明理學獨有註解《中庸》的關鍵詞，而這些關鍵詞背後也都各自象徵著當時理學思想的形成與演變軌跡，我們會在本文中加以羅列與綜合分析，這是注疏文字的新陳現象。而在代謝的部份，如我們發現在《中庸》原典與鄭玄注中，都以「性善」註解「唯天下至誠為能化」中的「化」字；但有趣的是，我們發現在往後的宋代契嵩、朱子，元代許衡，明代周汝登、智旭等人皆未用，可以說「性善」在後來的《中庸》注疏文字中被代謝了。我們從研究中得知「性善」觀念因為與後世理解《中庸》的脈絡不符，因此或被自然淘汰，或被選擇性避免使用，以上正是《中庸》註疏發展中所出現的「詞彙/觀念代謝現象」。如上述指出的「詞彙/觀念的新陳/代謝現象」，當可作為往後思想史研究時，配合數位人文方法下，能夠大量提出的新提問方向。

本文研究乃是從不同於人文精讀分析視野的角度出發，透過數位人文技術下的鳥瞰視野，從思想著作文本的資料結構面向進行考察與檢驗，觀察《中庸》注疏歷代發展中所透顯出的宋明理學形成與演變軌跡，並進一步與現有人文學研究成果對話，此即本文之研究價值。本文使用詞彙新陳代謝這一直觀且非數位人文技術難以達成的研究取徑，當能說明不同註解者，在註解時，其背後觀念系統結構的異同與變化。而本文所進行的研究嘗試，未來將繼續擴充注疏文本，也會加入日本與韓國部分的《中庸》注疏文字，藉以發展東亞的《中庸》注疏思想史研究；除此之外，也可擴充到其他經典的經學研究工作中，此即為本文未來推廣前景與發展所在。

關鍵字：中庸、經學、宋明理學、概念群權重計算、關鍵詞叢研究法

The Digital Humanities Study on the Exegesis of The Doctrine of the Mean—Focuses on the Formation and Evolution of the Neo Confucianism

Wei-yun Chiu[*], Guan-tao Jin[**]

Qin-feng liu[***], Wen-huei Cheng[****]

Abstract

This paper aims at two subjects. The first, in the respect of digital technology, is to construct the pre-filtered keyword candidates retrieval of classical Chinese. The second, in the respect of humanities study, is to apply the digital research ways regarding Chinese modern conducted by the Digital Lab for History and Thoughts in NCCU to Confucian classics.

This paper centers on the *Doctrine of the Mean(《中庸》)*, including Original Classic in *The Notes and Commentaries of the Book of Rites(《禮記正義》)*and the exegesis after Han Dynasty. From these, we try to observe the formation and evolution of the Neo-Confucianism (宋明理學). Therefore we will review respectively a representative exegesis from Han, Tang, Song, Yuan, and Ming dynasty, exploring the formation stages and evolution path of the Neo-Confucianism hidden in the development of the exegesis of the *Doctrine of the Mean(《中庸》)*. This is because previous research has pointed out, Overview of the Neo-Confucianism (宋明理學), that thinkers of the Neo-Confucianism view the *Doctrine of the Mean(《中庸》)*—the best representative of"Theory of Human Nature, (心性論)"the theme of Neo-Confucianism (宋明理學)—as the bible.

[*] Associate Professor, Department of Chinese Literature, Hubei University of Economics. Email: brianacwu@163.com.
[**] University Chair Professor, Department of Chinese Literature of National Chengchi University. Email: gtqf1908@gmail.com.
[***] Honor Researcher, Institute of Chinese Studies, Chinese University of Hong Kong. Email: 2869961913@qq.com.
[****] Professor, Department of Chinese Literature, National Chengchi University. Email: wenhuei_cheng@yahoo.com.tw.

Based on the research framework mentioned above, the following is our research process of Neo-Confucianism (宋明理學).In terms of formation stage, we would observe the Original Classic of *The Notes and Commentaries of the Book of Rites(《禮記正義》)*, the annotation of Zheng Xuan(鄭玄) in Han Dynasty, the exegesis of Kong Ying-da(孔穎達) in Tang Dynasty, and *Zhang Ju on the Doctrine of the Mean(《中庸章句》)* of Zhu Xi(朱熹) in Song Dynasty. From these we expect an outline and development of interpretation of the *Doctrine of the Mean(《中庸》)* in all ages.

As for the evolution stage, our study additionally involves in the *Literal Interpretation of the Doctrine of the Mean(《中庸直解》)* of Xu Heng(許衡) in Yuan Dynasty, who was the follower of Zhu Xi(朱熹) and by the comparison their different interpretation and thoughts, we can realize the path of inheritance and transformation on interpretation of the *Doctrine of the Mean(《中庸》)*. Besides, our study also involved in Zhou Rudeng's(周汝登)*The Purpose of Four Books*(《四書宗旨》) in Ming Dynasty. Mr. Zhou was the famous follower of Wang Yang-Ming(王陽明) so that we can see the continuity of "Learning of the Principle"(理學) in Song Dynasty and"Learning of the Mind"(心學) in Ming Dynasty, and the discontinuity of "Nature as Truth"(性即理) and "Hearts as Truth"(心即理) on the thought patters from his exegesis in Four Books study. Finally, in the part of *the Direct Interpretation of the Doctrine of the Mean(《中庸直指》) of Zhi Xu(智旭), that we can see the Buddhism how to use their influence to Buddhismlize(佛教化) the Doctrine of the Mean(《中庸》) in Ming Dynasty.* That is, through comparing the exegesis in Yuan and Ming Dynasty, we can depict the evolution direction of the Neo-Confucianism (宋明理學).

In terms of analysis technology, this paper will use methods including N-Gram, Keyword Cluster, and weight calculation of concepts, which were all applied in modern idea researches.

N-Gram can help us fix the problem of Chinese Word Segmentation. Over the past years we have found that N-Gram analysis is very useful to search for the pre-filtered keyword candidates of vernacular Chinese text. How to apply N-Gram to classical Chinese text is our latest concern. This will broaden the view of the digital humanities research so that we can analyze vernacular and classical Chinese text at the same time.

The second method, Keyword Cluster, is used to observe the Original Classic of

the Notes and Commentaries of the Book of Rites and exegesis in all ages, focusing on two things when mining the keywords cluster of Original Classic and exegesis in all ages. First, the outline of the data structure of every exegesis text composed in keywords. Second, the depiction and analysis of two phenomenon—continuous inheritance and discontinuous creation—among keyword development of exegesis of the Doctrine of the Mean. By doing so, we will be able to realize the development of exegesis of the Doctrine of the Mean and the formation change of Neo Confucianism.

Third, we will use concept-weight-calculation method to solve a problem regarding whether the Doctrine of the Mean(《中庸》) is the best representation of "Theory of Human Nature" among all canons of the Confucianism.

Based on the analysis mentioned above, we found that two phenomenon in the exegesis of the Doctrine of the Mean in all ages—continuous inheritance and discontinuous creation—called “the term metabolism.” For example, the term “unexpressed” in the original classics of the Doctrine of the Mean triggered a relative term “expressed” in Ming Dynasty. Basically this advent of unexpressed/expressed terms was based on Wang Yangming’s and Zhu Xi’s philosophy, which means that every term can reflect the change of thoughts orderly.

In conclusion, this paper aims to use digital humanities to review horizontally and vertically the data structure of classics and the development of the exegesis of the Doctrine of the Mean. And, more importantly, we expect to verify or correct former researches and even to create new researches. This study will be able to promote and strengthen future classics studies as well.

Keywords: The Doctrine of the Mean, Confucian classics, Neo Confucianism, concept-weight-calculation method, keyword cluster analysis

《全唐詩》顏色光譜學的數位人文研究

鄭文惠*、余清祥**、顏靜馨***、郁嘉綾****、邱偉雲*****

摘　要

顏色光譜是一種物質的客觀存在，卻由人的生理感知而來，甚或顯影為一種主觀存在，同時在傳統文化的時間長河裡，又承載著世世代代的觀念系統與價值信仰。正因顏色光譜複合著具象與抽象的物質屬性與心理感知，詩人從而使顏色光譜在物與色的共舞、光與影的交疊中更顯抒情音質。詩人在獨特的身體感知中，以詩歌的修辭技術，標記出本己的思想情感，呈顯為獨特的詩歌風格，從而蔚為一代的記憶表徵，也積澱了世代間不同的思想價值與文化風俗。

本文立基在去年〈情感現象學與色彩政治學:中唐詩歌白色抒情系譜的數位人文研究〉一文的研究基礎上，擬拓進數位人文技術方法並結合統計理論模型，更為整體全面而深入的研究唐詩顏色光譜學。

本文將唐詩顏色光譜分為白、黑、紫、赤、青、黃及金屬色等色系，再將每一色系構詞的詩作視為一個個文本群的同時，也進行詞彙篩選，將一義多字（如黑色系：黑、墨、皂、皁、烏、蒼、緇、玄、黔、黯、黝、涅、漆等）、一字多義（如「蒼」或為青色或為灰色或為黑色；「青」或為藍色或為綠色或為黑色）及常用字（如赤色系：紅、赤、絳、丹、粉、緋、赭、彤等）與罕用字（如黑色系：黲、黵、黴、黕、黮、驪等）的顏色詞整理出詞表，以供數據分析，除延續去年以數位技術方法勾勒出色系出現的位置及其構詞與構句方式的分子結構圖外，也同時關注各色系特別密集與特別離散的文本分佈狀況，以進一步釐析離群值的意義，如黑色系跟青色系、黑色系跟赤色系、青色系與赤

* 國立政治大學中國文學系教授，Email: wenhuei_cheng@yahoo.com.tw。
** 國立政治大學統計學系教授，Email: csyue@nccu.edu.tw。
*** 國立中正大學中國文學系博士生，Email: xing3325@gmail.com。
**** 國立政治大學統計系碩士生，Email: ja1234546@yahoo.com.tw。
***** 湖北經濟學院新聞與傳播學院中文系副教授，Email: brianacwu@163.com。

色系、黑色系與白色系……等的關聯，透過時間與顏色、顏色與顏色、顏色與光影的關聯性，以勾勒出唐詩的顏色光譜學及顏色光譜背後的深層意義。

大體而言，唐詩顏色光譜學除與詩人個人獨特的聯覺通感與視覺想像及心理情感與感覺結構息息相關外，還涉及佛道宗教信仰、經世與隱逸思想、園林文化、祭典儀式、身分地位、染織技術、彩繪技術、化妝術……等等，從中不僅可掌握詩人獨特的顏色修辭與詩歌主題風格的關係及其深層的顏色心靈光譜，也可深入理解透過初盛中晚唐詩多重性的顏色光譜所開展的隱喻系統，及所表徵不同時期宗教、思想、技術、政治、經濟、階級等社會文化的變革。

關鍵字：唐詩、顏色光譜學、隱喻、離群值、數位人文

A Digital Humanities Study of Color Symbolism in *Quan Tangshi*

Wen-huei Cheng[*], Jack C. Yue[**], Jing-xin Yen[***]

Jinny Yu[****], Wei-yun Chiu[*****]

Abstract

As a physical property, color is perceived by the sensory system of human being as a subjective experience. Different colors have been encoded with various symbolic meanings and values by the culture and tradition of different generations. As color is both objective as a physical property and subjective as perceptive experience, the poets are able to play with object, color, light and shadow in their works and hence imbue color with symbolic power and lyrical quality. By applying the rhetorical techniques for poetry, the poets transcribe their personal perceptive experience into a symbolic expression of feelings and thoughts, which not only forms the idiosyncratic style of individual poet but is also inscribed with the thoughts, value, cultural experience and convention of different generations.

This paper is based on the research of "Phenomenology of Emotion and Politics of Color: Digital Humanities Research on the Lyrical Genealogy of 'White' in the Poetry of Middle Tang" published in 2015. It aims to combine digital humanities with statistic model to give a more comprehensive study on the color symbolism of Tang poetry.

[*] Professor, Department of Chinese Literature, National Chengchi University. Email: wenhuei_cheng@yahoo.com.tw.
[**] Professor, Department of Statistics, National Chengchi University. Email: csyue@nccu.edu.tw.
[***] Ph.D. Student, Department of Chinese Literature, National Chung Cheng University. Email: Email: xing3325@gmail.com.
[****] M.A. Student, Department of Statistics, National Chengchi University. Email: Email: ja1234546@yahoo.com.tw.
[*****] Associate Professor, Department of Chinese Literature, Hubei University of Economics. Email: brianacwu@163.com.

This paper first classifies a few color groups such as white, black, purple, red, blue-green, gold, metal, etc., and then selects the poems that employ the words from these color groups to create a series of text sets for each category. It also identifies and lists all the color words of the same significance, the color words that have multiple significances and the most common words as well as the rare words for data analysis. In this article, the same digital humanities technology employed in the previous study is applied to draw molecular structure charts to locate the positions of the color groups in terms of the word-formation and the structure of sentence of the color words in the verses. It focuses on the clustering and dispersion of different color groups in the text sets and further investigates the significance of the outliers. It also examines the relationship between different color groups, time and color, as well as color, light and shadow in order to get a better understanding of the color symbolism in Tang poetry.

In general, apart from the synesthesia, visual imagination, psychological emotions and feelings of individual poet, the formation of the color symbolism in Tang poetry is also associated with various factors such as the religious beliefs of Taoism and Buddhism, ideas of governing and seclusion, ceremonial rituals, social identity and status, dyeing and weaving techniques, painting techniques, cosmetic arts and so on. Through the in-depth study this paper offers, we are able to gain a more comprehensive insight into the distinct rhetoric of color of individual poet, its relationship with the themes of poem, its cultural and psychological implications, as well as the metaphorical system formed by the color symbolism in Tang poetry which reflects the social and cultural changes in the religions, thoughts, techniques, politics, economy and hierarchy during different periods of time.

Keywords: poetry of Tang dynasty、color symbolism、metaphor、outliers、digital humanities

「新」的激變與交鋒：
中國現代性形成的數位人文研究

邱偉雲*、鄭文惠**

摘　要

中國近代的多元現代性如何形成，一直是人文學者不斷提問的重要問題，各領域學者莫不從自身專業學科出發，意圖去勾勒出現代性的形成與發展軌跡，而人文學者在這方面已經有諸多重要研究產出（Benjamin Schwartz, 1964；Cohen, Paul, 1974；Chang, Hao,1987）。本文主要研究目的，是想在有別於過去人文學者研究視域下，由數位人文視野出發，重探中國近代多元現代性的形成。本文將從近代中國思想轉型時期，觀念發生劇烈變化的特殊語言現象出發，考察近代中國思想轉型時代中，知識分子大量以「新」為詞族—一套以關鍵詞叢表述的話語系統—所建構與指涉的中國多元現代性議題，並從這些議題中，捕抓到中國現代性的發展歷程。

而為何要以「新」的詞族作為考察「現代性」的線索，主要原因在於近代中國的知識分子，是以建構一套以天演進化論作為合理化論據及正當性基礎的「新」之論述，去推廣西方及從日本引介而來的新思想、新制度、新事物，完成中國從千年傳統思維轉向現代的「現代性」過程，故「新」的論述可視為中國從傳統社會過渡到現代社會的重要表徵。正是在「物競天擇，適者生存」的天演進化論語境下，「新化」就等同於「現代化」，並且披帶上「文明」的色彩。在這套「新」的論述中具有一個普遍化的趨勢，即凡是冠上「新」者為「善」。其中「新」論述正是促成中國現代性形成與發展時的重要修辭策略。

在人文學者的過去研究成果中，大致歸結出兩個重點：其一，「新」的內容從清末至民國都不盡相同而與時俱進；其二，知識分子以「新」作為工具對傳統觀念進行再脈絡化。由上可見，「新」論述具有多元、複雜的特性，因此我們若

* 湖北經濟學院新聞與傳播學院中文系副教授，Email: brianacwu@163.com。

** 國立政治大學中國文學系專任教授，Email: wenhuei_cheng@yahoo.com.tw。

以傳統研究方法，我們將無法宏觀的勾勒出「新」的論述，在中國現代性形成軌跡上的意義。所以本文希望嘗試透過大數據視野，以數位人文技術，探勘近代巨量的歷史史料，掌握「新」的論述之資料結構與演變軌跡。本文以「中國近現代思想史專業數據庫（1830-1930）」中所收錄的近現代中國重要政治思想文獻為底本，該數據庫收錄歷史文獻時間橫跨 1830 年至 1930 年，資料底本時間段為中國近代觀念轉型最為重要的一百年。該數據庫具有一億兩千萬字的文獻量，所收集的資料涵括了清末民初近代期刊、晚清檔案資料、清季經世文編、清末民初士大夫著述、晚清來華外人中文著譯、西學教科書等文獻，這些文獻中幾乎包括了近代政治思想上重要的典籍，故具有資料相對完備的特點，正因如此，我們以這個資料庫作為我們研究的資料來源。（Jin et al., 2008）

在研究方法部分，首先我們以 N-Gram 技術，而後讓專家過濾關鍵詞(Liu et al., 2011)，最後得出" The Database for the Study of Modern Chinese Thought (1830-1930) "中的「新」的詞叢表。其次我們運用歷年出現比例累加圖（Cumulative proportion chart over period）方法，計算並描繪出「新」的詞族的歷時性變化 (Jin et al., 2012)。最後我們運用集群分析方法，勾勒出「新」的詞族的集群分布現象。我們藉由以上三個步驟，能勾勒中國近代「新」的觀念內涵以及演變過程，考察近代知識分子在「新」的各種論述中所透顯得世界觀與思維慣例，藉以理解中國近代多元現代性的形成步調。

本文在上述語料與理論方法的研究基礎上，進行了兩項研究，分別敘述如下：其一，我們以數位人文方法過濾得出中國近代「新」的詞叢表，可以使人文學者宏觀掌握中國近代自 1830 年到 1930 年之間，曾經出現過的「新」之關鍵詞叢，掌握百年歷史中「新」的論述圖像。其二，我們可以進一步由歷時性與集群性角度出發，描繪出百年之間「新」的關鍵詞叢的歷時發展與集群現象，這即可提供人文學者宏觀掌握不同時期的「新」的論述傾向與主軸。如中國 1840 年鴉片戰爭後，以「新兵」與「新例」為「新」之論述主軸；1860 年自強運動後以「新制」、「新軍」、「新法」為主軸，這符應中國近代自鴉片戰爭後至甲午戰前，以「制度改革」為啟蒙行動的主流思潮； 而 1895 年甲午戰爭後，則轉而以「新政」、「新理」、「新說」為主軸；1900 年庚子事變後則更聚焦在「新教育」、「新國家」、「新思想」等論述，符合以「思想改革」為救國行動的趨勢。而 1915 年後以「新文化」、「新文學」、「新經濟」為主軸，符應民國初年新文化運動「社會改革」的時代命題。人文學者當可從此圖示中快速地掌握每個時期「新」之論述的重點核心。

我們從以上兩類資料結構中，進一步掌握了中國近代「新」之論述的發展方

向。由此可見，數位人文技術確實能提供人文學者宏觀的歷史資訊，讓人文學者能快速且精確地掌握到時代重要的「新」之論述。本文研究成果計有以下幾項：其一是勾勒中國近代「新」與「舊」字詞族/觀念群的內涵、類型與發展；其二是描繪「新」的詞族如何受到近代重大事件之影響而改變其除舊佈新的方向；其三是指出在不同政治傾向中對於「新」字族的不同內容詮釋？最後本文透過中國近現代「新」之詞族的形成與演變之數位人文研究，掌握近現代中國現代性形成的發展歷程。本文也希望在數位人文研究的應用與研發中，能對數位人文學提供更多人文與技術層面的建議。

關鍵字：現代性、新、譬喻、概念、數位人文學

The Interaction and Conflict of "New": A Digital Humanities Study of Chinese Modernity

Wei-yun Chiu*, Wen-huei Cheng**

Abstract

It is a continuously raised question from humanities that how the multi-modernity formed in modern Chinese history. While scholars in diverse fields keep trying to depict the formation and development track of modernity from their point of professional views, humanities researchers have conducted many important research outcomes (Benjamin Schwartz, 1964；Cohen, Paul, 1974；Chang, Hao,1987). This paper aims to review the formation of the multi-modernity from the digital humanity view, which is distinguish from the past research horizon. This paper will observe the "new" word family in the period of thought transformation in modern China, analyzing how dose the intellectuals use those "new" word family to construct the multi-modernity in modern China.

