Facing
the Era of AI+DH

第九屆

數位典藏與2018 DADH

9th International Conference of Digital Archives and Digital Humanities

數位人文國際研討會

論文集

PROCEEDINGS

12/18(W2)-21(W5)

法鼓文理學院 大講堂GC001

Auditorium GC001,
Dharma Drum Institute of Liberal Arts (DILA)
新北市金山區西湖里法鼓路700號

Organizer 主辦單位

臺灣數位人文學會 TADH Taiwanese Association for Digital Humanities

法鼓文理學院 Dharma Drum Institute of Liberal Arts

Sponsors 贊助單位

MOST 科技部 Ministry of Science and Technology

國立政治大學 NATIONAL CHENGCHI UNIVERSITY

Contents 目次

Sessions 論文發表

Session 1: DH Infrastructure 數位人文基礎建設

SP11 DocuSky 與文本字詞關聯圖的視覺化應用
DocuSky and its Applications to the Visualization of Text-Term Relationship Graph…… 2

SP12 Research Infrastructure for the Study of Eurasia (RISE): Towards a flexible and distributed digital infrastructure for resource access via standardized APIs and metadata…… 21

SP13 開放工具、數位人文與圖書館的再思考 …… 38

SP14 以國際圖像互通架構為方法的佛教石窟與圖像之數位呈現與閱覽
The IIIF Approach to Digital Representation and View of Buddhist Caves and Images…… 55

Session 2: DH Research on Buddhist Literature 佛教文獻與數位人文研究

SP22 Tibetan-Chinese-Sanskrit Text Alignment using Intelligent Agents and Genetic Algorithms…… 76

SP23 自動標點的原理與實現
Principle and implementation of automatic punctuation…… 99

SP24 清文全藏資料庫工程之意義與規劃——以人工智慧技術為依託
Digitizaiton of the Manchu Buddhist Canon: Its significance and overall plan with AI as critical support…… 109

Session 3: Network Analysis 網絡分析

SP31 自動分群應用於傳記人物關係建立
The Application of Automatic Clustering in the Construction of Relationship between Historical People …… 125

SP32 第 18 屆到第 29 屆台灣金曲獎最佳國語男女歌手提名人創作角色分析
Creative Role Analysis of the Best Male/Female Mandarin Singer Nominees in Taiwan Golden Melody Awards (2007-2018) …… 136

SP33 數位人文科學語彙之生成與使用…… 147

SP34 古籍數位人文研究平台之史料人物關係圖工具發展與應用
Development and Application of an Ancient Books Digital Humanities Research Platform with Characters' Relationship Map Tool…… 173

Contents 目次

Session 4: GIS 地理資訊系統

SP41 大中華地區歷史地名系統之整合與開發
Integration and development of historical gazetteer system in Greater China 204

SP42 時空視角：文本挖掘〈李娃傳〉中唐長安城的空間感知
Time-Space: Urban Space Cognition of Chang'an in Tang Dynasty Through Text Mining of "The Tale of Li Wa" 221

SP43 歷史地理信息系統與空間人文研究——構建城市歷史地理學研究的時空 GIS 基礎框架
Historical GIS and Spatial Humanities —— Constructing a Spatio-temporal Infrastructure Architecture for Urban Historical Geographical Studies..... 253

SP44 兩宋鎮墓文物的地域分佈與變化：「遼宋金墓葬資料庫」的建置與應用… 270

Session 5: Ontology and Concept Maps 本體論與知識圖譜

SP51 基於本體的家譜数据可視化構建研究----以浙江"仙居高遷《吳氏西宅宗譜》"為例 270
Study on Visualization of Genealogies Based on Ontology----Deployed on the Genealogies of Wu's in Gaoqian,Zhejiang..... 287

SP52 Integration of a Chinese character ontology and Historical Glyph Examples..... 304

SP53 人文學基本 LOD 試論--以韓臺《職員錄》為例..... 318

SP54 近現代中國「新學」概念與新知識系譜的數位人文研究
A Digital Humanities Study on the Concept of "Xin Xue(新學)" and the Genealogy of Knowledge in Early Modern China 333

Session 6: Poetry 詩學研究

SP61 清人「拗救」說再審視——以《全唐詩》15290 首律詩爲樣本..... 347

SP62 現代詩的量化研究：發掘顧城詩的隱藏節奏
A Quantitative Research of the contemporary Chinese Poetry: the Discovery of the Underlying Rhythm in GU Cheng's Poems..... 359

SP64 「何處」的隱喻與轉喻——唐詩「空間詩學」的數位人文研究
The Metaphor and Metonymy of "He Chu" （何處）--A Digital Humanities Study on the " Poetics of Space" in Tang Poetry..... 379

Session 7: Text Analysis 文字分析

SP71 中日古辭書自動文本標注顯示工具 tagzuke

Contents 目次

Tagzuke: An Automated Markup Tool for Medieval China dictionaries and Early Japanese Dictionaries 384

SP72 傳記文本之資訊擷取——以續修台北市志人物志為例
NLP Methods for Information Extraction from Biographies: An Exploration with the Elite Biographies in the Extended Taipei Gazetteers..... 395

SP73 *A Yorkshire Tragedy:* Using Agent-Based Modeling to Suggest Authorship... 406

SP74 關鍵詞偵測方法的比較與應用..... 428

SP75 Science Fiction and Hyperchaos:Digital Humanities as Extro-Criticism..... 463

Session 8: Evolution and Edification in Digital Culture 數位文化之演進與培育

SP81 試析博物館數位藝術史的圖像需求─兼論國立故宮博物院圖像生產與利用方式對數位藝術史之影響..... 497

SP82 非物質文化遺產“雕版印刷”技藝的數位化保護研究..... 518

SP83 但丁數位地圖---現有數位專案及一項進展中的項目
A digital map for Dante -- current online projects and a work in progress..... 534

SP84 倫敦大學國王學院 30 年數字人文研究熱點分析..... 542

SP85 數字人文的本科教育實踐：成績與反思
How to Teach Digital Humanities: Achievements and Reflection 560

Panels 專題討論

Panel 2: 儒家經典注疏與數位人文研究
Confucian Classics and Their Commentaries: A Digital Humanities Approach 572

Panel 3: 墓葬風俗之流變：琉球群島台灣人、台港穆斯林與東南亞華人於墓葬、碑文銘刻、宗教及民間工藝等風俗的實踐與比對
Funerary Practices of Taiwanese in the Ryukyu Islands, Muslims in Taiwan and Hong Kong, and Chinese in Southeast Asia: Variations and Invariances of Tombs Epigraphic, Religious and Artistic Craftsmanship 622

Panel 4: 大數據與近現代韓國社會文化的宏觀趨勢
Big Data and Macro Trends of Society and Culture in Modern Korea 655

Panel 5: 中國古典文本探勘技術（自動標記、中文斷詞、事件擷取）
Text Mining in Traditional Chinese Text(Automatic Markup, Chinese Text Segmentation, and Events Extraction) 722

Contents 目次

Parallel Workshop 平行會議

WEDHIA Session 2:

WE21 數位人文跨域共授課程之群組討論行為模型分析
Analysis of Behavioral Patterns on Group Discussion of Digital Humanities Interdisciplinary Co-teaching Courses..... 752

WE22 Digital Political Science Learning Strategies 759

Posters 海報展示

PT01 Research on the Key Components of Construction of Historical Village Archives Digital Repository 768

PT05 人物檔案資源數據化研究——以上海交通大學錢學森圖書館館藏為例
Research on Arrangement & Datamation for Personal Archives and IT Application——A Case Study about SJTU QLM 776

PT06 《1960》：在衝突的年代討論自由與生命的意義
1960: Investigate the meaning of freedom and life in a conflicting era..... 789

PT07 以何「半部」《論語》治天下？——基於文檔向量相似性的論語篇章結構分析..... 817

PT08 華語文領域學者資料定義與資料使用的詮釋探討
Data in the Humanities: Humanists' Perception and Usage of Research Data in the Field of Chinese Language 824

PT09 Wikidata, a Low-tech Solution to Leverage Semantic Technologies? 830

PT11 "一花開五葉"——禪宗傳承視覺化平臺構建
A Visualization of Chinese Zen Lineage..... 834

PT14 基於文本挖掘的佛經人物畫像研究..... 842

PT15 聖經經節引用的擷取與應用
Extraction and Application of Bible Verse Citations 849

PT19 「曹錕賄選」與知識份子群體的政治選擇：以胡適日記（1923）為中心的觀察..... 859

Facing
the Era of AI + DH

第九屆 數位典藏與 2018 DADH

9th International Conference of Digital Archives and Digital Humanities

數位人文國際研討會

論文發表 SESSIONS



DocuSky 與文本詞彙關聯圖的視覺化應用

杜協昌

DocuSky 與文本字詞關聯圖的視覺化應用

杜協昌 *

摘要

數位人文研究的一項挑戰，是利用資訊科技，讓文史研究者能夠以宏觀的視野檢視並探勘大量的文本。傳統上，文史研究者必須透過精讀與略讀的方式，在掃描與審視文本內容的過程中反覆思考，從而形成研究者的深刻見解。當文本的數量增加，人力閱讀的成本就急速上升。因此，實務上研究者並無法利用精讀或略讀的方式對大量文本進行分析。他們必須藉助資訊科技，才能從大量文本中擷取感興趣的概念或關鍵字詞，並藉由擷取後的結果對文本進行分析與觀察。

本文將簡介一項利用資訊科技分析大量文本的方法：利用 DocuSky 平台分析文本，將結果繪製成視覺化的圖形，然後藉助這視覺化的呈現來幫助研究者檢視與理解這份分析結果。我們將說明如何利用 DocuSky 建構使用者個人的文字資料庫，並且利用這個平台所提供的 StatsTool 對資料庫進行字詞統計分析。我們將定義何謂文本字詞關聯圖，並利用史丹福大學所開放的 Palladio 工具，將 StatsTool 所產生的統計數據繪製成視覺化的關聯圖。我們也將舉出數個實例，說明文本字詞關聯圖在數位人文領域的可能應用。最後，我們期待可以開發出更多更好的工具，讓研究者可以對文本數據進行更深入的分析與觀察，從而推進數位人文領域的發展。

關鍵字詞：DocuSky、數位人文研究平台、詞頻統計、文本字詞關聯圖、視覺化。

* 國立臺灣大學資訊工程系博士後研究員。

DocuSky and its Applications to the Visualization of Text-Term Relationship Graph

One challenge in the research of digital humanities is to help humanists explore a large amount of texts with computer technology. Due to the limits of human power, it is impractical or even impossible to ask researchers to read texts intensively or even in a skimming way. One approach is to have computers extract keywords from text so that researchers can explore text properties from the analytic result.

In this paper, we introduce the DocuSky platform to help humanists build their own databases. We introduce a notion of text-term relationship graph (TTRG) to represent the relationship between texts and keywords, and adopt online tools to get the visualization of a TTRG.. We use several concrete examples to illustrate how a TTRG can help one explore properties in the text.

Keywords: DocuSky, collaboration platform, digital humanities, term statistics, text-term relationship graph, visualization.

一、導論

以文本為基礎素材的人文研究，傳統上必須透過人力以略讀或精讀的方式逐一瀏覽文本，由文字中提取精鍊後的內涵，並對內涵的解讀進行延伸應用。當文本數量逐漸增加，以單純人力進行文本研究的成本就會快速提高，終至無法負荷。在數位人文的時代，我們希望能夠藉由資訊科技的佐助，讓研究者能夠以相對省力的方式分析文本的特徵與脈絡。其中一種方式，是假設特定關鍵字詞代表了文本中的某類概念，而我們可藉由分析關鍵字詞在文本的出現情況，推論文本與這些概念之間的關聯。

具體的實踐方式，是在每篇文件[1]中將關鍵字詞逐一標記，最後就可以利用這些標記來統計關鍵字詞出現在各篇文件的頻率。雖然人們可透過 MARKUS[2] 之類的工具來加速標記工作的進行，但一般來說這項標記工作的成本仍然很高。一種替代的方式，是先將欲標記的關鍵字詞完整列出，透過數位工具去計算這些字詞在每篇文件的出現次數，然後直接將這樣的統計視為（關鍵字詞所代表的）概念出現頻率。這種替代方式的計算幾乎不涉及人力，因此可用電腦快速取得統計結果。必須留意的是，由於這種替代方式的工具，一般並不具備斷詞或一詞多義的判斷能力[3]，因此它的計算結果和正確的標記統計之間很可能存有誤差。經驗上，若關鍵字詞為專有名詞，且字數多於一字，那麼這項誤差經常可以被容忍，或者在事後透過人力對其進行必要的修正。本文在以下幾節所使用的實例，都是利用這種替代方式來進行的。

一旦有了關鍵字詞出現在每篇文件的頻率，我們就可以分析哪些關鍵字詞出現在哪些文本（文件的集合）、又有哪些關鍵字詞沒有出現在另一些文本。這些資訊通常被表示為一種表格結構。當文本和關鍵字詞的數量增加，要求人們從這樣的表格結構中看出文本和關鍵詞之間的關聯性，就越發顯得困難。此時，若我們能用視覺化的方式，將這樣的表格結構轉化成適當的圖形來呈現，在許多狀況下就可從中看出表格結構所不易表達的關聯性。

這篇論文的主旨，就是以具體的系統和可操作實例，闡釋以上的概念、流程

[1] 我們將文件 (document) 視為一個由文字內容所構成的單元。文件在實質上是一個字串 (string) 的資料型態，我們可以將其想像為硬碟中某個文字檔 (text file) 的內容。在本文中，我們將文本 (texts) 視為文件的集合。它可以僅包含一篇文件，也可以是多篇具有相同屬性文件的集合。

[2] MARKUS 是線上的文本標記工具，由荷蘭萊頓大學 (Leiden University) 所開發維護。網址 https://dh.chinese-empires.eu/markus/beta/，

[3] 例如，在字串「那裡有一台大電視機」中，缺乏斷詞能力的工具或程式就會以為這份文本包含「台大」這個關於學校的關鍵詞。此外，字詞「寶玉」在《紅樓夢》中，可能表示珍貴的玉石，也可能指涉主角「賈寶玉」。例如，第一回「上面字跡分明，鐫著通靈寶玉四字...」的「寶玉」並不是人名；但「賈蓉聽說，即同寶玉過會芳園來了」的「寶玉」指的當然是人名。

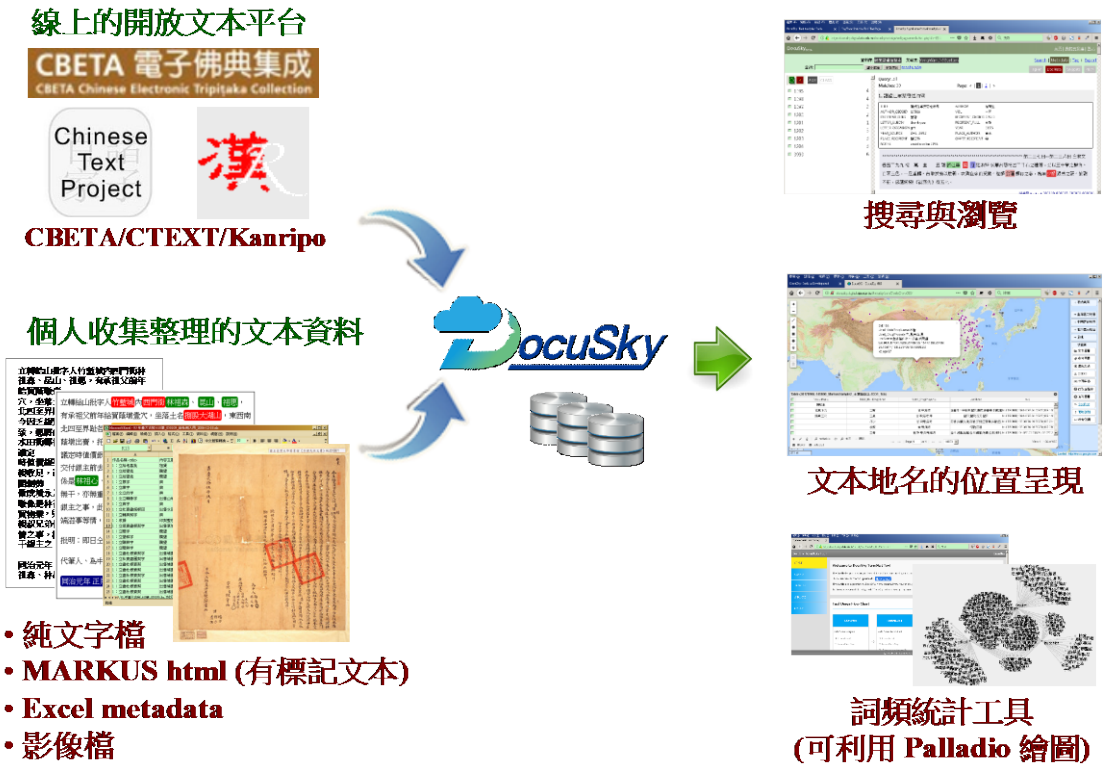


圖 1. DocuSky 功能示意圖。使用者可以從線上的 CBETA、CTEXT 和 Kanripo 直接下載文本來建庫，也可以利用個人收集整理的文本資料（這些資料可以是純文字檔、經 MARKUS 標記後的輸出檔、利用 Excel 所編輯的詮釋資料、也可以包含文件的附圖）來建庫。一旦建庫完成，使用者就可以利用 DocuSky 所提供的搜尋瀏覽工具檢視資料庫內容、透過 DocuGIS 呈現文本地名的地理分佈、也可以利用詞頻統計工具對資料庫進行分析，並將結果匯出進行後續的應用（例如匯入 Palladio 來繪製視覺化的文本字詞關聯圖）。

和後續可能的應用。首先，第二節將介紹 DocuSky[4] 這個平台，它可以讓使用者將自己的文本建構成資料庫。接著，我們將介紹 DocuSky 的 StatsTool，它是在 DADH 2016 論文（謝博宇，2016 年 12 月）所介紹字詞頻率統計工具的加強版，可以計算字詞在文本中的出現情形，並且以數種形式輸出統計的結果。我們將在第四節定義何謂文本字詞關聯圖 (Text-term Relationship Graph)，介紹如何利用工具來產生視覺化的文本字詞關聯圖，並探討它們的影響和可能應用。在最後一節，我們對這篇論文進行重點回顧，並總結數位人文的跨領域特性。

二 、DocuSky 與線上文本建庫

DocuSky 是個數位人文的學術研究平台。它的目標，是希望透過資訊工程師

[4] DocuSky 網址 https://docusky.digital.ntu.edu.tw/。

所開發的數位工具，讓人文研究者能夠從不同面向分析自己感興趣的文本，從而推進數位人文領域的發展。從使用者端的功能來說，DocuSky 是一個讓人文研究者可以上載自己感興趣的文本、建構個人資料庫、並對資料庫內容進行分析的研究平台（杜協昌，2016 年 12 月）。

自 2016 年底發佈以來，DocuSky 在資料流程與系統功能上都有顯著的進步。使用者除了可以上載純文字檔案來建構資料庫，更可整合文件的詮釋資料、內文的標記資訊、以及附屬於該文件的影像或圖檔（參考圖 1）[5]。由於篇幅限制，我們並不會在此介紹如何建構具有詮釋資料和標記資訊的文本資料庫，而是把重心放在如何運用 DocuSky、 StatsTool 和 Palladio[6] 來繪製視覺化的文本字詞關係圖。以下的說明，可參考圖 1 中，從左上方（線上開放文本）到右下方（詞頻統計、繪製關聯圖）的途徑。

當前網際網路上，已經有許多開放的文本儲藏庫 (text repository)，提供免費的文本讓使用者對其內容進行搜尋和瀏覽。例如，CBETA[7] 提供了完整的佛教經典文獻，而 CTEXT[8] 和 Kanripo[9] 則收集了大量的中文古籍文獻。為了便利研究者取用這些文本來進行研究，DocuSky 開發了一組小工具，可以列出這些線上資源的完整書目。使用者可以從書目清單中挑選感興趣的書籍，直接下載這些古籍的全文內容，並建構成 DocuSky 的個人資料庫[10]。圖 2 展示使用者如何利用這工具從 CBETA 下載五部戒律來建構 DocuSky 的個人資料庫。

一旦建庫完成，使用者就可以利用 DocuSky 所提供的工具對資料庫進行各種分析與應用。例如，使用者可以對資料庫的內容進行全文搜尋與瀏覽、可以透過 DocuGIS 工具呈現文本地名的地理空間分佈、也可以利用下一節所介紹的 StatsTool 對文本標記或關鍵字詞進行統計分析。有關 DocuSky 的基本理念與工具應用，可參考我們在 DADH 2016 所發表的論文（杜協昌，2016 年 12 月），或漢學國際研討會所發表的成果（杜協昌，2018 年 4 月）。

[5] 我們將另闢專文，從問題思辯到設計實作來介紹 DocuSky 如何看待「文字內容、標記資訊、詮釋資料、以及文件附圖」之間的資料整合問題。實務上的操作流程，可參考 DocuSky 網站的工具和說明，或者參加 DocuSky 的工作坊。

[6] Palladio 是史丹佛大學所開發的視覺化圖形呈現工具。只要在線上匯入表格資料，使用者就可以輕鬆繪製出多種視覺化圖形。網址 https://hdlab.stanford.edu/palladio/。

[7] CBETA 為中華佛典電子學會所維護的線上佛教文獻庫，網址 http://www.cbeta.org/。

[8] CTEXT 是 Donald Sturgen 所開發維護的漢籍文獻庫，網址 https://ctext.org/。

[9] Kanripo 是京都大學 Christian Wittern 所開發維護的漢籍文獻庫，網址 https://www.kanripo.org/。

[10] 這項便利的服務，需要文本儲藏庫提供書目資訊、開放線上取得文本內容的 API（Application Programming Interface，可以想成是兩組程式之間的互動窗口與溝通方式）、也需要了解這些內容的格式以便進行資料的彙整。這個例子說明了網路基礎設施 (cyber-infrastructure) 對數位人文工具環境的重要性，請參考（杜協昌，2018 年 4 月）的發表內容。

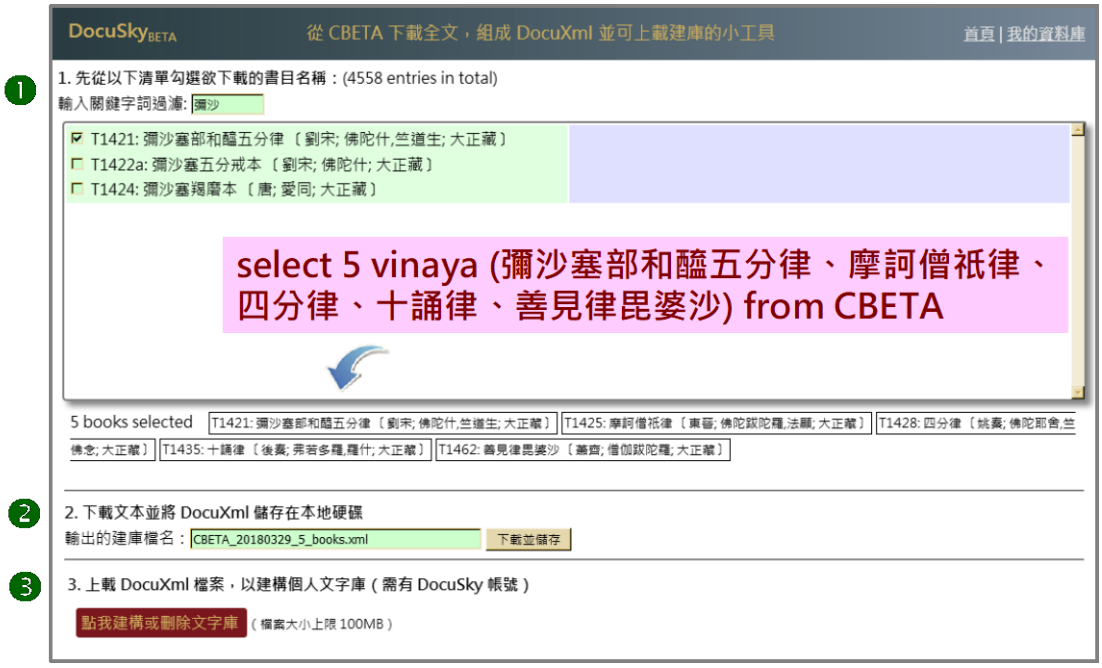
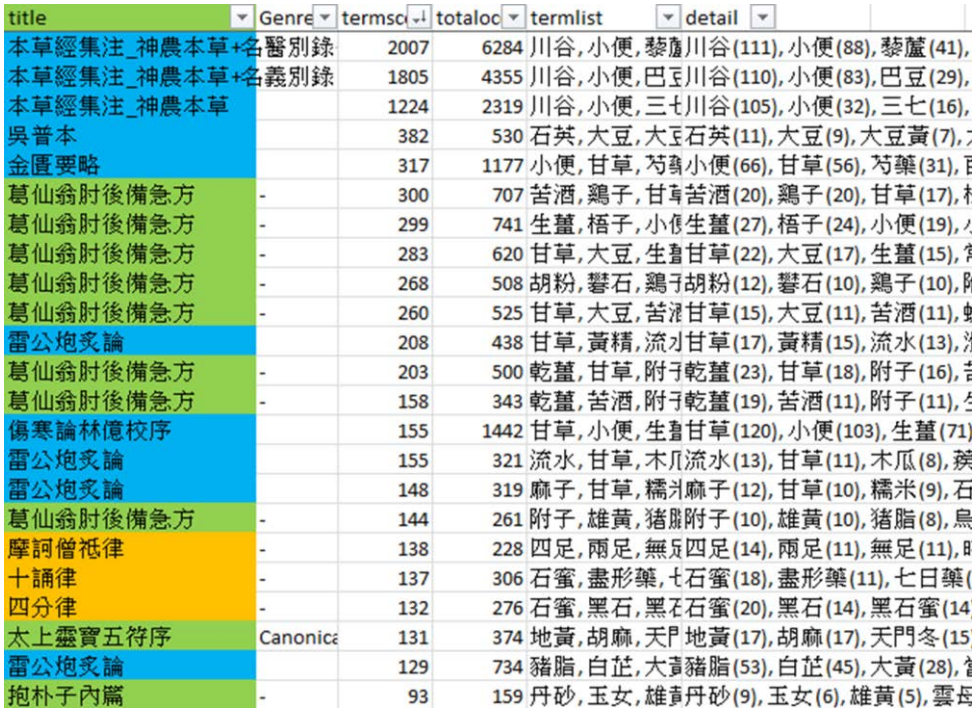


圖 2. DocuSky 所提供的的 CBETA 線上文本建庫工具。這工具會列出 CBETA 最新的完整書目列表讓使用者勾選。如本圖所示，使用者可以輸入關鍵字詞來進行書目篩選。在此例中，使用者輸入「彌沙」，而工具會列出 CBETA 書目（包含書名、作者、版本資訊）中含有「彌沙」的所有項目。使用者可反覆利用篩選機制來勾選欲下載的書籍。本例示範利用這個工具建庫的三個步驟。首先，使用者從書目清單中，挑選了《十誦律》、《四分律》、《摩訶僧祇律》、《彌沙塞部和醯五分律》和《善見律毘婆沙》等五部戒律（步驟❶）。接著，使用者需將下載後的文本以 DocuXml 格式儲存於本地硬碟（步驟❷）。最後，使用者就可將剛才儲存的 DocuXml 檔案上載到 DocuSky 來建庫（步驟❸）。

三、DocuSky 的 StatsTool 工具

如同第一節導論所述，我們認為特定領域的大量文本中，經常隱寓某些有趣的概念。這些概念因為不屬於文本主軸的脈絡，通常散落在文本的多個片段文字中，因而不易藉由傳統的精讀或略讀方式來進行觀察與推想。問題是，既然這些概念散落在文本四處，那我們該如何對它們進行觀察呢？一種方式是，假設散落的概念可透過關鍵字詞來觀察，那我們就可藉由關鍵字詞的分析，來捕捉這些概念在文本中的出現情況。例如，我們可以標記中國醫藥相關古籍中所出現的藥名，比較這些文本曾出現哪些藥名、又有哪些藥名並沒有出現，來觀察被這些文本所關注到、以及未關注到的中藥。藉由這類觀察，我們就有可能拓展我們的視野，從而對中醫藥物的起源有更深一層的認識。當然，由於精準的關鍵字詞標記需要大量的人工判讀，因此若欲分析大量的文本，實務上研究者也經常透過全自



title	Genre	termsc	totaloc	termlist	detail
本草經集注_神農本草+名	醫別錄	2007	6284	川谷, 小便, 藜蘆	川谷(111), 小便(88), 藜蘆(41),
本草經集注_神農本草+名	義別錄	1805	4355	川谷, 小便, 巴豆	川谷(110), 小便(83), 巴豆(29),
本草經集注_神農本草		1224	2319	川谷, 小便, 三七	川谷(105), 小便(32), 三七(16),
吳普本		382	530	石英, 大豆, 大豆	石英(11), 大豆(9), 大豆黃(7),
金匱要略		317	1177	小便, 甘草, 芍藥	小便(66), 甘草(56), 芍藥(31),
葛仙翁肘後備急方	-	300	707	苦酒, 鷄子, 甘草	苦酒(20), 鷄子(20), 甘草(17),
葛仙翁肘後備急方	-	299	741	生薑, 梧子, 小便	生薑(27), 梧子(24), 小便(19),
葛仙翁肘後備急方	-	283	620	甘草, 大豆, 生薑	甘草(22), 大豆(17), 生薑(15),
葛仙翁肘後備急方	-	268	508	胡粉, 礬石, 鷄子	胡粉(12), 礬石(10), 鷄子(10),
葛仙翁肘後備急方	-	260	525	甘草, 大豆, 苦酒	甘草(15), 大豆(11), 苦酒(11),
雷公炮炙論		208	438	甘草, 黃精, 流水	甘草(17), 黃精(15), 流水(13),
葛仙翁肘後備急方	-	203	500	乾薑, 甘草, 附子	乾薑(23), 甘草(18), 附子(16),
葛仙翁肘後備急方	-	158	343	乾薑, 苦酒, 附子	乾薑(19), 苦酒(11), 附子(11),
傷寒論林億校序		155	1442	甘草, 小便, 生薑	甘草(120), 小便(103), 生薑(71)
雷公炮炙論		155	321	流水, 甘草, 木瓜	流水(13), 甘草(11), 木瓜(8),
雷公炮炙論		148	319	麻子, 甘草, 糯米	麻子(12), 甘草(10), 糯米(9),
葛仙翁肘後備急方	-	144	261	附子, 雄黃, 猪脂	附子(10), 雄黃(10), 猪脂(8),
摩訶僧祇律	-	138	228	四足, 兩足, 無足	四足(14), 兩足(11), 無足(11),
十誦律	-	137	306	石蜜, 盡形藥, 七	石蜜(18), 盡形藥(11), 七日藥(
四分律	-	132	276	石蜜, 黑石, 黑石	石蜜(20), 黑石(14), 黑石蜜(14
太上靈寶五符序	Canonica	131	374	地黃, 胡麻, 天門	地黃(17), 胡麻(17), 天門冬(15
雷公炮炙論		129	734	猪脂, 白芷, 大黃	猪脂(53), 白芷(45), 大黃(28),
抱朴子內篇	-	93	159	丹砂, 玉女, 雄黃	丹砂(9), 玉女(6), 雄黃(5), 雲母

圖 3　透過 StatsTool 的 Categorized FileResult 輸出，可將字詞統計以文件為單位，產生 Excel 可讀取的 .csv 檔案 (每一列代表一篇文件)。輸出的基本欄位包含檔名、不同的字詞個數 (terms count)、字詞出現的總次數 (total occurrence)、字詞列表 (term list)、以及該列表中各字詞出現的次數 (detail) 等。從這張截圖可看到，除了這些基本的欄位，使用者還可指定輸出該文件的標題 (title)、以及其他自訂的詮釋資料欄位 (在此例為 Genre)。此截圖由 Michael Stanley-Baker 提供，他將 DrugsDb (其內容涵蓋中國六朝時期的醫藥相關文獻，參考第五節的說明) 的 FileResult 匯入 Excel 來進行觀察，用不同的底色標誌不同類型的書名，從而找出 DrugsDb 中包含多種藥名的文本。從這份截圖的倒數第四到六列可看出，《摩訶僧祇律》、《十誦律》、《四分律》這三部佛教戒律中，都含有超過 100 種中藥名稱。正是透過這樣的觀察，引發比較五份戒律在藥名詞彙分佈的實驗（請參考圖 2 和圖 5）。

動化的文本關鍵字詞擷取，來對關鍵字詞進行統計。

DocuSky 提供一個相當好的解決方案。使用者只要將文本建庫，就可以利用這個平台上所提供的工具對文字庫的內容進行檢索、瀏覽與分析，也可將工具的計算結果輸出，以利進行其他的加值應用。尤其是 DocuSky 提供 StatsTool[11] 工具，它可以計算字詞在文本中的出現頻率，並以數種不同的表格形式輸出統計結果。在這篇論文中，我們會將計算結果以 .csv 的表單格式輸出，以方便利用其他的工具匯入並進行後續應用。

[11] 這個工具的全名為 Tag/Term Statistics Tool。如同名稱所暗示，這個工具可以對標記後的文本 (tagged documents)進行標記字詞統計；也可用全自動的方式，計算未標記文本的關鍵字詞。

StatsTool 目前提供三種基本的輸出形式。第一種稱作 BasicTermResult，它是以字詞為單位，輸出各字詞在文本所出現的總頻率。BasicTermResult 可被應用於觀察關鍵字詞在文本中的使用頻率，例如以教育部成語典的成語作為關鍵字詞，觀察當代新聞中各成語的出現次數（謝博宇，2016 年 12 月）。第二種稱作 Categorized FileResult，它會以文件為單位，輸出各字詞在各文件中出現的頻率。這種形式的輸出，可讓我們得知哪些文件包含了最多不同的關鍵字詞，並從大量的文本中篩選出「含金量」較高的文件（參考圖 3）。第三種形式稱作 Categorized TermResult，它會將文件與其出現的字詞個別放在一列。使用者可將這種形式的輸出結果，直接匯入史丹福大學所開發的繪圖工具 Palladio，很快地繪製出下一節所介紹的視覺化文本字詞關聯圖 (visualized text-term relationship graph)。

四 、文本字詞關聯圖與視覺化呈現

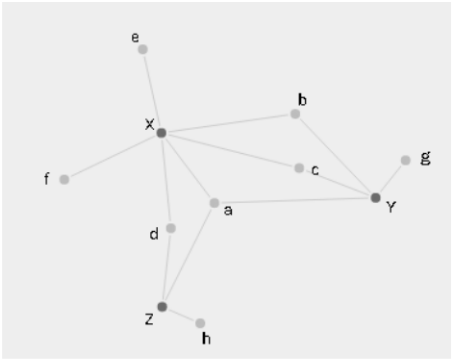
透過上一節所介紹的 StatsTool，使用者可以對文本中的關鍵字詞進行統計，並以表格的方式輸出。問題是，表格雖然能精準呈現每個字詞出現在哪一篇文件，但當表格逐漸變大（列數或行數增加），在有限視野內可呈現的資料將顯得繁雜且瑣碎，使得人們很難從中觀察出有趣的結果。此時，我們希望能透過資料視覺化 (data visualization) 的技術，讓研究者能夠從視覺化呈現中，更清楚地觀察到資料間的關聯性。視覺化技術涵蓋的範圍很廣，我們將只聚焦在能呈現文本和字詞關聯的圖形、將這種圖形視覺化的工具、以及視覺化圖形的案例討論。

E = {(X,a), (X,b), (X,c), (X,d), (X,e), (X,f),
(Y,a), (Y,b), (Y,c), (Y,g),
(Z,a), (Z,d), (Z,h) }
G = (N, E)，其中 N ⊆ D ∪ T

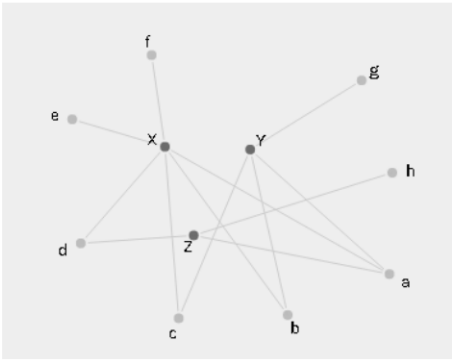
(a)

text	term
X	a
X	b
X	c
X	d
X	e
X	f
Y	a
Y	b
Y	c
Y	g
Z	a
Z	d
Z	h

(b)



(c)



(d)

圖 4　幾種文本字詞關聯圖的等價表達： (a) 使用正統的集合表達方式，集合 G 包含了節點集合 N 與連線集合 E，其中 E 的每一個元素 (x, y) 代表 y ∈ x，也就是字詞 y 出現在文本 x 中。 (b) 使用表格的方式，其中每一列 (row) 代表一條線。 (c) (d) 是兩份具有相同拓樸的視覺化呈現圖形。目前並沒有良好的準則可判定視覺化呈現圖的良窳，但我們認為好的呈現圖應該能讓人們很容易看出該圖所欲說明或彰顯的關聯性。例如，由於我們可以很容易地從 (c) 觀察出文本中共同出現的字詞（例如字詞 a 在三份文本中都有出現，字詞 b, c 共同出現於 X, Y 文本），我們認為相較於圖 (d)，圖 (c) 具有較佳的視覺化呈現效果。

我們先定義何謂文本字詞關聯圖 (Text-term Relationship Graph, TTRG)。給定文本集合 D 和詞彙[12] T，文本字詞關聯圖是一個圖 G=(N,E)，其中節點集合 N ⊆ D ∪ T[13]，連線集合 E={(d,t): t ∈ d, t ∈ T, d ∈ D}，其中 t ∈ d 表示字詞 t 出現於文本 d[14]。翻譯成白話：文本字詞關聯圖是以文本和字詞作為節點；若文本 d

[12] 詞彙 (term vocabulary) 是字詞的集合。由於「詞彙」在中文既可代表字詞的集合，又可指稱單一的字詞（例如我們可以說「詞彙 t 在某文本僅出現一次」），為了減少不必要的誤解，本文將避免將「詞彙」用於指稱單一字詞。

[13] 我們可以將 N 簡單定義為 N = D ∪ T，但這樣一來，將會導致圖 4(a)-4(d) 的表達方式並不必然等價（例如若有某份文本 d 並未包含 T 的任意字詞，或者某個字詞並未出現於任意文本，都會使得問題複雜化，參考註腳 18）。為了簡化這類困擾，我們將可 N 定義為 $D_0 \cup T_0$，其中 D_0 僅包含「出現於 E 的文本」，而 T_0 僅包含「出現於 E 的字詞」。也就是說，D_0 = {d: ∃t s.t. (d,t)∈E}，T_0 = {t: ∃d s.t. (d,t)∈E}。

[14] 在此 t ∈ d 有兩種等義的說法：「字詞 t 出現在文本 d 中」與「文本 d 包含了字詞 t」。

包含了字詞 t，就從節點 d 到節點 t 劃上一條直線。在一般的圖論 (graph theory) 中，若要視覺化呈現一份圖，(d,t) 會被畫成一條帶有箭頭的有向線 d→t。然而，大量帶有箭頭的線會降低圖形在視覺化呈現的效果，因此我們用較深的節點顏色代表文本 d，較淺的節點顏色代表字詞 t，然後將 (d,t) 繪製成沒有帶箭頭的線段 [15]。

我們用圖 4 的例子來說明文本關聯圖可以有多種表達方式：形式化的數學集合、表格的形式、以及視覺化的呈現圖。注意到這幾種表達方式，都傳遞了相同的訊息，它們在意義上等價的：從圖 4(a) 可以推導圖 4(b)，圖 4(b) 可以產生圖 4(c)，圖 4(c) 可以產生圖 4(d)，而從圖 4(d) 也可以推得圖 4(a)[16]。我們將圖 4(c) 與圖 4(d) 稱為視覺化文本字詞關聯圖 (visualized TTRG)，它們呈現了文本和字詞之間的關聯性。例如從圖 4(c) 中，我們可以看到文本 X 包含了字詞 a, b, c, d, e, f；也很容易觀察到字詞 a 在這三份文本中都有出現。我們還可以發現，字詞 e, f 僅出現於文本 X：它們都沒有出現在文本 Y 或 Z 中 [17]。

因此，一份文本字詞關聯圖的資料，可以表示成形式上的數學集合、類似 Excel 的表格結構、也可利用繪圖工具繪製視覺化關聯圖。這些表達方式各有其優點，可以讓人對相同的資料有不同的觀看角度與聯想視野。抽象的數學集合可以提醒我們回顧關聯圖的基本定義與性質 [18]、表格結構可以清楚展現文本和字詞之間的個別關聯 [19]、而視覺化關聯圖則可讓研究者從整體、宏觀的角度看見文本和字詞之間的關聯性。尤其是，當我們手頭有一份字詞列表，想知道列表中哪些

[15] 文本字詞關聯圖中的任意有向線 (x, y)，x 都必然指某份文本，而 y 必然是某個字詞，所以並不會有從字詞指向文本的線段。既然我們已經用深色節點代表文本，淺色節點代表字詞，兩點之間的線必然代表從文本指向字詞，因而可以略去箭頭而不會有影響。

[16] 我們需要註腳 13 對節點集合 N 的定義，才能從 4(d) 推回 4(a)。

[17] 在文本字詞關聯圖中，與某文本相連的字詞都「共同出現」在那份文本 (terms co-occurred in that text)，因此從關聯圖確實可以很清楚地看出有哪些字詞「共現」在某份文本中。然而，從這裡和第五節的實例中，我們可以發現文本字詞關聯圖的一項重要優點，是它可以展現「某個字詞僅出現在某些文本、卻沒有出現在其他文本」(a term occurred only in specific texts but not the others) 的情況。我們將本文所定義的圖形取名為「文本字詞關聯圖」而非「文本字詞共現圖」，就是為了避免讀者錯誤地將這種優點解釋為「某個字詞共現在某些文本」(從對應的英文可以清楚地看出錯誤之處:：a term co-occurred in specific texts)。

[18] 例如，注意到關聯圖的任意連線 (d, t) 必須起源於文本集合 D 和詞彙 T，而圖 4(a)-4(d) 都沒能完整表達出 D 和 T 的性質。考慮 D={X,Y,Z,W}，T={a,b,c,d,e,f,g,h}，其中 X, Y, Z 所包含的字詞都在圖 G 中，但文本 W 並沒有包含 T 的任何字詞。這種狀況下，這一節所定義的文本字詞關聯圖，就無從表達 D 其實還包含一份頗為特殊的文本 W。實務上，我們可以進行補救：在視覺化關聯圖中，繪製單一隔離 (stand-alone) 的深色文本節點 W 來表達「還有一份文本 W，但這份文本並沒有包含圖中的任何字詞」。欲完整表達出 D 和 T 的性質，最簡單的方式是修改關聯圖的定義，令 N = D ∪ T（參考註腳 13）。但這樣一來，表格結構會變得複雜，且視覺化關聯圖將不但包含單一隔離的文本，也會包含單一隔離的所有字詞，這結果通常並不是使用者所想要的（例如詞彙 T 可能包含了所有蒐集到的中藥名稱，但文本 D 其實僅包含其中一小部分藥名）。

[19] 表格的另一項重大優點是彈性。因為每一項關聯在表格中是以一列 (row) 來表示，我們可以在表格上彈性增添一些欄位，對這項關聯加上屬性或註記。

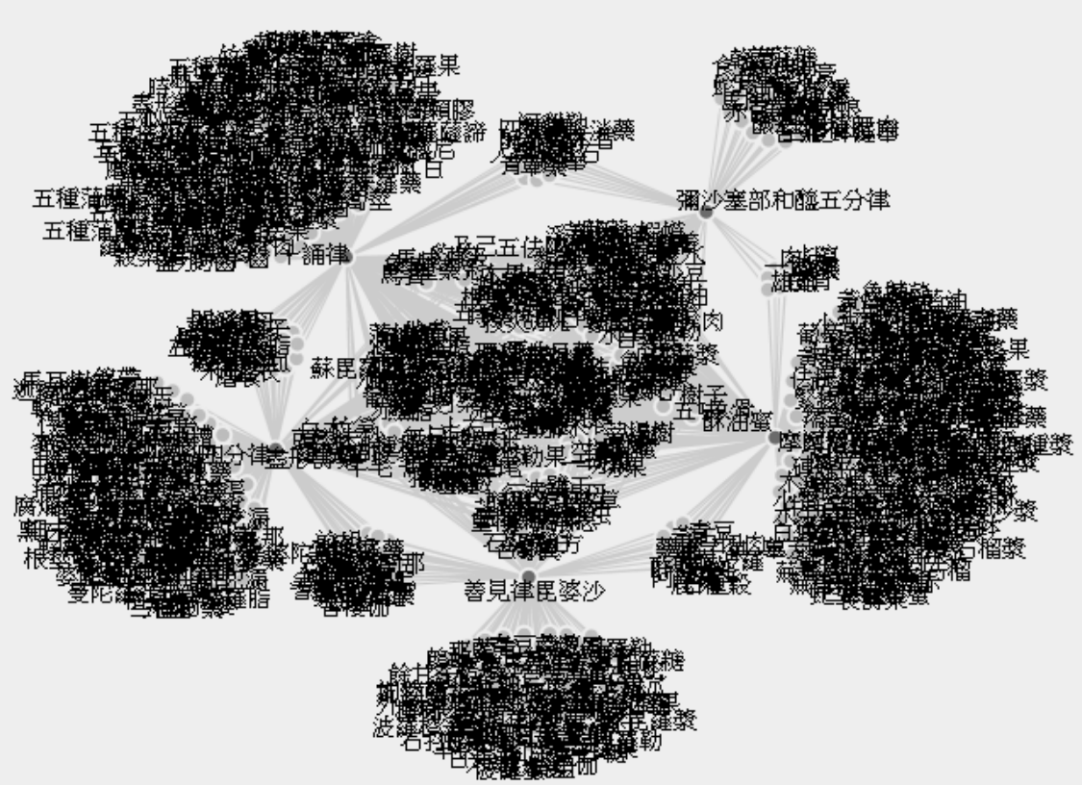


圖 5. 五份佛教戒律與藥名的關聯圖。Palladio 的演算法會嘗試將與多份文本有連接的字詞擺放在這些文本之間，它同時也允許使用者在相同的拓樸結構下拖曳節點到喜歡的位置。我們將五份戒律的節點拖放在外圈環形結構上（本例從左上方順時針方向依序為《十誦律》、《彌沙塞部和醯五分律》、《摩訶僧祇律》、《善見律毘婆沙》和《四分律》），共用的藥名字詞會被 Palladio 自動配置到環形的圈子裡面。從該五份戒律往外延伸的詞叢，明顯可看出它們都各自有相當多「僅出現於該戒律」的特殊字詞。例如，位於右側的《摩訶僧祇律》和左上方的《十誦律》各自具有向外延伸的大詞叢（數量可能比共用的字詞還多），表示它們有極多特殊字詞；右上方的《彌沙塞部和醯五分律》往外延伸的詞叢雖然較小，但也含有超過十個「僅出現於該戒律」的特殊藥名字詞。

字詞會共同出現於某幾份文本，又想比較有哪些字詞會共同出現於另幾份文本時，視覺化的文本字詞關聯圖可以提供相當有效的觀察方式。

圖 5 是視覺化文本字詞關聯圖的一個應用實例。它顯示梵文中譯後的五份佛教戒律，與其內文出現的中醫藥物名稱之間的關聯[20]。任何使用者都可透過

[20] 這個例子起源於 Michael Stanley-Baker (徐源) 在 2017 年以六朝 (魏晉南北朝) 時代的文本，對於中藥起源所進行的研究。徐源收集了六朝時代與醫藥相關的道藏、佛典、以及醫書，將它們建構成 DocuSky 資料庫，利用 Excel 檢視 StatsTool 的輸出，從而發現這幾份包含大量中藥名稱的五份戒律（參考圖 3）。透過視覺化關聯圖，徐源觀察到這五份戒律各自擁有很多特殊的藥名。接下來，經由比對這些戒律在印度的起源地，他察覺到一個不易從傳統閱讀方法發現的問題：這些特殊的藥名，是因為原本經文就已使用不同的當地藥材，還是因為譯者將藥名換成熟悉的中國藥材名稱? 我們藉由這項成果強調：合宜的視覺化關聯圖，加上適當

DocuSky 所提供的工具，直接從 CBETA 下載這五份戒律：《十誦律》、《四分律》、《摩訶僧祇律》、《彌沙塞部和醯五分律》和《善見律毘婆沙》來建庫（下載建庫的流程可參考圖 2）。接著，使用者只需另外準備好中醫藥物的名稱列表，就可利用 StatsTool 輸出[21]文本與藥名關聯的 .csv 表格。最後，把這份輸出的表格匯入 Palladio，就可輕易地繪製出具有相同拓樸[22]的視覺化圖形。從圖 5 中，我們可以很容易看出，這五份戒律雖包含不少共同使用的中藥字詞，但每份戒律都有為數眾多「僅出現於該戒律」的特殊中藥名稱。這個例子說明了視覺化圖形的優異性能：它可以清楚展現這五份戒律各自擁有為數不少的特殊藥名字詞（即使因版面大小的限制，這些字詞堆積成詞叢而無法清晰展現）。如果沒有經過視覺化處理，我們將很難從數學形式的集合、矩陣、或者龐大的表格中，輕易觀察出這項結果。

五 、幾個文本字詞關聯圖的視覺化實例

文本字詞關聯圖有兩項要素：文本集合和詞彙。理論上，給定任何文本集合 D 和詞彙 T，套用特定演算法[23]之後，我們都可計算出連線集合 $E=\{(d,t): t \in d, t \in T, d \in D\}$，並得到唯一的文本字詞關聯圖。然而在實務上，文本集合與詞彙並不該是任意的：任意的文本集合加上任意的詞彙，通常只會產生無法被人們解讀的關聯圖[24]。這暗示關聯圖一般都需要人文學者的參與，才能挑選出合適的文本集合與詞彙，並對視覺化的結果進行有意義的觀察與解讀。在這一節裡，我們將舉幾個視覺化關聯圖的例子，指出它們的有趣之處，並討論可能的後續應用。此外，如同在圖 4 解說中所討論，我們可有多種視覺化呈現關聯圖的方式。在這節的例子中，我們利用 Palladio 繪製視覺化關聯圖時，會盡量將文本節點拖放成一個環形，讓共同出現的字詞呈現在環內，而僅出現於特定文本的字詞則擺放到這個環形外。

的觀察與比較，具有引發新議題的潛力。

[21] 以 Categorized TermResult 形式輸出（參考第三節對 StatsTool 的說明）。這種形式的輸出，可將每份文件與其出現的字詞放在一列來進行輸出。由於文本可能包含多份文件（文本是文件的集合），我們可利用文件的詮釋資料欄位（例如「出處」）來指明文件屬於哪一份文本。Categorized TermResult 可額外輸出文件在這項詮釋資料的值。將這份輸出表格匯入 Palladio 後，只需選擇用這個詮釋資料的欄位作為「文本」，就可繪製出文本字詞關聯圖。

[22] 兩份視覺化圖形若都代表相同的資料，我們說它們具有相同的拓樸 (topology)。例如圖 4(c) 與圖 4(d) 就具有相同的拓樸。一般來說，我們可以任意移動視覺化文本關聯圖的節點，而不改變圖形的拓樸。

[23] 如同第一節所介紹，我們可以套用標記後的結果，或者採用全自動的演算法來計算字詞是否曾出現於某份文本。

[24] 這一節會多次提到「視覺化的文本字詞關聯圖」。在不會引發誤解的狀況下，我們將省略「視覺化」和「文本字詞」而僅稱呼為「關聯圖」。

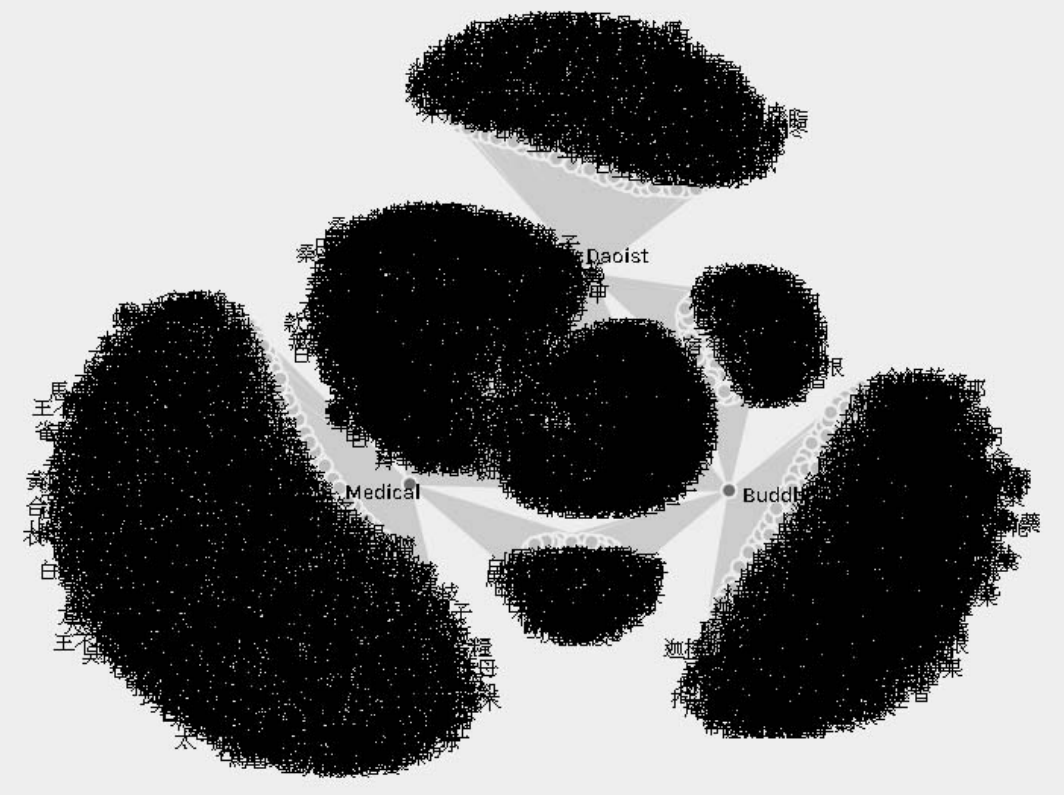


圖 6　利用 DrugsDb 的三個文獻集作為文本，套用中醫藥名作為詞彙，所得到的視覺化文本字詞關聯圖。從圖中可以很明顯地看出，Buddhist（佛典）、Daoist（道藏）和 Medical（醫書）都各自擁有為數眾多「未曾出現於其他文本的藥名」。由於 DrugsDb 已經包含絕大多數六朝時期與中醫相關的文獻，這張圖就強烈支持「在六朝時，佛道與醫書在使用的中醫藥名上有相當大的差異」。另外，我們也可觀察到，Daoist 與 Medical 之間也有一塊相當大的詞叢（僅出現於道藏和醫書，而沒有出現在佛典中的藥名）。這暗示六朝時期，民間的醫書與道教文本（或者引伸至民間與道教的藥學知識）之間的關係可能較為強烈。

第一個例子的文本取自 DrugsDb 資料庫。這個資料庫是德國馬克斯普朗克 (Max Planck) 科學史研究所贊助，由 Michael Stanley-Baker（徐源教授）透過網路資源所收集的六朝時代中國醫藥相關文本[25]。徐源將這些收集來的文件分成三個文獻集：Buddhist（佛典）、Daoist（道藏）以及 Medical（醫書）[26]，並且為其中多數文件加上出處、出版年代、以及自訂類型 (Genre) 等詮釋資料。徐源也另外收集了中醫藥材和疾病的名稱，因此我們可以透過前兩節所介紹的方式，以 DrugsDb 的文本和藥名詞彙（或病名詞彙）繪製關聯圖。

[25] 六朝是指中國魏晉南北朝時期的六個朝代。目前 DrugsDb 已置於 DocuSky 的公開資料庫，任何人都可對其進行檢索與瀏覽（從 DocuSky 首頁進入後，在「公開的資源」下可找到此資料庫的連結）。本論文所挑選的許多圖例（圖 2、圖 3、圖 5、圖 6、圖 7）都與 DrugsDb 的研究相關，請參看註腳 20。

[26] DrugsDb 包含有道藏 849 卷、佛典 2444 卷、以及醫書 113 卷，全文字數高達數千萬，檔案大小合計約 100MB。

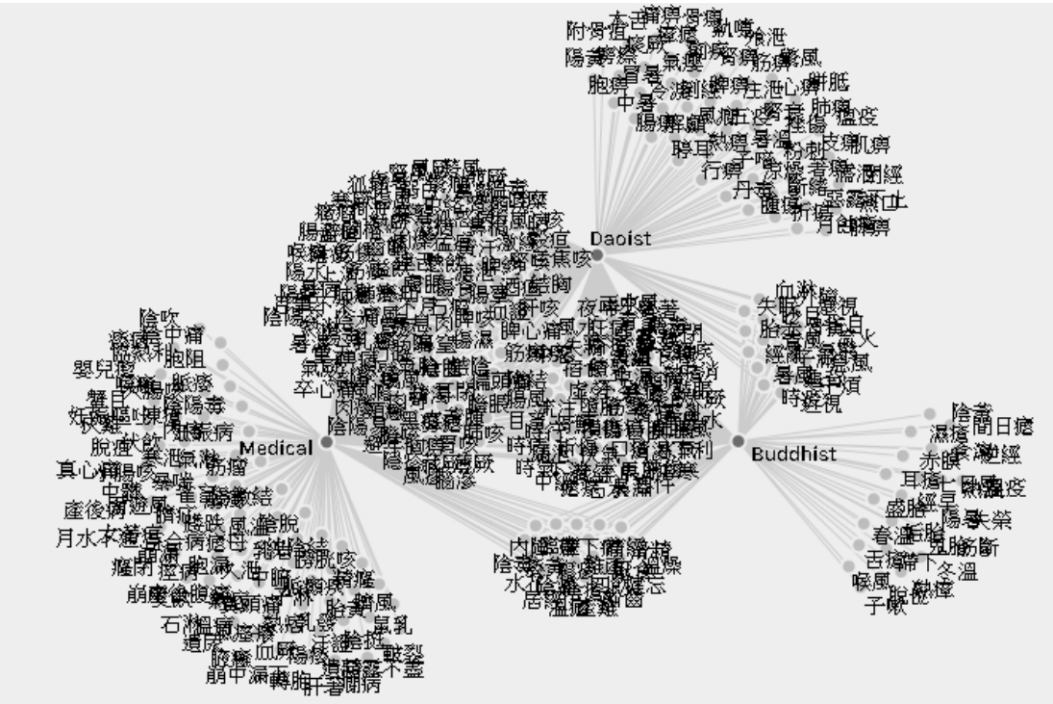


圖 7　以 DrugsDb 資料庫的文獻集作為文本，套用病名詞彙所繪得的視覺化文本字詞關聯圖。注意到它與圖 6 之間的相似性：三份文本 (Daoist 道藏、Buddhist 佛典、Medical 醫書) 都依然各自擁有為數不少（僅出現於該文本的）特殊病名。相較於佛典特殊藥名的詞叢，佛典特殊病名的詞叢顯得稀疏許多。

圖 6 是以 DrugsDb 三份文獻集作為文本，套用中藥詞彙所產生的文本字詞關聯圖。從這張圖中，我們可以很清楚地看到，文本 Daoist（道藏）、Buddhist（佛典）和 Medical（民間醫書）之間有個中等大小的詞叢，裡面的詞彙是這三份文本都曾出現的藥名。透過這個詞叢的字詞，我們就可以得知六朝時期佛道和民間所共同使用的藥名，並據而推想它們之間所共同認知的中藥知識[27]。除了展現共用字詞的詞叢，這張圖最有趣的地方是：我們可以清楚看到佛道和醫書各自擁有為數眾多的特殊藥名[28]。這項觀察支持[29]「六朝時代，中醫的知識散落在佛教、道教與民間；且這三個領域所記述的中醫知識，彼此有相當差異」的信念[30]。注意到，這個例子展現本文所介紹的數位人文方法，其實具有高度科學性質：它

[27] 可惜的是，目前 StatsTool 並沒有計算共用字詞的功能。我們目前也沒有其他的通用工具，能夠將這些共用藥名單獨輸出，以利進行更深入的檢視與探討。

[28] 在此「特殊藥名」是指僅出現於該文本，而沒有出現於其他文本的藥名。

[29] 或者反過來說，若佛道和民間對中醫的知識有高度的同質性，那麼我們將無從解釋為何它們的文本記錄竟各自擁有如此多的特殊藥名。

[30] 歷史學家可能很早就有類似的觀察或想法。然而，因為傳統精讀和略讀方法都缺乏對大量文本進行檢視和比對的能力，學者一般只能就少數文本和片段的文字來進行推想與論證。這個例子彰顯數位工具對數位人文研究的重要性：有了數位工具，我們才能對大量文本進行檢測，並對全貌提供較為客觀的線索。

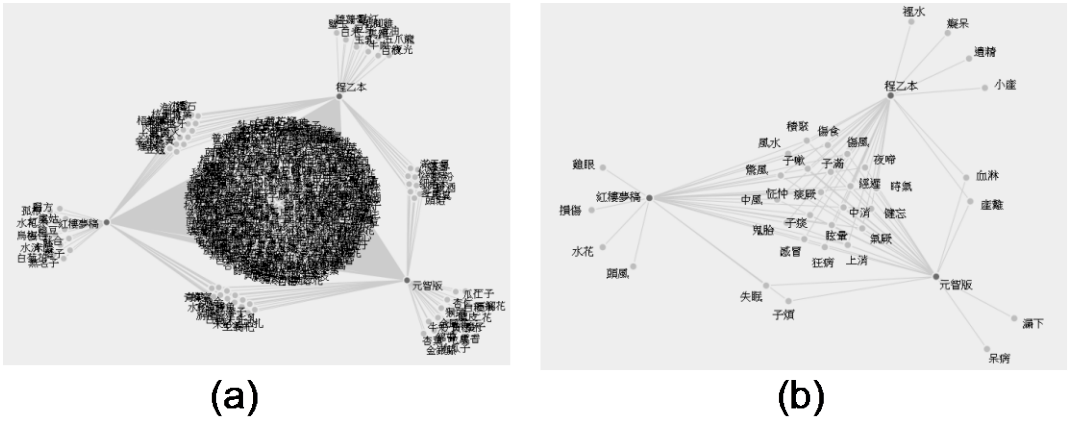


圖 8　以《紅樓夢》的三種版本（紅樓夢稿、程乙本、元智版）作為文本，套用藥名與病名詞彙所繪製的關聯圖（左圖套用藥名詞彙，右圖套用病名詞彙）。圖 (a) 外環左方、右上和右下方的點分別代表紅樓夢稿、程乙本和元智版。由於版本之間的內容應該相當類似，三份文本之間所顯示的大詞叢是可預期的。比較有趣的，是每個版本都仍有許多「僅出現在這個版本」的特殊藥名，且版本兩兩之間也各自有「僅出現於這兩種版本」的藥名。另外，從圖 (b) 可以看到，「紅樓夢稿」和「程乙本」之間並沒有兩者都曾出現，但「元智版」卻沒出現的病名。

讓我們可以進行實驗[31]，來檢驗某些假說或猜測。例如，我們可能會猜想，既然在六朝時期，佛道和民間關於中醫的知識是分散歧異的，那麼到了唐朝，這些知識應該會融合在一起。若我們收集了足夠多的唐朝醫藥文獻，就可利用它的關聯圖來驗證這項猜想：若佛道和民間的中醫知識，在唐朝已經融合在一起，那麼我們應可從唐朝文本的關聯圖上，觀察到相對少量的特殊藥名[32]。

我們也可以利用疾病名稱的詞彙，來檢驗從圖 6 所觀察出的推測：「在六朝時期，佛道和民間醫藥知識是分散且有歧異的」。若這項推測是合理的，那麼以相同的 DrugsDb 文本，套用病名詞彙所繪製的關聯圖，應該和圖 6 有相當程度的相似性。圖 7 就是在這種思想背景下所進行的實驗結果。比較圖 6 和圖 7，我們可以發現兩者在節點和詞叢的分佈上相當類似。例如，三份文本都有不少特殊病名，而道藏和醫書之間都曾出現的病名數，遠多於僅出現於佛典和道藏、以及佛典和醫書之間所共同使用的病名數。另一方面，比對兩張圖的詞叢密

[31] 在此「實驗」是指對不同的文本和詞彙內容來進行測試與觀察。當然，這些實驗是可重覆進行、可被他人進行檢驗的。

[32] 可能會有人質疑：許多迥異的藥名，其實都指涉相同的藥材。例如「芝麻」和「胡麻」通常指涉相同的物品，而「巴豆」也稱為「雙眼龍、大葉雙眼龍、江子、猛子樹、八百力、芒子」（參考《中國藥典》）。我們的分析並沒有考量到同義詞，會不會導致解釋和推論上的偏差？對此我們的回答是，雖然在此關聯圖並沒能展現文本和藥材（一種藥材可能有多種名稱）之間的關聯，但它依然清楚展現不同領域（佛道和民間）在藥名運用上的差異。當然，我們也期待能夠加強 StatsTool，讓它能夠加入同義詞的功能，以便對此提供更深入的分析。

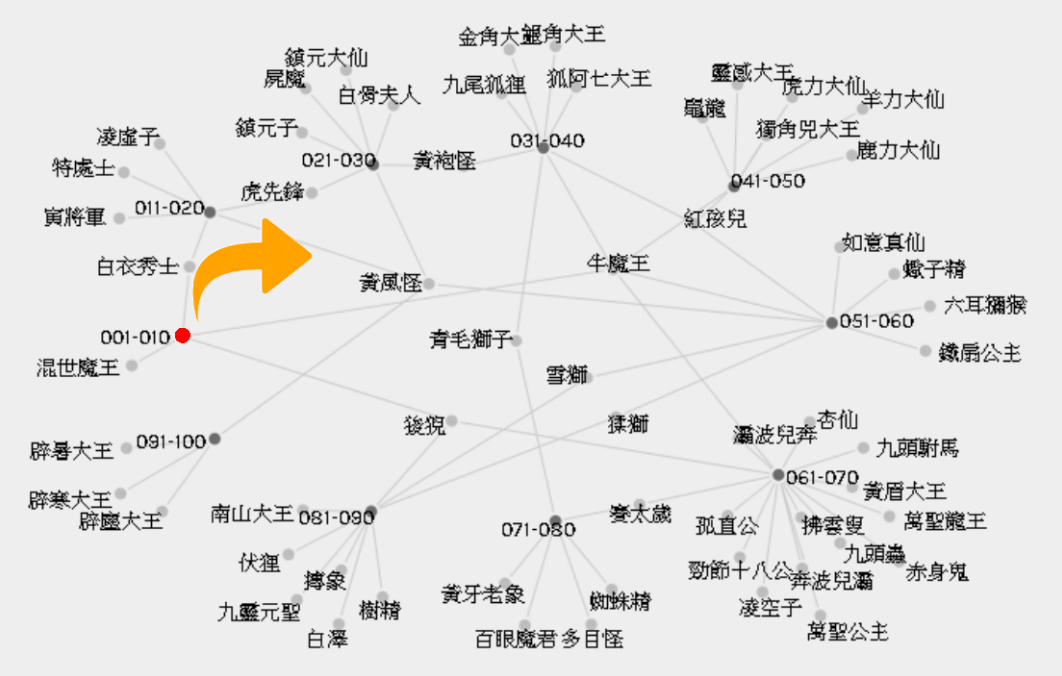


圖 9　將《西遊記》以每十回作為一個文本，套用妖怪詞彙所繪製的關聯圖。由於這些文本之間具有順序性，我們在視覺化的繪製過程中將它們以順時針方向擺放，這樣可以方便我們從第一個文本節點 (001-010，代表前十回）開始，依序檢視這些文本所出現的妖怪名稱。從這張圖中，我們可以很清楚地看出「黃風怪」曾出現在 001-010、021-030、051-060、091-100 等文本中。

度，我們發現佛典的特殊病名詞叢，遠較佛典的特殊藥名詞叢來得稀疏。可惜的是，目前 StatsTool 並無法以詞叢為單位來將字詞列出，因此我們不易對高密度詞叢的「特殊藥名」進行更深入的檢視與分析。我們猜想，由於六朝的佛典多半譯自梵文，這些高密度詞叢的特殊藥名，可能多半來自翻譯時的大量音譯藥名[33]。

對文本的不同版本繪製視覺關聯圖，有時也可得到一些有趣的結果。圖 8 是以《紅樓夢》的三個版本「紅樓夢稿、程乙本、元智版[34]」作為文本，套用藥名與病名詞彙所繪製的關聯圖。從圖 8(a) 中可以很清楚地看到三份文本之間有個龐大的詞叢：這是可以預期的，因為不同版本之間的文字內容應該非常相似。圖 8(a) 的有趣點，是它顯示各版本和兩兩版本之間，都仍有其特殊的藥名詞彙。圖 8(b) 是以這三份文本和病名詞彙所繪得的關聯圖，它顯示「紅樓夢稿」和「程

[33] 例如《善見律毘婆沙》提到：拘跋陀羅飯者，此是稊米飯也。修步者，此是青豆羹。吉羅者，此是竹[illegible]藥也。那。其中「拘跋陀羅飯」、「修步」、「吉羅」、「那」都被視為藥名，但它們一般並不會出現在道藏或民間的醫書中。

[34] 「元智版」是「用現存的十餘種本子互參互校，擇善而從」的《紅樓夢》版本，由浙江大學蔡義江教授校注。數位化文本取自元智大學的網站 http://cls.lib.ntu.edu.tw/HLM/home.htm。

乙本」之間並沒有兩者都出現，但「元智版」卻沒出現的病名。這個例子說明在適當條件下，關聯圖可以清楚呈現字詞共同出現在哪些文本，且並沒有出現在其他文本。這也是在呈現上，視覺化圖形優於數學集合或表格結構的良好實例。

圖 9 展示關聯圖的另一種可能應用。這張圖將《西遊記》切分為十份文本（每十回當作一份文本，例如 031-040 代表第 31-40 回），並顯示這些文本中各種妖怪出現的情形[35]。由於這些文本具有順序性，我們在視覺化的呈現中將它們以順時針方向擺放，這樣可以讓我們依照這些文本的順序來檢視妖怪出現的情形。這張圖清楚展示出，《西遊記》中妖怪多半僅與一個文本的節點（一個節點包含十回章節）相連。這項觀察並沒有帶來什麼新鮮的想法，因為隨著故事情節的發展，唐僧西行一路上所遇到的妖怪本就應該不一樣。相對有趣的，反倒是詢問為什麼「黃風怪」和四個文本節點有連結，也就是「黃風怪」為什麼在 001-010、021-030、051-060、091-100 都有出現。從這裡可以看出靜態關聯圖的不足：使用者若對於關聯圖的某個節點（或連結）感到好奇，靜態圖並沒有辦法提供互動的功能，讓使用者可以對該節點（或連結）進行更深入的檢視。換個方式說，我們應該提供互動的工具，讓使用者能夠從關聯圖上，直接對感興趣的節點或連結進行較深入的檢視（例如，點選「黃風怪」節點，系統就開啟另一個視窗，顯示所有黃風怪在文本出現的前後文）。當然，欲開發這類具整合性質的互動工具，將需要對資料間的互動協定進行更細密的了解與規範。我們相信，對這類互動與協定進行研究與開發，可以對數位人文的研究發展產生深遠的影響。

六 、結論與未來展望

在這篇論文中，我們簡述了利用關鍵字詞列表，來對文本內容進行分析比對的動機與方法。使用者可利用 DocuSky 所提供的工具，從 CBETA、CTEXT、Kanripo 等開放儲藏庫選擇所需的文本，將這些文本的內容打包下載、並建構成 DocuSky 的個人資料庫。接著，使用者只需準備關鍵字詞的列表，就可利用 StatsTool 來對建庫文本進行字詞的統計分析。將 StatsTool 的計算結果匯出儲存，再將該檔案匯入 Palladio，就可繪製出視覺化的文本字詞關聯圖。

這篇論文採取較為嚴謹的態度來看待文本字詞關聯圖。我們認為，視覺化關聯圖的最終目的，是讓研究者可以很清楚地從圖表的呈現中觀察到文本和詞彙間的關係。雖然相同的關聯圖可以有多種視覺化的呈現方式，且人們可以對最終視覺化的呈現圖進行不同解讀，但嚴謹的定義可以有效減少錯誤解讀的可能性。我們也嘗試利用幾個實際的例子，說明視覺化關聯圖可以展現傳統精讀略讀、甚至表格整理所難以達到的呈現效果。這些視覺化的優點，將有助於人們理解大量文本背後所隱藏的一些脈絡。

[35] 《西遊記》的妖怪詞彙，由台灣大學數位人文研究中心的胡其瑞博士提供。

總結地說，DocuSky 的個人資料庫，具有讓研究者自己操控文本內容的彈性。透過 DocuSky 的工具，使用者可以對建庫文本和詮釋資料擁有完整的控制權，並能透過數位工具對文本進行關鍵字詞的統計與比對。這篇論文所強調的文本字詞關聯圖，僅是這些方法中的一種統計與呈現方式。這種方法具有可重覆、可檢驗的科學性質：我們可以透過抽換文本和詞彙，來進行不同的實驗，並對呈現結果進行觀察和探討。

最後，雖然這篇論文著重於資訊工具的應用，我們不能輕忽文史學者對於數位人文工具開發的重要影響。憑藉著對時代與環境背景的深刻理解與感受，文史學者在文本選擇、詞彙篩選、以及分析結果的解讀上，都扮演著資訊科技所無法取代的角色。例如，文本字詞關聯圖的最初想法，就是來自於文史學者 Michael Stanley-Baker 利用六朝文本對中國醫藥源流的研究。此外，雖然理論上固定文本與詞彙，即可產生唯一的文本字詞關聯圖，但一份關聯圖還是可以有多種視覺化的呈現方式，此時也需文史學者從中挑選合宜的視覺化關聯圖。更重要的是，在多數情況下，僅靠分析結果的表格輸出和靜態關聯圖，也並不足以解決研究者的提問或疑惑。這暗示我們不應將視覺化的文本字詞關聯圖視為研究的最終產出：文史研究者通常需要進階的工具，幫助他們從詞彙的分析結果和關聯圖，回過頭來對文本進行更深入的檢視與再理解。從這裡的討論可以看出，數位人文工具的需求，經常是為了解決研究者在文本探究過程中所遭遇到的不便。我們相信，數位人文研究領域的推展，需要資訊與文史領域的專家學者共同努力，而 DocuSky 也將是促成這類跨領域合作的良好典範。

參考文獻

1. 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，台北。DADH 2016 國際研討會論文集 pp 25-38。
2. 謝博宇（2016 年 12 月）。〈以 DocuSky 為核心的工具開發與建置〉。第七屆數位典藏與數位人文國際會議，台北。DADH 2016 國際研討會論文集 pp 57-81。
3. 杜協昌（2018 年 4 月）。〈從 DocuSky 看跨平台的資源介接與應用〉，數位人文視野下的漢學研究國際研討會。



Research Infrastructure for the Study of Eurasia (RISE)

Towards a flexible and distributed digital infrastructure for resource access via standardized APIs and metadata

Sean Wang Pascal Belouin

Shih-Pei Chen Hou Ieong "Brent" Ho

Max Planck Institute for the History of Science

Research Infrastructure for the Study of Eurasia (RISE): Towards a flexible and distributed digital infrastructure for resource access via standardized APIs and metadata

Sean Wang, Pascal Belouin, Shih-Pei Chen, and Hou Ieong "Brent" Ho
Max Planck Institute for the History of Science

Abstract

Digital humanities (DH) is a burgeoning field of research in sinology and Asian studies more broadly, and its diversity and maturity necessitate a digital research infrastructure fit for DH-focused scholars' specific needs. In particular, the DH landscape evolved in a way that encourages fragmentation of both sources and tools, and these compartmentalized resources centered around disciplines and texts. **RISE** (formerly known as "Asia Network") is our solution to address this fragmentation across disciplines. It is a pioneering approach for resource dissemination and emerging data analytics (such as text mining and other fair-use, consumptive research techniques) in the humanities. It is a language-agnostic software that facilitates the secure linkage between third-party research tools to different third-party textual collections (both licensed and open-access ones) via application programming interfaces (APIs). Put more simply, it reduces the distances among DH resources not by duplicating them in a central repository, but by linking them together via flexible APIs. It revolutionizes how scholars can work with textual sources by promoting a flexible, networked approach to digital infrastructure development. Crucially, RISE is a loosely-coupled software with flexible topologies; it can enable both federated or centralized linkages, and it can even "disappear" as long as its API and metadata standards remain in place to facilitate communications among distributed databases and tools in the back-end. Thus, unlike large-scale infrastructural projects, RISE actively lowers the profile of centralized infrastructure and instead promotes existing tools and resources by enabling their interoperability in a flexible and distributed manner. As a result, it allows scholars to fully leverage the potential of material digitization and digital research tools without re-creating silos of resources in the digital realm. We believe that RISE, coupled with developing novel licensing models suited for digital research methods (including consumptive research like text mining), would significantly improve the infrastructure behind DH scholarly research in sinology and beyond.

Keywords

digital humanities, sinology, cyberinfrastructure

1. Introduction

Digital humanities (DH) is a burgeoning field of research in sinology and Asian studies more broadly. DH research techniques, including various databases from digitization efforts and growing numbers of digital research tools, have had an impact on Sinologist research communities globally. Stanford University, for example, has held an annual "Digital Humanities Asia" conference since 2016[1], and this venue is the ninth International Conference of Digital Archives and Digital Humanities. There are also collaborations across multiple regions, as evidenced by the enthusiastic participants at the International Conference on Cyberinfrastructure for Historical China Studies this March at Harvard Center Shanghai.[2] Such events demonstrate the diversity and maturity of DH in sinology globally.

Outside of sinology, DH has been grappling with issues such as long-term sustainability and interoperability. In response, many have proposed that DH needs basic infrastructures behind research projects to ensure its long-term success. In Europe, for instance, CLARIN[3] and DARIAH[4] are two such large-scale research infrastructures for humanities. While they have done a tremendous job in centralizing available digital resources, much of their infrastructures remain at the administrative level, and their generic coverage across the entire humanities meant that their utility for a specific discipline like sinology is limited. How can we, as DH scholars and Sinologists, design a cyberinfrastructure fit for our specific needs, taking past experiences with these large-scale infrastructural projects into consideration?

"RISE" (formerly known as "Asia Network") is our answer to this question.[5] It is a pioneering approach for resource dissemination and emerging data analytics (such as text mining and other fair-use, consumptive research techniques) in the humanities. It is a language-agnostic software that facilitates the secure linkage between third-party research tools to different third-party textual collections (both licensed and open-access ones) via application programming interfaces (APIs). It revolutionizes how scholars can work with textual sources because, under the current condition, it is impossible for scholars to use digital research tools to analyze licensed textual collections without downloading or scraping the full texts, which violates licensing terms. The RISE software can securely pass through these licensed texts to digital research tools, thus allowing scholars to work in a legal manner and ensuring commercial publishers the safety of their collections. Such flexible, networked approach to e-infrastructure development avoids re-creating silos of resources in the digital realm and allow scholars to fully leverage the potential of material digitization and digital research tools. Crucially, RISE is a loosely-coupled software with flexible topologies; it can enable both federated or centralized linkages, and it can even "disappear" as long as its API standards remain in place to facilitate communications among databases and tools in the back-

[1] http://dhasia.org/
[2] https://projects.iq.harvard.edu/cbdb/international-conference-cyberinfrastructure-historical-china-studies
[3] https://www.clarin.eu/
[4] https://www.dariah.eu/
[5] See https://asia-network.mpiwg-berlin.mpg.de/ for the web user interface for RISE's beta prototype.

end. Thus, unlike large-scale infrastructural projects, RISE actively lowers the profile of centralized infrastructure and instead promotes existing tools and resources by enabling their interoperability.

"RISE" stands for Research Infrastructure for the Study of Eurasia, and this name retains a degree of regional specificity even though RISE's technical set-up works with resources and tools in all languages. RISE's core developers have backgrounds in sinology and DH, and the ideas behind it grew out of disciplinary challenges there (especially with text mining and licensed textual resources). RISE's functionalities, however, address common infrastructural issues across the DH landscape regardless of disciplines, and RISE's development has also expanded to include multilingual resources and tools that pilot users at our institute work with in their research.

Since our inception in May 2017, RISE has progressed to the beta development stage and we plan to release it publicly by the end of this year. At the time of writing, RISE is linked via APIs to the following resources: Chinese Buddhist Electronic Texts (CBETA)[6], the Taiwan History Digital Library[7], the Kanseki Repository[8] (Kanripo), the Chinese Text Project (CText)[9], a small set of Staatsbibliothek zu Berlin's classical Chinese collections, and Perseus Digital Library (open-access Greek and Latin materials).[10] The only linked research tool is MARKUS[11], though DocuSky[12], Recogito[13], LERA[14] and other tools are on our immediate development horizon. It is important to note that RISE's current linked resources span the entire spectrum in terms of license and copyright restrictions. Some, like Kanripo and Perseus, are completely open-access. Others, like the Chinese Text Project, are generally open-access but require a license subscription for advanced functionalities. And then there are proprietary resources, whose licenses (regardless of read-only or text-mining) are very expensive to acquire. We are committed to the principle of open access, but we also recognize that it is our current reality, perhaps more so in sinology, that many resources are held in private hands. While RISE alone cannot solve this issue (and licensing restrictions are also not the main focus of this paper), we believe RISE's technical set-up provides a useful alternative for scholars to work with resources across various licensing restrictions and could induce some private publishers and database vendors to implement these technical standards.

Here in this paper, we outline (1) the current landscape of DH in sinology and what we see as the main challenges it presents to researchers; (2) a basic summary of RISE's functions and

[6] http://www.cbeta.org/
[7] http://thdl.ntu.edu.tw/index.html
[8] https://www.kanripo.org/
[9] https://ctext.org/
[10] http://www.perseus.tufts.edu/hopper/
[11] https://dh.chinese-empires.eu/markus/beta/
[12] https://docusky.digital.ntu.edu.tw/DocuSky/ds-01.home.html
[13] https://recogito.pelagios.org/
[14] https://lera.uzi.uni-halle.de/?lang=en

features design; and (3) a call for collaborators to develop common API and metadata standards for DH in sinology.

2. Landscape of DH in sinology

Digitization of historical materials has dramatically transformed how Sinologists gather research sources and approach research questions. As more and more archives, libraries, and other research institutions embrace digital technologies, DH-focused projects and initiatives in sinology would only continue to proliferate. This growth, however, cannot be assumed as a foregone conclusion, as the current landscape of DH in sinology is incredibly fractured and already presents many roadblocks to seamless access and sustainability. In this section, we briefly survey this fractured landscape and show how our RISE infrastructure bridges the fault lines within it.

The current landscape is fractured both geographically and thematically. It is not a hyperbole to say that sinology is a global discipline today, as China studies departments and research centers exist in many countries. However, a core-periphery relationship among the roles and foci of these nodes of global sinology research community persists. In Mainland China, where much of the primary sources still reside in various archives and institutions, the commercialization of resource digitization reigns supreme. Despite the fact that many of these sources originate in the public domain (or have long passed their copyright protection terms), commercial publishers build proprietary databases from the digitization and charges high royalties for access. Such commercial models are being copied by university libraries and presses as well. While open-access movements have been gaining steam in recent years and the Chinese DH community has grown dramatically, as seen in the third DH symposium at Peking University[15], this sector has continued to rely on this commercial model and shows little signs of movement. In particular, subscription access often does not include provisions for text mining, full-text access, and other standard DH techniques today, as the pricing model still prioritizes read-only access. It should be noted that we are not advocating for eradicating commercial database vendors from this landscape; rather, it is to point out that their existing business model (and the accompanying lack of technical improvements) make it difficult to leverage the full potential of their digitized materials, even for researchers who have subscription access to their materials.

In Taiwan, DH's first strong foothold in the Chinese-speaking world, intersecting scholarly expertise in humanities and computer science resulted in a strong environment for the development of research databases and research tools development. Many of these databases on driven by thematic interests, such as collections of Buddhist texts or Taiwanese historical documents, just to name a few. Also of note is the DocuSky platform developed by the National Taiwan University, which allows individual users to organize and work on their own set of materials in one place. While there are certainly monetizing tendencies in Taiwan as well, it is notable that many of these databases and tools are built on open-access principles

[15] https://www.lib.pku.edu.cn/portal/cn/news/0000001622

and, if there is a fee, it is usually only for high-usage clients and/or to cover basic maintenance costs. These practices encourage development and sharing of new tools and resources, and Taiwan hosted the first Asia-focused DH international conference in 2009.[16] Since then, the International Conference of Digital Archives and Digital Humanities (DADH) has become an annual event, and the Taiwanese Association for Digital Humanities became a constituent organization of the global Alliance of Digital Humanities Organizations in 2018.[17]

Elsewhere in the world, especially in North America and Europe, thematic DH research projects dominate the landscape. Close alliances between scholars and librarians, aided by international funding bodies like the Andrew W. Mellon, Chiang Ching-kuo, and Luce Foundations create diverse projects that largely consume primary sources for producing research results and presentations. Nonetheless, long-term preservation and sustainability remain serious issues, as is the linking of individual silos of project repositories and making them interoperable. In the context of sinology-focused DH projects, Harvard's China Biographical Database (CBDB) exemplifies both the success and the pitfalls, as its continued growth over almost two decades of open-access development was threatened by funding uncertainty and, just this year, sold its distribution rights in Mainland China to a commercial publisher.[18] In Europe, centralized governmental funding agencies like the European Research Council provides more stability, but similar stories exist as well. Nonetheless, many top research projects and tools (such as MARKUS, Ten Thousand Rooms[19], and Ming Qing Women's Writings[20]) come from Sinologists based in North America and Europe and continue to base at libraries and research institutions there.

So, how should an individual researcher in sinology approach this fractured landscape? If one is interested in primary sources, there are many individual databases that must be searched one by one. While there are open-access, full-text databases like the Kanseki Repository, the majority remain proprietary and read-only, making the usage of digital research tools on full texts incredibly difficult, as well as synthesis of sources from mixed copyright origins. In the rare cases where digital analyses and manipulation are possible via tools like MARKUS or LoGaRT[21], sharing of results is challenging tool. If one is interested in integrating research products from many DH projects, many employ static silos of project websites or repositories without appropriate technical linkages that enable interoperability. It is heartening that DH in sinology has progressed to such a point where a critical mass of research and researchers meant that we must consider cyberinfrastructure, and our "RISE" is a proposed solution that bridges these complex fault lines in this fractured global landscape.

3. Primary issues to be addressed

[16] http://www.dadh-record.digital.ntu.edu.tw/Scope.php?LangType=en&His=D09k2
[17] https://www.adho.org/announcements/2017/adho-welcomes-new-organizations-0
[18] https://projects.iq.harvard.edu/cbdb
[19] https://tenthousandrooms.yale.edu/
[20] http://digital.library.mcgill.ca/mingqing/
[21] https://www.mpiwg-berlin.mpg.de/research/projects/logart-local-gazetteers-research-tools

The current landscape of DH in sinology and beyond, in our opinion, severely restricts scholars' ability to conduct digital research. To summarize, we identified three primary issues that any basic cyberinfrastructure must address.

3.1. Unconnected resource silos

Many research resources in DH are thematic, and some have digital research tools specifically developed to match or fit particular texts. While this research method may be the most efficient way to address a particular research question, at a large scale it re-creates the same fractured, unconnected silos of resources from the physical world to cyberspace. For example, many museums and cultural heritage collections have digitized their materials and put them online. But even in cases where there are common metadata standards, there often still is not for resources to be accessed or shared across multiple collections. In effect, we have buckets and buckets of valuable resources that could only be accessed by scholars one by one.

3.2. Heterogeneous access and exchange data formats

We certainly do not pretend to be the only ones to recognize the previous issue, and many attempts have been made to develop common metadata formats to aid resource access and exchange. For texts, TEI has been one such popular format[22], and more recently IIIF has become widely used for images.[23] Since resources necessary for scholarly research are often held in different locations (both physically and digitally), such formats have greatly improved compatibility with a variety of databases and file systems. The introduction of these formats and their derivatives, however, is not fully exempt from the issue they are trying to solve, however. Multiplication of sub-formats and metadata fields developed for specific disciplines continue to proliferate. While we are certainly not advocating for a 'one-size fits all' approach to developing metadata standards, we believe that data formats for basic cyberinfrastructure must be sufficiently generic and flexible to enable common technical work.

3.3. Data import into (browser-based) digital research tools

For DH-focused sinologists, the tools they use are predominantly text-based. While some do use command line to code their analyses, most use some form of pre-made (browser-based) digital research tools for textual analysis. This research methodology necessitates importing (or uploading) the texts into the research tool itself, and this innocuous act of loading the text actually involves a number of thorny issues.

Our survey of various research tools available for DH research made apparent the fact that importing textual data is very complicated. Common tools such as MARKUS or Recogito provide the user with options to upload some text file or to 'cut and paste' into a textbox available on the browser. This methodology has several drawbacks, however. The scholars

[22] http://www.tei-c.org/
[23] https://iiif.io/

often must manually enter metadata of the text they upload or import into the tools. This input method is extremely problematic when working with large bodies of text. In cases where the texts are copyrighted, this method is next to impossible *even when* the user has a text-mining license. To circumvent copyright protections, many scholars simply do what they must in order to complete their analyses; that is, they do so illegally. In select situations, license holders permit the use of particular research tools in a closed environment by installing local copies of the tool and the texts together, such as the premises of a particular institutional library. In any case, these are imperfect (and occasionally not strictly legal) solutions to a serious problem to DH research.

Licensing issues are complex and require stakeholders beyond researchers and librarians to address comprehensively. We recognize that new licensing models must be developed to fit the cutting-edge DH research scholars are conducting every day, even though licensing issues are not the main focus of this paper. Instead, we think from our sinology-specific situation, where the proportion of copyrighted or protected texts is relatively high and sold at a high price, to develop technical solutions that could 'bridge the gap' until more comprehensive licensing solutions could be implemented. Indeed, working with current licensing restrictions in a legal manner is one of the main impetuses behind RISE.

4. "RISE": a basic cyberinfrastructure for DH research in sinology

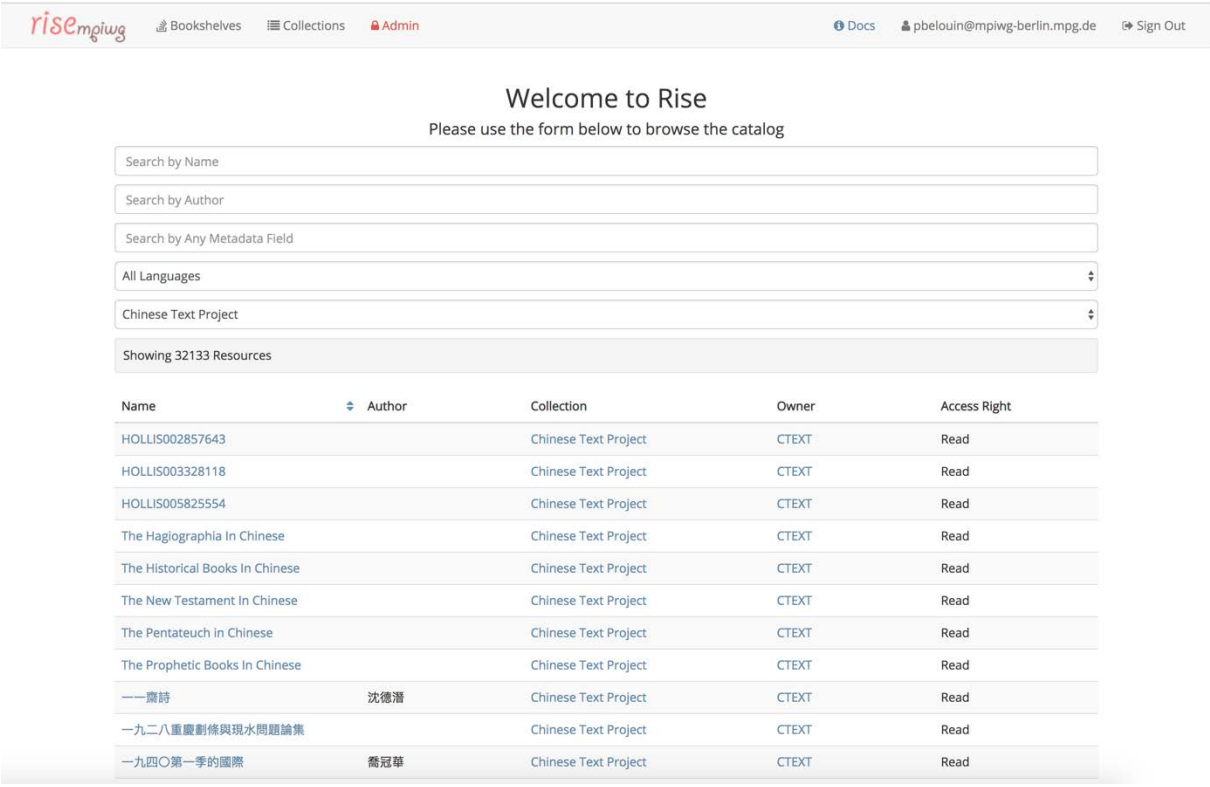


Figure 1. RISE's browser landing page

RISE is an infrastructure for DH projects (e.g., databases, tools, research platforms) to link with one another. Based on a collaborative process involving experienced stakeholders like

system developers, DH researchers in sinology, and librarians, our goal is to produce a general, reusable APIs that can cover common DH projects activities, such as log-in mechanism, contents discovery, tools discovery, contents and tools matching, and personal online research workspace (linked to researchers' individual storage). While there is a front-end web user interface that we have developed in-house, it is not essential for the API-linked ecosystem to function. This loosely-coupled infrastructural design and its flexible topologies are RISE's distinguishing feature from other large-scale, centralized research infrastructures. RISE's main instance is a codebase built on Ruby on Rails, and Figure 2 describes its primary components.

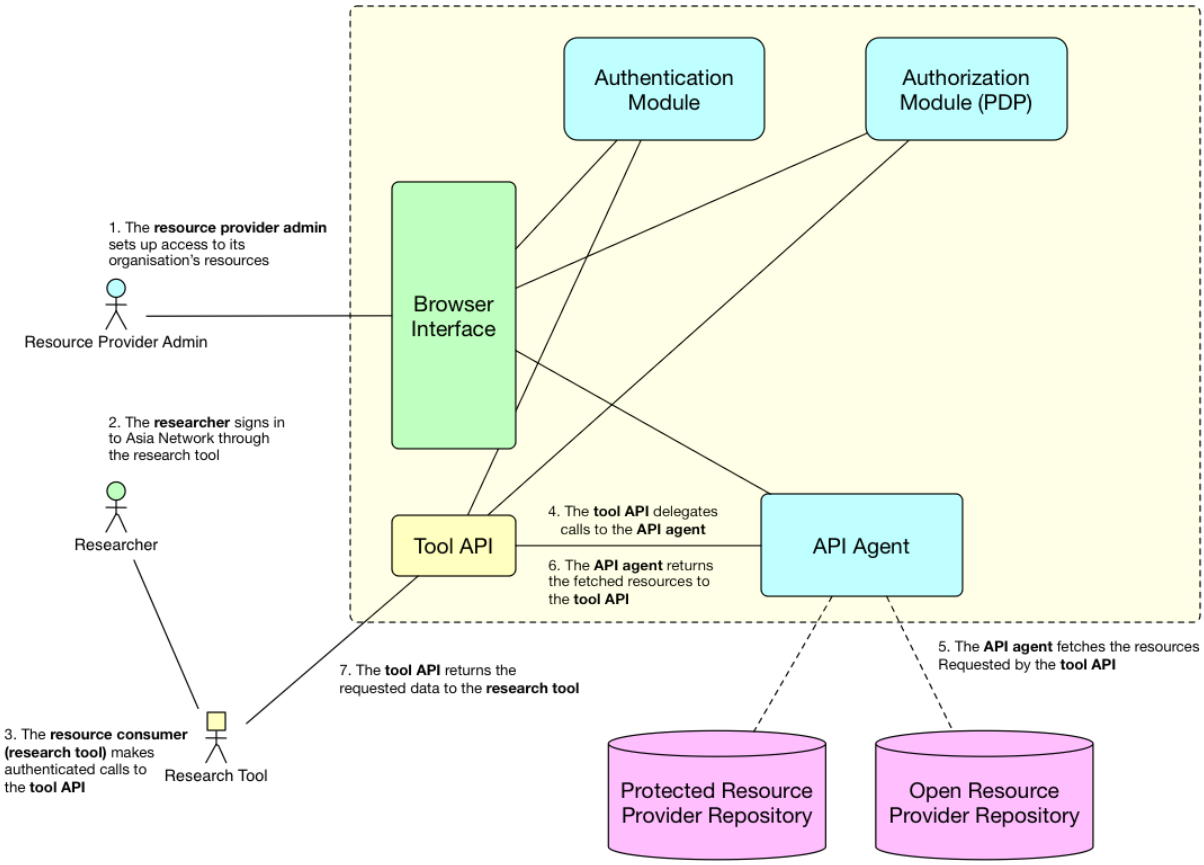


Figure 2. RISE's architecture overview

Despite its name, the RISE software can handle resources in all languages. Many projects and infrastructures have proposed similar ideas (including many European Commission-funded e-infrastructures), creating complex new initiatives like CLARIN and DARIAH. Ours, by comparison, is a modular solution that works, adapting and growing with research projects. Research and structural design remain intimately connected. This demonstrates the significant returns from our early investment into DH research in sinology.

The flexible topologies based on APIs enable diverse DH tools and contents development because they allow decentralized, role-based collaborative growth; said more simply, as long as individual stakeholders implement common API standards, everyone can just focus on their specific tasks knowing that results will be interoperable. This allows each DH project to focus on its own critical, unique contributions. At the same time, individual researchers can

still speed up their research using their existing research tools to search across multiple databases without necessarily going to a centralized portal. This workflow based on APIs apply to DH projects based on computational (rather than text-based) methodologies as well. The RISE APIs are designed to be flexible for different topologies (*ad hoc*, centralized, federated/distributed). In all cases, RISE maintains critical activities (i.e., audit, security check, transaction monitoring and encryption) for interactions among licensed materials and research tools.

The current API definition was designed to connect resource providers and tools developers. It requires resource provider to implement a minimum number of API endpoints in order to make their resources indexable and reachable through the RISE middleware. Resources can be protected by requiring the API client to provide a RISE-API-TOKEN authorization header. This allows resource providers to limit and monitor access to protected resources on a per-affiliation basis. The API allowing research tools to connect to the resources provided by the RISE middleware is similar to the resource provider API, but provides a number of extra features. These two sets of API endpoints adhere loosely to the REST standard and follows the data model presented in the following parts of this document. Below are a few examples of RISE API endpoints, and those who are interested in implementing our standard can find the full list online.[24]

[24] See all three tabs at https://asia-network.mpiwg-berlin.mpg.de/pages/doc_for_resource_providers for our current API definition.

[root]/collections/
Lists the collections available to the client

[root]/collections/[collection_uuid]/resources
Lists the resources that belong to a particular collection

[root]/resources/[resource_uuid]/sections
Lists the sections that belong to a particular resource

[root]/resources/[resource_uuid]/metadata
Returns the metadata for a particular resource

[root]/sections/[section_uuid]/content_units
Lists the content units that belong to a particular section

In an *ad hoc* topology, resource provider and tool developer must directly interact with each other's APIs. However, in most general cases (centralized and federated/distributed), RISE functions as a hub to maintain the most up-to-date API standards and to facilitate interactions among hooked-up resources and tools. Audit and other authorization actions are done via RISE's web interface as well. This infrastructure enables speedy back-end integration among existing DH resources and tools and does not reinvent the wheel at a large scale. It also enables researchers to freely manipulate and analyze resources they have access to in different research tools without violating licensing terms.

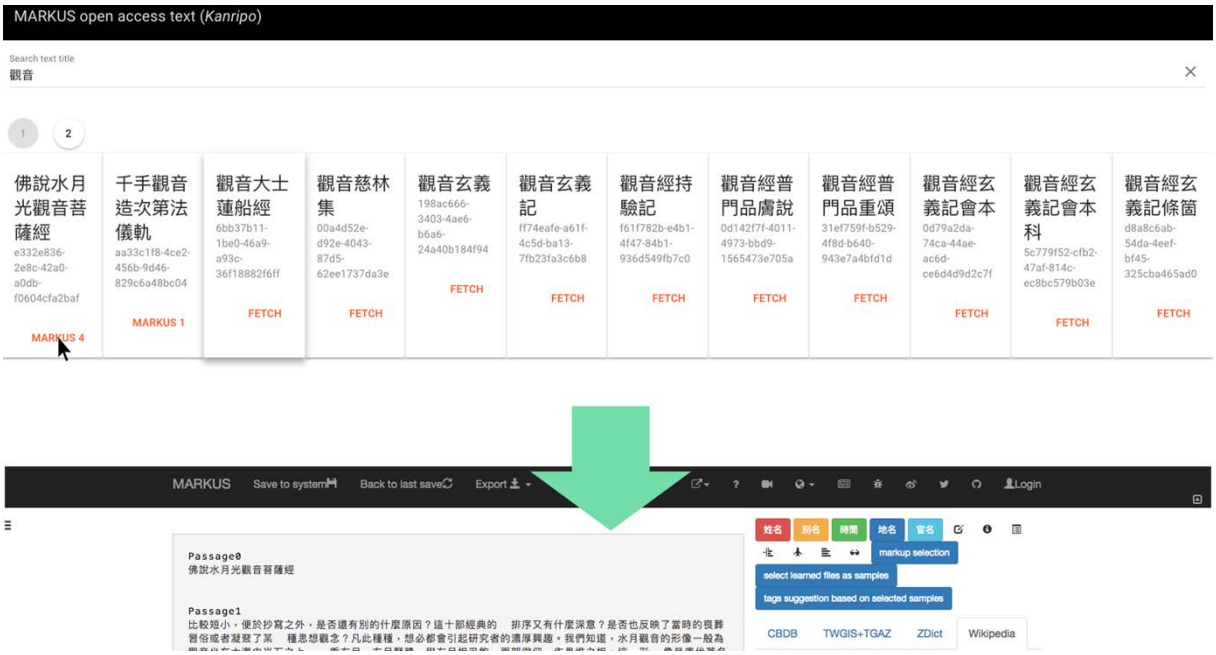


Figure 3. An example of direct link between resources and tools

4.1. Main concepts

To successfully design and implement the RISE software, we defined the following main concepts within the domain following requirements elicitation. Here we list their technical definitions by group. Furthermore, these main concepts are illustrated in Figure 4.

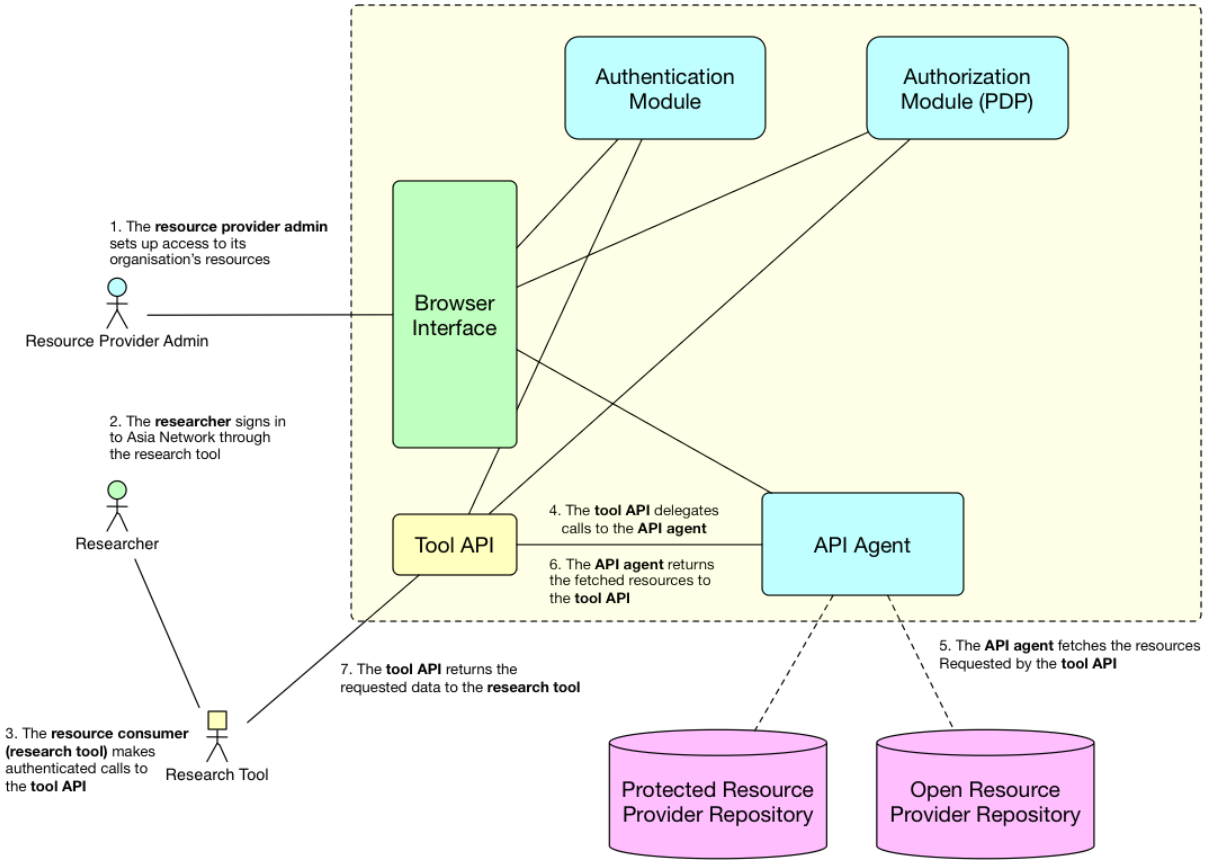


Figure 4. RISE's main concepts

4.1.1. Users

A user is a physical person who interacts with the RISE architecture, either through the browser-based interface or through the use of a research tool (in the case of the ‘researcher’ role). Authenticated users can manage their accounts via the browser interface and access resources according to the access rights defined in RISE’s authorization scheme. There are three possible user roles within RISE.

4.1.1.1. Researcher

Researchers are the default user role in the RISE architecture. Researchers can access the resources they have access to through RISE’s authorization scheme. They can also use research tools made available by the research organization they belong to.

4.1.1.2. Research Organization Administrator

Research Organization Administrators of research organizations can manage the collections and resources their research organization subscribes to via the browser interface. They can also manage the pool of users affiliated with their own organization and monitor their activities.

4.1.1.3. Resource Provider Administrator

Resource Provider Administrators moderate access to the resources their organization provides access through RISE’s browser-based interface.

4.1.2. Organizations

Although using RISE as a middleware by unaffiliated and even unauthenticated users is possible, the authorization mechanism relies on the fact that users and the resources they access belong to certain organizations. Therefore, there are also three types of organizations within RISE.

4.1.2.1. Research Organization

Research organizations have affiliated users. Their administrators must configure resource access and manage affiliated users via the browser interface. Research Organization Administrators also must make sure that resource access rights are kept up to date within RISE’s authorization mechanism.

4.1.2.2. Resource Consumer

A Resource Consumer is a software entity that consumes RISE-compatible resources by calling a RISE-compatible API, such as typically a research tool. However, we foresee that certain resource consumers such as NLP parsers would also make use of the resources provided by the RISE API and generate output in a certain format, which could then be in its turn consumed by another resource consumer.

4.1.2.3. Resource Provider Organization

A Resource Provider is an organization that makes protected or open access resources available to the RISE middleware or RISE-compatible resource consumers through a set of API endpoints. The RISE middleware software is capable of digesting and converting non-RISE-API-compatible API endpoints through the use of custom API mapping modules, which are components of the RISE middleware codebase.

4.1.3. Domain objects

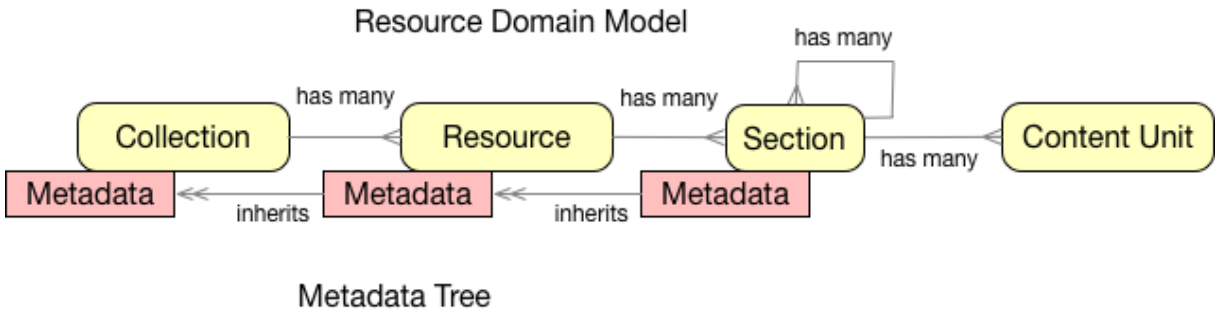


Figure 5. RISE's resource domain model

Texts and other resources accessed via RISE are modeled in a generic and flexible format to account for the heterogeneous types of resources and metadata standards provided by different resource providers. RISE's format is built in a hierarchical model and can be customized to fit different resource providers. It includes the following components below.

4.1.3.1. Collections

Collections are, as their name indicates, collections of resources. Access to these collections can be moderated by their owners on an organization-to-organization basis, so as to mirror licensing agreements made between institutions.

4.1.3.2. Resources

Resources represent items that can be accessed through the system (e.g. text, images or scanned pdfs, tables). These resources can be accessed through a unique uniform resource identifier by users according to access rights implemented in RISE's middleware. It is important to note that a particular resource (for example, a book) can be represented —and

therefore accessed— in various formats. For example, a resource's content can be made available both in computer-readable text format or as images (page scans).

4.1.3.3. Sections

Sections are the part of the RISE resource domain model that are used to represent the hierarchical structure often found in resources such as books in the form of chapters, subchapters, etc. The meaning of different section levels for a particular resource is expressed as part of the metadata tree.

4.1.3.4. Content units

Content units are the units of text and form the base layer of our resource domain model. In practice, a content unit can for instance represent a page or a line. What content units represent for a particular resource is expressed as part of the metadata tree.

4.1.3.5. Metadata tree

Metadata is made available by resource providers through their API and is reflected in the RISE resource domain model at the collection, resource and section level. This metadata is inherited through this hierarchical model; however, if a metadata value is set both at the collection and resource levels, the resource-level metadata definition supersedes the definition set at the collection level.

4.2. Progressive software architecture and development

We developed RISE's key features through a series of requirements elicitation with stakeholders, including DH researchers, sinologists, historians, research tool developers, and resource providers. We settled on a prototype middleware solution, and its accompanying browser interface was started in June 2017. As the development progressed, further feedback was gathered from these stakeholders and fed into the development cycle.

The heterogenous nature of resources and resource providers, combined with the fact that some providers do not have the ability nor the incentive to implement RISE's API standard, means that our current middleware solution needs to be flexible enough to adapt to existing API endpoints of resource providers. While this may not be a long-term solution, we now provide custom API mapping modules for select providers such as CText and Perseus.

As we link with more and more resource providers and make necessary alterations to our standard to cater for an ever-growing range of heterogenous resources, we hope that both the network effect and the efficiency of our standard will create a strong enough incentive for resources providers to adopt our API standard.

We see the middleware solution currently being developed as a necessary yet temporary 'placeholder' to facilitate the development of RISE's API standard. Indeed, the ultimate goal of this centralised middleware instance is to eventually vanish, replaced by a comprehensive standard allowing for the seamless integration of resource providers and resource consumers.

4.3. Authentication and access control

Many resource providers require access right management that is very granular, yet robust and simple, to handle access control to their contents. This management requires control not just at the collection level but also at the resource, and even sometimes even section, levels. In order to provide an adequate solution to this problem, our authorization scheme relies on the hierarchical inheritance of access rights across our domain model as illustrated below.

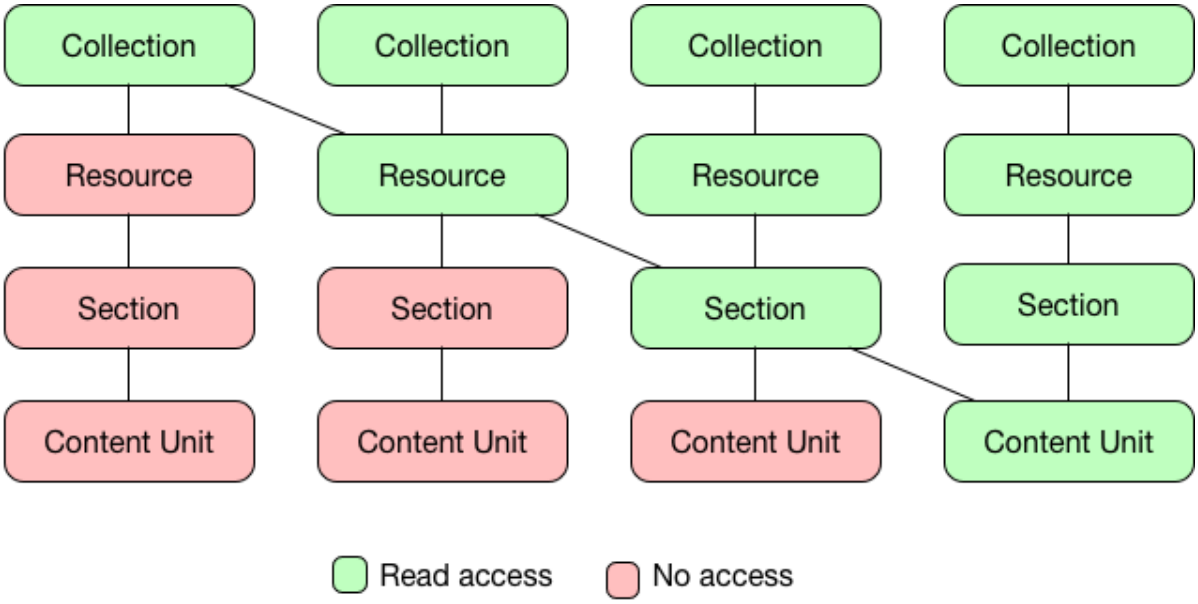


Figure 6. Granular access control

4.4. Dealing with various resource formats

In order to represent the resources linked via RISE as well as their associated metadata, we devised a multi-level data model made available as JSON objects by the middleware through RESTful API endpoints. However, some research tools prefer to access textual resources in richer formats than plain text, such as the ubiquitous TEI format and its derivatives. To cater for this issue, we also make available various formats through our resource REST API endpoints *if* the resource providers provide those formats themselves.

5. Conclusion

In RISE, we have built a middleware solution to shorten the distance between texts and research tools, provided secure linkage to facilitate digital research with licensed texts, and developed a generic yet flexible API standard for exchange between resource consumers

(often research tools) and resource owners that we hope to popularize. At the moment, RISE maintains a clearinghouse of access rights of different resources and institutions so that authentication and authorization could be done properly. In the future, we hope that adoption of RISE would gradually encourage the sinology community – especially the private vendors – to introduce novel licensing mechanisms (e.g., on-demand or consumptive text-mining) for the long-term sustainability of DH research. One of the driving ideas behind RISE is that even though such innovative licensing models do not exist yet – and the RISE team also works in parallel with our collaborators at the Staatsbibliothek zu Berlin on that – the technical mechanisms to implement these new models have now been put in place.

While open-access remains the ideal end goal in any DH endeavors whenever possible, the reality is that many digitized resources in the humanities are still sold by publishers or private vendors. In sinology especially, a fractured landscape as described above creates difficult conditions for DH scholars. We have had to navigate this complex licensing terrain during our everyday work, and RISE is now a prototype primed for transforming DH in sinology. Besides our core collaborators at Leiden University and the Staatsbibliothek zu Berlin, we have now linked up with additional collaborators in the Germany, United States, Taiwan, Japan, Mainland China, Singapore, and beyond. We look forward to launching RISE later this year, and we encourage collaborative development and constructive feedback from all those who wish to contribute to building basic cyberinfrastructure in sinology.

6. Acknowledgments

We thank Dagmar Schäfer, Matthias Kaun, and Hilde de Weerdt for their ongoing work and support for this project. We also thank all of our collaborators. Finally, we thank our student assistant Vitaly Lyapunov for his invaluable work in RISE's ongoing technical development.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

開放工具、數位人文與圖書館的再思考

涂豐恩*　楊麗瑄**

哈佛大學東亞語言與文明學系博士候選人*

哈佛燕京圖書館公共服務及電子资源部主任**

開放工具、數位人文與圖書館的再思考

涂豐恩
博士候選人
哈佛大學東亞語言與文明學系

楊麗瑄
公共服務及電子資源部主任
哈佛燕京圖書館

摘要

自 2000 年以來，「開放近用」（open access）的概念，成為學術界熱烈討論的話題。以科學界為首，不少學者紛紛投入推廣 OA 的運動當中，希望將學術研究的結晶從象牙塔中解放出來，促進知識的流動與傳播。到今天為止，我們已經看到為數可觀的單位，以不同方式支持這項運動。開放近用早期的目標，大多是將學者們的研究論文予以公開上網，便於各界人士取用閱讀，但隨著運動的發展，學者討論的開放近用，已不再只侷限於研究過後的成果，還包括了研究過程中所使用的原始資料以及研究工具。這項觀念上的改變，反映了當代數位科技對於學術研究更為全面性地影響。在本次的報告中，我們將以哈佛燕京圖書館近年來的實作為例，探討大學圖書館在這些轉變中扮演的角色與位置。

傳統上，大學圖書館為服務學者研究上的需求，會負責挑選並向提供商訂購適合的資料庫或相關研究工具。但近年來，在數位典藏與數位人文的影響下，不少學者與學術團體，不在倚賴商業公司，而是自己投身資料庫與工具的建置，因此創造了各種規模不一的數位研究資源。由於數位資源越來越多，研究者也更為需要圖書館提供導航，才不會迷失在茫茫的數位大海之中。

但從圖書館的角度，這樣的趨勢也引發了新的問題：第一，圖書館與資料庫之間不再只是單純的採購關係，很多時候，圖書館也會參與學者的計畫，共同參與資料庫與研究工具的建置。但在這樣的合作關係中，圖書館應該參與的程度以及可以提供的服務，是近年來經常被提出的問題；第二，學者所建置的資料庫或數位工具，雖然經常是免費開放使用，但也因為如此，在品質上與長期維護上，可能無法與商業資料庫相提並論，對於長期面臨經費壓力的圖書館而言，如何在這兩者之間決定資源的配置，成為了一個重要的新問題。

過去幾年，哈佛燕京圖書館參與了兩項哈佛校內學者所推動的數位人文計畫，其一是由包弼德（Peter Bol）領導的中國歷代人物傳記資料庫（CBDB）；其二是由德龍（Donald Sturgeon）一手建構的中國哲學書電子化計劃(Ctext.org)。圖書館所扮演的角色，包括取得建置資料庫與工具所需的資料，協助清理和統整，並釐清資料使用的權利範圍等。在本次的報告中，我們希望首先將分享上述實作的過程與經驗，也包括其中所遭遇的困難，以及尚未解決的問題。從這些實作經驗中，我們希望提出的論點是：圖書館不需要將商業與開放的資料庫或工具，視為非此即彼的對立關係，也不需要期待開放近用可以解決所有的問題；相對地，在多種數位資源並存的環境當中，圖書館可以採行的策略，是藉由積極地鼓勵和協助學者建置開放數位人文工具，同時要求商業公司提供更具有效益的服務，讓兩者相輔相成，創造出一個更理想的數位研究環境。

目次

1. 前言
2. 燕京圖書館及其收藏
3. 持續中的數位化工程
4. 學者的需求
5. 燕京圖書館的實踐
6. 數位人文與圖書館
7. 結語

關鍵詞

數位典藏、數位人文、開放近用、哈佛燕京圖書館

1. 前言

自 2000 年以來，「開放近用」（open access, OA）的概念，成為學術界熱烈討論的話題。以科學界為首，不少學者紛紛投入推廣 OA 的運動當中，希望將學術研究的結晶從象牙塔中解放出來，促進知識的流動與傳播。從 2002 年 2 月的布達佩斯開放近用提議（Budapest Open Access Initiative）到今天為止，我們已經看到為數可觀的單位，以不同方式支持這項運動。哈佛大學在其中也扮演著積極的角色，開放近用運動中的要角 Peter Suber 目前擔任哈佛學術傳播辦公室（Harvard Office for Scholarly Communication）主任，積極將全校學術研究成果予以開放，提供公眾利用，可為例證之一。

開放近用早期的目標，大多是將學者們的研究論文予以公開上網，便於各界人士取用閱讀，但隨著運動的發展，學者討論的開放近用，已不再只侷限於研究過後的成果，還包括了研究過程中所使用的原始資料以及研究工具。這項觀念上的改變，反映了當代數位科技對於學術研究更為全面性地影響。在本次的報告中，我們將以哈佛燕京圖書館近年來的實作為例，探討大學圖書館在這些轉變中扮演的角色與位置。

過去幾年，哈佛燕京圖書館參與了兩項哈佛校內學者所推動的數位人文計畫，其一是由包弼德（Peter Bol）領導的中國歷代人物傳記資料庫（CBDB）；其二是由德龍（Donald Sturgeon）一手建構的中國哲學書電子化計劃(Ctext.org)。圖書館所扮演的角色，包括取得建置資料庫與工具所需的資料，協助清理和統整，並釐清資料使用的權利範圍等。在本次的報告中，我們希望首先將分享上述實作的過程與經驗，也包括其中所遭遇的困難，以及尚未解決的問題。從這些實作經驗中，我們希望提出的論點是：圖書館不需要將商業與開放的資料庫或工具，視為非此即彼的對立關係，也不需要期待開放近用可以解決所有的問題；相對地，在多種數位資源並存的環境當中，圖書館可以採行的策略，是藉由積極地鼓勵和協助學者建置開放數位人文工具，同時要求商業公司提供更具有效益的服務，讓兩者相輔相成，創造出一個更理想的數位研究環境。

2. 燕京圖書館及其收藏

哈佛大學的圖書館系統目前 70 多個大小不同的圖書館組成，哈佛燕京圖書館是其中之一。如果以藏書量來計算，在校內，它目前是全校第三大的圖書館，僅次於總館懷德納圖書館（Widener Library）與法學院圖書館。它也是西方世界最大的東亞圖書館之一，截至 2018 年為止收藏超過 150 萬冊藏書，其中包含中文書 90 多萬冊、日文書 37 萬多冊、韓文書 20 多萬冊，另外還有西文書、以及越南文、藏文、滿文及蒙文等語言。

哈佛燕京圖書館的歷史可以追溯到 1879 年，一位名叫戈鯤化、來自浙江寧波的中國學者，應邀到哈佛大學開設中文課程，他為這門課程購買的書籍，後來留在校內，成為哈佛中文藏書的起點。至於日文藏書的起點，則是 1914 年，兩位東京帝國大學的教

授姉崎正治（Anesaki Masaharu）與服部宇之吉（Hattori Unokichi）到哈佛教授佛學與東方文化，他們所捐贈的一批日本漢學和佛教出版物。此後，哈佛校內對於東亞歷史與文化的研究逐漸增加，這也反應在校內的藏書之上。比如 1921 年知名的語言學家趙元任到哈佛任教，期間就為校內添購了不少圖書。

但哈佛燕京圖書館作為一個獨立的圖書館，則可以從 1928 年算起。當時哈佛燕京學社甫成立，同時建立了一座漢和圖書館，並由裘開明擔任第一任館長，成為了今天哈佛燕京圖書館的前身。當時哈佛燕京學社在美國鋁業大王查爾斯·馬丁·霍爾（Charles Martin Hall）遺產的支持下，得以快速發展藏書的規模，藏書量很快就超過了十萬冊，同時收藏的內容也更為多元豐富，這也有賴許多學者的協助，如知名的漢學家伯希和（Paul Pelliot）與鋼和泰（Alexander von Staël-Holstein），後者精通梵文與藏文，連陳寅恪都曾向他學習。鋼和泰為圖書館添購了藏文的資料，而他生前的許多書信，如今還保留在哈佛燕京圖書館內。圖書館收藏與校內學術研究的進展是分不開的，除了前述的藏文收藏外，後來擔任哈佛教授的著名內亞學者柯立夫（Francis Woodman Cleaves）與弗萊徹（Joseph Fletcher）也為哈佛燕京圖書館添購了滿文與蒙文的收藏。

起初，哈佛燕京圖書館的收藏集中在東亞的文史哲等人文學科，但在二次大戰之後，收藏的視野也逐漸擴充到社會科學方面的書籍，甚至是部分的自然科學書籍。這和美國社會科學界對於東亞研究興趣的增加是密切相關的。美國的福特基金會在 60、70 年代曾經提供大量經費，挹注這方面的研究，而哈佛是受到資助的重點單位之一，不僅在相關課程上有所進展，圖書館的購書經費也因此提昇。自 1965 年就任第二任館長的吳文津先生，對於推動這樣的發展也有著重要貢獻。自 1951 年起，哈佛燕京圖書館也開始收藏韓文書，而越南文的藏書，則要到 1973 年方才開始。

1965 年，原來的哈佛大學漢和圖書館正式改名為哈佛燕京圖書館，另外在 1976 年，並從哈佛燕京學社轉移到哈佛大學圖書館，雖然在經費與組織上仍然與哈佛燕京學社有著密切聯繫，但運作上則成為全校圖書館系統的一環。1998 年，現任鄭炯文館長接任後，除了持續推動館藏量與服務的提升外，同時，如我們以下要介紹的，也大力挹注電子資源的發展，包括館藏的數位化與相關服務。

哈佛燕京圖書館的館藏，除了數量龐大之外，也有許多珍貴或特殊的藏品，如中文善本就有 4000 多部，日文善本書 3000 多種。另外還有許多個人檔案或資料集，著名者如蔣廷黻資料集、洪業檔案、費吳生檔案（George A. and Geraldine Fitch Archives）等，此外還有許多零星的手稿、日記、照片等，其中不少都已經進行內容的編目，可供學者參照與進一步研究。[1]

3. 持續中的數位化工程

[1] 可見 https://guides.library.harvard.edu/c.php?g=310134&p=3847901

十多年前，哈佛燕京圖書館開始了幾個數位化的計畫，將重要館藏轉變為數位格式，以便世界各地的學者訪問和使用，這是屬於哈佛大學圖書館早期數位化工作的一環。早期的幾項計畫多半聚焦在館藏的照片，如莫里士中國老照片, 1933-1946 （Hedda Morrison Photographs of China）與畢敬士中國穆斯林老照片 （Rev. Claude L. Pickens, Jr. Collection on Muslims in China）。前者來自德國出生的攝影師 Hedda Hammer Morrison，他在 1930 與 40 年代居住在北京，透過影像的方式，他記錄了當時中國北方的風景，還有中國人的社會風俗與活動；後者則是 1920、30 年代之間，關於中國穆斯林的照片。畢敬士是一名傳教士，他在戰間期到了中國工作，同時也開始調查中國境內的穆斯林文化，留下了這些珍貴的影像記錄。

而後，哈佛燕京圖書館又開始與各地單位合作，包括中央研究院傅斯年圖書館、北京中國國家圖書館、韓國國家圖書館等，並尋求不同單位的資金，將館藏中日韓文的善本數位化，前前後後分成了 30 個大小不等、各有焦點的計畫。目前已經完成的計畫包括如下[2]：

- 寶卷（Chinese Bao Juan Collection）
- 中文善本方志（Chinese Local Gazetteers- Shan ben fang zhi）
- 中國舊海關資料（Chinese Maritime Collection）
- 中文善本特藏- 經,史部 (中国国家图书馆合作项目) （Chinese Rare Books- Classics & History）
- 中文善本特藏- 叢部（Chinese Rare Books- Collectaner Section）
- 中文善本特藏- 集部（Chinese Rare Books- Collected Works）
- 哈特教授藏書（Chinese Rare Books- Hart Collection ）
- 明清婦女著作(McGill University Library 合作項目)（Chinese Rare Books- Ming-Qing Women's Writings）
- 中文善本特藏- 特大尺寸 （Chinese Rare Books- Oversize ）
- 中文善本特藏- 子部（Chinese Rare Books- Philosophy）
- 韓南教授藏書（Chinese Rare Books- Prof. Hanan's Personal Collection）
- 齊氏兄弟 (齊耀琳, 齊耀珊) 藏書（Chinese Rare Books- Qi Brothers Collection）
- 齊如山藏書（Chinese Rare Books- Qi Rushan Collection）
- 中文善本稀見書目叢刊選目（Chinese Rare Books- Selected Literary Bibliographies）
- 中文善本特藏稿鈔本選輯（Chinese Rare Books- Selected Titles from Unique/ Manuscripts）

[2] 見 https://guides.library.harvard.edu/Chinese

- 中文善本特藏 稿、鈔、孤本 (傅斯年圖書館合作項目) （Chinese Rare Books-Unique/ Manuscripts ）
- 民國時期文獻（Chinese Republican Period Collection ）
- 大學講義（Chinese Republican Peroid Collection- Lecture Notes ）
- 中文拓片（Chinese Rubbings Collection）
- 基督教傳教士文獻（Christianity Collection）
- 大字報（Dazibao and Woodcuts from 1960s China）
- 莫里士中國老照片, 1933-1946（Hedda Morrison Photographs of China ）
- 滿文古籍（Manchu Rare Books ）
- 滿洲國明信片（Manchuria Postcards）
- 蒙文古籍（Mongolian Rare Books ）
- 納西東巴經 （Naxi Manuscripts ）
- 畢敬士中國穆斯林老照片（Rev. Claude L. Pickens, Jr. Collection on Muslims in China）
- 和刻漢籍 （Selected Titles from Japanese Rare Books in Chinese ）
- 韓文善本書（Harvard-Yenching Library Korean rare book digitization project，與奎章閣合作）
- 韓文善本書（Harvard-Yenching Library Korean rare book digitization project ，與 Minsokwon 合作）
- 韓文善本書（Harvard-Yenching Library Korean rare book digitization project，與 National Library of Korea 合作）
- 韓文善本書（Harvard-Yenching Library Korean rare book digitization project，Miyoung Lee and Neil Simpkins 支持）

以上所有數位化的成果，都可以在哈佛大學圖書館的線上目錄中查找，而且是採取 Open Access 的原則，無論是否具有哈佛大學校內人員的身分，都可以自由使用，我們也一直期待、並鼓勵來自世界各地的學者使用這些材料，這是根基於一種信念，也就是唯有這些珍貴的資料被多加利用，他們的價值才能夠更加彰顯。

在多年的努力下，在 2017 年我們宣布哈佛燕京圖書館的中文善本書已經全數數位化完成，其中包括 4000 個書目，共計約 50,000 冊，前後耗時十年，期間曾經資助過這個計畫的，包括蔣經國基金會、北京中國國家圖書館，廣西師範大學出版社，中國社會科學院和浙江大學等。這是一個重要的里程碑，消息在網路上發佈之後，更獲得了中文學界的熱烈響應。該消息不僅在臉書和微信上被大量地分享，同時被網路上的義工翻譯成中文，香港和中國的報紙或網絡媒體也都推出相關的報導。我們稟持著開放的態度，延續著開放近用的基本精神，將學術作為公有財，提供給更多學者與社會上對知識

有興趣的讀者使用。

在中文善本書的數位化完成之後，其他的數位化工作仍在持續進行，包括：

- 中文舊方志（Chinese Local Gazetteer- Jiu fang zhi）
- 中国穆斯林老照片（Carter D. Holton Photos Collection）
- 費正清與賴世和投影片 （John K. Fairbank and Edwin O. Reischauer Lantern Slide Collection）

以及其他的計畫，都會在未來數年中陸續完成。

當然，眾所周知，數位化只是圖書館面對新時代的眾多工作之一，事實上，我們經常聽到學者們抱怨，大量出現的數位資源，反而成為新的問題。的確，在信息超載的世界，研究者就和一般一樣，很容易迷失方向。我們經常遇到的另一個問題，則是這些工具並非針對東亞語言和相關資料設計的，因此時常需要額外的時間對此進行調整。這些新的問題也刺激著圖書館不斷改善他們的服務方向。

4. 學者的需求

在過去的幾年裡，越來越多的東亞研究學者投入數位人文的研究行列，因應著這樣的趨勢，哈佛燕京圖書館開始從數位化進入數位人文與數位學術（digital scholarship）的領域，協助學者在茫茫的數位資源大海中巡航，提供從資料的整理、分析到呈現等各個環節的支持。這是一個長期的工作，而目前哈佛燕京圖書館所做的是一個面對新時代長期轉型的一些起步，這一轉型的核心是繼續為學術界與研究者提供一流的服務，我們相信在數位時代的挑戰，不會歸結為一個或幾個簡單的解決方案，而是需要探索知識生產未來的形式，並作出回應。

我們的想法可以用兩個詞彙來概括：協作和連接。我們希望，圖書館 一方面可以作為連結的門戶（hub），讓對於數位學術研究感興趣的學者和學生可以相互聯繫，並在教學、學習和研究中，找到所需的資源 。另一方面，我們希望通過網路與數位媒體，我們希望跨越國界，擴大哈佛燕京圖書館在東亞研究的國際社群中能夠扮演的角色。我們相信數位人文應該也是全球人文（global humanities）。

2016 年，哈佛大學的東亞語言與文明學系，對其教師和學生進行了數位人文研究的相關調查，並彙整成一份報告。從這份報告中，我們得知，學者們常常覺得他們沒有充分了解他們可以使用的工具，以及這些工具提供的可能性。一些受訪者表示，學習數位人文的技能，是一個孤單的「試錯」（trial-and-error）過程，經常面對不知道投入的時間和精力能夠獲得多少回報的狀況。更值得注意的是，多數研究生在面臨到數位人文研究的相關問題時，並不會第一時間就轉向圖書館尋求協助，或許顯示了在多數人想像

中，圖書館的定位還是相對傳統，與新穎的數位人文似乎有些距離與落差。

隔年，哈佛大學圖書館也就類似的主題進行了更大規模的調查，根據該報告，學者們認為他們需要更多的協助，以便熟悉數位學術研究工具和方法。目前哈佛校內並沒有一個獨特的數位人文中心或數位學術中心，因此服務大多分散在各處，也因此引起一些學者的抱怨。投資報酬率的問題也一再被提起，有時學者會聽說一個有趣例子，但對於實踐過程的理解相對模糊，不確定投資需要多少時間才會帶來收益。或者說收益會是多少。許多人都提到，希望校內能有專家懂得各種技術性的細節，更重要的是能夠整合和消化，並適當地傳達給前來求助的研究者。一位受訪者說：「我不知道什麼是我不知道的，但我不能花三天時間去參加一個工作坊，只為搞清楚這個問題。」 好幾位受訪者則用「一站式服務」來形容他們的需求。此外，特定的工作坊或其他訓練課程，也是需要的。他們也強調，在這些訓練活動中，重要的不只是一些基本工具的描述和介紹，還需提供參加者數位人文背後的「概念框架」。

除此之外，一些學者則強調創造一個數位人文社群的重要性，由於哈佛大學龐大的組織，許多學者或學生發現自己並沒有很多機會了解同事或同行的工作成果。如果圖書館能夠適切地創造空間或創建網絡，用來展示校內各種研究成果，那會十分受歡迎。其他人則希望圖書館能擔負起保管數位人文計畫的責任。讓這些以數位化方式呈現的研究成果能夠永續保存下去。當然，用什麼方式保存，每位學者都有不同的意見。目前有一些學者還是採用自己的伺服器，亦或是向外部的商業伺服器租用（如亞馬遜），這些服務似乎更為靈活，或更容易使用。

當然也有一些學者提出了其他意見，比如傳統學術機構對於數位人文研究的懷疑、數位人文研究能否得到足夠的認可（這將牽涉著年輕的學者、特別是尚未拿到終身聘的學者，是否願意投入這個領域），其他諸如支援網路、資金等問題，也都在這次的調查中被提及。

5. 燕京圖書館的實踐

目前哈佛大學校內儘管有許多成員投入數位人文的研究，但並沒有一個集中的數位人文中心。在這樣的情況下，圖書館相當程度需要回應上述數位人文學者們的需求。近來，哈佛大學圖書館提出了以下策略，以因應數位時代學者們的需求，其中包括了五個方向，各個方向下面又有幾個具體的目標。（如附件）

作為哈佛大學圖書館的一員，哈佛燕京圖書館也依循著這樣大方向，推出或計劃推出多項新服務，都是希望推進東亞研究領域中的數位學術研究。如同本文前面曾經提及的，我們認為圖書館在推動東亞研究的數位學術研究方面，可以發揮更積極的角色。而且可以應該作為一個門戶，讓人們可以相互聯繫，同時圖書館應該鼓勵跨學科、跨國界的合作。2017 年 2 月起，我們啟動了哈佛燕京圖書館論壇（Harvard-Yenching Library

Forum）東亞數位人文系列（East Asian Digital Humanities Series）。該論壇專注於東亞研究的數位學術研究，迄今仍在延續當中。由於哈佛校內沒有為數位人文研究設置一個單一的機構，這個論壇成為了校內少數聚焦於這個議題的場合。

這個論壇的目的是為數位人文學者創建一個交流思想、分享經驗，同時展示各種進行中的計畫的平台。每次我們邀請一名不同背景的學者，針對他們當前的研究計畫進行 20 分鐘的演講，而後開放討論。活動迄今，每一場次都吸引了為數不少的與會者，而且背景十分多元。迄今舉辦過的講座主題包括了：

- Doing East Asian Studies in a Digital Age
- Introducing the Chinese Text Project
- Text Analysis for Literary and Historical Texts in Classical Chinese
- Data Visualization for Piecing and Reading
- Digital Humanist as Historical Detective
- The China Historical Information System—An Introduction
- Do We Need A Professional Certificate in the Digital Humanities?
- Experiments with DH Tools—Representations of Historical Figures in Two Northern Song Examples
- Mapping Unbeaten Tracks—A Digital Analysis of Travel Narratives to Japan, 1860-1920
- Maps, Graffiti, Kinship—The Use of GIS in the Spatial Analysis of a Sacred Mountain in Late Choson Korea (1600-1900)
- Taiwan Biographical Database—An Introduction
- Participatory Engagement with Japanese Digital Resources

涵蓋了廣泛的主題和領域。我們不僅邀請了教師，也邀請了研究生，還有許多來自世界各地的訪問學者。同時，我們也希望兼顧中國，日本和韓國、歷史，文學和視覺研究等不同的領域和範疇，以便人們可以從相關領域中學習和獲取靈感，跨學科的對話往往特別有成效。 我們的想法是：圖書館不再僅是人們查找書籍的空間，它也應該是人們交換意見和發現的地方。

除了系列演講外，我也建置了相應的網路空間，以提供重要的訊息給相關研究者。過去數年哈佛燕京圖書館的館員已經開發了數個線上的研究資源導航，包括：

- Research Guide for Chinese Studies
  （https://guides.library.harvard.edu/Chinese）
- Research Guide for Japanese Studies

（https://guides.library.harvard.edu/c.php?g=310291）

- Research Guide for Korean Studies
  （https://guides.library.harvard.edu/c.php?g=310159）
- Research Guide for East Asian Studies
  （https://guides.library.harvard.edu/EAS）

這些研究導航持續地更新，以協助讀者跟上最新的線上與線下資源。在前述的活動之後，我們則近一步建置了東亞數位人文實驗室（East Asian Digital Humanities Lab, https://guides.library.harvard.edu/EADH ），蒐羅相關的線上資料，並展示各地正在進行的、或已經完成的數位人文的計畫。與過去的研究導航有些不同，在這個新站中，我們根據人文學者一般研究的歷程來組織資源，也就是從搜集和管理資料與數據開始，而後進到分析與呈現等步驟。我們希望持續地擴充這個線上導航系統的內容，讓未來有意進行數位人文研究的學者與學生，都能在此處找到所需的資源或參考的資料。

我們也在這個站上展示了數個內部嘗試的小型數位人文計畫，比如將利用開放的時間軸工具，將莫里士中國老照片中的一些時間與空間資訊予以視覺化；又比如將日本修憲研究計畫所整理的重要時間點，轉繪為可以線上互動的時間軸形式。我們在以上計畫中所使用的工具，大多來自西北大學的 Knight Lab (https://knightlab.northwestern.edu/)，在過去數年內，這個團隊開發了數個開源(open-source)的工具。透過這些示範型計畫的展示，我們希望傳達的是，目前已經有許多免費、開源，而且易於使用的工具，一旦使用得當，可以產生出令人印象深刻的結果。

東亞數位人文實驗室原本定位為一個虛擬的、線上的實驗室，但自 2018 年開始，我們在哈佛燕京圖書館內部也成立了一個數位學術小組(Digital Scholarship Team)，並在圖書館內規劃一個新的空間給小組使用。這顯示了數位時代對於圖書館帶來的影響，並不局限於虛擬世界，而是從線上延伸到了線下，也代表有許多面向的工作都值得再重新思考。

6. 數位人文與圖書館

晚近隨著數位人文的發展，不少人文學者都投入了文本挖掘(text mining)或數據挖掘(data mining)等領域的研究，這些研究必須奠基於大量的數據或資料之上，但對學者而言，如何取得這樣大量的研究資料，本身就是一個挑戰。顯然， 如果不能先獲取到資料，那麼後續的數位人文研究是不可能開展的。如同前面所說，過去包含哈佛燕京圖書館在內的各大小圖書館，花了不少時間與精力將館藏數位化，但我們也意識到這樣的成果，對於數位人文學者而言是還不夠的。

傳統上，大學圖書館為服務學者研究上的需求，會負責挑選並向提供商訂購適合

的資料庫或相關研究工具。但近年來，在數位典藏與數位人文的影響下，不少學者與學術團體，不再倚賴商業公司，而是自己投身資料庫與工具的建置，因此創造了各種規模不一的數位研究資源。

從圖書館的角度，這樣的趨勢帶來的不同的變化：第一，圖書館與資料庫之間不再只是單純的採購關係，很多時候，圖書館也會參與學者的計畫，共同參與資料庫與研究工具的建置。但在這樣的合作關係中，圖書館應該參與的程度以及可以提供的服務，是近年來經常被提出的問題；第二，學者所建置的資料庫或數位工具，雖然經常是免費開放使用，但也因為如此，在品質上與長期維護上，可能無法與商業資料庫相提並論，對於長期面臨經費壓力的圖書館而言，如何在這兩者之間決定資源的配置，成為了一個重要的新問題。

從哈佛燕京圖書館的角度，我們將自己定位為個別的研究人員或計畫，以及內容或工具的提供方(包括商業公司和學術機構)，之間的橋樑。我們意識到，在數位人文的發展中，圖書館仍有一些獨特的、甚至是其他單位難以取代的角色可以扮演。以下我們介紹兩個晚近哈佛燕京圖書館所推行或參與的計畫。

第一個是「中國哲學書電子化計劃」(CText.org) ，這是由 Donald Sturgeon 博士所以一手創辦和建立的的線上開放電子圖書館，其中收錄中國歷代的文獻，根據網站的統計，目前收藏的文本「已超過三萬部著作，並有五十億字之多」。幾年前 Sturgeon 博士來到了哈佛大學工作，哈佛燕京圖書館也有幸開始與他展開合作。我們將數年來陸續完成的五百多萬頁中國古籍數位資料提供給 CText.org，再透過 OCR 的技術，將這些古籍圖像轉化為可以檢索的全文形式。由於古籍版型各不相同，字型也各有差異，為 OCR 的過程增添了許多困難，期間圖書館方曾經與 Sturgeon 博士多次開會討論，希望透過一部分人工的介入，提高全文的品質。同時 Sturgeon 博士也透過技術的改良，提高自動辨識的準確率。如今這個計畫的成果已經彙整進入 CText.org 的網站當中，包括燕京圖書館所提供的影像以及 OCR 全文，儘管 OCR 的結果仍然不能百分之百的準確，但已經足以提升 Ctext.org 的整體內容，也能提供基本的檢索功能。

對於 Ctect.org 的支持，也是表示我們對於開放式資料庫的支持，因為這與哈佛燕京圖書館的原則是一致，我們相信更多的開放可以提供更多學術探索的機會和可能。過去數位人文的研究者經常面臨需要大量資料，但卻無從取得的窘境，圖書館作為重要的內容擁有者，應該可以在這一點上使力。為了在這方面持續努力，從 2018 年開始，我們也與圖書館內的電腦技術人員開始合作，在上述的成果之上更進一步，在校內的目錄系統加入全文檢索之功能。

我們要提及的第二個例子是相當知名的中國人物傳記資料庫(China Biographical Database Project, CBDB)，這同樣是一個開放線上資料庫，其中包含超過 420,000 筆的人物傳記資料以及其人物關係。CBDB 背後的主要推動者是哈佛大學中國史教授包弼德(Peter Bol)，中央研究院歷史語言研究所與北京大學也都參與了此計畫。早期 CBDB 的

資料來自郝若貝(Robert Hartwell)個人蒐集的資料，但隨著資料庫的成長，CBDB 也將蒐集資料的範圍延伸地更廣。CBDB 的資料部分是由人工建立的，但近年來團隊工作人員也開始發展和增加自動化的技術，希望透過電腦科技，自動擷取大型文本中的訊息，作為建置資料庫的素材。

哈佛燕京圖書館的人員近年來也以不同方式參與 CBDB 的發展。顯然，要建置這樣一個龐大的資料庫，背後需要大量的原始資料支持。為了支持這個計畫，圖書館的人員開始與學校內部的法律相關人士合作，並與內容供應商就文本和數據挖掘權利進行討論，我們希望找到一個解決方案，在法律許可的範圍之內，讓學者與內容供應商都能夠同時接受並且從中獲益。對於圖書館和供應商，這實際上是一個非常新的領域，對雙方都是一個挑戰，我們需要來來回回反覆地討論，並確認彼此對於所謂文本和數據挖掘的認知是一致。雙方都需要確定這樣的權利與所訂的價格之間是否合理，並且在合約中加入條款，明文規範。這樣的嘗試，在哈佛內部也算是踏出新的一步。

當然，並非所有的內容供應商都能接受這樣新的合作方式，也曾經碰到供應商對於文本和數據挖掘的理解與學者和圖書館有所不同，因此產生了誤解的情形。但無論如何，圖書館有責任解決這樣的問題，透過持續與內容持有者的協商，確保學者能夠獲得他們需要的研究資源。這是圖書館在知識生產上一向以來扮演的重要角色，未來也將以這樣的方式繼續。

7. 結語

在結束本文之前，我們希望可以再談談「合作」，特別是學者和圖書館員之間的合作。當然，這議題有點老生常談，但我們應該繼續鼓勵和支持開放式的數據庫和工具，不能只是口惠而實不至，而是需要提供實質性和體制性的支持。我們需要建立一個生態系統，鼓勵從事開放資料庫和工具的學者。從前文中我們可以看到，從事這方面工作會面臨的實際問題，而哈佛燕京圖書館希望藉由實質的行動，解決其中的一部分問題，特別是在資料的取用上。

但我們要強調，支持開放的資料庫與工具，並不意味著我們應該拒絕商業資料庫或供應商，這兩者並非對立關係。事實上，哈佛燕京圖書館每年仍然有相當的經費投入在商業資料庫或相關工具的訂購上。但對於開放資料庫的支持，意味著我們對於商業和封閉型的資料庫與工具有著更高的期盼，當免費的工具已經可以解決基本的問題，付費的工具則應該解決更複雜的問題，才能彰顯其價值。過去幾年內，因為商業資料庫價格不斷地上升，已經在圖書館界造成不少討論，我們認為以這樣原則來規劃我們的工作，應該是比較合理的。

同樣的，我們也需要思考長期保存的問題。這是個非常重要的議題，因為數位人文的學術成果比想像中的更為脆弱。一本印刷書完成之後，進入圖書館收藏之後，可以

保存很久，但數位化的資源則非如此。過去，人們倚賴圖書館來保存書籍，也是藉此保存知識，但在未來越來越多研究成果以數位方式呈現時，圖書館是否還扮演這樣的角色，或者應該如何發揮其作用，仍然一個需要多方嘗試和討論的議題。 畢竟，沒有人希望看到他們的作品從世界上消失。

在過去十多年左右的時間裡，哈佛燕京圖書館將其許多稀有而獨特的藏品放在網路上，而且不多加限制，開放全世界訪問。數位化讓來自全世界的使用者，可以在不親自訪問圖書館的情況下使用這些館藏，也讓圖書館在能夠接觸大學校園以外的讀者，這意味著我們的服務範圍已擴展到哈佛以外，我們是與國際的東亞研究社群對話。數位技術帶來的新的發展，也帶來的新的挑戰，敦促著我們尋找新的工作和服務方式。

附錄

Harvard Library Digital Strategy 1.0[3]

Collections & Content

Building collections in support of University-wide research, teaching, and learning.

- GOAL 1 Develop a broad and diverse offering of networked digital resources for the Harvard community.
- GOAL 2 Increase the availability of electronic access to reformatted materials.
- GOAL 3 Enhance support for the selection, acquisition, organization, and access to born-digital materials.

Access and Discovery

Enable effective access to the world of knowledge and data through intuitive discovery, networks of expertise and global collaborations.

- GOAL 1 Provide seamless access to and use of all content purchased, licensed, and created by the library.
- GOAL 2 Enable users to find, access, and use the content acquired and created by the library without intervention.
- GOAL 3 Deliver library content to users regardless of the means of discovery.
- GOAL 4 Facilitate access to digital resources through a discovery platform that supports both novice and advanced researchers.

Research, Teaching, Learning

Deliver innovative and programmatic support for learning and research in partnership with faculty and other researchers.

- GOAL 1 Partner with others inside Harvard and externally to develop a suite of advanced services in support of research, teaching, and learning.
- GOAL 2 Clarify and supplement library support for research data management.
- GOAL 3 Open access to Harvard’s collections in support of global scholarship and public enrichment.
- GOAL 4 Create a rich online environment for instructors and students to use library content in classes and for scholars to work with in their research.

[3] https://dash.harvard.edu/handle/1/34830922

- GOAL 5 Blend the library's physical and digital spaces and make both vital to the research and teaching goals of the university
- GOAL 6 Partner with others inside Harvard and externally to develop a suite of advanced services in support of research, teaching, and learning.
- GOAL 7 Clarify and supplement library support for research data management.

Stewardship

Steward vulnerable and critical research information in partnership with academic and administrative functions across the University and beyond.

- GOAL 1 Ensure sustainability of digital objects through a program of stewardship which embraces international standards.
- GOAL 2 Create and maintain trusted and understood storage and preservation options for Library and University digital assets.

Professional Development

Support a learning organization for library staff to achieve the Harvard University mission.