We view the "new" word family as the clue of "modernity" mainly because intellectuals of modern China used a set of "new" discourse depending on the evolution theory to promote the west new ideas, new system, and new things, to complete modernization process. The "new" discourse therefore can be viewed as a pivotal manifestation of China's transition from tradition society to modern society. "New" is equivalent to modernization, and especially, is viewed as civilization in the context of Evolution Theory. This meaning expansion which views "new" equal to "good" is an universal phenomenon, which is the important rhetorical strategies enhancing the formation and development of Chinese modernity.

Notably, there are two key points from the past research results of humanities: 1) the "new" content was diversely different in the late Qing Dynasty but gradually

* Associate Professor of Department of Chinese Literature of Hubei University of Economics. Email: brianacwu@163.com.
** Professor of Department of Chinese Literature of National Chengchi University. Email: wenhuei_cheng@yahoo.com.tw.

changed over time in the Republic of China, and 2) the “new” was used by intellectuals as a means to re-contextualize traditional concepts. Based on that, the "new" discourse had the multiple and complex characteristics so that through traditional research method we will not be able to holistically outline the trace and meaning of “new” discourse in the modern China. Thus this paper tends to use digital humanities technology, comprehensively observing massive historical data and in turn seizing the data structure and evolution track of "new" discourse in modern China.

This paper based on the "The Database for the Study of Modern Chinese Thought (1830-1930) ", which includes the most important essay and books of political thought from 1830 to 1930, when is the most important period of thought transformation in modern China. The database totally contains one hundred and twenty million words, including the modern journals, archives of the late Qing Dynasty, Qing dynasty Confucian writings, translation works, and textbook of west—all the most valuable in the political thought of modern, meaning that this database is relatively comprehensive for this study. This is why it is chosen as our research source.(Jin et al., 2008)

In regard of research methods, the first we use N-gram method and then filter keyword by humanities scholars, gaining the word family of "new" in database. (Liu et al., 2011) Second, we use the "cumulative proportion chart over period" method to calculate and draw out the diachronic changes of the word family of "new" (Jin et al., 2012). And finally, we use cluster analysis method to outline the cluster distribution phenomenon of the "new" word family. By these three phases above, we could holistically grasp the content changing of "new" word family.

Two studies are conducted based on the corpora and theory above, as follows. First, we obtain the “new” word family in modern china through digital humanities method. Second, we further describe the diachronic development and the cluster phenomenon of top 50 high-frequency “new” words, tending to give humanists a holistic view of the trend and axis of the "new" discourse in different periods. We find that after the Opium War in 1840, the homophony of "new" discourse included "New recruit（新兵）" and "New precedents（新例）." Comparably, after Self-strengthening Movement in 1860, more like "new system（新制）," "New Army（新軍）" and "New law（新法）" became the spindle, suggesting the primary ideology majorly concerning “institutional reform” from Opium War to the Sino-Japan War. After the Sino-Japan War in 1895, it turned

out to be the "New Deal（新政）", " New axiom（新理）" and " Neodoxy （新說）." Instead, after the Gengzi Incident in 1900, it shifted to focus on "New education（新教育）", " New country（新國家）", " New Thought（新思想）" and so on, which perfectly accords with the primary ideological trend of "ideology reform" from the Sino-Japan War to New Culture Movement. Likewise, after 1915, the "new culture（新文化）", "new literature（新文學）", and "new economy（新經濟）"also accord with the primary ideological trend of "social reform" from the New Culture Movement. From these, humanists could quickly grasp the core of “new” discourse of every period.

Through two data structures above, this paper has seized the development of "new" discourse in modern China(1830-1930). Thus the digital humanity technology could indeed give a holistic historical history data structure for humanists, helps them grasp efficiently and precisely the critical discourses in different periods.

The research results as follows: First is that depicting the connotation, types and development of "new" words/concepts in modern China; the second is describing how the “new” word family to interact with major events in modern China; the third is depicting the different interpretations of the "new" word family among different political tendencies; and finally, through the digital humanities research of “new” word family, we grasped the development of multi-modernity in modern China. Moreover, this paper also expects to provide more humanities and technical advices for other digital humanities in light of application and development of digital humanity researches.

Keywords: modernity, new, parables, concepts, digital humanities

革命與中國當代“藝術”觀念的起源

彭卿*

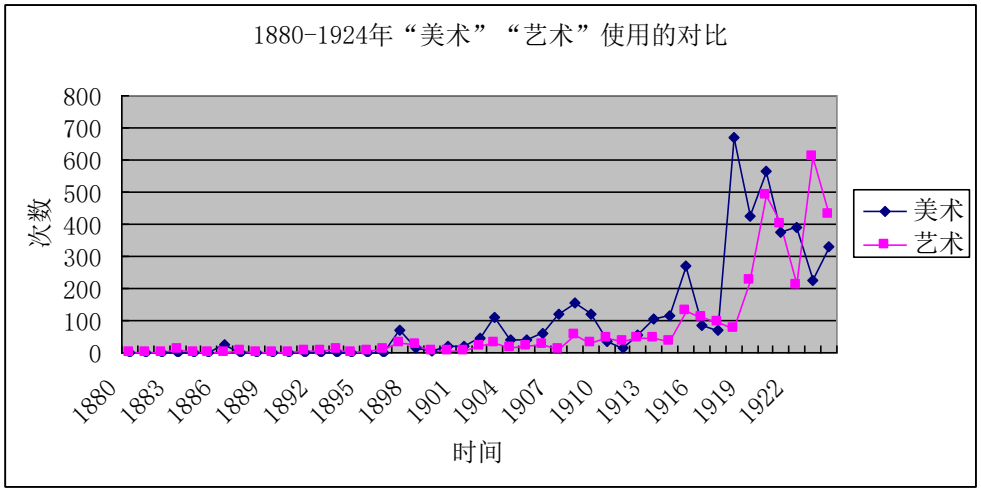
摘　要

本文研究的是革命思想如何影響當代“藝術”觀念的形成。是一篇以“藝術”為關鍵字切入考察，與思想史結合，運用數位人文研究法，探討當代“藝術”觀念起源的研究。

“藝”本意種植，《說文》中解釋“藝，種也”，引申意為技能，技術。“術”，本意道路，《說文》釋為“術，邑中道也”，引申為方法、技藝。中國傳統文獻中，“藝術”一詞，始見於南朝宋範曄《後漢書·卷五孝安帝紀第五》，“詔謁者劉珍及《五經》博士，校定東觀《五經》、諸子、傳記、百家藝術，整齊脫誤，是正文字。南朝至唐代的藝術觀念涵蓋了陰陽、蔔筮、天官、醫巫、音律、相術、技巧等“天文數術，方醫技巧”，“藝術”和“術藝”偶爾混用。北宋《新唐書·藝文志》子類出現了“雜藝術類”一類，所收錄書籍分為投壺、博戲、棋、射、繪畫幾種。元代《宋史·藝文志》子類中沿用了“雜藝術類”，所收錄書籍，含有射、棋、繪畫、草書、文房、酒事風俗、投壺、彩選、樗蒲、雙陸、葉子、漆、醫馬、醫駝等。清代《明史·藝文志》子類出現“藝術類”，把“雜”字去掉了，包含了文物賞鑒、雜錄、書法、金石、繪畫、琴譜、墨、印、文房等。中國傳統“藝術”觀念雖指涉的事物廣泛，但普遍重視技藝性，強調的是一項專門的技能。雖然隨著文人對藝術活動參與性的提高，藝術也從最初的卜筮等方術，演變成更具修身與娛樂性的活動。但是，從《論語·述而》“志於道，據於德，依于仁，遊於藝”始，即可看出“藝”的地位。正因為強調“技藝”中“技”的層面，洋務運動時期“藝術”的普遍意義代表著新技術。

本文論證了“美術”傳入中國後，對“藝術”產生了影響。在“中國近現代思想及文學史專業資料庫（1830-1930）”中檢索關鍵字“美術”’、“藝術”，依據時間，統計出“美術”、“藝術”在每年的使用量。同理，在“全國報刊索引”中檢索“美術”、“藝術”，得出每年的使用量。最後按照年份，將不同資料庫每年的使用量疊加，得出“1880-1924 年‘美術’‘藝術’的使用對比折線圖”。

* 浙江師範大學美術學院講師，Email：522131648@qq.com。



從該圖的資料走向看，“藝術”一詞在“美術”傳入以後，隨著“美術”的起伏而起伏。1897 年，“藝術”產生了一個微小的波動，資料主要出自維新派所創辦的《時務報》與《知新報》。但當時中國道在上藝在下的經世思想沒有改變，“藝術”的技術與技藝意，在普遍意義上幾乎沒有變化。

中西二分二元論時期（泛指 1900-1915 年），隨著“美術”理論引入中國，“藝術”有三種用法。一是與“美術”混著使用，與“美術”的意思相同。例如 1902 年楊度在《遊學譯編序》中說：“藝術者，使作者之感情傳染於人之最捷之具也。作者之主題當如何，則必以直接或間接向于人類同胞的結合，而求其好果，以為感情之用也。此處“藝術”與“美術”的美學意同，強調情感傳遞。二是“美術”對“藝術”的“技術意”產生影響，即美術與工業藝術（技術）相結合，誕生了“美術工藝”。例如 1907 年第 18 期《農工商報》刊《中國新聞：組織美術工藝所》。三是“美術”對“藝術”的“技藝意”產生影響，延伸出“美的技藝”之意，傳統技藝意被“美的技藝”所取代。例如 1908-1911 年，上海城東女學校創辦《女學生》校刊，有“藝術談”一欄。談各種女性生活技藝，如何做霜淇淋，如何製作香料，如何編制，如何做家務等，使技藝沾染上了美的意味，成為一種“美的技藝”。但是，“藝術”在二元論時期，“技術意”和“美的技藝意”很重，更強調“技”的層面，而不是“美”的層面。例如 1908-1909 年《萬國商業月報》開闢“新藝術”專欄，共發表 47 條“新藝術”，是新“技術”的意思。如煤油燈用法、無線電照相術、新發明簡妙之制冰機等。因此，“藝術”在二元論時期，整體的小波動雖與“美術”同構，但是幅度大不如“美術”。從整體看來，幾乎為一條直線。

五四新文化運動後，自由主義者試圖尋找一種建立在物質進化自然觀之上的新人生觀，因而賦予了“藝術”新的含義。從資料圖上看，1915 年新文化運動伊始，

“藝術”略微波動。此時，“藝術”與“人生”“理想”搭建出關係。例如 1916 年 9 月 1 日《新青年》刊王湜《時局對於青年之教訓》：“歐洲有自古傳來之三思潮，至今猶食其賜者，自由、平等、博愛是也。自由思想導源於希臘，希人富於想像力及愛美之精神。藝術、科學不囚拘於習慣，故能實現人生之新理想。1916，1917 年“美術”“藝術”的使用量差不多，而 1918 年對“美術”的討論到達頂峰，隨之 1919 年爆發了“美術革命”。

1919 年，“藝術”的使用量也與以往相比增加了，折線圖呈明顯上升趨勢。五四自由主義者秉持著科學主義和自由主義的思想，“藝術”與科學、人生相關，追尋著真美，注重創造真的和美的“新藝術”。例如 1919 年 3 月 15 日《新青年》刊魯迅《隨感錄：五三》：“畫《潑克》的美術家說他們盲目盲心，所研究的只是十九世紀的美術，不曉得有新藝術真藝術。1919 年 11 月 1 日《新青年》刊日本廚川白村著、朱希祖譯《文藝的進化》：“若是醜的，那麼就當做醜看，更就他極微奧的地方看出他一種的美來，所謂‘有生命的心中深伏一種內部的美’（the inner beauty that lies deep at the heart of life）。這就是新藝術生命所存的地方了。”

1920、1921 年“藝術”和“美術”的使用量不相上下，1922 年“美術”仍比“藝術”多，但是 1923 年後情況發生了變化，“藝術”的使用突然超過了“美術”。“美術”一詞，作為觀念詞彙，對應的是二元論時期的“公德”，五四新文化運動後則成為了學科名詞。（詳見中國美術學院 2016 屆博士論文彭卿《中國現代“美術”觀念的形成與演變——1895-1924 年的“美術”觀念》）本文研究的重點在於揭示作為觀念詞彙的“藝術”背後的觀念是“革命”，當代“藝術”一詞的使用與“革命”思潮的傳播是同構的。“藝術”對“美術”的取代，是革命一元論對中西二分二元論的取代。與馬克思主義者相比，五四自由主義者並沒有建立一種明確而又有說服力的新道德。1923 年，隨著科學與人生觀大討論科學派的勝利，馬列主義革命思想注入了人生追求，唯物史觀確立了霸權地位，形成新的革命道德意識形態。

本文選定《新青年》（1915-1924）為統計文本，檢索關鍵字“藝術”，截取包含有“藝術”一詞的上下文例句。使用眾果搜詞頻句頻統計網站 http://www.zhongguosou.com/zonghe/cipintongji.aspx 自動分析出與《新青年》“藝術”相關聯的詞彙頻率。發現 1923-1924 年與“藝術”關聯性高的詞彙裡，突然出現了“資產階級”“無產階級”“鼓吹”這樣代表著馬列主義革命的話語，從而搭建出“藝術”與“革命”的關係。例如 1923 年 6 月 15 日刊奚浈《共產主義之文化運動》：“革命運動之際，便能於藝術與科學方面大放光明。”1924 年 8 月 1 日刊蔣俠僧《無產階級革命與文化》：“為著明天，我們拋去藝術之花。”具有“革命”觀念的“藝術”

一詞，徹底反超具有“公德”意味的“美術”。

綜上，當代“藝術”觀念的形成，背後對應著“革命”。當革命成為新的道德追求，藝術亦脫去了技藝的古老定義，成為一種人生追求，藝術家的歷史地位也得以提升。隨著唯物觀的確立，人們的藝術人生觀從對真美的追求發展到幻化成一種對革命烏托邦式生活的嚮往。

關鍵字：藝術、革命、五四新文化運動

Revolution and the Origins of Contemporary Chinese "Art"

Qing Peng*

Abstract

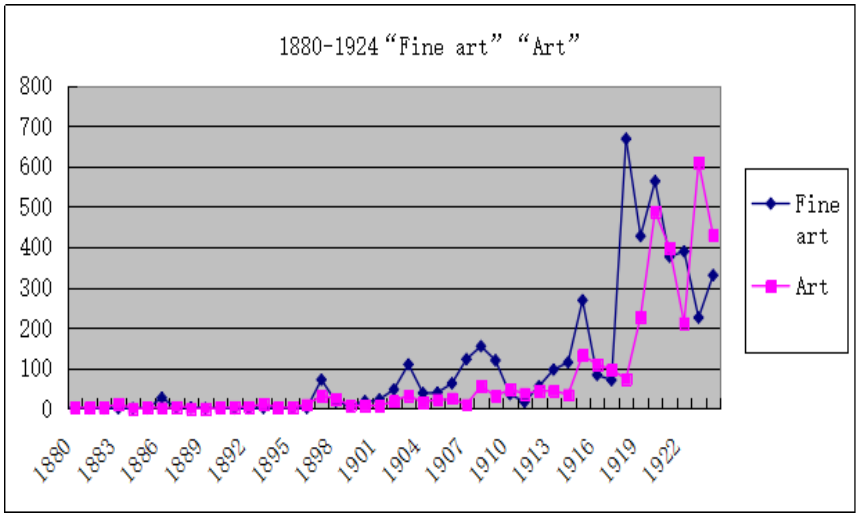
This paper studies the revolutionary thought how to influence the formation of the contemporary "Art". Is an "Art" as keywords into the investigation, combined with the history, using digital humanities research, discusses the origin of the modern concept of "art" research.