- GOAL 1 Reallocate staff time from the support of physical materials to support of user engagement with library digital content and tools

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

以國際圖像互通架構為方法的佛教石窟與圖像之數位呈現與閱覽

The IIIF Approach to Digital Representation and View of Buddhist Caves and Images

陳淑君*　王祥安**　凌宇謙***
中央研究院歷史語言研究所助研究員*
中央研究院數位文化中心技術總監**
中央研究院數位文化中心研究助理***

以國際圖像互通架構為方法的佛教石窟與圖像之數位呈現與閱覽

The IIIF Approach to Digital Representation and View of Buddhist Caves and Images

陳淑君	王祥安	凌宇謙
助研究員	技術總監	研究助理
中央研究院歷史語言研究所	中央研究院數位文化中心	中央研究院數位文化中心

摘要

數位人文學，是將數位化或原生數位的材料，運用數位科技方法與工具進行人文研究，當前多少文本的處理為主流，鮮少以圖像材料為基礎的數位人文系統發展。有鑑於此，本研究採用國際圖像互通架構(International Image Interoperability Framework，以下簡稱 IIIF)為理論基礎與技術規範，改善圖像資源囿限於各自機構的系統架構與使用權限，不易於跨機構之間的資源互通，造成數位人文研究時圖像資源被發現、比較、引用、再利用、交換的阻礙等問題。

本研究以佛教石窟與圖像為案例，首先，探索學者研究需求，並以「圖像的空間結構」及「圖像的分析研究」功能為主；其次，運用 IIIF 圖像(Image)、展示(Presentation)、搜尋(Content Search)及驗證(Authentication)等共四套技術規範（API）進行研究實作，並建構數位人文研究環境的圖像研究平台。本研究結果包括：（一）提供圖像檢視功能，諸如：縮放、旋轉、品質與格式，並檢視圖像的細部內容等作為圖像研究的基礎；（二）運用「展示規範」以結構化的方式整合零散、平面的圖像資源，使其有脈絡條理地展示，提供研究者掌握佛教藝術遺跡的空間結構、作品內圖像的空間相對位置、保持作品完整的脈絡資訊；（三）發展標註分類模型，提供學者在研究過程記錄圖像所見特徵，甚至交換、分享並可由學者專家共同標記，並針對特定學術標註進行主題研究，在領域專家協同合作的基礎下展開進階的研究；（四）可與各國研究與典藏機構交換研究材料；（五）以「驗證規範」管理合作研究團隊之間的使用權限，在保護研究材料之際，並讓圖像易於分享及取得，以達更高效率的交換資料。

目次

1. 研究動機
2. 研究設計與實施
2.1. 研究對象
2.2. 研究步驟
2.2.1. 確認、分析佛教藝術史學者研究需求
2.2.2. 解析作品的結構組成
2.2.3. 以瀏覽方式呈現作品空間結構
3. 研究結果與討論
3.1. 國際圖像互通架構（IIIF）各模組的研究實作結果
3.1.1. 圖像檢視與互通規範（Image API）
3.1.1.1. 圖像內容
3.1.1.2. 圖像資訊
3.1.2. 展示規範（Presentation API）
3.1.3. 驗證規範(Authentication API)
3.1.4. 搜尋規範(Search API）
3.2. 系統與 IIIF 瀏覽器(IIIF Viewer)
4. 未來的研究開發
5. 結論
6. 參考文獻

關鍵詞

International Image Interoperability Framework(IIIF)，國際圖像互通架構，佛教藝術，數位人文學。

1. 研究動機

數位人文學，是將數位化或原生數位的材料，運用數位科技方法與工具進行人文研究，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（項潔、涂豐恩，2011；林富士，2017）。以漢學研究為例，如：荷蘭萊登大學的 MARKUS 古籍半自動標記平台(DeWeerdt,2014)；法鼓文理學院的 CBETA 數位研究平台（洪振洲，2016）；台大數位人文研究中心的 Docusky 系統（杜協昌，2016）；國家圖書館的通用型古籍數位人文研究平台（陳志銘，2017）；中央研究院數位文化中心的數位人文研究平台（王祥安，2017）等，皆是致力於發展以文本為基礎的斷詞、標記、分析與視覺化等各種功能的數位系統或平台，以支持數位人文研究。相較而言，當前鮮少以圖像材料為基礎的數位人文系統發展。圖像在人文學研究中常作為歷史證據，可以揭示或暗示不同時期的思想或態度。雖然過去十多年的數位典藏與數位圖書館計畫，將大量的圖像材料數位化並建置資料庫，但是由於圖像資源囿限於各自機構的系統架構與使用權限，不易於跨機構之間的資源互通，造成數位人文研究時圖像資源被發現、比較、引用、再利用、交換的阻礙。

為此，2011 年開始由牛津大學圖書館、大英圖書館及史丹佛大學圖書館共同研究發展、並於 2015 年公布的國際圖像互通架構(International Image Interoperability Framework, 以下簡稱 IIIF)，旨在改善圖像資源的交換、互通與研究功能。IIIF 藉由將各機構圖像資源的開放資料，以及發展具標準化的應用程式介面(application programming interfaces (APIs))等方法，解決典藏系統之間圖像材料無法互通使用的問題。IIIF 核心架構是由四套 APIs 組成，包括：（Ａ）圖像(Image)、（Ｂ）展示(Presentation)、（Ｃ）搜尋(Content Search)及（Ｄ）驗證(Authentication)等模組。此套兼具理論架構、實務協定公開後，歐美及日本的各國學術機構、圖書館、博物館及數位人文學社群已展開各種研究與建構(Snydman, Sanderson, Cramer, 2015; The British Library, 2016)，包括數位人文學也開始以此架構，進行以圖像為基礎的研究(Delmas-Glass, 2016; Kitamoto, 2017)。本研究將以 IIIF 為理論基礎，嘗試以數位圖像材料為對象，結合人文學者的佛教藝術史研究需求，以中央研究院蒐集的佛教石窟與圖像為案例，探索並建構數位人文研究環境的圖像研究平台。

佛教藝術史研究者關注佛教藝術遺跡的圖像分析、時空脈絡與空間結構。其中，作品本身的空間結構與其造像具有重要關係性，董華鋒和寧宇（2010）提出南、北石窟寺中的佛本生、佛傳故事被繪於窟頂四披上，暗示這些造像題材非洞窟主題，圖像的空間分布與佛教藝術作品的主題息息相關。因此研究者需要建立一個保留佛教藝術作品完整脈絡的知識庫，統合整理分散各地、跨越千年的佛教遺跡，建立佛教藝術知識架構，分析作品因地域、時間演變而不同的風格；並解析佛教藝術作品的空間結構，拆解至最細微的圖像或造像單位的同時，亦讓研究者掌握完整遺跡的資訊，同時連結同屬相同作品的不同圖像，了解圖像之間的空間關係。

掌握佛教藝術作品圖像的空間位置後，佛教藝術史學者的下一步是針對作品描繪的圖像進行分析研究，辨識作品的描繪內容，內容闡述作品製作的信徒或僧人的思想與教派，不同教派將依據不同經典為主題雕刻作品；此外，作品圖像樣式不僅僅是宗教涵義，社會風俗同樣影響佛教藝術作品的表現形式，代學明 (2010)於〈須彌山石窟及其價值〉中描述：「高肉髻，面相清瘦，長頸溜肩，具有『秀骨清像』的特點。這種造像風格顯然是北魏孝文帝服飾改制在佛教文化方面的反映。」可知作品的風格內容與社會文化息息相關。在佛教藝術圖像上標註出圖像內容、風格以及描繪手法，方便研究者在數位平台上表示圖像特色以及整理歸納相同內容或風格的圖像，進行深入的分析和探討。

2. 研究設計與實施

2.1. 研究對象

本研究以中國河南省安陽市的小南海石窟為主要對象，為能參照與比較，本文另加入與該石窟圖像主題相關的其他石窟，包括：炳靈寺石窟、雲岡石窟、龍門石窟、麥積山石窟、南響堂山石窟、修定寺塔、大住聖窟，以及大留聖窟等，總計九個石窟（寺塔）、445 張相關的數位圖像（請詳表 1）。圖像的品質和取得方法是圖像研究的關鍵，本研究所使用之圖像為中央研究院歷史語言所顏娟英研究員提供之圖像，為其田野調查時委託攝影師實地拍攝之圖像，並經過精選而成的圖像集。

表 1 小南海石窟與相關主題石窟、寺塔的地點、開鑿年代與圖像數量

石窟題名	地點	開鑿年代（西元年）（出處）	圖像數量 (n=445)
小南海石窟中窟	河南省安陽市	北齊(550-560) (顏娟英, 1995: 563-564)	236
小南海石窟東窟	河南省安陽市	北齊(550-559) (顏娟英, 1995: 563-564)	81
小南海石窟西窟	河南省安陽市	北齊(550-559) （賓陽中洞，無日期）	61
炳靈寺石窟第 169 窟	甘肅省永靖縣	西秦建弘元年(420) (顏娟英, 1995: 580)	1
雲岡石窟第六窟	山西省大同市	北魏 (465-495) (宿白, 1978: 26-27)	1
龍門石窟賓陽中洞	河南省洛陽市	北魏(505-523) (法鼓文理學院, N.D.)	2
麥積山石窟 127 窟	甘肅省天水市	北魏(516-534) (董玉祥, 1983: 22)	1
南響堂山石窟第 1 窟	河北省邯鄲市	北齊(565) (顏娟英, 1991: 334)	1
修定寺塔	河南省安陽縣	唐(618-907) (楊寶順、孫德萱、孫士杰, 1982: 75)	23
大住聖窟	河南省安陽縣	隋 開皇九年(589 年) (河南省古代建築保護研究所, 1988: 1)	25
大留聖窟	河南省安陽縣	東魏武定四年(546 年) (河南省古代建築保護研究所, 1988: 9)	13

小南海石窟分為東中西三窟，建於北齊時代，為當時的佛教核心區域。而中窟又為三窟中保存得最為完善的一窟，由僧方法師創鑿、北齊國師僧稠重修而成，此窟雖小，但四壁浮雕內容精美豐富，且為北朝石窟難得有文字紀錄開鑿過程、刻錄經文表明禪修精神的石窟，並影響後世敦煌莫高窟等石窟，為佛教藝術史研究不可多得的材料。（顏娟英，1995）除此，本文也納入旁及與小南海石窟圖像主題相關的其他石窟，例如與小南海石窟中窟南壁同樣描繪「文殊維摩圖」的雲岡石窟第六窟、與西壁「往生淨土圖」相同主題的南響堂山石窟第一窟等，作為小南海石窟影像的參照與比較對象。

2.2. 研究步驟

2.2.1. 確認、分析佛教藝術史學者研究需求

本研究定期與佛教藝術史學者進行訪談與討論，確認以佛教藝術的時空發展為專題，具體的研究需求為「還原佛教藝術作品現場空間結構」，其包含整個作品的構造，諸如有幾面、幾道牆、幾扇門、幾根柱子等，以及單一細部圖像於空間整體的位置，圖像與圖像之間相對位置關係等。因佛教藝術作品中的圖像位置與作品建造者的思想理念息息相關，不同擺放位置皆有其佛理、學理上的意涵。如小南海石窟、大住聖窟中以盧舍那佛為正壁主尊佛，並在側壁安置彌勒佛與阿彌陀佛，此種尊像配置方式，是華嚴教的思想。囿於目前紙本書籍、電子資料庫的資料多為零碎資訊，不易了解圖像在空間內的位置，更無法理解不同圖像之間的相對位置。因此需要數位人文系統，幫助研究者還原作品的空間結構和圖像之間的相對位置，作為圖像學的基礎研究。

2.2.2. 解析作品的結構組成

為能還原作品現場結構，本研究先解析作品結構組成，釐清整體與部分關係，以及圖像的位置關係。本研究將作品解析為三層次結構，由整體到中間結構再到局部圖像，從巨觀到微觀，既能針對局部區域進行細緻研究，也能追溯其整體脈絡。經分析將作品分為第一層整體（石窟）、第二層中間結構（周壁）以及第三層圖像（壁面圖像作品），如圖 1 所示。

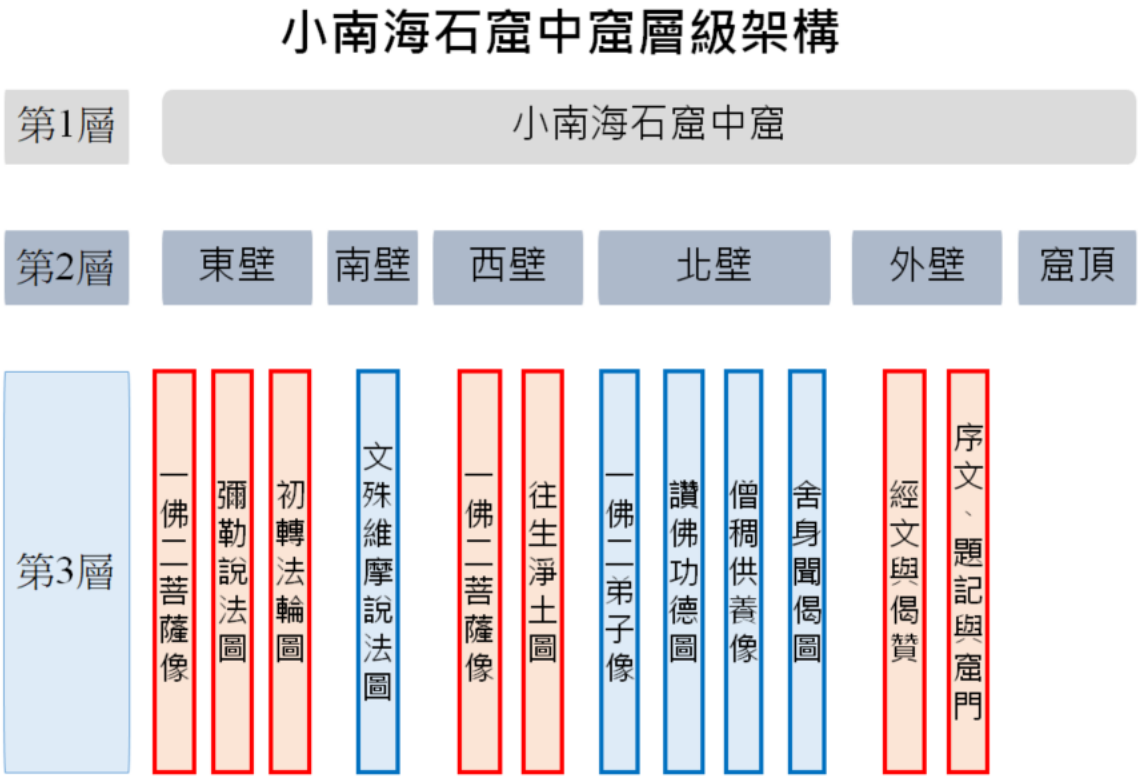


圖 1 小南海石窟層級架構

2.2.3. 以瀏覽方式呈現作品空間結構

本研究以 IIIF 為基礎開發系統，還原藝術作品三層次之間的關係。以層層頁面展示作品之間的關係，並展示作品的實際圖像。作品頁面提供結構座標位置圖，讓使用者明瞭結構分布情況；藉由 IIIF 瀏覽器提供圖像的展示、縮放、旋轉功能，並且利用 IIIF 的圖像結構性質，組合、排列和整合不同層次的圖像照片，使其按照分析後的作品結構典藏圖像。

3. 研究結果與討論

3.1. 國際圖像互通架構（IIIF）各模組的研究實作結果

本研究依 IIIF 的圖像(Image)、展示(Presentation)、搜尋(Content Search)及驗證(Authentication)等標準進行研究與開發，並完成以圖像為基礎的數位人文研究系統雛型，研究結果主要展示作品空間結構、圖像檢視、以及作品的詮釋資料等功能。在「展示作品空間結構」方面，本研究展現一個佛教藝術作品的完整結構，結合搜尋功能和分類檢視功能，供研究者篩選欲檢視的作品，並可連結至單一作品查看作品的詮釋資料（圖 2）。



圖 2 佛教藝術研究系統—展示作品空間結構（以第一層為例）

3.1.1. 圖像檢視與互通規範（Image API）

圖像檢視是佛教藝術史研究者亟需的核心功能，針對藝術作品的雕刻手法、主題、裝飾刻紋、配飾物件等圖像呈現的視覺效果進行分析研究，所有的圖像內容皆有其含義，如小南海石窟中窟西壁北側繪有三叢樹，可能對應至《觀無量壽佛經》的七重寶樹（顏娟英，1995），而非隨意繪製，因此佛教藝術研究者最注重是否能清楚展現圖像中的每一處細節，並針對兩張圖像進行比較和分析，找出兩者的潛在關聯，建立起佛教藝術跨時空背景下形成的區域特色。而圖像本身的所有權資訊亦是學者關切部分，因為圖像的所有權可以保護研究者的研究的權利，防止他人濫用。

在此需求之下，又再細分特定的圖像檢視需求，其中放大檢視是必備的功能，使研究者能清楚檢視圖像的每一處細節；同時須有查看特定範圍的功能，因放大必定導致無法在螢幕中呈現完整圖像；亦須具備縮小觀看圖像功能，才可掌握圖像整體布局；而呈現圖像的原始色彩則幫助研究者如同檢視真實作品。除了圖像內容，圖像本身的資訊也必須清楚展示，說明作品圖像的原始尺寸、所有者資訊。

為滿足上述圖像檢視的研究需求，本文採用 IIIF 的圖像規範（Image API）提供研究者瀏覽圖像的特定區域、縮放檢視、旋轉角度以及圖像色彩的功能，基於圖像可以在分散式的數位環境中進行協同合作與分享。以下，將分別就「圖像內容」、「圖像資訊」兩個部分闡述本文的研究實作與結果。

3.1.1.1. 圖像內容

運用統一資源識別碼（URI）和參數變化的設計，以滿足研究者可依不同範圍（region）、尺寸（size）、角度（rotation）、品質（quality）、格式（format）(Appleby, M., Crane, T., Sanderson, R., Stroop, J., & Warner, S. , 2017b)等不同條件的參數組合（請詳表 2），進行符合條件的圖像之查詢及檢視，並能與其他採用此標準的研究機構交換彼此擁有的圖像資源。符合圖像請求 URI 句法，形式如下：

{scheme}://{server}/{prefix}/{identifier}/{region}/{size}/{rotation}/{quality}.{format}。

表 2 圖像參數與功能列舉

參數	指令功能	意涵
範圍 (region)	square	該區域之寬度與高度皆與完整圖像之較短邊相同，於完整圖像之較長邊的裁切區段由伺服器自行判定，合理預設為置中。
	x,y,w,h	以絕對像素值指定回傳區域。 ● x：由橫軸 0 之位置起算 x 像素值的位置點開始選取。 ● y：由縱軸的位置 0 之位置起算 y 像素值開始選取。 ● w：選取區域的寬度 ● h：該區域的高度。
尺寸 (size)	w,	回傳圖像的寬度等於 w 值，高度則依照擷取區域之原始比例，配合給定的 w 值進行調整。
	pct:n	回傳圖像的寬度與高度皆調整為擷取區域寬度與高度的某個百分比值，回傳圖像的縱橫比例則保持與擷取區域相同。
角度 (rotation)	n	順時針旋轉之度數，n 值範圍為從 0 至 360。
	!n	圖像先經過鏡像翻轉後再依照給定的 n 度進行順時針旋轉。
品質（quality）	color	以全彩色回傳圖像。
	bitonal	回傳圖像為黑白，且其中每個像素若不是黑色即為白色。
格式 (format)	jpg, png	指定回傳圖片之格式。

展示小南海石窟中窟影像的圖像請求 URI（圖 3），改變 URI 參數後的圖像如下所示，並於表 3 中解讀改變 URI 參數造成的圖片呈現效果。

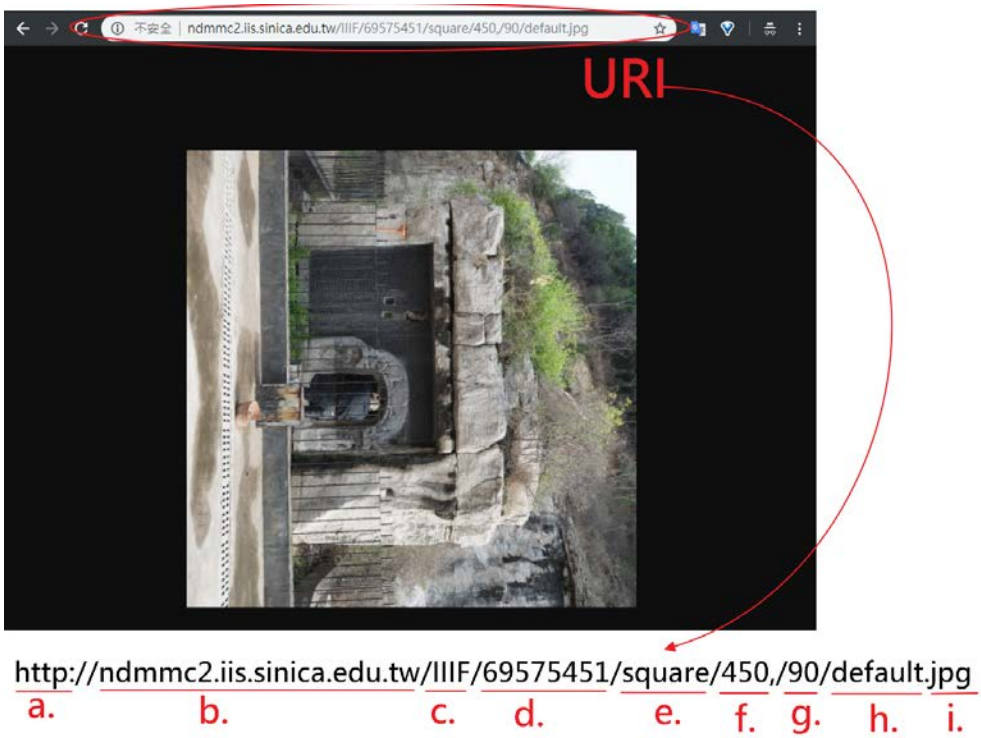


圖 3 以符合圖像請求參數的 URI 呈現圖像，並改變圖像呈現的區域、尺寸、角度

表 3 解讀圖 3 中圖像請求 URI 句法

編碼	實例	URI 組成要項	句法解讀
a	http	scheme	使用超文本傳輸協定(http)
b	ndmmc2.iis.sinica.edu.tw	server	伺服器位置
c	IIIF	prefix	使用 IIIF 作為表示
d	69575451	identifier	該影像的識別碼
e	square	region	圖像請求參數中的範圍，在此例是以 square 呈現，以影像中較短的邊決定另一條邊的長度，將原圖以正方形呈現，並改變 size 尺寸。
f	450	size	以 450 像素呈現(原圖為 5,472 像素)
g	90	rotation	90 度旋轉
h	default	quality	顏色保持原圖的彩色
i	jpg	format	格式為影像原始檔案格式

3.1.1.2 圖像資訊

圖像資訊以 JSON 資料交換語言表示，透過圖像資訊請求 URI，可以獲取圖像本身的原始寬度、高度、提供者等資訊，以及能夠提供圖像的尺寸大小、旋轉角度等規範。此外，針對圖像的使用權利，如：影像所有者，以保障研究材料歸屬權的需求等，將運用權利和許可屬性，說明圖像的擁有者、所屬機構商標、使用許可範圍等三個屬性，呈現所有者資訊。

綜整以上所述，佛教藝術研究者所需的圖像放大、縮小、旋轉、色彩變更等視覺效果皆可藉由 IIIF 圖像 API 呈現，且因圖像按照 IIIF 規範，使得圖像易以高效率方式交換，並且依靠改變圖像請求 URI 參數可以得到相對應範圍、大小、顏色、角度的圖像，配合符合 IIIF 規範的瀏覽器，研究者可以自由地放大與縮小圖像，以及平移放大後的圖像到特定範圍，查看局部或整體內容，這對佛教藝術研究者極有幫助，可觀看整體圖像掌握全局，亦可放大仔細審查細部內容。

3.1.2. 展示規範（Presentation API）

基於佛教藝術作品中的圖像位置與該作品建造者的思想理念息息相關，不同擺放位置皆有其佛理、學理上的意涵。但是現有紙本書籍、電子資料庫的方式不易讓研究者了解圖像在空間內的位置，更無法了解不同圖像之間的相對位置。本研究將佛教藝術作品分為第一層整體、第二層中間結構及第三層圖像共三個層次，解析整部關係以及圖像的位置關係，由整體到中間結構再到局部圖像，從全觀到微觀，既能針對局部區域細緻研究，亦可追溯其整體脈絡。此外，作品的創作時間、地點、創建人等後設資料也是佛教

藝術研究者在分析圖像時的重要依據，因此本研究在呈現圖像之際，也同時提供檢視後設資料。除架構層次的再現，圖像標記也是佛教藝術史學者的重要功能需求，以便在圖像上直接描述作品的特徵、內容、研究筆記和心得。直接於圖像上標註對個人研究以及教學解說皆有極大助益，可直觀展示研究者欲探討之範圍與區域，減少誤解的情況。

本研究以 IIIF 展示規範為架構，提供「結構」功能，經由完整的結構以表現佛教藝術作品圖像，將佛教藝術作品的每個層次數張至數十張不等的圖像，以有條理的方式展示作品的內部結構；除此，本研究也提供「排序」功能，當一個物件有數張不同角度的相片時，研究者可依據一定順序排列，使其由西到東、由上而下等順序逐一展示；最後，本研究提供「標註」功能，研究者得以直接在數位環境進行研究，以及顯示佛教藝術作品資訊。

在實作層面，本文分別以各層次的「Manifest」（整體物件）、「Sequence」（所有單一物件的順序）、「Canvas」（單一物件）、「Content」（數位內容—圖片本身或標註內容）等四種基本資源類型，建構出基本的資源結構，並加上「Collection」（整套藏品）、「Annotation」（標註）、「AnnotationList」（單一物件的標記清單）、「Layer」（標記清單組合」、「Range」（整體物件的內部結構）等五種額外的資源類型(Appleby, M., Crane, T., Sanderson, R., Stroop, J., & Warner, S. , 2017c)，呈現複雜的圖像物件架構（請詳表 4 及圖 4）。

表 4 資源類型與本研究實作情況

Basic Types		**Additional Types**	
Manifest	O	Collection	開發中
Sequence	O	Annotation	O
Canvas	O	AnnotationList	O
Content	O	Layer	開發中
		Range	

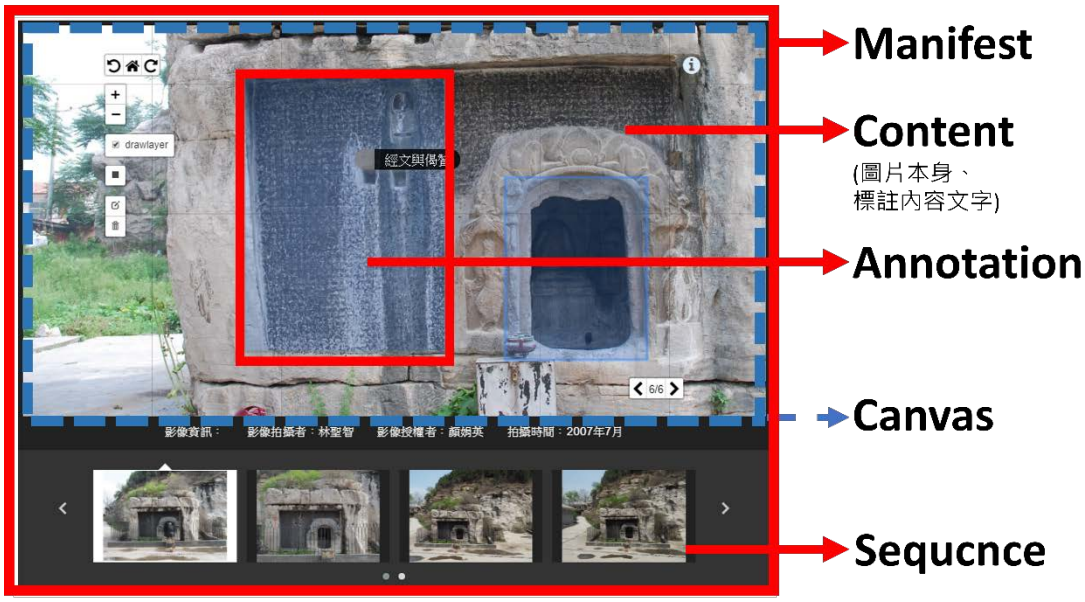


圖 4 IIIF Viewer 中展示的作品物件結構

本研究已開發的「展示功能」可處理包括Manifest、Sequence、Canvas、AnnotationList、Annotation、Content 等六種資源(resource)類型，並持續發展 Collection 和 Layer 的功能。

基礎類型：

1. Manifest：類似於相簿，描述一個物件（例如：小南海石窟中窟）的多張圖像，以及該集合物件的圖像資源、後設資料、圖像標註，是基本類型中的最高層級架構，也是一筆完整後設資料的描述單位。
2. Sequence：規範多張圖像呈現順序，一個 Manifest 中可能有多張影像，利用 Sequence 排列呈現的順序。
3. Canvas：類似畫布，儲存圖像以及標註內容。
4. Content：圖像內容或是標註內容都是 Content，和 Canvas 關聯。

增加類型：

1. Collection：可以包含其他的 Collection 或是 Manifest，形成樹狀結構，用以描述物件更複雜的結構。
2. Annotation：除了基礎的標註功能，為符合研究者對於標註的分類需求，本研究同時以 W3C Web Annotation 為基礎，發展標註分類功能，協助研究者查看特定類型的標註內容，並針對特定學術標註進行主題研究，例如：以佛傳故事為研究核心，檢視事件類型的標註。
    - 首先，每個標註都有其 URI（請詳圖 5-[a]）屬於 oa:Annotation 類別（請詳圖 5-[b]）；

- 其次，將標註進一步細分為標記動機及標記內容。其中，標記動機（請詳圖 5-[c]）可分為三類，包括：標記(tagging)、描述(describing)及評論(commenting)；標註內容（請詳圖 5-[d]）則區分為四類（請詳圖 5-[e]），並採用藝術與建築索引典（AAT）的詞彙表示標註類別，分別為人像/動物(figures)、物件/植物(objects)、地點(Built Environment)和事件(Events)等四類[1]；
- 最後，將標註和 IIIF Canvas（請詳圖 5-[f]）以「on」屬性建立關聯。

此項功能將可提供佛教藝術史學者，在研究過程中隨時記錄圖像所見特徵以及針對圖像的研究感想或分析評論，若交換或分享標註，則可以使其他研究者或是學習者藉由已有的標記直接了解圖像所述內容，促進學習效率以及理解研究者的研究思路。

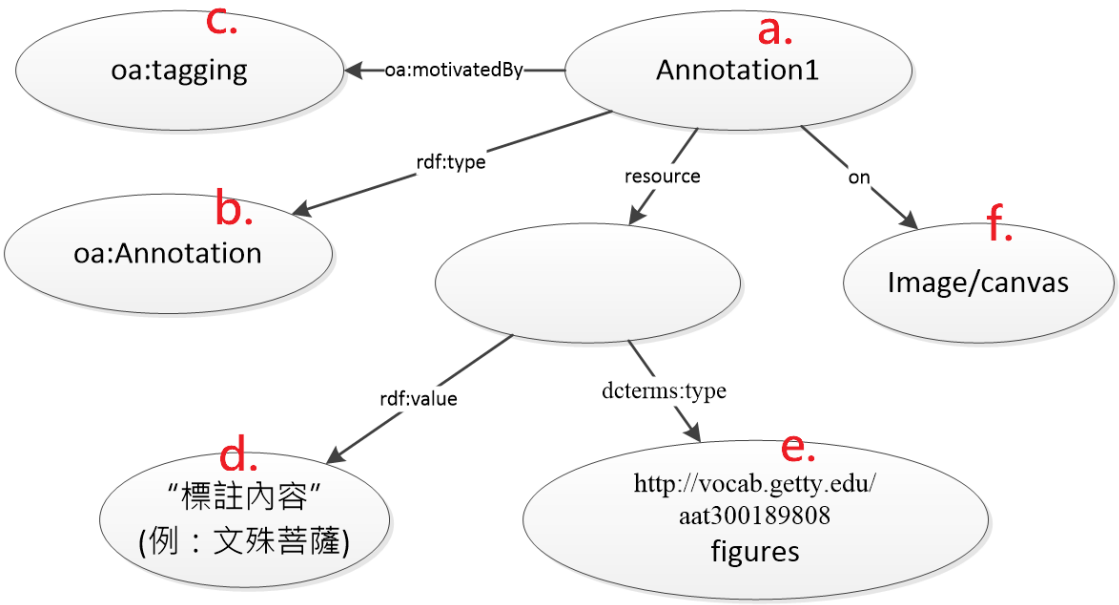


圖 5 以 Web Annotation 延伸 IIIF Annotation

3. Annotation List：可以集合同一個 Canvas 上的多個標註。
4. Layer：本功能尚在開發中，旨在提供不同組合的 Annotation List，這些 Annotation List 可以來自不同 Canvas 的 Annotation List，依靠 Layer 將不同 Canvas、但性質相同的 Annotation 結合在一起，例如：「小南海石窟中窟西壁一佛二菩薩像」有十六張不同角度照片，將這些照片的標註聚集為同一個層（Layer），可作為未來發展為自動分類功能的基礎。

[1] 使用 AAT 索引典中對於 figures、objects、Built Environment、Events 的定義

5. Range：在同一個 Manifest 中可區分不同的章節，例如標示出一本書的封面、書名頁、每個章節。目前本研究案例的層級架構圖如下所示（圖 6），每一個層級皆有一筆後設資料以描述此作品的內容，而每筆後設資料也會對應到一個 Manifest。以「小南海石窟中窟」為例，共有十九筆後設資料，因此總計十九個 Manifest，並以 Collection 整合這些零散的 Manifest。例如：小南海石窟西壁 Collection，包含一個西壁影像（第二層中間結構的影像）Manifest，以及西壁第三層圖像的 Collection，其中第三層圖像的 Collection 又包含兩個 Manifest，每個 Manifest 中的圖像以 Sequence 排序影像的呈現順序。並將實際資料放入 JSON，主要目的是將佛教藝術的資料模型對應到 IIIF 的資源結構。

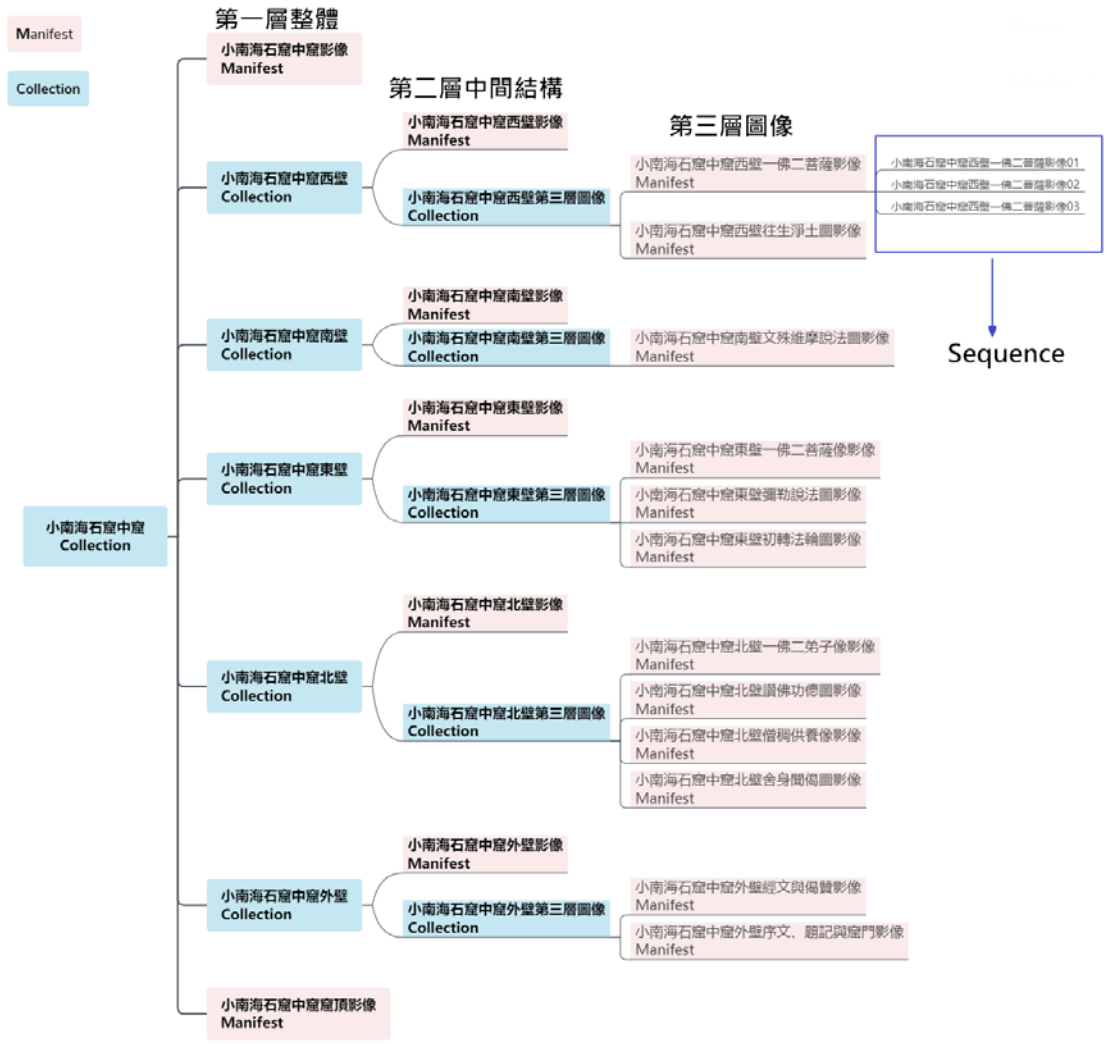


圖 6 佛教藝術 Data Model 對應至 IIIF 的資源結構示意圖

上述，展示規範的「Collection」和「Range」皆可以表示作品的結構，然而兩者差異在於「Range」是在一個 Manifest 下區分章節或部件，「Collection」則是多個 Manifest

和 Collection 的整合。Manifest 展示的是單一的物件，即是說可以做為單一物品討論、提供資訊，在交換圖像或取得圖像 URI，多以 Manifest 的 URI 作為交換，如由史丹佛大學以及哈佛大學共同合作開發的 Mirador[2]瀏覽器，即是由輸入 Manifest URI 建立新的物件。在這樣的架構下，Collection 以 Manifest 為單位，可以分享單一的 Manifest，但 Range 並不是完整的物件單位，只是 Manifest 的一部份，在 Sequence 中跳轉切換。本研究將以一筆後設資料描述的物件作為一個 Manifest，如「小南海石窟中窟(第一層作品)」或是「小南海石窟中窟西壁文殊維摩說法圖(第三層作品)」各自為一 Manifest，其中各有多張影像。

除上述圖像資源的架構描述，本研究並發展一套屬性作為描述資源類型的後設資料，目前已完成三大類型，分別為：描述屬性(Descriptive Properties)、權利和開放屬性(Rights and Licensing Properties)、技術屬性(Technical Properties)。「描述屬性」旨在提供使用者了解物件的背景資訊，諸如：佛教藝術作品的題名、建造時間、地點、建造者等資訊，讓研究者在利用圖像時可以隨時查看相關的作品資訊；「權利和開放屬性」則是標示圖像發布機構的資訊，以及建議標示資源的使用方式，讓圖像的使用者了解此資源的版權、防止濫用；「技術屬性」則是與建立 IIIF 資源結構相關的後設資料。

綜言之，展示規範界定如何呈現資源的結構。佛教藝術作品為立體物件，除了有不同層次與部件，相同部件擁有多張影像，若是混雜一處，研究者將難以系統性的檢視圖像，更遑論進行比較與分析。因此 IIIF 展示規範的建構，將可滿足研究者檢視作品的整體及局部圖像之功能需求，不僅了解作品局部，同時保留整體脈絡，而非破碎的圖像材料。

3.1.3. 驗證規範(Authentication API)

雖然 IIIF 支持資源開放，但考量機構典藏政策、法律規章、商業模式等等限制，可能要求使用者進行身份驗證並獲得與某些機構的授權。以本研究案例而語，由於圖像資源珍貴且多具有版權限制，加之學者於研究中的註記行為，認為只是研究中的思考過程、尚未定論，傾向不願全部公開、分享註記內容，因此需要控管使用者可以瀏覽圖像、圖像註記的權限。

IIIF 提供驗證規範的應用程式介面（API），使用網際網路協定位址（IP）位置限制、協議同意、使用者登入等多種使用者認證方式，以確保各機構或專案能保護其資源影像。在登入範型（Login）、點選連結（Clickthrough）、互動式多媒體資訊站（Kiosk）、外部（External）等四種互動範型（Patterns）(Appleby, M., Crane, T., Sanderson, R., Stroop, J., & Warner, S. , 2017a)中，本研究採用 Login 範型，限制可以進行標註的使用者身分，創立

[2] http://projectmirador.org/

兩種帳號身分，一為研究者，另一為研究助理，兩者沒有標註或瀏覽權限上的差異，其功用在於標明標註者的身分，以提供權威性的判斷依據。

3.1.4. 搜尋規範(Search API)

本規範旨在提供研究者於圖像上進行標註後，能夠使用搜尋功能以尋找相同標註的其他圖像。(Appleby, M., Crane, T., Sanderson, R., Stroop, J., & Warner, S. , 2016)相較於牛津大學、威爾斯國家圖書館、史丹佛大學圖書館等機構，由於其資源類型多為文字書籍，因此採用 Universal 瀏覽器（Viewer）提供內容文字搜尋的功能；本研究以研究者在田野拍攝的圖像為主，因此提供針對標註內容的檢索功能。

3.2. 系統與 IIIF 瀏覽器(IIIF Viewer)

本文根據「圖像規範」及「展示規範」，實作符合 IIIF 標準的圖像及圖像資源架構，並建構能與 IIIF 圖像兼容的圖像瀏覽器展示。有鑑於現有的瀏覽器開源軟體或開放原始碼眾多，但是無一完全符合佛教藝術研究者的需求。目前國際多數機構使用的 IIIF 兼容瀏覽器，一為 Mirador[3]，另一者為 Universal Viewer[4]，其中 Mirador 瀏覽器是由史丹佛大學以及哈佛大學共同合作開發的瀏覽器，使用者包括耶魯大學英國藝術中心(Yale Center for British Art) 、東京大學的 SAT 大正新脩大藏經資料庫 、牛津大學等機構採用 Mirador 瀏覽器展示館藏內容；另外 Universal Viewer 同樣是開源程式碼，亦受到廣泛的使用，包括牛津大學、威爾斯國家圖書館、史丹佛大學圖書館。

因無符合佛教藝術使用者需求的瀏覽器，本研究針對研究者需求開發必要的功能，期望能協助佛教藝術學者切中其最基礎的研究需求，提供符合 IIIF 功能的服務，並整合伺服器與前端使用者系統介面，以提供迅速便捷的圖像呈現服務。IIIF 瀏覽器的系統架構圖如下（圖 7）所示，包括：前端使用者介面（Client Site）、後端伺服器（Server Site），以及儲存資料庫（Image Repository）的運作：

[3] http://projectmirador.org/

[4] http://universalviewer.io/

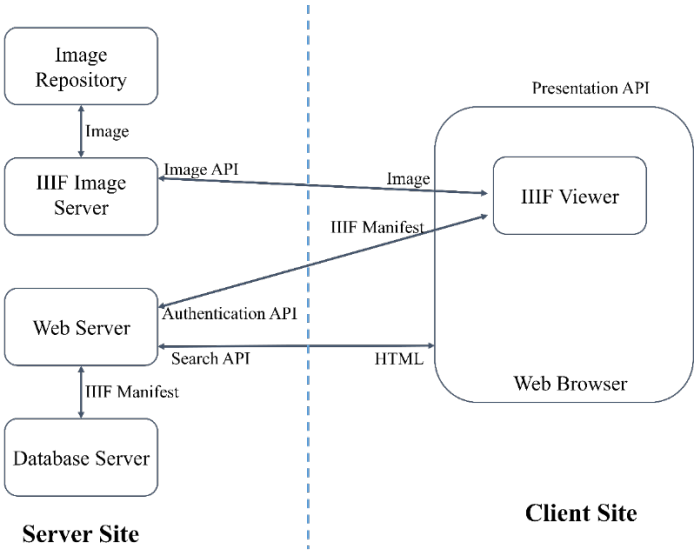


圖 7 本研究系統架構圖

相較於目前 IIIF 已公開的四個技術規範（API），在研究實作層面，官方網站中列出的開源軟體，經檢視尚無完整實作四個 API 的 IIIF Viewer 與 IIIF Image Server，其偏重的功能需求各不相同，依目標使用者需求進行開發，但無一與佛教藝術研究者需求完全吻合，因此本研究亦進行 IIIF Viewer 和 Image Server 的開發，使資料庫和網頁能串接（圖 7）。以使用者較多的 Mirador 和 Universal Viewer 兩者與本研究的 IIIF Viewer 進行功能分析：

表 5 IIIF 圖像瀏覽器比較(本研究、Mirador、Universal)

功能	功能說明	本研究	Mirador	Universal
1. 圖像縮放	放大檢視圖像細部內容，縮小查看整張圖像。	O	O	O
2. 圖像比較	可任意選擇兩張相同或不同的圖像，兩者並排檢視	O	O	X
3. 標註	在圖像上進行標註	O 開發中（分類標註功能）	O 無分類標註	X
4. 使用者驗證	Authentication API，需透過認證(例如使用規則同意、帳號登入等)才可以檢視圖像或使用 IIIF 功能(如:標註)	開發中	X	X
5. 檢索	可以檢索 IIIF Manifest 中的內容，如標註或是影像中的文字	O 開發中（檢索標註內容）	X	O 檢索影像內容文字(OCR)
6. Collection 列表	展示 Collection 樹狀列表	O 開發中	X	O
7.Canvas 切換	在單一 Manifest 依照 Sequence 切換瀏覽不同 Canvas	O	O	O

在本系統中研究者能輕易的放大縮小檢視圖像的內容，達到細部檢視圖像的效果，可以自由拖曳圖像瀏覽圖像觀察局部內容，通過身分認證的使用者可進行圖像標註並檢索標註內容。同時系統具備 IIIF API 未規範、但對於研究者而言至關重要的功能，例如：圖像比對的功能，將兩張圖像並排展示，提供研究者從兩者的比較中發覺新的觀點。

4. 未來的研究開發

現階段本研究安排佛教藝術作品的結構為：將佛教藝術作品分為三層結構，每一層結構都有其各自的一筆後設資料，一筆後設資料作為一個 Manifest，例如小南海中窟為一個 Manifest、小南海石窟中窟東壁為一個 Manifest、小南海石窟中窟東壁一佛二菩薩像為一個 Manifest，以網頁的分層以及視覺化圖型表示佛教藝術石窟的階層結構。

未來將持續開發 IIIF 展示規範中的 Collection，以呈現石窟的階層架構，並將佛教藝術作品的三層結構集合在一起，將同一壁面上的不同圖像以 Collection 組合，形成樹狀結構，使研究者能層層遞進，從最上層結構往下探尋。當未來研究者需求更加複雜時，將評估是否加入 Range 的功能，用以區分不同的細節，例如在第三層圖像(作品最小層次)中，不同影像有不同的側重點，如第三層圖像「文殊維摩圖」中有整幅圖像的影像、針對文殊菩薩的影像以及針對維摩詰的影像，可藉由 Range 將同一個 Manifest 下的所有影像區分為三個區塊，使研究者能一目瞭然一個作品底下拍攝的不同角度照片 。

5. 結論

本研究基於立體佛教藝術作品及其上雕刻繪製之圖像為研究材料，嘗試以國際圖像互操作架構（IIIF）為方法，佛教藝術史領域為例，發展以圖像材料分析為本的數位工具，使零碎、平面的圖像資料依其原有空間、時間脈絡串聯，以協助研究者解決研究過程可能需要的各種有意識的問題探究。IIIF 方法可使研究者以更高的效率交換資料，並且在研究過程中可以對圖像自由註記，將研究觀察與發現加以記錄，並進一步與其他研究者交流圖像及學術註記。基於 IIIF 建立在共通的標準，使圖像易於分享及取得，圖像可以由學者專家共同標記，研究者得以在領域專家協同合作的基礎下展開進階的研究。此外，圖像放大、縮小、旋轉、選擇圖像呈現範圍等功能，對於藝術史的圖像比較研究，擁有極大的優勢，幫助研究者進行細部圖像分析；其中，高品質圖像功能，可以提供研究者針對某個區塊深入觀察，也能將局部範圍分享或公布，使其他觀看者聚焦於局部，避免失去焦點。圖像的開放、互通，不只協助研究者在數位環境研究，也促進學術交流和傳播。圖像的開放、註記的公開，以及研究材料的開放，得以讓其他學者評估研究結果，或是提出疑問或是新的論點。未來，也能結合資訊科學領域的機器學習技術，讓標記好的圖像與內容做為訓練電腦的材料，以發展圖像內容偵測與比對等應用。

謝辭

本研究獲得中華民國科技部民國 106 年專題研究計畫（計畫編號：MOST 106-2420-H-001-021-MY2)補助，特此誌謝。此外，由衷感謝中央研究院歷史語言所顏娟英教授提供的研究材料、討論及協助，使研究得以順利進行。

參考文獻

（1） 專書

杜協昌（2016）。〈DocuSky：個人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研討會 DADH 論文集》。台北：國立台灣大學。

林富士（2017）。《「數位人文學」白皮書》。台北：中央研究究院數位文化中心。

項潔、涂豐恩（2011）。〈導論：什麼是數位人文〉，《數位人文研究與技藝》，頁 39-66，項潔（編）。台北市：臺大出版中心。

顏娟英（1991）。〈河北南響堂山寺石窟初探〉，在編者宋文薰《考古與歷史文化: 慶祝高去尋先生八十大壽論文集》，頁 331-356。臺北市 ： 正中書局。

顏娟英（1995）。〈北齊小南海石窟與僧稠〉，在編者釋恆清（編）《佛教思想的傳承與發展：印順導師九秩華誕祝壽文集》，頁 561－598。臺北市：東大圖書公司。

（2） 期刊論文

代學明（2010）。〈須彌山石窟及其價值〉，《絲綢之路》2010(20)，頁 16-17。

宿白（1978）。〈雲岡石窟分期試論〉，《考古學報》1978(1)，頁 25-38。

楊寶順、孫德萱、孫士杰（1982）。〈安陽修定寺唐塔雕磚的複製工藝〉，《文物》1982(12)，頁 75-79。

董玉祥(1983)。〈麥積山石窟的分期〉，《文物》，1983(6)，18-30。

董華鋒、寧宇（2010）。〈南、北石窟寺七佛造像空間佈局之淵源〉，《敦煌學輯刊》2010(1)，99-109。

（3） 會議論文

王祥安（2017）。〈中國傳統藥物資訊系統暨研究平台簡介〉，台灣中醫醫史文獻學會第二屆會員大會，2017/01/22，台北：中央研究院。

Kitamoto, A. (2017). “Creative scissors and paste: IIIF curation viewer and inquiry in digital humanities.” Pre-event of 15th International Conference of the European Association for Japanese Studies (EAJS2017).

（4） 網路資源

小南海石窟（2015 年 1 月 11 日）。維基百科，自由的百科全書。檢索自 https://zh.wikipedia.org/wiki/%E5%B0%8F%E5%8D%97%E6%B5%B7%E7%9F

%B3%E7%AA%9F
賓陽中洞（無日期）。法鼓文理學院。檢索自 http://www.dila.edu.tw/node/3705。
Appleby, M., Crane, T., Sanderson, R., Stroop, J., & Warner, S. (2016). “IIIF Content Search API 1.0” *International Image Interoperability Framework*. Retrieved from https://iiif.io/api/search/1.0/
Appleby, M., Crane, T., Sanderson, R., Stroop, J., & Warner, S. (2017a). “IIIF Authentication API 1.0” *International Image Interoperability Framework*. Retrieved from https://iiif.io/api/auth/1.0/
Appleby, M., Crane, T., Sanderson, R., Stroop, J., & Warner, S. (2017b). “IIIF Image API 2.1.1.” *International Image Interoperability Framework*. Retrieved from https://iiif.io/api/image/2.1/
Appleby, M., Crane, T., Sanderson, R., Stroop, J., & Warner, S. (2017c). “IIIF Presentation API 2.1.1.” *International Image Interoperability Framework*. Retrieved from https://iiif.io/api/presentation/2.1/
Delmas-Glass, E. (2016). “Yale Center for British Art’s Reformationto Restoration project: Applying IIIF Mirador technology to support digital scholarly collaboration and research.” MW2016: Museums and the Web 2016. Retrieved from: http://mw2016.museumsandtheweb.com/paper/yale-center-for-british-arts-reformation-to-restoration-project-applying-iiif-mirador-technology-to-support-digital-scholarly-collaboration-and-research/
Mirador.(n.d.). Retrieved October 8, 2018 from http://projectmirador.org
Sanderson, R. Ciccarese, P. Young, B. Web Annotation Data Model. W3C Recommendation, 23 February 2017, Retrieved from https://www.w3.org/TR/annotation-model/
Universal (n.d.). Retrieved October 8, 2018 from http://universalviewer.io/



Tibetan-Chinese-Sanskrit Text Alignment using Intelligent Agents and Genetic Algorithms

Christopher Handy
Principal Software Engineer, ERC Open Philology
Leiden University

Tibetan-Chinese-Sanskrit Text Alignment using Intelligent Agents and Genetic Algorithms

Christopher Handy

Principal Software Engineer, ERC Open Philology

Leiden University

Abstract

The problem of multilingual text alignment is a frequent concern in the study of Buddhist texts. Often we find ourselves in possession of several Chinese, Tibetan and Sanskrit versions of a given textual work, without a clear sense of exactly how each individual text relates to the others. Two texts may contain some material in common without sharing all of their content, or share the bulk of their content but with phrases in different orders, or have a common vocabulary but no shared content at all. These issues are well known to philologists, and the idea of using computer software to alleviate some of the mechanical legwork in comparing texts has revolutionized the ways that we do research, within the narrow field of Buddhist studies and also much more broadly on any texts. Yet ancient texts, and especially ancient Asian texts, pose difficulties that prevent some popular text analysis methods commonly used for modern European languages from working properly with Tibetan, Chinese and Sanskrit. One desired task that is reasonably complex is to compare any two texts across these three languages, quantifiably measure how similar they are, and align the texts based on regions of similarity. The method I describe here can theoretically achieve this goal for any set of input texts in any language, but my examples are restricted to a specific set of Buddhist works in Chinese, Tibetan and Sanskrit called the *Mahā ratnakūṭa Sūtra* (*MRK*). I demonstrate here a proof of concept on a few texts from this collection, and then discuss areas for improvement of the basic idea.

My method involves applying a genetic algorithm to intelligent agents to evolve the best alignments naturally from a given set of texts. Intelligent agents are computer programs designed to carry out a specific set of tasks using some kind of deterministic method and knowledge base. This type of system is useful when we know how to describe a decision process, but do not know all possible results of a decision. Genetic algorithms are information transmission schemes modeled on biological processes. They differ from biological processes

in that we tend to specify a quantifiable end goal for them to reach without specifying the means of getting to the goal. By stating this goal in terms of a fitness algorithm, we can promote reproduction of agent genes in our model for those organisms least unfit according to the desired output (i.e., consistently improving accurate text alignments). Over multiple generations of this promotion, the gene pool of agents approaches 100% fitness (normally, an unreachable ideal). Genetic algorithms are useful for applications in which we know what we want our output to look like but have no idea how to get the results. For our text alignment problem, we have target words in our text that will be "most interesting" in the mathematical sense. We do not care if the computer finds these in the most efficient way, only that it reliably reports them. But, what is most interesting could change based on additional input witnesses. So, our system must adapt as it analyzes more texts.

Our agents in this scenario are tiny grammatical engines that each do a sequence of short alignment tasks between strings of syllables encountered in the input texts based on training they receive from manual alignments. By stacking sequences of successful organisms together, we can achieve various alignment suggestions from the model.

Table of Contents

1. Introduction to the ERC Open Philology Project
1.1. The *MRK* Collection as a test project
1.2. The Buddhist Canon as a Digital Object: Resolution and Scope
1.3. The Problems of Current Software

2. Examples
2.1. Manual tests
2.2. Computer random tests
2.3. Assembling an organism
2.4. Massive population parallel problem solving

3. Conclusions
3.1. Interpretations of Data
3.2. Comparison of Automated and Human Alignments
3.3. Further research

4. Data

Keywords

Tibetan, Chinese, Sanskrit, alignment, genetic algorithm, intelligent agent

--- Paper begins from the next page ---

1. Introduction to the ERC Open Philology Project[1]

This paper presents a novel method for aligning Buddhist literature across multiple languages, with special attention to cases in which two witnesses do not preserve equally the contents of a text. I use large populations of very simple intelligent agents (decision-making digital processes) to approach a target alignment goal between two texts organically by evolving agent populations toward optimal solutions. In simple terms, this method treats grammatical rule systems as organic entities, tests these rule systems against real text samples, and allows or denies rule reproduction by promoting "least incorrect" solutions over many thousands of iterations.

In this model, I consider a language to be a system in which meanings are constructed through repeated forms of language particles that are more or less bounded as "syllables." This view of language is very inclusive and general, while at the same time also allowing us to find discrete points between units of meaning in a flexible way. The syllables that are most frequent often appear in "words" of one or more syllables, which themselves combine with each other and smaller particles to form "utterances." These utterances may or may not be grammatically well-formed sentences; as long as repetition occurs in some way, a meaning of some kind is assumed to be transmitted. In short, if a text is not random, then it must have some describable pattern, even if we do not know what that pattern is. A meaningful text cannot be entirely random, since its meaning must be preserved somewhere (possibly in the larger corpus of which it is a member). We refer to each text at various resolutions, as follows:

- Collections
-- Works
--- Witnesses
---- Segments

1 For full details about the project and its goals, visit the project website: openphilology.eu

----- Ngrams

This schema allows us to refer to a selection of text at whatever level is most convenient in any given situation. Each text is thus referenced in a database by its Collection (e.g., *MRK*), Work (e.g., *Gaṅgottaraparipṛcchā*), Witness (e.g., Taishō 310.31), and its constituent smaller Segments made up of Ngrams and individual Phonemes or syllables (1-grams). The main goal of our alignment engine is to produce strong relationships between the Segments of two source Witnesses, based on an analysis of their contents at higher resolutions (Ngram pattern matches).

We cannot know ahead of time if any particular alignment solution is even computable by linear methods. We therefore assume the worst, and approach the problem piecemeal. Each agent (a small computer program) only accounts for a small portion of a total alignment. So, the addition or removal of any single agent will in theory have no noticeable effect on the alignment production. In addition, this method allows for a continuously dynamic system in which new alignments are constantly being produced as the system incorporates new texts and continues into further iterations.

The texts to be examined are redundantly imported into our system as n-grams for every size n, with an n-gram defined as a sequence of n syllables. We determine statistically which sequences of various sizes occur most often within a text and set of texts, thereby establishing a baseline for where words may occur. We then check these n-grams against prepared dictionaries in order to obtain further understanding of waypoints of meaning within the text. This method requires a large amount of storage space, but yields interesting and useful results.[2]

2 See Handy Forthcoming.

After this initial preprocessing of the texts, we create a seed population of artificial organisms, small grammar robots that operate on rules unknown to us. Their understanding of the alignment between the source texts is at first completely random, but over successive iterations approaches a set of target alignments. The fulfillment of the goal is achieved by promoting successful agents (or more literally the least unsuccessful) in the system based on a fitness algorithm that compares agent results with manual human alignments. A given organism is thus defined as a collection of genes, which are themselves sequences of alignment rules, which are simply mappings between the n-grams of source texts:

- Organisms
-- Genes
--- Alignment rules
---- N-grams

Hierarchical multiresolution structure of agents

Since these agents are not coded by hand, they can evolve to account for alternate spellings of words, scribal errors, non-monotonic alignments, missing segments of text, and other issues that may throw off other alignment methods. Agents also do not need any part of speech tagging to be effective, although this feature can also be added to them to achieve more accurate results.

1.1. The *MRK* Collection

All examples here are extracted from various extant witnesses of the *Mahāratnakūṭa Sūtra* (*MRK*), a collection of 49 individual works on Mahāyāna Buddhism. This collection of texts appears to have its origins in Sanskrit texts, but we now have these works primarily in multiple Chinese and Tibetan translations. In order to do anything meaningful with the texts using a computer, we must first represent them digitally in a suitable way. The database

structure proposed here is suitable for the *MRK*, but also for any other collection of texts. Our system is thus adaptable to other problem sets (e.g., “all vinaya texts,” “all Lotus Sūtra texts”), even in languages other than those used for the examples in this document.

The ambiguous history of the *MRK* texts precludes us from representing the history of each text through a presumed ur-text and descending inheritors. So, we must have a non-hierarchical reference structure when talking about the different versions of what we define as a Work. In our database (a simple PostgreSQL relational database), a Work is a placeholder to be referenced by a specific Witness, and does not represent a real object in the world. In contrast, a Witness is a specific instance of a Work that can be stored as a text file, the digitized version of an actual physical document. The *MRK* Collection contains exactly 49 Works, but there is no limit to the number of Witnesses that we can add to our system and point to those Works. Works and Witnesses are easily represented as lines in a database table, which provides the file names of our various text collections. We then create other tables in our database at increasing levels of resolution to analyze each text as various smaller units (e.g., a page, a sentence, a word, a syllable).

A Witness, meaning simply an instance of a text, is a sequence of utterances. There are many different kinds of texts, but for our purposes we consider any text to be representable as one or more Segments ordered by a set numbering scheme. These Segments are collections of meaningful statements in a language, whether Chinese, Tibetan, Sanskrit or something else. They can be further divided into N-Grams, which are in this context combinations of syllables, and Phonemes, and which we can define as individual syllables. This schema allows us to consider texts at various levels of resolution about which we have different kinds of information, and also to draw quantifiable comparisons between any two texts at these different resolutions. So, for example, we might observe at the level of Witness that one version of a text uses the word Buddha 354 times more than the other version. At the

level of individual N-Gram, we might then see how the term “Buddha” is represented by a different sequence of syllables across various languages: in Chinese, 佛 (fó) but in Tibetan, □□□□□□□□□□ (sangs rgyas). In our system, 佛 (fó) is considered as a 1-gram and □□□□ □□□□□ (sangs rgyas) as a 2-gram. We want to draw a connection between these two terms such that we raise the alignment weight where our source texts contain them. Then, we want to combine these and other known matches, at increasing levels of resolution, to produce a complete alignment of two or more source texts in their entirety. Since we do not have reliable translations for most of the texts in the *MRK*, the system must also be able to suggest these alignments automatically, based on dictionary matches and statistical pattern matches. Dictionary matches rely on word to word correspondence of common terms in the Buddhist lexicon, including “Buddha”, “nirvāṇa”, and perhaps the name of the text itself. Statistical matches are generated by mathematical analysis of the text, are are thus context independent.

1.2. The Buddhist Canon as a Digital Object

Our database objects referencing each individual witness in our collection of Buddhist texts are part of a hierarchy of multiple resolutions stored as individual tables in the relational database:

Collections (*MRK*)

Works (49 *MRK* Works)

Witnesses (various individual *MRK* texts from Dergé Kanjur, Lithang Kanjur, Taishō Canon, Dunhuang manuscripts, etc.)

Segments (meaningful sequences within Witnesses)

Ngrams (strings of syllables within Segments)

Phonemes (individual syllables of Ngrams)

A unit of text can thus be addressed at any resolution desirable.

1.3. The Problems of Current Software

There are already many different software applications for text alignment, many of which are available free of charge. I evaluated a number of the most popular tools for alignment to see if the task at hand could be approached using CollateX, Juxta, GIZA++ or similar tools. While these applications are very useful on some texts, they tend not to be optimized for working with the unique problems encountered in classical Asian texts. The four most significant issues of this kind were the following:

1) Tibetan, Sanskrit and Chinese lack spacing between individual words, and therefore some other method than finding whitespace (as with English) must be used to find word boundaries.

2) Part-of-speech tags are not available consistently or are unreliable for our source texts.

3) Source texts often do not match up sentence by sentence as with monotonic alignments.

4) Digitized versions of source texts contain various and inconsistent editors' marks, computer codes (e.g., XML tags), transliteration schema, header information, transcription errors, and other "noise" that must be accounted for in some way.

I found that the simplest way to sidestep many of these issues was to use my own custom software utility, aks, to segment the source texts at all possible word boundary points, and then simply count the frequency of every single observed n-gram at every length to determine which strings occur in the set most frequently. This method is not a perfect way of finding words, but we do not need to find the exact boundaries of words for the genetic algorithm to produce a reliable match. The reason for this is that we are obtaining our results as the sum total of many thousands of agent "votes" in the system, and so incorrect boundaries at any step essentially drop off by not being well represented by the total population.

2. Examples

2.1. Manual tests

The Principal Investigator for Open Philology, Jonathan Silk, prepared a manual alignment of *MRK* 31 (*Gangottarā Paripṛcchā*) along with other *MRK* texts in advance of the current project. The alignment for MRK 31 became the initial sample data around which the segment hierarchy was constructed.

The most basic task at hand is to take Buddhist texts composed in Chinese and line them up with texts composed in Tibetan. Consider, for example, the following line from the *Gaṅgottaraparipṛcchā* (*Questions of Gaṅgottarā*, MRK #31, manual alignment and translation by Jonathan Silk):

Tibetan Derge *Kanjur* D75 (ADARSHA):

bcom ldan 'das kyis bka' stsal pa/ kun dga' bo chos kyi rnam grangs 'di ni dri ma med pa zhes bya ste/ 'di dri ma med pa zhes bya bar zung shig

The Blessed One said: “Ānanda, this exposition of the teaching is called ‘Stainless’; maintain it as ‘Stainless.’”

Chinese T310.31 from CBETA:

佛言。此經名為離垢清淨。以是名字汝當受持。

The Buddha said: “This sūtra is called ‘Stainless Purity.” Under this name you should maintain it.”

We can see here a number of keywords that could be used to link these passages together as “the same,” even though not all major words correspond. For example, the Derge Kanjur's “Blessed One” (**bcom ldan 'das**) is 佛 in the Taishō edition, and we likewise have an easy

match for **'di ni dri ma med pa zhes bya** and 離垢清淨. However, the appearance here of **'di ni dri ma med pa zhes bya** two times in the Tibetan version compared with the single appearance of 離垢清淨 in the Taishō edition means that we cannot simply map like numbers of n-grams together. Yet we can see from the above that there are many other potential strings in the above passages that could be used to link them together. Since we do not know which Tibetan strings match up with which Chinese strings, our chance of guessing an exact alignment should be the same regardless of which two strings we pick. If we were to perform a selection of two strings at random, and repeat this process 10,000 times, or 100,000 times, or an infinite number of times, in the manner of monkeys typing Shakespeare, we would eventually match each matchable point in the two witness, along with every possible wrong match. Each of these Shakespearean monkeys is an agent in our system, and our goal in using a genetic algorithm is to select, promote and mutate agents producing the least incorrect answers until we achieve the alignment 100%.

2.2. Computer random tests and scoring system

One strange feature in using genetic algorithms and other iterative approaches to solving problems is that the optimal starting conditions of the model are often unknowable. We can use this to our advantage by beginning with completely random alignments and then checking to see which alignments are least inaccurate according to a scoring system. As long as the scoring system is consistent, it should promote alignments that are closer to our target goals (based on manual training data) and allow these intelligent agents priority in reproducing offspring. So, we focus on defining what it means for an alignment to be correct, and allow the computer to decide which sequences of n-grams in one text map onto related n-grams in another text.

For Chinese and Tibetan, it is relatively easy to extract and examine texts at the level of individual syllables. In our system a 1-gram for Chinese means a single character, and a 1-gram for Tibetan is a single syllable (generally followed by a *tsheg* delimiter). In Sanskrit as well, a 1-gram refers to a single *akṣara* (syllable), but due to editing standards for roman character editions of Sanskrit texts, we put these through some special processing first.

I use my own custom software to prepare the texts, beginning with n sizes of 1 and storing the text as all possible 1-grams, then as 2-grams, 3-grams and up to a limit determined by the user. Larger values of n do increase processing time slightly, and required storage space greatly.[3] However, the large files generated by this approach are only needed briefly, in order to count the frequency of each n-gram throughout the entire text. Frequently-occurring longer n-grams should in theory be representative of stock words and phrases in a text. If texts in two different languages are equally consistent in employing these units of vocabulary, we can begin to imagine possible alignments based on sets of repeating n-grams in each text. As we collect more potential matches, natural alignment points become increasingly visible.

Each of our three languages of interest, Chinese, Tibetan and Sanskrit, may differ greatly with respect to which values of n represent the bulk of usable vocabulary. In Chinese, many words are only one or two characters, whereas Sanskrit and Tibetan are more likely to employ words of 3, 4 or 5 syllables.

After the initial text extraction up to n = 8 for our source texts, we can examine the results to determine whether or not higher values could potentially be useful. Again, Chinese texts predictably have less results at higher levels of n compared with Tibetan texts. For the latter, it can be useful in some cases to increase n to 100 or even higher. As an example of what we might find, let us consider our n-gram output for *MRK* 31 (see section 4.1) *Gaṅgottaraparipṛcchā* (Questions of Gaṅgottarā):

3 See Handy Forthcoming.

gang gA'i mchog / **恒河上** / Gaṅgottarā, the title character of the story

In our sample Tibetan text, gang gA'i mchog appears exactly 45 times. However, the string **恒河上** appears only 17 times. This discrepancy is due to different grammatical structures and artistic liberties employed in narrating the story. It is for this reason that we cannot rely on simple monotonic alignments for these texts, because there often is no 1:1 correspondence for each word in a set of two texts.

We can also see easily in the Tibetan that there are more long, meaningful n-grams at size 8 and above. In contrast, there are less than 10 significantly repeated 8-grams in the Chinese text. Analysis of this text at higher levels of n is not likely to be useful at all.

We also know from our manual analysis that there are several key conventions in this text that show up repeatedly. There are two major characters: the Buddha and Gaṅgottarā, who are more or less consistently called **bcom ldan 'das** and **gang gA'i mchog** in Tibetan, and **佛** and **恒河上** in Chinese. Since the entire text is a conversation between these two characters, we also find that the texts frequently include phrases such as “the Buddha said” and “Gaṅgottarā said,” which produce their own n-gram signatures we can use to find specific points in the texts.

2.3. Assembling an organism

Organisms are constructed from a simple Python framework that hosts a unique gene. We can really use any language for this task, but chose Python for its readability, its popularity in DH generally, and the relative ease with which we can manipulate strings in this language. A gene accounts for all of the specific traits of the individual organism, and is functionally just a record of one or more particular alignments between n-grams in Tibetan, Chinese and/or Sanskrit. We call it a gene because it is transmitted from parent to child in a crude model of

biological gene transmission. We can imitate either asexual or sexual reproduction in this model, and the mutation rate can be controlled by the user.

Initially, we want to make our agent organism as dumb as possible. The “intelligent” portion of the system will emerge first from the successful propagation of genes that do specific tasks well, not from the intelligent structure of any single organism. The instructions for our initial dumb organism are as follows:

-select 2 ngrams together at **random** from text A

-search for these ngrams in close proximity in text A

-select 2 ngrams together at **random** from text B

-search for these ngrams in close proximity in text B

This organism does not know why it matches the words, and it will likely fail. However, it may sometimes succeed. Even among the failed matches, some can be measured as less failed than others. So if we can call on the organism to do its task several thousand times, we can count on it getting a few good matches.

We replace the instruction above, “select 2 ngrams together at **random**” with “select 2 ngrams together **from this organism’s gene**”, to create a population of agents. In other words, we want a generic program whose behavior is determined at runtime based on the genetic information received from its parent agent. Each of the agents is a separate Unix process called by a parent program that maintains the background dependencies of the system. To make more complex organisms, we can string together multiple dumb organisms:

-select 2 dumb organisms at random

-run organisms together against test data and compare output

if successful alignment produced, continue production of complex organism

We can continue making increasingly complex organisms by repeating the above.

By increasing the complexity of these agents, we can make them do increasingly intelligent tasks, if we provide enough example matches. In this way, we can evolve agents that specialize in finding verb and noun indicators, proper names, vowel strengthening, end of text signifiers, and so forth. These rules are stored in a table of Rules, which are simply pairings in another relational database table (I used Postgresql).

We don't actually create the rules themselves, and simply knowing proper Tibetan, Chinese or Sanskrit grammar will not likely cover all cases. In fact, strict interpretations of grammar can even hide information about the text(s) that we want to have. So, our computer program will promote rules based on whether or not they seem to be capable of describing a variety of grammatical environments. Our rules are first trained on known *MRK* texts that have been aligned manually, then on *MRK* texts that have not been manually aligned, and then on other texts in the *Kanjur* and *Taishō* canons.

2.4. Massive population parallel problem solving

This task is determined to be naively parallel, meaning it is very easy to run all of our organisms at once as independent processes across as many processor cores as we are able to obtain at any given time. The naively parallel design of the system allows for the addition or removal of processor cores, even as the system is in operation, without negative consequences on the results of the data. This fact is easily verifiable through unit testing of the expected alignments in a controlled setup in which processors are methodically added and removed.

3. Conclusions

3.1. Interpretations of Data

Source texts used in these examples come from two Internet respositories:

- for Chinese, the CBETA Taishō Canon (http://www.cbeta.org/).

- for Tibetan, the Derge Kanjur files prepared by ADARSHA, available through the University of Vienna e-Kanjur Project (https://www.istb.univie.ac.at/kanjur/). This site, an amazing resource, has other Kanjurs available as well, and will likely continue to add to its collection.

I have not completed any testing of my entire method on Sanskrit texts yet, but plan to demonstrate a few examples during my presentation.

-alignments are more or less accurate, but not fine tuned enough

-initial data suggest that this method can achieve a useful level of automated preprocessing

3.2. Comparison of Automated and Human Alignments

- human alignments take far greater nuance into account, as expected

- computer can perform alignments much more quickly (effectively instantaneous)

- human-computer interactive environment can create better auto-alignments and faster human alignments through assisted suggestions.

3.3. Further research:

- increase number of dictionaries, witnesses, processor number -> higher organism population and or increased iteration speeds

- see website, openphilology.eu for more information on most recent data output

4. Data

4.1. *MRK* 31 (*Gaṅgottaraparipṛcchā*) ngrams and alignments

Taishō MRK #31 in T310.31, (Chinese) top 1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-grams:

1	2	3	4	5	6	7	8
如	河上	恒河上	恒河上言	趣涅槃界耶	豈不趣涅槃界	豈不趣涅槃界耶	汝豈不趣涅槃界耶

是	恒河	河上言	亦復如是	豈不趣涅槃	汝豈不趣涅槃	汝豈不趣涅槃界	言汝豈不趣涅槃界
言	涅槃	優婆夷	世尊告言	河上優婆夷	恒河上優婆夷	言汝豈不趣涅槃	當云何答世尊告言
上	如是	世尊告	趣涅槃界	汝豈不趣涅	不趣涅槃界耶	至涅槃亦復如是	恒河上言若一切法
說	世尊	斷流轉	豈不趣涅	恒河上言若	非思惟之所能	當云何答世尊告	乃至涅槃亦復如是
何	云何	者云何	若一切法	恒河上言如	言汝豈不趣涅	河上言若一切法	n/a
無	佛言	復如是	涅槃界耶	恒河上優婆	見身異於幻化	恒河上言若一切	n/a
法	上言	尊告言	河上言若	不趣涅槃界	至涅槃亦復如	亦復如是恒河上	n/a

Dergé D75 in Vienna/ADARSHA Dergé (Tibetan) top 1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-grams:

1	2	3	4	5	6	7	8
pa	ldan 'das	bcom ldan 'das	mya ngan las 'das	yongs su mya ngan las	yongs su mya ngan las 'das	bcom ldan 'das kyis bka' stsal pa	ldan 'das kyis bka' stsal pa gang gA'i
dang	bcom ldan	gang gA'i mchog	bcom ldan 'das kyis	su mya ngan las 'das	ldan 'das kyis bka' stsal pa	ldan 'das kyis bka' stsal pa gang	bcom ldan 'das kyis bka' stsal pa gang
ba	gang gA'i	ngan las 'das	pa gang gA'i mchog	stsal pa gang gA'i mchog	bka' stsal pa gang gA'i mchog	kyis bka' stsal pa gang gA'i mchog	'das kyis bka' stsal pa gang gA'i mchog
'das	gA'i mchog	mya ngan las	yongs su mya ngan	mya ngan las 'das pa	bcom ldan 'das kyis bka' stsal	'das kyis bka' stsal pa gang gA'i	lags bcom ldan 'das kyis bka' stsal pa
gang	pa dang	ldan 'das	su mya ngan las	ldan 'das kyis bka'	kyis bka' stsal pa	yongs su mya ngan las 'das	yongs su mya ngan las 'das pa la

		kyis		stsal	gang gA'i	par	
par	mya ngan	bka' stsal pa	stsal pa gang gA'i	bka' stsal pa gang gA'i	dge bsnyen ma gang gA'i mchog	lags bcom ldan 'das kyis bka' stsal	gang gA'i mchog gis gsol pa bcom ldan
de	ngan las	pa gang gA'i	ngan las 'das pa	bcom ldan 'das kyis bka'	'das kyis bka' stsal pa gang	yongs su mya ngan las 'das pa	gA'i mchog gis gsol pa bcom ldan 'das

(potential alignments for *MRK* 31)

Witness 1 (D75):	Witness 2 (T310.31):
zhes de skad bgyis na ji skad lan 'debs par 'gyur lags/ **bcom ldan 'das kyis bka' stsal pa**/ gang gA'i mchog sprul pa la ni 'greng ba med/ 'dug pa med/ nyal ba	T11n0310 p0549b24(01)‖即白**佛言**。世尊。若問化人汝從何來。
gA'i mchog khyod da gzod gang nas 'ongs zhes ci'i slad du smra/ **bcom ldan 'das kyis bka' stsal pa**/ gang gA'i mchog sprul pa ni ngan song du mi 'gro /mtho ris	T11n0310_p0549b27(03) 又問諸法豈不皆如化耶。**佛言**。如是如是。
'das par mchi ba ma lags te/ bcom ldan 'das dge bsnyen ma gang gA'i mchog kyang de lta bu'i dbyings can lags so/ /**bcom ldan 'das kyis bka' stsal pa**/ gang gA'i	T11n0310_p0549c06(09) 畢竟不復生善惡趣及般涅槃。我觀己身亦復如是。**佛言**。
par 'gyur lags/ **bcom ldan 'das kyis bka' stsal pa**/ gang gA'i mchog mi skye ba zhes	T11n0310_p0549c08(00) 應云何答。**佛言**。無生者即涅槃也。

bya ba ni mya ngan las 'das pa'i tshig bla dgas yin na de ci zhig lan 'debs	
bgyis na ji skad lan 'debs par 'gyur lags/ **bcom ldan 'das kyis bka' stsal pa**/ gang gA'i mchog dri ba 'di ni dmigs pa med pa'o/ /gsol pa/ ci bcom ldan 'das	T11n0310_p0549c09(03) 恒河上言。諸法豈不皆同涅槃。**佛言**。如是如是。
kyi slad du ji ltar dge ba'i rtsa ba yang dag par bsgrubs lags/ **bcom ldan 'das kyis bka' stsal pa**/ gang gA'i mchog gang dmigs pa de ni dge ba'i rtsa ba ma yin	T11n0310_p0550a21(01) 而作是言。以何因緣現此微笑。**佛言**。
slad du bsam gyis mi khyab pa la bsam gyis mi khyab pa zhes bgyi/ **bcom ldan 'das kyis bka' stsal pa**/ gang gA'i mchog chos 'di ni sems pas thob par bya ba ma yin	T11n0310_p0550a25(02) 於無餘涅槃而得滅度。阿難白**佛言**。當何名此經。
dang / 'khor ba dang / mya ngan las 'das pa zhes btags pa ji lta bu lags/ **bcom ldan 'das kyis bka' stsal pa**/ gang gA'i mchog 'di lta ste dper na bdag bdag ces	T11n0310_p0550a26(00) 我等云何受持。**佛言**。此經名為離垢清淨。

4.2. *MRK* #35 ngrams and alignments (MRK #35, *Acintyabuddhaviśaya*)

Taishō MRK #35 in T310.35, (Chinese) top 1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-grams:

1	2	3	4	5	6	7	8
諸	菩薩	殊師利	文殊師利	文殊師利言	如來應供正遍	如來應供正遍覺	阿耨多羅三藐三菩
菩	如是	文殊師	離垢清淨	菩薩摩訶薩	來應供正遍覺	阿耨多羅三藐三	耨多羅三藐三菩提
無	一切	諸菩薩	薩摩訶薩	言文殊師利	阿耨多羅三藐	耨多羅三藐三菩	釋迦如來應供正遍
是	文殊	須菩提	菩薩摩訶	時文殊師利	耨多羅三藐三	多羅三藐三菩提	迦如來應供正遍覺
薩	殊師	菩薩摩	須菩提言	應供正遍覺	羅三藐三菩提	一切聲聞辟支佛	如來應供正遍覺所
者	師利	師利言	彼諸菩薩	如來應供正	多羅三藐三菩	釋迦如來應供正	尊釋迦如來應供正
如	世尊	薩摩訶	善德天子	來應供正遍	切聲聞辟支佛	迦如來應供正遍	世尊釋迦如來應供
不	眾生	摩訶薩	言文殊師	聲聞辟支佛	一切諸天人故	來應供正遍覺所	詣世尊釋迦如來應

Dergé D79 in Vienna/ADARSHA Dergé (Tibetan) top 1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-grams:

1	2	3	4	5	6	7	8
pa	pa dang	bcom ldan 'das	yongs su dag pa	yongs su dag pa dang	'jam dpal gzhon nur gyur pa	lags so bka' stsal pa 'jam dpal	'dod chags dang zhe sdang dang gti mug
dang	yongs su	yongs su dag	byang chub sems dpa'	'jam dpal gzhon nur gyur	pas yongs su dag pa dang	'jam dpal gzhon nur gyur pa la	gzhon nur gyur pa la 'di skad ces
par	thams cad	byang chub	su dag pa dang	dpal gzhon nur gyur pa	so bka' stsal pa 'jam dpal	chags dang zhe sdang	dpal gzhon nur gyur pa la 'di

		sems				dang gti mug	skad
dag	ldan 'das	su dag pa	dpal gzhon nur gyur	pas yongs su dag pa	lags so bka' stsal pa 'jam	'dod chags dang zhe sdang dang gti	'jam dpal gzhon nur gyur pa la 'di
de	bcom ldan	chub sems dpa'	'jam dpal gzhon nur	gsol pa bcom ldan 'das	dpal gzhon nur gyur pa la	yang dag par rdzogs pa'i sangs rgyas	pa la 'di skad ces smras so 'jam
la	byang chub	sangs rgyas kyi	btsun pa rab 'byor	bka' stsal pa 'jam dpal	dang zhe sdang dang gti mug	shes pa yongs su dag pa dang	nur gyur pa la 'di skad ces smras
ba	'jam dpal	yang dag par	gzhon nur gyur pa	so bka' stsal pa 'jam	chags dang zhe sdang dang gti	nur gyur pa la 'di skad ces	mkhas pa'i rnam pas yongs su dag pa
kyi	sangs rgyas	dag pa dang	pas yongs su dag	lags so bka' stsal pa	'dod chags dang zhe sdang dang	gzhon nur gyur pa la 'di skad	la mkhas pa'i rnam pas yongs su dag

Bibliography

Handy, Christopher. Forthcoming. “A Context-Free Method for the Computational Analysis of Buddhist Texts.” In Daniel Veidlinger, Ed., *Buddhism and Digital Humanities* (De Gruyter).

Silk, Jonathan. Forthcoming. “Two Chinese Sūtras in Tibetan Translation: The Gaṅgottaraparipṛcchā and the Amituo jing.”

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

自動標點的原理與實現

釋賢超　方愷齊　釋賢迥　釋賢屈
釋賢碉　釋賢繼　釋賢大　釋賢奉　宋延淳
北京龍泉寺藏經辦公室

自動標點的原理與實現

釋賢超　方愷齊 釋賢迥 釋賢 diao 釋賢繼 釋賢大 釋賢 feng 宋延淳
法師　義工　法师　法师　法师　法师　法师　義工
北京龍泉寺藏經辦公室

摘要

古代漢語通常沒有標點，這給現代人閱讀、理解古籍文獻帶來極大困難。为古漢語文獻添加現代標點已成為古籍整理和研究的基礎，同時也是一項非常繁重的工作。歷史上，漢文大藏經的編修向來都是極為浩大的工程。在當今的智能科技時代，借助機器智能實現古籍文獻的自動標點具有現實意義。為瞭解決現代大藏經整理和校勘中面臨的具體困難，我們對大藏經基於人工智能（AI）輔助的自動標點方法進行了研究。應用 AI 技術在自然語言處理（NLP）領域的最新研究進展，通過兩種深度模型的訓練和測試，已獲得標點準確度最高達 94%的自動標點引擎，以此為基礎開發的自動標點系統（GJAP）現已上線運行 [1]。目前系統可提供七種現代標點（逗號、句號、問號、嘆號、頓號、分號、冒號）的古文線上標點服務。

本文將從深度模型原理、數據集的構建兩個方面來對自動標點的原理進行描述；使用總量超過五千萬個漢字的訓練數據和總量約一千萬的訓練數據，對兩種標點模型進行訓練；選取不同朝代的佛教古籍文本形成的測試數據集，對兩種自動標點引擎進行測試的比較；通過結果的分析討論，文章最後給出結論。

目次

1.　概述
1.1.　研究現狀
1.2.　技術範疇
1.3.　概念術語
1.3.1.　斷句
1.3.2.　標點
2.　模型原理
2.1.　**LSTM** 模型
2.2.　**CNN** 模型

[1] 自動標點平台：http://gj.cool/

3. 數據集
4. 模型比較
4.1. 模型訓練
4.2. 引擎測試 9
4.3. 結果討論
5. 結論

關鍵詞
自動標點，古籍文獻，數據集，LSTM，CNN

1. 概述

1.1. 研究現狀

自動標點是指在非人工幹預的情況下，根據特定演算法給沒有現代標點的古籍文本自動標注現代中文標點的技術，這項研究提出的時間並不長，只有十餘年的時間，研究報導也很少。有文章曾嘗試採用人工設計的規則庫或採用自然語言處理中的統計方法來進行自動標點研究[(1)]，也有探討利用循環神經網路對古文進行自動斷句[(2)]，如採用基於門控循環單元（Gated Recurrent Unit）的雙向循環神經網路，或採用條件隨機場下的雙向長短時記憶神經網路（Bi-LSTM-RNN-CRF）[(3)]，但是以上研究均未實現對古籍文獻的自動標點。英文自動標點的研究有不少文獻報導[(4)]，相對比較成熟，但與漢字標點研究有所區別。

1.2 技術範疇

基於人工智能的深度學習模型已廣泛應用於自然語言處理（NLP），如機器翻譯、文本分類、機器問答、自動摘要等。自動標點作為一個序列處理問題，也屬於 NPL 的應用範疇。本領域研究採用的主要模型架構有：基於循環神經網路（Recurrent Neural Network，RNN）的長短時記憶（Long Short Term Memory，LSTM）序列標注模型[(5)]，基於卷積神經網路(Convolution Neural Network，CNN)和多階注意力(Multi-step Attetion)序列到序列（Sequence to Sequence，Seq2Seq）模型[(6)]。

1.3 概念術語

本文涉及一些概念和術語，斷句和標點兩者的含義有區別。

1.2.1 斷句

斷句是指只標注句號，表示停頓；

1.2.2 標點

標點是指標注句號、逗號等多種現代標點符號，現代中文標點包括標號、點號兩類，標注這兩類標點屬於不同類型的自動標點問題。

（1）點號：點號表示語句中的語氣停頓，常見的點號有：句號、逗號、問號、感嘆號、頓號、分號和冒號。

（2）標號: 標號表示語句中的特定成分，常見的標號有：雙引號、單引號和書名號。

2. 模型原理

漢語古文序列的斷句和標點屬於序列標注問題，與自然語言處理中的命名實體標注

問題類似。根據古漢語的特點，標點標注問題可以分為兩類：一是點號標注，二是標號標注，它們分別對應於兩種不同類型的序列模型。為討論方便，以下兩類標點問題對應的模型稱之為 LSTM 模型和 CNN 模型。

2.1. LSTM 模型

點號的特點是，在同一個位置上最多只能出現一個點號，兩個點號不能同時出現在同一個位置。所以標注點號可以看作是一個序列標注問題，也就是由文本序列 T 生成一個具有同等字元長度的點號序列 C，標點序列中的第 i 個元素 c_i所對應的是文本序列中第 i 個元素 t_i與第 i+1 個元素 t_{i+1}之間的位置是否存在某種點號，如果存在，那麼 c_i便是七種點號中的某一種（“。”“，”“？”“！”“；”“：”“、”）；如果不存在，便為缺省常量 c。由文本序列生成點號序列是 N vs N 問題，適合採用 RNN 的經典結構及其變種，如 LSTM 模型。

經典的循環神經網路（RNN）模型由方程（1）表示 [6]，其中，(x_1，…，x_T）為輸入序列，($y_1, \ldots, y_T$)為輸出序列。

$$h_t = sigm(W^{hx}x_t + W^{hh}h_{t-1}) \quad (1)$$

$$y_t = W^{yh}h_t \quad (2)$$

RNN 結構（圖 1）的特點是每個時刻接受一個輸入向量，產生一個輸出向量，輸入與輸出之間存在至少一個隱藏層，每個時刻的輸出向量取決於當下的隱藏層狀態，每個時刻的隱藏層狀態則受上一時刻的隱藏層狀態和當前輸入向量的共同影響。原則上循環神經網路可以處理任意長度的序列，但是由於循環神經網路中每個時刻的狀態都受到上一時刻的影響，無法通過並行處理來加快訓練或推理的過程，所以運行效率通常比較低。

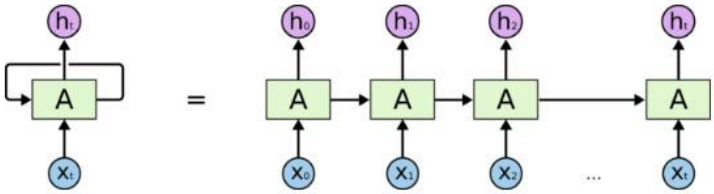


圖 1. RNN 經典結構示意圖

LSTM 是一種時間遞迴神經網路，它為解決 RNN 存在的問題而提出，其序列結構如圖 2 所示。基於 LSTM 的循環神經網路也可以很好地解決古典 RNN 方法難以適應多變數輸入的問題。

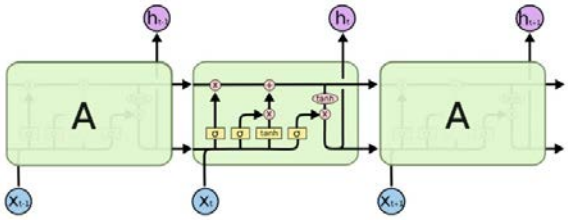


圖 2. LSTM 序列結構示意圖

2.2 CNN 模型

对于標號来说，其特點是成對出現的，即同一個位置可以出現多個標號。因此文本序列 T 的長度通常不等於標號序列 C 的長度，t_i與 c_i之間不能一一對應，屬於輸入序列與輸出序列長度不等的場景（N vs M）問題，適合採用“序列到序列”（Seq2Seq）模型[2]，其結構示意圖[3]見圖 3。

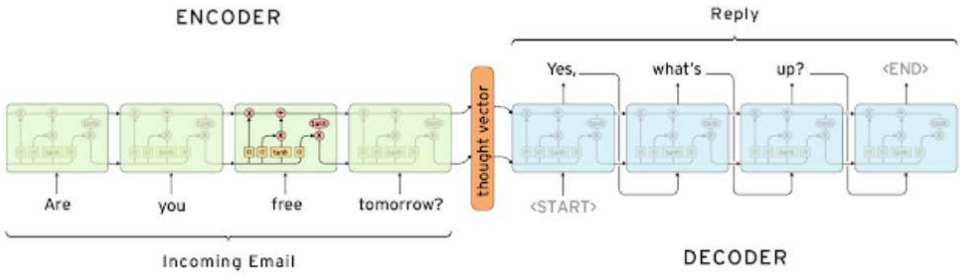


圖 3：Seq2Seq 結構示意圖

Seq2Seq 模型的原理是將原始序列“編碼”為一個背景向量（context vector），再從背景向量“解碼”為目標序列，從而實現不定長序列的轉換。編碼和解碼通常是基於循環神經網路而實現的。Facebook 發佈的 Fairseq 模型基於卷積神經網路實現編碼和解碼的過程，訓練和推理速度都大幅優於循環神經網路。

然而這種“編碼器-解碼器”（encoder-decoder）結構，由於輸入序列的數據被全部壓縮在一個背景向量之中，而背景向量的長度是有限的，因此背景向量的長度是提升序列轉換準確度的主要瓶頸。為瞭解決這個問題，允許背景向量在產生輸出序列的時候可以隨著時刻的不同而發生變化，改進後的結構被稱為注意力（Attention）[(7)]。CNN 架构同样适用于解决点号标注的问题。

[2] https://github.com/pytorch/fairseq
[3] https://github.com/nicolas-ivanov/tf_seq2seq_chatbot

3. 数据集

用於標點模型訓練和測試的數據集是通過大量電腦及人工輔助採集處理獲得的。內容包括不同朝代的歷史文獻數據，以及佛教、道教和儒家典籍等。訓練數據集總量約5400 萬古漢字，測試數據集總量約為 1190 萬古漢字。

數據集的構建主要步驟：

（1）標點文獻採集，主要通過開源的古籍數據收集和自行標點來完成的；

（2）電腦分類整理，通過軟件程序設計對各類文獻進行分類、篩選和存儲；

（3）標點校對審核，主要對電腦分選過的數據再次進行人工核查；

（4）數據格式標準化，則是通過電腦將人工核查確認的數據通過軟件程序設計進一步處理，將數據中多餘的符號標注等全部去掉，只保留標點和文字，並轉換成規範的自定义数据格式。

通過上述流程獲得的第一個古籍自動標點數据集版本（GJAP_Dataset_V1.0）已經在GitHub 平臺公佈[4]。圖 4 為標點系統的 GitHub 平臺頁面，用於數據整理和規範處理的Python 代碼一併公開。

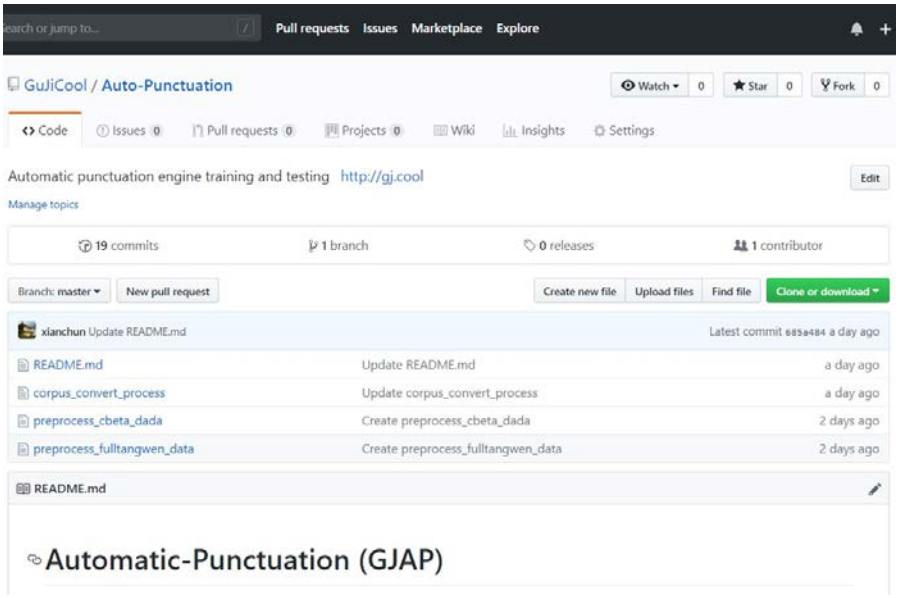


圖 4、 標點系統 GitHub 平臺頁面

4. 模型比較

4.1 模型訓練

基於 GJAP_Dataset_V1.0 數據集，LSTM 模型採用基於 ByTorch 修改的 LSTM 架構，CNN 模型為基於 Fairseq 的 CNN+Seq2Seq 訓練架構。引擎訓練之前還需經過一個特殊

[4] https://github.com/GuJiCool/Auto-Punctuation

的預處理，將原始数据集轉換為可用於時間序列的向量数据。圖 5 是基於 LSTM 模型訓練的引擎工作界面。

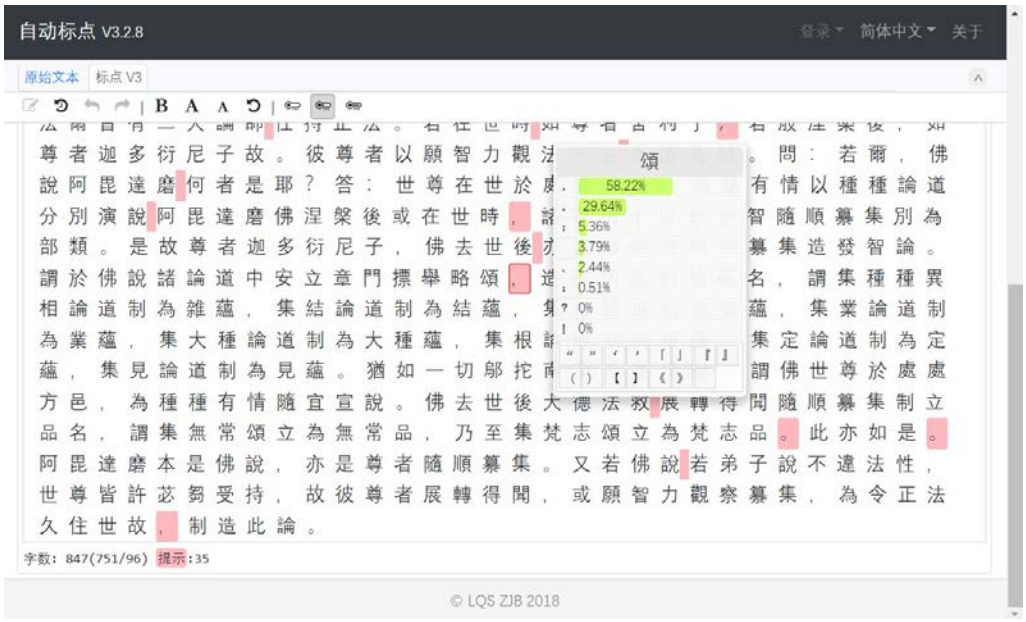


圖 5、LSTM 模型引擎標工作界面

4.2 引擎測試

測試數據分別選取南北朝、隋朝、唐朝、宋朝、遼朝和明朝等不同時期的佛教經典文獻，分別對二模型進行訓練和測試。表 1 是兩個標點引擎測試結果的對比。

表 1. 兩種標點模型比較

朝代	字數	有效標點數	LSTM 模型（%）	CNN 模型（%）
南北朝	1019	190.5	74.0	57.2
隋朝	689	103	90.3	53.4
唐朝	1020	194	94.3	69.1
宋朝	1020	180.5	75.3	51.8
遼朝	694	127	75.2	43.3
明朝	1014	169	65.7	46.7

4.3 結果討論

從表中的測試結果可以看出，LSTM 模型標點準確率在 65%-94%之間，CNN 的準確率在 43-70%之間。總體上，LSTM 模型的準確率高於 CNN 模型。標點準確率最高的是

唐代文獻數據，隋朝次之，明代文獻標點的準確率較低。

圖 6 為兩個模型準確率比較的柱狀圖。CNN 模型總體低於 LSTM，但二者趨勢一致，分析原因主要是訓練參數的優化有待進一步深入。不同時期文獻標點的準確度差異一部分來自於訓練集數據數量分佈的不平衡和品質上的差異。

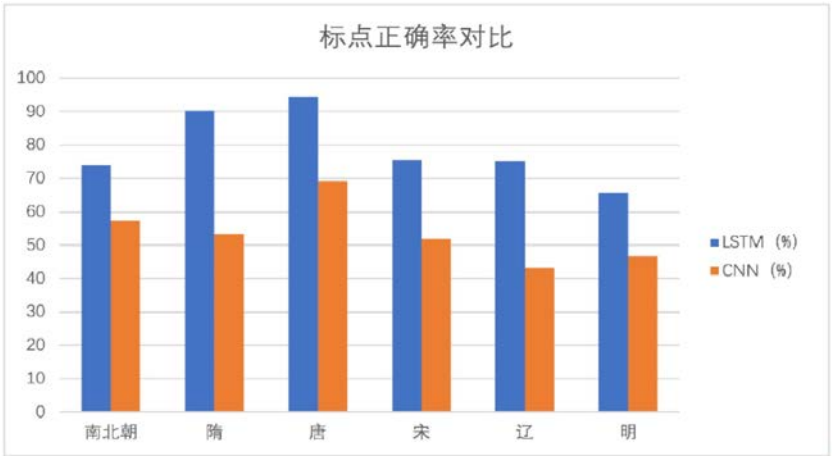


圖 6、LSTM 模型和 CNN 模型正確率比較

模型架構訓練實踐表明，LSTM 模型處理長文本序列的能力強，對 GPU 顯存要求較低，但因缺乏並行處理機制，計算速效率受到制約。Fairseq 可以進行並行處理，計算速度較快，但對於處理的文本序列長度則有限制，並且需要較大的 GPU 顯存，因此不同架構的模型對於標點訓練的適應性及模型參數的優化策略等，都需要進一步的研究和探索。

4. 結論

根據古漢語標點研究的特性，將自動標點的技術問題分別對應于兩種自然語言處理領域典型開源框架的應用問題；規範化流程構建的數據集，為模型訓練和測試打下基礎；兩種引擎的測試結果顯示，儘管自動標點引擎取得了最高 94%的標點準確度，但未來數據集結構的調整和數據品質的提升都等都將是下一代更優良標點引擎誕生的必要準備，同時訓練出適應不同歷史時期文獻、不同學術領域古籍文獻的高準確度獨立引擎，也是自動標點研究進一步努力的方向。跟蹤人工智能 NLP 領域的最新研究成果，不斷提升標點引擎的性能，讓智能標點工具更好地利益大眾閱讀，利益專業古籍工作者，這是我們研究的宗旨和目標。

參考

(1) 黃建年（2009）。《農業古籍的電腦斷句標點與分詞標引研究》。南京農業大學。

(2) 張開旭，夏雲慶，宇航 (2009)。 <基於條件隨機場的古漢語自動斷句與標點方法>，《清華大學學報》(自然科學版) 10，页 1733-1736。

(3) 王博立，史曉東，蘇勁松（2017）。<一種基於循環神經網路的古文斷句方法>，《北京大學學報》(自然科學版)2，页 255-261。

(4) Jachym Kolar, Lori Lamel, 2012. “Development and Evaluation of Automatic Punctuation for French and English Speech-to-Text”, INTERSPEECH ISCA's 13th Annual Conference Portland, OR, USA.

(5) Sutskever, I., Vinyals, O., and Le, Q. V. 2014. “Sequence to sequence learning with neural networks.” In Advances in Neural Information Processing Systems, pp. 3104–3112

(6) Gehring, Jonas, et al 2017. “Convolutional sequence to sequence learning”, arXiv, pp.1705.03122.

(7) Wang, L., Cao, Z., De Melo, G. & Liu, Z. 2016. “Relation Classification via Multi-Level Attention CNNs.” Acl pp.298–1307.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

清文全藏數據庫工程之意義與規劃

以人工智能技術為依托

张莉

中国第一历史档案馆研究馆员

清文全藏数据库工程之意义与规划

——以人工智能技术为依托

张莉

研究馆员

中国第一历史档案馆

摘要《清文翻译全藏经》是清代满文大藏经的原始名称，简称满藏。它是佛典法库中一部宝笈，是清朝盛世时期一项重大文化工程的结晶。本文旨在阐明满藏数据化的深广意义及总体规划。全文分三部分：

1、建立清文全藏数据库拟用版本

该部分利用中国第一历史档案馆档案，简述乾隆皇帝纂修清文翻译全藏经的缘由、目的、编纂过程以及清字经馆开闭馆时间、用印、馆址、官制设置及管理典制等。

2、建立清文全藏数据库的意义及规划

该部分从学术上及现代社会开发利用等方面，阐明建立乾隆朝清文全藏数据库的意义；并介绍拟试用人工智能技术手段迅速建立数据库的总体规划。

3、清文全藏数据库工程进度

介绍笔者利用满文藏经开展教学实践，为建立满文藏经数据库收集基础数据，以探讨人工智能识别技术引入满文藏经数据库进程。满文语音教学；语音教学实践。

关键词：满文，藏经，数据库

清文全藏数据库工程之意义与规划

——以人工智能技术为依托

《清文翻译全藏经》是清代满文大藏经的原始名称，简称满藏。它是佛典法库中一部宝笈，是清朝盛世时期一项重大文化工程的结晶。本文旨在阐明满藏数据化的深广意义及总体规划。

《清文翻译全藏经》是清代满文大藏经的原始名称，简称满藏。它是佛典法库中一部宝笈，是清朝盛世时期一项重大文化工程的结晶。本文旨在阐明满藏数据化的深广意义及总体规划。

上世纪七十年代，美国伯克利大学蓝卡斯特教授首先启动了高丽藏数据库的整理、著录工作。此后，日本、台湾等地佛教研究学者们纷纷加快对本地所藏汉文经藏的整理，并开展建立数据库，现已大见成效。

近年来，随着中国社会经济的发展，汉文藏经校勘、整理工作越来越为大陆专家、学者所重视。随着科学技术日新月异的发展，人工智能识别技术的日臻成熟，逐渐引入藏经整理校勘、建立数据库工程之中。这对加快藏经整理校勘、提高数据库工程建设的速度，将起到无法估量的作用。

然而某些客观因素，满文藏经数据库的建立，目前尚处在十分落伍的状态。台湾法鼓山文理学院师生在马德伟教授主持下，为满足满语言文学者们语言研究的需求，利用几年时间，对 2002 年紫禁城出版社以原本木雕模版刷印的满文大藏经进行扫描拍照，收集了含参考书目、满汉藏目录、满汉梵佛典语汇及 8 函影像照片，著录出数千条目。开了建立满文藏经数据库之先河。尽管如此，尚有 100 函满文藏经数据亟需收集、著录，并在此基础上对数据进行全方位研究。因此，加快建立满文藏经数据库工作，已迫在眉睫。本文分三部分向大会专家、学者汇报：

一、利用中国第一历史档案馆档案，简述乾隆帝纂修清文翻译全藏经目的及建立数据库拟用版本；

二、阐明建立乾隆朝清文全藏数据库的意义及迅速建立数据库的总体规划；

三、介绍笔者利用满文藏经开展教学实践，为建立满文藏经数据库收集基础数据，以探讨人工智能识别技术引入满文藏经数据库进程。

1、清文全藏数据库拟用版本

《清文翻译全藏经》是清乾隆五十五年内府刻本，全书共收 2535 部经卷。梵夹装，白口，每版页为 24.3*73cm，版框为 16.5*59cm; 四周印双边，朱丝栏 31 行，小字双行，版口刻有满汉文经名、卷次、页码。（见故宫博物院紫禁城出版社 2002 年出版印刷插图照片）



图 1、清文翻译全藏经外包装原样图



紅色暗花綢單袱
Red silk wrapping with veiled pattern



图 2、打开彩带去掉外包装、第二层图
图 3、去掉红包袱皮后的藏经



图 4、打开上下夹板后的藏经

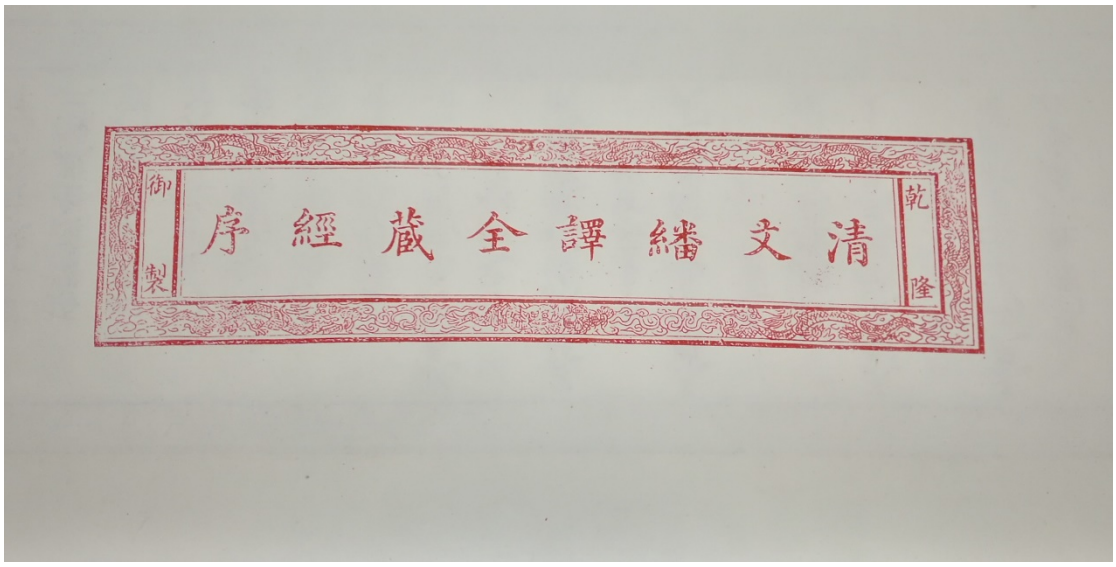


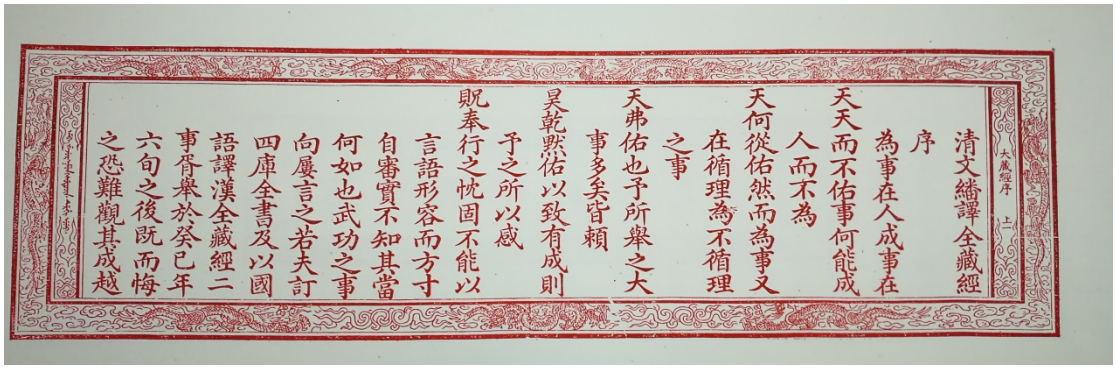
图 5、底板所刻佛菩萨图

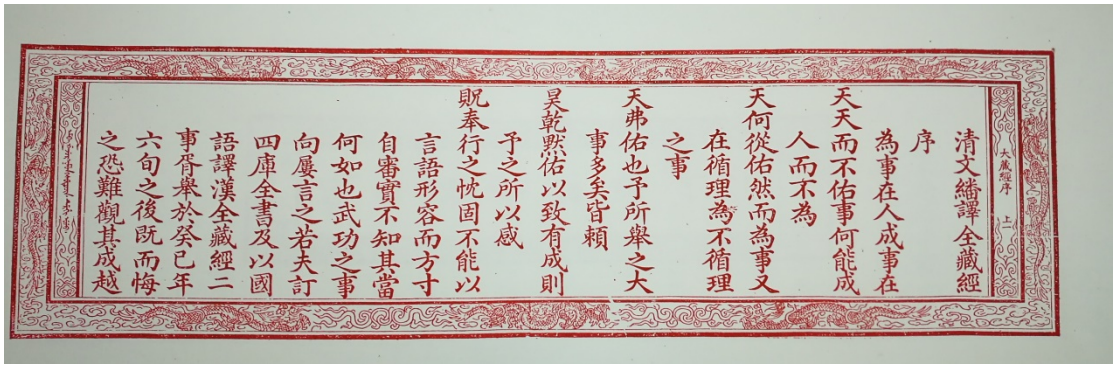
该部藏经存世虽已200余年,然因只印12份,且珍藏于宫中及满洲僧人所在庙宇,外人少能见到。2002年北京故宫博物院紫禁城出版社，为保护珍藏该院图书馆的清文翻译全藏经木雕刻版这一历史文物，利用所藏完整未受残损的部分藏经刻版，并用西藏布达拉宫所藏部分经书，补全故宫所藏经板之缺，重新刷印限量版满文大藏经20套，使原存不全之清文藏经得以凑全新版《满文大藏经》。在此基础上，为扩大发行，以惠更多社会需求，故宫博物院出版社即将再版2002版满文大藏经平装本。本工作室拟用该版本满文藏经，收集其的全部数据，以方便佛学及满学研究工作者及修行者们利用经典数据。

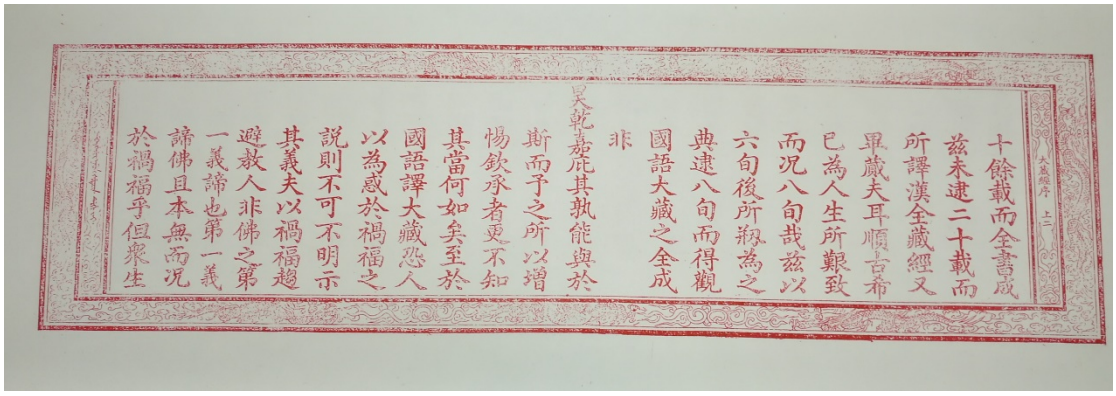
1.1 乾隆皇帝编纂满文藏经缘由

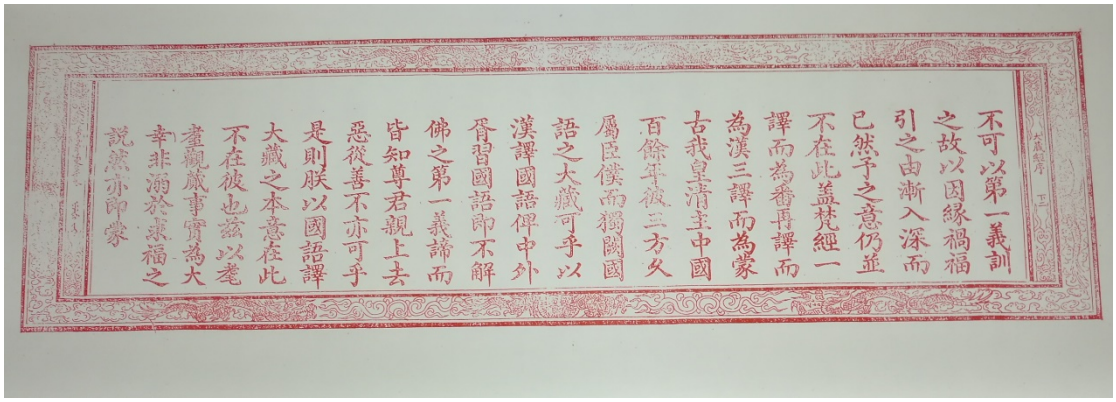
在清文翻译全藏经序中乾隆皇帝写道：盖梵经一译而为藏，再译而为汉，三译而为蒙古。我皇清主中国百余年，彼三方久属臣仆，而独缺国语之大藏，可乎？以汉译国语，俾中外胥习国语，既不解佛之第一义谛，而皆知尊君亲上，去恶从善，不亦可乎？是则，朕以国语译大藏之本意。（见乾隆序文照）

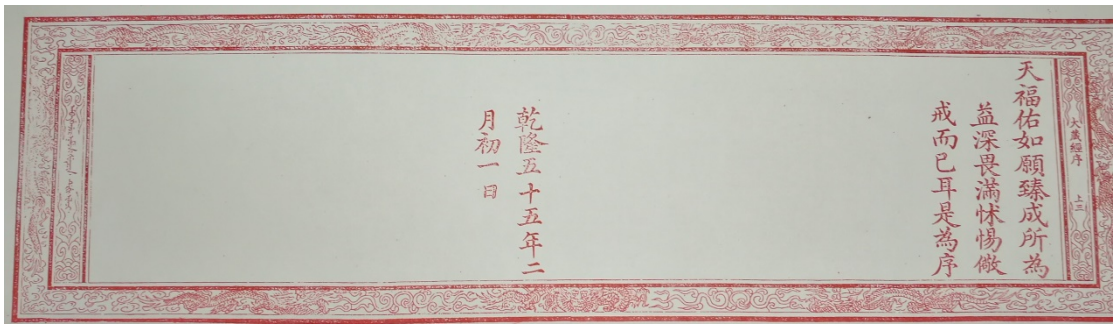





清文繙譯全藏經
序
為事在人成事在
天天而不佑事何能成
人而不為
天何從佑然而為事又
在循理為不循理
之事
天弗佑也予所舉之大
事多矣皆賴
昊乾默佑以致有成則
予之所以感
貶奉行之恍固不能以
言語形容而方寸
自審實不知其當
何如也武功之事
向屢言之若夫訂
四庫全書及以國
語譯漢全藏經二
事胥舉於癸巳年
六旬之後既而悔
之恐難覩其成越


十餘載而全書成
茲未逮二十載而
所譯漢全藏經又
畢藏夫耳順古希
已為人生所艱致
而況八旬哉茲以
六旬後所報為之
典逮八旬而得覩
國語大藏之全成
昊乾嘉屍其孰能與於
非
斯而予之所以增
惕欽承者更不知
其當何如矣至於
國語譯大藏恐人
以為惑於禍福之
說則不可不明示
其義夫以禍福趨
避救人非佛之第
一義諦也第一義
諦佛且本無而況
於禍福乎但衆生


不可以第一義訓
之故以因緣禍福
引之由漸入深而
已然予之意仍並
不在此蓋梵經一
譯而為番再譯而
為漢三譯而為蒙
古我皇清主中國
百餘年被三方文
屬臣僕而獨闕國
語之大藏可乎以
漢譯國語俾中外
胥習國語即不解
佛之第一義諦而
皆知尊君親上去
惡從善不亦可乎
是則朕以國語譯
大藏之本意在此
不在彼也茲以老
耄覩藏事實為大
幸非溺於求福之
說然亦即衆


天福佑如願臻成所為
益深畏滿怵惕儆
戒而已耳是為序
乾隆五十五年二
月初一日

不难看出，乾隆皇帝修撰翻译该经目的是国人均学汉语，却只知遵君亲上，去恶从善，不能明解佛之第一要义。故做翻译汉文经典而成满文。由此可知，以满文解释汉文佛语之真义，是乾隆皇帝翻译藏经的真实目的。

1.2 满文藏经编纂与翻译

根据中国第一历史档案馆所存档案记载，笔者总结概括满文大藏经编纂翻译的时间是：清乾隆三十七年（1772 年）至乾隆五十九年（1794 年）；该书由乾隆皇帝钦点组建清字经

馆，以汉、蒙两种文字的大藏经为底本，利用 22 年时间翻译、刊刻而成。清代又称《国语大藏经》或《御制清文翻译全藏经》。其目录、装帧形式及版框尺寸，均系内府朱色印本。

1.3 清字经馆

3.1 开闭馆时间：

据《清实录》及档案记载：清字经馆自乾隆三十六年十二月十五日（又乾隆三十七年正月初八日）开馆至嘉庆二十五年九月初六（己未）日裁撤，历时四十九年之久。

3.2 用印：

总理提调事务内阁学士博清额等案呈，前经军机处议奏条款内开，翻经事宜应咨各处之一切文移，俱钤用内务府印信。等语。于乾隆三十六年十二月十五日具奏，奉旨："依议。"钦此。钦遵在案。理合呈明知照内务府。所有本馆一切文移遵照原奏钤用内务府印信。[1]

由此可见，清字经馆对外办公使用总管内务府印章。

3.3 清字经馆馆址：

[咸安宫，门三间，殿五楹。左右配殿各三楹，为尚衣监。恭制御服于此。其后殿宇二层，旧为皇子所居。嗣为三通馆，为清字经馆，为恭修高宗纯皇帝实录馆，为皇清文颖馆][2]

由此可知，清字经馆设在西华门内咸安宫院内。（见照片）

3.4 清字经馆官制：

清字经馆设有总裁、副总裁、编纂、翻译、誊录、校对、刊刻经板、复勘、阅经总裁 、阅经副总裁、办理经咒喇嘛、校对经咒喇嘛、总校僧人、讲经僧人等官职，共计 96 员。以和硕质亲王永瑢、多罗仪郡王永璇、大学士和珅为总裁，吏部尚书金简及国师章嘉呼图克图为副总裁[3]负责管理该馆的编纂、翻译等事宜。

3.5 清字经馆管理制度：

据中国第一历史档案馆所存档案所藏档案，有关该机构的选官、任官、俸禄、奖罚等管理制度及修缮清字经馆房屋等内容的档案约有 250 余件，这些内容档案均有上述内容，言之有据，总括概述，在此不多考证，另文详论。

2、建立清文全藏数据库之意义及规划

[1] 中国第一历史档案馆藏 05-13-002-000027-0050

[2] 《嘉庆朝会典》卷 661 工部 26 页

[3] 乾隆三十八年二月初十日上谕

2.1 建立清文全藏数据库之意义

尽早建成清文全藏数据库具有十分重要的意义

首先，建立清文全藏数据库是对世界物质文化遗产的保护。

清代政权是以满洲贵族为核心的少数民族政权，至乾隆中期清王朝统治已逾百余年，时藏、汉、蒙等文字藏经均已刊行，惟缺满文藏经，乾隆皇帝认为这是颇感难堪的缺略。满洲统治者为确保本民族上层贵族的利益，使其政权得以永固，乾隆皇帝不断降下谕旨，强化本民族传统文化，将翻译、刊刻满文藏经与纂修《四库全书》、“十全武功”之记述视为同等重大之事。

为此，乾隆皇帝专设清字经馆，倾力调集各种专业人员参与翻译、刊刻藏经；据档案载，清字经馆翻办、刊刻、刷表、装潢全藏大盘若等经及律戒行经，共 2519 卷，计 108 套（函），每套（函）刷表十二分，共计 1296 套。自三十七年开馆起至五十八年计二十一年间，所有刻字、爆版、漆边、刷表、绘画佛像等項工价及办买纸张、笔墨等项物料并纂修、翻译、收掌、誊录、校对和尚、喇嘛、供事、苏拉等人员之饭食及恩赏闲散人员之钱粮，共用银 591000 余两。（中国第一历史档案馆藏军机处录副奏折：03-0195-3450-017）

清文全藏经翻译、刊刻的完成，是清乾隆时期的一大文化壮举，是目前传世唯一一部完整的满文藏经。然时印刷仅为 1 2 套，并深藏内宫及各皇家寺院，且受民族文字局限，故未在世间流通传播。故当时《清文翻译全藏经》在宗教、文化方面的意义远逊于政治意义，或者说该藏经的编纂，实际是为统治蒙藏需要，而藏经本身却成为维护清统治者民族自尊的文化象征物。时至今日，清文全藏的存世是国家、民族、人民大众的瑰宝；也是世界人民的珍贵宝藏，确有重要的历史价值和文物价值。

由此可知，清文全藏经的翻译、刊刻成书，在佛经译经史乃至清代文化史上占有重要地位。该经译纂工程浩大，从翻译到雕版印刷，经函装潢，无一不代表清代当代图书的最高水平。这一倾力之作，能够保存至今，在学界无不拍手称赞。

尽管如此，各种客观原因，致清文全藏至今尚未建立完整的数据库，远远落伍于汉文藏经数字库的建立。因此，清文全藏数据库之工程凸显其意义重大，刻不容缓。

其次，清文全藏是中华民族传统文化的一部分，建立该数据库是对中国传统佛教文化的继承和丰富。佛教经典自东汉传入中国以来，在流传过程中，被逐渐翻译成多种民族文字，由于满文藏经的翻译直接汲取蒙汉经典中的素材，且产生于清代国家统一多民族文化融合的背景之下，因而，满文藏经数据库的建立将为汉文藏经的深入研究，提供重要的印证和补充。以往佛典研究者们多数受限于不识满文字，故未能充分认识满文藏经对汉文经典的丰富、补充作用。总之，清文全藏数据库的建立在继承、丰富民族文化方面，将起到不可估量的作用；同时，促进深入探讨研究各民族间的文化融合。

第三，建立清文全藏数据库是对濒危语言满语的最好传承。

清文全藏翻译过程中，使用了多种民族语言的借词，其中包含梵、汉、藏、蒙古等多种语言文字，这些历史语言痕迹，为研究满语词汇，提供了不可或缺的丰富语料。且在数据库建立之后，可开发传承满文有声藏经，这将更加益于广大民众深入学习经藏。

2.1 建立清文全藏数据库之规划

综上建立清文全藏数据库意义，我们得知，做周密严谨的工作规划十分必要。为此，本工作室做如下规划并将在今后工作中，根据实际情况随时进行有效修订。

2.1.1 培养满文人才、积累以佛教经典为研究对象的满文教学经验，逐渐建立研究满

文藏经的团队。

开设满文培训班进行集中培训，是建立满文藏经数据库团队的基础。教材以扫描的满文藏经影像为材料，以罗马转写为语音练习，逐条核对清文全藏目录与经文；通过读诵经文，培养满语语感，尝试对扫描版原文进行校勘，逐字厘清藏经中“脱”“衍”“倒”等问题，制作校勘表。

在此基础上，培养新的教学人才，编写以满文佛典为教学资料的满文教材。设计课程、制作教学视频、分享教学成果。

2.2 建立数据库工作流程

2.2.1 扫描收集新版藏经电子影像数据； 对扫描文档进行逐版核查，必要时重新编目。

2.2.2 人工完成满文藏经罗马字符转写，准备人工智能引擎训练所需数据。

2.2.3 运用满文影像图片及转写数据训练人工智能引擎，识别满文字符图片。

形成包含佛教用语的满文输入法国际字符集，整理收集包括阿礼嘎利字在内的满文佛教词汇，形成完善的满文佛教用语字词库。同时，不断提高人工智能引擎识别及转写的准确率。

2.2.4 建立清文全藏数据库

本数据库包括满文藏经原文图像、满文字符文档及罗马字符文档、声音文档。建立该数据库步骤：

首先，利用人工将藏经原文图像转写成罗马字符，为智能识别提供可靠数据，通过智能识别满文影像，建立满文字符文档及罗马字符藏经文档数据库。

其次，设定满文识别条件。即将罗马字、满文藏经文档，以人工设置满文藏经翻译汉文识别条件，建立对应的汉文翻译系统，形成汉文数据，利用翻译数据训练人工智能，进行满汉文档机器翻译。而后利用人工，校对机翻结果，结合对人工智能的调参，提高机器翻译准确率。形成清文全藏汉文文档数据库，并与汉文藏经数据，实现对应检索数据库。

第三，以满文藏经汉译文档数据为核心，满文佛教用语字词数据为辅助，对满文藏经与汉文藏经文本数据进行标注。人工与人工智能互相结合，建立起满文与汉文藏经文本间的对应关系，为跨语言文本对读，准备好更高级别的数据库。

第四，着手收集满文藏经读诵，录制音频数据，形成音频数据库。同时将数据用于人工智能引擎训练，使机器具备识别满文语音的功能，如此，既可保留满文的语音资料，又能为学习满语文提供智能化学习材料。

第五，制作开放的翻译、句对、读诵、校对的满文佛教文献校勘网络众包平台。为满文佛教文献研究，不断持续发展，奠定基础并创造条件。

3 清文全藏数据库工程进度

如前所述，本工作室拟用北京故宫博物院新出简装版满文大藏经，然因该书尚未刷印完成，笔者先采用台湾法鼓山文理学院网络图书馆满文大藏经数据，进行教学实践。为建立清文全藏数据库收集基础数据，做前行准备。试图探讨人工智能识别技术，引入建立

清文全藏数据库，加快数据库完成，为此，笔者工作如下：

3.1 满文语音教学：

本年六月至七月开设满文教学实验班，参加该培训班有 10 人，通过教授满文语音课程，大家初步掌握了满文 108 个满文复合字及 24 个特定字以及 22 个辅音字中、字尾单独使用的用法。（附上课照）



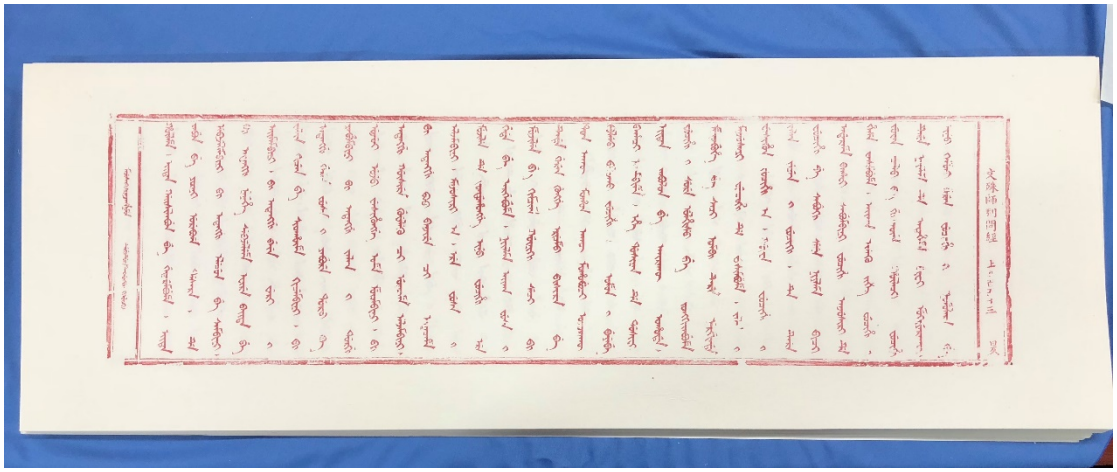




3.2 满文语音实践

本工作室学习宗旨是学以致用，在语音教学结束后，工作室于八月份开始实践。利用法鼓山文理学院收集的满文大藏经 52 函 PDF 文档，进行罗马字转写，以巩固所学知识。实践中，我们采取边实践边总结的形式，发现问题及时解决。在此过程中得到北京故宫博物院图书馆的鼎力支持，利用该馆所藏 2002 年版满文大藏经，进行逐页校对，以总结转写中存在的问题。（附校对工作照）





本工作室崔文治博士在总结转写体会时说道：满文学习从学语言角度看，可用易、难二字来概括。所谓易指满语属于音系简单、语法简单的语言，利于转写。所谓难——满文是辨认易错的文字，容易混淆倘若能若熟悉相关词汇，会极大提高转写效率。马欣先生则将崔文治上述的高度总结进行了分解，认为“转写的过程就是扩充积累满文数据库，开阔视野的过程；他用转写中若干例词指出其错误的原因。为提高转写速度，建议整理常用词及高频易错词汇对照表。（附易错词汇表）

综上教学情况，我们试将满文 52 函中金刚经、心经、救护日食经、部分央掘魔罗经转写成罗马字，以备建立清文全藏数据库提供必要数据。从而加快建立数据库速度。

简单结语

综上所述，建立清文全藏数据库意义重大。

从历史意义上分析，建立数据库是对清文藏经的继承和保护。

众所周知，满语是世界濒危语言之一，满文是清朝统治者使用现已成为基本无人使用的死亡文字。所遗留的满文藏经是我们从事佛学研究的最好素材之一；在从事清文全藏数据库工程过程中，参与者不仅可以深入经藏研究满文字，更重要可培养一批满文研究人才。当这些人才与人工智能技术相结合时，即可在较短时间内，突破专业人员匮乏及人工操作速度缓慢的问题。

同时，机器可替代人工重复性工作，可将清文全藏数据转化成多种数据库。并以该数据为基础，铺设校勘网络众包平台，以快速提升数据质量，促进满文佛教文献人才的培养。为后续人工智能技术应用、满文佛教文献研究、文化认同、文化传播等深层次研究，提供客观基础。本文讨论重点并非某种具体人工智能技术，在满文佛教文献数据化方面的应用；而是为今后快速完成清文全藏数据库工程跨出出一大步，探索人工智能技术，在满文藏经应用上的场景，呼唤科技与人文更广泛深入的结合。

不难看出，清文全藏数字化、智能化，是人工智能与人类智能紧密合作的过程及取得的结果。它并非是人类与机器的角逐，而是人与人的合作——即掌握满文的人与掌握机器语言的人之间的合作。这不是人类智能的危机，而是人类意识发展的新阶段——即以满文佛教文献研究为纽带，将民族文化、网络科技、佛教智慧以及现代智能结合为一体。这是科技时代人类文化的一大创举。

人工智能识别技术，在建立清文全藏数据库中的利用，为单纯以人力著录藏经数据库检索系统，提高了速度。倘若没有人工智能技术的支持，完成如此浩大的工程，将会推迟几十年。因此，人工智能技术的发展，必将给清文全藏数据库的建立提供有力支持。

近百年来，佛学界研究高度职业化趋势有增无已，各种版本的汉文藏经，陆续整理出版。在这一形势下，建立藏经数据库对保护传承古典藏经文献等具有非常重要的学术意义，参与者们可将整理成果作为研究成果，维持其学术生涯与世间生活。这将大大推动藏经文献整理事业，使几代学者投入其中，大量藏经文献得到整理。然而，当主要藏经文献基本整理完毕后，研究者们将片面求冷求僻，钻入象牙塔式研究，维系学术壁垒，这固然能在一段时期内维护部分研究者的利益，却不是学术研究真实意义所在。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

自動分群應用於傳記人物關係建立

The Application of Automatic Clustering in the Construction of Relationship between Historical People

謝順宏

國立臺灣師範大學圖書資訊研究所博士生

自動分群應用於傳記人物關係建立

The Application of Automatic Clustering in the Construction of Relationship between Historical People

謝順宏	柯皓仁	張素玢
博士生	教授	教授
國立臺灣師範大學	國立臺灣師範大學	國立臺灣師範大學
圖書資訊研究所	圖書資訊研究所	臺灣史研究所
mayh@ntnu.edu.tw	clavenke@ntnu.edu.tw	109692@ntnu.edu.tw

摘要

人物是歷史學研究的重要基礎，舉凡人物的個性、家庭背景、經歷、社會階層，甚至於整個社會的階層流動，以及親族與社會網絡等都是歷史研究的議題。本研究擬針對歷史人物的社會網絡進行探究，因社會網絡形塑歷史人物所處時代的背景、相關人物，甚至從中可探究其活動經歷，提供歷史學家評估歷史人物於何時、如何，及為何利用親族與非親族關係的參考。本研究擬以臺灣歷史人物資料庫(Taiwan Biographical Database, TBDB)中所收錄之 887 位彰化縣人物(簡稱傳主)的傳記全文(簡稱傳文)中探索傳主的社會網絡關係。考慮到僅由單一傳文中發掘，或透過已知之關係或共同參與之共同社群來建立起不同人物之間的連結，恐無法完善的描述傳主與其他人物之間的關係，或提供歷史學者另一個「未知」的檢視角度。有鑑於此，本研究旨在探討一種藉由共同特徵值，自動化合併多位傳主社會網絡的方法，透過自動階層分群以建立相關社會關係網路的方式，避免需經人工大量審視與建置之成本，同時又能確保自動建立網路具有語意上的關連。

關鍵詞

自動分群，臺灣歷史人物資料庫，文本探勘，社會網絡分析。

一、前言

人物是歷史學研究的重要基礎，舉凡人物的個性、家庭背景、經歷、社會階層，甚至於整個社會的階層流動，以及親族與社會網絡等都可是歷史研究的議題。本研究擬針對歷史人物的社會網絡進行探究，因為社會網絡形塑了歷史人物所處時代的背景、相關人物，甚至從中可探究其活動經歷，提供歷史學家評估歷史人物於何時、如何，及為何利用親族與非親族關係的參考。社會網絡關係分析家發現，任何人都須從不同的社會網絡關係中、不同的人身上，尋求情緒與經濟等各方面的支持。因此，僅研究人們如何於危機時刻利用親族關係已不足夠；相反地，歷史學的研究必須涵蓋過去人們如何為不同目的而利用親族與社會關係，以及此利用關係的優勢與限制(Wetherell, 1998)。

二、文獻探討

本研究擬以臺灣歷史人物資料庫(Taiwan Biographical Database, TBDB)中所收錄之887 位彰化縣人物(簡稱傳主)的傳記全文(簡稱傳文)中探索傳主的社會網絡關係。資料來源：《新修彰化縣志》本身考證嚴謹，且人物在擇選撰寫之初，即有明確的規範：[1]

1. 彰化縣出身或對彰化縣有貢獻，而且已經在清代《彰化縣志》或文獻中所記載，包括拓墾人士、官吏、平亂者、地方經理董事等。
2. 日治時期已列入人士鑑，或是政府公職人員，在《臺灣總督府公文類纂〉、《臺灣總督府專賣局檔案》、《臺灣總督府職員錄》有所記載，或會社要覽等等相關資料。
3. 戰後已收入志書、各類人物傳記、名錄、歷史辭典等等之人物。
4. 已收入各鄉鎮市方志之彰化人士，或編纂中的鄉鎮志已收錄的人物，並具有跨

[1] 張素玢、李毓嵐、顧雅文、李昭容(出版中)。總論。新修彰化縣志：人物志・文化人物篇。

鄉鎮影響力之人士。

5. 本縣各鄉鎮推薦之人士。

但考慮到僅由單一傳文中發掘，或透過已知之關係或共同參與之共同社群來建立起不同人物之間的連結，恐無法完善的描述傳主與其他人物之間的關係，或提供歷史學者另一個「未知」的檢視角度。有鑑於此，本研究旨在探討一種藉由共同特徵值，自動化合併多位傳主社會網絡的方法。在先前的研究中(謝順宏、柯皓仁、張素玢，2018)，研究者已藉由探勘傳主之傳文內容，嘗試建立起該傳主之社群網路圖，本研究試圖探索透過任二傳主之間的共同特徵或相關人物，建立起傳主之間之關係網絡，並擴大原本僅傳主與傳文內容提及人物之社群關聯，提供歷史學者另一個研究的角度。

本研究擬透過結合外部資訊與自動化概念階層分群(Automatic Concept Hierarchy Generation)，藉此建立任二相異傳主之關聯，擴大其社會關係網絡，並相對減低人工逐一審視並建立相同傳主網路之負擔。最後，再由專家檢視結果，以評估其關係建立正確與否。

人工建立傳主之社會關聯網絡是件繁重的工作，需逐一審視與歸納、統整各傳主之相關人、事、時、地後，才能開始繪製關聯網路；這個過程不但是耗時且可能會有不一致產生，亦無法被廣泛應用。自動階層分群法(Hierarchical agglomerative clustering)在近年有相當不錯的發展(Jain & Dubes, 1988; Jain, Murty, & Flynn, 1999)，而階層分群法(Widyantoro, Ioerger, & Yen, 2002)其最終可以產出分群結果，並以樹狀結構呈現，且仍可保持高擴展及一致性，唯僅依靠自動處理，可能會導致分類結果不具任何意義，而無法解讀。故若能結合二者之特長，透過給予特定之概念或特徵引導進行自動分群，並建立人物之間彼此的關聯，應可減少人力成本，同時具有可擴充與效率上的優點。

表 1 應社成員名單

應社成員名單
賴和、陳滿盈（虛谷）、楊樹德（笑儂）、楊木、陳英方（陳渭雄）、楊石華、王卻是（王克士）、吳蘅秋、石錫勳、楊添財（楊天佑）、楊松茂（楊守愚）、詹阿川（詹作舟）、楊宗城、詹阿本、施澄江（施江西）、曾未定（曾材庭）、高火順（高泰山）、王桂木

資料來源：本研究整理

三、方法

本方法所提出的方法，採先抽取傳文中的特定元素，例如：學經歷、社團、居住地點、傳文內提及人物。將傳文重新組成作為描述傳主 P 之特徵向量。部份內容如學歷、參與社團，可能已有既存之外部文件資源可供使用，例如詩社成員名單或創辦者名單……等等，如表 1 為應社成員名單。本研究中將引入這些外部資源，作為擴展資料來源，以增加現有資料內容，同時也可以讓自動分類結果朝向更具有語意上的關聯方向進行。透過計算不同傳主之特徵向量，計算其相似程度(similarity)，逐一進行合併，最終可合併成一階層分類結果。階層分類結果為一二元樹，本研究需另外訂定一個分割(partition)策略以進行結果分群，用以判斷最終二傳主之間是否需建立聯結關係，這時依其在分類結果樹上所處的距離長度，以做為判斷二相異傳主可否建立聯結關係，以產出之社群關係圖表。最後再交予歷史學者，檢視其正確性。具體實施的步驟詳細，如圖 1 所示，說明如下：

1. 概念的抽取與擴充：

   透過取得傳文之中，重要關鍵字、名詞片語、專有名詞……等等，並轉置為一傳文之特徵向量$\mathrm{P}\{f_1, f_2, f_3, \ldots, f_n\}$，f 表示其特徵值，而外部資源文件集$\mathrm{C}\{D_1, D_2, D_3, \ldots, D_k\}$則可視為一種詞彙的擴充，以擴展查詢詞的方式，提供並作為特徵f_n額外的補充資

訊，以豐富其內容。在引入並對應這些資源後，使得原有向量空間可擴充為

$\{f_1 = D_1, f_2 = D_2, f_3 = D_3, \dots, f_n = D_k\}$原先特徵向量則可為$P'\{W_1, W_2, W_3, \dots, W_n\}$，W為擴充後的結果。

2. 相似度計算與分群

透過計算任二相異傳主(P_a, P_b)之特徵向量，則

$P_a = \{W_{a,1}, W_{a,2}, W_{a,3}, \dots, W_{a,i}\}$，$P_b = \{W_{b,1}, W_{b,2}, W_{b,3}, \dots, W_{b,i}\}$透過 cos 夾角餘弦公式，則$\mathrm{sim}(P_a, P_b)$可計為

$\mathrm{sim}(P_a, P_b) = \frac{\sum_{i=1}^{k} W_{a,i} \times W_{b,i}}{\sqrt{\sum_{i=1}^{k} W_{a,i}^2} \times \sqrt{\sum_{i=1}^{k} W_{b,i}^2}}$，$0 \le \mathrm{sim}(P_a, P_b) \le 1$。

以二節點(a,b)之間相似度取最小值，作為自動分群的節點是否可合併的依據；最終會產出樹狀階層分類的合併結果，我們再依分類結果樹上之所處距離長度，作為判斷是否節點之間是否可以建立連結關係。

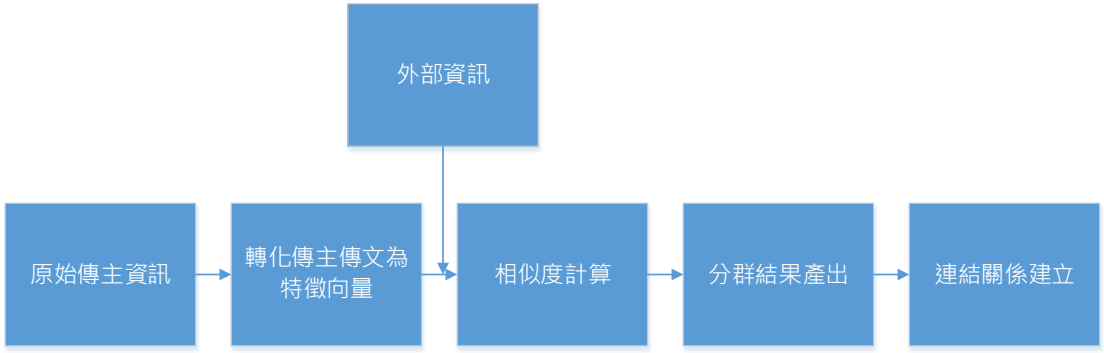


圖 1 實施步驟

針對分群結果，研究者透過抽樣比對原始傳文內容，以判斷結果之正確性外，另外特別針對分群結果，繪製網絡圖表，供歷史研究者分析，如圖 2。研究者除顯示傳主的關聯自動建立成果外，另外也標示造成關聯建立的特徵屬性為何，供人工判讀分析，以檢驗正確性。

四、結果與分析

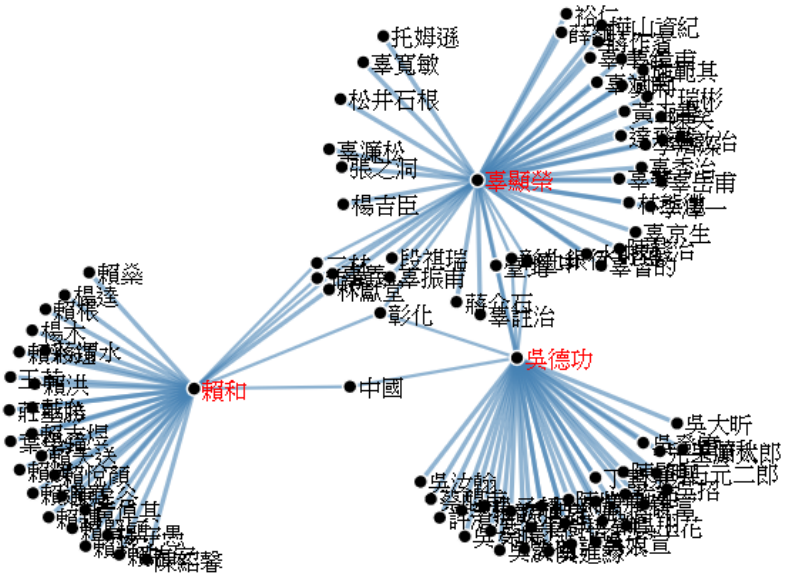


圖 2 自動關聯建立成果

如圖 3 所示，傳主間除了因活動地點具相同特徵，而建立起關聯外，針對特定人士或擔任職務也有一定關聯性，而產出最終的合併結果。但實際查閱傳文後可發現，傳主間的生活年代有一定的時間差；以顏志銘與陳質芬為例，生活的年代前者為 1950-2011，後者為 1873-1926，並沒有重疊的情況！但卻因有相似的特徵而產生關聯。造成這個合併的結果原因，起因為原始傳文中缺乏時間屬性，故無從得知傳主們實際活動的時間區間，例如：詩社參與的時間。而僅使用特徵的相似程度以判斷關聯可否建立，在缺乏時間維度的情況下，存在有一定的誤差。但本方法仍屬有效，且能快速的提供指定傳主間的網絡狀況，以供歷史學者進行檢視及概覽，或可進而決定研究的面向。在扣除地點資訊後，顏志銘與陳質芬之間網絡，可發現是經由李讀與吳在琨二人，其中李讀據傳文之描述，係邀請顏志銘至員林推廣拳擊運動，而與吳在琨均共同擔任過彰化縣議員，如圖 4。圖 5 則以施讓甫、朱啟南與王友芬及賴和，計四位傳主進行查詢，尋找其彼此間可

能關聯，可看出施讓甫與王友芬除了互相之間存有直接關聯之外，亦可能存在共同的朋友施梅樵，而王友芬與朱啟南，可能都有參與多個相同的詩社，也可能有共同的朋友洪寶昆，賴和則是因臺灣新民報的緣故與王友芬建立起關聯。

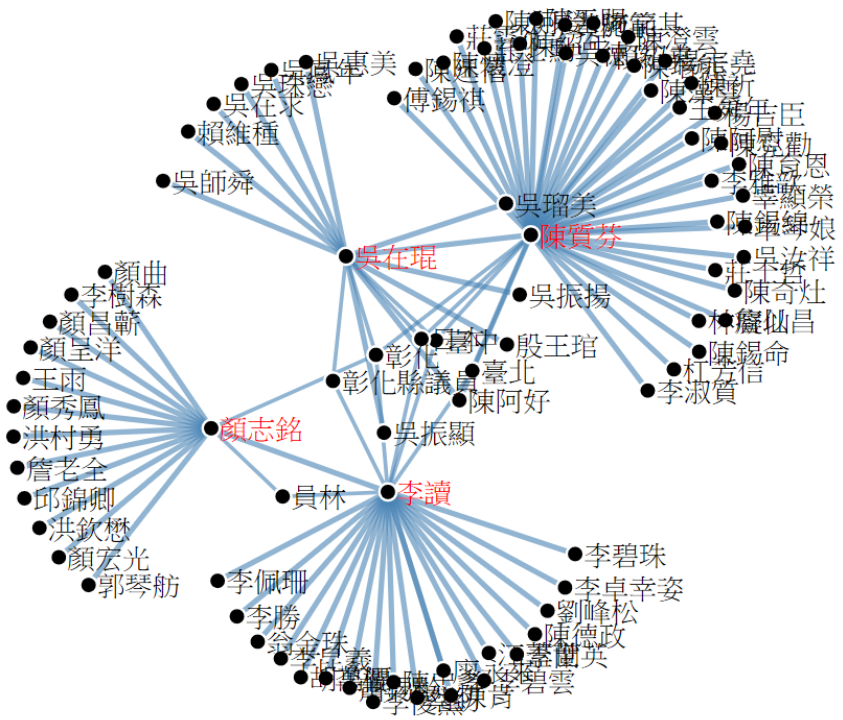


圖 3:傳主顏志銘、李讀、吳在琨及陳質芬之社群網絡

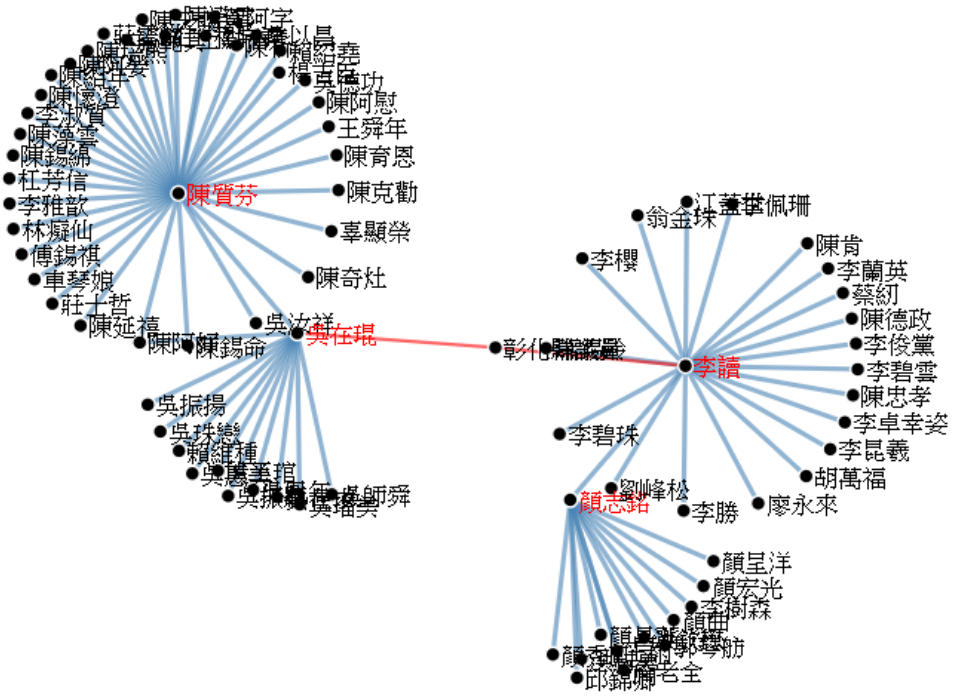


圖 4:扣除地點資訊後，傳主顏志銘、李讀、吳在琨及陳質芬之社群網絡

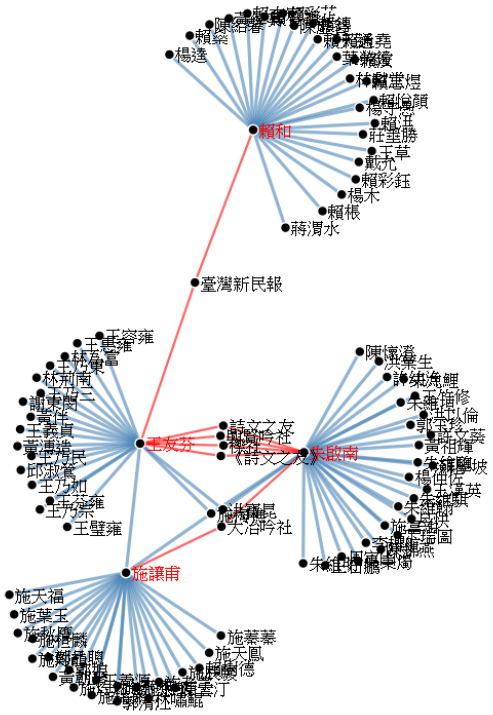


圖 5 詩社成員間的關係

五、結語

本研究嘗試提出一個透過建立傳主之特徵值向量，透過自動階層分類以建立相關社會關係網路的方式，避免需經人工大量審視與建置之成本，同時又能確保自動建立網路具有語義上的關連。其結果也成功且能快速的建立起指定傳主間的網絡圖表，供歷史研究者在研究這些傳主時的參考；另一方面，由於欠缺傳主的時間參與資訊，本方法僅能提供一個概觀的面向，但無法確切的掌握任二相異傳主，是否實際有所關聯或相識；但應足以提供另一種檢視的角度，供研究學者們參考。

歷史人物之社群關係為研究者所關注的焦點，而透過自動建立與合併社群，不但可以減少人力，同時也可以快速提供一個新穎的研究檢視角度供研究者掌握與解讀。針對後續研究，仍有部份議題尚待解決：

1. 時間序列：歷史的脈絡與時間習習相關，特徵的處理也是，在研究的下階段，或可搜集與合併時間因素於抽取的特徵向量之中，完善各事件與相關人物的參與時間區間，以增進合併的準確性。
2. 人名重覆：傳文中普偏存在人名重複的狀況，但在缺乏其它資料輔助的情況，尚無可靠的方式可以判斷是否為同一人物或僅僅為重名；為此，除希望能找到更多的資訊以輔助進行區隔人名之外，需再發展一可靠的方法，以進行人名除重的步驟，以讓形塑的傳主社群網路更加正確。

參考文獻

謝順宏、柯皓仁、張素玢 (2018)。〈臺灣歷史人物文本檢索與探勘系統之建置〉。《圖資與檔案學刊》(92)，頁 67-87。

Jain, A. K. & Dubes, R. C. (1988). “Algorithms for clustering data.” Prentice Hall.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). “Data clustering: a review.” ACM

computing surveys (CSUR), 31(3), 264-323.

Widyantoro, D. H., Ioerger, T. R., & Yen, J. (2002). “An incremental approach to building a cluster hierarchy.” In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on (pp. 705-708). IEEE.

Wetherell, C. (1998). “Historical social network analysis.” International Review of Social History, 43(S6), 125-144.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院 Dharma Drum Institute of Liberal Arts (DILA)

第 18 屆到第 29 屆台灣金曲獎最佳國語男女歌手提名人創作角色分析

Creative Role Analysis of the Best Male/Female Mandarin Singer Nominees in Taiwan Golden Melody Awards (2007-2018)

范蔚敏*　唐牧群**

國立台灣大學圖書資訊學系博士生*

國立台灣大學圖書資訊學系教授**

第 18 屆到第 29 屆台灣金曲獎最佳國語男女歌手提名人創作角色分析

Creative Role Analysis of the Best Male/Female Mandarin Singer Nominees in Taiwan Golden Melody Awards (2007-2018)

范蔚敏
博士生
國立台灣大學圖書資訊學系

唐牧群
教授
國立台灣大學圖書資訊學系

摘要

本研究以第 18 屆到第 29 屆最佳國語男女歌手獎提名人為研究範圍，想要分析金曲獎最佳國語男女歌手提名人的創作角色，從歌手在專輯歌曲上的創作型態切入，分析最佳國語男女歌手獎項提名人除擔任歌曲演唱者外，是否會更進一步獨立進行詞曲創作與音樂製作，成為身分多元的創作歌手。還有最佳國語男女歌手提名人與音樂工作者之間的網絡，以及歌手身分的多元性與其在音樂工作者社群中的位置之間的關係。

歸結欲探討之研究問題共有三點：(1)以社會網絡分析法探討台灣最佳男女歌手入圍專輯音樂工作者社群樣貌、(2)金曲獎最佳男女歌手提名人之身分多元程度與其參與音樂創作工作之類型、(3)金曲獎最佳男女歌手提名人之身分多元程度與網絡地位之相關性。

整體來說金曲獎最佳國語歌手音樂工作者社群網絡相當有規模(Diameter=7)，且社群成員連節程度高(Average Path Length=2.156)，但是有別於小世界網絡現象，整體網絡平均群聚係數數值相對較低（Avg. Clustering Coefficient=0.02），顯示音樂工作者彼此之間合作對象多元，且並未偏好與特定音樂創作人合作之情形。另以每位音樂工作者合作程度進行分群，顯示整體網絡有明顯分群，且各分群社群網絡凝聚程度較高（# of modules=22，Modularity Score=0.716）。

本研究統計結果顯示歌手參與音樂創作工作主要是以作詞為主（77.61%），作曲類型（76.12%）與監製類型（61.19%）次之，編曲類型（49.25%）最少。另外，本研究利

採用「赫芬達爾指數（Herfindahl-Hirschman Index）」計算歌手身分多元性，以演唱、作曲、作詞、編曲與監製五種音樂工作類型作為計算指標，分數越高則身分多元性越低代表歌手性格較為突出，分數越低則代表歌手身兼多重身分。並利用赫芬達爾指數對於中心性程度進行逐步回歸分析，結果呈現顯著負相關，表示歌手身分多元程度越高，越能居於整體社群網絡的中心性位置，換言之，身分多元的創作歌手在進行音樂創作時，可以更容易觸及其他音樂工作者，並且較會尋求其他音樂工作者共同創作音樂作品。

目次

1. 前言
2. 文獻探討
   - 2.1 金曲獎
   - 2.2 創作歌手
   - 2.3 流行歌曲創作工作類型
3. 研究設計
4. 結果分析
   - 4.1 金曲獎最佳國語男女歌手獲提名之專輯音樂工作者社群樣貌
   - 4.2 金曲獎最佳國語男女歌手提名人之身分多元程度與其參與音樂創作工作之類型
   - 4.3 金曲獎最佳國語男女歌手提名人之身分多元程度與網絡地位之相關性
5. 結論
6. 參考文獻

關鍵詞

金曲獎，最佳國語男女歌手，音樂工作者，社會網絡分析，中心性

1. 前言

台灣音樂產業市場的高度自由與多元開放之特性，使其成為華語流行音樂產製中心，加上台灣政府為鼓勵音樂產業設立金曲獎，並於 1997 年第 8 屆開始不限歌手國籍，開放受理任何於台灣市場出版之音樂作品，使得金曲獎成為其他地區華語音樂人共同關注之盛事。本研究想要了解金曲獎最佳國語歌手提名人的創作角色分析，以及最佳國語歌手提名人與音樂工作者之間的網絡關係，而提名人創作角色分析主要是從歌手在歌曲上的創造型態切入，想了解最佳國語歌手獎項提名人除擔任歌曲演唱者外，是否會更進一步獨立進行詞曲創作與音樂製作，成為身分多元的創作歌手。以及歌手身分的多元性與其在音樂工作社群中的位置，歸結欲探討之研究問題共有三點：(1)以社會網絡分析法探討台灣最佳男女歌手入圍專輯音樂工作者社群樣貌、(2)金曲獎最佳男女歌手提名人之身分多元程度與其參與音樂創作工作之類型、(3)金曲獎最佳男女歌手提名人之身分多元程度與網絡地位之相關性。

2. 文獻探討

2.1 金曲獎

本研究以金曲獎作為研究範圍，主要原因有三點，說明如下（文化部影視與流行音樂產業局，2018）：

(1) 行政院新聞局於 1990 年設立金曲獎發展至今歷史悠久，是具有傳統的評選比賽，加上聘請專業音樂人作為評審，評選方式有別於產業界普遍採用唱片銷售量或聽眾票選之評選方式，使其頒布結果具有公信力。

(2) 金曲獎項除最初設立之 11 種獎項，後續因應時代發展與產業需求陸續增設獎項，2003 年第 14 屆開始獎項分列國語、台語、客語、原住民語四種語言獎項，並依語言各別頒發最佳男女歌手，而 2007 年第 18 屆更將流行音樂細分為演唱類與演奏類，各類型音樂作品代表獎項發展成熟。

(3) 參賽資格公平且參賽作品數量具一定規模，其獎項提名作品與獲獎作品具有代表性。自 1997 年第 8 屆開放參賽者國籍限制，任何音樂作品於台灣出版皆可報名。

2.2 流行音樂創作工作類型

流行音樂工作依專業與職別可分為 12 類，如：企劃人員、歌手、製作人、作詞人、作曲人、編曲者、樂手、錄音師、混音師、母帶後期處理工程師、生產商、宣傳人員與配售單位（何佳娜，2017）。其中與歌曲創作有關的音樂工作包含詞曲創作者、製作人、編混音人員、錄音師、母帶處理與演奏樂手等（黃皓傑，2003；施韻茹，2005）。

有鑑於金曲獎流行音樂演唱類關於歌曲創作獎項類別有、最佳專輯製作人、最佳單

曲製作人作家作詞人獎、最佳作曲人獎、最佳編曲人獎共四項，可看出製作、作詞、作曲與編曲等工作對於流行音樂創作之重要性。而各項創作工作之重要性說明如下：

(1) 關於製作人，製作人作為歌曲的總籌，需要按照唱片企劃的設定並掌握藝人的特性，發稿邀歌並加以監製（何佳娜，2017）。且曾裕恒（2007）更指出製作人是音樂創作工作中最重要的角色，是掌握唱片成敗的關鍵人物，。

(2) 關於詞曲創作人，詞曲創作人主要依據製作人訂製主題與意象描述創作歌詞或歌曲，（謝鴻源，2004）表示作詞者與作曲者為音樂創作提供內容的基礎，亦為歌曲產製的關鍵人物，是流行音樂產業的核心，除了為歌曲的創作工作指定懺製團隊，還要與歌手討論演唱的方式與形式，更需參指導各項相關行政工作，如挑選主打歌等。

(3) 關於編曲者，編曲工作通常是以詞曲創作的旋律為基礎配上伴奏，以及編排歌曲的前奏、間奏、尾奏各段，是一項相當專業的音樂創作工作（謝鴻源，2004；陳建銘，2002）。尤景仰（2018）說明製作人與編曲人的關係是相當緊密的，製作人向出資者描述其工作團隊於經常會以編曲人作為募資亮點。綜合上述可知編曲在音樂創作上是一項專業工作且有其不可忽視的重要性。

3. 研究設計

本研究流程先以文化部影視及流行音樂產業局公告之歷屆金曲獎資料作為清單，蒐集金曲獎第 18 屆到第 29 屆金曲獎最佳國語男歌手獎與最佳國語女歌手獎提名歌手與專輯歌曲創作者代表華語流行音樂工作者社群，其中包含 67 位歌手、130 張專輯（1,428 首單曲）。再從國家圖書館國際標準錄音錄影資料代碼查詢系統（ISRC）檢索與下載歌曲製作工作者相關資料，其中歌曲製作工作類型以作詞、作曲、編曲、監製四類型為主，再以人工查找方式補齊系統未收錄專輯資訊或專輯資料不全部分，總共蒐集到 6,733 音樂工作人次資料（958 位音樂工作者），並輔以統計分析軟體與社會網絡分析軟體進行資料分析。本研究以歌曲製作工作者對應歌手的方式繪製社會網絡圖，例如：林夕為林俊傑「只要有你的地方」這首歌作詞，林夕會因為林俊傑這一首歌而產生有向連結，以林夕作為來源（Source）、林俊傑作為目標（Target），兩人的邊權值（Edge Weight）會以兩位合作次數而產生不同的數值。

4. 結果分析

4.1 金曲獎最佳國語男女歌手獲提名之專輯音樂工作者社群樣貌

金曲獎最佳國語男女歌手獎之提名音樂工作者社群整體網絡分析數值，詳如表 1 所示，音樂工作者社群節點總共 966 個節點，代表最佳國語男女歌手獎之提名音樂工作者

總人數，主要成份規模（Main Component Size）為 963 個節點，主要成份所佔百分比（Main Component Percentage）為 99.69%，顯示多數音樂工作者有合作關係且僅有極少數音樂工作者獨立創作專輯。再者從平均程度（Average Degree）可以看出每一位音樂工作者合作人數為 1.645 位，且加權平均度（Avg. Weighted Degree）代表每位音樂工作者合作人次為 6.963 人次，顯示有少數節點在網絡中占有重要位置，如圖 1 所示。

更進一步看網絡直徑（Diameter）為 7，代表整體網絡規模有相當規模，平均路徑長度（Average Path length）為 2.156，代表社群成員彼此連結度高，但是有別於小世界網絡平均路徑短且群聚係數較大的特質，平均群聚係數（Avg. Clustering Coefficient）只有 0.02，群聚係數數值相對較低，顯示音樂工作者彼此之間合作對象多元，顯示音樂工作者並未偏好與特定音樂創作人合作之情形。另外，依據每位音樂工作者合作程度進行分群，模組（# of modules）可分 22 組，且模組分數（Modularity Score）為 0.716，代表整體網絡有明顯分群，且各分群社群網絡凝聚程度較高。

表 1 金曲獎最佳國語男女歌手獎之提名音樂工作者社群樣貌

社群網絡節點數（# of node）	966
主要成份規模（Main Component Size）	963
主要成份規模百分比（Main Component Percentage）	99.69%
平均程度（Average Degree）	1.645
加權平均度（Avg. Weighted Degree）	6.963
網絡直徑（Diameter）	7
平均路徑長度（Average Path length）	2.156
平均群聚係數（Avg. Clustering Coefficient）	0.02
模組分數（Modularity Score）	0.716
模組（# of modules）	22

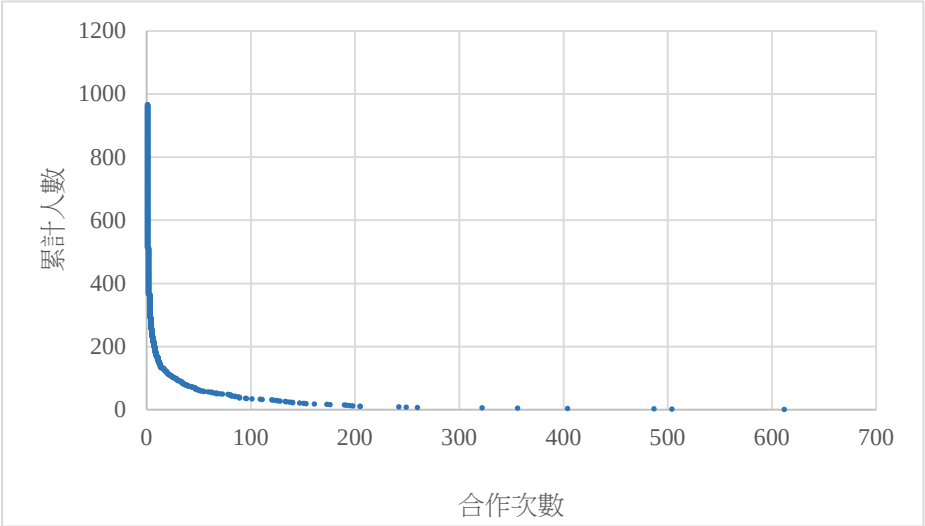
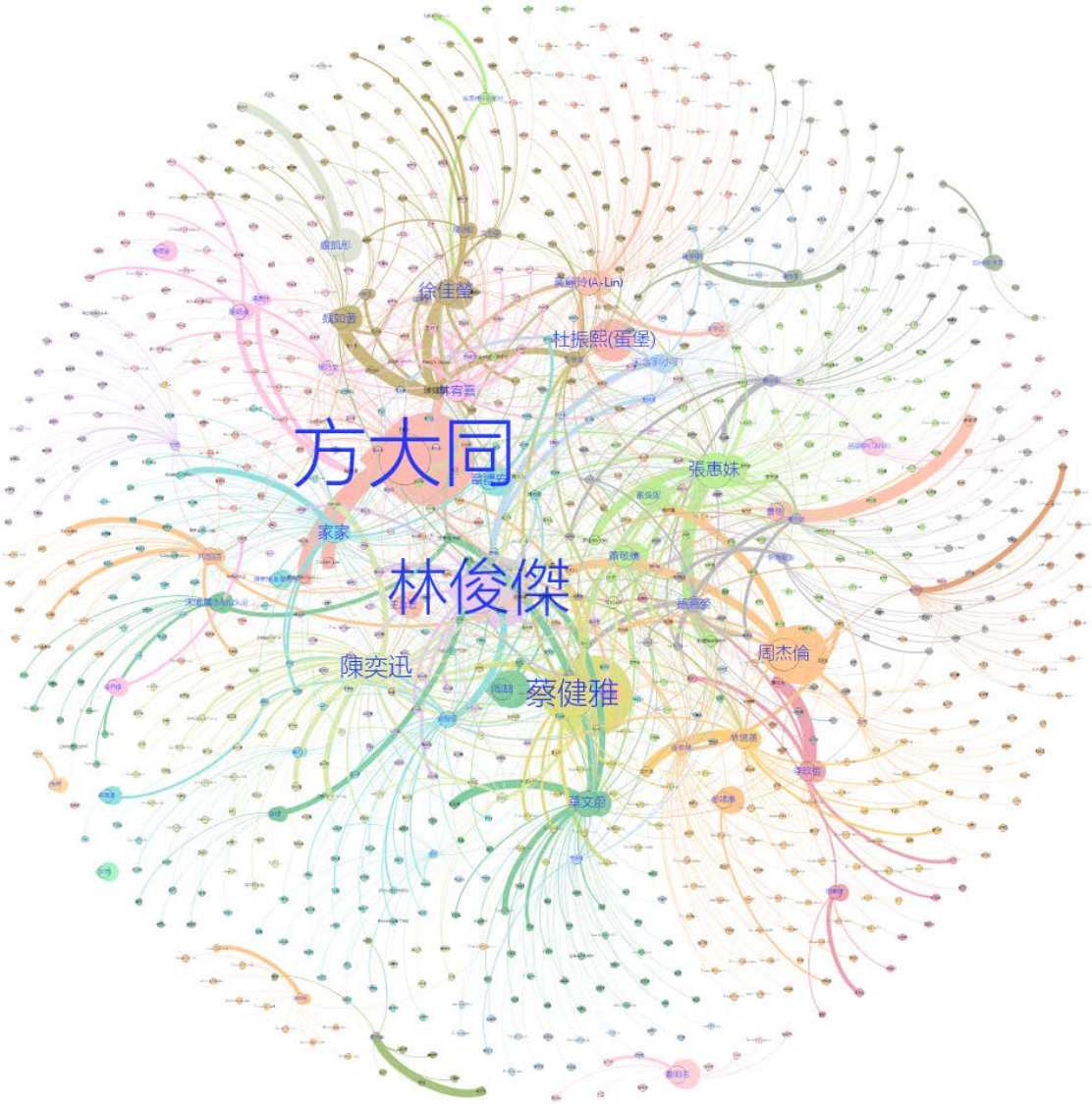


圖 1 金曲獎最佳國語男女歌手獎之提名歌手與音樂工作者合作情形分布圖

華語流行音樂工作者社群樣貌如圖 2 所示，由節點大小可以看出加權程度前五名歌手為方大同、張惠妹、林俊傑、陳奕迅、蔡健雅，主要原因是歌手提名次數高，累積合作人次高。而字體大小可以看出歌手的中介中心性，其中張惠妹雖在加權程度方面數值較高，但未曾替其他歌手跨刀製作，故其接近中心性與中介中心性兩指標皆為零，前五名歌手的社會網絡指標數值如表 2 所示。

表 2 合作人次加總最高前五名歌手之社會網絡指標數值

社會網絡指標數值	方大同	張惠妹	林俊傑	陳奕迅	蔡健雅
提名次數	6	7	6	6	6
加權程度（Weighted Degree）	612	404	487	322	504
接近中心性（Closeness Centrality）	0.5	0	0.5	0.6	0.6
中介中心性（Betweenness Centrality）	997.5	0	975.5	323.5	234



節點顏色：模組分組（modularity）

字體大小：中介中心性（betweenness centrality）

節點大小：加權程度（weighted degree）

藍色字體：最佳男女歌手提名者

圖 2 金曲獎最佳國語男女歌手獎之提名音樂工作者社群樣貌

4.2 金曲獎最佳歌手提名人之身分多元程度與其參與音樂創作工作之類型

本研究利用「赫芬達爾—赫希曼指數（Herfindahl-Hirschman Index，簡稱赫芬達爾指數）」計算歌手身分多元性，越多則身兼多重身分，赫芬達爾指數是一種測量產業集

中度的綜合指數，用某特定市場上所有企業的市場佔有率之平方和，來計量市場企業佔有率的變化，即市場中廠商規模的平均程度，其值介於 0 至 1，數值越接近 0 代表分布越平均，數值越接近 1 代表分布越不平均。N代表產業內的企業數，Xi代表i企業的規模，X代表市場的總規模數，可透過公式（1）計算：

$$\mathrm{HHI} = \sum_{i=1}^{N}(Xi/X)^2 \qquad (1)$$

本研究藉由赫芬達爾指數了解歌手身分多元性程度，分數越高則身分多元性越低代表歌手性格較為突出，分數越低則代表歌手身兼多重身分，也是就說作為創作者的性格較強烈，計算指標有演唱、作曲、作詞、編曲與監製五種音樂工作類型次數，另為避免歌手因提名次數多寡產生影響，因此每位歌手之赫芬達爾指數須再除以歌手金曲獎提名次數進行正規化運算，研究統計結果顯示如表 3 所示。更進一步分析 67 位歌手分別參與四種類型音樂創作工作之情形，統計結果如表 4 所示，統計結果顯示歌手參與音樂創作工作主要是以作詞為主（77.61%），作曲類型（76.12%）與監製類型（61.19%）次之，編曲類型（49.25%）最少。

表 3 金曲獎最佳國語男女歌手獎之提名歌手身分多元程度

平均值	0.326
中位數	0.24
最大值	1
最小值	0.035

註：小數點以下第二位，四捨五入

表 4 金曲獎國語流行歌曲最佳男女歌手之提名歌手創作類型統計

創作類型	**作曲**	**作詞**	**監製**	**編曲**
歌手百分比	76.12%	77.61%	61.19%	49.25%

4.3 金曲獎最佳歌手提名人之身分多元程度與網絡地位之相關性

本研究以歌手的提名次數與赫芬達爾指數最為歌手的自變項，以兩項社會網路中心性程度為因變項，代表歌手在社群網絡中地位，兩應變項分別以中介中心性（Betweenness Centrality）、接近中心性（Closeness Centrality）。中介中心性是測量單一節點位於網絡中其他兩節點對的中介的程度，可以用來觀察音樂工作者連結其他音樂工作者之合作程度，接近中心性是計算網絡內單一節點要到達所有節點的最短路徑總和，數值越大可表示社群網絡節點之間距離小，代表音樂工作者越容易觸及彼此。

表 5 本研究四變項偏度與對數正規化之最大值與最小值

研究變項	偏度	對數正規化最大值	對數正規化最小值
提名次數	1.272	2.85	2
赫芬達爾指數	3.151	2	0.55
接近中心性（Closeness Centrality）	-0.993	2	0
中介中心性（Betweenness Centrality）	16.128	5	0

上述四項研究項數值未呈現常態分佈，分布圖有左傾（數值小於零）或右傾（數值大於零）之情形，為避免變項數值影響分析結果，將四變項之數值進行正規化處理，將每一變項乘以一百並取對數計算，研究四變項之偏度與正規化後之最大值及最小值如表 5 所示，再利用逐步回歸分析法排除解釋力較小的變項，統計結果顯示提名次數對於中介中心性與接近中心性兩變項解釋力顯著性並未顯著，故於模式中排除。

而赫芬達爾指數對於兩中心性程度逐步回歸分析結果如表 6 所示，可以看出赫芬達爾指數與兩中心程度指標呈現顯著負相關，表示歌手身分多元程度越高，越能居於整體社群網絡的中心性位置，換言之，身分多元的創作歌手在進行音樂創作時，可以更容易觸及其他音樂工作者，並且較會尋求其他音樂工作者共同創作音樂作品。

表 6 歌手身分多元程度與中心性程度之相關分析

逐步回歸分析法	赫芬達爾指數	顯著性	說明
接近中心性（Closeness Centrality）	-0.480***	.000	n=67, ***$p < .001$
中介中心性（Betweenness Centrality）	-0.429***	.000	n=67, ***$p < .001$

5. 結論

綜和資料分析結果說明近十年華語流行音樂歌手普遍參與音樂創作工作，合作過程中未有與特定音樂工作者合作之情形，主要音樂創作類型以作詞與作曲為主，監製次之，說明金曲獎國語最佳歌手提名人多數為創作型歌手，且身分越多元的創作歌手，越容易觸及其他歌曲創作人產生合作關係，並且 較容易尋求其他音樂工作者共同創作流行音樂作品 。本研究以台灣金曲獎國語最佳歌手提名人作為分析範圍，未來可以進一步分析歌手跨刀創作之情形或以音樂工作者為研究對象進行分析，更能顯示台灣華語流行音樂工作者社群之全貌與歌手多元程度分布之情形。

6. 參考文獻

文化部影視與流行音樂產業局。(2018)。金曲獎沿革。取自 https://www.bamid.gov.tw/content_175.html

尤景仰。(2018)。製作人與編曲人的關係。取自 http://www.yusmusic.com/page1.aspx?no=10011011015104790 5&step=1&pno=100110307140004664

何佳娜（2017)。臺灣華語流行音樂編曲者的默會知識與工作歷程。國立臺灣大學圖書資訊學研究所碩士論文，台北市。 取自 https://hdl.handle.net/11296/yy24bn

施韻茹（2005)。由節拍旋律到娛樂商品：台灣流行音樂產業產銷結構轉變研究。國立交通大學傳播研究所碩士論文，新竹市。 取自 https://hdl.handle.net/11296/avcfw9

曾裕恒（2008)。台灣唱片產業之研究：主流與非主流之比較分析。國立政治大學企業管理研究所碩士論文，台北市。 取自 https://hdl.handle.net/11296/8w4g32

謝鴻源（2004)。本地唱片業者如何因應當前主要問題之研究。國立交通大學傳播所碩士論文，新竹市。 取自 https://hdl.handle.net/11296/u9b4q8

黃皓傑（2003)。兩個樂團的產銷分析--以交工及閃靈為例。國立中正大學電訊傳播研究所碩士論文，嘉義縣。 取自 https://hdl.handle.net/11296/5fkrx4

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

數位人文科學語彙之生成與使用

陳光華*、吳孟家**

國立臺灣大學圖書資訊學系教授*

國立臺灣大學圖書資訊學系教授研究生**

數位人文科學語彙之生成與使用

陳光華、吳孟家

教授、研究生

國立臺灣大學圖書資訊學系

{khchen, r04126011}@ntu.edu.tw

摘要

觀察或是探究一個研究領域的發展，一個可能的方法是透過書目計量的方法，分析該領域眾多研究產出的書目資料；另一個可能的方法則是透過內容分析的方法，分析該領域眾多研究產出的學術論文。本研究以自動的方法進行數位人文學術論文的內容分析，觀察跨年代間不同地區數位人文的研究，特別是以科學語彙概念的生成的角度，探討數位人文學術研究的發展

本研究採用語料庫分析為基礎之研究取徑，藉由不同時間學術詞彙的使用變遷，展現各地區在數位人文學科術領域上研究主題的特徵與差異。研究者挑選數位人文研究相關的七本國際期刊與會議論文集做為研究對象，分析 2009 年至 2013 年所有學術論文的學術論文摘要，透過比較分析，分別由時間歷線與空間範疇，探討臺灣、日本、美國、加拿大、英國數位人文研究發展的共性與異性，檢視數位人文主題概念或學術語彙的生成與變遷。研究者在蒐集完整的資料之後，扣除非正式學術文獻，以 588 篇文獻摘要作為分析對象。

研究結果顯示五地區的核心詞彙為“digital”、“information”、“humanities”、“technology”，呈現出數位人文學科領域核心主題內容仍然是資訊科學與人文學；臺灣以在地化研究為主，其他四個地區反映各自的研究特色，日本以技術、架構層面為主；英國經常進行語言學、資訊軟體相關研究；美國重視當代藝術與當代文學的加值應用；加拿大則是以中世紀文物、典藏手稿等第一手資料為主，亦包含許多跨語言、跨地域的研究。整體的研究主題傾向以資訊科技相關的主題詞彙較多，若以各地區主題詞彙分布來看：英國、美國、加拿大整體傾向科技類別為主，包含資料科學、網際網路、巨量資料等領域；臺灣與日本的主題詞彙在科技類別與人文、社會科學類別的分佈較為平均，且重視圖書館學、歷史、地理與宗教學科領域的加值應用。

關鍵詞

科學語彙，社會網絡分析，語料庫，庫博

1. 前言

當前全球化、數位化、資訊科技快速的發展，廣泛應用、結合不同學科並予以實踐的數位人文（Digital Humanities）的各項研究議題，近年來已形成重要的學術領域。由於數位人文一詞之涵蓋範圍仍持續快速擴展，諸多學者認為目前仍不易於對整體數位人文研究範疇做一清楚定義（Liu, 2012; Warwick, Terras, & Nyhan, 2012）。

Schreibman、Siemens 與 Unsworth（2004）於《A Companion to Digital Humanities》首度提出數位人文一詞，試圖更廣泛的定義傳統的人文計算（humanities computing）領域；項潔與涂豐恩（民 100）則是界定數位人文為運用數位材料結合資訊科技，以此從事人文研究，其特點在於借助數位科技發掘出以往隱藏於巨量文獻而難以直接觀察到的現象。以此可知，數位人文之重要性在於應用當代之資訊技術與方法，輔佐傳統人文學科之研究，藉由有別於傳統方法之不同視角，從而拓展新的研究議題並增進研究價值。

當前眾多學術機構已相繼加入數位人文研究發展行列，國際性聯盟組織諸如：「國際數位人文組織聯盟」（The Alliance of Digital Humanities Organizations；ADHO），成立於 2005 年，其宗旨在於促進世界各地的數位人文研究組織之學術交流與協同合作，其以社群為基礎，為全球最大的數位人文傘狀架構中介組織；2007 年成立的「國際數位人文中心網絡」（centerNET）由全球各國數位人文中心共同組成之世界網絡，藉由該網絡的建立，各國數位人文學者有機會相互分享交流與共同建置合作計畫。

國外機構如美國的「計算與人文協會」（ACH），是最早由數位人文學者們設置的專業協會，催化數位科技與人文研究的互動與交流。加拿大的「數位人文學會（CSDH/SCHN）」創立於 1986 年，其附屬於加拿大人文與社會科學聯盟，由當地的學院與大學共同組成研究機構，以英文及法文語系推動數位人文工作。歐洲人文學跨領域的代表為「數位人文歐洲協會（EADH）」，協助人文學各領域的數位人文建置需求，提供各項資料運算與技術的支援。日本數位人文學會（日本デジタル・ヒューマニティーズ学会）則是於 2012 成立，旨在促進日本數位人文學者與國際的學術交流合作。表 1 簡要羅列重要的國際性數位人文相關學術或研究機構。

臺灣的數位人文領域亦有許多的重要組織，2008 年「國立臺灣大學數位人文研究中心」為臺灣首個數位人文研究機構，致力於重要文化資產的典藏數位化與資料庫建置。爾後，科技部推動以數位人文為主題的研究計畫，在國家與計畫的推展支持下，各研究機構陸續建立數位人文相關組織，例如中央研究院「數位文化中心」、政治大學「數位人文中心」、中興大學「中臺灣數位文化中心」等（項潔、陳麗華，民 103）。此外，社會團體如「臺灣數位人文學會」（TADH）以培育數位人文專業人才、舉辦各類學術交流活動為其主要項目，並時常舉辦國際研討會促進數位人文國際交流合作，擴展臺灣的全球視野。國外內相關研究組織之相繼成立，足證數位人文研究活動之蓬勃發展。

表 1：國際重要數位人文研究組織

	成立年	類型	國家	出版刊物
計算與人文協會（ACH）	1978	協會	美國	協會成員各別出版
加拿大數位人文學會（CSDH/SCHN）	1986	學會	加拿大	Digital Studies/Le champ numérique
國際數位人文組織聯盟（ADHO）	2005	國際中介組織	全球性	聯盟組織成員各別出版
國際數位人文中心網絡（centerNET）	2007	各國數位中心聯合組成	全球性	DHCommons
數位人文歐洲協會（EADH）	2011	協會	歐洲會員國	協會成員各別出版
日本デジタル・ヒューマニティーズ学会	2012	學會	日本	Journal of Japanese Association for Digital Humanities

資料來源：研究者自行整理

數位人文研究經由結合人文社會科學與電腦資訊科學的研究取徑，跨足於多個學科領域之間，並開創多元的研究可能性，眾多新興研究議題隨之孕育而生。由於其領域的特殊背景，數位人文涉及的主題概念亦變得愈加豐富且複雜。因此，本研究試圖以語料分析方法，透過數位人文相關的學術論文，對數位人文研究主題詞彙的衍生進行較為全面的分析，以了解數位人文研究領域的學術發展。

我們藉由分析與比較數位人文領域國際期刊與國內會議論文集之摘要全文，並嘗試回答以下三個問題：（1）2009 年至 2013 年，數位人文研究的主題趨勢，以及研究主題增長或式微的情形；（2）觀察國際間共同關注的研究主題，主題之間是否存有關聯性；（3）進一步比較不同地區於數位人文研究重點的異同，以及影響研究主題改變的原因。

2. 數位人文的發展

數位人文自 2004 年領域學術名稱的確立以來，已然累積了近十餘年來的發展，過程中不斷影響眾多學科領域固有的學術生態（項潔、陳麗華，民 103），它的應用範疇廣泛，不僅是致力於人文學研究領域的技術輔助，更是推動人文學與社會科學於應用資訊科技層面的深化。

回顧數位人文在國際的發展，林顯明（民 104）梳理數位人文過去的研究成果與發展，並將其大致分為三個階段：第一階段為「數位典藏階段」，係以 1990 年代至 2000 年針對大量的歷史史料與檔案、手稿等進行數位化工作，著重在建立規模龐大的數位資料庫，傾向以量化資料進行研究分析；第二階段的關鍵時期為 2009 年，自提出「數位人文 2.0 倡議」（The Digital Humanities Manifesto 2.0）後，數位人文研究朝向發展新的研究方法、數位工具，著重在創造一個生產及開發知識互動的環境。直至網路科技與數位時代的快速發展，第三階段的數位人文概念產生，海量的資料、資訊與知識改變人文及社會科學研究和分析的途徑，亦創造出跨領域研究的新的內涵及發展方向。

當前數位人文研究大多是以研發新的數位技術或實作數位工具，輔助人文研究學者以打破時間與空間的框架，對文本進行較為宏觀的資料分析、加值應用與深入研究。像是國家圖書館將特藏文獻「明仁文集」為首批文本，建立國圖「通用型古籍數位人文研究平臺」，以建構古籍全文資料環境，提供學者後續應用文獻分析工具創造更多古籍的應用價值（呂姿玲、莊惠茹，民 107）。

中央研究院臺史所檔案館的「臺灣日記知識庫」計畫，即是以數位工具結合專業社群協作創作機制，將原本第一手資料的私人日記典藏，轉變成開放公眾查詢與取用的數位服務（王麗蕉，民 104）。還有以協助自然語言處理、建置文本語料分析工具的平台，諸如 Docusky、Corpro、MARKUS 等平台，能給予沒有資訊專業背景、沒有資訊人員協助支援的人文學者能夠自由運用數位資源至個人的研究。

不過探討數位人文本質與特性的研究相對較少，目前僅有少數研究採用書目計量方法探究數位人文的現況與特性，例如鄭允人（民 104）藉由引文資料分析探究數位人文領域跨學科知識整合的變化，以及學科合作的現況；Chen & Hsueh（2015）則是觀察大量數位人文學術論文，以基於圖書館編目工作的記敘分析及主題分析的實務方法，解釋數位人文研究領域的研究樣貌。Wang & Inaba（2009）則以兩本數位人文期刊及四本年度會議論文集做文本，嘗試透過對應分析（correspondence analysis）和共字分析（co-word analysis）分析數位人文學科領域的結構和演變，探索數位人文仍不斷在拓展其研究領域的範圍。

但目前沒有從時間、空間的面向，並以語料庫分析為研究方法，將數位人文學術論文設定為文本標的，從而探討數位人文研究的學術研究發展情形與科學語彙生成狀況。闕河嘉（民 107）認為若以語料庫（corpus）分析為研究方法探討學科領域，極可能從中挖掘研究對象或研究主題潛在的意義與新興研究議題。本研究嘗試以摘要作為標的語料，以語料庫分析方法觀察 2009 年至 2013 年不同地區數位人文學術的發展，並以社會網絡分析勾勒出數位人文研究主題的變化。

3. 研究方法

本研究係以語料分析方法及社會網絡分析方法，分析數位人文研究領域自 2009 年至 2013 年間，在四個不同地區之數位人文研究在研究主題上的差異，並探索此領域研究者於詞彙使用上的變化趨勢。本節將簡要說明研究之施行步驟、目標對象、使用工具、資料蒐集與分析方法。

3.1 目標對象

本研究以數位人文研究領域之學術出版物為目標對象，研究者檢視 ADHO 本身及其成員機構出版的學術出版品，以經過同儕審查程序且為開放取用之出版品為依據。在此標準下，本研究挑選出八種數位人文學術出版品，其中包含七本國際期刊與一本國際會

議論文集。依各出版品出版國別，列舉說明如下。

- 美國

  Journal of Digital Humanities（JDH）刊載文章的特色以最佳的學術、工具與數位人文社群對話為主。Digital Humanities Quarterly（DHQ）為 ADHO 出版之期刊，符合國際開放資料標準，便於社會大眾於線上閱覽。

- 加拿大

  由加拿大數位人文學會出版的 Digital Studies / Le champ numérique（DS/CN）著重將技術應用於歷史、文化與社會議題，鼓勵學者將研究實踐於跨領域、跨語言及歷史方面的研究。Digital Medievalist（DM）為線上計畫的參考刊物，文章性質主要接受原創研究與計畫型研究，討論該領域的評論文章、書目資料以及項目報告。

- 英國

  英國牛津大學發行的 Literary and Linguistic Computing（LLC）涵蓋電腦運算與資訊相關的國際期刊，為該領域歷史最為悠久的代表性期刊。LLC 收錄內容是以文學、語言研究及教學等應用層面為主，收錄內容極為廣泛。目前該期刊已更名為「Digital Scholarship in the Humanities」，不再侷限於人文科學，應用更加廣泛。International Journal of Humanities and Arts Computing（IJHAC）為英國首屈一指的多學科、同儕評鑑的學術論壇，內容關注藝術與人文運算相關議題。

- 日本

  日本數位人文學會於 2011 首次舉辦數位人文學術會議，2013 年正式加入 ADHO，其每年固定出版的學術刊物 Journal of Japanese Association for Digital Humanities（JJADH），是日本數位人文研究的重要刊物，故選用 JJADH 為日本數位人文研究研究的代表期刊。

- 臺灣

  臺灣數位人文研究的相關學術出版品則是挑選歷年於臺灣舉辦之重要數位人文國際研討會「International Conference of Digital Archives and Digital Humanities」（DADH）出版的會議論文集。

3.2 資料蒐集與整備

確認研究目標對象之後，本研究先取得各出版品於 2009 至 2013 年所有刊登文章之摘要內容。為確保蒐集的文本能清楚地識別其出版地區、期刊名稱與出版年代，研究者將每篇摘要以所屬地區、刊名和年代命名，並依期刊編排的順序依次編號（例如：TW_DADH_2009_1、UK_LLC_2010_2）。每篇文章以 UTF-8 編碼格式儲存成純文字檔（.txt），並經人工校正確保網路下載內容之正確，包含修正亂碼（例如：含有程式碼符號之文章）、確認特殊符號（例如：梵文、歐語系特殊文字）、修訂數字跳脫等問題。除

此之外，由於本研究之目標在於分析數位人文研究之研究主題，因此亦排除不符此條件之文本，排除類別包含僅有書目資料、開源程式、文獻回顧描述之摘要等。

最終蒐集文獻的簡要資料如表 2 所示，共計有 588 篇摘要。其中以英國 LLC 篇數最多，達 175 篇；以美國的 JDH 篇數最少，僅 11 篇。此外由於 JDH、JJDAH 於 2011 年以前尚未出刊，因此僅有 2011 年至 2013 年文章；臺灣 DADH 則是 2013 年未舉辦，故僅收錄 2009 年至 2012 年文章。

3.3 研究工具與資料分析方法

本研究使用的程式工具主要以國立臺灣大學闕河嘉教授與陳光華教授及國立海洋大學林川傑教授開發維護之「庫博（CorPro）中文語料庫分析工具」進行前期資料分析（闕河嘉、陳光華，民 104），後續則以社會網絡分析工具「Gephi」及書目計量網絡視覺化工具「VOSviewer」進行網絡分析。

表 2：2009 年至 2013 年目標期刊與會議論文集統計資料

期刊名稱	出版年	收錄時間	篇數	地區	備註
Journal of Digital Humanities (JDH)	2011	2011-2013	11	USA	2009-2010 未出刊，無資料
Digital Humanities Quarterly (DHQ)	2007	2009-2013	120	USA	
Digital Studies / Le champ numérique (DS/CN)	1992	2009-2013	47	Canada	
Digital Medievalist (DM)	2005	2009-2013	20	Canada	
Literary and Linguistic Computing (LLC)	1986	2009-2013	175	UK	現改名為 Digital Scholarship in the Humanities
International Journal of Humanities and Arts Computing (IJHAC)	1994	2009-2013	61	UK	
Journal of Japanese Association for Digital Humanities (JJADH)	2011	2011-2013	75	Japan	2009-2010 未出刊，無資料
International Conference of Digital Archives and Digital Humanities (DADH)	2009	2009-2012	79	Taiwan	2013 年並未舉辦 DADH

資料來源：研究者自行整理

為了擷取出代表該文本的關鍵主題（key topics）與術語，以利後續比較不同地區在數位人文研究主題的異同，本研究採用 JSTOR Labs 開發的文字分析工具 Text Analyzer 0.14.1 版本，擷取每篇文章所陳述的研究主題、相關術語。由於單篇文章可能同時會被賦予多個主題詞彙與相關術語，為確保 JSTOR 最終識別的主題詞彙能涵蓋文章陳述之

內容，本研究優先挑選 JSTOR 工具判斷權重在 5 以上的主題詞彙，以高權重之主題詞彙代表該篇文章的研究主題。

在資料分析方面，研究主要以庫博（CorPro）語料庫分析文本的詞頻與不同地區在辭彙使用上的異同，並以 VOSviewer 和 Gephi 進行資料之視覺展現。接著，本研究採用餘弦相似度計算文本相似程度，再透過 Gephi 與 VOSviewer 進行模組化分析。進一步說明資料分析方式如下：

3.3.1 餘弦相似度（cosine similarity）

餘弦相似度是以計算向量空間中兩向量夾角的餘弦值，以此衡量兩個物件的相似程度。一般應用於計算文件間或者詞彙間的相似性，為資訊檢索常用的相似度計算方式之一。在計算文件相似度前，須先將文件表達成向量形式，亦即將文件中所有重要詞彙視為個別向量維度，而以詞彙的權重（詞頻）代表該維度的值，並組成一個向量代表該篇文件。假設文件 X 與文件 Y 為 n 維向量，文件 X 可以表示為$D_x = [X_1, X_2, \dots\dots, X_n]$，文件 y 為$D_y = [Y_1, Y_2, \dots\dots, Y_n]$，則兩個文件的餘弦相似度計算公式如下：

$$\cos(Dx, Dy) = \frac{\sum_{i=1}^{n}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2} \times \sqrt{\sum_{i=1}^{n}(Y_i)^2}}$$

餘弦相似度的數值範圍介於 0 到 1 之間，若代表兩個文件之餘弦相似度越趨近於 0，相似程度越小；反之，若其值越趨近於 1，相似程度越大。

本研究將八種目標對象出版品以出版地區作為依據，區分成美國、加拿大、英國、日本與臺灣等五個地區，並結合五個年代，探察不同地區在 2009 年至 2013 年間，數位人文研究領域之文本相似性。為比較文本內容的相似程度，研究者將 588 篇摘要以相同年代且相同出版地區為依據做分類（例如：英國 2013 年文件、美國 2012 年文件）。由於資料本身之限制，英國、美國與加拿大各自分為 5 個時間區間；臺灣 2013 年並未出版數位人文會議論文集，故分為 4 個時間區間；日本因僅有 2011 至 2013 年的資料，故分為 3 個時間區間。最終研究者以 22 個時間區間進行餘弦相似度計算。