According to the "Art" meant to planting and city middle path, explained in "*Shuo Wen"*, implied meaning skills and technology. In Chinese traditional literature, the word "art" was beginning in the "*Hou Han Shu Fifth volumes* Biography of Emperor Xiaoan". From the Southern Dynasties to the Tang Dynasty, "Art" concept covers Yin-Yang, divination, astrology, medicine witch, music, Fortune-telling skills etc. The Northern Song Dynasty "Art History of the new Tang Dynasty history" sub class appeared "miscellaneous art" category, the collection of books is divided into game, chess, shot ,painting. The Yuan Dynasty "Art history of the Song Dynasty history" sub class follows the "miscellaneous art". The Qing Dynasty "Art history of the Ming Dynasty history" sub class "art", which removed the word of "miscellaneous" , including cultural relics, miscellany, calligraphy, stone appreciation, painting, music, ink, seal,paper etc. Chinese traditional "Art" concept refers to a wide range of things, but generally pay attention to skills, emphasis is a special skill. Although art is more and more literati, but its status in the traditional culture is relatively low. Because of emphasizing the skills, during the period of Westernization Movement universal, "Art" means a new technology.

This paper demonstrates that the "Fine art" was introduced into China, and it has influenced the "Art". Searching the keywords "Fine art" "Art" in the"Chinese modern thought and literature history professional database (1830-1930)" , then according to

* Lecturer of Academy of Fine Arts, Zhejiang Normal University. Email：522131648@qq.com.

the time, statistics "Fine art" "Art" in the annual usage. The same to the "national newspapers and Periodicals Index". In the end, we could get the 1880-1924 year contrast line chart.



From the chart, we find the words of "Art" follow the "Fine arts" ups and downs. In 1897, the "art" produced a slight fluctuation , the data is mainly from the newspaper which founded by reformers .But there is no change in the meaning of “Art”.

In 1900-1915 years,"Art" has three meanings. First, it is the same to the“Fine art”; Second, it means Art craft; The last one means the beauty of skills. But “Art” is more emphasis on technology.

After the new culture movement, the liberals are trying to find a new way of life based on the concept of natural evolution, which gives the new meaning of "Art". At the beginning of the New Culture Movement, "Art" and "Life" "Ideal" build a relationship. 1916-1917 years of "fine art", "art" is almost the same amount, and in 1918 the "fine arts" discussion reached its peak, followed by the 1919 outbreak of the "Fine arts revolution". In 1919, the use of "art" also increased compared with the previous, the line chart showed a significant upward trend. "Art" is related to science and life, which is creating True beauty art. 1920-1921 years "Art" and "Fine arts" is similar, in 1922 "fine arts" is still more than "art". But after 1923 the situation has changed, the use of "Art" suddenly passed "Fine art".In 1900-1915 years,“Fine art”means public morality.(The 2016 doctoral dissertation of China Academy of Art:The formation and evolution of the modern concept of “Fine arts”in China:

The concept of “Fine arts”in 1895-1924, by the Dr. Qing Peng) The emphasis of this paper is to reveal “Art” means revolution. The use of the word "art" and the spread of the thought of "revolution" are isomorphic. The "Art" replaced "fine art" means the revolution replaced public morality. Compared with the Marx doctrine, the 54 liberals did not establish a clear and persuasive new morality. In 1923, the historical materialism established the hegemony status, and formed the new revolutionary moral ideology.

In this paper, selected "New Youth" (1915-1924) as the statistical text, we could find out the sentences which are containing the word "Art". 1923-1924 suddenly appeared "Bourgeois" "Proletariat" "Advocate", which represented the Marxist Leninist revolutionary discourse, to build the relationship between "Art" and "Revolution".

In summary, the formation of the contemporary "Art" concept, corresponds with the "Revolution". When the revolution has become a new moral pursuit, art has also taken off the old definition of technology, become a pursuit of life, the artist's historical status can be improved. With the establishment of historical materialism, people's outlook on life from the beautiful artistic pursuit of development becomes a kind of revolutionary utopian life.

Keywords: art, revolution, the may fourth new culture movement

20世紀初韓國、中國的文化觀念及文化運動

宋寅在*

摘　要

一、導論

本文關注1919年左右所謂“新文化運動”共同進行在韓國與中國的知識界，探討經過政治社會上激變的時期，“文化”觀念起的作用及其歷史意義，比較兩國狀況的同一性及聯係性。筆者認爲，瞭解同一時期韓中兩國的類似運動，會有助於掌握東亞現代性的跨國特點。其次，這一時期在韓中兩國新文化運動的核心基地也是該時期各國比較有代表的雜誌《新青年》、《開闢》，因此本文也可以説明1910年代末1920年代初中韓兩國的重要雜誌對於重要觀念、行動的作用。再次，《新青年》與《開闢》都已經數位化，研究人比較方便地算出自己有興趣的分析結果。總之，本文的意義不僅在於貢獻於近代韓中觀念史的交流及比較研究的發展，而且進行兩國數位化資料及系統的比較，還成爲探索兩個系統聯繫的基礎。

二、研究方法

本文爲了討論20世紀初東亞的“文化”觀念的演變，“文化”觀念與文化運動的關係，并行運用國立政治大學「中國近現代文學及思想史專業數據庫（1830-1930）」及韓國翰林大學「翰林概念史語料庫」（收藏一共19種近代期刊），分析20世紀初“文化”觀念的語義演變，詞彙共現關係、作者、期刊、文章等各種語境。再加上，比較在1920年左右韓中兩國進行的文化運動及這一時期“文化”觀念出現、運動的特點，闡明這一時期兩國文化運動、“文化”觀念的聯繫。

三、《開闢》資料簡介

(一) 運動後殖民時期韓國新文化運動的中心基地

1919年韓國與中國都發生了重要政治運動，就是3.1運動、5.4運動。雖然這是政治運動，但文化也成爲此時期的重要觀念，並且跟政治運動形成聯繫，特別是韓中兩國的歷史上都發現所謂“新文化運動“的現象。除了在中國《新青年》爲主的新文化運動進行以外，1919年後韓國的一群知識分子還進行“朝鮮新文化建設運

* 翰林大學翰林科學院教授，Email: progsong@gmail.com。

動”，其中心期刊就是《開闢》。《開闢》是韓國近代宗教“天道教”爲主的，天道教是從“東學”發展的，即甲午年農民運動的重要力量。因此1920年代韓國的文化運動也跟近代政治運動有關。並且《開闢》的編輯部得知關於中國新文化運動的信息，跟胡適等部分新文化運動人士聯繫過。這説明近代韓中新文化運動不祇是一國性事件而是東亞的區域性事件。因此此時期兩國新文化運動的比較與疏通值得研究。本文爲了實現這種意義積極運用數位人文方法，進行初步討論。

四、在《開闢》關鍵詞‘文化’、‘新文化’的語境、語義

《翰林概念史語料庫》檢索結果表明，20年代的文化觀念大體上在建設朝鮮固有的文化或新文化的語境上使用了。頻率第一到第十的貢獻詞是這樣：文化、運動、我們、東方、西方、朝鮮（韓文）、社會、政治、朝鮮（中文）、民族。這表示，文化在討論區域類別、人類生活因素的語境下出現。仔細來說，前邊出現跟區域相關的共現詞，後邊大多出現建設、運動、事業、做、製造等形成新文化的動詞的共現詞。結果表示，《開闢》頻繁地談論，“東方、西方、朝鮮文化的比較”、“建設、製造朝鮮文化”，讓人判斷：1920年代韓國的文化這一次從在運動的脈絡上出現而具有爲改變韓國有所作用的動態性。

五、1919年左右中國的文化/新文化

「中國近現代文學及思想史專業數據庫（1830-1930）」的檢索結果得知，在1920年關鍵詞‘文化’、‘新文化’的有關頻率，期刊、作者、文章的信息。這不限於《新青年》，還顯示在其他期刊。在年度頻率上，1919年次數突然增加，1920年達到高峰。有趣的是，“新文化”一詞在1919年前一次也未出現過，可是在這兩年顯示跟‘文化’一樣的趨勢。據此可判斷，1919、1920這兩年在關鍵詞的頻率上也是“文化/新文化”的高峰時期。另外，“新文化”一詞的文章語境告知有意思的情況。頻繁使用“新文化”的文章大部分討論界定“新文化”，表示當時“新文化運動”已成爲流行詞。更具體來説，新文化成爲了跟近代中國青年的生活、中國前途有關的術語。文化/新文化的頻率、新文化的文章語境表示，在文本脈絡上也可以成立。

六、結語

有關“文化/新文化”的文本資料的分析表示，在1920年左右文化觀念帶上前綴“新”一邊對應時代激變，一邊做出貢獻於韓、中兩國的近代轉型。因此，關於“文化/新文化”的信息分析告訴我們，當時文化觀念跟運動形成了緊密的關係，成爲了具有動態性的運動觀念，不祇是形容人類生活因素或方式的靜態性觀念。

關鍵字：文化、新文化、建設、運動、《開闢》、《新青年》

The Idea of Culture and Cultural Movement in Early 20c

In-jae Song *

Abstract

About 1919 New Culture Movement occurred simultaneously in Korea and China. The Journal *Xinqingnian*(新青年) and *Gaebyeok*(開闢) leaded it at each country. *The Database for the Study of Modern Chinese Thought and Literature* (1830-1930) and *Hallym Corpus of Conceptual History* can be the utility of Text Mining for humanistic analysis. The frequency and co-occurrence information of keyword, the information of bibliography is the main data for the analysis of this paper. Digital data New Culture Movement of provides us information as follows. Through this movement the concept of "culture" realize its unique meaning, that not just about the way of human life or itself, but represent aggressive action to change society of China and Korean. Therefore, we can say that "culture" was dynamic concept that have political and social significance. The basis about this information can draw from digital humanistic method. The construction of database of the text of modern age of China and Korea can be helpful for studies on the comparison and communication of the modernity of two countries.

Keywords: culture, new culture, construction, movement, *Gaebyeok, Xinqingnian*

* Professor at Hallym Academy of Sciences, Hallym University. Email: progsong@gmail.com.

臺北市日治時期建築古蹟知識本體之建置與應用

鄭有容*、張映涵**

摘　要

古蹟是國家與生俱來的豐沛文化資產、先人智識傳承的文化寶庫。在臺灣，絕大多數的古蹟皆屬建築類別，而建築的特色往往反映了興建時之社會文化背景與技術成就。近年在政府的推動與重視下，有越來越多專家學者投入文化資料數位典藏工作，彙整許多豐富的詮釋資料。另一方面，伴隨開放資料運動的全球化浪潮，臺灣亦於 2013 年建置「政府開放資料平台」，截至目前為止已累積超過 2 萬餘筆開放資料，鼓勵民眾、專家學者、技術專業人員取用。儘管目前已有文化部建置的「iCulture 文化資料交換平台」，與民間團體自行開發之「Find 古蹟」、「古蹟在這兒」等應用軟體，但運用古蹟類開放資料的範例仍不多，更遑論將開放資料進一步語意化、鍵結化呈現。有鑒於此，本研究以首善之都臺北市中的日治時期古蹟為研究對象，運用語意分析技術將文化類開放資料轉置為鍵結資料，並以數位人文工具與技術將古蹟資料分析與應用，使舊有文化資產賦予新意，促進文化領域與不同領域間交流及對話。具體而言，本研究目的為**（一）建置建築古蹟知識本體與視覺化關聯地圖。（二）推論建築古蹟知識本體中建築物與建築技師之語意關係。（三）結合地理資訊系統探究建築古蹟分佈情形。**

為建置臺北市建築古蹟知識本體，本研究首先透過文化部所建置的「文化開放資料服務網」取得的 824 筆古蹟開放資料（包含名稱、古蹟級別、創建年代、經緯度等欄位），為使資料內容更為豐富，次利用同是文化部建置的「國家文化資產資料庫」擴充古蹟資料內容（包含建築類型、建築材料、建築風格等欄位），另透過中央研究院臺灣史研究所建置的「臺灣總督府職員錄系統」，獲取相應之建築技師基本資料（包含服務單位名稱、官職名等欄位）。在彙整資料後，透過文獻分析探究建築古蹟相關研究，建立「日治時期建築類型與風格分類綱要」，用以辨別建築古蹟資料描述詞彙間的語義關係（例如紅磚自由古典風格又稱為辰野金吾風格等），並萃取建築技師之間的人際關係（例如交友關係、師生關係、

* 國立臺灣大學圖書資訊學系暨研究所博士生，Email: ron0816@gmail.com。
** 國立臺灣大學圖書資訊學系暨研究所博士生，Email: inmiddleman88808@gmail.com。

聘僱關係等），整合不同類型之數位典藏詮釋資料，豐富古蹟知識內涵，更成為建置知識本體之基礎。

基於上述古蹟與建築技師資料，本研究續透過資源描述綱要（Resource Description Framework，簡稱 RDF）中的三元組（Triple）概念，以主體（Subject）、屬性（Predicate）、客體（Object）描述結構來建構建築古蹟的語意關係，並以蓋提研究中心（Getty Research Institute）所發展的鍵結開放資料詞彙（Getty Vocabularies as Linked Open Data，簡稱 Getty Vocabularies）做為 RDF 詞彙描述依據。由於 Getty Vocabularies 包含了藝術與建築索引典（Art & Architecture Thesaurus, AAT）、蓋提地名索引典（Getty Thesaurus of Geographic Names, TGN）、藝術家人名權威檔（Union List of Artist Names, ULAN），以及文化物件名權威檔（Cultural Objects Name Authority, CONA）等四套有關人文藝術作品之知識組織詞彙，較適用於描述本研究中建築古蹟與建築技師的語意關係，故本研究予以採用。研究者利用 Getty Vocabularies 建立「建築古蹟鍵結資料與詮釋資料語意對照表」後，再以知識本體工具 Protégé 建置建築古蹟知識本體。

在知識內容的呈現上，一般常用 Protégé 內建軟體（如 OWLViz）建立關聯圖表，然而實際呈現上往往不能清楚、脈絡化的讓研究者推論知識本體的語意關係，因此研究者嘗試利用社會網絡分析軟體 Gephi 建立知識本體視覺化關聯地圖。本研究認為，在本質上儘管知識本體與社會網絡分析（Social Network Analysis, SNA）中以節點（Node）與屬性（Edge）描述結構網絡關係方式有所不同，然而在圖形呈現上，RDF 的 Subject 與 Object 概念與 SNA 中的 Node 概念相似，而 RDF 的 Predicate 概念以 SNA 的 Edge 概念詮釋，也恰如其分。因此本研究自 Protégé 匯出知識本體檔案，將 Protégé 檔案與 Gephi 檔案進行資料欄位對照與資料轉置，並以 Gephi 呈現建築古蹟知識本體的視覺化關聯地圖，以便進行語意推論。最後，為釐清建築古蹟知識本體之歷史時空演變，本研究以人文地理學（Human Geography）概念，將建築古蹟中的經緯度與分布時代等資料匯入地理資訊系統（Geographic Information Systems），便於呈現日治時代古蹟的時空脈絡。

倘能彙整知識本體產生之鍵結資料，妥善運用視覺化工具以呈現視覺化關聯知識地圖，將能有助於人文學者進行語意推論。審視本研究結果，主要有以下三點：（一）**歷任總督府技師出自同門，但建築風格多元**：日治時期臺灣總督府歷任技師（如井手薰、近藤十郎、野村一郎、松山森之助）皆畢業於日本東京大學，然而在建築式樣的設計上卻十分多元（包含紅磚自由古典、紅磚折衷、馬薩風格、混和風格、仿羅馬風格等）。（二）**辰野金吾建築風格在臺傳播路徑來自於學生與**

同學：辰野金吾為日本建築名家，建築特色為紅磚外貌與橫貫立面的白色水平飾，雖本人從未踏足臺灣，但在臺北市建築古蹟中卻不乏辰野式建築（例如總統府、公賣局、建中紅樓等）。建築學者認為，日治時期來臺擔任建築技師者，多數曾於日本求學時受教於辰野金吾。透過語意本體之探究，發現主要來自其學生（森山松之助）、與同學（近藤十郎）所建造，呼應了建築學者的考究。另外，其助手井手薰雖曾受雇於辰野金吾之建築事務所，但似乎在臺未設計辰野式樣建築；（三）**日治時代建築古蹟分布集中在現今臺北市西區**：在臺北市日治時期古蹟中，分布密度以現中正、萬華區比率最高。

本研究提出政府開放資料的應用模式，透過數位人文工具（包含知識本體工具、資訊視覺化工具、地理資訊系統工具），將臺北市日治時期古蹟詮釋資料加以語意化、鏈結化呈現，進而促進古蹟開放資料的活化與運用。儘管上述結論是基於古蹟鏈結資料所產生的推論，仍需進一步諮詢領域專家以獲取更多有力資料佐證，然本研究藉由客觀的開放資料運用，初步建置出臺北市日治時期建築古蹟之知識本體與應用模式，可供從事人文研究之專家學者參考，並喚起不同領域學者對於在地文化資產的投入與重視。未來還有許多研究課題尚待努力，具體包含：（一）**拓展在地文化資產的全球視野**：嘗試將本研究所的古蹟資料匯入英文版維基百科（Wikipedia），建立 Wikipedia 與 Dbpedia 資料自動對應機制， 增加取用臺灣文化資產開放資料的便利性與整合性，另建立建築古蹟知識本體的 Endpoints，與世界鏈結資料串聯，使古蹟鏈結資料符合開放資料之五星國際交換標準格式；（二）**擴大並深化研究範疇**：利用 SPARQL 查詢語法至國際上的 Endpoints 進行檢索，取得整合性的鏈結資料進行跨區域、時代背景與主題的人文研究；（三）**評估詞彙標準及語意視覺化工具**：由於國內外建築古蹟之知識本體範例不多，盼能輔以深入訪談或領域分析等研究方式，評估適合我國建築古蹟特色的國際性詞彙描述標準與視覺化工具，以符合實際學術研究需求。展望未來，本研究期盼能將具有在地特色的臺灣建築古蹟鏈結資料進一步與國際大型、主流知識本體串連，除拓展文化資產之國際視野外，更重要的是讓國際研究學者能共同參與，發掘潛在的脈絡與議題。

關鍵字：鏈結資料、知識本體、文化資產、建築古蹟、資料視覺化

Construction and Application of the Ontology on Taipei's Historical Sites During Period of Japanese Rule

Yu-jung Cheng*, Ying-han Chang**

Abstract

Historical sites are the rich cultural assets of the nation, as well as the cultural intellectual of ancestors. In Taiwan, most of the historical sites are in the architectural categories. The architectural features often reflect the construction of social and cultural background and technical achievements. With the globalization of the movement on "Open Government", Taiwan has developed a platform of " DATA. GOV.TW" in 2013, till now, more than 20,000 open data have been accumulated to encourage people, experts and technical professionals to access, use and compiled lot of interpretations. In this paper, we choose the historical sites which are built during the period of Japanese rule in Taipei City as the research object, and used semantic analysis technique to transpose the cultural open data into linking data, finally we applied some digital humanities tools to present the results. Our goal is to bring the new ideas into the old cultural assets and try to connect the different cultural fields. The purpose of this study is: (a) Build the historical sites ontology and visualize with the knowledge map. (b) Surmise the semantic relation between historical building and architect. (c) Combine with geographic information systems to explore the distribution of historical buildings.