3.3.2 模組化分析（Modularity）

模組化分析是用於衡量社會網絡結構的方法，將廣大的巨型網絡有效分割成多個不同的子社群（communities）、群組（group）或者族群（clusters）。模組化分析的數值範圍介於-1與+1之間，根據一個群組內部會比群組外部具有高密度連結的原則，切割不同的群體（Blondel , Guillaume, Lambiotte, & Lefebvre, 2008）。Modularity 的計算方式是將網絡先行分群過後，群體內部邊數（Edge Number）的比例減去群體內部隨機分配的期望邊數的比例。反覆進行前述動作，直到獲得最高 Modularity 分數作為最後結果。

本研究使用 Gephi 社會網絡分析工具提供之功能，可自動辨識出內部連結緊密的群

體，再由此觀察不同群體間的互動關係。本研究及適應用這項功能觀察不同地區在數位人文研究主題上的分群。

4. 研究結果

4.1 高頻詞分析

本研究將分析結果先依據地區歸類，整合成五大類別，分別為臺灣、美國、英國、加拿大及日本。接著再以詞頻分析與顯著詞分析兩種方式，分析不同地區於數位人文研究領域的常用詞彙與特殊詞彙。以下簡述高頻詞分析結果：

首先，本研究移除停用詞、虛詞、感嘆詞、特殊符號、程式符號與網址等詞項（token）。再以兩個標準選出高頻詞，首先詞彙必需頻繁出現，即以詞頻（term frequency）做為標準；再考慮詞彙的廣泛使用度，即文件頻率（document frequency）做為標準。本研究判斷詞彙是否廣泛使用的標準，為該詞彙需至少於該地區 10%的文件均有出現，例如文件總數為 79 篇時，則該詞彙需出現於 8 篇不同文件中。挑選高頻詞時，先以詞頻由高至低依序檢視是否符合廣泛使用的標準，若符合則選為高頻詞，反之則否。最終結果各地區前 20 名高詞頻，總計出現不重複詞彙 54 個，重複詞彙有 46 個。分析結果如表 3 所示。

表 3：五個地區前 20 名高頻詞與詞頻

地區	五個地區前 20 名高頻詞
TW	Chinese(121), historical(115), information(82), database(71), digital(49), system(48), Qing(36), Japanese(35), documents(33), land(32), GIS(30), materials(29), archives(26), digital libraries(26), government(24), technology(23), humanities(21), traditional(21), map(20), THDL(15)
Japan	digital(117), information(113), database(83), system(75), documents(72), metadata(72), Japanese(71), scholarly(58), historical(47), Chinese(47), humanities(41), structure(40), tools(38), model(36), interface(34), TEI(32), corpus(31), cultural(29), resources(29), technology(28)
USA	work(78), humanities(75), digital(104), literary(64), scholarly(52), technology(46), media(38), tools(38), model(28), critical(27), textual(25), information(24), historical(23), experience(20), potential(20), system(19), contemporary(17), network(17), language(17), academic(16)
Canada	digital(57), tools(31), historical(25), work(22), pour(22), information(21), humanities(21), sources(18), online(18), documents(17), manuscripts(16), textual(16), Web(16), medieval(15), database(14), technology(14), public(14), approach(11), scholarly(11), user(11)
UK	digital(146), humanities(73), corpus(73), work(66), information(60), language(58), system(47), model(47), tools(44), cultural(43), documents(41), linguistic(39), resources(39), technology(34), software(34), traditional(32), scholarly(31), user(29), particular(26), nature(21)

資料來源：研究者自行整理

本研究進一步分析出現於多地區之高頻詞，並觀察各地區之特有的高頻詞。地區與高頻詞間之關係，以 VOSviewer 繪製成圖 1。如圖 1 所示，節點（node）分為兩類：地區與詞彙。圖一共有五個地區節點，分別以五種顏色對應：臺灣（TW）為紅色、日本

（JP）為黃色、加拿大（Canada）為綠色、美國（US）為藍色、以及英國（UK）為紫色。詞彙節點則為前項統計中出現之高頻詞，其顏色意義與地區點一致。若出現其他顏色，則代表該詞彙同時出現在多個不同地區。連結兩個點的線條（edge）則代表某地區在該詞彙的詞頻數量，當詞頻數越高，線條越粗。

圖 1 以環狀圖形呈現詞彙整體分佈，中央為核心聚集詞彙，越往外圈為各地區獨自擁有的詞彙。依高詞頻詞彙連結程度，詞彙可區分成高度、中度、低度三種連結等級。高度連結詞彙為涵蓋五個地區之詞彙；中度連結詞彙為蓋三至四個地區之詞彙；低度連結詞彙則為連結兩地區之詞彙。若僅有個別地區特有之高詞頻詞彙，則歸類為無連結詞彙。最後我們再依照其詞彙分布的狀態，將高度連結詞彙視為中央核心，而中、低度連結詞彙則為半邊陲，無連結詞彙歸為邊陲。根據前述整理如表 4 所示。

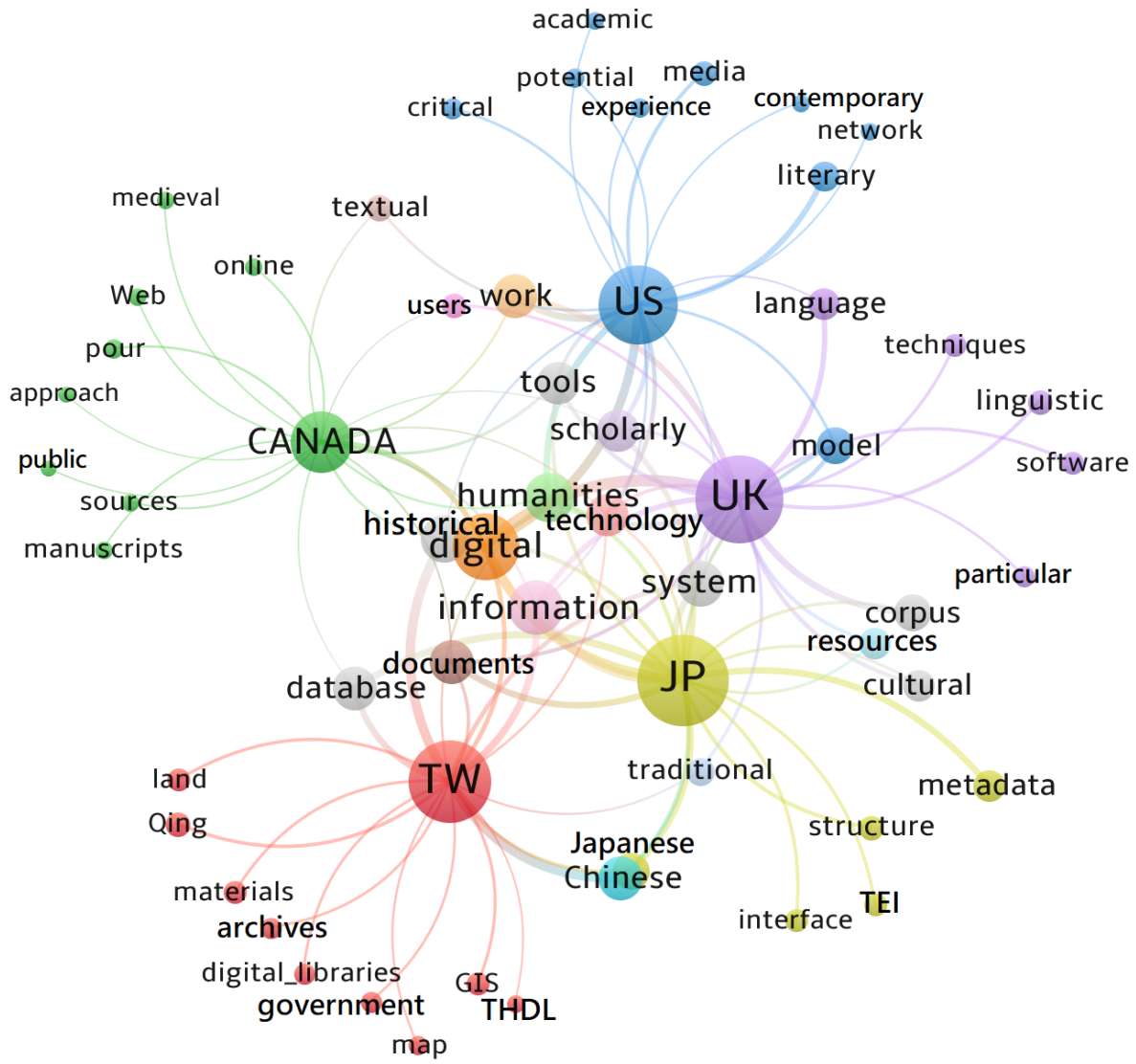


圖 1：地區與高頻詞之關係網絡

表 4：高頻詞彙連結程度與分布

連結程度	涵蓋地區	核心程度	詞彙
高度連結	5	中央核心	information, digital, technology, humanities
中度連結	3~4	半邊陲	historical, system, documents, tools, scholarly, database, model, work
低度連結	2	半邊陲	Chinese, Japanese, traditional, corpus, cultural, resources, language, textual, user
無連結	1	邊陲	THDL, Qing, land, GIS, materials, archives, digital libraries, government, map, metadata, structure, interface, TEI, sources, manuscripts, medieval, approach, public, pour, Web, online, literary, media, critical, academic, potential, contemporary, network, experience, techniques, software, linguistic, particular

資料來源：研究者自行整理

4.1.1 中央核心詞彙

位處中央區塊的詞彙，為五大地區皆有連結之高詞頻，顯示出數位人文整體核心概念主要是由“digital”、“information”、“humanities”、“technology”等詞彙所構成。此結果呼應了數位人文最初定義，Schreibman 等人（2004）認為數位人文即是「以數位資訊科技」進行「傳統人文」領域之研究；或是以人文研究的觀點和方法進行數位科技或現象的研究。

4.1.2 半邊陲詞彙

半邊陲詞彙主要為連結超過兩個地區以上之高詞頻，區分為中度與低度連結。首先，中度連結詞彙的“historical”、“documents”彼此經常共同出現使用，代表該領域常著眼於人類社會過去的事件、活動的檔案紀錄或史學研究，善用數位化資源觀察新的文本脈絡並衍生出新的問題意識。

“database”、“model”與“system”在文本代表兩種意義：一種為使用新研究工具或方法；一種為使用的研究材料或分析對象。在前期，數位人文研究取向仍以計算、量化為主，偏重資料庫的建置、搜尋、檢索能力、自動化語料庫語言學等（Schnapp & Presner, 2009）。後期部分學者則經常以既有資料庫做為文本對象，進行深入的解釋、分析。

低度連結詞彙可以呈現出兩個地區在數位人文領域詞彙使用上的共性。根據分析結果顯示，低度連結詞彙有明顯的地域性劃分。首先“Chinese”、“Japanese”為臺灣與日本高強度連結之詞彙，進一步使用搭配詞分析可以發現大多連接“culture”、“database”、“Buddhism”、“historical document”等詞彙。詞彙間連結的強度與密度，顯示兩地區在數位人文應用於文化、宗教與歷史檔案等主題上有密切學術合作與交流。

“language”、“textual”與“user”為英國、加拿大以及美國三個地區之間兩兩相互連接

詞彙。前兩個詞彙突顯這幾個地區在數位人文研究裡，經常從事文本分析與語言變異研究。Burdick、Drucker、Lunenfeld、Presner 與 Schnapp（2012）曾說明藉由網際網絡的發展，影響數位人文研究發展重視使用者介面、使用者經驗研究。因此，“user”於文本中顯示其代表意義為「使用者經驗研究」、「使用者友善介面」。

日本與英國的相互連結詞彙，則有“corpus”、“cultural”和“resources”三者。從兩地區 2009 年至 2013 年的文本可以看出，在文化資源應用及開創新的語料庫的研究上著墨很多。文化一詞本身代表了兩種意義，一種為文化的本質，研究文化的現象、文化的多樣性；另一種為與文化相關的人、事、物的研究，如文化遺產研究、文化活動相關研究。日本與英國藉由拓展數位資源、開發新的數位工具，期望能夠開創該地區文化的新價值。

4.1.3 邊陲詞彙

分布在整體網絡最外圍的詞彙為單一地區常用詞彙，這些詞彙突顯各地區的詞彙使用特色，亦呈現各地區數位人文研究的特性，以下以地區分別說明。

4.1.3.1 臺灣

首先，臺灣數位人文的常用詞彙相較其它地區獨樹一格。由常用詞彙來看，臺灣數位人文研究常用詞彙多以“land”、“Qing”（清朝）、“map”、“materials”、“archives”。本研究認為這些詞彙反映出臺灣數位人文研究者著重於創造歷史保存與想像的價值，呈現過往史料、特藏檔案與古地圖等歷史資料的不同面貌。臺灣數位人文較偏向地區性、專指性的研究，而不似其他地區之研究者多數採用的較多大範圍、大主題的全面性研究。

此外，由項潔教授團隊開發的「臺灣歷史數位圖書館」（Taiwan History Digital Library, THDL）以及其他地理資訊系統（GIS）亦被多次提及，顯現臺灣應用資訊科技工具投入數位人文研究的數量愈來愈多。除可得知此這類工具的重要性外，亦印證經由資訊工具挖掘更多人文研究未來的可能性。有趣的是，“digital library”與“government”等詞彙僅在臺灣被大量提及與研究。“digital library”（數位圖書館）於文本中代表的是數位人文工具（THDL）與研究亞洲文化與歷史的研究機構「東洋文庫」（TōyōBunko）；“government”則代表研究對象的所屬機構或是出版資料的發行單位。蘇信寧（民 104）研究國內數位人文研究計畫分布領域，發現計畫數量前三名的學科分別為：資訊工程、歷史語言、圖書資訊。研究計畫的支持與開展有助於數位人文研究活動的進行，明顯可看出數位圖書館、資訊科技和歷史研究等主題在臺灣的數位人文研究佔有重要的角色。

4.1.3.2 日本

從詞彙類型來看，日本常用的單一詞彙多傾向於技術、架構層面，包含：“metadata”、“structure”、 “interface”與“TEI”（Text Encoding Initiative）。日本數位人文研究著重於資料庫建置與檢索、建立自動化詮釋資料檢索系統，並提出新的資訊架構。這些高頻詞橫跨日本 2011 年至 2013 年所進行的數位人文研究，持續被使用的詞彙代表了在這個領域

中的核心趨勢。除此之外，本研究也發現 JJADH 期刊收錄文章的學科範圍，有傾向計算科學與資訊技術領域的情形，主要刊載類別為資訊科學、跨學科科學與人文社會科學，凸顯出此期刊刊登之文章特色。

4.1.3.3 英國

英國高頻詞為“techniques”、“software”、“linguistic”，為五大地區中單一地區特殊詞彙最少的國家。這三個詞彙在數位人文研究領域為通用、且大範圍的研究主題。本研究認為英國數位人文研究較偏向主題廣泛、多元與全面，所以高詞頻詞彙大多已囊括於前述核心詞彙、半邊陲詞彙裡面。至於「語言學」詞彙被大力突顯出來，應為受到英國數位人文期刊收錄特色所致。

本研究於英國地區挑選的期刊為 LLC 與 IJHAC，LLC 為數位人文研究領域中發展極為成熟之期刊，期刊特色係以文學、語言研究及教學等應用層面為主。現已改名為“Digital Scholarship in the Humanities”（DSH），期望期刊未來刊載文章的研究主題能更為全面，不要僅侷限在文學與語言學領域的應用上。

4.1.3.4 美國

美國數位人文領域高頻詞包含：“literary”、“media”、“critical”、“academic”、“potential”、“contemporary”、“network”、“experience”，主要偏向以數位人文工具結合當代藝術、多媒體與出版品的應用。其中「當代」（contemporary）於文本中多以文學、藝術品等詞彙做搭配，學者經常使用幾世紀以前的著作作為文本語料進行研究。美國期刊收錄特色亦反映在詞彙上面，像是美國出版的 JDH 主要以數位工具應用在保存、分析與文化遺產出版品之研究，故許多藝術領域的相關專有名詞會被大力突顯出來，諸如：“artifacts”、“comic”、“Ekphrasis poetry”（讀畫詩）等。

4.1.3.5 加拿大

加拿大的高頻詞彙包含：“sources”、“manuscripts”、“medieval”、“approach”、“public”、“Web”、“online”，可以發現除了與網際網絡相關詞彙經常使用外，與中世紀相關的詞彙，還有研究資源相關的詞彙被突顯出來。

研究者認為加拿大的高頻詞彙與其 Digital Medievalist 期刊之刊載文章、期刊特色有所關連，DM 的文獻性質傾向中世紀學者感興趣之相關研究主題。同樣以顯著詞分析了解該加拿大 DH 文獻特性可以發現，多以手稿（manuscripts）、卷宗，或者古憲章（Cartulaire）等一手資料和珍稀資料為文本對象。以此可見研究取向明顯與英國、日本偏向技術創新、應用發展有較大區別。此外，加拿大的研究內容亦展現數位人文學科跨語境、跨地域的研究精神，舉凡分析冰島語“Icelandic”、古方言盎格魯ㄧ諾曼語“Anglo-Norman”的文學著作，亦或是探究聖歌“Cantus”的旋律調，皆展現數位人文跨學科研究合作的精神。

整體而言，在單一地區高詞頻詞彙使用方面，日本與英國的詞彙數量明顯少於其他地區，各自僅有四個特殊詞彙。可看出兩地區常用詞彙較為大眾，為一般地區皆使用之詞彙。臺灣的詞彙數量最多，詞彙使用也傾向個別地區、朝代、或特殊工具的使用。

4.2 各時期顯著詞分析

除了以地區做為觀察視角外，本研究繼以時間的面向，分析同時期出版的文本語料。其目的在於確認特定時期文本使用詞彙的特徵，從而由詞彙特徵掌握該時期數位人文研究可能之主題概念。於分析各時期的文本時，本研究先以 CorPro 進行顯著詞分析，並刪除停用詞、虛詞、數字、特殊符號與感嘆詞，接續篩選出 Chi-square 檢定顯著（$p < .001$），且 Keyness（顯著性）前 15 名之顯著詞。Keyness 的定義為：「文本中以相對於參考語料庫而言異常頻率出現的單詞；並非詞彙頻率高即代表異常頻率高，而是與計算關鍵詞偶而發生的頻率差異有關」（Scott, 2011）。藉由顯著詞分析，可突顯相較於參考語料庫，該文本語料之特殊詞彙。

我們進一步將各個顯著詞彙予以分群，並分析各群詞彙所代表的涵義。在整合五個時期所有出現的顯著詞並予以歸納分類後，發現大致可分為 6 大類：「主題」（topic）、「內容（content）、「研究工具」（tools）、「檔案與資料庫」（archives/database）、「語言學」（linguistic）與「計畫」（project）。其中，「主題」與「內容」的界定可能較容易混淆，本研究將「主題」視為能涵蓋該篇文章的研究核心，通常為作者定義之關鍵字或是學科領域；「內容」則是可能於文本中大量使用的詞彙，它不一定完全是代表此文章的意涵，可能為與主題相關的詞彙或是重要的人、事、物的名稱。

最後，以線上資訊視覺化工具 Plotdb，將顯著詞分析結果繪製成圖 2 至圖 6。除了針對各時期之詞彙分析外，本研究亦比較分析各不同時期之詞彙特徵，以觀察 2009 年至 2013 年間數位人文研究之主題概念的變遷。

4.2.1 個別時期顯著詞

2009 年的詞彙主要以新的數位人文工具與方法為研究主軸（請參見圖 2），例如“oral”（擷取口述歷史影音檔之數位工具）、“AFED”（高層級數位傳真模式）等詞彙較為顯著，顯示各國陸續開發新工具或是發展新模式，應用於人文學科的研究。

其次是研究主題與內容上，傾向計算機領域相關之詞彙，像是“e-science”、“electronic”；臺灣方面的研究相對其他地區較有極大的差異，故相關詞彙容易達到顯著；“HPC”（高效能運算平台） 為加拿大的社會科學家和人文科學（SSH）研究人員提供了豐富的研究可能性。另外在 2009 年有兩項大型 DH 計畫被大力突顯，“ACKU”為學術機構合作的數位項目，為了創造與保存阿富汗文化、政治史料；“CCRI”為基於人口普查的大型數位計畫，是第一個獲得加拿大基金會贊助的產、官、學合作項目，在 DH 跨領域合作的實踐上具有重要的代表意義。

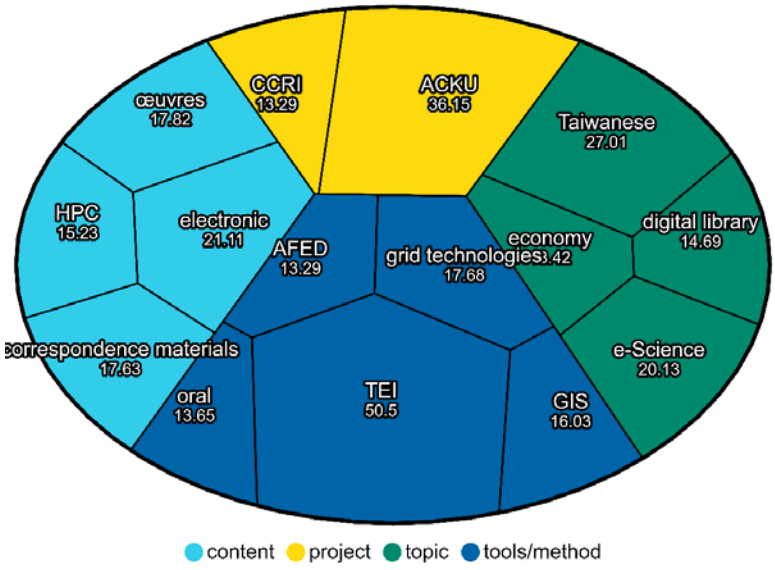


圖 2：2009 年前 15 名之顯著詞

2010 年 DH 學科較活用資料庫與檔案庫進行研究(請參見圖 3)，例如英國運用 ODIN（線上多元語言資料庫）進行語言學相關研究。此外學者更增進 DH 應用的廣度，將數位技術應用到語言學、聲韻學領域，像是識別大量歷史文獻的「loanwords」(外來；借用語)，以分析語言的聲音變化；亦或開發新的自動處理拼寫變體「spelling variation」的計算模型，幫助人文學科深入了解詞彙的演變。

在主題方面，研究中國與日本的文化、歷史史料研究明顯比例升高，諸如 “ancient China”、“Heian-Kyo”（平安京）。此外莎士比亞的劇作在 2010 年英、美 DH 研究被高度關注，無論是透過新的數位工具分析文本風格，或者繪製作品出場角色的社會網絡等，皆可讓專家學者與一般民眾探索隱藏在文本之間的連結關係。

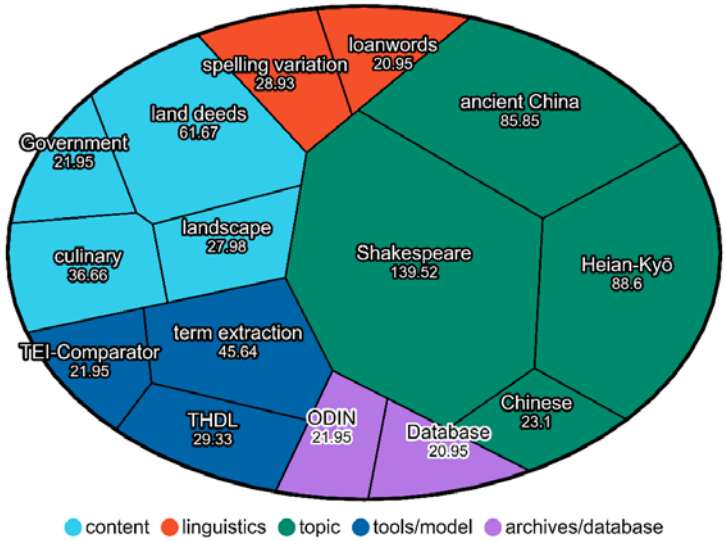


圖 3：2010 年前 15 名之顯著詞

2011 年的顯著詞顯示研究主題語文本對象愈加多元（如圖 4 所示），諸如古文字學、編纂學、翻譯譯本（translation）、辭典、經典著作、語料庫和手稿研究經常作為 DH 領域的分析對象。此外如何改進與開發新的數位工具以協助人文研究的進展一直是 DH 發展的動力目標，因此可以明顯發現 DH 研究類型在新工具和新的研究方法方面著墨許多，例如京都大學信息學研究院開發新的詞性和型態分析器“Mecab”可促進更多人參與數位人文研究活動。

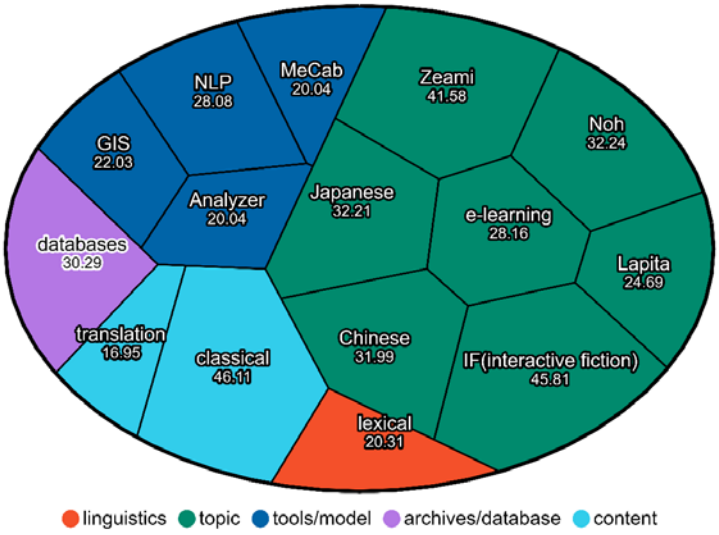


圖 4：2011 年前 15 名之顯著詞

2012 年 DH 領域之研究對象明顯擴展至多元媒體，舉凡像是以圖片導向的有聲書視覺媒體“MANGA”、“comic books”，或者是涵蓋各項領域與全球議題的 TED 會議；亦有以宗教經典《聖經》為研究主軸（如圖 5 所示）。在 DH 學科裡愈來愈多學者突破研究主題的侷限。至於計畫項目的開展可以顯示更多學者重視手稿的數位典藏與編目，像是“Bentham Project”為英國 UCL 大學的計畫之一，旨在典藏眾多 Jeremy Bentham 撰寫的原創手稿與著作。透過 DH 計畫的合作讓專家與公眾參與這些未經過進一步研究之手稿的線上轉錄語翻譯。

2013 年顯著詞分析結果顯示，與 DH 相關的計畫項目急遽增長。有以培育數位文化機構管理人職業教育的“DigCurV”，以及以考古和手稿為研究主題，建置歐洲人文藝術領域數位研究，並促進各國資源共享的“DARIAH”等，跨地區乃至於跨國界的協同合作，積極提高各領域專家學者的交流，挑戰既有的理論架構，並發展新的問題與研究方法。

數位人文研究內容亦傾向與社群媒體、使用者接觸，例如透過追蹤 Twitter[1]上使用者之間相互 tweets 次數，建立社群連結關係。此外，藉由數位技術與工具挖掘文章潛在寓意、語義的潛在語義分析“Latent Semantic Analysis”（LSA）的相關研究逐漸變多。

[1] Twitter 為全球性的社群網絡與微網誌服務的平台，提供使用者更新日常訊息「推文」，為現今網際網路上瀏覽量最大網站之一。

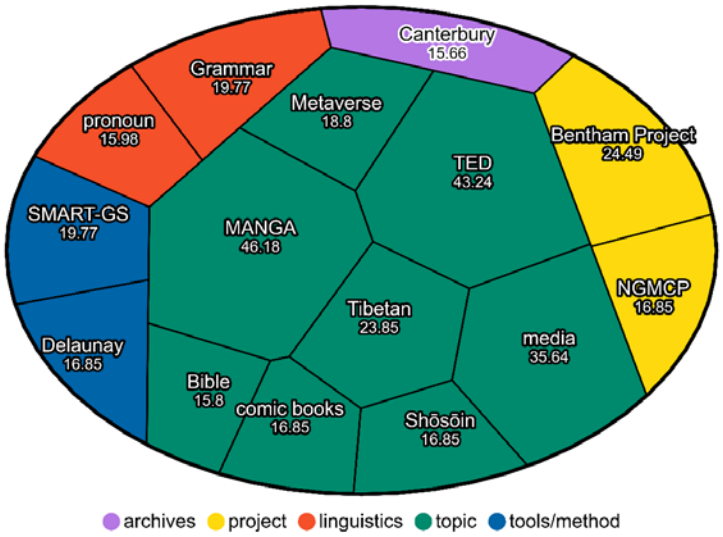


圖 5：2012 年前 15 名之顯著詞

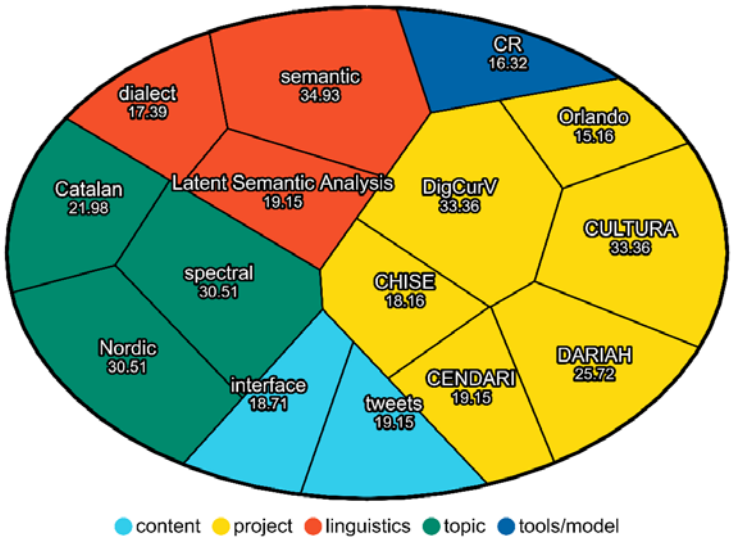


圖 6：2013 年前 15 名之顯著詞

4.2.2 整體顯著詞分析

若以顯著詞分類狀況來觀察 2009 年至 2013 年數位人文研究顯著詞變化，如圖 7 所示，可以發現數位人文研究在計畫項目（project）的詞彙於 2011 年上漲幅度顯著，可預期越來越多領域的專家、學者與機構有意願藉由協同合作和資源共享的方式，促進數位資源的整合，並提高公民參與計畫的推展。語言學（linguistic）與檔案及資料庫（archives/database）在 2009 年至 2011 年變化幅度相同，不過自 2011 年後，語言學相關的詞彙討論愈多；相反的是主要討論檔案庫與資料庫的研究漸少。

此外，探討研究主題（topic）與研究工具（tools/method）的詞彙在五個時期皆為前 15 名顯著詞。從前述研究主題的內容來看，明顯可發現許多與臺灣、日本相關聯的詞彙達到高顯著，可能的原因除了近年來跨地域的 DH 研究增多之外，英國、美國、加拿大文化背景和研究內容較為相似，詞彙整體無法突顯出來；而臺灣與日本的地理、歷史專

有名詞較特殊，研究的內容也偏向當地的風土民情，與其他地區的研究方向不同，詞彙則容易被突顯出來。

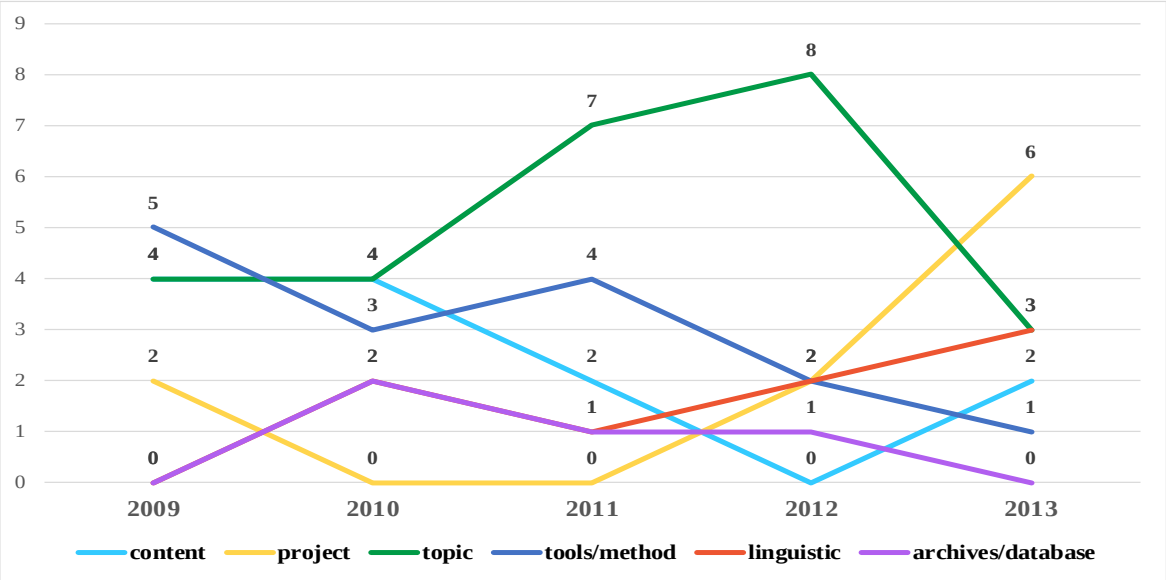


圖 7：2009~2013 年顯著詞數量趨勢

4.3 研究主題之社會網絡

數位人文研究社群的特色在於，「不同於傳統建以學系建立研究社群的方法，而是藉由跨領域、跨科系方式不斷的討論與交流」（施伯燁，民 106），學者藉由跨領域知識的分享與合作，產生更多的探索，也擴大數位人文研究領域的範疇。

因此為了瞭解數位人文研究主題的異性與共性，我們先以量化方法「餘弦相似度」檢視整體文本的相似性，初步了解不同地區在數位人文研究的異同；接著，在以模組化分析觀察各地研究主題的群聚結果；最後以社會網絡分析工具繪製出五大地區研究主題的共現結果。

4.3.1 五大地區研究領域文本相似性

文本相似性係以詞彙為基礎的餘弦相似性，計算不同地區數位人文文本之間的相似程度，結果如表 5 所示。日本 2009 年至 2013 年數位人文領域的研究內容及主題與其他地區皆大多中度相關。在日本 2013 年的研究內容與英國 2013 年（0.737**）、美國 2013 年（0.712**）高度相關，顯示日本地區於數位人文的研究內容普遍與其他國家相近。經檢視日本 2011 至 2013 年 JADH 的摘要可以發現，研究內容大多廣泛且取材多元，經常結合不同國家之學術資源作為研究主題的創新應用。例如日本學者在中國研究方面，曾針對中國佛教（Buddhism）術語辭典「法寶義林（Hobogirin）」進行數位典藏研究。

至於日本 2013 年文本與英國 2013 年、美國 2013 年文本高度相似，本研究認為其原因與日本數位人文學會（JADH）於 2013 年正式成為 ADHO 旗下成員亦有關聯。JADH 為促進更多的數位人文研究，致力於與歐美國家進行國際合作，並持續以成為 ADHO

會員為目標。其出版刊物 JJADH 的刊載標準也多以資訊科學、跨領域科學為主。

表 5：各地區文本相似性

	CA_09	CA_10	CA_11	CA_12	CA_13	JP2_11	JP_12	JP_13	TW_09	TW_10	TW_11	TW_12	UK_09	UK_10	UK_11	UK_12	UK_13	US_09	US_10	US_11	US_12	US_13
CA_09		0.153	0.224	0.230	0.181	0.331	0.324	0.317	0.159	0.221	0.229	0.222	0.245	0.274	0.184	0.159	0.299	0.400*	0.209	0.351	0.383	0.396
CA_10			0.261	0.231	0.261	0.291	0.308	0.300	0.191	0.236	0.207	0.217	0.312	0.281	0.308	0.267	0.298	0.316	0.126	0.306	0.204	0.253
CA_11				0.362	0.205	0.414*	0.351	0.328	0.275	0.313	0.240	0.162	0.343	0.213	0.212	0.136	0.247	0.421*	0.197	0.400	0.390	0.411
CA_12					0.272	0.512*	0.396	0.568*	0.319	0.363	0.306	0.335	0.396	0.289	0.352	0.200	0.398	0.586*	0.314	0.564*	0.543*	0.573*
CA_13						0.457*	0.329	0.498*	0.394	0.449*	0.406*	0.437*	0.464*	0.331	0.434*	0.289	0.425*	0.276	0.148	0.309	0.272	0.298
JP_11							0.595*	0.772**	0.509*	0.604*	0.576*	0.550*	0.593*	0.482*	0.513*	0.329	0.591*	0.657*	0.367	0.599*	0.614*	0.672*
JP_12								0.674*	0.434*	0.458*	0.475*	0.390	0.504*	0.530*	0.484*	0.448*	0.616*	0.524*	0.254	0.441*	0.438*	0.467*
JP_13									0.518*	0.594*	0.565*	0.586*	0.630*	0.598*	0.600*	0.455*	0.737**	0.707**	0.418*	0.672*	0.664*	0.712**
TW_09										0.501*	0.654*	0.427*	0.439*	0.442*	0.474*	0.420*	0.516*	0.365	0.163	0.330	0.287	0.315
TW_10											0.572*	0.507*	0.457*	0.354	0.436*	0.295	0.491*	0.437*	0.268	0.458*	0.418*	0.456*
TW_11												0.538*	0.503*	0.420*	0.476*	0.324	0.520*	0.431*	0.209	0.390	0.366	0.390
TW_12													0.439*	0.470*	0.553*	0.365	0.569*	0.403*	0.211	0.415*	0.371	0.415*
UK_09														0.463*	0.458*	0.345	0.534*	0.549*	0.298	0.472*	0.482*	0.503*
UK_10															0.741**	0.805**	0.842**	0.414*	0.188	0.348	0.314	0.345
UK_11																0.699*	0.797**	0.378	0.199	0.379	0.302	0.332
UK_12																	0.752**	0.227	0.079	0.229	0.139	0.170
UK_13																		0.524*	0.253	0.467*	0.421	0.452*
US_09																			0.474*	0.727**	0.798**	0.818**
US_10																				0.397	0.493*	0.494*
US_11																					0.696*	0.720**
US_12																						0.815**
US_13																						

註：中度相關(0.4~0.699)*　高度相關(0.7~0.999)**

臺灣文本相似程度普遍落在中度相關程度，且大多為內部一致，或者與日本一致程度高，與美國、英國及加拿大文本的相似性並沒有很高。從前述高詞頻分析、顯著詞分析研究可觀察出，臺灣數位人文研究傾向有在地化的趨勢，這是和其他地區的數位人文研究的較大區別。歐美地區的研究方面，美國與英國在 2009 年至 2013 年內部研究大部分達到高度相關，顯示各自文本研究高度一致性。加拿大則是與日本、美國文本呈現中度相關。研究者認為加拿大文本與日本文本相似，由於兩者皆為研究主題範疇廣，能夠出現相同主題詞彙的比例高。而加拿大文本與美國文本相似，取決於兩者地理位置相對較近，且便於兩地區學術合作。

4.3.2 研究主題之分群

為探究數位人文研究主題在不同地區的社會網絡關係，研究者嘗試以 JSTOR Labs 開發的在線文字分析工具 Text Analyzer 取得每篇文件的關鍵主題，再使用社會網絡分析（SNA）軟體進行模組化分析，比較不同地區在 2009 年至 2013 年研究主題的分群結果。為檢驗其分群結果的可信度，研究者使用兩種不同的 SNA 工具 Gephi 和 VOSviewer，進一步比較其分群結果的差異。

兩種 SNA 工具針對數位人文研究主題的分析結果如表 6 所示，根據研究主題的相近程度，越多相近研究主題的文件彼此會被歸類在同一群體。最終 VOSviewer 歸類成九群，而 Gephi 歸類成八群。整體來看，分群結果大致一致，除了美國 2010 被 VOSviewer 判定單獨形成一群，變動最大的主要為加拿大與日本在 2011 年到 2013 年的分群結果。

研究者綜合前面文本相似性分析結果和兩者研究主題內容，認為加拿大與日本地區

在兩種不同 SNA 工具分群不一致的原因在於：加拿大和日本在數位人文領域的研究內容皆相當廣泛，且自 2011 年後研究方向更傾向於英國與美國數位人文，使得分群上較難以被歸類在哪一個特定群體，僅能依據比例上較偏向於哪一個地區來做分類。整體上 Gephi 和 VOSviewer 兩項工具在研究主題的模組化分析結果相當一致，顯示其分群具有可信度。

表 6：Gephi 與 VosViewer 分群之差異

	VOSviewer	Gephi
第一群	UK_2013	UK_2013, CA_2013, JP_2012, JP_2013
第二群	US_2009, US_2011, US_2012, US_2013, CA_2011, CA_2012, JP_2013	US_2009, US_2010, US_2011, US_2012, US_2013, CA_2011, JP_2011
第三群	UK_2012	UK_2012
第四群	UK_2011	UK_2011, CA_2012
第五群	JP_2011, JP_2012, TW_2009, TW_2010, TW_2011, TW2012	TW_2009, TW_2010, TW2011, TW2012
第六群	UK_2010	UK_2010
第七群	CA_2010, CA_2013, UK_2009	CA_2010, UK_2009
第八群	CA_2009	CA_2009
第九群	US_2010	X

資料來源：研究者自行整理

大致了解各地區數位人文研究領域摘要文本的相似程度後，本研究進一步探究文本所呈現之研究主題的共現狀況。

4.3.3 各地區核心研究主題與共現關係

總計 588 篇摘要經由 JSTOR Text Analyzer 分析，扣除重複詞彙、權重在 5 以下的詞彙後，共獲得 910 個主題詞彙。本研究將每篇文章被 JSTOR 賦予的主題詞彙建立關係，以同一期刊、同一出版年為基礎，將所有主題詞彙聚集，再以主題詞彙建立各地區的共現關係，最終以 VOSviewer 繪製詞彙共現圖。

研究主題之共現結果，如圖 8 所示。圖中節點代表意義共有兩種，一種為文件節點，以地區和時間命名，代表該時期某一地區之文本總和（例如：UK2009）；另一種為主題詞彙節點（例如：Digital humanities）。文件節點越大，顯示連結的主題詞彙越多；詞彙節點越大，其意謂著該主題詞彙同時被多個文件提及，代表越多文件節點有該詞彙所代表的研究主題。連接文件節點與主題詞彙節點的線條，代表該文件包含該主題詞彙，連接的線條越粗，代表連接的權重強度越大。

整體而言，數位人文學科核心研究主題圍繞在「數位人文」、「文本分析」、「資訊科學」、「記錄管理」、「教育科技」、「資訊科技」等（如表 7、圖 8 所示）。不同地區數位人文研究主要經由這幾個核心主題詞彙互相連結形成共現網絡。前 20 名主題詞彙顯示數位人文研究主題整體傾向科技（Technology）類型，較少數屬於社會科學與人文（Social science & Humanities），像是「語言學」、「歷史批判」等。從集群來看，前 20 名主題詞

彙主要出自第一群、第二群與第六群，詞彙主要來自英國與美國的學術論文。

研究者參考 Chen & Hsueh（2015）以「社會科學及人文」（Social science and Humanities, SSH）和「科技」（Technology, T）兩個領域，觀察個別地區核心主題詞彙的共現情形，探索數位人文主題詞彙的分布狀態。以下說明各地區的主題詞彙。

表 7：前 20 名研究主題共現詞彙

排名	主題詞彙	所屬群體	權重（整體連結強度）
1	Digital humanities	2	181
2	Text analytics	1	146
3	Information science	1	85
4	Records management	3	68
5	Educational technology	3	46
6	Information technology	2	34
7	Linguistics	1	33
8	Higher criticism	6	30
9	Databases	5	30
10	Project management	2	27
11	Machine learning	4	26
12	XML	7	25
13	Reading instruction	6	22
14	Dictionaries	6	22
15	Mapping	5	22
16	Syntax	4	22
17	Applied linguistics	4	22
18	Systems analysis	1	21
19	Spacetime	1	20
20	Information search	6	20

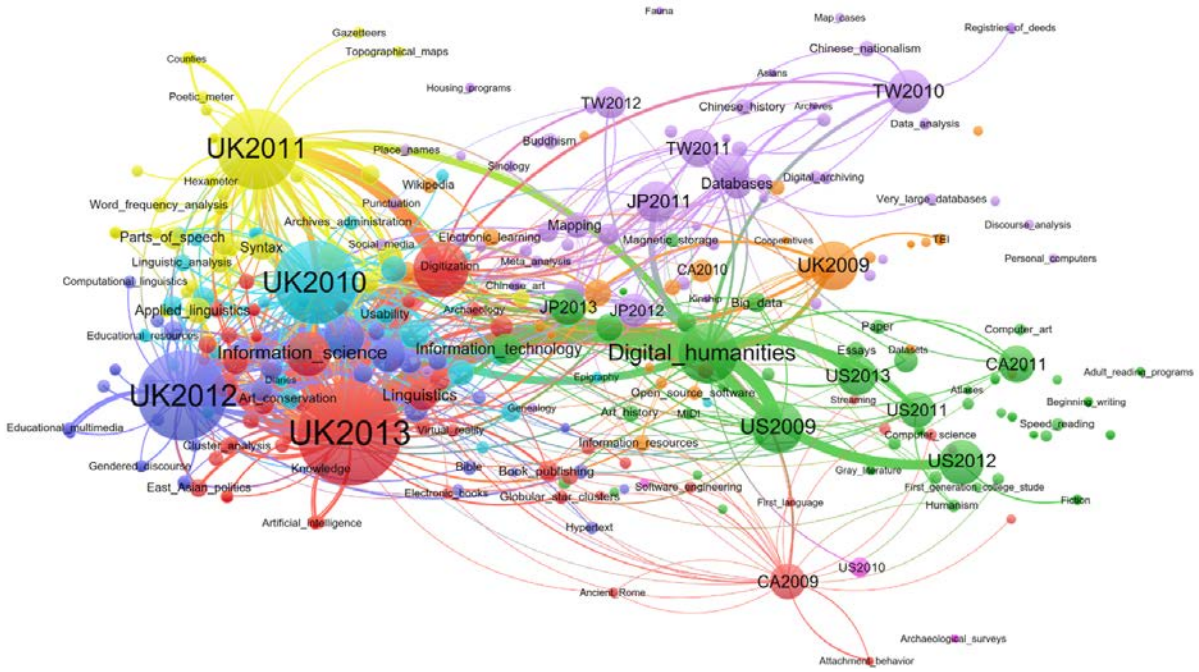


圖 8：五大地區 2009-2013 年研究主題詞彙共現

4.3.3.1 英國地區主題詞彙

英國地區主題詞彙分群結果顯示，詞彙主要分布在第一、三、四、六與七群，如圖 9 所示。

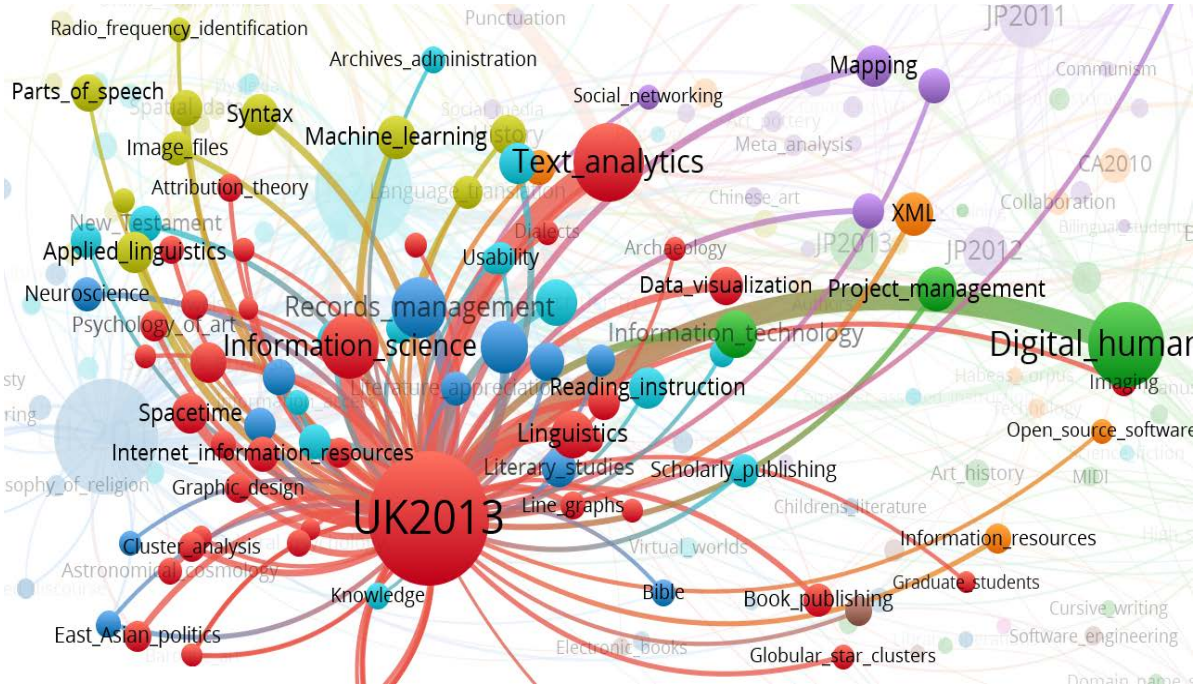


圖 9：英國研究主題詞彙共現

英國 2013 年核心主題（第一群）以「文本分析」、「系統分析」、「資訊科學」、「語言學」和「資訊視覺化」等為主，主題傾向科技類型。在社會科學與人文領域方面，藝術心理學「Psychology of art」與藝術保存「Art conservation」占據最大比例，顯示英國重視數位人文在藝術學門的應用。

英國 2012 年的主題詞彙主要是以「記錄管理」、「教育科技」、「第二語言學習」、「博物館研究」、「文學賞析」、「神經科學」和「計算機網路」等（第三群）。顯示英國亦偏重將教育、檔案結合數位人文進行研究，開發創新的教學方法或者是檔案資料庫等。英國 2011 年（第四群）的主題詞彙主要是「機器學習」、「句法研究」、「應用語言學」、「詞性」、「圖像文件」、「語言翻譯」和「社會科學史」等，相比於其他時期的研究，2011 年的數位人文與其他時期有明顯差異，著重在語言學的相關應用。

英國 2010 年主題詞彙以「歷史批判」、「資訊檢索」、「閱讀指導」、「學術出版」、「英國文學」等為主（第六群）。其中宗教經典「新約全書《New Testament」》於該時期被學者多次探討。英國 2009 年的主題詞彙以「XML」、「資料模型」、「電子學習」、「資訊資源」、「合作」等（第七群），偏向資訊技術領域，社會與人文科學領域的主題詞於 2009 年並沒有展現。

綜上所述，英國 2009 至 2013 年數位人文研究主題整體傾向科技技術類別，尤其在網際網路、資訊科學領域；社會科學與人文方面偏向藝術、語言、教育、歷史與圖書館科學領域，其分析結果與 Chen 與 Hsueh（2015）的主題標籤分布結果大致相同。

4.3.3.2 美國與加拿大地區主題詞彙

美國地區主題詞彙分群主要分布在第二群與第九群，此外加拿大 2011 至 2012 年與日本 2013 等地區之研究主題詞彙與美國相近，故被分群在同一類別下。整體而言，因地理位置與文化背景相近，核心主題為「數位人文」、「資訊科技」、「計畫管理」、「線上社會網絡」、「巨量資料」、「藝術史」、「遊戲」、「人文主義」，還有中世紀宗教藝術的「泥金裝飾手抄本」（illuminated manuscript）等，如圖 10 所示。

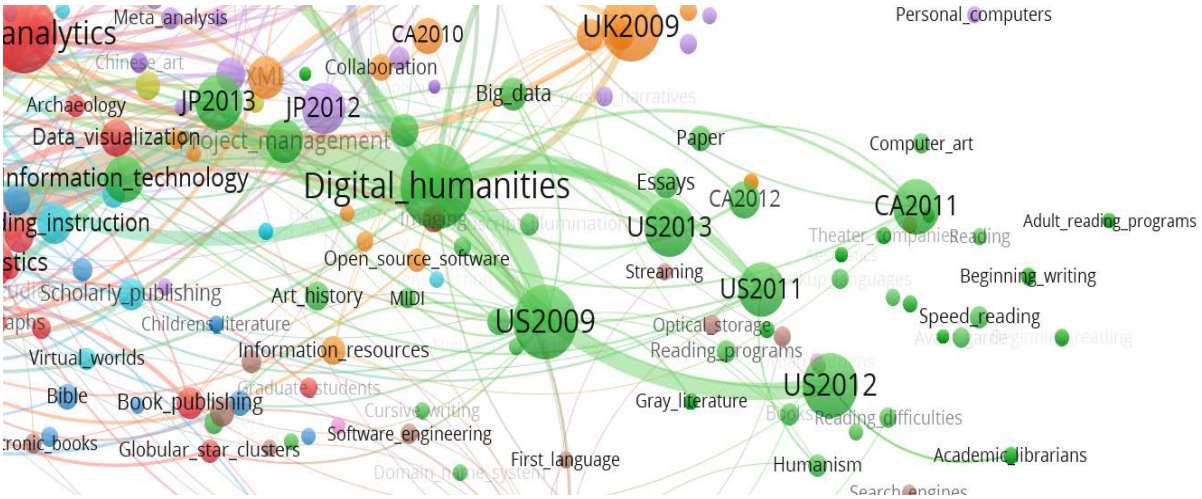


圖 10：美國研究主題詞彙共現

研究方向同樣以科技類型為主，大致傾向「數位人文」（Digital humanities）、「資料探勘」（Data mining）和「網際網絡」方面，亦有許多數位圖書館、資訊科學領域相關的主題詞彙；社會科學與人文方面則著重在人文學與藝術領域，亦結合數位人文與多媒體、宗教的發展應用。至於美國 2010 年主題詞彙偏向「文字處理軟體」、「考古調查」、「書目紀錄」、「都市農業」，研究取向明顯與其他時期有較大的差異。

4.3.3.3 臺灣與日本地區主題詞彙

2011 至 2012 年臺灣以及日本的數位人文核心主題主要以「資料庫」、「地圖」、「東亞文學」、「數位圖書館」、「圖書館館藏」、「資料庫管理系統」、「佛教」、「中國民族主義」、「圖書館」、「中國歷史」等，請參見圖 11。

明顯可看出，臺灣與日本數位人文研究內容帶有濃厚的地方性與文化特色，且偏重數位人文應用在人文與社會科學的文本語料研究。此外與英國、美國和加拿大數位人文研究主題的取向有所差異，明顯著重圖書館學、歷史、地理與宗教學科領域的加值應用；至於科技類型的研究主題，主要傾向資料庫管理與資訊科學的發展。

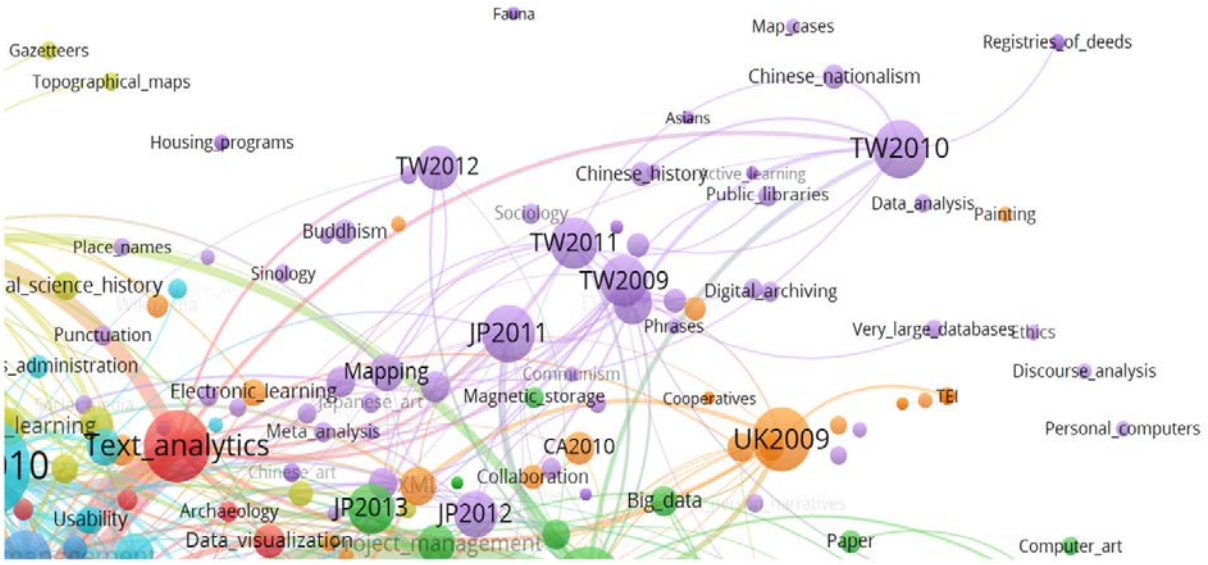


圖 11：臺灣與日本 DH 研究主題詞彙共現

5. 結論與建議

數位人文領域的概念引發眾多面向的探討與論辯，更多的跨領域探索，為數位人文研究帶來新的見解、型態、交流活動與現象，展現出多元的面貌。臺灣與國際已有許多數位人文研究成果使用了內容分析的方法，探討許多其他研究領域或是子領域的特定課題，然而卻沒有將數位人文學術論文視為標的文本語料，探討數位人文研究本身的學術研究發展與科學語彙的生成。本研究以庫博（CorPro）語料分析軟體檢視不同地區於數位人文文本高頻詞彙，並建立出中央核心詞彙、半邊陲詞彙與邊陲詞彙，呈現各地區之研究特性，觀察顯著詞彙的差異可以反映出各時期於數位人文研究的特色趨勢變化，藉由社會網絡分析建立核心研究主題網絡以及共現關係，則可以探索數位人文整體的學科生態。以下為本研究發現：

(1) 五大地區數位人文詞彙使用網絡呈現環狀分佈，依照高頻詞彙連結程度，可以將其分為中央核心詞彙、半邊陲詞彙與邊陲詞彙。中央核心詞彙為五地區皆常用之詞彙，由“digital”、“information”、“humanities”、“technology”等詞彙構成，呈現出數位人文學科領域核心主題內容仍然是資訊科學與人文學。

(2) 半邊陲詞彙呈現不同地區在詞彙使用上的共性，通常為數位人文領域重要的研究主題或研究方法，如“database”、“corpus”、“system”等，或者呈現出兩地區密切學術交流的研究取向，或共同合作之研究計畫。

(3) 各地區高頻詞彙顯示，臺灣以在地化研究為主；其他四個地區反映期刊收錄特色，日本以技術、架構層面為主；英國經常進行語言學、資訊軟體相關研究；美國重視當代藝術與當代文學的加值應用；加拿大則是以中世紀文物、典藏手稿等第一手資料為主，亦包含許多跨語言、跨地域的研究。

(4) 2009 年至 2013 年整體數位人文研究方向從著重數位檔案、資料庫的量化資料分析，走向以電腦技術輔助語言、歷史與人文藝術研究，重視將數位科技加值應用於數位文本分析、文學評論等研究主題。整體研究內容取向呈現質性、歷史批判性與詮釋性。此外各項數位人文計畫的不斷開展，揭示許多學者機構重視跨地區性的數位人文研究，並提高國際的合作交流。

(5) 數位人文研究的相似性受其地理位置、文化與國際機構組織的影響，臺灣與日本在 2012 年以前的研究相近；加拿大研究主題廣泛再加上地利之便，與美國、英國的研究相似度高。日本 2013 年後致力於與歐美國家進行數位人文計畫合作，文本內容傾向與英美國家相近。

(6) 整體研究主題結以資訊科技相關的主題詞彙較多。若以各地區主題詞彙分布來看：日本 2013 年與英國、美國、加拿大整體傾向科技類別為主，包含資料科學、網際網路、巨量資料等領域；臺灣與日本 2009 年至 2012 年主題詞彙在科技類別與人文、社會科學類別的分佈較為平均，且重視圖書館學、歷史、地理與宗教學科領域的加值應用。

本研究雖然使用多地區、長時距、多文本的學術文獻，企圖探討數位人文學科語彙的使用與變化。然而，數位人文仍是持續快速發展的學術領域，本研究的成果僅是初期數位人文研究的景況，但是本研究使用的方法可以持續應用於探索數位人文的發展，未來可以擴大空間面向的範疇，例如歐洲大陸德國以及中國大陸目前致力於數位人文的記基礎建設，相信對於數位人文的發展，都是可以期待的。此外，仍有一些問題必須要克服的，各地區的數位人文學術論文的蒐集，是類似研究方法重要的挑戰，這有待數位人文領域本身學術資訊傳播管道的良好規劃，以及學術文獻資料庫的建設。

致謝

本研究感謝科技部研究計畫的支持，計畫編號：MOST 105-2420-H-002-056-MY2。

參考文獻

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 1-12.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & J. Schnapp (2012). Digital Humanities. Cambridge, Mass: MIT Press.

Chen, K.-H., & Hsueh, B.-S. (2015). Exploring Research Characteristics and References Patterns of Scholarly Literature in Digital Humanities. (master's thesis). Retrieved June 20, 2018 from https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi?o=dnclcdr&s=id=%22103NTU05

448009%22.&searchmode=basic

Liu, A. (2012). “The State of the Digital Humanities: A Report and a Critique.” Arts and Humanities in Higher Education, 11(1-2), 8-41. DOI: 10.1177/1474022211427364.

Schnapp, J. and Presner, P. (2009). “Digital Humanities Manifesto 2.0”, Retrieved July 16, 2018, from http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf

Scott, M. (2011). WordSmith Tools Manual, Version 6. Liverpool: Lexical Analysis Software Ltd.

Wang, X., & Inaba, M. (2009). Analyzing Structures and Evolution of Digital Humanities Based on Correspondence Analysis and Co-word Analysis. アート・リサーチ, 9, 123-134.

Warwick, C., Terras, M., & Nyhan, J. (2012). *Digital humanities in practice*. London: Facet Publishing in association with UCL Centre for Digital Humanities.

王麗蕉（民 104）。數位人文系統的建置與加值應用： 以臺灣日記知識庫為探討中心。**漢學研究通訊，34（4）**，30-39。

呂姿玲、莊惠茹（民 107）。國家圖書館通用型古籍數位人文研究平台。**國家圖書館館訊，156**，26-27。

施伯燁（民 106）。數位時代的人文研究：數位人文發展沿革、論辯與組織概述。**南華社會科學論叢**，3-19。

項潔、涂豐恩（民 100）。什麼是數位人文。載於項潔（主編），**從保存到創造：開啟數位人文研究**（9-28 頁）。臺北市：臺大出版中心。

項潔、陳麗華（民 103）。數位人文－學科對話與融合的新領域。載於項潔（主編），**數位人文研究與技藝**（9-23 頁）。臺北市：臺大出版中心。

鄭允人（民 104）。**數位人文學科知識整合趨勢之研究**。（未出版之碩士論文）。國立臺灣大學，臺北市。

闕河嘉（民 107）。庫博中文語料庫分析工具的數位人文價值。人文與社會科學簡訊，

闕河嘉、陳光華（民 104）。中文獨立語料庫分析工具之開發與應用。**在第六屆數位典藏與數位人文國際研討會論文集**。臺北：臺灣大學。

蘇信寧（民 104）。數位人文主題研究計畫分析與管理。**人文與社會科學簡訊，17（1）**，89-100。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

古籍數位人文研究平台之史料人物關係圖工具發展與應用

Development and Application of an Ancient Books Digital Humanities Research Platform with Characters' Relationship Map Tool

陳志銘*　張鐘**

國立政治大學圖書資訊與檔案學研究所教授*

國立政治大學華人文化主體性研究中心 IT 工程師**

古籍數位人文研究平台之史料人物關係圖工具發展與應用

Development and Application of an Ancient Books Digital Humanities Research Platform with Characters' Relationship Map Tool

陳志銘

國立政治大學圖書資訊與檔案學研究所教授

張鐘

國立政治大學華人文化主體性研究中心 IT 工程師

摘要

本研究旨在開發支援數位人文研究之「古籍數位人文研究平台之史料人物關係圖工具」，能自動識別文本中的人名，同時提供易上手的即時互動介面，透過人機互動方式協助人文學者更有效率且正確的建立擬分析文本之人物社會關係，以探索複雜的人物社會網絡關係，找到有用的研究發現。本研究以準實驗研究法比較有無使用本研究發展之「史料人物關係圖系統」在支援人文學者解讀文本中的人物與人物關係成效，以及科技接受度是否具有顯著差異，並輔以半結構式深度訪談了解人文學者對於本研究發展之「史料人物關係圖工具」的看法與感受，也使用滯後序列分析(lag sequence analysis)分析受測者使用「史料人物關係圖」解讀人物與人物關係的行為歷程。實驗結果發現，採用本研究發展之「史料人物關係圖系統」，在解讀人物與人物關係的成效上高於使用無史料人物關係圖工具的系統，但未達統計上的顯著差異。採用「史料人物關係圖」系統的技接受度顯著高於無史料人物關係圖的系統。此外，由訪談結果歸納得知，受訪者對系統介面的整合與操作的流暢度及外部查詢功能給予正面的評價，並認為史料人物關係圖能方便他們了解整體文本中的人物脈絡。本研究發展之史料人物關係圖中所提供的人物與人物關係資訊若能更加精準與廣泛，將更具有實用性。

目次:

1. 緒論 4

2. 文獻探討 6

2.1. 數位人文 6

2.2. 社會網絡分析 6

2.3. 科技接受模式 8

3. 系統設計 9

3.1. 系統架構 9

3.2. 系統元件 11

3.3. 系統介面與功能 12

4. 研究方法 17

4.1. 研究設計 17

4.2. 研究對象 17

4.3. 研究工具 17

4.4. 實驗流程 18

5. 實驗結果與分析 20

5.1. 有無史料人物關係圖系統解讀人物數量之成效比較分析 ...20

5.2. 有無史料人物關係圖系統解讀人物關係數量成效比較分析 20

5.3. 有無史料人物關係圖系統科技接受度之差異分析 20

5.4. 史料人物關係圖系統歷程行為資料統計 21

5.5. 史料人物關係圖系統使用行為序列分析 22

5.6. 訪談分析 23

6. 結論與未來研究 24

6.1. 結論 24

6.2. 未來研究方向 26

關鍵字: 數位人文，社會網絡分析，人機互動，文本探勘，資訊視覺化

1. 緒論

隨著資訊科技的進步與發展，資訊科技被廣泛應用於各種領域之中，當然也包括人文研究領域。傳統上人文學者多半以紙本形式的文本做為研究對象，而現今結合資訊科技，許多研究機構開始將歷史檔案、文本資料進行數位化，建立其數位典藏資料庫，使得人文研究領域的研究環境與知識獲取管道產生了巨大的改變，人文學者使用資料庫輔以研究已逐漸成為研究的必經過程[1]。而如何將典藏的龐大數位資源進行有效的支援人文研究應用，已成為人文學者必須思考的問題。因而近幾年來產生「數位人文」之新興研究領域，透過全文數位資料及數位工具的輔助，幫助人文學者觀察不易透過解讀傳統紙本文本所能觀察到的現象[2]。

在數位人文研究領域中，透過資訊科技的輔助，再搭配運用文本探勘(text mining)或資料探勘(data mining)的技術，可以幫助人文學者發現隱含在大量文本資料中的關係[3]。 陳詩沛、杜協昌與項潔結合數位化的臺灣土地契約文書與明清中央政府的行政檔案，運用自動化的技術，擷取台灣土地契約文書和明清中央政府的行政檔案中個別文件的詞彙特徵，重建這兩種檔案之間的特殊關係，並利用此方法找出隱含在這些檔案中的相關脈絡，人文學者可依據這些脈絡，進一步進行學術分析或討論[4]。此一方法雖然無法完全取代人文學者的角色，但能夠縮減人文學者在原本研究工作中重複或無趣的部分，以更有效率的進行人文研究。

此外，透過資訊科技的輔助，也可以協助人文研究者找出隱含在大量史料、文本中複雜的人物網絡關係。哈佛大學與中央研究院歷史語言研究所及北京大學所合作之中國歷代人物傳記資料庫(CBDB)，就是透過累積大量的中國歷代人物傳記資訊，利用群體傳記學（prosopography）的方式輔以人文研究[5]。群體傳記學是找出某一個群體所共享的身分，比如教育程度、出身背景乃至宗教信仰等，藉此對社會現象進行分析[6]。研究清代的學者 Guy 利用群體傳記學的方式研究清朝巡撫，比對、統計巡撫的籍貫、教育背景與科舉成績等，發現巡撫職位多半為這些人升官的中繼站，由此方法所得出的歷史解釋，有數據作為佐證，使論點能夠更加穩固[7]。而中國歷代人物傳記資料庫群體傳記學的另一研究面向為人際網路的分析，主要針對許多一對一關係組構而成的複雜網絡進行分析，而非人物群

[1] Hockey, “The History of Humanities Computing. In A Companion to Digital Humanities,” pp. 1–19.
[2] Schreibman, S., Siemens, R., & Unsworth, J. *A Companion to Digital Humanities.*
[3] 項潔、涂豐恩。(導論--甚麼是數位人文)，收於項潔編，《從保存到創造: 開啟數位人文研究》，頁 9- 28 。
[4] 陳詩沛、杜協昌、項潔，(史料整體分析工具之幕後-介紹「台灣歷史數位圖書館」的資料前置處理程序)，收於項潔編，《從保存到創造: 開啟數位人文研究》，頁 51 - 66 。
[5] China Biographical Database (CBDB). Retrieved September 14, 2018, from https://projects.iq.harvard.edu/cbdb/how-cite-cbdb
[6] Stone, L. “Prosopography, ” Daedalus, 100(1), pp. 46–79.
[7] Guy, R. K. Qing Governors and Their Provinces: The Evolution of Territorial Administration in China, 1644-1796. University of Washington Press.

體共享的特性[8]。以往人工分析方式僅能侷限於有限數量的文本，或明顯、單純的關係，透過資訊科技的輔助，人文學者可觀察人際網路中不同節點的互動，可能會發現隱含其中更為複雜的人際關係，並提出新的論點。

台灣於 2002 年發起「數位典藏國家型科技計畫」後，許多學術機構開始將其所典藏的紙本典藏品進行數位化典藏，10 多年間累積大量數位資料庫，但 Rosenzweig 提出研究者面臨的並不是資料匱乏，而是面對過於龐大的資料要如何處理，並讓這些資料產生意義，因此近幾年來許多對於資料內容分析的數位工具被發展出來[9]。台灣大學發展的「台灣歷史數位圖書館」，其資料庫典藏內容包含「淡新檔案」、「明清台灣行政檔案」、「古契書」等超過十萬筆的全文資料，並且提供許多分析資料的數位工具，但此一系統缺少與使用者進行即時性互動的介面，主要發展的數位工具仍以檢索功能協助查詢資料為主。而另一平台「DocuSky」提供使用上傳文本，並可使用平台發展之文本分析工具輔以人文學者進行文本解讀[10]。法鼓山的「CBETA 數位研究平台」提供了漢文佛經的線上閱讀介面，亦提供人名、地名及社會網絡關係圖讓人文學者參考。

綜合上述，數位人文學者和科學家都相當依賴「工具」，運用數位工具能解決過去的人文研究問題，但是目前尚缺乏一個跨平台的即時性社會網絡分析工具，並且提供友善的互動性介面，可以有效輔助人文學者進行人物社會網絡分析之數位人文研究。因此，本研究於古籍數位人文研究平台上發展支援數位人文研究之「史料人物關係圖工具」，能自動識別文本中的人名，同時提供易上手的即時互動介面，透過人機互動方式協助人文學者更有效率且正確的建立擬分析文本之人物社會關係，以探索複雜的人物社會網絡關係，找到有用的研究發現。為了驗證此一系統支援數位人文研究的效益，本研究以文本解讀理解人物與人物關係之成效、科技接受度、使用者行為分析及訪談，評估人文學者使用此一工具支援數位人文研究的成效與看法，並根據研究結果提出對此一系統的改善建議方向。

[8] 同註 3

[9] Rosenzweig, R. “Scarcity or Abundance? Preserving the Past in a Digital Era,” The American Historical Review, 108(3), pp. 735–762.

[10] DocuSky. 2017. Retrieved September 14, 2018, from http://docusky.digital.ntu.edu.tw/DocuSky/ds-01.home.html

2. 文獻探討

2.1. 數位人文

傳統人文研究係指人文學者透過大量紙本文獻的解讀並進行分析後，從中找出文本中的思想內容及時空脈絡。而隨著資訊科技的進步，數位化的技術逐漸成熟，人文學者開始藉由資訊技術的輔助進行文本內容的蒐集、分析和應用，逐漸形成一門新興的研究領域－「數位人文」(Digital Humanities)。Berry 將數位人文定義為運用資訊科技技術輔助人文學者從事人文研究，乃資訊技術在人文領域的應用 [11]。

數位人文的發展基礎為大量具有全文之數位化典藏資源，以及基於這些數位化資源所設計的數位分析工具。例如國立臺灣大學圖書館的「台灣歷史數位圖書館」資料庫，典藏包含「淡新檔案」、「明清臺灣行政檔案」、「古契書」三個文獻集，提供人文學者豐富的數位化典藏研究資源 [12]，該資料庫除了提供全文檢索介面，並在檢索後顯示「檢索結果年代分佈圖」、「檢索後分類」功能。另有針對古地契的「契約文書買賣角色分析」和針對系統的「前後綴詞分析工具」等數位研究工具，協助研究者理解文本中的社會關係脈絡及文件詞彙分佈關係 [13]。此外，中國歷代人物傳記資料庫(China Biographical Database，簡稱 CBDB)」收錄超過 360,000 中國歷代人物傳記和譜系資料，研究者可以透過此資料庫檢索歷史人物之籍貫、官職等資料，並且也開始嘗試發展社會網絡分析工具支援數位人文研究。此外，DocuSky 為國立臺灣大學發展的數位人文平台，此一平台可讓研究者透過工具自行建立資料庫，上傳自身所典藏的文本，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，幫助人文研究者查找資料、進行內文比對，以及進行字詞相關統計等，增加人文學者對於文本內容的掌握 [14]。

綜合上述，提供友善、容易操作的數位人文研究工具，輔助人文學者進行研究，是支持數位人文發展的關鍵。本研究因此發展「古籍數位人文研究平台之史料人物關係圖」，以支援人文學者進行文本中之文物關係脈絡解讀，希望提升人文學者進行數位人文研究的效益。

2.2. 社會網絡分析

社會網絡(Social network)一詞起源於 Barnes 的 Human Relations 的文章中，主要係統整社會科學家或大眾一般認知的社會關係概念，像有限制的群體，如部

[11] Berry, D. 2012. Understanding digital humanities. London: Palgrave Macmillan.
[12] Chen, S. P., Hsiang, J., Tu, H. C., & Wu, M. “On Building a Full-Text Digital Library of Historical Documents. In Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, pp. 49–60.
[13] Taiwan History Digital Library (THDL). 2009. Retrieved September 14, 2018, from http://doi.org/10.6681/NTURCDH.DB_THDL/Text
[14] Tu, H. C. 2016. “DocuSky: a DH Platform in the Making”. Presented at the The 7th International Conference of Digital Archives and Digital Humanities, Taipei.

落、家庭，或者社會分類，如性別、種族，並以系統化方式呈現其關係模式[15]。Wasserman 和 Faust(1994)兩位學者對於社會網絡分析的定義為，社會網絡分析是社會學中的一個方法，主要透過分析關係之間的模式與社會活動者之間的互動模式，並從中尋找潛在的社會網絡結構[16]。

社會網絡的組成主要包含三種要素；行為者(actors)、關係(relationships)及連結 (linkages)[17](Hanneman & Riddle,2005)，行動者為社會網絡中的節點(nodes)，一個節點可代表個人、組織或公司，為構成社會網絡的基本單位，同時也是社會網絡的主體；關係指在行動者間因為交流、屬性等因素，形成某種型態的連結，在社會網絡中以線表示點與點之間的連結，而線表示一種關係。關係亦可分成有向性(directed)及無向性(undirected)，隨著不同的關係類型，會建構出不同的社會網路；連結指行動者之間連接的途徑，當行動者彼此之間建立關係時，可透過路徑(path)直接或間接產生關係。

Moretti(2011)將社會網絡理論應用於文本的情節分析中，利用社會網絡分析將敘事性的文本中內容的空間與時間結構、核心角色等轉換成網絡結構圖，用邊與節點形成情節中的社會網絡，而隨著情節的發展，網絡圖也會有所變化，能讓研究者輕易了解文本中關係與情節的變化，進行更深一層的探究[18]。Moretti(2011)藉由莎士比亞的名著《哈姆雷特》的社會網絡圖找出關鍵人物 Horatio，並從 Horatio 的社會網絡關係探討故事的脈絡。許多經典或著名文學研究，都有許多學者運用社會網絡分析對文本中的人物關係、故事情節進行探討，趙薇也用此概念研究李劼的《大波》三部曲中的人物關係，以及故事的情節變化[19]。

網絡分析結合多來源文本也能呈現人物的生平交往或遷移過程，以及整個大環境時空脈絡。So 和 Hoyt 曾分析跨太平洋的文學社團與文學活動，從中整理出 1920 年代現代主義詩歌在全球傳播的路徑[20]；劉吉軒等人分析台灣海外左派刊物，用社會網絡圖呈現出海外台灣人民部分的政治思想輪廓[21]；金觀濤等人等人(2016)研究《新青年》雜誌中群眾觀念的變化，分析重要關鍵詞變化，並利用概

[15] Barnes, J. A. “Class and committees in a Norwegian Island parish”, Human Relations, pp. 39–58.
[16] Wasserman, S., & Faust, K. *Social Network Analysis: Methods and Applications*.
[17] Hanneman, R. A., & Riddle, M. *Introduction to Social Network Methods*.
[18] Moretti, F. Network theory, plot analysis.” New Left Review, pp 68.
[19] 趙薇。(「社會網路分析」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)，收於項潔編，《數位人文──在過去、現在和未來之間》，頁 398- 426。
[20] So, R. J., & Long, H. “Network Analysis and the Sociology of Modernism,” Boundary 2, 40 (2), pp 147–182.
[21] 劉吉軒、柯雲娥、張惠真、譚修雯、黃瑞期、甯格致。(以文本分析呈現海外史料政治思想輪廓)，收於項潔編，《數位人文要義: 尋找類型與軌跡》，頁 83-114。

念網絡圖呈現《新青年》雜誌從自由主義轉向馬列主義的過程[22]。史丹福大學的「Mapping the Republic of Letters」網站呈現啟蒙時代知識份子透過書信交流所形成的社會網絡[23]。而「中國歷代人物傳記資料庫」提供開放版本的資料庫供研究者下載，研究者能以群體傳記的角度，將資料匯入社會網絡分析軟體進行分析，以了解各人物間的社會關係；Yeh 也同樣利用群體傳記的方式，運用社會網絡分析國民黨刊物-《婦女共鳴》與商業刊物《婦女雜誌》作者群體的差異，解釋國民黨內部的分化與合作情形[24]。

從以上之文獻探討顯示社會網絡分析對於數位人文研究具有重要助益，本研究因此發展支援數位人文研究之「古籍數位人文研究平台之史料人物關係圖」，可以人機互動合作方式萃取文本中的人物社會網絡關係圖，除了提供人文學者即時互動性的操作介面外，也會提供文本中的人物關係，協助人文學者進行文本中的人物社會關係及情境探討，本研究也透過實際之實驗，驗證此一系統對於支援數位人文研究的效益。

2.3. 科技接受模式

科技接受模式（Technology Acceptance Model, TAM）是 Davis Jr 依據理性行為理論(Theory of Reasoned of Reasoned Action, TRA)，以及使用者的情感因素與資訊科技的使用提出，希望藉由使用者的態度(attitude)、行為意圖(behavior intention)等因素，預測使用者的資訊科技使用行為[25]。在理性行為中，主張使用者會依據其行為意圖來決定使用資訊科技的行為，其中使用者的態度和主觀規範(subject norm)則是影響其行為意圖的關鍵因素。Davis 指出在資訊系統使用中，使用者的態度會比主觀規範對使用者的行為意圖更具有影響力，因此科技受模型主要著重於使用者態度對於使用者行為意圖的影響[26]。

在科技接受模型中，使用者的使用態度會受到其「認知有用性(perceived usefulness)」與「認知易用性(perceived ease of use)」的影響。認知有用性是指使用者主觀認為使用特定科技產品會提升其工作績效，當科技產品越容易使用，使用者將能完成越多工作，其工作績效就會越高；認知易用性是指主觀認為特定科

[22] 金觀濤、邱偉雲、梁穎誼、陳柏聿、沈錳坤、劉青峰。(觀念群變化的數位人文研究——以《新青年》為例)，收於項潔編，《數位人文──在過去、現在和未來之間》，頁 427- 463。
[23] Mapping the Republic of Letters. Retrieved September 14, 2018, from http://republicofletters.stanford.edu/
[24] Yeh, W. C. "Journal, Gender and Mobilization: Social Network Analysis of "Women's Resonance" (1929-1944)," Presented at the The 8th International Conference of Digital Archives and Digital Humanities, Taipei.
[25] Davis, F. D. *A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results. Sloan School of Management*. Massachusetts Institute of Technology.
[26] 同 25 註

技產品操作的難易度，兩者合稱為使用者的內部信念(internal beliefs)[27]，也是使用者個人接受資訊科技的主要決定因素。此外，認知有用性和認知易用性通常會受到系統特性、個人特質等一些外部因素影響。結合上述，整個接受模型大致可分為外部變數、內部信念、態度、行為意圖及實際使用科技。

科技接受模型用於數位人文領域方面，多半以評估資料庫及輔助人文研究的數位工具為主，Kemel 利用科技接受模型及半結構訪談方式探討歷史學者使用歷史資料庫的使用態度，並歸納出 15 項影響歷史學者對於使用歷史資料庫的認知有用性和認知易用性因素 [28]；Hong 使用科技接受模型評估台灣數位典藏系統，結果得出系統介面的呈現，是影響使用者使用數位典藏系統的最重要因素 [29]；Chen 等人(2018)使用科技接受模型評估人文學者使用該研究所開發之「支援數位人文研究之文本自動標註系統」與 MARKUS 文本半自動標註系統差異，結果發現在認知易用性方面，自動標註系統顯著優於 MARKUS 文本半自動標註系統，代表相較於 MARKUS 文本半自動標註系統，自動標註系統系統功能更容易使用 [30]。本研究希望藉由科技接受模型，從使用者的認知有用性與認知易用性兩大面向，了解人文學者使用「史料人物關係圖工具」支援數位人文研究的態度及意願。

3. 系統設計

3.1. 系統架構

本研究於古籍數位人文研究平台上發展的「史料人物關係圖工具」架構如圖 1 所示，整體架構分為二個部分，包括處理文本斷詞，依據人名權威檔從中抽取人名，建立文本中人名社會關係矩陣的人名關係分析模組；將人名社會關係視覺化建立社會網絡分析圖，以及讓使用者可參考史料人物關係圖和史料文本進行人物關係編輯的視覺化互動介面。

[27] Davis, F., Bagozzi, R., & Warshaw, P. “User Acceptance of Computer Technology: A Comparison of Two Theoretical Models”, *Management Science*, *35*(8), 982–1003.

[28] Kemell, K.-K. *Technology acceptance of digital historical record database systems among historians* University of Jyväskylä, Jyväskylä. Retrieved from https://jyx.jyu.fi/handle/123456789/52037

[29] Hong, J. C., Hwang, M. Y., Hsu, H. F., Wong, W. T., & Chen, M. Y. “Applying the technology acceptance model in a study of the factors affecting usage of the Taiwan digital archives system,” *Computers & Education*, *57*(3), pp. 2086–2094.

[30] Chen, C. M., Chen, Y. T., & Liu, C. Y. “Development and evaluation of an automatic text annotation system for supporting digital humanities research,” *Library Hi Tech*. (in press)

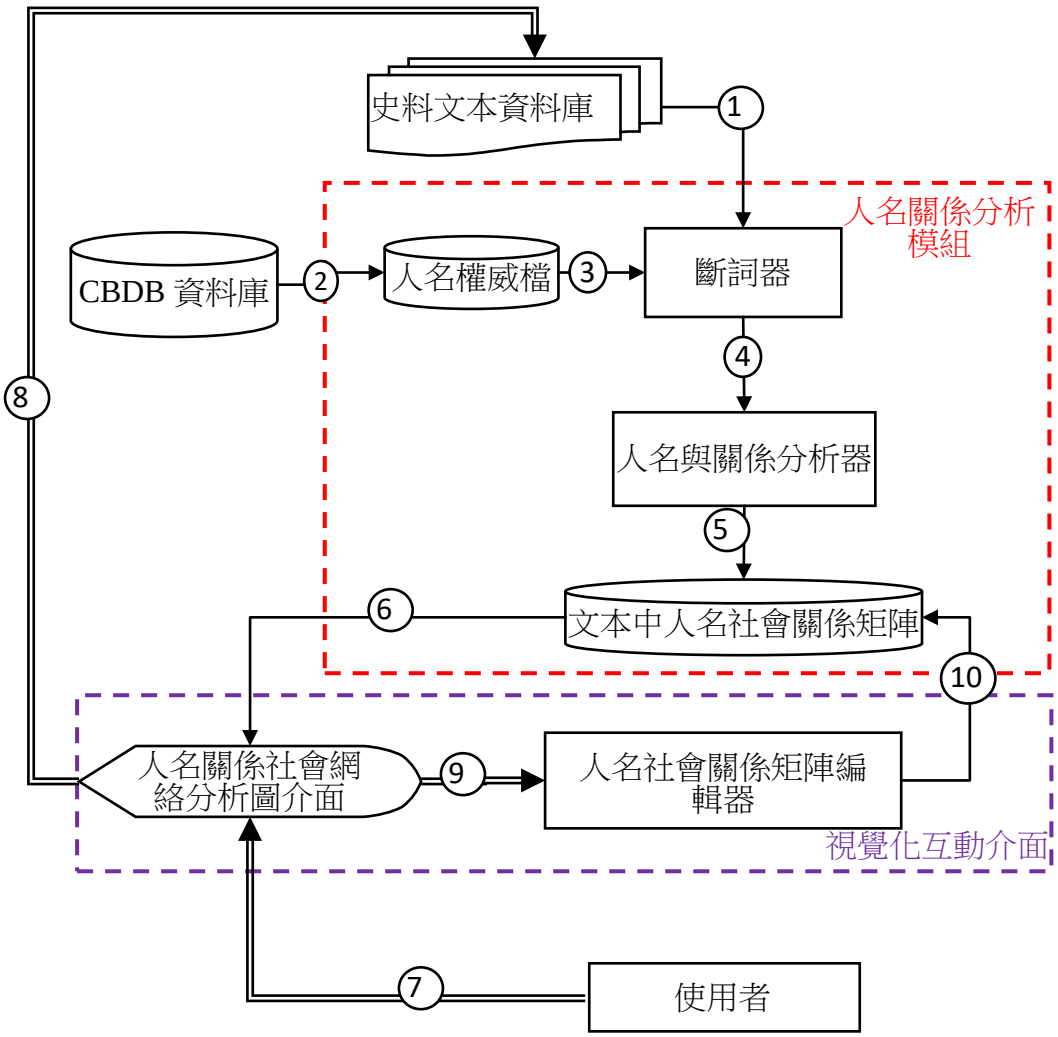


圖 1 史料人物關係圖工具系統架構圖

系統運作流程說明如下:

(1) 中文並無明顯的分隔符號或空格區分詞與詞之間的分界，因此須借助斷詞器的輔助將文本中內文進行斷詞處理，識別出文本中的詞彙，以利後續人名與社會關係的分析。

(2) 本系統以中國歷代人物傳記資料庫（CBDB）中所收錄之人名詞彙作為識別文本中人名詞彙的依據，從中國歷代人物傳記資料庫中抓取資料庫內的人名資訊，如人名於 CBDB 中的 ID、本名、別名、官名及諡號等，並據此建立人名權威檔。

(3) 斷詞器進行文本斷詞時，斷詞器詞庫中的詞彙及詞性會作為識別文本中詞彙的主要依據；本系統將上一階段所建立的人名權威檔加入斷詞器詞庫中，並給予其專屬的詞性代號，以提升史料文本中人名辨識的準確性和權威性。

(4) 經過斷詞器處理的史料文本，可將文本中識別出的人名輸出至人名與關係分析器上，進行人名與社會關係分析判斷。本系統目前將社會關係分為兩大類，

分別為著述關係（Writings）、不明關係（unknown），其判斷方法於下一小節說明。

(5) 將人名與關係分析器中所建立的人名社會關係轉換成人物社會關係矩陣，並輸出至資料庫內儲存。

(6) 依照上階段所建立之人名社會關係矩陣繪製成人物關係分析圖，以視覺化的方式呈現史料文本內人名之間的社會網絡關係，提供人文學者參考。

(7) 使用者可選擇要瀏覽文集，進入史料人物關係圖系統，解讀文本中的人物與人物關係。

(8) 使用者可同時瀏覽史料人物關係圖和史料文本資料庫中的閱讀介面，方便正確依據文本解讀社會網絡關係。

(9) 人文學者可使用人名社會關係矩陣編輯器新增、修改人名社會關係矩陣的資訊，以確認正確之人名社會網絡關係。

(10) 人名社會關係矩陣編輯器編輯過的資訊，會更新人名社會關係矩陣內容，同時視覺化介面也會重新產生新的人物關係分析圖。

3.2. 系統元件

本研究於古籍數位人文研究平台發展之「史料人物關係圖工具」的各個元件，說明如下：

3.2.1. Jieba 中文斷詞器

本研究使用 Jieba 斷詞器進行中文斷詞器之實作，Jieba 係採用開放原始碼形式釋出的中文斷詞器，使用者可根據自身的需求自定義所需的詞彙，增加詞庫中的詞彙，而 Jieba 採用的斷詞原理是基於前綴樹結構生成句子中所有可能成詞的詞彙後，使用隱馬可夫模型來判斷字詞組合的機率，並使用動態規劃演算法找出最大機率的詞彙組合路徑，此路徑就是基於詞庫的最佳斷詞結果。本研究在 Jieba 中文斷詞器詞庫中加入中國歷代人物傳記資料庫(CBDB)的人名權威表，經斷詞處理後，能識別出史料文本資料庫中文本的人名。

3.2.2. 中國歷代人物傳記資料庫(CBDB)

中國歷代人物傳記資料庫(CBDB)為中央研究院、哈佛大學、北京大學合作開發，主要收錄中國歷史上所有重要的人物傳記資料，此資料庫的內容免費公開作為學術用途使用，可作為人物傳記參考、統計分析與空間分析之用。目前主要收錄中國七世紀至十九世紀間人物傳記資訊，約 370,000 人，本系統的人名權威表乃採用中國歷代人物傳記資料庫中的人物資訊建立而成。

3.2.3. 人物社會關係分析器

本元件對經過斷詞處理後的史料文本進行人名社會網絡關係的分析，此階段系統以自動化方式建立四種判斷人名社會網絡關係的規則，說明如下：

3.2.3.1. 史料文本作者與文中所辨識的人物關係

本系統會抓取史料文本資料庫文集資訊中作者欄位，以及史料文集中斷詞器所識別的人名，並據此建立著述關係。

3.2.3.2. 為作者撰寫序跋之人與史料文本作者關係

本系統會抓取史料文本中撰寫序跋之人與史料文本作者，並建立彼此之間為著述關係。

3.2.3.3. 為作者撰寫序跋之人與序跋段落中所辨識的人名社會網絡關係

本系統會抓取史料文本中撰寫序跋之人與史料文本中之序跋段落中斷詞器所識別的人名，並建立彼此之間為著述關係。

3.2.3.4. 史料文本中人物與人物之間社會網絡關係

本研究針對史料文本中一定字數之內同時出現的兩個人名先建立不明關係，以待人文學者作進一步確認，本系統目前將判斷的字數設定為 10 個字之內同時出現的人名先建立不明關係。

3.2.4. 人物關係網絡視覺化介面

此一介面主要將文本中人物關係網絡以視覺化方式呈現，人文學者可在此介面進行文本人物關係網絡圖的觀察，例如選擇要觀看的史料人物關係圖中的人物與人物關係，同時也可和史料文本資料庫中的文本內容互相參照。

3.2.5. 史料文本閱讀介面

此一介面可顯示人文學者所選擇的文集內文，當使用者於史料人物網絡關係圖中點選人名時，可於史料文本文本閱讀介面中顯示該人物於文本中出現的位置。

3.2.6. 人物社會關係編輯器

此一編輯器主要作為編輯人物社會關係之用，人文學者可根據自身的知識或閱讀文本後，針對有疑慮的人物關係進行修正，修正過後之內容，會同步更新顯示於人名關係網絡視覺化介面。

3.3. 系統介面與功能

本研究於古籍數位人文研究平台發展之「史料人物關係圖工具」提供友善之使用者介面與功能，讓使用者在解讀文本的過程中，可搭配本系統提供之史料人物關係圖、閱讀介面、內文與外部搜尋組及記事本四個模組，進行人物與人物關係的解讀，整體系統介面如圖 2 所示，各功能介面說明如下：



圖 2 整體系統介面

3.3.1. 史料人物關係圖介面

使用者可於該介面觀看文本段落中的人物關係，如圖 3 所示。圖中橘色與藍色的節點為此卷文集中系統所判斷出的所有人物，橘色的人物節點為目前閱讀介面所在的文本段落中出現的人物，紫色的人物節點為文本其他段落中出現的人物。如圖 4 之史料人物關係圖中節點分別顯示出人名、關係屬性、連結人名的對應，系統預先判斷出的人物關係為著述或未確認關係。人物關係編輯方式如所圖 5 所示，點擊史料人物關係圖中的人物關係線條，即可以開啟人物關係編輯器，可依自身所判斷關係的結果，編輯兩人物間的人物關係，編輯後的關係，會立即更新於史料人物關係圖介面。

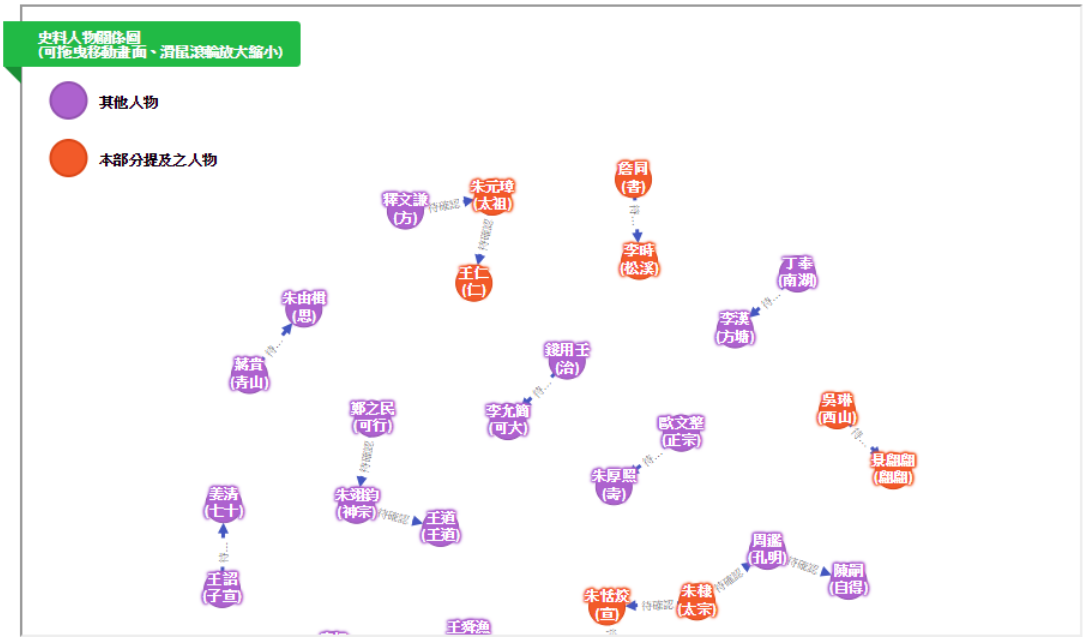


圖 3 史料人物關係圖介面



圖 4 人物關係網絡

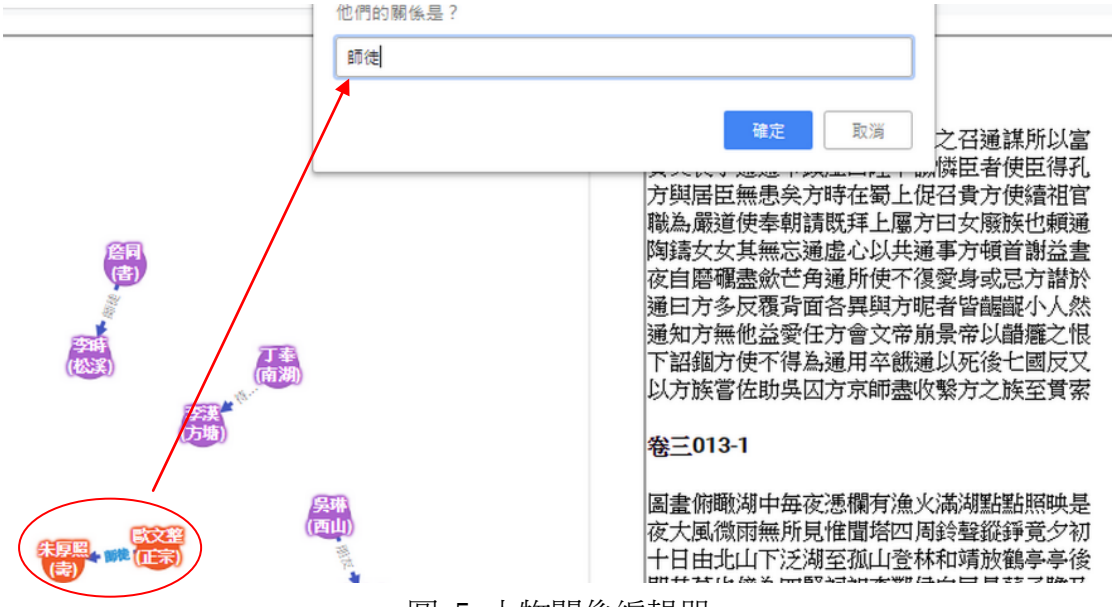


圖 5 人物關係編輯器

3.3.2. 閱讀介面

閱讀介面如圖 6 所示，閱讀介面與史料人物關係圖及外部查詢功能整合在一起，史料人物關係圖會根據閱讀介面中的文本段落不同，進行相對應的變化，而當使用者點擊史料人物關係圖中的本段落出現人物的節點時，閱讀介面上會標亮將該人物於文本中的位置，方便使用者分析人物與人物關係。若使用鼠標將要查詢的文字或段落標註，即可與外部搜尋的功能連結，可利用外部搜尋功能進行查詢，詳細的運作機制於內文搜尋與外部搜尋中進行說明。



圖 6 史料人物關係圖閱讀介面模組

3.3.3. 內文搜尋與外部搜尋介面

內文搜尋主要可讓使用者在文本中搜尋想要查詢的人物或字彙，並於文本段落中標亮顯示。外部搜尋的功能則能讓使用者利用外部資料庫查詢所需的資訊，以利其解讀文本中的人物與人物關係，提供人物資料庫、線上百科、搜尋引擎、字典等各式外部連結，包含 CBDB、萌典、維基百科、百度百科、Google、漢典、異體字字典及康熙字典，如圖 7 所示。使用者只要劃記文本中所要搜尋的段落文字，該段文字即會自動填入外部搜尋框，即可進行搜尋。



圖 7 外部搜尋連結

3.3.4. 筆記欄介面

筆記欄介面模組能讓使用在解讀文本時，記錄所發現的資訊，如圖 8 所示。將筆記欄模組擺放在閱讀介面旁的用意，是希望使用者能更專注在解讀文本中的人物與人物關係上，減少於視窗間來回切換的時間，能在單一介面下完成其解讀人物與人物關係的任務。



圖 8 筆記欄介面

4. 研究方法

4.1. 研究設計

為了驗證本研究於古籍數位人文研究平台發展「之史料人物關係圖工具」是否有助於人文學者進行數位人文研究，本研究採用準實驗研究法比較古籍數位人文研究平台上有無使用「之史料人物關係圖工具」輔助解讀文本人物與人物關係，在成效上是否具有顯著的差異。此外，本研究於實驗對象完成實驗後，邀請實驗對象填寫科技接受度量表，以瞭解使用者對於本研究發展之「史料人物關係圖工具」，以及僅以人工閱讀文本解讀文本人物與人物關係的科技接受度，希望藉此了解實驗對象使用此系統支援數位人文研究，在認知有用性及認知易用性上的感受，是否具有顯著的差異。最後，本研究於實驗對象完成科技接受度量表量測後，進一步以半結構深度訪談的方式，來補充量化分析上的不足。訪談內容如：「使用系統中的各項功能時，有無遇到困難或疑問」、「系統整體版面設計是否清晰易懂」、「系統人機互動設計是否直覺易用」、「針對未來系統持續發展有何看法或建議」。

4.2. 研究對象

本研究之研究對象為具備解讀古文的中文、歷史相關科系學生，考量成本、時間、地點限制等因素，本研究以方便取樣且有意願配合實驗的臺北市某國立大學中國文學、歷史學相關系所或具備古文解讀能力之大學生和研究生 21 人為研究對象。

4.3. 研究工具

4.3.1. 古籍數位人文研究平台之史料人物關係圖工具

本研究於古籍數位人文研究平台所開發之「史料人物關係圖」，可從文本中識別人名，並初步建立人物關係，並提供社會關係編輯器輔助人文學者修改及確認文本中的人物關係，本研究評估使用此一工具，對於輔助人文學者透過文本解讀人物與人物關係之成效。

4.3.2. 科技接受度量表

本研究參考 Hwang、Yang 與 Wang (2013)所編製之科技接受量表，並修改語句以符合本研究之需求，量表樣式採用李克特六點量表，量表包含兩大構面，分別為「系統認知有用性」，共 6 題；「系統認知易用性」，共 7 題，總計 13 題。在量表的信度上，「系統認知有用性」構面之 Cronbach's α 值為 0.95；「系統認知易用性」構面之 Cronbach's α 值為 0.94，皆具有良好的信度。此一量表於實驗結束後，邀請研究對象填寫，以了解研究對象對於支援數位人文研究之「古籍數位人文研究平台之史料人物關係圖」的科技接受度感受。

4.3.3. Google 分析(Google Analytics)

Google 分析(Google Analytics)為 Google 公司所發展的網站分析平台，可於要分析的網頁內部嵌入特定的程式碼，記錄該網站的使用歷程、流量分析等資訊(Google Analytics, 2017)。本研究採用 Google 分析紀錄研究對象使用「古籍數位人文研究平台之史料人物關係圖」解讀人物與人物關係的歷程紀錄，並將該記錄匯出於資料分析階段進行滯後序列分析(lag sequence analysis)。

4.4. 實驗流程

本研究透過程式計算的方式評估明代文集各卷文集中，所包含的人物和人物關係數量，並以人物關係為未確認的關係為主，文本段落的長度也盡可能控制一致，依計算結果選擇明代文集中的《校刻具茨先生詩集》作為本研究之實驗的文本，文本中的每一段落系統所計算的人物、人物關係及段落長度如表 1 所示。本研究將實驗分成兩階段，每一階段分別需要解讀兩個段落文本中的人物與人物關係。為了避免實驗結果受到使用系統的先後順序及閱讀文本的先後順序影響，本研究將 21 位受測者分為甲乙兩組，於實驗活動階段一、階段二交錯使用兩個系統與兩段文本。

本研究實驗流程如圖 9 所示，實驗開始先告知實驗對象本次實驗的流程，告知其所屬組別，並開始進行第一階段實驗，甲組的實驗對象會先進行有史料人物關係圖系統操作說明，接著在 40 分鐘內使用有史料人物關係圖的系統輔助解讀文本中的人物與人物關係;乙組的實驗對象則先進行無史料人物關係圖系統的操作說明，接著在 40 分鐘內使用無史料人物關係圖的系統輔助解讀文本中的人物與人物關係，並於時間結束後進行第一篇文本內容的人物與人物關係評估，並填寫科技接受度量表。第二階段實驗開始前，甲組的實驗對象會先進行無史料人物關係圖系統操作說明，接著在 40 分鐘內使用無史料人物關係圖的系統輔助解讀文本中的人物與人物關係，乙組的實驗對象會先進行有史料人物關係圖系統的操作說明，接著在 40 分鐘內使用有史料人物關係圖的系統輔助解讀文本中的人物與人物關係，並於時間結束後進行第二篇文本內容的人物與人物關係評估，最後進行半結構訪談，實驗時間共約 140 分鐘。

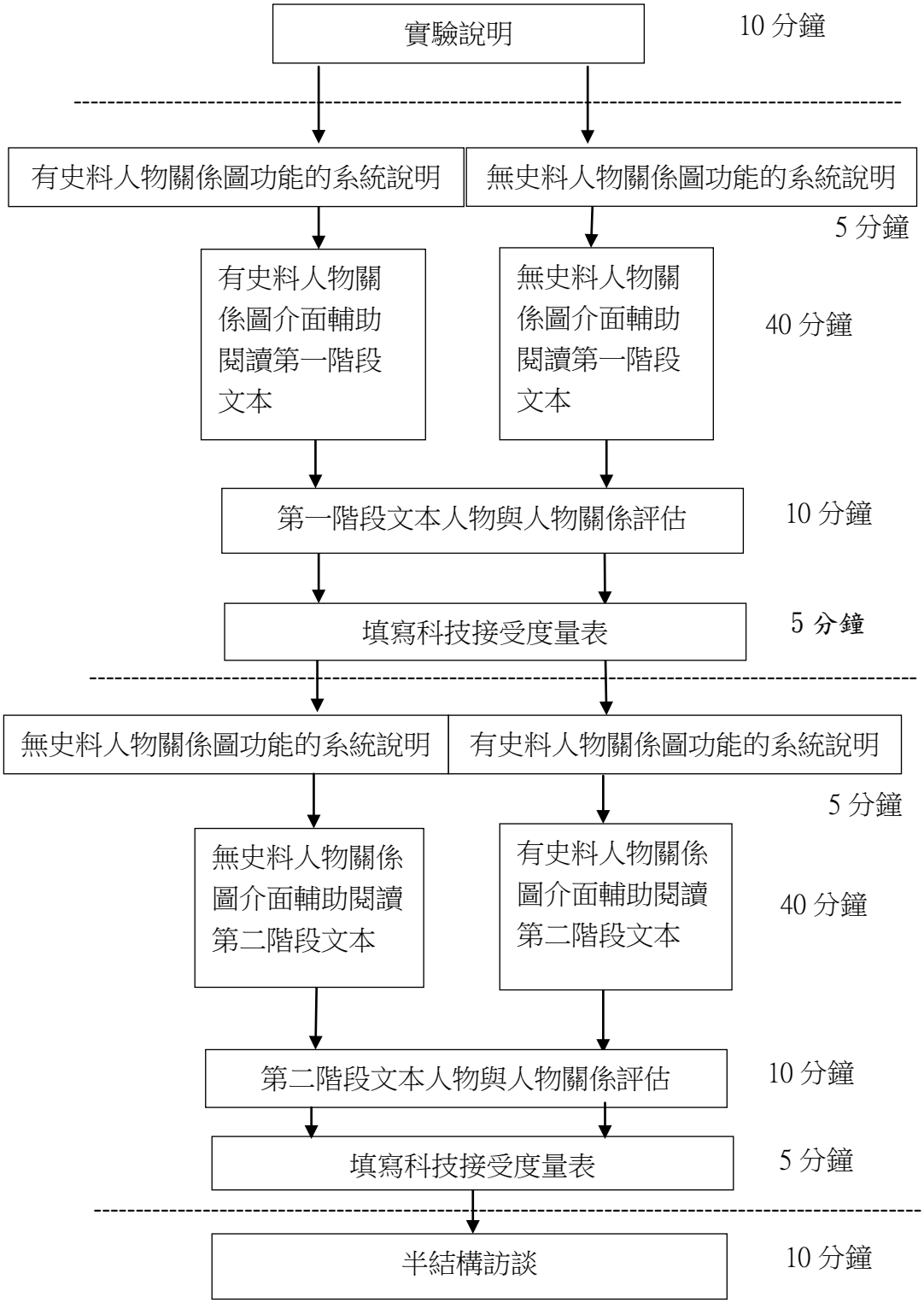

實驗說明
10 分鐘
有史料人物關係圖功能的系統說明
無史料人物關係圖功能的系統說明
5 分鐘
有史料人物關係圖介面輔助閱讀第一階段文本
無史料人物關係圖介面輔助閱讀第一階段文本
40 分鐘
第一階段文本人物與人物關係評估
10 分鐘
填寫科技接受度量表
5 分鐘
無史料人物關係圖功能的系統說明
有史料人物關係圖功能的系統說明
5 分鐘
無史料人物關係圖介面輔助閱讀第二階段文本
有史料人物關係圖介面輔助閱讀第二階段文本
40 分鐘
第二階段文本人物與人物關係評估
10 分鐘
填寫科技接受度量表
5 分鐘
半結構訪談
10 分鐘

圖 9 實驗流程圖

5. 實驗結果與分析

5.1. 有無史料人物關係圖系統解讀人物數量之成效比較分析

表 1 為受測者分別使用兩系統所解讀出的人物數量之獨立樣本 t 檢定分析結果，結果顯示兩系統在解讀出的人物數量上並未呈現顯著差異(t=1.489,p=0.152>0.05)，但在平均數方面，有史料人物關係圖系統解讀出的人物平均數量為 12.43，高於無史料人物關係圖系統的 9.19。

表 1 兩系統解讀人物數量之獨立樣本 t 檢定

系統	個數	平均數	標準差	t 值	顯著性(雙尾)
有史料人物關係圖系統	21	12.43	10.948	1.489	.152
無史料人物關係圖系統	21	9.19	7.534		

5.2. 有無史料人物關係圖系統解讀人物關係數量成效比較分析

表 2 為兩系統解讀人物關係數量之獨立樣本 t 檢定結果，結果顯示為兩系統在解讀出的人物關係數量上並未呈現顯著差異(t=0.249,p=0.806>0.05)，但在平均數方面，有史料人物關係圖系統所解讀人物關係數量之平均數為 3.14，高於無史料人物關係圖系統的 3.00。

表 2 兩系統解讀人物關係數量之獨立樣本 t 檢定

系統	個數	平均數	標準差	t 值	顯著性(雙尾)
有史料人物關係圖系統	21	3.14	2.651	.249	.806
無史料人物關係圖系統	21	3.00	1.732		

5.3. 有無史料人物關係圖系統科技接受度之差異分析

表 3 為兩個系統科技接受度之平均分數獨立樣本 t 檢定結果，結果顯示有史料人物關係圖系統之科技接受度顯著優於無史料人物關係圖系統(t=3.270,p=0.004<0.05)，此一結果顯示受測者對於有史料人物關係圖系統輔以解讀文本中人物或人物關係，抱持更正向的接受態度。表 4 為對兩系統科技接受度的認知有用性面向獨立樣本 t 檢定結果，結果顯示所示兩系統在科技接受度認知有用性面向達顯著差異(t=3.078,p=0.006<0.05)，有史料人物關係圖系統顯著高於無史料人物關係圖系統。表 5 為對兩系統的科技接受度的認知易用性面向獨立

樣本 t 檢定結果，結果顯示兩系統在科技接受度的認知易用性面向達顯著差異(t=2.234,p=0.037<0.05)，有史料人物關係圖系統組別顯著高於無史料人物關係圖系統組別。

表 3 兩系統科技接受度之平均分數獨立樣本 t 檢定

系統	個數	平均數	標準差	t 值	顯著性(雙尾)
有史料人物關係圖系統	21	4.67	.746	3.270	.004
無史料人物關係圖系統	21	3.99	.791		

表 4 兩系統於科技接受度之認知有用性獨立樣本 t 檢定

系統	個數	平均數	標準差	t 值	顯著性(雙尾)
有史料人物關係圖系統	21	4.03	.984	3.078	.006
無史料人物關係圖系統	21	3.04	.941		

表 5 兩系統於科技接受度之認知有用性獨立樣本 t 檢定

系統	個數	平均數	標準差	t 值	顯著性(雙尾)
有史料人物關係圖系統	21	5.20	4.383	2.234	.037
無史料人物關係圖系統	21	4.80	6.611		

5.4. 史料人物關係圖系統歷程行為資料統計

本次參與實驗的 21 位受測者的行為統計資料如所示，由表 6 中可發現使用最多的前三名功能依序為內文搜尋功能(search_raw_text)、點擊史料人物關係圖人物結點(graph_node_search)、編輯筆記(edit_note)，而外部查詢連結前三名為 Google(search_google)、維基百科(search_wiki)、CBDB (search_cbdb)。

表 6 史料人物關係圖系統歷程行為統計

事件編碼	出現次數	出現百分比
內文搜尋(search_raw_text)	917	35.7%
點擊史料人物關係圖人物結點(graph_node_search)	888	34.5%
編輯筆記(edit_note)	266	10.3%
搜尋 Google(search_google)	194	7.5%
切換文本段落(select_volume)	115	4.5%
搜尋維基百科(search_wiki)	60	2.3%
編輯人物關係(edit_relationship)	36	1.4%
搜尋 CBDB(search_cbdb)	35	1.4%
搜尋萌典(search_moedict)	23	0.9%
搜尋百度百科(search_baidu)	14	0.5%
搜尋漢典(search_zdic)	14	0.5%
搜尋異體字字典(search_chardb)	6	0.2%
搜尋康熙字典(search_kangxi)	4	0.2%
總數	2572	100.0%

5.5. 史料人物關係圖系統使用行為序列分析

本研究將史料人物關係圖系統中的各種操作行為，依照時間順序進行編碼，得到受測者的行為序列觀察觀察樣本，並利用滯後序列分析(lag sequence analysis)得出具有顯著行為轉移之行為轉換圖，如圖 10 所示。

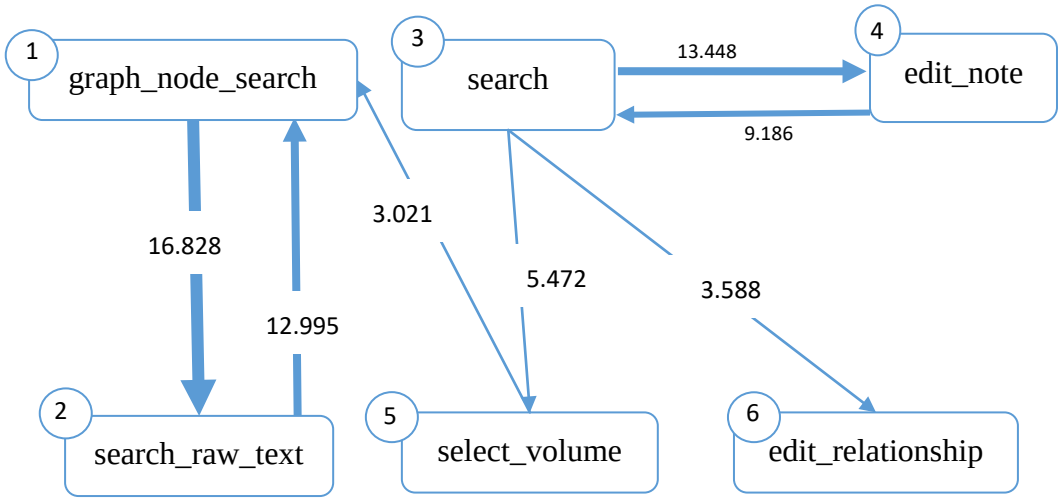


圖 10 史料人物關係圖系統使用行為轉換圖

圖中的編號 1~6 表示史料人物關係圖系統中的各種系統操作行為，箭頭表示行為與行為之間的轉移，行為轉移越明顯，則所表現的線條就越粗，箭頭線上的數值為 z 分數，若大於 1.96 則表示行為與行為之間的轉移達顯著水準，以下針對各行為之序列轉移結果進行說明:

5.5.1. 點擊史料人物關係圖中人物節點(graph_node_search)

內文搜尋(search_raw_text)的行為轉移至點擊史料人物關係圖中人物節點(graph_node_search)達顯著水準(z=12.995 > 1.96)，代表許多受測者在使用內文搜尋後，會接著使用史料人物關係圖。此外，選擇文本段落(select_volume)的行為轉移至點擊史料人物關係圖中人物達顯著水準(z=3.021>1.96)，代表許多受測者在選擇文本段落後，會接著點擊史料人物關係圖中的人物節點。

5.5.2. 內文搜尋(search_raw_text)

點擊史料人物關係圖使用行為轉移至內文搜尋(search_raw_text)達顯著水準(z=16.828>1.96)，代表許多受測者在點及史料人物關係圖中的人物節點後，會使用全文搜尋功能找尋人物在段落中的位置。和上述第一點一起結合，可以發現大部分受測者在內文搜尋及點擊史料人物關係圖的人物節點這兩者行為間會來回轉換。

5.5.3. 外部搜尋(search)

筆記功能(edit_note)使用行為轉移至外部搜尋(search)達顯著水準(z=9.186>1.96)，代表許多受測者在記錄完筆記後會接著利用外部搜尋解讀文本或有疑問的地方。

5.5.4. 筆記功能(edit_note)

外部搜尋(search)行為轉移至使用筆記功能(edit_note)達顯著水準(z=13.448>1.96)，代表許多受測者在使用外部搜尋的功能之後會使用筆記功能，紀錄所得到的結論或答案。和上述第三點一起分析，可以發現大部分受測者在進行人物關係分析時，會交互使用外部搜尋與筆記功能記錄有用的資訊。

5.5.5. 選擇文本段落(select_volume)

外部搜尋(search)行為轉移至使用選擇文本段落(select_volume)達顯著水準(z=5.472>1.96)，代表許多受測者在使用完外部搜尋後，會切換文本的段落。

5.5.6. 編輯史料人物關係圖中的人物關係(edit_relationship)

外部搜尋(search)行為轉移至編輯史料人物關係圖中的人物關係(edit_relationship)達顯著水準(z=5.472>1.96)，代表許多受測者在使用外部搜尋後會編輯史料人物關係圖中的人物關係。

5.6. 訪談分析

5.6.1. 整體系統介面簡潔明瞭，區塊標示清楚，操作過程直覺、人性化

多數受訪者認為本研究發展之輔以史料人物關係解讀之社會網絡分析工具的整體系統介面簡潔明瞭，功能區塊之間標示清楚，各功能之間又能互相參照，不用離開所在頁面，能在同一頁面下同時進行文本閱讀、人物關係的解讀、確認或進行外部查詢及進行筆記。

5.6.2. 人物史料關係圖功能所提供的資訊有助於使用者解讀文本中的人物及關係

史料人物關係圖功能能呈現文本內的可能人物與彼此關係，有助於使用者在解讀文本前大致瀏覽史料人物關係，以了解文本可能提及的年代、事件等，讓使用者在解讀文本中過程中不易遺漏人物及彼此關係。

5.6.3. 跨朝代人名歧義的問題

許多受訪者表示史料人物關係圖中有些別的朝代的人物會被辨識成明代的人物，例如: 某段文本中的太祖指的不是明太祖，而是宋太祖，這會嚴重影響解讀文本中人物與人物關係的效益。

5.6.4. 文本中的人名代稱及別稱問題

有些受訪者表示文本中有些人叫熊君、張君等，這些人物關係在目前發展的史料人物關係圖中是無法呈現的，但這些人可能在人物關係上扮演重要的角色。

5.6.5. 解讀人物關係時能透過外部查詢功能更加了解人物資訊及人物關係

許多受訪者表示系統能透過外部查詢功能至包括維基百科、CBDB 等多種外部資料庫查詢人物，對理解文本中的人物或人物關係有著極大的幫助。

5.6.6. 內文搜尋功能有助於讓使用者快速找到所需的人物或詞彙

系統所提供的內文搜尋功能，可讓使用者點擊史料人物關係圖中的人物，即能快速標示出該人物於文本中所在的位置，也能讓使用者在文本中找到所需的詞彙，有助於人物關係解讀。

5.6.7. 系統提供的筆記功能，有助於提升解讀文本的速度

本研究所發展系統將筆記功能區塊和閱讀介面呈現於同一介面，讓使用者能一邊閱讀文本一邊進行筆記，節省了使用者在文本和其他編輯器不斷來回切換的步驟，因此更能專注於解讀文本人物關係上。

6. 結論與未來研究

6.1. 結論

本研究透過獨立樣本 t 檢定，分析有無史料人物關係圖系統在輔以解讀文本中人物數量與人物關係數量的成效上，雖然在敘述統計上，有史料人物關係圖系統輔以人物關係解讀之成效高於無史料人物關係圖系統，但未達統計上的顯著差異。本研究推測係因為史料人物關係圖所提供的人物與人物關係資訊還未完善，

致使兩者差異未達顯著差異。此外，本研究透過獨立樣本 t 檢定分析受測者使用有無史料人物關係圖系統輔以文本人物關係解讀，在科技接受度上是否具有顯著差異。結果顯示有無史料人物關係圖在科技接受度上具有顯著差異，有史料人物關係圖系統的科技接受度高於無史料人物關係圖系統。在科技接受度「認知有用性」和「認知易用性」兩個構面上，有史料人物關係圖系統在這兩個構面上皆高於無史料人物關係圖系統，並且兩個構面皆達到統計上的顯著差異。此結果與許多受測者在訪談中表示，史料人物關係圖對解讀文本中的人物與人物關係具有幫助，系統介面操作起來相當有善、易上手，各功能模組之間標示明確的結果一致。

本研究發展之史料人物關係圖所提供的人物資訊及關係，係以 CBDB 中明代人物為主，並以詞庫比對的方法來識別人名，某些受測者表示部分的人物名稱會被錯誤解讀，因而產生干擾。但有些受測者表示文本中有些人物是明代以前的人物，但系統並沒有顯示出來，然而這些人物可能在文本中是關鍵人物或有著重要的人物關係。此外，跨朝代同名的問題亦可能產生錯誤，例如訪談中一位受訪者所提到文本中的太祖其實指是宋太祖，而系統將宋太祖識別成明太祖。另外，CBDB 所收錄人物明代人物以進士或官場人士居多，而明代文集所收錄的文集廣泛、多元，所以明代文集內文中有些人物在 CBDB 上並未被記錄，這點在訪談及使用者行為歷程記錄上得到證。許多受訪者表示 CBDB 查不到的人物資訊，他們會以查詢 Google 或維基百科輔助，在行為歷程記錄器中的外部搜尋連結次數最多的前兩名即為 Google 和維基百科，CBDB 則排在第三位。上述問題可能是導致史料人物關係圖系統在解讀人物和人物關係成效上，並沒有顯著優於採用無史料人物關係圖系統的可能原因，這也凸顯出史料人物關係圖系統所提供的人物與人物關係資訊必須更加精準、廣泛，以符合文本的特性、類型以及人文學者的需求。

本研究藉由使用滯後序列分析受測者於史料人物關係圖系統的行為歷程，結果發現點擊史料人物關係圖的人物節點與內文搜尋的行為之間，轉換達顯著差異，代表受測者點擊完史料人物關係圖後會接著使用內文搜尋查看該人物的所在的文本段落；而外部搜尋的行為轉換至記錄筆記行為及編輯史料人物關係圖中的人物關係行為皆達顯著差異，代表受測者使用完外部搜尋功能時，會利用筆記欄進行記錄所查到的資訊或編輯史料人物關係圖中的人物關係，相關分析結果也顯示外部搜尋次數與編輯史料人物關係圖中的人物關係次數之間達顯著正相關。受測者在訪談中也表示他們需要檢視文本內容及外部連結所查詢的資訊來輔助他們解讀人物與人物關係。

許多受訪者表示本研究發展之史料人物關係圖系統之使用者介面好上手，功能標示清楚，各功能之間也整合的非常好。此外，在科技接受度量表中的「認知

易用性」構面也得到相當高的分數，特別是點擊史料人物關係圖中的人物節點，即能透過閱讀介面的內文搜尋功能找到人物的所在位置功能，獲得相當高的評價。再則，微歷程行為記錄器之微歷程分析也顯示這部份的行為轉移具有非常高的顯著水準，是使用者最常使用的功能之一。

許多受訪者表示外部搜尋功能被設計在同一頁面，在使用上非常方便，不用離開所在的文本頁面進行查詢，而且能在筆記欄直接做筆記或編輯史料人物關係圖中的人物關係，微歷程行為記錄也顯示筆記功能和外部搜尋兩者之間的行為轉換皆達非常高的顯著水準，外部搜尋的行為轉換至編輯史料人物關係圖中的人物關係也達顯著水準。有些受訪者認為史料人物關係圖類似心智圖的概念，其呈現之視覺化史料人物關係圖，能讓人文學者快速掌握了解整個史料的人物或人物關係狀況。而微歷程行為分析也顯示切換文本後，點擊史料人物關係圖的行為轉移達顯著水準。

6.2. 未來研究方向

根據上述研究結論，本研究提出未來史料人物關係圖發展社會網絡分析工具，應提供包括中心度(centrality)、結構洞(structure hole)及最有影響力節點等社會網絡測度分析功能，並可任意搭配不同文集進行人物關係分析，以更有效協助使用者找出隱藏在文本中難以察覺的人物關係或脈絡。此外，也應發展明代文集之實體命名模型，這部份可使用機器學習方法，找出人名辨識的規則，建立明代文集的實體命名模型，並透過人機合作的方式，將有助於改善模型的正確率。文集和文集之間的脈絡關聯也值得探討，若能利用書目計量學分析文本之間引用或參照關係，或許也能從中得到文本內容中不易被發覺的人物關係。再則，建構以明代為主的語意網路(semantic network)，例如上位詞、廣義詞、相關詞等，並建立相關的知識本體(ontology)，將有助於使用者依據文本中語意、情境或找出隱藏在文本內容背後的意涵，進行更深一層的分析。

參考書目

(1) 專書

Berry, D. 2012. *Understanding digital humanities. London: Palgrave Macmillan.*

Davis, F. D. 1986. *A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results. Sloan School of Management*. Massachusetts Institute of Technology.

Fishbein, M., & Ajzen, I. 1975. *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research. Reading*. MA: Addison-Wesley.

Hanneman, R. A., & Riddle, M. 2005. *Introduction to Social Network Methods*. Riverside, CA: University of California, Riverside.

Kemell, K.-K. 2016. *Technology acceptance of digital historical record database systems among historians*. University of Jyväskylä, Jyväskylä. Retrieved from https://jyx.jyu.fi/handle/123456789/52037

Schreibman, S., Siemens, R., & Unsworth, J. 2008. *A Companion to Digital Humanities*. Wiley Publishing.

Wasserman, S., & Faust, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge ; New York: Cambridge University Press.

Guy, R. K. 2010. *Qing Governors and Their Provinces: The Evolution of Territorial Administration in China, 1644-1796*. University of Washington Press.

(2) 期刊論文

Adams, D. A., Nelson, R. R., & Todd, P. A. 1992. “Perceived usefulness, ease of use, and usage of information technology: A replication,” MIS Quarterly, 16(2), pp. 227–247.

Aires, V. P., Almeida, T. G., Nakamura, F. G., & Nakamura, E. F. 2017. “A Social Network Analysis of the The Lord of The Rings’ Trilogy,” In Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web, pp. 469–472. New York, NY, USA: ACM. https://doi.org/10.1145/3126858.3131567

Barnes, J. A. “Class and committees in a Norwegian Island parish”, *Human Relations*, pp. 39–58. http://dx.doi.org/10.1177/001872675400700102

Chen, C. M., Chen, Y. T., & Liu, C. Y. 2018. “Development and evaluation of an automatic text annotation system for supporting digital humanities research, ” *Library Hi Tech*. (in press)

Chen, S. P., Hsiang, J., Tu, H. C., & Wu, M. 2007. “On Building a Full-Text Digital Library of Historical Documents. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pp. 49–60. Springer Berlin Heidelberg.

Davis, F., Bagozzi, R., & Warshaw, P. 1989. “User Acceptance of Computer Technology: A Comparison of Two Theoretical Models,” *Management Science*, *35*(8), pp. 982–1003.

Duling, D. C. 2000. “The Jesus Movement and Social Network Analysis: (Part II. The Social Network),” *Biblical Theology Bulletin*, *30*(1), pp. 3–14. https://doi.org/10.1177/014610790003000102

Ha, S., & Stoel, L. 2009. ”Consumer e-shopping acceptance: Antecedents in a technology acceptance model,” *Journal of Business Research*, *62*(5), pp. 565–571. https://doi.org/10.1016/j.jbusres.2008.06.016

Hockey, S. 2004. “The History of Humanities Computing. In *A Companion to Digital Humanities*,” pp. 1–19. Wiley-Blackwell. https://doi.org/10.1002/9780470999875.ch1

Hong, J. C., Hwang, M. Y., Hsu, H. F., Wong, W. T., & Chen, M. Y. 2011. “Applying the technology acceptance model in a study of the factors affecting usage of the Taiwan digital archives system,” *Computers & Education, 57*(3), pp. 2086–2094.

Hu, P. J., Chau, P. Y. K., Sheng, O. R. L., & Tam, K. Y. 1999. “Examining the Technology Acceptance Model Using Physician Acceptance of Telemedicine Technology. *Journal of Management Information Systems*, *16*(2), pp. 91–112.

Moretti, F. 2011. “Network theory, plot analysis.” *New Left Review*, pp *68*.

Rosenzweig, R. 2003. “Scarcity or Abundance? Preserving the Past in a Digital Era,” *The American Historical Review*, *108*(3), pp. 735–762. https://doi.org/10.1086/ahr/108.3.735

So, R. J., & Long, H. 2013. “Network Analysis and the Sociology of Modernism,” *Boundary 2*, *40* (2), pp 147–182.

Stone, L. 1971.“ Prosopography, ” *Daedalus*, *100*(1), pp. 46–79.

Tu, H. C. 2016. “DocuSky: a DH Platform in the Making”. Presented at the The 7th International Conference of Digital Archives and Digital Humanities, Taipei.

Vijayasarathy, L. R. 2004. “Predicting consumer intentions to use on-line shopping: the case for an augmented technology acceptance model,” *Information & Management*, *41*(6), pp. 747–762. https://doi.org/10.1016/j.im.2003.08.011

Yeh, W. C. 2017. “Journal, Gender and Mobilization: Social Network Analysis of “Women’s Resonance” (1929-1944),” Presented at the The 8th International Conference of Digital Archives and Digital Humanities, Taipei.

金觀濤、邱偉雲、梁穎誼、陳柏聿、沈錳坤、劉青峰 (2016)。(觀念群變化的數位人文研究——以《新青年》為例)，收於項潔編，《數位人文──在過去、現在和未來之間》，臺北:台大出版中心，頁 427- 463。

陳詩沛、杜協昌、項潔 ，2011。(史料整體分析工具之幕後-介紹「台灣歷史數位圖書館」的資料前置處理程序)，收於項潔編，《從保存到創造: 開啟數位人文研究》，臺北:台大出版中心，頁 51 - 66。

項潔、涂豐恩 (2011)。(導論--甚麼是數位人文)，收於項潔編，《從保存到創造: 開啟數位人文研究》，臺北:台大出版中心，頁 9- 28。

趙薇 (2016)。(「社會網路分析」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)，收於項潔編，《數位人文──在過去、現在和未來之間》，臺北:台大出版中心，頁 398- 426。

劉吉軒、柯雲娥、張惠真、譚修雯、黃瑞期、甯格致 (2012)。(以文本分析呈現海外史料政治思想輪廓)，收於項潔編，《數位人文要義: 尋找類型與軌跡》，臺北:台大出版中心，頁 83-114。

(3) 網路資源

Taiwan History Digital Library (THDL). 2009. Retrieved September 14, 2018, from http://doi.org/10.6681/NTURCDH.DB_THDL/Text

China Biographical Database (CBDB). Retrieved September 14, 2018, from https://projects.iq.harvard.edu/cbdb/how-cite-cbdb

DocuSky. 2017. Retrieved September 14, 2018, from http://docusky.digital.ntu.edu.tw/DocuSky/ds-01.home.html

Mapping the Republic of Letters. 2008. Retrieved September 14, 2018, from http://republicofletters.stanford.edu/

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

大中華地區歷史地名系統之整合與開發

Integration and development of historical gazetteer system in Greater China

林農堯　何浩洋　郭俊麟

大中華地區歷史地名系統之整合與開發

Integration and development of historical gazetteer system in Greater China

林農堯、何浩洋、郭俊麟

從目前現有的大中華區域歷史地名資源來看，由哈佛大學 China Historical GIS placename database 所提供的的 TGAZ API 雖涵蓋了整個大中華地區，但對於臺灣的歷史地名的收錄其實並不豐富，讓許多進行臺灣歷史研究的學者，無法真正享受到地理時空資訊技術運用在人文研究上的好處。因此，本研究實作臺灣歷史地名與地籍 API，搭配哈佛大學的中國歷史地名 API，此研究成果可補足 TGAZ API 在臺灣歷史地名上的缺憾，且提供一個整合大中華地區的地名查詢平台，使漢學研究者可以獲得更完整的地名查詢功能。

臺灣歷史地名資料的來源，主要是中央研究院的台灣歷史文化地圖中採用創用 CC 授權的圖資，並對地名資料中的中文缺字及日文字進行補充及編輯修正。目前整理臺灣地名時期主要有 1684-1894, 1895-1926 與 1953-1999 等三個主要時期，其中地名行政階層分為四階，分別為廳、堡、街庄及土名。由於中央研究院的臺灣地名東部較為缺乏，在本研究中整合東華大學提供的東部地名資料，讓臺灣的地名資料可以更完備。地名的資料編輯中，加入唯一的 ID 編碼，收錄地名、時間、行政區級別及上一階層行政區名稱。經由這樣的設計，可透過指定地名 ID 或地名名稱取得位置的資料，並以 JSON 回傳，有利其它程式整合介接與呈現。

臺灣較完整的土地丈量始於劉銘傳擔任臺灣巡撫期間。劉銘傳 革新土地制度，設立清丈總局展開全臺清丈工作。日治時期實施戶口調查，按戶就田賦實施土地清丈，使戶籍與地籍資料結合，為臺灣地籍管理之開端。由於清代及日治時期的地籍成果與臺灣現代的地籍地段大致都可對應。因以，本研究進一步以臺灣歷史地名架構 API 為基礎，配合科技部「運用 THDL 重現台灣土地秩序計畫」的歷史地籍圖空間定位成果，將原本的地名查詢，再精進定位至土地等級。歷史地籍的資料格式為堡名+街庄+地號或堡名+街庄+土名+地號。其中地號又分成母號與子號。目前地籍的定位已可精準到母號，待日後全台歷史地籍建構完成，即可使用本研究開發的地籍 API 進行整合運用。

透過本研究成果開發臺灣歷史地名與地籍 API 及介接哈佛大學的 TGAZ API，已建立一個跨平臺的的大中華地區歷史地名查詢系統。本研究成果目前已提供荷蘭萊登大學(Leiden University)MARKUS-古籍標記平台與臺灣大學 DocuSky-數位人文學術研究平臺等開發符合自身需求的臺灣地名查詢系統。在 DocuSky 上的 DocuGIS 文本地理資訊系統中，可使用本研究成果，檢索大中華地區的歷史地名位置並產生新的圖層以供後續的分析使用。

關鍵字：臺灣歷史地名、臺灣歷史地籍、歷史地名檢索、MARKUS，DocuSky，DocuGIS

一、緣起

此大中華地區歷史地名系統建置與研究團隊在執行科技部的「運用 THDL 重現 17-20 世紀台灣土地秩序-古契書的時空定位與邊區官有原野土地開發的地緣整合分析」(MOST105-2420-H-259-013-MY2)有密切的關係。其中一個研究重點是要針對台灣歷史數位圖書館(Taiwan History Digital Library, THDL)中的「古契書」文獻集收錄四萬餘件全文資料及影像，來進行空間對位。 古契書本文獻集由國立台中圖書館與國立臺灣大學圖書館等單位合作，將臺灣總督府檔案抄錄契約文書、岸裡大社文書等內容加以數位化。 內容含括明清以降，臺灣的各類契約，其中又以土地契約為最多，但亦包括相關之公私文書，如契尾或婚姻契等(項潔, 陳詩沛, &杜協昌, 2009)。這些龐大的契約書檔案若沒有完整的歷史地名資料對照，將很難進行後續的分析處理。

THDL 古契書中較為較明確的空間位置的描述是臺灣總督府檔案抄錄契約文書，其中約有一萬餘筆已使用 1904 台灣堡圖圖資定位，但其它的契書在本計畫執行前則尚未定位處理。人文學者在 THDL 系統中使用關鍵字搜尋後，僅呈現臺灣總督府檔案的資料。這樣的地圖呈現資料是偏頗，尤其是 THDL 的古契書資料是較完整滿且有系統地收錄全台的契書。使用上僅有四分之一的資料呈現空間上是很可惜的。人文學者將資料在空間上呈現，會受到資料限制，造成讀圖的誤解。這樣的限制主要是資料的完整度、資料定位的精準度及兩者的相互影響(林農堯, 2018)。於是透過資訊技術與結合中研究的臺灣歷史圖資，將古契書更多的地名資訊找出更多契書的的空間位置，乃是當下重要的研究課題。

研究團隊因此分為以下幾個階段進行臺灣歷史地名與古契書空間定位的資訊處理。首先重新整理各時期臺灣歷史地名將其分為 1684-1894, 1895-1926 與 1953-1999 等三個主要時期，將中央研究院的台灣歷史文化地圖中採用創用 CC 授權的圖資，並對地名資料中的中文缺字及日文字進行補充及編輯修正。其次則是將將 THDL 中古契書全文的中文地名進行資訊擷取，接著利用修正後的台灣歷史地名資料庫，進行古契書的地名比對，此時使用的歷史地名行政階層分為廳、堡、街庄及土名等四階來進行比對。最後則是反覆修正 GeoPort 介面，希望能將 THDL 中的古契書空間定位的完整度及精準度提升，讓古契書檢索的分佈的呈現更有意義。若能進一步將此計畫系統性整理臺灣歷史地名及地籍資訊建立 Application Programming Interface(簡稱：API)，可提供其他平臺或程式介接使用。本研究因此提出整合臺灣歷史地名 API、臺灣歷史地籍 API 及哈佛大學所提供的之中國歷史地名 TGAZ API 做為大中華區歷史地名整合查詢的研究構想。

二、建立臺灣歷史地名資料庫與 API

(一)臺灣歷史地名資料

「地名」代表某一空間之符號，任何地名皆反映一個地方之歷史，地理狀況，聚落發展，開拓的沿革。臺灣歷史地名資料最重要的來源是中央研究院建置之《臺灣歷史文化地圖系統》，近來該系統已使用創用 CC 授權的方式供一般使用者下載使用。其公告的授權條款為「創用 CC 授權條款：姓名標示—非商業性—相同方式分享」，允許國內針對國內研究、教學等非商業使用者，重製、散布、傳輸以及修改資料，但仍須使用相同的授權規範來散布衍生作品。[1]

我們整理臺灣歷史文化地圖系統中的圖層[2]，其時間範圍包含 1654 荷西時期至 1926 的日治時期，126 類型，約有 263 個圖層資料。首先將其中地名的資料進行日文及缺字編輯；在坐標的校正上，若是多邊形的資料則取其質心坐標來定位。這些圖資資料最重要的是台灣堡圖，裡面有 1904 台灣堡圖年左右的廳、堡、庄及土名資料，這圖層是所有資料中數位化最詳細且最具參考價值的資料。本研究因此整理出以下三階段的地名時間範圍：1684-1894，1895-1926，1953-1999，總共約 4 萬多筆的臺灣歷史地名資料，資料樣本如圖 1 所示。其中罕見字、日文缺字已使用 unicode 的字碼重新修正。主要的欄位設計有 8 個，羅列於下：

1. id：前綴為 twgis_代表為臺灣地名的唯一編號
2. NAME_SIM:為地名
3. BEG:為地名開始時間
4. END:為地名結束時間
5. X:為 WGS84 之經度坐標
6. Y:為 WGS84 之緯度坐標
7. TYPE_SIM:為該地名的行政階層
8. PARTOF_SIM:為上層行政區之名稱

id	NAME_SIM	BEG	END	X	Y	TYPE_SIM	PARTOF_SIM
twgis_13865	ウライ社	1904		121.5483814	24.8605307		
twgis_6100	カピヤン社	1904		120.6435911	22.59560583		
twgis_38064	一七三	1999		120.5012688	24.16824748	四等城市	伸港鄉
twgis_42084	一坑	1999		121.7143284	25.0216067	四等城市	平溪鄉
twgis_8929	一甲	1904		120.2565976	22.97730662		
twgis_22066	一甲	192007	192607	120.2523602	22.9791465	大字	仁德庄
twgis_22129	一甲	192007	192607	120.2619123	22.8777366	大字	路竹庄

圖 1 地名資料庫內容

（二）臺灣歷史地籍資料

[1] 詳細授權條款請參閱：創用 CC 法律條款(CREATIVE COMMONS, 2018)

[2] 臺灣歷史文化地圖的歷史圖層的清單可以參考 https://thcts.ascc.net/view.php 網址。

臺灣的地名資料雖然在中國大陸與歐美的研究單位已有提供相關線上地名資料庫服務，但資料仍不夠完整且多無法深入比對歷史地籍相關檔案。由於地名的特性與應用有其地域性，臺灣因面積狹小、人口稠密，地名的密度與複雜度亦非常的高。由於臺灣的地籍資料在日治時期已建立完整且系統性的架構，若能將臺灣歷史地籍資料納入地名資料庫系統中，將更有助於歷史文獻的空間定位與探索。因此本研究以臺灣歷史地名架構與其 API 為基礎，首先利用研究團隊在科技部「運用 THDL 重現 17-20 世紀台灣土地秩序」整合型計畫所取得新竹廳、苗栗廳、桃仔園廳地區三千餘筆的土地申告書的番租口糧資料進行研究測試[3]。另一方面，在搭配 1953 年臺灣實施耕者有其田政策時，為掌握臺灣地籍狀況，將全省各地政事務所保管由地籍正圖描繪並用 106 磅圖紙製作的地籍藍曬底圖(簡稱「160 磅地籍藍曬圖」)作為歷史地籍資訊抽取的基礎圖資。經校正比對後，共生 3 千多筆地籍圖地號坐標資料。

(三)臺灣歷史地名地籍 API

依目前的臺灣現況而言，相關的地名學研究仍屬區域性文獻研究與收集為主，著重於村里級之地名沿革探討，雖有涵蓋所有臺灣地區正在使用的地名資訊，但各系統資料庫間仍多無法相互溝通運用。因此建立公開的臺灣歷史地名 API 有其刻不容緩之必要性。

本研究因此嘗試將中研院「臺灣歷史文化地圖系統」與科技部「運用 THDL 重現 17-20 世紀台灣土地秩序計畫」的歷史地籍兩種資料來源，整合成臺灣歷史地名資料庫，並進一步加值建置成臺灣歷史地名 API。如此，各種系統需要用到臺灣的歷史地名，均可以使用 API 規範的標準方式，呼叫及取得坐標且整合呈現在自己的 GIS 系統中。API 查詢地名主要的程式是用 PHP 開發；Web Map 則是使用 Javascript、JQuery 與 Leaflet 開發。API 查詢其使用方法及網址說明如下：

1. 系統與 API 網址

表 1. 系統及 API 網址一覽

說明網址	https://goo.gl/t5K6nU
系統網址	http://docusky.digital.ntu.edu.tw/DocuSky/docutools/geocode/map.html
地名 API	http://docusky.digital.ntu.edu.tw/DocuSky/docutools/geocode/tw.php
地籍 API	http://docusky.digital.ntu.edu.tw/DocuSky/docutools/geocode/twcada.php

原始資料來源:

1. 中央研究院，《臺灣歷史文化地圖系統》第一版，（台北，中央研究院，2003 年 9 月）。
2. 臺灣大學數位人文研究中心。
3. 科技部「運用 THDL 重現 17-20 世紀台灣土地秩序」整合型計畫。

[3] 該資料由整合型計畫總計畫暨子計畫一主持人彰化師範大學歷史所李宗信副教授提供。

2. API 介接及參數設定說明

本研究所建置的臺灣歷史地名及歷史地籍資料庫之 API 網址及相關參數設定方法詳見下表。

表 2. API 介接說明

<table>
<tr><td colspan="2">ID 編碼</td><td>地名編碼 twgis_1
地籍編碼 twcada_1</td></tr>
<tr><td rowspan="2">API 與 UI 說明</td><td>臺灣歷史地名</td><td>預設回傳 JSON
http://docusky.digital.ntu.edu.tw/DocuSky/Docutools/Geocode/tw.php
預設回傳 Map
http://docusky.digital.ntu.edu.tw/DocuSky/DocuTools/Geocode/map.html</td></tr>
<tr><td></td><td>參數 urlencode 的中文完整地名或地名 ID，範例如下：
?n=台北縣
?n=twgis_1
設定回傳地圖的中心位置
map.html
?cs=23.8，121.7，7 //臺灣為中心
?cs=28，107，3 //大中華為中心</td></tr>
<tr><td rowspan="2"></td><td>臺灣歷史地籍</td><td>預設回傳 JSON
http://docusky.digital.ntu.edu.tw/DocuSky/Docutools/Geocode/twcada.php
預設回傳 Map
http://docusky.digital.ntu.edu.tw/DocuSky/DocuTools/Geocode/map.html</td></tr>
<tr><td></td><td>參數 urlencode 的中文 廳、堡、街、庄、土名、地號，地籍編號
?t=新竹廳
&b=竹南一堡
&z=海口庄
&d=海口
&n=167
?n=twgis_1
回傳參數
?n=twcada_1</td></tr>
</table>

備註：*紅色為必要參數

由於文獻檔案中的地籍資料常不完整，或因文獻繕寫方式與資料記錄有差異而導致查詢時不易使用。本研究根據三千多筆土地申告書資料，簡化並建議地籍輸入及查詢方式如下:

第一類:街庄+地號

原件記載例如:「竹南一堡頭份庄六三四ノ一」，則需輸入為「竹南一堡，頭份庄，，634-1」因該筆資料並未有土名，因此在街庄及地號間，需用逗號空一欄位，且地號母號 634 與子號 1 號間以「-」區隔。

第二類:街庄+土名+地號

原件記載例如:「竹北一堡濫仔庄土名薯園五三|二」，此筆資料有完整街庄、土名、及地號，因此輸入方式為「竹北一堡，濫仔庄，薯園，53-2」，其中五三為地籍母號，二為子號。

藉由地籍資料庫的整合，本系統可將地名的空間定位精度提升到地籍的母號，並透過 API 可做為其他系統平台的介接服務。目前此服務簡稱為 twgis 資料庫及 twgis API，希望藉由此 API 的公開，在日後整合更多臺灣地名研究成果或相關服務。

三、運用實例

透過本研究建立的公開的臺灣歷史地名資料庫與 API，使用者已可自行整合運用。以下就列舉(1) 大中華地區歷史地名系統、(2) MARKUS 整合、(3) THDL 古契書之空間坐標等三個具體應用實例，說明其應用的價值與潛力。

1. 大中華歷史地名查詢

服務網址:
http://docusky.digital.ntu.edu.tw/DocuSky/docutools/geocode/map.html

目前現有的大中華區域歷史地名資源來看，由哈佛大學 China Historical GIS placename database 所提供的的 TGAZ API 雖涵蓋了整個大中華地區，但對於臺灣的歷史地名的收錄其實並不豐富，讓許多進行臺灣歷史研究的學者，無法真正享受到地理時空資訊技術運用在人文研究上的好處。因此，使用本研究實作臺灣歷史地名與地籍 API，加上哈佛大學的中國歷史地名 API，我們的研究成果可補足 TGAZ API 在臺灣歷史地名上的缺憾，且提供一個整合大中華地區的地名查詢平台，使漢學研究者可以獲得更完整的地名查詢功能。

臺灣的歷史地名行政階層分為四階，分別為廳、堡、街庄及土名。地名的資料編輯中，主要會加入唯一編號 ID，收錄地名、時間、行政區級別及上一階層行政區名稱。其中編號的前綴為 twgis_開頭，之後接阿拉伯數字的號碼，如此就賦予每個時期的地名唯一的 ID 編碼。透過這樣的設計程式可透過指定地名 ID 或地名名稱取得臺灣地點的資料。API 傳回的結果是 JSON，以利其它程式整合介接與呈現。

本查詢系統以整理臺灣歷史地名及地籍資訊，建立 API 且提供其他平臺或程式介接使用。同時也整合臺灣歷史地名 API、臺灣歷史地籍 API 及哈佛大學之中國歷史地名 TGAZ API 提供大中華區歷史地名整合結果查詢之地圖網站。目前查詢系統建立在台大數位人文中心的 DocuSky 平台，提供全球漢學學者查詢歷史地名。



圖 2. 大中華歷史地名查詢系統

在系統查詢「公館」地名，同時整合哈佛大學 TGAZ API 及本系統之 API 回傳一共有 42 筆的公館地名呈現在地圖上。地圖適當的放大縮小後，可以發現 37 筆在臺灣、5 筆在中國。系統左下的搜尋結果也會出現所有地名的清單，其中有地名編號、地名、坐標及時間資訊。點擊地圖上單一地名的圖示會出現相關連結，可以使用 API 或維基百科尋找該地名的沿革。



圖 3. 公館地名整合查詢結果

歷史地籍的資料是由 1953 年「160 磅地籍藍曬圖」空間定位產生。利用這些歷史地籍資料，可將原本的地名查詢，再精進定位之精準度至土地等級。臺灣歷史地籍 API 會將定位到母號且傳回坐標值。目前此地籍 API 架構完備，未來只待補充地籍資料即可。透過此精確到地號等級的查詢系統與 API 完成，可以讓之後其它大中華地區的研究參考此計畫的成果。

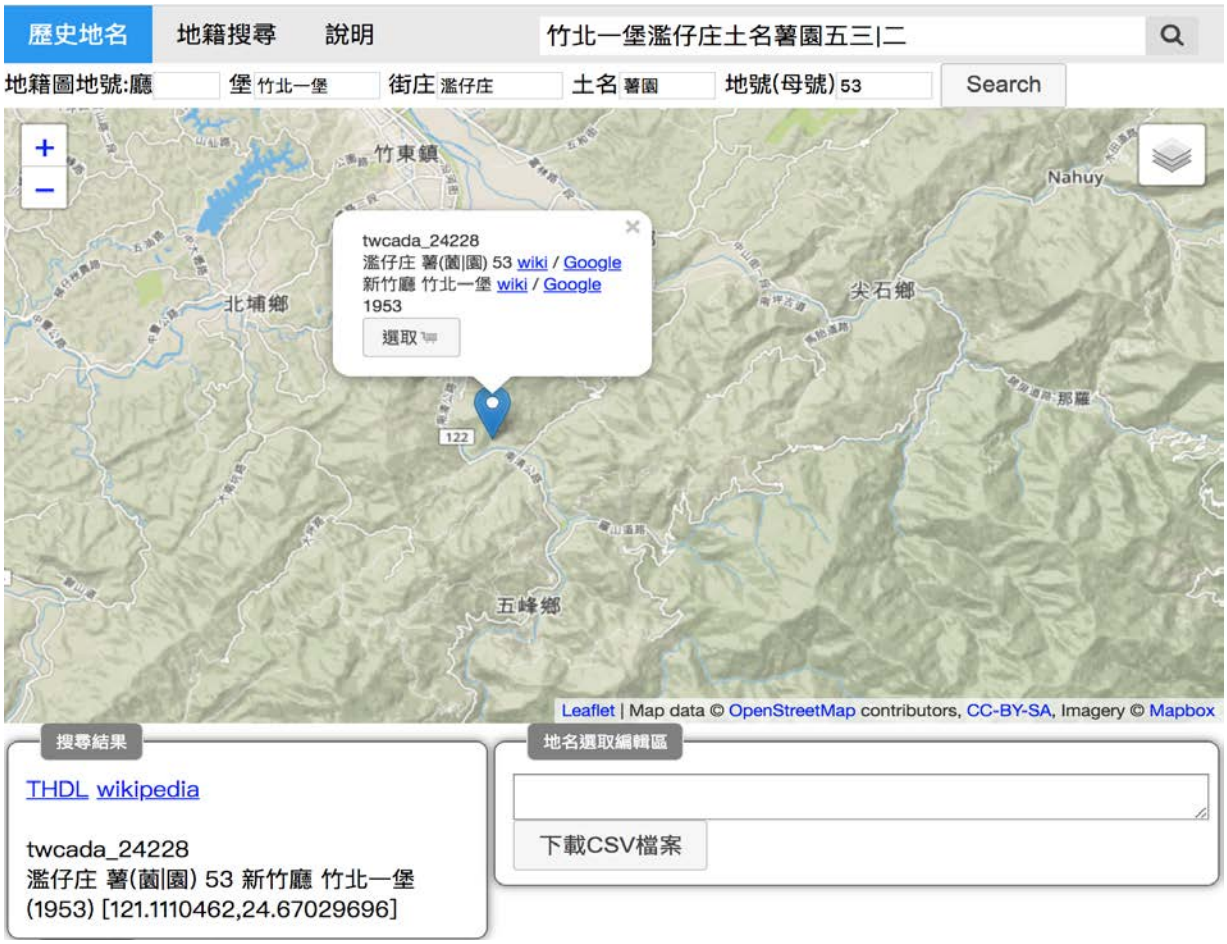


圖 4. 歷史地籍查詢

本計畫運用科技部「運用 THDL 重現 17-20 世紀台灣土地秩序」整合型計畫的圖資建置成果，特別是中研院向臺灣的各地的地政事務所取得的「160 磅地籍藍曬圖」成果後，由計畫團隊進行地籍圖影像空間定位後的圖資，以及中研院 GIS 中心建立歷史地籍圖等相關 WMTS 圖磚服務。使得本系統得以進行相關資料的空間整合查詢及圖磚套疊的服務，如圖 5。



圖 5. 歷史地籍圖層套疊

另一方面，DocuSky 中的 DocuGIS 整合大中華地區的歷史地名地籍的查詢系統。主要是透過透過 API 及操作介面的整合，在 DocuGIS 文本地理資訊系統中，則可使用地名查詢系統，檢索大中華地區的歷史地名位置並產生新的圖層以供後續的分析使用，如圖 6。

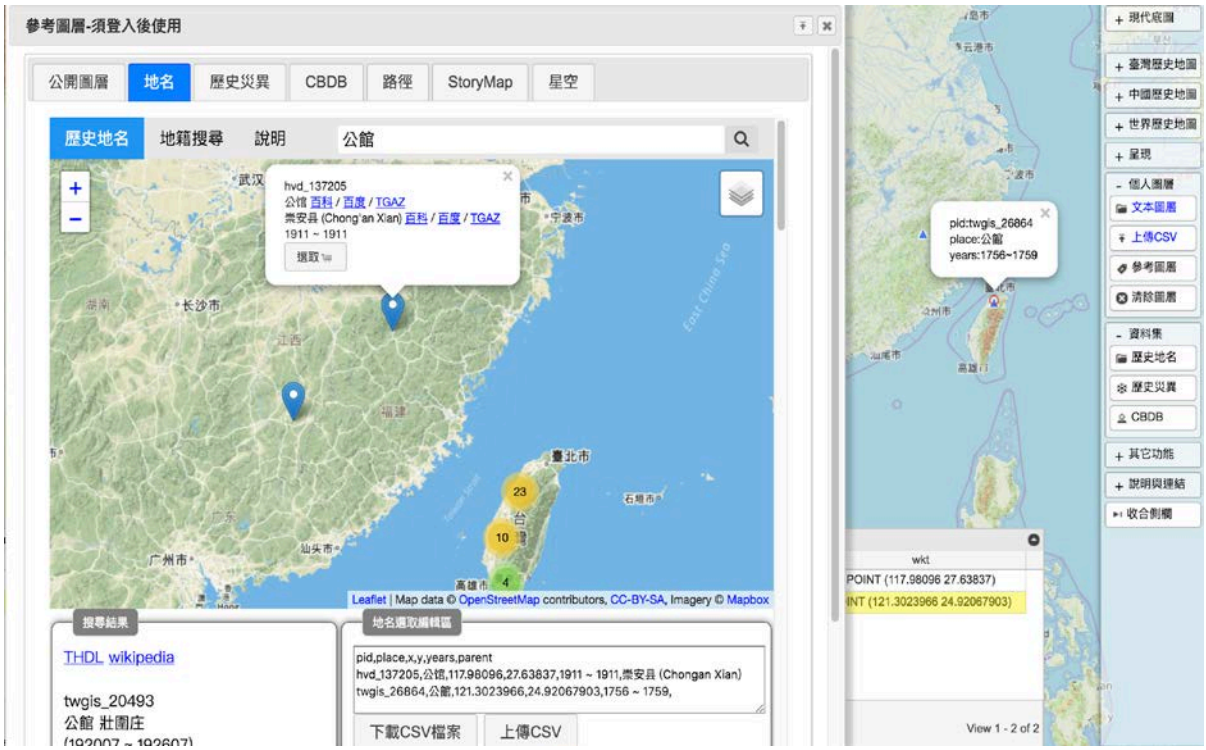


圖 6. DocuGIS 整合使用大中華歷史地名查詢系統產生新的圖層

2. MARKUS-古籍標記平台

服務網址： https://dh.chinese-empires.eu/markus/beta/

荷蘭萊登大學(Leiden University)的 MARKUS 為古籍標記平台，主要處理漢學的全文中的人事時地標記。但在此計畫之前 MARKUS 的地點標註只有支 TGAZ，也就是中國地區的歷史地名及極當少數的臺灣地名。在本研究之臺灣歷史地名地籍資料庫、API 及大中華整合查詢的成果之已提供 MARKUS 整合及使用後，在 MARKUS 中標記臺灣文本時，可自動或手動標記臺灣的歷史地名。圖 7 中的 2.a Automated markup 裡的 Taiwan GIS(TWGIS)即是整合此計畫的地名成果。

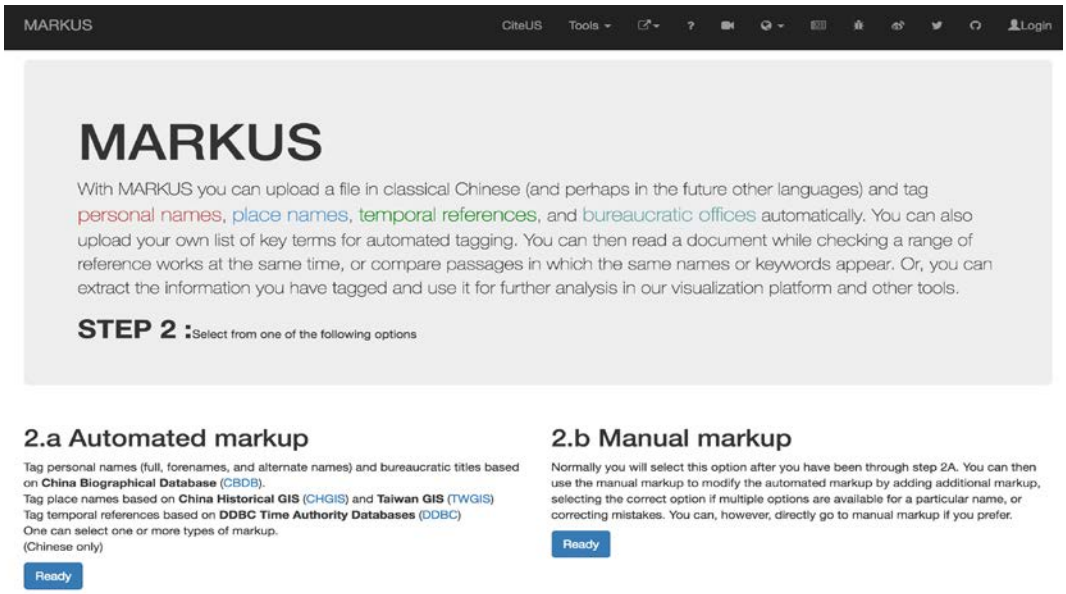


圖 7. MARKUS 標記時可以使用 Taiwan GIS 地名資料

在使用自動標記時，可以將勾選臺灣歷史地名以自動找出文本中的臺灣歷史地名，如圖 8。

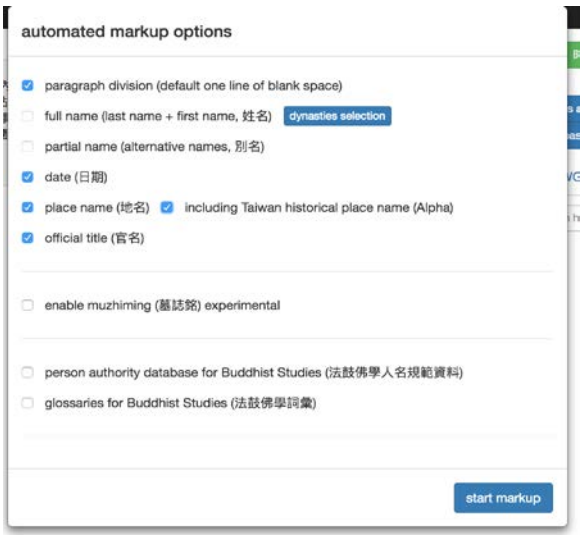


圖 8. MARKUS 自動標記時可使用臺灣歷史地名

在找出文本中的臺灣地名或手動標記時，可以指定地名唯一 ID。在圖 9 的畫面中，右側欄的 TWGIS+TGAZ 即是使用本計畫的大中華歷史地名整合查詢成果。讓人文學者透過一次查詢，即可以找中國 TGAZ 與臺灣 TWGIS 的歷史地名。



圖 9. MARKUS 手動標記時可以使用大中華整合查詢介面

3. 尋找更多 THDL 古契書坐標

本研究結合臺灣歷史地名及地籍 API，可將 THDL 中更多古契書的坐標定位，彙整更完整的契書空間資料庫，此工作主要可包含(1)將 THDL 古契書之全文中地名截取、(2)使用 twgis 資料庫比對古契書地名、(3)人工編修介面 GeoPort 等三大部分。

第一部分主要是用正規表達式及詞夾子演算法(張尚斌, 2006)找出契書中的地名。透過正表示式及詞夾字技術的運用，可找出契書裡大部分的地名。包括重要的地契位置地名及次要的其它地名。這些地名資料要依重要性，找出坐標即可以在地圖上定位出契書的位置。

第二部分是使用臺灣歷史地名資料庫後比對古契書中的地名。執行畫面如下。在 Conent 中有全文，藍色的字是前一步驟找到的地名。地契清單中會列出載入地契的檔名。點選每一篇地契時會列出計算好的坐標給使用者檢視。

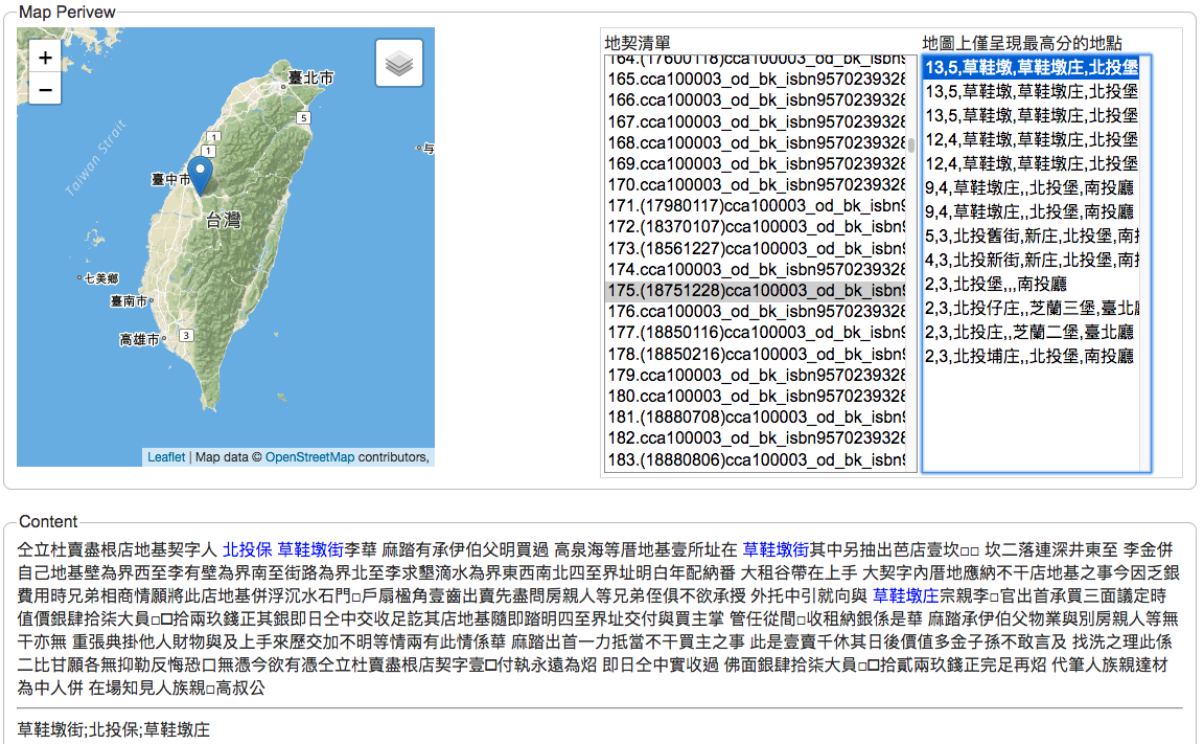


圖 10. 臺灣歷史地名資料庫比對 THDL 古契書

最後可以將此計算結果下載成檔案，以利之後的校對及編輯。

實作的作法如下：

(1)將 twgis 地名資料庫載入記憶體中

(2)每篇的文本中的地名進行正規化

(3)將正規化的地名與與資料庫中的地名進行比對。

(4)比對符合的地名均加入候選地名

(5)再設計調整權重演算法，將候迴地名依將權重調整。

地名經過繁簡及異體字正規化處理後，即可進行文本地名及 twgis 資料庫地名比對。比對的方式是將每篇文本的地名，依序至地名資料庫中尋找符合的地名。包含同名及各皆層級的地名，所有找到的地名均保留至地名候選清單中。最後再將地名候選清單進行權重調整。調整的權重的原則有三準則，第一是 LCS 值要高；第二精確度要高；第三常見地名減分。

LCS(Longest Common Subsequence)是最長公共子序列，它是是一個在一個序列集合中（通常為兩個序列）用來尋找所有序列中最長子序列的問題。這與尋找最長公共子串的問題不同的地方是：子序列不需要在原序列中占用連續的位置 。透過文本地名及匹配的 twgis 地名的最長字串總數，即可以簡單顧及文本出現所有出現的地名及初步的排序(維基百科, 2018)。第二是找到同樣名稱的地名，以行政層級較小的為優先。以提高定位的精確程度。第三常見的地名，如吳厝、舊厝，會影響排序結果，所以設計參數以減低其出現的權重。

透過以上三準則方式調整權重，即可以將候選地名的清單調整至較符合文本的狀況。不過本研究的成果是要提供人文學者使用，為了謹慎起見。我們設計一介面提供專人校正及確認。

第三部分是使用人工編修介面 GeoPort，古契書是研究臺灣早期土地的重要資料，雖然有資訊技術找出坐標資料。但為了提供優質的服務水準，我們也請相關領域專家協助驗證查核，也可以驗證程式的水準。我們在 DocuSky 的平台下，設計可以編輯 Metadata 坐標的程式-GeoPort。可以載入各式個人資料庫，例如古契書。畫面右方會出現契書列表，在每一筆契書的右上方，有[修改坐標]，如圖 11。



圖 11. 契書列表及查驗介面

GeoPort 編輯古契書坐標，有兩種方式。第一種使用候選坐標，在前一步比對地名資料庫的結果，如圖 12 在綠框中的資料。

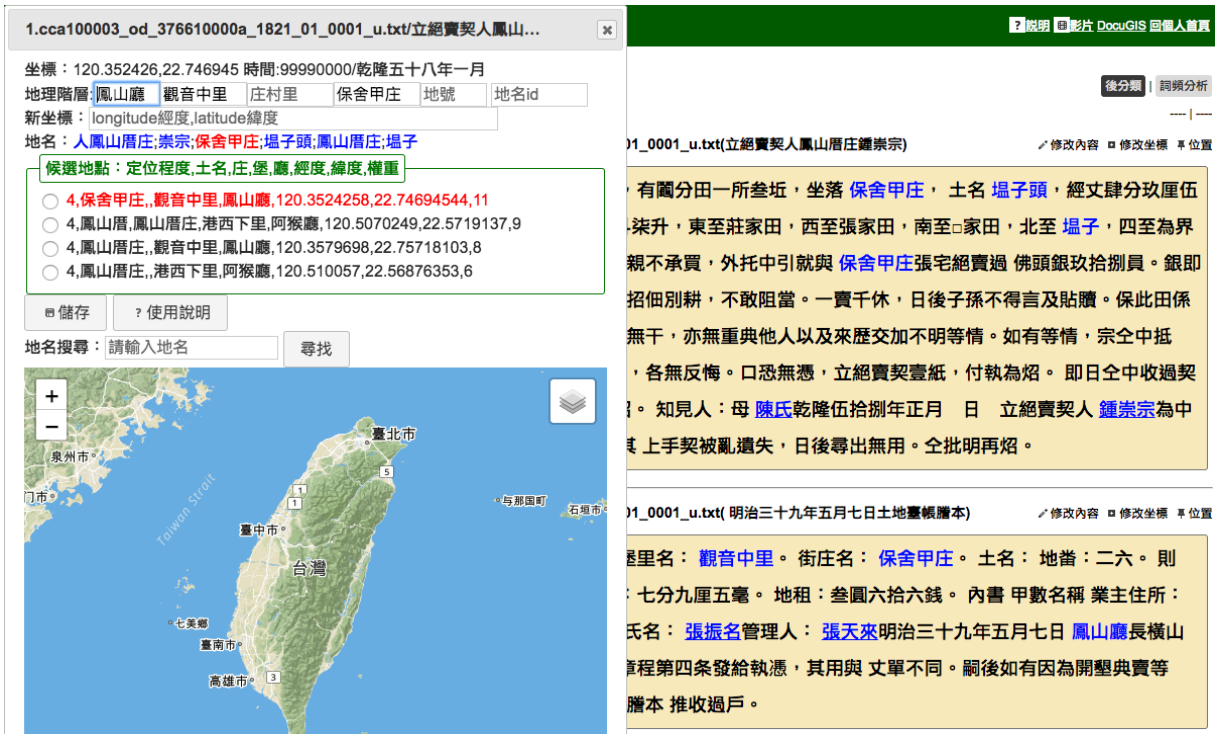


圖 12. 古契書坐標編修介面

第二種方式是使用歷史地名搜尋。這是直接整合本研究的成果的 API，在編修坐標時提供臺灣歷史地名搜尋。在實作上編修者可以使用契書文本裡的地名或藍色地名來搜尋，如圖 13 中的人鳳山厝庄;崇宗;保舍甲庄;塭子頭;鳳山厝庄;塭子。點擊藍色地名，會自動將該地名填入地名搜尋欄位中。以實際的例來說，已確定此契書在保舍甲庄，但找不到在下一階層的塭子頭。用塭子頭搜尋後發現竹南鎮有兩筆，但不是在保舍甲庄中。所以即使有找到，但還是不行使用竹南鎮的塭子頭。



圖 13 使用歷史地名搜尋

使用臺灣歷史地名資料庫與 API，我們可以大量且自動的處理 THDL 古契書的坐標資料。再加上專人的編審可以大幅度的提高 THDL 古契書的服務水準。

四、結論

臺灣的政府組織或學術單位都有臺灣的歷史地名資料庫系統，但是臺灣的歷史研究上一直缺乏一個公開可用的歷史地名的 API。透過本計畫的成果之一 - 建置成臺灣歷史地名資料庫與 API，並且能夠實際公開提供臺灣歷史地名 API 服務，不僅能符合臺灣歷史研究的應用需求，同時示範整合哈佛大學的 TGAZ，如此也能兼顧中國的歷史地名。

我們也將此成果與萊登大學一起分享，並將我們的成果整合至 MARKUS 中。如此對國際上研究漢學的學者或開發者提供的一個完整的漢學地名服務，為將來全球化漢學資訊整合預先做好準備。本研究的執行成果，不僅是在完成本計畫成標的古契書之坐標定位，也由於公開及開放成果與 API，擴展學術公享與實際國際合作。

參考文獻

項潔、陳詩沛、杜協昌，〈台灣古契約文書全文資料庫的建置〉，《第三屆台灣古文書與歷史研究學術研討會論文集》（臺中：逢甲大學出版社，2009），頁 243-269。

林農堯，〈文本地理資訊系統與運用〉，《臺北：國立臺灣大學資訊網路與多媒體研究所博士論文，2018》，頁 129。

中央研究院，《臺灣歷史文化地圖系統》第一版（台北，2003 年 9 月），http://thcts.ascc.net/。

CREATIVE COMMONS，姓名標示—非商業性—相同方式分享 3.0 台灣，https://creativecommons.org/licenses/by-nc-sa/3.0/tw/legalcode。

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。

維基百基，〈最長公共子序列 LCS〉，https://zh.wikipedia.org/wiki/%E6%9C%80%E9%95%BF%E5%85%AC%E5%85%B1%E5%AD%90%E5%BA%8F%E5%88%97，2018 年。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

時空視角

文本挖掘〈李娃傳〉中唐長安城的空間感知

馬昭儀*　劉帥帥*　何捷**

天津大學研究生*

天津大學副教授**

時空視角：文本挖掘〈李娃傳〉中唐長安城的空間感知

馬昭儀，劉帥帥，何捷
研究生，研究生，副教授
天津大學

摘要：

針對唐傳奇〈李娃傳〉的研究歷來將文本和城市空間的結合作為探討的重點，我們從敘事學的時空視角運用數位工具對其進行了重構，探索了唐人白行簡的敘事手法和其構建的故事世界。從時間線上觀察作者的創造文學世界的手法、文學世界的場景活動，最後回歸空間使得時間線被壓縮在平面內觀察感知語境中的歷史城市。