In order to construct the historical sites ontology, this article obtained 824 historical data from the "Open Data Services for MOC" website, which is made by the Ministry of Culture, and we added the architectural descriptions, such as architectural types, materials, styles for expanding the content by searching the " Office of the Governor-General Staff Records Database", which is built by the Institute of Taiwan History of Academia Sinica. After collected all the information, we established a "classification schema of architectural style during period of the Japanese rule" to

* Ph.D. Student, Department and Graduate Institute of Library and Information Science, National Taiwan University. Email: ron0816@gmail.com.
** Ph.D. Student, Department and Graduate Institute of Library and Information Science, National Taiwan University. Email: inmiddleman88808@gmail.com.

identify the semantic relationship between historical data and description terms, for instance, Western free Style also known as Tatsuno Kingo Style. And extracted the interpersonal relationship between architects (such as friendship, teacher-student relationship and employment relationship). To integrate different types of digital collection data not only to depth the knowledge of historical sites but also to become the ontology basis. Moreover, based on the above data, this article continued to describe the structure by adapting to the Triple concept of Resource Description Framework (RDF), including Subject, Predicate and Object. Then we choose the Getty Vocabularies (abbreviated form a name of Getty Vocabularies as Linked Open Data) as the RDF description references, which is developed by Getty Research Institute, in the end, we successfully built the historical sites ontology with Protégé and Gephi.

The results show as following three points: (a) Successive architects from the Office of the Governor-General are out of the same door, all graduated from Tokyo Imperial University, however, their architectural style are full of variety. (b) In Taiwan, the distributing paths of Tatsuno Kingo Style mostly from his students (such as Matsunosuke Moriyama) and schoolfellow (such as Kondo Juro). (c) The area of the historical sites is concentrated in the present-day Taipei City West district (Zhongzheng District and Wanhua District). This article puts forward the application way of government open data, through using digital humanities tools (including ontology tools, information visualization tools and geographic information system tools) to analyze the data of historical sets during period of the Japanese rule in Taipei City semantically, and promote the activation and utilization of historical information. Although the conclusions are based on the inference of historically linked data, we still need to consult domain experts to get more powerful data as testimony. The study can be used as a reference for experts and scholars in the field of humanities research and arouse the investment and attention on the value of local cultural heritage in different fields. Looking to the future, this study expects to link the information of local characteristic's Taiwan historical sites with the international large-scale and mainstream ontology, in addition to expanding the international perspective of cultural assets, and more importantly, making international scholars participate and explore the potential context and issues.

Keywords: linked data, ontology, cultural heritage, historical sites, data visualization

運用社會網絡分析探究臺灣政黨合縱與連橫現象：以臺北市議會第 11 屆議員提案資料為例

萬麗慧*、鄭有容**、曾蘭棋***、陳嘉勇****

摘　要

議事紀錄詳載了民主發展的軌跡，是研究臺灣政治史不可或缺的史料，近年由於數位人文技術蓬勃發展，研究者紛紛以不同視角擴展人文研究的面向。在數位人文相關技術中，社會網絡分析適用於探究公共決策過程、政黨乃至於個人關係網絡情況。其中議員提案被視為瞭解政黨網絡關係的重要管道，因為在議員提案的過程有一定連署人數要求，且政黨間亦需保持既競爭又合作的關係，方能順利完成提案，上述議員提案連署關係可交織成動態且複雜的社會網絡圖像，有助於瞭解在議員提案中的權威角色與中介角色。有鑑於此，本研究以臺北市議會第 11 屆議員提案資料作為分析文本，運用社會網絡指標探究不同政黨網絡間的合縱連橫關係。具體而言，本研究有以下目的：**（一）瞭解不同政黨在議員提案網絡結構特性。（二）瞭解不同政黨在議員提案網絡交流情形。（三）運用多重網絡指標驗證不同政黨在議員提案網絡的權威角色與中介角色。**

基於上述研究目的，本研究擷取臺北市議會第 11 屆（自 2010 年 12 月 25 日至 2014 年 12 月 25 日止）的議員提案資料，自 62 席議員中挑選出有登記參選第 12 屆議員選舉的 53 席議員作為研究對象，以議員的選區、性別、年齡、擔任議員屆數、政黨、得票率、選舉結果作為觀察項目，分析上述對象所提出的 103 則議案中之議員連署關係。考量以議員與議案連署關係進行分析之原因，係由於現代民主代議制度中，議員是代表人民行使政治權利的公職人員，而議員的重要職權之一為議員提案，身為引領公眾議題的重要意見領袖，議案為反映選區地方事務、選民心聲的重要管道，且經大會審議通過的提案議決結果，市政府有法定義務去落實與建設，往往直接影響人民未來的生活品質與現況。

在研究方法部分，社會網絡分析乃運用圖論（Graph theory）與社會計量法（Sociometry）方式，從網絡關係中找出行動者（Actor）在社會層次網絡、組織

* 閩南師範大學新聞傳播學院副教授，Email: wanlihui2005@gmail.com。
** 國立臺灣大學圖書資訊學系暨研究所博士生，Email: ron0816@gmail.com。
*** 國立臺灣大學圖書資訊學系暨研究所碩士生，Email: black53719@gmail.com。
**** 國立臺灣大學圖書資訊學系暨研究所博士生，Email: d04126001@ntu.edu.tw。

層次網絡、人際層次網絡所扮演的角色，過往研究常以程度中心性（Degree Centrality）、中介中心性（Betweenness Centrality）、接近中心性（Closeness Centrality）作為測量依據，在上述中心性指標的基礎上，本研究另外發展適用於議員提案網絡的「權威角色指標」與「中介角色指標」，以瞭解不同政黨網絡結構特性與交流情形。在權威角色指標部分，除程度中心性指標外，本研究另以特徵中心性（Eigenvector Centrality）指標、核心邊陲（Core-Periphery Analysis）指標來瞭解不同政黨網絡結構間，在議員提案上扮演權威角色的議員；在中介角色指標部分，基於中介中心性指標基礎，研究者納入結構洞（Structural Holes）指標來探究在議員提案中扮演中介角色的議員，最後利用 E-I（External-Inernal）指標、群聚係數（Clustering coefficient）指標與網路總體密度（Density）指標來討論不同政黨間之交流情形。

綜上所述，本研究運用多重社會網絡指標，來驗證不同政黨在議員提案網絡，發現不論在權威角色或中介角色上，均有相似研究結果，綜整為以下三點：**（一）國民黨內有較多的權威角色，且黨內聯繫較黨外聯繫密切**：國民黨籍議員的程度中心性、特徵中心性以及核心邊陲的程度比率均高，且群聚係數比例亦高；**（二）民進黨較無權威角色，但黨外聯繫較黨內聯繫密切：**民進黨籍議員程度中心性、特徵中心性以及核心邊陲的程度比率均不高，但 E-I 指標之組間聯繫（External）高於組內聯繫（Internal）；**（三）小黨（包含新黨、親民黨、無黨籍、臺灣團結聯盟）在議員提案上扮演重要中介角色，且與大黨間交流密切**：小黨議員的中介中心性、結構洞程度比例高，且 E-I 指標中組間聯繫亦高於組內聯繫。

本研究透過數位人文相關技術，分析不同政黨間議員的權威角色與中介角色，研究結果顯示，在第 11 屆議員提案中扮演權威角色與中介角色的議員，均與下屆議員選舉結果無關，亦即在議員提案上辛苦耕耘的議員，往往不是得票較高的，或許對議員而言，透過議員質詢或進行選民服務，較議案更能提升曝光度與選票。本研究提出未來發展方向如下：**（一）探究議案內容以檢視議員政見落實程度：**建議針對議案內容進行社會網絡分析，探討核心議案網絡分布情形，並進一步探究核心議案與議員選舉政見之落實程度。**（二）輔以多重研究方法瞭解現象背後的成因：**倘能輔以其它研究方法（例如深入訪談法、問卷調查法等）近行調查，將有助於解釋相關政治現象背後成因之解讀。**（三）綜合不同議事資料進行整合性分析：**議事資料種類繁多，倘能運用議員質詢紀錄、新聞資料、預算審查資料、會勘協調會等重要議事資料來進行分析，將使運用社會網絡分析等相關數位人文技術，在政治、公共行政領域的研究面向更為寬闊。

關鍵字：社會網絡分析、議員提案、政黨網絡、權威角色、中介角色

Using Social Network Analysis to Investigate Political Parties in Taiwan: A Case Study of the Eleventh Proposal of Bills of Taipei City Council

Li-hui Wang [*], Yu-jung Cheng[**]

Lan-chi Tseng[***], Chia-jung Chen[****]

Abstract

Proceedings which is recorded the trace of democracy development, which are important data for studying Taiwan's political history. Recently, because of the development of digital humanity, researchers explore the different part of humanity. The proposal of bills is taken as a way to see the relationship of the political party network, because there is a rule of certain people countersign in the proposal of bills, and each party should maintain the competitive and cooperative relationship, so they can successfully finish the proposal of bills. Councilors' countersign shows dynamic and complicated social network, which is helpful for understanding authoritative roles and betweenness roles in the proposal of bills. In digital humanity, social network analysis is used to explore public decision, political parties as well as councilors' relationship. Therefore, this study's dataset draws from the 11th proposal of bills of Taipei City Council and using social network indicators to explore different political parties' relationship. The purpose of this study is: (a) exploring the network structure of different political parties in the proposal of bills; (b) exploring the network interaction of different political parties in the proposal of bills; (c) using multiple network indicators to verify the authoritative roles and betweenness roles in different political parties with which in the proposal of bills.

[*] Associate Professor, School of Journalism and Communication , Minnan Normal University. Email: wanlihui2005@gmail.com.

[**] Ph.D. Student, Department and Graduate Institute of Library and Information Science, National Taiwan University. Email: ron0816@gmail.com.

[***] M.A. Student, Department and Graduate Institute of Library and Information Science, National Taiwan University. Email: black53719@gmail.com.

[****] Ph.D. Student, Department and Graduate Institute of Library and Information Science, National Taiwan University. Email: d04126001@ntu.edu.tw.

On this account, based on the existent data in the proposal of bills of Taipei City Council (2010-2014). There are 53 of the 62 councilors who registered the 12th councilor election were selected as the subjects in the study. We analyze 103 proposals of bills proposed by these 53 councilors with consideration of their electoral districts, gender, age, sessions, parties, percentage of votes, and election results. Analyzing 103 proposals of bills and countersign relationships of councilors. In the representative democracy, councilors are civil servants who represent people to exercise political right, and one of the councilors' duties is the proposal of bills. The proposal of bills is important ways to reflect voters' opinions as well as something important in the constituency. Besides, the government has the legal obligation to implement the proposal of bills' resolution. Therefore, this study analyzes the relationship between councilors and the proposal of bills.

Regarding the research method, social network analysis is about using graph theory and sociometry to detect the roles of actors in social level, organization level and relationship level networks. Most previous studies employ degree centrality, betweenness centrality, and closeness centrality as basic measures, while in our study, the authoritative role index and betweenness role index are proposed specifically for councilor network to analyze the network structure and communication of different parties. The authoritative role index is a combination of degree centrality, eigenvector centrality, and core-periphery analysis to represent the key councilors in the proposal of bills. Based on betweenness centrality, betweenness role index focuses on the councilors who act as mediation in the process of the proposal with analysis of structural holes. Eventually, we make use of E-I (external-internal) index, clustering coefficient, and density to observe the communication between different parties.

Overall, by the application of multi-level social network measures, we find authoritative roles and betweenness roles mainly share three attributes: (a) Pertaining to the authority role in the network structure of the councilors' countersign, the Kuomintang (KMT) councilors are in majority, and the connections inside the party are more intimate compared with the outside: The degree centrality, eigenvector centrality, level of core-periphery, and clustering coefficient of the KMT councilors are all high; (b) There are fewer Democratic Progressive Party (DPP) councilors in the coalition network, yet the connections outside the party are more closely than the

inside: The degree centrality, eigenvector centrality, and level of core-periphery of DPP are not high, but the external connections were higher than the internal. (c) Small parties (including the New Party, the People First Party, nonpartisan politician, the Taiwan Solidarity Union) play an important betweenness role in the proposal of bills, there are intimate interactions between the small and major parties: The betweenness centrality and the structural hole degree of the councilors from small parties are higher than major parties, and moreover, the external connections were also higher than the internal.

We analyze authoritative role and betweenness role of the councilors in different parties with digital humanity technologies. The research result shows that the councilors who play authoritative role or betweenness role in the 11th proposal of bills are irrelevant with the next election results. That is to say, councilors who spent a lot of time in the proposal of bills are not likely to win the election. Interpellation or voter service might have a greater influence in attracting public attention and increasing the votes. Our future work will have the following three directions: (a) Investigate the contents of the proposals to examine the extent of the implementation of core proposals and political opinions of the councilors; (b) Enhance the research by introducing other research methods such as interviews and questionnaire surveys. (c) Conduct the comprehensive analysis of relevant proposal materials and make the social network analysis techniques more applicable in political and administrative areas.

Keywords: social network analysis, proposal of bills, political party network, authoritative role, betweenness role

大數據與社會學

蕭煒馨*

摘　要

使用者經驗（User experience）在各種領域都佔有一席之地，包含設計、商業行銷、醫療、遊戲、軟體等等產業，都會在產品推出前後，執行相關的調查。從傳統的問卷調查、訪談到眼動追蹤（Eye tracking），這些方式都能夠間接或直接了解人使用某種技術或產品後，所產生的反應、感想、意見、使用目的等（維基百科「使用者經驗」條目）。這類以「研究者」或「觀察者」出發的經驗研究，非常依賴研究者本身的背景與訓練。在調查使用者經驗時，時常是在調查時，使用者才開始思考自己為何使用該項技術或產品，以社會學的方式來說，這是一種事後的建構。因此，觀察者本身也成為使用者經驗的一部份，因為他們與使用者之間的問答，也會是影響調查結果的一個因素，研究者當下的每個問題，都讓使用者為過去的自身經驗重新賦予意義，或者回憶與建構自己當初使用產品的目的。

臉書（Facebook）是目前擁有最多使用者數量的社群網站，也累積了巨量的使用者資料。相較於上述由觀察者預先擬訂的問卷和選項，臉書上的使用者在表達意見時，是遵循著臉書的二元思維模式，是否對一篇文章或一則動態按下「讚」、是否點擊一篇長文下的「顯示更多內容」超連結、是否特定「社團」顯示有新文章時，就進入其中閱讀等等。這些是非題，都是臉書用來累積並分析使用者的工具。如果單就使用者是否對一則動態、一篇文章、一個超連結產生反應來看，二元思維模式必定無法描繪出使用者的輪廓。但如果是二元思維，加上長久時間的累積，資料的巨量增長和演算法不斷改寫，改變運算資料的公式，就可以消除二元思維模式過度簡化的盲點。例如《物聯網革命》一書中所引用的例子：零售通路的感測器能通知銷售及行銷部門，讓他們知道顧客察看或觸摸了哪些品項、把哪些品項放回貨架，或已購買什麼品項等，從而協助評估消費者的行為。（Jeremy Rifkin,2016:21）在臉書上也是如此進行，分析使用者讀了含有哪些關鍵字的文章、搜尋了哪些社團名稱、臉書上使用的語言等等，都有助於臉書向各

* 德國維藤/海德克大學文化反思學院社會學系博士候選人，Email: weihsinhsiao@gmail.com。

大企業說明，這些數據是「直接」來自使用者，而得以避開研究者篩選樣本時，可能遭遇的困難與盲點，如樣本數不足、研究倫理上的困境等。

從上述零售業通路和臉書的案例中，我們可以引入「媒介」的概念。過去，我們依賴研究者進行的使用者經驗分析，描繪出使用者輪廓，並進一步預測這群使用者未來的趨勢。現在則是直接收集使用者所產生的資訊，不論是他們觸碰商品的類型、停留在貨架前的時間；或是他們在臉書上最常點閱的文章，含有什麼關鍵字、哪個使用者的動態最受歡迎等，都是透過使用者的滑鼠點擊與演算法的運算，才得以呈現為分析結果。過去，研究者是媒介，他們「共同參與」了使用者經驗所呈現的結果；現在，使用者自己就是媒介。隨著他們在網路上的時間越來越長，閱讀臉書上的新聞、文章和動態，在購物網站上用關鍵字搜尋折扣優惠等這些行為，就不斷累積成資料，這同時也形塑了使用者自身的輪廓。對社會學來說，最大的挑戰正好就發生於此：如果使用者透過自己再加上大數據、演算法、電腦、手機和網路，這幾種媒介就可以描述自己的話，也無需社會學者從一個外在與建構的角度來描述使用者。那麼，研究者或觀察者的位置在哪裡？或者，我們還需要社會學嗎？