敘事視覺化中大部分的探討屬於文本結構的探討。從文本時間與真實時間的對比來看小說的故事時間安排，可以看到前密後疏以及開頭倒敘的寫法，結尾虛化。全文情感曲線 N 型“走向成熟”類小說模型。人物情感中能夠看到的是滎陽生的人生深究起來其實被政治（父）、世俗（姥）、愛情（娃）所影響。〈李娃傳〉人物情感線中間大段李娃的缺失和最後的突然出現，表現出的是一種男性視角下的傳記寫法。地點情感曲線中，地點對情節的推動作用也表現的異常明顯。空間敘事視覺化可以看到被精心安排的情節是怎樣在空間和話語時間裡展開的。

對故事世界的評價建立在將其視為一個世界的基礎之上進行觀察分析。人物軌跡統計從一種街道統計視角來看也許會發生在長安城的日常生活和形色人等。地點網路連接分析從拓撲結構的角度挖掘出四類社群、東市「名人」、宣陽坊「傳

播瓶頸」、宣陽-安邑-平康「八卦傳播」的隱含資訊。關鍵字地圖注重對環境的直觀反映，東區安邑坊、平康坊、宣陽坊這三個高檔居住區的細微差別被察覺，作者對東區的瞭解被洞悉。情感地圖的不同地塊的情感傾向呈現兩對倒三角的平衡狀態，這也許就是當初世界的最初設置。這些從空間層的深度挖掘已經脫離了文本而進入世界，體現出了作者潛在的城市感知。

目次

1. 〈李娃傳〉研究中文本與空間的結合

2. 〈李娃傳〉文本到空間的新方法探索

2.1 文本預處理

2.1.1. 詞層級

2.1.2. 句層級

2.2. 敘事視覺化

2.2.1. 文本時間下的敘事分析

2.2.1.1. 文本時間和真實時間

2.2.1.2. 全文情感

2.2.1.3. 地點情感

2.2.1.3. 人物情感

2.2.2. 三維時空製圖

2.2.2.1. 故事時間和文本時間下的敘事三維視覺化

2.2.2.2. 文本時間下的人物軌跡三維模擬

2.3. 空間統計

2.3.1. 模擬軌跡的統計

2.3. 2. 地點網路連接分析

2.3.3. 關鍵字地圖

2.3.4. 情感值地圖

3 討論

致謝

關鍵詞

李娃傳，空間敘事，故事空間，城市感知

1 〈李娃傳〉研究中文本與空間的結合

唐傳奇小說〈李娃傳〉由白行簡（776 年－826 年）所撰，敘述了滎陽大族滎陽生赴京趕考，卻邂逅長安娼女李娃與之熱戀，在屢經波折、歷經艱辛後終獲美好結局的故事（圖 1）。該小說因其不凡的藝術魅力和文學價值，一直受到歷代學者的青睞和重視。

創作背景與思想主題、人物形象、藝術特色等文學類探討是〈李娃傳〉歷來的研究熱點。其中從情節結構量化視角和空間敘事角度的探索尤其具有啟發意義。日本學者妹尾達彥在〈唐代後期的長安與傳奇小說—以〈李娃傳〉的分析為中心〉一文中極具創見性地指出文本情節、人物軌跡與長安坊裡的緊密連接，小南一郎《唐代傳奇小說論》沿用前者的論點進一步對情節和長安城空間的起伏和層級進行了論述[12]。另有朱明秋〈〈李娃傳〉情節數理批評〉、吳淑鈿〈門之內外:李娃故事的敘述結構〉、李雙紅〈淺論唐傳奇〈李娃傳〉的故事情節〉等嘗試對小說情節進行簡單量化解讀和空間解讀[345]。除了對小說的空間場景與敘事關係的探討，還有很多研究以長安城空間為主要著墨點，探討小說的城市背景。李效傑等、王歡分別對小說反映的唐代長安的商業發展狀況、房舍租賃現象進行了專門的探討。[67]朱玉麒的〈隋唐文學人物與長安坊裡〉指出隋唐文學往往將長安坊裡填充為其作品虛擬生活場景的真實外殼，探討了〈李娃傳〉中的公共空間與庶民信仰建築[8]，文美英在〈唐人小說中的長安城——以傳奇為主〉中將〈李娃傳〉等唐

[1] 妹尾達彥，〈唐代後期的長安與傳奇小說——以〈李娃傳〉的分析為中心〉，《日本青年學者論中國史·六朝隋唐卷》，頁 519
[2] 小南一郎（著），童嶺（譯），《唐代傳奇小說論》。頁 111
[3] 朱明秋，〈〈李娃傳〉情節數理批評〉，《桂林師範高等專科學校學報》3，頁 83
[4] 吳淑鈿（2006）。門之內外:李娃故事的敘述結構，《古籍研究》2,194-201
[5] 李雙紅（2007）。〈淺談唐傳奇〈李娃傳〉的故事情節〉，《三峽大學學報（人文社會科學版）》S2，頁 1
[6] 李效傑,張紅雲（2017）。〈從〈李娃傳〉看唐代的商業競爭〉，《山東工商學院學報》6,頁 1-5+36
[7] 王歡（2007）。〈〈李娃傳〉中的房舍租賃，《河南農業》 2，頁 56-58
[8] 朱玉麒（2003）。〈隋唐文学人物与长安坊里空间〉，荣新江主编，《唐研究》第九卷。北京：北京大学出版社

傳奇作為研究考證長安物質生活和精神意識的資料。這些研究繼承了《唐兩京城坊考》的作者徐鬆開創的小說證史的方法，部分承認了小說作為長安城地理景觀和生活文化還原依據的價值。

這些基於〈李娃傳〉小說文本的探索，洞見性地將小說文本和城市空間進行結合，揭示了這一名篇所蘊含的文學藝術和城市史料價值，初步萌現了對文本的量化和城市日常生活史的重視，契合了近年來社會科學領域的“人文計算”/“數字人文”和“空間轉向”的新方向。因此在此基礎上，本文將進一步從“數字人文”和“空間人文”視角來探討，運用數位化技術重新深度認識和解讀小說中的城市、景觀（landscapes）和場所（place）。通過〈李娃傳〉這一媒介，本文將運用文本處理和文學製圖的途徑，探索由白行簡這一在唐長安社會文化背景下長久活動的文學家所構建的真實與虛構交互的文學世界所映射出的嵌入了其個人經歷的城市地理感知狀態。

情节序号	情節
1	滎陽生出身官宦家庭，是家裡的千里駒，滎陽公非常器重他，並以豐厚盤纏送其赴考。
2	滎陽生在長安與李娃邂逅，兩情相悅，與李娃一起生活。
3	滎陽生為其散盡盤纏，被老鴇嫌棄，李娃與老鴇設計拋棄滎陽生。
4	滎陽生輾轉進入西肆，靠唱挽歌維持生計，滎陽生在挽歌大賽中贏得比賽。
5	偶遇父親，因玷辱門庭被父親鞭打數百。
6	被東肆同伴所救，後又被拋棄，淪為行乞者。
7	滎陽生偶遇李娃，李娃懺悔之前的過錯。
8	在李娃細心照料陪護下，滎陽生身體恢復了健康，考取了功名。
9	李娃因出身原因欲與滎陽生分離。
10	滎陽生偶遇父親，父子和好如初，與李娃結為夫妻。李娃治家有方，子嗣均有大成。

圖 1 〈李娃傳〉基本情節

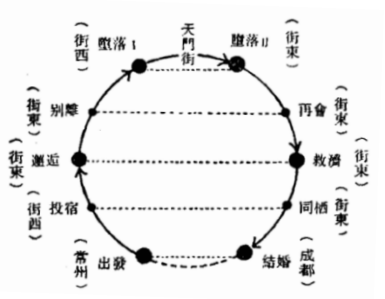


圖 2 妹尾達彥對〈李娃傳〉結構的圖解

2 〈李娃傳〉文本到空間的新方法探索

〈李娃傳〉從文本到空間的探索，涉及到文本和空間，而文本和空間的結合點是敘事。因此，我們將首先對文本進行結構化處理，在此基礎上重新量化描述整個故事，再用統計性視角描繪這個世界（見圖 3）。描繪故事時以呈現平面寫作手法、展現空間敘事過程、復原文學空間活動為目標。在這些和人文有關的資

訊重新被構建于時空之後，我們將從空間分析的視角來觀察此世界中的場所，並借此探討作者體驗下的長安城和創作筆下的長安城的關係。

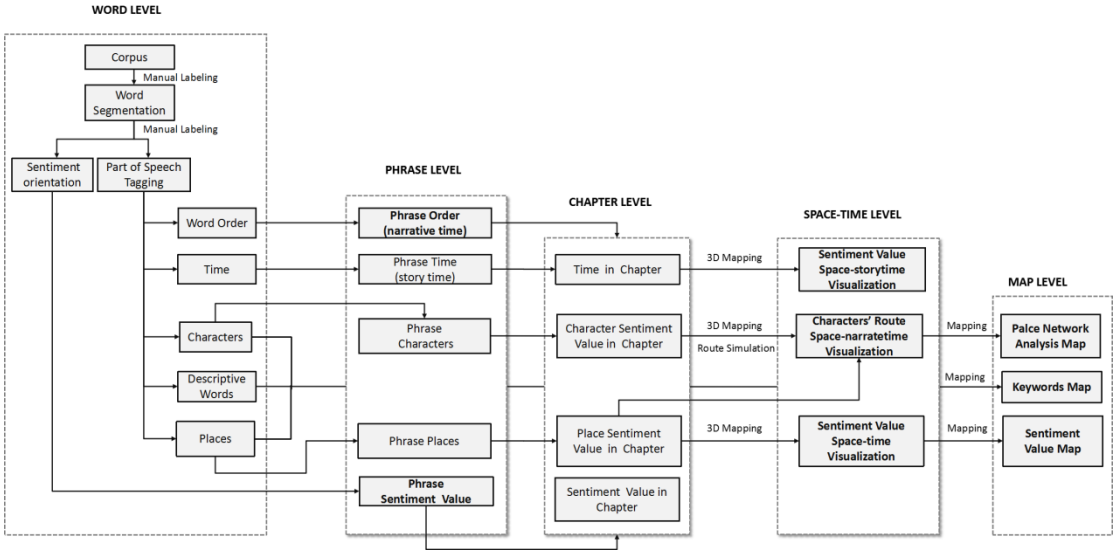


圖 3 技術流程圖

2.1 文本預處理

文本挖掘作為數位人文領域中重要的研究基礎，主要目的是將大量的非結構化文本源結構化，通過資料採擷技術發掘先前未知的、隱含的、潛在有用的資訊。但是由於中古語言詞典的缺乏、古文一詞多譯等現象以及文本量比較少，本文考慮以人工標注為主的方法建立具有“詞-短句”結構的文本資料庫。

2.1.1 詞層級

1.分詞。我們通過中國哲學書電子化計畫（CText）資料庫中獲取李娃傳的原始文本，並通過影印版的古籍進行校準，以保證原始資料的準確性。然後對文本資料進行分詞，在這裡我們首先採用北京師範大學中文資訊處理研究所開發的古代漢語語料庫線上分詞平臺對原始文本進行分詞，然後結合國學大師網站的詞典搜索庫對分詞結果進行校準。分詞的標準參考王曉玉等提出的從寬原則、詞典原則、詞義透明原則，即主張從合不從分、收錄在“詞庫”中的語義單位從合、詞

性發生轉變從合[9]。

2.標注。在分詞後，根據研究問題不同的需求對詞語進行標注，從開始最基本的文本人物資訊、極性情感值、文本進程到後來的空間資訊、時間資訊、描述資訊等。其中詞性的資訊標注主要參考詞典的解釋，並依據 GB/T 20532—2006《資訊處理用現代漢語詞類標記規範》，在此基礎上添加了部分類別（如表 1）。人物資訊的標注主要是依據小說中詞語的描述對象來確定的，在文中的人物主要涉及滎陽生、李娃、李娃母、滎陽公、李娃姨母、狎昵等人物或人群。空間資訊的標注主要是依據小說故事發生的場所，由於對於長安的地圖考古尚不完善，故在空間的資訊標注的過程中主要是以坊裡級別或街道級別來確定，所涉及的空間資訊包括布政坊、平康坊、安邑坊、宣陽坊、東市、西市、通義坊、崇仁坊、天門街、通善坊、尚書省、興慶宮等。文本進程資訊的標注，是對分詞後的詞語在整個文本中的排序序號。時間資訊的標注包括文本中的“月余”、“旬餘”、“累月”等等詳見 2.2.1.1。情感資訊的標注主要是依據極性情感的分類劃分成正面描述、中性描述和負面描述，所對應的極性情感值為 1、0、-1。描述資訊的標注主要是結合處所名詞、地名來提取形容詞。

表 1 分詞表

n	nt	nd	nl	nh	nhf	nhs	ns	nn	ni	nz
普通名詞	時間名詞	方位名詞	處所名詞	人名	姓	名	地名	族名	機構名	其他專名
no	nhh	v	vd	vl	vu	a	f	m	q	d
官名	指代性人名	動詞	趨向動詞	聯繫動詞	能願動詞	形容詞	區別詞	數詞	量詞	副詞
r	p	c	u	e	o	i	g	w		
代詞	介詞	連詞	助詞	嘆詞	擬聲	慣	語	標點		

[9] 王晓玉，李斌（2017）。〈基于 CRFs 和词典信息的中古汉语自动分词〉，《数据分析与知识发现》5，页 62-70

					詞	用語	素字	符號		

2.1.2 句層級

在短句級別的預處理上，我們主要沿用對李娃傳的句讀劃分，並對短句標注其下屬詞語的標籤。其短句標注與詞語不同的是，短句是以一個短句的屬性進行界定的，如短句情感值的確定需要對詞語情感值執行演算法。通過統計求和或者定性的方式來計算或界定短句級別的文本的資訊屬性值，形成“詞-短句”兩個文本細粒程度的文本資料結構，方便之後的各類量化重現。

2.2 敘事視覺化

在對〈李娃傳〉文本進行初步的結構化處理和統計之後，我們將著重從時空視角來還原其營造的唐長安的文學世界。這一還原的靈感以及視覺化標準的構建很大程度上都來源於西方 20 世界 60 年代末興起的的敘事學理論。敘事學主張對敘事作品進行抽象性的研究，從“結構主義敘事”或稱之為“經典敘事學”到後來的“後經典敘事學”或“新敘事理論”之後，其更加注重讀者以及社會、歷史、文化語境的作用。故事與話語、情節結構、人物和手法、敘事時間、敘事空間、敘事交流、敘述視角等成為其主要的研究物件[10]。在對〈李娃傳〉的文學世界進行解讀時，主要從故事時間、敘事時間、情節、人物、軌跡等方面進行還原。

2.2.1 文本時間下的敘事分析

2.2.1.1 文本時間和真實時間

敘事學（narratology）作為一個學科誕生之後，時間就一直是其探討的重點。敘事是一種語言行為，是線性的、時間性的，所以敘事的重現首先與時間發生關係。敘事尤其是小說具有時間的雙重性，敘事學中的「故事時間」“story time”是指所敘之事的自然發生時間，而「話語時間」“discourse time ”則是指敘事時事件被重新安排的「偽時間」。參考以上理論，在對〈李娃傳〉文本時間進行量化

[10] 申丹，王麗亞。《西方敘事學：經典與後經典》

時，我們創建了兩個欄位，一個是文本的進程時間，一個是進程的真實時間。但這兩個時間和前述兩者並不是一回事，文本的進程時間是一種讀者進入文學世界的時間，真實時間是每句話的然排序才是所謂「故事時間」，和真實時間的對比才能反映所謂「話語時間」[11]。

在對真實時間進行標注之前，我們對文本中的時間性詞（年、季、月、旬、天）語進行了重新判讀（見圖 4），並結合人物的生命歷程，確定了事件的“故事時間”：752（天寶十一）滎陽生去長安，759（乾元二年）授官，780（建中元年）李娃授封，795（貞元十一年）作文。將故事時間進行文本的重新排序之後能得到文本中每句話的“真實時間”。

我們將文本時間下和故事時間下的情節做了分別的呈現（圖 5、圖 6），可以看到兩者的明顯差別。在 y 軸同為情感值的情況下，x 軸為文本時間（單位：句）時呈現的是連續的上下波動，而 x 軸為真實時間（單位：天）時則呈現出前密後疏的狀態，說明真實時間後半段（美好結局）的敘事真實時距很大，呈切片式概述。單獨觀察圖 5 中藍線的趨勢，則能發現小說的話語時間發展，開頭時倒敘李娃受封的結局，中部場景性敘事，末端的真實時間呈斷崖式增長為跳躍式概述。

常州—>（月馀）长安布政坊—>东市、平康—>问友人—>（他日）再见李娃留宿—>（第二天）长住—>（岁馀）竹林神—>（第二天）宣阳坊—>（第二天）平康坊—>（第二天）宣阳坊、布政坊—>（三日）布政坊绝食—>（旬馀）西肆—>（后）病愈—>（日）假之—>（累月）唱歌—>东肆—>（累月）有名—>比赛被发现—>（乘机会）老竖来寻见父亲被打被救—>（第二天）活—>（月馀）抛弃—>（十旬）杖策而起—>（自秋徂冬）乞食见娃被救—>（旬馀）水陆之馔—>（未数月）肌肤稍腴—>（卒岁）愈—>（异时）问话买书—>（二岁）业大就—>（更一年）登甲科—>（其年）策名第一—>授成都府参军—>三事以降皆其友—>将之官—>（月馀）剑门—>（浹辰 12 天，或许是指从常州到剑门 12 天而确是当他）遇见父亲—>（翌日）留娃剑门—>（明日）迎亲—>（岁时伏腊）妇道甚修—>（向后数岁）持孝—>本道上闻—>宠锡加等—>（终制守孝期满）累迁清显之任—>（十年间）至数郡—>四子为大官—>（伯祖三任）与生为代—>（贞元中）为传。

圖 4 文本時間詞流

[11] 申丹，王麗亞。《西方敘事學：經典與後經典》

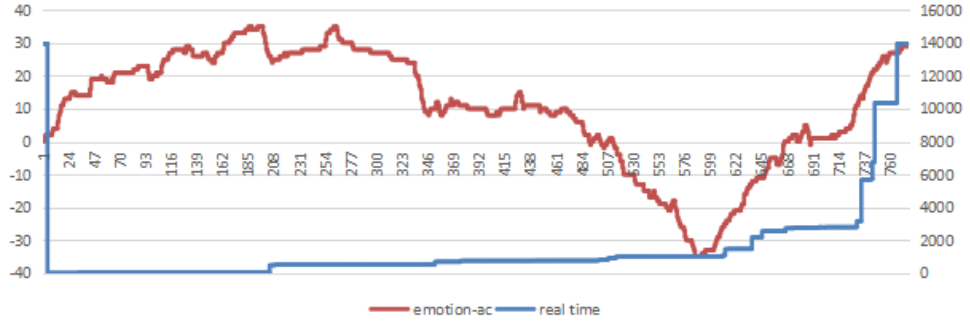


圖 5 文本時間下的情節發展與真實時間（x 軸為文本時間句，y 軸藍色為真實時間天，y 軸紅色為情感累加值）



圖 6 故事時間下的情節發展（x 為真實時間天，y 為文本的情感累加值）

2.2.1.2 全文情感

Matthew Jockers 和 Jodie Archer 所著的《暢銷書密碼》曾就小說的情感分析做了專門的研究。本研究受詞啟發，以文本短句進程作為變數，文本中每個短句累加情感極性值作為因變數，將作者所體驗到的閱讀情感進行量化分析，公式如下，得出《李娃傳》中不同情節的情感變化結構，如圖 7 所示。

$$e_j = \sum_{i=1}^{n} e_{ji}$$

e_j means the emotional value of *phrase j*, e_{ji} means the emotional value of *word i* in *phrase j*, n means the numbers of words in *phrase j*

$$E_m = \sum_{j=1}^{m} e_j$$

E_m means the cumulative emotional value of the first m phrases, e_j means the emotional value of *phrase j*

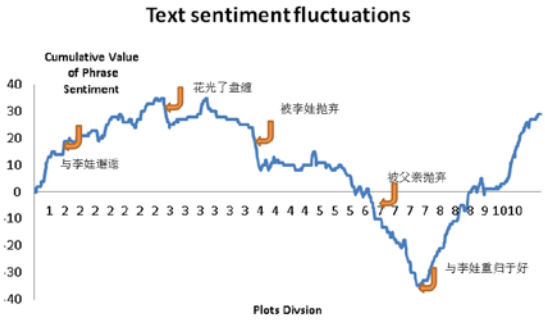


圖 7 《李娃傳》的情感曲線

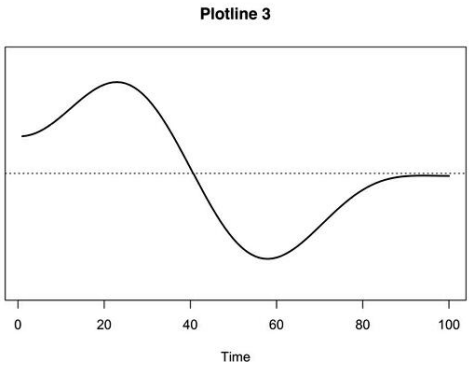


圖 8 橫 N 型情感模型

通過圖 7 可以看出《李娃傳》這篇文章的情感跌宕起伏。從變化曲線來看，更符合符合 Matthew Jockers 和 Jodie Archer 所提出的橫 N 型曲線(見圖 8)。此情感曲線在暢銷書密碼中被解讀為“走向成熟”類的小說情感模型，一般是主人公先獲得成功，後遇到挫折，最後克服困難的故事。人物情感

我們在對全文的情感進行統計後，將其涉及人物的部分單獨抽離來看（見圖 9）。就分別的情感統計值來看，滎陽生經歷了一個 N 型的情感波動，李娃經歷了一個波動上升的情感波動，符合各自的人生軌跡。從兩者之間的關係來看，前期兩者基本呈負相關關係，後期兩者基本呈正相關關係，反映出兩者的愛情曲折的真正原因。

在小說開篇介紹了滎陽生優越的家庭背景和才華、李娃的貌美，兩人邂逅後李娃與滎陽生過起了貌似甜蜜的生活，直至滎陽生耗盡了盤纏，滎陽生的情感曲線陡然下滑，而李娃的情感值處於穩定狀態。而在加入其它人物情感時故事線得以重新解讀（見圖 10）。前期李娃姥（老鴇）的情感才是和滎陽生正相關，滎陽生生出的也許只是某種意義上的單相思。而後李娃在整體的情感線上可以說是出現了斷檔，而在文本中所體現出的是著重描寫了滎陽生的兩次低谷，滎陽生耗盡家財被李娃拋棄之後，人生的軌跡被東西肆的貧民所影響獲得了新生，卻又逢滎陽公（男主角父親）打擊跌入穀底。而後滎陽生與李娃的重逢，經歷過兩次低谷

的滎陽生慘絕人寰，李娃後悔前文中對他的欺騙，而後雙方對彼此真正相愛愛走向美好結局，在情節 7 到情節 10 的情感曲線中體現出更加協同的情感趨向。

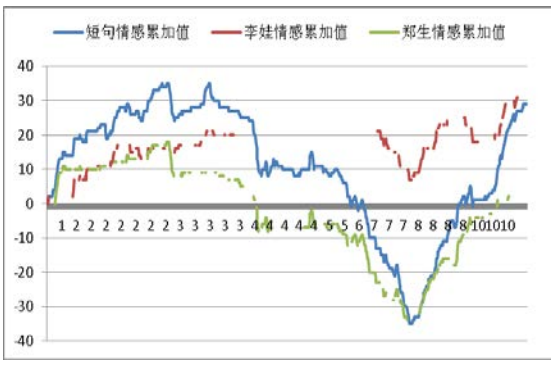


圖 9 男女主人公的情感曲線（中斷點統計）

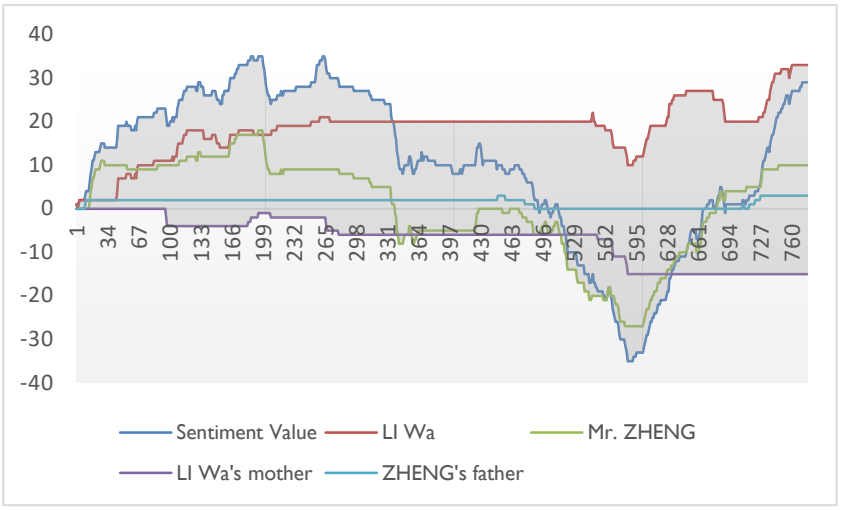


圖 10 主要人物的情感曲線（連續統計）

2.2.1.3 地點情感

地點情感曲線是對全文的情感中涉及不同地點的部分拆分開來看（見圖 11）。初步的統計結果顯示為，李娃傳中明確提及或通過考察其他文本得知的地點有 12 個，其中位於長安街東萬年縣的有平康坊、崇仁坊、東市、宣陽坊、安邑坊、通善坊；位於長安街西長安縣的有布政坊、西市、通義坊；位於皇城的有興慶宮和尚書省；街道性質的天門街。

根據地點的進程長度可以看出故事的舞臺以平康坊、安邑坊、宣陽坊為核心展開。在文中體現出了數個情節滑落時出現了地點的轉換，第一個滎陽生資材蕩

盡，舞臺從平康坊轉換到通義坊；第二個榮陽生被騙拋棄，舞臺從宣陽坊轉換到布政坊；第三個榮陽生遇父遭打，舞臺從崇仁坊轉換到曲江；這三個事件的發生都是伴隨著不同屬性地點的轉換而發生的情節。再之從地點的整體情感累加值來看，處於正面舞臺的有平康坊、宣陽坊、天門街，文中描寫的是榮陽生和李娃在平康坊邂逅並相愛，榮陽生在天門街靠感人的挽歌贏得比賽，李娃在宣陽坊見到的華麗的場景；處於負面舞臺的有布政坊、東市和通善坊，在文中描寫的是榮陽生被李娃拋棄後回到布政坊的絕望和被客棧老闆送到了西肆，榮陽生被父親毒打並拋棄的通善坊，在東市要飯的淒慘形象；安邑坊最獨特，情感總和先是負值，而後又變回了正值，在文中體現出的是李娃哭訴之前對榮陽生的欺騙，而後對於榮陽生百般呵護，支持輔助榮陽生考取了功名。

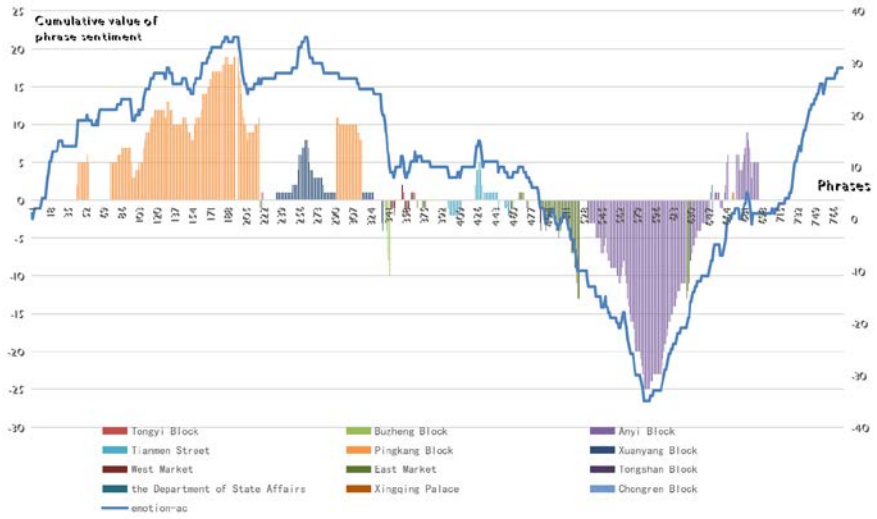


圖 11 地點情感曲線和全文情感曲線

2.2.2 三維時空製圖

基於 Arcgis 平臺，我們以陝西師範大學西北歷史環境與經濟社會發展研究院 GIS 實驗室提供的「數字歷史黃河·城市聚落資料集」的唐長安考古向量地圖為底圖，將文本挖掘得到的各類發生在長安城的屬性資訊映射到空間中以重新觀察其

空間敘事和構建的世界。

2.2.2.1　故事時間和文本時間下的敘事三維視覺化

在考慮對〈李娃傳〉敘事進行視覺化時，除了單純考慮時間、運用二維的折線圖來展現文本時間線上發展的各個物件，加入空間視角也成為展示其空間敘事最有效的方式。如圖 12、圖 13 所示，以 z 軸為時間，xy 面為長安城平面空間，我們繪製了〈李娃傳〉空間敘事 3D 圖。原本圖 5、圖 6 的 x 軸時間變數用 3D 空間的 z 軸展現，原本 y 軸上的情感值高低被用冷暖顏色展現在 z 軸的時間線上，每個時間間隙中的空間轉換用紫色的歐式距離連線來連接。兩圖的唯一不同在於 z 軸時間線上展開的情感值的變化：在自然的故事時間中情感值是從低到高的，而在給讀者重新呈現的文本時間中的情感值則是高低迴旋的，足可見〈李娃傳〉敘事的精妙性。

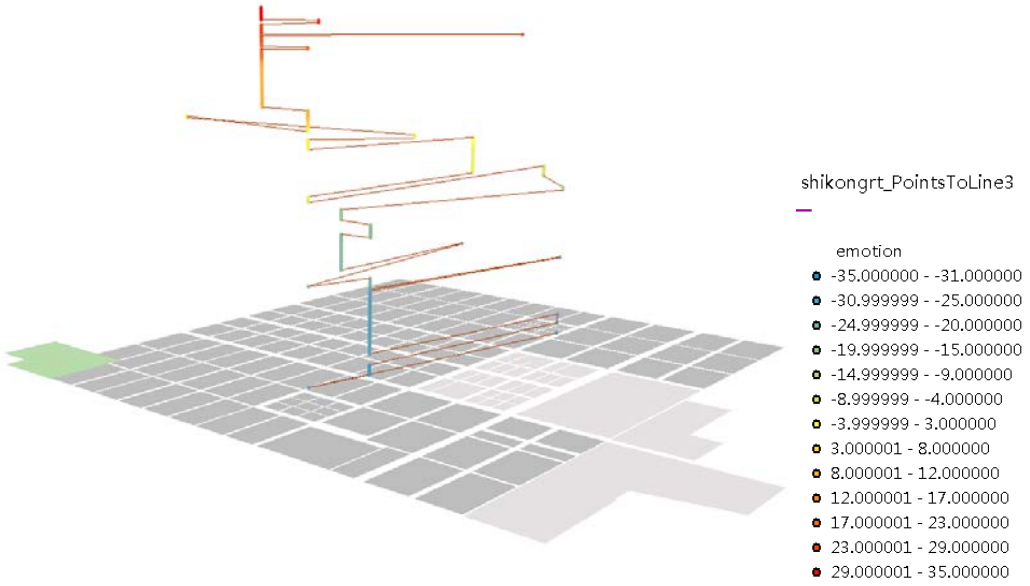


圖 12 故事時間的敘事三維視覺化
（z 軸=故事時間，z 軸顏色為情感累加變化，紫線為空間轉換連線）

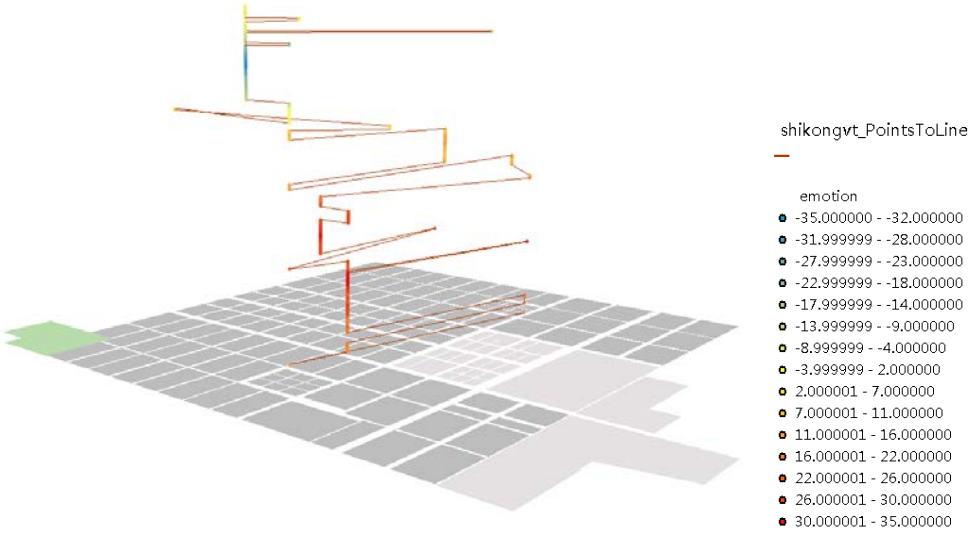


圖 13 文本時間下的敘事三維視覺化（z 軸=文本時間）

2.2.2.2　文本時間下的人物軌跡三維模擬

在空間視角觀察時，我們討論的其實不僅僅是空間敘事手法了。在某種程度上我們承認了小說天地作為一個小世界的觀點，在〈小說的時間形式和空間形式——歷史詩學概述〉中巴赫金說，“文學中已經藝術地把握了時間關係和空間關係的重要聯繫，我們將之稱為時空體。”[12]由白行簡構建的世界裡，一色人物在長安的坊裡、街和十字街裡穿行和停留。在停留空間發生著擁有大量話語的情境，然而其穿行空間卻屬於想像的縫隙。在故事世界裡這些行跡佔據著時間和空間，于長安城街道之上構建著人物的日常生活。這是只屬於作者白行簡本人的潛意識，甚至是他也不能想像完整的。而抱著對這一想像世界的探索目標，我們決定去復原人物在真實長安城中的行跡。選擇唐長安城考古地圖作為人物軌跡底圖的方法並完全準確，因為文學的故事地圖和考古地圖是有很大差別的，但這一模擬至少提供了一個可以研究話語時間縫隙的觀察平臺。

在復原人物軌跡時，我們確定了最基本也最簡單的路徑類比原則：最短路徑。但由於長安城的方正的格局，最短路徑在某種程度上處於失效狀態，我們增加了

[12] Jodie Archer, Matthew Jockers,2016 “The Bestseller Code: Anatomy of the Blockbuster Novel”.New York: St. Martin’s Press,p.141

一條原則：可達性最高。我們採用用以描述空間的拓撲、幾何、實際距離關係的空間句法去分析長安街道的整合度，得到每個街道（包含十字街）的可達性值。針對每一對起訖點“origin-destination”我們進行了路徑選擇分析，繪製了不同角色的文本時間間隙之間的 3D 行跡圖（圖 15）。

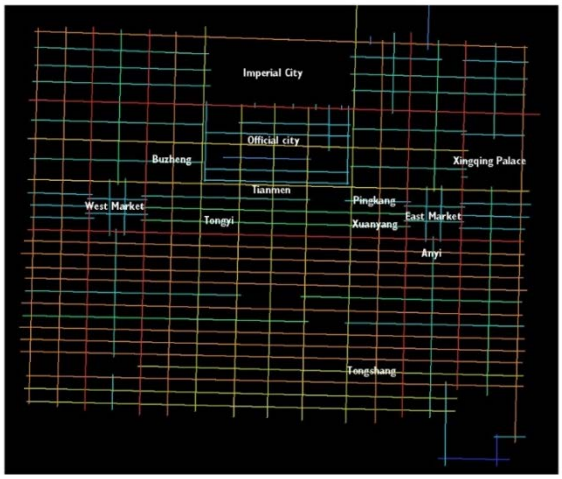


圖 14 空間句法空間街道整合度分析（顏色越暖表示可達性越高）

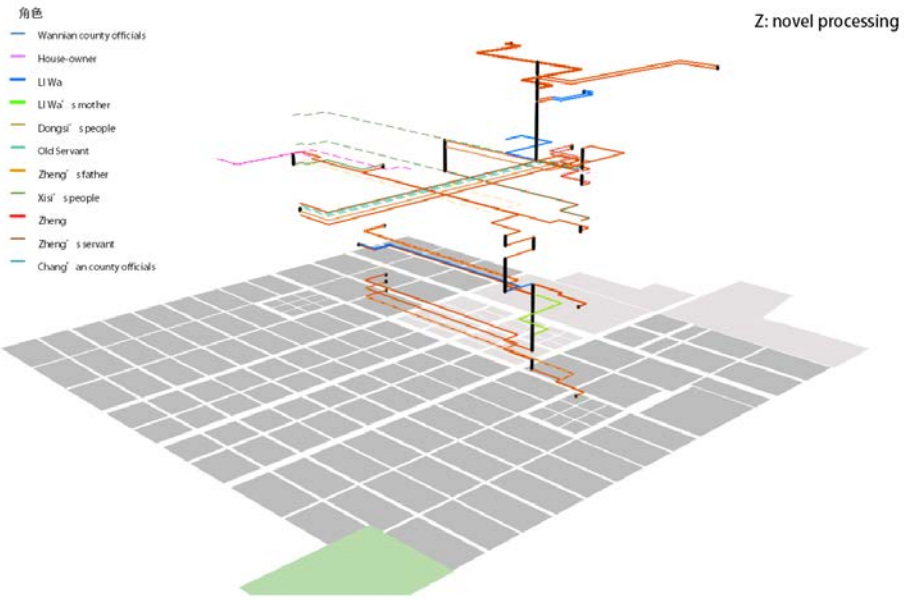


圖 15 不同角色人物的 3D 行跡（z 軸=文本時間，紅色為滎陽生軌跡）

在對人物行跡進行解讀時，我們注意到在不同街道上活動的人物的社會階級（見表 2）有所不同，並且在地點轉換之後的每個地點單位內的情感和$\boldsymbol{E}_z$的變化其實是可以伴隨著路徑而展現的，這樣的解讀和展現也許會使這種軌跡具有某些歷史和社會意義（見圖 16、圖 17），這些軌跡的解讀將在空間統計小節討論。但

是我們從圖 16 中將男主人公榮陽生提出來單獨觀察（見圖 18），事情變得容易解讀，可以看到榮陽生的在不同區域的身份轉換。

表 2 本文定義的〈李娃傳〉中角色的社會階級等級

社會階級	1	2	3	4
角色	乞丐、僕人、妓女	商人 、平民、老鴇	裡胥、秀才、賊曹、舉子	外地官員、官員

$$e_j = \sum_{i=1}^{n} e_{ji}$$

e_j means the emotional value of ***phrase j***, e_{ji} means the emotional value of ***word i*** in ***phrase j***, ***n*** means the numbers of words in ***phrase j***

$$E_z = \sum_{j=1}^{m} e_{zj}$$

E_z means the cumulative emotional value of the phrases belongs to ***stage z***, e_{zj} means the emotional value of ***phrase j*** belongs to ***stage z***, ***m*** means the number of phrases that ***stage z*** owns

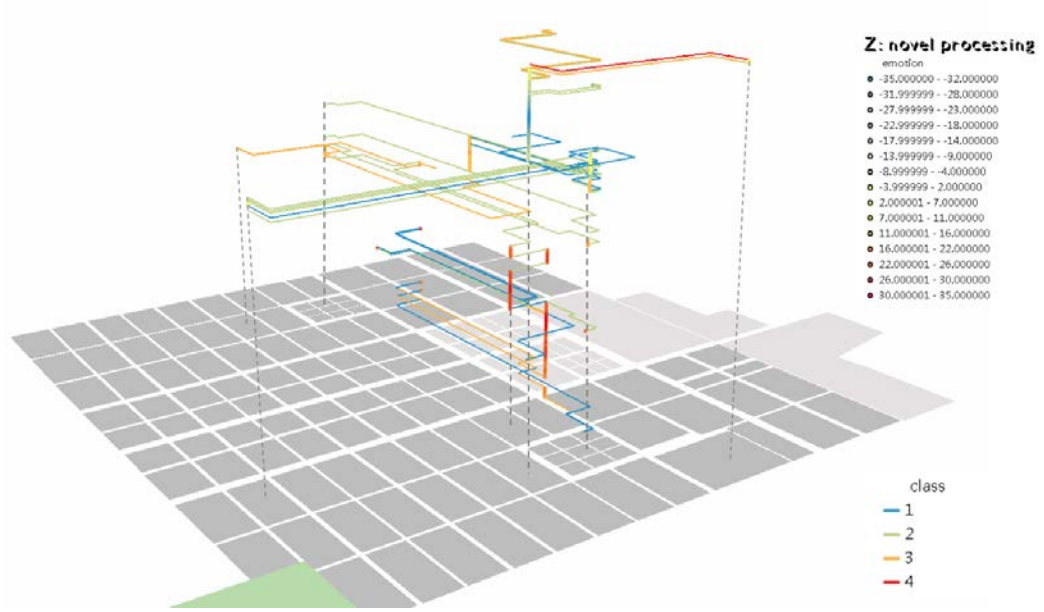


圖 16 不同社會階級的 3D 軌跡

（z 軸=文本時間，情感累計值在 z 軸展開，水平面上為 1-4 級人物軌跡）

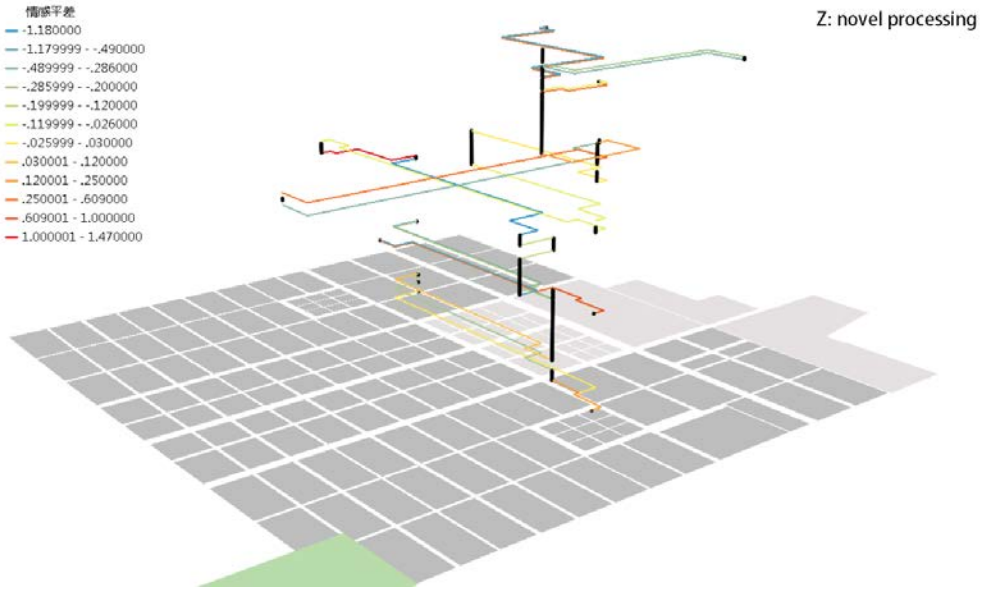


圖 17 地點之間情感變化的 3D 軌跡化
（z 軸=文本時間，水平線段為E_z的差值）

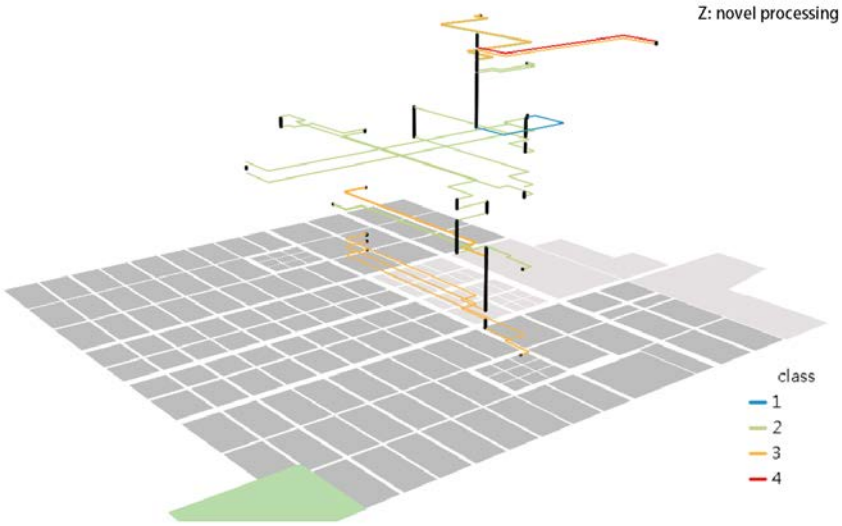


圖 18 滎陽生在長安城中變化的階級
（z 軸=文本時間，水平面上為滎陽生 1-4 級的階級變化）

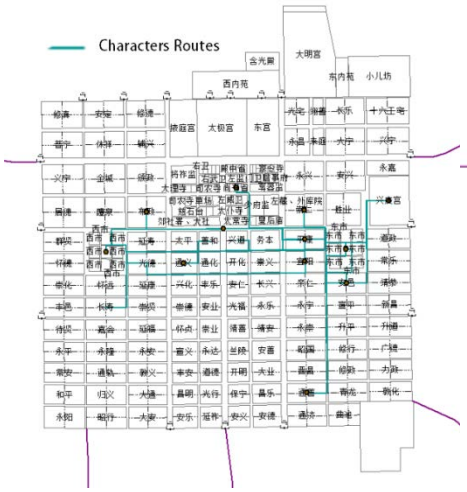
2.3 空間統計

在巴赫金的時空體“chronotopic”理論中，小說世界中的時空關係在象徵意義上和愛因斯坦的相對論類似。在這一設定下，時空體可以被用作研究文本和它所在的時代之間的關係，乃至作為廣泛的社會歷史分析的一個基礎工具。故事的人物、行動、空間是一個整體，人物的每一個行動背後有一個思想立場，和一個相

關聯的空間，這一空間的描述又和人物的內心變化息息相關。從一個空間分析統計的角度來看，在對待直接從文本挖掘或是通過一些演算法、類比而復原的世界時，對它進行時空資訊的平面化統計處理，不再關注過程而是統計角度，將能夠讓我們更加直觀的看到〈李娃傳〉中的社會。這將有助於我們理解白行簡個人化的情感和真實長安城歷史社會文化的關係。

2.3.1 模擬軌跡的統計

對人物的軌跡線的疊合統計時，我們採用了 Arcgis 軟體的空間分析包的核密度分析輔助其視覺化（見圖 19、圖 20）。該分析用於計算每個輸出柵格像元的鄰域內的線狀要素的密度，並假定每條線上方均覆蓋著一個平滑曲面，隨著與線的距離的增大此值逐漸減小，該分析常見於城市管理規劃中。在對軌跡附帶的社會屬性值包括社會階級和情感變化時進行觀察時，我們對這兩個要素賦予了更大的權重，增強效果後的圖如圖 21、22。階級圖反映出東市東側-興慶宮的社會階層最高，長壽-宣陽的社會階級較高（底層官僚），通義-平康、安邑西側、東市東北角的社會階層最低。情感變化圖反映出西市容易是容易滿足之地，而興慶宮是容易沮喪之地。



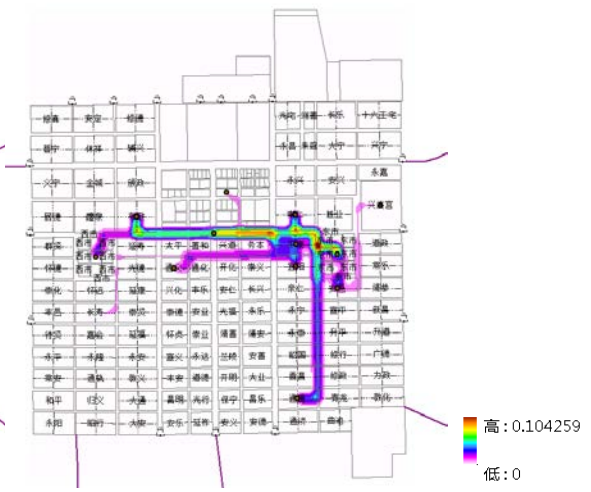


圖 19 所有人物的路徑痕　圖 20 所有人物路徑的核密度分析（顏色越紅密度越高）

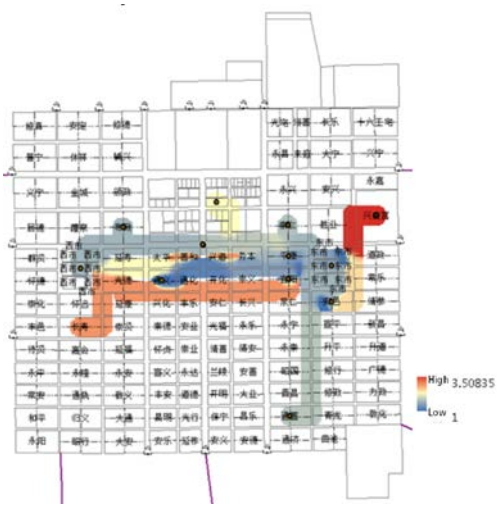


圖 21 行走於街道上不同階級

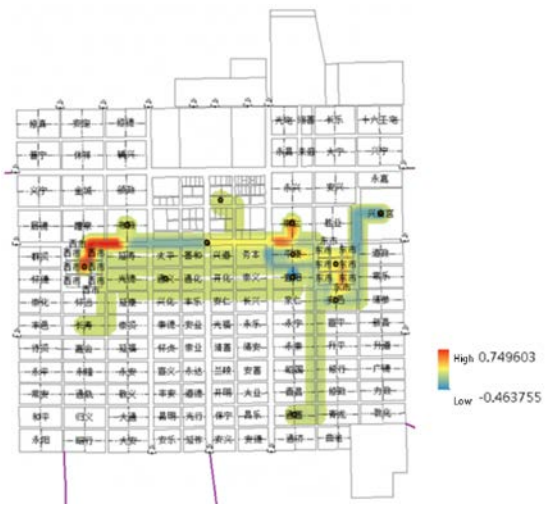


圖 22 行走於街道上的人的情感變化

2.3.2 地點網路連接分析

如果說對模擬軌跡的統計還具有相當的主觀性的話，當我們將路徑抽離，只留下節點、邊和權重，這樣的拓撲網路視角將更加有洞見性和可信性。這樣的複雜網路理論已經被證明適用於自然界和人工界，點度中心性、路徑長度、群集性、匹配性等統計量是主要的研究指標。但是另外一方面，地理網路（保留地理位置的拓撲網路）的空間距離、時間距離、感知距離也是必須被考慮在內的度量（如圖 23）。本文嘗試構建了空間節點之間的連接權重：聯繫頻次/歐式距離，我們認為它似乎可以被認為是一種感知距離。我們在 Gephi 中製作了無向圖，並就模組化“Modularity Class”、中心度“Degree”、仲介中心性“Betweenness Centrality”、接近中心性“Closeness Centrality”進行了視覺化。

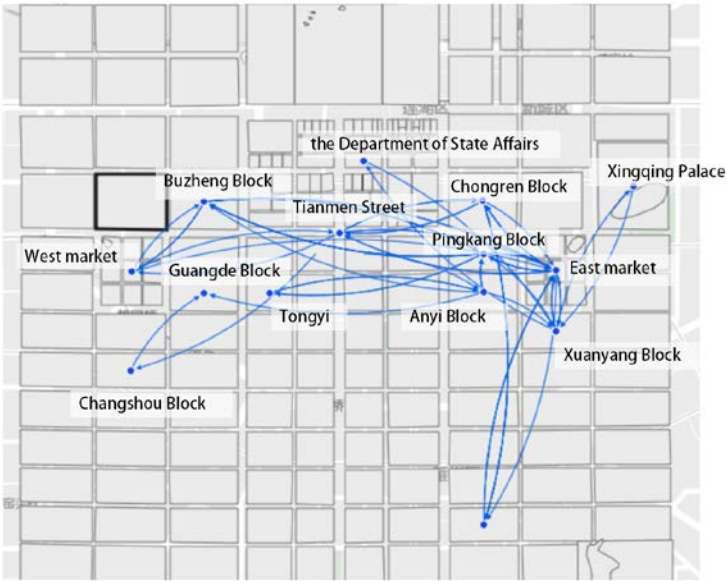


圖 23 地點之間的人物連接網路

模組化分析經常被用於社區發現中，圖 24 的四個大類顏色的觀察可以看到隱藏在文學活動背後的連接社區或者可以認為是文學城市中具有地理中心的人群組團。藍色社群的安邑坊、興慶宮、尚書省代表了權利中心，紅色社群的外地赴京人士徘徊于西市、平康坊、布政坊，綠色社群（東市、天門街、崇仁坊、通善坊）代表了外地赴京官員，橙色社群代表著本地官員集中于宣陽坊、通義坊、光德坊、興道坊、長壽坊、善和坊。

中心性是刻畫節點中心性“Centrality”的最直接度量指標，節點度越大就意味著這個節點的度中心性越高，該節點在網路中就越重要。圖 25 反映出在所有地點的度中心性：東市>>宣陽> 布政=安邑>...。東市所佔有的文本話語是比較少的，東市卻成為了文本中的隱藏「名人」，這樣的對比反而能夠讓我們觀察到在事件背後的運作邏輯。

仲介中心性是以網路中經過某個節點的最短路徑數目來刻畫，圖 26 中宣陽坊以遠高於其他地點的值成為了「傳播瓶頸」，這也許可以被解讀為中央官員對整個城市的控制力。

接近中心性反映在網路中某一節點與其他節點之間的接近程度，宣陽-安邑-平康這一組鐵成為了「八卦傳播」的鐵三角組合。

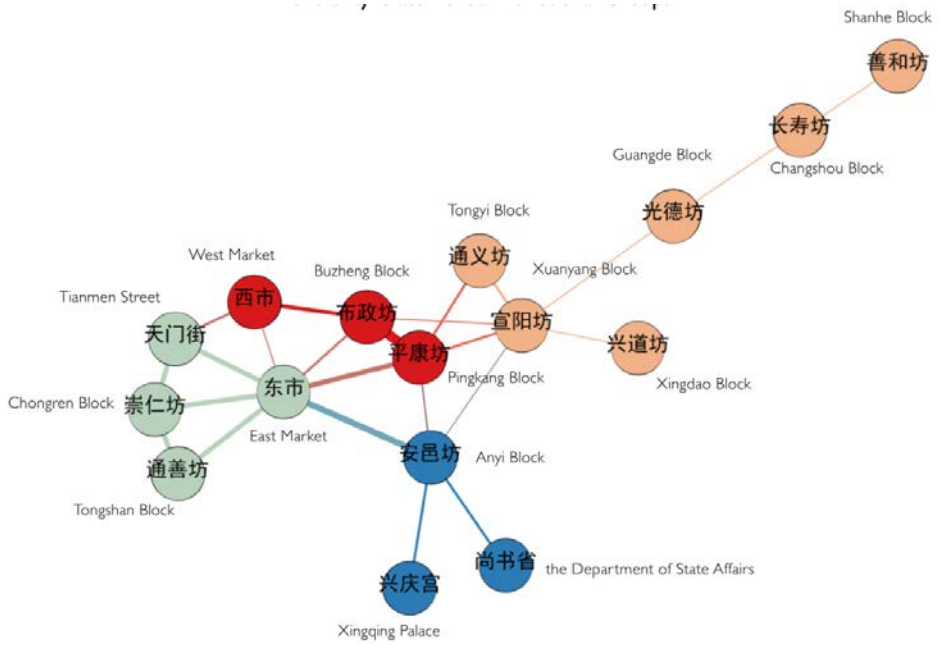


圖 24 地點連接網路 Modularity Class 分析（resolution=0.8、Q=0.322）

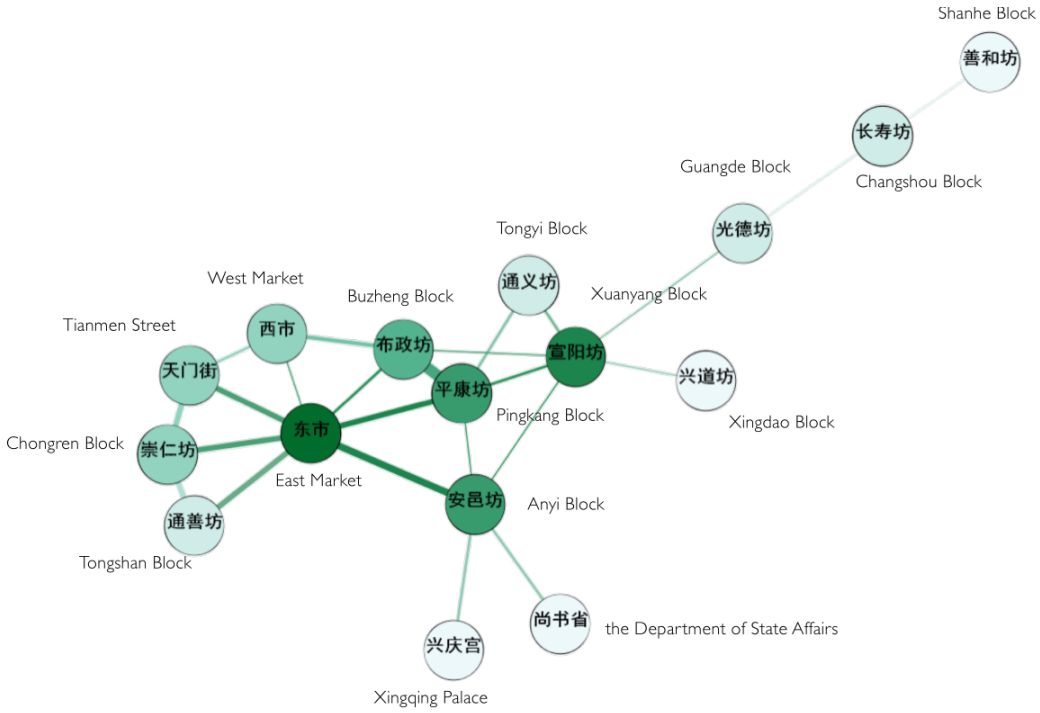


圖 25 地點連接網路 Degree 分析 （Weight=頻率/歐式距離）

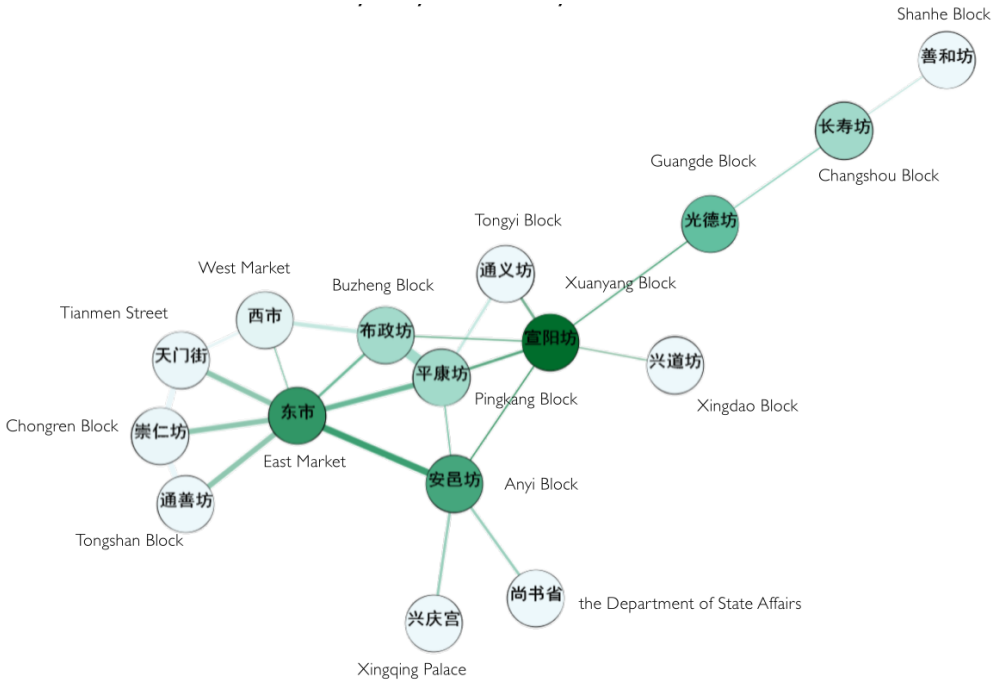


圖 26 地點連接網路 Betweenness Centrality 分析（Weight=頻率/歐式距離）

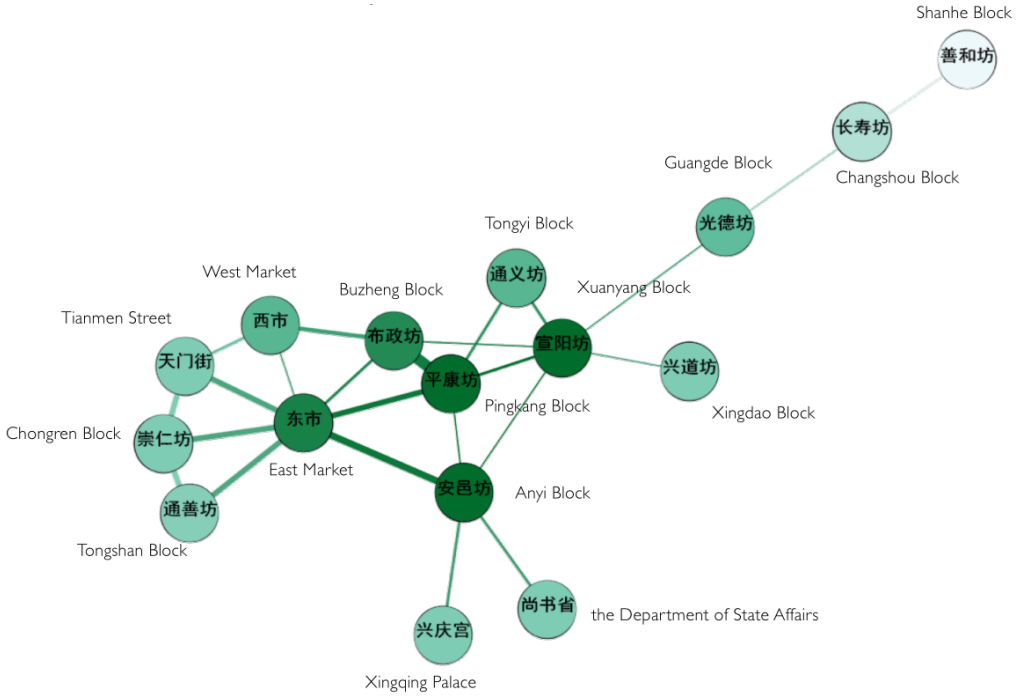


圖 27 地點連接網路 Closeness Centrality 分析（Weight=頻率/歐式距離）

2.3.3 關鍵字地圖

如上文所述，作者在對長安不同坊裡中發生的情節的描述中，在字裡行間都隱藏著作者自身對城市的認知狀態。特別是文中對城市空間的描述，某些程度上反應了當時的城市風貌。我們將作者在文本中描述的建築場所類名詞、空間環境

形容詞、環境形容和人物形容詞等詞語類別提取出來映射到長安城的不同坊裡上，可以得出場所類別主題詞地圖、環境形容類主題詞地圖、環境形容和人物形容類主題詞地圖。

依據全文中關於長安城景觀和建築名詞所繪製的主題詞地圖表現了作者筆下的長安城物質上的繁盛（見圖 28）。東區萬年縣尤其繁榮： “肆、鬻墳典、旗亭、南偏門、窟室、糞壤”等可見東市複雜的功能結構，有店面、書店、管理區、市場入口、地下室； “西戟門、偏院、榭、山亭、車門、竹樹”等反映了宣陽坊居住環境的優越和品位，有高門、偏院、亭子、水池、竹林；平康坊中“門庭、館宇、西堂、簾榻、院”等反映出平康坊的風流和舒適，有門前敞地、會客樓、會客房、簾榻、院子；“門、室、院、垣”等反映出安邑坊封閉的居住功能，有門、房間、院子、牆。西區長安縣功能呈現較為單一：布政坊“邸”意為旅舍；西市“凶肆、兇器”意為殯儀館，通義坊“祠宇”意為寺廟。總體來看，白行簡對於東區的萬年縣描繪較豐富，而對於西區的長安縣只有寥寥。不同地點的描繪詞地圖（見圖 29）中的東區萬年縣呈現出物質的富足。地圖中對萬年縣描繪較多，如平康坊 “嚴邃、僻陋、煥然奪目、侈麗”等主題詞表現出了一個精心佈置的小空間居住環境；宣陽坊的“蔥蒨、幽絕、珍奇、弘敞”等主題詞表現出的是一個高檔華貴的大空間居住環境。這些主題詞直觀體現了文中人物在不同地點的不同生活品質，也反映了萬年縣不同居住區的差別。將文中不同地點與人物的描述詞映射到長安地圖上（見圖 30），可以看出只有殯儀館所在西市的人物精神狀態最好：“無與倫比、聰敏、妙”等。擁有精緻小空間的平康坊中大多是正面的精神描述，而東市、宣陽坊、安邑坊、天門街、布政坊在描述美麗環境的同時有大量關於人物的負面描述。平康坊的鮮衣怒馬，宣陽坊的惶惑，布政坊的憤懣，

西市的得意，東市的低賤，天門街的恥辱，安邑坊的康復，這些人物與環境的描述詞反映出不同環境下的人物的生活狀態。這些環境的表徵在很大程度上是為了烘托人物經歷而設置，但不能否認這一設置反映出作者對於真實長安城市空間的一種認知狀態或者是場所帶來的反應，當然作者的生活經歷也會影響到這個世界的呈現。也許可以做出合理的猜想，作者住在街東，對於萬年縣的瞭解更勝於長安縣，萬年縣描繪時更加詳細和充滿情緒，而對長安縣則充滿了美好想像。



圖 28 場所類主題詞地圖

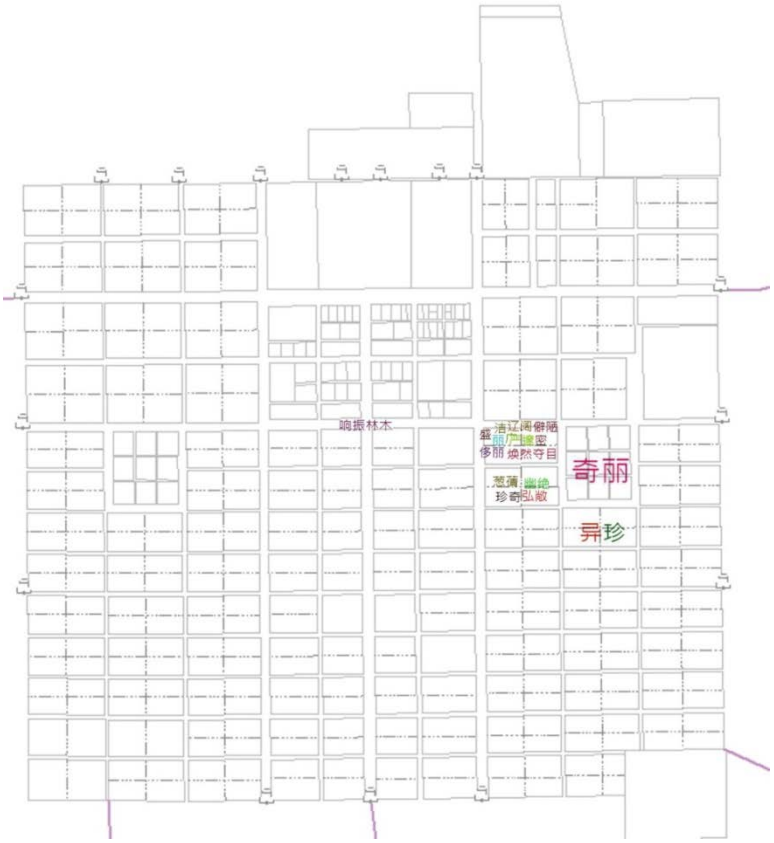


圖 29 環境形容類主題詞地圖



圖 30 環境形容和人物形容類主題詞地圖

2.3.4 情感值地圖

軌跡三維模擬章節中已經闡釋了每個地點的情感和統計，這個篇幅我們基於地點的短句情感和使用克裡金插值的方法將其映射到長安城市歷史地圖上，以空間的視角審視城市的情感。圖 31 可以看出，街東萬年縣的情感值明顯偏高於街西長安縣。這一地圖更傾向于反映一種讀者視角下的不斷累積的長安認知：街東萬年縣是美好生活的場所，在文中體現的是邂逅、中舉的美好情節等；而街西長安縣是底層階級的去處，在文中體現的是失魂落魄回到以前的客棧，而客棧老闆將他拋棄於西肆等情節。

而如果採用每個地點的情感平均值$\bar{E}_z$來做克裡金插值分析，則會產生截然不同的地圖。因為在做了平均之後，這個值更體現出作者視角下對每個坊裡的情感認知，將情感值與文本中涉及的短句數相比，得出了的是作者的書寫話語時帶出的情感。

$$\bar{E}_z = \frac{E_z}{m}$$

$\bar{E}_z$ means the emotional value of average phrase in ***stage z***, $\boldsymbol{E_z}$ means the cumulative emotional value of the phrases belongs to ***stage z***, ***m*** means the number of phrases that ***stage z*** owns

文中關於城市描述的情感傾向狀態基本分成了六個區域（見圖 32）。以官城為中心的區域、以通義坊為中心的宗教區域和以興慶宮為中心的宮城，這個三角呈現出正面的情感，反映了對於科舉仕途的渴望、對神明的尊敬以及對於當時皇權空間的尊崇，這也許正是構建小說情節邏輯的核心；在布政坊、崇仁坊、曲江的三角呈現出了明顯負面的情感，和作者對人物遭遇的設定有很大關係，這也許也和作者的經歷有關；在平康坊、宣陽坊、東市、安邑坊、西市和天門街的地點情感比較溫和。

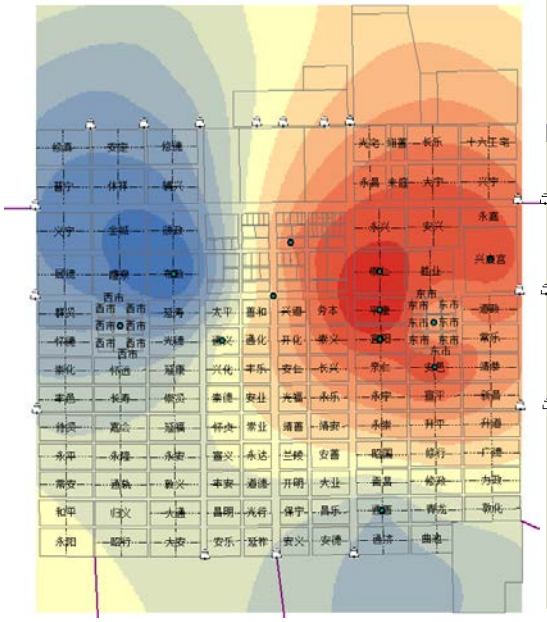


圖 31 短句情感和地圖

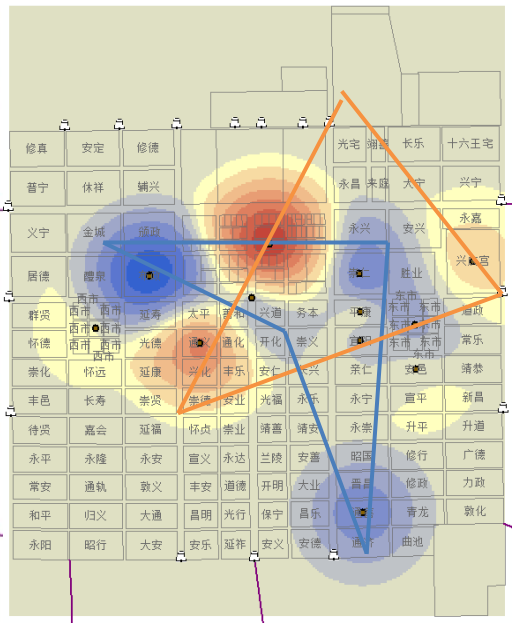


圖 32 平均情感地圖

3 討論

在量化探究〈李娃傳〉的話語、故事和故事世界之後，我們發現了很多現象並對其資料進行了初步的解讀。在針對話語、故事的敘事視覺化時，這些現象和解釋也許是完全個性化的獨屬於〈李娃傳〉的現象；而在對故事世界的評價的時候，因為融合了真實城市的體驗，這些現象也許就具有了一定的共通地可比性。下面將分別對敘事視覺化和空間統計進行探討。

敘事視覺化中大部分的探討屬於文本結構的探討。從文本時間與真實時間的對比來看小說的故事時間安排，可以看到前密後疏以及開頭倒敘的寫法，這其實也是唐傳奇的常用敘事手法，即在開頭或結尾表明作者身份，用第一人稱言明故事來歷，結尾虛化，也有學者探討過〈李娃傳〉的無聲結局蘊含的禮教意義[13]。全文情感曲線實際上在《唐代傳奇小說論》有所提及，小南一郎認為小說前半部分“下降”，而在後半部分“上升”，形成了一種 V 字型的展開方式，但忽略了更前段中對滎陽生以及李娃的正面描述。在前人對李娃傳的研究中，普遍認為滎陽生的兩個低谷是被李娃拋棄和被父親拋棄，而情感曲線在其被李娃拋棄前的下

[13] 李云飞，王引萍，〈论〈李娃传〉的“无声”结局及李娃形象，《牡丹江教育学院学报》2，页 1-2

降早已暗示出在平康坊花光盤纏就是其走向低谷的開始。人物情感中能夠看到的是滎陽生的人生深究起來其實被政治（父）、世俗（姥）、愛情（娃）所影響，這一點在前人研究人物關係時並沒有能夠被拆解的這麼清晰。〈李娃傳〉人物情感線中間大段李娃的缺失和最後的突然出現，表現出的是一種男性視角下的傳記寫法。地點情感曲線中，妹尾達彥提出的地點對情節的推動作用也表現的異常明顯。空間敘事視覺化可以看到被精心安排的情節是怎樣在空間和話語時間裡展開的。

對故事世界的評價建立在將其視為一個世界的基礎之上進行觀察分析。人物軌跡統計從一種街道統計視角來看也許會發生在長安城的日常生活和形色人等。地點網路連接分析從拓撲結構的角度挖掘出四類社群、東市「名人」、宣陽坊「傳播瓶頸」、宣陽-安邑-平康「八卦傳播」的隱含資訊。關鍵字地圖注重對環境的直觀反映，東區安邑坊、平康坊、宣陽坊這三個高檔居住區的細微差別被察覺，作者對東區的瞭解被洞悉。情感地圖的不同地塊的情感傾向呈現兩對倒三角的平衡狀態，這也許就是當初故事世界的最初設置，這一設置與當時的“三個中心，一個聯絡點”的政治格局有相似之處的（見圖 33）[14]。這些從空間層的深度挖掘已經脫離了文本而進入世界，體現出了作者潛在的城市感知。

從〈李娃傳〉文本的選擇、到文本處理和空間處理，過程中存在偶然性也存在共通性。我們已經展開了對更多唐傳奇文本的研究，驗證了其中很多方法的普適性，但因文學是一個涉及個人、社會與環境的問題，更多的探討將值得期待。

從研究方法上來說，將定量分析的方法在用中國古代文學作品上，將時間和空間融入，從時間線上觀察作者的創造文學世界的手法、文學世界的場景活動，

[14] 李永，〈從州邸到進奏院:唐代長安城政治格局的變化〉，《南都學壇》2，頁 31-34

最後回歸空間使得時間線被壓縮在平面內觀察感知語境中的歷史城市，這一視角是非常具有開創性的。

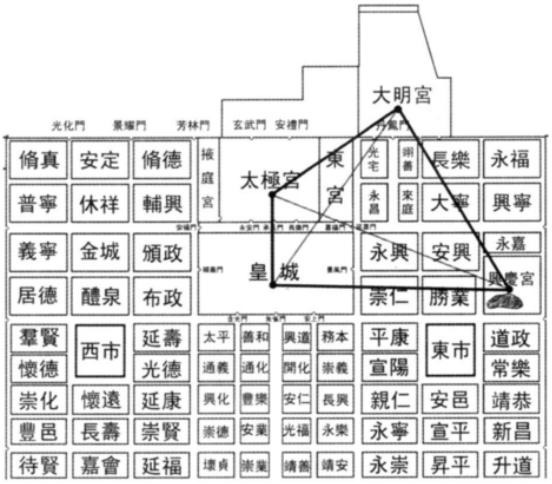


圖 33 唐代長安城政治空間示意圖

致謝

本研究得到了本研究得到了“高等學校學科創新引智計畫”（項目編號 B13011）及國家自然科學基金（專案編號 51478300）的支持。

参考文献

李效傑,張紅雲（2017）。〈從〈李娃傳〉看唐代的商業競爭〉, 《山東工商學院學報》6,頁 1-5+36

王歡（2007）。〈李娃傳〉中的房舍租賃，《河南農業》 2，頁 56-58

申丹，王麗亞（2010）。《西方敘事學：經典與後經典》。北京：北京大學出版社

Barabási, Albert-László. 2002. Linked: The New Science of Networks.

劉勇強，潘建國，李鵬飛（2017）。《古代小說研究十大問題》。北京：北京大學出版社

小南一郎（著），童嶺（譯）（2015）。《唐代傳奇小說論》。北京：北京大學出版社

張同利（2015）。《長安與唐五代小說研究》。北京：人民出版社

曾大興（2012）。《文學地理學研究》。北京：商務印書館

程國斌（1997）。《唐代小說嬗變研究》。廣東：廣東人民出版社

王曉玉，李斌（2017）。〈基於 CRFs 和詞典資訊的中古漢語自動分詞〉，《資料分析與知識發現》5，頁 62-70

妹尾達彥（1995）。〈唐代後期的長安與傳奇小說——以〈李娃傳〉的分析為中心〉，《日本青年學者論中國史·六朝隋唐卷》，頁 509-553

程國斌（1993）。〈〈李娃傳〈嬗變研究〉，《南京大學學報（哲學社會科學版）》3，頁 111-117

Jodie Archer, Matthew Jockers,2016 “The Bestseller Code: Anatomy of the Blockbuster Novel”.New York: St. Martin’s Press,p.141

李双红（2007）。〈浅谈唐传奇〈李娃传〉的故事情节〉，《三峡大学学报（人文社会科学版）》S2，页 109-110

朱明秋（2013）。〈〈李娃传〉情节数理批评〉，《桂林师范高等专科学校学报》3，82-84+89

吴淑钿（2006）。〈门之内外:李娃故事的叙述结构〉，《古籍研究》2,194-201

朱玉麒（2003）。〈隋唐文学人物与长安坊里空间〉，荣新江主编，《唐研究》第九卷。北京：北京大学出版社

李云飞，王引萍（2013）。〈论〈李娃传〉的“无声”结局及李娃形象，《牡丹江教育学院学报》2，页 1-2

李永，〈從州邸到進奏院:唐代長安城政治格局的變化〉，《南都學壇》2，頁 31-34

歷史地理信息系統與空間人文研究

構建城市歷史地理學研究的時空 GIS 基礎框架

陳剛

南京大學地理與海洋科學學院

歷史地理信息系統與空間人文研究
——構建城市歷史地理學研究的時空 GIS 基礎框架

陳　剛 *

（南京大學地理與海洋科學學院，210023）

摘　要

本文立足於新興的「空間人文」研究領域，探討其學術緣起及其在數位人文、歷史信息系統發展脈絡的中地位與意義；進而指出發展空間人文研究，需要研製新一代的歷史地理信息系統——時空 GIS 基礎框架。接著，探討時空 GIS 基礎框架的定義、功能及其學術意義。再次，面向城市歷史地理學研究，以六朝建康歷史地理信息化研究為例，提出構建城市歷史地理時空 GIS 基礎框架的理論方法與關鍵技術。本研究的學術願景是：藉由時空 GIS 基礎框架的支持，六朝建康歷史地理信息化研究，將可以有效集成與整合歷史文獻、考古資料與野外考察等歷史地理學研究內容與方法，幫助學界開展多層次、多維度的綜合研究，重建六朝建康歷史地理圖景。

目　次

1. 引言 1
2. 新一代歷史 GIS 平臺——時空 GIS 基礎框架 2
3. 時空 GIS 基礎框架的定義、功能與挑戰 3
4. 構建城市歷史地理時空 GIS 基礎框架 5
4.1. 工作基礎 6
4.2. 六朝建康時空 GIS 基礎框架設計 9
4.2.1. 功能與特徵 10
4.2.2. 統一時空基準 10
4.2.3. 實證案例：六朝人物關係時空數據建庫研究 11
5. 結語 13
參考文獻 13

關鍵詞

空間人文，歷史地理信息系統，城市歷史地理學，六朝，南京

* 南京大學地理與海洋科學學院副教授、博士。

Historical GIS and Spatial Humanities —— Constructing a Spatial-temporal GIS Infrastructure for Urban Historical Geographical Studies

Gang Chen*

（School of Geographic and Oceanographic Sciences，Nanjing University，210023）

Abstract

Focusing on the emerging research field of "Spatial Humanities", this paper discusses its academic origin and its status and significance in the development of Digital Humanities and Historical GIS (HGIS), and further points out that the development of Spatial Humanities research requires the development of a new generation of HGIS —— Spatial-temporal GIS Infrastructure (STGI). Next, we discuss the definition, function and academic significance of STGI. Thirdly, facing the research of urban historical geography, taking the informationization of Jiankang historical geography in the Six Dynasties as an example, the paper puts forward the theoretical method and key technology of constructing the basic framework of urban historical geography spatial-temporal GIS. Thirdly, for the study of urban historical geography, taking the historical geography informationization of Jiankang in the Six Dynasties (Liuchao-Jiankang) as an example, the theoretical methods and key technologies for constructing the STGI of urban HGIS are proposed. The academic vision of this study is as follows: with the support of the STGI, the historical geographic informationization of Liuchao-Jiankang will effectively integrate and integrate historical research contents and methods, such as historical documents, archaeological materials and field investigations, and help the academic community to carry out multi-level and multi-dimensional comprehensive research to reconstruct the historical and geographical landscape of Jiankang city in the Six Dynasties.

Key Words: Spatial Humanities, Historical GIS, Urban Historical Geography, Six Dynasties, Nanjing

* Associate Professor，School of Geographic and Oceanographic Sciences，Nanjing University.

1. 引言

近年來，隨著信息技術的飛速發展、「數位人文」（Digital Humanities）研究領域及理念的不斷深入，在中國歷史地理學及信息化研究領域也促發著重要進步。根據筆者的觀察與思考，大體分為以下幾方面。其一，湧現出新型研究成果，產生了一批具有重要學術影響力、普惠學林的歷史 GIS 基礎數據庫及信息系統。[1] 其二，推動了學科發展，產生出新的學科增長點，壯大了學科研究隊伍。[2] 其三，革新了研究方法與研究理念。應用 GIS 技術，建設歷史地理數據庫、編制電子地圖集和研製歷史地理信息系統，已成為歷史地理學研究不可或缺的新技術手段。[3] 在大數據時代，歷史地理學又不斷審度研究方法與理論創新，注重與國際學術的對話與接軌；借助地理信息系統（GIS）、文本分析（Text Analysis）、社會網路分析（Social Network Analysis）等技術，逐步融入「數位人文」發展潮流，進一步促進了跨學科合作與國際學術交流，成為歷史地理學研究方法創新的主要趨勢。

另一方面，伴隨人文社會學科領域興起的「空間轉向」（Spatial Turn）研究思潮，如在歷史學領域，Knowles（2000）首先闡述了在歷史學研究中「空間轉向」的意義與歷史地理信息系統的應用。包弼德（Peter K. Bol，2013）指出歷史學已進入「空間轉向」時期，「空間」成為理解歷史過往的現代思考（modern thinking about the past）中的核心詞彙。王汎森（2014）指出：「在 GIS 工具的幫助下，歷史研究者可以從空間的角度去思考問題」。同時，地理信息科學領域的學者也開始把自己的研究重心從資源環境等傳統研究領域，逐步向人文社科研究方向延伸，通過多學科合作與交叉，逐步形成了以地理信息技術為支撐的「空間綜合人文社會科學」（Spatially Integrated Humanities and Social Science, SIHSS）跨學科研究領域。[4] 這一趨勢，或可稱為「數位人文學」（Spatial

1 在中國歷史地理學研究領域，GIS 技術的應用大約開始於 1990 年代後期，以譚其驤主編《中國歷史地圖集》為基礎，學界集中力量開始研建中國歷史基礎地理信息庫及信息系統。復旦大學與哈佛大學等機構於 2001 年啟動「中國歷史地理信息系統」（CHGIS）（http://yugong.fudan.edu.cn/）；中央研究院 GIS 團隊則在 1997 年開始研製「中華文明之時空基礎架構」（CCTS）（http:// ccts.ascc.net/），取得了重要成就，也奠定了中國歷史地理信息系統建設的理論、方法與應用基礎。2003 年召開的「21 世紀中國歷史地理學暨兩浙歷史時期人地關係」研討會上，中國大陸歷史地理學界已形成「將地理信息系統引入歷史地理學已勢在必行」的共識。近十年來，GIS 技術在歷史地理學領域取得快速發展，積極探索與研發「絲綢之路歷史地理信息系統」、「清史 GIS」等專題型或區域性歷史 GIS。

2 從 2015 年開始，中國歷史地理專業委員會連續舉辦四屆中國歷史地理信息系統（HGIS）學術沙龍，2017 年組建歷史地理信息科學與技術工作分委會，通過了跨學科合作與交流，逐步吸引地理信息科學、測繪、地理學、電腦科學等相關領域的專家投入到歷史地理學研究中來。

3 2018 年中國歷史地理學術研討會（北京，8 月 10-12 日）的主題是「新時代、新技術、新思維」，會議主題之一就是「新技術手段在歷史地理研究中的應用（「數位人文」）」，也成為會議討論的熱點，參見：強慧婷 ，〈新技術助力歷史地理學發展〉，《中國社會科學報》，2018 年 8 月 29 日。其實，在近年來兩屆國際歷史地理學者大會（International Conference of Historical Geographers, ICHG）（2015，倫敦；2018，華沙）上，歷史 GIS 研究佔據著重要地位。關於 GIS 在歷史地理學研究中的應用，可參：陳剛，2014，〈「數字人文」與歷史地理信息化研究〉，《南京社會科學》，第 3 期，頁 136-142。

4 以國際華人地理信息科學協會（CPGIS）、中國區域科學協會（RSAC）為主體，自 2009 年所發起的「空間綜合人文學與社會科學論壇」，已連續舉辦 9 次，吸引了地理、歷史、考古、人類學、社會、城市領域的重要學者。比如，第九屆空間綜合人文學與社會科學國際論壇（上海，2018）的主題是探討地理信息系統在人文科學與社會科學領域的應用研究，探討建立空間綜合人文學與社會科學的理論、方法與平臺體系，分享促進人文社會科學與計算科學、地理科學融合創新，發展新的社會智慧科學的思想與共識。

Humanities）的興起，成為 「數位人文」的重要組成部分。

近年來，臺灣中研院範毅軍研究員所領導的研究團隊在「地理資訊數位元典藏與空間人文學發展計畫」（ https://ascdc.sinica.edu.tw/）中就明確提出，「將持續運用既有地理資訊數位典藏與技術發展成果來發展空間人文學（Spatial Humanities）」。在 2017 年的多場學術報告中，範毅軍研究員闡明籍由歷史地理信息系統（HGIS）建設和推進數字人文領域的廣泛合作，研建虛擬時空框架，進而發展空間人文，「重回歷史現場」的學術倡議與構想；提出“Geo-Humanities（Spatial Humanities）= HGIS ＋ Digital Humanities”的基本看法。[5] 簡而言之，以筆者看來，「空間人文」作為「數位人文」的重要學術分支，重視「空間研究」，[6] 並通過 GIS 及其他數位工具的支援，促進跨學科整合研究；在技術層面，「空間人文」的核心是新一代歷史地理信息系統，需要集成多源時空數據，構建時空 GIS 基礎框架（平臺）。[7]

2. 新一代歷史 GIS 平臺——時空 GIS 基礎框架

歷史地理信息系統（Historical GIS，HGIS）是空間人文學研究的主要技術手段；它是將地理信息技術 [8]用於歷史地理學研究，包括建設歷史地理數據庫、編繪數位地圖（Digital Map）和研製專題信息系統。HGIS 不僅能將歷史上傳統的地理要素表達方法（紙質地圖）與內容，轉移到以現代地理坐標系統為參照系的數字地圖上，方便歷史地理文獻的數位化管理及信息挖掘。同時，歷史地理信息系統還可以把傳統地圖的表現手法與電腦自動製圖、數據庫管理與信息查詢等現代化手段緊密地結合起來，進而開展空間分析與信息挖掘。在現代歷史地理學研究中運用地理信息技術，構建時空 GIS 基礎框架（平臺），不僅可以輔助開展歷史地理學研究，也拓寬歷史地理學的研究方向、擴展學術視野、豐富成果形式和推進多學科合作研究。[9]

[5] 近年來，筆者與范毅軍研究員有著多次學術交流與合作，獲得極大幫助與教益。2017 年 4 月 7-9 日，在廣州舉行的第三屆全國歷史地理信息系統學術沙龍上，范先生作題為「地理資訊與空間人文——構建虛擬時空框架的設想、具體實踐與應用」的主題報告，提出虛擬時空框架的設想，展示了中研院正在研製與完善的時空資訊整合平臺——SinicaView 平臺，並提出發展「空間人文」（Geo-Humanities），推動跨學科整合研究。2017 年 11 月 30 日，范先生在南京大學做題為「從時空 GIS 到空間人文學—構建虛擬時空框架的設想、具體實踐與應用」的學術講演，明確提出發展空間人文的學術構想。之後，在南京大學舉行的「空間人文（Geo-Humanities）跨學科交流會」上，邀請來自歷史、考古、地理、GIS、測繪、信息管理、文化遺產、城市規劃等領域的學者，也就「空間人文」主題展開熱烈討論。

[6] 此處的「空間研究」，不僅指「從空間的角度去思考問題」，也包括「在數位工具的支持下，空間本身便成為值得思考的課題」。參見：王汎森，2014，〈數位人文學之可能性及限制——一個歷史學者的觀察〉，收于項潔編《數位人文研究與技藝》，臺北：台大出版中心，頁 25-35。

7 事實上，GeoHumanities 一詞最早出現在 2011 年出版的英文專著 *GeoHumanities: Art, history, text at the edge of place* 中，通過 30 篇文章，集中反映地理學與人文學科相結合所形成的跨學科、新興知識領域，這一概念的範疇及所指，不同于范毅軍先生所宣導的空間人文研究。見：Dear, M. J. (2011). *GeoHumanities: Art, history, text at the edge of place*. London: Routledge.

8 地理信息技術（Geographic Information Technologies，簡稱 GIT）是以電腦技術為基礎，以具有空間資料為處理對象，運用系統工程和信息科學的理論、方法，採集、存儲、管理、顯示、處理、分析和輸出地理信息及其產品的現代綜合技術體系。地理信息技術包括地理信息系統（GIS）、遙感（RS）、全球衛星導航系統（GNSS）等多種信息技術。一般地，學界多採用 GIS 這一術語，本文中地理信息技術與 GIS 並用，但所指略有不同。

9 法國學者安克強（Christi an Henriot）對於 GIS 的觀點頗值得重視。他說：「地理信息系統技術是一個提問的工具，不能代替學問本身。…GIS 在傳統資料、資料和空間之間建立了聯繫，幫助我們發掘尚待研究的問題。」參見：任思

當前，HGIS 的發展方興未艾，國內外學界圍繞各類研究課題，已構建了大量歷史 GIS 數據庫與專題信息系統。但也存在一定的局限與不足，特別是在數據庫建設方面，大略可表現為以下幾個方面：（1）目前圍繞專題研究所形成的眾多歷史地理數據集或數據庫，大多缺乏統一規劃，在建設上因陋就簡，規模大小也不一，存在一定的重複建設問題；（2）在數據庫設計方面往往存在不足，特別是缺乏統一的時空數據架構設計；數據結構簡單（很多就是 excel 表格數據）、數據冗餘大，數據錯漏多，數據更新與維護難；（3）忽視數據規範性與標準化設計，很難開展數據交換與數據共用，形成「數據孤島」；（4）大多數據庫隨著軟硬體環境的變化，很難進一步升級與維護；（5）一些數據庫的生命週期短，開放度不夠、利用率不高，特別是隨著項目任務結束，缺乏後期維護與升級，成為無人問津的死庫。由於當前歷史地理數據庫建設還存在一定不足，基於數據庫所建設歷史地理信息系統，就成為歷史地理信息化的瓶頸，更遑論更好地開展歷史地理空間分析、知識挖掘和專題製圖研究。

事實上，新一代的 HGIS 發展需要借助新的信息技術，構建基於統一時空基準的歷史地理基礎信息平臺（即：時空 GIS 基礎框架），集成多源、多媒體歷史地理數據，提供更易用的使用者介面及更豐富的數位工具，以助力開展定性與定量的綜合集成研究。同時，還需積極借鏡與汲取相關學科的理論與方法，面向具體科學問題與應用實踐，強調團隊合作，重點解決如下問題：如何充分挖掘與集成各類資源（歷史文獻、考古資料、古舊地圖、老航照等）、歷史文獻數位化、歷史地理數據的不確定性與精度評價、歷史地理數據時空建模、歷史地圖編繪、歷史地理場景重建與可視化等問題。

隨著互聯網時代及雲計算技術的成熟，基於時空 GIS 基礎框架理念研製新一代 HGIS 已成為重要趨勢。臺灣中研院所研發的新一代 HGIS——SinicaView，就是基於統一時空架構的 4D GIS 平臺，可以視作「中華文明時空基礎架構」（CCTS）全新升級版，它既整合了中研院數十年來所建置的豐富歷史地理數位資源，又通過 API 及 Web Service 服務，可以包容不同來源的時空信息，從而建成了兼具社會服務性質與跨領域學術研究應用的綜合型歷史地理信息服務平臺。[10] 另一個例子是哈佛大學地理分析中心在互聯網上建立的數字化合作研究平臺——WorldMap。它是基於亞馬遜雲計算環境，採用面向 Web 服務（Amazon Web Services）架構，運用開源和開放存取模式，大大便利了學者們在互聯網上使用基礎地理信息服務，並能方便地進行歷史地理信息的可視化查詢、數據存取、合作共用和線上製圖。

3. 時空 GIS 基礎框架的定義、功能與挑戰

時空 GIS 基礎框架（Spatial-temporal GIS infrastructure）是多源、多媒體歷史地理數據的軟硬體集成環境（平臺），同時也包括數據獲取、加工、分析、交換及 Web 服務所

蘊，〈法國里昂第二大學特級教授安克強：地理信息系統是關涉到想像力的技術〉，《文匯報》，2013 年 7 月 1 日。

[10] SinicaView 平臺網址：http://3dgis.rchss.sinica.edu.tw/。

涉及的標準、技術、設施、機制等的總稱。它是由基礎歷史地理數據庫、歷史地理數據目錄與交換體系、歷史地理信息公共服務體系等構成。其中：(1)基礎歷史地理數據庫：是時空 GIS 基礎框架的核心，包括統一測繪基準（現代大地測繪基準）、基礎地理信息數據（現代測繪地理資料）、歷史地理專題數據、面向服務的歷史地理信息產品（系統定制）、管理系統和支撐環境。(2) 歷史地理數據目錄與交換體系：是歷史地理數據共建共用的關鍵，包括歷史地理數據標準分類體系和分類代碼，以及數據目錄、中繼數據、管理交換系統和支撐環境。(3) 歷史地理信息公共服務體系：是時空 GIS 基礎框架應用服務的表現層，包括線上數位地圖、線上空間分析工具、線上查詢系統及其支撐環境等，在互聯網、雲計算技術支援下，提供更為方便的信息查詢、地圖顯示與空間分析服務。（4）其他：包括歷史地理數據及服務的相關技術標準及運行保障體系，這是時空 GIS 基礎框架建設與服務的支撐和保障。[11]

簡單來看，時空 GIS 基礎框架不同于傳統的歷史地理數據庫，也不同於專題型歷史地理信息系統，它是一個綜合型歷史地理信息基礎設施（平臺），是基於互聯網環境所開發，提供數據錄入、數據管理、數據處理、數據交換、空間分析、專題製圖、成果展示、信息服務、線上協作及其多種工具支援的集成研究環境；特別是隨著雲計算、大數據技術的發展，時空 GIS 基礎框架也將成為架構在雲端的數位人文研究平臺。[12] 同時，根據包弼德（Peter K. Bol）的觀察，目前數位人文研究平臺可分三種模式，即：個人化平臺、聯邦式平臺，以及提供使用者探索後設數據來源的平臺。[13] 依此觀之，時空 GIS 基礎框架主要是一種聯邦式的數位人文研究平臺。

構建時空 GIS 基礎框架對於推進歷史地理學發展具有積極意義。一般而言，時空 GIS 基礎框架，具備統一時空基準、統一數據標準、統一信息服務、統一使用者介面等優點。採用統一時空基準，可以將不同來源的歷史地理數據，分為不同專題要素圖層，疊置後表達在統一時空坐標系中，便於歷史地理要素的古今對照分析。採用統一數據標準，又包括統一要素分類與編碼、統一數據結構、統一數據目錄、統一數據交換結構等內涵，為歷史地理數據規範化建庫、共建共用等帶來極大便利。採用統一信息服務模式，則充分遵循 OGC 相關標準及 Web Service 技術，採用統一服務介面與信息服務技術，為網路數據訪問與信息共用，提供標準化介面與規範化服務。統一使用者介面，是基於面向使用者友好的原則，提供統一易用的視窗介面和工具集，可提高系統使用效率、提升

[11] 時空 GIS 基礎框架的定義及組成的設計，參照國家標準《GB/T 30317-2013 地理空間框架基本規定》的相關內容。

[12] 時空 GIS 基礎框架可視作一種數位人文研究平臺。可與中研院所研發的「數位人文研究平臺」相參照，它是中研院數位文化中心所研發的雲端平臺，兼具資料開放存取與多人協同研究機制。研究者不僅能上傳文本與權威詞，更可自由加入不同主題的研究群組，結合平臺內部的既有豐富史料與其他研究者匯入的開放資料，運用文本自動標記、詞頻統計、相似內容比對、關聯分析、時空整合呈現、資料可視化等工具，進行文本資料探勘，發現新的知識脈絡。參見：http://dh.ascdc.sinica.edu.tw/。

[13] 包弼德（Peter K. Bol）將目前數位人文研究平臺，分為三種模式，即：個人化平臺（研究者可通過平臺建立自己的資料庫，如臺灣大學數位人文研究中心所研發的 DocuSky）、聯邦式平臺（由不同機構聯合而成），以及提供使用者探索後設資料來源的平臺。參見：房翠瑩，2018，〈盼共創臺灣人文研究新典範 中研院數位人文研究平臺 10 月吹響集結號〉，《數位文化志》，vol.1，線上地址：http://digitaldigest.ascdc.tw/。

使用者體驗和更好地服務學術研究。

按照中研院近年來所推進的「地理資訊數位元典藏與空間人文學發展計畫」（ https://ascdc.sinica.edu.tw/），就展示了以 SinicaView 為代表的時空 GIS 基礎框架，將具有的功能特徵及其學術追求：

> 將持續運用既有地理資訊數位典藏與技術發展成果來發展空間人文學（Spatial Humanities），將嘗試結合群眾外包（Crowdsourcing）技術，建立一個開放式地圖加值應用及空間敘事平臺，以「時間-空間-主題」資訊為框架，提供臺灣各類型文史研究空間化、可視化的網路基礎設施（Cyber-Infrastructure）；透過線上地圖校正讓使用者自行擴充歷史 GIS 圖層，並透過故事地圖介面讓使用者自由進行空間敘事（Spatial Narratives），進而促進鄉土研究、社區發展及環境改善。[14]

當然，構建時空 GIS 基礎框架也存在著一定挑戰。歷史地理信息往往是包括「人、事、時、地、物」五要素的時空整合信息，[15] 具有時空模型複雜、歷史地理要素空間定位難、時空精度差、信息不確定性強、數據多源異構等多方面問題，因此還有待進一步探索。特別是，歷史地理數據主要來自古文獻及古地圖的考證、解讀、定量與定位，數據的可獲得性、完整性、可靠性、歧義性、時段性等與現代地理數據存在很大差異，從而對歷史地理信息的時空定位提出新的難題，歷史地理數據的不確定性與時空尺度問題、古今地理數據的時空配准與整合應用等已經成為歷史地理信息化研究的基礎科學問題。

4. 構建城市歷史地理時空 GIS 基礎框架

基於以上思考，本文將在六朝建康歷史地理信息系統研製的現有基礎上，進一步探索時空 GIS 基礎框架的構建。[16]

[14] 參見網址：https://ascdc.sinica.edu.tw/，檢索時間：2018 年 10 月 2 日。

[15] 關於歷史地理信息的五要素特徵，可參見范毅軍研究員近年來的相關學術報告，如前述第三屆全國歷史地理信息系統學術沙龍中，主題報告「地理信息與空間人文——構建虛擬時空框架的設想、具體實踐與應用」中的相關表述。

[16] 南京大學歷史地理信息化研究團隊，近十多年來致力於運用 GIS 技術開展南京歷史地理研究，通過學科交叉、兩岸合作等方式，研製六朝建康、民國南京主題數據庫 7 個，專題信息系統 6 個，開發了面向學界開放的六朝建康歷史地理信息系統（http://hdgis.nju.edu.cn/jk/），結合《南京古舊地圖集》整理與出版，逐步推進 GIS 技術支援下的《南京歷史地圖集》編繪研究、發展南京城市歷史地理信息系統。

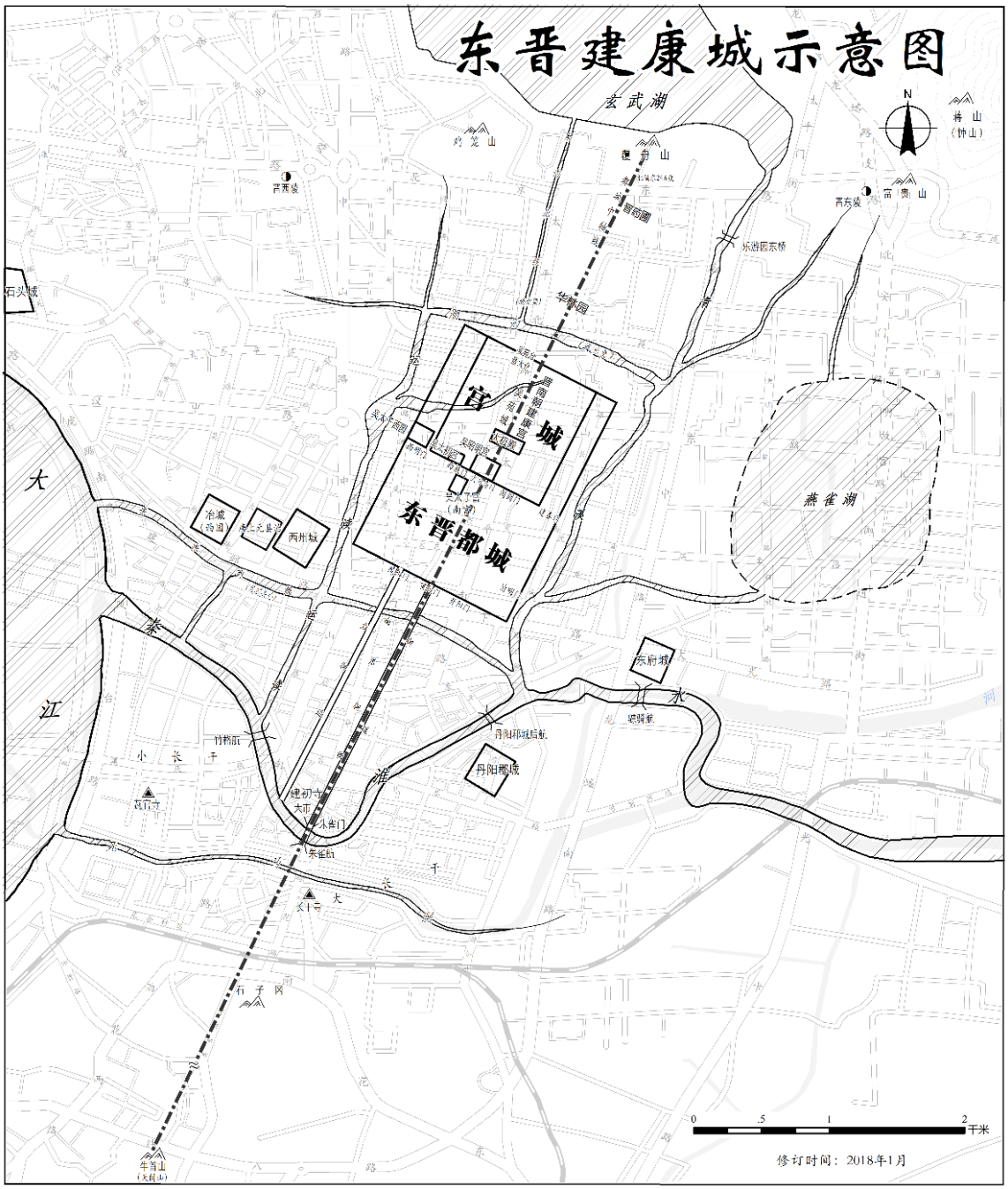


圖 1 東晉建康都城遺址復原圖（合作編繪：張學鋒、陳剛）[17]

4.1. 工作基礎

面向六朝建康歷史地理研究，本研究通過十多年的努力，系統收集南京歷史地理文獻，開展文獻數位化工作（以《建康實錄》為中心），同時收集與整理了近現代以來的南京城市六朝考古發掘與研究文獻（以考古報告、考古簡報為主，文獻資料跨度近 70 年，約 320 餘篇）。在此基礎上，結合正史、地方史志及其他學術文獻等的比勘分析，

[17] 自 2014 年起，與張學鋒、胡阿祥教授合作，以歷史文獻、考古資料為中心，結合實地踏勘與 GIS 製圖，完成「東晉建康都城遺址復原圖」（2014 年初稿、2108 年修訂），相關研究，可詳見：張學鋒，2015，〈所謂「中世紀都城」——以東晉南朝建康城為中心〉，《社會科學戰線》，第 8 期，頁 71-80。

借助野外踏勘及 ArcGIS 製圖技術，在現代城市電子地圖基礎上編繪六朝建康城市復原圖（圖 1），作為六朝建康歷史地理研究的基礎工作底圖。同時，考證與定位六朝考古遺址（墓葬、石窟、石刻等遺存）約 250 餘處，按孫吳西晉、東晉、南朝宋、南朝齊、南朝梁、南朝陳等時代及遺址類型進行分類表示，繪製六朝建康城市考古遺址專題地圖，進而研製專題數據庫（見下）與六朝建康歷史地理信息化展示平臺（圖 2）。



圖 2 六朝建康歷史地理信息化展示平臺（http://hdgis.nju.edu.cn/，2015 年 9 月上線）

① 《建康實錄》全文檢索系統（http://hdgis.nju.edu.cn/seo/）：系統收錄《建康實錄》全文 36 萬餘字，掃描圖像共 893 頁。同時，利用國際電子文獻標準 TEI/XML，進行文本標準化與語義提取，從中提取地名、人名、事件、事件等信息。線上系統提供對《建康實錄》的全文檢索服務，並提供了電子文本與古籍掃描圖像之間的參照對讀功能；系統提供關鍵字精確查詢及任意字詞快速檢索，保證了線上查詢的快速回應與準確性。

② 六朝建康城市歷史地名信息系統（http://hdgis.nju.edu.cn/HG）（圖 3）：包含歷史地名 1388 條（城市地名 455 條）。在城市歷史地名分類體系與城市歷史地名時空數據模型基礎上，結合古舊地圖與重要歷史文獻資料實現城市歷史地名屬性與空間信息的有機整合，最終形成地名時間、空間、屬性、沿革的可視化展示。

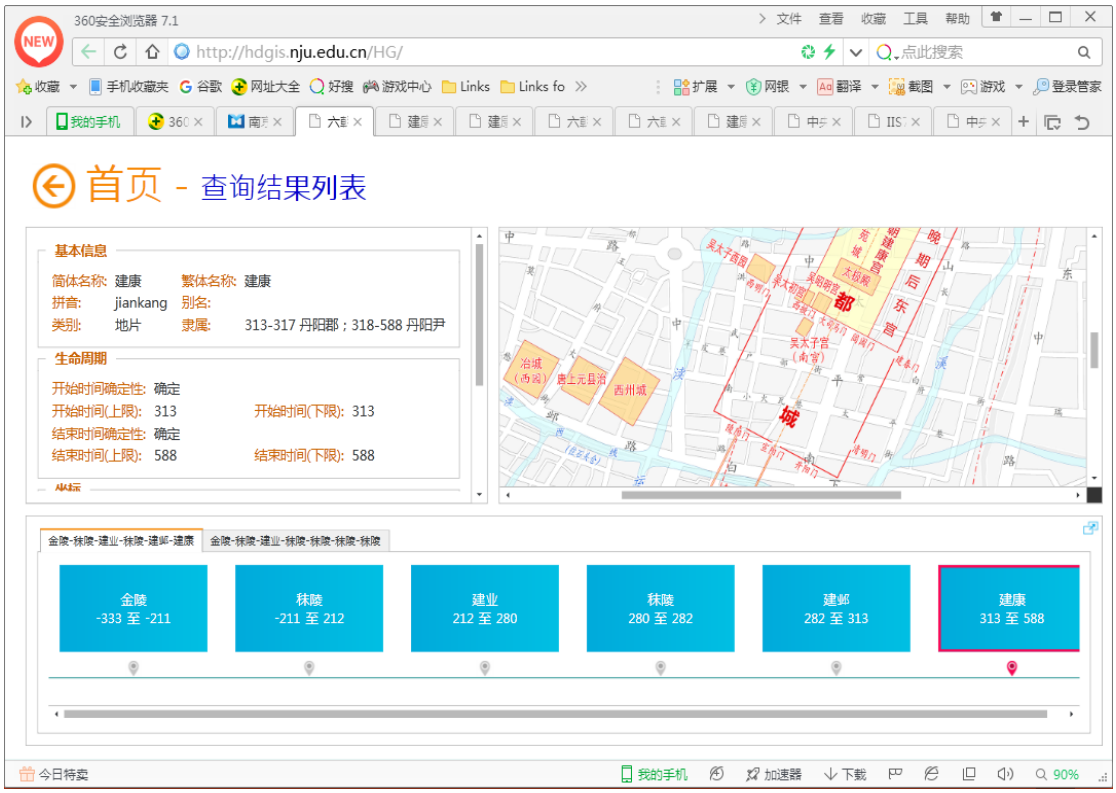


圖 3 六朝建康城市歷史地名信息系統（http://hdgis.nju.edu.cn/HG，2015 年 9 月上線）

③ 六朝建康城市考古遺址信息系統（http://hdgis.nju.edu.cn /NjArchaeologicalSite）（圖 4）：收錄六朝南京考古遺址 283 處、文獻資料 320 篇、多媒體資料 2188 張，構建了「考古遺址-文獻資料-多媒體資料」的關聯式結構，基於 MVC 框架，在網路地圖介面上將考古遺址及文獻資料進行整合，實現了考古遺址信息的時空展示與檢索。系統通過朝代、區域、媒體內容和文獻來源四種分類方式進行檢索，得到每種類型下的或遺址、或媒體物件、或文獻資料信息，對於每個特定的遺址、媒體物件或文獻信息均可展開詳細內容進行查閱；同時，將不同朝代的文物遺跡分類展現至現代電子地圖上，有利於表現其空間分佈特徵。



圖 4 六朝建康城市考古遺址信息系統

（http://hdgis.nju.edu.cn /NjArchaeologicalSite，2015 年 9 月上線）

④ 六朝建康文物圖像信息系統（http:// hdgis.nju.edu.cn/NjImage）：收錄六朝建康文物圖像 1366 幅，涵蓋目前所見的六朝時期出土文物，並將其分類、匯總、整合，將數據與時空屬性關聯，實現對文物數據的查詢、定位、流覽功能。

⑤ 六朝人物關聯式數據庫系統（http:// hdgis.nju.edu.cn/people）。對《建康實錄》中六朝人物關係數據進行採集，基於六朝人物關係時空數據模型，建設六朝人物關聯式數據庫並展開實證研究，可滿足對於六朝人物基本信息的查詢，對親屬關係的家族樹展示，人物社會關係的網路構建與人物地緣關係的地理空間可視化等歷史地理研究的需要。

⑥ 六朝建康陵墓石刻三維展示系統（http://hdgis.nju.edu.cn/sk）。基於無人機和三維重建技術，對分佈在南京、丹陽地區的 7 處六朝陵墓共 16 件石獸和神道柱等石刻進行了數據獲取和三維重建，獲得其實景三維模型，這些模型具有紋理清晰、精度較高（釐米級）、帶有地理座標等優點，可以直接作為三維空間展示的數據來源。

4.2. 六朝建康時空 GIS 基礎框架設計

據上，六朝歷史地理信息系統主要還是若干專題數據庫及信息系統的鬆散整合，接下來的工作，將研究與建設六朝建康時空 GIS 基礎框架，進一步整合歷史文獻、考古資料與野外考察為傳統的歷史地理學研究方法與內容，開展多層次、多維度的綜合研究。

4.2.1. 功能與特徵

（1）首先，面向城市歷史地理研究的時空 GIS 框架，需要有機融合人、時、地、事、物等多種要素，是構建在統一時空坐標系上的多維度、多來源、多尺度、多媒體的信息網路。

（2）其次，基於時空基礎框架的六朝建康歷史地理信息系統，需要構建有機關聯的多個數據庫，包括基礎歷史地理、歷史文獻、考古資料、歷史地名、歷史人物、歷史事件、歷史圖像等，具有多源異構等大數據特徵，需要涵括六朝建康研究的所有可用資源。

（3）再次，基於時空基礎框架的六朝建康歷史地理信息系統，具有統一的時空坐標系，在 GIS 平臺上來組織與整合各類資料；因此，需要在歷史文獻數位化、歷史地名考證與時空定位、歷史地圖編繪、多媒體信息展示等關鍵技術方面，開展深入研究。

（4）基於時空基礎框架，進一步推進六朝建康歷史地理信息化建設，對目前已完成《建康實錄》全文檢索系統、六朝建康城市歷史地名信息系統、六朝建康城市考古遺址信息系統、六朝建康文物圖像信息系統、六朝建康陵墓石刻三維展示子系統等數據庫與信息系統，進行進一步整合和升級。對正在開展的六朝人物關係時空數據庫，將在時空 GIS 基礎框架上，結合社會關係網路分析方法與可視化技術，構建六朝人物關係時空數據模型，進而研製六朝人物時空數據庫及其應用的實證案例。

4.2.2. 統一時空基準

根據六朝時期紀年方式的特點，採用年號紀年與西元紀年兩種方式。六朝時期朝代更迭頻繁、年號使用時間短，同時數據中對時間的記錄有模糊性，採用中國歷代人物傳記資料庫（CBDB）中的年號（nianhao）與朝代（dynasty）表的數據作為年號紀年的基本數據。同時設計了年號紀年與轉西元紀年的相互轉換工具，對時間數據進行歸一化處理，將所有的年號紀年全部轉換為西元紀年，便於時間上的統一管理。但同時將年號紀年保留在數據的屬性中，可以方便對照查閱。

以中國歷史地理信息系統（CHGIS V4.0）數據、譚其驤《中國歷史地圖集》和本研究所繪製的六朝建康城市復原圖作為基礎空間數據底圖。其中使用 CHGIS 數據中六朝數據作為主要行政區底圖數據，其包括了六朝時期各個朝代的郡縣向量數據（點狀以及面狀數據），可以為研究提供基礎地理數據。對於空間框架基礎數據，由於各個朝代不同時期的疆域變化、地名更名、區劃變化等原因，在不同時間點上同一個描述性地名可能對應的是不同的地點，所以對於空間框架需要引入時間維度，本文將其分多個時間斷面做出系列底圖。同時參照《中國歷史地圖集》與《中國古今地名大詞典》作為空間定位依據以及地名分類分等級的依據，歷史文獻中的描述性地名有多種不同的尺度，對地名進行等級劃分可以有效的對其空間精度進行管理。利用現代網路地圖作為對照地圖，在空間上提供一種直觀的古今對照關係，對於理解空間有著重要的學術意義與實用價

值。

4.2.3. 實證案例：六朝人物關係時空數據建庫研究 [18]

考慮到六朝建康時空 GIS 基礎框架研究與建設需要一個較長的探索時期，本文以正在開展的六朝人物關係時空數據建庫研究為例，探索構建六朝建康時空 GIS 基礎框架設計與建設的關鍵技術。該案例作為時空 GIS 基礎框架建設的重要內容，主要是結合社會關係網路分析方法與可視化技術，設計六朝人物關係時空數據模型，研製六朝人物時空數據庫及其信息系統。

（1）統一時空基準

建立統一的歷史時空框架，將六朝人物放置于特定歷史時空場景中。歷史時期不同朝代或相同朝代不同時間，疆域、政區與地名都會發生變化，導致同名異地、同地異名的情況，利用不同的時間與地名標準對於歷史人物研究會造成困難，有必要構建統一的歷史時空框架，為歷史人物關係提供研究的時空場景。

利用譚其驤主編《中國歷史地圖集》與網路電子地圖作為基礎數據，構建統一的時空框架。六朝時期分朝代選擇六個標準年，從《中國歷史地圖集》中選擇對應時間的歷史地圖，按照研究區範圍將其基本要素（疆界、政區、地名）數位化，共同構成六朝時期的基本時空框架基準。分別是：吳永安五年的〈三國時期全圖〉(西元 262 年)、晉太元七年的〈東晉十六國時期全圖〉(西元 328 年)(圖 5)、宋元嘉二十六年的〈宋、魏時期全圖〉(西元 449 年)、齊建武四年的〈齊、魏時期全圖〉(西元 497 年)、梁中大同元年的〈梁、東魏、西魏時期全圖〉(西元 546 年）和陳太建四年的〈陳、齊、北周時期全圖〉(西元 572 年)。選擇西元 328 年東晉圖作為基礎地理參照，利用六張數字歷史地圖（空間）與西元紀年（時間），構建出六朝人物關係人物研究的時空框架基準。



[18] 此處工作由筆者和研究生曹飛飛共同完成。具體參見：曹飛飛，2018，《六朝人物關係時空資料建模研究》，南京大學碩士論文。

圖 5 東晉太元七年（西元 382 年）歷史地圖

（2）六朝人物時空數據模型設計

參照中國歷代人物傳記數據庫（CBDB）數據模型，[19] 根據六朝人物關係數據的特點，改進 CBDB 人物關聯式結構，本研究以西元紀年與古今地名編碼方式，依據六朝人物關係數據的特點，分別對親屬關係、社會關係與地緣關係的組織方式進行改進，然後將三種人物關係模型整合到統一的時空框架下，構建六朝人物時空數據模型。

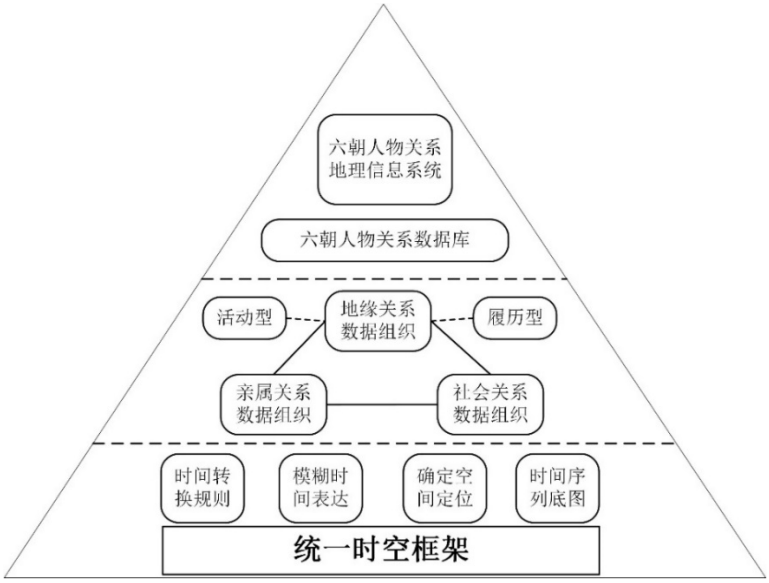


圖 6 基於同一時空框架的六朝人物關係時空數據模型

（3）六朝人物關聯式數據庫設計與建庫

六朝人物關聯式數據庫根據內容可以分為幾個組成部分，分別為人物基本信息、親屬關係、社會關係、地緣關係、時間系統、空間系統以及使用者註冊信息。人物基本信息主要由以下表組成：人物表（person）、人物別稱表（altname）、人物頁碼索引表（pages_index）和人物文本信息表（person_text）。人物表記錄了人物姓名、朝代、生卒年等信息，別稱表記錄了人物的字、諡號等信息；人物頁碼索引表記錄了人物在《建康實錄》原文中出現的頁碼，人物文本信息表記錄人物在《建康實錄》中出現的句子原文。

在六朝人物時空數據模型的基礎上，以《建康實錄》為基礎文獻，利用 VBA、Python 以及規則運算式，對《建康實錄》中六朝人物關係數據進行採集；採用 MySQL 作為數據庫平臺，建立六朝人物關聯式數據庫，利用 MVC 框架結合線上地圖 API 研製六朝人物關係地理信息系統。系統為使用者提供了人物基本信息查詢、親屬關係查詢與家族樹可視化、社會關係查詢與網路圖可視化、地緣關係查詢與 Web 地圖展示，初步滿足了用戶的使用需求。在此基礎上展開六朝人物關係研究。

[19] UNIVERSITY HARVARD. 中國歷代人物傳記資料庫（CBDB）[EB/OL]. https://projects.iq.harvard.edu/chinesecbdb.



圖 7　六朝人物關係地理信息系統（內測中，暫未開放）

5. 結語

本文首先討論「空間人文」的緣起及其在數位人文、歷史信息系統發展脈絡的中地位與意義，進而探討作為新一代歷史 GIS 平臺——時空 GIS 基礎框架的定義、功能及其學術意義。本文所指時空 GIS 基礎框架，是多源、多媒體歷史地理數據的軟硬體集成環境（平臺），同時也包括數據獲取、加工、分析、交換及 Web 服務所涉及的標準、技術、設施、機制等的總稱，是新一代的歷史 GIS 平臺。接著，面向城市地理學研究，以六朝建康歷史地理信息化建設的現狀分析為基礎，提出構建城市歷史地理學研究的時空 GIS 基礎框架的理論與方法，並以六朝人物關係時空數據建庫為例，初步探索構建六朝建康時空 GIS 基礎框架設計與建設的相關關鍵技術。在未來研究中，在時空 GIS 基礎框架支持下，六朝建康歷史地理化研究將進一步整合與集成歷史文獻、考古數據與野外考察等歷史地理學研究文獻與研究方法，開展多層次、多維度的綜合研究，以期重建六朝建康歷史地理圖景。

參考文獻

Knowles A K. 2000. “Special Issue: Historical GIS: The spatial turn in social science history”. *Social Science History*, 24(3)，pp.451-70.

Baker A. R. H. 2003. *Geography and history: bridging the divide*. Cambridge: Cambridge University Press.

Ian N. Gregory, Richard G. Healey. 2007. “Historical GIS: structuring, mapping and analyzing geographies of the past”. *Progress in Human Geography*, 31(5), pp.638–653.

Bol P.K. 2013. “On the Cyberinfrastructure for GIS-Enabled Historiography”. *Annals of the*

Association of American Geographers, 103(5), pp.1087-92.

Guan W. W, Bol P K, Lewis B G, et al. 2012. “WorldMap–a geospatial framework for collaborative research”. *Annals of GIS*, 18(2), pp.121-34.

陳軍（2002）。〈論數字化地理空間基礎框架的建設與應用〉,《測繪工程》第 3 期，頁 1-6。

王汎森（2004）。〈歷史研究的新視野：重讀「歷史語言研究所工作之旨趣」〉,《中央研究院歷史語言研究所七十五周年紀念文集》。臺北：中央研究院歷史語言研究所，頁 161-176。（《古今論衡》，第 11 期，頁 1-12）

范毅軍、廖泫銘（2008）。〈歷史地理資訊系統建立與發展〉,《地理資訊系統》（季刊）第 1 期，頁 23-30。

廖泫銘、范毅軍（2012）。〈中華文明時空基礎架構：歷史學與資訊化結合的設計理念及技術應用〉,《科研信息化技術與應用》第 4 期，頁 17-27。

項潔、翁稷安（2012）。〈多重脈絡——數位檔案之問題與挑戰〉，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，頁 25-59。

王汎森（2014）。〈數位人文學之可能性及限制——一個歷史學者的觀察〉，收于項潔編《數位人文研究與技藝》。臺北：臺大出版中心，頁 25-35。

陳剛（2014）。〈「數字人文」與歷史地理信息化研究〉,《南京社會科學》第 3 期，頁 136-142。

林富士主編（2017）。《「數位人文學」白皮書》。臺北：中央研究院數位文化中心。

（致謝：中央研究院人文社會科學研究中心地理資訊中心執行長范毅軍研究員、廖泫銘研究助技師對本文研究提供很多幫助，特致謝忱。本文研究得到國家自然科學基金「基於新型超媒體 GIS 技術的城市歷史地理研究」（41271160）資助；同時，南京大學數字人文與超媒體 GIS 工作室曹飛飛、于靖、段淼然等研究生為本工作付出了許多努力，也借此表示感謝。）

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

兩宋鎮墓文物的地域分佈與變化

「遼宋金墓葬資料庫」的建置與應用

許雅惠*　黃庭碩**　黃皇堯***　杜協昌****

國立臺灣大學歷史學系副教授*

國立臺灣大學歷史學系博士候選人**

國立臺灣大學資訊工程學系碩士***

國立臺灣大學資訊工程學系博士後研究****

兩宋鎮墓文物的地域分佈與變化：「遼宋金墓葬資料庫」的建置與應用

許雅惠（國立臺灣大學歷史學系副教授）
黃庭碩（國立臺灣大學歷史學系博士候選人）
黃皇堯（國立臺灣大學資訊工程學系碩士）
杜協昌（國立臺灣大學資訊工程學系博士後研究）

「遼宋金墓葬資料庫」完整收錄遼宋金時期的墓葬，彙集各期刊與專刊中的考古報告，分類整理，以供搜尋。內容包括：墓葬結構、墓主資訊、墓室裝飾與隨葬品等。每則資料均附書目，以便查找原始出處。目前已建置完成之資料近四千筆，其中兩宋墓葬約 2,200 筆（不含漏澤園），遼金元墓葬約 1,500 筆，未來也將持續更新資料。本資料庫由於完整收錄十至十三世紀各地墓葬，可讓研究者鳥瞰大範圍的地區分佈或長期的發展趨勢，進一步從中定義有意義的研究課題。

資料庫建置過程中，與臺大資工所項潔教授的 DocuSky 團隊合作，開發表格工具，讓使用者可對 Excel 表格資料進行標註、群組與筆記，以利資料整理與分析。此表格工具也介接 DocuGIS，讓使用者得以迅速掌握資料的地理空間分佈。本論文利用「遼宋金墓葬資料庫」與此新開發的表格工具，一方面展示計畫成果，一方面思考其研究上的應用。

一、前言

古代中國的鎮墓習俗源遠流長，各地區作法不盡相同，隨著人群的移動、信仰的變遷，形成複雜的歷史樣貌。綜觀十至十三世紀宋墓，常見的鎮墓之物包括買地券、五色卵石、神煞俑、鐵豬鐵牛。買地券的使用可上溯到漢代，一般在方磚上書寫購買葬地的契約。[1]一方面警告地下之邪精故炁，不得干犯騷擾，以此祈求「官家富貴，男女昌熾」；[2]另一方面令死者在地下安居，與生者永不相見，「若要相見，直至海變桑田」。[3]五色卵石、神煞俑與鐵豬鐵牛則承襲自唐代，卵石塗以象徵五方的五色，置於墓室四角與中央。鐵豬與鐵牛經常成對出現，擺放於墓室四角。神煞俑多以陶、木製成，包括造型奇特的人俑與動物俑，在墓中有一定的擺放位置，見於《大漢原陵秘葬經》記載。[4]

[1] ﺎ ﻛﻜﻎ ﻗﻘﻐﻌﻨﻊ ﻉ ﻟﻠﻜ

本研究利用科技部數位人文計畫所建置之「遼宋金墓葬資料庫」，對兩宋墓葬中的鎮墓文物做整體分析。探討上述鎮墓文物的地理空間如何分佈？如何搭配使用？從十至十三世紀，地理分佈有何變化？對資料進行大規模鳥瞰分析，掌握其長時段的地理分佈之後，進一步結合墓葬隨葬特點與文獻記錄，勾勒兩宋時期鎮墓習俗的發展，並探討其變化的原因。

二、鎮墓文物的類型分佈

1. 地券

地券是以死者名義，向地下官僚購買葬地的陰間買賣契約，出現於東漢，一直到十九世紀都還可見，使用時間很長。[5]一般認為地券的使用者為平民，能反映有別於士大夫階層的民間流行信仰。[6]宋代地券使用以南方為多，華北地區也使用，山東、河北基本不見。

北宋仁宗年間曾令王洙（997-1057）校正《地理新書》，費時二十一年，於仁宗嘉祐元年（1056）書成。[7]書中明文提供地券的文字範本，應是參酌既有規範，釐定而成：

> 某年月日，具官封姓名，以某年月日歿故，龜筮協從，相地襲吉，宜於某州某縣某鄉某原安厝宅兆。謹用錢九萬九千九百九十九貫文，兼五綵信幣，買地一段。東西若干步，南北若干步，東至青龍，西至白虎，南至朱雀，北至玄武。內方勾陳，分掌四域。丘丞墓伯，封部界畔；道路將軍，齊整阡陌。千秋萬歲，永無殃咎。若輒干犯訶進者，將軍亭長，收付河伯。合以牲牢酒飯，百味香新，共為信契，財地交相分付，工匠修營安厝，已後永保休吉。知見人歲月主，保人今日直符，故氣邪精，不得忤悞。先有居者，永避萬里。若違此約，地府主吏，自當其禍。主人內外存亡，悉皆安吉。急急如五帝使者女青律令。[8]

實際出土例證中，地券中的知見人與保人有幾：除《地理新書》所列歲月、今日直符外，最常見的有張堅固、李定度。分析宋代地券，兩者分佈地域都以南方為多，集中在四川與江西。張堅固與李定度流行的時間較早，歲月、今日直符流行於十二世紀之後，應該與《地理新書》的頒布有關。

[5] 關於歷代買地券的整理、錄文與討論，見魯西奇，《中國買地券研究》（廈門：廈門大學出版社，2014）。

[6] ا ككذ ع ذ ضغضغ ق ج ملا ع غ ش ت د ففغف ق ف ك ت م ك ر ف ك ,□۶۷۹-714.

[7] 影鈔金明昌三年本《地理新書》王洙序有小字注，其所記校正成書年代有誤，關於《地理新書》之編纂過程，見沈睿文，〈《地理新書》的成書及版本流傳〉，《古代文明》，第 8 卷（2010），頁 313-336。

[8] 王洙等撰，《圖解校正地理新書》，影抄金明昌三年（1192）本（臺北：集文書局，1985），頁 455。

隨葬買地券的宋墓共 180 餘座，南北宋各八十座左右，分佈如下：

圖 1a、兩宋地券分佈圖（北宋）

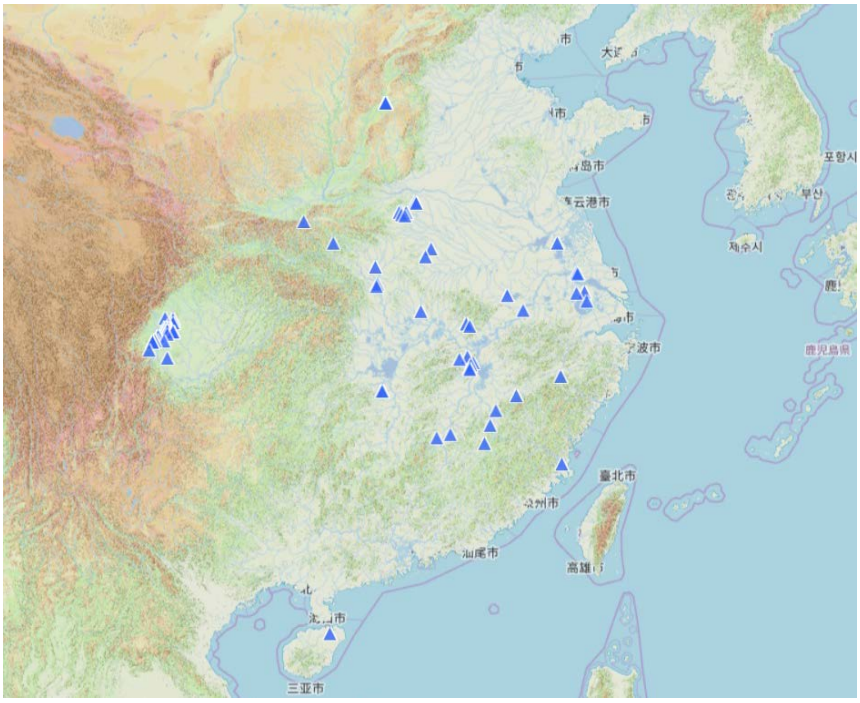
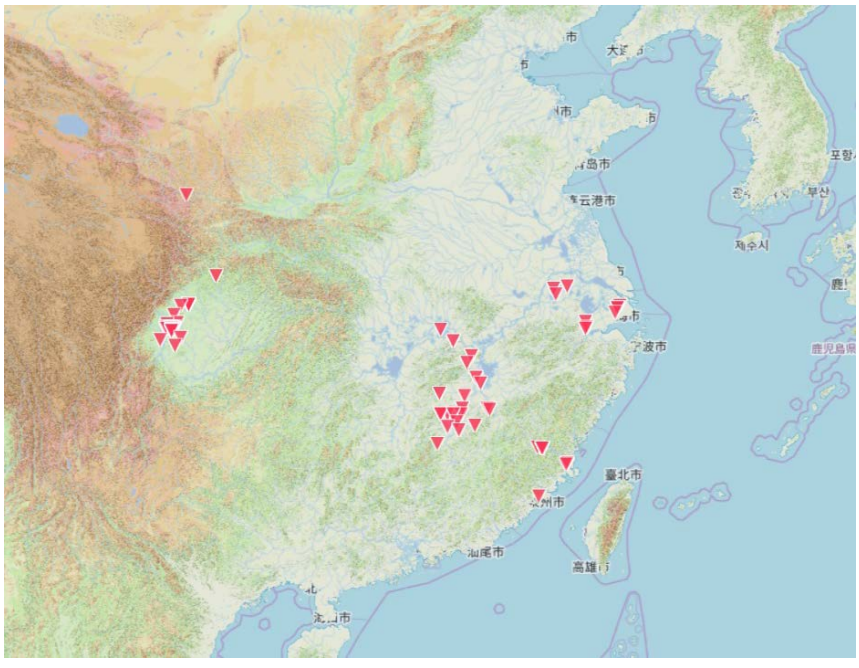


圖 1b、兩宋地券分佈圖（南宋）



宋代地券分佈範圍甚廣，北至山西太原、西至甘肅定西、南至海南島皆有發現，而在傳統核心區中，尤以四川、江西、河南、江浙等地最為密集。比較兩宋分佈，南宋地券更加集中於四川、江西與江浙，福建出土的地券也增加。

就墓主身份而言，隨葬地券者多為平民，士人官僚的比例不高，在本資料庫所收約一百八十座宋墓中，可確知出自官僚階層的僅十九座，約占一成強，而且都出現在地券流行的地區，包括江蘇、四川、江西、福建、上海，應該是遵循地域的葬俗傳統。此外，地券材質有磚、石、木、鐵，平民階層一般書寫在磚、石之上，江蘇地區北宋墓葬常用木地券，至於鐵地券的使用者大半為士人官僚。

2. 鎮墓石

鎮墓石是以完淨的卵石，按五方方位塗上青、白、赤、黑、黃色，放置於墓室四角及中央，稱為「五精石」或「五色石」。根據文獻記載，宋真宗（997-1022在位）永定陵便用五色石鎮墓法，將來河南鞏義北宋皇陵若發掘，應可發現鎮墓遺物。[9]

以五色石鎮壓五方是宋代官方規範的鎮墓法，見於仁宗朝完成的官書《地理新書》：

> 鎮墓古法有以竹為六尺弓度之者，亦有用尺量者。今但以五色石鎮之於冢堂內，東北角安青石，東南角安赤石，西南角安白石，西北角安黑石，中央安黃石，皆須完 婷大小等，不限輕重。置訖，當中央黃石南，祝之曰：「五星入北，神精保佑，歲星居左，太白居右，熒惑在前，辰星立後，鎮星守中，辟除殃咎，妖異災変，五星攝授，亡人安寧，生者福壽，急急如律令。」泓師云：「凡葬墓中用豆黃完淨者一斗，及錢紙五百張，安於墓內吉更多，此數尤佳。」[10]

從這段文字可知，五色石代表五星神精，歲星、太白居左右，熒惑、辰星居前後，鎮星居中。藉由五星鎮守，使「亡人安寧，生者福壽」。

河北石家莊市鹿泉區 M40 發現三塊朱書礫石，出土時緊貼於東、北、西三壁底部，上面文字漫漶，頭端鎮墓石書「□星」、東側書「□〔歲〕星居左」、西側為「太白居右」。[11]鎮墓石上的文字與擺放位置與《地理新書》記載相合，可見《地理新書》之施行。該墓為小型土坑豎穴夫婦合葬墓，銅錢最晚者為「紹聖元寶」，可知墓葬年代在哲宗紹聖（1094-1098）之後。

《地理新書》所載的五色石鎮墓法於兩宋使用情況如何？根據分佈圖，在北宋時期主要使用於華北地區，包括河北、河南與陝西西安，不見於長江流域。進入南宋，五色石鎮墓法之使用集中於長江下游的江蘇與浙江二地。從兩宋的分佈

[9] 徐松輯，《宋會要輯稿》（臺北：世界書局，1964），禮 29-25，頁 1076。
[10] 王洙等撰，《圖解校正地理新書》，頁 457。
[11] 四川大學歷史文化學院考古系等，〈河北鹿泉市西龍貴墓地唐宋墓葬發掘簡報〉，《考古》，2013 年 5 期，頁 42-43 (29-54)。

看來，《地理新書》的五色石鎮墓法可能是靖康之難時，隨著朝廷南遷，將此華北習俗傳播至東南地區。

圖 2a：兩宋五色石分佈圖（北宋）

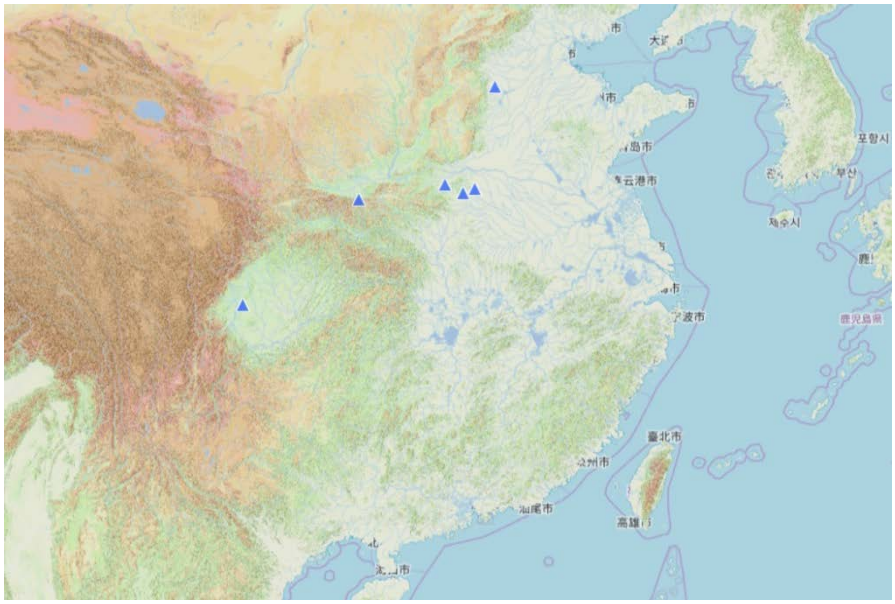
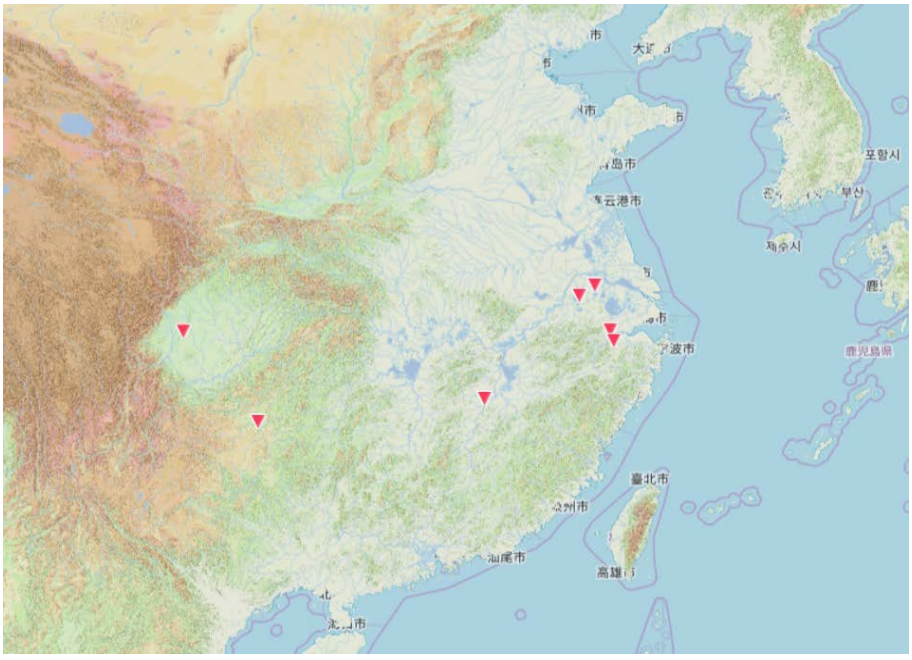


圖 2b：兩宋五色石分佈圖（南宋）



分析使用五色石的墓主身份，不少屬於士人與官僚階級。就其實際使用，北宋時與《地理新書》的規範較接近，在墓室四角與中央安放卵石。南宋時期出現

一些變通的作法，有在卵石上直接書寫「黃」字，代表中央黃石；[12]也有的僅以一顆卵石鎮壓於墓室中央。[13]

除了《地理新書》規範的五色石或五精石鎮墓法，還有一種方形的鎮墓石，碑石以秘篆書寫而成，分東、西、南、北、中五方，屬道教靈寶派。鎮墓石之使用儀式與鎮文內容見於《太上洞玄靈寶滅度五鍊生尸妙經》之「安靈鎮神天文」；真書文字則來自《靈寶無量度人上品妙經》卷一〈元始靈書中篇〉「大梵隱語」。[14]在此道教靈寶派的煉度儀式中，道士分別向東方青帝、西方白帝、南方赤帝、北方黑帝、中央黃帝祈願，煉度墓主亡魂，修煉升天。這類五方鎮墓真文券出現於唐代中宗時期，集中於陝西關中地區，使用者均為身份極高之皇室貴族，不少為中宗皇后韋氏（?-710）之家族成員。[15]入宋後僅見於四川，且集中出土於成都，使用者身份明顯降低，幾乎都是一般平民，其原因尚待探討。

3. 明器神煞俑

鎮墓文物中有一類神煞俑，多半以陶、木製成，造型作人形或動物形，也有人與動物混合的怪異造型，經常被稱作「神怪俑」。[16]這類墓俑見於唐宋官方禮書規定，從天子、品官到庶人，根據社會身份之高低，隨葬的種類、數量、材質均有等差。

唐代禮書規定的明器神煞俑種類不多，主要是四時十二神。根據《唐六典》（成書於 739 年）：

> 凡喪葬則供其明器之屬，三品以上九十事，五品以上六十事，九品已上四十事。當壙、當野、祖明、地軸、鞬馬、偶人，其高各一尺，其餘音聲隊與僮僕之屬，威儀、服玩，各視生之品秩所有，以瓦、木為之，其長率七寸。[17]

[12] 南京市博物館、江寧區博物館，〈江蘇南京南宋周國太夫人墓〉，《東南文化》，2010 年 4 期，頁 54-59。

[13] 趙一新、趙婧、蔣金治，〈金華南宋鄭繼道家族墓清理簡報〉，《東方博物》，2008 年 3 期，頁 54-61。

[14] 王育成指出鎮墓真文文字來自《靈寶無量度人上品妙經》，Carole Morgan 進一步指出鎮文內容來自《太上洞玄靈寶滅度五鍊生尸妙經》，往後之討論基本不出此兩部道教經典。參王育成〈唐宋道教秘篆文釋例〉，《中國歷史博物館館刊》，總 15-16 期（1991），頁 82-94、46；Carole Morgan, □ق د ققلاب م ل كمپ غگ ذ ځ ځ غ ځ ر ع گ ځ ح ځ ا ، ل گ م ذ ځ ځ ق ل ا ع ا ل ک □ *T'oung Pao*, second series, vol. 1, no. 4/5 (1996): 317-348；加地有定，《中国唐代鎮墓石の研究：死者の再生と崑崙山への昇仙》（大阪：かんぽう，2005），頁 114-124。

[15] 加地有定，《中国唐代鎮墓石の研究：死者の再生と崑崙山への昇仙》（大阪：かんぽう，2005），頁 40-112。

[16] 張勛燎、白彬，〈隋唐五代宋元墓葬出土神怪俑與道教〉，《中國道教考古》（北京：線裝書局，2006），頁 1611-1750。

[17] 李林甫等撰，陳仲夫點校，《唐六典》（北京：中華書局，1992），卷 23，頁 5970。

文中提到的明器有當壙、當野、祖明、地軸、鞮馬、偶人，共六種，以陶、木為之，隨官階數量不同。其中當壙、當野、祖明、地軸合稱為「四神」，在有些禮書中與十二神並列，為墓中必要的明器。[18]

《唐會要》記載元和六年（811）訂文武官及庶人喪葬品秩，共四等：三品以上、五品以上、九品以上、庶人，各等使用之儀物不同：

> 三品以上，明器九十事，四神十二時在內……庶人明器一十五事，共置三舁。喪車用合轍車，幰竿減三尺、流蘇減十道、帶減一重，幃額、魌頭車、魂車准前，挽歌、鐸翼、四神十二時各儀，請不置。所用明器竝令用瓦，不得過七寸。[19]

《唐會要》還規定，四神十二時限品官使用，庶人不置。十二時即為十二時之神，作十二生肖動物形象，也稱十二支獸。以天干、地支相配計時之法，在商代便已發展完成，見殷墟出土的甲骨卜辭。不過將十二地支配上十二種動物的做法可能較晚，目前可追溯至戰國晚期，有本土起源與外來說兩種看法，尚無定論。北朝開始出現陶製的十二支獸俑，造型隨時代有別，至唐代大致定型，有動物首人身，也有整體人形的表現。[20]

考古資料顯示，唐代的長安與洛陽兩京地區的確流行以陶明器隨葬，有些表面罩上低溫鉛釉，即為一般熟知的唐三彩。這些陶俑在墓中的佈排頗有規律，墓口通常是一對鎮墓獸，一作單角人面、另一作雙角獸面，鎮墓獸後方跟著一對天王或武士俑，再往墓室內部則是各類文、武、侍俑，以及家禽與家畜等動物俑。王去非認為墓口的兩件鎮墓獸為祖明、地軸，而其後的天王或武士俑則爲當壙、當野，即為唐代禮書中所稱的「四神」。[21]河南鞏義康店磚廠採集的加彩陶獸面鎮墓獸，背部墨書「祖明」二字，可證實王去非的看法。[22]

四時俑之外，唐墓中還陪葬一些《唐六典》所載的「偶人」，包括：文官俑、武官俑、侍僕俑。河南鞏義唐墓與河南偃師咸亨三年（672）楊堂墓出土一些帶有墨書題字的陶俑，包括「文官」、「武官」、「奴」、「奴益錢」、「奴典樂」等，說明這類陶俑在地下世界中的職能。[23]

[18] 王去非，〈四神、巾子、高髻〉，《考古通訊》，1956 年 5 期，頁 50-54。
[19] 王溥撰，《唐會要》（北京：中華書局，1955, 1990），卷 38，頁 695-696。
[20] 秦浩，〈南方唐墓的形制與隨葬品〉，《南京大學學報》，1982 年 1 期，頁 72-78；謝明良，〈出土文物所見中國十二支獸的形態變遷〉，《故宮學術季刊》，第 3 卷第 3 期（1986），頁 59-105；鄧菲，〈形式與意涵的多元性：論兩宋考古資料中的十二生肖像〉，《民族藝術》，2015 年 6 期，頁 90-100。
[21] 王去非，〈四神、巾子、高髻〉，《考古通訊》，1956 年 5 期，頁 50-54。
[22] 鄭州市文物考古研究所，《中國古代鎮墓神物》（北京：文物出版社，2004），頁 181。
[23] 張新月等，〈鞏義出土的唐代題字俑〉，《中原文物》，2007 年 2 期，頁 81-87；偃師商城博物館，〈河南偃師縣四座唐墓發掘簡報〉，《考古》，1992 年 11 期，頁 1005。

唐墓還出現一些不見於《唐六典》與《唐會要》的神煞或神怪俑，自徐蘋芳於 1963 年指出《大漢原陵秘葬經》（以下稱《秘葬經》）在研究宋元華北墓葬的重要性後，研究者經常將考古出土的神怪俑與《秘葬經》「盟器神煞篇」相比對。[24]唐代的神怪俑集中出土於遼寧、河北與長江中游一帶，如遼寧朝陽的初唐、盛唐墓出土不少徐蘋芳稱作「伏聽」（匍匐人形）、「儀魚」（人首魚身）、「墓龍」（人首蛇身）的陶俑。[25]安徽、江蘇、湖南、河北曾出土被稱作「觀風鳥」的人首鳥身俑。[26]河南鞏義唐墓也曾出現奇特的動物造型俑，如「地吞」陶獸俑。[27]這類人、獸混合的奇特造型似不見於首都長安地區，已發掘的幾座唐代中、晚期太子、公主陵墓也均未見。不過長安地區曾出土數例跪拜文官俑，其尺寸與身份相稱，如玄宗之兄「讓皇帝」李憲（679-742）墓所出便達 102 公分，很可能是《秘葬經》「伏聽」的形象淵源。[28]

宋初制度大體沿襲唐代，如《宋史》「詔葬」載：

> 又按《會要》：勳戚大臣薨卒，多命詔葬……其明器、牀帳、衣輿、結綵牀皆不定數。墳所有石羊虎、望柱各二，三品以上加石人二人。入墳有當壙、當野、祖思、祖明、地軸、十二時神、誌石、券石、鐵券各一。殯前一日對靈柩，及至墳所下事時，皆設敕祭，監葬官行禮。熙寧初，又著新式，頒于有司。[29]

宋初官方制度明言的明器種類不多，僅較唐代增加「祖思」一項。至宋真宗永定陵又新增一些之前未載的明器，包括仰觀、伏聽、清道、蒿里老人、鯢魚，[30]其中仰觀、伏聽、蒿里老人三者名稱見於《秘葬經》「盟器神煞篇」記載。

《秘葬經》「盟器神煞篇」出現大量明器，這些名稱多不見於唐宋官方禮書記載。書中明確規範墓中使用的明器種類與擺設位置，按身份分為四級：天子山陵、親王、公侯將相、大夫以至庶人。身份愈高，陪葬的種類與數量也愈多，以下為大夫至庶人為例：

> 十二元辰，長一尺二寸，按十二方位。五呼將，長一尺二寸，鎮墓五方。五精石鎮五方。祖司、祖明，長一尺二寸，安棺後。仰觀、伏聽，長一尺二寸，安埏道中。當壙、當野長一尺二寸。五穀倉一尺二寸，三漿水高九寸，安棺頭。金雞，高一尺二寸，安酉地。玉犬長二尺九寸，高

[24] 徐蘋芳，〈唐宋墓葬中的「明器神煞」與「墓儀」制度〉，頁 87-106；張勛燎、白彬，〈隋唐五代宋元墓葬出土神怪俑與道教〉，《中國道教考古》，頁 1611-1750。
[25] 吳炎亮，〈試析遼寧朝陽地區隋唐墓葬的文化因素〉，《文物》，2013 年 6 期，頁 50-56。
[26] 耿超，〈唐宋墓葬中的觀風鳥研究〉，《華夏考古》，2010 年 2 期，頁 112-119。
[27] 張新月等，〈鞏義出土的唐代題字俑〉，頁 86。
[28] 張蘊，〈關於李憲墓隨葬陶俑的等級討論〉，《考古與文物》，2005 年 1 期，頁 60-63；李奕周，〈唐代跪拜俑與伏聽俑考辨〉，《文物鑒定與鑒賞》，2017 年 8 期，頁 14-20。
[29] 脫脫，《宋史》（臺北：鼎文，1980），卷 124，頁 2909-2910。
[30] 徐松輯，《宋會要輯稿》，禮 29-20，頁 1073。

一尺，安戌地。蒿里老公，長一尺五寸，安堂西北角。天關二箇，長一尺二寸，安堂西〔南〕北界上。地軸二箇，長一尺二寸，安堂東西界上。天喪、刑禍一對，長二尺，安墓？。墓龍長三尺，高一尺二寸，安辰地。金牛，長二尺，高一尺二寸，安丑地。玉馬，高一尺，安午地。鐵豬，重三十斤，安亥地。四廉路神，長一尺九寸，安西〔四〕角。以上皆大夫庶人用之吉。凡大葬後，墓內不立盟器神，亡靈不安，天曹不管，地府不收，恍惚不定，主人不吉，大殃咎也。[31]

根據《秘葬經》，墓中若不立明器神煞，將使「亡靈不安，天曹不管，地府不收，恍惚不定，主人不吉，大殃咎也」，可知這些神煞俑的用途是壓勝鎮墓，以安亡靈。

唐代的南北各地雖出現一些神煞俑，但不見於長安地區的高級墓葬。五代十國時期，位於江蘇江寧的南唐李昇欽陵（建於 943 年）與李璟順陵（建於 962）均出土不少神煞俑，墓主為國君，等級相當高。徐蘋芳認為這些俑是：蒿里老人、鎮殿將軍、儀魚或鯢魚、墓龍。[32]入宋之後，明器神煞陶俑集中於四川與江西二地，不僅數量多，而且有一定的擺放位置，呈現出系統化的發展趨勢。除了區域性顯著，分析使用者身份也可知，神煞俑的使用者多為一般平民。

四川與江西二地墓葬中根據方位佈排神煞俑的概念，與《秘葬經》十分接近。少數神煞俑還帶有題記，可知其名稱，有些與《秘葬經》與《地理新書》記錄相合。以下按年代整理帶題記的神煞俑：

出土墓葬	《秘葬經》	《地理新書》	其他
四川成都新津元豐四年（1081）王府君墓陶俑[33]	「文」、「武」、「老人」		「地精」、「萬歲女」
江西南豐古城政和八年（1118）墓瓷俑[34]	「仰俑」、「伏聽」		
江西南豐桑田鄉南宋墓瓷俑[35]	「金雞」、「玉犬」		「大小二耗」
江西撫州臨川區溫泉鄉濟南朱公（1140-1197）墓陶瓷俑[36]		「□□童子」？	「張堅固」、「李定度」、「張仙人」、「王公」、王母、「指路」、「引路」、帶獸俑「王」1、「東□□□」、「南□」、「中□」、「□

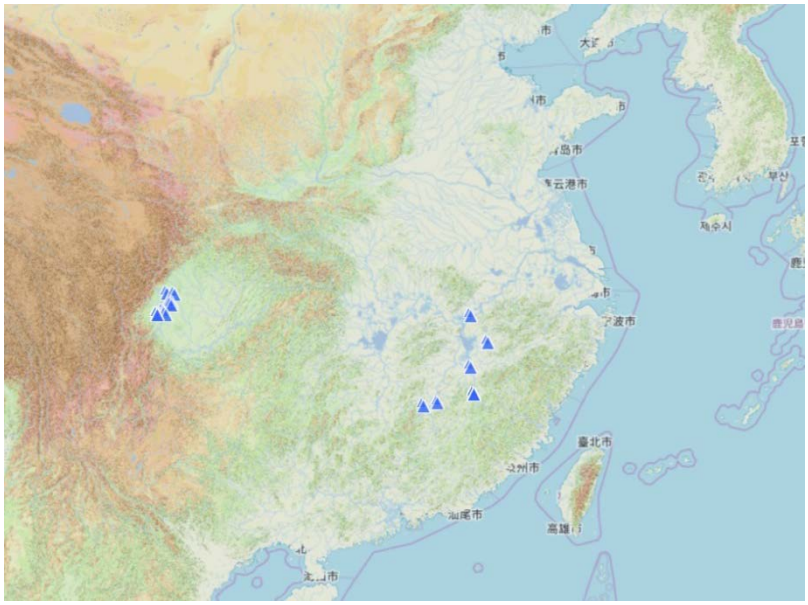
[31] 不著撰人，《大漢原陵秘葬經》，永樂大典本，收入藏外道書（成都：巴蜀書社，1994），第 1 冊，頁 159。
[32] 徐蘋芳，〈唐宋墓葬中的「明器神煞」與「墓儀」制度〉，頁 91-94。
[33] 成都文物考古研究所、新津縣文物管理所，〈新津縣邛雙鄉北宋石室墓發掘簡報〉，《成都考古發現》，2002 年，頁 384-401。
[34] 陳定榮，〈論江西宋墓出土的陶瓷俑〉，《江西歷史文物》，1986 年 S1 期，頁 89-95、88。
[35] 江西省文物工作隊、南豐縣博物館，〈南豐縣桑田宋墓〉，《江西歷史文物》，1986 年 1 期，頁 28-43；陳定榮，〈江西南豐縣桑田宋墓〉，《考古》，1988 年 4 期，頁 318-28。
[36] 臨川縣文物管理所，〈臨川溫泉鄉宋墓〉，《江西歷史文物》，1986 年 2 期，頁 43-48；陳定榮、徐建昌，〈江西臨川縣宋墓〉，《考古》，1988 年 4 期，頁 329-34。

			禮□長」、「□□持□神」
甘肅武威西夏劉德仁（1131-1198）墓木板彩畫[37]	「蒿里老人」、「天關」、「金雞」	「童子」、「二童子」	「大六」、「太陽」、「南柏人呼北柏人」、「柏人」
廣東海康水鬼嶺元墓出土磚刻神煞[38]	「地軸」、「覆聽」、「蒿里父老」、「金雞」、「玉犬」	「青龍」、「朱雀」、「玄武」	「勾陳」、「墓門判官」、「張堅固」、「左屈客」、「右屈客」、「東叫」、「西應」、「喚婢」、「川山」、「伏屍」

根據以上題名例證，形象確定的神煞俑有《秘葬經》中的地軸、伏聽、蒿里老人、金雞、玉犬，與《地理新書》的「童子」，其餘可能是各地的區域葬俗。上述六例，只有一例出土於甘肅武威，墓主劉德仁原籍彭城（江蘇徐州），有學者推測這種作法應該是按照墓主故鄉的傳統葬俗。[39]

神煞俑當中，十二神與伏聽最為常見。十二時神俑，《唐會要》僅限品官使用，庶人不置，至《秘葬經》也適用一般庶人。伏聽俑在唐代的遼寧朝陽、山西長治、陝西西安等地均可見，宋元時期持續使用。若以十二神與伏聽為指標，北宋神煞陶俑的出土地點集中於四川、江西二地，南宋延伸至福建福州。見以下分佈圖：

圖 3a：兩宋十二神與伏聽分佈圖（北宋）

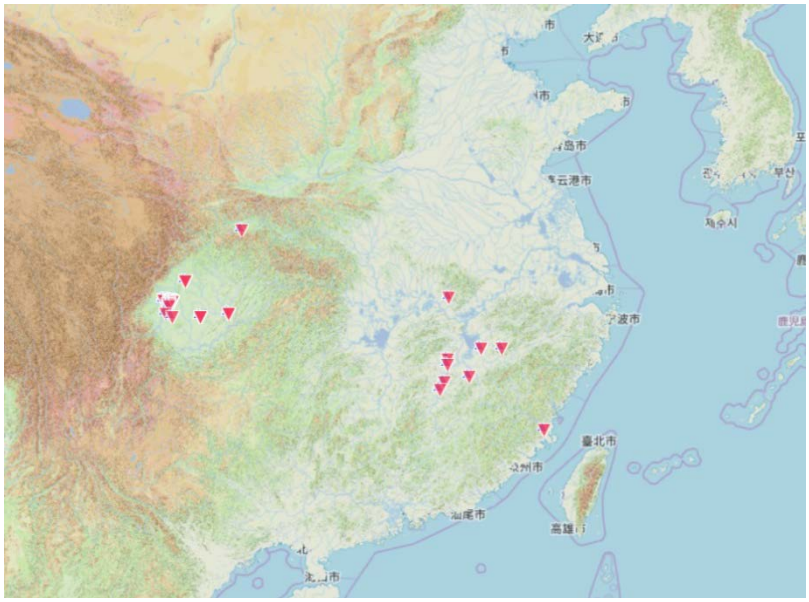


[37] 甘肅武威地區博物館，〈甘肅武威西郊林場西夏墓清理簡報〉，《考古與文物》，1980 年 3 期，頁 63-66；于光建，〈武威西郊西夏 2 號墓出土木板畫內涵新解〉，《西夏研究》，2014 年 3 期，頁 67-72。

[38] 曹騰騑、阮應祺、鄧杰昌，〈廣東海康元墓出土的陰線刻磚〉，《考古學集刊》，2（1982），頁 171-80。

[39] 張勛燎、白彬，〈隋唐五代宋元墓葬出土神怪俑與道教〉，《中國道教考古》，頁 1676。

圖 3b：兩宋十二神與伏聽分佈圖（南宋）



要特別指出的是，唐代三彩明器流行的陝西、河南一帶基本已不用陶俑隨葬，很可能與造紙普及，紙紮明器流行有關。《東京夢華錄》記載開封有紙馬舖，[40]馬可波羅於十三世紀末到杭州，也見到喪禮時「取紙製之馬匹、甲冑、金錦等物並尸共焚之」，可知至少在兩宋首都地區紙明器的使用代替了傳統陶俑。[41]

四川與江西特別盛行以明器神煞俑陪葬，可能與當地的道教流行有關，特別是救濟亡靈的黃籙齋。舉行齋儀時，要請求諸神靈協助，包括：勾陳、東王公、西王母、太陽帝君、太陰星君、五星真君、五月蒿里相公、二十八宿，這些神靈也常見於南方神煞俑。[42]從江西出土例證可知神煞俑在墓中的安排：墓主人一般位於墓室後壁，與四靈（青龍、白虎、朱雀、玄武）居上層龕位，十二時神則分居下層，環繞男室周圍。[43]自上而下，神祇將塚墓層層包圍，讓墓室自成一個小宇宙，以安鎮亡靈。[44]

4. 鐵豬與鐵牛

鐵豬與鐵牛出現於晚唐的洛陽與長安兩京地區，一般認為，當鎮墓獸與天王、武士俑□□也就是「四神」□□消失之際，鐵牛、鐵豬興起並取而代之。長

[40] 孟元老撰、伊永文校注，《東京夢華錄》（北京：中華書局，2006），頁 626。
[41] 馬可波羅著，馮承鈞譯，《馬可波羅行紀》（北京：東方出版社，2011），第 151 章，頁 373。
[42] 張勛燎、白彬，〈隋唐五代宋元墓葬出土神怪俑與道教〉，頁 1742-1750。
[43] 陳定榮，〈論江西宋墓出土的陶瓷俑〉，《江西歷史文物》，1986 年 S1 期，頁 90。
[44] 王洙等撰，《圖解校正地理新書》，頁 441-442。

安地區發生的時間在九世紀，河南地區稍早。[45]陶明器於二京地區消失的原因還不清楚，有認為被紙明器取代，也應考慮唐晚期動亂造成二京殘破這個巨變。

以鐵豬、鐵牛陪葬之習俗，見於晚唐記錄。劉肅於《大唐新語》（序於 807）記錄一則唐玄宗開元十五年（727）之事：

> 集賢學士徐堅（？-729）請假往京兆葬其妻岑氏，問兆域之制於張說（667-730）。說曰：「……長安（701-704）、神龍（705-707）之際，有黃州僧泓者，能通鬼神之意，而以事參之。僕常聞其言，猶記其要：『墓欲深而狹，深者取其幽，狹者取其固。平地之下一丈二尺為土界，又一丈二尺為水界，各有龍守之。土龍六年而一暴，水龍十二年而一暴，當其隧者，神道不安。……鑄鐵為牛豕之狀像，可以禦二龍。玉潤而潔，能和百神，實之墓內， 以助神道。』僧泓之說如此，皆前賢所未達也。[46]

這條資料提到的張說、徐堅與僧泓均活躍於盛唐時期之長安，徐堅葬妻時，當時的宰相張說（667-730）向他轉述僧泓之葬法。根據僧泓，在墓中置鐵豬、鐵牛，可鎮壓地底下的土龍與水龍，避免二龍暴起，導致地下諸神不安。由此條記錄可知，鐵豬、鐵牛陪葬之法，自八世紀開始，在僧泓的提倡下，開始流行於京兆地區的達官貴人之間。

僧泓，來自齊安（唐嶺南道），是中、晚唐之際活躍於首都長安地區的僧人，善擇陰陽宅與葬法，被稱為中宗、睿宗之「國師」，當時的達官顯要頗重其說。[47]宋初成書的《太平廣記》收錄多條僧泓為達官貴人相地之事，《地理新書》中也收錄泓師說法，可見僧泓之說對晚唐至北宋的上層階級具有影響力。[48]

五代時期，鐵豬與鐵牛出現在規格相當高的墓葬中：陝西寶雞市陵塬鄉陵塬村大唐秦王李茂貞（856-924）墓與四川成都前蜀王建（847-918）墓。北宋皇陵尚未發掘，不知是否也隨葬鐵豬與鐵牛。整理已發布的資料，隨葬鐵牛、鐵豬的宋墓共三十餘座，分佈如下：

圖 4a、兩宋鐵牛鐵豬分佈圖（北宋）

[45] 鄭州市文物考古研究所，《中國古代鎮墓神物》（北京：文物出版社，2004），頁 21；程義，《關中地區唐代墓葬研究》（北京：文物出版社，2012），頁 146-150。

[46] 劉肅撰，許德楠、李鼎霞點校，《大唐新語》，唐宋史料筆記叢刊（北京：中華書局，1984），卷 13，頁 195。

[47] 僧泓傳見劉昫，《舊唐書》（臺北：鼎文，1985），卷 191，頁 51130；大藏經刊行會，《宋高僧傳》，收入《大正新脩大藏經》（臺北：新文豐，1983），第 50 冊，頁 889-890。

[48] 李昉，《太平廣記》（北京：中華書局，1961），第 8 冊，卷 389，頁 3108-3109；王洙等撰，《圖解校正地理新書》，頁 457。

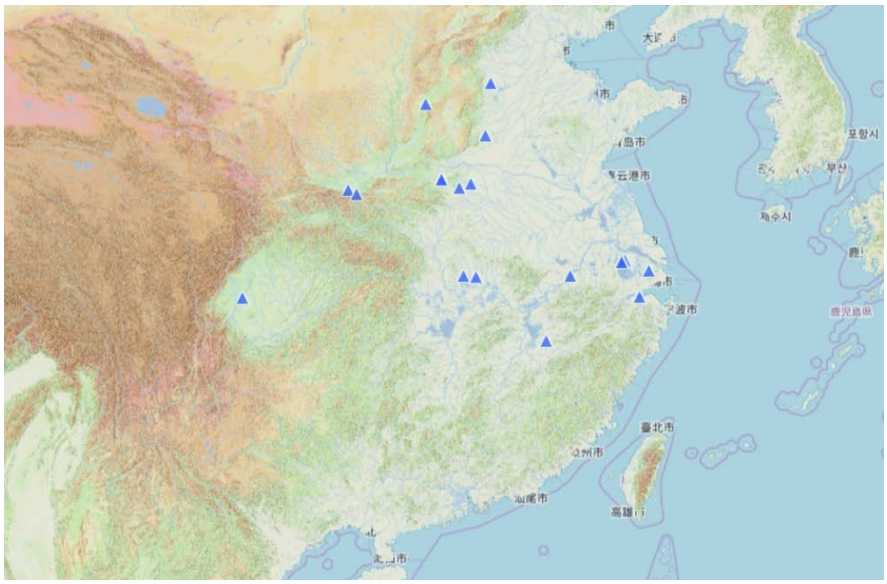
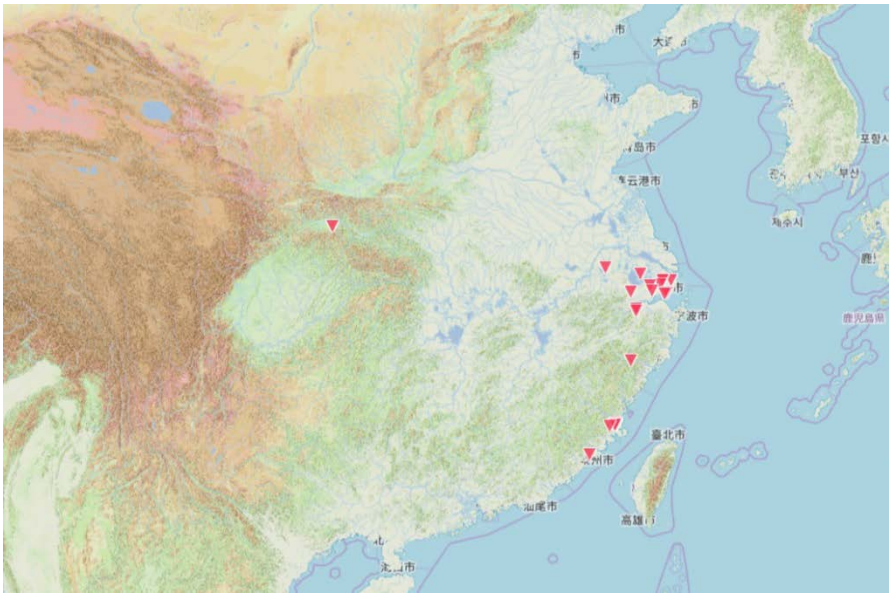


圖 4a、兩宋鐵牛鐵豬分佈圖（南宋）



由此分佈圖可知，北宋不少地區均出現鐵豬、鐵牛，包括華北的河南、陝西、山西，與長江中、下游的湖北、安徽、江西、江蘇、浙江等地。進入南宋，只有一例出現在陝西漢中（南宋利州路），其餘均在上海、浙江、江蘇、福建，有向東南地區集中的趨勢。其中以福建地區最值得注意，此地在北宋時沒有鐵豬鐵牛陪葬之例，南宋時期開始出現這種葬俗，並且集中在福州與泉州，應該是從外地傳入，可能與兩宋之際，人群大規模南遷有關。

分析墓主的社會身分，出土鐵豬、鐵牛的三十餘座墓中，有明確紀年訊息者二十四座，其中北宋七座，南宋十七座。北宋七座紀年墓中，五座來自士大夫官

僚家庭，包括：江蘇常州胡宗愈（1029-1094）墓、[49]湖北安陸時知默（1061-1103）墓、[50]江西鄱陽施氏（1042-1110）墓、[51]陝西藍田呂大臨（1040-1093?）夫婦墓、[52]河南安陽韓治（?-1124）夫婦墓。[53]其中品秩最高者是胡宗愈，累階至從二品右銀青光祿大夫；韓治與呂大臨官階雖沒那麼高，但兩人都來自宰相家庭，分別是仁宗宰相韓琦（1008-1075）的長孫與元祐宰相呂大防之弟。至於少數沒有任官的墓主，包括上海嘉定縣城廂鎮的趙鑄（984-1062）夫婦墓[54]與湖北省孝感的杜氏一娘（?-1120）墓，從墓誌可知他們是地方上的經濟菁英。[55]

南宋的十七座紀年墓中，墓主本身具有官僚身份者有十四座，品秩自被追贈太師者（浙江麗水 1209 年何倜墓[56]），到高階文官（1226 年福建福州故吏部尚書朱著墓[57]）、中階武官（1192 年陝西漢中故武功大夫彭杲墓[58]），再到從九品的迪功郎（福建福州 1208 年陳元吉夫婦墓[59]）或承信郎（江蘇吳江 1195 年葉奘夫婦墓[60]、上海 1213 年張珒夫婦墓[61]）皆有。有的來自士人官僚家庭，如江蘇南京 1271 年周國太夫人楊善慶墓、[62]1186 年福建泉州恭人蔡氏墓。[63]其餘三座墓的墓主或不屬官僚階層、或仕宦履歷不明，不過從墓葬形制與隨葬品來看，應該都屬地方菁英，包括：上海浦東黃俁夫婦墓、福建福州嘉定三年（1210）墓，以及浙江嘉興陶違父子墓。

從北宋到南宋，當隨葬鐵豬、鐵牛的習俗從北方向南方傳播之時，也出現簡化的情況，先是從鐵豬、鐵牛簡化成只有鐵牛，接著從一組四件簡化成以一件代替。北宋鐵豬、鐵牛多為一組四件，放於墓中四角，傳達出鎮墓的功能。標準的葬式為鐵豬、鐵牛各二，包括：韓治夫婦墓、杜氏一娘墓；也有四件全為鐵牛者，包括：施氏墓、趙鑄夫婦墓。南宋時期幾乎不見鐵豬，全以鐵牛陪葬，江浙地區

[49] 陳晶，〈宋胡宗愈墓誌〉，《文博通訊》，1981 年 5 期，頁 53-55。
[50] 安陸縣文化館，〈安陸毛家山一號宋墓清理簡報〉，《江漢考古》，1983 年 1 期，頁 63-64。
[51] 余家棟，〈江西波陽宋墓〉，《考古》，1977 年 4 期，頁 286。
[52] 陝西省考古研究院等編，《異世同調：陝西省藍田呂氏家族墓地出土文物》（北京：中華書局，2014），頁 226-229。
[53] 安陽市文物考古研究所等，〈河南安陽市宋代韓琦家族墓地〉，《考古》，2012 年 6 期，頁 41-53；河南省文物局，《安陽韓琦家族墓地》（北京：科學出版社，2012），頁 38-48。
[54] 何繼英編，《上海唐宋元墓》（北京：科學出版社，2014），頁 46。
[55] 孝感市文化館，〈湖北孝感大灣吉北宋墓〉，《文物》，1989 年 5 期，頁 69-70。
[56] 浙江省文物管理委員會，〈麗水青瓷調查發掘記〉，《浙江省文物考古研究所學刊》，第 7 輯（2005），頁 509-537。
[57] 福建省博物館，〈福建福州郊區清理南宋朱著墓〉，《考古》，1987 年 9 期，頁 796-802。
[58] 李燁、周忠慶，〈陝西洋縣南宋彭杲夫婦墓〉，《文物》，2007 年 8 期，頁 57-70。
[59] 張煥新，〈福建博物院藏南宋陳元吉墓出土器物〉，《文物》，2011 年 7 期，頁 71-84、87。
[60] 蘇文，〈江蘇吳江出土一批宋瓷〉，《文物》，1973 年 5 期，頁 68。
[61] 何繼英編，《上海唐宋元墓》（北京：科學出版社，2014），頁 53-58。
[62] 南京市博物館、江寧區博物館，〈江蘇南京南宋周國太夫人墓〉，《東南文化》，2010 年 4 期，頁 54-59。
[63] 王洪濤，〈泉州、南安發現宋代火葬墓〉，《文物》，1975 年 3 期，頁 77-78。

墓葬仍維持一組四件，安置於墓室或木槨四角；福建福州地區的墓葬多僅陪葬一件鐵牛，作為象徵，包括福建福州的陳元吉、朱著、許峻墓，而且經常與明器神煞俑共出。

三、綜合討論

自漢代以降，對於亡靈不安的焦慮，使得鎮墓文物普遍出現在各地墓葬當中，使用者從帝王到一般平民，幾乎涵蓋所有的社會階層。隨著宗教信仰與儀式的發展，遺留在墓葬中的文物也有所不同。本研究討論的四類鎮墓文物：地券、五色石、明器神煞陶俑、鐵豬鐵牛，為宋墓常見之物，過去也有不少研究，不過多半僅針對其中一類文物進行討論，難以掌握整體面貌。由於「遼宋金墓葬資料庫」的建置，使得跨越文物類別的大規模分析成為可能。本研究首先按類別整理這些文物在兩宋的分佈，觀察分佈特點與變化；接下來利用墓中出土的墓誌銘與隨葬品，分析使用者的社會階層與移動軌跡；最後觀察文物在墓中的隨葬特點，探討不同鎮墓習俗的長期發展，以及傳播至不同地域時的流變。

整體而言，這些鎮墓文物絕大部分出土於中等以上規模的墓葬，有能力營建這類墓葬的家庭，多半具有一定財力與社會地位，有些墓主屬士大夫官僚階級，有些沒有官職，是地方上的地主或富商豪強。當中鐵豬鐵牛的使用者平均身份最高，主要是士人官僚家庭，包括高階士大夫。除了鐵豬鐵牛，士大夫官僚階級也常使用五色石，使用者身份最高的是馮京（1021-1094），不過五色石的使用並不限於官僚階級，也見於華北地區的平民墓中。

這四類文物中，數量最多、分布最廣的是地券，使用者多為平民階層，也有一些士人官僚家庭，以四川出土數量最多，其次是江西。平民階層的地券一般書寫在磚、石之上，鐵地券的使用者大半為士人官僚。出土明器神煞陶俑的墓葬數量，僅次於地券，不過分佈地點集中在四川與江西，南宋時期福建的墓葬也有使用，屬於地區性的葬俗，使用者以平民為多。

觀察文物的共出情況，鐵豬鐵牛有時與五色石、鐵地券共出，均較常見於士人與官僚階層墓中，使用者社會身份較高。地券的使用者社會身份較低，而且在四川與江西地區，經常與明器神煞陶俑共同出土，應該是這兩個地區中間階層的陪葬組合。

以地理分佈而言，鎮墓文物從北宋至南宋似有集中的趨勢，以最普遍的地券為例，南宋明顯集中至四川、江西、杭州與福建沿海，其餘地區出土零星。另外，觀察鐵豬鐵牛、五色石、鐵地券於兩宋之分佈，明顯可見從中原地區向長江下游與福建傳播的趨勢，這些文物的使用者身份較高，應該與兩宋之際趙宋皇室與北方士大夫家族南遷有關。這個推測可從福建泉州蔡氏（1134-1161）墓得到佐證。蔡氏是宗室趙士琣之妾，趙士琣在宋金戰爭時有功於朝廷，紹興年間知南外宗正

事，管理趙姓宗室，生平見《宋史》。[64]蔡氏為泉州人，應該是趙士琣遷居泉州後所納，紹興三十一年（1161）卒，淳熙十三年（1186）葬於福建泉州南安，墓中出土墓誌銘、鐵鑄陽文地券、小鐵牛四。[65]鐵地券的內文與《地理新書》幾乎全同，買地的見證人「歲月主」與保人「直符」，也符合《地理新書》記載。由此可知，泉州人蔡氏的墓中出現《地理新書》規範的鐵地券以及華北地區使用的鐵牛，應該與其身為趙士琣侍妾有關，是北方習俗經由趙宋宗室帶到福建的有力例證。

宋金戰爭時，除了大量北方人口南遷之外，也有不少人群向華北其他地區遷徙，應該也同樣帶動喪葬文化的改變。接下來也將把金代納入討論，以釐清十二至十三世紀的整體發展圖像。

[64] 脫脫，《宋史》，卷 247，頁 8752-8753。
[65] 王洪濤，〈泉州、南安發現宋代火葬墓〉，《文物》，1975 年 3 期，頁 77-78。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

基於本體的家譜數据可視化構建研究

以浙江“仙居高遷《吳氏西宅宗譜》”為例

祝振媛　梁继红

中國人民大學信息資源管理學院

會議論文

基於本體的家譜数据可視化構建研究
----以浙江“仙居高遷《吳氏西宅宗譜》”為例

祝振媛，梁继红

（中國人民大學信息資源管理學院 ）

摘　要：結合數字技術和傳統家族歷史文化解析，將傳統家譜轉變為立體和多維整合的數字資源，将是當前家譜整理和研究數字化轉向的時代趨勢，尤其通過家譜中人物數據可視化可以更直觀展現傳統家族歷史文化。本研究以浙江“仙居高遷《吳氏西宅宗譜》”作為研究素材，構建吳氏家譜本體，探究本體技術在傳統家譜視覺化展示研究中的可行性。

关键词：家譜；可視化；本體構建；仙居高遷；吳氏宗譜；

Study on Visualization of Genealogies Based on Ontology----Deployed on the Genealogies of Wu’s in Gaoqian,Zhejiang

Zhu zhenyuan， Liang Jihong

（ School of Information Resource Management, Renmin University of China）

Abstract：Combined with the analysis of digital technology and traditional family history and culture, the transformation of traditional genealogy into three-dimensional and multi-dimensional integrated digital resources, will be the trend of genealogy research. Especially through the visualization of figures in the genealogy, it will be more intuitive to display the traditional family history and culture, This research takes Gaoqian "Wu's West Residence genealogy" as the research material, constructs the Wu's genealogy ontology, explores the research feasibility of ontology technology in the traditional genealogy visualization demonstration.