由德國社會學家 Dirk Baecker 所提出的「後設資料」（Metadata）概念，已經為資訊時代的社會學研究提供一些工具。《大數據》一書的作者，麥爾荀伯格和庫基耶提醒我們，應該更注意資料之間的相關性（Mayer-Schönberger&Cukier, 2013:74-79）。因此，過去社會學或使用者經驗研究所強調的因果關係，已經逐漸稀釋它的重要性，反而能夠描繪各種資料的不同模式，以及觀察者在這些模式中不斷轉換其觀察位置，甚至是設計出不同的資料連結模式，才能對於描述或分析巨量資料有所貢獻，而這也是社會學能夠有所助益之處。（Baecker, 2014:184）

首先，使用者利用不同的媒介將自身視覺化，也讓使用者彼此可以相互觀察，並進一步與彼此產生關係。在這樣的意義上，使用者可以看成是「訊息」（information），它可以不斷的生產出新的訊息。不斷自我生產新訊息的方式，就是透過上述的二元思維模式：點擊超連結看完全文；或是將螢幕繼續往下滑動（點擊或不點擊），按讚或不按讚，留言或不留言、購買與不購買、不購買的商品是否仍留在購物車裡等等。臉書或購物網站上的使用者，並不是人以為基本單位，而是以「訊息」為單位，它的特色在於，能夠不斷地自我再製。此時，研究者開始可以想像，使用者是隨時在改變自身狀態的，他們隨時透過滑鼠展現他們喜歡什麼、不在意什麼、瀏覽過什麼等等。過去的使用者經驗研究，一旦回收問卷和結束訪談後，就等同於將使用者框限在特定的時空脈絡下，再由研究者進行

分析，這種研究方式預設使用者在某個時段中是穩定的狀態，也因此研究者可以觀察到並分析他們。而臉書累積大數據的機制，則是可以無時無刻捕捉到使用者的動態，也開展了另一種對於使用者的想像：使用者是動態且跳躍式思考的。使用者在每個時刻的思考，如果都展現在他們使用臉書時的滑鼠點擊，那麼臉書就成為一個可以描繪出動態使用者的平台。這種思考上的轉向，同時也能反饋（feedback）到社會學中，讓使用者所呈現的樣貌，不再只是由特定時空脈絡形塑的單面向輪廓，而是更多元且動態的描述。由「區分」使用者與否、按讚與否、購買與否等等二元式的思考方式出發，再藉由社會學的「後設」思考：也就是從資料彼此的關係、模式或秩序來觀察使用者，不但能更有效地呈現使用者的動態樣貌，也為社會學自身持續注入研究的動力。

本論文透過兩種特別重視使用者的案例－使用者經驗與臉書出發，來討論新型態媒介－電腦與網路及隨之而來的數位化，對於社會學產生的挑戰，以及社會學的回應方式。使用者經驗代表著以統計學所呈現穩定且靜態的世界輪廓，而臉書則代表著當下或者未來所面臨的急遽變動的數位世界。藉此，社會學不但能夠反省自身學科中，長久以來的質、量化研究方式之間的對立，並提供解決之道；也能夠貢獻自身學科中豐富的理論工具、後設思考模式和長久以來對統計資料的嫻熟分析，為其他領域提供各種理解或詮釋資料的方式。當然，社會學也能夠從其他領域對數位化的應用，擴展自身的研究視野，並激起跨領域研究的眾多火花。

關鍵字：大數據、媒介、訊息

Big Data and Sociology

Wei-hsin Hsiao*

Abstract

With the appearance of the computer and the internet, various organizations, industries and academic disciplines are forced to dealing with these two media and solving consequently rising problems. Recently the concept of Big Data is challenging the sociology which concentrates specifically on finding cause and effect relations by quantitative or qualitative methods. Following Viktor Mayer-Schönberger and Kenneth Cukier, Big Data reveals a transformation from causality to relations: not knowing *why* but only *what*. (Mayer-Schönberger and Cukier, 2013:7) The concept of Big Data refers not merely to the enormous amount of data but also to patterns, relations and organization of data. Furthermore the concept of Big Data reflects the instability of data which are always changing their status. In order to analyse these dynamic Big Data, researchers have to keep themselves flexible as well. The uncertainty and instability between observers and data or between data themselves will shape and illustrate our next society.

Most of user experience surveys try to explain, why users are using or buying products. Researchers collect questionnaires from their research objects and interview them, namely users and customers. In this context users and customers are involved in the complex of assumptions from researchers, such as the concepts of action, cause, reason, purpose, intention. They are usually featured as rational actors who are able to make decisions. Accordingly the researchers are designing their questionnaires, arranging the interviews and collecting these data. They also notice their research objects are so unstable that the surveys will be regularly conducted and the questions will keep refreshing. As the concept of Big Data appears, the user experience surveys can't elaborate the instability of users. Instead of those conceptual assumptions the

* PhD Candidate in the Fakultaet fuer Kulturreflexion at the Universitaet Witten/Herdecke, Germany. Email: weihsinhsiao@gmail.com

concept of Big Data concentrates on if users and customers act or not rather than why they act. Now researchers' task is turning to collecting and analyzing data produced by users instead of asking what users really think.

For example, Facebook collects if their users click "like" or not. If they click, Facebook will keep them continuously clicking by feeding similar articles. If they don't click, it will keep feeding them various news and contents to motivate their clicks. Instead of the traditional assumptions to users Facebook characterizes its interface design and algorithms as approaches to collecting the data and sketching users. In contrast to these classic surveys of user experience in which users are recognized as temporally stable and reasonable (so that they are observable), the collecting data by Facebook will reveal users who are always reacting and thus unstable. Users are stable only when they click hyperlinks or act and immediately change their status. Hence the instability of users is normal state and plays a central role. Obviously the regularly arranged interviews or refreshed questionnaires are too stable to describe the endless varying state of users.

As an empirical case of Big Data we can observe Facebook more precisely. It uses numerous hyperlinks to collect data from users including "Like" button or "See more" and so on. All kinds of these functions are oriented towards arousing and collecting users' clicks (actions). Applying this idea to the user experience surveys is firstly motivating research objects to act. From a different perspective users are always changing their state. Including touching products, discussing about them, talking with friends, standing in front of the products and walking around the products might be new variables for conducting these surveys of user experience. Even if users and customer are not buying any product, they actually pay their attention to those products. Most sociological researchers take their research objects as actors. An actor will not be noticed until he changes his status or acts. Under this condition non-actions of research objects are easily ignored by researchers. Statistically these collecting data can't illustrate dynamic states of research objects. In this case the researchers can't play a role as medium between users and companies. The users can directly reveal their preference or attention to some products without any surveys conducted by the researchers.

Unlike other user experience surveys which are only concerning visible actions

or statistically significance Facebook bases on its classifications, collects data as much as possible and then analyses them. Depending on Metadata Facebook determines its classifications and categories to collect data from users. At first Facebook demands its users registering and distinguishes them from non-users who will be taken as potential users and possibly later register. In Dirk Baecker's sense metadata is finding concepts to distinguish, compare, list and sort data. (Baecker, 2013:161) The distinction between users and non-users is thereby a kind of metadata. After users register on Facebook, they will be recognized as dynamic. Even if they idle on Facebook, the algorithms will collect their idle time and save it in the database. Continuously these data will be analyzed in which contexts users idle. Hence the idea of Big Data and metadata can capture all status of users in various contexts and analyze their relations. Without asking the reasons why users idle the results of calculation by algorithms can still reveal the idle status in various contexts, namely the diversity of idle. Depending on different viewpoints of observers and researchers Facebook can further decide how it deals with the idle status of users. It might keep feeding users fresh news and articles with different keywords to draw users' attention and encourage their clicks. It might also keep collecting and calculating the idle time of users to estimate if they are reading some articles or just doodling. Facebook can offer its analyses to companies which can improve their fan pages. In comparison to user experience surveys which design specific questions to investigate user experience in individual context, metadata means the order of data which depends on the observer how and from which viewpoint to organize and analyze data. According to these metadata Facebook also has chance to modify or extend its algorithms.

Apparently the methods applied by user experience surveys and Big Data are different. The former assumes that users are so static research objects and rational actors that researchers merely conduct interviews and surveys to describe how users look like. The latter is major in the dynamic status of users. Now their unstable status can be visualized by the technologies and is empirically observable for researchers. Referring to system theory and form theory the concepts of communication, form and metadata are more suitable to explain the phenomenon and solve the problems in the next society.

In the example of Facebook metadata bring Big Data into order, algorithms

calculate and analyze the relations of data, the interface design visualizes the results of calculation by algorithms and all of these arouse more new data inputted by users. The more data Facebook receives, the longer this system survives. Under system theory and form theory in sociology the society will be collapsed by none or few of data rather than mass of data. Users are produced by Facebook which always demands internet surfers registering. Data are produced by users who select clicking “Like” and by non-users who still anonymous surf on Facebook. Following system theory and form theory data or users are constructed by the distinctions. When the distinctions are made either by users or by Facebook, it means that information emerges and will continuously reproduce itself. Therefore drawing a distinction, which comes from the mathematician George Spencer-Brown, will properly illustrate our next society and cope with the problems in the current sociology. Actually the sociological system theory and form theory are always getting ready to confront with the challenges in the information age. The concepts of Big Data and metadata will feedback to the sociology which is stuck in the controversy between qualitative and quantitative methods.

Keywords: big data, distinction, information and medium

實名中的匿名：臉書上的告白／黑特風潮

陳韻如*

摘　要

2014 年五月初，熱門社交網站臉書上突然出現許多以告白／黑特為開頭的匿名粉絲專頁，又在 2016 年逐漸沒落。本研究首先區分匿名與化名的概念，並檢閱過去文獻中討論匿名制下的使用者心態，但這只能說明使用者對匿名環境有需求，卻無法解釋不同匿名平台之間的差異。且此解釋路徑亦無法顧及技術物在其中的角色，僅將匿名平台當做一不需加以說明的背景，不但沒有討論相關社會團體（程式設計者、行銷者）的影響，更未觸及技術物本身對其他行動者的影響。本研究採用行動者網絡理論的視野，強調互相影響、共同建構的網絡，在此網絡中，物與人被放在同樣的重要程度下來檢視。透過觀察主要行動者如何轉譯其他行動者的旨趣而建立網絡，以及其他的人或非人行動者又是怎麼影響網絡，企圖了解匿名粉絲專頁興起與沒落的原因。

本研究訪問了 AnonyMonkey Inc.的亞太區負責人與黑特政大的粉專管理員，藉以詳細了解不同匿名粉專創始的初衷與契機。筆者亦從匿名粉專興起時便以使用者的身分深入參與在此現象當中，從 2014 年五月到 2016 年二月之間持續進行參與觀察，除了理解匿名粉專的使用者為何被此平台所吸引之外，也對 AnonyMonkey 和臉書這兩個非人行動者的技術特性與這段期間的演變有更深入的了解。

本研究發現 AnonyMonkey Inc.此團隊在這當中扮演著主要行動者的角色。該團隊因認同匿名能夠帶來正面效果而開始參與美國大學匿名平台的運作，加上受訪者稱為「I want to do something big!」的創業精神，團隊成員便開始以非營利模式推廣匿名粉絲專頁。他們在台灣的推廣，受助於臉書的技術特性與人們在其上社會網絡所產生的「一個拉一個」效應，也非常順利進入許多大專院校與高中，成功徵召一般使用者與粉專小編，使他們加入網絡。隨著使用者增加，原本運用 Google 表單作為匿名遞交媒介的管理方式開始顯得沒有效率，幾經波折後

* 國立臺灣大學社會系碩士生，email: r05325005@ntu.edu.tw。

AnonyMonkey Inc.的工程師自行寫出 AnonyMonkey 匿名遞交軟體，除了完整嵌入臉書、方便管理，AnonyMonkey Inc.能夠根據所收集到的使用者回應不斷改良軟體功能，吸引更多行動者投入。

但 AnonyMonkey Inc.創辦人的畢業卻導致該團隊突然解散，這個穩定的網絡可說是直接失去了主要行動者。若以第一代 ANT 的單一霸權行動能力而論，該網絡應會崩解。在告白系列粉專雖萎縮但仍持續運作的事實中，我們可以看見相對弱勢的行動者如何因應告白台灣團隊的消失以及缺乏維護導致故障甚至完全停擺的匿名程式。在他們不同的實作中，本來一致性相當高的告白系列粉專，開始出現更不同的樣貌。

另一方面，以黑特政大為首的黑特系列粉專自發性地成立了各自的粉絲專頁，使用 AnonoyMonkey 為匿名遞交程式，卻完全獨立於 AnonyMonkey Inc.的運作。而黑特政大小編於 2014 年 5 月的台北捷運隨機殺人案後創立黑特政大匿名粉專，秉持著「每個聲音都是珍貴、值得聆聽」的信念管理。其接近放任的管理風格使得此網絡的豐富樣態成為可能，但黑特政大豐富的內容之所以能夠形成，必須放置在多個行動者的不同實作之中才能理解，也就是說，各個行動者彼此吸引、互相動員，而無法簡單歸因於某個主要行動者的旨趣轉譯工作。

另外，匿名粉專鑲嵌在實名制的臉書中這個事實，也使得網絡的樣貌更加複雜。透過正文中的詳細剖析，我們發現匿名與實名的混雜使網絡中的互動樣貌十分豐富。且臉書上以個人為節點的人際網絡發揮了動員力以及串聯力，再加上粉專貼文公開以及匿名粉專貼文強烈的情緒色彩，使情緒的共感共應得以發生，成為與私密談話性質的匿名平台之間的區隔。在與 2011 年佔領台北的匿名粉專的對比當中，我們可以看見 AnonyMonkey 在匿名粉專之所以可能的網絡中扮演重要的角色，以及臉書粉專的管理員在這樣的匿名粉專中握有掌控言論的權力。對 AnonyMonkey Inc. 而言，言論的管制即是能夠消弭網路霸凌的關鍵所在；但對黑特政大的管理員而言，每個想法都是珍貴、值得聆聽的。而情緒的共感共應所激發的對既有規範價值的挑戰，也使鑲嵌在實名制臉書中的匿名粉絲專頁在關於網路民主的研究中能夠佔有獨特的地位。

本論文之貢獻除了在於具體描繪出告白與黑特匿名粉專的聚散過程，勾勒出網絡中人與非人行動者的交織互動，證明行動者網絡理論在網路社會學的領域能夠發揮分析效力之外，亦提供了一個討論的基礎。未來的研究可在這樣的基礎上去再思索爭議已久的線上民主、網路霸凌之可能出發點。而本論文討論焦點著重

於使用 AnonyMokey 匿名程式的告白、黑特 X 大的匿名粉絲專頁，但實際上，網路上還存在許多其他主題、使用不同遞交程式的匿名平台，或是像 PTT 這類雖屬化名，但與匿名粉專有多項重疊特質之平台。未能詳盡探究、比較這些其它的網路平台是本論文的主要限制之處，也是未來的研究能夠持續發展的方向。

關鍵字：行動者網絡理論、匿名、臉書

Anonymity within Real Name System： The Crush/Hate Fashion on Facebook

Yun-ju Chen*

Abstract

At the beginning of May 2014, anonymous fan pages by the name of “Crush” and “Hate” start popping up on Facebook. However, the number of users of these fan pages started to decline sharply in 2016. This study starts by distinguishing the concept between “anonymity” and “pseudonymity”, and reviews how former studies discuss the users’ mentality of performing anonymity. However, discussing users’ mentality only explains that anonymity is desirable. We can neither know how exactly some anonymous platforms succeed while others didn’t, nor discuss what roles artifacts play in the “Crush/Hate fashion”. If we only take the user’s mentality into consideration, the influence of relevant social groups and artifacts will be ignored.

This study adopts ANT (Actor Network Theory) as the analytical framework, emphasizing on the inter-influence and co-construction of networks. Human and artifacts are examined under the same amount of emphasis within this network. By observing how main actors translate the interests of other actors and how human and non-human actors influence the network, this study tries to understand why these anonymous fan pages came into fashion and why they declined.

In order to understand how and why people set up anonymous fan pages, this study interviewed the chair of Asia of AnonyMonkey Inc. and the administrator of “Hate NCCU”. From the time crush and hate fan pages popped up, the researcher also took part in anonymous fan pages as a user, practicing the participant observation method. By exercising these methods, we discover more details about why the crush and hate fan pages attracted so many users. Moreover, we can learn about the characteristics and development history of the two primary artifacts, AnonyMonkey

* M.S. student, Department of Sociology, National Taiwan University. Email: r05325005@ntu.edu.tw.

and Facebook.

This study explores that AnonyMonkey Inc. is the main actor of this network. Because the incorporation believed in the bright side of anonymity, they devote themselves to running anonymous fan pages of several universities in the USA. With the so-called “I want to do something big!” mindset, the incorporation started to promote anonymous fan pages without taking profit into consideration. Supported by the characteristic of Facebook and social networks people have already established, the idea of anonymous fan page was widely accepted by college and high school students in Taiwan, and successfully recruited editors. But as the number of users expands, the usual way of submitting Google sheets to editors became very inefficient. After some struggle, the engineers of AnonyMonkey Inc. designed an online anonymous service, AnonyMonkey. This service not only improves their efficiency but also allows the incorporation to collect feedbacks from users. They can then adjust their service accordingly and mobilize more actors to join the network.

However, after the founder of the incorporation graduated from college and left the group in 2015, the company dissolved and the robust network lost its main actor. According to the first generation of actor network theory, losing the main actor would have caused the breakdown of the network. The crush fan pages did decline after 2015, some of them, however, are still running today. In these cases, we can find out how the relatively weak actors respond to the dissolution of the incorporation and the malfunction of AnonyMonkey. And with their different ways of responding, the uniform of crush fan pages became increasingly diverse.

The anonymous hate fan pages popped up autonomously. They have used the service of AnonyMonkey but are independent of the incorporation. The administrator of “Hate NCCU” created the fan page in May 2014, right after the Taipei Metro attack. He believed that every thought is precious and worth listening. With this belief, the administrator managed the fan page with high tolerance. The high tolerance makes the fan page become rich and colorful. But it is not the only reason to explain the vibrant content of Hate NCCU. In other words, in order to make this network possible, enrollment and mobilization between different actors are necessary.

In addition, a real name system, Facebook, is embedded in anonymous fan pages.

This makes things more complicated. The researcher from this study found out that the combination of anonymity and real name system allows the interaction in this network to be very complex and thus interesting. The social network on Facebook also showed its ability of mobilization. Besides, posts on the fan pages are open to the public, and usually with strong emotion. Under this condition, collective emotion happened and showed its power. Moreover, this study compares “Occupy Taipei”, an anonymous fan page established in 2011, to crush and hate fan pages, and hence shows the importance of AnonyMonkey. We can also discover the crucial role of the administrator and editor of these fan page plays. They monitor all the content of anonymous fan pages and decide whether they want to publish them. To prevent cyberbullying, AnonyMonkey Inc. monitors the content; however, the administrator of Hate NCCU values the freedom of speech more. Besides, the collective emotion arouses provocative actions against norms and even laws. The dispute of freedom of speech and the provocative actions aroused by collective emotion makes the study of anonymous fan pages important to cyber democracy.

This study described how the crush and hate fan pages came into fashion and declined, also outlined the interaction of human and non-human actors in the network. At the same time, it provides a stepping stone. Future research can discuss the long-standing dispute of cyberdemocracy and cyberbully based on this study.

Keywords: ANT, anonymity, Facebook

看得見的公眾：從運算取徑中的閱聽人談起

李長潔*、邱慧仙**

摘　要

政治評論家 Walter Lippmann 在 1927 年出版的《幻影公眾》中描述了他對「公眾」的想像，是一個「不再抱有幻想的人」，像是坐在劇院後排的聾啞觀眾，本應對舞台上的演出充滿熱情，但實在無法使自己保持關注，一切發生的事情似乎都與他無關。這個尖銳諷刺的批評，將自由主義思想中神聖無比的「公眾」，拉進了他著名的社會哲學結論，公眾輿論不是上帝的聲音，也不是社會的聲音，而只是旁觀者的聲音。Lippmann 的「社會學想像」（sociological imagination）困擾了社會科學研究者們許久，尤其是傳播學領域。從 1948 開始，Harold D.Lasswell 等諸多傳播學者便對「公眾」、「民意」、「輿論」等相關派生的概念進行窮盡畢生之力的探索，了解這些主題也幾乎是傳播研究的知識核心（hard core）之一。

然而，隨著媒體環境的鉅變，信息技術、網際網路、公民媒體、人機互動等媒體新領域持續地增加，資料創造、蒐集、傳遞、儲存的功能也日益強大，造成人類傳播社會重大之轉變，無論是傳播的方式、內容或是閱聽都有著劇烈的變化。新的傳播科技，也帶來了新的「公眾」。我們的生活全面進入到網絡環境中，我們檢查電子郵件（e-mail）、致電他人、用「台北等公車」APP 查詢公車班次、上 Facebook 幫友人的生日活動按讚（like），我們幾乎可以宣稱「數據運算無所不在」。從 Google Trends 的關鍵字搜尋趨勢分析圖也可以見到，10 年前對複雜資料處理重視，隨著新媒體科技的突飛猛進，燃起了近年在「大數據」（big data）發展的信號。

一個「運算的社會科學」（computational social science）的時代已然來臨。運算社會科學因資料處理方式的創新而生成，透過大數據與新媒體的輔配，得以重新架構社會研究，傳統社會科學哲學中所著力的個人與集體的辯證，在運算的社

* 世新大學創新傳播與數據智慧實驗室執行長、淡江大學未來學研究所兼任助理教授，Email: makcem_9@hotmail.com。

** 世新大學教學卓越中心博士後研究員、世新大學公共關係暨廣告學系兼任助理教授，Email: cindyc@mail.shu.edu.tw。

會科學中已然展開了各種新的可能性，像是描繪鉅量的傳播模式、測量細緻的趨勢動態、揭露人類行動軌跡等。儼然形成一個跨學科的隱現領域，且亟待新興研究者投入發展。

在這波「運算轉向」（the computational turn）的浪潮中，可以見到社群媒體（social media）佔著關鍵的成分。社群媒體已經成為公共論述與社會溝通的重要環節，越來越多的公眾議題是透過社群媒體來建立起豐沛的交流討論。社交網站（social network site），如臉書（Facebook），有著高度社會參與潛力，在臉書上傳遞地不只是一般與私人資訊，同時諸政府單位、新聞媒體等機構也紛紛運用臉書粉絲專頁與民眾溝通，鼓勵直接對話。透過社群媒體來達到蒐集意見、掃描環境、大資料分析、歸納現象、並且將傳遞訊息視覺化之需求，可以「數據智慧」概念來統攝。在過去被認為是難以處理的跨平台、大量、複雜資訊之社群媒體分析，透過文字探勘（Text Mining）、情感分析（Sentiment Analysis）等資料處理技術，已經可以用一個有別於傳統傳播研究的方式來「描繪公眾」。

「公眾」的概念與意涵，透過現今的大數據分析技術，將可能窺見其全貌。傳統上，傳播學門中的「閱聽眾」研究，乃針對各種媒體的使用者，或稱「受眾」，進行其媒體使用行為之分析。Denis McQuail 曾歸納出以透過不同而彼此重疊的方式來定義的閱聽人概念：藉由「地方」，好比地方性媒介的情況；藉由「人群」，當媒介的特色是要吸引特定的年齡層、性別、政治信仰或收入範圍；藉由「特定媒介或管道形式」，尤其是技術和組織的結合；藉由「訊息的內容」，指涉文類、主題事物、風格；藉由「時間」，例如當我們提及「白天時段」或「主要時段」的閱聽人，或是稍縱即逝的閱聽人。同時，也衍生出多種研究方法，大抵包括量化及質化兩種取徑，或以行為典範與接收分析典範做為區分。

行為取向源於心理學及社會心理學。一般來說，行為取向的目的主要在於個別的人類行為，尤其是若干和傳播訊息之選擇、處理與反應相關的旨趣上；大眾傳播的使用被視為一種具備特定功能的理性、動機性行為，或是為了個人目的而使用。文化取向則發源於人文學、人類學及語言學等，它主要應用在特定社會脈絡與文化經驗的細節上；運用在媒介研究時，對於媒介、媒介生產與接收的環境之間的差異則較為關注，而它對特定的、獨特的個案與情況興致盎然，甚於對「通則化」的興趣。當典範預設科學真理可以跨越時空環境而存在，研究者必須摒棄任何價值判斷、使用純粹中立的語言來描述真實時，作為被研究觀察之客體對象的閱聽人，便被視為個別的、獨立的、原子化的，可以相互加總並運用數學統計法則加以運算。這也構成傳播學領域中量化研究的後設基礎。然而，今日的數位

科技及資料蒐集、分析工具的日新月異，不僅「閱聽眾」定義蛻變，傳統上尋找閱聽人的途徑、研究方法也面臨著挑戰與新契機，我們究竟看到更多或是更少？

故此，在運算轉向下重看「幻影公眾」，有學者認為，大數據消解了公共與私人的界線，其後果是每一個人都成為「赤裸裸的人」，無法建構私己領域，也無法捍衛主體性。正如 Lippmann 所描述的「幻影」（phantom）般，公眾的主體成為隱蔽的運算。但亦有學者對大數據抱持著堅定的熱誠與信心，認為「運算化」（computation）讓社會研究邁向過去無法察覺的面向，其哲學亦隨之汰新，從根本上重新回答了「公眾如何可能」的問題。

本文將針對傳播研究中的運算取徑（computational approach）進行整理，尤其了解其在台灣的發展現況；其次，展開「公眾」在傳播研究中的系譜，並放置於數據智慧下來理解；並且透過手上進行的健康傳播與社群媒體研究，審視社群媒體在數據研究中的方法體現；總結而言，在數據時代的運算轉向中，「描繪公眾」將依然是傳播研究的核心，數據智慧將成為傳播應用與研究中立即且實用工具手段，但同時也是更貼近人性的一種社會學想像。

關鍵字：方法論、公眾、社會學想像、運算取徑、閱聽人

The Visible Public：
Discussion of Computational Audience

Chang-chieh Lee [*], Hui-hsien Chiu [**]

Abstract

The drastic changes of media environment have led to significant changes of human communication, including the vehicle, the content, and the audience of communication. New communication technology brings out a new group of "public", and we can claim data computing is omnipresent, which triggers the research and development of big data. The era of computational social science has arrived where social media plays a crucial part in the computational turn. Social media has become an important link in public discussion and social communication, and more public issues trigger heated debates via social media. Computational social science is born through the renovation of data management, bearing the foundation of social research on big data and new media. Social media analysis of cross-platform, massive and complicated information used to be considered difficult, but through text mining and sentiment analysis, it is now possible to adopt an unconventional method to profile the public.

Keywords: methodology, public, sociological imagination, computational approach, audience

[*] CEO, CIDI, Shih Hsin University; Adjunct Assistant Professor, Graduate Institute of Futures Studies, Tamkang University. Email: makcem_9@hotmail.com.
[**] Post-Doctoral Fellow, Center for Teaching Excellence, Shih Hsin University; Adjunct Assistant Professor, Department of Public Relations and Advertising, Shih Hsin University. Email: cindyc@mail.shu.edu.tw.

新新媒介中歷史事件的再現：
以敘事分析「台灣吧」快評 228 事件為例

金姵君*

摘　要

自 1990 年代對普羅大眾開放，到現今進入 web2.0 時代，網際網路的發展除了帶來更加快速的訊息傳送之外，也導致不同媒介數位的匯流，豐富了使用的經驗。有當代麥克魯漢之稱的保羅·萊文森（Paul Levinson）[1]認為，如今我們更進一步身處在「新新媒介（New New Media）」的時代，使用者間的互動性不僅增強外，其本身亦是生產者，其中影片分享的媒體也包含在內，網友由過去的資訊接收者，變成資訊發布者。這不僅打破了時間的限制，也讓更多元的聲音得以呈現；而其中，針對某一作品或事件所做娛樂的敘述與評論，為近幾年在影分片分享網站很熱門的呈現方式。這種以娛樂為主的的呈現方式目前並沒有統一的稱謂，筆者暫時先以「娛樂的快速評論(簡稱娛樂快評)」來代稱。同樣的，針對歷史事件的陳述方式產生了變化，除了主流媒體(報紙、電視)對歷史事件的敘述之外，也可以從「娛樂快評」中知道年輕一代對於歷史事件的陳述方式。

在台灣，四位年輕人組成的「臺灣吧（Taiwan Bar）」新媒體公司在 2014 年於影片分享網站上開播一系列講述台灣歷史的影片，運用幽默口白，在約 10 分鐘的時間內，回顧台灣的重要歷史，其中 2015 年以 228 事件為主題的影片-「全球瘋傳，臺灣人不告訴你的，228 事件」在台灣引起了正反兩極的爭議。針對同一個歷史事件為何會有如此的差異，其特殊性為何？這樣的敘事方式是讓我們對歷史事件的觀點帶出不同視角的省思，還是更加凝聚對其事件特定敘事的認同？

本研究針對臺灣吧（Taiwan Bar）的此部 288 影片做為文本，以敘事批評影片內容，探索現今新新媒體下，快評的形式如何呈現嚴肅的歷史事件，以及為何被如此表達。

關鍵字：新新媒介、臺灣吧、娛樂快評、228、敘事分析

---

* 世新大學口語傳播學系碩士二年級，Email: stellasusanjules@gmail.com。

[1] 出自 Levinson, P. (2014)。新新媒介(何道寬譯)。上海:復旦大學。

The Resurgence of Historical Event in New New Media：A Narrative Analysis of "Taiwan Bar" Entertainment Quick Review 228 Incident

Pei-jyun Jin*

Abstract

Since the opening to public use in the 1990s till the web2.0 era nowadays, the development of Internet has not only brought about faster information transmission, but also lead to the convergence of different digital media. Also, enriched the user experience. Paulo Levinson, a professor of communication known as “contemporary Marshall McLuhan”, believes that we are living in a time surrounded by “New New Media”. Not only the interactivity among consumers (users) has been strengthened, consumers (users) also can possibly become a producer. Such phenomena includes the media of video sharing, which turns Internet users from information receivers in the past into publishers now. This not only broke the time limit, but also allow more voices to be presented. Among which a form of entertaining narrative and comment developed based on a piece of work or event has gain popularity on video sharing sites in recent years. Such form of entertaining presentation hasn’t got a uniform terminology yet, so I would like to call it as "Entertainment Quick Review". Except for the presentation, entertaining messages of historical events has changed too. In addition to the narratives of historical events in the mainstream media (newspapers and television), it is also possible to know how the younger generation is describing historical events by the Entertainment Quick Review.

In Taiwan, four young men established a new media company called “Taiwan Bar” and launched a series of videos about Taiwan's history on YouTube in 2014. Reviewing important events of Taiwanese history by means of humorous narration in about 10 minutes. One of the videos featuring the notorious 228 event (with a title of “Go viral

* M.S. Student, Department of Speech Communication, Shih Hsin University. Email: stellasusanjules@gmail.com.

globally – The 228 Incident that Taiwanese won't tell you") has caused a great controversy island-wide. Why is there such a big difference for the same historical event? What is its particularity? Will this way of narrative allow us to view historical events from different perspectives of reflection, or help us cohere the specific narrative recognition on certain event?

This project focuses on the 228 video made by Taiwan Bar as the content, criticizes the contents of the video, and explores the historical narration under the influence of New New Media, how such historical narration reflect the particularity of Entertainment Quick Review, how it is narrated and why it is said in such way.

Keywords: new new media, Taiwan Bar, entertainment Kuài píng, the 228 incident, narrative analysis

文學史 Web2.0：臺中文學地景 LocalWiki 塊莖現象的文學史書寫辯證性

解昆樺*

摘　要

「臺中文學地景 LocalWiki」為 Wiki 發展出的 Web2.0 網路平台，在建置上以對應現實之臺中網路地圖進行文學地景詞條插標，其文學地景詞條在地圖上的塊莖分布現象，具體再現了台中文學地景的地方知識架構。此一地方知識文本本身就是一種對傳統地方文學史的辯證，特別是其針對傳統文學史的作者書寫，其更以 Web2.0 的數位共寫性，刺激著傳統文學史書寫進行數位人文的辯證。

文學史涉及文學之時間與空間發展的敘述，然而數位人文對過往文學史的反省，不單只是空間、時間，而是「之間」的課題。在什麼之間，不只是彼此之關係，中文的「之」，更有前往意思，在彼此概念間的前往、往返，使「之間」形成一個意義流動擴張的辯證範疇。

傳統對文學史書寫的反省，主要在對其敘述框架的思考，如何制訂有效地通體適用不同時代文學之敘述框架，以容納經典作家、文本、流派、現象等。而此敘述框架最主要引動的問題主要有二：第一、如何訂立框架的分期標準，特別是不為政治史概念，以及制式化地十年分期法所左右；第二、框架看似「廣博容納」文學史材料，但本身是否在有意識的擇選中，無意識地排除了什麼。

對於上述文學史敘述框架的兩個主要問題，需要新文學史觀念的介入才有能予以有效突破。王德威〈百年來中國文學的鉅變與不變—被壓抑的現代性〉便曾以「城／鄉座標」，並與文學「國／家」與「群／我」創作相辯證，呈現有別過往，特別是適合現代文學史敘述的文學場域、文字形式與文化生產概念。

在理論研究之外，王德威所指「城／鄉座標」如何落實？正由一九九〇年代臺灣區域文學史書寫，到二〇一〇年代「臺中文學地景 LocalWiki」數位地圖插標

* 國立中興大學中文系副教授，Email: fung682002@gmail.com。

的發展所呈顯。

文學史，顧名思義乃是文學的歷史，但不可能無所邊際通論，是以往往以一國家作為空間範疇。國家／族文學史在十九世紀歐洲大量發展，在浪漫主義與民族主義的驅動下，試圖透過對國／民族文學史的編寫，完成文學共同體─亦即對等國家層次，一個大寫、共同，可自然地讓國民接受這就是「我們」的精神體─之建構。由此，如何在敘述上統合詮釋國土範疇中，居住於各區域之作家文本，辯證「個／群體作家」與「城／鄉區域」之間的精神風格變成為重要課題。辯證意味兩個層次間存在著的異同，在臺灣一九九０年代起各縣市區域文學史編寫推出後，將各區域文學史並列通看，此一辯證更顯迫切。

許俊雅〈建構與新變／敞開與遮蔽－台灣區域文學史的意義與省思〉便指出：「這些都淵源於現下文學史的編撰先天上就有許多的困難，不僅僅是史料、史觀、史識的問題，其操作本身很難面面俱到，顧了文類發展線索，對於多種文類兼具的作家就難安放其位置；以作家作品為主的敘述，便欠缺時間歷史感；以主題或區域為導向的書寫，便切斷作家文學的全貌。」可以發現區域文學史與國家文學史所既存之問題相交集，更因為作家本身出生、定居而有該歸屬何區域的流動性，與以行政區作為區界方式是否得當等問題。具新觀點的區域文學史若限縮於傳統書寫方式，其對時空間敘述之調配，並不能超出過往模式。就此而觀，「臺中文學地景 LocalWiki」聚焦於「文學地景」，本身即鎖定在「作家文本－地景」關係，克服作家流動問題；但更重要的，還在於其數位人文平台載體特性、運轉性能，對於文學史書寫新意義皺摺的刺激。

文字之書寫工具、載體與出版傳播方式對文學發展之影響，在 2002 年 David Loewenstein、Janel Mueller 主編《劍橋英國早期現代文學史》被彰顯，其關注手抄本流通方式、印刷工具運用之看法，為 2013 年 Wilt L. Idema〈關於中國文學史中物質性的思考〉所延續。Wilt L. Idema 在該文中更進一步指出：「在數字化革命中我們正在經歷一個可以預見文學書寫，傳播和消費方式發生劇變的時代。我們將不僅在網絡上擁有文學，而且這個網路文學將利用我們的電子傳媒獨特的可能性來創造前所未見的文學載體。」說明了有別傳統的數位平台對文學創作甚至文學場域的革命性影響，「臺中文學地景 LocalWiki」數位人文平台的書寫載體特性，也同樣衝擊著區域文學史傳統書寫模式。「臺中文學地景 LocalWiki」的網路平台在文學史書寫上所提供的積極皺摺，不僅止在用電腦螢幕呈顯文字，這是傳統紙媒就能做到的；也不在能不斷伸展拉長頁面能便利地容納很多文字資料；也不在克服傳統文學史因書寫內容膨脹，形成厚重不利流通的印刷書籍，以