Key words：Genealogies；Visualization； Ontology construction；Gaoqian；Genealogies of Wu’s

1 引言

家譜記載一個家族的繁衍生息的歷史，以及社會關係、風俗習慣等地方史內容，蘊含著豐富的歷史資訊。家譜對一個傳統家族而言，首先是一個制度化的文本，規範族人的倫理秩序和日常生活，凝聚家族認同的力量，使之跨越時間和空間的阻隔。一個家族世代相承的精神文化是家譜的重要內容。它直接體現在家譜所載的家規、家訓、家禮之中，更體現在家族

先賢的行為事蹟之中，而先賢事蹟為後世子孫所樹立的楷模作用，與純粹的言說相比，更為深切著明。因受紙媒的制約，傳統家譜中的世系、世系圖和傳記等人物資料不得不分篇展開，其記載方式是平面和分列式的。

傳統家譜蘊含著大量的家族相關的資訊，承載著家族的歷史、精神、文化，家譜的內容有著值得挖掘的多種層面。我們現代人在面對古代傳統家譜時，一個核心的問題是如何將這些紙質的、平面的、局部的、單緯度的資料轉變為數位的、立體的、全面的、歷時性、共識性的多維度的資源。結合數字技術和傳統家族歷史文化解析，將傳統家譜轉變為立體和多維整合的數字資源，尤其是通過家譜中人物數據可視化，更直觀展現傳統家族歷史文化，是當前家譜整理和研究數字化轉向的時代趨勢。

同時，中國傳統大家族傳承千年，脈細深遠，家族中不乏產生不同領域中的精英人物，這些精英文化的展示與傳承也是一個普遍的文化現象，值得研究。本研究旨在借助本體構建技術，通過展示家譜中的代系人物與事蹟，挖掘家譜中所反映的社會關係，實現傳統家譜的視覺化展示，並突出展示家族中的精英人物，挖掘家族精英文化。

浙江省台州市仙居縣吳氏家族，自五代（梁）開始，子孫綿延千年，是典型的江南望族。這個家族歷史上曾湧現北宋龍圖閣直學士吳芾、南宋左丞相吳堅、明代左都御史吳時來等精英人物。本研究以高遷《吳氏西宅宗譜》作為研究素材，構建吳氏家譜本體，呈現吳氏家族中精英人物的源流，探究本體技術在傳統家譜視覺化展示研究中的可行性。

2 家譜本體的構建設計

本體是關於特定主題的規範說明，它把現實世界中某個應用領域抽象概括成一組概念及概念之間的關係，形成該領域共用的、概念化、形式化表示的知識體系，本體的構建可以極大地說明電腦對領域的知識進行處理【1】。

根據不同的分類視角，本體具有多種分類方式：依照領域依賴程度，本體可以分為頂級本體（Top-level ontology）、領域本體（Domain ontology）、任務本體（Task ontology）和應用本體（Application ontology）四類；其中，領域本體描述特定領域中的概念及概念之間的關係，是專業性的本體，用以描述特定學科領域的知識。按照是否具備推理功能，本體可以劃分為羽量級本體（Lightweight ontology）、中級本體(Middle ontology)和重量級本體(Heavyweight ontology)；其中，中級本體具有簡單的邏輯推理功能，系統可以識別一階謂詞邏輯的運算式。

本文通過對仙居高遷吳氏西宅宗譜中的世系內容與傳記內容進行分析，總結家譜中主要要素的以及要素之間的關係，構建家譜領域中級本體，通過對家譜中相關概念及概念間關係的描述實現對家譜領域知識的表示和組織。在構建之前需要對家譜本體的需求分析、設計原則以及構建流程進行闡釋。

2.1 家譜本體的需求分析

家譜在當下面臨著資源整理與數位化的要求，業界已經形成了一些家譜總目、家譜網站、家譜資料庫，而這些大多是將家譜進行掃描成圖像，文字進行數位化全文，不可否認這些工作有著重大的價值和意義，是家譜文獻資源保護和挖掘的起點。可是不能止於此，我們希望挖掘傳統家譜中蘊藏的大量文獻價值、歷史價值、社會學價值。本體作為一種知識組織的有效方式，具有層次性、繼承性、擴展性、聯結性等特點，使其成為家譜整理與展示的有效工具。 本體能夠反映相互關聯的多種關係，這種特性可以很好地將家譜所記載的家族資訊反映出來，包括家族在一定時期的社會、政治、經濟、軍事、文化、教育等諸多方面的情況。

2.2 家譜本體設計原則

在設計本體時遵循相應的原則是保障本體科學構建的前提，本文的家譜本體在構建過程中將遵循以下幾個原則：

- 家譜本體中的術語詞彙應儘量專業化，符合家谱整理與分析的工作實際。
- 本體構建過程中應對相關術語做出明確的定義。
- 具體類的下面不應有抽象類別。
- 不同層級類的整體分佈應該比較均勻。
- 類的命名唯一。
- 本體應具有可擴展性，可以添加家譜中的新術語。

2.3 家譜本體構建流程

當前比較成熟的本體構建方法主要包括“骨架法”【2】、TOVE 法”【3】、METHONTOLOGY 法【4】、KACTUS 法【5】、SENSUS 法【6】、IDEF5 法【7】、AFM 法、七步法【8】等，不同的方法在構建流程上存在一些區別，在綜合比較以上幾種本體構建方法的適用範圍與流程特點的基礎上。本文以斯坦福大學醫學院開發的本體構建的七步法為基礎，借鑒已有的相關研究的七步法框架，並進行調整，構建家譜本體，構建流程如圖 1 所示。

如圖 1 所示，首先對高遷《吳氏西宅宗譜》紙本作圖像和全文本加工處理，借助於古籍識別軟件進行文字轉錄，經標點斷句等基本古籍整理，形成進行家譜知識整理和分析的素材，此為預處理的階段。本體的具體構建流程包括 1）明確家譜的知識範疇，明確家譜本體的用戶及需求 2）從高遷《吳氏西宅宗譜》的世系圖、傳記、仙居地方誌等資料中提煉出人物傳記資料的知識內容 3）梳理相關資料中包含的家譜術語，構建術語表 4）對家譜術語表中的術語進行分析與歸類，確定類中不同級別的內容 5）定義家譜中不同類的屬性，描繪概念間的關係 6）借助 protégé 作為本體編輯工具創建家譜本體中的類、屬性、關係等 7）借助 protégé 的 ontograf 實現家譜本體視覺化。

在最終的本體編輯與視覺化方面，本文選用 protégé 4.3 作為本體編輯工具創建家譜領域本體，protégé 4.3 這個版本的優點在於具有 ontograf（本體關係圖）功能模組，可以支援中文圖像結果的視覺化顯示。家譜本體借助 protégé 中的 ontograf 本體關係圖可以將家譜中的每一個世系節點根據需要而展開，節點中包含實例，實例之間通過對象屬性產生聯繫，形成家族內部或外部的各種關係，達到突出展示家族中精英人物的目的。

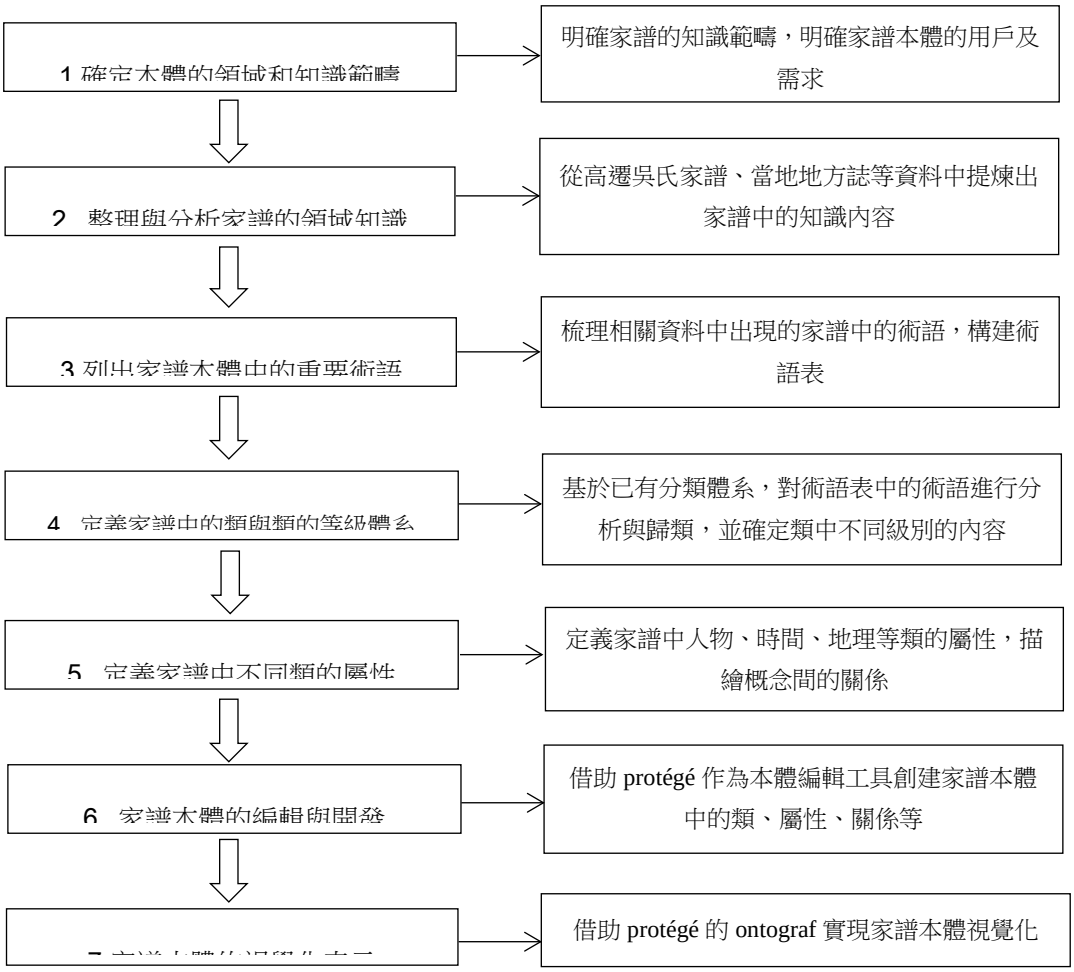


圖 1　家譜本體的構建流程

3　家譜知識組織體系構建

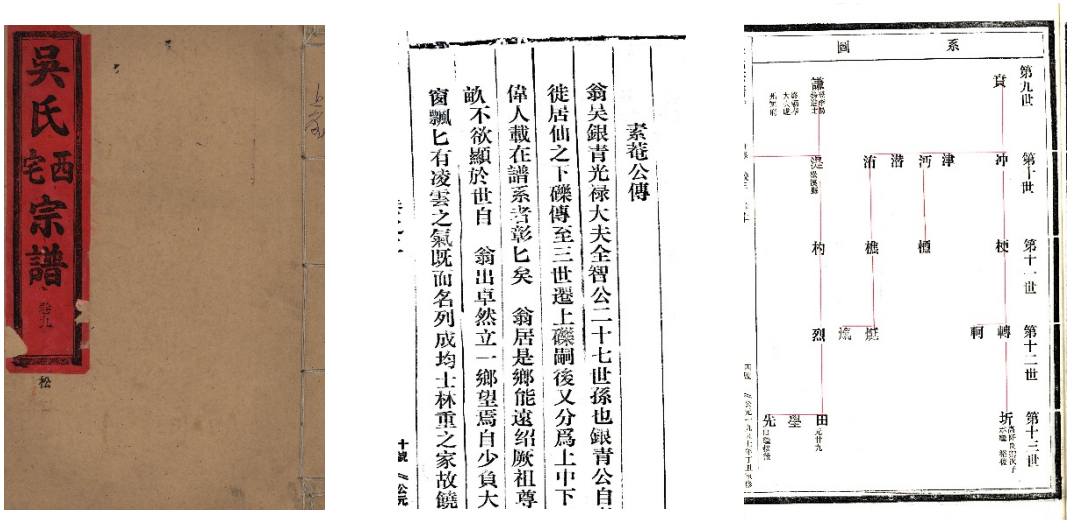
在本體構建前，需要從家谱知识的搜集与整理、家譜本體概念體系、家譜概念間關係的建立三方面完成家譜知識組織體系本體的構建。

3.1　家譜知識的搜集与整理

首先將高遷村委會提供的《吳氏西宅宗譜》進行掃描，形成圖像文件；其次，對掃描圖像進行圖片處理，利用 OCR 文字識別軟體對掃描影像檔進行文字識別轉錄；最後進行校對、

標點斷句等基本古籍整理流程，形成最後的進行家譜知識整理和分析的族譜文本。吳氏宗譜的前 3 卷中為吳氏家族中的精英人物的傳記，而後是吳氏的世系圖，圖 2 為《吳氏西宅宗譜》的書影，包括傳記部分和世系圖部分節選。

图 2　《吳氏西宅宗譜》的書影



3. 2　家譜本體的概念體系构建

概念的清晰界定是本體研究中必不可少的內容，概念體系的建立是構建領域本體框架的首要步驟。家谱本體構建中涉及人物、時間、地理資訊、社會身份、世系等多個大類，需要對這些類中的概念進行界定，建立家譜本體的概念體系。

基於家譜中重要要素的類型分析和家譜術語表構建，本文的家譜本體中的概念分為人物、世系、社會身份、德目、時間、地理資訊等 6 大部分或稱基本類，其中的概念皆是基本類的成員或實例。構建的家譜本體的重要概念體系如圖 3 所示。

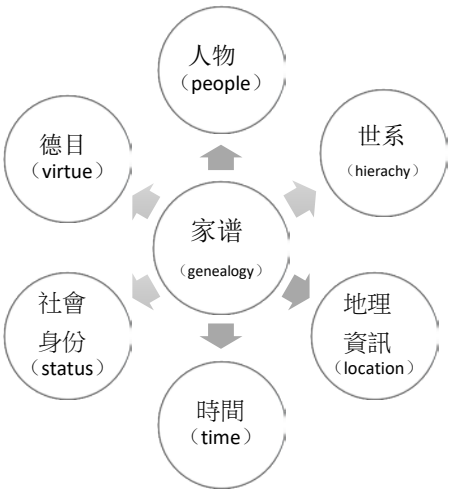


圖 3 家譜本體概念體系圖

大類概念可以根據需要再細分為不同的子概念類，如“社會身份”類可以細分為“功名、

職官、鄉賢、夫人”等子類，其中地理資訊類還可以再繼續細分為“住址、任職地、流放地”等多個子類，圖 4 為六大類概念的基本類示例。六個大類的定義和包含子類情況如下：

（1）人物（people）：人物是家譜記載的主要內容之一，是家族發展的主體，也是家族研究描述分析的主要物件。家譜中記載的人物主要包括傳記主人及家族相關人物等， 包括男丁、傳主、譜外相關人物等基本類。

（2）世系（hierachy）：世系是家譜記載的主要內容之一，是家譜中最主要的部分，展現了一個家族延續發展的脈絡，從第一世開始往後計代。

（3）社會身份（status）：社會身份是家譜中記載人物在社會中擔任的角色以及取得的成就、榮譽，是描述人物的重要一項。社會身份包括人物所取得的功名、擔任的職官以及所獲得的榮譽等，包括功名、職官、鄉賢、夫人等基本類。

（4）時間（people）：時間是指族譜中人物的生卒年及所處朝代，對人物進行描述及時間定位，包括後樑、後唐、宋、元、明、清、民國等基本類。

（5）地理資訊（hierachy）：地理資訊是指族譜中人物所處地理位置的描述，包括了居住地、任職地等資訊，包括居住地、任職地、流放地、墓地等基本類。

（6）德目（status）：德目是對一個人所具備的道德品質特徵的歸納，體現人物的品行，包括循吏、忠義、孝友、為善、誠信、剛勇、勤儉、貞正、勤勉、自律、隱德等基本类。

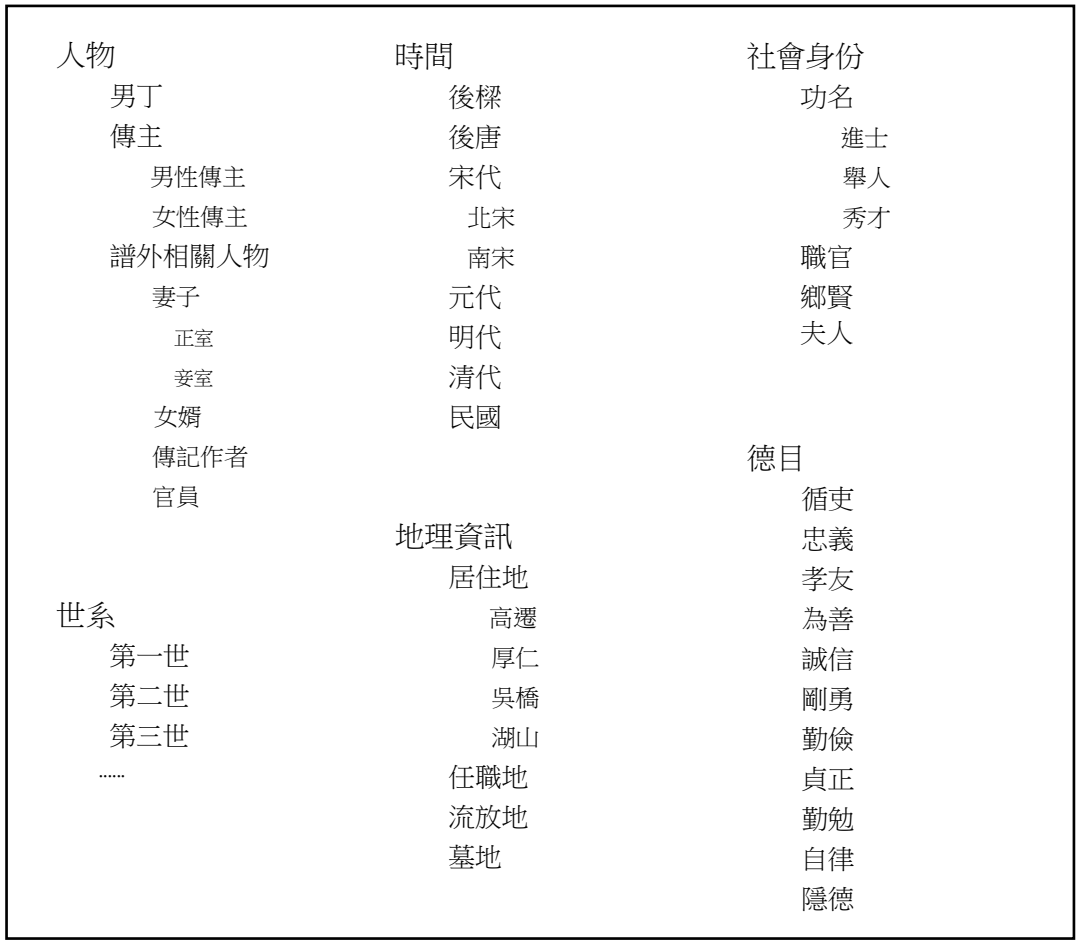


圖 4　家譜本體的基本類示例

3.3　家譜概念間關係的建立

在本體構建中，除了類之外另外一個重要的組成部分就是屬性（Properties），一方面表示實例的基本資料情況，一方面表示個體之間的關係，是連結實例间的“紐帶”。本體中有兩種主要的屬性類型，一種是對象屬性（Object Properties），描述實例與實例之間的關係；一種是數據類型屬性（Datatype Properties），描述實例與基本數據類型之間的關係。屬性的定義可以實現本體中概念之間的關係映射，基於對人物、社會身份、時間、地理資訊等本體中的概念的屬性分析，本文構建的家譜本體主要包含 24 種關係，包括數據類型屬性（如“字”、“號”）7 種與對象關係（如“出生於”、“排行”等）17 種，如圖 5 所示。具體的關係與實例如下：

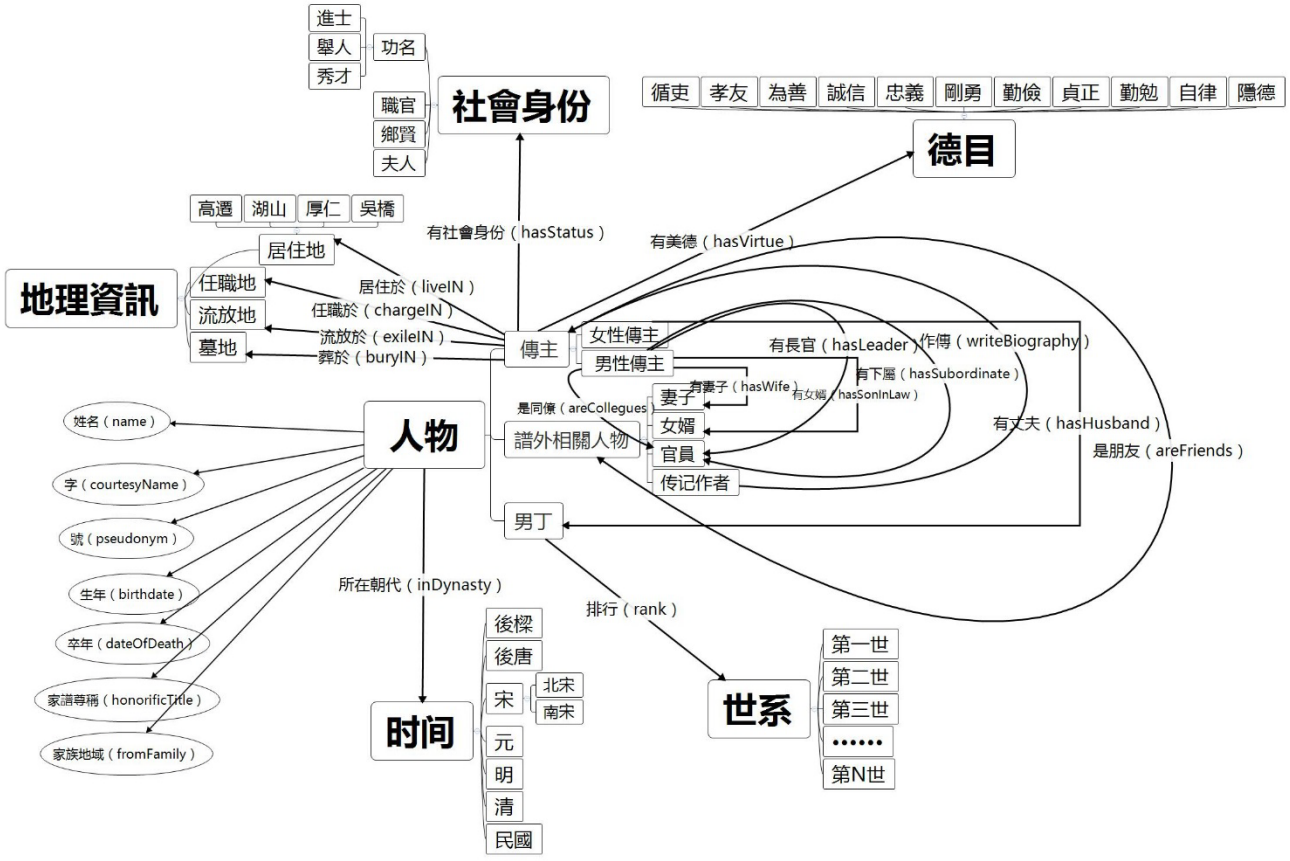


圖 5　家譜本體概念及語義關係

1）數據類型屬性

① 姓名（name）：描述“人物”的姓名，實例如下：

- 朱熹_姓名：朱熹

② 字（courtesyName）：描述“人物—傳主”的字，實例如下：

- 吳時來_字：惟修

③ 號（**pseudonym**）：描述“人物—傳主”的號，實例如下：

- 吳芾_號：湖山居士

④生年（**birthdate**）：描述“人物—傳主”的出生時間，實例如下：

- 吳堅_生年：西元 1912 年

⑤卒年（**dateOfDeath**）：描述“人物—傳主”的去世時間，實例如下：

- 吳堅_卒年：西元 1976 年

⑥家譜尊稱（**honorificTitle**）：描述“人物—傳主”在家譜中尊稱，實例如下：

- 吳植楨：靜齋公

⑦家族地域（**fromFamily**）：描述“人物—譜外相關人物”的所屬家族與地域情況，實例如下：

- 張開泰：城西張家

2） 對象屬性

① 排行（rank）：描述“人物-男丁”與“世系”實例如下：

- 吳時來　排行　第二十三世

② 所在朝代（inDynasty）：描述“人物”與“時間”之間的關係，實例如下：

- 吳芾　所在朝代　北宋

③ 有妻子（hasWife）：描述“人物--傳主—男性傳主”與“人物--譜外相關人物—妻子”之間的關係，實例如下：

- 吳炳台　有妻子　王氏

④ 有丈夫（hasHusband）：描述“人物--傳主—女性傳主”與“人物—男丁”之間的關係，實例如下：

- 節婦楊氏　有丈夫　吳廷謀

⑤ 有兒子（hasSon）：描述“人物—男丁”與“人物—男丁”之間的關係，實例如下：

- 吳炳台　有兒子　吳時來

⑥ 有女婿（hasSonInLaw）：描述“人物--傳主—男性傳主”與“人物--譜外相關人物—女婿”之間的關係，實例如下：

- 吳漢儒　有女婿　蔣大楚

⑦ 有長官（hasLeader）：描述“人物—傳主—男性傳主”與“人物--譜外相關人物—官員”之間的關係，實例如下：

- 吳時來　有長官　徐階

⑧ 有下屬（hasSubordinate）：描述“人物—傳主—男性傳主”與“人物--譜外相關人物—

官員”之間的關係，實例如下：

- 吳堅　　有下屬　　賈餘慶

⑨ **是同僚（areCollegues）：**描述“人物—傳主—男性傳主”與“人物--譜外相關人物—官員”之間的關係，實例如下：

- 吳時來　　是同僚　　張居正

⑩ **是朋友（areFriends）：**描述“人物—傳主”與“人物--譜外相關人物”之間的關係，實例如下：

- 吳芾　　是朋友　　朱熹

⑪ **有社會身份（hasStatus）：**描述“人物—傳主”與“社會身份”之間的關係，實例如下：

- 吳堅　有社會身份　南宋左丞相

⑫ **有美德（hasVirtue）：**描述“人物—傳主”與“德目”之間的關係，實例如下：

- 吳敬　有美德　孝友

⑬ **居住於（liveIN）：**描述“人物—傳主”與“地理資訊-居住地”實例如下：

- 吳堅　　居住於　　厚仁

⑭ **任職於（chargeIN）：**描述“人物—傳主”與“地理資訊-任職地”實例如下：

- 吳芾　　任職於　　臨安

⑮ **流放於（exileIN）：**描述“人物—傳主”與“地理資訊-流放地”實例如下：

- 吳時來　　流放於　　橫州

⑯ **葬於（buryIN）：**描述“人物—傳主”與“地理資訊-墓地”實例如下：

- 吳堅　　葬於　　城郊西嚳

⑰ **作傳（writeBiography）：**描述“傳記作者”與“人物—傳主”實例如下：

- 朱位男　　作傳　　吳餘之

通過上述 24 種關係映射可以家譜中涉及的人物、時間、地理資訊、社會身份、世系等要素的實例聯繫起來，形成關係網絡。

4　家譜本體的構建——以吳氏家譜為例

4.1　家譜本體的開發與編輯

Protégé 是由斯坦福大學醫學院生物資訊研究中心基於 Java 語言開發的本體編輯與開發工具[9]，因其介面友好、操作簡單、多種文體支持等優點成為最受歡迎的本體編輯工具之一。本文選用 protégé 4.3 作為本體編輯工具創建情報分析領域本體，protégé 4.3 這個版本的

優點在於具有 ontograf（本體關係圖）功能模組，可以支援中文圖像結果的視覺化顯示，非常適用于傳統家譜的視覺化展示。通過前文的準備工作，利用 protégé 4.3 對情報分析領域本體進行編輯，最終建立的一級類目體系如圖 6 所示：

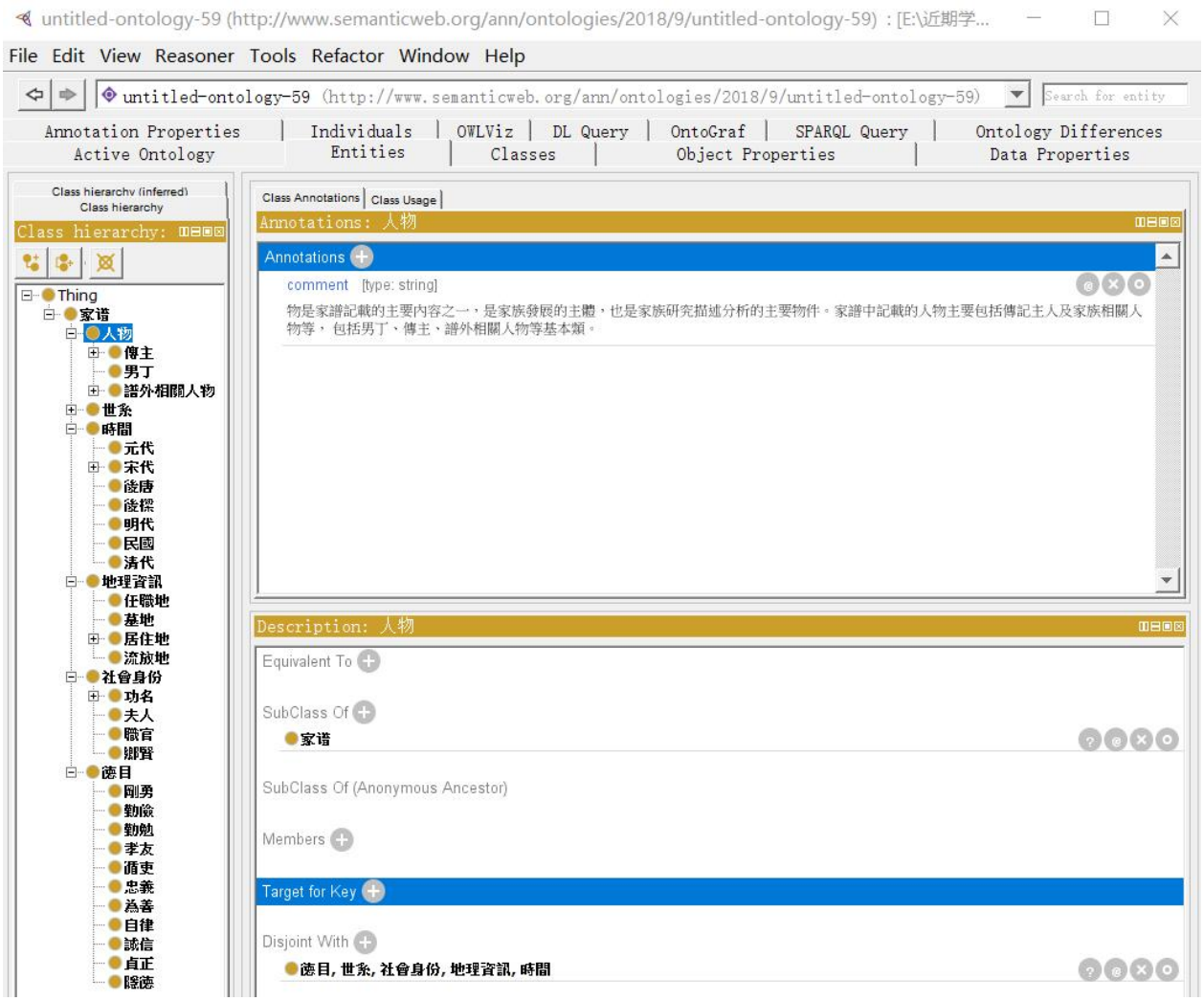


圖 6　protégé 編輯下家譜本體各類一級類目編輯

類與類之間通過 SubClassOf、DisjointClasses、EquivalentClasses 實現類的三大公理，類中的層級關係即為 SubClassOf 關係。DisjointClasses 是指類之間相互不相交，即類之間不存在相同的實例，本文的本體構建中人物、世系、社會身份、德目、時間、地理資訊 6 類為 DisjointClasses。EquivalentClasses 表示類與類之間的等價關係，在本文本體中不涉及 EquivalentClasses。

二級類目如圖 7 所示：

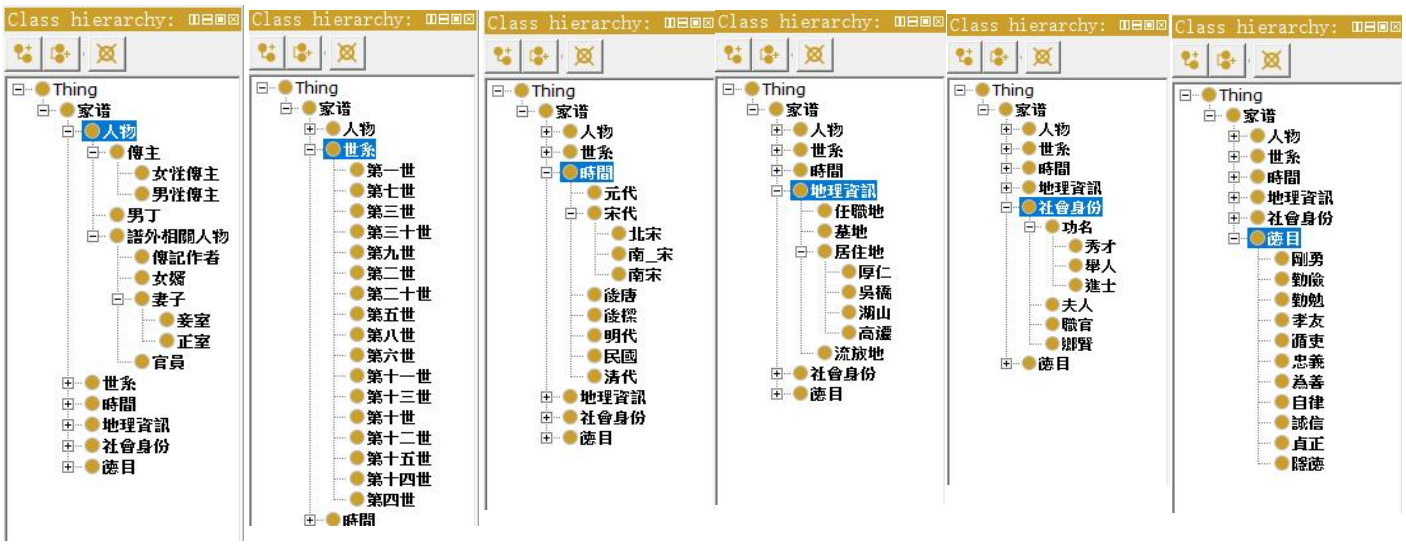


圖 7　protégé 編輯下家譜本體各類的二級類目編輯

本體中除了類之外，另外一個重要的組成部分就是屬性（Properties），用以表示個體之間的關係，是連結兩個實例的“紐帶”。本體中有兩種主要的屬性類型，一種是物件屬性（Object Properties），描述實例與實例之間的關係；一種是資料類型屬性（Datatype Properties），描述實例與基底資料型別之間的關係【10】。基於上文構建的關係，本體的物件屬性編輯面板如圖 8 所示。

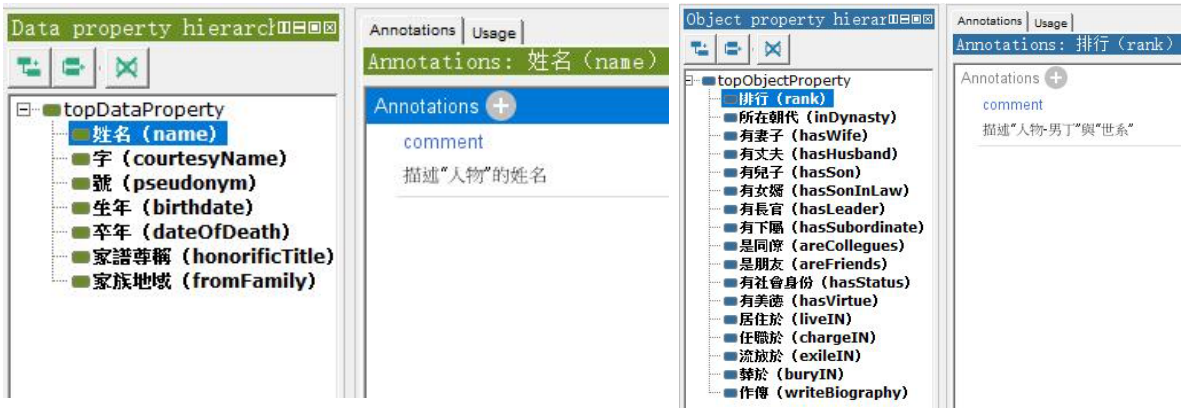


圖 8　protégé 編輯下情報分析領域本體的屬性編輯

如圖 8 所示，根據上文分析得到的本體概念間的 24 種關係，構建 7 個數據類型物件，分別為：姓名（name）、字（courtesyName）、號（pseudonym）、生年（birthdate）、卒年（dateOfDeath）、家譜尊稱（honorificTitle）、家族地域（fromFamily）；17 個對象屬性，分別為：排行（rank）、所在朝代（inDynasty）、有妻子（hasWife）、有丈夫（hasHusband）、有兒子（hasSon）、有女

婿（hasSonInLaw）、有長官（hasLeader）、有下屬（hasSubordinate）、是同僚（areCollegues）、是朋友（areFriends）、有社會身份（hasStatus）、有美德（hasVirtue）、居住於（liveIN）、任職於（chargeIN）、流放於（exileIN）、葬於（buryIN）、作傳（writeBiography）。建立好物件屬性類型，就可以將實例聯繫起來。家譜領域本體的 ontograf 本體關係圖如圖 9 所示，其中每一個節點可以根據需要展開，節點中包含實例，實例之間通過物件屬性產生聯繫，形成各種關係。

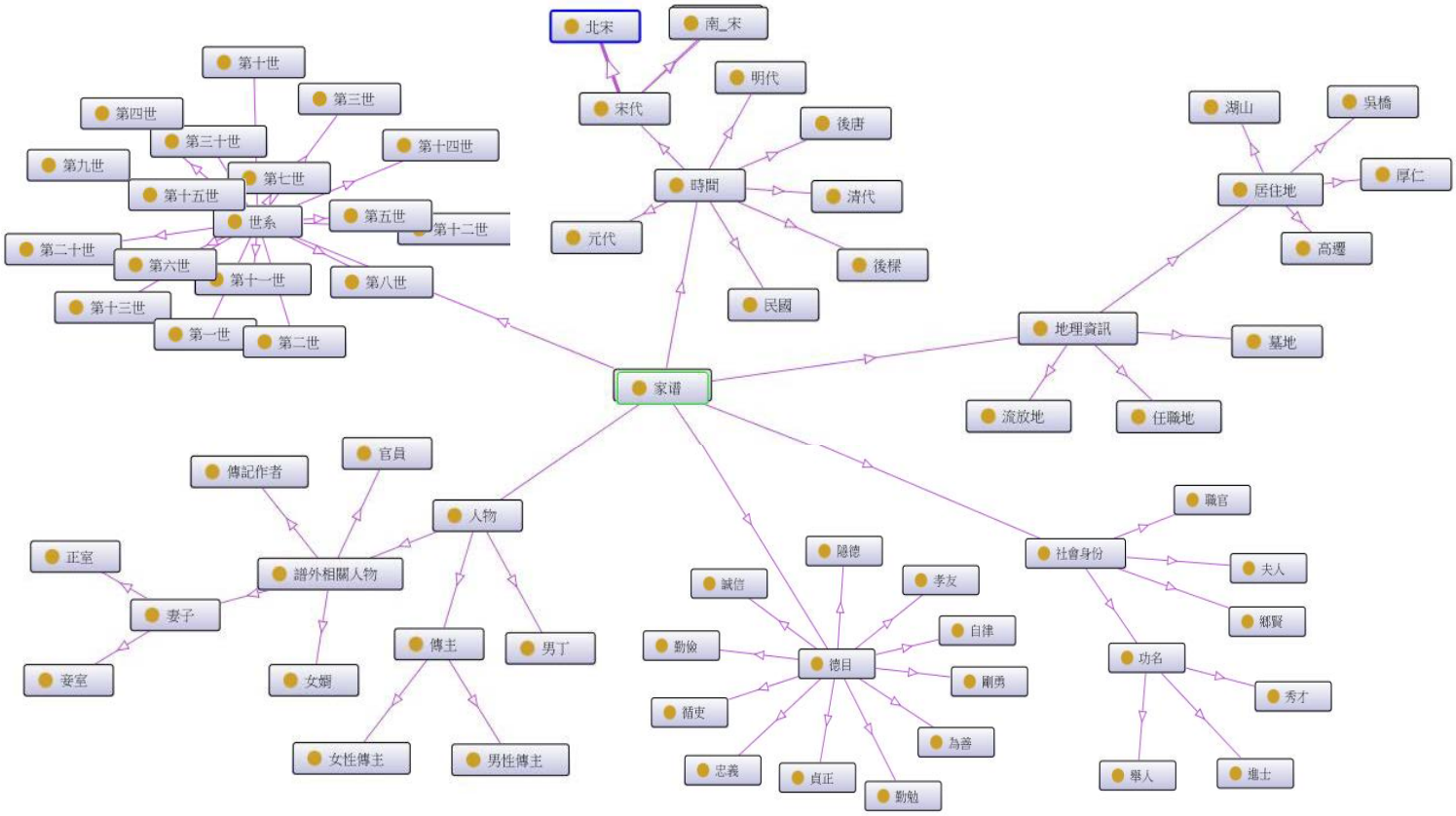


圖 9　protégé 編輯下家譜本體的 ontograf 本體關係圖

4.2　家譜本體的 OWL 表示

本體的標記語言是表示本體的語言工具，主要包括基於 Web 的 RDF、RDFS、OWL 等和基於 AI（Artificial Intelligence）的 KIF、Flogic 等語言。本文選用 OWL（Ontology Web Language）作為情報分析領域本體的本體表示語言，OWL 提供三種子語言：OWL Lite、OWL DL、OWL Full，根據構建本體的需要，本文選用 OWL Lite 語言進行家譜本體的 owl 表示。

4.3　家譜本體的視覺化實例

本文以吳氏第二十三世先祖吳時來作為視覺化實例進行展示。吳時來（1527—1590），字惟修，號悟齋，浙江仙居縣白塔鎮厚仁上街村人，官至左都禦史。吳時來任職於嘉靖年間，長官有嚴嵩、徐階等人，同僚有張居正等，病逝於北京，贈太子太保，謚忠恪，墓葬於橫溪大林。如圖 10 所示，“吳時來”是“傳主-男性傳主”的實例，與“譜外相關人物-官員”實例“張居正”為“是同僚”關係，“譜外相關人物-官員”實例“徐階”與“吳時來”為“是長官”關係。

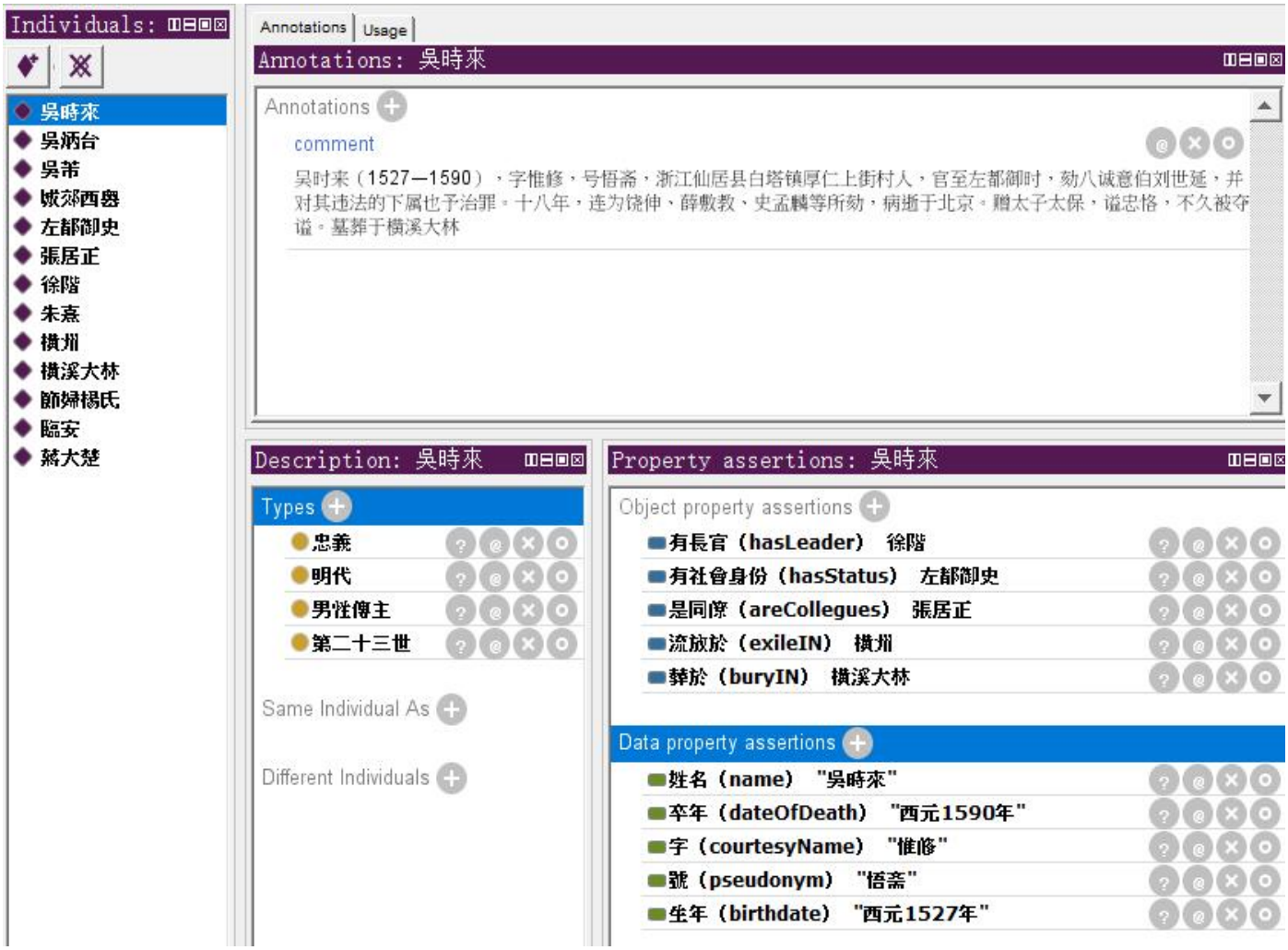


圖 10　protégé 編輯下實例“吳時來”的編輯示例

如圖 11 所示，吳時來作為家譜中有傳記的男性傳主，通過物件屬性與不同的人物、時間、身份、地點等方面產生了聯繫，將其生命的重要方面的內容凸顯出來，尤其是他的社會關係可以更清晰的體現出來，形成了多維度的、立體的視覺化展示。

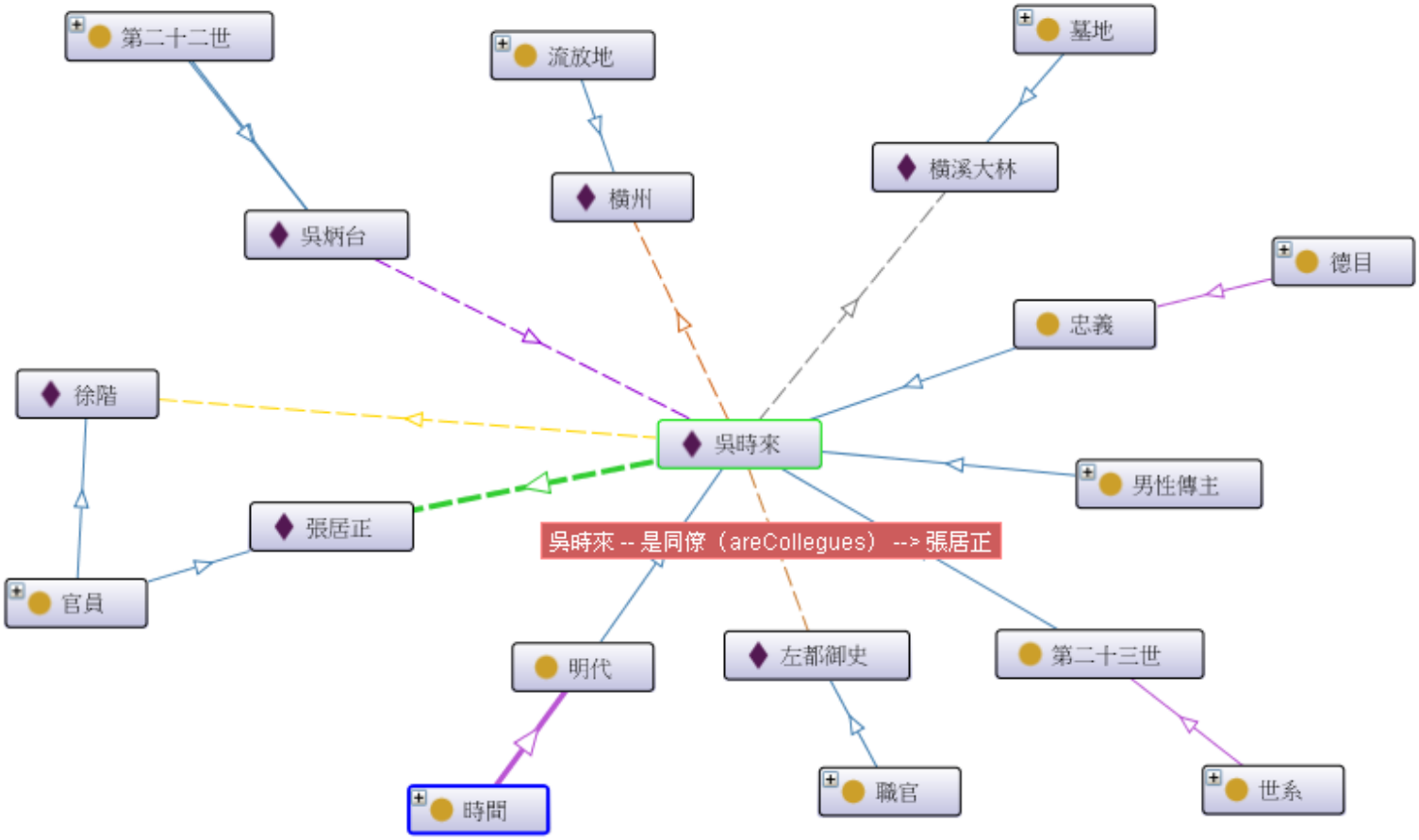


圖 11　protégé 編輯下實例“吳時來”的关系图

5　結語

本文從家譜知識組織體系構建的視角出發，在對家譜領域本體的需求進行調研分析的基礎上，梳理總結了家譜中的重要知識要素，構建了家譜領域本體。通過家譜本體概念體系建立、家譜本體概念間關係描述等方面對家譜領域本體進行了構建，並借助 protégé 4.3 實現了家譜本體的開發與編輯，從而探究出傳統家譜的視覺化構建的一套方法，通過對明代左都禦史吳時來的視覺化實現，印證了基於本體的傳統家譜視覺化的可行性。

本文的家譜領域本體的建立為家譜的知識組織體系建設提出了一種新的嘗試，這一家譜領域本體將更好地展示家譜中的代系人物與事蹟，挖掘家譜中所反映的社會關係，實現傳統家譜的視覺化展示，並突出展示家族中的精英人物，挖掘家族精英文化。

参考文献

[1] 胡兆芹. 本体与知识组织[M]. 中国文史出版社, 2014.

[2] Uschold M, Gruninger M. Ontologies Principles, Methods and Applications[J]. Knowledge Engineering Review, 1996 , 11 (2): 1 -62.

[3] Gruninger M , Fox MS. The Design and Evaluation of Ontologies forEnterprise Engineering [C] . In : Workshop on Implemented Ontologies, European Workshop on Artificial Intelligence, Amsterdam ,NL, 1994 ,105 - 128.

[4] Fernandez M, Gormez - Perez A, Juristo N. METHONTOLOGY:From Ontological Art Towards Ontological Engineering The Designand Evaluation of Ontologies for Enterprise Engineering [C]. In :Workshop on Implemented Ontologies, . AAAI - 97 Spring Symposiumon Ontological Engineering , Stanford University , March 24 - 26th ,1997 , 33 -40.

[5] Bernaras A, et al. Building and Reusing Ontologies for ElectricalNetwork Applications[C]. In : Proceeding of the European Conference on Artificial Intelligence. Budapest, Hungary, 1996 ,298 -302.

[6] ISI Natural Language Processing Research Group , Ontology Creationand Use: SENSUS [EB /OL]. [2017- 8 -12] .http : / /www. isi.edu /natural - language/projects/ONTOLOGIES. html.

[7] IDEF5 ontology Description Capture Method. [EB/OL].[2017-08-12]. http://www.idef5.com/

[8] A Guide to Creating Your First Ontology [EB /OL]. [2017 -8 -12]. http://protege.stanford.edu/publications/ontology_development/ontology101.pdf.

[9] Protégé 简介.[EB/OL].[2017-8-16.http://www.zhixing123.cn/shijian/39008.html.

[10] A Practical Guide To Building OWL Ontologies Using Prot´eg´e 4 and CO-ODE Tools Edition 1.3[EB/OL].[2016-12-28]. http://mowlpower.cs.man.ac.uk/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

Integration of a Chinese character ontology and Historical Glyph Examples

Morioka, Tomohiko
Ph.D / Assistant Professor
Center for Informatics in East Asian Studies
Institute for Research in Humanities
Kyoto University

Integration of a Chinese character ontology and Historical Glyph Examples

Morioka, Tomohiko

Ph.D / Assistant Professor

Center for Informatics in East Asian Studies

Institute for Research in Humanities

Kyoto University

Abstract

This report describes an attempt to integrate the “CHISE” (“Character Information Service Environment”) character ontology and the “HNG” (“Hanzi Normative Glyphs”) database / dataset.

The CHISE character ontology is a large scale character ontology which includes 357 thousand character-objects including Unicode and non-Unicode characters and their glyphs, etc. It was developed for CHISE which is a character processing system not depended on character codes. The framework of CHISE is based on a graph storage named “CONCORD”. We developed a Web service to display and edit objects of CONCORD, called “EsT” (or “CHISE-wiki”).

The CHISE character ontology uses the “Multiple Granularity Hanzi Structure Model” to support various glyphs and multiple unification granularity of Chinese characters. This model works fine for modern glyphs of Chinese characters. However, before we started the study to integrate CHISE and HNG, it was not clear that the model is sufficient for premodern Chinese characters. In addition, to design reasonable unification rules for each unification granularity, we need various glyph examples of Chinese characters. In these senses, the CHISE character ontology should integrate glyph database and/or glyph corpus. Therefore, we tried to integrate HNG and the CHISE character ontology.

When viewed from the HNG side, this integration has the following significance. The original HNG web service has been stopped since the spring of 2015. Therefore, we applied research on the integration of CHISE and HNG, we provided HNG search function and data browsing function on the CHISE Web service.

Table of Contents

1. INTRODUCTION 3

2. HNG 4

3. DATA STRUCTURE OF HNG 7

4. CHISE CHARACTER ONTOLOGY 7

4.1. Multiple glyph granularity 7

4.2. Multiple Granularity Hanzi Structure Model 7

5. REPRESENTATION OF HNG GLYPHS IN CHISE 8

5.1. Integration of HNG glyphs into the CHISE character ontology 8

5.2. Encoding of HNG glyph image object 9

6. IMPLEMENTATION 9

6.1. Classification of HNG glyphs 9

6.2. Integration without classification 10

6.3. Web applications 11

Keywords

Chinese character, glyph, linked data, database integration, dataset preservation.

1. Introduction

This report describes an attempt to integrate the “CHISE” (“Character Information Service Environment”) character ontology (1) and the “HNG” (“Hanzi Normative Glyphs”) database (2) / dataset (3).

Glyph database and glyph corpus including historical glyph examples of Chinese characters, such as “HNG (Hanzi normative glyphs) database” or “Character Database of Digital Rubbings” (「拓本文字データベース」) [4], are useful tools in considering the historical transition of the Chinese character glyphs and their normative consciousness. In particular, HNG is designed to display a relatively small number of typical examples of glyph-images of *kaishu* (楷書) to demonstrate that each time period and geographical region (country, state) had its own orthographic standard which differed from that of other time periods and geographical regions.

HNG is designed and developed based on various (deep) knowledges about Chinese characters and codicology, however these background knowledges may be tacit knowledges and/or they are not machine-readable knowledge. HNG does not have machine readable knowledge about unification granularity (namely how glyph-images are classified into abstract glyphs (字體)) and structure of characters (compositions of components described by IDS or other formats). Therefore, users can search through abstract characters instead of directly searching for glyphs (although HNG has its own criteria for abstract glyphs).

On the other hand, we have a character information service named “CHISE” (“Character Information Service Environment”) (5) that can complement HNG. “CHISE IDS Find” (「CHISE IDS 漢字検索」) (6) is a one of the most powerful tools to search complicated Chinese characters. In a result page of CHISE IDS Find, the first column of each line indicates an entry for a character. It has a link for a page in the “CHISE-wiki” (“EsT”) to display details of the character. CHISE-wiki has links for Chinese character related Web services such as “GlyphWiki” and “UniHan database”, or other service such as the “Classical Chinese Morphological Linked Open Data” (「古典中国語形態素 LOD」) and the “Bibliography of Oriental Studies” (「東洋學文獻**類**目」) database [7]. CHISE-wiki is a good tool to find information of each character and it also displays usage of each character.

These Web services of CHISE are based on the “CHISE character ontology”. It is a large-scale character ontology which includes 357 thousand character-objects including Unicode and non-Unicode characters and their glyphs, etc. It uses the “Multiple Granularity Hanzi Structure Model” to support various glyphs and multiple unification granularity of Chinese characters. This model works fine for modern glyphs of Chinese characters. However, before we started the study to integrate CHISE and HNG, it was not clear that the model is sufficient for premodern Chinese characters. In addition, to design reasonable unification rules for each unification granularity, we need various glyph examples of Chinese characters. In these senses, the CHISE character ontology should integrate glyph database and/or glyph corpus. Therefore, we tried to integrate HNG and the CHISE character ontology. In addition, the original HNG web service has been stopped since the spring of 2015. We urgently needed to set up an alternative service for the original HNG. Therefore, we applied research on the

integration of CHISE and HNG, we provided HNG search function and data browsing function on the CHISE Web service.

2. HNG

“Ishizuka Register of Chinese Character Standards of Writing” (「石塚漢字字体資料」) is a set of paper cards [Figure 1] comprising 400,000 character instances from 69 manuscripts. It is created by Emeritus Professor Harumichi Ishizuka of Hokkaido University during his tenure in Japanese linguistic seminar or other educational and research programs.

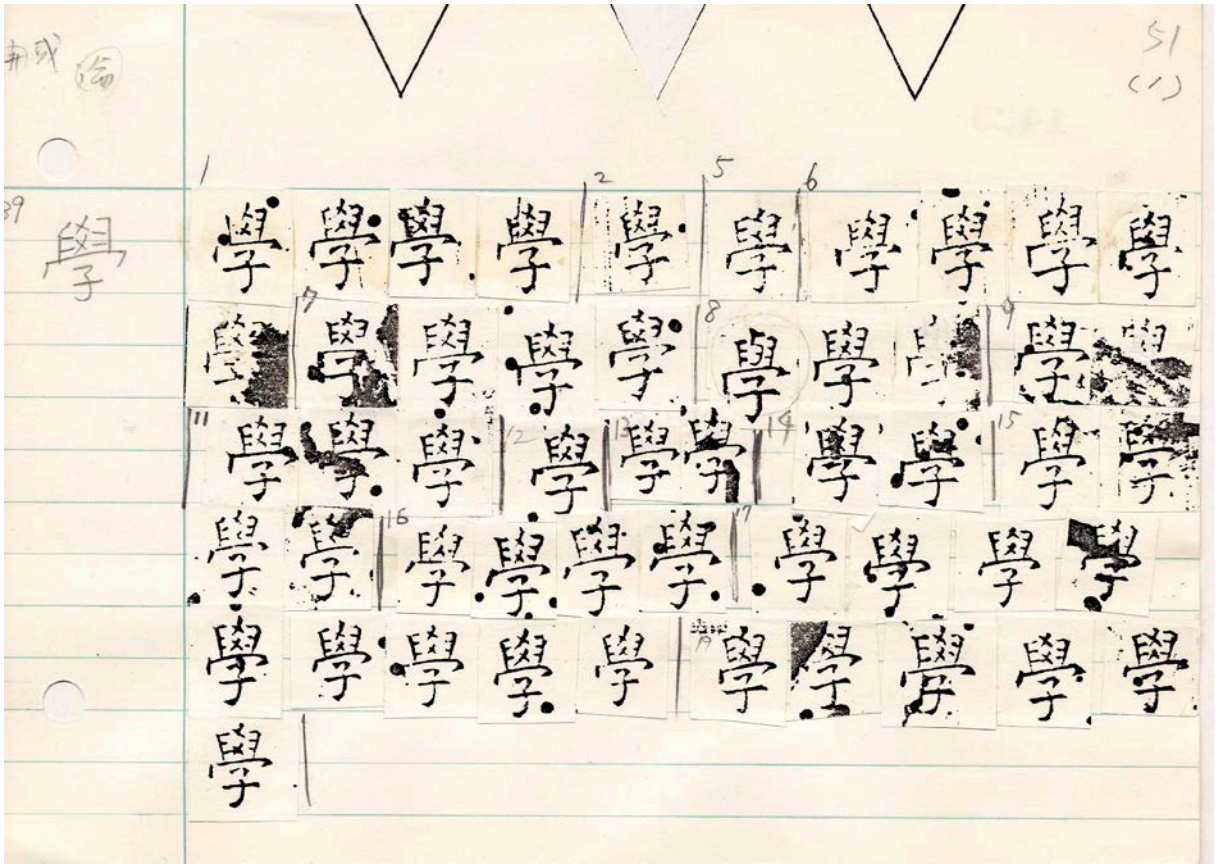


Figure 1: A sample of paper card (「學」 at 「開成石経論語」)

In order to deal with the deterioration problem of paper cards and to provide the data to the academic circle, it was digitized by researchers and students in Linguistic Sciences, Graduate School and Faculty of Letters, Hokkaido University. The result of this project was released on the Internet in 2005 under the name “Hanzi normative glyphs (HNG) database” (「漢字字体規範史データベース」) (2).

The original HNG database continued its Web service for ten years, but it stopped in the spring of 2015. In order to solve this problem, we have offered alternative service using CHISE technology, we concluded that Web-based search service alone is insufficient to maintain stable data availability over the long term. Therefore, we started new project to release HNG as a dataset, named “Hanzi normative glyphs (HNG) dataset” (「漢字字体規範史データセット」) (3). Currently, it contains 48 sources listed in Table 1.

In the original HNG database, not all glyph examples in each source of the specified character were displayed, only the representative glyph image was displayed (if a source has two or more glyphs for the specified character, two or more representative glyph images were displayed). However, we received many requests from users wanting to confirm all examples. Therefore, images of paper cards of the "Ishizuka Register of Chinese Character Standards of Writing" were also added to the HNG dataset.

No. (folder code)	ID	Type	Name of source	Abbrev	Age
1 (H03)	dng	南北朝写本	S81 大般涅槃経巻十一	S81	506
2 (H02)	keg	南北朝写本	S2067 華厳経巻十六	S2067	513
3 (H01)	jou	南北朝写本	P2179 誠実論巻八	P2179	514
4 (H05)	mam	南北朝写本	P2160 摩訶摩耶経巻上	P2160	586
5 (H06)	drt	隋写本	P2413 大楼炭経巻三	P2413	589
6 (H07)	kgk	隋写本	賢劫経巻二	賢劫経二	610
7 (H08)	myz	隋写本	P2334 妙法蓮華経巻五	P2334	617
8 (H10)	khi	初唐写本	今西本妙法蓮華経巻五	宮廷今西	671
10 (H11)	khm	初唐写本	守屋本妙法蓮華経巻三	宮廷守屋	675
11 (H13)	hok	初唐写本	S2577 妙法蓮華経巻八	S2577	7C 末
12 (H14)	kyd	初唐写本	上野本漢書揚雄伝	漢書揚雄	初唐
13 (H15)	sok	則天写本	守屋本花厳経巻八	華厳守屋	則天期
14 (H16)	yhk	盛唐写本	S2423 瑜伽法鏡経	S2423	712
18 (H09)	kda	高昌写本	大品経巻二十八	京博大品	高昌期
19 (H22)	sys	吐蕃写本	S5309 瑜伽師地論巻三十	S5309	857
20 (H49)	wan	大和寧写本	花厳経巻三十八	和寧花 38	9-10C
24 (H18)	kar	開成石経	開成石経論語	開成論語	837
25 (H19)	kae	開成石経	開成石経周易	開成周易	837
26 (H17)	kak	開成石経	開成石経孝経	開成孝経	837

29 (H25)	tzj	北宋版	東禅寺版阿毘達磨大毘婆沙論巻百七	東禅毘婆	1100
30 (H29)	jhk	北宋版	開禅寺版道神足無極変化経巻四	開元神足	1126
31 (H24)	tsu	北宋版	通典巻一	通典巻一	11C
32 (H26)	hos	北宋版	斉民要術	斉民要術	12C 初
34 (H28)	nak	南宋版	華厳経内章門等雑孔目巻一	華厳孔目	1146
35 (H30)	hod	南宋版	法蔵和尚伝	法蔵和尚	1149
36 (H31)	gok	南宋版	後漢書光武帝紀	光武帝紀	1198
37 (H74)	smk	西夏版	妙法蓮華経巻一	西夏法華	1149
38 (H50)	okd	日本写本	小川本金剛場陀羅尼経	金剛小川	686
39 (H54)	wad	日本写本	和銅経大般若経巻二百五十	和銅 250	712
40 (H55)	kmi	日本写本	高山寺本弥勒上生経	弥勒上生	738
41 (H56)	zkd	日本写本	守屋本五月一日経続高僧伝	五一続高	740
43 (H57)	doh	日本写本	高山寺本大教王経巻一	金剛大教	815
45 (H60)	tzs	日本写本	東禅寺版写大教王経巻一	仏説大教	12C
47 (H64)	kss	日本写本	明恵自筆華厳信種義	華厳信種	1221
48 (H66)	kyo	日本写本	親鸞自筆教行信証巻四	教行信証	1224
49 (H58)	jyu	日本版本	寛治二年刊本成唯識論巻十	成唯識 10	1088
52 (H33)	ink	日本書紀写本	岩崎本日本書紀巻二十四	岩崎紀 24	10C
53 (H34)	nto	日本書紀写本	図書寮本日本書紀巻二十四	図書紀 24	1142 頃
55 (H39)	nkk	日本書紀写本	兼方本日本書紀巻二	兼方紀 2	1286
56 (H36)	nkm	日本書紀写本	兼右本日本書紀巻二十四	兼右紀 24	1540
57 (H41)	kcc	日本書紀版本	慶長勅版日本書紀巻二	勅版紀 2	1599
58 (H42)	kcj	日本書紀版本	慶長十五年版日本書紀巻二	慶長紀 2	1610
59 (H43)	kbk	日本書紀版本	寛文九年版日本書紀巻二	寛文紀 2	1669

60 (H37)	k24	日本書紀版本	寛文九年版日本書紀巻二十四	寛文紀 24	1669
61 (H44)	sik	韓国写本	新羅本花厳経巻八	華厳新羅	754-755
62 (H46)	skk	韓国印刻本	晋本華厳経巻二十	古麗華 20	10C
63 (H47)	kyu	韓国印刻本	高麗初彫本瑜伽師地論巻五	初麗瑜 5	11C
64 (H48)	ksk	韓国印刻本	高麗再彫本華厳経巻六	再麗華 6	13C

Table 1: List of sources

3. Data structure of HNG

Currently, the HNG dataset is published in a Git repository: https://gitlab.hng-data.org/HNG/hng-data hosted by GitLab Community Edition. In the Git repository, each source is stored if own folder named `<folder code>_<name>`. For example, “10_妙法蓮華經卷五（今西本）” is the folder for 「今西本妙法蓮華経卷五」, and “10” is the folder code. Each folder has two subfolders, “`cards/`” and “`glyphs/`”. The subfolder “`cards/`” stores images of paper cards. The another subfolder “`glyphs/`” stores representative glyph images selected from paper cards and cropped out.

Each card is numbered in decimal four digits. Each representative glyph image cut out from a paper card is given an ID composed of a main code and a subcode. The main code is the same as the card number. The subcode is used to distinguish between variants if they exist. If a card (corresponding with a character example of a source) does not have any character/glyph variants, subcode is empty. Otherwise, subcodes “`a`”, “`b`”, “`c`”, ... are assigned to each variant. In addition, the relationship between each ID crossing the source is managed in a table of CSV format.

4. CHISE character ontology

The “CHISE character ontology” is a lightweight ontology developed by the authors for character processing. The CHISE character ontology contains information of characters included in Unicode, and other information for Chinese characters.

4.1. Multiple glyph granularity

As for Chinese characters, in addition to Unicode's unification rules (to define abstract characters of CJKV Unified Ideographs), it has information related to the plural glyph granularity of Chinese characters such as super abstract character, unified glyph, abstract glyph (字體), abstract glyph form and glyph image (字形).

4.2. Multiple Granularity Hanzi Structure Model

In the syntax of Ideographic Description Sequence (IDS) defined in ISO/IEC 10646, only coded ideographs (Chinese characters included in UCS) and radical characters can be used as

terminal components (leaf nodes of a IDS tree). However, theoretically, any component can be used as a leaf node. CHISE can represent and process characters not included in UCS. Therefore, in CHISE, Chinese characters and special components not included in UCS are also available as components of extended IDS represented by the ideographic-structure feature.

CHISE supports inheritance of character definition and CHISE character ontology uses this to represent relationships among different unification granularity, such as abstract character, abstract glyph, and glyph image. If we use abstract characters as terminal components of an IDS, the IDS represents a structure of an abstract character. If we use abstract glyphs, the IDS represents a structure of an abstract glyph. If we use glyph images, the IDS represents a structure of an abstract glyph image. Thus, the extended IDS of CHISE can represent unification coverage (granularity) of a character object (Figure 2).

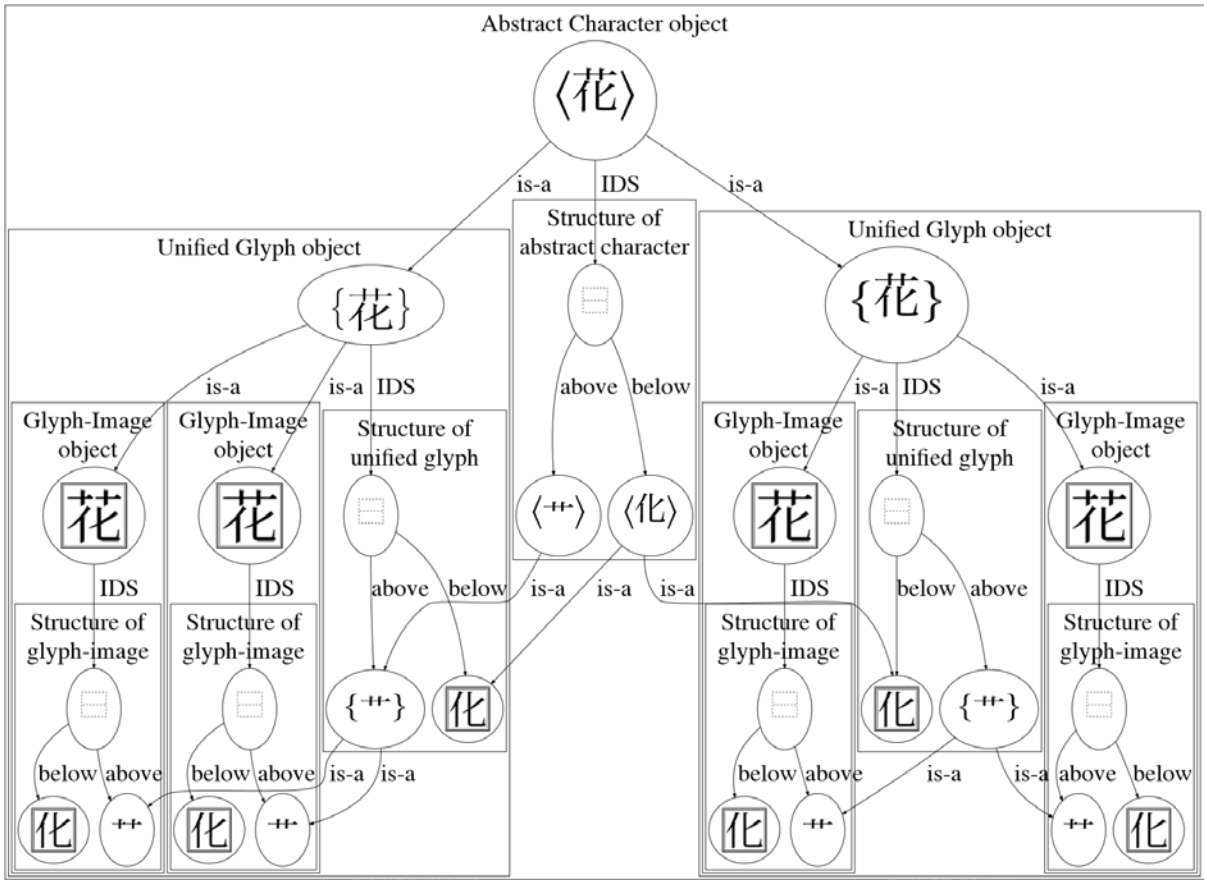


Figure 2: Conceptual graph of multiple-granularity IDS(花)

5. Representation of HNG glyphs in CHISE

5.1. Integration of HNG glyphs into the CHISE character ontology

Several methods can be considered to incorporate HNG information into the CHISE character ontology, but the following method was adopted here: Each representative glyph image of HNG is regarded and defined as a glyph image object of CHISE, and it is attached to any existing abstract glyph form object of CHISE if possible. Otherwise, it is attached to any existing abstract glyph object of CHISE if possible. Otherwise, it is attached to any existing

abstract character of CHISE if possible. Otherwise, it is regarded as new abstract character or new abstract character object will be defined to attach it. In this way, every HNG glyph object can be placed somewhere in the CHISE character ontology.

In addition, if an HNG glyph object to be newly added can be subsumed into an already existing abstract glyph object (If it can be subsumed into an abstract glyph form object, the abstract glyph form object is subsumed into an abstract glyph object. Therefore in the case, the HNG glyph object can be subsumed into the abstract glyph object which subsumes the abstract glyph form object which can subsume it.), there is no need to newly describe Hanzi structure (IDS).

5.2. Encoding of HNG glyph image object

A glyph image object of HNG is represented by a pair of glyph ID (see Section 3) at its source and ID feature of glyph image granularity indicating its source.

In HNG, an ID consisting of three alphanumeric characters is attached to each source. Using this, in CHISE, each source of HNG is represented by ID features with the prefix '===hng-' ("===" is the prefix for glyph image and "hng-" is the prefix of HNG) before the three alphanumeric ID to indicate the source of HNG.

For example, in the case of 「開成石經孝經」, the source ID is "kak", so the ID feature in CHISE is '===hng-kak'.

For the glyph ID, use the value obtained by adding 10 times the card number to the number corresponding to the suffix (0 for no suffix, 1 for suffix "a", 2 for suffix "b", and so on) as the feature value. For example, a glyph ID is "0100", the corresponding feature value is 1000; a glyph ID is "0123b", the corresponding feature value is 1232.

6. Implementation

6.1. Classification of HNG glyphs

Currently, in the CHISE project, we are formulating a guideline to detect unifiable range of glyph or abstract glyph form objects named "CHISE Guidelines for Glyph Granularity of Chinese characters" (CHISE-GGG) (8). In Section 5.1 we have discussed how HNG glyph objects are arranged in the graph of the CHISE character ontology based on Multiple Granularity relations. In the procedure, this guideline is used in detection of unifiability for abstract glyph objects or abstract glyph form objects.

In principle, "IRG Working Document Series (IWDS) 1: List of UCV (Unifiable Component Variations) of Ideograhs" (IWDS-1) (9) is used for determining the unifiable range of abstract characters.

Such relatively rigorous HNG glyph classification work is currently manual worked and requires a lot of labor. For the reason, it is difficult to do with all HNG data at present. In order to obtain the maximum effect from minimum samples, we chose the following three sources: 「今西本妙法蓮華経巻五」(khi), 「守屋本妙法蓮華経巻三」(khm) and 「開成石経論語」(kar). The first two (belong to Early Tang Court Sutras) are sources representing

an early Tang standard, and the third source representing normative *kaishu* glyphs of the Kaicheng Stone Classics (開成石經).

6.2. Integration without classification

To avoid taking a lot of effort, we decided to integrate HNG into CHISE using only the mechanically convertible part from the information of HNG for the remaining parts other than those classified by hand described in Section 6.1.

First, we defined HNG glyphs as glyph image objects of CHISE by the method described in Section 5.2.

Next, linking between the HNG glyph image object and the UCS abstract character object by using mapping information to UCS (or mapping information to "Daikanwa" (「大漢和辭典」(10))) in HNG. Each link from HNG glyph image object to the corresponding UCS abstract character object is represented by relation feature '`<-HNG`'. In CHISE, there is a mechanism to automatically generate reverse links for relation features, so the relation feature '`->HNG`' from UCS abstract character object to HNG glyph object is automatically generated.

However, in order to bring it closer to the original HNG, we are now using the following relation features shown in Table 2 with domain specifiers attached for each source type instead of relation features '`<-HNG`', '`->HNG`'.

Source Type	HNG glyph to entry	Entry to HNG glyph
Chinese manuscripts	`<-HNG@CN/manuscript`	`->HNG@CN/manuscript`
Chinese printed books (including Stone Classics)	`<-HNG@CN/printed`	`->HNG@CN/printed`
Japanese manuscripts	`<-HNG@JP/manuscript`	`->HNG@JP/manuscript`
Japanese printed books	`<-HNG@JP/printed`	`->HNG@JP/printed`
Korean sources	`<-HNG@KR`	`->HNG@KR`
Other sources	`<-HNG@MISC`	`->HNG@MISC`

Table 2: Relation features between HNG glyph image objects and entry objects

By this method, when displaying a CHISE-wiki page for a UCS abstract character, it is now possible to display unclassified HNG glyph image objects as the same as classified objects.

However, in the method, the search target of CHISE IDS Find is limited to entries characters (may be UCS abstract characters) of HNG glyphs, and it is not possible to search HNG glyphs which have different structures from their entries characters.

Conversely, in the three sources that integrated with the classifications described in Section 6.1 (「今西本妙法蓮華経巻五」(khi), 「守屋本妙法蓮華経巻三」(khm) and 「開成石経論語」(kar)), each HNG glyph image object knows own Hanzi structure even if its structure and abstract glyph are different from its entry character. For example, by searching Chinese

characters having 「十」 and 「刀」as components in CHISE IDS Find, you can find 「切」(U-0002D0C4) in the search result and you can see 「切」in its CHISE-wiki page.

6.3. Web applications

CHISE-HNG IDS Find (CHISE-IDS HNG 漢字検索) (11) [Figure 4] is a Web service to search Chinese characters included in the HNG dataset. It is like CHISE IDS Find (CHISE IDS 漢字検索), if a user specifies one or more components of Chinese characters into the "Character components" window and runs a search, the characters that include every specified component are displayed. However, unlike CHISE IDS-Find, only Chinese characters with examples of HNG are displayed.



Figure 3 CHISE-HNG IDS Find

In a results page of the CHISE-HNG IDS-Find (or CHISE IDS-Find), the first column of each line indicates an entry for a character. It has a link for a EsT page [Figure 5] to display details of the character. In the EsT page, links for HNG examples are displayed in the "→ [HNG] …" field. (cf. Table 2)

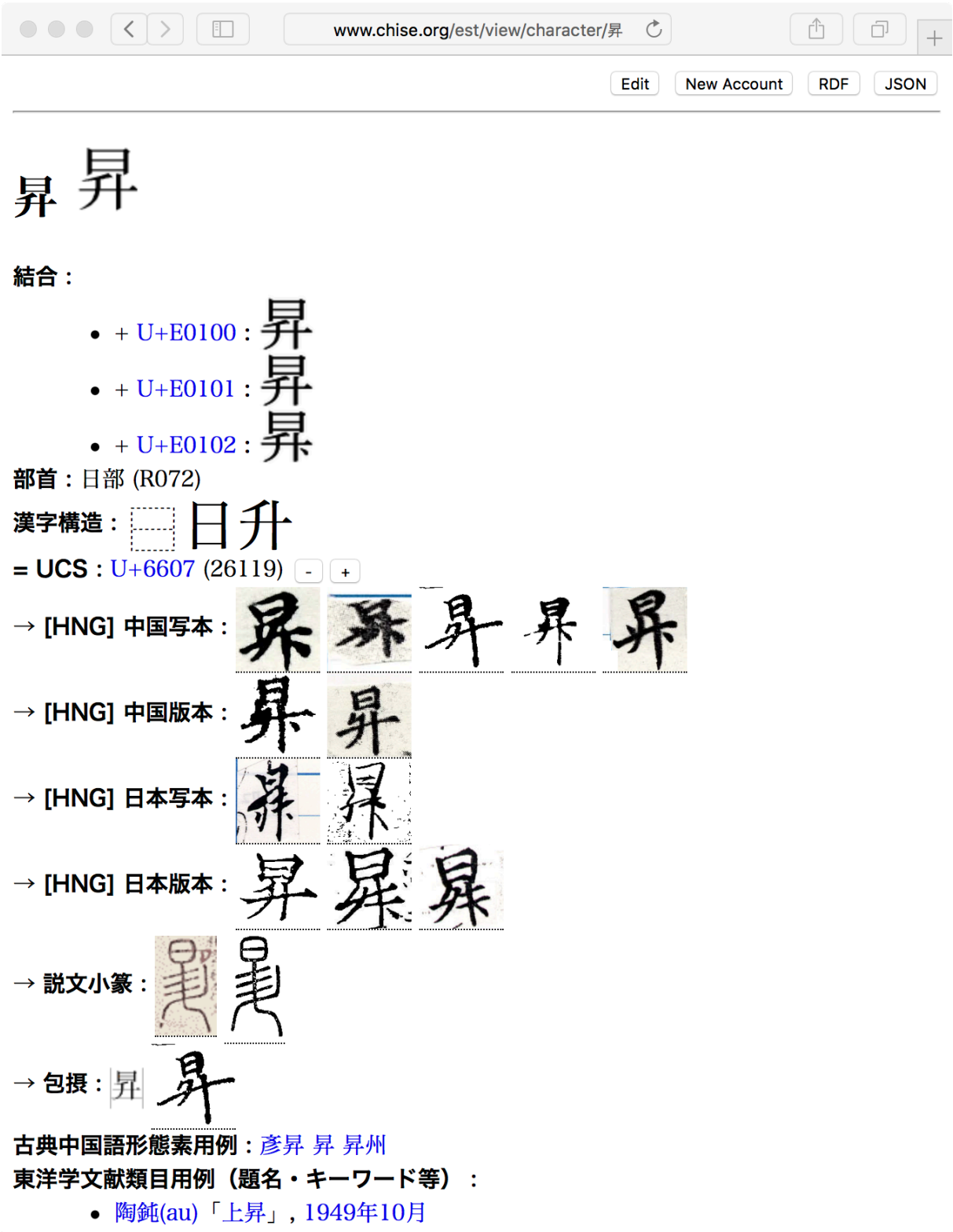


Figure 4: A sample CHISE-wiki (character object viewing in EsT) page (昇)

7. Conclusion

This report outlined the integration of the HNG dataset and the CHISE character ontology, and also introduced the outline of the HNG dataset.

By realizing this integration, it became possible to browse information of HNG via CHISE Web services such as “CHISE IDS Find” or “CHISE-Wiki” (“EsT”), and we were able to strengthen the search function for HNG data. In addition, by realizing “CHISE-HNG IDS-Find” that limits the search range of CHISE IDS Find to the range of HNG data, it has become possible to search for glyph examples of HNG having common components.

References

1. Morioka, Tomohiko, “Multiple-policy character annotation based on CHISE,” Journal of the Japanese Association for Digital Humanities 1(1), p. 86–106.

2. Ishizuka, Harumichi, “Current status and future prospects of the Hanzi Normative Glyphs (HNG) database”, http://idp.bl.uk/downloads/hng_translation.pdf.

3. *HNG dataset*. https://gitlab.hng-data.org/HNG/hng-data.

4. *Character Database of Digital Rubbings*（拓本文字データベース）. http://coe21.zinbun.kyoto-u.ac.jp/djvuchar.

5. *CHISE Project***.** http://www.chise.org.

6. Morioka, Tomohiko**,** *CHISE IDS find*（CHISE IDS 漢字検索）. http://www.chise.org/ids-find.

7. *Bibliography of Oriental Studies database*（東洋学文献類目検索 [第 7.4 α 版]） http://ruimoku.zinbun.kyoto-u.ac.jp/.

8. Morioka, Tomohiko. 2016. *CHISE Guidelines for Glyph Granularity of Chinese characters*（CHISE 文字オントロジーのための漢字字体・字形粒度の情報記述に関するガイドライン） *[Ver.0.9.1]*. http://www.chise.org/specs/ggg_v0.9.1.pdf.

9. *IRG Working Document Series (IWDS)*. http://appsrv.cse.cuhk.edu.hk/~irg/irgwds.html.

10. Morohashi, Tetsuji (諸橋轍次) et al. *Dai Kanwa Jiten* (大**漢**和辭典). Tokyo: Taishūkan.

11. Morioka, Tomohiko. *CHISE-HNG IDS find*（CHISE-IDS HNG 漢字検索）. http://www.chise.org/hng-ids-find.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12 18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

人文學基本 LOD 試論

以韓臺《職員錄》為例

金把路

研究教授

韓國中央大学

人文學基本LOD試論
- 以韓臺《職員錄》為例

金把路
研究教授
韓國中央大学

摘要

本發表提出了可鏈接韓國史編纂委員會的《朝鮮職員錄》與臺灣中央研究院的《臺灣職員錄》的Ontology（本體）設計模型與LOD（Linked Open Data， 鍵連公開資料)服務以及人文學基本LOD系統概念。

首先在數據搜集上，根據韓國《公共數據法》搜集了《朝鮮職員錄》，通過網頁抓取(Web scraping)搜集了《臺灣職員錄》。根據搜集到的韓臺職員錄數據，進行數據來源的敘述而數據結構的分析。根據職員錄數據結構的分析結果，設計了職員錄Ontology而建設了各自獨立的韓臺職員錄LOD示範服務。在設計職員錄Ontology中， 筆者提出了可描述內含著時間、空間，人物、行為與文獻的“事件（Event）"Ontology模型。

其後，為了鏈接各自獨立的韓臺職員錄LOD服務，設計了可鏈接韓臺職員錄的Ontology。在設計鏈接職員錄Ontology中，筆者提出了反映人文學研究特質的“人文學基本Ontology”，它以數據形式可敘述了各個研究者對同一對象的不同判斷與其根據。而根據韓國國史編纂委員會《韓國史數據庫》之《韓國近現代人物資料》，找出韓臺《職員錄》的共同人物，並建設了可鏈接韓臺職員錄的LOD示範服務。

最後 ,為了各自主題的多種人文學LOD之间的鏈接 ,提出以核心要素(時間、空間、人物、事件）為中心的人文學基本LOD系統之概念。

目次

1. 绪论
2.《职员录》数据的搜集与分析
3.《职员录》Ontology
4.人文學基本LOD
5.人文数据共享方法比较
6.结论

關鍵詞

Digital Humanities、數位人文、Ontology、數位本體、LOD、鍵連公開數據、《職員錄》

1. 绪论

全世界人文學的危機與衰落已經不是新聞的時代了。與此相反有超速發展的領域 - 人工智能。人工智能的輝煌發展通常以硬件發展、大數據形成、 深度學習技術的實用化為由。但是人工智能發展的內在基礎是共享。人工智能領域享受SQuAD[1]，Imagenet[2]，OpenML[3]等的大數據共享平臺與TensorFlow[4]、GitHub[5]等的公開開發平臺以及arXiv[6]為代表的論文共享平臺。相反、人文學領域的共享有著一定的限制。因此為了突破人文學的呆滯，需要數位人文（Digital Humanities）方法為基礎的共享。

數位人文是對計算與人文學科之間的交叉領域進行學習、研究、發明以及創新的一門學科[7]。它不僅力求傳統文本的數位化，還考慮到文本挖掘（Text mining）、社會網絡分析（Social Network Analysis）、空間分析（Spatial analysis）等的數位分析方法與多媒體（Multimedia）、增强现实（Augmented Reality，AR）、虚拟现实（Virtual Reality，VR）等的數位視覺化在歷史、哲學、文學等學科的應用。簡單地說，數位人文是傳統人們學研究的基礎上導入數位的研究方法。於是數位人文的本質還是與對人類的探究，與傳統人文學研究在同一個線上。只是傳統人文學以紙張為基礎，但是數位人文以數位為基礎。而且數位人文借用計算機的能力，可實現人類無法實現的情報搜集、分析、共享，還可實現紙張無法提供的多媒體視覺化。

[1] SQuAD：https://rajpurkar.github.io/SQuAD-explorer/

[2] Imagenet：http://www.image-net.org/

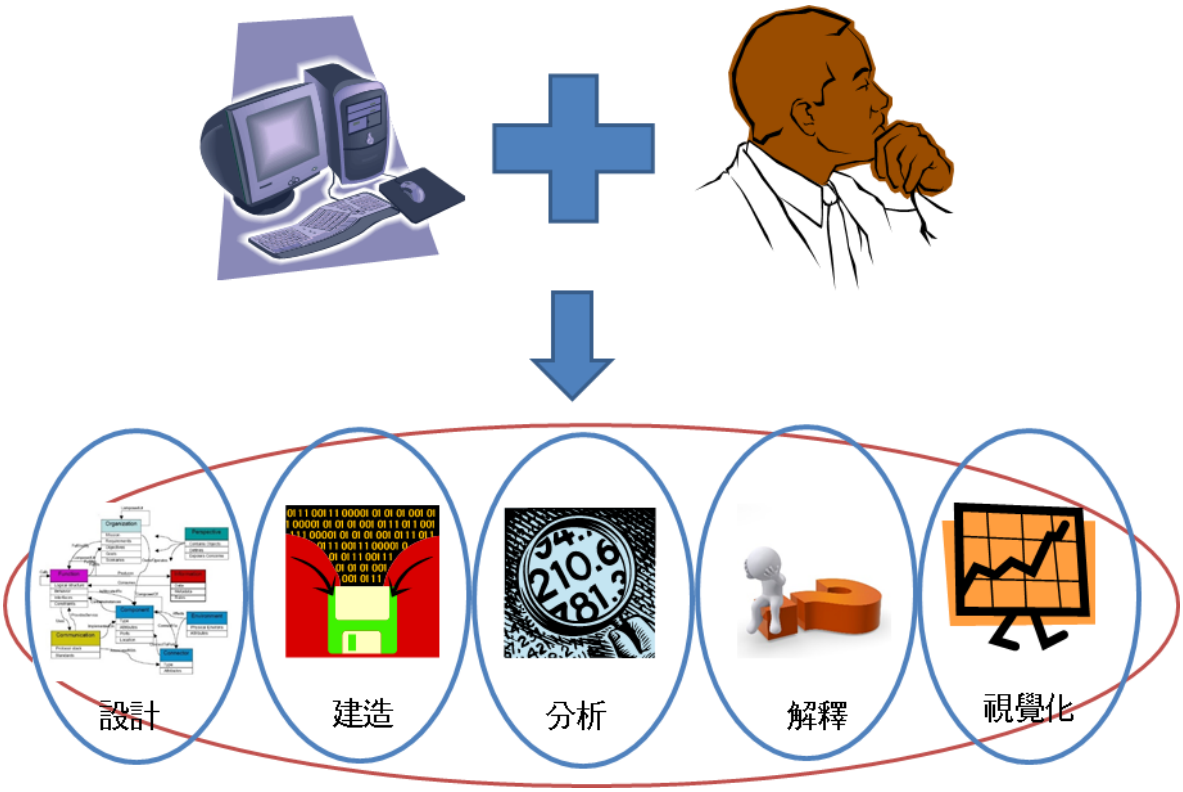
[3] OpenML：https://www.openml.org/

[4] TensorFlow：https://www.tensorflow.org/

[5] GitHub：https://github.com/

[6] arXiv：https://arxiv.org/

[7] 金炫、林永尚、金把路，《數位人文入門》。

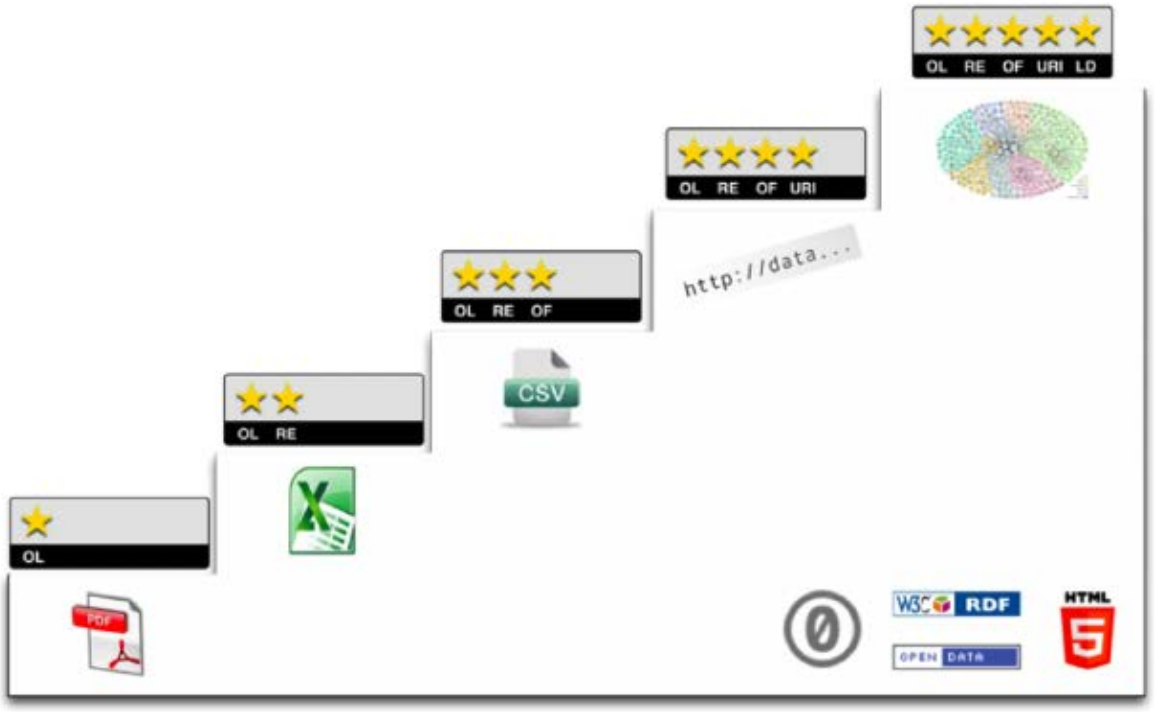


圖一、數位人文領域概念圖

論者把數位人文研究領域分為設計、建造、分析、解釋、視覺化。設計是為了對象人文學情報移植在數位上，研究對象的結構與內容。建造是按照設計結果、新造、再編、運營人文數據庫。分析是以人文數據為基礎，傳統人文學研究方法結合文本挖掘、社會網酪分析、空間分析等的數位分析方法導出分析結果的領域。解釋是通過人類的觀點與思維，在分析結果上賦予意義的領域[8]。視覺化是設計結果、人文數據、分析結果、解釋結果變為合適與人類的領域。其中在人工智能時代，人文學可占優勢的是數據設計與數據解釋領域。因為數據設計與解釋需要針對對象的深刻理解，並且人文學已在數千年積累了史料搜集與整理以及解釋的多種方法論。只是與往前不同，為了人類與人工智能的合作，人文學者需要脫離人類可讀數據（Human-readable data）的束縛，得走向機器可讀數據（Machine-readable data）的未來。

設計與建造都是數據的領域。人文學領域已經有豐富的紙張情報，所以目前人文學領域的比較關心數位化（digitalization）。但是很多人文學領域的人士忽略數據的品質。根據鍵連公開數據（Linked Open Data，LOD），數據可分析五等級。

[8] 金把路， <制度與人事的關系性數據數位檔案的建立及應用 - 以近代學校資料(1895~1910)為中心 ->。

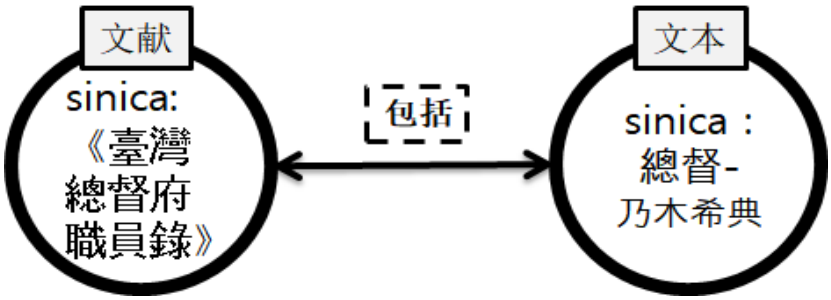


圖二、鍵連公開數據（Linked Open Data，LOD）的數據等級概念圖

第五等級的條件是制作權公開的（Open Licence），我們一般接觸的PDF文件屬於這一類。數位網絡的最大強點在情報的共享，如果某一個情報無法共享等於沒有數據的。第四等級的條件是制作權公開與可再用的（Reusable），我們一般接觸的Excel文件屬於這一類。為了借用計算機的力量，必須建造機器可讀數據（machine readable data），雖然最近PDF文件通過光學字符識別（Optical Character Recognition，OCR）可以變成文本（TEXT），但是其導出的文本還是機器可有限的讀出其內容。機器有限的讀其內容等於人工智能有限的處理其內容的。第三等級的條件是制作權公開、可再用的、自由文件格式（Open format），其代表文件形式為CSV格式。雖然我們常用Excel，但是Excel文件只能在微軟的Office上才能運行的，所以我們為了保障自由性，得采用自由文件格式。第二等級的條件是制作權公開、可再用的、自由文件格式、統一資源標誌符（Uniform Resource Identifier，URI）。URI是為了同時保障多樣性、平等性而產生的一種出處表明手段。第一等級的條件是制作權公開、可再用的、自由文件格式、統一資源標誌符、鍵連數據（Linked data）。世界人人皆有自己的想法，按照自身的想法，建立滿足第二等級的數據，而各自的人文數據庫互相連接的才是滿足第一等級的。但是現在大部分的人文數據連第五等級都達不到的。重點在於機器可讀性數據與數據共享。

現在最高級的機器可讀性數據概念是語義網（Semantic Web）。語義網是由萬維網聯盟的蒂姆·伯納斯-李（Tim Berners-Lee）在1998年提出的一個概念，它的核心是：

通過給萬維網上的文檔（如: HTML）添加能夠被計算機所理解的語義（Meta data），從而使整個互聯網成為一個通用的信息交換介質。語義萬維網通過使用標準、置標語言和相關的處理工具來擴展萬維網的能力。不過語意網概念實際上是基於很多已有技術的，也依賴於後來和text-and-markup與知識表現的綜合[9]。為了實現語義網需要設計數位本體（digital ontology）。現在數位本體的基本要素為類（Class）、屬性（Property）、個體（instance），基本形式為論域（domain）- 關系（relation) - 定義域（range）。



图三、语义网概念示范图

“《臺灣總督府職員錄》”，“總督:乃木希典”是個體。《臺灣總督府職員錄》屬於“文獻”類，“總督-乃木希典”是“文本”類。“包括”是個屬性。我們可以敘述“文獻”類的“《臺灣總督府職員錄》個體 - “包括” - “文本“類的“總督-乃木希典”，這可抽象化而設計為“文獻”類 - ”包括“ - ”文本“類的。而且通過統一資源標誌符記述了其個體的出處，比如上圖的“sinica”是筆者自定的“臺灣中央研究院”的前缀（Prefix），“sinica:《臺灣總督府職員錄》”意味著“臺灣中央研究院的《臺灣總督府職員錄》”[10]。雖然數位本體的結構是比較簡單的，但是語義網（Semantic Web）已經成功的實現了各種大型人文數據庫。如歐洲數位圖書館(Europeana)[11]統合了歐洲各個圖書館、美術館、博物館的文化遺產情報的。

註意的是人文情報的結構與語義網的結構比較相似。從前人文情報的數位化依靠了可擴展標記語言（Extensible Markup Language，XML）與關系數據庫（Relational database，RDB）。雖然XML與RDB是商業上得到地位的有效方法，但還是無法完全包含著數位情報的全部內容。相反，語義網以數位本體為基礎，可輸入、運營、輸出多層次的人文情報，還可以按照以往的人文數據進行倫理推論而找到新的情報。大膽的說，我們通過語義網，把人文學的思維移植到數位上了。

[9] 语义网，维基百科：https://zh.wikipedia.org/wiki/语义网

[10] 如记述为https://www.sinica.edu.tw/臺灣總督府職員錄

[11] 歐洲數位圖書館： https://www.europeana.eu/portal/

論者以上述的數位人文觀點為基礎 ,提出可鏈接韓國史編纂委員會的《朝鮮職員錄》與臺灣中央研究院的《臺灣職員錄》的Ontology（本體）設計模型與LOD（Linked Open Data， 鏈連公開資料)服務以及人文學基本LOD系統概念。

2. 《職員錄》數據的搜集與分析

首先在數據搜集上，根據韓國《 關於公共數據提供與利用活性化的法律 》[12]搜集了《朝鮮職員錄》[13]，通過網頁抓取(Web scraping)搜集了《臺灣職員錄》[14]。根據搜集到的韓臺《職員錄》數據，進行數據來源的敘述而數據結構的分析。

《職員錄》是日本帝國是關於國家公務員的記錄文件，是每年一兩次發行包括官制以及公務員的姓名，職名，等級等的情報。所以不僅可研究日本帝國的制度，還可以研究當代人物。

韓國國史編纂委員會《朝鮮職員錄》是綜合了1908年《大韓帝國職員錄》，1910年至1943年《朝鮮總督府職員錄》，1952年《大韓民國職員錄》的201033條記錄。論者按照韓國《關於公共數據提供與利用活性化的法律》得到《朝鮮職員錄》的RAWDATA。其RAWDATA形式為"|"為分隔符號的TXT文件。而其數據屬性（Attribute）包括，調查時期，人名，所屬，官職，官等，功勛，參考事項[15]。

表一、《朝鮮職員錄》的數據標本

LEVEL_ID	MAIN_TITLE	SERIES_TITLE	PERIOD	POSITION	OFFICIAL_RANK	MERITS	SUBJECT_CLASS
jw_1935_5995_0010	伊東良夫	朝鮮總督府及所屬官署職員錄1935年	1935	教諭	7等6等级	NULL	地方官署>咸鏡北道>公立學校>羅南中學校
jw_1935_5995_0020	須崎正義	朝鮮總督府及所屬官署職員錄1935年	1935	教諭	7	NULL	地方官署>咸鏡北道>公立學校>羅南中學校

[12] <关于公共数据提供与利用活性化的法律>（韓國法律第11956號）

[13] 职员录资料, 韩国史数据库, 国史编纂委员会： http://db.history.go.kr/item/level.do?itemId=jw

[14] 臺灣總督府職員錄系統，台湾中央研究院台湾史研究所：http://who.ith.sinica.edu.tw

[15] 兼职情报等的备注叙述在此.

臺灣中央研究院臺灣歷史研究所《臺灣職員錄》是綜合了1896年至1944年的總共49個年，51本，43163條[16]。論者通過網頁抓取（Web scraping）技術，搜集了臺灣總督府職員錄系統所提供的數據。其RAWDATA數據屬性（Attribute）是序號,姓名,日本紀年,西元紀年,單位名稱,官職名,頁碼[17]。

表二、《臺灣職員錄》的數據標本

序號	姓名	日本紀年	西元紀年	單位名稱	官職名	頁碼
0	乃木希典	明治二十九年	1896	臺灣總督府	總督	573
1	木村新九郎	明治二十九年	1896	臺灣總督府總督官房	副官	573

3.《職員錄》Ontology

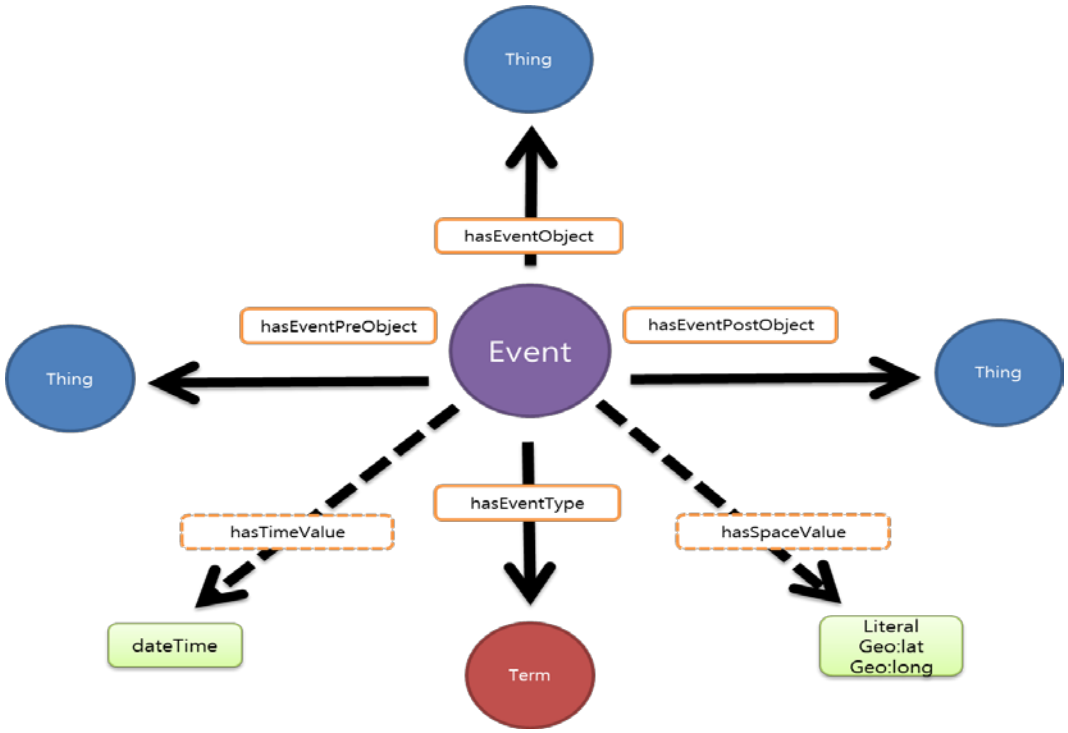
根據職員錄數據結構的分析結果，設計了職員錄Ontology而建設了各自獨立的韓臺職員錄LOD示範服務。在設計職員錄Ontology中，筆者提出了人物、制度、人事記錄互相連貫的職員錄Ontology。特別是，筆者提出了可描述內含著時間、空間，人物、行為與文獻的“ 事件(Event)"Ontology模型。並且為了反映各個人文學機構對自身人文學數據的著作權利與維持義務，根據職員錄Ontology，建設了各自獨立的韓臺職員錄LOD示範服務。

筆者把事件（EVENT）定義為“人類生活中的一切活動與現象”。事件的基本構成要素是主體、時間、空間、行為、對象。事件的結構性特征是以特定時間為基點，發生由行為引發的變化，再說，事件是一種變化，其變化的對象是或是某人物或是某物質。[18]

[16] 由于地方改革，1909年与1920年是各个年度存在2本《职员录》

[17] 本文没有搜集了影像与延伸查询的数据。

[18] 雖然已有The Event Ontology（http://motools.sourceforge.net/event/event.html），但是其Ontology把事件定義為獨立事件，沒有反應由事件而發生的變化。



圖四、事件模型設計概念圖

筆者以上述的事件概念為基礎，設計了敘述可以敘述歷史事件的“事件模型”。事件模型的Class為Event（事件）、Term（概念）、Person（人物）、Group（團體），Property為hasEventObject[19]、hasEventPreObject[20]、hasEventPostObject[21]、hasEventType[22]、hasTimeValue[23]、hasSpaceValue[24]。

學術情報需要編纂者、編纂時間、情報判斷根據。但是因為以往的人文情報整理是以個人或者單一機構為主導，基本上沒有特意記錄了每一條情報的編纂者，而且從紙張時代無法修改書籍上的單一條信息，只能改版才能修改，所以基本上沒有記錄每一條的編纂時間。反而共享為基礎的人文數據的的設計需要包括針對各個情報的編纂者、

[19] hasEventObject是链接事件与事件主体。Domain为Event，Range为Thing。

[20] ha Event，

Ran

[21] ha 为Event，

Ran

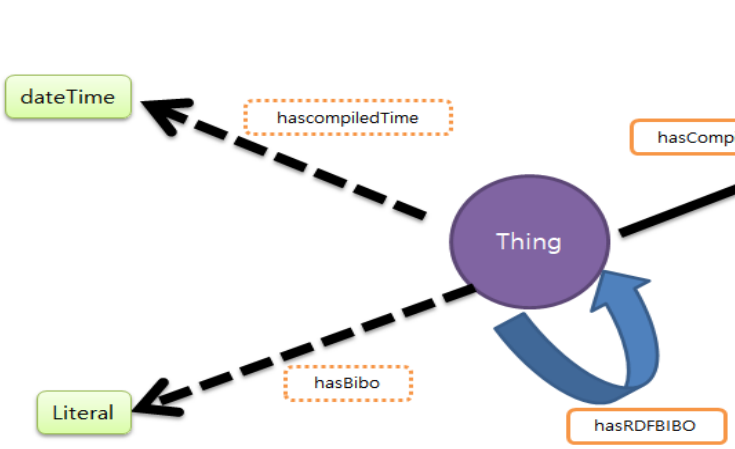
[22] ha 。Domai

n为E

[23] ha e。

[24] ha eral。而

为了



編纂時間的。而且從前的人文情報通過註釋或參考文獻記錄了其判斷的根據，是在紙張的界限上最有效的記錄方法。但是人文數據可以記錄各個情報的判斷根據。

圖五、基本學術模型設計概念圖

筆者以上述的學術情報的分析為基礎，設計了敘述滿足學術要求的“基本學術模型”。基本學術模型的Class為Person（個人）、Geroup（團體），Property為hasCompiler[25]、hascompiledTime[26]、hasBibo[27]、hasRDFBIBO[28]。

最後，筆者綜合了事件模型與學術模型，設計了韓臺《職員錄》Ontology。韓臺《職員錄》Ontology的Class為Agent（行為者），Concept（概念），Bibliography（文獻），Event（事件），Property為hasEventObject, hasEventPostObject, hasEventPreObjct, hasEventType, hasBibo，hasRDFBIBO, hasCompiler， hasCompiledTime，hasTimeValue, hasSpaceValue。

[25] hasCompiler是链接编纂者。Domain为Thing，Range为Person，Group。

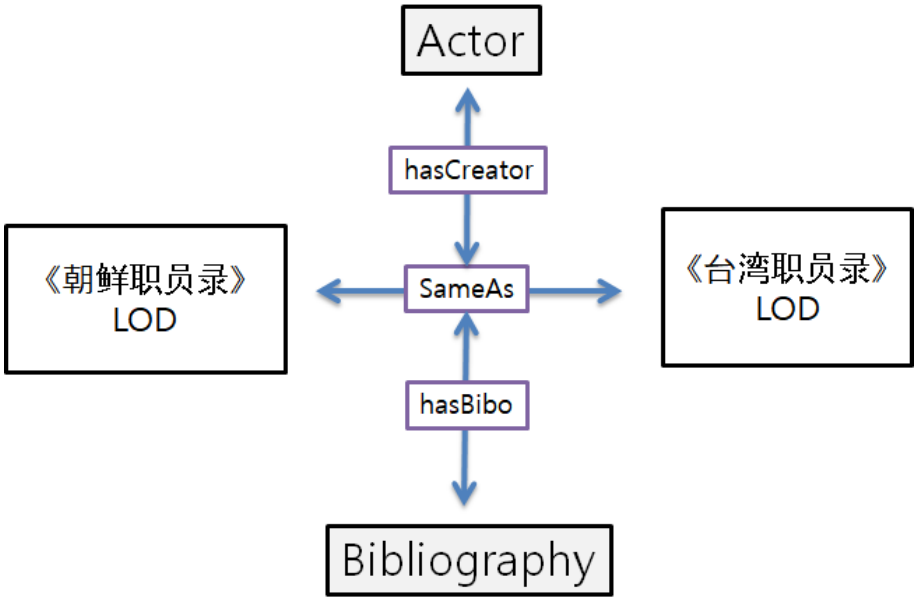
[26] hascompiledTime是链接编纂时间。Domain为Thing，Range为xsd:dateTime。

[27] hasBibo是链接编纂根据。Domain为Thing，Range为rdfs:Literal。

[28] hasRDFBIBO是链接RDF形式的编纂根据。Domain为Thing，Range为Thing。

4. 人文學基本LOD

為了鏈接各自獨立的韓臺職員錄LOD服務 ， 設計了可鏈接韓臺職員錄的Ontology。在設計鏈接職員錄Ontology中，筆者提出了反映人文學研究特質的“人文學基本Ontology”，它以數據形式可敘述了各個研究者對同一對象的不同判斷與其根據。而根據韓國國史編纂委員會《韓國史數據庫》之《韓國近現代人物資料》，找出韓臺職員錄的共同人物，並建設了可鏈接韓臺職員錄的LOD示範服務。

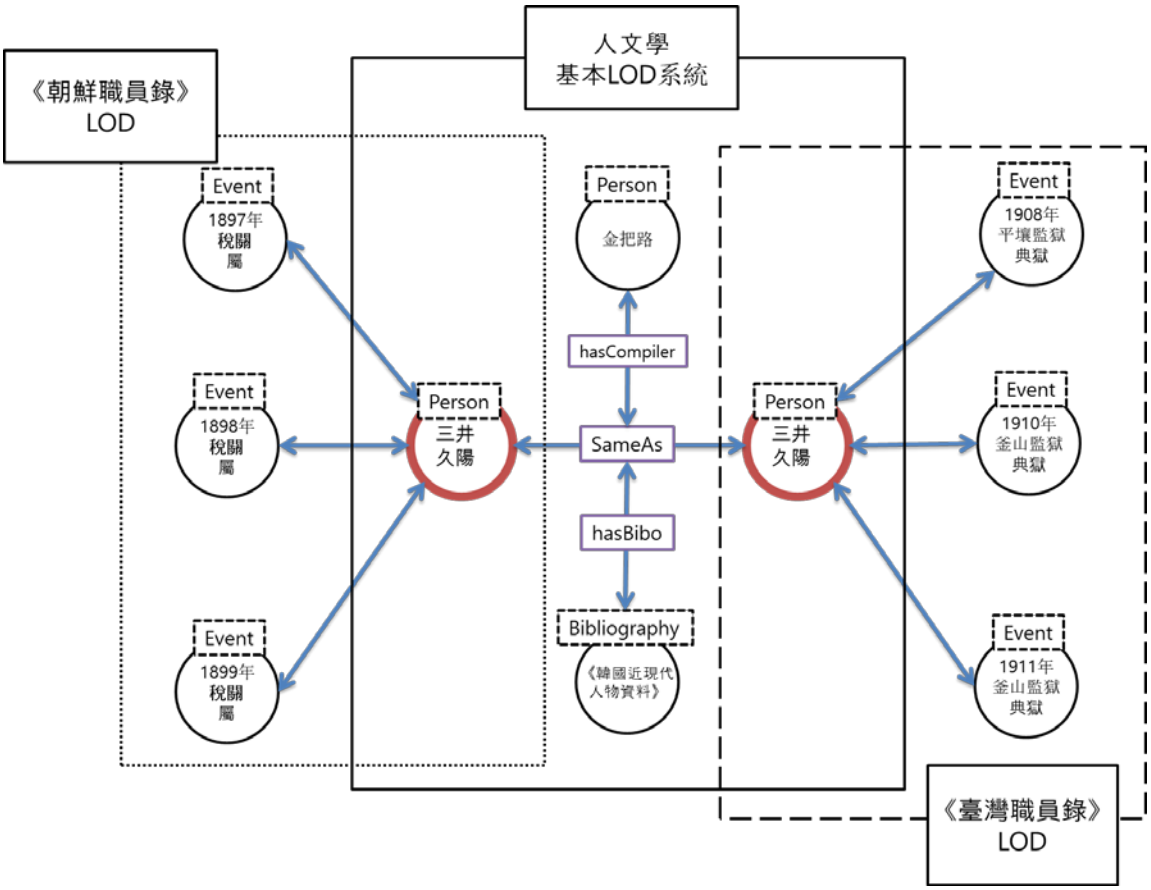


圖六、人文學基本數據LOD的數據連接概念圖

雖然人文學基本LOD可以涉及到諸多內容，如人際關系、註解內容，但是為了系統的簡明性與初步統合過程的現實性，先復雜的內容歸去各自獨立的LOD上處理，而人文學基本LOD集中於各自不同個體之間的相同關系，特別是時間、空間、人物、事件等的人文學基本要素之間的相同關系。

筆者以上述的現狀與學術模型 ， 設計了鏈接各自人文LOD的人文學基本Ontology。人文學基本Ontology的Class為Agent（行為者），Bibliography（文獻），Judgment（判斷），Property為Owl:SameAs[29], hasCompiler, hasCompiledTime，hasBibo，hasRDFBIBO。

[29] Owl:SameAs，*OWL Web Ontology Language Reference*， W3C Recommendation 10 February 2004：https://www.w3.org/TR/owl-ref/#sameAs-def



圖七、人文學基本數據LOD的數據連接示範

《朝鮮職員錄》LOD與《臺灣職員錄》LOD的基礎上，參考《韓國史數據庫》的《韓國近現代人物資料》數據，論者找到《朝鮮職員錄》與《臺灣職員錄》的同一人物26名。例如：石塚英藏出現在1898年至1905年的《臺灣職員錄》，1910年至1916年的《朝鮮職員錄》，1929年至1930年的《臺灣職員錄》。根據《韓國近現代人物資料》，石塚英藏是“聘請韓國議政府顧問官，而中間歷任臺灣總督府慘事官, 關東州民政長官等。當任朝鮮總督府農商工部長官指導開發很多朝鮮殖産興業”[30]。論者根據上述的資料，《朝鮮職員錄》的石塚英藏與《臺灣職員錄》的石塚英藏為同一人物，並且石塚英藏在韓臺兩國都擔任著比較特殊的以土地調查為中心的經濟相關工作而其官階逐步上升。

三井久陽根據《臺灣職員錄》在1897年至1899年擔任安平稅關的屬，根據《朝鮮職員錄》在1908年至1919年擔任釜山監獄與京城監獄的典獄。根據《韓國近現代人物資料》,“1887年任命為神奈川縣看守部長，歷任看守長, 監獄署記；1897年5月辭職後，渡臺灣擔任臺灣總督府稅關屬；1899年1月擔任東京警視廳監督書記, 沖繩縣典獄, 廣島

[30] “韓國議政府 顧問官으로 초빙되어, 중간에 臺灣總督府 參事官, 關東州民政長官 등을 역임하였다. 朝鮮總督府 農商工部長官이 되어 조선의 殖産興業을 지도 개발한 바가 다대하다.”，《朝鮮功勞者銘鑑》

監獄；‘1908年5月由於韓國政府的囑托渡鮮歷任平壤監獄典獄, 統監府典獄, 朝鮮總督府典獄’; 歷任永登浦監獄, 釜山監獄後 ,移到京城監獄至1917年”[31]。因此我們可以判斷《臺灣職員錄》的三井久陽與《朝鮮職員錄》的三井久陽是同一人物。遺憾的是，雖然根據1909年6月18日的《日本官報》的三井久陽的減奉記錄[32]可以查看三井久陽在日本的活動內容，但是日本國立國會圖書館[33]的《日本職員錄》以及有關史料以圖像為中心提供服務，只能針對少數的關鍵詞提供檢索，事實上針對個別人物的檢索是不可能的。

最後，沒有直接的考證證據下，我們可以通過從事土地調查或者軍事的特殊職圈為間接證據。例如：日高仙吉根據《臺灣職員錄》在1902年至1904年擔任臨時臺灣土地調查局測量課的技手，根據《朝鮮職員錄》在1908年至1917年臨時財源調查局與臨時土地調查局的技手與監查官。雖然論者沒有發現證明《臺灣職員錄》的日高仙吉與《朝鮮職員錄》的日高仙吉是同一人物，但是土地測量的特殊技術的同名異人的概率極少，可以判斷為同一人物。

上述的考證是傳統歷史學的基礎工作，已經傳統史學界積累了諸多內容與方法。問題是其內容分散在論文、書籍等的紙張上面，越來越專門化、越來越急速擴張的歷史學界的一位研究者已經無法完全掌握學術情報，或許搜集限定為範圍也需要大量的時間。本論只是歷史學最基礎的工作移植在數位上，而可解決現在歷史學苦惱。

5. 數據共享方法比較

人文數據分散在世界各國各機構與個人。過去為了統合人文數據，大部分以國家力量為背景下建構了各種人文數據統合系統[34]。但是其系統雖然向用戶提供統合檢索等的方便功能，但是事實上，難以改變統合系統的Metadata，無法完全使用各自數據內容，限制新建數據以及開發應用等的諸多問題。與此相反，LOD同時提供各自系統的獨立性

[31] “1887년 12월 神奈川縣 看守部長으로 임명, 이어서 看守長, 監獄署記 역임; 1897년 5월 사직 후 臺灣으로 건너가 臺灣總督府 稅關屬이 됨; 1899년 1월 東京警視廳 監督書記, 沖繩縣 典獄, 廣島監獄 근무; 1908년 5월 한국 정부의 촉탁으로 渡鮮하여 平壤監獄 典獄, 統監府 典獄, 朝鮮總督府 典獄 역임; 永登浦監獄, 釜山監獄 등을 거쳐서 京城監獄으로 옮기고 1917년 현재에 이름”，《在朝鮮內地人紳士名鑑》

[32] 《日本官報》. 1906年03月22日。

[33] 国立国会図書館デジタルコレクション：http://dl.ndl.go.jp/

[34] 如韩国的韩国历史情报统合系统（http://www.koreanhistory.or.kr/）与台湾的中央研究院數位典藏資源網（http://digiarch.sinica.edu.tw/）。

與各自系統之間的連接，如本文所提到的《朝鮮職員錄》與《臺灣職員錄》的各自LOD建設以及韓臺《職員錄》之鏈接。

表三、數據層面的網絡服務、RAWDATA、API、LOD 比較

	网络服务	RAWDATA	API	LOD
数据再加工	X	0	0	0
数据接近程度	△	0	△	0
机器可读性数据	△	△	0	0
数据的维持管理	0	X	0	0
数据之间的连接	X	X	△	0
数据著作权保障	△	△	△	0

如上表，LOD優越於從前的服務。 網絡服務在數據再加工與數據之間的連接層面有問題，網絡服務所提供的數據難以使用在數位分析方法上，而且網絡服務的數據是孤獨的，雖然臺灣總督府職員錄系統提供延伸查詢等的手段，支持與其他人文數據的鏈接，但本質上沒有標準主鍵，因此還是關鍵詞檢索。RAWDATA在維持管理與數據連接層面有問題，RAWDATA無法即時反應數據的更新。目前共享數據最常用的手段是API。但API是管理者向用戶的單方向數據共享技術，所以用戶無法接近全數據，而且數據之間只能間接的鏈接，並且還是無法完全保障盜用的可能性。LOD還是完全保障上述的API之問題。最重要的是，擁有人文數據的機構通過數據之間的鏈接得到數據著作權的保障，因為某人或某機構企圖盜用數據也無法被鏈接其他數據，無法得到網絡的公認。

6. 結論

雖然論者為了各自主題的人文學LOD之間的鏈接，提出以核心要素（時間、空間、人物、事件）為中心的人文學基本LOD之概念。但是韓臺《職員錄》LOD與人文學基礎LOD設計是本質上虛構的。是因為論者只是利用《朝鮮職員錄》的RAWDATA與《臺灣職員錄》的網頁抓取數據。再說，論者本質上不足維護管理《朝鮮職員錄》與《臺灣職員錄》的“明示權限”。各個《職員錄》應當由於韓國國史編纂委員會與臺灣中央研究院來建設、管理《職員錄》LOD。而且人文學基礎LOD也由於國際共同體來建設、管理才能得到數據的可靠性、信賴性以及運營的現實性。

因此本文是為了擴大人文學數據的共享文化，敘述了人文學數據共享的技術背景與共享效果。而我相信史料如命的人文學者近早建設各自的人文數據，並為了提高人文數據的價值進行數據的共享。

參考書目

《朝鮮功勞者銘鑑》， 1935年，http://db.history.go.kr/id/im_215_12008

《在朝鮮內地人紳士名鑑》， 1917年，http://db.history.go.kr/id/im_215_21701

《日本官報》. 1906年03月22日， 國立國會図書館デジタ兒コレクション，http://dl.ndl.go.jp/info:ndljp/pid/2950154/7

<關於公共數據提供與利用活性化的法律（공공데이터의 제공 및 이용 활성화에 관한 법률）>（法律第11956號，2013.7.30.， 制定， 法律 第12844號，2014.11.19.，它法改定，法律 第13723號，2016.1.6，部分改正）； 網絡參考：國家法令情報中心，法制處， http://www.korealaw.go.kr/lsEfInfoP.do?lsiSeq=179039

金炫、林永尚、金把路（2016）， 《數位人文入門》， HUEBOOKs（김현·임영상·김바로, 『디지털 인문학 입문』, HUEBOOKs, 2016）

金把路（2017）， <制度與人事的關系性數據數位檔案的建立及應用 - 以近代學校資料(1895~1910)為中心 ->. 韓國學中央研究院韓國學大學院博士學位論文，(김바로,「제도와 인사의 관계성 데이터 아카이브 구축과 활용 - 근대 학교 자료(1895~1910)를 중심으로 -」, 한국학중앙연구원 한국학대학원 박사학위논문, 2018)

職員錄資料, 韓國史數據庫, 國史編纂委員會：http://db.history.go.kr/item/level.do?itemId=jw

臺灣總督府職員錄系統，臺灣中央研究院臺灣史研究所：http://who.ith.sinica.edu.tw

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 12/21 (W5)
Venue 地點 / 法鼓文理學院 Dharma Drum Institute of Liberal Arts (DILA)

近現代中國「新學」概念與新知識系譜的數位人文研究

A Digital Humanities Study on the Concept of "Xin Xue(新學)" and the Genealogy of Knowledge in Early Modern China

鄭文惠*　葉昱廷**　梁穎誼***
國立政治大學中國文學系特聘教授*
國立政治大學統計學系碩士生**
私立逢甲大學風險管理與保險學系助理教授***

近現代中國「新學」概念與新知識系譜的數位人文研究

A Digital Humanities Study on the Concept of “Xin Xue(新學)” and the Genealogy of Knowledge in Early Modern China

鄭文惠※　葉昱廷※※　梁穎誼※※※

摘要

本文擬透過數位人文方法分析「新學」作為重要關鍵詞，所呈現近現代中國的新知識概念與新知識系譜的建構。本文語料取自於「中國近現代思想史專業數據庫（1830-1930）」，係包含中國近現代報刊、雜誌、傳教士和西方著作中譯本及各種文集，計有88647篇文章，字數達1.2億字的數據庫。[1]

本文主要從詞彙與概念入手，藉由數位技術方法，分析近現代中國面對全球知識流動所產生的文化雜交與文化斡旋等有關文化交涉與符號混成(semiotic syncretism)的過程裡，概念、話語、事件、行動之間的關係與互動，希望能勾勒出近現代中國新知識的傳播與接受過程裡多層向的概念生成與情感樣態，及知識產製、知識結構、學校體制、學科體系等現代轉型的大歷史圖像。

為了有效勾勒近現代中國新知識概念與新知識系譜，本文結合觀念史／ 概念史理論方法與數位技術與統計分析，帶入新問題新視野，冀望更多元更多向的挖深主題。本文主要考察近現代中國在跨語際實踐(translingual practice)的過程裡，如何面對自我與他者的異置與共在、共性與殊性的交錯與互動，及透過文化交涉與斡旋，文化母體如何蛻變再生；依違在「新學」之新知識／世界知識的接受與抗斥的多重影響焦慮下，知識生產與話語爭奪如何展演在文化交涉的符號混成過程裡；跨文化流動或全球知識流動，在近現代中國如何呈現或達致某些層次意義上的文化斡旋與融攝或轉化，從而促成新概念的形成與觀念的內塑，進而使文化母體革命性的蛻變與再生；為了深化議題，本文以「詞彙」為研究焦點，透過「新學」一詞所指涉近現代中國新知識概念與新知識系譜建構的重要「關鍵詞」之詞頻、共現詞叢、概念關係網絡及年代分布等觀察，嘗試勾勒這一系列關鍵詞所指涉的概念形成與變化軌跡，及概念、事件、話語、行動互動的大歷史圖像。

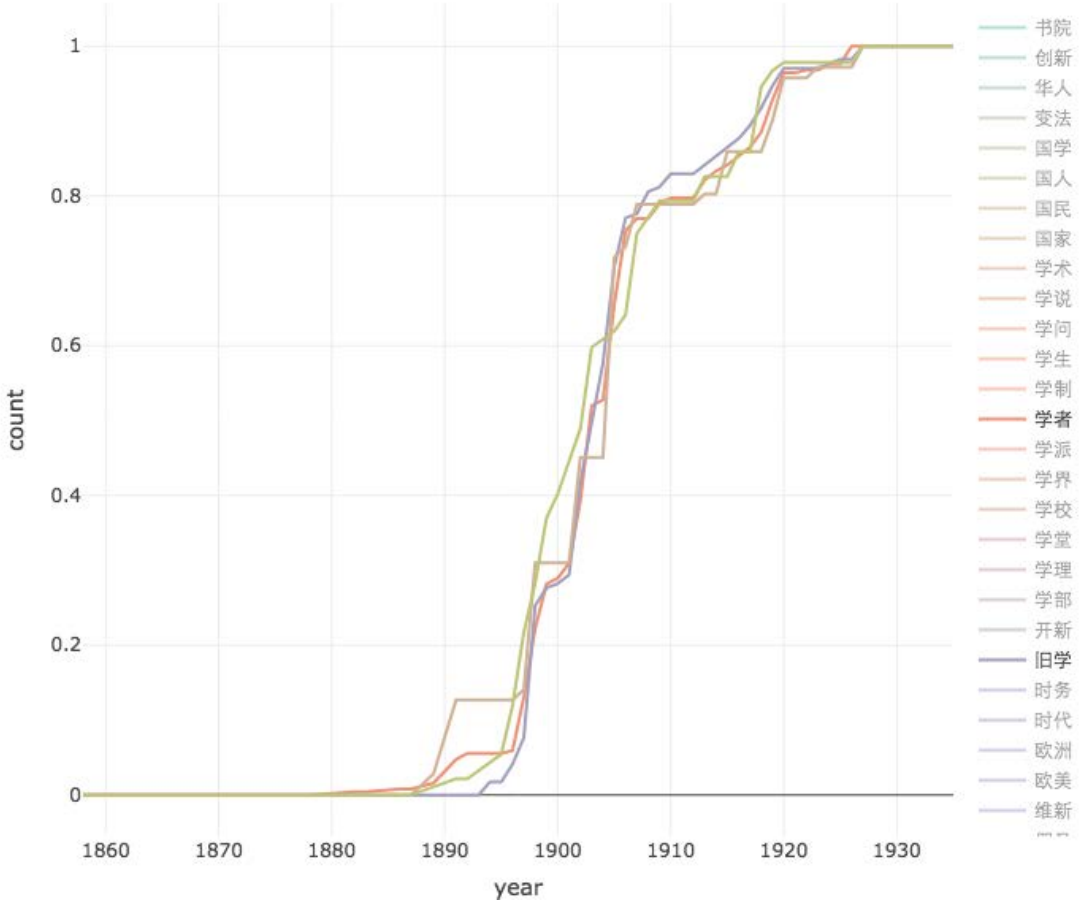
[1]本文研究中關於「新學」詞彙資料，係取自「中國近現代思想史專業數據庫(1830-1930)」 (香港中文大學中國文化研究所當代中國文化研究中心開發，劉青峰榮譽研究員主編)，現由台灣政治大學高教深耕計畫「東亞文化傳統及其現代轉型」國際拔尖計畫(主持人：金觀濤講座教授、鄭文惠特聘教授)項下「中國近現代思想及文學史專業數據庫(1830-1930)」持續開發功能與完善數據庫並提供檢索服務。

本文將「中國近現代思想史專業數據庫(1830-1930)」視為一個巨型文本結構體，係能反應近現代中國轉型的總體社會政治與文化變貌。本次分詞以 N-gram 為基礎，因研究所取用的文本為近現代中國文獻，因而分詞的基本單位為音節(syllable)。本文先將作為分析對象的關鍵詞詞條，提取前後 50 個字元，並篩選過濾噪音詞，共獲致「新學」關鍵詞 ，共 2063 個詞條。先將文本中所有出現「新學」字詞，提取前後 50 個字元，篩選分離出共 2063 個「新學」詞條，再針對這些詞條以 N-gram 分詞(N=1~6)，建立「新學」的共現詞表。若以新學為主體，以舊學作為對比，除了各別觀察兩者的共現詞以外，又進一步擷取同時出現「舊學」字詞的「新學」詞條，再進行斷詞，藉此了解當「舊學」出現於「新學」詞條時，其共現詞與各別出現的差異；反之亦然。此外，也運用 CUSUM 研究方法進行系列關鍵詞的分群研究，除利用關鍵詞歷時性線性增長時間相似性進行分群外，也加入統計學中關聯相似性進行分群研究，突出眾多關鍵詞在時間上的脈絡，以突出關鍵詞所指涉的概念如何隨著時間、事件、話語、行動而消長而變化。近現代中國邁向現代性歷程中，知識分子在跨語際實踐過程中援引中國傳統及歐西與日本等各方思想資源，由天下進入世界體系中，在與帝國霸權文化擴張主義進行政治頡頏、文化協商的同時，一方面也隨著西學／東學等世界知識的輸入而進行價值改造；價值的改造又與國族認同，新型國家的創制、國民的素質息息相關，因此或可說「新學」關鍵詞中，深刻的積澱了也刻劃出歷史轉變的軌跡，諸如西學東漸或東學西漸；或是中學為主、西學為輔；以中學包羅西學；西學淩駕中學；中學為體、西學為用、體用兼備；或是以新學接於舊根、以舊學為新學之根等「移花接木之法」……等爭辯。或是新政新學、新學新器、格致新學、新學新理、新法新學等涉及學理、物質、政治等各層面的詞彙組合。其中也因著時間的遞嬗及局勢的轉變，「新學」又與新政、新法、變法、維新、政治、革命與國家富強、求新息息相關，又指涉於精神、文明、思想、自由、科學等；又與人才深切相關，而有設立新學部，以廣其教，以擴民智，以產製新知識，因而，教育、人才、學生、學堂、學校、書院……等等往往是共現詞彙，而學者、學士、學說、學術、學問、學界、學風、學派……是直指西學／洋學／新學／夷學／東學的核心人物與學理、派別。至於中國與西人、日本、印度、泰西、西國、歐洲、歐美、羅馬、希臘、世界等，及與舊學、孔子、國學、守舊、八股、古學、中學……等對舉；或因歐戰爆發、日俄戰爭、新文化運動及無產階級革命等政治局勢或文化轉向，「新學」往往被調動為論述的理據。其他如講求、提倡、振興、設立、輸入、發明、創新；或不可、不能、不知、不足、未能；或可以、莫不、無不、未嘗、往往、足以、無一、皆有、必有等常見動詞、肯定詞或否定詞使用所呈顯的行動方案或心理情感；乃至於今日、十年、今之、年前、每年、年來、數年、至今、之後、一年、三年、以後等時間詞之頻繁使用所表徵的當下急迫性……。

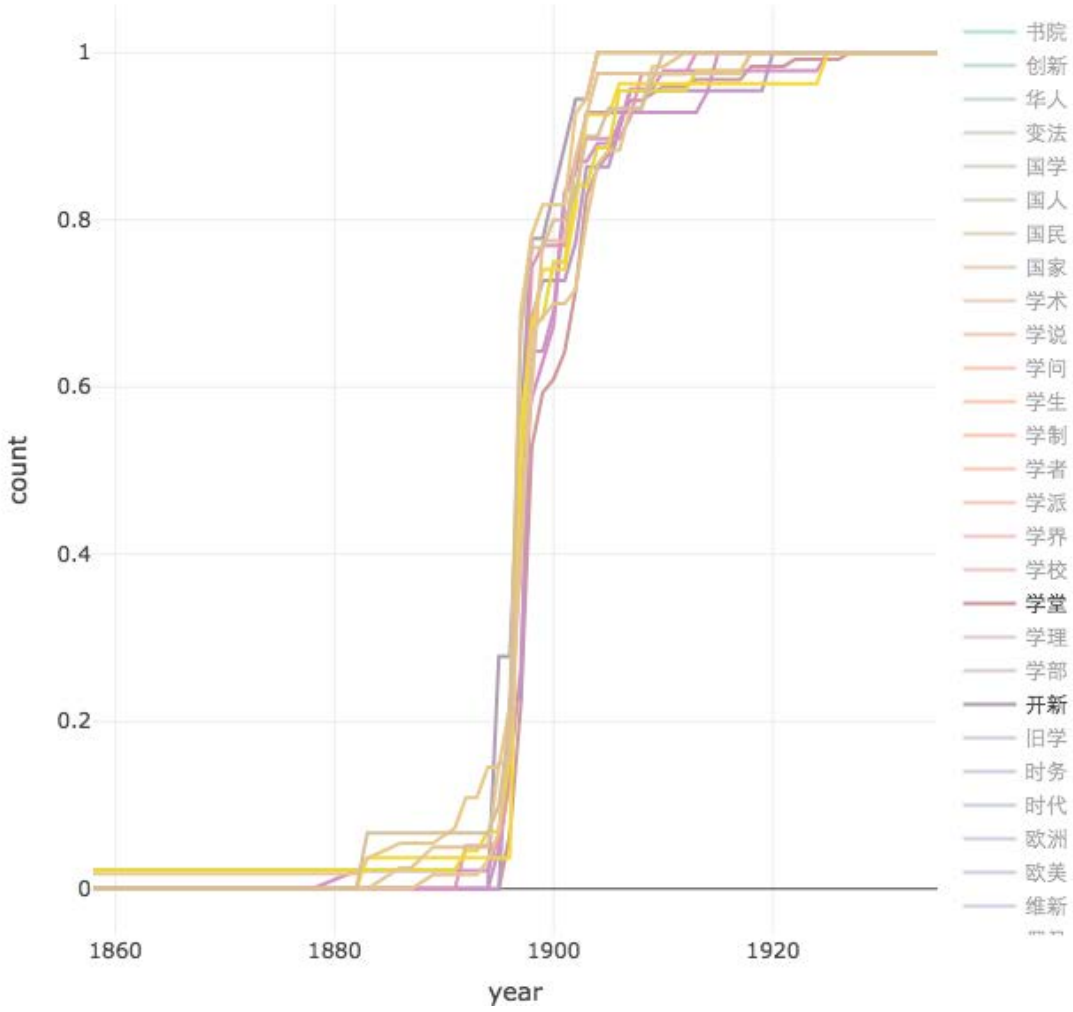
本文除論述「新學」概念為何集中於甲午戰爭後到新文化運動前的二十年（1895-1915）間之外，為有效掌握「新學」共現詞彙所呈顯的概念的集群性，又採用 CUSUM計算，製作這些共現概念之歷年共現比例累加圖。將共現的眾多詞叢加以聚落區分，考察為何某一群與「新學」共現的概念詞叢會在某一時間點同步陡增，進而從大量資料中，過濾出重要的概念分群結構，以考

掘「新學」論述系統的各個概念構面，以及「新學」概念與事件、話語、行動的互動軌跡。透過比對共現關鍵詞歷年升降比例變化的模型，也可以看到具有相同升降模式的共現關鍵詞，大致可視為同一個論述主題模型。

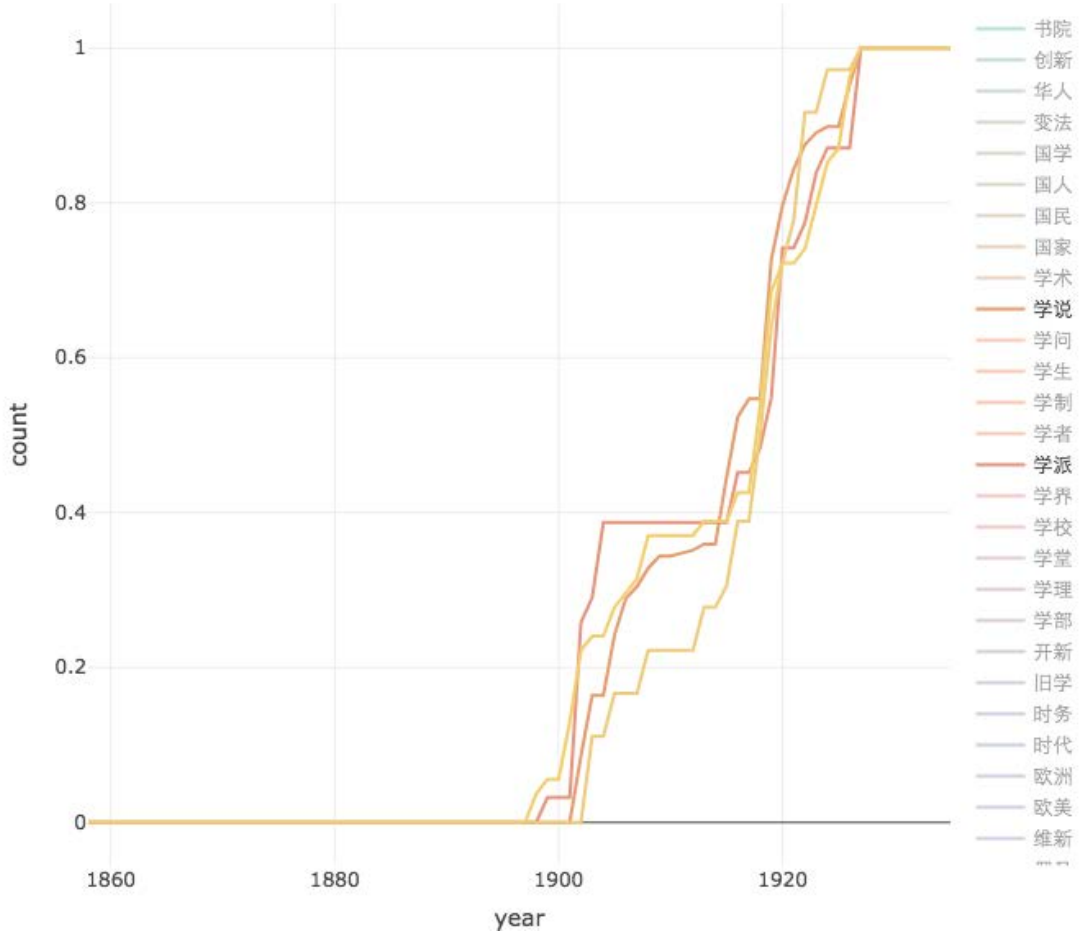
本文透過CUSUM計算，製作「新學」共現概念之歷年共現比例累加圖，著眼於年代位置相似性、共現累加相似性，分析、勾勒「新學」如何及為何作為一個新概念，及其與事件、話語、行動的相互關係與互動軌跡，從中論述國學、學者、舊學、歐洲、孔子、日本；學堂、開新、譯書、風氣、人才、大臣、西人、希臘、格致、泰西、教習、富強；文明、自由、我國、革命、國學、國民、學界、學者等分群概念結構所表徵的主題論述核心。或是學說、學派、研究、科學等分群概念於1915年同步陡升；學說、學派、學術、學問、大學、世界、研究、科學、智識、精神等分群概念於1917-1918年同步陡升；吾國、時代、學理、道德等分群概念於1903-1905年同步陡升；外國、本國、吾國、國人、學理、時代、輸入、歐美等分群概念於1918-1920年同步陡升；國家、維新、八股、科舉、新政、新理、變法等分群概念於1905年密切相關，科舉、八股、維新等分群概念於1907年緊密相關；書院、華人、學部、開新、羅馬、格致、教皇、富強、人才、大臣、希臘、泰西、教習、朝廷、新法等分群概念於1896-1907年緊密相關；學生、大學、智識、小學、學理、時代、輸入、政治、新舊等分群概念於1906-1924年緊密相關的原因及其表徵的意涵，除建立分群概念與主題論述的研究模型外，也冀望探勘近現代中國一切的變革，以極短時間高度濃縮了跨文化知識流動與文化擴張主義下文化斡旋所生成的劇烈的社會文化轉型，在跨語際實踐的過程，新知識的產製與新知識系譜的建構，如何生成新概念，而有效內塑，進而形成新的話語體系與行動策略而與原文化母體的觀念系統磨合出新的蛻變與再生。



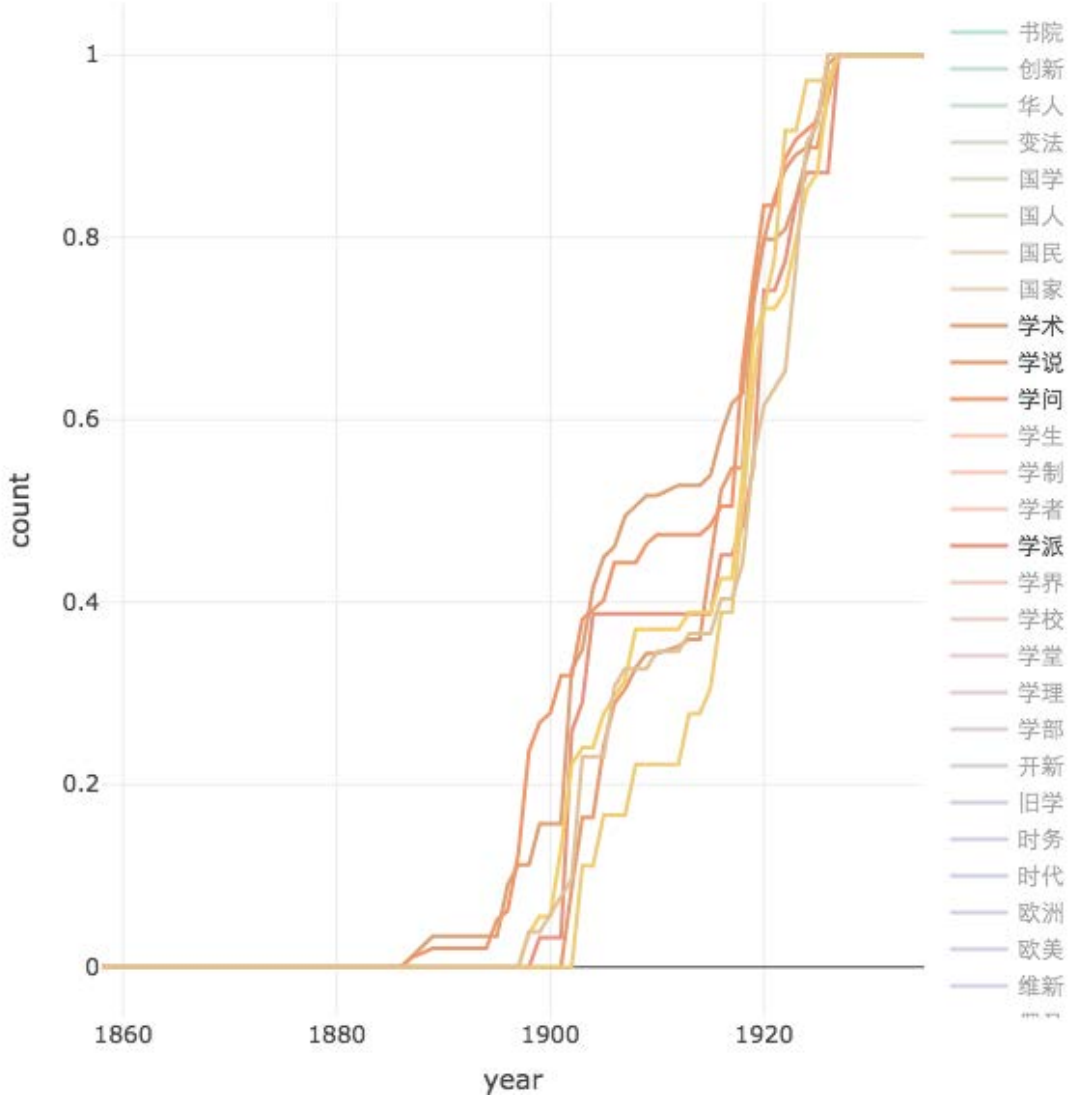
圖一　「新學」共現詞叢與分群概念之一：國學、學者、舊學、歐洲、孔子、日本



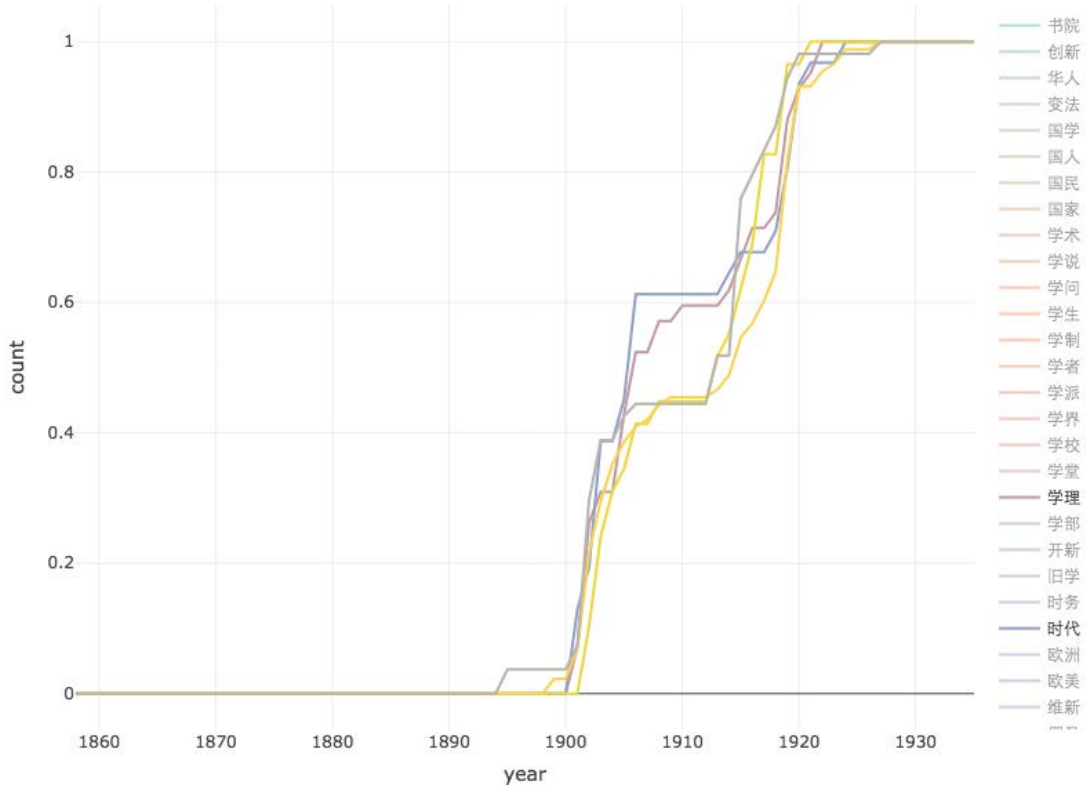
圖二　「新學」共現詞叢與分群概念之二：學堂、開新、譯書、風氣、人才、大臣、西人、希臘、格致、泰西、教習、富強



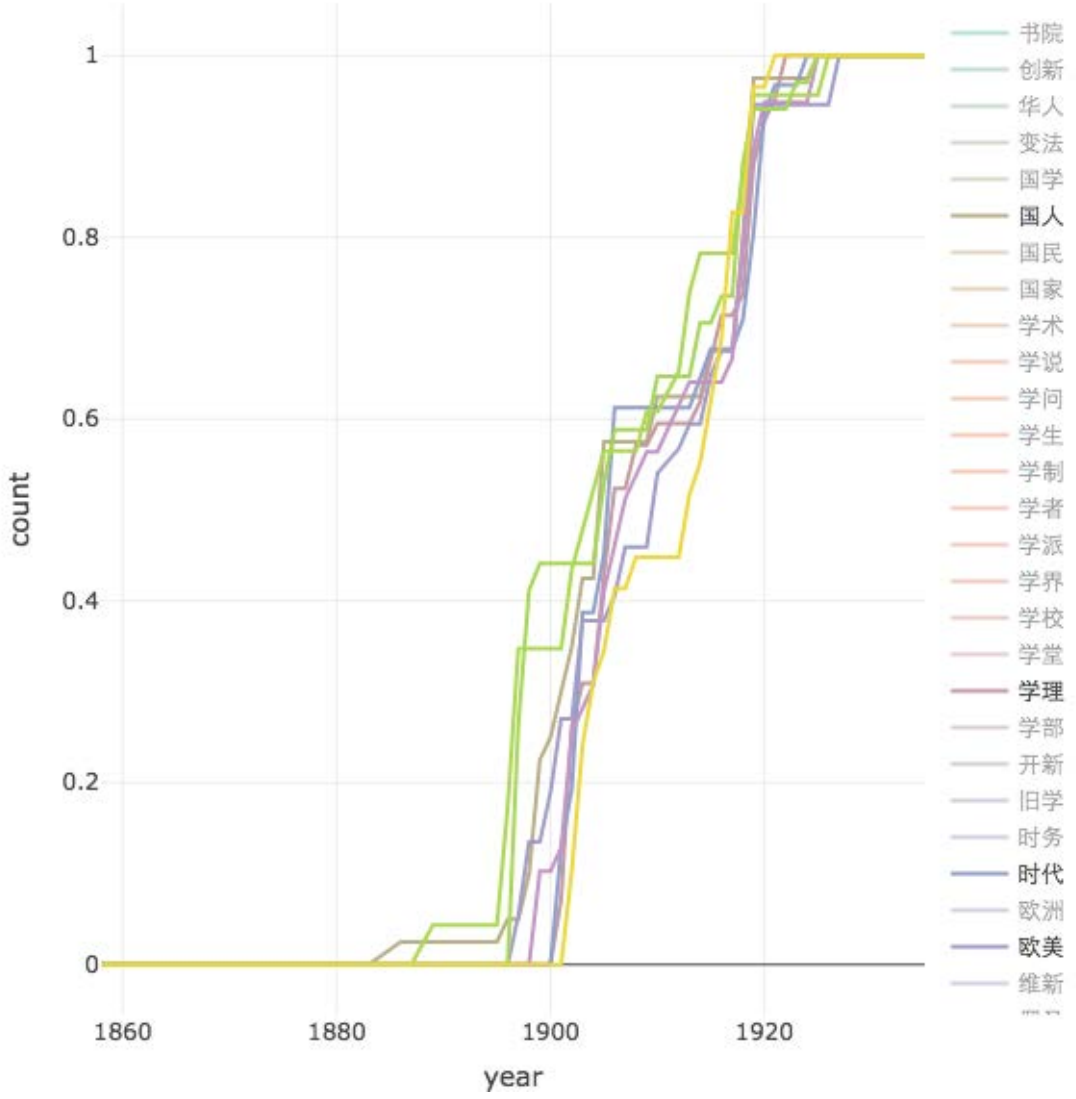
圖三「新學」共現詞叢與分群概念之三：學說、學派、研究、科學



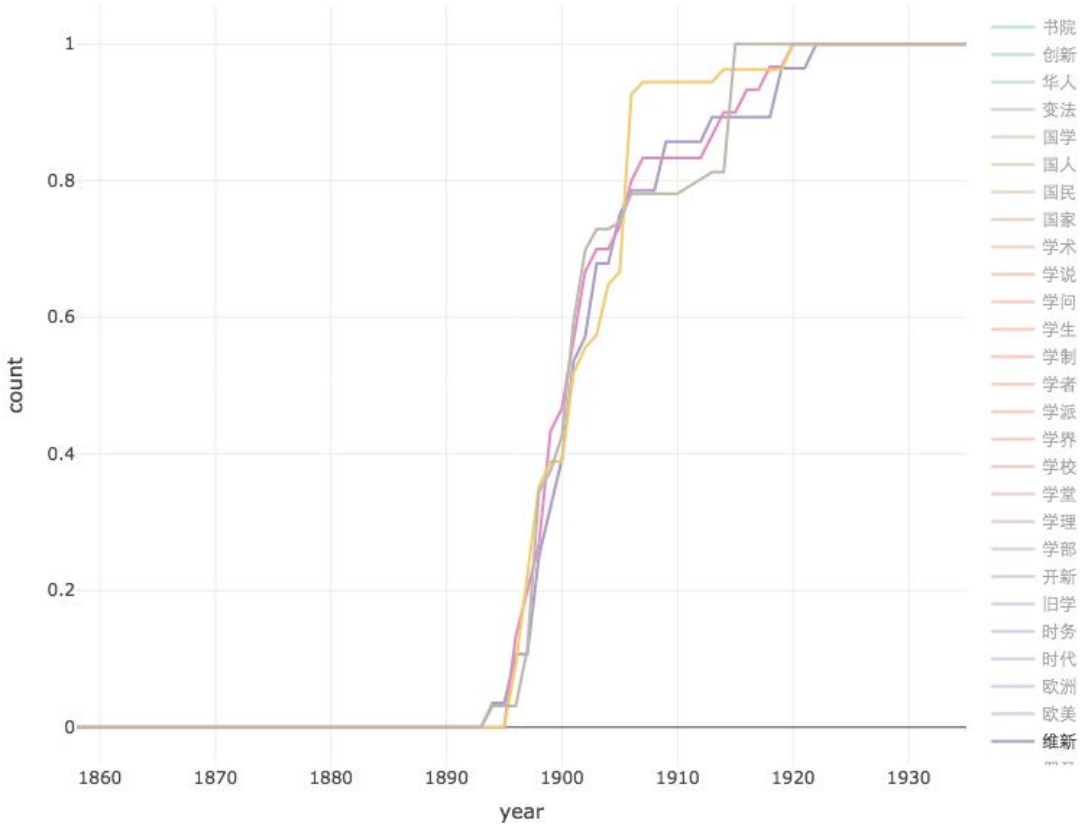
圖四 「新學」共現詞叢與與分群概念之四：學說、學派、學術、學問、大學、世界、研究、科學、智識、精神



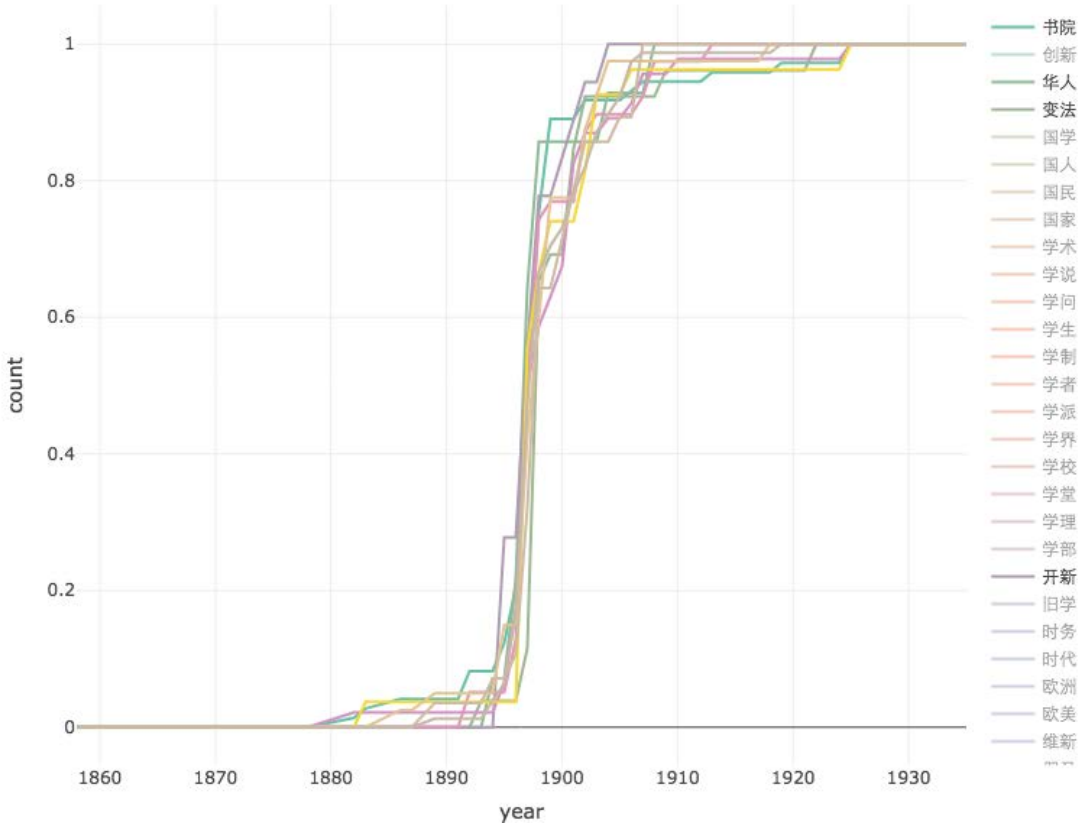
圖五 「新學」共現詞叢與分群概念之五：吾國、時代、學理、道德



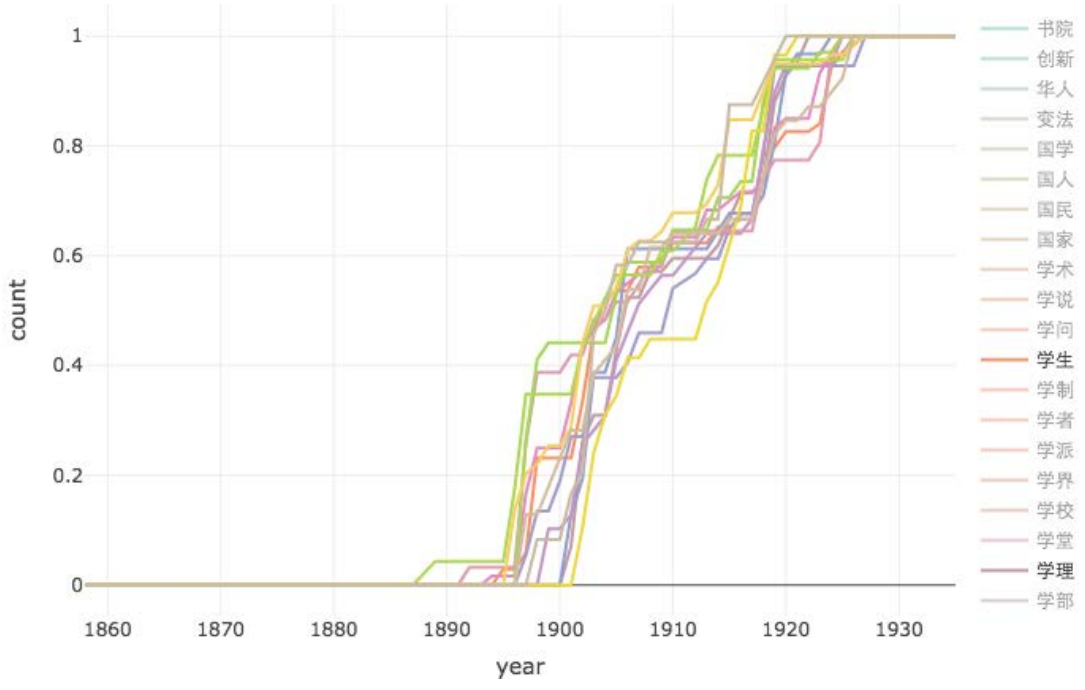
圖六 「新學」共現詞叢與分群概念之六：外國、本國、吾國、國人、學理、時代、輸入、歐美



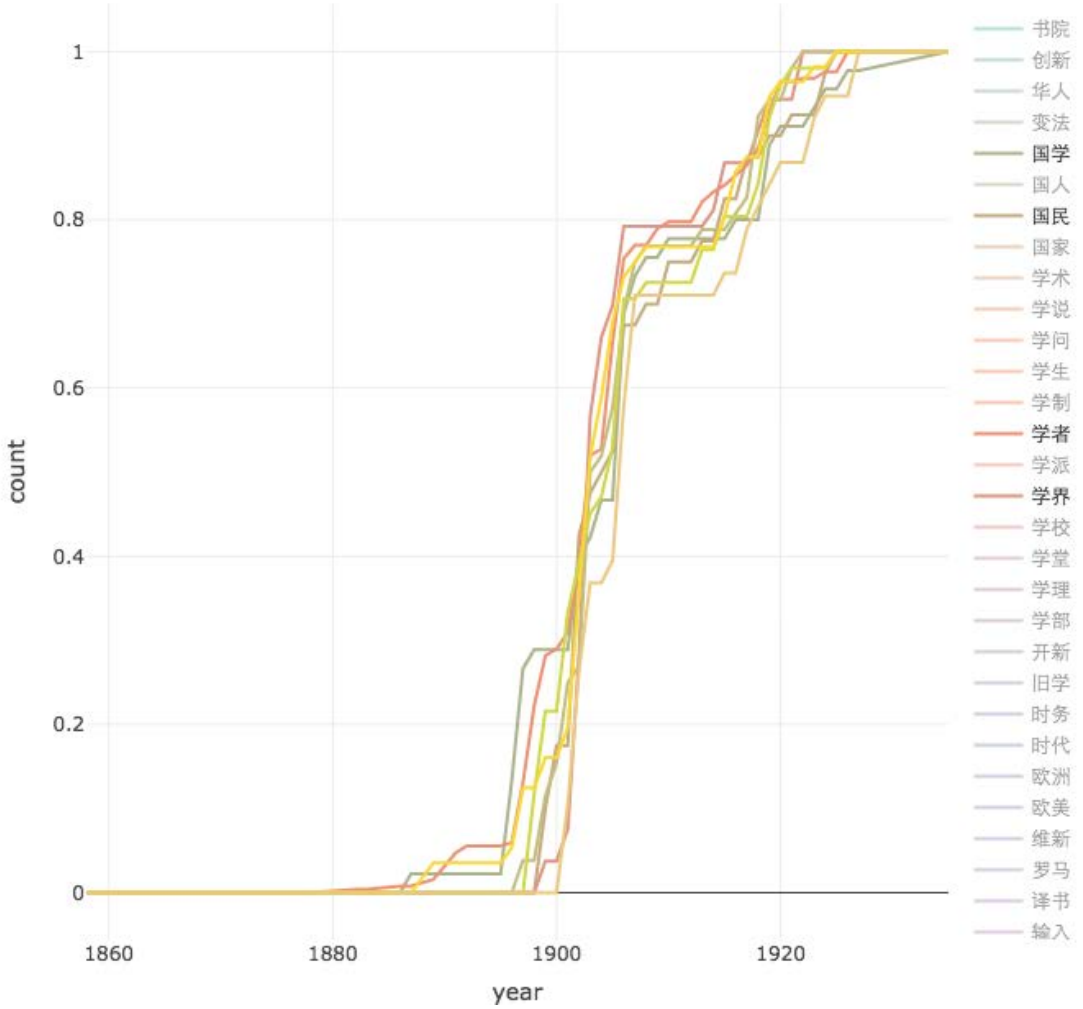
圖七 「新學」共現詞叢與分群概念之七：國家、維新、八股、科舉、新政、新理：變法



圖八 「新學」共現詞叢與分群概念之八：書院、華人、學部、開新、羅馬、格致、教皇、富強、人才、大臣、希臘、泰西、教習、朝廷、新法



圖九 「新學」共現詞叢與分群概念之九：學生、大學、智識、小學、學理、時代、輸入、政治、新舊



圖十 「新學」共現詞叢與分群概念之十：文明、自由、我國、革命、國學、國民、學者、學界

關鍵字：近現代中國、新學、非監督學習、相似度測量、集群方法

※ 鄭文惠 國立政治大學中國文學系特聘教授 whcheng123@gmail.com

※※葉昱廷 國立政治大學統計學系碩士生 yeyuting0307@gmail.com

※※梁穎誼 私立逢甲大學風險管理與保險學系助理教授 yyleong@fcu.edu.tw

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12 18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

清人「拗救」說再審視
以《全唐詩》15290 首律詩為樣本

諸雨辰 胡韌奮
北京師範大學文學院中文信息處理研究所講師

清人「拗救」說再審視
——以《全唐詩》15290 首律詩爲樣本

諸雨辰 胡勒奮

講師

北京師範大學 文學院 中文信息處理研究所

摘要

本文基於計算機自動標注技術，對《全唐詩》15290 首律詩中不合格律規範的「拗句」進行分析，標注共探測出 13683 句單句拗句、2011 聯成對拗句。以這些統計數據爲參照，可以梳理清代王士禛《律詩定體》、趙執信《聲調譜》以及今人王力《詩詞格律》、啟功《漢語現象論叢》等著作中提出的唐詩「拗救」說，辨析其合理性與局限性。通過統計分析，最終可將「平平平仄仄」句式的「四拗三救」，以及對句「仄仄平平仄，平平仄仄平」句式的「三拗三救」認定爲唐詩中常見的單句拗救與對句相救。而「平平平仄平」與「平平仄仄仄」則屬於唐詩中常見的特殊句式。「拗救」現象體現了唐詩格律的「常中之變」，而對「律句」與「拗句」的統觀，亦可總結出其中的「變中之常」，當以辯證的視角加以審視。

目次

1.數據來源及統計結果

2.清人「拗救」說及其局限

3.唐詩格律的再審視

關鍵詞

拗救，《全唐詩》，格律，平仄

關於唐詩的格律，素有「拗救」之說。一般認爲始于清代王士禛（1634—1711）的《律詩定體》以及趙執信（1662—1744）的《聲調譜》。《四庫全書總目提要》稱：「執信嘗問聲調于王士禛，士禛靳不肯言，執信乃發唐人諸集，排比鉤稽，竟得其法。」[①]謂趙執信的理論是受王士禛詩學影響而發，《聲調譜》也確在《律詩定體》基礎上大大細化了對詩律的分析。其後，又不斷有詩律論著產生，將「拗救」說系統化、條目化、複雜化。今人研究中，則以王力的《詩詞格律》爲代表，他總結了三類拗救法：本句自救、對句相救和半拗[②]，將拗救視爲一種特殊的格律規則。

中國古代文論一旦體系化了，往往就會引起學者的質疑，「拗救」說也不例外。比如趙克剛[③]、袁慶述[④]、鐘如雄[⑤]等人分別以杜律、《唐詩三百首》等文本爲例，反駁了「孤平拗救」之說。龔祖培通過考察宋元以來「拗」的概念以及王士禛與趙執信的理論傳播情況，並結合對杜律的統計分析，全面反駁了「拗救」理論，認爲「拗句」現象的出現主要是基於仿古的修辭之風[⑥]。此外，霍松林以《唐詩別裁集》近百首詩歌爲例，提出清人的「拗救」說遠不足以解釋詩律的變例[⑦]。

對於唐詩豐富多樣的文本寫作來說，古人不難從一定範圍內的文本中總結出某種規律，今人當然也可以用其他詩例反駁某種規律。還是尚永亮的評價比較客觀：要確立或駁倒某種理論，單靠幾十首、甚至幾百首的詩例是遠遠不夠的，必須以統計學的方法衡量其百分比，結果才有意義[⑧]。傳統的詩學研究因爲精力所限，不得不以選本作爲研究對象，而無論以杜詩、《唐詩三百首》還是《唐詩別裁集》，都難以避免因選擇文本而造成的樣本偏差，無法全面回答圍繞「拗救」說而形成的理論問題。現代技術的發展則使我們有可能對更大量的文本進行地毯式的普查，本文將研究對象放寬到《全唐詩》中，以《全唐詩》（含補編）中的全部律詩作爲研究對象，應用信息技術統計其平仄分佈規律，以期全面而有效地辨析清代以來關於唐詩「拗救」的詩歌理論問題。

1.數據來源及統計結果

本研究抓取《全唐詩》全文數據庫[⑨]作爲統計研究的電子文本，共計 42863 首，並補充《全唐詩補編》收錄的唐詩文本，共計 6327 首，總計 49200 首詩作爲總體研究對象。首先篩選出其中字數爲 40、56 字的詩（即符合律詩字數），並刪去屬於《樂府雜曲》、《橫吹曲辭》、《相和歌辭》等卷中的樂府詩，再根據題目，刪去題爲《戰城南》、《巫山高》等樂府體詩，並對其中因作者歸屬存在爭議而重複的詩進行去重處理。需要說明的

① 永瑢等，《四庫全書總目》，頁 1784。
② 王力，《詩詞格律》，頁 34—38。
③ 趙克剛，〈關於犯孤平〉，《重慶師院學報哲社版》4，頁 46-53。
④ 袁慶述，〈從〈唐詩三百首〉看「拗救」問題〉，《中國文學研究》2，頁 41-43。
⑤ 鍾如雄，胡娟，〈論唐詩「孤平拗救」說之不成立〉，《雲南師範大學學報》6，頁 77-82。
⑥ 龔祖培，〈漢語詩歌「拗救」說辨偽〉，《文史哲》5，頁 77-96。
⑦ 霍松林，〈簡論近體詩格律的正與變〉，《文學遺產》1，頁 104-117。
⑧ 尚永亮，〈唐人作詩是否拗救？這是一個問題〉，《中國韻文學刊》4，頁 117。
⑨ http://www16.zzu.edu.cn/qts/ ，2017 年 3 月 27 日訪問。

是，近體詩包括絕句與律詩，近體絕句同樣須符合格律規範，也應該是考察「拗救」問題的對象。但是從唐詩體制上說，古體詩一般很少四聯八句，因此可以通過字數標準較爲快速地提取出律詩，而近體絕句則與古體絕句沒有字數上的區別，不易提取。爲了回避甄別古體絕句的技術問題，本文暫不將絕句納入考察對象。

我們使用計算機自動標注技術，對所選律詩用字的平仄進行標注。因《平水韻》一般被認爲代表了唐人作詩的押韻情況，所以本研究使用「搜韻網」的《平水韻》電子文本[①]作爲標注詞典。自動標注主要分爲兩個步驟：

第一步，依據《平水韻》將唐詩中的所有漢字分爲四類：(1) 無爭議的平聲字，記爲 A；(2) 無爭議的仄聲字，記爲 B；(3) 因多音字現象導致既被平聲韻收錄、又被仄聲韻收錄的字，比如「分」、「長」、「中」、「更」等，記爲 C；(4)未被《平水韻》收錄的字，如「杓」，記爲 D。D 類字數量很少，可人工標注。

第二步，對詩句中的每個字進行平仄標注。A 類和 B 類字的平仄所屬沒有爭議，但 C 類字則很難界定。爲了解決這個問題，我們提出一個假設：詩人在作詩時，在保證詩意順暢表達的情況下，會儘量追求使其用字符合律詩的標準格律。（譬如唐詩有借音對、扇面對、錯綜對等特殊的對仗形式，既實現了詩意的流暢表達，亦遵守了格律規範，正可視爲對此假設的間接證據。）在此假設下，對於一句之中出現多音字的情況，若某字在歸入某韻部時，可以匹配上律詩的基本格律，則自動視爲律句。比如：「天邊看淥水」（李白《廣陵贈別》，頁 1787）[②]中「看」字分屬平聲十四寒和仄聲十五翰兩個韻部，視爲平聲則符合格律，視爲仄聲則出律，那麼就默認將其視作符合格律的十四寒。計算機進行標注時，依次將句中 C 類替換爲 A、B 類字，如替換後可使整句詩合乎格律，則取替換後的標注結果。對於無論如何劃歸韻部也無法匹配格律的詩句，則進行人工標注。本步驟設定的假設不僅可以大大減輕人工標注的工作量，還可以規避一些因具體詩例判斷而產生的爭議，比如王力所舉對句相救的例子：「野火燒不盡，春風吹又生。」（白居易《賦得古原草送別》，頁 4847）後人就以「不」可爲平聲、「吹」可爲仄聲爲由說明此句根本不拗，進而推論對句相救的不成立[③]。而一旦我們將所有可平可仄的字都默認取合律的平仄，自然可以避免陷入這些爭議。

此外，本文研究律詩的平仄格律，而無意於糾纏「拗體」的問題。因而在對所選律詩進行平仄標注後，我們假設一首詩中若有五句以上（含）的句子不符合基本格律，可以視爲詩人寫作該詩時無意於遵守詩律規範，則該篇當屬古體或拗體，這類詩均不納入本研究的討論範圍，亦從樣本數據中刪除。

經過以上對《全唐詩》文本數據的預處理，最終得到 15290 首律詩，其中五言詩 9297 首（60.8%）、七言詩 5993 首（39.2%）。從中提取出 13683 句不合律的拗句，占總句數

① http://sou-yun.com/QR.aspx ，2017 年 3 月 27 日訪問。
② 彭定求等，《全唐詩（增訂本）》，頁 1787。下引唐詩，皆據此本，僅注作者、篇名、頁數。
③ 龔祖培，〈漢語詩歌「拗救」說辨偽〉，《文史哲》5，頁 86。

的 11.19%，其中五言 10024 句（73.26%）、七言 3659 句（26.74%）。據此計算，五、七言律詩使用拗句的比例分別爲 13.48%與 7.63%，前人提出五律較七律更多使用拗句的結論[①]，於此可以得到驗證。爲了考察唐詩格律中的拗救現象，需要對這 13683 句拗句的平仄規律及其分佈情況進行詳細統計。

前人討論拗救，一般分單句與對句兩類，本研究也分別統計單句和對句兩種情況下的平仄組合。首先看單句拗救。在 13683 句拗句中，包括五、七言共 134 種平仄排列組合方式，當然其中大多數的出現頻率都很低，沒有討論價值。而因爲七言律詩首字平仄可不論，所以同一種句式，在五言只有一種，在七言則有兩種，比如與五言「平平仄平仄」對應的七言句式就包括「仄仄平平仄平仄」和「平仄平平仄平仄」兩種，在統計時應當把相應的七言句合併處理。處理後五、七言拗句占比高於 1%的句式如下：

表 1.《全唐詩》中本句單拗頻次表

序號	平仄組合類型	頻次	占比	序號	平仄組合類型	頻次	占比
1	（仄）仄平平平仄平	929	25.39%	12	（仄）仄平平仄仄仄	94	2.57%
2	平平仄平仄	2542	25.36%	13	仄仄平仄仄	198	1.98%
3	（仄）仄平平仄平仄	927	25.33%	14	（仄）仄仄平仄仄平	67	1.83%
4	平平平仄平	1877	18.73%	15	仄平仄平仄	147	1.47%
5	平平仄仄仄	1382	13.79%	16	仄仄平平平	116	1.16%
6	（平）仄仄平平仄平	503	13.75%	17	（平）平仄仄平仄仄	42	1.15%
7	仄平平仄平	1243	12.40%	18	仄平仄仄仄	112	1.12%
8	仄仄仄平仄	957	9.55%	19	（仄）仄平平仄平平	39	1.07%
9	（平）平仄仄仄平仄	282	7.71%	20	仄平仄仄平	102	1.02%
10	平仄仄平仄	737	7.35%	21	（平）平仄仄平平平	37	1.01%
11	（仄）平平仄仄平仄	268	7.32%				

再看對句相救。所謂對句相救是指若出句「拗」了，就在對句相應位置變換平仄去「救」，所以只有一聯之內的兩句皆不合律時，才可稱爲對句相救。在所取樣本中，一聯內兩句皆不合律的總計 2011 聯，其中五言聯 1506 聯（74.89%）、七言 505 聯（25.11%），仍取將七言首字平仄合併以後，占比高於 1%的組合類型，其出現頻次如下：

表 2.《全唐詩》中對句雙拗頻次表

序號	平仄組合類型	頻次	占比
1	（平）平平仄仄平仄_（平）仄仄平平仄平	134	26.53%
2	（平）平仄仄仄平仄_（平）仄仄平平仄平	109	21.58%
3	仄仄仄平仄_平平平仄平	319	21.18%
4	仄仄仄平仄_仄平平仄平	305	20.25%
5	平仄仄平仄_仄平平仄平	280	18.59%
6	（平）平仄仄仄平仄_（仄）仄平平平仄平	72	14.26%

① 袁慶述，〈從《唐詩三百首》看「拗救」問題〉，《中國文學研究》2，頁 41。

7	平仄仄平仄_平平平仄平	206	13.68%
8	（平）平平仄仄平仄_（仄）仄平平平仄平	51	10.10%
9	仄仄平仄仄_平平平仄平	51	3.39%
10	仄仄平仄仄_仄平平仄平	51	3.39%
11	仄仄仄仄仄_仄平平仄平	36	2.39%
12	仄仄仄仄仄_平平平仄平	27	1.79%
13	（平）平平仄仄仄仄_（仄）仄仄平平仄平	10	1.98%
14	（平）平仄仄平仄仄_（仄）仄平平平仄平	9	1.78%
15	（平）平仄仄平仄仄_（平）仄仄平平仄平	8	1.58%
16	平仄平仄仄_仄平平仄平	21	1.39%
17	平仄平仄仄_平平平仄平	19	1.26%
18	平仄仄仄仄_仄平平仄平	17	1.13%

以上是對《全唐詩》律詩中不合格律的句子的平仄組合類型及出現頻次的基本統計，結合單句與對句兩類失律句的分佈情況，我們就可以進一步審視清人關於「拗救」的種種說法。

2.清人「拗救」說及其局限

拗救之說，世所公認以清初王士禛和趙執信的影響最大，梁章鉅（1775—1849）說：「自漁洋、秋谷之書行，此說幾於家喻戶曉矣。」（《退庵隨筆》）[①]今天王力等人的說法，也大多不出二人的框架，因而首先需要總結王士禛與趙執信的說法，以此作爲衡量統計數據結果的基本參照，進而辨析其合理性與局限。

王士禛在《律詩定體》中提出兩種拗救。一種是對句相救，即五言出句「仄仄平平仄」的第三字若拗爲仄聲，對句「平平仄仄平」的第三字就換以平聲來救，形成「仄仄仄平仄，平平平仄平」的句式。第二種是單句拗救，即五言「仄平平仄仄」的第四字若拗爲平聲，則第三字換以仄聲救，形成「仄平仄平仄」的句式，後人習慣將此概括爲「四拗三救」。至於七言句，則可由五言類推：「所謂單句第六字拗用平，則第五字必用仄以救之，與五言三四一例。」此外，王士禛還指出「五律，凡雙句二四應平仄者，第一字必用平，斷不可雜以仄聲，以平平止有二字相連，不可令單也。」七言句則是第三字必平，因爲「凡平不可令單。」[②]王士禛的說法後來演變爲啟功的孤平說：「止有五言 B 句式首字不能更換，是因爲它如果換用仄聲，則下邊一字便成爲兩仄所夾的‘孤平’，聲調便不好聽。」[③]

從統計數據看，王士禛所說的對句相救確爲唐詩中較爲常見的現象，表 2 中的句式（3）與句式（7）均可視爲出句第三字拗，對句第三字救的詩例，比如「石路枉回駕，山家誰候門」（王維《酬虞部蘇員外過藍田別業不見留之作》，頁 1268）、「誰念北樓上，

① 梁章鉅，《退庵隨筆》，郭紹虞編，《清詩話續編》，頁 1863。
② 王士禛，《律詩定體》，丁福保輯，《清詩話》，頁 115、117。
③ 啟功，《漢語現象論叢》，頁 177—178。

臨風懷謝公」（李白《秋登宣城謝朓北樓》，頁 1845）。而句式（6）與句式（8）則是對應的七言句式，比如「寒林葉落鳥巢出，古渡風高漁艇稀」（杜牧《冬日五湖館水亭懷別》，頁 6069）、「身無拘束起長晚，路足交親行自遲」（劉禹錫《和留守令狐相公答白賓客》，頁 4072），應當說王士禛的這一論斷是可靠的。

但是王士禛的單句「四拗三救」理論卻不一定完全合理。他認爲「仄平平仄仄」第四字若「拗用平則第三字必用仄救之」，形成「仄平仄平仄」句式。但是反之「平仄仄平平」則不能拗救成「平仄平仄平」。其實無論「仄平仄平仄」還是「平仄平仄平」，在《全唐詩》中出現的都很少，「平仄平仄平」的僅 24 次，占 0.24%，即便是「仄平仄平仄」也僅 147 次，占 1.47%，可見單句四拗三救的說法不能推廣。至於其原因，王士禛倒是總結得很充分：「以平平止有二字相連，不可令單也」，無論是「平仄平仄平」還是「仄平仄平仄」，都形成了單平、單仄，聲律上都是不和諧的，某種程度上說，王士禛的理論也存在一些矛盾。

趙執信也講「四拗三救」但他限定了範圍，認爲只有「平平平仄仄」的句式可以拗救爲「平平仄平仄」，比如「行人碧溪渡」（杜牧《句溪夏日送盧霈秀才歸王屋山將欲赴舉》，頁 6010）、「芳心向春盡」（李商隱《落花》，頁 6215）。此外，趙執信還提出另一種單句拗救，即將「平平仄仄平」拗救成「仄平平仄平」，如「繭蠶初引絲」（杜牧《句溪夏日送盧霈秀才歸王屋山將欲赴舉》，頁 6010）。從表 1 來看，這兩種拗救的確是常見的拗救法，分別出現了 2542 次和 1243 次，而對應的拗而不救的「平平平平仄」和「仄平仄仄平」只出現了 43 次和 102 次，形成拗救之例遠高於不救之例。

至於對句拗救，趙執信基本延續了王士禛出句三拗、對句三救的說法。而由於他提出了「仄平平仄平」的單句救法，所以表 2 中句式（1）、句式（2）以及句式（4）、句式（5）產生的原因也可以理解了。因爲出句第三字拗，所以對句第三字必救，而只要對句第三字救爲「平」，則第一字即便拗爲「仄」也沒關係。此外，趙執信還擴展了王士禛的說法，對句第三字不僅可救出句第三字拗，亦可救第四字或三、四字同時拗，比如「暗暗春籍滿，輕輕花絮飛」（杜甫《宴胡侍禦書堂》，頁 2554）、「粉壁正蕩水，緗幃初卷燈」（李商隱《僧院牡丹》，頁 6280），對應的七言句則如「撫躬道地誠感激，在野無賢心自驚」（李商隱《贈田叟》，頁 6284）、「因吟郢岸百畝蕙，欲采商崖三秀枝」（陸龜蒙《奉和襲美抱疾杜門見寄次韻》，頁 7221）。標準放寬之後，趙執信的理論已完全涵蓋了表 2 的所有句式，在對句相救方面，他的歸納完全符合《全唐詩》中律詩的句式習慣。有論者以王士禛、趙執信之說缺乏宋元以來的理論傳承爲由，斷言「拗救」是謬說，這種見解恐怕是不甚科學的。

趙執信很好地解釋了怎麼救的問題，但是救不救的問題卻仍然無法得出一個完滿的解答。一個典型的矛盾是，趙執信總結說：「上句第三字平，下句第三字可仄。若上句第三字仄，下句第三字斷宜平。」這個總結對於前面的那些句式都是合理的，但卻可能有一個反例，即出句「平平平仄仄」的第三字若拗爲「仄」，那麼對句本應變成「（仄）仄

平平平」來救，但實際上這種句式極少出現，在《全唐詩》中一共只出現了 20 次。但是從表 1 可見，「平平仄仄仄」卻出現了 1382 次之多，換言之只有不到 2%的句子遵守了拗救規律，絕大多數則是拗而不救。爲此，趙執信只好給他的理論單獨加了一句補丁說：「平平仄仄仄，下句仄仄仄平平，律詩常用。」[①]

趙執信對於王士禛理論矛盾的處理是一面限制理論的適用範圍，一面又補充其他拗救模式，後人再修補趙執信的理論也是相同的邏輯。比如王力再講對句拗救，就不再提對句「三拗三救」的普遍規律，而把對句相救嚴格地限定在「仄仄平平仄」一種句式，即表 2 中的句式（9）至句式（18）幾種。而句式（1）至句式（8）則被王力認定爲可救可不救的「半拗」，這是在分類討論的邏輯基礎上，進一步限定了對句相救的範圍。至於本句自救，王力放棄了「四拗三救」說，只承認「仄平平仄平」一種單句救，同樣是把單句拗救的範圍縮小了。但是王士禛、趙執信所說的「平平平仄仄」拗爲「平平仄平仄」又是無法忽視的拗救現象，所以王力與趙執信一樣，也加了個補丁說「平平仄平仄」是「特定的一種平仄格式」，「在唐宋的律詩中是很常見的，它和常規的詩句一樣常見。」[②]而如此一來，各家討論拗救要打的補丁也就越來越多，拗救說漸趨複雜的原因正在於此。

相比之下，啟功的說法就比較圓滿了，既然無論怎麼歸納拗救的規律都會存在反例，那麼乾脆指出：「拗句在篇中，共拗幾句和拗哪一句或哪幾句，都沒有限制。」[③]拗救是一種詩律變體，既然是變體，那麼就「法無定法」。但是「變」中又有「常」，所以啟功也指出出句第三字、第四字拗，對句第三字救屬於常見的拗救，這樣就把趙執信解釋得清楚的句式全部納入，而解釋不清的句式，代之以「拗無定體」處理了。啟功以「常」與「變」的辯證邏輯理解拗救，以格律爲常，而以拗救爲變，但是在拗救中卻不乏某些常見規律，又是變中之常，這深刻體現了中國傳統詩學的核心觀念：「具體法則是固定的，是謂定法；法的原理是靈活的，是謂活法。相對具體法則而言，更重要的是法的原理。對法的原理的把握同時也就是對法的超越，所以說‘有定法而無定法’。」[④]

3.唐詩格律的再審視

如前所述，根據唐詩平仄組合總結歸納其變化類型比較容易，而一旦將其上升爲普遍規律就難了，學者們只好不斷縮小規則的適用範圍，並將其他常見拗句式補爲特例。此即古人所謂「法無定法」，或啟功所謂「拗無定體」。但是畢竟在單句與對句中也存在一些出現頻率較高的拗救現象，存在某些「變中之常」，因而一概稱之「拗無定體」，似乎也有可能遮蔽一些詩律的問題。因此，有必要確定一個大致合理的標準，在「拗救」

① 趙執信，《聲調譜》，丁福保輯，《清詩話》，頁 335。
② 王力，《詩詞格律》，頁 31—32。
③ 啟功，《漢語現象論叢》，頁 196。
④ 蔣寅，〈至法無法：中國詩學的技巧觀〉，《文藝研究》6，頁 69。

說的諸種規則中發現那些真正爲詩人們普遍採用的句式，從而探究唐詩平仄組合的規律性。

首先明確單句拗救成立的標準。依次取表 1 中各句式出現的頻次減去其存在對句相救的頻次，重新統計單句句式中「拗而不救」現象的出現頻次，結果如下：

表 3.《全唐詩》中拗而不救頻次表

序號	平仄組合類型	頻次	占比	序號	平仄組合類型	頻次	占比
1	（仄）仄平平平仄平	797	21.78%	12	（仄）仄平平仄仄仄	94	2.57%
2	平平仄平仄	2542	25.36%	13	仄仄平仄仄	96	0.96%
3	（仄）仄平平仄平仄	927	25.33%	14	（仄）仄仄平仄仄平	67	1.83%
4	平平平仄平	1255	12.52%	15	仄平仄平仄	147	1.47%
5	平平仄仄仄	1382	13.79%	16	仄仄平平平	116	1.16%
6	（平）仄仄平平仄平	242	6.61%	17	（平）平仄仄平仄仄	25	0.68%
7	仄平平仄平	533	5.32%	18	仄平仄仄仄	112	1.12%
8	仄仄仄平仄	333	3.32%	19	（仄）仄平平仄平平	39	1.07%
9	（平）平仄仄仄平仄	101	2.76%	20	仄平仄仄平	102	1.02%
10	平仄仄平仄	251	2.50%	21	（平）平仄仄平平平	37	1.01%
11	（仄）平平仄仄平仄	83	2.27%				

與表 1 相比，句式（1）、（2）、（3）、（5）的頻次未發生明顯變化，句式（4）中「拗而不救」的頻次仍比「拗而救」的頻次高一倍，此五種句式可以被認定爲單句拗救或特殊句式。句式（6）至句式（11）在減去對句相救的頻次之後，單句拗救的頻次都下降了一半以上，比如趙執信與王力所舉句式（6）、（7）「一拗三救」的例子，其作爲單句拗救出現的頻率僅爲 6.61%和 5.32%，而在對句相救的頻率則是 51.7%和 32.6%，是明顯的對句相救而非單句拗救。而這也證實了袁慶述所謂「‘孤平拗救’的出現並不是爲救孤平，而是構成這種句式，用來作對句以便與出句構成完美的平仄相對」[①]的判斷。至於句式（12）之後的句式，出現頻率已經比句式（5）以前者低了一個數量級，出現了較爲明顯的差異。因此，我們可以大致將第五句設爲判斷兩句之間是否形成拗救關係的分水嶺，在數據上反映爲頻率高於 10%。

在前五種句式中，句式（1）、（4）、（5）僅有一字平仄發生變化，且對句相救也較少出現，屬於「拗而不救」，當以「特殊句式」視之。句式（2）、（3）即王士禛的「四拗三救」，則可視爲單句拗救。趙執信最大的貢獻，就在於從眾多句式中發現了句式（5）的特殊性。據此分析拗救的原因，有學者提出唐詩的「拗救」實爲詩人因爲「仿古修辭」而形成的慣例，是對「永明體」格律的模仿[②]。但是永明體當避免「蜂腰」之病，即第二字、第五字不得同聲，句式（2）、（3）、（5）在趙執信、王力的「拗救」說中已是定例，它們確實符合永明體的聲律，但句式（1）、（4）這種前人未認定的特殊句式，卻犯了「蜂

① 袁慶述，〈從《唐詩三百首》看「拗救」問題〉，《中國文學研究》2，頁 42。
② 龔祖培，〈漢語詩歌「拗救」說辨偽〉，《文史哲》5，頁 82-85。

腰」之病，因而「仿古修辭」之說，似乎也不能完全涵蓋唐詩的拗救情況。

再看對句相救。很明顯以表 2 前八種句式爲分水嶺，句式（9）之後的出現頻次，較前八種句式已有明顯的下降。反應在數據上，則與單句的情況一致，均可以 10%爲界。而這八種句式都是出句第三字（七言第五字）由平拗爲仄，對句第三字（第五字）由仄救爲平，即王士禛的對句拗救。至於趙執信與王力提出的第四字拗和三、四字同時拗的例子，則占比不高，其原因或許可以從節奏與平仄的角度審視。

一般認爲，漢語詩歌以雙音節音步爲基本節奏單位，故多以第二、四、六字作爲關鍵節拍點，其平仄須保證相對，而第一、三、五字因爲不在關鍵節拍點上，故多爲可平可仄。比如松浦友久有「二四不同」、「二六對」等論[①]，俗語亦有「一三五不論，二四六分明」之說，此說雖不乏反例，但大致上符合律詩的基本格律特點。而對句相救的前八種高頻句式，無一例外地遵循這條規則，可見在對句相救的情況下，詩人更偏愛在一三五字上拗，使其更接近一般格律的用法。趙執信與王力以對句救出句的第四字拗，實已破壞了詩律的常態，出現率自然不高，而最終可以被認定爲對句相救的，則只有王士禛「（平）平平仄平」一種句式而已。

如果說突破格律的拗句是一種「常中之變」，則單句「平平平仄仄」句式的「四拗三救」和對句「仄仄平平仄，平平仄仄平」句式的「三拗三救」這兩種拗救現象，以及「平平平仄平」與「平平仄仄仄」兩種特殊句式則可以視爲「變中之常」。至於其他拗救現象，在《全唐詩》中並不十分普遍，不應求之過深。而反過來看，當我們把各種拗句與律句統觀時，進一步尋找唐詩平仄組合的「變中之常」，亦可以尋找某些唐詩格律的基本規則。

比如前人提出漢語詩歌的雙音節音步，音步間的平仄須保持平仄相間，從原理上說很有道理，但從數據上看似乎又不完全普適，畢竟《全唐詩》中還有 2542 句「平平仄平仄」和 927 句「（仄）仄平平仄平仄」。導致這個現象出現的原因，或許還在於五、七言詩句尾的三音節組合的節拍問題。松浦友久將其視爲兩個節拍（單音節的一拍視同襯了一個休音），所以二、四、六是關鍵節奏點。但我們同樣可以將結尾的三個音視爲一個節拍，特別是當詩句的意義節奏爲「二一二」時，比如「何由問香炷」（李商隱《哀箏》，頁 6246）、「月夜歌謠有漁父」（劉禹錫《自江陵沿流道中》，頁 4091），或者三個字語義粘合度較強時，比如「從茲匣中劍」（張九齡《眉州康司馬挽歌詞》，頁 596）、「二百年來霸王業」（皮日休《南陽》，頁 7117）。這種情況下就未必是二、四平仄相反，而有可能是二、五平仄相反更合適，永明體所避之「蜂腰」即是如此。李鎁《詩法簡易錄》謂：「大凡聲調之高下，必附氣以行，而平仄因之，以成節奏。故離平仄以言音節不得也，泥平仄以言音節亦不得也。」[②]當我們以音節論平仄時，似乎也不當泥於某些固定的字位。

[①] 松浦友久，《中國詩歌原理》，頁 209。
[②] 李鎁，《詩法簡易錄》，道光二年刊本。

在這個意義上，我們就可以跳出「拗救」說的理論框架來審視唐詩的格律。無論如何形成變例，其中可歸納的一條基本規則是各音步結尾的平仄須相對，此是「變中有常」。而因爲句尾三音節的靈活性，其節奏點不一定落在第四或第六字，也可能在第五或第七字，這樣就有可能形成某種「拗救」，其實也是一種「常中有變」。

另一條普遍規律是，除了句尾的三音節外，唐詩句中一般避免出現「仄平仄」的「孤平」現象，即王士禛所雲「凡平不可令單」。從表 1 看，除了句式（9）可能在開頭出現「仄平仄」以外，其餘高頻句式均避免了兩仄夾一平的平仄組合，而句式（9）之所以可能形成「仄平仄」又是因爲七言詩的首字平仄不論的結果。直到句式（13）、（14）、（15）才出現「仄平仄」現象，但其頻次已經很小了，這也側面說明了爲什麼趙執信提出的出句第四字拗、對句第三字救並不常見的原理，因爲這種拗救有違唐詩的基本格律規則，即便可以「救」，也往往爲詩人所避免。而在此兩條規律下，真正突破律句的「常中之變」，則很明顯集中於句尾三音節的平仄變化，即形成「平仄平」與「仄平仄」、「仄仄仄」的結尾，這或許可以視爲辨別唐詩拗救的標誌之處。

結論

綜上所述，若以「常中之變」的視角辨析拗救，則只有單句「平平仄平仄」和對句「仄仄仄平仄，平平平仄平」兩種拗救，以及「平平平仄平」與「平平仄仄仄」兩種特殊句式具備一定普適性，其餘拗救皆不常見，不宜求之過深。如此論「拗救」更接近唐詩的基本面貌，也可大大簡化「拗救」說的理論框架。

而若以「變中之常」的視角審視唐詩中的律句與拗句，則唐詩平仄的基本原則可以概括爲句中各音步單位的尾音節奏點（並非機械地以二、四、六字爲准）平仄須相對，同時儘量避免在前四字（七言前六字）中出現「仄平仄」的孤平組合，這又是唐詩寫作萬變不離其宗之處。

當然，本文使用計算機標注與統計的方法，研究清人的唐詩「拗救」說，只是初步形成對唐詩格律的整體性認識。文學現象是一個不斷發展的「生命體」，律詩在唐代甚至唐以後的發展還具有一定過程性，數據分析的方法目前還難以深入挖掘唐詩在不同發展時期的階段性特徵，更難以從文學創作心態上揭示唐人對拗救的自覺意識。因此，本研究採用計算機與統計的方法辨析清代的唐詩「拗救」理論，仍然只是一種方法論的探索，其結論亦不是絕對的。而對古典詩歌形式美學的探討，也依賴于更多文本分析、美學理論以及統計測量的深度結合，如何在學融合基礎上形成中國特色的詩歌形式美學理論，則有待於進一步的探索與討論。

參考書目：

永瑢等（1965）。《四庫全書總目》。北京：中華書局。

王力（2000）。《詩詞格律》。北京：中華書局。

趙克剛（1998）。〈關於犯孤平〉，《重慶師院學報哲社版》4，頁 46-53。

袁慶述（2007）。〈從〈唐詩三百首〉看「拗救」問題〉，《中國文學研究》2，頁 41-43。

鍾如雄，胡娟（2014）。〈論唐詩「孤平拗救」說之不成立〉，《雲南師範大學學報》6，頁 77-82。

龔祖培（2015）。〈漢語詩歌「拗救」說辨偽〉，《文史哲》5，頁 77-96。

霍松林（2003）。〈簡論近體詩格律的正與變〉，《文學遺產》1，頁 104-117。

尚永亮（2013）。〈唐人作詩是否拗救？這是一個問題〉，《中國韻文學刊》4，頁 115-117。

彭定求等（1999）。《全唐詩（增訂本）》。北京：中華書局。

梁章鉅。《退庵隨筆》。郭紹虞編（2016）。《清詩話續編》。上海：上海古籍出版社。

王士禛。《律詩定體》。丁福保輯（2015）。《清詩話》。上海：上海古籍出版社。

啟功（1997）。《漢語現象論叢》。北京：中華書局。

趙執信。《聲調譜》。丁福保輯（2015）。《清詩話》。上海：上海古籍出版社。

蔣寅（2000）。〈至法無法：中國詩學的技巧觀〉，《文藝研究》6，頁 68-74。

松浦友久（1990）。《中國詩歌原理》。沈陽：遼寧教育出版社。

李鎁（1822）。《詩法簡易錄》。道光二年刊本。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

現代詩的量化研究

發掘顧城詩的隱藏節奏

廖學盈

法國波爾多蒙田大學 TELEM 研究中心研究員

現代詩的量化研究: 發掘顧城詩的隱藏節奏

廖學盈

研究員

法國波爾多蒙田大學 TELEM 研究中心

E-mail : shuehyingliao@gmail.com

摘要

本研究運用組合分析探勘詩人顧城的語法，全面調查《顧城詩全編》作品中「字型」、「音節」、「輔音」、「元音」和「聲調」在每一行詩句中的高頻使用模式，最後對作品的音樂性和節奏感進行描述和分類。

目次

引言

1. 詩人創造的語法體系
2. 語法體系的探勘方法
3. 語法體系的探勘成果
4. 結論

關鍵詞: 現代詩、量化分析、節奏分析、詩人語法、顧城

A Quantitative Research of the Contemporary Chinese Poetry: the Discovery of the Underlying Rhythm in GU Cheng's Poems

Dr. Shueh-Ying LIAO
Research fellow
University of Bordeaux TELEM Research Team (France)

Abstract

The present work explores GU Cheng's creative syntax by enumeration of combinatorial structures of sinograms, phonetics, consonants, vowels, and tones. On the basis of colocation frequency, we extract underlying musical and rhythmic patterns from every lines. This allows to categorize the musicality and the rhythmic feeling in reading GU Cheng's poems.

Keywords: contemporary Chinese poetry, quantitative analysis, rhythmic analysis, poetic syntax, GU Cheng

引言

漢語的精神世界不只在文字所指之處，也不局限於形象所表之列，它是文字與事物肌理相連的整體。字可指物，字本身亦是物，內容與形式並無主次之分。在這樣的客觀條件下，漢詩抒情境界的維度，單從語義的線索羅織，容易理解得有情無境。然而，非語義的氛圍和氣質，卻又不容易感知，即便有所觸動，也很難形容。這些微弱的動態信號，在古典詩歌中，尚能以音韻對位的格律進行機械論式的描述；而在現代詩歌中，則成為難以定義的節奏感或音樂性。

事實上，任何微小的悸動，都可以對其外部源頭，進行機械論式的探索：即便答案或無。更何況，如前面所述，漢詩的內容與形式，要共同構成抒情的維度，不但需要機械式的銜接，還需具備機體式的聯動。換言之，它的形式與它的內容可以等值。

漢語現代詩能不能登大雅之堂，榮冠詩的尊稱，許多古典學者仍持保留態度：白話閑雜淡無味，無韻更是不成詩。矛盾的是，在漢語詩歌發展的歷史中，押韻的古雅文言，也不就因此成詩。顯然，詩人著眼之處，不限文言抑或白話，亦不單就音韻對仗與否。詩乃是詩人識力的結構和推展識力的節奏，它以語文的形式留下光影和餘音。

1. 詩人創造的語法體系

在靜謐的文本世界，語法深刻體現詩人對事物秩序的認識。句長句短的變化、實詞虛字的遷移、空隔換行的織紋，創作和閱讀的進程中，人們隨之逸出常規的界域。遠在詩經的時代，人們就已經創造了「直寫實物」和「營造虛景」之外的第三類章法：「實與實並列產生的虛」[1]。文字既不敘事也不抒情，實物景象只是引發內心故事的線索，最後呈現的將是主詞謂語在時間終點消逝後的實相世界 [2]。然而，為了解讀實與實並列產生的詩象背後透露的

[1] 翁文嫻（2007）：＜《詩經》「興」義與現代詩「對應」 美學的線索追探 ——以夏宇詩語言為例探研＞，中國文史哲研究集刊，31：121-148，頁 130。

[2] 翁文嫻（2004）：＜顧城詩「呈現」界域的存在深度——「賦」體美學探討系列之一＞，當代詩學年刊， 1 ： 181-201，頁 195-196。

消息，要不緣著注釋旁敲側擊，要不涵泳音韻沈潛聆聽。然而，實物隨興並列產生的意義斷裂，有時還是難以重新連接。[3]

現代詩也有與此相同的類型。描述實物的詩易懂，講的是一種情懷，藉由典故或者經驗得以進入。營造虛景的詩乍看之下不明白所以，還能憑藉分析性的語法輔助推論。至於「並列實物」以達到本然實相的書寫方式，則觸碰到了漢語精神世界相通的一些經典原則。當代詩學家翁文嫻就曾經使用裴普賢提出的詩經「興-應」模式來講解現代詩中思想跳躍的段落[4]。所謂的「興-應」模式，從句式的層次來看，像是一種「題-釋」的結構[5]。兩兩並列的事物，有時因果相生，有時虛實互補，有時矛盾對立。然而，在文字層面上，這些關係都沒有明確定義。

顧城後期的詩作即有這種現象。顧城作品的分類和編年，有四個時期，前三期「有我」，後三期「無我」[6]。根據當代詩家翁文嫻教授的分析，「無我」時期的作品無論從句式、章法和思維來說，都有一種返回詩經原型的趨勢[7]。她很早就注意到現代詩人語法創造的問題，且將這種語言實驗，放置在「東西方詩人語法比較發展史」的高度上[8]。她曾以現代符號訊息與結構主義的精讀法釋出興體詩中「綢繆」一首「綢繆束薪，三五在天」與下文「良人」的關係，並主張以「興應」作為批評語彙，推測詩歌的景物與心意之間的關聯[9]。較之古典文學界多用傳統語彙，現代文學界直引西方理論，此實則貫通古今漢詩的原生方法論。「詩人語法」是一種強調同語系內也會產生異質語法的創作理論，然而，要基於「興-應」結構解讀詩作，至少要先能掌握章內每一句詩的意思。換言之，「興-應」作為文學批評的詞彙，可更進一步探入「句內」用詞模式的層面。針對這點，單就詩經興體詩的釋讀，為了定義「興句」和「應句」之間的關聯，筆者曾進一步提出「高頻用詞模式」來辨別詩句內隱藏的「節奏模式」，進而以「節奏模式」的「重複」、「對稱」和「互補」來連通語義

[3] 廖學盈(2016)：＜詩經的量化研究：發掘興體詩的隱藏節奏＞，2016 數位人文國際研討會論文集，頁 403。
[4] 翁文嫻（2007），頁 135-136。
[5] 高友工（2004）：《中國美典與文學研究論集》。臺北，臺灣大學出版。頁 179-180。
[6] 顧工(1995)：《顧城詩全編》。上海，三聯書局。頁 1-6。
[7] 翁文嫻（2004），頁 192。
[8] 翁文嫻（2007），頁 129，註 23。
[9] 同上註，頁 128。

看似斷裂、節奏卻緊密匹配的上下文[10]。這種節奏模式不能像文字指稱事物，但是，它提供了詩人情緒擾動的信號，提示音響和氣氛在文章中的變化[11]。

取得「高頻用詞模式」的基礎是組合分析和頻率計算，初始的結果是靜態的句式結構，表面上不具備節奏所需的時間條件。其實不然，頻率效應會改變語言單位書寫/閱讀的時值，一系列重複循環的時值差能產生節奏感。筆者所定義的頻率效應如下：「詞組頻率的高低與[書寫/閱讀]時[書寫工具/眼睛凝視]落在詞組上的時間是負相關的，即詞組頻率越高，上述活動停留在這個詞組上面的時間越短，反之，停留時間越長。」也就是說，作為組合軸元件的高頻詞組，如同事先設計的閘道，調校組合軸上嵌入低頻詞的時值間隔，形成所謂的節奏[12]。

因此，語法體系並非只是選字連文的靜態結構，還具備調校時值的動態功能。漢語又因為具備聲調，語法體系不只傳遞詩人的節奏感，還能描述詩作的音樂性。

2. 語法體系的探勘方法

本研究採用「顧城詩全編」九百一十七首詩的內容進行語法體系的探勘。每首詩都將分別以四個漢字音節的長度，自第一個漢字音節起，一個個順序皆向後取四個漢字音節切成組合分析的片段。每個切片都將個別以字型、發音、輔音、元音、聲調等資訊進行文本標記、組合分析和頻率計算。根據計算結果，我們嘗試比較前期和後期作品這五個元件在所有切片中的組合情況。以下舉「頌歌世界」組詩「離」一篇的句子「灰暗的兔子眼神如火」為例，首先文本先進行標記：

[10] 廖學盈(2016)，頁 405-406。
[11] 廖學盈(2016)，頁 412。
[12] 在神經語言學的領域，閱讀的部份，目前已知且被反覆驗證的現象有：a. 內容詞(content word)比起功能詞(function word)較常被凝視。b. 詞的字符越多，被凝視的時間越長。c. 上文鋪陳越長，下文被凝視的時間越短。文本的節奏感理應可以透過作家的筆觸和讀者的閱讀得以實現。筆者的研究假定：熟練的作家是以浮動的筆觸書寫，成熟的讀者是以跳視的方式進行閱讀。以母語[書寫/閱讀/說話/聆聽]的時候，[工具/眼睛/口舌/耳朵]周遊在某個詞的時間，可以定義為[勾畫/凝視/發音/理解]某個詞的落點，以及在落點上停留時間的長短。詞頻較高的詞組元件，[勾畫/凝視/發音/理解]的時間越短，反之， 時間越長。以上是筆者與陽明大學神經科學研究所林依禎 2015 年在法國 CEA-Neurospin 研究院設計模擬漢字閱讀核磁共振實驗時的討論內容。參考：RAYNER, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372–422.

字型	灰	暗	的	兔	子	眼	神	如	火
音節	hui	an	de	tu	zi	yan	shen	ru	huo
輔音	h	*	d	t	z	y	sh	r	h
元音	ui	an	e	u	i	an	en	u	uo
聲調	1	4	0	4	0	3	2	2	3

接著進行切片得到['灰暗的兔', '暗的兔子', '的兔子眼', '兔子眼神', '子眼神如', '眼神如火']。這樣的切片能以每四個字符為單位，調查每個字符與鄰近字符各類元素共現的情況，例如：字型、音節、元音、輔音、聲調的組合變化[13]。

接下來，我們將每個切片都進行組合分析，以字符「〇」標記所有可以作為選擇軸的位置，所有可以作為組合軸的位置則保留原來的字符，看看有哪些可能的情況。例如，第一切片「灰暗的兔」，從字型組合分析得到：

['〇〇〇〇', '〇〇〇兔', '〇〇的〇', '〇〇的兔', '〇暗〇〇', '〇暗〇兔', '〇暗的〇', '〇暗的兔', '灰〇〇〇', '灰〇〇兔', '灰〇的〇', '灰〇的兔', '灰暗〇〇', '灰暗〇兔', '灰暗的〇', '灰暗的兔']；

從音節組合分析得到：

['〇 〇 〇 〇', '〇 〇 〇 tu', '〇 〇 de 〇', '〇 〇 de tu', '〇 an 〇 〇', '〇 an 〇 tu', '〇 an de 〇', '〇 an de tu', 'hui 〇 〇 〇', 'hui 〇 〇 tu', 'hui 〇 de 〇', 'hui 〇 de tu', 'hui an 〇 〇', 'hui an 〇 tu', 'hui an de 〇', 'hui an de tu']；

從輔音組合分析得到：

['〇 〇 〇 〇', '〇 〇 〇 t', '〇 〇 d 〇', '〇 〇 d t', '〇 * 〇 〇', '〇 * 〇 t', '〇 * d 〇', '〇 * d t', 'h 〇 〇 〇', 'h 〇 〇 t', 'h 〇 d 〇', 'h 〇 d t', 'h * 〇 〇', 'h * 〇 t', 'h * d 〇', 'h * d t']；

從元音組合分析得到：

['〇 〇 〇 〇', '〇 〇 〇 u', '〇 〇 e 〇', '〇 〇 e u', '〇 an 〇 〇', '〇 an 〇 u', '〇 an e 〇', '〇 an e u', 'ui 〇 〇 〇', 'ui 〇 〇 u', 'ui 〇 e 〇', 'ui 〇 e u', 'ui an 〇 〇', 'ui an 〇 u', 'ui an e 〇', 'ui an e u']；

從聲調組合分析得到：

[13] 本實驗使用漢語拼音進行標記。

['〇〇〇〇', '〇〇〇4', '〇〇0〇', '〇〇04', '〇4〇〇', '〇4〇4', '〇40〇', '〇404', '1〇〇〇', '1〇〇4', '1〇0〇', '1〇04', '14〇〇', '14〇4', '140〇', '1404']。

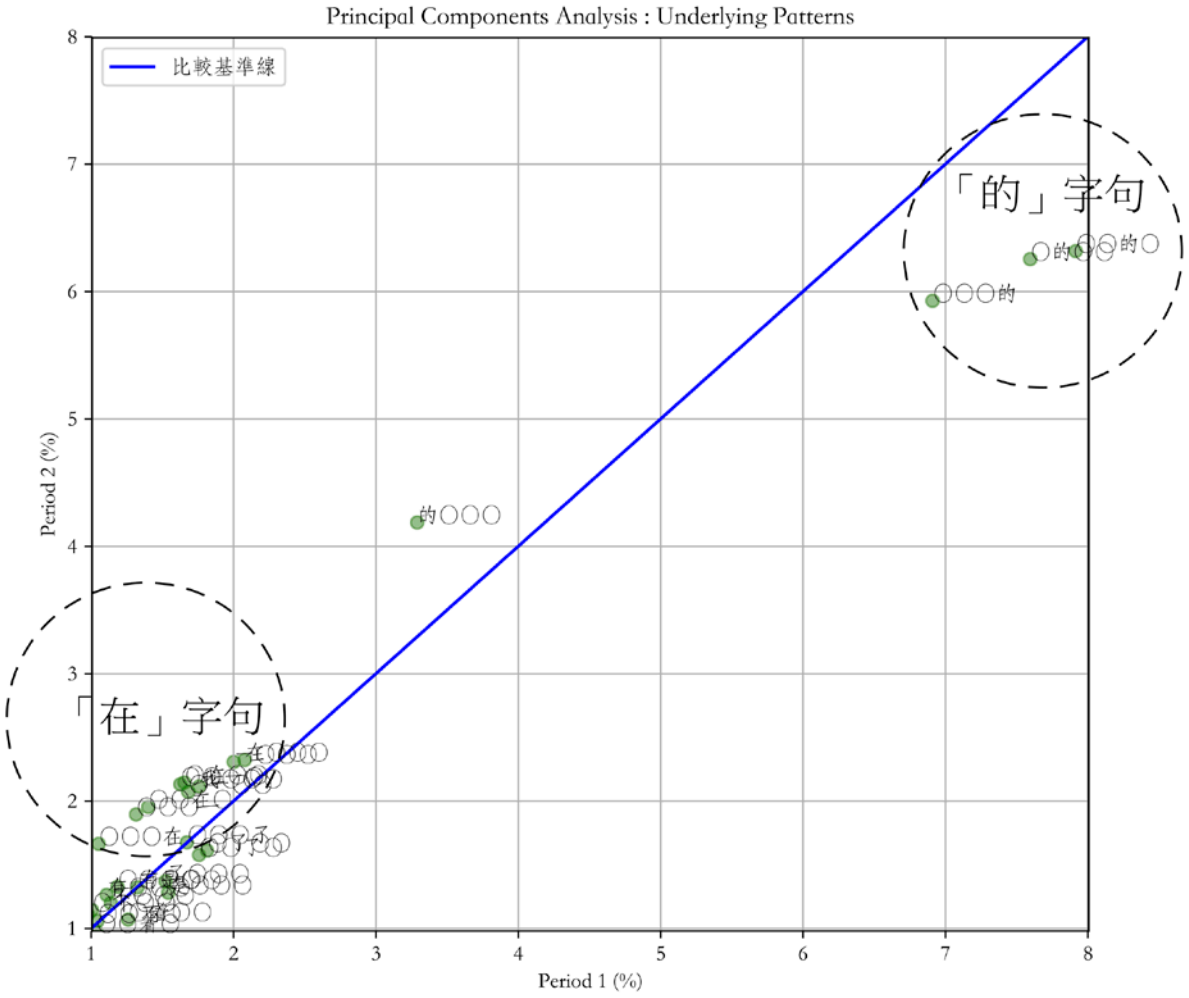
最後，進入文本反查這些組合出現的次數和位置，即可得出各種字型、音節、輔音、元音、聲調在顧城詩中組合的頻率和分佈的情況。

3. 語法體系的探勘成果

3.1. 字型組合

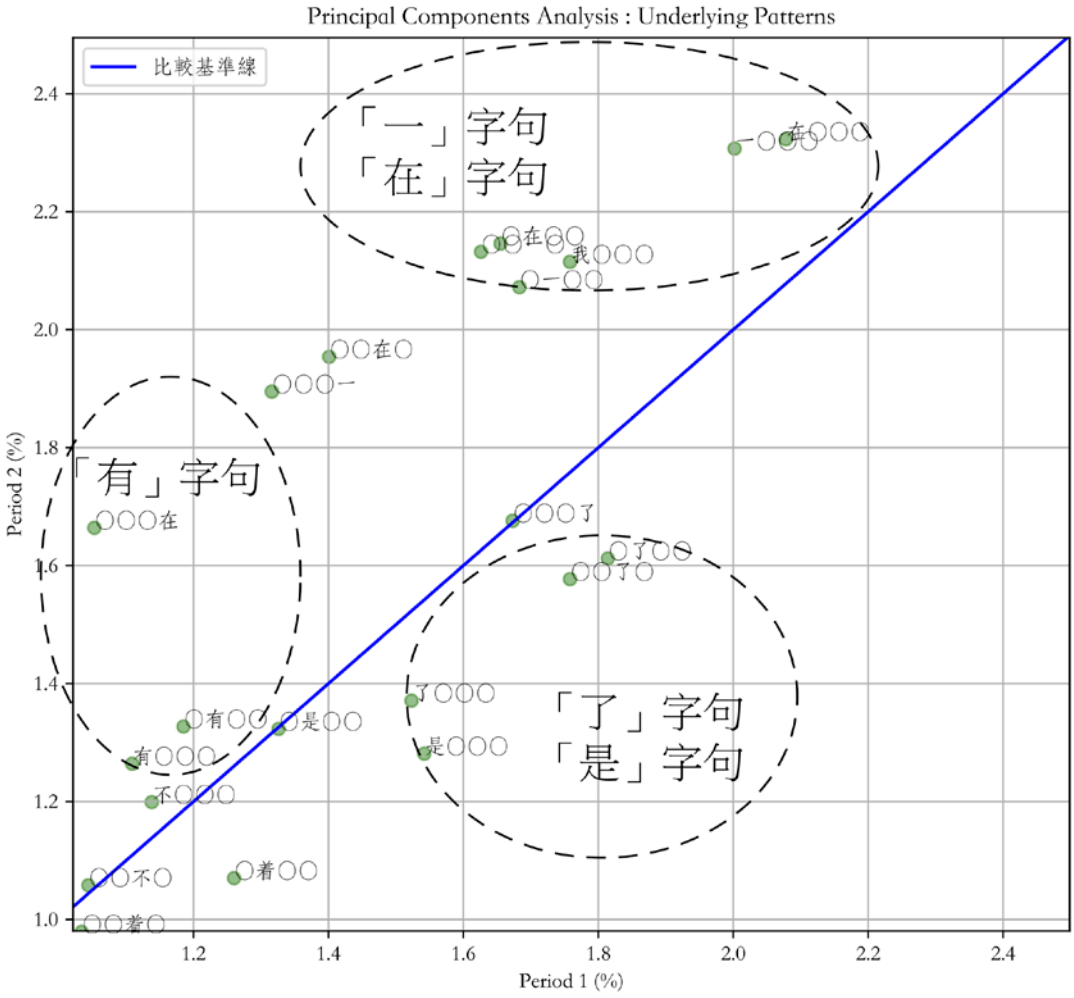
顧城前期作品（Period 1）較後期作品（Period 2）含有更多比率的「的」字句；後期作品則開始朝「在」字句發展（*圖表 3.1.1.*）：

圖表 3.1.1. 顧城詩中的「在」字句和「的」字句



圖表左下角的區域，還可以發現，表示動作的「了」字句和表示判斷的「是」字句較為突出，而表示齊一的「一」字句、表示位置的「在」字句和實存的「有」字句後期開始增加（*圖表 3.1.2.*）：

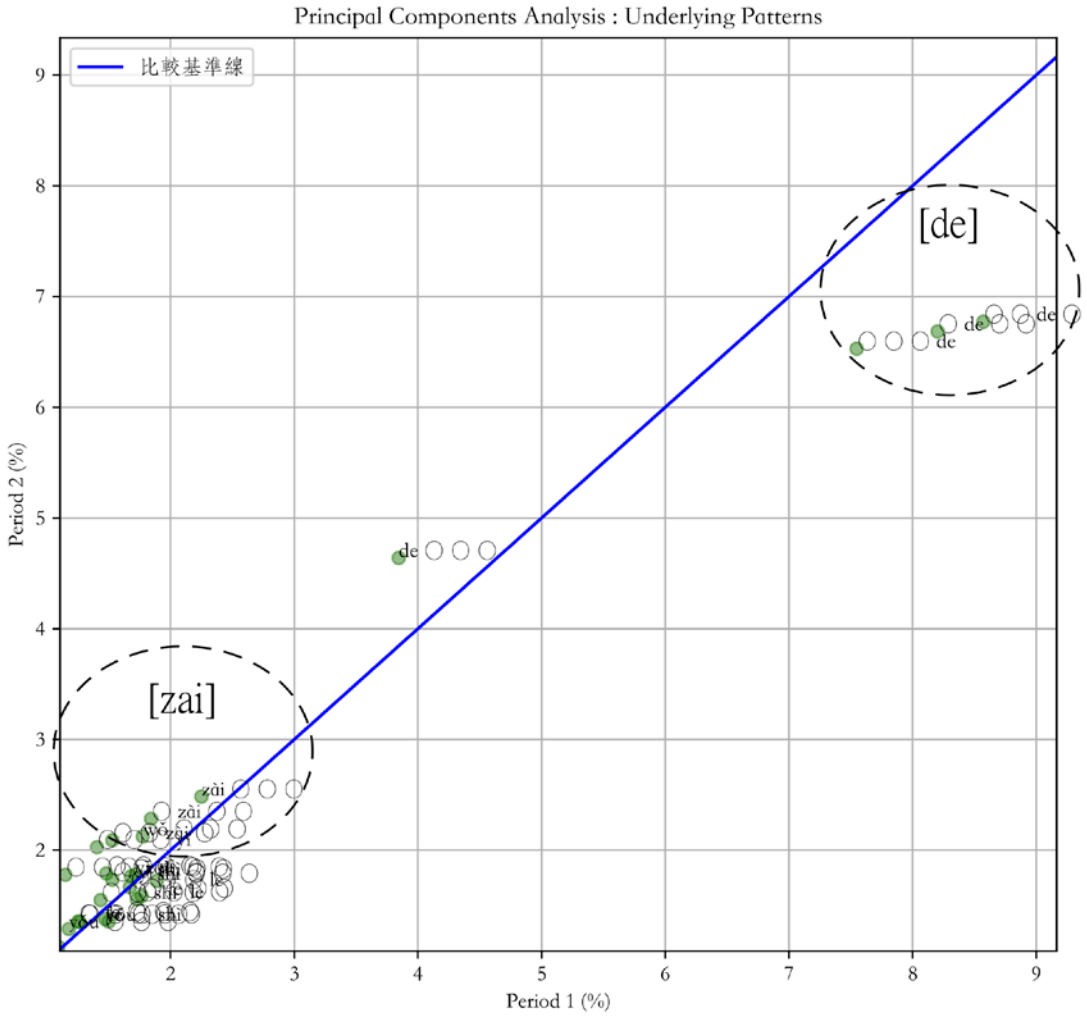
圖表 3.1.2. 顧城詩中的「一」/「在」/「有」字句和「是」/「了」字句



3.2.音節組合

[de]音節和[zai]音節的調查結果幾乎與「的」字句和「在」字句重疊，但表達的意義不同。音節組合表達的是一種口語遣詞的慣性，字型組合表達的是一種書寫選字的偏好。兩種情況應該分別計算（*圖表 3.2.1.*）：

圖表 3.2.1. 顧城詩中的音節：[de]和[zai]

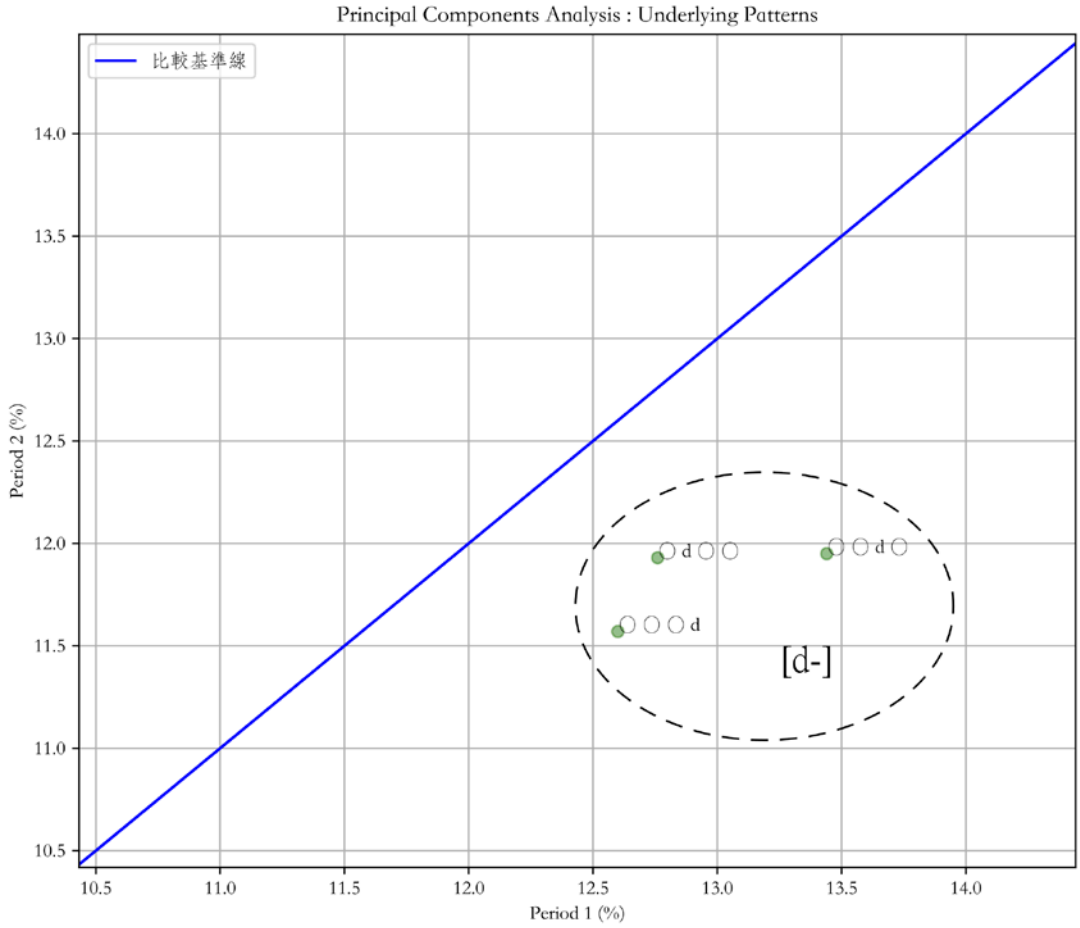


同音字的情況下，口語和書寫的差異會對比出來，因為，同樣的音節不會因為書寫字型的差異而另外計算使用頻率。這種情況在單音節詞彙繁多的古典詩歌中會比較明顯。由於「顧城詩全編」不像「顧城詩全集」納入顧城大量的古詩創作，現階段的音節探勘不能良好反應這種差異。筆者尚未數位化「顧城詩全集」的內容，有待日後的探勘補足這一方面的缺漏。

3.3. 輔音組合

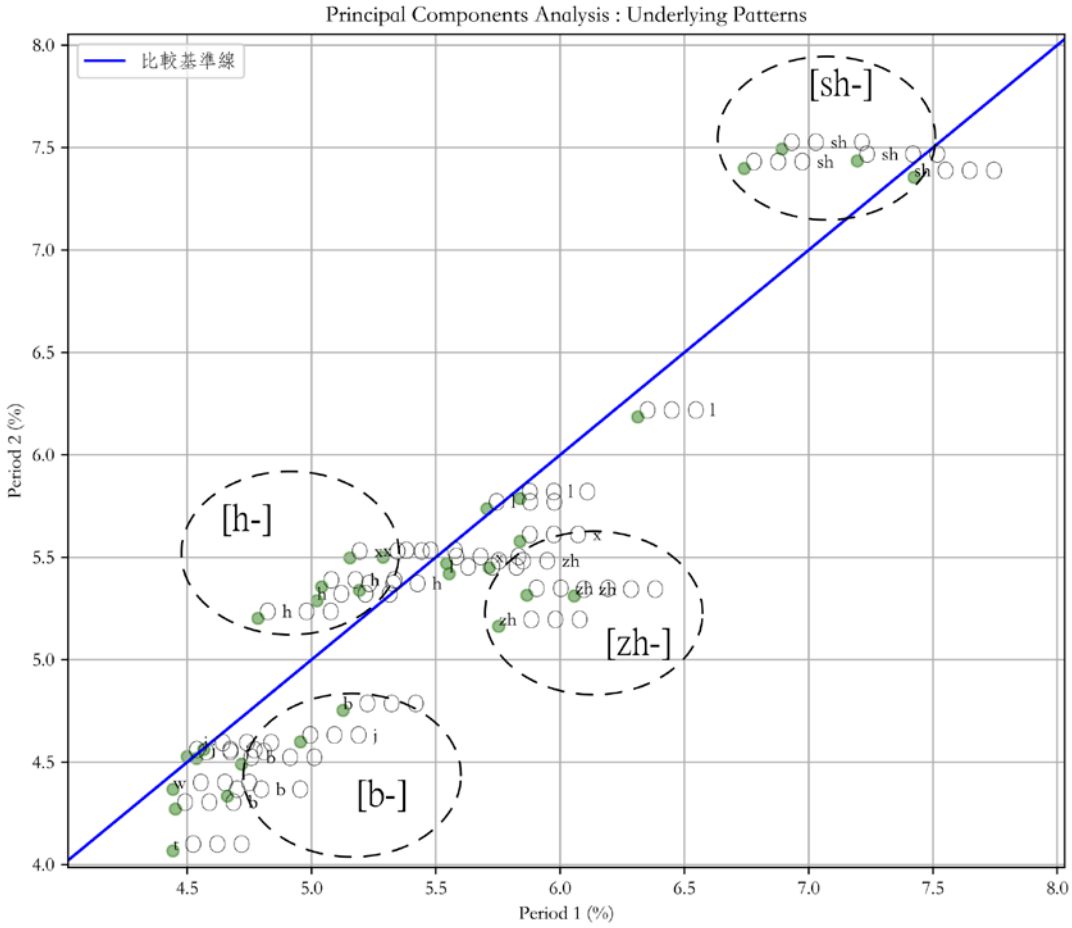
由於「的」字句「de」音節的大量使用，使得顧城前期作品咬字時必然會出現較大量的「d-」輔音，齒齦音（或作舌尖中音）非常鮮明，後期作品中沒有與之對應的咬字模式（*圖表 3.3.1.*）：

圖表 3.3.1. 顧城詩中的輔音：[d-]



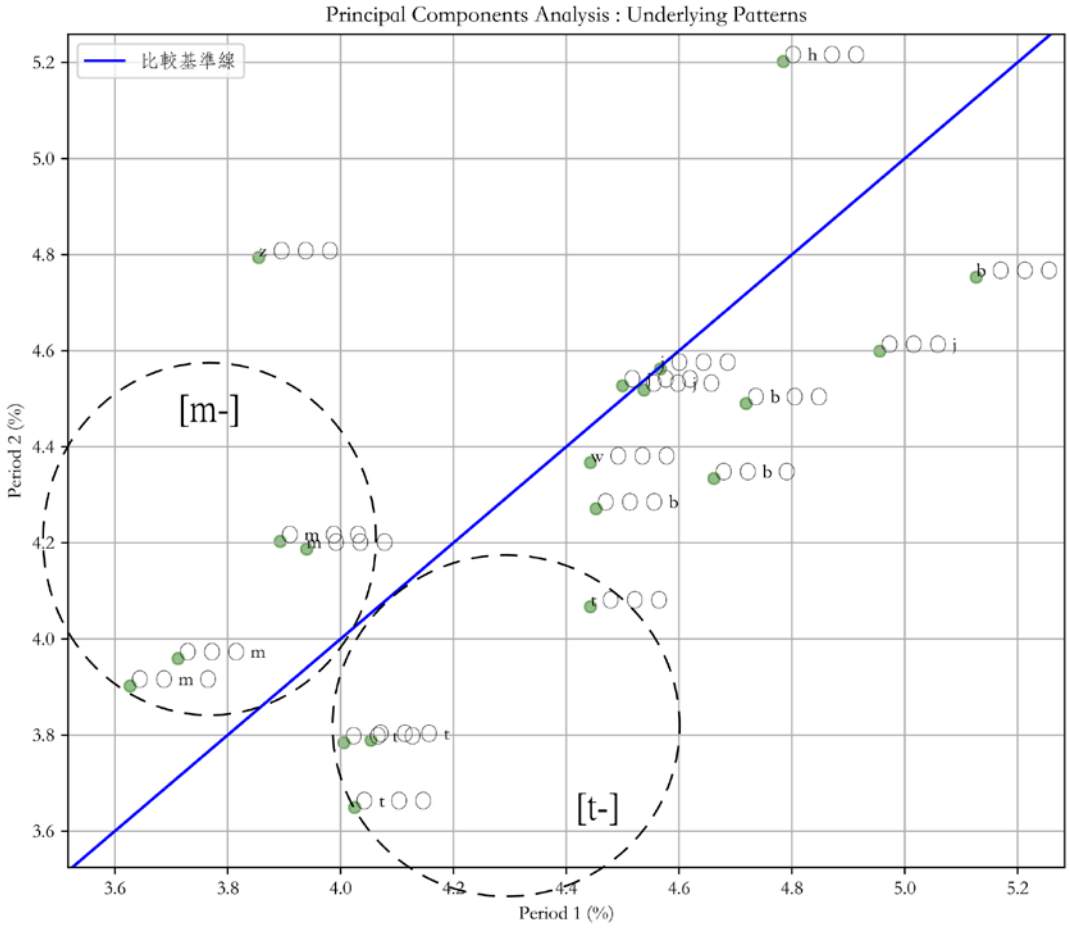
相反地，顧城後期的詩歌出現了大量的[sh-]輔音，捲舌音以擦音的方式輕易滑出口腔，不像前期的塞擦音[zh-]較為吞吞吐吐。後期作品的舌面後音[h-]，亦屬擦音，比起前期作品雙唇音[b-]的阻塞感，咬字更為自在（*圖表 3.3.2.*）：

圖表 3.3.2. 顧城詩中的輔音：[b-]、[zh-]、[sh-]和[h-]



顧城前期還有一個慣用的塞音，就是位置略深的舌尖中音[t-]，與之對比的是後期發聲方式較為沉著、位置前緣的雙唇鼻音[m-]（*圖表 3.3.3.*）：

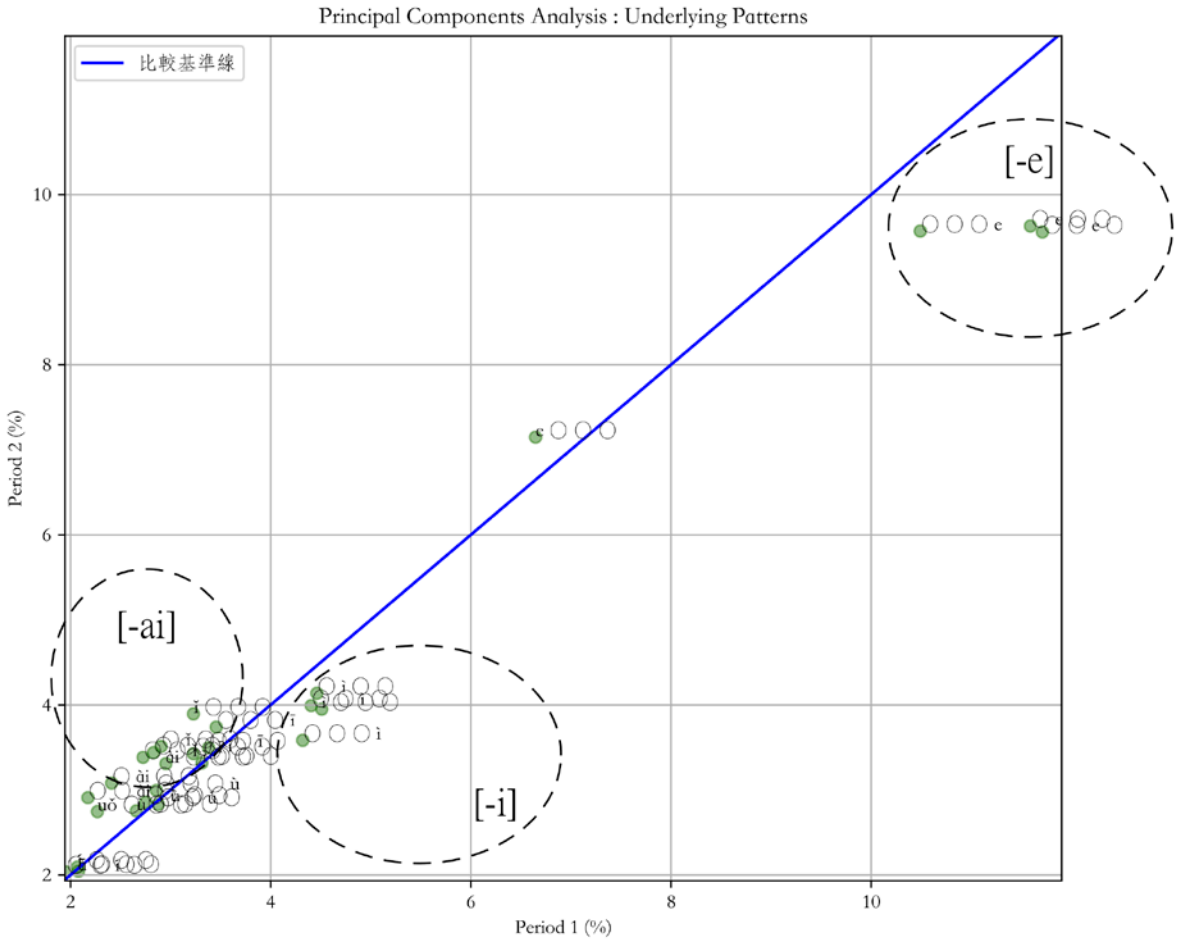
圖表 3.3.3. 顧城詩中的輔音：[t-]和[m-]



3.4. 元音組合

顧城詩中的元音使用似乎比較單純。前期有大量舌頭位置較前，開合對比不明顯的閉口元音[-i]和處於中段的半開口元音[e-]，後期開始出現較多舌頭發音位置由後往前，開合對比較明顯的的的雙元音[-ai]（*圖表 3.4.1.*）：

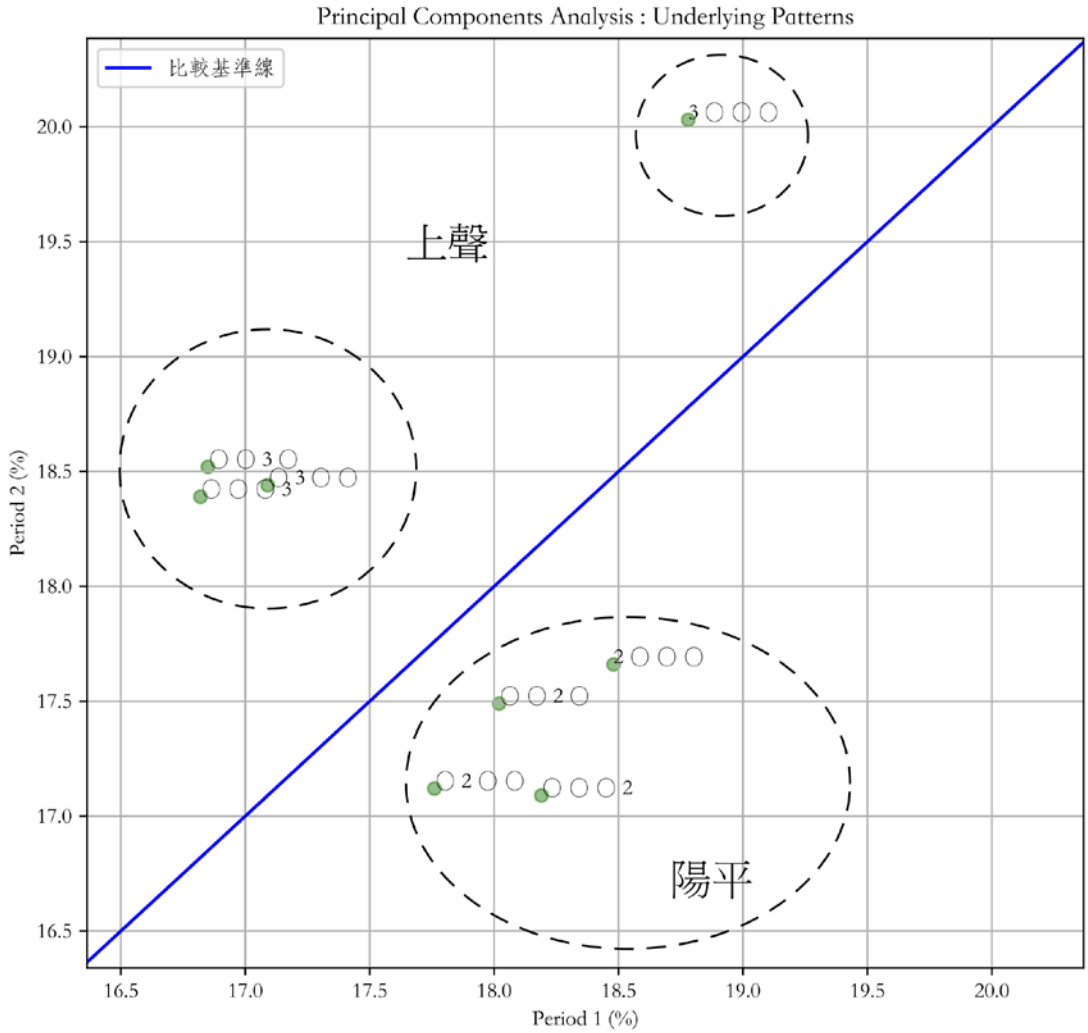
圖表 3.4.1. 顧城詩中的元音：[i-]、[e-]和[-ai]



3.5. 聲調組合

顧城詩歌前後期聲調的對比較明顯。前期以逐步高昇的陽平聲為特色，後期以先抑後揚的上聲為主調（*圖表 3.5.1.*）：

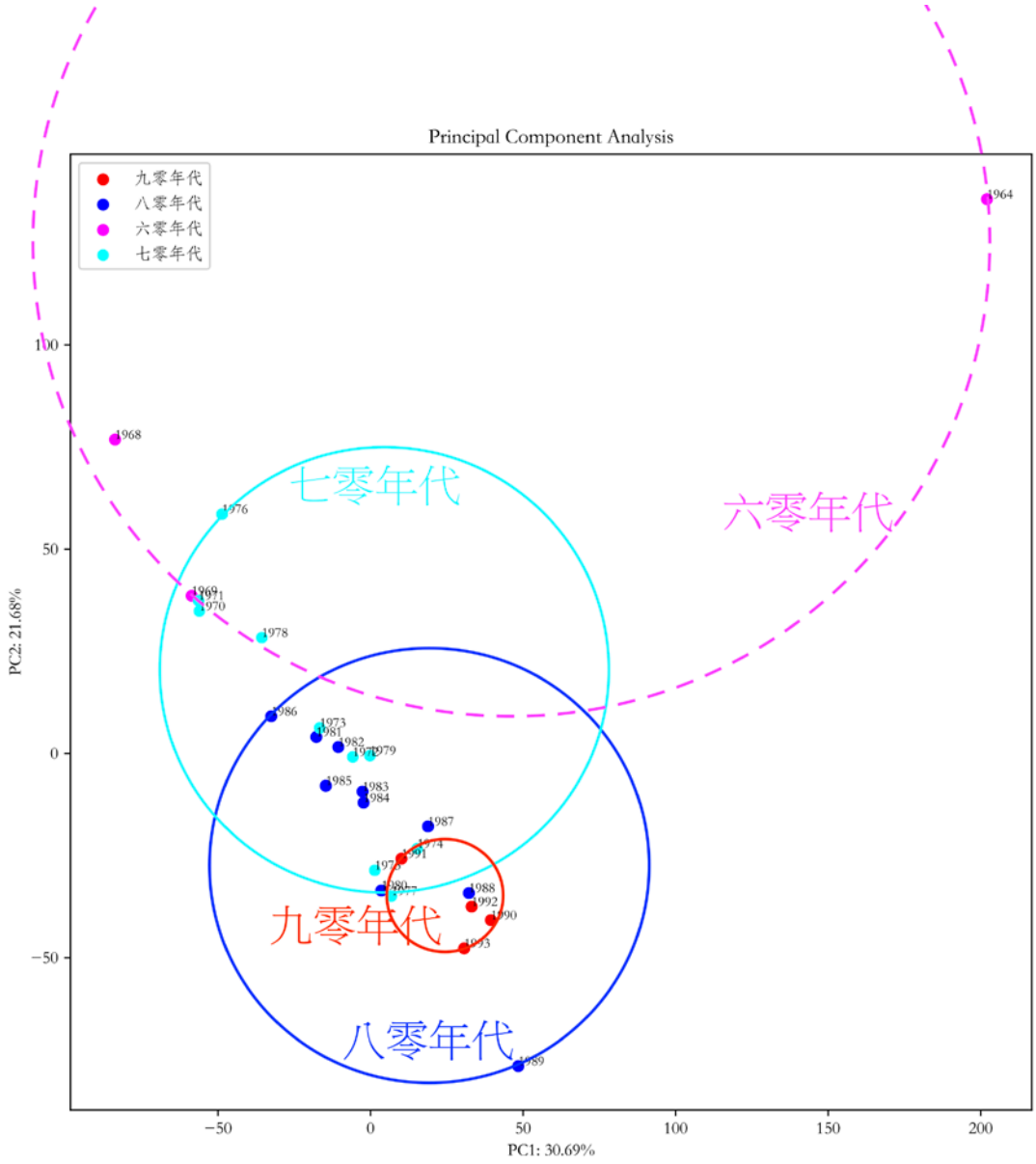
圖表 3.5.1. 顧城詩中的聲調：陽聲調和上聲調



3.6. 組成要素分析

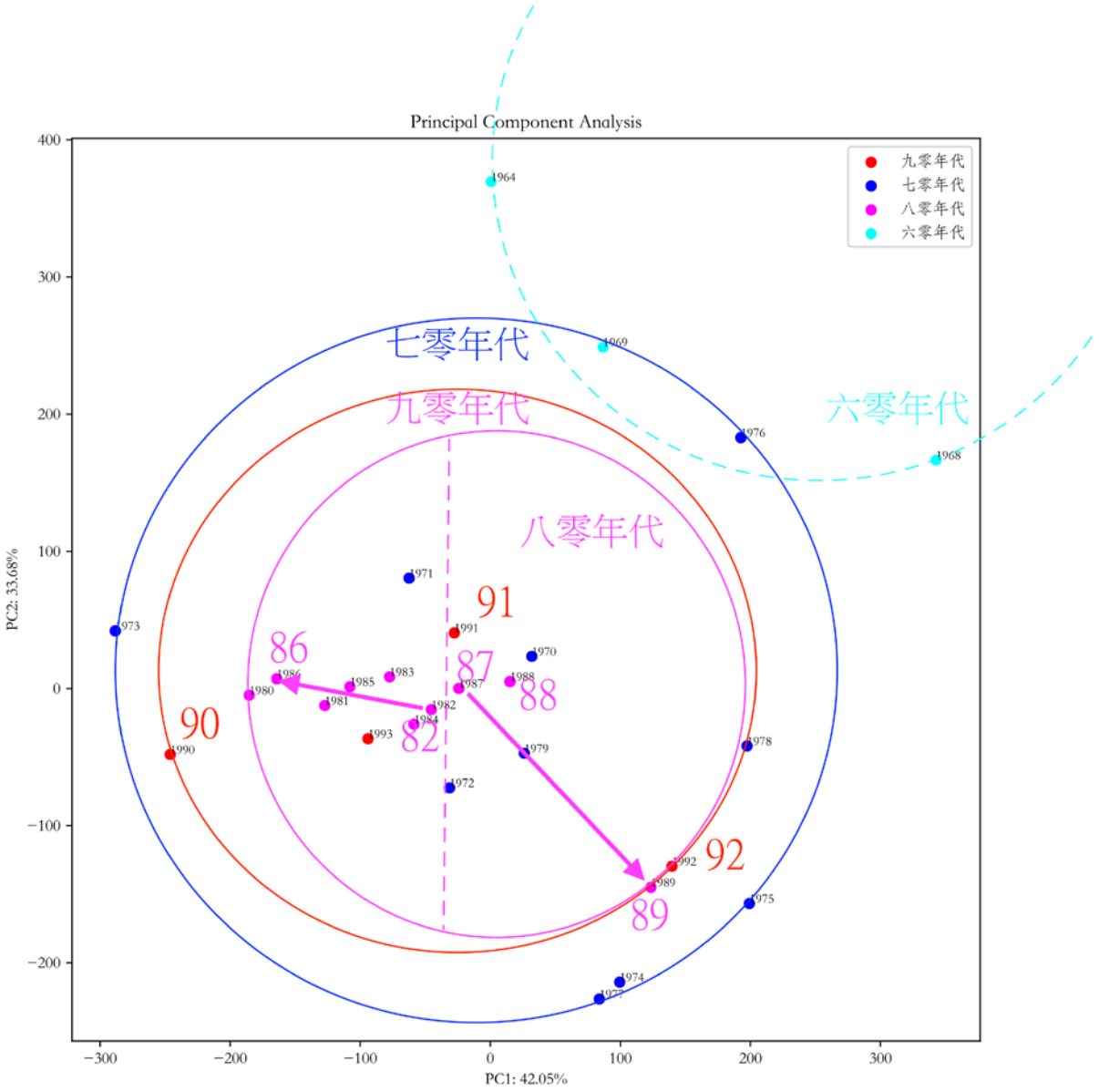
將顧城詩作所有共通的詞彙定義為特徵，進行組成要素分析，並以十年為單位進行分期。*圖表 3.6.1.*中，一個點代表一個年度的作品，同時期的作品著上相同顏色，最後以圓圈定為群組。我們發現，愈後期的作品，彼此之間的親緣關係愈近，風格似乎逐漸統一（*圖表 3.6.1.*）：

圖表 3.6.1. 顧城詩組成要素分析（根據共通的詞彙）



但是，這不是學者最關心的問題，因為，上述這種趨勢對於詩家來說幾乎是一種自然現象。比較值得注意的是，顧城詩通常被認為帶有「齊物論」的色彩；根據顧城自己的分類，八六年以前的詩作屬於「有我」時期，八六年以後則屬於「無我」時期。如果我們以相反詞作為特徵，並特別考慮「我」字的有無，進行「組成要素分析」，結果令人眼睛為之一亮（*圖表 3.6.2.*）：

圖表 3.6.2.顧城詩組成要素分析（根據選定的詞彙）



82-86 年之間完成的「頌歌世界」組詩是「有我」風格的極致，86-89 年的「水銀」組詩則是「無我」實驗的顛峰。這兩組詩分別朝兩個不同的方向，劃定了八零年代顧城詩風格的疆域。「水銀」的筆法和風格奠定以後，詩人雖然延續這種寫作方式，但是語言較為明朗易讀 [14]，這與上圖中九零年代時期作品離散的情況相符：似乎有一種返回相容 73-76 年風格的趨勢。

4、結論

[14] 翁文嫻 （2004），頁 191。

根據現階段探勘結果，顧城作品的「音樂性」和「節奏感」可分別從兩個方面來定義：

a. **以聲韻模式定義音樂性：**前期作品(60 年代-70 年代)慣用逐步高昇的陽平調，常使用合口和半合口元音發聲，氣流活動範圍較窄，踟躕在雙唇音和舌尖後音之間，鼻音較不明顯;後期作品(80 年代-90 年代)慣用先抑後揚的上聲，常使用開合反差大的雙元音，氣流活動範圍較寬，游走在雙唇音和舌面後音之間，鼻音較明顯。根據頻率效應，我們接下來可以假定，頻率較高的聲韻模式也許可以在背景調和詩句的音色，烘托頻率較低的重點音節。

b. **以語法模式定義節奏感：**前期作品較常出現「的」來定義關係、性質和狀態，補足語和說明語的使用應該較為頻繁，構築寓意的條件較充足;後期作品較常出現「在」來描述動作、時間和地點，補足語和說明語的使用應該較為零散，即席興發的氛圍較濃厚，正巧應證了古典解釋學的精讀結果 [15]。根據頻率效應， 我們接下來可以假定，頻率較高的語法模式則也許能夠在底部推動詩句的節奏，突出頻率較低的關鍵詞彙。

未來的研究將以上述兩個模式進行詩行之間音樂性和節奏感的配對分析，嘗試連接意義斷裂的段落中彼此對應的重點音節和關鍵詞彙。

本研究展示了漢語現代詩風格量化研究的基礎框架。現代詩的流變，及至古典詩和現代詩的親緣關係，都可以就此方法進行探勘、測定和詮釋

[15] 翁文嫻（2004），頁 190-191。

參考書目

MORETTI, F. (2007). *Graphs, maps, trees: abstract models for literary history*. London, Verso.

RAYNER, K. (1998). “Eye movements in reading and information processing: 20 years of research”. *Psychological Bulletin*, 124(3), 372-422.

顧工編。1995。《顧城詩全編》。上海，三聯書店。

顧城。2010。《顧城詩全集》。南京，江蘇文藝出版社。

高友工。2004。《中國美典與文學研究論集》。臺北，臺灣大學出版中心。

廖學盈。2016。＜詩經的量化研究：發掘興體詩的隱藏節奏＞，《第七屆數位人文國際研討會論文集》，頁 401-414。

翁文嫻。1998。《創作的契機》。臺北，唐山出版社。

_____。2004。＜顧城詩「呈現」界域的存在深度——「賦」體美學探討系列之一＞，《當代詩學年刊》，第 1 期，頁 181-201。

_____。2007。＜《詩經》「興」義與現代詩「對應」 美學的線索追探 ——以夏宇詩語言為例探研＞，《中國文史哲研究集刊》，第 31 期，頁 121-148。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 12/21 (W5)
Venue 地點 / 法鼓文理學院 Dharma Drum Institute of Liberal Arts (DILA)

「何處」的隱喻與轉喻

唐詩「空間詩學」的數位人文研究

The Metaphor and Metonymy of "He Chu" (何處)

A Digital Humanities Study on the "Poetics of Space" in Tang Poetry

鄭文惠*　葉昱廷**　顏靜馨***　余清祥****

國立政治大學中國文學系特聘教授*

國立政治大學統計學系碩士生**

台中科技大學通識中心兼任講師***

國立政治大學統計學系教授****

「何處」的隱喻與轉喻——唐詩「空間詩學」的數位人文研究

The Metaphor and Metonymy of "He Chu" (何處)--A Digital Humanities Study on the " Poetics of Space" in Tang Poetry

鄭文惠*、葉昱廷、顏靜馨、余清祥

摘 要

由於中國古典詩歌的典型特徵，大多表現為以具體可感的形象描摹詩人內在抽象的心靈情感，乃至表徵現實世界的紛繁事物與萬端情狀；因而，詩歌中的一個個構詞，實質建構出一個個深具隱喻與轉喻的意象叢軸，從而承載了象徵著詩人乃至一代巨大的心理情感與深刻的思想觀念系統。自 2014 年始，本團隊開始以數位方法先後就《全唐詩》中的色彩詞、樂器詞及其與政治、社會、宗教、文化……等等之關聯進行研究。2015 年，借用高分子化學的「分子鏈」概念，施作於意象叢及主題研究等，提出以「構詞」原則擷取詞彙的數位技術，立基於詩歌的構詞與句鏈結構，分從情感現象學與色彩政治學兩向度，探討中國社會階層結構產生劇烈移位的中唐時期，詩人白色抒情系譜所再現出深刻而繁複的觀念系統與文化景觀，及其所表徵中唐文士獨特的深層心理結構及社會文化的變革，進而驗證國際學界唐宋轉型或唐宋變革相關論述。2016 年更擴大對顏色詞的數位人文探索，納入顏色詞的同義字，在原有技術基礎上，結合 R 進行文本探勘，運用統計理論模型，更為全面且深入地研究唐詩顏色光譜學於文化傳統意義上的轉換狀態，以及各色系離散現象所傳達的情感的、社會的、政治的複雜且豐富的指涉意涵。2017 年，則以弦樂器為主，進行《全唐詩》音樂詩的數位人文研究，分析弦樂器所載承著自遠古以來天人交流、先王治道、聖賢情志、貞女行操等記憶；中原文化與異文化協商、融合的軌跡，及和唐代音樂文化、城市文化、工藝技術、儒釋道思想觀念……等的關係與連結，並凸顯在郊廟祭典儀式、樂府詩中樂器配置的獨特性與功能屬性等。而多年來藉由統計方法 Bi-gram 所得出的高頻詞，「何處」列居《全唐詩》榜首，卻一直無暇研究，因此，本文再次借助 R 進行相關嘗試，希望透過關鍵詞與概念關係網絡所指涉的主題與意象的深層關係，及意象叢軸的隱喻與轉喻系統，建立一套唐詩「空間詩學」的分析方法與理論模型。

首先，本文以 N 元語法(N-gram)解析詩的句構，採取半監督模式篩選出《全唐詩》中最常和「何處」伴隨出現的前 100 個高頻詞，以及與這 100 個高頻詞伴隨出現的共現詞。接著，取共現詞中前 100 個高頻詞為特徵詞（非指定變數）作為統計上的特徵變數(feature)，以階層式集群分析(Hierarchical Clustering)的方法，進行探索性資料分析(Exploratory Data Analysis, EDA)。以「何處」作為關鍵詞，

從《全唐詩》中提取詩歌樣本約有 1,636 份，詞頻為 1,667 次。下二表分別列舉「何處+X」、「X+何處」共構次數大於 2 的共構字表。在詩歌中，「何處」一詞或實指或虛指，大致呈現出下列幾種類型或意涵：或肯定表述，如：「春江潮水連海平，海上明月共潮生。灩灩隨波千萬里，何處春江無月明」（張若虛〈春江花月夜〉）；「聖明無一事，何處讓堯年」（無名氏〈晉朝饗樂章・群臣酒行歌〉）；「出處全在人，路亦無通塞。門前兩條轍，何處去不得」（聶夷中〈雜曲歌辭・行路難〉）。或否定表述，如「築城畏不堅，堅城在何處」（陸龜蒙〈築城曲〉）。或更大量表述為疑問或惶然不寧的情狀，如：「何處接長波，東流入清渭」（于濆〈隴頭水〉）；「矯翼知何處，天涯不可窮」（杜牧〈別鶴〉），均呈顯茫然不可知不可行之感。或表述身體或心靈的承受性，如：「何處力堪殫，人心險萬端。」（薛能〈行路難〉）。至於指涉範疇空間：或閨怨空間，如「誰家獨夜愁燈影，何處空樓思月明。更入幾重離別恨，江南岐路洛陽城」（柳中庸〈聽箏〉）；或家鄉空間：「山川路長誰記得，何處天涯是鄉國」（劉商〈胡笳十八拍・第四拍〉）；或情欲空間：「何處期郎游，小苑花臺間」（李暇〈怨詩三首〉之二）；或功名空間：「朱顏日漸不如故，青史功名在何處」（白居易〈浩歌行〉）；或戰場空間、或理想空間……等等。而「何處」抽象的隱喻與轉喻：「行路難，行路難，何處是平道」（顧況〈行路難〉之一）；「一聲何處送書雁，百丈誰家上水船」（杜甫〈十二月一日三首〉之一）更是包覆著巨大的個人的、社會的、家國的、政治的、文化的現實與情感的意義與象徵。

表一：何處＋X（T=次數，表列>2 的字詞）

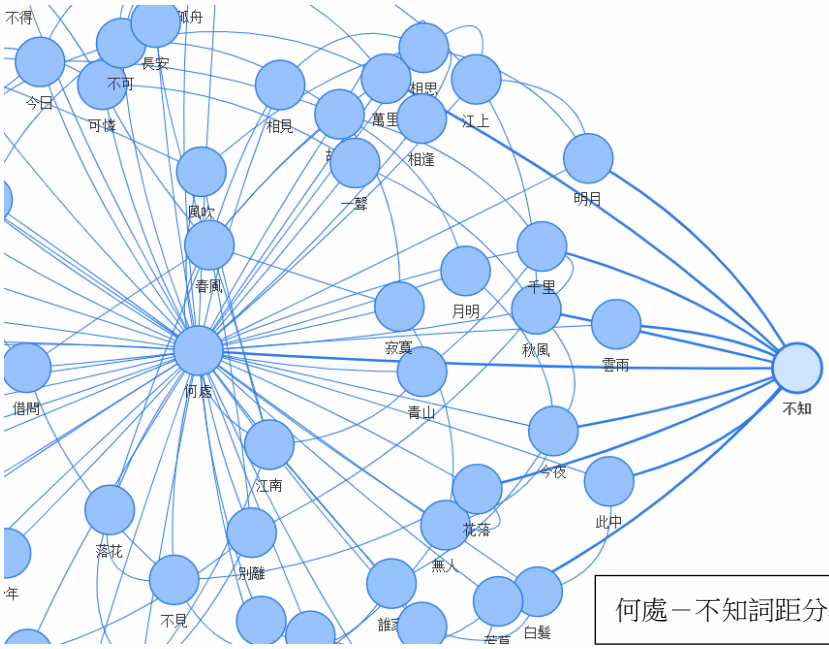
X	去	來	在	尋	宿	是	歸	好	所	客	期	邊	望	盡	落	遊	人	行	問	飛	覓	傳	說			
T	65	28	27	25	22	21	19	17	12	11	10	8	7	7	6	6	5	5	5	4	4	4	4			
X	火	有	老	住	村	見	流	起	笛	逢	發	樓	聲													
T	3	3	3	3	3	3	3	3	3	3	3	3	3													
X	入	山	水	生	寺	別	定	泊	看	眠	砧	酒	得	船	無	開	雲	圓	葬	路	夢	聞	遠	還	隱	斷
T	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

表二：X＋何處（T=次數，表列>2 的字詞）

X	在	知	歸	向	落	今	生	來	遊	人	風	從	雲	聲	帆	舟	君	花	魂	投	於	飛	宿	情	心	門	書
Freq	60	40	35	21	14	15	13	13	9	8	8	8	8	8	7	7	7	7	7	6	6	6	6	6	5	5	5
X	光	衣	到	為	寄	居	思	郎	將	船	通	離	川	功	去	自	吟	更	見	言	身	依	宜	城	泉	軍	宵
Freq	4	4	4	4	4	3	3	3	3	3	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
X	家	留	迷	移	復	朝	琴	源	猿	詩	塵	歌	聞	鳴	墳	樓	醒	應	霜	鐘	靈						
Freq	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2						

而與「何處」共現頻率最高的詞是「不知」(142 次)，其次分別為：萬里、千里、春深、春風、不可、人家、不見、誰家、悠悠、相思、今日、白雲、此時、天涯……。以「何處－不知」為例，其共現狀況常見有：(一)「不知何處」同句連用(60/142)，如：「不知何處恨，已解入箏弦」（柳中庸〈寒食戲贈〉）；「不知何

處去，月照玉樓空」(竇鞏〈遊仙詞〉)；「送客自傷身易老，不知何處待先生」(李端〈贈道者〉)；(二) 又或者「不知……何處」同句共現(12/142)，在這一組句式中，「歸何處」(5/12)的詞頻高於其它。或 (三)「不知……，何處……」同聯共現的形式，如：「不知精爽歸何處，疑是行雲秋色中」(李羣玉〈題二妃廟〉)，「不知移舊愛，何處作新恩」(白居易〈怨詩〉)；也有「……何處……，……不知……」，如「明日薄情何處去，風流春水不知君」(高蟾〈偶作〉之二)。



何處－不知詞距分析圖

在 tri-gram 中，前五組高頻詞分別為「在何處(86) 」、「何處是(67) 」、「何處去(65) 」、「歸何處(57) 」、「何處來(28)」、「知何處(28)」，其中「知何處(28)」(扣除「不知何處」的次數) 往往也是「不知」的狀態。凡此，均可再進一步解析其中的構詞方式與句型結構，及其藉由「何處－不知」所建構出的意象叢與主題內涵，乃至於勾勒出唐代詩人群體透過隱喻與轉喻所呈現的「空間詩學」。

數位人文研究不僅是「數位」+「人文」，本次進行非指定變數與指定變數的嘗試，是希望能理解人文學者在初步資料分析的介入度會產生哪些影響，同時也希望能尋找出盡可能降低主觀成分的「有意義分析」方法，進而讓人文學者進行更有意義的解讀。在統計中，集群分析能將特徵相近的樣本歸類為同一群，不同群之間則有明顯的差異，類似生態系中的不同生物族群。由於變數之前可能存在著共線性(Multicollinearity)，因此主成分分析得以透過變數之間的線性組合構成主成分，使得主成分之間相互獨立，消除共線性的問題，並達到維度縮減的效果(Dimension Reduction)。由於共現詞之間可能有著雷同的語意，加上詞彙種類眾多，因此本次將 PCA 技術引入共現詞叢的分析，萃取分群所需的有效變數。

本文將高頻詞視為詞彙物種，特徵詞視為語意環境，並判斷哪些詞彙物種存在於相同的語意環境，而語意環境背後的意義與象徵，正是文學的精粹所在。索緒爾語言二軸理論所提出的組合(syntagmatic)與聚合(paradigmatic)兩個概念，應用在統計分析上，共現詞及其頻率呈現出語序關係中話語選擇的元素，即組合關

係，階層式集群分析討論相似性、範疇化，以及其它聚合關係；在認知語言學中，前者為轉喻(metonymy)系統建構元素，後者則因聯想形成隱喻(metaphor)系統。2015 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，本團隊藉由詩歌的意象造境與詩人的心理情感及文化隱喻等多重層關係，勾勒出中唐知識階層心理結構與身分認同變異的象徵圖譜。而本文在可考範圍內，中唐以降的「何處」詩作佔了將近 2/3 的比例，其中實質表徵著中唐社會階層位移及家國劇變下詩的現實與抒情產生了巨大的改變，不知、質疑、迷茫、困惑、失落、焦慮、不安、猶疑、徬徨……一種心靈歸屬的急切的扣問——屬於人類「原鄉」宇宙，一種主體足以安身立命的「棲居」的情感樣態與理想空間的扣問。就在現實多變的遭逢，或歷史的追憶與文化的召喚中，透過現實、記憶與空間的意指實踐，及其外延與裂變的構合關係，表述了一種如何與文化歷史綿延相續的時間存在性共感及不斷追尋一個主體「棲居」的理想空間。也或許詩之所以為詩，將詩人帶入了「棲居」的理想樣態；在詩的宇宙中，詩人既進入現實也擺落了或超脫於現實之上，而回到了原鄉。

關鍵詞：唐詩、何處、空間詩學、隱喻轉喻、數位人文研究

*鄭文惠，國立政治大學中國文學系特聘教授
葉昱廷，國立政治大學統計學系碩士生
顏靜馨，台中科技大學通識中心兼任講師
余清祥，國立政治大學統計學系教授



中日古辭書自動文本標注顯示工具
tagzuke

tagzuke
An Automated Markup Tool for Medieval China dictionaries and Early Japanese Dictionaries

劉冠偉

日本北海道大學博士課程學生

中日古辭書自動文本標注顯示工具 tagzuke

tagzuke: An Automated Markup Tool for Medieval China dictionaries and Early Japanese Dictionaries

劉冠偉

博士課程學生

日本 北海道大學

摘要

近年，隨著 Open Data（開放資料）與 Open Science（開放科學）的興起，典籍的數位化已成為人文研究的重要方法之一。中國與日本的古辭書作爲研究古代漢語、日語的重要資源，亦有很多資料被公開。其中，以全文文本形式公開的中日古辭書資料庫不在少數，而如何在研究中運用這些資料庫也成爲了一個重要的課題。

日本古辭書的結構較爲複雜。在部首分類體的古辭書中，每個條目由字頭與釋文構成，用於解釋漢字的形、音、義。釋文中又大致可以分爲字音注、字體注、漢文義注、和訓四種要素。每種要素都有不同的表現形式，不同學者所關注的資料也各有不同。如果對文本資料庫中的相關要素進行標注，就可以准確地提取出相應資料來進行研究比對，對於文獻研究具有重要的意義。

在中文古籍研究領域，已有的文本標注顯示工具較難應用在辭書類古文獻中出現的上述要素之上。因此，在日本古辭書的資料庫化過程中，需要一款能夠與之對應的文本標注顯示工具。

因此，作者提出了一套專用於日本古辭書文本標注的簡化標簽集（tag set）以及與之對應的文本標注工具 tagzuke。隨後，作者主要圍繞著操作性、汎用性、維持性三個方面對本工具做了改善。

本次發表，主要集中於改進釋文中要素的自動預標注，提高預標注的準確性；以及嘗試對中國部首分類體古辭書《大廣益會玉篇》進行標注，使 tagzuke 成爲一個可支援中日古辭書資料庫的自動文本標注顯示工具。

最後，本工具已於開源平台 GitHub 公開所有代碼。歡迎各界人士的使用與指正。

目次

1. 前言
2. 中日古辭書與其數位化

2.1 古辭書的數位化
2.2 古辭書的結構特徵
2.3 釋文要素標籤集
3. 自動標記工具 tagzuke
3.1 tagzuke 的目的
3.2 tagzuke 的開發過程
3.2 tagzuke 的操作方法
3.3 tagzuke 的開發
4. 使用 tagzuke 標注《大廣益會玉篇》
4.1 《大廣益會玉篇》文本資料庫
4.2 《大廣益會玉篇》的要素特徵
4.3 《大廣益會玉篇》標記結果
5. 結語

關鍵詞

類聚名義抄，篆隸萬象名義，大廣益會玉篇，XML

1. 前言

近年，隨著 Open Data（開放資料）與 Open Science（開放科學）的興起，典籍的數位化已成為人文研究的重要方法之一。古辭書作爲研究古代漢語、日語的重要資源，亦有很多資料被公開。

其中，以全文文本形式公開的中日古辭書資料庫不在少數，例如“漢字資料庫（**漢字データベース**）”中的宋本《廣韻》，“平安時代漢字字書綜合數據庫（**平安時代漢字字書総合データベース**，HDIC）”中公開的高山寺本《篆隸萬象名義》、觀智院本《類聚名義抄》等。因此，如何在研究中運用這些資料庫也成爲了一個重要的課題。

部首分類體日本古辭書的釋文中大致可以分爲字音注、字體注、漢文義注、和訓四種要素。不同的要素擁有不同的表現形式，不同學者所關注的要素也各有不同，如何準確快捷的從資料庫中抽取出這些要素對於研究者具有重要的意義。另一方面，在校正資料庫時，不同的要素有時會採用不同的方針，例如字體註應儘量貼近原本所書寫的字體，而為了方便檢索，字音註又應使用後世音韻資料中相同的字體。

在中文古籍研究領域，已經有了 Markus 這一文本標注顯示工具被大家所熟知。Markus 可針對文獻中出現的人名、地名、年代等進行標注。但是較難應用在辭書類古文獻中出現的上述要素。

因此，發表者認為亟需一款能夠將古辭書中的釋文要素分類、標記的工具。對於日本部首分類體古辭書，發表者開發了一套專用於古辭書文本標注的標簽集（tag set）以及與之對應的文本標注工具 tagzuke，並以 HDIC 中公開的日本古辭書觀智院本《類聚名義抄》全文資料庫爲例，實際驗證了本工具的可用性。

本次發表，主要集中於改進自動預標注，提高預標注的準確性，以及對同時代的中國部首分類體古辭書《大廣益會玉篇》進行適配，使 tagzuke 成爲一個可支援中日古辭書資料庫的自動文本標注顯示工具。

2. 中日古辭書與其數位化

2.1 古辭書的數位化

以全文文本形式公開的中日古辭書資料庫有“**漢字データベース**”中的宋本《廣韻》，“平安時代漢字字書綜合數據庫（**平安時代漢字字書総合データベース**，HDIC）”中公開的高山寺本《篆隸萬象名義》、觀智院本《類聚名義抄》等。

2.2 古辭書的結構特徵

日本的部首分類體古辭書內容多來源於中國古辭書，其條目之結構如圖一所示。而

中國的部首分類體古辭書除去沒有和訓這一要素外，結構基本相同。另，在部首分類體古辭書中，條目以外的結構多較為簡單，故本發表不做討論，下文內容均針對釋文文本。

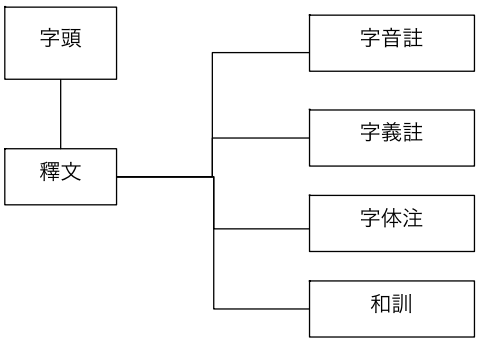


図 一、日本部首分類體古辭書結構

2.3 釋文要素標籤集

在構建大型資料庫時，通常由數名乃至數十名研究人員分擔工作。但不論是 XML 語言還是 TEI（Text Encoding Initiative）標準的使用都需要一定的前置知識，在人數較多的項目中會造成一定的負擔。

在 TEI P5 中 9. Dictionary（辞書類）部分中記載的標籤集（Tag Set）應用於日本古辭書時難點甚多。例如，釋文中的和訓不知應適用表示定義的<def>標籤還是表示發音的<pron>標籤。或者適用表示註釋的<note>標籤，使用屬性@type 來表示釋文要素的類別，然而<note>標籤下層可以使用的標籤種類具有一定限制，難以處理釋文中的埋字[1]條目。因此，本研究將不依據 TEI 標準，而採用單獨設計的標籤集。

釋文中不僅有多樣的要素，還有複雜的結構。為了更高的實現可能性，在本研究中將從淺層結構到深層結構，分步驟設計標籤集與執行標記。特別是在處理字音註與日本古辭書特有的和訓時還有較多未被研究透徹的內容，需要能夠跟隨最新研究解讀而修改的柔軟性。

更多標籤集相關內容請參考劉等人（2017）。本次發表僅限使用最淺層的要素標籤，如下表一所示。

表 一、要素標籤

標籤	對應要素
<ziti>	字體註
<ziyin>	字音註
<ziyi>	字義註

[1] 指出現在釋文內的條目。

3. 自動標記工具 tagzuke

3.1 tagzuke 的目的

日本古辭書大量引用了由中國傳入的古辭書內容，對於研究已在中國散佚的辭書資料具有很大的價值。在部首分類體的古辭書中，每個條目由字頭與釋文構成，用於解釋漢字的形、音、義。日本古辭書在參考了中國古辭書之後，釋文中加入了日語釋義（和訓），表示音調的聲點等，使得結構變得比原本的中國辭典更加複雜。

部首分類體日本古辭書的釋文中大致可以分爲字音注、字體注、漢文義注、和訓四種要素。每種要素都有不同的表現形式。不同學者所關注的資料也各有不同，例如日本國語學研究者關注釋文中的和訓，中日漢語學者則多會關注字音注、字體注。如果對文本資料庫中的相關要素進行標注，就可以准確地抽出相應資料來進行研究比對，對於文獻研究具有重要的意義。

3.2 tagzuke 的開發過程

作者在 2017 年 12 月舉辦的“2017 年人文科學與計算機討論會（人文科学とコンピュータシンポジウム 2017）”上，提出了一套專用於古辭書文本標注的標簽集（tag set）以及與之對應的文本標注工具 tagzuke，並以 HDIC 中公開的日本古辭書觀智院本《類聚名義抄》全文資料庫爲例，實際驗證了本工具的可用性。tagzuke 由作者一人開發，在最初開發之際，tagzuke 是一款專為日本部首分類古辭書中的文本進行自動化標注顯示的工具。這款工具可以將翻刻好的文本以 CSV 文件格式導入，通過滑鼠點擊釋文中的要素來進行標注，并將標注好的要素施以相應的顏色來顯示，最後可以將標注好的文本以獨有的文本格式輸出保存（圖二）。

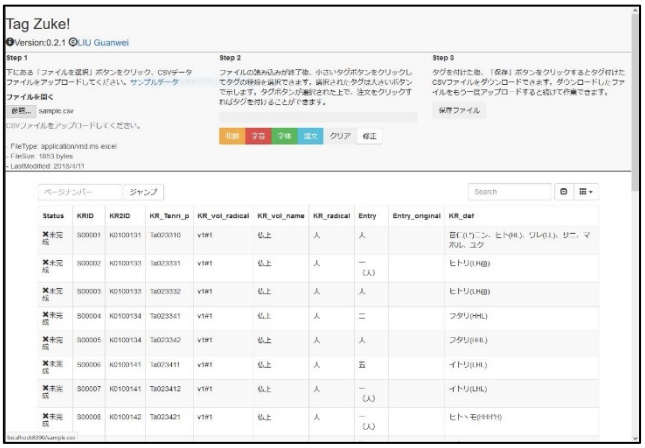


図 二、舊版 tagzuke

隨後，作者主要圍繞著操作性、汎用性、維持性三個方面對本工具做了改善，并在2018 年 5 月舉辦的“第 117 次人文科學與計算機研究會發表會（第 117 回人文科学とコンピュータ研究会発表会）”上做了相關匯報。具體改善方面如下所示。

操作性：加入了鍵盤方向鍵操作，自動預標注。

汎用性：支援 Excel 文件格式，可自定義字頭、釋文的欄目。

維持性：導入開源框架。

藉此三方面的改進，可以更高效率地對日本古辭書的釋文要素進行標注。此外，還嘗試了對與中國古辭書形態上比較接近的高山寺本《篆隸萬象名義》進行標注（圖 2、圖 3）。



図 三、使用 tagzuke 標記《篆隸萬象名義》（一）



図 四、使用 tagzuke 標記《篆隸萬象名義》（二）

3.2 tagzuke 的操作方法

Tagzuke 的操作順序如下所示。

1 打開檔案

2 設定使用列，要素之間的分隔符

3 使用鍵鼠進行標註

4 確認與保存

古辭書中釋文要素具有一定模式。利用其模式特徵實現了釋文要素的自動標記，詳細內容將在後文描述。Tagzuke 的操作界面分為左右兩部分，左半為菜單欄，右半為操作欄。根據使用終端的分辨率，會自動設定是否顯示菜單欄，以擴大操作界面顯示面積。



図 五、打開檔案

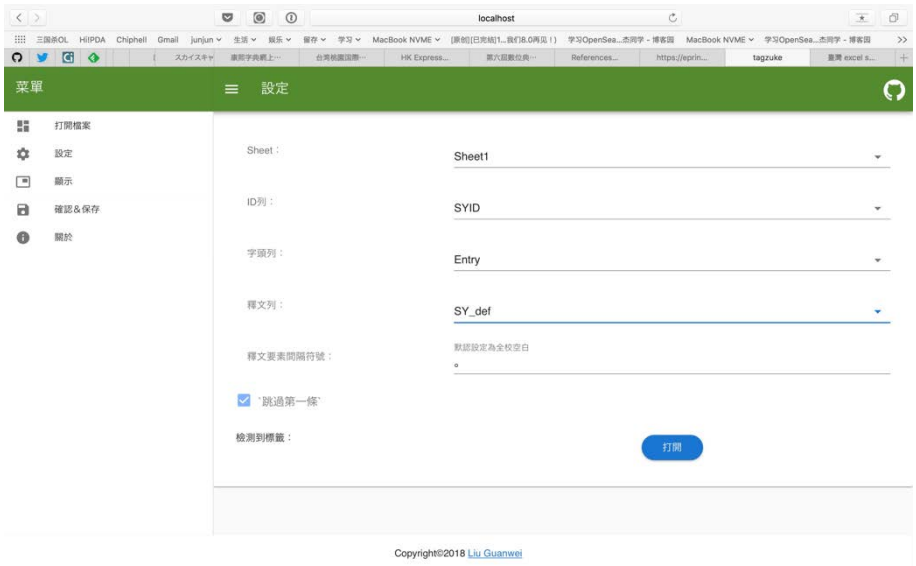


図 六、設定界面

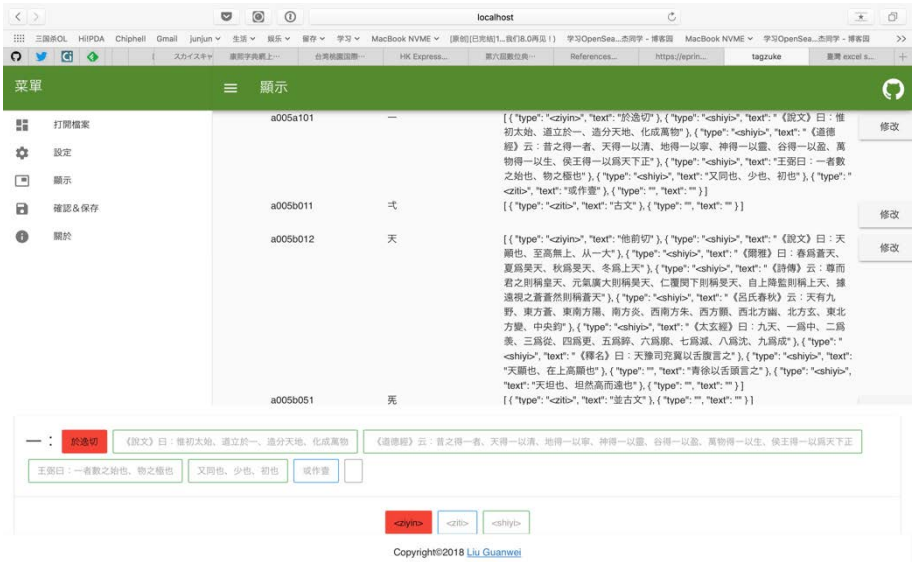


図 七、標記界面

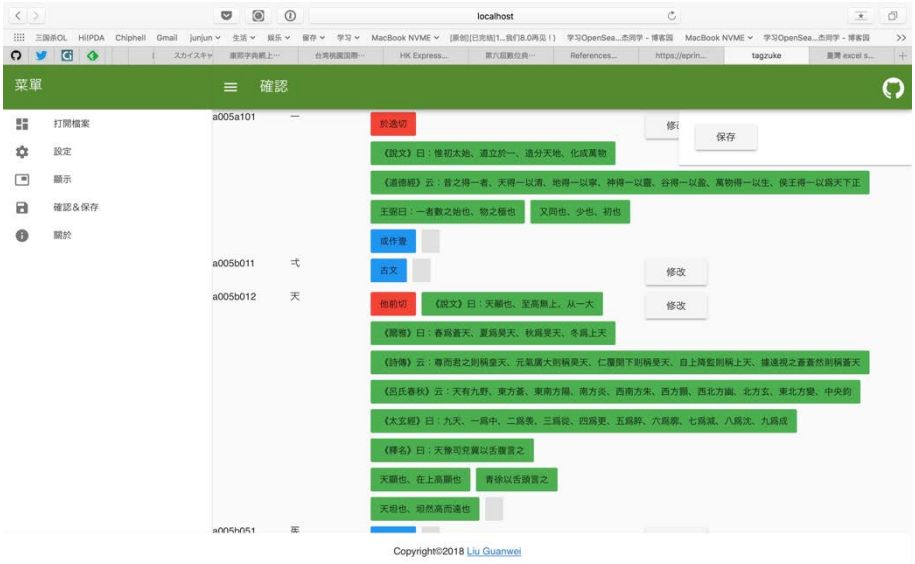


図 八、確認保存界面

3.3 tagzuke 的開發

Tagzuke 是一款在瀏覽器使用的 Web APP，使用了開源框架 VueJS，Vuetify，XLSX-JS 等開源資料庫。本工具已於開源平台 GitHub 公開所有代碼，代碼倉庫為 https://github.com/toyjack/tagzuke。

4. 使用 tagzuke 標注《大廣益會玉篇》

4.1 《大廣益會玉篇》文本資料庫

本次發表使用“平安時代漢字字書綜合數據庫（平安時代漢字字書総合データベース，HDIC）”中公開的《大廣益會玉篇》文本資料庫。此資料庫基於日本宮內廳書陵部所藏的宋板，與澤存堂本製作而成。資料庫的結構如下表。

表 二、資料庫結構

列	說明
SYID	位置 ID
SY_vol_radical	所在卷數與部首號碼
SY_radical	部首
Entry	字頭
Entry_original	原本形近字頭
UCShex	Unicode 位置碼
SY_def	釋文
KSY_diff	宮內廳本區別
SY_remarks	備註

該資料庫現已於 GitHub 平臺開放下載[2]。

4.2 《大廣益會玉篇》的要素特徵

《大廣益會玉篇》的釋文可大致分為字音註，字體註，釋義三種要素。其中，各要素的特徵之總結如下所示。〇代指漢字。

字音：〇〇切，〇〇〇〇二切，音〇，又音〇，亦音〇，〇〇二音等。

字體註：俗作〇，俗〇字，亦作，或作，本作，籀文作，篆文作，本亦作，原作，正作，今作，俗以琦瑁作玳，古文等。

釋義：〇〇也，〇〇皃（貌）等。

另外，參照 2.3 節，“同上”“同〇也”等須參照其他條目之釋文在本次發表中暫不處理。

[2] 開放下載地址：https://github.com/shikeda/HDIC

4.3 《大廣益會玉篇》標記結果

限於篇幅，此處將在發表會場展示。

5. 結語

本次發表介紹了古辭書中日古辭書自動文本標注顯示工具 tagzuke 的理念以及使用方法。本工具可以說尚處於原型開發階段，標籤集的設定也需更多討論，權當拋磚引玉，期待與會學者們的建議與指正。

引用文獻

池田証壽（2014）。〈**平安時代漢字字書総合データベース：現状と課題 2014 夏**〉，《**漢デジ 2014—デジタル翻刻の未来**》，頁 3-43。

劉冠偉，李媛，鄭門鎬，張馨方，池田証壽（2017）。〈**部首分類体日本古辞書の項目構造の多様性に対応したマークアップ・ツールの開発**〉《**じんもんこん 2017 論文集**》，頁 97-102。

劉冠偉（2018）。〈**日本古辞書マークアップ・ツール tagzuke の課題−操作性・汎用性・維持性の改良−**〉，《**研究報告人文科学とコンピュータ**》2018-CH-117(11)，頁 1-4。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 12/21 (W5)
Venue 地點 / 法鼓文理學院 Dharma Drum Institute of Liberal Arts (DILA)

傳記文本之資訊擷取

以續修台北市志人物志為例

NLP Methods for Information Extraction from Biographies

An Exploration with the Elite Biographies in the Extended Taipei Gazetteers

王億祥* 陳維睿** 劉昭麟**

國立政治大學企業管理學系*

國立政治大學資訊科學系**

傳記文本之資訊擷取 — 以續修台北市志人物志為例

NLP Methods for Information Extraction from Biographies: An Exploration with the Elite Biographies in the Extended Taipei Gazetteers

†‡王億祥　　†陳維睿　　†劉昭麟

†國立政治大學資訊科學系

‡國立政治大學企業管理學系

{103305079, 105753015, chaolin}@nccu.edu.tw

摘要

本研究旨在探索數位技術在抽取傳記資訊上的新方法或程序，內容包括人名辨識的加強方法、關係抽取的輔助方法、加權的共現關係、自動產生年譜的嘗試。我們以續修台北市志人物志為例來展現，但是多數方法也可以套用或類推到其他文本。經驗顯示以數位技術自動擷取傳記資訊的深度雖仍不及人工分析，但是精心設計的計算程序和新方法展現了輔助人文研究的未來潛力。

關鍵詞(Key Word)

文字探勘、關係擷取、年譜自動建構、人名辨識、傳記資料庫

text mining, relation extraction, automatic construction of personal chronicle, name entity recognition, biographical databases

一、緒論

傳記為歷史和文學研究的重要一環，對歷史和文學都提供相當豐富的資料[1]。但傳記研究常需梳理大量的文獻，支持一個論點也需要理清大量人物關係、生平經歷，極為費時。隨著研究範圍擴大和對象變多，研究的網絡也會變得更複雜，使得研究範圍受限於人工可以處理的程度。

因此從過去就有許多先驅，探索數位技術在傳記研究上的應用。「CBDB」分析社群關係網路[2]，自動畫出關係圖，並與地理資訊系統結合等等。而在台灣，相似性質的「TBDB」也正在起步。本研究也試著提出新的方法，來加強或輔助過去的數位技術。

首先我們說明，利用合併複數個人名辨識工具，以增強效用的方法。再說明如何《續修台北市志》人物志的撰寫特性，以依存關係和正則表達式來輔助抽取人物間的關係。接下來介紹在傳統的共現關係之上，進行加權計算的方法，及說明其意義。最後嘗試自動產生人物年譜，並利用依存關係進一步產生精簡版年譜。

2. 語料

台北市是台灣的政經中心，對台灣歷史有舉足輕重的影響，尤其目前台灣地方志的研究還比較少，所以我們以《續修臺北市志》[3]的人物志作為語料，其分為「政治與經濟篇」、「社會與文化篇」，共319篇傳記，729,182字。

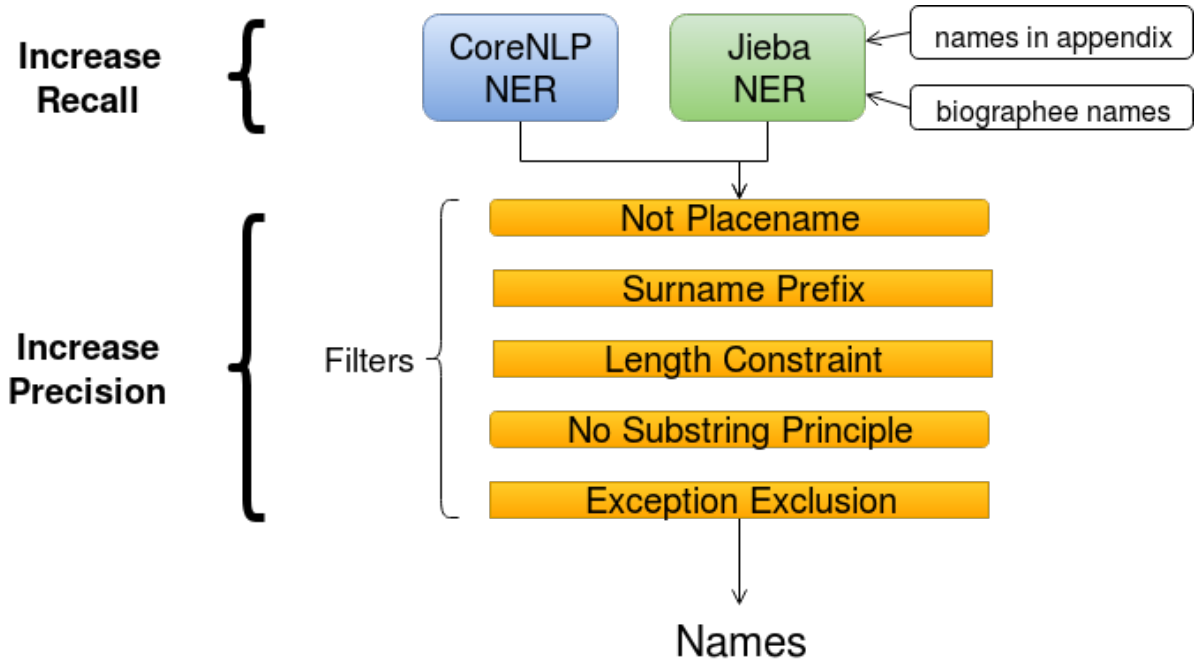
3. 人名辨識

在以人為主的傳記研究中，辨識出其中的人名自然變成重要的工作。在此我們提出一個容易實現的手法，提升辨識到的人名數量和準確程度。

[1] 張素玢(2018).<建置 「臺灣歷史人物傳記資料庫」 (TBDB)的嘗試與初步成果>.《人文與社會科學簡訊》, *19*(2),頁102.

[2] Cho Wonhee. 2016. “DigitalDigital Humanities and Yuan Studies: An Introduction to the CBDB and its Potentials，” CENTRAL ASIAN STUDIES *21*(1), ｐｐ． 117－138.

[3] 《續修台北市志》網址：<https://www.chr.gov.taipei/cp.aspx?n=EEBCFC3935F53838>



圖一、 人名辨識加強手法的概念圖

相比普通僅使用單一人名辨識工具，我們將使用「CoreNLP」[4]和「Jieba」[5]這兩個工具的人名辨識結果合併。針對「Jieba」，我們還將人物志的附錄和目錄中的人名作為其輔助資訊。另外針對親屬人名，用正則表達式作有效的擷取，此手法的詳細將在下一章節一併說明，此三舉皆是為了提升辨識到的人名數量。

但是這樣的方法同時會造成錯誤辨識的增加，因此我們設計了五個過濾器來過濾結果。「非地名」排除被誤認為是人名的地名、「姓氏存在」規定有效的人名必須有姓氏(以百家姓和日本前七千常見姓氏為主)作為前綴、「長度限制」則限定人名長度在2~4字間、「不為子字串」則源自於人名的部分常被辨識成另一個人名(例如人物王喬峰卻有王喬峰和喬峰兩個人名結果)、「例外處理」則排除掉高頻率的漏網之魚，例如「於民國」(原文為「於民國19xx年...」)。

此同時增加辨識到的人名數量和準確度的手法，在隨機抽樣十個傳記的評估中，取得了 precision 0.82、recall 0.78 作為其效用的證明[6]。

[4] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. 2014. "The Stanford CoreNLP natural language processing toolkit." In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55 – 60.

[5] Jieba - Chinese text segmentation: built to be the best Python Chinese word segmentation module. https://github.com/fxsjy/jieba

[6] precision 和 recall 是評比資訊檢索結果的常見指標，可參考下列經典自然語言處理書本之定義：Christopher D. Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. The MIT Press. 1999.

但是其仍有不夠完善的地方，首先有「同名同姓」的問題，當兩人物同名同姓時，此手法並沒有涵蓋到如何分辨成兩人物。再來是「外國譯名」，如「蘿絲」並沒有有效的姓當前綴。

作為可能的改善方法，由於傳記常有豐富的時間資訊，我們可能可以加上時代來辨別同名同性但活躍時代有差的人物。而在台北市人物志，外國譯名通常會括號標註原文名字，因此最簡單的方法就是加上外國的常見的姓氏到事先定義的有效姓氏，但這就須更多的額外的努力來列出外國常見姓氏。

4. 人物關係擷取

人物之間的關係是許多人文研究的重點，自動關係擷取也累積了許多研究，在此將針對傳記體提出兩個自動關係擷取的「輔助方法」。

首先我們可以發現，在以中文敘述句型，主詞(S) – 動詞(V) – 受詞/補語(O)，的傳記體中，在提到傳主時，常省略其名，導致主詞或受詞的省略。而利用史丹佛大學的依存關係分析器，我們可以找出一個句子裡的主詞、動詞、受詞，進而根據圖二的規則，我們可以找出某人物與傳主的人物關係。

《王世慶傳》

Principle	Text	Result Relation
V + O	V O O 民國56年(1967)先後**訪問** *陳逢源*、*黃旺成* 兩位先生。	王世慶 **訪問** 陳逢源 王世慶 **訪問** 黃旺成
S + V	S S V 在 *林熊祥*、*林衡立* 等人 **鼓勵** 下	林熊祥 **鼓勵** 王世慶 林衡立 **鼓勵** 王世慶

圖二、 以依存關係擷取人物關係的規則 (以王世慶傳為例)

此外針對親屬關係，台北市人物志常以「育有x子x女」為開頭，後面羅列子女名，如圖三所示，或「父xxx」，等等特殊的寫法，因此我們可用正則表達式來擷取其名字和親屬關係，如圖三所示。



圖三、左下內容為數個傳記關於子女的描寫，左上為正則表達式，彩色反白的地方為正則表達式對應到的文面

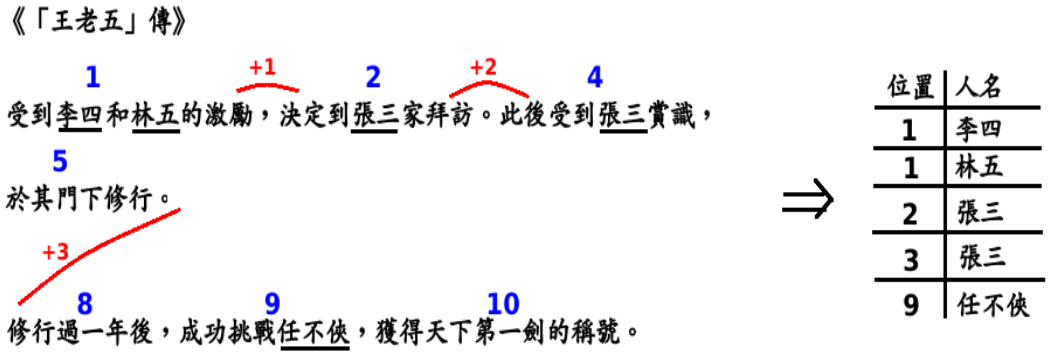
這兩個關係擷取的輔助手法的共通缺點為，依賴於傳記的特殊寫法，使得可能難以推廣。而進一步分開來說，利用句型的手法，十分依賴依存解析器的表現，但精準的中文依存解析是項困難的任務。另一方面，利用正則表達式的手法，則是在於發展正則表達式需要大量的心力和時間觀察寫法規則。

但是我們仍可以期待，此二手法可以與主流的關係擷取方法並用或混用，取得更好的結果。

5. 加權共現關係排名

當人物數量眾多，很容易就有超過破百的共現關係(cooccurrence)。在這裡我們將在傳統的共現關係上，以一個新的計算方式來加權，使得我們能夠將有限的心力關注在比較重要的共現關係上。

假設兩人名在文面上的距離和兩人間的關係深淺有關。 我們將每句子都標示位置，從1開始，經過逗點時，下一句位置+a，經過句點則+b，經過段落間隔則+c, 且a<b<c，並且以句子的位置標示其句中人名的位置，則我們可得到考慮不同間隔下會有不同影響的人名位置，如圖四。則進一步也可得到任兩人名之間的距離，至此做好了加權計算共現關係的鋪墊，如圖五。



圖四、標示句子的位置，並進而標示人名的位置之圖。 在此例a=1, b=2, c=3。

人名	人名	距離	人名	人名	距離
李四	林五	0	林五	任不俠	8
李四	張三	1	林五	王老五	2
李四	張三	2	張三	任不俠	7
李四	任不俠	8	張三	任不俠	5
李四	王老五	2	張三	王老五	2
林五	張三	1	任不俠	王老五	2
林五	張三	1			

圖五、根據上述方式決定出任兩人名間的文面距離。

再來假設隨著距離愈遠，距離變化對兩人關係深淺的影響越小，我們將常數k (k>0) 除以距離， 將距離轉化成關係深淺的分數，使得距離愈遠，距離差對分數的影響愈小。而人名重複出現的話，兩人物間會有多個分數相加，但隨著出現次數增加，我們讓分數折舊越多，使得分數雖然會受到人名重複出現的影響而增加，但是不會過份增加。公式如下

$$\text{Score} = \frac{k}{\triangle p_1+1} + \alpha \times \frac{k}{\triangle p_2+1} + \alpha^2 \times \frac{k}{\triangle p_3+1} + \cdots$$

$$= \frac{k}{\triangle p_1+1} + \sum_{i=2} \alpha^{2^{i-2}} \times \frac{k}{\triangle p_i+1}$$

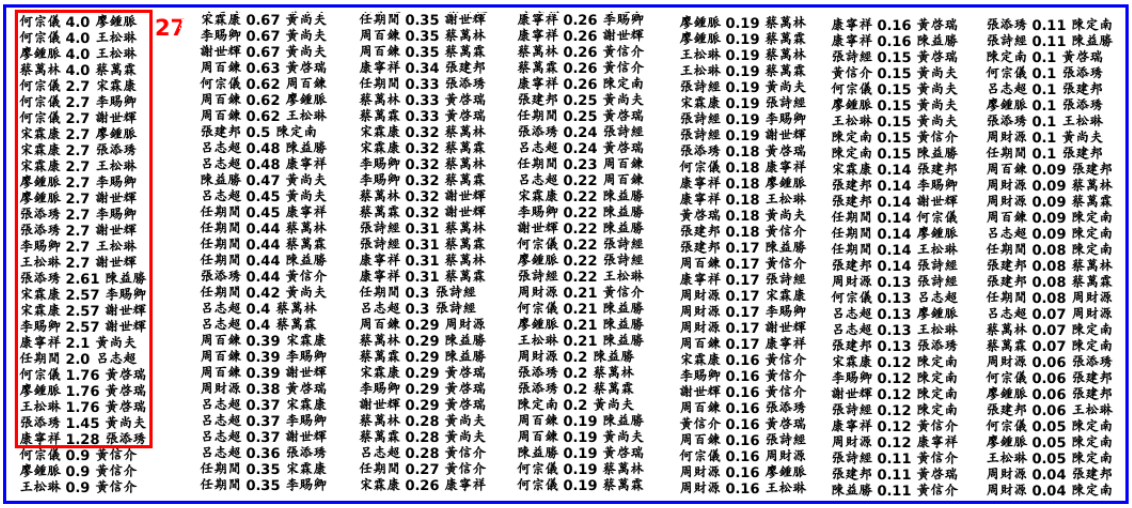
$\triangle p_i$:= the i-th small difference in positions of sentences where the two person names appear

α := depreciation factor, $0 < \alpha < 1$

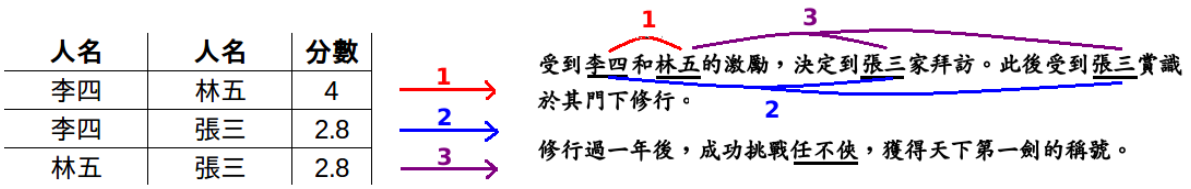
k := distance to raw score factor, $k > 1$

圖六、加權的計分公式

而其結果的意義不在數值本身，在於數值的大小排名。 即使一小群人，兩兩之間有關係下，共現關係可能有非常多個，此時共現關係排名可以協助研究者濾出可能比較重要的關係，如圖七。二來，文本中有些人物關係是沒有在文面上明說的，但是透過共現關係，我們可以協助研究者找出這些隱藏的人物關係，如圖八。



圖七、以張詩經傳為例，在共210個共現關係中，當僅取分數>1的共現關係，範圍將縮小至27個。(在a=1,b=2,c=3,k=4, α =0.65的參數下)



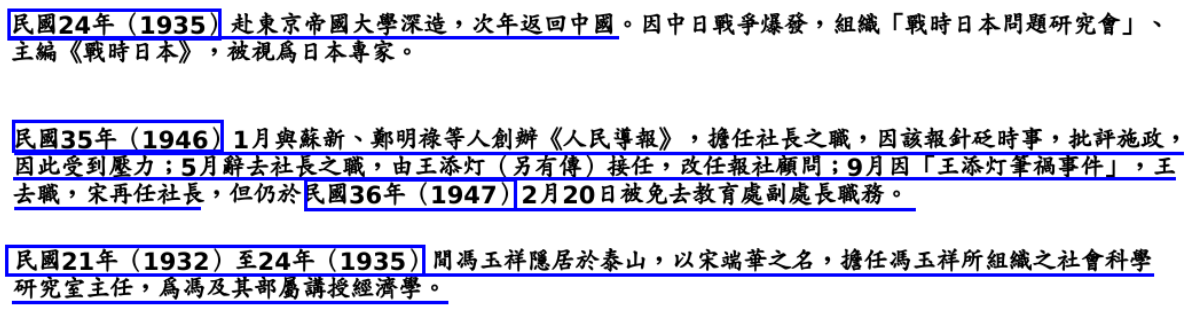
圖八、延續圖四、五的例子。共現分數前三名的人物組合，在文中雖沒有明示，但可推測出有可能是認識或擁有某種聯繫。

6. 自動年譜產生

除了理解人物的關係， 了解人物活動在時間上的分佈也是人文研究重要的一環。過去鮮少有人研究自動建構年譜，在此我們將其實踐，並說明需特別注意的地方和難處。

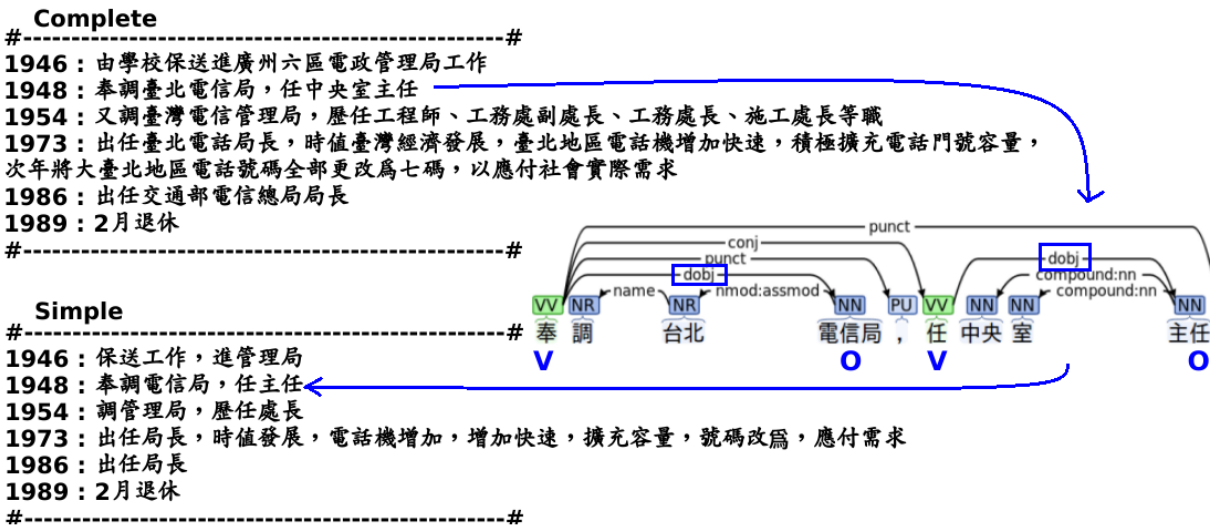
這裡自動產生年譜的手法為，利用傳記對年代的寫法特性，用正則表達式找出所有年代，並將年代後面所有文字依序抓回，直至遇見另一個年代或句點，作為該年代的事件，如圖九。

但是要特別注意，例如「民國46年（1957）至47年（1958），《自由中國》發表...」，雖包含兩個年代，但其實是一個持續了數年的事件。另外就是年代不一定是順敘或是倒敘，因此要將年代作排序。



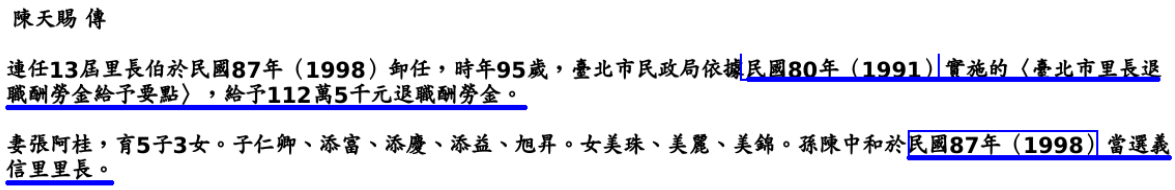
圖九、自動產生年譜手法的概念圖

雖然產生的年譜看起來大都不錯，我們仍進一步嘗試產生「精簡的年譜」，將事件敘述精減以便更快了解全貌。我們用依存關係分析器只取出敘述句型的基本要素(S V O)，去掉了修飾的部分，如圖十所示。



圖十、完整年譜和精簡年譜的結果。並以1948年的事件為例展示產生精簡年譜的手法。

不夠完美的地方在於，某些年代事件可能是環境或者其他人物的事件，跟傳主沒有直接的關係，如圖十一。事後我們發現，在台北市人物志中，若非以傳主作為主體的年代事件，則通常會在年號前面有關於此年代事件主體的信息，或可利用此特性，過濾掉跟傳主非直接相關的年代事件。另外精簡版的年代表雖然在依存關係分析器精準時，能產出精簡的有效敘述，但是對分析器的表現不佳時非常敏感，會產出難以看出意義的敘述。



圖十一、跟傳主無直接關係但被抽出的年代事件。

雖然我們已經證明了自動產生年譜的潛力，並開創不一樣的年譜，但似乎離真正能實用仍差那麼一步。

7. 結論

本研究探索了自然語言處理技術對傳記分析的幫助，我們先說明了如何合併工具結果和以正則表達式擷取人名，並以幾項規則作過濾，且證明能得到有效的成果，也說明了此手法不足之處和可能解法。隨後我們提出兩項擷取人物關係的輔助手法，也各說明其難處，並提出可能的改善法。之後提出全新的計算方法來改善傳統的共現關

係，並展示可藉此找到沒有明說或前階段沒找到的人物關係，和幫助研究者濾出較重要的共現關係。最後嘗試仍很少人做的自動年譜產生，精簡年譜產生，並說明其難點和接下來可能的改善方法。

自動分析固然還比不上人工分析的深度，但是在廣度和節省時間心力的部份有很大的幫助，另外重要的一點是能夠發現在人工分析難以找到的一些啟發。

本研究提出一些新的方法、方向和觀點，利用自然語言處理技術幫助研究傳記(或可推廣至其他類型)，協助數位人文開拓新的領域和方法，同時也介紹了各種難點，為數位人文後面的研究者提示了研究的方向。

8. 特別致謝

本研究工作起始於政治大學「文本分析與數位人文」課程，在學期之中，感謝班級同學虞夢夏(北京大學中國語言文學系應用語言學專業)、李冠霖(政治大學心理系)的共同討論與資源分享有助於我們的研究工作。本研究另外承蒙科技部研究計劃MOST-104-2200-E-004-005-MY3 與政治大學高教深耕拔尖計畫107H121-08 的部分補助，特此致謝。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

A Yorkshire Tragedy

Using Agent-Based Modeling to Suggest Authorship

Brian Kokensparger

Asst. Prof. of Computing

Creighton University

Omaha, Nebraska, U.S.A.

A Yorkshire Tragedy: Using Agent-Based Modeling to Suggest Authorship

Brian Kokensparger
Asst. Prof. of Computing
Creighton University
Omaha, Nebraska, U.S.A.
bkoken@creighton.edu
Telephone: 402-280-2878

Abstract

In this paper, *A Yorkshire Tragedy* (AYT) is used as a focus for development of an agent-based modeling application, using a Java swarm-type modeling library called Mason. This paper utilizes data generated by stylometric algorithms within the modeling application to visualize interactions and affinities, employing plays, playwrights, and theaters all as agents against a simple backdrop of Early Modern London. During whole-play analysis (WPA), though the title page of *A Yorkshire Tragedy* attributes authorship to Shakespeare, and first performance to the Globe, this model suggests that it certainly was not written by Shakespeare. Scene analysis (SA) of the play, employing scenes as independent agents, suggests that none of the scenes appear to be written by Shakespeare, and suggest that the primary author may have been William Davenant, but considering that Davenant was reputed to have been born the year AYT was first produced, results are inconclusive. Further analysis with different stylometric dimensions may provide more clarity in answering this question.

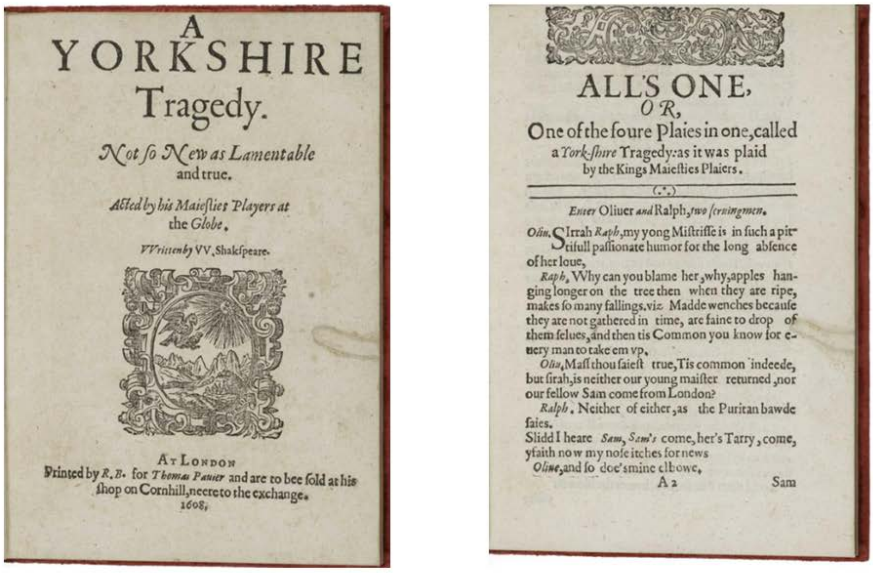
Introduction

Stylometry has been employed historically to answer questions of author attribution (Craig & Kinney, 2010). The history of stylometry predates the computer but blossomed under the advent of the ubiquity of computers with famous stylometry cases, such as Mathew Jockers' work with the Book of Mormon, and a list of other notable cases (Juola 2006).

Since Early Modern English drama is known for a great deal of collaboration among its playwrights, often the author listed on the title page is not the only playwright, and sometimes has not contributed to the play's writing at all. Historically, stylometry, as an analytical technique, has been applied to the plays, in an attempt to determine authorship. This has been difficult due to these problem areas:

- It is difficult to determine which parts of any given play were written by a given playwright
- It is difficult to combine all of the "forces" involved with suggesting which playwright, or playwrights, likely contributed to the play script, as different versions may feature the work of playwrights after the original production of the play
- There has been little done so far regarding the theater manager's input into which plays were performed in the space, and which playwrights were hired to create them.

A Digital Anthology of Early Modern English Drama, hosted by the Folger Shakespeare Library (2018), provides not only access to Early Modern English Drama scripts, but also to a database of records regarding those scripts. This is the first time that the complete records referencing a great number of the plays, including the latest scholarship, have been successfully brought together all in one place.



A
YORKSHIRE
Tragedy.

Not so New as Lamentable
and true.

Acted by his Maiesties Players at
the *Globe*.

Written by VV. Shakspeare.

AT LONDON
Printed by *R. B.* for *Thomas Pauier* and are to bee sold at his shop on Cornhill, neere to the exchange.
1608.

ALL'S ONE,
OR,
One of the foure Plaies in one, called
a *York-shire* Tragedy: as it was plaid
by the Kings Maiesties Plaiers.

Enter Oliuer *and* Ralph, *two seruingmen.*

Oliu. Sirrah *Raph*, my yong Mistriſſe is in such a pitifull passionate humor for the long absence of her loue,

Raph, Why can you blame her, why, apples hanging longer on the tree then when they are ripe, makes so many fallings, viz. Madde wenches because they are not gathered in time, are faine to drop of them selues, and then tis Common you know for euery man to take em vp.

Oliu, Mass thou saiest true, Tis common indeede, but sirah, is neither our young maister returned, nor our fellow Sam come from London?

Ralph, Neither of either, as the Puritan bawde saies.
Slidd I heare *Sam*, *Sam's* come, her's Tarry, come, yfaith now my nose itches for news

Oliu, and so doe's mine elbowe,

A 2 Sam

Figure 1: Title and first page from the play, *A Yorkshire Tragedy*, from the Folger Shakespeare Library's A Digital Anthology of Early Modern English Drama, directly accessible through the link: https://www.folger.edu/the-yorkshire-tragedy#page/front+endleaf+2v/mode/2up

One of the big questions regarding collaboration and attributed authorship in Early Modern English drama is the case of the play, *A Yorkshire Tragedy*. This play, attributed to William Shakespeare as the author on the title page (see Fig. 1), is clearly not the work of William Shakespeare. Several critical sources have indicated this, including Sykes (1917), who makes a case for George Wilkins as the author. This claim, as well as claims that Heywood was the actual author, were laid to rest by Logan & Smith (1965).

More recent scholarship has indicated that Middleton is most likely the author of the play (Lake, 1975), supported with computer analysis evidence more recently by Hartmut Ilsemann (2018). So this all appears to have been resolved in the 1970s, and validated shortly thereafter. Why delve into it again?

Because we did not have ubiquitous modeling software in the 1970s, and we also did not have EMED. The combination of this rich resource (EMED) with this hardy tool (agent-based modeling) allows us to re-open this investigation in a way that we have never been able to do before.

The purpose of this study is to use the wealth of information available in EMED and to build an agent-based model that is driven by that information, to validate and – if necessary – refute the information in current Early Modern English Drama scholarship. As just one small portion of this overreaching goal, we focus on modeling *A Yorkshire Tragedy* to validate or refute current scholarly thought in terms of its attribution.

Using this problem as a focus for both the development and validation of this modeling application, this paper is the record of early attempts by the author to bring together the script itself, as well as the records related to the script, in a model that allows the play to gravitate both towards the theaters where (as stylometric analysis suggests) it is most likely to “feel at home,”

as well as to attract playwrights who (again, as stylometric analysis suggests) have the highest level of affinity to both the writing styles of plays produced by a specific theater AND the writing style involved in the play itself.

Methodology

The first step was to develop software to stylometrically analyze the corpus of plays in various groupings, based on venue (theater), playwright, and data recorded in and about the play itself. We developed a Python script using the NLTK (Natural Language Toolkit) library to do an initial stylometric analysis of the corpora, employing simple word frequencies (with no stop words) using Burrow's Delta to find corpora stylistically closest to the given individual play. Francois Laramee's (2018) excellent tutorial provided some of the base code for this analysis which, like the tutorial, used a single dimension in its analysis (i.e., a frequency distribution of the 30 most common words in the overall corpus, not eliminating stop words). This study also used the NLTK for easy application of sophisticated algorithms and other stylometric methods.

A number of XML files for Early Modern English plays were downloaded from the EMED website, and we developed a script to transform the spoken lines only of the XML files into text files (a file of "given" text, and a file of "lemma" text). Analysis was performed on the "lemma" versions of these plays (which substitute standardized English words when possible and appropriate). Altogether, 125 plays were downloaded, transformed, cleaned, and initially grouped for analysis by venue and by playwright.

Of the 125 plays gathered, 82 listed a theater of first production on their title page, so they were grouped according to venue (theater) of first production. 6 venues were identified as having a sufficient number of noted performances among the title pages of plays downloaded during the time period of focus. The venues and number of plays are included in Table 1.

Table 1. Venues (theaters) and number of plays analyzed for each theater

Venue (Theater)	Number of Plays Analyzed
Blackfriars	29
Cockpit (aka Phoenix)	19
Globe	14
Redbull	8
Salisbury Court	8
Whitefriars	4

The number of plays within each venue's collection of plays is normalized during analysis by averaging over the entire corpus, and then over each venue's corpus, to reduce inconsistencies and false positives introduced by the theaters with smaller numbers of downloaded plays.

After stylometric analysis was performed for *A Yorkshire Tragedy* against the plays from these venues, a set of delta values by venue were generated. The delta value represented the average difference between the target text (in this case, the whole play and scenes from *A Yorkshire Tragedy*), and the average of the collected texts in each grouping (like an individual theater). The results are reported in the Findings section of this paper.

After venue analysis was completed, additional analysis was done with plays grouped by playwright. 122 plays were grouped in folders for a total of 20 playwrights. The playwrights' works analyzed and number of plays in their subcorpora are provided in Table 2. After stylometric analysis was performed for *A Yorkshire Tragedy* against the plays attributed by current criticism to these playwrights, a set of delta values of the play and its scenes by playwright was generated. The results are also reported in the Findings section of this paper.

Table 2: Playwrights and number of plays analyzed for each playwright (plays recorded as collaborations on the title page were included in both playwrights' groups of plays)

Playwright	Number of Plays Analyzed
Beaumont	6
Brome	4
Chapman	9
Davenant	3
Dekker	5
Field	3
Fletcher	8
Ford	5
Heywood	7
Jonson	7
Marmion	2
Marston	4
Massinger	12
Middleton	18
Nabbes	2
Rowley	8
Shakespeare	8
Shirley	6
Webster	2
Wilkins	3

After performing these stylometric analyses, the results were used to do agent-based modeling. The Java swarm-type modeling library called Mason was used to visualize these interactions and affinities, employing plays, playwrights, and theaters all as agents against a simple backdrop of Early Modern London. A still screenshot of the model is provided below, to illustrate the model's usefulness in affirming and questioning some of the theories regarding collaboration among playwrights in Early Modern English Drama.

Whole Play Analysis (WPA)

After producing the delta values for the whole play of *A Yorkshire Tragedy*, against plays attributed to be performed at specific venues and written by specific playwrights, the agent-based

modeling application was applied to the whole play, and the model was run several times with randomized initial placements of the play. Those findings are included below.

Scene Analysis (SA)

However, it is widely known among Early Modern English Drama scholars that playwrights of the era in which *A Yorkshire Tragedy* was produced often collaborated on plays, so identifying a single author of the whole play may not yield the most accurate results. Therefore, it was decided to apply these methods to scene analysis, where *A Yorkshire Tragedy* was broken down into scenes, and then each scene was analyzed against the entire corpus to produce a delta value, with all scenes then modeled together in the agent-based model application. It was hoped that the scenes would reveal their authors, and that some pattern of authorship may emerge from the effort, which could provide a much more realistic view of authorship for the play.

In the Folger Anthology of Early Modern English Drama edition of *A Yorkshire Tragedy*, there were no scenes delineated in the published version. However, a later scholarly edition of the play (Shakespeare 1918), scanned and made available via Project Gutenberg, identifies 10 scenes for the play. The decision was made to adopt these scene breaks for this project. The play with the individual scenes is available for everyone to view at the Project Gutenberg website (Project Gutenberg 2018).

Therefore, a class diagram was produced to show the objects included in the agent-based modeling application. This class diagram is provided in Figure 2.

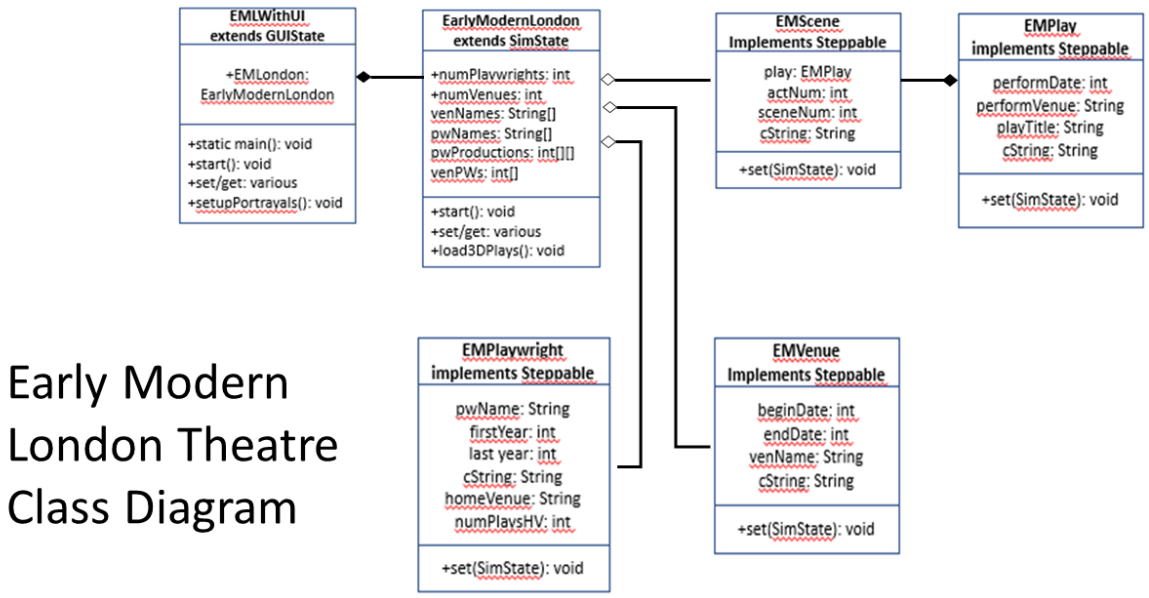


Figure 2: Class Diagram for the agent-based modeling code for the AYT project.

Classes of objects are defined in object-oriented programs (OOP) to allow each object to contain its own data relevant to the individual object being modeled, as well as the methods, or behaviors and tasks that each object is able to do. The class diagram reveals two classes that are required for the agent-based modeling application to run (EarlyModernLondon, which provides the field upon which all of the agents act, and EMLWithUI, which provides the visualization of this field so that users can view the interactions). The other four classes are individual object types in the model, including EMPlaywright, EMVenue (the theaters existing and producing plays at the time), and EMScene (which is the unit of analysis in this application, i.e., a collection of 10 objects of type EMScene that make up an EMPlay). Note that in whole play analysis (WPA), the EMPlay object is the agent in the field, whereas in scene analysis (SA) the EMScene objects are the agents in the field. The EarlyModernLondon object creates and manages all of the objects and agent objects in the model, including several EMPlaywright objects (such as Shakespeare or Rowley), several EMVenue objects (such as CockPit or Globe),

and either one EMPlay object (like AYorkshireTragedy) or a number of EMScene objects (like YTS1 or YTS7).

Employing scene analysis, some additional findings were produced, which vary somewhat from the findings produced by whole play analysis. Both are provided below.

Findings

A number of findings were produced using both whole play analysis (WPA) and scene analysis (SA). As reported in the Methodology section above, very few changes needed to be made between WPA and SA. Some additional stylometric analysis needed to be done to provide delta values for all of the agent objects provided either as the whole play (AYorkshireTragedy) or each scene in the play (YTS1 through YTS10). Then the agent-based modeling application needed to be updated to employ 10 EMScene objects instead of one EMPlay object. Otherwise, the model code remained the same during both phases of analysis.

Whole Play Analysis

Figure 3 shows the results of the stylometric analysis performed on spoken scripts of plays grouped by theaters that the plays had listed on their title pages. Using Burrow's Delta, which reduces the effect of unique words that are used in great frequencies within a specific text by utilizing z-scores, the lowest value displayed shows the venue with plays that have the least amount of distance from the target play (in this case *A Yorkshire Tragedy*). In other words, the lower values show the higher similarity between the target text and that grouping of known texts. Using the frequency distributions of merely the 30 most common words, it is clear in Figure 3 that the plays performed at the Cockpit (aka the Phoenix theater) were least different from *A Yorkshire Tragedy*. Though these results are not conclusive, they definitely call into question the

play's title page, which claims its first performance at the Globe theater. If Shakespeare as the author is not true, which current scholarship contends, then this study suggests it also may not be true that the play was first performed at the Globe theater. However, a keen eye might notice that *A Yorkshire Tragedy* was first reportedly produced before 1608 (the publication date of the play), and the Cockpit as a theater did not exist until about 1612. So how can a play stylometrically have an affinity for plays produced at a theater that did not exist until 6 years after it was first performed? This is a good question, but as style does not typically change a great deal for a specific playwright, it could be that the style of writing that gave birth to *A Yorkshire Tragedy* is similar to the style of writing used for later plays written and performed at the Cockpit.

```
Delta score for venue Blackfriars is 2.451721017586747
Delta score for venue Cockpit is 2.301749589006921
Delta score for venue Globe is 2.7885745886055853
Delta score for venue Redbull is 2.419228902262909
Delta score for venue SalisburyCourt is 2.802031278192 56
Delta score for venue Whitefriars is 2.5044093708172537
Zscore feature average for Unlisted play is 0
>>>
```

Figure 3: Results of stylometric analysis for plays grouped by title page records of venue of first performance, against *A Yorkshire Tragedy*. Note the smaller value means a smaller delta, or a smaller difference between the target play and the subcorpus of a specific theater's productions

Figure 4 shows the results of the stylometric analysis performed on spoken scripts of plays grouped by playwrights as attributed by current scholarship (using the EMED database as arbiter of disputes). Using Burrow's Delta on the frequency distributions of the 30 most common words, it is clear in Figure 4 that the plays attributed to William Rowley (1.14) were least different from *A Yorkshire Tragedy*. Middleton (1.19) and Dekker (1.19) are close seconds. Interesting enough, in the records of first productions, where both the first production venue and playwright are indicated, Rowley and Dekker both had most of their performances in the Cockpit

theater. Middleton, though listed as having his most performances at the Blackfriars theater, collaborated with Rowley on many of his later plays. It is conceivable that stylometric analysis points the play *A Yorkshire Tragedy* to Thomas Middleton, but perhaps this analysis draws its strength from Rowley's input. Though these results are not conclusive, they also definitely call into question the play's title page, considering that Shakespeare's plays have the highest delta value, or the most distance from *A Yorkshire Tragedy*. If Shakespeare as the author is not true, which current scholarship supports (and this analysis validates), then this study throws Rowley into the mix as a possible true author of *A Yorkshire Tragedy*.

```
Delta score for playwright Beaumont is 1.411921030937791
Delta score for playwright Brome is 1.5738402504189495
Delta score for playwright Chapman is 1.4162479258538936
Delta score for playwright Davenant is 1.249233422741804
Delta score for playwright Dekker is 1.1855480596923018
Delta score for playwright Field is 1.4401628833879614
Delta score for playwright Fletcher is 1.3407036849921996
Delta score for playwright Ford is 1.3113327419136516
Delta score for playwright Heywood is 1.506869379505154
Delta score for playwright Jonson is 1.4087088945973707
Delta score for playwright Marmion is 1.6340246684084612
Delta score for playwright Marston is 1.3493815478266702
Delta score for playwright Massinger is 1.5801154146403824
Delta score for playwright Middleton is 1.1893538086389988
Delta score for playwright Nabbes is 1.5720507503154093
Delta score for playwright Rowley is 1.1400910788075784
Delta score for playwright Shakespeare is 1.8380135369046866
Delta score for playwright Shirley is 1.5536246813339845
Delta score for playwright Webster is 1.3650673148608985
Delta score for playwright Wilkins is 1.8045751537752905
Zscore feature average for Unlisted play is 0
```

Figure 4: Stylometric analysis results for plays grouped by playwright, against the frequency distribution for *A Yorkshire Tragedy*. Note the smaller value means a smaller delta, or a smaller difference between the target play and the subcorpus of a specific playwright's attributed plays

Figure 5 shows a screen shot from the Mason agent-based model showing the result of movement from originally random placements of the play, *A Yorkshire Tragedy*, towards objects which, according to stylometric analysis, produced or wrote plays that were most similar to it.

The model basically takes the stylometric data provided, in the venue and playwright analyses, and uses the data as forces to step the play agent towards or away from the other agents.

When the model begins to run, the play agent is randomly placed somewhere among the theaters and playwrights. With each time interval (or step) of the model, the play agent is moved according to the various forces acting upon it. In the case of *A Yorkshire Tragedy* as a play agent, it moves towards the Cockpit theater, and towards Rowley, Middleton, and Dekker. Since Rowley and Dekker are more closely associated with the Cockpit theater, the weight of all three of those agents exceeds the weight of the Middleton force, and the play agent hovers around the Cockpit theater.

For whole play analysis, the agent-based modeling application is used more as a visualization tool than an analysis tool, so it does not reveal anything startling or different from what the NLTK stylometric analyses have already revealed. But we will see, in scene analysis, that the model itself allows novel analysis of individual scenes against playwrights and venues in Early Modern English Drama.

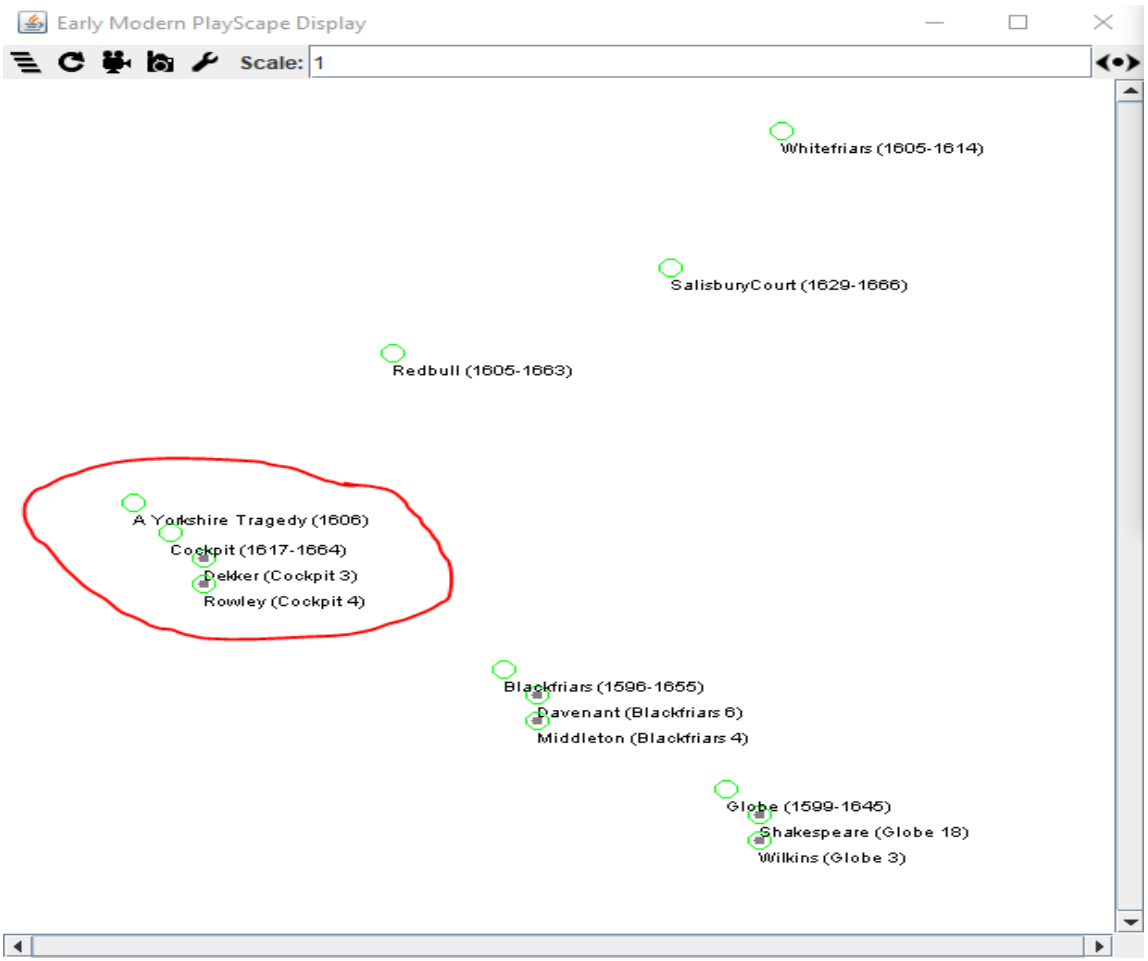


Figure 5: Example Mason agent-based model showing the result of movement from originally random placements of the play, *A Yorkshire Tragedy*, towards agents which, according to stylometric analysis, produced or wrote plays that were most similar to it.

Scene Analysis

Whole play analysis provides one lens through which to consider authorship for *A Yorkshire Tragedy*. But Early Modern English drama scholarship indicates that playwrights in the early 17th century often collaborated on scripts (Coughlan 2012) and that sometimes specific scenes in a play attributed to a specific playwright were sometimes written by different playwrights, such as the witch scene in Shakespeare's Macbeth often being attributed to Middleton (Taylor & Lavagnino 2007).

Therefore, scene analysis was also done for this study, producing delta values for the same 30 most frequent words in the corpus as was done for whole play analysis. For scene analysis, the same playwrights and venues were featured, so that a direct comparison could be made for each scene against the whole play itself.

Table 3 shows the stylometric delta values for the six playwrights of focus for each scene of *A Yorkshire Tragedy*. Different scenes produced different ranges of values, but all ranges consisted of tenths of a point differences between two integer values. For example, YTS2 delta values ranged from 1.30 to 2.02, while YTS6 ranged from 6.40 to 6.92. This is not a result of strength or significance of results, but primarily a result of the relative mathematical relationships between the corpus averages and the total word count provided within individual scenes. Therefore, it made sense to normalize all of the deltas to values ranging from 1.05 (Middleton for YTS9) to 2.09 (Wilkins for YTS8). This would allow the agent-based modeling application to give each scene equal weight in its behaviors within the model.

Table 3: Normalized scene deltas by playwright for *A Yorkshire Tragedy*. Yellow highlights indicate lowest delta (most similarity) among selected playwrights for each scene, whereas blue highlights indicate the highest delta (least similarity).

	Davenant	Dekker	Middleton	Rowley	Shakespeare	Wilkins
YTS1	1.39	1.37	1.37	1.53	1.64	1.80
YTS2	1.70	1.41	1.30	1.33	2.02	1.83
YTS3	1.07	1.36	1.26	1.27	1.58	1.31
YTS4	1.12	1.36	1.30	1.17	1.56	1.28
YTS5	1.26	1.13	1.18	1.17	1.56	1.80
YTS6	1.56	1.88	1.66	1.52	1.92	1.68
YTS7	1.18	1.52	1.36	1.47	1.34	1.44
YTS8	1.61	1.63	1.90	1.97	1.63	2.09
YTS9	1.29	1.10	1.05	1.22	1.59	1.81
YTS10	1.32	1.23	1.35	1.26	1.40	1.51

This analysis indicates most significantly that Shakespeare and Wilkins were most likely not the authors of *A Yorkshire Tragedy*, a supposition that is supported by whole play analysis as

well. Shakespeare's lowest delta was for YTS7 (1.34), but still not nearly as low as Davenant (1.18), and relatively close in value to Middleton (1.36). It is notable that Shakespeare and Wilkins (1.44) also had lower deltas for YTS7 than Dekker (1.52). So if Shakespeare did write one of the scenes, the most likely scene would have been YTS7. But Davenant still had a significantly lower delta than Shakespeare for that scene as well.

Of secondary significance is that the scenes were stylometrically similar to those plays written by a number of playwrights, and in some cases, the deltas for second-place finishers were not too far off from first place finishers. Table 3 reveals a good mix of scenes where Davenant, Dekker, and Middleton had the lowest deltas in the focused group of playwrights. This could be construed that all three of those playwrights may have collaborated in their writing, and perhaps influenced each other stylistically to have styles similar to that of *A Yorkshire Tragedy*, but more analysis would need to be done to more fully support that thesis.

In the need to see if these results point to any one playwright as being above the others in possible authorship of *A Yorkshire Tragedy*, it is better to produce some line charts visualizing the delta values against each scene. In that way, close second-place finishers in terms of delta values will be revealed with their low delta values in a more comprehensive way. Figure 6 provides these line charts.