網址進行展示傳播。「臺中文學地景 LocalWiki」突破傳統文學史的新書寫潛能，表現在「插標」、「引文」、「共寫」三者。

Wiki 百科乃是以設建詞條的方式，積累知識並形成其體系結構，由此發展出的 LocalWiki 則同時需將詞條，歸屬插標於世界地圖上的一處。「臺中文學地景 LocalWiki」即是創建者規劃出臺中區域後，以此為地圖界面範疇，由使用者在其上進行臺中文學地景的詞條插標。可以發現相較前文所述王德威於一九九０年代末所提出「城／鄉座標」，「臺中文學地景 LocalWiki」更進一步地以「插標」進行實踐。從「座標」到「插標」，意謂著對區域文學史的書寫，將不只是一對城／鄉文學的文字描述，而必須在數位區域地圖上以詞條徵實定標。初步來說，呈現了人文科學更細密的考證研究；更進一步來說，在「臺中文學地景 LocalWiki」上之詞條插標，也具有由點而進行畫線以及區塊的功能，這多層次圖層以及自然形成的塊莖意象，正是一臺中地方知識的具體皺摺，凸顯了文學地景詞條內在涉及的作家生活、地方記憶的網佈強度。

LocalWiki 既為一網路平台，其網路鍊（連）結本身就是一種強度「引文」功能的展現。Gilles Louis René Deleuze 在《游牧思想》「塊莖」中便指出：「一本書的理想就是把一切都展示在這種外部的平面上，展示在同一頁紙上、同一本書中：歷經過的事件，歷史的決定因素，概念，個人，團體，社會構形。」傳統文學史書籍之引文方式僅能透過抄寫、註解方式，使得既成知識概念得以介入文字敘述。但傳統引文受到紙本載體界面限制，同時也必須規避過度引文干擾原本文字敘事脈絡。1995 年離世的 Deleuze 理想的那本能共同展示、帶豐富引文的書本，在現實中 2000 年後高度發展的網路平台以虛擬方式實踐。「臺中文學地景 LocalWiki」的「引文」不只是對「文字」的徵引，更是對「數位多媒體文本」的引文，將知識經驗從文字層次解放出來，使得隨智慧手持、穿戴電腦裝置，更影響人們資訊、知識接收方式的影音文本，亦得以被引文而入。

「臺中文學地景 LocalWiki」的 Web2.0 特性，也使得文學史書寫不再拘泥於單一作者，而得能以地方生活群體共寫的方式去想像與推動。群體使用者可利用電腦甚至穿戴智慧裝置，實際在文學地景現場，即時進行地圖詞條插標共同、即時的閱讀與編寫。如此群體經驗與地方體驗的現場結合，豐富了地方與群體共同感間的鍊結。比其一般紙本著作，LocalWiki 的平台有隨寫隨儲的「活動記錄」，翔實紀錄網頁編寫歷程，也使得在群體共寫文學史上能清晰地呈現群體編寫細節生成與脈絡運轉，呈現群體地方知識生成的意義跡軌。由此，「共寫」稀釋了原本傳統文學史內在的政治意圖，形成了地方經驗資料庫的積累。

2016 年 7 月「臺中文學地景 LocalWiki」已有 156 個插標，在其「插標」、「引文」、「共寫」特性下，詞條據點之聯繫、聚合與圖層形成了塊莖圖景。每一時代都有自身表／追述時空間的方式，若這個表／追述時空間方式未見穩定，正反顯一時代經驗結構及其知識系統尚處未轉型的狀態。我們該如何凝視「臺中文學地景 LocalWiki」在網路、螢幕共構的實虛載體上的塊莖意象，並以語言敘述之，正呈現文學史 Web2.0 的特性－非單一譜／樹系，而是在敘述上容納異質，在任何點任何方向能與其他點相連結；非單純只是蹤跡，而是帶有地圖結構感的塊莖。如此，也正能完成我們交織數位、地方的新區域文學史視野，並真正完成傳統紙媒到數位人文時代知識系統轉型。

關鍵字：LocalWiki、Web2.0、文學地景、文學史書寫、共寫、詞條插標

The History of Literature Web 2.0:The Historical Literary Written Dialectic of the Taichung Literature Landscape's Localwikituberation Phenomenon

Kun-hua Hsieh*

Abstract

"Taichung Literary Landscape on LocalWiki" has been established through the marking of the literary landscape entries on Taichung's network map in correspondence with the reality on the Web 2.0 wiki-based network platform created by DavisWiki. The rhizomatic distribution of the literary landscape entries on the map has specifically represented the local knowledge structure of Taichung literary landscape. This local knowledge text itself is a kind of dialectic of traditional local literature history, especially in the authors' writing of traditional literature history, which even further stimulates the dialectics of digital humanities of the writing of traditional literature history with the digital co-writing feature of Web 2.0.

The history of literature involves the narration of the temporal and spatial development of literature. Nevertheless, the reflection of digital humanities on the past history of literature not only involves times and spaces, but much more an issue of "between". More specifically, it is not just a relationship "between" each other. In Chinese, "between" has an underlying implication of "going and coming between two things", which makes the word "between" seem to form a dialectical scope with a flowing and expanding meaning in it.

Traditionally, the reflection on the writing of literature history mainly focuses on the narrative framework, exploring how to develop a narrative framework that can effectively suit various literatures at different times in order to accommodate the classical writers, texts, genres and phenomena. This narrative framework mainly raises two issues: first, how to set a periodization standard of the framework that would particularly not be affected by the concept of political history and the stereotyped 10-year periodization method; second, the framework seems to "extensively accommodate" all kinds of materials in the history of literature, but it should be examined whether it

* Associate Professor, Department of Chinese Literature, National Chung Hsing University. Email: fung682002@gmail.com.

has excluded anything by an unconscious selection.

These two major issues of the narrative framework of literature history require the intervention of new concepts of literature history to achieve an effective breakthrough. In *Great Changes and Unchangeableness of Chinese Literature in the Last Century—The Repressed Modernity* (百年來中國文學的鉅變與不變—被壓抑的現代性), David Der-wei Wang (王德威) has represented the dialectics by comparing the "urban/rural coordinates" with the literature creation of "state/home" and "group/I", resulting in different literary fields, text forms and the concept of cultural production that are particularly compatible with modern literature narrative.

Beyond theoretical research, how have the "urban/rural coordinates" brought up by David Der-wei Wang been implemented? The answer is: the writing of Taiwan's local literature history in the 1990s can be represented with the marks on the digital map of Taichung Literary Landscape on LocalWiki developed in the 2010s.

To avoid being too indefinite and unbounded, the discussion of the history of literature tends to be confined to a certain spatial scope, mostly based on the definition of state. The history of state/nation literature had been adequately developed in the Europe of the 19th century. Driven by romanticism and nationalism, the common literary community, which is a great common spiritual community that can be accepted by the citizens to represent as "we" at the state level, had been completely constructed through the editing and writing of state/nation literature history. On this account, how to conduct the dialectical analysis of the spiritual styles between "individual/group writers" and "urban/rural areas" based on the texts of writers living in various areas within a specifically defined and united state territory has become an important issue. The term "dialectical" refers to the similarities and differences existing between two levels. Since the 1990s, the compilation of regional literature history in each county and city in Taiwan has regarded the regional literature history of each area as a whole, which makes the dialectical analysis even more urgent.

In *Construction and Reformation, Openness and Closeness: Reflections on the Meaning of the History of Taiwan Regional*, (建構與新變／敞開與遮蔽－台灣區域文學史的意義與省思), Chun-ya Hsu (許俊雅) has pointed out that: "The compilation of the history of modern literature has its own difficulties. Apart from the issues of historical materials, historical perspectives and historical knowledge, the operation itself is difficult to attend to each and every aspect of a matter—when focusing on the development of literary genres, it would be difficult to place writers with more than one genre; when focusing on the narration of individual writers' works, it would be difficult to possess a sense of time and history; when focusing on the theme- or regional-oriented writing, it would be difficult to see the whole picture of writers' works." Therefore, it

can be seen that in addition to the existing issues between the history of regional and national literature, there are other issues, such as how to determine the fluid belonging of writers who have different regions of birth and residence, as well as whether the zoning method based on administrative district boundaries is proper. If the history of regional literature with new ideas is limited to the traditional way of writing, its deployment of temporal and spatial narratives would be not able to exceed the previous model. In this context, since the focus of Taichung Literary Landscape on LocalWiki is on the literary landscape, the issue of writers' fluid belonging can be well addressed within the writer' text-landscape relationship. More importantly, yet, the carrier properties and operating performance of digital humanities platform will trigger a significant stimulation to the writing of folds with new meaning of literature history.

The impacts of writing tools, carriers and the ways of publication and dissemination of words on the development of literature have been highlighted in *The Cambridge History of Early Modern English Literature* edited by David Loewenstein and Janel Mueller (2002). The opinions on the circulation modes of manuscripts and the application of printing tools in this book have been continued by Wilt L. Idema in *Material Technology and the Periodization of Chinese Literary History* (2013). In this article, Wilt L. Idema further stated: "In the digital revolution, we are experiencing an era with predictable upheavals in literary writing, dissemination and consumption. We will not only have literature on the network, but such kind of network literature will even exploit the unique possibilities of our electronic media to create an unprecedented literary carrier." This statement has clearly identified the revolutionary impact of unconventional digital platforms on both literature creation and literary fields.

Similarly, the characteristics of digital humanities platform of Taichung Literary Landscape on LocalWiki as the writing carrier also have an impact on the tradition writing mode of the history of regional literature. The active folds provided by the network platform Taichung Literary Landscape on LocalWiki for writing the history of literature do not lie on the presentation of the text on the computer screen, which can be achieved by traditional printed media, the unlimited text information that can be easily contained in single page with a constant extension, or the convenient display and dissemination of text information through the websites which address the problem of thick and heavy printed books of traditional literature history that are unfavorable for circulation. In fact, the new writing potential of Taichung Literary Landscape on LocalWiki that breaks through the traditional literary history lies on the following three aspects: marking, citing and co-writing.

The systematic structure of Wikipedia is formed by the knowledge base accumulated with a number of entries. As for LocalWiki, which is developed from

Wikipedia, both the establishment of entry and the marking of location in the world map have to be done simultaneously. Taichung Literary Landscape on LocalWiki is built by the creator who first laid out the area of Taichung City, and the users later mark the entries of Taichung literary landscape within the scope of the map. Compared to the previously mentioned "urban/rural coordinates" proposed by David Der-wei Wang at the end of the 1990s, it can be found that Taichung Literary Landscape on LocalWiki has further implemented the "marking" practice. The transition from "coordinates" to "marking" means that the writing of regional literature history is no longer the text description of urban/rural literature but the marking of entries on the digital map.

Preliminary, this mode presents a more detailed exploration and textual research on the humanities; further, the marking of the entries on Taichung Literary Landscape on LocalWiki also has a function of drawing lines and blocks instead of only dots. The multiple layers and the naturally formed rhizomatic image are exactly the specific folds of Taichung's local knowledge, highlighting the writers' life involved in the literary landscape entries and the strength of the distribution of local memories.

As a network platform, LocalWiki's network links themselves are the representation of citation function at strength levels. Gilles Louis René Deleuze explained in the part of *Rhizoma* in *Nomadic Thought*: "The ideal of a book is to show everything on the external plane, *i.e.* the same page in the same book, including the events that pass through, as well as the historical determinants, concepts, individuals, groups and social configurations. In traditional literature history books, the citation can only be done with transcription and annotation, giving the opportunity for existing knowledge concepts to intervene in the text description. However, limited to the size of paper carrier, the traditional citation method also has to avoid excessive citations that tend to intervene in the original text narrative context.

Deleuze passed away in 1995; nevertheless, his perfectly ideal book that could display all of the information with abundant citations has been finally presented on the highly developed network platform in the virtual way in 2000. The citations on Taichung Literary Landscape on LocalWiki are no longer restricted to the textual format but can be presented as digital multimedia texts, freeing the knowledge and experience from the text level. Moreover, along with the popularity of smart handheld, wearable computing devices, the video texts that have a significant influence on the information and knowledge receiving modes can be also included in the citations.

The Web 2.0 feature of Taichung Literary Landscape on LocalWiki also makes the writing of literature history no longer rigidly adhere to a single author but can be promoted and imaged through the co-writing of a group actually or virtually living in the same local area. The users in the group can use the computer or even wearable smart

devices to conduct the instant marking of entries on the map and the instant reading and writing of the entries at the actual literary landscape scene. In this way, the combination of group experience and local experience can enrich the links between the local area and the group's sense of community. Compared to general printed publications, LocalWiki has a platform with an activity log can be written and saved at anytime, allowing users to fully and accurately record the compilation history of the webpage. Moreover, it also provides a clear vision of the generation of group writing details and the context operation in the group's co-writing of the literature history, presenting the significant track of the generation of a group's local knowledge. Thus, such a co-writing approach can dilute the inherent political intentions in the original traditional literature history, and further form the accumulation of local experience and knowledge base.

In July 2016, a total of 156 entries have been marked on Taichung Literary Landscape on LocalWiki. With its characteristics of marking, citing and co-writing, the connection, aggregation and layers of each entry mark have formed a rhizomatic picture. Each era has its own expression and retrospection approaches; if these expression and retrospection approaches are unstable, which means the experience structure of this era and its knowledge system are still in a state of non-transition. The way that we gaze at the rhizomatic image of Taichung Literary Landscape on LocalWiki on the actual and virtual carriers commonly constructed both by the network and the screen and make a further description in languages is exactly the evidence that the features of Web 2.0-based literature history—it is not a single pedigree or family tree, but a description that is able to accommodate heterogeneous media and link with other dots in any direction; it is not a simple track but a rhizoma with a sense of map structure. Therefore, the new perspective on regional literature history interweaving with digital and local content can be fully established, and the transition from traditional printed media to the digital humanities knowledge system at modern times can be truly completed.

Keywords: LOCALWIKI, web2.0, literary landscape, the writing of literature history, co-writing, term construct

數位人文跨國合作專案之省思：
以日本四國文史數位學習系統開發為例

黃國鴻*、林妍彣**、莊盛宇***、湯茹婷****、涂智鈞*****

摘　要

本專題是以日本四國為主題的數位學習系統，藉由和香川大學數位學習中心合作，希望能製作出讓使用者能夠更深入了解到四國地區地理歷史的數位教材，過程中包含實地到四國參訪，邀請當地的學科專家並蒐集線上圖書館的資料，設計並開發教材，最後對大學生及一般社會民眾共 106 位實施滿意度調查，其中抽取 30 位做教材成效調查測試使用教材後的進步狀況，經過一年的製作過程中，總結最大的疑慮是自身的能力不足以處理相關資料的正確性問題，常常需要詢問學科專家或指導老師。專題學生成長方面，包含對於日本有更深的認識、培養團隊合作精神、溝通和問題解決能力上都有所進步。

關鍵字：數位人文、數位學習、四國、專題導向學習

* 國立嘉義大學數位學習設計與管理學系教授，Email: kuohung@mail.ncyu.edu.tw。
** 國立嘉義大學數位學習設計與管理學系暨研究所研究生，Email: serena8329@gmail.com。
*** 國立嘉義大學數位學習設計與管理學系暨研究所研究生，Email: nxikking@gmail.com。
**** 日月光半導體製造股份有限公司人力資源服務處管理師，Email: ruby8478@gmail.com。
***** 一甲一光學有限公司內勤，Email: king821013@gmail.com。

Reflection on A Cross-nation Project of Digital Humanities: A Case of Developing the E-learning System on Humanities and History of Shikoku

Kuo-hung Huang[*], Yen-wen Lin[**], Sheng-yu Chuang[***]
Ju-ting Tang[****], Chih-chiun Tu[*****]

Abstract

This article described the process of developing the e-learning system on humanities and history of Shikoku, as well as the reflection on students' learning experience. This one-year project, a collaborative work with Kagawa University, aimed to design a web site introducing Shikoku in Chinese language. The activities consisted of field trips to Shikoku, data collection through online resource and library, communication with a Kagawa University professor who served as the subject matter expert, design content, implementation, and evaluation. In spite of encountering difficulties such as insufficient information and language gap, the team members worked together to complete this project on time. More importantly, students enhanced their understanding about Japanese history, as well as improved their team spirit, communication and problem-solving skills through this project.

Keywords: digital humanities, e-learning, Shi-koku, project-based learning

[*] Professor, Department of E-learning Design and Management, National Chiayi University. Email: kuohung@mail.ncyu.edu.tw.
[**] Graduate Student, Department of E-learning Design and Management, National Chiayi University. Email: serena8329@gmail.com.
[***] Graduate Student, Department of E-learning Design and Management, National Chiayi University. Email: nxikking@gmail.com.
[****] Manager, Department of Human Resource Service of Advanced Semiconductor Engineering, Inc. Email: ruby8478@gmail.com.
[*****] Staff, 一甲一光學有限公司. Email: king821013@gmail.com.

一、背景

日本的四國地區相較於日本名勝地區而言是屬於偏鄉地帶。卻也因為環境未遭受戰爭及觀光產業的重大影響，保存了不少文化遺產，人文氣息豐富。日本香川大學與四國的其他三所國立大學在政府的經費補助下，共同開發出「四國學」的數位課程，教導學生與四國地區相關的主題。本研究是透過大學畢業專題，由學生與香川大學數位學習中心合作，開發中文版本的四國主題的數位學習系統，並實際使用評估其成效。本文則從開發團隊的立場描述與省思此一跨國專題合作之過程，以提供給未來數位人文合作之參考。

二、 主題

四國地區由香川、愛媛、高知、德島四個縣所組成的，隔著瀨戶內海與本國相對應。由於該地區具備海上戰略優勢，在歷史事件上也扮演著重要的角色。近年來國人赴日旅遊人數越來越多，惟四國仍然是國人較為陌生的地區。為與觀光旅行資訊作區別，本專案內容以四國地方的地理歷史作為主要核心，由此發展為主幹，向外延伸為分支介紹人文特色習慣，深刻了解四國背後的歷史文化。有鑑於中文資料中對於四國地區的瞭解僅止於表面一般的旅遊資訊或粗淺的歷史資料，因此期望製作出能讓使用者能更深入了解到四國地區地理歷史的數位學習教材。

本專案的進行為期一年，工作任務包括至日本實地探訪收集資料、邀請香川大學四國學計畫主持人林敏浩教授擔任學科專家、線上及圖書館收集資料、設計開發教材、內容審查及設計滿意度調查、最後為使用者成效評估。

三、教材內容

本數位學習的範圍規劃為四大部分的課程，以四國地區地理歷史當主軸貫穿，並以此作為延伸出其他的學習課程。

(一) 地理歷史

以四國地區—香川、愛媛、高知、德島的地理歷史為主軸概念。而時間軸事件則大致上從飛鳥時期到近代平成時期(如表 1)。在地理歷史的動畫製作上，會以時間軸串起整個四國歷史，分別有歷史事件的發生與結束、歷史人物的誕生與足跡、歷史古蹟的建造與毀滅呈現在整張四國地圖之中。

表 1 四國地區從飛鳥時期到近代平成時期之中日對照西元年表

<table>
<tr><th>日本</th><th>中國</th></tr>
<tr><td>飛鳥時代（593-710）
奈良時代（710-794）</td><td rowspan="3">唐朝（618-907）</td></tr>
<tr><td rowspan="2">平安時代（794-1185）</td></tr>
<tr><td>五代（907-960）</td></tr>
<tr><td rowspan="2">鎌倉時代（1185-1333）</td><td>宋代（960-1279）</td></tr>
<tr><td rowspan="2">元代（1271-1368）
（蒙古人統治）</td></tr>
<tr><td rowspan="2">室町時代（1338-1573）</td></tr>
<tr><td rowspan="2">明代（1368-1644）</td></tr>
<tr><td>安土桃山時代（1568-1600）</td></tr>
<tr><td rowspan="2">江戶時代（1603-1868）</td><td rowspan="2">清朝（1644-1911）
（滿清統治）</td></tr>
<tr><td>明治時代（1868-1912）</td></tr>
<tr><td>大正時代（1912-1926）</td><td>中華民國（1911-1949）</td></tr>
<tr><td rowspan="2">昭和時代（1926-1989）</td><td rowspan="2">中華人民共和國（1949 年至今）</td></tr>
<tr><td>平成（1989 年至今）</td></tr>
<tr><td></td><td>中華民國（1949 年至今）</td></tr>
</table>

(二)人文景觀

分為四個部分—名城、名人、自然、古蹟，主要講解四國地區的古蹟名城〈史跡、寺廟等〉還有當地特別的景觀以及在四國地區誕生的名人名士跟地理歷史的關聯性。名人動畫的製作，以影響四國乃至日本的歷史人物為主軸，分別是日本現代化推手—坂本龍馬、土佐的能人—長宗我部元親、真言密教第八代祖—弘法大師。

圖 1　歷史人物之動畫點擊入口畫面

(三)風俗藝品：

分為三個部分—風俗祭典、特產、藝術工藝，結合四國地區的風俗習慣以及四國地區最有名的藝術工藝品—阿波本藍染、丸龜扇、香川漆器、砥部燒〈陶瓷〉來做介紹。

圖 2　風俗藝品之內文介紹畫面

(四)悠遊嬉戲遊戲區

綜合上述課程，並結合其中的知識，設計出四款教育互動遊戲，給予使用者進行娛樂學習活動。

1. 互動遊戲（1）激鬥！風雲兒：以 4 選 1 的選擇題方式來答題，答對則玩家可以增加集氣量，集滿五點後點擊並正確答題可以啟動暴擊，也可以透過防禦減低傷害，透過於遊戲內的打怪通關，藉此格鬥方式來增加使用者對內容的理解並讓學習

不無聊，增加主動學習的動機。

2. **互動遊戲（2）戰功報賞：**使用者需要看著遊戲出現的任務道具來判別該物品產出自哪裡，找出那些地方特色，藉此機會讓使用者了解四國地區不同鄉鎮不同的文化。

3. **互動遊戲（3）進擊的忍者：**遊戲是以電流急急棒的方式，使用者操控忍者根據題目，找出敵軍密書，也就是答案，透過另類的方式把一般的問答題形式改頭換面，增加趣味性。

4. **互動遊戲（4）跑吧！龍馬：**以坂本龍馬為主角的遊戲，主角後方會有一士兵追殺龍馬，使用者須透過鍵盤獲得增益物件，增加其分數及跑速，反之若獲得減益物件，則會使士兵追擊距離縮短，其中龍馬也能搭配其武器，破壞障礙物來通關。

圖 3　遊戲之點擊入口畫面

四、困難與解決之道

在專案開發的過程中，所遇到的困難可以統整成以下幾項：

(一)歷史內容資料量不足

四國紀事的一大主軸就是地理歷史，在這方面希望能把四國地區所發生過大大小小歷史事件，都包含到教材之中。後來發現日文資料太多且未必是國人有興趣的內容；而台灣方面的資料中，比較少去提到有關於四國的東西。因此，導致搜尋資料的過程中，無法取得足夠的歷史資料。所幸透過日本香川大學方面給我們提供了

許多細節的資料，教材內容豐富度才能圓滿充實。

(二)時程規劃不完善

前期的規劃中資料的收集整理超前了進度，但是到了後期的設計開發進度才發現時程其實很迫切。在最後又加上人物動畫增加工作量之後，明顯超出範圍，必須要在最後的階段日夜趕工，才能趕上規畫的進度。

(三)技術上的困難

有時候遇到一些依現階段無法處理的程式問題，需要求助專家或找尋相關知識，通常改變表現方式或轉個方向來設計程式，就能使專案順利進行。

五、成效評估

數位學習系統開發完成後，進行使用者的滿意度評估及學習成效評估。本專案的施測對象設定為一般學生及社會人士，學生及社會人士方面的先備知識希望是能知道日本這個國家，還有對於日本歷史至少有基本的了解。而經調查施測後，總共有 104 名參與測試的對象。他們適用過教材之後，對於內容及視覺呈現提出建議，並由開發者據以修改。修改完之後進行使用者學習成效評估。參與的施測對象為有填寫過上一份的滿意度以及使用狀況問卷的 104 人之中，抽出 30 名來做教材成效調查。教材成效測驗共 10 題，每題 10 分，總分為 100 分，目的在觀察前、後測的成績變化。受試者先接受前測，再使用四國紀事數位教材，之後接受後測。以兩次施測的成績來做對比，來測試課程教材，是否能夠使學習者透過使用這套數位教材之後，成績會有所進步。t 檢定的結果顯示前後測並無顯著差異。

六、 專題學生的成長

(一)對於日本人文的深入了解

在教材內容的選擇部分，雖然組員個人的主觀意識會影響所要傳授的知識面，但是在討論後會以客觀描述，輔以台灣、日本、西元對照年表來相互映證。此外，學科專家為由香川大學的教授，來檢核我們所找尋的資料正確性，及提供更加細部的四國歷史的內容。而指導教授會抓準專題的大方向，適時調整專題內容。小組之間定期開會檢討彼此進度，並依照教材內容的深度及廣度來看要不要對內容有所增減。這樣的實作歷程幫助所有成員對於四國的歷史人文都有更深刻的了解。

(二)溝通與解決問題能力的提升

小組之間的主要溝通管道有三個階層：

首先，透過定期的專案小組會議，當面講出每個人所負責的進度如何，避免因在網路上間接傳遞所造成的資訊缺失或語意誤解，接著再根據大家的進度，決定下次的進度以及要開會的日期。把握定期定量的方式，持續追蹤彼此組員間的聯繫，來制定出下次方案執行內容，就不會發生有人的進度特別落後的情況。

其次是一旦遇到內容上面的疑慮是自己的能力不足以找到資料或者是對於正確性有所疑問的話，就會寄電子郵件給學科專家，請他在日本方面來協助提供給我們正確且詳細的歷史內容，並且傳達要改進的地方。

最後就是當對重大方向不確定時，就必須與專題指導老師開會，共同完成決定。例如，曾經對於要不要推廣至平板或手機上使用這套課程教材有過爭論，後來指導老師明確指出要在一開始下決定做適合的版型畫面，通用於多個平台的教材效果不佳。因此小組就決定專心致力開發網頁版本。

七、結論

學生為期一年的畢業專題，透過指導老師和日本香川大學合作的一起跨國專案，讓不熟悉日本四國歷史文化的同學來介紹四國地區的歷史文化。這本來就是數位學習專業人員的基本能力：透過自我學習及學科專家的協助，在期限內將學科知識整理開發成數位課程。學生另外遇到的困難就是溝通的問題，因為日文不是很精通，而必須用英文交談與日本的學科專家溝通，在連絡上就必須透過文字上的交流。此外，在專案執行過程中，學生彼此同心協力，培養團隊整體合作的精神，學生自認為整個專題已經精彩的畫下句點。

傳統戲曲藝術口述歷史資料庫建置及擴增實境文創應用

孫劍秋[*]、黃一峰[**]、楊曉菁[***]

摘　要

「國立臺灣戲曲學院」是臺灣唯一的一所十二年一貫制戲曲專業學府，自民國四十六年前校長王振祖先生創立「復興劇校」至今約六十年，目前設有京劇學系、歌仔戲學系、客家戲學系、民俗技藝(特技)學系、戲曲音樂學系及劇場藝術學系等六個教學單位，本校設立的目的即是為了傳承、保存及推廣傳統戲曲藝術，創新發展並與時代接軌，以提升戲曲藝術的價值，誠如本校校訓「承先啟後，精益求精」，使戲曲藝術傳承綿延不絕。本校歷史悠久，整個學校的歷史幾乎成為近代臺灣戲曲藝術發展的縮影，例如本校歌仔戲學系前主任廖瓊枝教授，她獲得無數獎項，由於對傳統戲曲教育的使命感，教學富有愛心及耐心，至今退而不休，仍在本校擔任兼任教師。口述歷史通常是針對當代具有歷史價值的人物，根據其一生與史實有關的部分來設計問題，加以深度訪談並作成紀錄，以作為未來研究的史料。對於本校的退休教師或是在傳統戲曲表演藝術界享有盛名的畢業校友，許多也都陸續凋零，因此為傳統戲曲藝術界建立口述歷史紀錄是件刻不容緩的工作。目前正是本校創校 60 週年前夕，校內規劃發行 60 週年口述歷史專刊，專刊訪談人數 60 人，訪談對像包含對本校發展有卓越貢獻的師長，或是在傳統戲曲專業領域上有傑出成就的畢業校友，這些人選先是由各學系推薦，再經校級編審委員會依其貢獻事跡來核定是否列入訪談名單。訪談前由計畫主持人及採訪教師依受訪對像性質修正訪談的問題；訪談時由採訪教師主持採訪及紀錄，再做成文字稿，隨行並有錄影人員協助錄影及收音工作，後續再由影音後製人員協助將口述歷史錄影片段編製為影片。口述歷史訪談之後，相關資料將進行求證並整理，可提供教學研究使用或是有興趣民眾查詢，因此建立傳統戲曲藝術「口述歷史資料庫」網站。拜現代資訊科技之賜，訪談內容將全程錄影，影片後製修剪並加上

[*] 國立臺灣戲曲學院通識教育中心教授，Email: sun0761@tcpa.edu.tw。
[**] 國立臺灣戲曲學院通識教育中心副教授，Email: ifeng@tcpa.edu.tw。
[***] 國立臺灣戲曲學院通識教育中心助理教授，Email: bennieyy@tcpa.edu.tw。

字幕及背景音樂後，每位受訪影片長度約 30 分鐘。「口述歷史資料庫」網站除了放置口述錄影影音資料之外，也會存放訪談文字紀錄，另外會請受訪者提供歷史照片、影音資料、文宣、演出資訊、劇本及樂譜等，經計畫工作人員整理後，在每份檔案加上後設資料，再放置於網站上，以豐富本網站內容，增加網站的實用性。實體網站將建置在本校「圖資中心」所提供的網站空間中，此網站空間與校網站是放置在同一組伺服器中，但是操作上是獨立於校網站之外，網站功能僅包含文字及圖片的管理與維護，但不包含影音資料。口述歷史訪談之錄影剪輯影音及訪談時所獲得的影音資料等，將放置在本校另行建置的「雲端影音分享平台」中，影音資料上傳本平台後再聯結回「口述歷史資料庫」網站中，此影音分享平台使用 Web 2.0 的管理與分享機制，可開放不同使用單位獨立管理與分享的權限，從使用者與平台互動的角度思維，提供各單位使用者依需要自行增刪網站資料的功能，可以豐富平台內容，並分散系統管理者負擔。由於智慧型行動裝置的使用已經相當普遍，本平台也提供行動裝置瀏覽及播放的功能。重要訪談資料將集結成冊出版為 60 週年口述歷史專刊或是校史室展覽空間中展示，加入擴增實境功能，可使用智慧型手機或是平版電腦掃描出版品或展示板上的 QR Cod 來載入「口述歷史資料庫」網站中的訪談影片或是受訪者提供的相關圖片或影音等，使得出版的書籍或是展示資料的內容更豐富且生動。QR Code 是一種二維條碼，可以儲存網址資訊，使用手機或是平版電腦的 QR Code 掃描軟體，即可以顯示該網頁所呈現的影音或圖片內容，使成為專刊或是展場的一項特色。本文主要是闡述本校如何建置傳統戲曲藝術「口述歷史資料庫」網站，及利用 QR Cod 的功能來達成擴增實境的效能，並應用於後續出版品中，以提升出版品的趣味性及可讀性，增添出版品的動態內容。未來本資料庫網站的訪談對像將擴及到國內重要的傳統戲曲藝師，網站內容也再增加 GIS 定位資訊，並整合過去本校數位典藏的成果，以提供傳統戲曲研究及學習者更豐富的研究資訊內容。

關鍵字：戲曲藝術、口述歷史、擴增實境

Establishing Oral History Database of Traditional Xi Qu Arts and Augmented Reality Applications

Jian-qiu Sun*, I-feng Huang**, Xiao-jing Yang***

Abstract

The National Taiwan College of Performing Arts (NTCPA) is the only 12 years education system of formal institution in the country that fosters talent in “Xi Qu” – the various forms of traditional Chinese folkloric dramas. NTCPA has been established for almost 60 years since 1957 by philanthropist Mr. Wang Zhen-Zu to found the Fu Hsing Dramatic Arts Academy. Up to now, there are 6 academic departments as follows：The Jing Ju, The Acrobatic, The Xi Qu Music, The Ge Zi Xi, The Theatre Arts, and The Hakka Opera. Over the years it has become the cradle of first-class performers, and continues to be the locomotive in the education, preservation, and propagation of these marvelous yet endangered art forms. Inheriting from the Past and Carrying on for the Future; Relentless Strive toward Perfection” is the School Motto of NTCPA, and in fact the ultimate goal of Xi Qu education. Because NTCPA has been established for long time, the history of NTCPA possibly represents that of Xi Qu development in Taiwan. For example, The Professor Qiong-Zhi Liao, who was the previous chairperson in the department of Ge Zi Xi, and she also got several honors from the government. She always recognizes that Xi Qu education is her mission; therefore, she doesn’t retire from NTCPA even though she is over 80 years old. Oral history usually records the people who are valuable of the day. The interviewer will design questions according to the historic fact of the interviewee. After a deep interview, the records will become some research materials in the future. Some retired teachers from NTCPA or some NTCPA alumni who had been famous in the Xi Qu stage are gradually vanished. Therefore, it is urgent to build the Oral History Database in the Xi Qu field. Before the 60 anniversary of NTCPA, an oral history book is going to publish. The amount of interviewee is limited to 60 of outstanding teachers and alumni. Before an interview,

* Professor of General Education Center, National Taiwan College of Performing Arts. Email: sun0761@tcpa.edu.tw.
** Associate Professor of General Education Center, National Taiwan College of Performing Arts. Email: ifeng@tcpa.edu.tw.
*** Assistant Professor of General Education Center, National Taiwan College of Performing Arts. Email: bennieyy@tcpa.edu.tw.

the project host and interviewer will modify the interview questions together according to the situation of interviewee. The interviewer asks the questions and records answers to make oral history report. An interview team should include a video recorder to record video camera and record the voice. After the interview is finished, all records should be verified. For the video files from interview, they should be montaged and then add caption and background music. The length of video files for each interviewee is about 30 minutes. The website of Oral History Database is established to store all files about interview video files, historic pictures, and other related documents. These files should be defined their own metadata, and then uploaded to the website to enrich its contents. The entity website is supported by the Library and Information Center of NTCPA. The website is located in the same space as NTCPA official website; anyhow, its operation is independent from there. The function of the website includes management and maintain of text files and picture files, but the video files are not included. The video files from oral history interview will be uploaded to the "Platform of Audio and Video Sharing in Cloud", and then link to the website of Oral History Database. The platform is applied Web 2.0 technology for each user to manage and to share. The platform supports users to increase or delete their own files. This function enriches the content for the platform and distributes the loads from system manager. Due to the intelligent mobile devices are very common now, the platform also support the function of browse and play for the mobile devices. The published book of oral history for NTCPA 60 anniversary or exhibition space in the School History Room are added the function of augmented reality. People use intelligent mobile device with QR Code scanner to load interview films or some related records from Oral History Database. This feature abounds the published book or exhibition room. QR Code is one of two dimensional barcode which is capable to store URL information. When people use QR Code App in their intelligent mobile device, the device will show the film or picture of the webpage from the URL. It will become the feature of the book or exhibition space. The purpose of this paper is to describe how NTCPA builds the website of Traditional Xi Qu Art Oral History Database. We use the function of QR Code to work as augmented reality and apply in the published book to increase interesting, readability, and vividness. For the future studies, the interviewee for the Oral History Database will be expanded to domestic Xi Qu artists. The GIS information will be increased to the Oral History Database. Furthermore, we will integrate previous digital archive website to support more abundant resource for Xi Qu researchers.

Keywords: Xi Qu Arts, Oral History, Augmented Reality