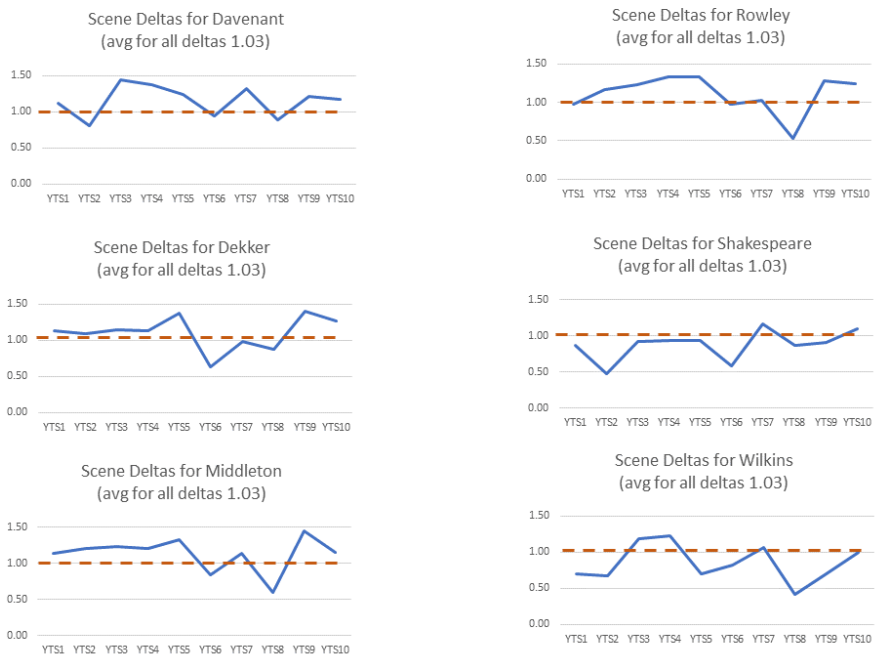


Figure 6: Line graphs of normalized and inverted scene deltas for *A Yorkshire Tragedy*

To be properly displayed as strength values, the deltas had to be inverted. This was due to the fact that the most significant values (or strongest values) were the lower deltas, but it is difficult to convey strength using lower values on a chart. Inversion was done by subtracting all delta values by 2.5. Since the highest delta value was 2.09, this produced a point on the line graph at 0.41, and the lowest delta value (1.05) produced a point on the line graph at 1.45. The overall average for all normalized and inverted delta values was 1.03, which provided a line for comparison. All points above the average line would be considered more likely for authorship, and points below the average line were considered less likely for authorship.

Viewed in this way, there is a large area in the line chart in Figure 6 above the average line for Davenant, as well as for Rowley. Davenant's three points below the average line (YTS2, YTS6, and YTS8) were not too far below. All other playwrights, including Rowley (YTS8), had points much more dramatically below the average line. This points to consideration of Davenant as the likely author of *A Yorkshire Tragedy*, perhaps in collaboration with Rowley. This,

however, is problematic, because William Davenant was reputed to have been born in 1606, the very year that *A Yorkshire Tragedy* was reported to have premiered on London stages.

It is also clear from Figure 6 that Shakespeare and Wilkins were most likely not the authors of *A Yorkshire Tragedy*, as most of their data points are below the average line. This is yet additional quantitative evidence that Shakespeare was not the author of the play, despite the title page of the printed play suggesting otherwise.

But these findings only point to the stylometric analysis of the scenes against the work attributed to the playwrights. Where do the venues come into the picture? Additional stylometric analysis was done scene-by-scene with plays attributed to the selected venues. The scenes, playwrights, and venues were all brought together into the agent-based model to view their interactions. Figure 7 shows a still shot of the model, which generally stabilized after about 100 steps, to reveal consistent findings, as shown.

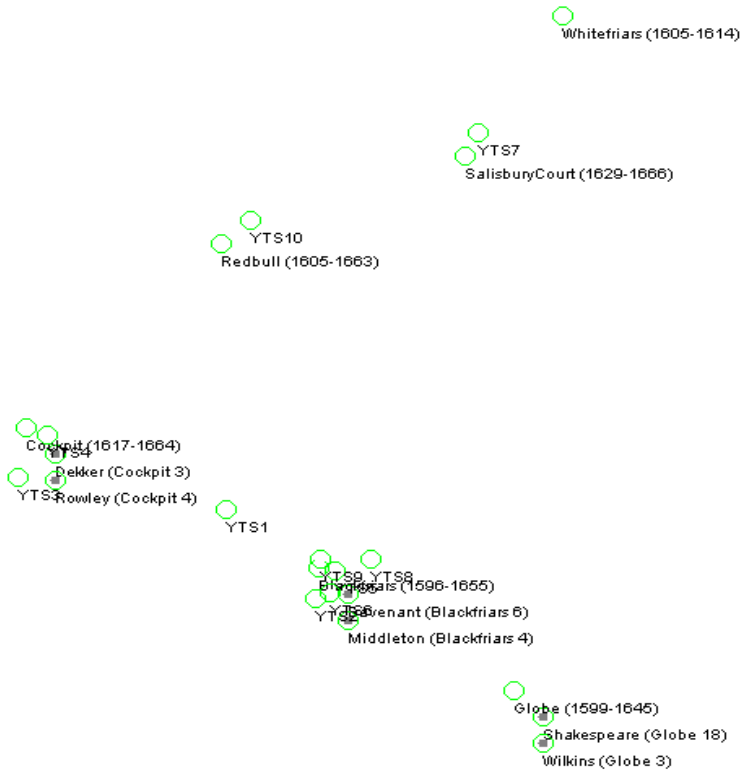


Figure 7: Agent-based model still shot of stabilized visualization of scenes as agents against a backdrop of playwrights and theaters as stationary agents

In the model, scenes YTS3 and YTS4 tended to converge on the CockPit theater, where Dekker and Rowley had their most plays produced (according to title page records, which we now understand may be suspect). Scenes YTS2, YTS5, YTS6, YTS8, and YTS9 converged on the Blackfriars theater, where Davenant and Middleton allegedly had most of their plays produced, and YTS1 hovered between the two theaters.

As outliers, YTS10 always headed for the Redbull theater, and YTS7 headed for Salisbury Court. What all of this means, is that mixing in the attributed first production venue data into the model muddies it, but still supports the supposition that Davenant has a high likelihood of being the primary author of *A Yorkshire Tragedy*, though likely with a number of collaborators. (And again, overshadowed by the problem that Davenant may have been born the

year the play was produced.)

Conclusions

Historic records say that *A Yorkshire Tragedy* was written by William Shakespeare, and had its first performance at the Globe theater. Early critics, who took on the authorship question for that play, said that it was written by Wilkins or Heywood. Current critical scholarship says that *A Yorkshire Tragedy* was written by Middleton. Where it was performed, if not the Globe, has not been conclusively considered in the popular literature.

When running stylometric algorithms on the whole play, my agent-based model suggests that *A Yorkshire Tragedy* was not written by William Shakespeare, but rather was written by William Rowley and has an affinity with other plays performed at the Cockpit theater (though the play was written before the CockPit theater was created). Since the play also has an affinity for Middleton, and Rowley and Middleton collaborated on a number of works, this could have been an early play of Rowley's, which stylistically reflects some of Middleton's later plays that featured collaboration with Rowley.

When running analysis on individual scenes of the play, my agent-based model supports existing literature that William Shakespeare did not write the entire play, and likely did not write any of its scenes as well. Stylometric analysis using the 30 most frequent words in the corpus indicates that William Davenant was most likely to have written a majority of the scenes of the play and combining that analysis with comparative analysis between playwrights and venues supports those findings, but with additional subtlety and nuance. This is in contradiction of Davenant's historically-accepted birth year of 1606. Besides asserting that William Shakespeare

did not write the play, and most likely did not write any of the individual scenes in the play, there is not enough strong and clear evidence to support additional assertions.

For future study on the *A Yorkshire Tragedy* question, we hope to do further analysis with other dimensions (beyond the simple common word frequencies we used for this paper), and perhaps to eventually model all of the existing plays independently of the information provided on their title pages to suggest true authorship and places of original performance. This would provide more information about the culture of drama in Early Modern London, and the personalities of the playwrights and theater managers, and how they all merged to provide entertainment and edification for the people of Early Modern London.

References

A Digital Library of Early Modern English Drama. Folger Shakespeare Library. https://emed.folger.edu/. Accessed 11 July 2018.

Craig, Hugh; Kinney, Arthur F., eds. (2010), *Shakespeare, Computers, and the Mystery of Authorship*, Cambridge University Press, ISBN 978-0-521-51623-5

Coughlan, Sean. (2012) *"Shakespeare's 'co-author' named by Oxford scholars"*. BBC News. Retrieved 11 October 2018.

Hope, Jonathan (1994), *The Authorship of Shakespeare's Plays: A Socio-linguistic Study*, Cambridge University Press, ISBN 978-0-521-41737-2

Shakespeare Statistics. Hartmut Ilsemann. http://www.shak-stat.engsem.uni-hannover.de/eauthoryork.html. Accessed 30 July 2018.

Taylor, G., & Lavagnino, J. (Eds.). (2007). *Thomas Middleton: The Collected Works*. OUP Oxford.

Juola, Patrick (2006). *"Authorship Attribution"* (PDF). *Foundations and Trends in Information Retrieval. **1**: 3.* doi:10.1561/1500000005.

Lake, David J. *The Canon of Thomas Middleton's Plays.* Cambridge, Cambridge University Press, 1975. pp. 163-174.

Laramée, F.D. 2018. Introduction to stylometry with Python. *The Programming Historian*. Downloaded from https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python. Accessed 11 July 2018.

Logan, Terence P., and Denzell S. Smith, eds. 1965. *The Popular School: A Survey and Bibliography of Recent Studies in English Renaissance Drama.* Lincoln, NE, University of Nebraska Press, 1965. pp. 231-232.

Love, Harold (2002), *Attributing Authorship: An Introduction*, Cambridge University Press, ISBN 978-0-521-78948-6.

Mason. George Mason University's Evolutionary Computation Laboratory. https://cs.gmu.edu/~eclab/projects/mason/. Accessed 30 July 2018.

Natural Language Toolkit. NLTK Project. https://www.nltk.org/. Accessed 30 July 2018.

A Yorkshire Tragedy. Project Gutenberg. https://www.gutenberg.org/ebooks/4255. Accessed 11 October 2018.

Shakespeare, William. 1918. *A Yorkshire Tragedy. Not So New as Lamentable and True.* In C.F. Tucker Brooke, ed., The Shakespeare Apocrypha. (Oxford).

Sykes, H.D. (1917). The authorship of "A Yorkshire Tragedy". *The Journal of English and Germanic Philology*, 16(3), pp. 437-453.

Van Droogenbroeck, Frans J. (2016) "Handling the Zipf distribution in computerized authorship attribution".

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

關鍵詞偵測方法的比較與應用

余清祥*　許承恩*　梁穎誼**
政治大學統計系*　逢甲大學風保系**

關鍵詞偵測方法的比較與應用



余清祥、許承恩：政治大學統計系

梁穎誼：逢甲大學風保系

Dec. 18, 2018

數量分析

- 透過數理模型描述觀察結果：

觀察現象 = 模型 + 誤差

或是

$y = f(x) + \text{error}$；觀察值 = 訊號 + 雜訊。

- 數量化模型的關鍵：

→量化目標值 y ：定義問題！

→選取關鍵變數： $x_1, x_2, \ldots, x_p$

→建立量化模型：統計學習、機器學習。

量化資料分析的範例

- 入學考試或申請（信用卡、文章作者）

→量化目標值：適合（0或1）就讀某科系的學生（如何量化及評估目標？）

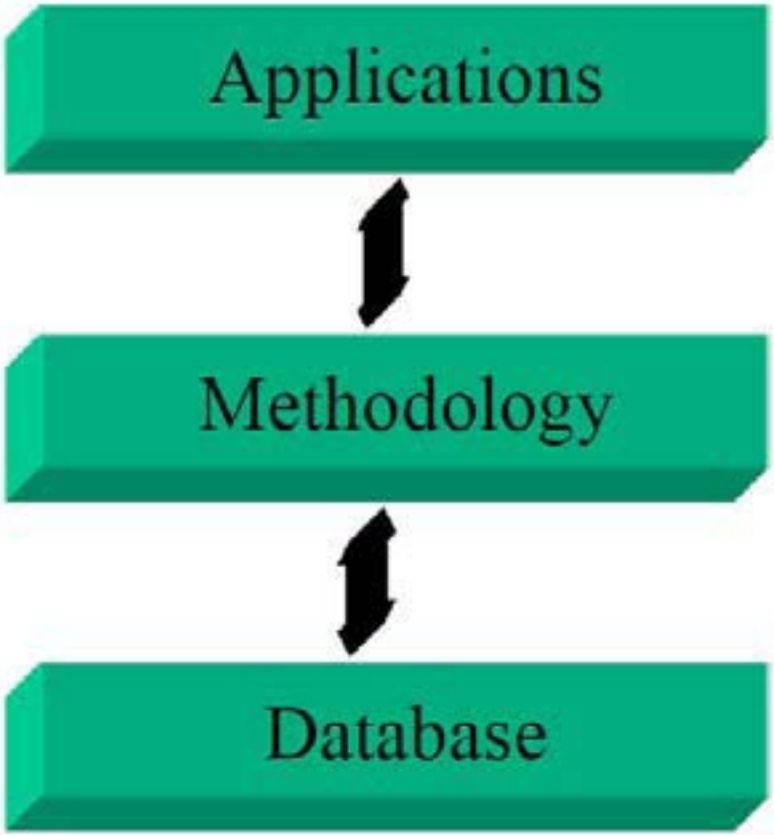
→關鍵變數：在校成績、競賽表現、課外活動等。（決定原則、範圍？）

→量化模型：羅吉士迴歸、群集分析等。（模型準確性、可用性、解釋性？）

註：近幾年不少數位人文研究聚焦於機器學習（或深度學習）等模型研究。

跨領域合作

- 透過數量化分析，篩選出人文社會研究的重要訊息及知識。（三者配合！）



文字分析的相關研究

- 趙岡與陳鍾毅(1980)

→使用統計分析判定作者用字習慣（前後各抽一百頁，每頁720字，小說全文有738,024字 ）

→考慮五個虛字的使用習慣。

	前80回	後40回	t-檢定值
兒	2.60 (2.36)	3.99 (2.95)	3.677
在	2.98 (2.08)	3.95 (1.96)	3.392
了	7.62 (5.54)	17.71 (5.48)	0.166
的	12.25 (5.00)	14.81 (5.66)	3.391
著	4.83 (3.02)	6.57 (3.27)	3.910

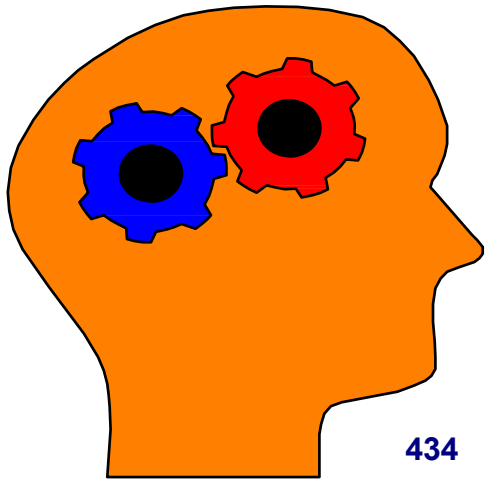
文字分析的相關研究（續）

- Mosteller and Wallace (1984)

→運用貝氏分析，探討擁護聯邦主義的論文(The Federalist Papers)作者。

→77篇中有12篇文章沒有定論

(可能是Hamilton或Madison所作)



文字分析的相關研究（續）

- Efron and Thisted (1976)

→估計莎士比亞(Shakespeare)的字彙總數；

→用Poisson Process估計字彙，估計方法與Alan Turing有關(Turing's Estimate)；

→推論1985年在莎翁故居附近發現一首詩，應是莎士比亞所作(1987)。



常見的文字分析議題

- 數位研究多半著重在描述資料基本特性（Exploratory Data Analysis；探索性資料分析），及簡化目標值的模型分析。

→不同作者用字的主要差異；

→寫作風格式否相似、作者是誰（二選一）。

- 分類(Classification)較為容易，具啟發性、開創性的想法很難由模型中發掘。

→文言文、白話文的主要差異？

基因圖譜與關鍵詞

- 關鍵詞（Keywords）猶如統計分析的變數（Variable），根據研究目的及問題定義，從原始觀察值找出適當的變數，可提高量化分析的效率和準確性。
- 關鍵字詞的研究大致可分為三個方向：

→偵測及判斷「潛在關鍵詞」；

→仿照生物資訊，把文字問題類比成基因圖譜 (如：關鍵基因組合)

→關鍵詞間的關聯（基石關鍵詞？）

研究方法及目標

- TFIDF (Term Frequency–Inverse Document Frequency)為文字探勘的常用方法，可用於決定關鍵詞。

→根據文本特性、個人經驗等，使用者挑選篩選門檻。

- 本文以《人民日報》及電腦模擬方法探討透過TFIDF挑選關鍵詞。

→討論加入其他統計方法的可能。

研究流程圖(Float Chart)

龐大字詞數量
(原文)

1971 ~ 1989年
人民日報報導

- 1. 以數位方法決定準關鍵字。
- 2. 人文學者協助去除不相關的關鍵字。
- 3. 引入生物多樣性、物種演化概念，進一步篩選出核心關鍵字詞。
- 4. … … (Iteration…)

核心關鍵字詞

TFIDF的想法及特性

- TFIDF為採納統計思維的文字探勘方法，分為兩個步驟分析：TF、TDF。

→TF（詞頻）：類似「期望值」的概念，出現次數愈多、是關鍵詞的機會愈高（至少不應太少）。

→IDF（逆向檔案頻率）：類似「變異數」，某個詞彙出現愈有規律（前後一致），可能是常用而非關鍵詞彙的機會愈大。

1979年人民日報報導，根據出現次數排序的前100名關鍵詞（節錄）：

關鍵詞	出現次數	理論次數
[1,] "人" "民"	"236"	"34.292311191992 7"
[2,] "民" "主"	"196"	"24.700864422202"
[3,] "權" "利"	"171"	"6.2657971893640 7"
[4,] "代" "表"	"154"	"4.0958447073096 8"
[5,] "主" "義"	"141"	"6.3535031847133 8"
[6,] "選" "舉"	"141"	"3.6991709635021 7"
[7,] "我" "們"	"137"	"4.1328985946820 3"
[8,] "國" "家"	"119"	"8.1783439490445 9"
[9,] "問" "題"	"118"	"1.2423415225963"
[10,] "社" "會"	"98"	"7.9433828733191 8"

TFIDF的參數挑選

- TF期望值、IDF變異數的挑選門檻，除了與文本特質有關，我們覺得也和文本每篇文張的字數，以及關鍵詞、非關鍵詞（常見詞）的特性有關。
- 分析《人民日報》 1971~1989 (19年)，與「人權」相關報導535篇報導。

→人文學者挑選出48個關鍵詞；

→變動參數數值，探討TFIDF的挑選結果。

電腦模擬與設定

- 採用交叉驗證(Cross-validation)，將關鍵詞及非關鍵詞區分為訓練組、測試組，由訓練組資料找出參數代入測試組。

→訓練組、測試組比例為3：1；

→關鍵詞、非關鍵詞等比例抽出；

→評估兩個因素的影響：

(1) 非關鍵詞的個數

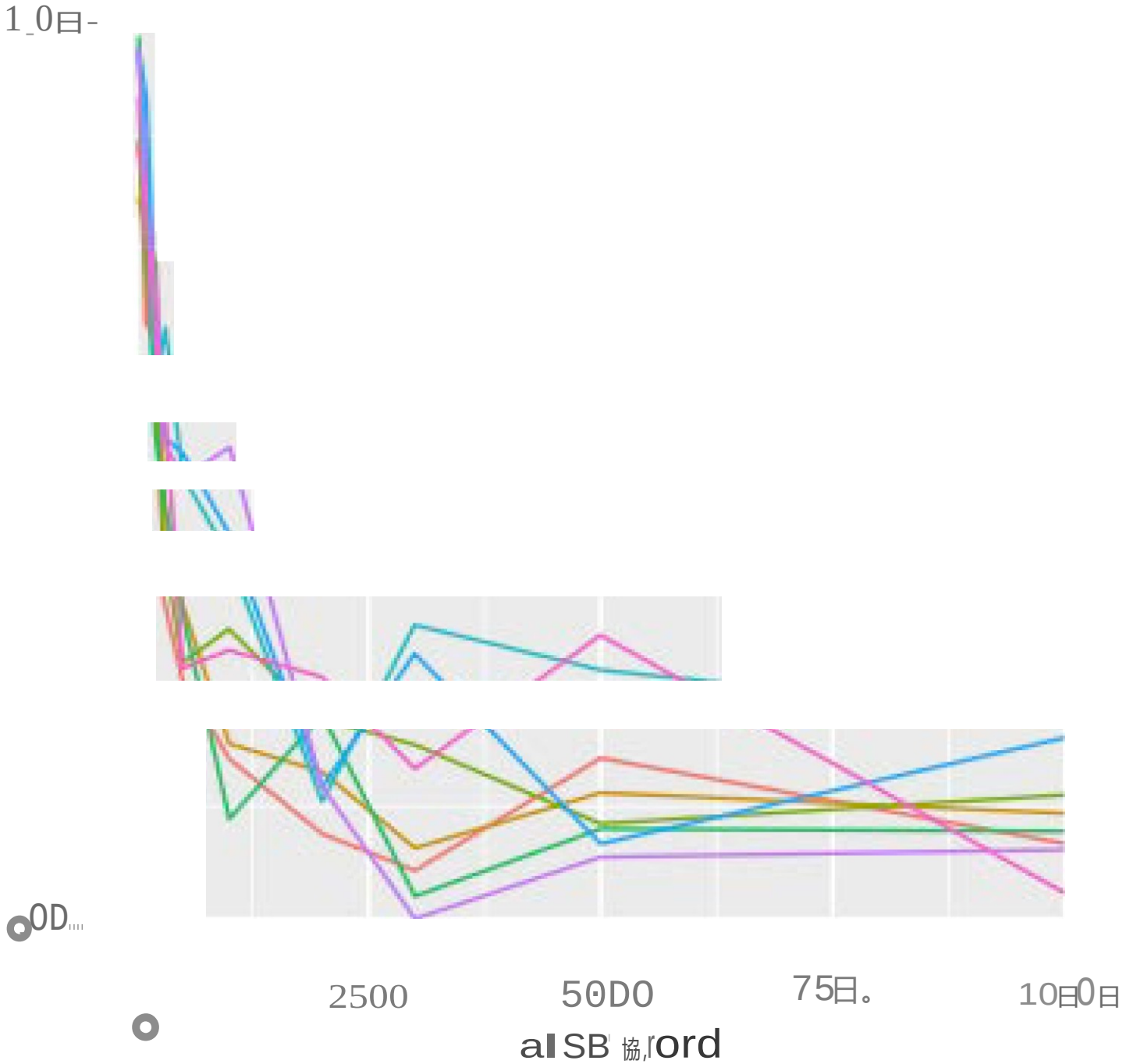
(2) 每篇文章的字數

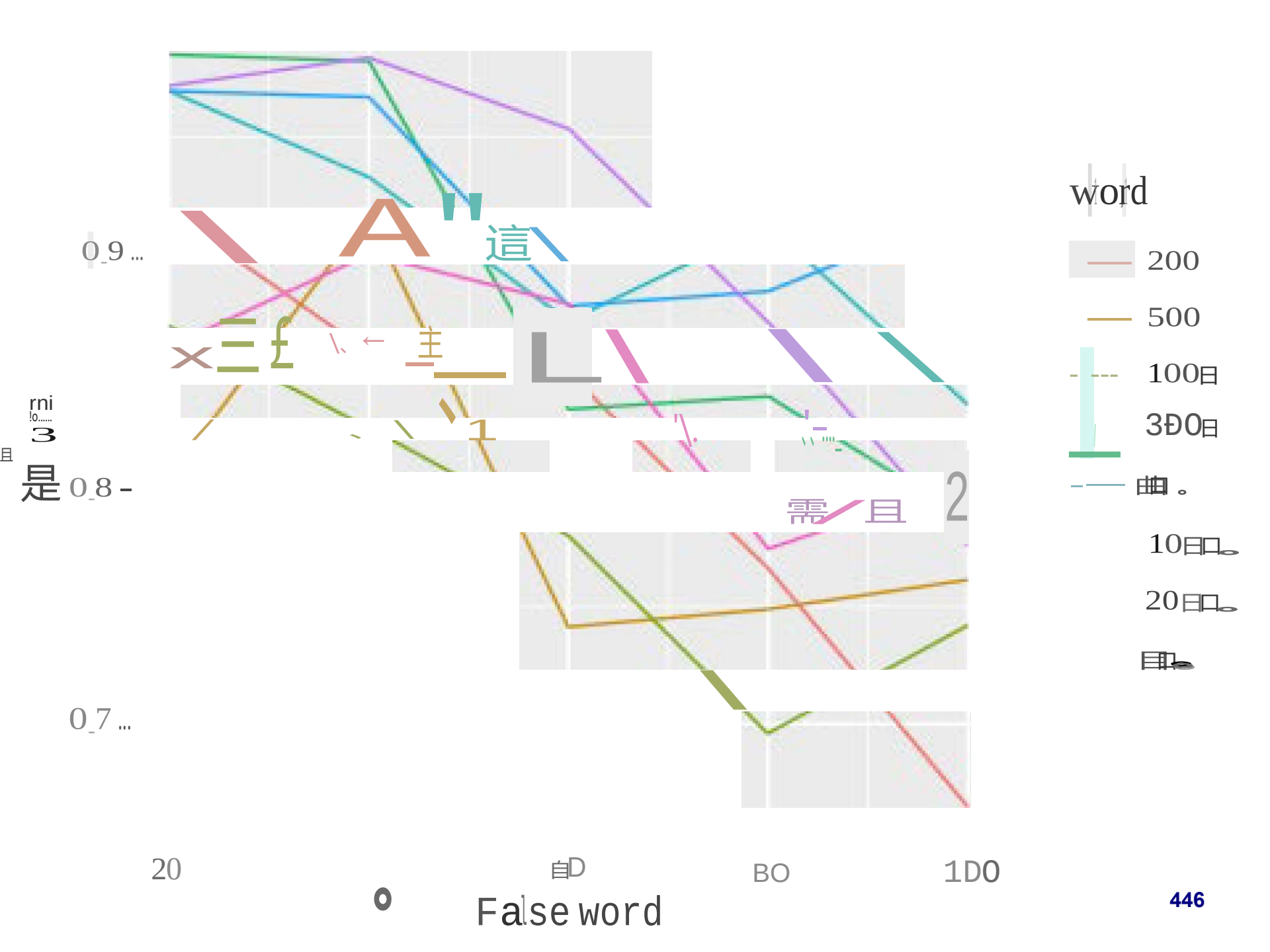
挑選關鍵詞的評估標準

- 仿造醫學檢查，以測試組的關鍵詞正確性為標準，由兩個角度評估優劣：

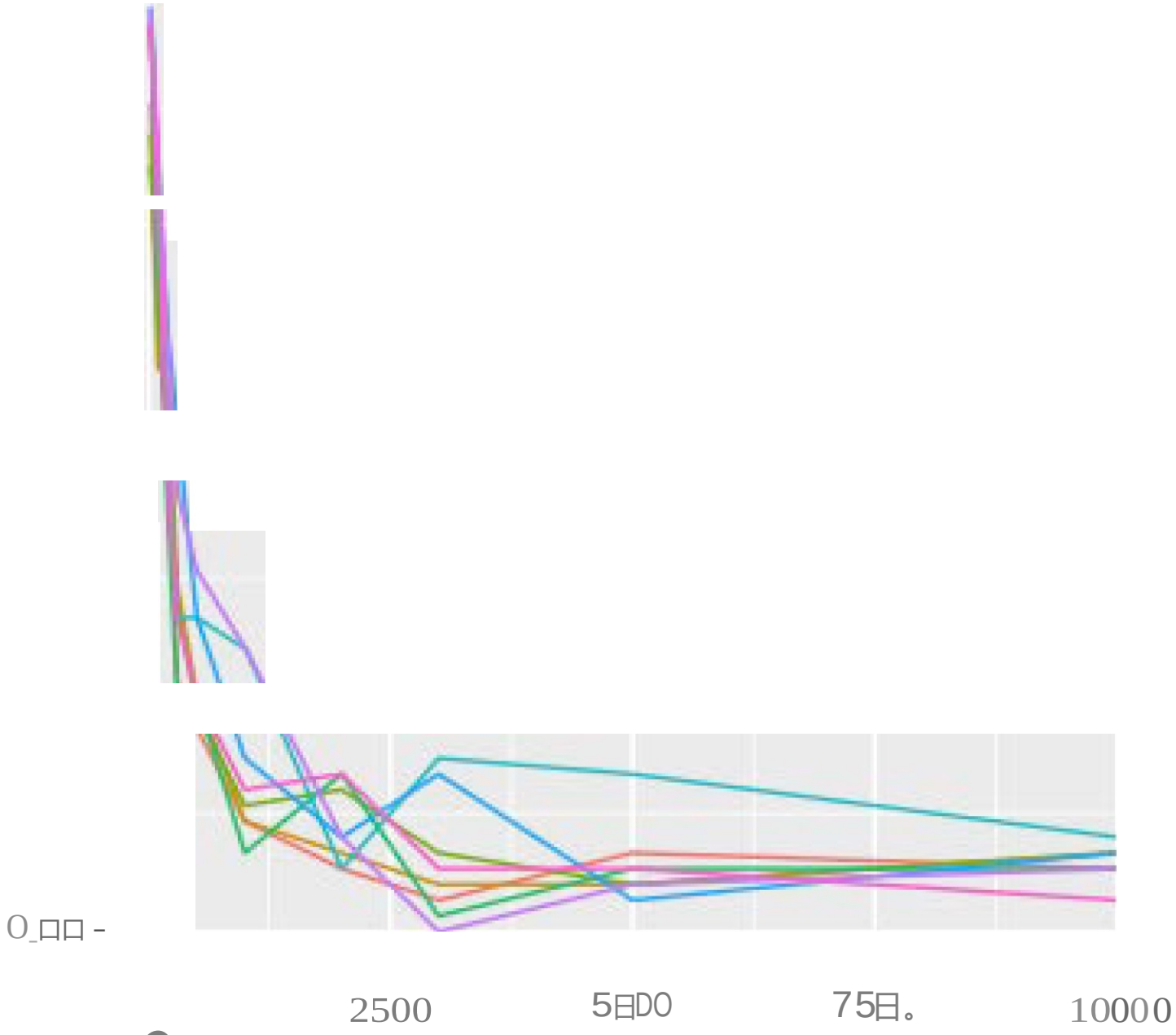
→偽陰性(False Negativity)：對應於統計的型一誤差，就是關鍵詞被忽略，有時會以準確率(Accuracy)表示。

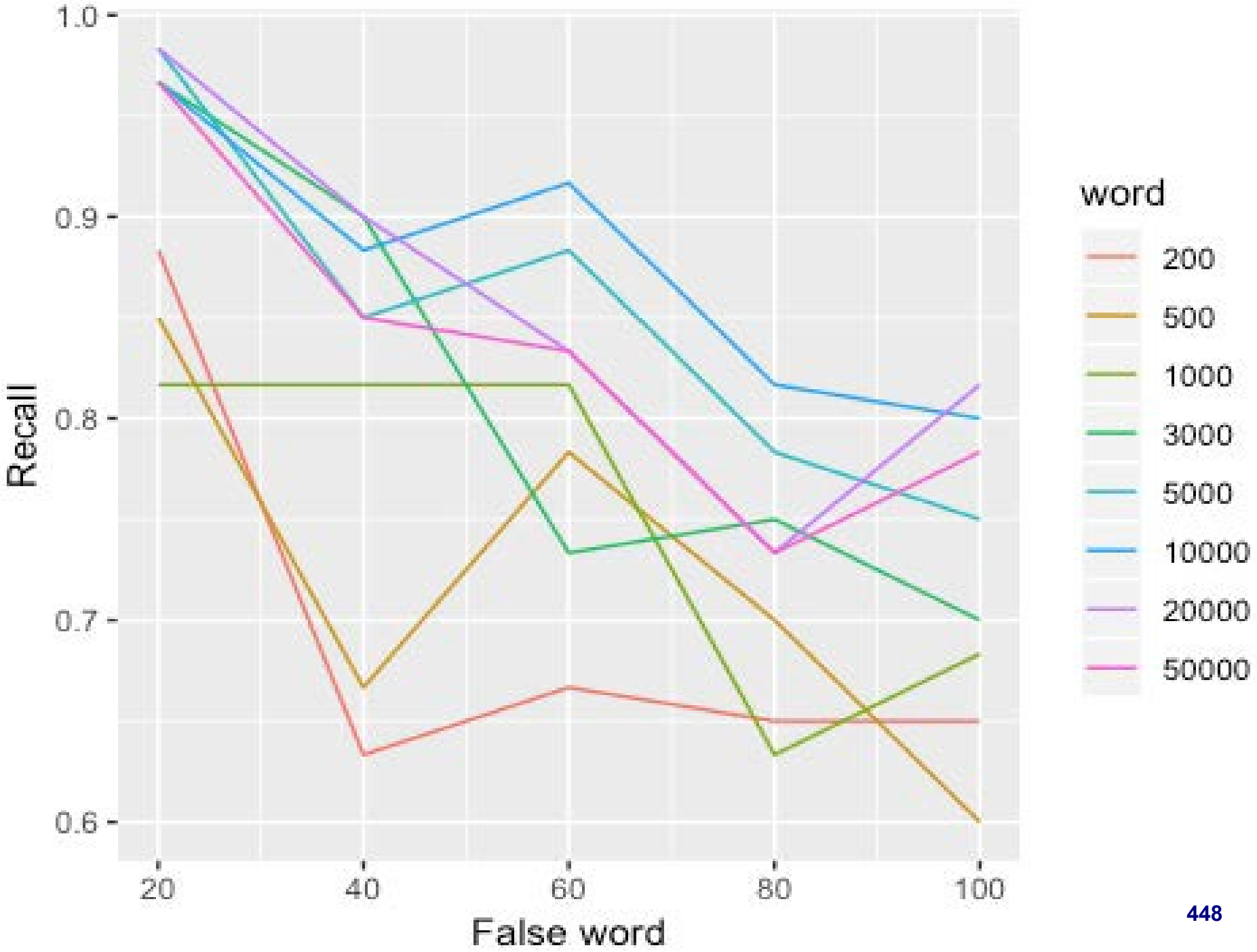
→偽陽性(False Positivity)：對應於統計的型二誤差，就是非關鍵詞被誤判，有時也以召回率(Recall)呈現。

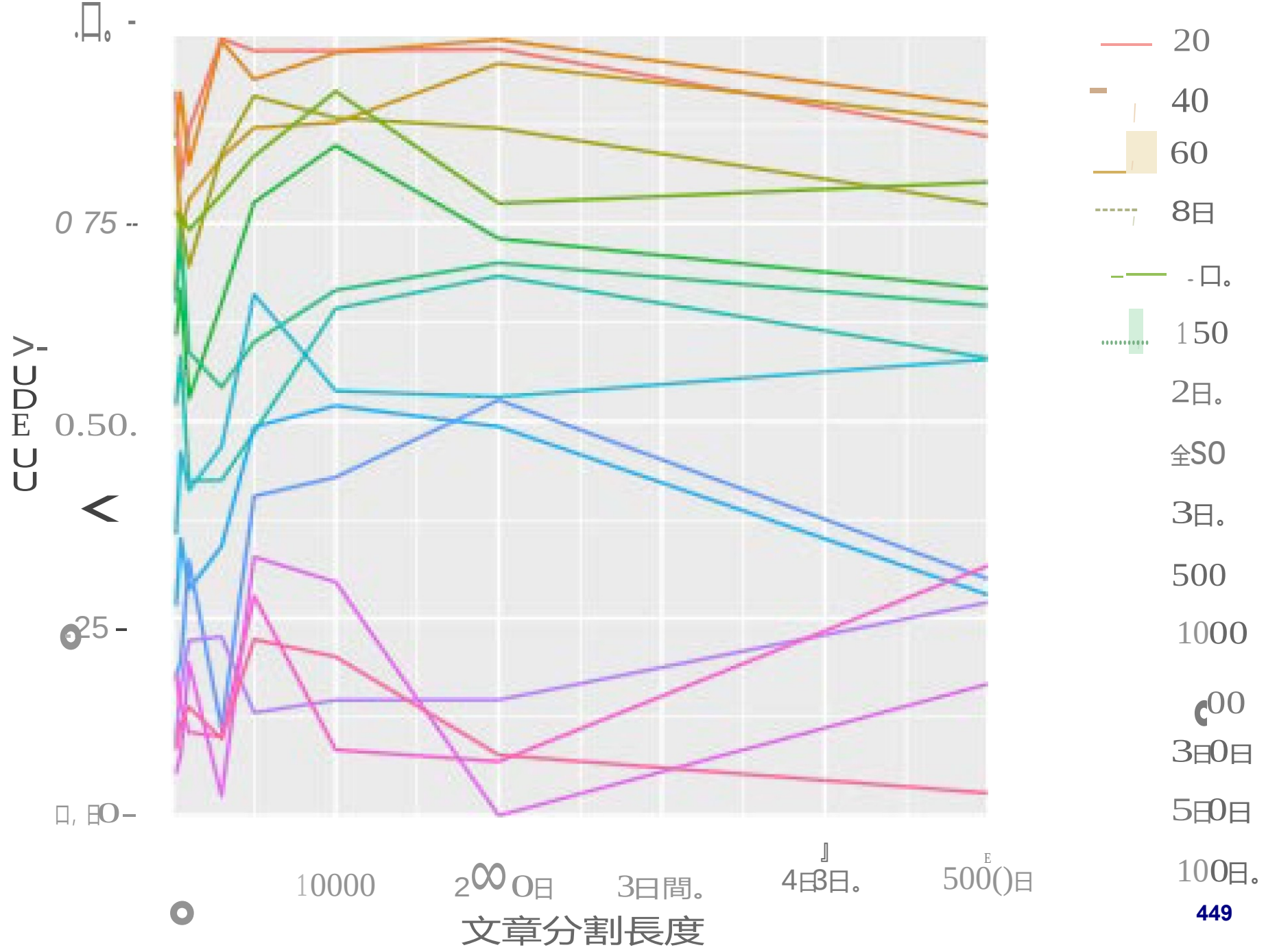

1_0日-
OD
2500
50DO
75日。
10日0日
aI SB 協,rord

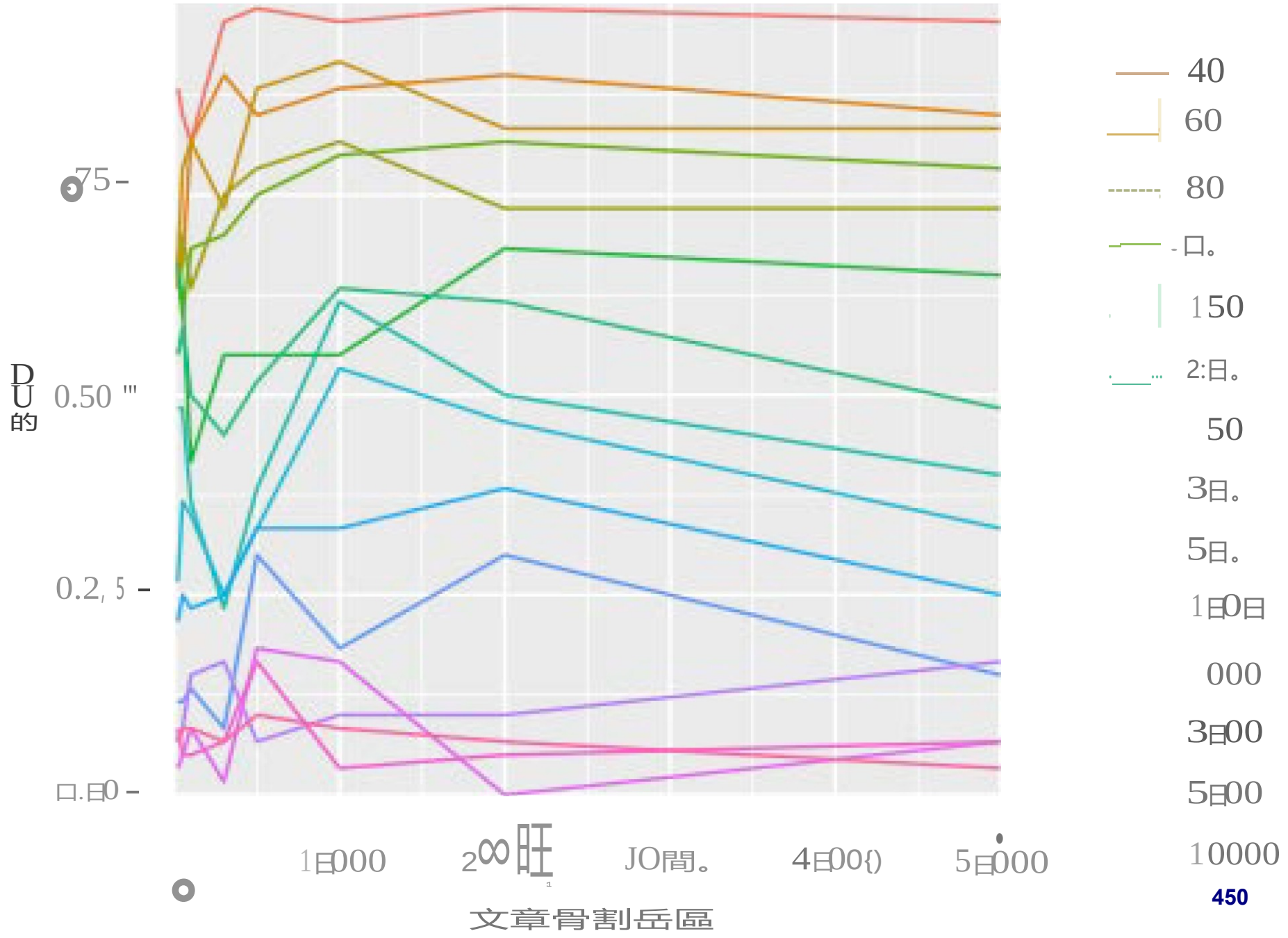


0.9
0.8
0.7
20
BO
1DO
False word
word
200
500
100日
3DO日
10日
20日


2500
5日DO
75日。
10000
alSB word


1.0
0.9
0.8
0.7
0.6
Recall
20
40
60
80
100
False word
word
200
500
1000
3000
5000
10000
20000
50000


0 75
0.50.
25
10000
3日間。
文章分割長度
20
40
60
8日
150
2日。
3日。
500
1000
3日0日
5日0日
100日。



DU的
0.50 '''
0.2, 5
口.日0
1日000
2∞旺
JO間。
4日00{)
5日000
文章骨割岳區
40
60
80
口。
150
2:日。
50
3日。
5日。
1日0日
000
3日00
5日00
10000

模擬研究小結

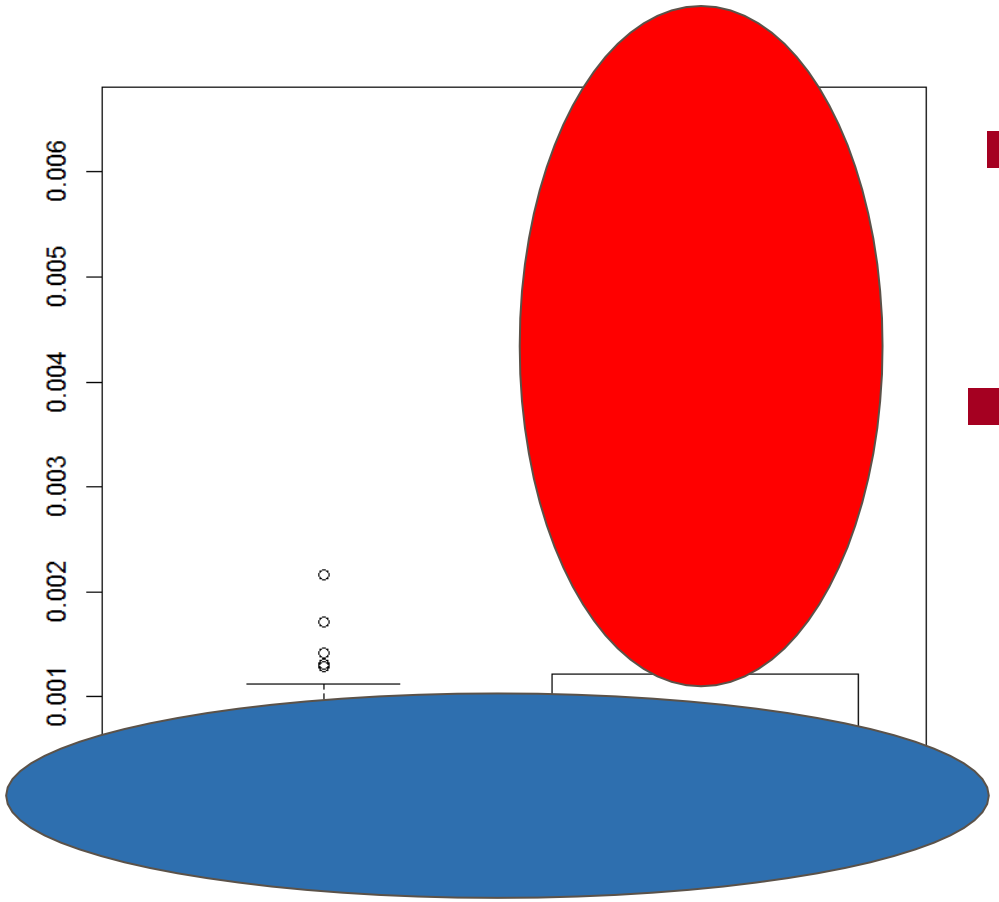
- 電腦模擬的結果顯示：

→非關鍵詞愈多、準確率和召回率愈低；

→每篇文章字數多未必會有好效果，大約3千字至兩萬字較佳。

- 關鍵詞未必有單一判斷標準，依賴由TF、IDF無法獲得很好的效果，48個關鍵詞 明顯有不同特性。

1、頻次



- 高頻基本為重要關鍵詞
- 低頻也有重要關鍵詞，但不確定性很高!

2、不均度指標

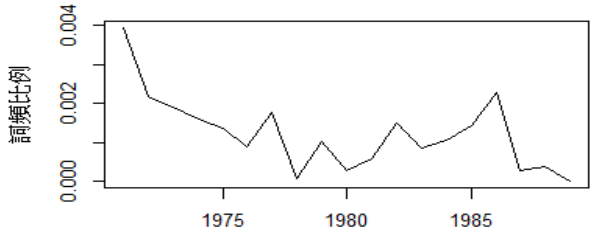
常用關鍵字				核心關鍵字			
	Simpson	卡方檢定值	KL距離		Simpson	卡方檢定值	KL距離
我們	0.0634	94.0*	0.0569	階級	0.1775	352.7*	0.354
問題	0.0702	152.2*	0.0590	資本	0.1448	136.5*	0.330
政策	0.0685	65.8*	0.0743	帝國	0.1630	355.1*	0.390
自己	0.0580	36.4*	0.0224	殖民	0.2382	380.2*	0.482
任何	0.0655	54.0*	0.0555	主義	**0.0736**	**249.4***	**0.098**
繼續	0.0692	61.4*	0.0794	柬埔	0.2345	1009.9*	0.546

分析結果:

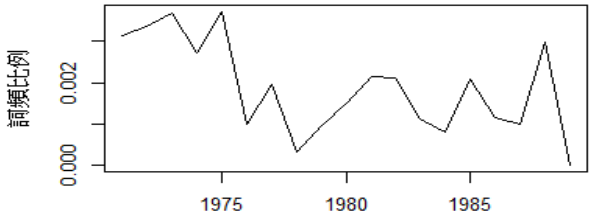
表一、常用詞與重要關鍵詞的交叉表

詞數	非常用詞	常用詞
[redacted]		
非重要關鍵詞	114	54

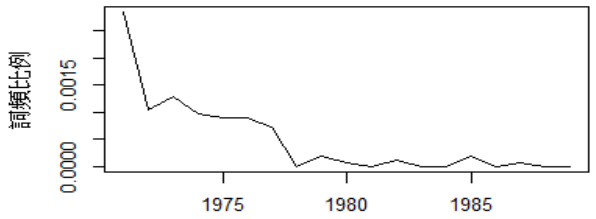
3、高頻集中



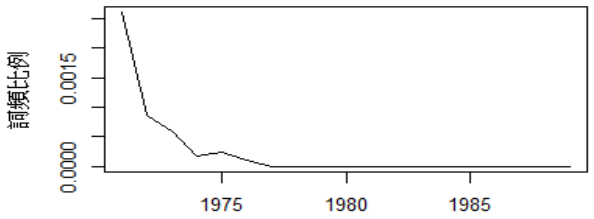
政府



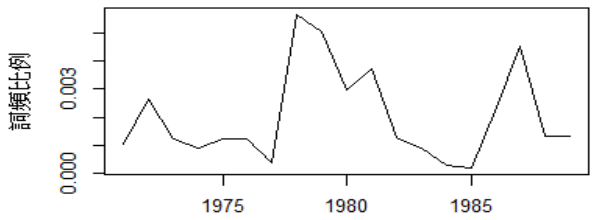
美國



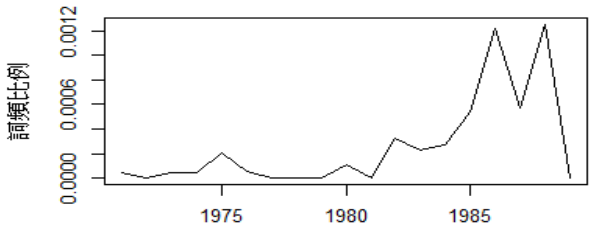
帝國



美帝



民主



改革

- 與重要概念相關的關鍵字可能具有兩個特性:

→1. 出現頻次 / 比例高。

→2. 連續不中斷的出現。

- 類似的論述在各領域都曾出現，如鞍型期理論、創新擴散理論。

- 能使用Cluster的想法進行檢定。

分析結果:

表二、高頻集中字與重要關鍵詞的交叉表

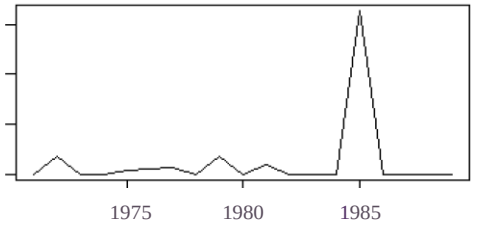
詞數	高頻集中詞	非高頻集中詞
[redacted]		
非重要關鍵詞	40	128

4、尖峰高頻

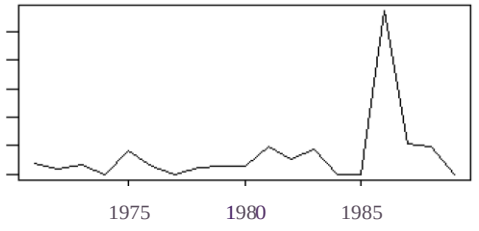


祖國

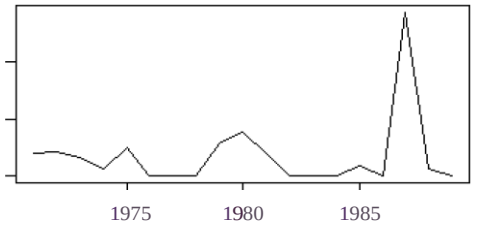
王國



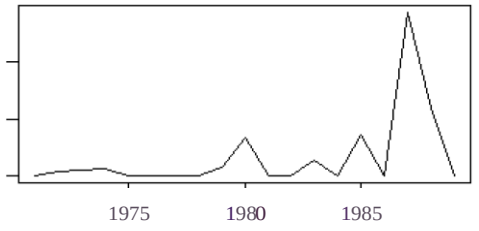
科學



投票



黑人



農村

分析結果:

表三、尖峰高頻字與重要關鍵詞的交叉表

詞數	尖峰高頻詞	非尖峰高頻詞
[redacted]		
非重要關鍵詞	45	123

表六、各種方法在文本的實驗結果

尋找重要關鍵字決策	F- measure	召回率	準確率	選出詞數
平均每 扭頭tt倒:				
選擇前 30 名	0.33	0.21	0.83	12
選擇前 50 名	0.40	0.27	0.72	18
選擇前 100 名	0.42	0.39	0.45	42
邏輯斯迴歸(本研究方法).	0.59.	0.58.	0.60.	47.

結論與討論

- TFIDF的關鍵詞準確率受到文本、每篇文章字數等因素影響。
- 加入統計思維可加入準確率：

→發展由數位資料庫找出關鍵字詞的數量化方法，建立分析檢定的統計理論。

→建立核心關鍵字詞的分析技術，套入生物 棲息地的概念，以物種多樣性的角度判斷 寫作風格、觀念等之特性是否隨時間變動。

報告完畢，

敬請指教！



第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

Science Fiction and Hyperchaos

Digital Humanities as Extro-Criticism

Graham Joncas* Nora Li**

Fudan University, School of Economics*

Shanghai International Studies University, School of English Studies**

Science Fiction and Hyperchaos: Digital Humanities as Extro-Criticism

Nora Li[1] and Graham Joncas[2]

[1] Shanghai International Studies University, School of English Studies
[2] Fudan University, School of Economics

摘要

The two major awards for science fiction are the Hugo Award and Nebula Award. Their main difference is that the Nebula Award is selected by a committee of writers, while the Hugos are chosen by vote. This paper quantitatively investigates systematic differences between the two awards.

Our main variables are word count, hapax legomena percentage, average word length, average word frequency, and average sentence length, as well as three dummy variables for author gender, tense (first person vs. third person), and whether an author has won an award more than once. After extracting these for each winner, they are used as inputs for logistic regression, estimating their importance for each award. Suppose Hugo winners have value ‘1’ and Nebula winners have ‘0’. Then a positive coefficient for a variable means stories with higher values of that variable will more likely win a Hugo award, and vice versa.

To motivate this project, this paper draws from the philosophy of Quentin Meillassoux, who initiated the `speculative realism' movement in avant-garde continental philosophy, to examine how science fiction lets us better understand the structure of digital humanities as a discipline. Meillassoux isolates ‘extro-science fiction’ (XSF) as a parasitic anti-genre that violates the fundamental notion of ‘science’ on which the genre of science fiction (SF) is predicated. Whether at the borders of the galaxy, in hyper-advanced civilizations, or in extreme conditions such as black holes, SF's narratological force derives from the efficacy of science, embodied in constant laws and repeatable experiments. In XSF the laws of nature can change at any time—a state of radical contingency that Meillassoux calls ‘hyperchaos’.

Following Barthes, a text is seen as a tapestry of intertwined extra-literary codes, and literary criticism as the unweaving thereof. This is somewhat problematic in the case of SF, whose internal codes are judged the more exciting the less they correspond to extant ones. From a critic's view, then, the ‘codes’ of SF occur along a spectrum, from banal allegories

(straightforward adoption, with some slight twist) to creating a whole new ‘world’.

Framed differently, SF relies far less than other genres on conventions, inhibiting criticism qua decoding. A great SF novel constructs an autonomous system of codes, strictly incommensurate with any other such system. The creation of such a system is thus an ‘event’ in the philosophical sense—radical contingency. Thus, the challenge of SF criticism is mapping a genre (universe) whose conventions (laws) may change at any time. In hyperchaos, conventional criticism fails.

This paper frames digital humanities as an ‘extro-criticism’ able to operate upon radically contingent literary material. Analogous to Meillassoux’s concept of the arché-fossil, digital humanities operates upon arché-texts—that which is asemic within the literary sign. In this way, reducing stories to numbers allows comparison of disparate code-systems while refusing any commensurability among them (e.g. reduction to common themes). As shown by the applicability of text mining to works that are strictly unreadable, such as the Voynich manuscript, digital humanities analyzes the ‘untext’ within the literary text.

目次

Introduction
1. Text Mining & Extro-Criticism
1.1. Extro-Science Fiction
1.2. The Arché-Text
1.3 Extro-Criticism
2. Hugos vs. Nebulas - Variables
2.1. Word Count
2.2. Average Word Length
2.3. Average Word Frequency
2.4. Sentence Length
2.5. Hapax Legomena
2.6. Dummy Variables
3. Results
3.1. Logistic Regression
3.2 Next Steps
Conclusion
References
Notes

關鍵詞

Science fiction, digital humanities, postmodern literary criticism, genre analysis, computational hermeneutics.

----------正文自下一頁起始----------

Science Fiction and Hyperchaos:
Digital Humanities as Extro-Criticism

Introduction

The two major awards for science fiction are the Hugo Award and Nebula Award. Their main difference is that the Nebula Award is selected by a committee of writers, while the Hugos are chosen by vote. While in some years, such as 2018, the same story wins both awards, an obvious question is whether the two awards systematically differ in any way, as well as how winning stories differ across time.

This paper uses digital humanities to compare winners of both awards for short stories. The Hugo award began in 1955, the corresponding Nebula award in 1966, giving a large corpus. The key advantage of using quantitative metrics is being able to analyze all winners for each award at once (cross-section), within the same award over time (time series), and across awards over time (panel).

Metadata such as word count are extracted from each winner, then used as inputs for logistic regression. If Hugo winners have value ‘1’ and Nebula winners have ‘0’, then a positive coefficient for a variable means stories with higher values of that variable will more likely win a Hugo award, and vice versa.

This paper analyzes various lexical parameters, how science fiction authors vary them for stylistic effect, and their interrelations. Regression analysis shows that three variables—average word length, average word frequency, and author gender—account for 20% of variation between Hugo winners and Nebula winners.

Such ‘distant reading’—with a *quantitative* appeal of reading many books—is seen as the *raison d’être* for digital humanities. For science fiction, however, comparing stories is harrowing not only quantitatively, but *qualitatively*, as no common ground exists among disparate narratives. Thus the next section develops a qualitative motivation for digital humanities as ‘extro-criticism’ that, unlike conventional criticism, remains operative even in a state of absolute contingency.

1 Text Mining & Extro-Criticism

The philosophy of Quentin Meillassoux, who initiated the ‘speculative realism’ movement in avant-garde continental philosophy, lets us link science fiction to the structure of digital humanities as a discipline. Meillassoux (2015) isolates ‘extro-science fiction’ (XSF) as a parasitic anti-genre that violates the fundamental notion of ‘science’ on which the genre of science fiction (SF) is predicated.[1]

Whether at the borders of the galaxy, in hyper-advanced civilizations, or in extreme conditions such as black holes, SF’s narratological force derives from the efficacy of science—embodied in constant laws and repeatable experiments.

Drawing from prior philosophical work, Meillassoux asks whether science fiction is possible under conditions of complete contingency—a state of ‘hyperchaos’ in which natural laws may change without warning, eliminating all grounds for science. This as-yet-unknown genre, Meillassoux names XSF.

1.1 Extro-Science Fiction

Asimov’s story “The Billiard Ball” shows how SF’s dramatic tension arises precisely from its internal notion of science, whose expectations are lost in XSF.

A theoretical physicist, Priss, always lived in the shadow of his rival Bloom, who became rich

by applying Priss's theories. Priss develops a theory of anti- gravity, earning him a second Nobel, but claims it is impossible to realize in practice. Bloom swears he will find a way to apply it, and eventually invites the entire press to witness a public demonstration of his machine.

Bloom has not tested the machine, but is certain it will work. To humiliate Priss, he insists for his demonstration that Priss shoot a billiard ball into an antigravity ray atop a billiard table. Bloom predicts that on hitting the ray, the ball will rise, weightless. Priss aims, then strikes the ball. It bounces in a complex trajectory, then hits the ray. A thunderous noise is heard, and the onlookers see to their astonishment that the billiard ball has pierced through Bloom's heart.

Meillassoux (2015: 22) notes how the story only 'works' as SF, not XSF:

> If the story were Humean, i.e., extro-science fiction, there would be nothing more to say about this aberrant event, and the plot would leave us unsatisfied. But fortunately it is a story of science fiction, i.e., Popperian, and the plot finds a brilliant denouement.

Priss exposits at the story's end how this unforeseeable catastrophe arose: an object disconnected from gravity does not move with calm weightlessness, but "can only move at the speed of a massless object, i.e., the speed of a photon, the speed of light" (ibid.). The story ends as the protagonist speculates—what if Priss (otherwise famous for thinking slowly) had instantly understood what would happen, and Bloom's death had not been accident, but in fact, murder?

Meillassoux (2015: 49-56) later identifies a real XSF novel: René Barjavel's Ravage. In this story, electricity one day simply stops existing. A SF novel would, in the course of its plot,

attribute the disaster to some ‘meta-law’; by contrast, Barjavel simply describes its consequences. No metaphysical ‘closure’ occurs.

Meillassoux ends by speculating upon even more radical XSF—whether a narrative in a state of hyperchaos, as the laws of nature mutate to the point of eliminating all extant life, could satisfy the barest conditions of a coherent plot.

1.2 The Arché-Text

Following Barthes, a text is seen as a tapestry of intertwined extra-literary codes, and literary criticism as the unweaving thereof. This is somewhat problematic in the case of SF, whose internal codes are judged the more exciting the less they correspond to extant ones. From a critic’s view, then, the ‘codes’ of SF occur along a spectrum, from banal allegories (straightforward adoption, with some slight twist or other) to creating a whole new ‘world’.

No doubt, much of SF is boilerplate. Yet, award-winning stories rule this out. Instead, ‘great’ SF qualitatively differs both from SF potboilers and great works of other genres, in that it relies far less on conventions, inhibiting criticism qua decoding. A great SF story constructs an autonomous system of codes, strictly incommensurate with any other such system (extant or fictional). The creation of such a system is thus an ‘event’ in the philosophical sense—radical contingency.

The challenge of SF criticism is mapping a genre (universe) whose conventions (laws) may change at any time. In hyperchaos, conventional criticism fails.

Meillassoux’s philosophy arose as a way to answer how science can still be meaningful in a hyperchaotic world; his thoughts on SF are largely an allegory of this larger project. Thus, to construct an ‘extro-criticism’ for contingent literary matter, this section will sketch out a solution parallel to Meillassoux’s own.

⁎ ⁎ ⁎ ⁎ ⁎ ⁎ ⁎ ⁎

If there is a text, but nobody is around to read it, does it—so to speak—make a sound? This literary formulation of Berkeley's classic query is instructive in that a strict postmodernist ("It means whatever you want it to!") must answer no.

For Berkeley's idealist, a tree falling in the woods may well generate physical sound waves, but with no witnesses, a 'sound' as signifier cannot strictly exist. In the same way, a 'text' as semiotic entity only exists as a correlation between a reader and a written material. Such philosophical rhetoric may seem silly. Yet, suppose one day a book appears from nowhere. No-one has ever read it. Nor can one appeal to authorial intent. One need not be a hardline postmodernist to say it makes no sense to talk about the 'meaning' of this unwitnessed text.[2]

Pierre Bayard's *How to Talk About Books You Haven't Read* takes this Berkelian view even further, arguing for the radical non-equivalence of work and text. In one's internal impression of a work, it is simply impossible to remember every word, or even every feeling or event that the words evoke. Thus, given that any 'reading' of a text must be partial and incomplete, we cannot rigorously demarcate those who have 'read' a text from those who have not. As Bayard wryly notes, many French intellectuals have strong opinions about literary works they have not read—often stronger than the opinions of those who have read them!

Extending Bayard's view, a 'text' can exist even in the absence of a work—as with Abdul Alhazred's *Necronomicon*, or Ts'ui Pên's garden of forking paths. In many ways, these non-existent texts have more meaning than most real texts.

Our Berkelian query concerns the opposite claim: whether a text (work) can be said to have meaning even in the complete absence of a 'text' (social idea). Let us refer to this 'text'-less text as an **arché-text**.

A similar idea, called the *arché-fossil*, is what inspired Meillassoux. Scientific methods such as carbon dating make claims about matter that existed prior to any experiencing subject. Schools of thought such as phenomenology are irreconcilable with such claims, which imply "that manifestation itself emerged in time and space, and that consequently manifestation is not *the givenness of a world*, but rather an intra-worldly occurrence," or that the arché-fossil is "the givenness *in the present* of a being that is *anterior to givenness*" (2008: 15).

Much like carbon dating, methods from digital humanities make claims about texts, irrespective of readers' internal 'texts'. Recalling our unwitnessed Berkelian book, such claims would apply *even to a text that no human has ever read*. That is, digital humanities occupies itself exclusively with arché-texts.

1.3 Extro-Criticism

Meillassoux's project is not merely an exercise in odd premises, but is meant as a radically new answer to the question of how mathematics can provide an absolute description of the real (2011: 18). Unsurprisingly, this still a work-in- progress, yet his strategy for tackling this problem maps onto digital humanities in a striking way.

Crucial to Meillassoux's approach is the concept of the *kenotype*, or sign- devoid-of-meaning (2016: 166). In his view, to view mathematics as signifiers entirely misses the point. After all, mathematical models from game theory, for instance, can be applied to humans, computers and bacteria, without any change to the model's form. Thus, instead of saying that the model 'signifies' any or all of these objects, we can go a step further.

The kenotype "refers to neither meaning nor reference, but only to itself as a sign" (2011: 22). By so doing, it unifies the sign within its contingency (ibid.). For Meillassoux, the efficacy of mathematics arises due to its nature as kenotype.

In a similar way, digital humanities foregrounds that which is asemic within the literary sign. Within the signifier–signified relation of words evoking images, qualities such as letter count enter in only as asignifying detritus. This is shown by the applicability of such methods to texts that are strictly unreadable, such as the Voynich manuscript or *Codex Seraphinianus*.

Digital humanities, seen as extro-criticism, operates at the barest degree of code *as* code. In this sense, by treating literary matter solely at the level of arché-text, reducing stories to numbers allows comparison of disparate code-systems while refusing any commensurability among them (e.g. common themes).

The remainder of this paper aims to illustrate this approach, deciphering a series of literary parameters within an institutionally-imposed state of hyperchaos. The next section outlines each variable, followed by a quantitative analysis.

2 Hugo vs. Nebulas – Variables

The Nebula Award was founded in 1966 by Science Fiction and Fantasy Writers of America, founded by Damon Knight in 1965. This organization has over 2000 members, where membership requires having published science fiction, being professionally involved in science fiction or fantasy, representing a group based on science fiction or fantasy (e.g. a university library), or being a legal representative for a deceased science fiction author's estate. These members decide the Nebula Award by vote. Unlike the Hugo Award, it is more welcome to avant-garde material, so that less popular works ('boring' or 'abstruse') can win the prize.

The Hugo Award was set up by the World Science Fiction Society in 1953, and is named after the author Hugo Gernsback, considered the father of science fiction. Its award for short stories began in 1955. In contrast to the Nebulas, the Hugo Award is decided by public vote at the

World Science Fiction Convention. In other words, anyone with a ticket to the convention has the right to vote. Accordingly, Hugos tend to be awarded to more popular works, such as those with more exciting scenes and plots.

Measuring differences between awards requires extracting textual variables from each story. Regression analysis produces numbers called coefficients that identify each variable's sign (positive or negative), magnitude (large or small), and significance (whether the coefficient is noticeably different from zero). In general, variables with large positive coefficients favor the Hugos, and variables with negative or small positive coefficients favor the Nebulas.[3]

The present section describes each variable used in our regression, giving summary statistics, illustrating stories where a given variable plays a large role in its style, and predicting the coefficient's sign and magnitude.

2.1 Word Count

For both awards, a short story is defined as having less than 7,500 words, with no lower bound. Naturally, stories with fewer words cannot fit as much information, so all else being equal, require more technique to produce a 'complete' story. Thus we expect winners of the Nebula award, which favors technique, to have fewer words, giving the coefficient a negative value.

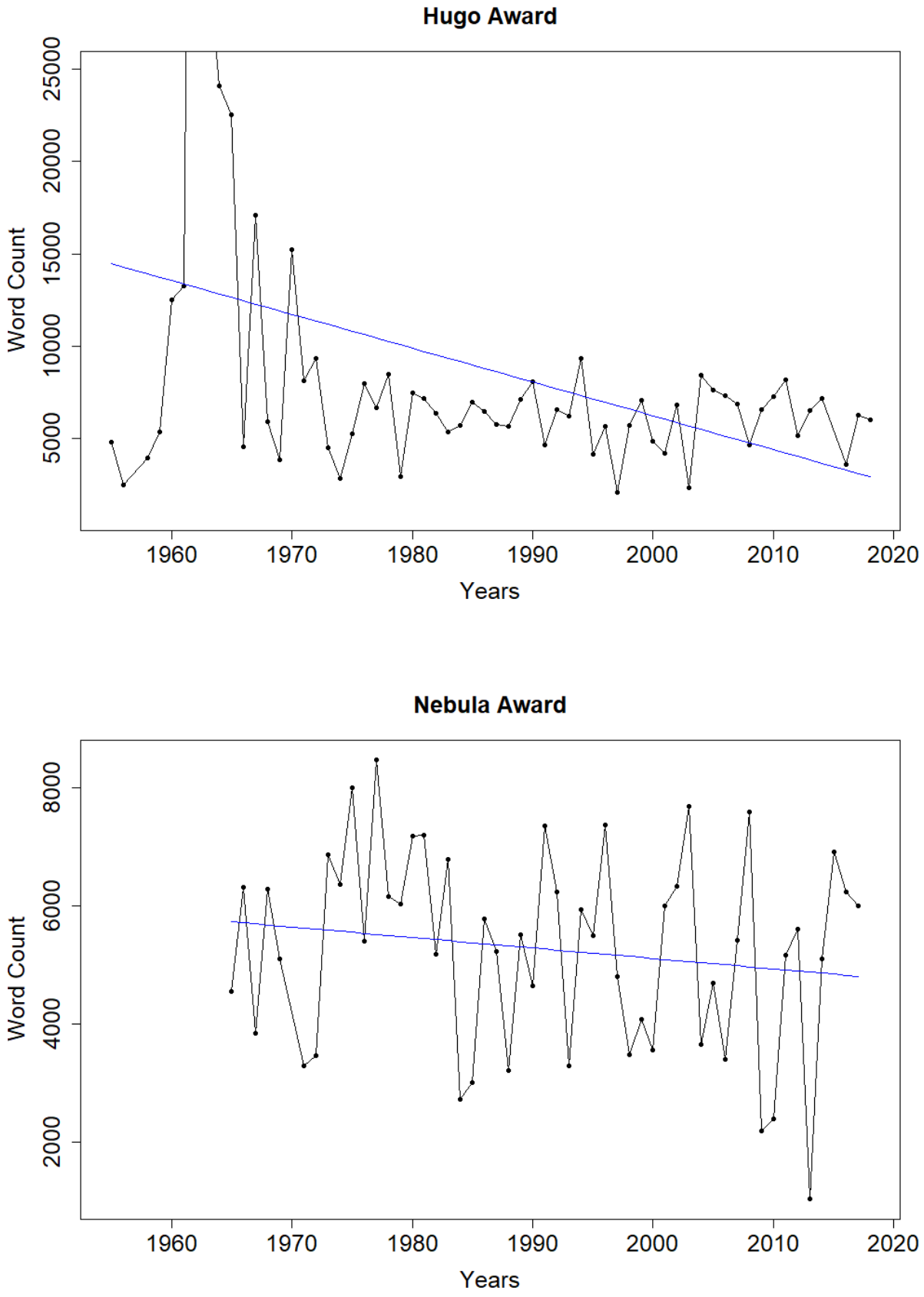


Fig. 1: Word Count (note: different *y*-axes)

On average, Nebula winners have 5, 250 words, with a standard deviation of 1, 700.[4] Early in its history, the Hugo awards had several winners with very high word counts, the most obvious being the 1962 winner *Hothouse* by Brian Aldiss, with 82,600 words. Even after dropping the latter from the sample, several other early winners were more novellas than short stories, raising the mean to 7, 500 and standard deviation to 5,200. Beyond these early stories, word count has been largely stable over time, with a slight downward trend.

Also noteworthy is the exceptionally short 2013 Nebula winner, "If You Were a Dinosaur, My Love" by Rachel Swirsky, at a mere 1035 words. The story begins in the style of a whimsical nursery rhyme, growing sillier and sillier until—not to ruin the ending—it becomes much more than a nursery rhyme. Here, shortness is crucial to the work, which would otherwise feel belaboured.

2.2 Average Word Length

Mean word length helps to measure a text's general verbosity, e.g. using 'obtain' or 'acquire' instead of 'get', or using complicated scientific terminology. We expect this variable to decline over time for both awards. Likewise, we expect the committee-selected Nebula awards to have higher mean word length, and the popular-vote Hugos to have less, giving a negative regression coefficient.

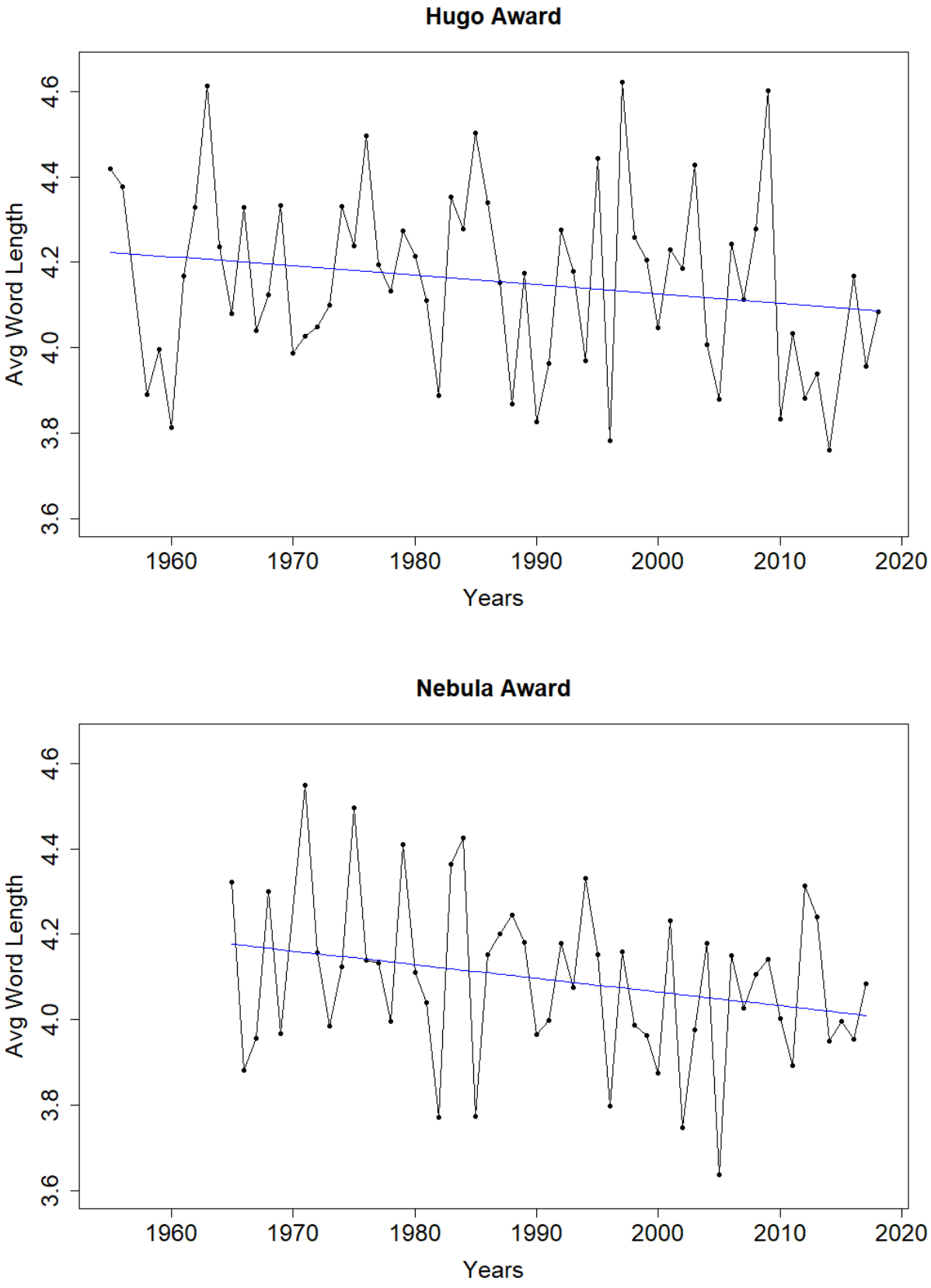


Fig. 2: Average Word Length

Hugo winners tend to have slightly longer words (4.15 letters, $\sigma = 0.21$) than Nebula winners (4.09 letters, $\sigma = 0.19$), though the difference is not significant ($p = 0.16$). For both awards, word length is approximately normally distributed.

The story with the smallest mean word length (3.64) is Carol Emshwiller's "I Live with You", the 2005 Nebula winner. It's hard to imagine how any story can have such small words, but this story makes frequent use of "I" and "you", as well as contractions, which our program splits into two words (e.g. "don" + "t").

The Hugo winners for 1997, 1963, and 2007 have the highest word length. The 1997 winner is a faux literary analysis of Emily Dickinson in light of H.G. Wells' *War of the Worlds*, using verbosity to establish the "author's" personality. Likewise the 1963 story takes place in a medieval-esque setting, using long words to establish an 'archaic' ambiance. Yet, the 2007 story has nothing odd about it—the author simply likes long words, though the story has no sense of verbosity.

2.3 Average Word Frequency

Mean word usage measures how often words are re-used. Since we already control for story length, this variable has the straightforward interpretation of word repetition, and thus of (lack of) lexical variety. We expect the popular Hugos to have higher mean word usage, giving a positive coefficient, though this variable's evolution over time is unclear.

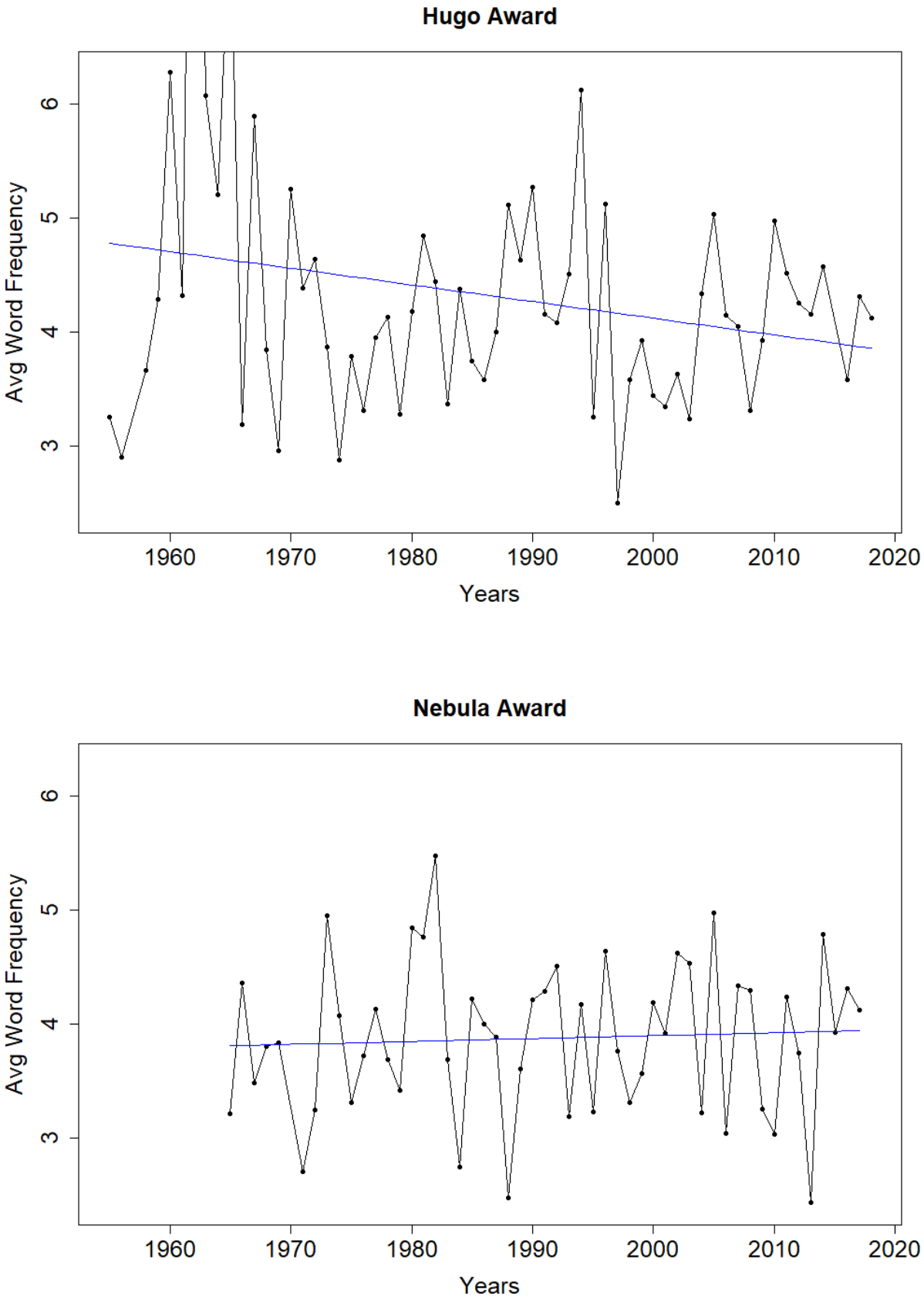


Fig. 3: Average Word Frequency

Hugo winners re-use words slightly more frequently, with a mean of 4.2 and standard deviation of 0.94, compared to Nebula winners with mean 3.9 and standard deviation 0.67. Again, the difference fails to be significant ($p = 0.12$).

A slight problem is that this index (as shown in fig. 3) shows the average of *absolute* word frequency—that is, not proportional to text length, so that longer stories tend to have higher values, hence the very large values for early Hugos. It would be easy enough to divide each value by the word count in that story, but this new index gives poor values in our regression later on, so we omit it.

Lowest word frequency for Hugos is found in the aforementioned review of Dickinson, doubtless as a conscious literary device. For the Nebulas, this occurs in 1988: James Morrow's "Bible Stories For Adults, No. 17: The Deluge". The story takes place in a Noah's ark-like setting, and is written in King James-style biblical English, which is crucial for setting up its tone and twist ending.

2.4 Sentence Length

Sentence length is a well-known parameter for setting the tone of a piece of writing, with short, clipped sentences giving a far different atmosphere than long, flowing ones. Further, literary works abound with stylistic abuses of sentence length, such as Joyce's 4,391-word sentence in *Ulysses*, and Faulkner's 1,288-word sentence in *Absalom, Absalom!*

Thus, one variable worth experimenting with is maximum sentence length, to encode how many authors indulge in this literary flourish, and to what extent. The largest sentences in the Hugos have on average 70.5 words ($\sigma = 35$), while the Nebulas' largest sentences have on average 66.5 words ($\sigma = 36.8$).

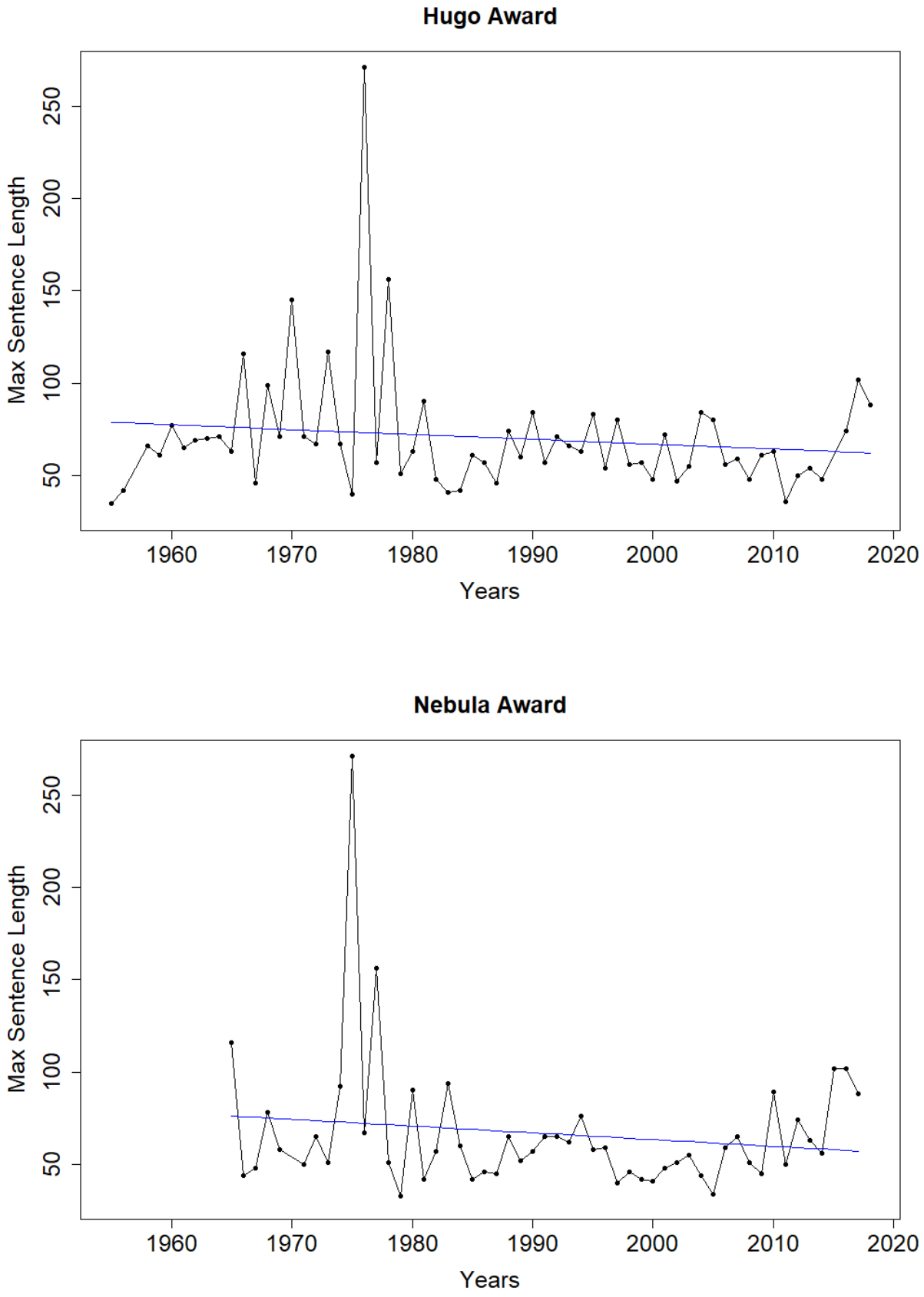


Fig. 4: Maximum Sentence Length

The longest sentence out of all the stories is 271 words, from “Catch That Zeppelin!” by Fritz Leiber, which won both the 1975 Nebula award and 1976 Hugo award.[5] Here the excessively long sentence is meant to give a panoramic view of the culture on a German zeppelin, in an alternate timeline (circa 1937) where this is a primary means of travel. One can imagine how breaking this into smaller sentences would give the dull impression of an itemized list.

The runner-up for longest sentence (156) is also a dual-winner, of the 1978 Hugo and 1977 Nebula award: “Jeffty is Five” by Harlan Ellison. Here, the long sentence occurs when the main character has entered the magical ‘world’ of the titular character (an ageless five-year-old), which involves watching TV and radio shows cancelled long ago (or involving actors who died long ago) as if they have run up to the present day. Its length in part mimics a childish garrulousness, mixed with the main character’s sense of nostalgia for a time that never was.

Perhaps a more general picture, however, is given by average sentence length, which encodes the general atmosphere given by either terseness or loquacity. In the Hugos, the average sentence has 12.7 words, with a standard deviation of 3.7, while for the Nebulas the average sentence has 12.16 words ($\sigma = 3$).

The clearest outlier is Ted Chiang’s “Exhalation”, the 2009 Hugo winner. Given its magnitude, Chiang’s use of long sentences is clearly a deliberate literary device—the main character is part of a race of sentient robots, so this loquaciousness forms an implicit contrast with robots’ typical terseness in other science fiction stories. Interestingly, there are no outliers for short sentences.

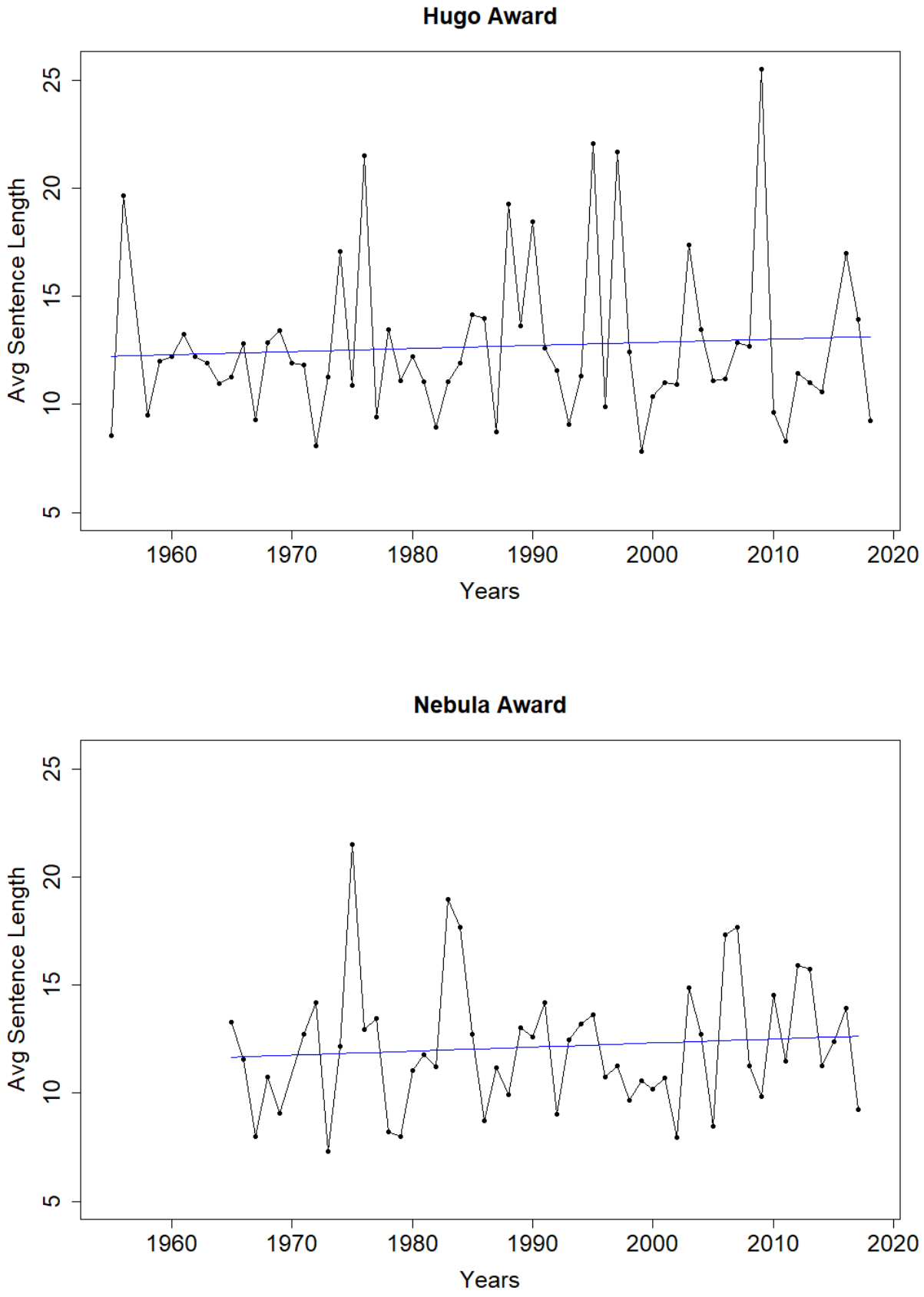
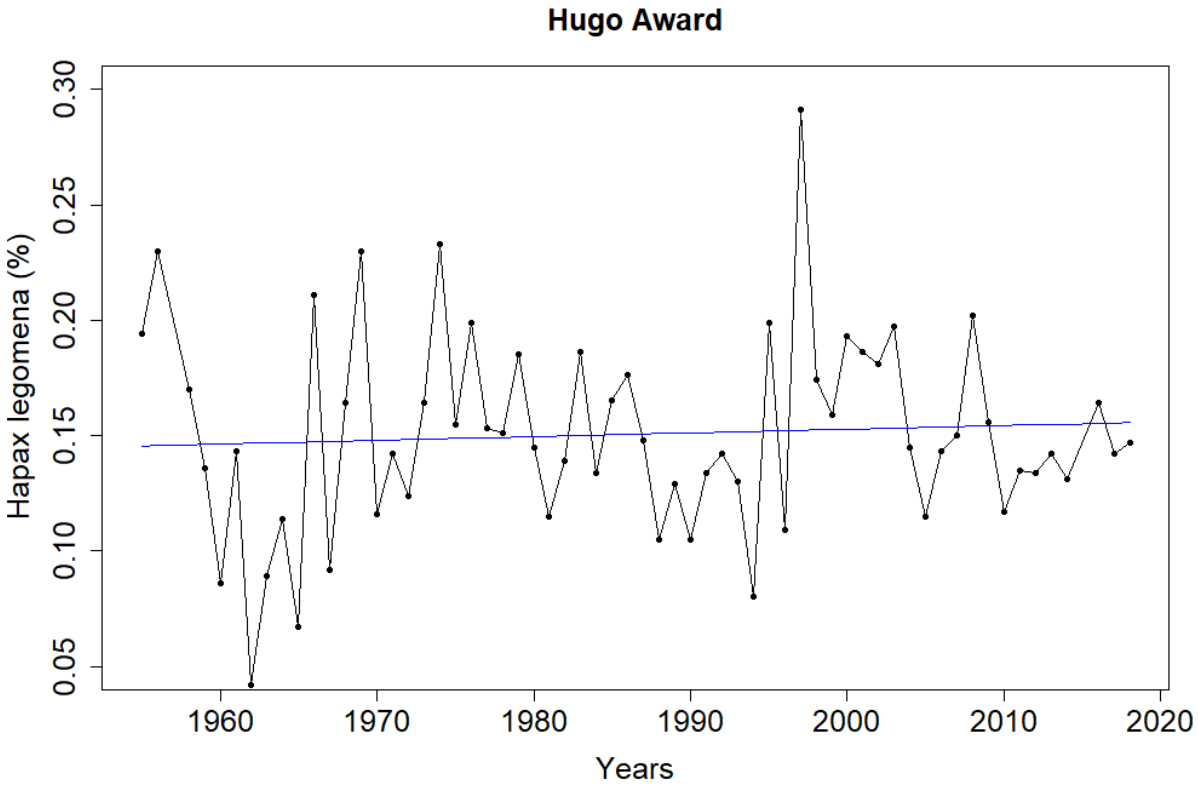


Fig. 5: Average Sentence Length

2.5 Hapax Legomena

Hapax legomena are words that appear in a text only once. This serves as a broad index of a text's lexical variety, such as diction or non-standard spelling. Longer texts tend have more unique words, so this index can be normalized by dividing by the text's length (Jockers, 2014: 69), yielding the percentage of hapax legomena as a fraction of the text. We expect the less 'popular' Nebula awards to have higher lexical variety, giving a negative coefficient. Yet, this variable's evolution over time is unclear, as the relative verbosity of older writing styles may be counterbalanced by stylistic irregularities in spelling (e.g. for accents).



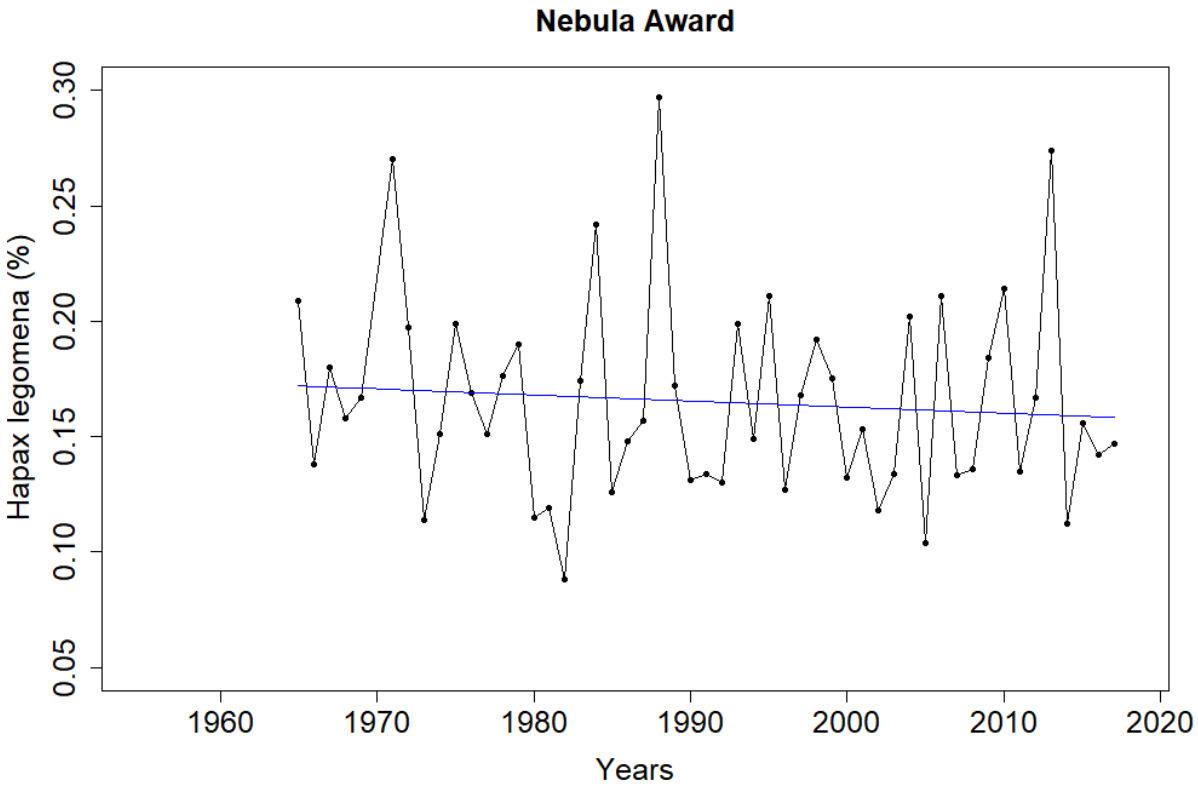


Fig. 6: Hapax Legomena

Hapax legomena make up a larger part of Nebulas (16.5%, $\sigma = 0.044$) than Hugos (15%, $\sigma = 0.042$). While this difference is visible via the regression line of figure 6, it is not strong enough to be statistically significant ($p = 0.2$).

The Hugo with the smallest hapax count is “Soldier, Ask Not” by Gordon R. Dickson, the 1965 winner. The main character is a military-like journalist whose speech and internal monologues are cold and laconic. Yet, the story is quite long (22,551 words), so perhaps words in a longer story are more likely to be re-used.

We encountered the highest-hapax Hugo before—the excessively verbose analysis of Emily Dickinson, where the large variety of words again helps en- force the story’s naïvely pedantic tone. Likewise for the highest-hapax Nebulas, which is the Noah’s ark story written in grandiloquent biblical prose.

2.6 Dummy Variables

Dummy variables encode binary (either ‘this’ or ‘that’) properties, such as labeling stories as Hugo winners (1) or Nebula winners (0). It is likewise possible to include dummies as explanatory variables, though since our dependent variable is itself a dummy, the sign is easy to predict just by cursory inspection.

For example, only 13 Hugo winners (21%) have been women, whereas 29 Nebula winners (56%) are women. Such a large difference implies a strong and significant negative coefficient for gender. The first female author nominated for a Hugo was Pauline Ashwell in 1961, but it was not until 1974 that Ursula Le Guin became the first female Hugo-winning author.

Conversely, the author Jane Beauclerk was nominated during the Nebulas’ opening year of 1965 (albeit among 31 stories authored by men), and in 1968 the Nebula award was won by Kate Wilhelm. As a further curious fact, since 2001 only two Nebula winners have been male. It would be interesting to see if this reflects changes in the Nebula committee’s gender composition over time, but such data are currently unavailable.

Another dummy variable is tense, with 1 for first-person and 0 for third- person. One way to automate this dummy would be to check whether “I” is in the top 20-or-so most frequent words. However, this is liable to error in stories such as Terry Bisson’s “macs” (2000 Nebula winner), which consists of monologue-type responses to the main character, with little use of the personal pronoun.

Curiously, the 2018 winner (Rebecca Roanhorse) is the only author to use the second-person tense, though for reasons of practicality this is coded as 0. Unlike for gender, choice of tense is approximately even for both awards: 33 Hugos (53%) are first-person, while 31 Nebulas

(59%) are first-person. Thus we should expect a negative but small and insignificant coefficient for tense.

Another dummy variable is whether the author has won an award multiple times. This is encoded using two lists of authors for each story (one for Hugos, one for Nebulas), and setting '1' if the author appears more than once on the list, '0' otherwise. For the Hugos, 26 stories (42%) are by repeated winners, while for the Nebulas this occurs for 21 stories (40%). The most frequent winner of both awards is Harlan Ellison, who has won 4 Hugos and 3 Nebulas. Since the difference between awards is minuscule, we expect an insignificant coefficient.

Other possible dummy variables might include the main character's gender, the venue in which the story was first published, data about competing stories, or data about nominations. However, the large corpus size limits dummies related to content, such as whether the story takes place on Earth.

3 Results

Our corpus spans 62 Hugo-winners and 52 Nebula winners. Our dependent variable is the award won by each story, and our regression has 9 regressors, 3 being dummy variables. Figures 1–6 give time series diagrams for word count, average word length, average word frequency, average sentence length, maximum sentence length, and hapax legomena percentage, showing the awards side-by-side.

The advantage of regression analysis is to show how much variation in award can be explained by a given variable, holding the other variables constant. For example, all else being equal, longer stories can be expected to have more hapax legomena; thus, to use it as an index for lexical variety, we should control for word count. Note that after controlling for word count, both hapax legomena and average word frequency act as indices for lexical

variety, hence including both dulls their respective effects, so it is preferable to only include one.

3.1 Logistic Regression

Here, we use logistic regression, designed for binary dependent variables such as award.[6] Once again, award is coded as 1 for Hugo winners and 0 for Nebula winners, which implies positive coefficients if a quality is associated with Hugos, negative coefficients if linked to Nebulas. Thus our initial regression is as follows:

$$\text{award} = \beta_0 + \beta_1 \cdot \text{word\_count} + \beta_2 \cdot \text{hapax\_legomena} + \beta_3 \cdot \text{word\_length}$$
$$+ \beta_4 \cdot \text{word\_frequency} + \beta_5 \cdot \text{avg\_sentence} + \beta_6 \cdot \text{max\_sentence}$$
$$+ \beta_7 \cdot \text{author\_gender} + \beta_8 \cdot \text{first\_person} + \beta_9 \cdot \text{multiple\_wins} + \varepsilon$$

Coefficients are shown in table 1. Regression (1) includes all variables at once, and the only significant coefficient is for author_gender at the 1% level, though word_length ($p = 0.07$) and word_frequency ($p = 0.095$) are significant at the 10% level.[7] Using the McFadden pseudo-R^2, we find that these regressors explain 22% of variation between Hugo and Nebula winners.

Regression (2) drops word_count, hapax_legomena, and multiple_wins due to their lack of significance, as well as max_sentence because its coefficient is so small. It is somewhat surprising that word_count has little value as a control variable, but this is likely due to its high variance.

As noted above, hapax_legomena and word_frequency both act as indices of lexical variety, which explains the strong effect on word_frequency of dropping hapax_legomena. Auxiliary regressions (not shown) make clear that word_length's sharp increase in

significance is due to dropping hapax_legomena.

Regression (3) further drops avg_sentence and first_person, both of which were insignificant in regression (2). The only difference from (2) is in word_length; only dropping max_sentence from (2) does not change increases its coefficient, but dropping only first_person raises it up to 4.2 (σ = 1.6). This may reflect how first-person pronouns such as "I" and "my" have fewer letters than third person pronouns ("he/she", "his/her", etc.), which thereby reduces the average.

We are left with three highly significant variables—average word length, average word frequency, and author's gender—which together explain a fifth of the variation between Hugo and Nebula winners. As we expected simply from observing the summary statistics, author_gender is strong and negative. However, the coefficient for word_length is quite high, contradicting our earlier hypothesis that popular-vote Hugos would tend to have smaller words.

Table 1: Logit Regression – Hugos (1) vs. Nebulas (0)

	(1)	(2)	(3)	(4)
word_count	0.1 (0.2)			
hapax_legomena	1.4 (1.8)			
word_length	3.5* (1.9)	3.7** (1.6)	4.5*** (1.5)	4.5*** (1.4)
word_frequency	2.2* (1.3)	1.5*** (0.5)	1.5*** (0.5)	1.3*** (0.4)
avg_sentence	0.12 (0.1)	0.07 (0.08)		
max_sentence	-0.01 (0.01)			
author_gender	-1.8*** (0.5)	-1.7*** (0.5)	-1.7*** (0.5)	
first_person	-0.55 (0.5)	-0.5 (0.5)		
multiple_wins	-0.25 (0.5)			
pseudo-R^2	0.22	0.21	0.20	0.11

Note: *** – 1%; ** – 5%; * – 10%

Recall that for Hugos the mean word length is 4.15 (σ = 0.21), and for Nebulas the mean is 4.09 (σ = 0.19). Likewise, the Hugos' mean word frequency is 4.2 (σ = 0.94) and the Nebulas' mean is 3.9 (σ = 0.67). Given these small differences, it's surprising that these variables turn out to be the most important.

To isolate the interrelations among each variable, we run another series of regressions (table 2). One striking observation is the change when word_length and word_frequency are put together, rather than separately. This likely reflects how short words such as "the" or "a" tend to be most frequent. Thus the two variables control for one another in a complementary way—word_length becomes an index for verbosity, while word_frequency becomes an index for lexical variety.

Table 2: Logit Regression – Effects of Main Variables

	(1)	(2)	(3)	(4)	(5)	(6)
word_length	1.54 (0.95)			4.5*** (1.4)	1.11 (1.02)	
word_frequency		0.53** (0.25)		1.3*** (0.4)		0.68** (0.29)
author_gender			-1.56*** (0.42)		-1.5*** (0.4)	-1.7*** (0.4)
pseudo-R^2	0.017	0.037	0.095	0.115	0.103	0.140

Last, although author_gender is strong and highly significant in regression 1.3, regression (4) drops it to see how this affects the other two variables. Curiously, word_frequency decreases, suggesting that female authors reuse words at a higher rate than men. A simple regression of word_frequency on author_gender yields insignificant results (not shown), but adding word_count as a control variable indeed gives a positive coefficient (0.31, $\sigma = 0.12$) significant at the 1% level.

Our hypotheses about the signs of these coefficients relied on the Hugos being chosen by popular vote, versus the Nebulas being chosen by a team of experts. We initially predicted that more 'popular' qualities would tend to be favored in the Hugos, i.e. have positive coefficients. The variable for word frequency, taken as indicating lack of lexical variety, appears in line with this hypothesis. However, the positive coefficient for word length (interpreted earlier as an index of verbosity) contradicts this hypothesis. Last, with the exception of author gender, the other variables seem not to be strongly favored by either award.

3.2 Next Steps

Several methodological issues arise that we hope to fix as the project progresses.

First, using a binary dependent variable means that qualities favoring the Hugos ('1') will

have positive coefficients, those favoring the Nebulas (‘0’) will have negative coefficients. In this framework, however, that a coefficient near zero can mean that a quality is either helpful for both awards, or for neither. It would be preferable to disaggregate these two effects.

Second, 6 stories have won both awards, and our solution (the easiest) is to re-use the text twice, as if they were separate texts. A common view is that dual winners are (or will be) classics, so this may allow an auxiliary analysis comparing dual winners to single winners. It’s not clear, however, whether sample size of dual winners is large enough to yield meaningful results.

Last, our analysis has the advantage of covering the entire population of winners. Nevertheless, our scanned versions are not perfect, and so our analysis faces standard issues involving transcription errors. Still, we care more about the relative magnitudes of our regression coefficients rather than literal magnitudes, and since typos are random, they will simply be accounted for by the error term.

⁎ ⁎ ⁎ ⁎ ⁎ ⁎ ⁎ ⁎

The goal of our analysis is to find systematic differences between winners of the two awards, using regression analysis to make these differences explicit.

Each of these stories is a universe unto itself. Receiving a Hugo or Nebula award means that a story is like nothing else the world has ever seen. Almost by definition, there can be no common factors on the order of *content*. Nevertheless, despite the myriad reasons for selecting a given text as a winner, this study finds that a fifth of the variation between these two awards can be explained by three variables: average worth length, average word frequency, and the author’s gender.

Further extensions to this project can take three directions. The first is to develop more detailed variables for each story, such as topics or sub-genres; the key here is figuring out how to automate these tasks, e.g. by topic modeling.

A second extension is to examine each award's evolution over time using time series and panel methods. The difficulty is that such methods do not work for binary dependent variables, which instead are used to group the data. Rather, we can only ask questions using our regressors (e.g. explaining word_count with word_length), which is far less intuitive than simply comparing the two awards.

A third extension is to see whether a support vector machine can successfully classify stories according to the award they won. If it cannot, this will be a major negative result; if it can, it will be worthwhile to inspect any cases that it misclassifies, as well as how it classifies stories that have won both awards.

Conclusion

The premise behind this research project is that science fiction's marginalization is due not to any inherent quality as 'low-art', but because its structure as a genre resists the theme-based analysis of conventional literary criticism. Conversely, digital humanities let us make general statements about a corpus of stories noted for their singularity. Thus, we hope that our project can give insight both into the nature and evolution of science fiction, and also into the structure of digital humanities as a methodology that can open up vistas of inquiry that would otherwise be neglected.

References

1. Bayard, P.; Mehlmann, J. (trans.). (2007). *How to Talk About Books You Haven't Read*. New York: Bloomsbury.

2. Jockers, M. (2014).*Text Analysis with R for Students of Literature*. Heidelberg: Springer.

3. Meillassoux, Q.; Brassier, R. (trans.). (2008). *After Finitude: An Essay on The Necessity of Contingency*. New York: Continuum.

4. Meillassoux, Q.; Lozano, B (trans.). (2011). ``Contingency and the Absolutization of the One." Retrieved from https://www.scribd.com/document/81307810/Contingency-and-Absolutization-of-the-One

5. Meillassoux, Q.; Edlebi, A. (trans.). (2015). *Science Fiction and Extro-Science Fiction.* Minneapolis, MN: Univocal.

6. Meillassoux, Q.; Mackay, R. (trans.). (2016). "Iteration, Reiteration, Repetition: A Speculative Analysis of the Meaningless Sign," in Malik, S. & Avanessian, A. (Eds.) (2016). *Genealogies of Speculation: Materialism and Subjectivity Since Structuralism*. New York: Bloomsbury.

Notes

[1]Meillassoux's French title is *Métaphysique et fiction des mondes hors-science,* which literally translates to: "Metaphysics and Fiction of Worlds Outside/Beyond Science." This *hors* is famously difficult to translate, as in Derrida's "Il n'y a pas de hors-texte" (There is nothing outside the text), whence the awkward neologism 'extro-science'.

[2]A less esoteric example might involve the reams of prose generated by spambots and AI, which will be less and less distinguishable from human texts as NLP improves.

[3]Taking word count as an example, a story obviously must have some words, but since Nebulas tend to have fewer words the coefficient will be positive, but small.

[4]These figures are rounded, to account for transcription errors in the text files.

[5]This 271-word sentence is part of an internal monologue early in the story:

> I stole another glance up at the Ostwald, which made me think of the match- less amenities of that wondrous deluxe airliner: the softly purring motors that powered its propellers—electric motors, naturally, energized by banks of lightweight TSE batteries and as safe as its helium; the Grand Corridor running the length of the passenger deck from the Bow Observatory to the stern's like-windowed Games Room, which becomes the Grand Ballroom at night; the other peerless rooms letting off that corridor—the Gesellschaftsraum der Kapitän (Captain's Lounge) with its dark woodwork, manly cigar smoke and Damentische (Tables for Ladies), the Premier Dining Room with its linen napery and silverplated aluminum dining service, the Ladies' Retiring Room always set out profusely with fresh flowers, the Schwartzwald bar, the gam- bling casino with its roulette, baccarat, chemmy, blackjack (vingt-et-un), its tables for skat and bridge and dominoes and sixty-six, its

> chess tables presided over by the delightfully eccentric world's champion Nimzowitch, who would defeat you blindfold, but always brilliantly, simultaneously or one at a time, in charmingly baroque brief games for only two gold pieces per person per game (one gold piece to nutsy Nimzy, one to the DLG), and the supremely luxurious staterooms with costly veneers of mahogany over balsa; the hosts of attentive stewards, either as short and skinny as jockeys or else actual dwarfs, both types chosen to save weight; and the titanium elevator rising through the countless bags of helium to the two-decked Zenith Observatory, the sun deck wind-screened but roofless to let in the ever-changing clouds, the mysterious fog, the rays of the stars and good old Sol, and all the heavens.

[6]We also repeated the regression using a probit model, which gives essentially the same results. Although the absolute magnitude of the probit coefficients differ, their sign, relative magnitudes, and significance are basically identical to logistic regression.

[7]Both robust standard errors and clustering by years make little difference, so for the present analysis we simply opt for homoskedastic errors.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 12/21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

試析博物館數位藝術史的圖像需求

兼論國立故宮博物院圖像生產與利用方式對數位藝術史之影響

張志光
國立故宮博物院助理研究員

試析博物館數位藝術史的圖像需求—兼論國立故宮博物院圖像生產與利用方式對數位藝術史之影響

張志光
助理研究員
國立故宮博物院

摘要

數位時代來臨對藝術史學者最大的影響就是研究材料—圖像的取得、應用、發佈與分享的方式，有了巨大的改變。大量用來從事藝術史研究與教學的幻燈片被數位化，數位圖像在網路上幾乎隨手可得—只要知道如何尋找，這樣的數位環境應該有利於藝術史的研究與發展，但藝術史學者面對數位科技的變革，適應的速度似乎遠較其他人文學者緩慢，其原因值得探討。

本文試以參與觀察法(Participant Observation)觀察故宮藝術史研究者的日常生活行為，分析博物館藝術史日常工作的圖像需求。其次聚焦博物館藝術史研究，從藝術史研究方法討論研究者之圖像需求，並分析圖像管理工具需求與取用數位圖像的阻礙，以及博物館在數位藝術史發展扮演的角色。最後介紹故宮圖像生產與利用方式對數位藝術史之影響。

目次

1. 前言：發展遲緩的博物館數位藝術史
2. 故宮藝術史研究者日常生活的圖像需求
2.1. 典藏管理
2.2. 研究
2.3. 展覽
2.4. 出版
2.5. 演講
2.6. 文物徵集
3. 博物館數位藝術史研究之圖像需求
3.1. 藝術史研究方法與圖像需求
3.2. 數位藝術史研究的圖像管理工具
3.3. 取用數位圖像的阻礙
3.4. 博物館與數位藝術史發展
4. 故宮博物院圖像生產與利用方式對數位藝術史之影響

4.1. 1924-1936 故宮成立初期各類文物的照相與編輯出版
4.2. 1961-1964 年高居翰(James Cahill)促成的中國藝術文物拍攝計畫
4.3. 1989-1991 年利用「全院藏品文物總清點計畫」時對院藏所有文物拍攝紀錄照
4.4. 2002 年院藏文物全面數位化的規劃
5. 結論

關鍵詞：數位典藏、數位人文學、數位藝術史、博物館、圖像

---------正文自下一頁起始----------

1. 前言：發展遲緩的博物館數位藝術史

相較於其他學科悠久的歷史傳統，數位人文學的「數位」源自於計算機的發展應該是可以被接受的說法，計算機真正對人文學科起到幫助的作用卻是在20世紀中期之後，「數位人文學」(Digital Humanities)的誕生源自於「計算人文學」(Humanities Computing)以及「電子文本」(E-Text)與「數位典藏」(Digital Archives)之後。[1]但是，「數位藝術史」(Digital Art History)雖然也冠上了「數位」一詞，但是在學科發展程度上卻不如數位人文學朝氣蓬勃的樣貌，相較之下顯得有些遲緩。「數位藝術史」的學科發展在最近幾年開始加快腳步，2014年德國慕尼黑大學藝術史研究所開辦了數位藝術史方向的博碩士項目；2015年9月第一個以「數位藝術史」為名的期刊「國際數位藝術史(International Journal for Digital Art History)」在才正式在德國創刊，[2]吹起「數位藝術史」學科地位逐漸成形的號角。另外，2015年美國大學藝術協會(College Art Association)出版的期刊(caa.reviews)新增「數位人文與藝術史」專欄等作為，還有許多會議，以及為藝術史研究生與學者舉辦的工作坊，[3]說明數位藝術史學界加快發展腳步的努力。

學校與博物館雖然同樣是教育與研究機構，但是藝術史學者的研究定位與方向仍然有些許差異。學校藝術史學者需要對大學生或研究生授課，而且學校的學術氣氛較為濃厚，也比較容易接觸到資訊系所的數位技術專家，對數位藝術史的發展有環境優勢。博物館的藝術史學者通常具有另一個身分—策展人(cutator)，需要為展覽而研究，而且研究對象通常與院藏品相關，展覽教育的對象要老少咸宜，除了考慮研究的深度之外(針對藝術史學者)，研究成果也需要透過展覽、出版圖錄以及演講對外展示與呈現(針對一般觀眾)，應用數位工具於研究工作較少，但是展覽則不乏透過各種數位技術的組合，提供參觀者更好的博物館體驗。就博物館藝術史研究者為何願意在展覽上使用創新的數位科技進行展示陳列，而在研究上卻停留在掃描機的使用與圖像檢索，值得另外深入研究。[4]

圖像是全球文化遺產的主要載體，數位圖像是最廣泛的國際觀眾共享和解釋重要文物的媒介。[5]越來越多的博物館希望提供盡可能多的資訊：不僅包括高解析數位圖像，還包括保護說明、技術分析、研究出版物、出處和其他歷史資訊，以及資源本身的後設

[1] 林富士，〈「數位人文學」概論〉，《「數位人文學」白皮書》，頁8。
[2] 何峰、梁诗昕，〈德语国家艺术史研究动态（2010-2015）〉，《艺术学界(第十六辑)》，頁184-185。有關International Journal for Digital Art History 可透過 http://dah-journal.org/線上閱覽。
[3] 陳淑君、江婉綾，〈數位藝術史研究系統的功能需求之環境掃描〉，《圖書館學與資訊科學》42(2)，頁65-66。
[4] James Cuno 認為藝術史學者、策展人和藝術保護者並沒有利用新技術進行不同的研究。參見 James Cuno, "How art history is failing at the Internet in The Daily Dot". Retrieved from https://www.dailydot.com/via/art-history-failing-internet/ (2018/10/08 瀏覽)
[5] Stuart Snydman, Robert Sanderson, Tom Cramer, “The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images”, Paper presented at the Archiving Conference, pp. 16-21.

資料等等，資助機構努力通過開放獲取框架提供研究結果。[6]數位資源的開放來自於需求，藝術史研究者需要圖像，網路世界使用者也需要圖像，博物館圖像開放就是來自使用者的需求，藝術史研究者可以說是圖像開放最大的受益者，這樣的改變有利於營造藝術史研究環境。

本文試以參與觀察法(Participant Observation)觀察故宮藝術史研究者的日常生活行為，分析博物館藝術史日常工作的圖像需求。其次聚焦博物館藝術史研究，從藝術史研究方法討論研究者之圖像需求，並分析圖像管理工具需求與取用數位圖像的阻礙，以及博物館在數位藝術史發展扮演的角色。最後介紹故宮圖像生產與利用方式對數位藝術史之影響。文中討論範圍限於藝術類博物館，並將「圖像」一詞的定義限縮在館藏品衍生或複製之圖像，有關藝術史之圖像研究，其牽涉範圍之廣，遠非本文篇幅及筆者學力所能論及與涵蓋。此處僅以博物館藝術品之分身—圖像，作為本文探討與分析之標的。

2. 故宮藝術史研究者日常生活的圖像需求

本文以參與觀察法觀察故宮器物處研究人員的研究行為[7]，特別是圖像需求與應用情形。筆者直接以參與者的身分涉入被觀察者的日常生活，將圈內人的日常生活世界與外界不易了解的情況，綜理歸納後參照與檢驗目前國內外數位藝術史的發展理論，筆者認為博物館數位藝術史發展遲緩的主要原因可能來自研究者圖像需求未被滿足。觀察故宮藝術史研究者在日常工作或多或少都需要圖像的參與，這些圖像或許來自內部產生，或許是研究者自行產生，甚至是經由網路蒐集，圖像種類與等級也隨著使用目的的不同而異，圖像取得的難易程度不同，而且受到研究者的資訊技術能力、數位技術與網路等因素之影響很大，擁有圖像機構的開放政策也影響圖像的取用。

故宮典藏單位的研究者專長大多都屬於人文學，如藝術史、歷史、考古學、人類學、中文或外語等，進入故宮後調整以藝術史方式研究藏品，背景雖然不同，但是經過一段時間的練習與訓練，均能順利銜接既有的工作方式。故宮在器物處任職的藝術史研究者，其日常工作包括：典藏管理、研究、展覽、出版、演講、文物諮詢鑑定與徵集，對於圖像的需求主要以研究、展覽與出版工作為核心，其他典藏管理、演講與文物諮詢工作較屬於研究成果或學識的應用或具體展現，但是都有程度不同的圖像需求。

觀察故宮研究者主要的圖像來源是館內的藏品管理系統，各種低階圖檔可直接下載於研究或簡報使用，系統提供大圖瀏覽功能，可以將圖檔無限放大，但是超出原始數位圖檔尺寸時便呈現馬賽克，即便如此，也已經滿足大部分研究者之需求。館外的圖像需

[6] Jorge Sebastián Lozano, "Digital Art History at the Crossroads: Critical Approaches to Digital Art History", In A. Dressen & L. Markey (Eds.), kunsttexte.de (Vol. 2017, pp. 14): kunsttexte.de. Retrieved from www.kunsttexte.de. (2018/10/13 瀏覽)

[7] Danny L. Jorgensen 著,王昭正、朱瑞淵譯,《參與觀察法》，頁 29-34。

求除了自網路上下載之外，有時候也會透過研究者彼此間的研究情誼交換互通，而圖錄或是期刊抽印本便是研究成果交流的禮品。

2.1. 典藏管理

故宮器物處依藏品質材分科管理，分成陶瓷、珍玩、銅玉等三個科，每個科約有 3-5 位研究人員，每次需要研究人員與技工友等至少三人才能進入藏品庫房工作。典藏管理的目的是妥善規劃文物進出文物儲位或庫房，並在藏品管理系統上紀錄，以掌握文物所在位置。系統上的每件文物都有圖像，可供研究人員查找文物或比對文物時參考。文物圖像除了本身的檔案名稱直接鏈結文物的編號之外，部分圖像會跟拍 1 張號碼牌作為標記，除了比對圖像與文物的細節之外，文物編號的比對也有助於釐清外表長得很相似的文物。由於庫房面積很大，桌上型電腦不方便移動，而筆記型電腦又太重，目前都是以平板電腦採無線 wi-fi 方式登入系統存取文物圖像進行工作。

目前藏品管理系統上的圖像來源，部分是舊有照片或底片的掃描，更多是來自進行中的大量藏品數位化計畫—以高階數位相機拍攝的數位圖像，其他圖像種類包含：線描圖、拓片、X 光片，然而囿於缺乏經費之故，系統無法提供影片與 3D 建模圖檔的儲存與瀏覽，而且對於關鍵字詞的檢索功能不佳，也無法滿足習慣於使用 Google 搜尋的便捷與效能的研究人員。

2.2. 研究

觀察故宮藝術史研究人員從事研究的方式：除了「確認命題」之外，從「資料蒐集」、「分析與討論」到「發表」等步驟，均須接觸、處理與利用大量的圖像資料。「資料蒐集」階段，資料來源包括：出版品、期刊、古籍檔案、文物圖像、考古出土報告、網路、電子資料庫、其他展覽與到其他博物館的特別參觀等，須要蒐集大量與命題及館藏文物有關之圖像，利用拍照、掃描或下載方式將圖像儲存於電腦設備。「分析與討論」階段，則是透過同儕討論、演講時之研究者間交流、出席研討會與特別參觀等場合，就特定主題文物及其圖像進行研討。「發表」階段分成書面與口語，書面發表以論文、專書或圖錄出版，更是需要印刷品質之圖像。口語發表如研討會發表或專題演講，利用簡報將文物圖像投影到螢幕。然而，找到、看到與獲得圖像，並不意謂可以直接使用，圖像使用權的取得是藝術史學者進行研究的麻煩事。

圖檔取得後的管理也是一個問題。有研究者辛苦的將自己從出版品掃描下來的圖檔，以非常詳細的方式命名，包含出版品名稱、時間、作者、頁碼等等訊息，以便將來單獨使用於其他研究後還能回查取得該圖像的出版品，做完整的引用紀錄。而同 1 張圖檔可能複製了幾十次儲存在不同的研究主題資料夾中，除了沒有圖檔的管理意識之外，也缺乏合適的圖檔管理工具，或是沒有人來告訴研究人員採用更方便好用的圖檔管理軟體。

研究用的圖象一般由數位攝影所掃描產生，但是在研究過程，也可能產生特殊圖像。例如以非破壞方式檢測文物，如 X 光攝影、顯微攝影或紅外線攝影等特殊攝影，均需要特殊的儀器裝備來進行檢測分析，生產的圖像也需要學者專家進行判讀。

2.3. 展覽

博物館藝術史研究人員與學校藝術史學者一個最大的不同就是，博物館藝術史研究人員很多是為了展覽而研究，也就是為了能幫藏品說一個好故事所做的背景研究。展覽時需要圖像大部分在於展覽場地的布置、展覽的補充說明、海報、網站與導覽摺頁，在展覽計畫通過後，需要將額外的文物拍攝作業排入時程，拍攝針對本次展覽所需要的圖像。有些器物雖然因執行「數位典藏與數位學習國家型科技計畫」已經數位化，但是通盤性數位化所拍攝的器物圖像一般是標準套路，例如：器物正視圖、後視圖、底部等角度，與展覽目的所需要特定的圖象角度不同，因此部分器物圖像需要再優化以滿足展覽用途。如果需要借助其他博物館的文物照片輔助說明，需要另外向該博物館提出圖像申請使用。

透過展覽策劃來進行研究也是博物館藝術史研究者最大的幸福，因為可以直接提出展品上手研究，如果有需要可以進行科學檢測分析，例如使用電腦斷層掃描文物，或以紅外線拍攝文物，或電子顯微攝影術，或以拉曼光譜分析等方式，取得文物特殊圖像或檢測資料，作為文物斷代、製作技法、使用情形或材質等之參考資料，這些有助於藝術史研究的重要線索也是科技不斷發展所帶來的好處。

除了靜態的圖像外，展覽也可能有動態的影片輔助說明，並在展覽網站上提供完整的展品清單與圖像，同時在社群網站貼上精彩展品圖像作為宣傳。3D 圖像與虛擬展覽也有許多博物館嘗試，預料也將是未來的趨勢。不管展覽運用多少數位科技，將展品妥善組織呈現，並說一個好故事，才能吸引博物館的參觀者。

2.4. 出版

出版展覽圖錄是博物館研究人員最重要的工作之一，展覽有期限但圖錄是永久的，研究成果透過出版圖錄可以維持更長遠的影響力。圖錄中收錄的圖像比展出的文物多，展覽受到場地大小的限制，難免有遺珠之憾。但是圖錄可以收錄較多的展品圖像，以及其他用來對照比較說明的其他博物館藏品圖像，對研究成果可以做比較完整與嚴謹的論述。大部分的藝術史研究人員都傾向出版實體書，目前還沒有展覽的圖錄是電子式的。而由於研究是不斷進行，圖錄出版的再版也可能有很大的修改意見，電子書的流傳速度太快，相對的未臻完美之地的內容也被擔心會快速流傳。

獲得期刊出版所需的圖像使用權是藝術史研究者的最後一哩路，研究者在藝術史學術期刊發表論文時需準備適當品質的圖像，並取得圖像合法授權，這是論文發表人的義務。故宮為落實開放資料(Open Data)與博物館公共化政策，於官網成立開放資料專區，

不限用途對象免費提供 20MB 大小的 Tiff 圖像檔，成為最好的中國藝術圖像來源，以及最好的博物館圖像開放使用典範。

實務上許多研究論文的發表可能因為沒有好的期刊呈現方式，而無法完整地傳遞論述或是清楚的說明研究過程或論點。例如：以古地圖、書畫或是 3D 藝術品為對象的研究論文，紙本期刊只能使用靜態圖像，若是在電子期刊刊登，將可提供更佳的視覺效果與更好的溝通方式。例如在國際數位藝術史期刊 2018 年第 3 期的論文 PDF 檔就有動態的呈現方式，[8]提供令人驚豔的視覺傳達效果。

2.5. 演講

配合展覽的推出也會舉辦演講活動，利用簡報軟體依展覽單元彙編展品圖像，或比對或放大或以動態呈現，都是為了展覽觀眾說更好的故事。此外，學術研討會上台報告也需要整理許多圖像資料，論證所提的觀點與研究發現，非營利性的論文發表或演講也許有較寬鬆的圖像使用機制，但是演講者與發表人仍應盡力取得圖像合法授權或合理的使用，避免侵權。

另外，從演講的聽眾來說，吸收藝術新知，或為參觀展覽做準備(很多展覽前都會舉辦演講活動介紹展覽內容)，聽眾也有演講內容的圖像需求。一般來說，藝術史的演講或研討會的舉辦單位大都以尊重著作權的說法而不允許聽眾拍攝或錄音。實際上講者的演講屬公開形式，聽眾如果拍攝講者簡報畫面僅僅是自己閱讀研究，而未再次傳播、轉貼或編輯，是否屬於合理使用範圍？講者進行學術性非營利的演講所蒐集的圖像，也可能是合理使用的範圍，但也需注意是否為未經許可的使用方式，圖像的使用權利一直是博物館界複雜的議題。

2.6. 文物徵集

故宮藏品研究人員在接受外界收藏家的捐贈或是購買文物時，需要對標的物進行評審，常需要上網蒐集最新的交易價格資料，作為徵集資料的價格資訊參考，也需要同時蒐集該文物的圖像資料以衡量品相與價格之關係。受贈或收購文物的徵集評審程序大致是一樣的，除了價格之外，也需要蒐集該文物在其他博物館是否有相同的收藏做作為評審的參考。

3. 博物館數位藝術史研究之圖像需求

[8] 國際數位藝術史期刊 2018 年第 3 期的下載版 PDF 檔，參見 Harald Klinke, Liska Surkemper, Justin Underhill, "Creating New Spaces in Art History", International Journal for Digital Art History(3), p 8.，下載點：https://journals.ub.uni-heidelberg.de/index.php/dah/issue/viewIssue/3471/806 (2018/10/12 瀏覽)

每個博物館都有其意義、使命與藏品的特殊性，博物館藝術史研究的對象通常是館藏品或是與其相關的文物，但是即便在博物館內的研究工作，也不可能將所有藏品一一上手研究。因此，在研究資料的蒐集上，相當依賴「文物」的替代品或複製品—「文物圖像」，而數位時代的網路世界提供了更方便與更多元的圖像來源，但是也產生了複雜的數位圖像蒐集、分析與使用問題。

藝術史研究方法將影響圖像的需求。例如以類型學方法研究時，需要蒐集該文物之所有圖像進行分類，特別是年代較明確的遺址出土文物，可以做為同類型文物斷代的參考依據。因為無法直接上手細看文物，越大的圖像越能提供研究者觀察細節，觀察故宮藝術史研究者對於館內藏品管理系統的圖像閱覽效果感到滿意，即是系統提供了比在現場肉眼看還要清楚的圖像。但是圖像的管理與分析工具能做到遠不只是搜尋圖像或是將圖像放大。研究者通常採用檔案管理方式管理自行掃描或從網路下載的圖檔，許多軟體提供資料庫方式的管理功能，例如為圖像添加關鍵字，允許樹狀結構多層代碼管理圖像與分類，在不變動圖像原始路徑的情況下群組呈現所需要的研究圖像。網路藝術史圖像研究平台甚至提供更多結合地理資訊與群體協作方式的功能，或是其他視覺化呈現方式。

從電腦螢幕可以顯示的圖象都可以透過擷取功能儲存於電腦，但是取得之圖像在使用上仍然受限於圖像提供者的規定，而藝術品圖像最大量提供者除了圖像授權公司之外就是博物館、美術館或檔案館等機構，因此博物館對於藝術史研究關係密切。

3.1. 藝術史研究方法與圖像需求

藝術史究竟是什麼樣的學科？許倬雲院士在「2003 海峽兩岸藝術史與考古學方法研討會論文集」序言表述：「藝術史、考古學與歷史，都是呈現人類社會演變的學科。在發展過程上，三者血脈相繫；在研究方法上，又必然相互扶掖。」而且「三個學科的領域，重疊交叉，也很難明確分割。」[9]簡單道出藝術史、考古學與歷史之間的關聯性。張忠培(2003)則認為:「藝術史學是考古學的一個特殊門類,也是史學的一個組成部分。」[10]故藝術史學者需要同時懂得歷史與考古學的研究方法，才能正確地使用考古學資料與史料。藝術史自身的學科定位與學科整合，一直是藝術史學者關心的核心課題。目前許多藝術史學者的研究大量引用考古資料，尤其是博物館藝術史學者，有鑑定學上的需要，將出土文物與相似風格的傳世器對照，有助於了解館藏品的時代與風格。汪聞賓(1993)在〈從科際整合的觀點談藝術史的研究方法〉中提到藝術史研究的六種基礎方法:鑑定、藝術批評、風格造形的方法、社會學的方法、圖像學的方法以及符號學與結構主義的方

[9] 許倬雲，〈三為一體的學科〉，黃翠梅編《2003 海峽兩岸藝術史學與考古學方法研討會論文集》序一。
[10] 張忠培，〈關於古代藝術史學問題的討論—在「2003 海峽兩岸藝術史學與考古學方法研討會」上的總結發言〉，黃翠梅編《2003 海峽兩岸藝術史學與考古學方法研討會論文集》序二。

法，[11]這些方法特別是鑑定、風格造形與圖像學的研究方法相對依賴藝術品，但除了少數藝術品有機會可以直接上手研究外，大部分只能透過圖像來進行研究。

考古學中的類型學是藝術史最常借用的研究方法之一，常被用來分析某一特定時間所出土的特定文物類型。有別於考古學者是以挖掘場址所有出土的文物進行分類，藝術史學者分類的對象，通常範圍包括全世界所有博物館，從網路上並無法完整收集這些文物的圖像，而從圖書館的圖錄中掃描圖像為藝術史學者最常用的方法，但是對這些下載或自己數位化的圖像，學者多以檔案夾方式分別儲存相同來源的圖像，除了數位圖像取用阻礙帶給學者研究資料蒐集的困擾之外，數位圖像的管理與利用也是影響數位藝術史推展的主要因素之一。

最近二個可歸類為數位藝術史研究案例值得提出來比較。首先是上海博物館的「董其昌數位人文」專題研究計畫，其中利用機器學習讓計算機學習董其昌的書法與繪畫，[12]培養計算機認識董其昌的風格，以智慧方式分析中國古代繪畫的元素與特徵，首開機器在中國書畫鑑定之先例。另一個案例是五代北宋山水畫的數位人文研究，[13]以「漁隱」主題為例進行圖像分析與文本挖掘互相參照，追溯「漁隱」主題的原型與流變。前者直接使用機器學習分析圖像，減少人為的操作，後者則是將圖像進行後分類，由研究者分析北宋繪畫中的主題類型，以構建圖像資訊數據模型。從研究材料上來說，二者都需要大量的圖像來進行研究，而且圖像蒐集的齊全與否影響研究成果的效度。從計算機的利用程度來說，前者由計算機直接對圖像進行分析，如何分析如同「黑盒」，分析結果學者只能事後解釋；後者需要學者參與圖像分析，但是不同的學者是否會做出不同的研究結果，產生「白盒」推理偏見的影響，是數位藝術史研究者進行研究命題設計需要關注之處。

3.2. 數位藝術史研究的圖像工具

Drucker, Helmreich, Lincoln& Rose (2015)等人就數位工具對藝術史影響分成四大類：文本分析、空間分析、網絡分析與圖像分析，[14]這些工具的單獨或組合使用取決於研究者的命題，數位藝術史學者雖然因其命題之不同，可以採取不同研究工具，但是圖像分析仍為多數藝術史學者較迫切需要技術或工具。根據 KRESS 基金會於 2012 年所做的報告顯示，圖像分析工具是藝術史學者最需要的工具之一。[15]陳淑君與江婉綾的研究認為

[11] 汪聞賓，〈從科際整合的觀點談藝術史的研究方法〉，《新竹師院學報》7，頁 145-164。

[12] 刘健、张彬，〈博物馆与数字人文—董其昌数字人文项目的实践与思考〉，《第三届北京大学数字人文论坛—孵化与实践：需求驱动下的数字人文项目》。報告中提到有關上海博物館的「董其昌數字人文」專題研究計畫。

[13] 王平、鈕亮、金觀濤&劉青峰，〈五代北宋山水畫的數位人文研究（二）－以「漁隱」主題為例〉，《數位典藏與數位人文》(1)，頁 127-147。

[14] Johanna Drucker, Anne Helmreich, Matthew Lincoln, Francesca Rose, “Digital art history: the American scene”, Retrieved from http://journals.openedition.org/perspective/6021 (2018/10/13 瀏覽)

[15] Diane M. Zorich, "Transitioning to a digital world: art history, its research centres and digital scholarship",

圖像閱覽功能是數位藝術史平台的核心課題。[16]因圖像分析的技術與工具發展較文本分析、空間分析、網絡分析工具緩慢，也是影響數位藝術史的發展的可能原因。另一方面，Cuno (2012)認為在博物館和學院的藝術史研究對新技術的擁抱是緩慢的，他說“Imagine what Panofsky or Aby Warburg could have done with our technology.”，指出研究者並未深入了解新技術如何應用於藝術史研究。[17]

圖像研究工具依使用情境可區分成個人的圖像管理工具以及網路的圖像研究平台。個人圖像管理工具允許研究者在自己的電腦查詢、瀏覽、編輯與標註等操做，以管理蒐集到的數位圖像，甚至可以提供 2 張圖像在同一畫面進行細節的比對，以及圖像的引用格式的匯出，目前的看圖軟體都不是針對藝術史學者量身打造，無法吸引學者使用。網路的圖像研究平台系統則由博物館或大型研究機構支持建置，不同的系統因研究對象的差異而有不同的功能需求，但在圖像閱覽功能構面是大部分系統都會提供，在原件內容檢視上可視需求提供影像導覽、細節放大、分割視窗檢視與關聯藏品推薦；在多媒體互動檢視上的需求則是：數位年表(時間軸)、地理關係模擬、場景模擬、關係網絡模擬。[18]3D 圖像建模與虛擬場景的呈現也是很重要的需求，但相對的技術門檻也較高，短時間內不易實現。

另外還有二個知名的圖像平台案例：IIIF 與 Ionclass。IIIF 國際圖像互操作性框架(The International Image Interoperability Framework, http://iiif.io)是一個由學術和國家圖書館，研究機構，博物館，檔案館，非營利組織和商業組織組成的社群，致力於在網絡上實現可互操作的圖像傳輸。[19]Iconclass 是一個專為藝術和圖像設計的分類系統。它是用於描述和檢索圖像（藝術作品，書籍插圖，複製品，照片等）中所表示的主題的最廣泛接受的科學工具，並且被世界各地的博物館和藝術機構使用。[20]前者透過 API 提供圖像與其後設資料的相關利用，後者則為可視化藝術數位資料資源的分類系統，雖然系統目的不同，但是都是基於圖像的研究平台。

Goolgle 除了是網路搜尋引擎的龍頭，在藝術與文化的保存與推廣也耕耘多年。Google Art Project 上已經收錄了很多全球知名博物館的展場環景，[21]以及知名藝術高清晰圖像，允許使用者直接透過網頁參觀這些博物館陳列室與瀏覽藝術作品。在搜尋引擎

Retrieved from https://s3.amazonaws.com/academia.edu.documents/31018497/Zorich_TransitioningDigitalWorld.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1537802779&Signature=AklMQjjxZj1m1cqVbXRXa7IN9iM%3D&response-content-disposition=inline%3B%20filename%3DTransitioning_to_a_Digital_World_Art_His.pdf (2018/10/13 瀏覽)

16 同註 3，頁 71。

17 同註 4。

18 同上註。

19 同註 5。

20 Rijksbureau voor Kunsthistorische Documentation, "about Iconclass", Retrieved from http://www.iconclass.nl/ (2018/10/10 瀏覽)

21 https://artsandculture.google.com/ (2018/10/13 瀏覽)

首頁除了可以用文字或語音輸入搜尋圖像，也提供「以圖搜尋」查找圖片的方式，正確使用這項工具可以提升藝術史研究資料蒐集的效率。此外，Google Arts & Culture 實驗也做了很多藝術與技術結合的新嘗試，例如 X Degrees of Separation，[22]允許使用者任意拖拉 2 個圖像，系統會自動將 2 個圖像間填滿有關聯性的圖像。另外一個 T-SNE 地圖的計畫，[23]利用機器學習映射藝術作品的實驗，也是訓練計算機以 t-SEN 算法組織數千件藝術品圖像，作品越相似則越接近，創建一個導航式的交互虛擬空間，提供使用者悠游其中。這二個實驗提供使者以不同角度與方式觀看藝術品，並以藝術品的數位圖像為媒介，進行圖像分析處理，提供給使用者在藝術品搜尋、比對與想像的驚奇呈現。

撰寫文章時在文獻引用上有書目管理軟體，但是參考的圖像來源，除了部分圖像來源網頁會提供引用格式之外，似乎沒有專門的圖像引用管理軟體，雖然圖像的引用有格式可供參考，例如：美國心理協會(APA)、芝加哥風格手冊(CMS)與現代語言協會(MLA)等格式，如果有圖像引用管理工具將可以提升藝術史研究發表的效率。

數位工具的應用典範也是影響數位藝術史研究發展的原因之一，由於藝術史研究面向廣泛，目前藝術史期刊的研究案例尚未能引起效尤，仍停留在所謂的「數位化的藝術史」研究階段，典範移轉的動力不足。

3.3. 取用數位圖像的阻礙

博物館的編目卡片約消失在 90 年代，當電腦與網路科技廣泛在博物館應用，藏品管理系統接管了原本卡片提供的功能，而以更有效率的方式提供藏品資料查詢服務，由研究人員桌上的螢幕取代了卡片實體，包含研究人員的研究卡片(研究書目)。網路發展對研究的影響更大，除了連上網路的方式從有線變成無線，頻寬也從文字資料的速度進展到數位圖像，音頻、視頻與虛擬等多媒體，實現如水、瓦斯和電流般從遠方通到自家，開機就能上網取用諸多數位資料。[24]即便如此，藝術史學者要從網路上取用數位圖像，仍然面臨各種阻隔。

典藏機構圖像開放程度與其公共化政策及數位博物館發展策略有關，雖然近年來博物館於網路上開放大量的藏品圖像，但是開放的數量與館藏品的比例仍佔少數。除了因為許多藏品尚未數位化，無法提供數位圖像之外，提供的數位圖像品質不佳，圖像尺寸與檔案都太小，無法滿足藝術史學者「看」的需求。網路上的圖像材料來源分散且未整合，許多典藏機構採取策略聯盟，匯聚圖檔與後設資料於單一平台，並增加圖像放大縮小旋轉瀏覽、註記與留言等功能。

[22] https://experiments.withgoogle.com/x-degrees-of-separation (2018/10/13 瀏覽)
[23] https://experiments.withgoogle.com/t-sne-map (2018/10/13 瀏覽)
[24] 連俐俐，〈雲端上的博物館：自媒體與博物館藝術傳播〉，收錄在顏上晴 & 蕭國鴻編，《博物館科技運用與數位契機》，頁 262。轉引自 1928 年法國詩人保羅・瓦樂希(Paul Valéry)的文章，2018/10/14 網路重新檢索：Paul Valéry,"La conquête de l'ubiquité", Retrieved from http://www.rae.com.pt/valery_conquete_ubiquite.pdf

圖像取用的另一個阻礙是圖像版權問題，取得圖像不意謂可以任意使用，即使是學術非營利的使用，都有侵權的可能。網路世界無國界，但各國的著作權法規與各圖像所有權機構釋出圖像的使用規定卻不盡相同。為了解決網路時代的數位資料取用問題，Lawrence Lessig 與具相同理念的先行者，於 2001 年在美國成立 Creative Commons 組織，透過公眾授權條款，用以標註網路上的數位資料權利，增加資料被利用便利。[25]儘管如此，要取得所有自網路上下載的數位圖像使用權，仍非易事，消彌數位圖像取用阻礙依舊是致力於數位藝術史學社群努力的目標。

3.4. 博物館與數位藝術史發展

美國博物館電腦網路協會(Museum Computer Network, MCN)成立於 1967 年，該協會成立之目的就是致力協助其博物館會員應用電腦改善博物館運作，依此觀點，博物館應用電腦的歷史已經超過 50 年。在科技發展與不同時空及社會政經環境的背景，博物館應用電腦歷經「電子化」、「自動化」、「電腦化」、「資訊化」與「數位化」等等的發展概念。網際網路發展之後，則「數位博物館」、「虛擬博物館」、「網路博物館」、「無牆博物館」、「eMuseum」與「智慧博物館」等博物館在網路世界的分身，大大增加博物館與參觀者接觸的機會。而博物館為了運作需要以及數位博物館需求，紛紛進行藏品的數位化，從數位典藏到數位人文的發展趨勢，似乎也是順應數位科技應用的潮流。美國新媒體聯盟 2016 年《地平線報告》(NMC Horizon Report:2016 Museum Edition)認為數位人文技術是博物館教育與詮釋的重要發展方向，[26]且近年來博物館研究人員與資訊技術專家也開始共同合作，嘗試發展與應用數位人文技術與工具，為博物館多元學科研究奠定基礎。

一般咸認為數位人文學的發展是植基於數位典藏的大量數位化內容之上，博物館的數位典藏則是因數位科技的發展而開始，然而博物館從數位典藏到數位人文，這條路走得並不如圖書館或是學校研究所來得順利。博物館研究有多種討論面向，從博物館業務運作與功能角度區分為：典藏、展覽、教育、保存修護、體驗以及娛樂等研究主題，分別與不同領域的學科有關聯，部分也已經有許多應用數位人文工具的案例。大英博物館館長尼爾‧麥葛瑞格(Neil MacGregor)在《看得到的世界史》一書說過：「**用物品說歷史，正是博物館存在的理由。**」[27]那麼，如果將數位人文工具視為研究藝術史的工具，進行數位藝術史的研究發展，增加博物館的藝術史創新研究的可能。

[25] 中央研究院資訊科技創新研究中心的台灣創用 CC 計畫介紹，Retrieved from http://creativecommons.tw/explore (2018/10/10 瀏覽)

[26] Alex Freeman, "The NMC Horizon Report > 2016 Museum Edition: Mobiles as Catalysts for Change".，收錄在顏上晴、蕭國鴻編，《博物館科技運用與數位契機》，頁 3-43。

[27] 尼爾‧麥葛瑞格(Neil MacGregor)著，劉道捷、拾已安譯，《看得到的世界史 ：99 樣物品的故事 你對未來會有 1 個答案》，頁 21。

4. 故宮博物院圖像生產與利用方式對數位藝術史之影響

故宮博物院的歷史與中國藝術史的研究發展有關，而研究面的拓展有賴於故宮文物圖像的流傳。故宮博物院正式成立於 1925 年 10 月 10 日，從遜帝溥儀出宮後，國務院組織了「清室善後委員會」點查故宮所有物品，自 1924 年起至 1930 年止，清點清宮遺留下來的物品達 117 萬件之多，[28]是當時全世界數量最龐大、最精美的中國歷代藝術文物的寶庫，「故宮博物院之建立，意謂著這個豐富的資料庫對研究者及其他一般大眾的開放，而對那些藝術文物的任何討論，也終於擺脫了傳統『古董』式的訴求，有機會就作品本身進行學術性的探討。」[29]透過舉辦展覽與出版，故宮影響了藝術史學界對中國藝術史的興趣，特別是大量院藏品圖像的生產與利用，提供了藝術史學者研究中國藝術史的材料。

故宮史上共有四次大量圖像生產的時期，分別是：1924-1936 故宮成立初期的各類文物編輯出版、1963-1964 年高居翰(James Cahill)促成的拍攝中國藝術文物的計畫、1989-1991 年利用「全院藏品文物總清點計畫」時對院藏所有文物拍攝紀錄照、以及 2002 年院藏文物全面數位化的規劃，均對中國藝術史的研究產生莫大的影響。

4.1. 1924-1936 故宮成立初期各類文物的照相與編輯出版

故宮成立於風雨飄搖的動盪時期，當時國民政府雖然成立，但是軍閥政權更迭頻仍，清廷復辟聲音不斷，政府財政拮据，為了籌措維持故宮運作的經費，以及對藝術的廣為宣傳，除了開放故宮供一般民眾參觀之外，也出版許多定期刊物，並印製許多精美的書畫與銅器上的拓片。「為使所藏達到宣揚中華文化的目的，除了將各類文物大量展出以外，攝影出版，更能普遍發行，深入民間」。[30]

1842 年(達蓋爾法攝影術發明後 3 年)法國人埃笛爾(Jules Itier, 1802-1877)是第一個在中國拍照的人，[31]而珂羅版印刷技術傳入中國大約是在清光緒（1875-1908）年間。珂羅版印刷技術又稱為玻璃板印刷，雖然技術是由德國慕尼黑攝影師阿爾貝特在 1870 年左右創造，但是中國的珂羅版技術卻是由上海有正書局自日本引進，這是攝影與珂羅版印刷技術傳入中國的時間點。[32]故宮則在 1924 年就開始應用攝影術記錄文物狀態，當時故宮尚未正式成立，「清室善後委員會」為了點查清宮文物，分成許多組同時點查，每一組都配有一位照相人員負責攝影，[33]這些攝影成果很多都出現在日後的出版刊物上。故

[28] 清室善後委員會，《故宮物品點查報告》第一輯，〈編輯說明〉頁 1。
[29] 石守謙，〈臺灣的藝術史研究〉，《漢學研究通訊》31(1)，頁 9。
[30] 昌彼得、馮明珠，《故宮七十星霜》，頁 83。
[31] 吳钢，《摄影史话》，頁 144。
[32] 趙珩，〈有正書局與珂羅版〉，Retrieved from https://hk.saowen.com/a/2502527f192087385ed80c3dda0be3e892642e71d9740c883eecb572bf4368e0 (2018/09/24 瀏覽)
[33] 同註 28，〈清室善後委員會點查清宮物件規則〉第七條頁 1。

宮第一個文物攝影棚於 1928 年春季完成設置，當時在慈寧宮附近的院落，裝修了一個利用陽光做為拍攝光源的日光照相室。1929 年春夏間，又另闢一個使用燈光做為拍攝光源的電光照相室。[34]1931 年於北上門之東雁翅房成立印刷所，內設鉛印、石印、玻璃版(珂羅版) （collotype printing）、凹版等各種印刷機具，在圖像來源無虞的情況下，陸續出版《故宮週刊》、《故宮月刊》與《史料旬刊》等定期出版業務。精美的書法與繪畫則以玻璃版(珂羅版)印製，例如：《故宮書畫集》、《故宮名扇集》、《郎世寧畫幀專集》等之書畫集，印刷品質精良，極負盛名。[35]

除了故宮利用攝影術生產圖像提供文物藝術研究之外，民間人士利用照相珂羅版印刷方式出版青銅器圖錄，改變了中國金石圖錄出版的傳統。黃睿文探討 1900 年到 1940 年間的青銅器圖錄，分析編者如何運用照相與新式印刷技術，在圖錄中呈現青銅器的不同視角，用來說明 1930 年代整個中國學界在青銅器研究上的轉向。[36]珂羅版印刷的品質精良，不像現在的四色印刷還有網點，珂羅版印刷沒有網點，即便現在高仿中國山水畫仍然使用珂羅版印刷。王正華研究清末民初時期代表「國粹派」的《國粹學報》，即是使用代表西洋技術的珂羅版印刷印製，藉由接近原作品複製程度的品質，以求傳達國粹—古物原有的感覺。[37]

出版雖然可以使研究資料—圖像大量傳播，但是資料品質好壞也是影響研究的重要因素。例如使用珂羅版印刷與其他印刷方式所產生的出版品，品質相差甚大，比較《故宮周刊》、《故宮月刊》與《故宮書畫集》印製的文物圖像，即可證明，珂羅版印刷品提供的圖像細節是其他印刷術所不能比擬的。

4.2. 1961-1964 年高居翰(James Cahill)促成的中國藝術文物拍攝計畫

著名美國學者高居翰是將中國藝術傳播到西方最重要推手之一，除了他本人就是中國繪畫史研究的大師級學者之外，也與他促成將故宮文物透過分身—幻燈片展現在世人眼前有關。1961-1962 年故宮「中國古藝術品展覽會」到美國巡迴展覽，策展人即是高居翰，在華盛頓特區國家藝廊開展前，一項大型的幻燈片攝影製作計畫在此展開，繪畫作品拍攝的選擇都是由高居翰負責指揮，「這些幻燈片以成本價提供給許多博物館翰積極投入中國藝術研究與教學的機構。它們改變了中國藝術的教學，特別是繪畫」。[38]由於這個拍攝計畫的成功促成了另一個更大的攝影計畫推展。1963-1964 年由弗瑞爾藝廊(Freer Gallery of Art)與密西根大學籌畫拍攝中國藝術文物的計畫，高居翰擔任繪畫拍攝

[34] 張志光，〈故宮文物攝影的發展歷程與未來挑戰〉，《故宮文物月刊》353 期，頁 120-127。
[35] 同註 30，頁 83-84。
[36] 黃睿文，《民初青銅器圖錄複製觀念的轉變：以容庚《武英殿彝器圖錄》為中心》碩士論文。
[37] 王正華，〈清末民初「古物」的發現、展示文化與國族意識〉，《「玩古・賞新－明清的賞玩文化」國際學術討論會》論文集。
[38] 高居翰(James Cahill)著，王靜靈譯，〈國立故宮博物院在我學術生涯中的位置〉，《故宮文物月刊〉272，頁 94。

工作的監督，在當時故宮所在地台中霧峰鄉北溝拍攝了大量的文物照片(底片)，這批文物照片除了留一份在故宮供研究或出版之用外，另一份被帶回美國，這批大量的圖像奠定了西方研究中國藝術史之開端。因為這些圖像的公布，「構成了美國各大學對中國古畫有系統性研究的基礎，同時也很快地成為美國大學『世界藝術史』計畫中不可或缺的一部分」。[39]

本次拍攝計畫除了產出大量圖像提供藝術史研究之外，也提供藝術史學者一次親近原作，以及相互交流討論的機會。就筆者所知的學者與畫家，例如：王季遷、李鑄晉、艾瑞慈、席克曼、方聞、葉公超、黃君璧、王士杰、江兆申、傅申、莊喆、劉國松、陳其寬、羅覃等，都先後在拍攝計畫期間造訪北溝。[40]雖然大部分藝術史研究都是以圖像為對象，但是能親眼目睹原作仍是藝術史研究者最渴求的願望。

4.3. 1989-1991 年利用「全院藏品文物總清點計畫」時對院藏所有文物拍攝紀錄照

第三次文物圖像的大量生產為 1989-1991 年利用「全院藏品文物總清點計畫」期間對院藏文物除文獻檔案資料外均拍攝 35mm 底片，並沖洗成 3*5 吋照片裝訂成冊，成為院內藝術史研究的重要參考資料。「此次利用清點機會，拍攝每一文物照片建立檔案資料，為現代博物館必備業務措施，也是本次清點計畫中重要項目，此項目決策也造福了本院許多後續業務發展的根基」。[41]以往故宮僅就特定展覽或出版相關文物攝照，在本次全面攝照建檔之後，每件文物均有照片可供閱覽，對於典藏管理、展覽策劃與研究等有很大的幫助。唯本次所拍照片使用於出版上仍有不足，此一遺憾成為下一次院藏文物全面數位化規劃的重要目標。

4.4. 2002 年院藏文物全面數位化的規劃

第四次時間最長且尚未結束，2002-2012 年故宮因執行「數位典藏與數位學習國家型科技計畫」，利用當時發展日益成熟的數位攝影術，開始大量拍攝文物靜態數位圖像，包含一個實驗性質的 3D 圖像計畫，完成 5 件最受歡迎的文物立體圖像。當時的期望是將院藏所有文物都拍攝高品質並可提供出版之用的數位圖像，為此故宮器物處與書畫處陸續建置數位攝影棚，並招募數位攝影師，陸續整理與清潔文物並安排數位攝影。

「數位典藏與數位學習國家型科技計畫」與 2012 年結束後，還未完成院藏所有文物數位化工作，2013-2017 年改以公務預算勉強維持數位攝影，直到 2017 年 10 月起參與

[39] 方聞著，邱士華譯，〈感念國立故宮博物院〉，蔡玫芬編《八徵耄念：國立故宮博物院八十年的點滴懷想》，頁 140。

[40] 李鑄晉，〈早年的國立故宮博物院〉，蔡玫芬編《八徵耄念：國立故宮博物院八十年的點滴懷想》，頁 65-68。羅覃(Thomas Lawton)著，林品樺譯，〈人生難得的機遇〉，蔡玫芬編《八徵耄念：國立故宮博物院八十年的點滴懷想》，頁 69-76。

[41] 林勝安，〈故宮文物帳務管理系統之今昔〉，《第一屆博物館資訊管理學術暨實務研討會》論文集。

「前瞻基礎建設計畫」，獲得經費挹注後，重啟大規模的數位化工作，並採用更高規格的數位化設備與技術，在世界各博物館紛紛進行 3D 圖像建模的潮流下，故宮也開始進行 3D 圖像建模，為參觀者提供更好的數位文物瀏覽體驗。2016 年響應政府開放資料(Open Data)以及博物館公共化政策，對外界開放大量文物數位圖像，包含 300dpi 約 20MB 的中階數位圖像數千張，與超過 7 萬張約 200K 的低階圖像，受到外界很大的關注與讚譽，為中國藝術史研究注入極大的能量。

雖然已經開放大量的數位圖像，但是對於數位藝術史研究尚未見到應用這些圖像的案例。這也就回到本文希望討論的，數位藝術史的圖像需求是否需要圖像分析工具？提供數位圖像或許也可以引起藝術史學者的問題意識，但是數位人文工具或圖像分析工具，提供視覺化或數據化資料，透過不同角度觀察圖像資料的方式，是否能提供藝術史學者數位思維模式？值得進一步的研究。

5. 結論

本文以參與觀察法觀察故宮藝術史研究者的日常生活行為，歸納得出研究人員的工作與圖像息息相關，特別是博物館內的藝術史研究主要是為了展覽，文物的斷代與鑑定真偽都需要分析大量的圖像，或以類型學、風格分析或圖像學等方式進行研究，因此，現階段的數位藝術史對圖像的需求仍然是排在第一位。故宮雖然只是眾多藝術博物館之一，但其規模與知名度在某種程度上可以反映出博物館藝術史研究者共同的研究需求，觀察發現數位藝術史較其他數位人學科發展遲緩的主要原因之一在於研究者的圖像需求未被滿足，從藝術史研究方法與數位圖像取用的阻礙情形，以及相關圖像研究工具未臻完善，都是數位藝術史發展較為低迷的成因。建議數位藝術史的發展應循序漸進，先滿足目前研究者的圖像需求，才可能移動板塊形成典範轉移。

故宮的歷史與中國藝術史的發展有莫大的關係，故宮圖像生產與利用對世界藝術史發展有很大的貢獻。除了貢獻本身的圖像資源外，如果能利用良好的藝術史研究傳統，加強與國內的數位人文中心合作，從「數位典藏」基礎進化成「數位藝術史」研究，[42] 故宮應該具備發展成為數位藝術史研究中心的條件。

自 1924 年迄今，攝影與印刷之技術及工具不斷的演進，圖像生產與傳播方式也隨之改變，使得藝術史學者更容易也更快速取得圖像進行研究。故宮四次大量圖像生產都是應用當時較先進的技術與工具，數位藝術史研究離不開圖像，而博物館等機構是保存最多圖像的地方，數位工具開發的機構能增加對數位藝術史研究的關注，與藝術史學者合作開發適合的研究工具，透過大型研究計畫案，與舉辦數位藝術史研究工作坊，促進數位人文技術與工具的應用，數位藝術史的典範移轉將水到渠成。

[42] 項潔教授在 2010-2012 呼籲台灣學界應該重視數位典藏資料的人文研究，從數位典藏到數位人文。筆者呼應提出，博物館應該積極思考「從數位典藏到數位藝術史」的可能發展。項潔，〈發刊詞：從數位典藏到數位人文〉，《數位典藏與數位人文》(1), 頁 i-v。

參考書目

高居翰(James Cahill)著，王靜靈譯 (2005)。〈國立故宮博物院在我學術生涯中的位置〉，《故宮文物月刊》272，頁 92-99。

Cuno, J. (2012). “How art history is failing at the Internet in The Daily Dot.” *The Way We Think.* Retrieved October 8, 2018, from https://www.dailydot.com/via/art-history-failing-internet/

Rijksbureau voor Kunsthistorische Documentation. (2006). “about Iconclass” Retrieved October 10, 2018, from http://www.iconclass.nl/

Drucker, J., Helmreich, A., Lincoln, M., & Rose, F. (2015). “Digital art history: the American scene.” Retrieved September 24, 2018, from http://journals.openedition.org/perspective/6021

Freeman, A. (2017). “The NMC Horizon Report > 2016 Museum Edition: Mobiles as Catalysts for Change.” In 顏上晴、蕭國鴻(Ed.)，《博物館科技運用與數位契機》。高雄：國立科學工藝博物館，頁 3-43。

Jorgensen, D. L.著，王昭正、朱瑞淵譯 (1999)。《參與觀察法》。台北：弘智。

Klinke, H., Surkemper, L., & Underhill, J. “Creating New Spaces in Art History.” *International Journal for Digital Art History*(3), pp. 8-17.

羅覃(Thomas Lawton)著，林品樺譯 (2005)。〈人生難得的機遇〉In 蔡玫芬 (Ed.)，《八徵耄念：國立故宮博物院八十年的點滴懷想》。 台北：國立故宮博物院，頁 69-76。

Lozano, J. S. (2017). “Digital Art History at the Crossroads: Critical Approaches to Digital Art History”. In A. Dressen & L. Markey (Eds.), *kunsttexte.de* (Vol. 2017, pp. 14): kunsttexte.de. Retrieved October 13, 2018, from www.kunsttexte.de.

Snydman, S., Sanderson, R., & Cramer, T. (2015). “The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images.” Paper presented at the Archiving Conference.

Valéry, P. (2003). “La conquête de l'ubiquité”, J.-M. Tremblay. Retrieved October 14, 2018, from http://www.rae.com.pt/valery_conquete_ubiquite.pdf

Zorich, D. M. (2012). “Transitioning to a digital world: art history, its research centres and digital scholarship”. Retrieved October 13, 2018, from https://s3.amazonaws.com/academia.edu.documents/31018497/Zorich_TransitioningDigitalWorld.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1537802779&Signature=AklMQjjxZj1m1cqVbXRXa7IN9iM%3D&response-content

-disposition=inline%3B%20filename%3DTransitioning_to_a_Digital_World_Art_His.pdf

中央研究院資訊科技創新研究中心 (2003)。〈台灣創用 CC 計畫〉。Retrieved October 10, 2018, from http://creativecommons.tw/explore

方聞著，邱士華譯 (2005)。〈感念國立故宮博物院〉In 蔡玫芬 (Ed.)，《八徵耄念：國立故宮博物院八十年的點滴懷想》。台北：國立故宮博物院，頁 137-150。

王平、鈕亮、金觀濤、劉青峰 (2018)。〈五代北宋山水畫的數位人文研究（二）－以「漁隱」主題為例〉，《數位典藏與數位人文》(1)，頁 127-147。doi:10.6853/dadh.201804_1.0005

王正華 (2004, 01-15~01-16)。〈清末民初「古物」的發現、展示文化與國族意識〉 Paper presented at the 「玩古•賞新－明清的賞玩文化」國際學術討論會。台北：國立故宮博物院。

石守謙 (2012)。〈臺灣的藝術史研究〉，《漢學研究通訊》31(1)，頁 7-15。

何峰、梁诗昕 (2016)。〈德语国家艺术史研究动态（2010-2015）〉In 王廷信 (Ed.)，《艺术学界(第十六辑) 》。南京：凤凰美术，頁 166-192。

吴钢 (2006)。《摄影史话》。北京：中国摄影出版社。

李鑄晉 (2005)。〈早年的國立故宮博物院〉In 蔡玫芬 (Ed.)，《八徵耄念：國立故宮博物院八十年的點滴懷想》。台北：國立故宮博物院，頁 65-68。

汪聞賓 (1993)。〈從科際整合的觀點談藝術史的研究方法〉，《新竹師院學報》7，頁 145-164。

昌彼得、馮明珠(1995)。《故宮七十星霜》。台北：臺灣商務印書館。

林勝安 (2004)。〈故宮文物帳務管理系統之今昔〉，Paper presented at the 「第一屆博物館資訊管理學術暨實務研討會」，台北。

林富士 (2017)。〈「數位人文學」概論〉In 林富士 (Ed.)，《「數位人文學」白皮書》。台北：中央研究院數位文化中心，頁 1-35。

張志光 (2012)。〈故宮文物攝影的發展歷程與未來挑戰〉，《故宮文物月刊》353，頁 120-127。

清室善後委員會 (Ed.) (2004)。《故宮物品點查報告》第一輯。北京：線裝書局。

連俐俐 (2017)。〈雲端上的博物館：自媒體與博物館藝術傳播〉In 顏上晴、蕭國鴻 (Eds.)，《博物館科技運用與數位契機》。高雄：國立科學工藝博物館，頁 3-43。

陳淑君、江婉綾 (2016)。〈數位藝術史研究系統的功能需求之環境掃描〉，《圖書館學與資訊科學》42(2)，頁 65-82。

尼爾・麥葛瑞格(Neil MacGregor)著，劉道捷、拾已安譯 (2012)。《看得到的世界史：99樣物品的故事 你對未來會有 1 個答案》。台北：大是文化。

項潔 (2018)。〈發刊詞：從數位典藏到數位人文〉,《數位典藏與數位人文》(1)，頁 i-v。doi:10.6853/dadh.201804_1.0000。

黃睿文 (2017)。《民初青銅器圖錄複製觀念的轉變：以容庚《武英殿彝器圖錄》為中心》(碩士論文)。台北：國立臺灣師範大學。

黃翠梅編 (2005)。《2003 海峽兩岸藝術史學與考古學方法研討會論文集》。臺南縣官田鄉：國立臺南藝術大學藝術史學系。

趙珩 (2017)。〈有正書局與珂羅版〉，Retrieved September 24, 2018, from https://hk.saowen.com/a/2502527f192087385ed80c3dda0be3e892642e71d9740c883eecb572bf4368e0

An analysis of the image requirements in museum's digital art history—Also discussing the influence of digital art history based on the image production and utilization of the National Palace Museum

Cheung, Chi-Gwong
Assistant Curator
National Palace Museum

Abstract

The greatest impact of the advent of the digital age on art history scholars is the enormous change in the way of image materials acquired, applied, published and shared. A large number of slides used to study and teach art history have been digitized. Digital images are almost readily available on the web - just know how to find them. Such a digital environment should be conducive to the research and development of art history. However, art history scholars face the transformation of digital technology, and the speed of adaptation seems to be much slower than the other humanities. The reason is worth exploring.

This paper tries to observe the daily life behavior of curators in the National Palace Museum by Participant Observation method, and analyzes the image needs of the daily work. Secondly, it focuses on the study of museum art history and discusses the image needs of curators from the art history research methodology. It also analyzes the needs of image management tools and the obstacles of the use of digital images as well as the role of museums in the development of digital art history. Finally, the influence of the production and utilization of images in the National Palace Museum on the digital art history is introduced.

Keywords: digital archives, digital humanities, digital art history, museums, images

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

非物質文化遺產“雕版印刷”技藝的數位化保護研究

鄧抒揚*　　葛懷東**　　舒鐵*

金陵科技學院講師*　金陵科技學院副教授**

DADH 2018

非物質文化遺產“雕版印刷”技藝的數位化保護研究

鄧抒揚　葛懷東　舒鐵
講　師　副教授　講師
金陵科技學院

摘要

2009 年以扬州“广陵古籍刻印社”、南京“金陵刻经处”为代表的“中国雕版印刷技艺”被联合国教科文组织列入“人类非物质文化遗产代表作”名录。本研究将以两家单位为依托，以雕版技艺、古籍版片、雕版典籍为保护对象，开展全面的数字保护研究探讨。

数字化技术在“广陵刻书”、“金陵刻经”技艺保护中的应用，主要体现在四个方面：

⑴ 创建“雕版印刷”技艺数据库，借助数字记录和存储技术为“雕版印刷”的工艺流程建立永久的数字档案；⑵ 开展“雕版板片”再造计划，应用计算机辅助设计、三维建模、数控加工等技术，对古籍版片和佛像经版进行逆向复制，开发“文化衍生产品”；⑶ 建立“雕版典籍”数字内容资源库，利用数字化技术，揭示“雕版典籍”所蕴涵的文化及符号资源；⑷ 建设“雕版印刷”数字博物馆，借助三维动画、虚拟现实技术探讨雕版印刷知识可视化表达，为非物质文化遗产传播提供平台。

目次

1.雕版印刷概述
2.“雕版印刷技藝”的數位化保護
3.創建“雕版印刷”技藝資料庫
4.推進“雕版板片”再造
5.開發“雕版典籍”的數位資源庫
6.構建“雕版印刷”為主題的數位博物館
7.“雕版典籍”再造案例

關鍵詞

雕版印刷，广陵刻书，金陵刻经，数字化保护，文化传承

傳統雕版印刷業從創始至今已有 1300 多年，經歷了初創、發展、繁盛、衰落階段。我們在肯定傳統雕版印刷業的歷史功能和社會價值的同時，也要正視它退出印刷歷史舞臺的現實。作為一種民族遺產，雕版印刷術不僅是中國的，也是世界的。傳統雕版印刷技藝是民族優秀文化的重要組成部分，在延續傳統工藝歷史文化和原有的藝術風格基礎上，如何為之注入更多更好的現代元素和活力，使之更能展示中國傳統特色文化，更能適應現代人的價值取向和對傳統文化的需求，是當下一個現實而又緊迫的文化命題。

1.雕版印刷概述

雕版印刷是運用刀具在木板上雕刻文字或圖案，再用墨、紙、絹等材料刷印、裝訂成書籍的一種傳統技藝，是活字印刷術和現代印刷的技術源頭。它開創了人類轉印技術的先河，承載著難以計量的歷史文化資訊。2009 年以揚州“廣陵古籍刻印社”、南京“金陵刻經處”為代表的“中國雕版印刷技藝”被聯合國教科文組織列入“人類非物質文化遺產代表作”名錄。[①]

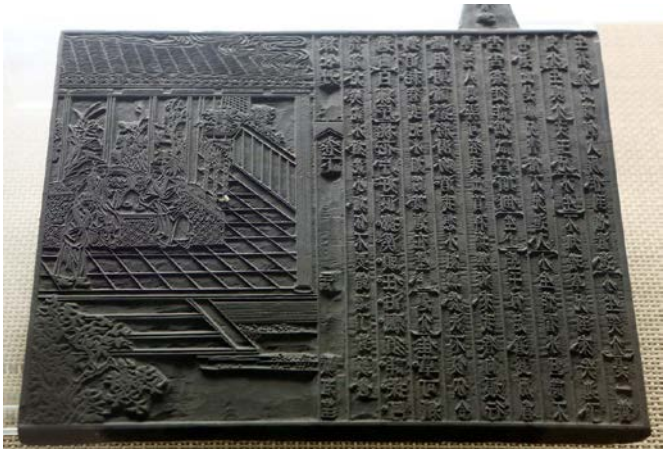


圖 1 雕版板片

“廣陵古籍刻印社”和“金陵刻經處”的“雕版印刷”技藝傳承源於同一譜系，兩家機構創建初期的寫工、刻工、印工、裝訂工都來自江蘇揚州東南郊的杭集鎮[②]。創建於 1960 年的“廣陵古籍刻印社”，從 1962 年起開始全面整理、修補、重印古籍，逐漸恢復了傳統手工雕版印刷的規模生產。目前刻印社擁有陳義時[③]等一批雕版印刷技藝傳人，並運用傳統工具手工操作，整理、雕刻、出版了一大批珍貴古籍，其印行的《桃花

[①] 2009 年 9 月 30 日联合国教科文组织保护非物质文化遗产政府间委员会第四次会议在阿联酋首都阿布扎比作出决议，由扬州广陵古籍刻印社、南京金陵刻经处、四川德格印经院代表中国申报的雕版印刷技艺正式入选《人类非物质文化遗产代表作名录》。

[②] 扬州杭集镇，清代以来这一带雕版艺人众多，以陈开良、陈正春、陈礼环、陈开华、王龙、刘文浩、陈兴荣等为代表的“杭集扬帮”，写工、刻工、印工、装订工齐全，世代相承。

[③] 陈义时，中国雕版印刷技艺的代表性传承人，也是第一批国家级非物质文化遗产项目代表性传承人。生于雕版世家，祖父陈开良、亲陈正春世代从事雕版作坊，从十四岁起随父刻苦钻研雕版技艺。

扇》、《四明叢書》、《西廂記》、《楚辭集經》等 50 餘種線裝古書，繼承了一千多年的技術成果，工藝十分考究。揚州中國雕版印刷博物館也是國內唯一的一座雕版印刷博物館，保存有近 30 萬片古籍版片，能夠原汁原味地展示了雕版印刷技藝。



圖 2 廣陵古籍刻印社的老門牌

由近代著名佛教學者楊仁山居士創建的“金陵刻經處”（1866 年)，是目前世界最大的漢文木刻佛像經版的收藏機構,也是延續傳統方法開展雕刻、刷印漢文佛教典籍的唯一官辦機構。其雕版藝術成就，全面而深刻地體現在佛經刊印的版本、版式、紙張、字體和墨色等各方面，被稱作“金陵本”。金陵刻經處的刻版是經疏會合，另加句逗圈點，劃分段落，並經精校細勘，極少訛錯。金陵刻經處刻印之書籍，紙質綿柔，字大如錢，版式疏朗，清淨莊嚴；刻印之佛像，刀法細膩，章法謹嚴，形神兼備，層次分明；完美地將中國傳統雕版印刷技藝同佛教文化與佛教藝術結合起來，形成了宗教性、藝術性、文物性兼具，獨樹一幟的刻印風格。目前，金陵刻經處保存有 12 萬 5 千餘塊經像版，呈現了內容豐富的佛學經典。



圖 3 金陵刻經處

2.“雕版印刷技藝”的數位化保護

深藏於“廣陵刻書”、“金陵刻經”背後的技藝及傳承模式，強調的是以傳承人為核心，集雕版技藝、經驗和人文精神為一體，具有活態、傳承等特殊性質。顧炎武在《日知錄》中曾提出:“天下無不可變之風俗”。[④]伴隨著社會文化的變遷，人們的生產、生活方式有著不可逆轉的改變，當下的雕版傳統手工藝的生存形式日益嚴峻。20 世紀七十年代以來，雕版印刷術因受到鉛字印刷等原因影響而陷入生存邊緣。再加上各種高科技印刷手段的運用，雕版印刷術幾乎完全退出了現在市場的競爭。如何確保傳統文化不再被丟失，確保傳統手工不再受到冷遇，亟需採取有針對性的方式、方法。

以中國揚州雕版印刷博物館來為例，目前館內的文物庫房中保存著近三十萬張的古老版片，但缺少學者、研究人員對這些版片進行歸類、整理。另外，珍貴的雕版資料被封存、隔絕在庫房裡，研究人員既看不到相關的可視性資料，也買不到專業書籍和資料，面對的只是站在陳列架上的版片，也就談不上很好地傳承和發揚。

記憶與傳承是非物質文化遺產保護的重要手段。利用強大的數位技術對文化遺產進行保護是一條符合時代發展脈絡的途徑。國務院辦公廳印發《關於加強我國非物質文化遺產保護工作的意見》中明確提出：“要運用文字、錄音、錄影、數位化多媒體等各種

④（明）顾炎武《日知录》卷十三《宋世风俗》

方式，對非物質文化遺產進行真實、系統和全面的記錄，建立檔案和資料庫。”[⑤]

對於廣陵古籍刻印社以及金陵刻經處來講，無論是傳統的雕版印刷環節，還是極富藝術價值的版片，都具有特定文化空間的歷史傳承性，並深刻體現出雕版藝術特徵，是中國特有的文化記憶。綜合運用數位技術，包括立體掃描、大資料存儲、檢索、動態攝影、數位建模、雕刻等，針對雕版工藝進行全景式記錄，建立詳細資料庫，通過數位化集成手段，可以為其珍貴雕版的保護與修復提供全面服務，保護與傳承並重，並為其非遺因數的當代活化提供新的路徑，以弘揚不朽的中國傳統雕版印刷技藝。

針對“廣陵刻書”、“金陵刻經”的文化遺產保護及傳承需要，本研究所提出的數位化保護技術路線主要包括四個方面：

⑴ 創建“雕版印刷”技藝資料庫，借助數位記錄和存儲技術為“雕版印刷”的工藝流程建立永久的數位檔案；

⑵ 開展“雕版板片”再造計畫，應用電腦輔助設計、三維建模、數控加工等技術，對古籍版片和佛像經版進行逆向複製，開發“文化衍生產品”；

⑶ 建立“雕版典籍”數位內容資源庫，利用數位化技術，揭示“雕版典籍”所蘊涵的文化及符號資源；

⑷ 建設“雕版印刷”數位博物館，借助三維動畫、虛擬實境技術探討雕版印刷知識視覺化表達，為非物質文化遺產傳播提供平臺。

3.創建“雕版印刷”技藝資料庫

數位記錄具有客觀性、互動性和數位化的優勢，對於雕版印刷技藝、技能和技巧的記錄完全符合原真性的要求。為了強調遺產的“非物質性”，數位記錄重點強調“技藝”的完整性和真實性。借助數位設備和數位工具，採用“直接觀察方法”，可以取得調查的第一手原始資料。這種以追溯為目的記錄，綜合運用了數碼相機、錄音筆、攝像機、智慧手機終端等設備將原始資料收集並轉化成數位文本、圖像照片、視頻影像等客觀的“證據”，再經後期軟體編輯合成，實現“遺產”資訊的記錄和存檔，並最終形成“雕版印刷”工藝流程的原生性數字資源。

[⑤] 2005年3月26日国务院办公厅以国办发〔2005〕18号印发《关于加强我国非物质文化遗产保护工作的意见》。



圖 3 江蘇雕版技藝會傳人 沈樹華（左） 李江民（右）

利用錄音筆和其他語音辨識設備對“雕版印刷”技藝傳承人和知情人進行採訪錄音，實現採訪者與訪談物件的直接互動和雙向交流。並通過語音辨識系統將語音加工成文字檔，提高數位記錄的效率和品質，形成口述歷史的數位文本。利用數碼相機、攝像機可以全程跟蹤拍攝“廣陵刻書”、“金陵刻經”的相關工藝，將深藏于文化遺產背後的流程、細節與經驗，通過攝錄真實素材，聲像並茂地展現出來。

以現場採樣為例，使用數碼攝錄一體機對“刻書”工藝流程進行全程跟蹤拍攝。“刻書”的工序較多，記錄內容包括寫樣、上版、刊刻、刷印、裝訂等將近 30 道工序。其中僅裝訂環節就有抽頁、對折、齊欄、上紙撚、貼封面、切邊、打眼、線裝、貼簽條、上函套等多個步驟。此外，課題組還使用攝像機自帶的話筒和吊杆話筒進行同期音訊採樣，對以下幾段聲音進行了重點記錄：刻板工作時，刻刀與木板相觸的聲音；印刷過程中、墨刷在宣紙上“唰唰”掃過的聲音；分頁、折頁、齊欄過程中，紙張摩擦的聲音；裝訂時，機器與眼釘碰撞的聲音。以上聲音均發自於安靜的工作環境中，顯得尤為珍貴。後期編輯中，同期聲的渲染更能襯托出工序步驟的細節、工作的辛勞、工人的堅忍。前期採集的數碼照片和影像資料經專業軟體編輯合成，處理為相應的圖像和視頻，生成“雕版印刷”技藝資訊的原生性數位資源並永久保存。

隨著互聯網的迅猛發展，數位記錄的解釋性與多樣性能夠在更多的媒介平臺上進行展示，使原本平面的內容動態立體，也讓這些原生性數位資源能夠借助三維、虛擬實境等新興技術實現二次創作，進一步形成再生性資料資源。

4.推進“雕版版片”再造

雕版資源數位化最大的收益，不僅僅在於記錄非物質文化遺產的相關資訊，而且還通過科技手段和方法介入到文化遺產的再生產和再創造，延續“雕版印刷”傳統文脈，實現其在現代社會可持續發展的時代要求。

歷經了成千上萬次的油墨浸泡和印刷，加之發黴、蟲蠹，很多雕版都不可避免地出

現了龜裂和破損。一塊近 800 字的雕版，需要一個技藝精湛的刻工半個月的時間才能完成，而雕版的修復工作對刻工的技藝要求更是嚴苛。儘管如今想學習雕刻技藝的年輕人看似比原來有所增多，但是能夠堅持下來的卻寥寥無幾。目前，“揚州中國雕版印刷博物館”藏有 30 萬塊古籍版片，“金陵刻經處”藏有 12.5 萬餘塊經版，其中 20%～30%已嚴重損壞。而現階段人工複刻書版的成本高、效率低、難度大。同時，鐫刻過程中手工工藝要求也非常嚴格、加工時間長，更不利於新產品的研發。



圖 4 金陵刻經處板片庫房

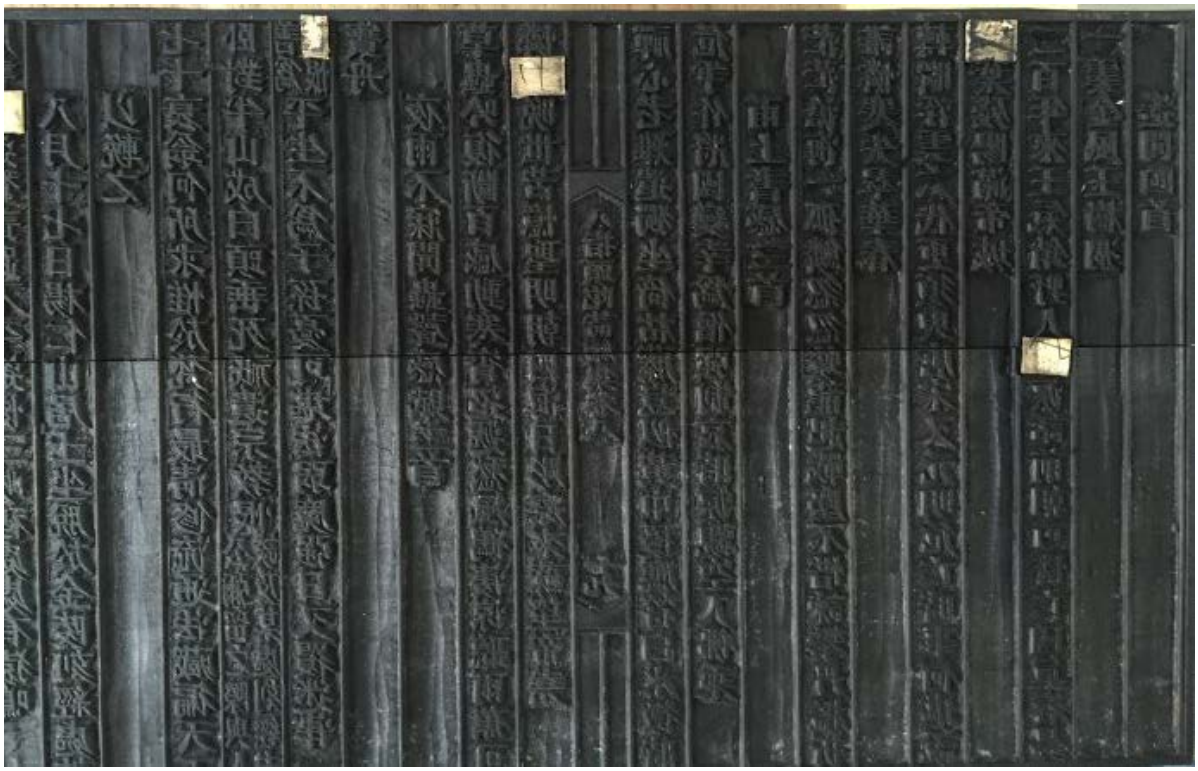
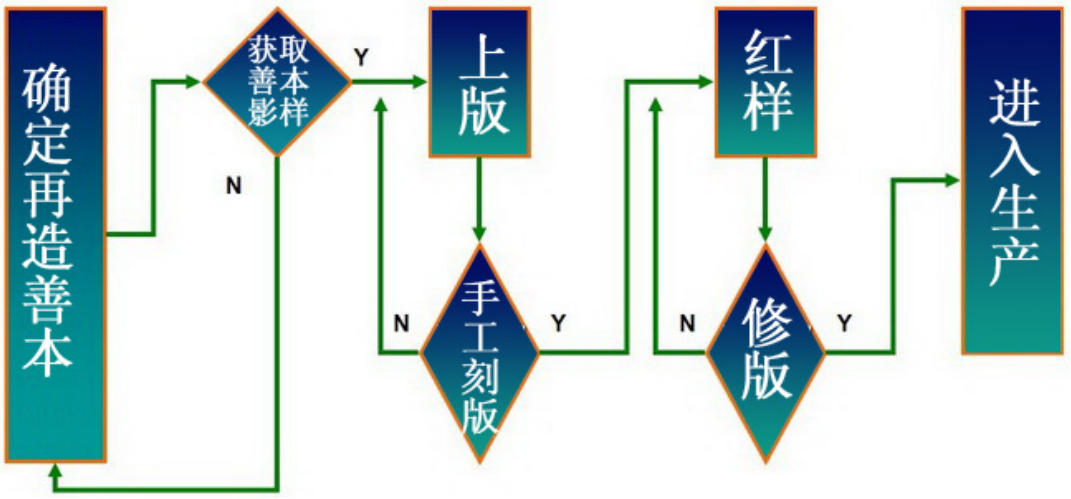


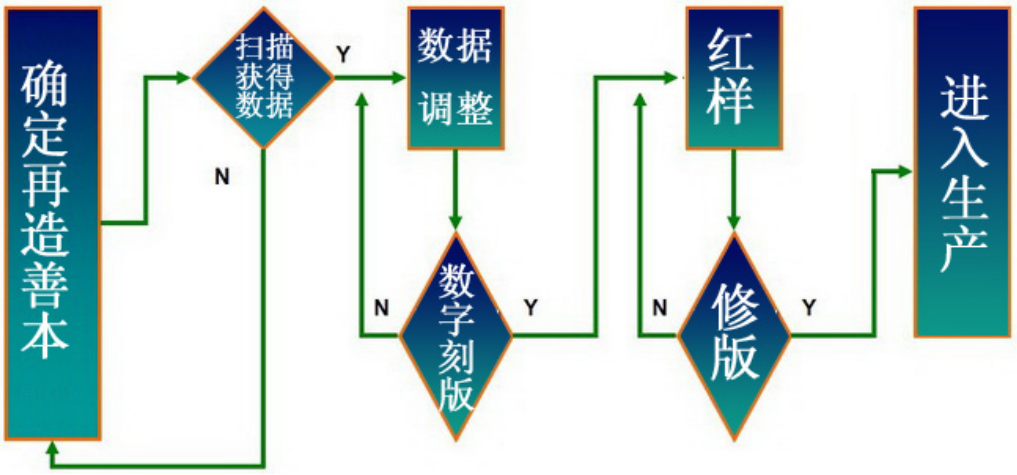
圖 5 補板後的古籍板片

電腦輔助設計和鐳射雕刻成形技術的應用，大大提高了版片的刻制效率和雕刻精度，有快速、柔性、高精度、雕刻成本與模型複雜程度無關聯等優點，從而提升了經版再造的品質。同時，數位化逆向複刻技術，與傳統雕版印刷中分版等關鍵工藝的數位化將大大降低了原來因手工勞動而產生的巨大製作成本，使原本高雅、精美卻價格昂貴，並讓普通民眾望而卻步的傳統雕版印刷藝術品，能夠真正成為多數人能夠擁有的文化收藏。“逆向複刻技術”是根據正向雕版印刷所產生的雕版原件或保存善本原件資料進行資料重建，通過對其可見或可預見的資訊錯誤進行資訊資料的調整修改、試驗和分析得到相對理想的結果，然後再根據修正後的資訊通過打樣，修版、資料再調整等一系列數位化類比生產或部分必要的手工工作得到最終的雕版實物。

採用逆向再造技術所得到的雕版產品，因其以實物精確資訊與各種試驗來完善原有資料，因此得到的結果相對于人為完成和資訊類比更接近原件，從而能迅速找到並確定產品的正確形態縮短產品開發週期。



传统复刻流程



“逆向再造技术”流程

5.開發“雕版典籍”的數位資源庫

“金陵刻經處”刊刻的佛教典籍中有很多為古代失傳佛典，出品佛教經卷以注疏詳實，段落劃分，句讀明確，校勘嚴格著稱，加之精湛的刻印技術，被譽為最精善之佛典版本。其鼎盛時期選印佛典共計 465 種，3300 卷之多，印刷佛像 lO 萬余張，因此“金陵本”佛教典籍文獻價值極高。“金陵刻經”中的一些珍本、善本、孤本，除了具有文獻價值外，其本身也是彌足珍貴的文物，刻經處視這些文獻為鎮館之寶，部分佛像佳刻更是密不示人。相關部門對文物的保藏和展示具有嚴格的規定，普通用戶一般很難查閱和使用。藏而不用，已經成為文化遺產保護的無奈之舉。其實，刻經處作為非物質文化遺產傳承單位有其不得已的苦衷，一些珍貴佛典文獻和佛像版畫由於刊刻年代久遠，大部分紙張已經酸化脆化，平時收藏需要謹慎小心、加倍愛護，頻繁的翻閱使用將會對其造成極大傷害，文獻和版畫的文物價值損失巨大。



圖 6 准提菩薩像（金陵刻經處鏤版，1878 年）



圖 7 准提菩薩像（金陵刻經處鏤版板片，1878 年）

建立金陵刻經處“佛教典籍數位內容資源庫”是解決庫藏古文獻版畫收藏與利用之間矛盾的有效途徑。資源庫收集的主要內容是金陵刻經處刻印的紙質佛教典籍和紙質佛、菩薩畫像，其中還應包括楊仁山居士于清末從日本和朝鮮覓回的佛經。雕版典籍數字資源庫是對雕版印刷文獻資源的內容再現，是雕版印刷再生性保護的重要途徑。開髮雕版典籍數位資源庫，就是將紙質文獻的特點與資訊技術的優勢相結合，讓經過加工後的雕版典籍數字資源保持原有的文化特徵與內涵，實現從影像的數位再現到內容的分析、聚類，從單一雕版印刷內容的處理到海量文獻的資訊重組，從簡單的文本轉換到知識挖掘。可以說，建成後的雕版典籍數位資源庫是基於內容整合、有序的數位資源集合，從而説明研究者更好地進行雕版印刷資源的知識建構，實現文化遺產資源分享。

利用當前的數位及多媒體技術對“金陵本”佛教典籍文獻進行採集、加工，建立數位資源庫，以實現佛教典籍資源的共用。從數位化佛典文獻利用的角度看，珍本佛典文獻一經數位化，就可以無限量地複製和無限制的使用，而不必擔心會有任何損毀，有效解決了收藏與利用之間的矛盾。其次，建立文獻資源庫可以促進人文學術研究和佛學研究手段的更新，檢索起來更方便，為人們節省大量的時間和精力，形成更符合現代社會生活節奏的文化傳播方式。再次，文獻數位化也改變了傳統佛經印刷製品的流通方式，形成數位化的出版和發行管道，使用戶獲得了更為廣闊的閱讀、參研空間，並最終實現了文化遺產社會效益最大化。

6.構建“雕版印刷”為主題的數位博物館

“廣陵古籍刻印社”和“金陵刻經處”作為世界級非物質文化遺產傳承單位，近年已被廣大群眾熟知，每年都要接待眾多考察團體。隨著參觀人數的增加，自然免不了對生產環境的干擾與破壞。基於數位媒介與虛擬實境技術而建立的數位博物館，將多種媒介形式的非物質文化遺產資訊整合在一起，借助電信、無線通訊、互聯網、有線電視以及各種數位電視網路進行傳播，打破特定間、場所的限制，使之成為現代技術條件下適合於大眾傳播的一種新的應用平臺。建立數字博物館，再現“雕版印刷”生產方式和工藝流程，讓雕版印刷技藝的展示、傳播與利用擺脫物理場地的制約，使其得到最大限度的共用利用。



圖 8 揚州中國雕版印刷博物館

雕版印刷技藝數位化展示與數位博物館建設主要包括以下方面：⑴ 將文本、聲音、圖像、視頻、動畫等前期積累的數位資源進行整合，將記錄成果詳細地嵌入虛擬博物館之中；⑵ 利用三維場景建模、特效渲染、虛擬場景協調展示等動畫技術，對雕版印刷技藝特別是傳統手工藝的生產環境、生產方式、消費方式、流通方式、傳播傳承方式等進行情景復原；⑶ 綜合情景建模、行為交互技術、知識視覺化技術、動作綁定技術、WB 技術等技術方案，構建非物質文化遺產多媒體交互體系平臺。

在這其中，情景建模及行為控制技術是數位博物館建設的技術難點，可以綜合利用真實感角色生成技術、動作綁定技術、Multi—Agent 的群控制技術、場景生成技術等快速生成非物質文化遺產中三維場景、角色以及動作。應用 Agent 模型表示虛擬環境中所有的角色，重點解決非物質文化遺產內容製作中的三維表達及互動技術。此外，還要利

用 Motion capture，Motion Analysis 的動作捕捉技術，通過身體部位定點、紅外線掃描等特殊技術手段，捕捉人物肢體的動作資訊生成動作的三維空間座標，經模型修復及動畫處理，形成清晰的動作畫面，並支援交互使用，為全方位、多角度地觀察、分析、研究動作提供方便。還可實現對雕版印刷遺產傳承人單體動作的三維採集和整理，建立雕版印刷工藝流程動作資料庫。

7.“雕版典籍”再造案例

⑴ 廣陵古籍刻印社藏（清）道光《觀無量壽佛經》卷首圖頌

使用三維掃描技術獲得雕版點雲資料。通過 CAD／CAE／CAM 系統分析、進行數控程式設計、數控加工等工程完成點雲資料修復。以梨木為雕版，按圖像三維雕刻完成雕版。

《觀無量壽佛經》卷首圖頌雕版已是國家保護文物，在測量的過程中不允許有任何的損傷，根據文物保護的原則，要保證文物完好情況下獲取到最為精確、完整、細微的三維資料。雕版實物上的雕刻的文字與圖像各異，除了豐富的構圖變化外，還有因雕刻時刀法不同形成的高低錯落，這樣複雜的物體表面，普通的精密型三維測量設備很難在短時間內獲取到所有完整的資料，操作時效性差，也造成延長了工作時間與文物保護的難度。

根據古籍雕版的實際狀態，設置三維掃描方案，對文物進行完整、高精度、無損地測量。無論是複雜的構圖與內容，還是細微的層次落差，都能夠全面，高精度，清晰的將所有資料進去採集和表現，最小點距可達 0.02mm。足以保證再造雕版上的纖毫畢現。

傳統的手工雕版製作方式與生產效率有限，現代人工成本極高，無法形成量產，以及雕版技能高低不定等生產問題，三維掃描與數控雕刻機密切配合，在快速獲取高保真度的雕版三維資料後，通過程式設計軟體生成 CNC 刀路，再輸入雕刻加工中心，數控雕刻機就能自動雕刻加工，在極短的時間內就可完成高精度的雕刻工作。

⑵《蘿軒變古箋》

刊印于明天啟丙寅（西元 1626）的《蘿軒變古箋譜》目前被認為是中國現存最早的一部箋譜，存於上海博物館，被考訂為海內孤本。《蘿軒變古箋譜》的重要價值之一是在於其譜中所用到極為成熟的“飯版、拱花”之“木版浮水印”技藝。

以明代套色印刷古籍為原本，對其彩色全息掃描，可以得到物體表面三維資料和 r、g、b 三原色資料，通過圖像軟體對套色部分進行數位虛擬分版。再轉換為點雲資料庫，通過 CAD／CAE／CAM 系統分析、進行數控程式設計、數控加工等工程完成點雲資料修復。最後，以梨木為雕版，按圖像三維雕刻完成雕版。彩色全息掃描獲得雕版分色與點雲資料整合運用，完成對傳統木版套色技術的逆向再造。



圖 9 《蘿軒變古箋譜》

參考書目

⑴ 專書

揚州中國雕版印刷博物館（2013）。《雕版印刷》。濟南：山東友誼出版社。

羅琤（2010）。《金陵刻經處研究》。上海：上海社會科學院出版社。

劉維芳（2012）。《金陵刻經印刷技藝》。南京：南京出版社。

⑵ 期刊論文

李霞,萬豔華(2012)。〈揚州市非物質文化遺產“雕版印刷”的生產性保護與發展〉，《城市發展研究》5, 頁 55-60。

趙子君,倪建林（2014）。〈揚州雕版印刷技藝的調查與研究〉,《創意與設計》2，頁 74-82。

朱秋婷（2018）。〈產業資訊數位化服務與清代揚州雕版印刷插圖的結合與運用〉，《美術教育研究》13，頁 32-33。

何煒（1991）。〈雕版印刷術的“活化石”——訪揚州廣陵古籍刻印社〉,《圖書館》1，頁 69-70。

王罡（2016）。〈金陵刻經處載籍藝術的數位化保護與傳承〉,《藝術教育》12，頁 138-139。

陳筱嬌（2011）。〈現代技術下的金陵刻經處〉,《大眾文藝》7，頁 260。

陳治國.（2016）。〈金陵刻經處與金陵刻經印刷技藝保護〉,《法音》7，頁 25-30。

王罡（2016）。〈數位化背景下中國雕版藝術的傳承新路徑——以南京金陵刻經處為例〉,《美術教育研究》22，頁 42-43。

黃常倫（1994）。〈譽滿中外的金陵刻經處〉,《江蘇地方誌》1，頁 60。

三、碩博士论文

朱燕（2017）。〈揚州雕版印刷技藝的傳承探究〉，合肥：安徽財經大學。

趙子君（2012）。〈揚州雕版印刷技藝調查與研究〉。南京：南京師範大學。

陳筱嬌（2011）。〈金陵刻經處雕版印刷技藝傳承研究〉。南京：南京藝術學院。

李朵朵（2012）。〈數位化工作方式下的非物質文化遺產活態記錄與傳播研究〉。南京：南京藝術學院。

四、其他文献

DB32/T 2038-2012,雕版印刷技艺[S].

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

A digital map for Dante
current online projects and a work in progress

成沫
罗马大学罗曼语文学博士
同济大学人文学院、欧洲思想文化研究院讲师
中意学院院长助理

演讲题目： **A digital map for Dante -- current online projects and a work in progress**

演讲者：成沫，罗马大学罗曼语文学博士，同济大学人文学院、欧洲思想文化研究院讲师，中意学院院长助理。电话：187 21353627, 021-65983561。电子邮箱：chengmo@tongji.edu.cn

关键词：罗曼语文学、但丁、数字人文

演讲论文初稿：

在当前数字人文学蓬勃发展态势下，欧洲与美国在**古典学、中世纪与文艺复兴**研究方面，大量采用了数字人文研究技术。从最早的中古经院哲学方面的阿奎那数字化索引研究，到希腊罗马古典学研究方面极重要的 Persus 数字图书馆，直到普林斯顿大学《但丁 2.0》计划、哥伦比亚《数字但丁》计划等，在欧美中古及但丁研究方面，存在着丰富的数字资源和大量的优秀案例。本文将通过对相关数字项目的梳理及其在语文学研究方面的意义，呈现但丁研究中的数字人文化进展，以及其对中国的欧洲中古研究者的启示。

数字人文技术在文献目录学方面的意义：

西方“语文学”（philology）的概念来自希腊语 φιλολογία，即对语言、词语的热爱。学科结合文学评论、历史学及语言学，通常以文学文本为研究对象，具体运用版本学（ecdotics）、古文书学（paleography）等研究手段，对文本进行考据校勘（textual criticism），注疏释义。[1] 对于从事中世纪研究的语文学家来说，在确立研究主题后，首先需要确定的是对应的手稿文本及相关文献信息。以文本为核心的研究方法对于手稿资料的收集、整理，提出了很高的要求。在前数字时代，这样工作，或依靠前人已有的研究，若是无人涉猎的作家作品，则结合各藏书馆的藏稿目录及二手文献，进行梳理。在中世纪研究领域内获得长足发展的数字化平台，为这些研究提供了大量的便利，并从某种程度上说更改了语文学研究的方式。

Les Archives de littérature du Moyen Âge (ARLIMA，中古文学档案库) 项目就是一个很

[1] 但丁研究所涉及的语文学分支为罗曼语文学，早在欧洲民族国家尚待行成之时，但丁以拉丁文撰写的 *De Vulgari Eloquentia*（论俗语）即可被视为是这门学科最早的论述作品。浪漫主义与民族主义推动下，F. Raynouard 的 *Éléments de la grammaire de la langue romane*; K.Lachmann 的文本考订法（Timpanaro 2003）; J.Bedier 的 *La tradition manuscripte du "Lai de l'ombre". Réflexions sur l'art d'éditer les anciens textes* 等论著在古典语文学的基础上，构建出了罗曼语文学的大致框架。意大利人 G.Pasuqali 的 *Storia della tradizione e critica del testo* 在校勘考据方面对 Lachmann 方法与 Bedier 方法进行了总结与提升。

好的例子。项目于 2005 年启动，由渥太华大学法语系 Laurent Brun 教授牵头，由近百人的团队编纂了中古文学作品的文献学信息并提供了大量的参考资料。以《马可波罗游记》为例，《游记》不同语言的早期文本关系错综复杂，文献繁多，而该词条则由《游记》意文赖木学本的校订者及马可波罗专家 Eugenio Burgio 和 Serena Modena 撰写，项目总协调人审订，词条内容包括生平简介、不同语言版本梳理，相关文献汇总[2]。单就法文版《游记》（*Le devisement du monde*）文本，词条提供文本基本信息，包括标题、日期、语言、类别（游记）、形式（诗或散文）及开篇文字（Incipit）、终章文字（Explicit）转写等，另外还有抄本信息、现代版本、译本、研究文献、参考文献等[3]。对于西方文本学研究者来说，最大的便利在于抄本信息的部分。手稿所收藏的图书馆及其编号，抄本内容分段介绍，抄本所有者历史追溯，抄本相关参考文献等均有罗列，而部分抄本则可以直接转至如法国国家图书馆 BNF-Gallica 项目等的数据库，在线阅读网络扫描本。在文本学领域，原稿的阅读对于中古文本编纂是必须的工作，而研究者针对原稿进行影印（称为 facsimile 或 photographic edition）或誊写（diplomatic edition），随后进行文献校对工作，形成校本（critical edition）。在非数字时代需要奔波往返于多个国家不同图书馆才能获得的信息与复印本或誊写本，通过这样的数字平台以及图书馆的数字化可以较轻松获取，打破了研究者地域上的限制，对于中古语文学研究者来说，意义重大。对我国的研究者来说，也提供了直接参考源头文献的机会与路径。当然，依靠扫描影印本来进行文本研究，也会有潜在的问题，如资源提供方没有标明，那么在古文书学（paleography）方面的某些必要信息在不接触原本的前提下则无法获得，如抄本的大小、纸质、抄本保存及损毁状况、抄工笔迹（一人或多人）、污痕等，在影印版中无法很好地体现，还是需要学者亲自现场考察。

中古语文学领域涉及到的很多图书馆，均陆续将其资源数字化，并面向公众开启。梵蒂冈图书馆的文献在几十年前尚未向学术界开启，而今天其 DIGIVATLIB项目已将部分馆藏数字化并发布上网供公众查询，其中包括手抄本、印刷本，以及近 700 多份摇篮本 Incunabola（十五世纪中叶古腾堡发明印刷术至十六世纪初这五十年间的早期印刷本）。而通过 gallica.bnf.fr，Archive.org 等网站，今天的人文学者可以轻松接触到数字手稿，而不受地域、差旅经费、图书馆时间等因素的制约。

文本、译文、图像、声音—全方位的但丁：

[2] Eugenio Burgio et Serena Modena, https://www.arlima.net/mp/marco_polo.html. 2017 年 10 月 29 日
[3] Irène Fabry-Tehranchi, https://www.arlima.net/ad/devisement_du_monde.html. 2017 年 10 月 26 日

对于非语文学研究者，即其学术研究工作不包括直接查阅手稿，而是依赖于已有的现代转写版本或译本进行文本分析的外国学者，包括大部分中国的研究者来说，更值得利用的资源则可能是一些提供了现代版本、译本及评论的平台。

以但丁研究为例，美国学者在数字化文本资源平台建设方面，做出了突出贡献。最先开始将但丁作品数字化入库的是上世纪八十年代普林斯顿大学但丁学家 Robert Hollander 教授推进的两个重要项目：达特茅斯但丁计划（Dartmouth Dante Project）和普林斯顿但丁计划（Princeton Dante Project）。两个平台的项目互为补充，共同收录但丁的《神曲》及其他作品。

达特茅斯但丁计划基于 Petrocchi 的意文校对本（田德望教授的但丁汉译本也依此为底本），最突出的特点则在于计划收集了近 80 种不同作者编写的但丁《神曲》注释，为但丁研究者提供了大量的参考及研究资料。注释者从最早的但丁之子 Jacopo Alighieri（1322），到薄伽丘（1373-75）、塔索（1555-1568）、《神曲》英译者 Longfellow（1867）、Hollander 教授本人（2000-2007），直到最新的 Fosca（2003-2015），如此庞大的注释库数据是任何纸质版《神曲》注释集都无法比拟的[4]。达特茅斯但丁计划开始于 1982-1988 年间，注释数据库不停完善，同时 2013 年也推出了新的用户界面。Dante Lab Reader允许读者同时打开四个独立窗口，便于用户在同一个屏幕上比对意文原本、英文译本，以及不同的评论注释本。

普林斯顿但丁计划与达特茅斯计划类似，但相比达特茅斯仅收录《神曲》，普林斯顿还包括了但丁的其他作品，如《论俗语》、《帝制论》（拉丁原文及英译），以及《诗篇》、《飨宴》（意、英版本）等。Hollander 教授在他的论文中介绍了项目的始末[5]。在普林斯顿但丁计划的 2.0 更新版本中，也加入了如图像、英意文音频、地图、链接等资源。

这两个项目之后，随着个人电脑的普及和数字化进程的发展，但丁研究相关的网站开始增加，其中很典型的例子包括 1994 年哥伦比亚大学的 Jennifer Hogan 主持的数字但丁（Digital Dante）项目。在此项目中，文本互文分析（intertextualité）的研究方法得到了凸显，在 Intertextual Dante平台中研究者将但丁《神曲》的《地狱》篇原文的部分段落与奥维德的多部作品相应部分进行比对，并附上简短评价，以清晰而直接的方式呈现不同文本之间的互文性。通过这个研究呈现出的但丁与奥维德之间的联系，尤其是图像学方面的关联，反过来也可以通过奥维德研究的平台 ICONOS来体现。这个平台结合了奥维德作品的文本，相

[4] https://dante.dartmouth.edu/commentaries.php. 2018 年 5 月 10 日引用

[5] Robert Hollander，《The Princeton Dante Project》. Humanist Studies & the Digital Age 3，期 1 (2013 年 4 月 24 日). https://doi.org/10.5399/uo/hsda.3.1.1596.

关联图像作品，包括考古出土文物照片、古典时期、中世纪及文艺复兴时期著名画作等等。如奥维德《变形记》第四册关于美杜莎的章节，除相应参考书目外，另提供了相关图像学展示[6]，以及该母题所对应的古典时期、中古时期及文艺复兴时期的涉及文本，其中自然包括但丁《地狱》篇的第九章中与美杜莎相对应的描述[7]。

在但丁项目中引入多媒体资料也成为数字化但丁平台建设的趋势。新的项目大量结合图像、音频、地图、时间表和其他资源，为学习和研究者提供便利。德克斯萨大学的但丁世界（Danteworlds），以及意大利但丁协会的但丁在线（Dante Online）项目，尤其是后者新推出的阅读但丁 (Leggere Dante)计划，结合了意大利但丁协会的资源优势，提供了大量评论但丁和阅读但丁的视频音频文件。

数字人文在罗曼语文学研究手段及方式方面的影响：

除去在文献目录以及文本和多媒体资料方面的贡献外，数字手段对于语文学研究人员来说，在研究方式上也有重大影响。在前数字时代需要耗费掉一位学者或一个团队数年时间才能整理出来的索引目录（concordance）[8]，在文本数字化后，可以由一位学者通过简单的后台实现[9]。当然后者在所采纳版本、体例等方面并无说明，其参考价值和科学性会有打折。

意大利国家研究委员会资助下的 Gattoweb-OVI平台（意大利早期文本档案管理网）则是个极好的词条及文本检索数字工具案例。对于罗曼语文学、但丁及中世纪研究者来说，所牵涉的古典及中古文献无数，语种也包括古法语、普罗旺斯语、加泰罗尼亚语、意大利语等多种。单意大利语方面，佛罗伦萨、米兰、威尼斯、比萨、罗马等方言导致转写的抄本上词条形态（form）有多种变化，这从《马可波罗游记》意文版各地抄本之间有的内容类似但形态各不相同上即可看出。传统或简单的搜索方式并不考虑这些变化，会导致搜索结果的缺失，而通过将检索词词条化操作（lemmatisation），则可以一并包含同一个词的不同变体，包括因书写习惯不同（variante grafica），或因发音习惯不同（variante fonetica）而引出的变体。用 Gattoweb-OVI 以词条而非词形方式搜索“主”（signore）一词，词频较高的结果中则会包括诸

[6] http://www.iconos.it/le-metamorfosi-di-ovidio/libro-iv/perseo-e-medusa/fonti-medievali/, 2018 年 5 月 17 日引用

[7] “三个染着鲜血的地狱女妖/躯体、神态和一般女人无异/身上被一条条鲜绿的多头蛇缠缭/头发则是一窝小毒蛇和角蛇/虬然盘结在她们可怕的鬓角。”但丁，《神曲》《地狱篇》第九章第 38-42 行。黄国彬译本，外语教学与研究出版社，北京，2009 年 9 月。

[8] 如但丁《神曲》索引目录：G.A.Scartazzini,《Concordanza della Divina Commedia》. Leipzig: F.A.Brockhaus. 1901.

[9] 比如 Terrill Souls 为个人高中英语教学和但丁《神曲》翻译而制作的索引网站：http://tsoules.com/dante/concordance/。

如 segnor（81 条），segnore（145 条），seignor（110 条），signor（277 条），signore（508），signori（358 条）等多种词形变体。

在但丁搜索（DanteSearch）项目中，也有类似的词条化功能。所有的但丁文本，都进行了数字词条化处理，在搜索引擎方面，还加入了 AND，OR，NOT 等不同词条之间的逻辑关系处理，便于准确找到需要的信息。

在 Gattoweb-Ovi 中，最重要的利用方式还是在文本库里针对不同关键词的组合，比对搜索后分析不同作者文本之间的关联性。如搜索《神曲》开篇首句"方吾生之半途"中的"吾生"（vita），"半途"（cammino）两词，除得到但丁《神曲》的出处外，还能看到在他的《飨宴》篇中，也有"在人生之路行走"的类似表述，另外 Jacopo Alighieri 但丁之子的注释，Ottimo 注释本，以及薄伽丘在多处文本中，均对此段话语有过评论。这样的"博览群书"式的研究结果，在之前只有满腹经纶的资深学者凭记忆或借助 Concordance 类索引册的仔细考证，才能以文本实证印证自己的推测，而当下的搜索模式在这类研究速度和准确度上均有了大量的提高。这在中古文本分析、作品流派及互文分析、作者归属考证方面，均有深远的意义，但同时语文学研究工作对学者学术能力的要求，并不会随着工具的便利和流程的简化而降低标准。

跨学科的挑战：

笔者在罗马一大就学期间，作为实验参与者参加了由罗马大学分子医学系 Cartocci 教授以及欧美及跨文化研究系 Canettieri 教授等联合开展的神经但丁计划（NeuroDante）[10]。实验主体为共 46 位 22-33 岁的研究生，分成人文背景组和非人文背景组，除一中一俄两位外国研究生外，其他均为意大利语母语学生。在实验中，让受试对象聆听约 20 分钟的录音，包括日常对话、但丁《神曲》中《地狱》、《炼狱》及《天堂》篇的部分章节，以原诗形式以及散文转写形式念出，最后以日常对话结尾。在此过程中，监控实验对象的脑电图（EEG）、心率（HR）以及皮电反馈（GSR）。通过对情感趋避性测量（approach withdrawal, AW），脑力负荷（cerebral effort）以及情感指数（EI）[11]的测量，发现人文组相对非人文组，在聆听但丁作品朗读时，体现出较高的 AW 和 EI 参数。换句话说，熟悉但丁文本的受试者在聆听时存在更多的情感反应，而并不熟悉文本的受试者则体现出较少的情感"共鸣"。"审美体验是

[10] 以下内容均节选自相关研究论文：Cartocci, Giulia, Anton Giulio Maglione, Enrica Modica, Dario Rossi, Paolo Canettieri, Mariella Combi, Roberto Rea 等. 《The "NeuroDante Project": Neurometric Measurements of Participant's Reaction to Literary Auditory Stimuli from Dante's "Divina Commedia"》. Symbiotic Interaction. Cham:Springer International Publishing, 2017. pp.52-64.； Luca Gatti, 《Poesia (e non poesia) alla luce delle neuroscienze: il progetto NeuroDante》.Cognitive Philology. No.10 (2017)

[11] Cartocci,《The "NeuroDante Project"》, p.54.

先前知识的一项功能”[12]。针对但丁《神曲》文本的独特性，也有些有趣的发现。对于意大利人来说，无论人文组或非人文组，在高中语文中已大量学习过《地狱》篇相关章节，因此，在《地狱》篇章节阅读时两组的情感指数差别不大，而在《炼狱》及《天堂》篇，两组之间显现出较大的差异。在聆听《地狱》第五章的时候，两组受试者相比散文体而言，对诗歌体的诵读有较高的情感指数，另外非人文组诗体的指数，还高于人文组的散文体情感指数。可以有这样的假设，受众在并不了解完整内容的条件下，仍可以较好地欣赏诗歌这样的艺术形式。鉴于样本量偏小以及实验的探索性本质，还有很多的后续工作需要展开，但这样的交叉学科实验也是但丁研究数字化实践的一个有趣的案例。

（以下为本人构想中的项目，其尚不成熟的示例如有机会，将在演讲时展示）

笔者本人在策划中的《塔罗牌与欧洲中世纪数字地图》项目，则希望借自十四世纪风靡意大利与欧洲的塔罗牌，尤其是 22 张大阿卡纳牌背后所蕴含的文化及图像学意义，介绍欧洲中世纪与意大利文艺复兴文化历史文学等知识。特定牌的命名及图像与中世纪及文艺复兴研究中的主题有深切的契合，由此引发学生联想，培养开拓性思维方式及综合性人文研究与理论思维能力。具体说来，如《愚者》一牌，根据其图像学特征，可以设置如“旅行者”、“朝圣者”、“迷路”等关键词，以 xml 语言建立 tag，再进一步与相对应的中古历史、绘画、文学、宗教文本进行链接。如“旅行者”，或“迷路”关键词，可与《神曲》中的但丁（迷失于人生半途的森林），《奥德赛》中的奥德修斯（十年归家之旅途）进行意义关联。“朝圣者”这一关键词，又可以与但丁《天堂》篇、乔叟《坎特布雷故事集》相关联。而与此同时，这些不同的文本、图像之间，又有互文对应，如《神曲》《地狱》篇第二十六章便是奥德修斯的相应章节。如此，通过互相之间的链接、互文，相互跳转，配以一个直观清晰的呈现界面，可构建出中古文学文化的知识体系网络，为中古研究学习者与爱好者提供一个新的探索模式。

通过对欧洲中世纪及但丁研究过程中涉及到的数字人文学项目案例的分享与分析，对于

[12] Cartocci,《The “NeuroDante Project”》, p.61.

中文世界的欧洲中古文学及罗曼语文学研究者来说，但丁研究的种种数字化项目为学生学者接触原本、原文及译本、评注本提供了极大的便利，而丰富的多媒体资料无论是在教学层面还是在研究层面，都具有极重要的辅助价值。在此基础上，不同背景的学者联合展开的跨学科研究方式，则在如认知语文学等交叉学科领域有所突破，这样的经验与方式方法也可以供文学研究者参考借鉴。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院 Dharma Drum Institute of Liberal Arts (DILA)

倫敦大學國王學院 30 年數字人文研究熱點分析

包晗* 张力元**
伦敦大学国王学院数字人文系博士生*
北京大学信息管理系博士生**

伦敦大学国王学院 30 年数字人文研究热点分析

包晗 博士生 伦敦大学国王学院数字人文系
张力元 博士生 北京大学信息管理系

摘要

数字人文（Digital Humanities），也被称为人文计算，是针对计算与人文学科之间的交叉领域进行学习、研究、发明以及创新的一门学科。伦敦大学国王学院(KCL)的数字人文研究最早可以追溯到七十年代初期，是当时世界上少数几个在艺术人文和社会科学学科中，应用计算方法和工具的大学之一。2002 年 KCL 数字人文系（DDH）正式成立，在英国教学排名(REF)排名第一，处于世界领先地位，主要涉及领域有：数字文化与社会中的数字媒介、数字方法与数字设备的发展、数字社区参与的平台与渠道等研究方向。KCLDDH 发展至今，学院设立一个本科专业：数字文化（Digital Culture），五个硕士专业:数字人文（MA DH）、数字资产媒介管理（MA DAMM）、大数据（MA Big Data）、数字社会文化（MA DSC）、数字策展（MA Digital Curation）和一个博士专业数字人文(DH)。伦敦大学国王学院数字人文热点研究领域包括：文化生产与创新型经济、数字和日常文化以及记忆和知识环境等研究方向。本文通过文献计量学、聚类分析、社会网络分析研究以及深度访谈方法，论述伦敦大学国王学院三十年来数字人文研究的趋势、和交叉领域学科分布、以及领域的细分分析。

目次

1.引言
2.研究方法
2.1 数据来源
2.2. 数据分析
3. KCL 数字人文的基础统计分析
3.1. 时间分布
3.2. 语言分布
3.3. 期刊分布
4. KCL 数字人文的研究热点分析
4.1. 词云
4.2. 关键词词频统计
4.3. Gephi 关键词共现网络分析
5. 分析与讨论

关键词

数字人文，伦敦大学国王学院，数字文化，演化路径、热点分析

1.引言

数字人文（Digital humanities, DH）,源于人文计算（Humanities Computing），是在计算技术、网络技术、多媒体技术等新兴技术支撑下开展的人文研究新型跨学科研究领域，早期的人文计算发端于文学和语言学领域，在新兴技术发展的强劲支撑下，伴随着人文资料的数字化以及网络分享，人文研究的方法发生了重大变革。学科领域的许多学者都致力于将技术融合进入学术研究，例如基于文本的学科分析（古典文学、文学、哲学和思想史）、地理信息系统（GIS）、互动游戏和多媒体等在历史、 哲学、文学、宗教学和社会学等学科的应用。基于认识论，数字人文可以被归纳为两个问题：如何了解到那些我们无法了解的知识；如何推测那些我们不知道的信息。基于方法论，它被归纳为利用知识产生、分散、收集的手段来来对人文学科进行补充。

伦敦大学国王学院教授威拉德麦卡蒂(Willard McCarty) 认为这些在原则上都可以归类为在已存在的客体与我们推测的模型之间建模。罗伯托布萨(Roberto Busa) 认为，计算科学主要作用并不是加速人文学科的进步，而是为人文学科领域中长期存在的问题提供的的方法，从广度和深度两个维度重新定义和解构人文学科研究。Willard McCarty 教授及其同事 Harold Short 对图 1 进行了深入阐释，他们认为 “图中央区域指的是数字人文研究的方法论共同基础，它们是数字人文的核心，包括各种可计算的基础数据对象，如自由文本、格式化数据、图像、声音等。针对这些数据而进行的计算活动包括文本分析、数据库设计、数字绘图、音乐检索等。”

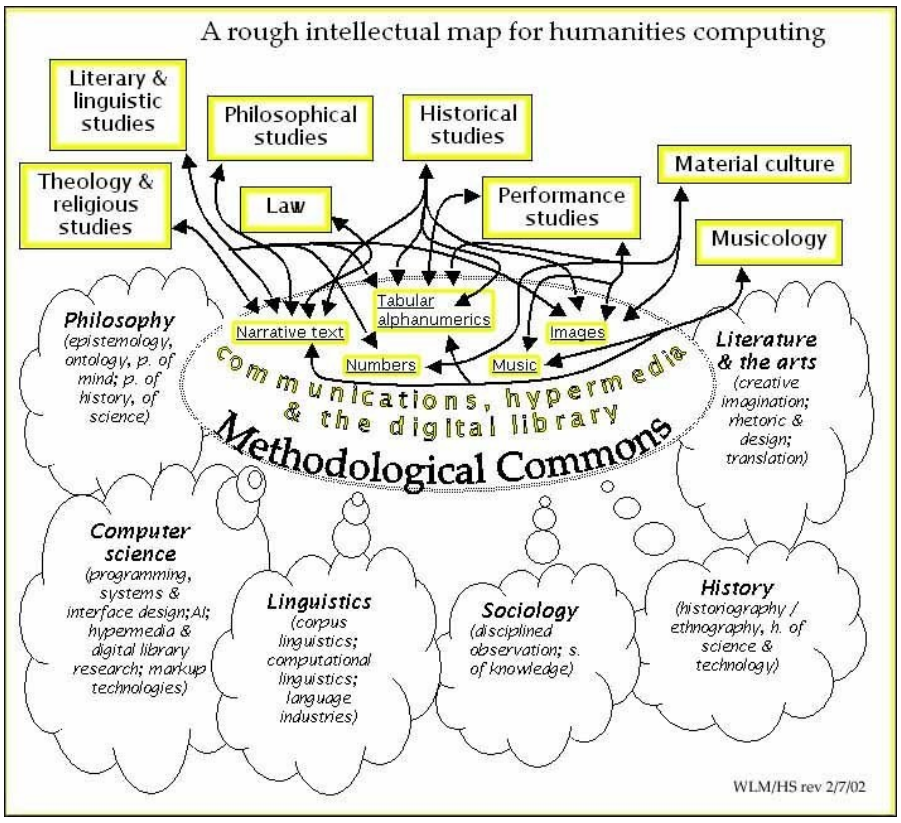


图 1：A Rough Intellectual Map for Humanities Computing

在这样的一个背景下，本文希望能够对伦敦大学国王学院的数字人文发展进行一项综述性的研究。伦敦大学国王学院的数字人文（Digital Humanities）由人文计算（Humanity

Computing）发展演变而来。伦敦大学国王学院（King's College London） 由英国国王乔治四世建于 1829 年，是世界顶尖的综合研究型大学。伦敦大学国王学院以其优质的教学科研实力享誉世界，2017QS 综合排名世界第 21 位。艺术人文学院（Faculty of Arts & Humanities）成立于 1989 年，是全英最前沿的人文科学学术机构之一，2018 年 QS 世界排名第六位，学院提倡多样化的学科及课程设置，涉及人类文化及历史的广泛学术领域，伦敦大学国王学院人文和艺术学院下面由 15 个学院组成，包括古典学、语言学、历史学、哲学、神学与宗教学等学科，目前共有在读学生 4828 人。数字人文系（Department of Digital Humanities）是艺术人文学院的下设系所，是目前世界上在数字人文领域规模最大、最具声望的学术机构之一。该系致力于探索前沿科技应用于人文社会科学研究的可能性，通过联合英国内外不同学科的学术伙伴，相继开展包括艺术文化领域移动应用发展、数字公民赋权等超过 30 项重要研究项目。

伦敦大学国王学院的数字人文研究最早可以追溯到七十年代初期，是当时世界上少数几个在艺术和人文学科中应用计算方法和工具的大学之一。2002 年，数字人文系正式成立，拥有数字文化与社会中的数字媒介、数字方法与数字设备的发展、数字社区参与的平台与渠道三个主要研究方向。以数字人类（Digital Human）作为核心视角，在理论、研究方法及社会参与三个方面结合大数据分析，具体研究包括通过档案构建原始数据库（EHRI），与牛津、剑桥等大学共同开展历史与人文研究，以及推动公共政治参与、争议性知识等方面的创新性人文研究，为人文学科提供了新的研究视角和理论方法。数字人文项目是数字人文建设的关键，一般具有六大特征：面向主题（Subject Oriented）、关注本体（Ontology Focused）、基于数字仓储（Digital Repository Based）、跨学科整合（Cross Disciplinary Integration）、多机构协同（Collaboration between Multiple Institutions）、持续性开发（Long-term Development）。

本文将利用文献计量分析方法，结合共词分析法和社会网络分析法以及深度访谈，对 KCL 数字人文 30 年来的研究文献进行统计分析和内容挖掘，分析 KCL 数字人文研究的前沿和热点领域，并结合内容分析法，对现有研究的主题和主要观点做详细分析，以期对国内外数字人文的研究和发展提供借鉴参考。

2.研究方法

2.1 数据来源

本文共有三个数据源，分别为 KCL 的 Arts & Humanities 中 DH 研究小组官网，Web of Science，和人工统计数据。其中 KCL DH 官网的数据和 Web of Science 的数据应用于定量分析中，而人工统计的数据应用于定性分析及对定量结果的解读。因为本文的研究重心是在 KCL 的 DH 中心三十年的演化，因此是以第一部分，即 KCL 官网的数据为主。采用了 python3 + scrappy 的爬虫模式，爬取了 1029 条记录，经去重等初步清理，获得有效记录 854 条，占比 83%，涵盖论文、著作、书籍章节等出版物。每条记录包括：’year’, ’title’, ‘abstract’, ‘Journal’, ‘language’, ’author’, ’url’七个字段，并以 Mysql 数据库存储。数据清洗工作包括对年份的对齐，对缺失项的补充等。鉴于官网上的数据可能存在缺失，因此本文以 Web of Science 的数据资源加以补充。通过多次试检索，兼顾查全率和查准率，本文选取了 Web of Science 上的全部数据库，以”digital humanities”和”humanities computing”作为主题检索词，以”king’s college London”作为地址检索词，时间截止至 2018 年。具体逻辑关系检索式为：AD=king’s college London AND (TS=digital humanities OR humanities computing)。并辅以如下表 1 中所列举的可能和 DH 相关的检索词对搜索结果进行补充。经过人工筛选，剔除与数字人文主题无关文献后共获得 89

篇文献（论文）。本文通过人工审核，将这 89 条记录与在 KCL 官网上爬取的 854 条记录进行整合，以’year’, ’title’, ‘abstract’, ‘Journal’, ‘language’, ’author’, ’url’七个字段提取 Web of Science 上获得的记录的相关信息，补全到数据库中，剔除重复项，最终获得 894 条有效记录，作为本文定量研究的基础数据。

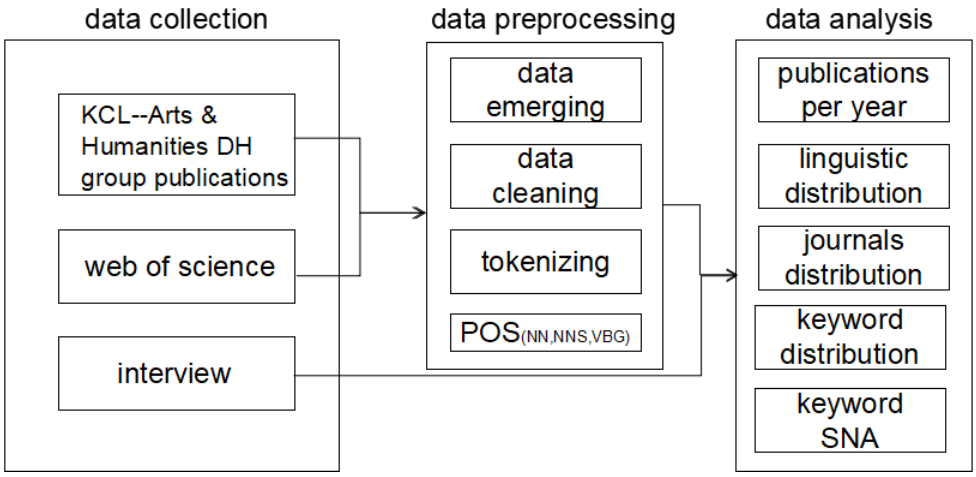


图 2：数据收集、预处理、分析流程图

此外本文还结合了定性访谈方法，进行人工统计的数据统计和整理。目前，国王学院数字人文学院共有学者专家 43 人、教学研究员 13 人、博士研究生 37 人，通过采访终身教授、数字人文学科奠基人 Willard McCarty、院长 Tobias Blanke 教授、副院长 Simon Tanner 教授、数字人文高级讲师、执行院长 Stuart Dunn、大数据讲师、大数据中心主任 Mark Coté、数字文化与社会讲师、数字文化研究中心主任 Paolo Gerbaudo，对 6 人进行了取样深度访谈，对其他 30 余人讲师学者进行了非结构化访谈，例如讨论出版论文领域方向、自身研究兴趣和数字人文结合等。问题针对受访者来 KCLDH 的时间以及所属的系和近期开展的数字人文项目的教学与研究工作，来收集伦敦大学国王学院 30 年数字人文研究的主题演化路径与热点领域分析，并征求了受访者对数字人文学科发展和 KCL 学科建设方面的建议和看法。采访结合结构化（Structured in-depth interview）、半结构化（Semi structured in-depth interview）和非结构化（Unstructured Interview）的混合形式，获得丰富的细节和相关事实之间的意义关联，访谈时间为 2018 年 8 月 20 日到 2018 年 10 月 10 日， 每次访谈持续约 1 小时。在分析部分，会对访谈结果进行原话引用、转述、情境描述等真实再现，对访谈结构进行数据统计和文本分析，结合 KCL 出版物的基础统计分析对数字人文的历史进行回顾。

表 1：web of science 检索词

主要检索词	辅助检索词
digital humanities	digital asset & media management
humanities computing	digital culture & society
	big data in culture & society
	digital curation
	digital history
	computational linguistics
	digital art
	digital archaeology
	digital classics

	digital comparative literature
	digital culture
	digital media and creative industries
	digital film studies
	digital theology& religious studies
	digital liberal arts
	digital linguistics

2.2 数据分析

本研究将从学术论文发表数量和研究主题的发展与变化分析 KCL 数字人文方面三十余年来研究的演化路径。首先，学术论文数量的时序变化是衡量某一研究领域发展的重要指标。通过对数字人文文献的发表数量，绘制相应的时间分布曲线，能够明晰其所经历的发展过程，所处的发展阶段，并预测未来发展动态。关键词是表述论文中心内容的实质词汇，是论文的精髓。通过对关键词词频与中心性的统计．能够掌握数字人文研究领域的主题分布，发现隐藏在真实关系背后的关系网络，探究研究主题的成熟度、知识结构等状况。

本文使用关键词共现网络分析 KCL 数字人文方面相关研究的热点。本文将绘制数字人文关键词共现图谱，并加以聚类，每个聚类可以反映数字人文领域的一个研究方向。采用关键词作为聚类标识。以具体化的名词短语标识各聚类。以便客观归纳数字人文的研究热点，并对每个聚类内的具体文献进行内容分析，梳理研究热点问题。

在解释研究热点的时候，本文首先使用词云的方式对获取到的标题和摘要进行粗粒度的高频词的可视化展示。主要针对语料使用词云模型可视化显示数字人文领域研究热点词汇：1）所有出版物的标题 2）具有摘要信息的文章的摘要

由于本文的数据是多源异构数据融合得到的数据集，因此无法利用常见的 citespace 或 SATI 等软件进行数据处理。因为源数据没有关键词，仅有 43%的数据有摘要信息。因此为了避免结果产生偏差，本文以文章为单位，使用 python 对标题进行关键词的提取和生成。笔者考虑了 TF-IDF，TextRank，Word2Vec，RAKE 等几种常用的文本关键词提取方法，但由于这些方法对语料的规模均有一定的要求，对于本文这种单文档字数很少的情况得到的效果不佳。因此本文决定使用 NLTK 对每条记录的文本进行分词、去除停用词和标点、标注词性的预处理，并且仅保留具有名词词性（NN,NNS,VBG）的词语作为每条记录的代表性词汇，并对关键词的阈值设定为，即关键词需要在全文本里出现 1 次以上，以此去除那些很少见或错误的词汇。然后对生成的关键词构建关键词共现矩阵。并将关键词共现矩阵转化为由 Source 到 Target 的点对形式导入 Gephi0.9.2 进行可视化展示，模块化聚类以实现对热点领域的探究。Gephi 的模块化功能进行社区聚类，基于一种启发式快速社区发现算法。这个算法的大概思路是：首先为网络中的 N 个节点各自分配一个社区，然后将节点移动到使模块度正向增加最大值的相邻社区中，重复这一过程直至最后没有可变项。然后将每个社区视为一个节点，新节点之间的链路的权重即内部链路权重之和，重复上一步。在不断的迭代过程中，元社区的数量不断减少，直到无法改变时得到最终结果。设置随机+使用权重+解析度等于 1 的模式。

3.KCL 数字人文的基础统计分析

3.1 时间分布

本文通过统计出版物数量随时间的分布来探究 KCL 数字人文发展的趋势和所处阶段，具体分布如图 3 所示。

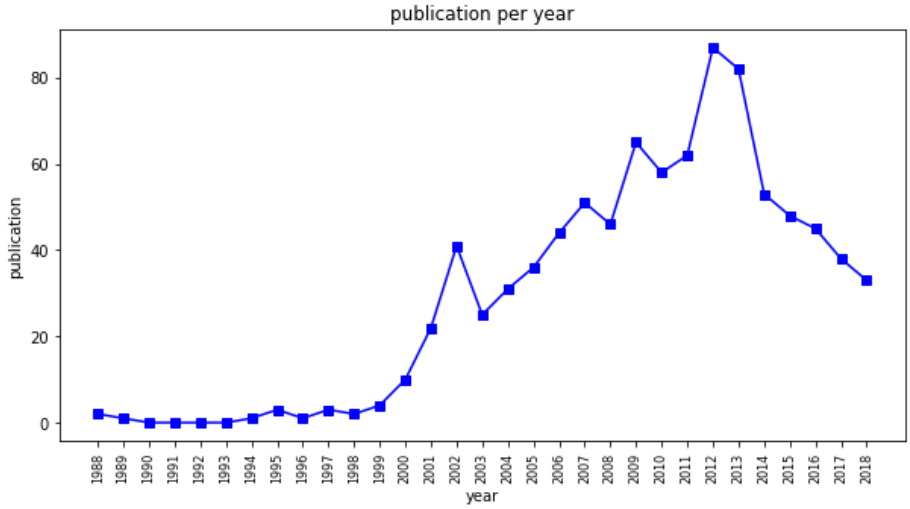


图 3: KCL DH1988-2018 年发表物数量统计

通过统计分析发现最早的一篇关于数字人文的论文 Tim Jordan 和 Paul Taylor 1988 年发表在《社会学评论》（The Sociological Review）的题名为《黑客社会学》A sociology of hackers，该文主要介绍了信息社会中，通过阐释黑色社会学和黑客社区的人口统计数据和社区成员的构成流行性，讨论了黑色攻击文化的本质。KCL 数字人文研究论文在近 30 年间虽各年略有波动，但整体仍处于上升状态。如图 3 所示，1999 年是 KCL 数字人文研究的重要转折点，之前有关数字人文研究的论文数量缓慢上升，而之后的论文数量快速增长，这与国际数字人文领域刊物的增长模式是一致的。KCL 数字人文的研究发展大致可以分为 4 个阶段：萌芽阶段（1988 年-1999 年）。这一阶段的发文量平稳，数字人研究的焦点在于人文计算，对象从电子文本逐步扩展到超文本、图像、视频、音频、数字地图、网页、虚拟现实、3D 等多媒体技术，计算的领域也不再单单只在语言学领域，像考古、古典学、地理学等拓展。起步阶段（2000-2012 年），该阶段的研究文献出现大幅的增长，研究主题主要集中于考古、文化遗产、文学、历史、音乐、艺术等多个领域。发展阶段（2012-至今），主要特点是 KCLDH 和大英图书馆、大英博物馆以及英国艺术与人文社科基金合作，开展数字人文项目科研与实践，利用数字出版、数字可视化技术、地理空间展示、模拟空间以及大数据网络信息管理系统，阐释数字社会即数字化社会在新媒体环境中的全球化等特点。

按照数字人文学者苏珊·霍基教授的划分，数字人文的发展（主要在北美及欧洲地区）可以分为四个阶段。[参见《数字人文指南》（A Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth, Oxford: Blackwell, 2004）。

1970 年代到 1980 年代中期是“联合”阶段；
1980 年代中期到 1990 年代早期被霍基教授称为 “新发展”阶段；
1990 年代早期到现在的 “互联网”时期是数字人文的成熟阶段；
在 20 世纪 90 年代末，数字人文项目开始利用数字可视化技术、地理空间展示、

模拟空间以及复杂系统的网络分析开始分析。

在时间分布上，由“数字人文组织联盟” (成立于 2005 年)（Alliance of Digital Humanities Organizations) 组织的、全世界最大的数字人文大会讲数字人文的边界扩展到了一个非常宽泛的范围，2002 年国王建立数字人文学院 Digital Humanities Department 文献发表达到一个小峰值,2005 年到 2012 年快速发展，“新媒体研究”与“互联网研究”的出现为数字人文的繁荣发展以及学科融合注入了动力，此外，GIS 技术以空间数据库为基础，通过原始数据按照时间顺序，空间关系在历史地理研究中发挥优势。2012 年 KCL 文献达到了峰值，人文与艺术学科采用运算系统和计算媒体发生的整体学科范式“计算转向”（Computational Turn）,研究始从超文本等介质被数字化转换性研究材料逐渐扩展，并聚焦到了原生性数字材料，并增加了原生性数字材料的历史维度，比如媒介考古学、互联网历史研究等,以及可再生材料。此外基于自然语言分析、统计方法的“传统”的文本分析已经进入方法的扩散和多样化的阶段，特别是随着 R 语言、Python 语言在人文学者中的日益普及，学者们越来越多地自行开发各种“定制化”的工具包来解决人文研究中的特定问题。而 GIS 技术和 HGIS 也被大量应用，成为数字人文中的基本方法与途径之一。相对较新的 VR 和 3D 建模也开始从考古、建筑走向了历史、文学与艺术领域。图像识别及数据可视化也成了这几年的热点议题，“数字人文” 作为“人文和艺术学科”的本质，“数字之术”和“人文之道”相结合，积极致力于数字人文认识论创新界定以及方法论的创新实践，相比于国际的数字人文整体发展的趋势，可以看到 KCL 数字人文学科整体发展是处于国际领先地位。

3.2 语言分布

本文对出版物的语言进行统计，获得如表 2 排名。

表 2：出版物语言分布

语言	计数
English	834
Undefined	31
Italian	12
Spanish	6
French	5
German	4
Finnish	1
Multiple languages	1

从语言分布的统计中，可以看到占大部分的是英语发表物，这可能与发表人的母语大部分是英语，以及刊物的接受语言多为英语相关。此外，意大利语，西班牙与，法语，德语，芬兰语都有所涉及。这表现了 KCL 的研究人员背景多样性，以及其在超越了工具与对象的关系，是一个相互渗透、彼此强化的学科。

3.3 期刊分布

本文对收集到的记录进行期刊的统计，获得如表 3 排名。

表 3：学术论文期刊分布

期刊	计数

Literary and Linguistic Computing (DSH)	34
CLASSICAL REVIEW	15
Digital Medievalist	9
Journal of Digital Information	8
Digital Humanities Quarterly	6
Information Communication & Society	5
Ariadne	4
The Sociological Review	4
INTERDISCIPLINARY SCIENCE REVIEWS	4
D-Lib Magazine	4
International Journal of Humanities and Arts Computing	4

期刊分析统计记录所示，前三位的文献分别是文学与语言学，古典学，数字中世纪历史学，分别 34，15，9 次记录。人文学科的大厦深深根植于古典语文学的相关领域，例如考古学、艺术史和逐渐脱胎于文本研究的语言学。DSH（digital scholarship in the humanities）是最受欢迎的发表刊物。DSH 是 SCI 三区的期刊，在数字人文领域拥有十分权威的地位。DSH 的前身为 Literary and Linguistic Computing，是在数字人文领域比较偏向于文本的自然语言处理的技术和应用的期刊。Classical review 则是侧重于人文评述方向的期刊（需核实）。这说明了在 KCL 数字人文的研究中，既有偏向于技术处理方向的研究，也有侧重于人文的研究。

4. KCL 数字人文的研究热点分析

4.1 词云

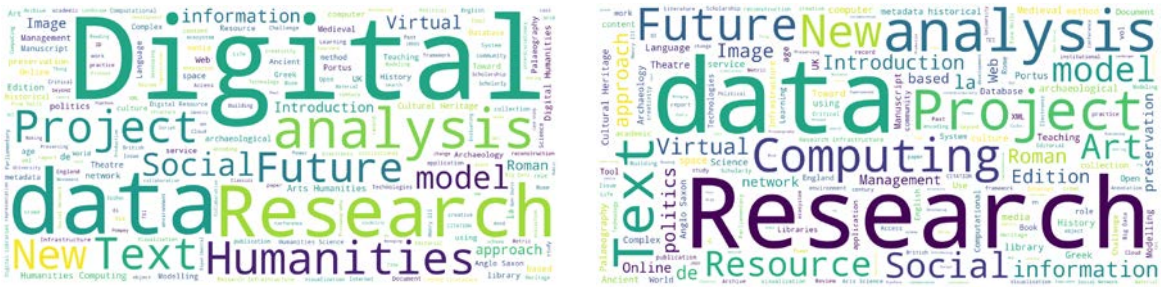


图 4：表题词云（左）、摘要词云（右）

从词云的统计中，可以看到在 research, data，text, information, practice, society, humanities 等概念是最为常见的。这表明了在 KCL 的研究中，具有将社会科学领域的某些研究方法引入人文领域，通过信息技术工具软件或规模化数据为人文研究提出问题、界定问题和回答问题提供新的视角的特点。美国伊利诺伊州立大学香槟分校图书馆和信息科学研究生院教授约翰·昂斯沃斯（John Unsworth）的观点，认为数字人文的主要范畴是改变人文知识的探索（discovering）、标注（annotating）、比较（comparing）、引用（referring）、取样（sampling）、阐释（illustrating）与呈现（representing），实现人文研究方法论上的创新发展

4.2 关键词词频统计

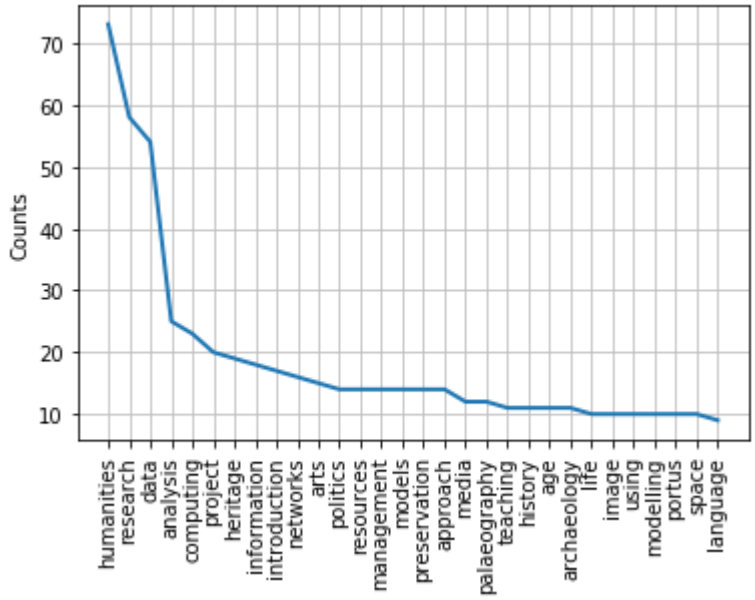


图 5：频次 Top30 关键词

表 4：TOP30 关键词

高频词（1-10）	计数	高频词（11-20）	计数	高频词（21-30）	计数
humanities	73	arts	15	history	11
research	58	politics	14	age	11
data	54	resources	14	archaeology	11
analysis	25	management	14	life	10
computing	23	models	14	image	10
project	20	preservation	14	using	10
heritage	19	approach	14	modelling	10
information	18	media	12	portus	10
introduction	17	palaeography	12	space	10
networks	16	teaching	11	language	9

本文保留了标题中的有效语义信息，并没有将 humanities, research, data 等词汇作为停用词去除，是为了保留其原始语义，洞察 KCL 研究人员的研究与用词习惯。如图 5 和表 4 所示，KCL 数字人文出版物的关键词中，具有通用含义的 humanities, research, data, analysis 最常出现在标题中，这样的标题可以有效的表达研究是属于数字人文领域的。Computing, information, networks, models 等词汇强调了研究的“数字”属性。而 heritage, arts, politics, history 等词强调了研究的“人文”属性。

4.3 Gephi 关键词共现网络分析

表 5：关键词共现矩阵（部分）

	humanities	research	data	analysis	computing	project	heritage	information	introduction	networks	arts	politics
humanities	0	14	12	0	15	0	0	1	0	0	12	0
research	14	0	11	1	3	1	1	3	0	0	1	0
data	12	11	0	2	0	1	1	2	0	0	3	0
analysis	0	1	2	0	0	1	0	0	0	0	0	0
computing	15	3	0	0	0	1	0	0	0	0	0	0
project	0	1	1	1	1	0	1	0	0	1	0	0
heritage	0	1	1	0	0	1	0	1	0	1	1	0
information	1	3	2	0	0	0	1	0	0	0	0	1
introduction	0	0	0	0	0	0	0	0	0	0	0	0
networks	0	0	0	0	0	1	1	0	0	0	0	0
arts	12	1	3	0	0	0	1	0	0	0	0	0
politics	0	0	0	0	0	0	0	1	0	0	0	0

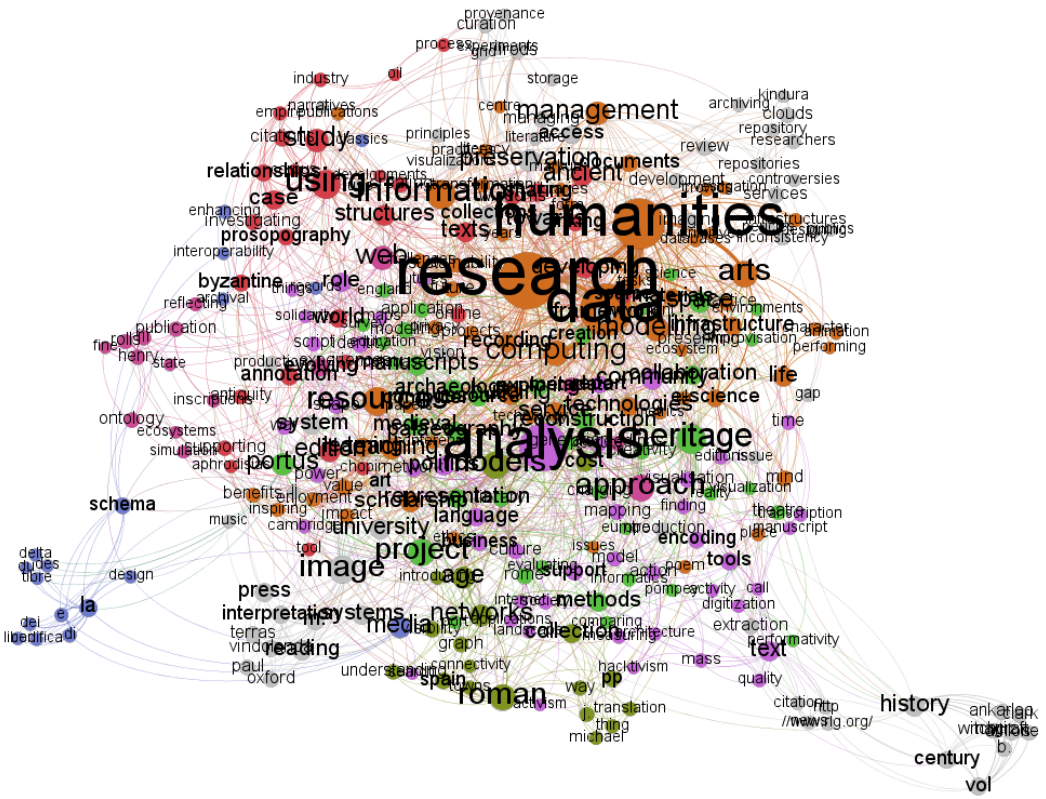


图 6：关键词聚类

本文按照 Gephi 自动化共探测到 24 个社区，选取其中最大的最具有代表性的 5 个社区作为领域热点讨论对象。具体选择了每个社区中的度排名在前 30 的关键词，如表所示。

表 6：聚类结果

Cluster	Keywords(1-10)	Keywords(11-20)	Keywords(21-30)	Description
	research	life	e-science	
	humanities	scholarship	materials	
	data	building	value	
	information	technologies	libraries	
NO1	resources	infrastructure	mind	General
	computing	learning	practice	
	arts	service	benefits	
	management	future	sustainability	
	modelling	documents	infrastructures	

	teaching	recording	projects	
NO2	analysis	network	model	Qualitative Analysis
	community	culture	power	
	text	business	war	
	role	identity	maps	
	visualisation	time	digitization	
	politics	support	finding	
	language	script	society	
	report	action	texture	
	tools	shape	futures	
	encoding	mapping	measuring	
NO3	heritage	representation	charging	Humanity
	project	archaeology	generation	
	portus	resource	theatre	
	space	exploring	informatics	
	palaeography	modeling	england	
	manuscripts	creation	evaluating	
	reconstruction	rome	reality	
	methods	library	visualization	
	collaboration	application	survey	
	medieval	creativity	science	
NO4	case	creating	industry	Text Analysis
	texts	prosopography	empire	
	world	antiquity	making	
	developing	investigating	aphrodisias	
	sharing	citations	tool	
	annotation	online	oil	
	source	process	developments	
	relationships	supporting	archives	
	linking	wisdoms	control	
	byzantine	experiences	corpus	
NO5	media	interoperability	zoom	Foreign Research
	schema	delta	della	
	records	tibre	school	
	design	regarding	turn	
	enhancing	progetto	digitales	
	classics	modelli	connecting	
	liber	scrittura	experiment	
	archival	carolina	borders	
	codifica	finance	example	

从关键词聚类分析可以看出，关键词主要集中在人文学（Humanity）、文本分析(Text Analysis)、定量分析(Qualitative Analysis)等。人文学者较多采用定性研究，其创作或研究较多基于基础材料进行思辨、演绎推理、解释、叙述，而且个体性、主观性、灵活性强、无固定的范式，社会科学学者较多采用定量研究，其创作或研究结果较多基于对基础数据的统计分析得出，具有一定的规律性且可以在一定条件下重复再现。量化、大数据的方法更新迭代是非常快的，定量与定性不可能做到严格的泾渭分明，定量的方法或者看上去神秘的“大数据”，并不会自动转化为有效的“信息”进行关系型网络数据库的深度学习运算，都需要社科人文学者进行定性解读，定量方法能够拓宽艺术人文社科的问题同时，也要需要和专业领域学者合作，提出专业的分析和解构阐释，才能够更高效和精准地处理更多资料，挖掘出隐藏在文献背后的信息，提高人文研究的工作效率。此外，在西方，“数字人文”迅速取代“人文计算”成为一个广泛传播的新兴跨学科研究领域的代名词。

5.分析与讨论

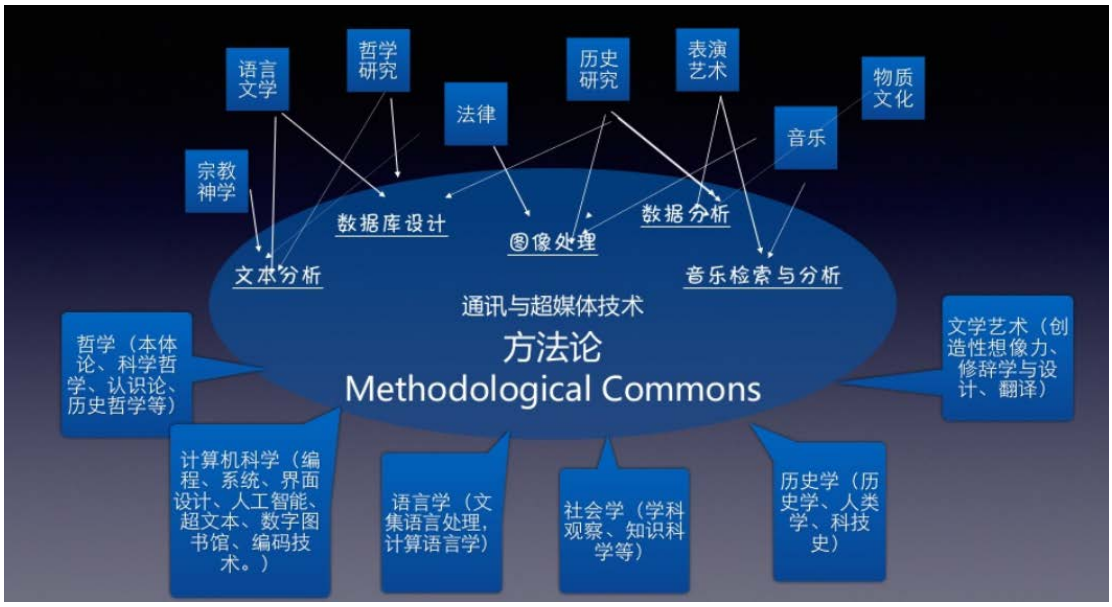


图 7：Willard MaCarty 数字人文全景图（2002.5 ）

笔者采访了伦敦大学国王学院的威拉德麦卡蒂（Willard McCarty）教授，Willard McCarty 是人文计算领域的知名教授。Willard McCarty 教授强调数字人文学不是简单的数字化加人文学科的叠加。西方人文研究和人文学的基础可以追溯至中世纪的文化、逻辑、修辞学三大科目和算数、几何、天文和音乐四大科目，现代人文学的根源在于文艺复兴，从中世纪统治的神权世界转向以人为中心的世界观。早期 KCL DH 人文领域学科的发展得益于早期主义，古典时期作品语料库的恢复编辑，这类语料库保存在拜占庭和伊斯兰学习中心的希腊文和阿拉伯文手稿中，随着中世纪第一批大学建立，大学和修道院的缮写开始出现，谷登堡印刷机的发明，第一次在欧洲应用了活字印刷术，培养了大批的对科学、艺术、文学、宗教以及世俗文化感兴趣的读者。随后，通俗文化以及文学形式的发展，进一步扩大了人文表达的范围，拉丁语、希腊语、古英语、挪威语、法语、德语以及其他语言的翻译本开始流行。印刷术推动了人文主义文化资料的标准化和传

播，同时促进了编辑技术的发展和完善，与文艺复兴时期以及后文艺复兴时期的印刷文化蓬勃发展类似，数字革命带来了文化资料开始像数字媒体迁徙。 McCarty 教授用腓尼基商人经商之道来比喻，数字人文发展过程中学者的实践经验应用价值， 他反对数字人文仅仅是一个学科定义，而是一个高度交叉融合的领域 ，借鉴美国理论物理学家哈佛大学教授戴森（Peter Galison）的观点，人类学独特的 “交易区”（Trading zone）[1]来比喻，实验家、工程师、数学家、理论家之间的分歧与合作，科学界与工业界及政治家的关系，实验室与大学之间的权力斗争等, 但是有别于社会建构论者们所关注的意识形态斗争——右派对左派、资本主义对社会主义、男性对女性等，应该关注于更物质化的因素的影响。数字人文的发展如同在一块崭新的画布上绘画，而不是从一个文化走向另一个文化，从一个技术走向另一个技术（McCarty 1999）。 数字人文学者和出版物可以被比喻成为一个“群岛”，其最显著的特点是有群岛的疏远感和也可以创造出“人类学核心事件的碰撞”，例如计算应用于人文科学，要超越目录、年代表以及英雄历史，基于量化数据的分析，拥有更多创新研究视角。

在传媒学研究方面， 大数据研究中心主任 Mark Cote 是加拿大多伦多传播 学派（Toronto School of Communication）[2], 多伦多传播学派一向以胆大睿智和最具原创性的媒介理论而闻名，他们考察媒介与文明之间的关系，分析传播媒介的偏向性和基于虚拟现实的赛伯空间和赛伯文化给人们带来了的媒介新体验，强调传播媒介的时间偏向和空间偏向之间的平衡对社会稳定的影响，提出了“媒介即信息”、“媒介是人体的延伸”的新主张，研究新技术和新媒体的发展对文化和社会所产生的影响。 提出了一种在人类文化结构和人类心智中传播居于首位的新理论和新的媒介分析技术，对后来的传播学发展产生了重大的影响。 “媒介即信息” ，如麦克卢汉所言，媒体决定了我们的处境。海量数据并不能引导我们发掘更多的真相，相反它只会引出更多的问题。海量数据虽然可以提高洞察能力，可以让学者从中找到 “是什么” ，却不一定能让我们找出 “为什么”。“大数据以人为中心”、“拥抱不可预测性”、“数字主义”等理念逐渐成为人文学术研究的主流。人文追逐着数字远行，GDP 或经济效益考量的是数字结果的现实利益，而这些理性冰冷的数字逻辑背后，如何平衡生存价值、道德尺度和环境代价，则需要温情的人文关怀。数字人文是一个典型的文理交叉学科，由于人文学宽泛广博的特性，一般不适用简单的文字表述和界定。在西方，数字人文并不是“数字之术”和“人文之道”两者之间更不存在相互依附的关系。

2002 年，数字人文学院（Digital Humanities）由人文计算学院（Humanity Computing Center）发展成立，拥有数字文化与社会中的数字媒介、数字方法与数字设备的发展、数字社区参与的平台与渠道三个主要研究方向，至今发展为一个本科专业，五个硕士专业和一个博士专业。根据 Gephi 关键词共现网络分析和词云显示，人文 (Humanities) 和文化（Culture）为高频词，集合笔者访谈，KCLDH 有大量研究考古、历史、古典学的教授学者。下面就数字资产媒介管理、考古和文化遗产、大数据和社会、数字文化等几个方面进行举例阐述。

在数字资产媒介管理（Digital Asset Media Management, 简称 DAMM）研究方面，主要研究如何加强管理能力和应对市场变化与客户需求应变能力。数据资产管理理论上是

[1] 彼特·盖里森提出了独特的“交易区”(Trading Zone)理论,试图通过结合实例对科学活动的机制(Dynamics)、区位(Locality)和场所(Place)的分析,描绘出构成科学的不同亚文化之间、信念和行为之间的局部性协调(Local Coordination)。

[2] 多伦多传播学派（Toronto School of Communication）是上世纪 60 年代由加拿大多伦多大学学者哈罗德·英尼斯（Harold Innis）、艾瑞克·亥乌络克（Eric Havelock）和马歇尔·麦克卢汉（Marshall McLuhan）共同创立的。

对于数据资产从采集，输入，（用元数据）描述，存储，检索，分析，循环利用，到安全隐私管理、知识产权管理的，贯穿其生命周期的长期的管理模式；在实践中，还会运用到配套的数据资产管理软件，以自动化流程化可控化的管理流程来应对不同数据类型、满足企业与部门的整体发展需求（Regli, 2016）。代表人物有 Simon Tannar 教授 和 Tobias Blanke 教授。Tobias 是数字人文系社会和文化信息学教授，欧洲艺术与人文数字研究基础设施（DARIAH）主任。他领导和管理大型国际跨学科研究项目，并在教学本科和研究生，致力于开发数字方法和大数据的新颖教学方法，以了解文化和社会, 例如，Tabias 和 Stuart 于 2006 年发表了《英国的艺术与人文科学电子科学倡议》（The arts and humanities e-Science initiative in the UK），2016 年发表《管理：异常和安全的主题算法》（Governing others: Anomaly and the algorithmic subject of security）。

在考古和文化遗产研究方面，在 2000 年初期 DH 的出版物开始出现并蓬勃发展，学者对该交叉领域的研究兴趣也日益浓郁，随着进化论的“理论转向”、媒体考古学（Zielinski 2006）和平台研究（Bogost and Monfort n.d）的兴起，公众参与和公众考古开始流行，知识的力量不仅取决于其本身价值的大小，更取决于它是否被传播及传播的深度和广度。大众传媒和考古、文化遗产学科的合作受到的学界的关注，考古不再是一门默默无闻的人文学科，也不再局限于专业领域人员，互联网的发展也为考古学者和传媒学者，计算科学学者的合作具备了可能性和必要性。例如，Mark Hedges， Stuart Dunn 和 Tobias 在 2009 年也出版了《艺术与人文科学电子科学-当前的实践和未来的挑战》（Arts and humanities e-science—Current practices and future challenges）， 提供了对英国艺术和人文电子科学的实践分析，以及如何适应调整电子学科的议程挑战， 在 2007 年讨论的第一阶段，2007 年以后更系统地研究数字艺术和人文科学研究中的重大挑战以及提供相应的解决方法和实践技术。但随着网络改变协作和沟通并模糊学术和非学术世界之间的界限，2013 年 《众包作为人文研究基础设施建设的组成部分》（Crowd-sourcing as a Component of Humanities Research Infrastructures），提出了描述和分析学术人文人群的框架，并利用这种“原始”框架，探索众包和人文科学研究基础设施之间潜在关系的基础。

在大数据和社会学研究方面，大数据研究作为一种新的范式尚不完善，其中包含着没法调和的矛盾。随着社交媒体、智能手机与可穿戴设备的普及和发展，媒介使用过程会产生海量的、可得的、充满意义的自然用户数据（Naturally occurring data），这与学科体制转变的速度很慢有关。在智力上，这与对这些非言语类的知识客体缺乏认识有关。非言语类的知识如何能被长期保存的问题。例如针对如何研究以大数据为特征的社交媒体，伦敦国王学院数字文化与社会讲师、数字文化研究中心主任 Paolo Gerbaudo 提出了数据诠释学（Data Hermeneutics）的研究方法，在使用量化方法进行整体分析的基础上，结合诠释学，对社交媒体中对话的深层语义结构展开探究。理解社交网络不能只依赖数据分析学（Data analytics），将数据仅仅看作行为结构进行研究。研究者还需要将数据视为文本、视为意义之网在特定的文化与意识形态语境中进行细读。为此，有必要提出一套看似简单却实在可行的系统性指南。Paolo Gerbaudo 对于 2011 年埃及与西班牙抗议活动的比较研究证实了线下活动与线上分享的动态双重性。成千上万人占领广场、露宿街头、公开辩论，这些抗议活动的强烈情感属性，通过社交媒体传递给全世界，进而构建起一种情感上的公共空间。

不仅仅是诠释学，质化研究者善用的各种工具（如民族志、话语分析、对话分析、视觉分析），都需要考虑如何与海量网络数据进行对接。也许有了一套数字化的工具箱（digital tool-kit），质化研究者才能更好地理解与批评“大数据”的诸般现象。 2015 年 12 月 3 日到 2016 年 20 日，在伦敦萨默塞特府 (Somerset house) 举办的 BIG BANG

DATA 展览中，Paolo Gerbaudo 和 Mark Coté。以生态系统和表达平台的展示形式，传递了交互式 Persona Non-Data[3]的表达，没有人知道如何处理大量的数据，或者如何了解看到的数据实际上对人意味着什么。

在数字文化方面，KCL 数字人文专业和文化创意产业合称 CMCI(创新媒体文化产业)，两个专业的学生可以自主选择课程模块，KCL 有一些来自英国对当代文化影响深远的伯明翰学派(Birmingham School)的学者，结合了社会学、文学理论、媒体研究与人类学在工业社会中的文化哲学现象，例如数字民工的本体论和认识论以及人工智能代的“卢德运动”等。代表人物有视觉文化（Visual culture）Howells Richard 教授，通过特定的视觉文本来观照世界外物的生活方式， 数字人文的思想与当前在文化中使用的工具（如多媒体、纪录片、交互式游戏、其他可视化的媒体）的低下地位相冲突。他认为这种认识是不可避免的，需要特别关注的是对数字知识客体的理解和保存。2016 年 KCLDH 增加了本科生数字文化专业，学者研究热点体现在全球化、分散权力、赋予权力等热门话题，建构数字文化发展的人文生态。

伦敦大学国王学院数字人文发展也面临诸多挑战，目前人类的发展已经处于一个崭新阶段，人工智能、基因技术、数字技术对人类产生重大影响，人文社科生产过程都有可能面临重大变化。而 19、20 世纪形成的知识范式越来越难以解释社会变迁中的诸多问题，人文学科形成历史过程中的局限性逐渐暴露。数字人文知识分类、知识图谱需要重新思考现代的学科制度跟知识传统、文化传统之间的关系，形成自己的世界观和价值观，也要应当警惕数字文化转向（Culture Turn）中的“文化本质主义”（Culture Essentialism）。

[3].http://bigbangdata.somersethouse.org.uk/persona-non-data-is-transforming-big-bang-data-into-a-data-generator/

Reference

A Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth, Oxford: Blackwell, 2004

Aradau, C., (a1) and Blanke. T.，2018，Governing others: Anomaly and the algorithmic subject of security.

Blanke, T., Dunn, S. and Hedges, M., 2007,"Arts and Humanities e-Science From Ad Hoc Experimentation to Systematic Investigation", e-Science and Grid Computing IEEE International Conference on, pp. 103-110.

Blanke, T, Hedges. M., 2013, Scholarly primitives, Future Generation Computer Systems, v.29 n.2, p.654-661, February.

Dunn, S. and Hedges. M., 2013, International Journal of Humanities and Arts Computing, vo. 7, No. 1-2: pp. 147-169.

Jordan, T. and Taylor,P. 2008, A sociology of hackers.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

數字人文的本科教育實踐

成績與反思

王濤*　孫豔**

南京大學歷史學院教授*　無錫科技職業學院助理研究員**

數字人文的本科教育實踐：成績與反思

王濤 教授 南京大學歷史學院
孫豔 助理研究員 無錫科技職業學院

摘要
「數位化轉向」(digital turn)將如何影響對歷史的認知、描述與研究，在「數位化生存」的時代需要反思。本文將重點介紹南京大學歷史學院的「數字史學研究中心」在數位人文人才培養方面的實踐活動，主要包括常規的本科教學實踐，以及不定期進行的短期培訓課程，用案例說明大陸的數位人文社群在數位人文教育方面的各種嘗試。數字史學研究中心於 2016 年開設面向歷史研究的「數位人文」課程——「數位工具與世界史研究」，系大陸高校的首次嘗試。同時，數位史學研究中心與學術微信公眾號「人文社科新方法」聯合舉辦了以「社會科學新方法」為主題的短期培訓課程，以線上直播平臺「南播玩」為依託，進行了數位人文線上課程的嘗試。通過教學實踐，數字人文的方法和理念得到了推廣，我們積累了一些人才培養的經驗，但也暴露出了一些問題。我們將在此文對各種形態的教學實踐活動進行評述，希望以此為契機，總結經驗，重新起航。

目次
1．常規教學之緣起
2．常規教學之課程設置
3．成果展示
4．常規教學之問題與解決
5．線上課程實踐
6．結語

關鍵字
數字人文，本科教學，歷史研究，南京大學

在當下，數字人文理念愈來愈切入傳統人文學科領域，追隨與質疑同時上演。在最新的討論中，批評的聲音試圖瓦解數字人文存在的意義，認為經過幾十年的發展，有說服力的成果極其有限，而各種回應文章則針鋒相對地意圖捍衛數字人文的價值。[①] 拜技術條件所賜，國內學界幾乎可以毫無時差地與國際數位人文的發展同步，卻有意無意地忽略了數字人文獲得持續發展的根基：人才培養。一個新的學科需要得到認可，雖然沒有馬克斯·普朗克那樣詛咒般的預言，但確實需要有持續的新鮮血液融入。在國內的數字人文圈，已經有學者對國外數位人文的數位人文中心進行了調查，[②] 但主要偏重研究環境的分析；也有學者對數位人文的研究生教育進行了梳理，[③] 並對「資料科學」課程進行了調研；[④] 在實踐層面，國內不少高校不定期地舉辦各種涉及數字人文方法的培訓或者工作坊，[⑤] 但鮮有系統化的課程實踐。

南京大學歷史學院系大陸同類高校中較早成立「數字史學研究中心」（以下簡稱 DH 中心）的高校。從 2016 年秋季學期開始，筆者在南京大學開設了面向本科生的數字人文課程，系大陸高校的首次嘗試。[⑥] 與此同時，DH 中心與學術微信公眾號「人文社科新方法」聯合舉辦了以「社會科學新方法」為主題的短期培訓課程，以線上直播平臺「南播玩」為依託，進行了數位人文線上課程的嘗試。通過常規的本科教學實踐，以及不定期的短期培訓課程相互結合的方式，我們試圖讓年輕的本科生以及數字人文的愛好者們瞭解數字人文的歷史、理念和方法，為他們提供進入數字人文領域的機會，用實在的行為回應數字人文對傳統歷史研究的衝擊。截止 2018 年初，常規課程已經開設了兩輪，選修課程學生共計 49 名，第三輪課程正在進行中。短期培訓也進行了若干場次。通過教學實踐，數位人文的方法和理念得到了推廣，筆者積累了一些人才培養的經驗，但也暴露出了一些問題。筆者將在此文對教學實踐活動進行評述，希望以此為契機，總結經驗，重新起航。

1 常規教學之緣起

數字人文的一個應有之義是合作。人文學者要擺脫「獨狼式」的工作方式，學會跟不同學科背景的學者共同完成研究項目，而不再想當然地認為孤獨地在檔案館查閱資料才是合格歷史學家的惟一指標。[⑦] 然而，要促成真正

[①] Timothy Brenna, “The Digital-Humanities Bust”, The Chronicle of Higher Education, October 15, 2017. 回應的文章見 Eric Weiskott, “There is no such thing as the Digital Humanities”, The Chronicle of Higher Education, November 1, 2017; Sarah Bond, Hoyt Long and Ted Underwood, “Digital is not the Opposite of Humanities”, The Chronicle of Higher Education, November 1, 2017.

[②] 鄧要然, 李少貞，〈美國高校數字人文中心調查〉，《圖書館論壇》3，頁 26-34。

[③] 吳加琪, 董梅香, 趙子菲，〈國外數字人文專業研究生教育調查〉，《圖書館論壇》6，頁 42-48。

[④] 蘇日娜, 肖鵬, 林毅鴻，〈圖書館與資訊科學（LIS）的資料科學（Data Science）課程體系設置——以 iSchools 高校課程調研為中心〉，《圖書館論壇》1，頁 1-6。

[⑤] 據筆者所知，2017 年哈佛大學博士後徐力恒在北京大學人文社會科學研究院舉辦了基於 CBDB 的「數位人文研究技能與方法」讀書會；北京大學圖書館在朱本軍主持下，有不定期的基於各種數位人文方法的講座。廈門大學於 2018 年初舉辦了基於中國哲學書電子化計畫（CTEXT.org）工作坊，由哈佛大學費正清中國研究中心德龍（Donald Sturgeon）主講。值得提及的是，臺灣地區在數位人文的教育實踐起步較早，也有內容更加豐富、形式更加靈活的工作坊，具體資訊可見 http://tinyurl.com/dhintaiwan/。臺灣清華大學中國文學系的祝平次於 2015 年開設有「數位工具與文史研究」課程，廣受歡迎。

[⑥] 參見南京大學陳靜、哈佛大學博士後徐力恒等主持的「數位人文」主題微信公眾號「零壹 Lab」的報導《國內首個數位史學課程：獨家推送課程資料！》，2016 年 12 月 12 日。

[⑦] Alun Munslow, *The New History*, Harlow: Pearson, 2003, p. 93.

有價值的合作專案，需要參與的各方能夠用大家都能夠理解的語言描述需求、問題以及解決方案。這意味著不同領域的學者需要有一些專業上的共識：計算專家能夠理解人文學科的常識，熟悉人文學科的話語體系，即所謂「人文素養」；人文學者也需要瞭解技術背後的原理，主動跟蹤技術反覆運算的步伐，即所謂「數位素養」。但在傳統的人文學科培養體系中，並沒有意識到「數字素養」的價值，或者將其簡單地理解為電腦的使用技能。在數位人文進一步融合到人文研究的趨勢下，個體研究者的數位素質的提升被排上了日程，未來歷史學工作者可能會被要求掌握諸如 HGIS、文本挖掘、建構資料庫等專業技能。為了應對這樣的挑戰，我們需要在人文學科的本科課程體系中開設相關課程，讓年輕學生的數字素養得到更新，從容地應對尤瓦爾・赫拉利所預言的未來。

南京大學 DH 中心於 2016 年 9 月的秋季學期推出了「數位工具與世界史研究」課程（以下簡稱「數字歷史」）。課程名的設置有兩個方面的考慮：

首先，沒有使用「數字人文」的字樣，不僅因為數字人文的概念本身沒有定論，而且不希望學生以為這是一門泛泛而談數字人文概念的理論課程。我們強調工具性的價值，希望學生意識到新方法、新工具對歷史研究的推動作用，讓他們在研究實踐之後再去構建自己對「數位人文」的理解，甚至昇華到哲學層面的探討，反思資訊技術是否能夠改造歷史學的理論問題。[①]

其次，強調「世界史」是為了引起國內的世界史研究同行對數位人文的關注。數字人文在國內雖然起步很晚，但中國史領域對它的接受度較高，已經有了比較豐碩的成果。[②] 相反，國內的世界史學術圈對數字人文的關注稍顯滯後，但筆者深刻地意識到，在數字人文的方法和理念武裝下的世界史研究，一定可以推出更多具有原創性的成果。[③] 同時，我們也希望強調「世界史」其實是包含了中國史的世界史，從而讓學生用「全球史」或者「整體史」的思路來理解數字人文背景下的歷史研究，也更加呼應「歷史學宣言」的訴求。[④]

2 常規教學之課程設置

本課程以「數位人文」為核心，可以劃分為三大板塊：理論、方法與實踐。

理論討論主要偏重「數位人文」的歷史演進，以及基本議題。考慮到選修該門課程的大多數學生此前沒有接觸過數字人文，試圖通過這個板塊的教學讓學生對數字人文有一個感性的認知。在教學環節特意安排了讓學生調查現有數字人文專案的小作業，並鼓勵學生評判數字人文專案的優缺點。這個環節的安排能夠讓學生迅速進入數字人文的天地，激發他們的興趣，為後續

[①] 薑萌，〈中國史學在數位化時代的變與不變〉，《史學月刊》5，頁 14-19。

[②] 代表性成果包括:趙思淵，〈地方歷史文獻的數位化、資料化與文本挖掘〉，《清史研究》2016(4)；徐永明，〈《全元文》作者地理分佈及其原因分析〉，《復旦學報》2017(2)；申斌, 楊培娜，〈數位技術與史學觀念——中國歷史資料庫與史學理念方法關係探析〉，《史學理論研究》2017(2)；陳靜，〈數位檔案化廣告蜉蝣：以中國商業廣告檔案庫（1880～1940）為例〉，《江海學刊》2017(2). 另有《史學月刊》編輯部，《大資料時代的史料與史學》，北京：人民出版社，2017；舒健，《大資料時代的歷史研究》，上海：上海譯文出版社，2018，是國內中國史領域在數位人文方面成果的集體展示。

[③] 王濤，〈數字史學：現狀、問題與展望〉，《江海學刊》2，頁 172-176。

[④] 喬.古爾迪，《歷史學宣言》，上海：上海人民出版社，2017。

的學習鋪路。

第二個板塊是方法，介紹了數字人文相關方法的原理、工具，並手把手教授基本的操作過程。與數位人文相關的方法和工具異常豐富，出於教學實踐的可操作考慮，以及歷史研究的針對性需求，主要涉及文本挖掘、社會網路分析、HGIS 以及量化研究方法。

每一個方法作為獨立的教學單元，劃分為背景介紹與實踐課堂兩個部分。背景介紹主要向學生講解某個方法背後的理論邏輯、演算法原理，並配合代表性案例分析作為應用示範。比如，在講授 HGIS 的實際價值時，我們會舉斯諾（John Snow）利用空間分析的方式找出倫敦霍亂源頭的例子。有許多方法涉及複雜的演算法，需要很強的數理邏輯才能理解。我們試圖用平實的語言解釋原理，而不太涉及演算法本身。在初始階段，我們的預設是：學生們只需要把工具想像成一個黑箱，能夠理解在什麼狀況下使用，如何解讀結果就可以了。

對工具的選擇也考慮到學生的基礎水準。比如，盡可能使用開源的工具，盡可能推薦不需要程式設計基礎的獨立軟體，或者網路平臺。這樣，學生不需要花費太多精力去學習一門新的技能，而將更多的時間分配到對方法本身的體驗、對結果的解讀層面。

第三個板塊是實踐環節，也是「數字歷史」課程最大的亮點所在。在各種反思數位化時代的歷史教學的研究中，許多學者主張通過「做歷史」來讓學生掌握歷史思維的技能。[①] 數字人文本身是很具實踐性的學科，讓學生在實踐中學習，也就成為貫徹數位人文理念的最佳手段。所有選修課程的學生被要求根據個人的研究興趣自由組合成不同的研究小組，展開一個具體的「數字人文」專案。為了對專案的進展進行監控，「數字歷史」課程特意安排三次報告時間，分別是開題、中期考核以及結項，讓學生及時彙報課題運行的狀態，也是督促研究小組持續工作的方式。三次彙報時間被安排在課堂教學的不同時間節點，學生們有機會不斷修正專案的方向和內容，保證他們最終提交的方案盡可能完整。為了滿足學生更多的技術需求，本課程還設置了課後答疑時間，每週安排固定的時間解答學生們在專案進展中遇到的問題。



[①] 王濤，〈關於數字時代歷史教學的思考〉，《歷史教學》4，頁 61-66。

圖 1 「數字歷史」課程體系

3 成果展示

在老師和學生的共同努力下，參與課程的學生提交了完成度極高的數位人文成果。兩輪的教學實踐下來，近 50 名學生一共組建了十多個研究團隊，專案主題涉及社會史、文化史、經濟史、思想史等各個方面。每個研究團隊參與者有多有寡，但都體現了不同的分工，因應了數位人文的合作訴求。

學生的項目可以劃分為兩大類型：歷史的網路書寫，以及用數位工具進行歷史議題的研究。前一種類型主要是對具體的歷史問題進行梳理，並用網站的形式呈現出來，內容創新不是追求的要點，而是用非線性的敘述，將傳統議題進行了重塑。學生的網路書寫都注意了板塊劃分的歷史邏輯，特別留意引入圖片、視頻等多媒體手段提高內容的可讀性，網站的佈局、設計等美化環節還需要提升。這個類型的專案包含了中國服飾史、遊戲史、甲午海戰、二戰中國勞工、明信片研究等主題。

第二種類型將數字工具作為歷史研究的新方法，對具體的歷史議題進行了嘗試。學生通過在課堂教學中學會的文本挖掘、HGIS 等技術，對很多研究課題進行了分析。有學生用文本挖掘工具分析了亞當·斯密的著作，試圖用遙讀理解斯密的經濟思想；有學生挖掘了《人民日報》中關於「女權」的概念演變；還有學生基於對文本的情感分析研究了當事人對一戰的主觀感受。

學生們的專案雖然略顯稚嫩，但從最終呈現的樣態來看，都較好地體現了數字人文的維度，並證明了他們對工具的熟悉程度。

4 常規教學之問題與解決

經過兩年的教學實踐，「數位歷史」課程的內容愈來愈豐富，在學生中受歡迎的程度也愈發明顯。從學生對本課程的回饋來看，大部分學生對這個課程還是比較認同的。當然，兩年在教學前線的實踐也讓筆者意識到了本課程存在的問題。

（1）缺乏技術背景專業的學生。本課程設立的初衷是讓人文學者與技術學者在合作交流中能用雙方都理解的術語對話。雖然本課程主講老師具有歷史學背景，大部分選修的學生也來自歷史系，但也有相關專業的同學選修了本課程，包括社會學、外語、商學、法學等。本課程試圖讓社會科學背景的學生掌握技術話語，也讓筆者意識到讓技術專業出身的學生瞭解人文社科研究者的需求同樣重要。所以，如果有更多技術背景的學生加入進來，或許能夠更加促進不同學科之間的交流。要解決這個問題，需要繼續推廣數位人文的概念，讓人文社科背景的學生和技術背景的學生都意識到數位人文的重要性，從而更加自覺地加入到這個新興的學科中來。

（2）內容的深度與廣度。數位人文涉及的內容非常寬泛，我們要在一個學期的課程中將對數字人文可能毫無概念的學生，引入到數字人文的門徑，是一個很大的挑戰。為此，我們只能盡可能壓縮內容，向學生推薦最流行和最成熟的概念和方法。這也是跟教師本身的技術水準與課程要求的廣度存在矛盾。筆者雖然對數位人文的各種技術有所涉獵，但也有自己擅長的領域和需要學習的短板，很難憑一己之力全面地介紹數字人文。另一方面，數位人文技術反覆運算迅速，為教師的知識更新提出了更高要求（本課程雖然只開

設了兩年，但有些技術已經出現了升級換代，筆者需要在第二次上課時隨時調整內容）。而為了課程的完整性，筆者不得不去教授自己並不擅長的內容，甚至要面臨現學現賣的窘境。要解決這個問題，該課程需要進行合理的人力資源配給，讓某個領域的專家來講授某個領域的方法，把本課程建設成合作制經營的體系。

（3）通用性與個性問題的落差。教學只能向學生展示最基本的方法，數位人文的大量內容還需要學生在興趣的驅使下自己去探索。這樣也帶來一個困境，學生通過本課程的學習，瞭解到最基本的技術，但這些技術只能解決通用性的問題，而無法解決自己在研究過程中個性化的問題。這個問題會直接關係到數字人文在多大程度上介入學生們今後的研究，也將影響到數位人文的發展的後勁。為此，需要鼓勵學生不斷學習新的技術和工具，並有意識地在自己的學年論文甚至畢業論文中使用數位人文的理念，讓數位人文成為研究活動中自然而然的一部分。[①]

（4）**專業性與工具性的融合。**本課程強調學生通過「做歷史」來掌握基本的技能。但通過兩年的教學實踐發現，學生在專案的展開過程中存在一個共性的問題：許多成果呈現出技術活躍，但是專業性分析薄弱的毛病。學生對工具都抱有積極的心態，在各自的專案展開中躍躍欲試，然後把更多精力安置到了方法的實現層面，而對結果的解讀稍有鬆懈。有一些學生在專案推進中，清洗資料佔據了絕大部分時間，留給結果分析的時間不得不一再壓縮。實際上，重工具，輕分析是很多數字人文專案的通病。要正視這個問題，一方面需要開發更有效率的工具，把研究者從繁瑣的前期準備中解放出來，另一方面還需要研究者轉換思路，強調論證驅動，而不是工具驅動。

（5）基礎設施的服務。從學生的報告中，我們遺憾地瞭解，學生在專案的研究中有很多想法，但苦於發佈平臺的局限，無法全部得到實現。特別是基於網路平臺的歷史書寫，他們只能使用提供網路服務的基礎版（免費版），限定了功能與容量，從而讓學生的項目呈現效果大打折扣。我們期待學校能夠改善數位人文教學的基礎環境，增設更多的伺服器或者技術平臺，來滿足教學與科研不同層級的需求。

5 線上課程實踐

常規課程能夠為數位人文的愛好者提供系統化的知識體系建構，但缺點是時間固定、學習週期漫長，受眾也拘泥于南京大學註冊的本科生。為了滿足更廣大年輕學生的需求，DH 中心於 2018 年與學術公號「人文社科新方法」聯合舉辦了以「社會科學新方法」為主題的短期培訓課程，以線上直播平臺「南播玩」為依託，進行了數位人文線上課程的嘗試。微信公眾號「人文社科新方法」由高校大學生和高校青年教師共同發起，旨在分享人文社會科學最新定量研究方法，為從事人文社科定量研究的學者搭建的交流平，於 2017 年 11 月 10 日正式註冊上線，保持每日更新的頻率。

我們以 2018 年 5 月 21 日的一次線上直播為例進行說明。

網路直播的方式受眾面更廣，不受地域限制，線上網路直播方式比較靈

[①] 值得肯定的是，南京大學歷史學院 2013 級朱哲慧在本科畢業論文中，用數字人文方法研究了蘇格蘭啟蒙運動；錢超峰用量化方法研究了晚清的人事制度與政治格局，都是有益的嘗試。

活，更節約了學員時間成本和經濟成本；課程內容可以反復觀看，有助於學員提高學習效果。更重要的是，我們堅持了資源分享的原則，線上直播的課程、資料全部免費，面向所有愛好者開放。

從流程上看，線上直播課程的工作流可以劃分為內容設計、宣傳推廣、線上直播以及後期總結等幾個環節。

- **內容設計** DH 中心與微信公眾號共同商討線上課程的內容，重點突出數位人文工具性的一面，試圖用實際的案例帶動方法論的探討，並把有效性與有用性貫徹課程內容的始終。我們計畫用一個可重複、可再現、可模仿的數位人文案例為課程內容（此次直播課程主要是處理超過 10000 封電子郵件的案例），讓學員感知什麼是數位人文研究？數位人文研究如何應用？可以使用哪些技術方法？需要經歷哪些步驟？使用什麼工具等一系列問題，讓學員通過模仿來嘗試再現這項研究並啟發自己的研究興趣。我們認為，對於初學者而言，模仿是一個很重要的過程。從課程最終呈現效果看，我們基本上實現了初衷。

- **宣傳推廣** 我們通過微信公眾號平臺「人文社科新方法」進行課程宣傳。前期在策劃本次活動前數月，我們開始通過微信公眾號，轉載了授課老師在數位人文研究領域的兩篇文章《歷史學家的新技藝：如何處理成為史料的電子郵件》和《歷史學家的新技藝：遙讀作為史料的電子郵件》，讓讀者大致瞭解到利用數位手段處理電子郵件的可能性和趣味性。在本次線上直播課程正式招生宣傳的 20 天時間內，共進行了三輪宣傳，第一輪是直播預告：將開設本次數字人文線上直播課程的初衷、內容和形式做了一個大致介紹，並且通過加微信群和 QQ 群的方式招募有意向的學員；第二輪是對話學者：通過採訪問答的形式，由授課老師對學員在數位人文學習中的經驗、看法、數位人文學習中的困惑進行解答；第三輪是直播前的終極預告：對直播的時間、形式、內容重複精確宣傳。期間，我們還通過與其他學術公眾號合作轉載的方式，共同進行了這三次宣傳。

- **線上直播** 我們與直播平臺「南播玩」進行合作，由後者提供技術支援。這樣做的目的可以讓內容提供者專注於內容的優化，而讓技術人員負責輸出環節的流暢，讓學員具有更好的直播觀看體驗。

- **後期總結** 課程結束後，我們對學員發出了問卷調查，希望能夠瞭解大家觀看後的感受，以便我們在後續的課程中揚長避短。

本次線上直播的效果，超出了我們的預期。5 月初，我們通過微信公眾號第一次發出正式直播預告後，24 小時內便吸引了超過 300 人加入我們的直播微信群和 QQ 諮詢群。在課程開始之前，有 376 人加入了微信直播群和 500 人的 QQ 諮詢群。5 月 21 日當晚的課程，有 1452 人次觀看了直播，學員中不僅有國內學生，還有來自海外的學員。當晚的直播歷時將近三個半小時，直播過程中，我們即時在微信群內和 QQ 群分享課件和需要用到的資料和代

碼，學員在群內也可以提問。直播結束後，學員反響熱烈，在他們的強烈要求下，也為了更好的鞏固和理解課程效果，我們將原本重播三天的直播視頻延長至一個月，最終的觀看量達到了 3178 人次，總觀看時長為 37475 分鐘。另外，有很多學員強烈要求能夠繼續開設線下培訓課程，提升學習的效果。

我們通過「南播玩」的後臺資料，發現了參與本次直播的學員狀態。

觀看量前十名的省份為江蘇、上海、北京、廣東、陝西、湖北、山東、雲南、天津和河南。分佈如下：



排名	省份	观看量
1	江苏	727
2	上海	314
3	北京	291
4	广东	256
5	陕西	178
6	湖北	172
7	山东	113
8	云南	102
9	天津	94
10	河南	90

圖 2 學員來源地

學員的背景也體現出多樣性的特色，研究生的群體（包括碩士生和博士生）佔據了主導地位，但也吸引了本科生，甚至青年教師的關注。正是由於來源廣泛，學員的學習基礎也存在很明顯的分化，對課程難度感覺「適中」和「有點困難」的比例勢均力敵。考慮到超過六成的學員是第一次嘗試線上課程的學習，學員在「數字人文」的基本掌握上呈現兩極化的分佈，也是情有可原的現實了。

令人意外的是，學員的專業背景呈現了極其多元的狀態。本次直播課程的講師是歷史專業出身，所採用的案例也是歷史學的問題意識，預期聽眾是歷史專業的學員，但同樣招徠了新聞、經濟、管理等學科背景的聽眾。可見「數位人文」的方法論具有通用性，從各種專業都具有接受度。

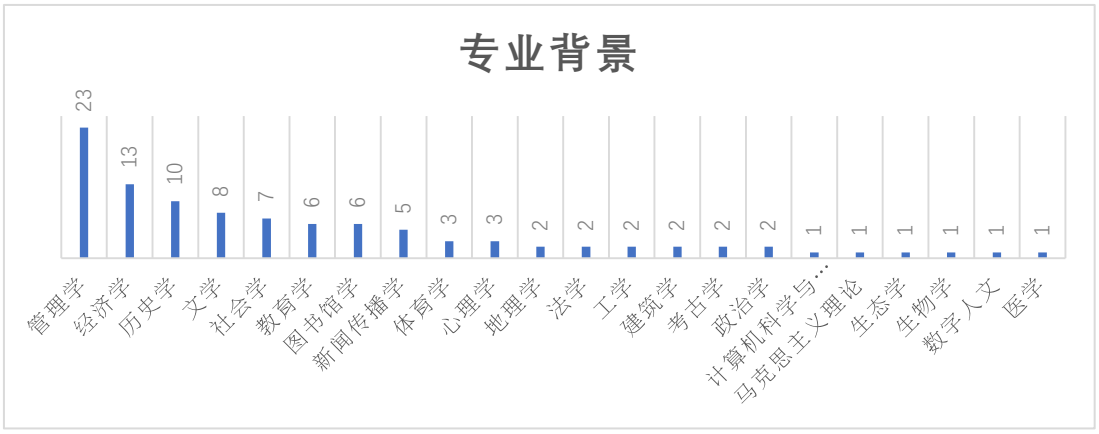


圖 3 學員專業背景

問卷和直播平臺即時討論環節的詞頻統計，也從側面表達了學員對線上課程以至於數位人文教學的心聲。回饋最多的意見是希望線上平臺能夠堅持下來，提供內容更加豐富多彩的課程。這將成為我們努力的方向。

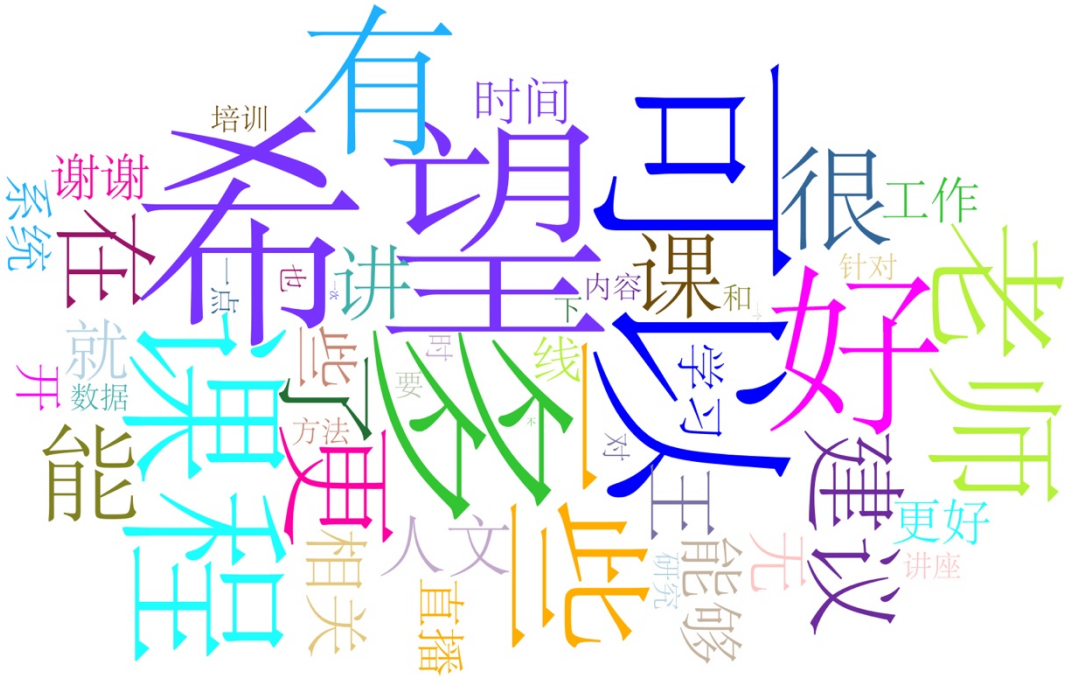


圖 4 學員回饋的文字雲

6 結語

把「數字人文」的概念納入本科培養體系是一個全新的嘗試。國內高校，尤其是人文學科的培養體系還是新鮮事物。而與此同時，國外高校已經在這個方面取得了長足進步，比如英國倫敦大學國王學院已經建立了完整的課程體系；[①] 德國許多高校已經在廣泛招聘「數位人文」的教席。[②] 本科層次的數位人文課程體系非常重要，因為數字人文對傳統人文學科已經提出了越來越嚴峻的挑戰，未來的發展也無法估量。我們需要借助完備的課程和實踐，讓學生為

[①] 包晗，〈英國國王學院的數字人文教學〉，《零壹 Lab 公眾號》，2017 年 12 月 31 日。
[②] 截止 2016 年，德國各大學有近 50 個「數位人文」研究中心，參見 http://dhd-blog.org/?p=6174.

未來的數位人文挑戰做好準備，甚至期望能夠通過類似課程的學習，為學生提供更加多元化的就業管道。DH 中心的常規課程具有顯著的實驗性質，內容全面，自成體系，但囿於課堂教學場所，受眾有限；短期線上課程內容靈活，受眾面廣，但也存在弊端，比如對網路環境要求較高，學員缺少與教師互動，有可能出現注意力不集中或中斷的行為，影響學習效果，同時學員成分複雜、水準參差不齊，也會對教學效果產生影響。總而言之，我們通過教學實踐的總結，希望為其他有志于數位人文教育的兄弟院校和同行提供借鑒，讓數位人文的人才培養事業取得更多進步。

Facing
the Era of AI + DH

2018

第九屆
數位典藏 與 DADH

9th International Conference of Digital Archives and Digital Humanities

數位文國際研討會



專題討論 PANELS

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 12/21 (W5)
Venue 地點 / 法鼓文理學院 Dharma Drum Institute of Liberal Arts (DILA)

歷史古城之數位人文創新探討

以臺南府城為例

Innovation of digital humanities in historical cities

a case study of old Tainan city for example

陳姿汝　楊智傑　楊斯嵐　徐芳真
南台科技大學視覺傳達設計系

Panel 提案

Title

歷史古城之數位人文創新探討：以臺南府城為例

Innovation of digital humanities in historical cities: a case study of old Tainan city for example

Keywords

府城、數位創新、歷史遊戲、創意街區、臺灣當代藝術、兒童文學

Proposal

臺南作為臺灣近代四百年的歷史發展起點，從荷據時期開始，古地名大員(現今安平)的紅毛人熱蘭遮城、漢人聚集的大井頭與十字街遺跡附近的市街，以及鄰近的麻豆社、蕭壠社等原住民社群，更是十七世紀大航海時代中，世界重要的貿易據點。歷經鄭氏家族的屯田擴張，引入更多漢人移民入臺開墾後，由於清代對於臺灣處於帝國邊疆的放任統治態度，漢人移民的勤勉加上商業的發達，在當時形成府城維繫對外商貿的重要樞紐。進入日治時代之後，遍及全臺灣的現代化建設，在臺南也留下許多痕跡，包括街道的截彎改直、水利與鐵道的建設，以及日本以作為殖民者所做的許多政治、產業、文化的政策。進入民國時期之後，雖然臺南府城已不具有首都的地位，在經濟逐漸發展的情況下，位處於臺南當地高密度的文化資產、廟宇信仰、民俗習慣仍保留的相當完整，可做為以現代的數位科技與設計創新最好的立足點與資源。而本研究群組在臺南府城豐富的人文脈絡下，從歷史遊戲、老街人文、當代藝術與兒童閱讀等四個面向來發展出各自不同的古城文化再造之相關研究議題，依此從事數位人文創新之研究工作。

而有關「數位人文」之討論，起源於在資訊科技高度的發展之下，對於傳統的人文與社會學科造成新的挑戰與契機。近年臺灣數位人文相關研究當中，以「設計科學」營造數位新世界的此研究主軸下，期望透過數位科技與設計專業的輔助下，尋找創造嶄新文化的機會。也因此，本研究群組將此方向擴展至「數位人文創新」的研究框架，不僅只有透過設計科學來營造數位世界，而是探究如何透過設計思考與創造力，並綜合歷史、文學、藝術的人文底蘊，透過數位媒材來進行數位敘事並產生數位體驗。以此框架為核心理念，並以具有豐富文化底蘊之臺南古城為實驗與研究場域，透過數位科技為憑藉，依此來進行人文創新各面向議題之討論。

本研究群組之四篇論文各自以「歷史」、「藝術」、「文學」為範疇，透過「設計思考」或「創造力」為工具，進行「數位敘事」、「數位設計」、「數位媒材」、「數位體驗」之應用。四篇論文各自的研究素材範疇不盡相同，或有交集，或各據一地，

實際上卻都進行了「數位敘事」、「數位設計」、「數位媒材」、「數位體驗」的應用。如「用遊戲講述看不見的府城」此篇論文論述有關歷史遊戲的設計，從歷史故事的敘事，到遊戲介面的設計，以及如何應用數位媒材，進而形成玩家的數位體驗；「歷史街區的文化保存與創意開發之互動觀察」這篇論文則是透過數位媒材的呈現與設計，進行老街歷史的敘事，並觀察拜訪歷史場域的人文體驗到進駐創意街區的創意活動之間的互動關係；而「臺南當代藝術訊息重植之跨世代媒材敘事轉譯研究」是以當地藝術的離根現象為契機，文化的著根為目的，研究傳統媒體、社群媒體、實境技術之間的敘事模式如何轉換，數位媒材如何透過自動化的數位設計，來形成閱聽人的數位體驗；「不同文化下的動畫角色色彩之研究」則是探討數位工具如何輔助兒童閱讀，抓取文學的敘事特性，運用數位敘事方法研究，透過數位工具設計，整合數位媒材，引導兒童進行說故事，形成了歷史與文學的數位體驗，本篇論文先以角色設計為初探。本研究群組透過對府城的各種面向的探討，一筆一筆的描繪出臺南特有的歷史人文，讓這種融合多種族、經歷遷移、抗爭、對立等種種歲月轉換，透過歷史文學、遊戲、藝術、新創等各種融合數位科技的手段，將臺南特有的人文內涵以研究的角度，獲得重新詮釋的機會。

在數位人文的研究範疇中，如何透過設計科學與數位科技的結合來創造新文化，是重要的研究與實務課題之一。然而，在現有的人文與設計科系之間的藩籬與隔閡之下，需要透過學術研究與跨界整合，來創造嶄新的視野與可能性。臺南市在活絡的藝術文化與觀光旅遊發展下，雖然已經大幅吸引國內民眾的目光，在假日休憩之時，常有來自國內外的旅客造訪這個歷史悠久的古都。雖具有豐富的文化資產，但大部份的民眾對於臺南的印象僅停留在以古蹟景點為中心的旅遊，以及能夠大快朵頤的美食小吃。雖然近年來官方與民間齊力發展此區域的城市特色，然而，臺南在過往四百年的歷史中消失的記憶，以及目前散佈其中的藝文能量，仍尚未讓民眾有深入認識。為了讓府城有關的歷史、文化與藝術內涵，能夠繼續影響未來的古城發展，並透過各種媒介與數位科技讓大眾熟悉，也因此，本研究目的則是希望能透過遊戲設計、互動科技與影像媒體作為數位媒介，以臺南時空脈絡中與往昔有關的「移民．殖民．遺民」之歷史、現今尚存的老街人文，以及街區中特有的藝術人文根離化現象等為議題，其此能夠成為作為臺灣發展數位人文研究的特色主題之一。

Paper 1

Title

歷史遊戲的概念框架與啟發式設計程序之建構：以「進攻台灣府」實境遊戲設計為例

歷史遊戲的概念框架與啟發式設計程序之建構：以「進攻台灣府」實境遊戲設計為例

楊智傑
副教授
南臺科技大學 多媒體與電腦娛樂科學系

摘要

本研究以清代府城五條港的文史脈絡作為內容，探討歷史遊戲設計過程當中，對於形成「歷史－遊戲」之間的中介知識進行觀察，並且透過麥柯(McCall，2012a)提出的「歷史問題空間」框架概念，實際應用於遊戲開發過程中。另外，我們將歷史遊戲開發視為一種「啟發式的設計程序」(Flanagan & Nissenbaum，2014)，透過「發現－轉譯－驗證」的過程中，將文史資料與田調知識，透過跨領域的異質團隊，轉化為可遊玩的遊戲體驗。我們透過執行小規模的遊戲轉譯練習，以及一個實境遊戲「進攻台灣府」的設計過程，來觀察歷史遊戲設計所需的關鍵知識如何成形？以及思考這些知識如何能夠透過工作坊與教學活動，培養未來更多的歷史遊戲「轉譯者」。研究結果顯示能夠形成「歷史－遊戲」中介知識的方法論，以及有助於構想發展過程中的「概念工具」，有助於歷史遊戲的設計。

目次

1. 研究背景
2. 理論基礎
2.1 啟發式遊戲設計程序
2.2 「歷史問題空間」概念框架
3. 遊戲設計與實作
4. 結果與討論
5. 結論與建議

關鍵詞

價值意識遊戲，歷史問題空間，轉譯，歷史思維，五條港，蔡牽

1. 研究背景

本研究屬於數位人文研究領域中的「人文世界之數位遊戲化」之主題範疇之內，針對人文領域中的歷史內容，如何轉譯為遊戲體驗所需之理論、方法與工具進行探究。我們認為歷史遊戲設計的主要困難點，在於如何將歷史的「內容」,「轉譯」為有趣而吸引人的遊戲「形式」。「轉譯」的工作，在前期需要能夠研讀史料與進行田野調查的研究員。與撰寫文章或報告的傳統做法一樣，研究員的意識形態與對議題的熟稔與否，具有關鍵的影響力。遊戲設計與撰寫文章不同的是，需要熟知遊戲能展現的手法限制，以及創造出的體驗對玩家/使用者的影響。

大部份採用人文主題的遊戲開發者，通常依賴自身經驗與創造力，每個遊戲有如藝術品一般，皆有團隊自身的獨特性和詮釋。本研究探究歷史遊戲設計「內容—形式」之間的轉譯過程，建立過程所需的概念框架與設計程序，並且以工作坊動手操作的方式，讓遊戲設計構想能夠容易被辨認、評估並不斷改善。著眼於前述的高教環境形成的困境，以及遊戲開發本身所需的高度橫向整合需求，本研究針對歷史遊戲設計所需的概念框架，以及如何透過啟發式設計程序來發展遊戲雛形進行研究。

本文的撰寫的結構介紹如後，第二章是相關理論基礎，第三章是遊戲設計與實作的過程與成果介紹，第四章是結果與討論，最後，第五章是結論與建議。

2. 理論基礎

2.1. 啟發式遊戲設計程序

Flanagan and Nissenbaum (2014)認為創造數位遊戲的過程既複雜，有與許多相互交織的尺度有關。與專案開發的不同角色的成員，都會以自己的意願影響遊戲的產出。要將「價值」這種高層次的思考或意識形態，加入遊戲設計的過程當中，極為容易讓遊戲體驗變得抽象與含糊。因此，她們提出的 “Value at Play”（VAP）的設計方法，採用「啟發式/捷思」（heuristic）的思考模式。此類方法在心理學領域或處理「使用性」（usability），在面對問題情境難以有清楚與全盤理解時，依據個人或團隊經驗採用直觀的推論方式。當遊戲專案中，最終目標不明確時，仍能夠把設計團隊有關社會、道德和政治的價值，將某些啟發式的準則（criteria）引導專案進行。

VAP 的啟發式方法包含「發現」（discover）、「轉譯」（translate）和「驗證」（verify）三個部份(Flanagan & Nissenbaum，2014)：

(1) 「發現」是鎖定與專案有關的價值，並以遊戲的脈絡來給予定義；

(2) 將價值「轉譯」到遊戲元素，包括規格訂定、美術圖形和程式碼。這步

驟是設計的核心，以遊戲中基本的實作要素實現價值的過程；

(3) 「驗證」設計師用於發現與轉譯價值的努力是否有效？也可視為一種品質控制。

VAP 的三個步驟不見得要依序執行，而是不斷反覆（iterative）的過程。也就是在遊戲開發過程中，需要重複概念產生、製作原型、測試、分析到修正這些步驟，直到獲得滿意的結果。

2.2.「歷史問題空間」概念框架

歷史遊戲設計的「轉譯」所需的中介知識，常與歷史與遊戲這兩者之間混成而來知識有關。目前在國內外的歷史遊戲相關研究之中，以美國學者麥柯（Jeremiah McCall）提出的「歷史問題空間」（historical problem space）最符合此項需求。以歷史教學為出發點，他以下列觀點解釋歷史模擬（historical simulation）的意義。麥柯認為模擬遊戲可透過「問題空間」（problem space）的設計，詮釋過往發生的事件(McCall，2012a)。在教育和認知研究領域中，「問題空間」指的是人在試圖達到目標或不同狀態時，由可行選擇形成的心智地圖（mental map）。此概念和物理或實體空間（physical space）無關，也有別於電玩研究者對遊戲中空間的看法，例如「透過遊戲空間來進行敘事」(Jenkins，2004)，或將遊戲空間視為「讓玩家互相競逐的場所」(Squire & Jenkins，2002)。

麥柯提出的「歷史問題空間」有下列的三大定義(McCall，2012a)：(1)玩家，或是實體世界中的行動者（agent），在空間中其角色和目標應廣泛形成脈絡；(2)透過施行選擇或策略，玩家能付出努力以達成目標；(3)選擇和策略的結果（特別是成功的）由情境支持（affordance）和限制（constraint）來形成，呈現的形式與有限的或會匱乏的可數量化的資源（quantifiable resources），以及文化框架（cultural framework）與心理趨向（psychological tendencies）有關。

麥柯認為以「歷史問題空間」這個框架來思考，能夠達成二十一世紀的歷史教育的某些目標。他認為基於過去如何能夠以問題空間的形式下，有意義地被探索。並且建立合理懷疑與人文思考的健全感受，以下的這些特徵可被歷史遊戲的開發者來遵循(McCall，2012a)：

(1) 扮演不同角色的玩家：在建立歷史觀感的規範時，了解過去是充滿各式不同選擇的人物，做出抉擇和扮演角色；

(2) 具有目標的玩家：雖然遊戲中的目標是很明確，但真實生活中的目標卻不容易有清晰面貌。可以這樣來看，真實生活中的行動者的目標通常是分岐的、不清楚的、衝突的、不理性和難以達到的。當我們把過往視為問題空間來考量時，應該一直察覺這些真實面貌。盡管如此，並不代表

遊戲中給予行動者目標是無意義的過度簡化。這點是解釋非群體的個別人類行為時，考量這些角色的意圖，辨認出人類的確會有其追尋的目標；

(3) 玩家與其動作發生在實體空間：教師和學生很容易忘記過去（和現在）的人類是處於實體和空間的脈絡中。就算是最理智的（intellectual）、情感的（emotional）、精神上（spiritual）的目標都是體現在實體和空間的脈絡。理解這些脈絡有助於理解行動者的角色、目標、選擇、情境支持和限制；

(4) 具有選擇和策略的玩家：雖然哲學家會辯稱是否每個人都真的具有選擇，歷史學者的確常以選擇和抉擇來探討問題。而作為人類的我們是透過自己和他人的選擇，來做出行動並讓世界變得完整。就算是當我們認為自己成為受害者時，仍會對加害者對象做出某些選擇；

(5) 情境支持與限制：過去（和現在）的行動者會擁有機會和遭遇障礙、餘裕和匱乏、才能與弱點、接觸與排除，這些情境支持和限制形塑他們的選擇、目標和角色；

(6) 空間脈絡：人類的動機、目標和行動和實體的脈絡有關，也會受到很多情境支持和限制的影響。心理、情感、精神和知性等層面都扮演關鍵角色。人類的目標和動作無法從所在的環境抽離仍具有完整意義。

其他關於麥柯論述遊戲如何應用於歷史教學的內容，可參酌他的專書 *Gaming the Past* (McCall，2011)，以及對於「歷史問題空間」的其他細節(McCall，2012b)。

3. 遊戲設計與實作

本研究共進行三項遊戲設計，皆由清代府城五條港的文史脈絡發想而來。第一項是研究助理與學生團隊合作構思的「划水仙」桌遊雛形，第二項是與台南塾合作開發的「進攻台灣府」實境遊戲(四天活動)，第三項是「魂牽夢瀛」實境遊戲，在靜態的展場空間中呈現。

在第一項桌遊雛形的發想階段，在專任助理閱讀五條港的文史脈絡 2 個月後，並進行 2-3 次的田野調查，開始與美術與遊戲設計專長的學生合作構思遊戲。為了讓不是文史背景的設計人員，能夠快速進入五條港的時空情境。我們透過繪製一系列的情境圖，做為建立共識的開始。美術人員依照專任助理設定的規格與參考圖片，所繪製的包括街道場景、街屋內部、做十六歲、龍舟競渡等情境圖。此部份工作與一般以歷史為題材的影視作品製作方式類似，所有相關人員透過視覺的「再現」，得以觀看目前已不復存在的過往場景。另外，研究助理亦透過「歷史問題空間」框架來討論「龍舟競渡」與「划水仙」背後的歷史脈絡，並據此進

行遊戲設計的發想。

接下來，團隊挑選了以「划水仙」這個帶有傳奇故事色彩的主題，繼續構思遊戲。在移民社會中漢人漂洋過海來台，由於氣候難料與迷信風氣盛行，在船難危急時刻，船員要解開髮辮、趴在甲板並拿筷子作狀划水的荒謬動作，此項情境在郁永河的《裨海記遊》（又名《採硫日記》）中被記載。當時船上的行動者可以含括船家（船長/船員）、郊商（主人/隨從）、移民（官員/書生/偷渡客），在遭遇船難的特殊困境下，橫跨海峽時不可預料的事件，以及在自身利益與宗教信仰的力量，促使玩家扮演的行動者根據理性與情緒之間作出抉擇。

後續考慮呈現的遊戲形式包括桌遊與虛擬實境，在經費有限的情況下，採用桌遊的玩法繼續發展，並以「怒海求生」（Lifeboat）的遊戲玩法做為參考，划水仙的情境圖與桌遊圖版雛形請見（圖 1）。故事的主題是船上的羅盤水被打翻，迷失航程，船上的人要想辦法求生。遊戲的規則以回合制進行，3-6 名的玩家依據抽卡決定自己扮演的角色，（圖 1）（右）所示的圖版上，不同角色有其活動範圍，例如偷渡客可在暗艙中躲避追查，在甲板上可打撈海中貨物。

在約兩個月的遊戲開發過程中，參考既有桌遊玩法的構思結果，似乎限縮了我們想透過遊戲來表現歷史情境中行動者抉擇的意圖。另外，跟一般的桌遊常見的限制相同，通常歷史的脈絡僅能夠形成氛圍，透過美術和角色設定，來增強其主題性。雖然在諮詢過台南地區的桌遊設計專家之後，商業桌遊亦有能夠引發歷史思考的其他表現形式，但由於計畫進行的時限，並未繼續再進行後續探究。

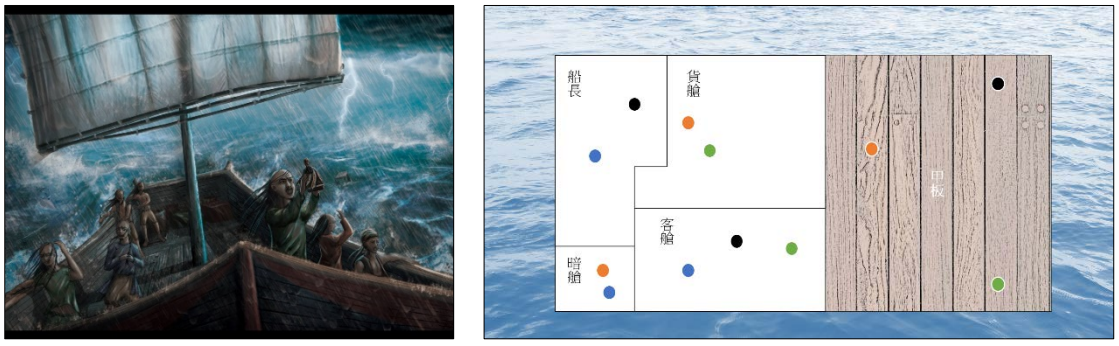


圖 1 划水仙的情境圖(左)與桌遊圖版構想(右)。

本計畫的第 2 項（進攻台灣府）與第 3 項（魂牽夢瀛）實境遊戲的實作，是以清代第一個從海上攻擊台灣的內部勢力－海盜蔡牽作為發想起點，在初期由移民社會的商貿視角不同，希望透過遊戲體驗喚起台南（台灣）被遺忘的海洋想像。玩家不是扮演郊商或工人，而是替準備攻擊府城的海盜收集情報的探子。實境遊戲的玩法會在台南與嘉義的實際地景中遊玩，相較於相對靜態的桌遊，可以讓遊戲參與者對歷史情境有更多想像與「神入」（empathy）。另外，為了能夠透過故事來「再現」行動者所處的時空背景，開發成員需要更細膩地安排故事劇情與地

景之間的關聯，並賦予玩家在遊戲中進行活動的特殊意義。在將近半年的開發過程與活動執行，得以讓我們透過開發日誌來檢視其中的「啟發性」，以及團位成員的特質對遊戲設計過程形成的影響，細節的論述我們在第四章進行說明。

「進攻台灣府」的故事腳本的撰寫，以蔡牽與蔡牽媽相關的描述，以及與其攻擊府城戰役中相關的地點，組織起來的情節。這個故事腳本也由南台科大的學生團隊製作成 2D 動畫，於營隊舉辦時讓學生觀看。另外在四天的營隊過程中，使用「芝麻出任務」App 讓學員在地圖上觀看題目，並回答謎題，並配合實體道具和地圖遊玩。

由於「進攻台灣府」牽涉的地景較多，關卡設計較為複雜。「魂牽夢瀛」實境遊戲簡化為在一個實體展場空間內，僅設計五道最關鍵的謎題，並且亦吻合先前撰寫的故事脈絡，只是再透過另一種編劇巧思（路上遭遇到蔡牽的鬼魂），來引導遊戲的進行。「魂牽夢瀛」的展場佈置以室內牆面，對應真實的城市空間方位，在牆面上張貼情境氛圍圖。與關卡相關的地景分別是赤嵌樓、媽祖樓、鹿耳門天險與水仙宮。

4. 結果與討論：「進攻台灣府」設計「轉譯」過程與其中的「啟發性」

本研究將歷史遊戲的創作與開發，視為由「價值意識遊戲」(Flanagan & Nissenbaum，2014)框架：「發現→轉譯→驗證」的程序中發生，並進一步可拆分為文史資料解讀、轉譯、企劃構思、雛形產生、驗證與調整等細節步驟。針對清代府城五條港的文史脈絡，我們希望透過理論與實務的結合，觀察遊戲設計過程中的樣貌。跟一般的商業遊戲開發不同，我們無法投入完整的美術、程式、編劇、音樂製作人員。在開發的人力資源有限的情況下，我們以具有人文背景的研究員作為小組核心，搭配學生團隊，打造能夠足以進行評估的企劃樣貌與遊戲雛形。

我們透過觀察計畫中進行的幾個專案，其中遊戲設計程序中進行的狀態，來探討其中「啟發性」產生的原因和特色。首先，用於評估這些遊戲的「價值」的標準，在於其遊戲體驗是否能引發玩家的「歷史思考」（historical thinking），理解這些歷史時空脈絡中的行動者，因其生存所需所做出的行動背後的脈絡。而設計程序中企劃方向的成形或改編，若能具有引導出前述特質的遊戲體驗，便將其視為「啟發性」出現之處。

在針對本研究之設計程序進行細節說明之前，以下幾個關鍵點是影響遊戲產出的重點：

(1) 文史資料解讀能力：台南文史工作者羅○○先生田調經驗豐富，解讀能力屬

於專家（expert），專任助理許○○小姐屬於熟手（intermediate），協力進行美術與數位工具的學生團隊是新手（novice）。

(2) 遊戲形式的限制：例如設計者決定採用桌遊、實境遊戲或混合多種不同的形式來進行遊玩，將影響實作的方式。

(3) 設計者的意識形態：從一些以台灣史為範疇的電玩遊戲中觀察的到，設計者本身的意識形態，包括對特定政黨/政權的觀點，或是對於社會經濟狀態的價值觀判斷，明顯會對遊戲企畫的方向產生影響。

(4) 角色扮演：以歷史為題材的遊戲，勢必與特定時空情境中的行動者有關，遊戲挑選的「角色扮演者」的視角，是形成遊戲主要體驗的重點。

如（表 2）所示，我們將實境遊戲「進攻台灣府」的設計程序，以及對應的工作描述與重大設計決定，在「發現→轉譯→驗證」三階段來進行觀察。從「發現」階段探詢可能方向，確定明確方向後進行「轉譯」階段，「驗證」階段則是透過實際遊戲的體驗並反思設計過程。在表中的「工作描述」中可看到田調、史料範疇、遊戲開發、活動舉行的主要發生位置。「重大設計決定」則是我們能夠觀察設計過程中，「啟發性」發生之處與其原因。

遊戲構思初期從 2016 年 11 月開始，由田調活動出發，透過辨認實際地景與史料範疇之間的相關性，擬定為探討清代府城五條港脈絡常見的商貿視角，而核心的遊戲玩法訂定為小人物經商之路，與小說《水神》的題材類似。玩法並具有策略遊戲的味道，讓玩家可以管理資源，從事活動來擴張商業規模。然而，在邀請台南文史工作者羅○○先生加入團隊之後，在 2017 年 3-5 月之間的讀書會，以及實驗室團隊與政大「轉注藝遊」團隊合辦工作坊活動，研讀由故宮研究員提供的「同安船」史料範疇之後，遊戲主題由商貿視角轉變為考量海盜外敵入侵府城，進而重新組織包括港道變遷、軍工廠、蔡牽攻台行動、府城城防與王爺信仰等史料範疇的脈絡。遊戲的核心玩法也轉變為扮演「海盜攻台的情報蒐集者」角色。「轉譯」過程中在專案一些限制下，例如在實際城市空間遊玩、使用「芝麻出任務」App 結合實體道具、繪製美術與製作動畫，將蔡牽與蔡牽媽的故事做為主軸，搭配設計謎題與安排線索。將文史範疇進行「轉譯」時，團隊對於故事劇情鋪陳與當時時空背景有關的地點，需做細節的考察，並呈現在遊戲的玩法之中。最後的「驗證」階段則是透過觀察實境遊戲活動，並透過拍攝紀錄片，探討遊戲參與者的收穫與心態轉變，進而反思與檢討日後的遊戲將如何修正改善。

綜觀「進攻台灣府」實境遊戲約半年的開發過程，我們討論其中具有前述所定義「啟發性」發生之處，以及其發生原因，在（表 2）中以「★」標示，並以其數量表示其影響程度。首先是在「發現」階段 1-3 中，團隊加入具有鮮明而自主的意識形態的文史工作者羅○○，羅先生在台南市經營獨立教育工作室已將近

四年，而且在文史資料解讀能力屬於專家等級，與專任助理許○○互相配合的情況下，很快就確立遊戲調性。

另一個有趣而具有「啟發性」的轉變，在階段 1-1 過渡至 2-1，團隊從常見的以「商貿視角」解讀五條港史料，轉換為考量外敵威脅府城的「海盜視角」，並且決定將「角色扮演」由小人物經商之路，改變為替海盜工作的情報收集者。其中由於故宮研究員在與許○○合作舉辦工作坊活動時，提供了明確而有趣的「清代同安船」史料（階段 1-4），描述了許多在清代「官府－海盜」在東亞海域上的競合關係，包括有關蔡牽、李長庚、王得祿等人的事蹟。此一筆新的史料範疇產生的外部刺激，在團隊內部成員的意識形態判斷上，明確定調為找尋「台南被遺忘的海洋想像」（階段 2-1）。不僅在題材上較有新意，與許多動漫和電影中引人入勝的海盜情節有關，容易在遊戲玩法上創造特點。以歷史為題材的遊戲開發，與純娛樂與僅依賴想像力的企劃方向不同，有賴於根據某些已消化並彙整的文史知識脈絡（例如同安船史料），再據此進行遊戲企劃發想。若不是故宮研究員的外部刺激，在有限的計畫執行期間中，不可能再從頭進行完整的文史調查。

最後在遊戲企劃方向確認之後，「轉譯」階段 2-2～2-4 需考量的設計細節，則是如何在故事劇情（階段 2-2）、決定採用的遊戲類型與其玩法（階段 2-4）、科技工具（階段 2-3）的限制之下，透過關卡與謎題設計挑戰玩家，創造有趣的遊戲體驗。其中在階段 2-2 與 2-4 設計的困難程度較高，如何在關於蔡牽攻台的故事情節下，將地景有關的文史脈絡融入其中。例如讓玩家辨認出西外牆的輪廓、找尋軍工道廠的確切位置、在軍事戰略上考量欲採取的行動。在設計時不僅要熟稔田調的地點，對歷史事件（虛構或真實）的事態發展，也要有明確而清晰的全貌。這項工作的轉譯困難度，明顯高於坊間常見的實境解謎遊戲其中的謎題設計，無文史背景或對田調地點不熟悉者，無法勝任。

（表 2）「進攻台灣府」實境遊戲的設計程序

階段	編號	工作描述	重大設計決定
發現	1-1	● 進行田調：作為構思起點 ● 《穿越五條港》文學踏查地圖	● 意識形態（預設）：商貿視角
	1-2	● 參與《水神》小說走讀活動 ● 史料範疇：郊商、工人、廟宇信仰、商業活動	● 核心玩法：具不同陣營的策略遊戲 ● 角色扮演：小人物成為大商人的成長之路
	1-3	● 開始與台南塾讀書會 ● 進行田調：普濟殿至信義街（屎溝墘區域）	● 成員增加：文史資料解讀專家進入，具有明顯而自主的意識型態（★★★）
	1-4	● 史料範疇：故宮同安船（蔡牽、李長庚、王得祿） ● 與政大「轉注藝遊」計畫合作	● 史料範疇擴充：在商貿範疇之外，加入戰爭與海盜行動者的考量（★）
轉譯	2-1	● 進行田調：台灣道廠、台灣府廠 ● 史料範疇：港道變遷、軍工廠、蔡牽攻台、府城城防、王爺信仰	● 意識形態：台南被遺忘的海洋想像（★） ● 角色扮演：仿蔡牽的行動，玩家扮演海盜的探子
	2-2	● 撰寫故事劇情：蔡牽與蔡牽媽的故事 ● 安排故事分段中的地點與任務 ● 進行田調：小北門、赤崁樓、媽祖樓、軍工廠、鹿耳門文館、府城天險、洲仔尾、北汕尾、西外城、柴頭港	● 透過故事來「再現」行動者所處的時空背景 ● 考量在故事的限制下，發展玩法，並將構想精緻化（★★）
	2-3	● 美術繪製與數位工具：製作動畫、芝麻出任務 App、實體道具	● 在科技工具的限制下發展構想（★★）
	2-4	● 關卡設計：設計謎題與安排線索	● 團體的腦力激盪（★★）
驗證	3-1	● 舉辦活動：四天營隊 ● 拍攝紀錄片	● 參加遊戲參與者的收穫和心態轉變
	3-2	● 反思與檢討	● 設計者獲得的反饋，日後如何修改遊戲？

5. 結論與建議

本研究欲探究歷史遊戲設計所需的概念框架與設計程序，在清代府城五條港的文史脈絡下，透過實際籌組團隊來開發遊戲，並從開發過程的日誌紀錄加以觀察其中的現象與活動。我們將「歷史遊戲設計」此一活動置於公眾史學與大眾史

學的發展脈絡來觀察。此類的遊戲設計不僅需要異質而專業的各項人才一同進行，對於設計過程中由「歷史－遊戲」融合而成的中介知識，亦是目前研究者難以觀察並探討的議題，本研究的結論闡述如下。

(1) 「歷史問題空間」概念框架對設計過程的助益與其限制：

作為本研究發想的主要起點，麥柯(McCall，2012a)提出的「歷史問題空間」概念框架。由於其具有「歷史－遊戲」中介知識的特質，在遊戲設計過程中的初期「發現」階段，的確具有某種程度的引導性。能夠透過系統性的討論來理解歷史行動者背後的心理趨向、文化框架與情境支持，由本研究的「划水仙」桌遊設計過程可觀察到的確有其效用，並且能夠據此進行後續的遊戲設計構思。然而，團隊對於遊戲設計的熟稔程度，以及是否決定受到現有遊戲的玩法限制，例如在本研究中決定參酌「怒海求生」桌遊之設計手法。遊戲的玩法本身似乎對於形成的遊戲體驗有重大的影響。相較之下，「進攻台灣府」實境遊戲的發想，並未由「歷史問題空間」的框架下做為思考起點，在有經驗的文史工作者帶領下，即便是相對單純的遊戲玩法（找尋線索、回答謎題），並非透過像桌遊常具有的相對複雜規則，創造出來的遊戲體驗，仍具有很高的啟發性，並且對於遊戲參與者的歷史思考有其助益。

(2) 以遊戲設計工作坊導向的活動需要更精緻的活動支持

不管是前述所探討的「歷史問題空間」和「遊戲設計概念工具」的效用與侷限性，進行歷史遊戲設計的核心關鍵點，仍在於是否能夠對於歷史情境進行思考並產生洞見。在本計畫進行的初步成果當中，我們已窺見透過跨領域的工作坊活動，在更細緻的安排之下，能夠發揮這些研究框架與工具的效用。例如工作坊活動的進行，至少應該以四週以上的連貫性課程安排，包括前期的文史資料選材與研讀，到後期設計出來的遊戲試玩與回饋修正，皆需要再使其更精緻化。

(3) 經驗豐富的文史專業者，其意識形態與人文洞見是引導遊戲企劃方向的關鍵

在「進攻台灣府」實境遊戲的設計程序中，我們可以觀察到對於文史資料解讀與田野調查經驗豐富的參與者，對於創造歷史遊戲體驗的重要性。在遊戲的開發過程中，文史工作者羅○○先生實際上在不斷理解既有的數位工具或遊戲玩法的可能性，並藉由其意識形態與對於特定歷史議題的觀點，來引導遊戲設計的過程，並使其逐漸成形。若我們的目標是將歷史遊戲設計的過程，視為一種能夠揭露並重新演示學習的過程。前述的更細緻的工作坊活動，肯定要思考人文領域專家的意識形態與觀點，如何在活動過程中一併被培養與導引出來。

(4) 建構「歷史－遊戲」中介知識的必要性

在目前國外的歷史遊戲研究中，在後現代史學與歷史教育此兩大議題的引導下，已有學者能夠辨明歷史遊戲作為一種特別的大眾媒介的獨特性質。不僅在於遊戲本身具有的主動性，而非被動的閱讀的特性。體驗一個設計良好的歷史遊戲的意義，肯定不是在其「娛樂性」而已。我們所謂的「歷史－遊戲」中介知識的定義，會接近如查普曼所提及的「重建主義－寫實模擬」與「建構主義－概念模擬」這兩大類歷史遊戲形式(Chapman，2016)。這兩類形式的成份之一（重建主義/建構主義）是由後現代史學中對於知識論的討論而來，另一個成份（寫實模擬/概念模擬）則是將歷史遊戲視為「模擬」（simulation）或「再現」（representation）的意義。我們認為這樣的中介知識，是未來發展歷史遊戲設計的重要關鍵。

在國外的歷史遊戲研究學者已普遍接受的觀念之一，例如 Uricchio(2005)極具啟發性的文章中提及的，歷史遊戲做為容易產生「反歷史」和「虛構的歷史」的特性。現代主流的歷史教育觀點，也早已跳脫純粹記憶與背誦歷史知識，肯定如何透過學習歷史，進行歷史思考，並考量歷史的「當代性」。在歷史普及的書籍中，我們也常看到對於史實以外的思考(Ferguson，2005)或「歷史的『如果』（What If）」(Cowley，2011)。在計畫執行的過程中，我們不禁好奇如寓言或虛構的文類的敘事手法，是不是也能引發吾人對於特定時空情境的思考？未來在本研究團隊持續進行的數位人文主題計畫當中，也將持續對於這些議題進行深究。

參考書目

Chapman, A. 2016. *Digital Games as History: How Videogames Represent the Past and Offer Access to Historical Practice*. Oxford, UK: Routledge.

Cowley, R. 2011。《*What If？*史上 20 起重要事件的另一種可能》（王鼎鈞譯）。台北市：麥田出版社。

Ferguson, N. 2005。《反事實的思考－虛擬歷史》（楊豫譯）。台北市：知書房 。

Flanagan, M., & Nissenbaum, H. 2014. *Values at Play in Digital Games*. Cambridge, MA, USA: The MIT Press.

Jenkins, H. 2004. Game design as narrative architecture. In N. Wardrip-Fruin & P. Harrigan (Eds.), *First Person: New Media as Story, Performance, and Game*. Cambridge: MIT Press, pp. 118-130.

McCall, J. 2011. *Gaming the Past: Using Video Games to Teach Secondary History*. New York: Routledge.

McCall., J. 2012a. "*Historical simulations as problem spaces: criticism and*

classroom use, ” Journal of Digital Humanities, 1(2), http://journalofdigitalhumanities.org/1-2/historical-simulations-as-problem-spaces-by-jeremiah-mccall/.

McCall., J. 2012b. “*Navigating the problem space: the medium of simulation games in the teaching of history*,” History Teacher, 46(1), pp. 9-28.

Squire, K., & Jenkins, H. 2002. The Art of Contested Space. In L. King (Ed.), Game On: *The History and Culture of Video Games*. London: Laurence King Publishing, pp. 64-75.

Uricchio, W. 2005. Simulation, history, and computer games. In J. Raessens & J. Goldstein (Eds.), *Handbook of Computer Game Studies*. Cambridge MA, USA: The MIT Press. pp. 327-338.

Paper 2

Title

歷史街區的文化保存與創意開發之互動觀察：以臺南五條港街區為例

Observing the interaction of cultural preservation and creative development: a case study of old five channels cultural zone in Taiwan for example

歷史街區的文化保存與創意開發之互動觀察：以臺南五條港街區為例

陳姿汝
助理教授
南臺科技大學 視覺傳達設計系

摘要

隨著五十年來數位技術發展迅速，衍生了各種社會問題，如文化保存受到漠視、全球化浪潮壓縮在地生態的永續生存等，為因應這些困境，「創意城市」的議題受到廣泛的關注，創意城市一般被認為與多元文化、創意人才、科技等三要素有關，而現今臺灣老街大多具備此三個條件。然而，臺灣老街近年來因為都市更新、過度商業化之因素，使得許多老街原有樣貌被破壞，且青年人口外移、新住民進駐等問題，使得傳統街道文化、原始生活模式流失，另外，也因為國人長期以來對自己生活土地的過往歷史不重視，對在地文化認同感不足，缺乏對本土文化的深度體驗與認知，使得對歷史街區之創新開發上，多著重於經濟發展和觀光產業之面向，缺少對文化創新應用思維之強調，因此老街場域中的創意氛圍難以有效彰顯。本研究將觀察「創意群體」的形塑與發展過程，並提出一些觀察現象。研究方法採「個案研究法」，並與此創意商家合作開發「街刊」，藉此串聯信義街之創意群體，透過對臺南老街人、事、物的參與觀察、訪談和實踐，分別從信義街中的三個族群—「新住民」、「老居民」和「拜訪者」來個別分析、討論。研究結果即提出創意群體建立之觀察與流程，希冀此能做為未來文化創意區域開發的參考。

目次

1. 前言
1.1 研究動機與重要性
1.2 研究問題與目標
2. 先前相關文獻
2.1 全球創意城市的興起
2.2 臺灣的創意街區與創意群體

3. 研究方法與步驟
3.1 案例選擇
3.2 訪談與街刊發行
4. 研究結果
5. 結論
6. 參考文獻

關鍵詞

創意群體，數位媒材，歷史街區，文化保存

1. 前言

1.1 研究動機與重要性

近年來全球興起一股創意城市(creative city)的熱潮，起因於這五十年來數位技術發展迅速，對全球很多區域造成很大的衝擊與變化，各區域過度追求科技進步的同時，卻也衍生了各種社會問題，如全球經濟發展停滯、文化保存受到漠視、全球化浪潮壓縮在地生態的永續生存等，為因應這些困境，創意城市的議題受到廣泛的關注，而在 2004 年聯合國教科文組織正式建立起「全球創意城市網絡」(the creative cities network)，並將創意城市分為媒體藝術、文學、工藝品、音樂、美食、電影及設計等七類，目前已有來自 54 個國家，共 116 個城市已正式成為該網絡的會員(UNESCO, 2017)。創意城市一般被認為與多元文化、創意人才、科技等三要素有關，而現今臺灣老街大多具備此三個條件，本研究觀察，由歷史街區轉型成為創意街區的發展型態，或許已悄然成形。

臺灣是全球蘊含豐富中華文化的代表區域之一，其自由、活絡的創作環境，曾歷經不同時代的殖民和不同民族的統治，使得這個小小的東亞島國成為一個文化薈萃之地，同時也造就了今日文化創意活動的蓬勃發展。其中，老街文化意識之抬頭極具代表性，老街風貌保留了都市發展過程中的歷史痕跡，更形塑出一種由當地居民共同構築的「文化生活圈」，具備豐富之人文內涵。有關老街區的建築與文化之創意開發與產業創新之探討，在全球皆是重要命題，老街區因為富含豐富歷史文化要素，吸引越來越多的創意人才進駐，而這些創意人才一方面為吸取此原生之人文養分，而試圖沉浸於此歷史環境中，也因而各種歷史體驗活動和科技因應而生；而另一方面，這些創意人才為從此養分中尋求能安身立命之機會，導致各種創意活動蓬勃發展，形塑出屬於此類創意街區的產業型態。

1.2 研究問題與目標

而在文化資源的考量下，本研究選擇具有豐富歷史資源的「臺灣臺南五條港街區」為主要研究場域，五條港是十七世紀東亞港口商業貿易中心，人文內涵以民間經濟活動為主，並是以「港口市街」的型態孕育而生，目前很難得的尚能保留部分舊港街市的景象和建築之樣貌，包含港灣、碼頭、媽祖廟、暗街、官道、驛站和閩南式的古厝和街屋等內容(趙文榮，2006；張溪南，2007)。而五條港區也是全臺祭祀廟宇密度最高之區，由於五條港是先民來臺最早上岸的地點，因此保留許多大陸移民所帶來的原鄉文化和信仰(鄭道聰，2013)。從古至今，五條港地區一直是庶民可以自發發展的特殊場域，雖然經歷清領、日治和民國三個時代的變遷，但都於其中扮演著文化引領之角色，五條港現今仍是臺南的首善之區，延續著五條港的活力與風華，五條港區特殊的港口人文現象，直到今日都還深深

影響當地居民(黃婉玲、李孟哲，2007；鄭道聰，2013)。

觀察到現今五條港老街中仍有多座古蹟建築和文化空間（百年廟宇）座落於居民之生活場域內，而各種經濟活動和創意經營團隊大多圍繞在此歷史建築周圍而生，形塑出一種特殊的共生關係，可見歷史古蹟與現代經濟活動是有機會於相同時空下並行。然而，臺灣老街近年來因為都市更新、過度商業化之因素，使得許多老街原有樣貌被破壞，且青年人口外移、新住民進駐等問題，使得傳統街道文化、原始生活模式流失，另外，也因為國人長期以來對自己生活土地的過往歷史不重視，對在地文化認同感不足，缺乏對本土文化的深度體驗與認知，使得對歷史街區之創新開發上，多著重於經濟發展和觀光產業之面向，缺少對文化創新應用思維之強調，因此老街場域中的創意氛圍難以有效彰顯。因此，本研究即是希望能針對現今創意群體(creative clusters/groups)在臺南歷史街區中的形塑與發展過程，並提出一些觀察現象。

2. 先前相關文獻

2.1 全球創意城市的興起

2008 年 TIME 雜誌以「Ny-lon-kong」為封面主題，討論國際城市競爭下創意經濟與創意氛圍的重要性，並提出當前世界城市競爭的重點不再是興建重大地標建築或是創造企業主優良的投資環境，而是如 Florida(2005)在《城市與創意階級》(cities and the creative class)一書中所提及的區域發展，即如何創立一個能吸引各種創意人才的環境氛圍，提出當一個城市具備注重生活環境、創意、品質、酷魅力(coolness)、創意氛圍、文化設施(cultural amenities)等特性，並且彰顯如自由、都會主義、波西米亞和多元文化等都市意象時，才能夠吸引科學家、工程師、高級服務業、知識經濟專業人士等創意階級進駐，進而達到創意經濟、提升城市競爭力之目的，也就是說，當一座城市的經濟發展是透過聚集創意人才與高科技產業來達成，同時也具備開放多元的生活空間時，便可稱之為「創意城市」(creative city)。

創意城市一般被認為在都市發展策略上，應重視知識文化和相關的硬體建設，打造優良的創意氛圍，吸引創意人才匯流，並以促進跨領域的創意思考作為重要目標(Landry, 2008)。而如何創立一個能吸引各種創意人才的環境氛圍，Florida(2002)認為創意人比較偏好多元化、包容力大、對新觀念開放的地方，而一個地方想要吸引創意人、激發創新能力與刺激經濟成長，需具備科技(technology)、人才(talent)和包容(tolerance)三個條件，創意與創意階級會在具有此三 T 的地方生根。另外，Landry(2008)指出，創意城市的建構，需透過文化節慶、都市意象、文化創意群聚等方式、刺激文化經濟驅動、整體經濟發展，達成

都市再生的目標。然而，Scott(2006)指出大城市可能擁有空前的創新、創意能力，惟社會中仍存在著文化、經濟等的不平等情況，不僅是資源分配問題，同時還牽涉到基本公民意識及民主問題，至於這些根深蒂固的老問題，往往使得創意城市難以真正的落實。綜觀先前學者對創意城市形塑的觀點，多與文化、創意人才、科技等三要素有關。而其中，文化的重要性在於它是創意的平臺與資源，文化能提供創意所需的素材(raw materials)，創意是實踐觀念與思維的力量(the power of thinking and ideas)，而此一力量需要文化所蘊藏的豐富資源如多元的價值、生活方式等，文化可視為是創意源源不絕的溫床(Landry，2008)，可見「文化」此要素在創意城市觀點中的重要性。

2.2 臺灣的創意街區與創意群體

臺灣近年也因文化創意產業的推行，有越來越多的「創意街區」在臺灣各地便地開花，2008 年當時的臺北文化局長劉維公就曾提出：「創意街區強調的是生活美學、生活價值觀、生活風格等，創意街區的宗旨就是讓更多人認識自己生長的地方，進而找到自己的生活價值與生活風格。」(陳歆怡，2012)，而臺灣對「創意街區」指認，多偏重於「巷弄美學」與「生活風格」的營造(賴彣華，2014)。也因此，臺灣老街是具有豐富文化資源的場域，也吸引越來越多的創意人聚集，而形成「創意群體」(creative group)，創意群體是指有一群具有創意才能的人，能從事創意產出，而創意產出通常是和環境及社群互動之累積而設計出來的，這些創意活動通常聚集於固定的地點，因而將之稱為「創意群聚」(creative clusters) (Propris & Hypponen, 2008)。邱詠婷和余倩瑋(2014)整理創意群聚，認為其有「群體」與「空間」兩層意義，並採用偏向「群體」的定義，即創意群聚是「一群具有創意才能、從事創意產出、喜好主題不同但都對新奇事物具有高度興趣的人，聚集於特定地理空間所形成的群體」。而其中，創意氛圍指的是一個讓群聚能夠持續創新的環境溫度，也是能夠讓聚落持續運作的能源(祁政緯，2012)。

而觀察臺灣這百年來社會經濟的發展脈絡，臺灣老街的發展與演變是與其所在的地理環境、商業行為和人文活動有絕大的關係，是整個大時代的縮影(張溪南，2007)，觀察臺灣老街成形之脈絡，可窺知臺灣經濟、社會、文化發展史，而這些老街場域中最主要的歷史內涵，是創意街區發展的基礎，然而觀察臺灣現今在此類歷史街區所營造的創意氛圍，多數皆忽略或甚至不知其歷史脈絡，僅強調當今世代的仿古與流行文化，在這樣的情境下進行創意街區的開發，則很可能會造成歷史斷層與文化遺失之危機。

3. 研究方法與步驟

3.1 案例選擇

本研究方法採「個案研究法」，選擇『臺南市信義老街』為觀察與分析對象，信義街位於臺南舊市區，目前尚保有獨特社區氛圍與老街文化，蘊藏豐富的老屋故事，其曲折的小巷與民宅彼此交錯，只有在巷弄交會處或廟宇前，才有較大的空地廣場出現，狹小的巷道，只有行人和少數機車、腳踏車能夠通行，是一條具有人性尺度和空間趣味的巷弄，請見圖 1。也於在明清時期此地為五條港海域，有便捷之水運經過，因此，信義街保有豐富的五條港文化內涵，隨處可見的歷史遺跡如集福宮、媽祖樓、百年老屋等古蹟，融入居民的一般日常生活中，然而，由於信義街位於臺南市中西區，此區曾為臺南商業中心，但在都市更新計畫後，新建圓環和道路，開闢出許多崎零小地，使得信義街被後來新建的金華路和康樂街所截斷。也因為當時政府只以商業因素作考量，尚未有都市文化保留之概念，導致老街社區紋理被破壞；又因信義街位於臺南舊市區，公共設施服務水準相較低，以至於信義街有青年人口外移之問題，相較於其他多數過度商業化的老街，目前信義街社區尚保有獨特社區氛圍與老街文化，且近年陸續有青年進駐進行街區改造，具備老街文化保存與文化創新之內涵。

此外，信義街之街頭有一座二級古蹟—「兌悅門」，兌悅門現位於由文賢路轉向信義街之路口，臺灣府城大西門的外城門，因位處城西，在八卦中屬「兌」方，由此命名而成，此是目前臺南城門中唯一尚供通行使用的城門，也是臺南市唯一現存的外城門，於道光 16 年（西元 1938 年）丙坤元月興建完成，有關兌悅門之使用現況，請見圖 2。由於城門多由砧硓石搭成，又稱「砧硓石城門」，是為「甕城」，又稱「子城」，門高四公尺、厚三公尺，門洞寬三公尺，而城門牆寬度僅在 30 公分至 60 公分之間，上面並未建有城樓，城面題有「兌悅門」石橫額，及「道光拾陸年元月建」之落款。為方便內開城門，其城門結構外窄內寬，城門內左側有階梯可登城台(黃秋月建築師事務所，2009；黃宜靜、王明雪，2003)。其城牆形式外表幾乎為紅磚造，台度為硓砧石，一圈圓拱貼石片，上為垛牆及鎗口則砌紅磚，其他部分則砌磚或以石塊勾縫，材料以石灰、黏土、砂、牡蠣殼粉末拌合而成(黃秋月建築師事務所，2009)。



圖 1 臺南信義街街景（圖片來源：本研究記錄）



圖 2 兌悅門實景圖（圖片來源：黃宜靜、王明雪，1995）

3.2 訪談與街刊發行

五條港區在經過各年代的改朝換代之下，仍不減其對區域地理發展之重要性，現今的五條港文化園區就如同一座「生活文史博物館」，傳統與現代建築、藝術、文化共存於同一個空間中，也讓許多的文史工作者在此進行重要的田野考察工作，讓更多的歷史記憶能重新被拼湊起來，在此豐富文化資產的環境中，也同時吸引了一群從事創意工作的藝術家、建築師、設計師等進駐於此找尋創作素材。在上述各種條件下，使得五條港文化園區的創意生活型態悄然成形，而另外再透過數位工具、新媒材的輔助，或許有機會讓五條港此類歷史老街有機會轉型為創意街區，讓歷史與創意能同時並存於其中，豐富五條港之多元面貌。也因此，在這些創意商家部分，挑選部分在地指標人物之深度訪談紀錄與結果，獲得受訪者同意後，已將部分受訪者之訪談結果初步整理為刊物形式提供大眾免費取閱，其內容包含店家資訊、人物介紹、人物故事等，見圖 3，並藉此串聯信義街之創意群體，並將數位互動科技，如 QR Code、擴增實境等技術，應用於街刊之內容中，見圖 4，透過對臺南信義老街人、事、物的參與觀察、訪談和實踐，分別從信義街中的三個族群—「新住民」、「老居民」和「訪問者」來個別分析、討論。研究結果即提出創意群體建立之觀察與流程，希冀此能做為未來文化創意區域開發的參考。



圖 3 臺南信義街刊，訪談四位在地店家（圖片來源：本研究記錄）



圖 4 將擴增實境技術應用於街刊設計（圖片來源：本研究記錄）

4. 研究結果

而有關創意群聚的實際觀察，本研究以臺南信義街為例，這兩年來信義街吸引越來越多的各式創意人才進駐，造就當地創意氛圍的形成，也讓小巷弄原本濃厚的歷史人文氛圍，逐漸添增越來越多的創意活力氣氛，讓古老的歷史基調和新興的創意思維二者界線越來越模糊，如圖 5，本研究初步整理信義街區之新舊景點介紹，發現在地居民和特色店家多圍繞歷史景點而生，歷史景點多是過去重要生活場域，如廟宇、城門、廣場等，此吸引著新住民之好奇與認同，此也成為新

住民在此停留、活動、進而生活之重要原因，新舊融合之現象在臺灣此類歷史街區亦為常態。

而進駐的創意人類型多元，如隱身在巷弄內的手工文創小店「**莎普萊囍**」是一間以各種數位製造工具來開發商品的店家，而「**能盛興工廠**」是一群充滿熱情與理想的青年以獨樹一格的方式學習與在地共榮，另外，目前提供餐飲服務的「**筑馨居**」則是被當地老屋所吸引進駐，因此對老屋的維護不遺餘力，還有在當地經營檳榔攤生意已經超過三十年的「**馬路楊檳榔會社**」老闆馬路楊，已經是當地重要的文化指標人物，見圖 6。近年可觀察到上述創意街區的氛圍，在臺南信義街已隱然成形，然而，創意群聚的現象是動態的，創意人才的變化也是流動的，而這群新住民的進駐與在地文人的互動過程與結果，正以極快的速度、極大的影響力決定此歷史街區的未來走向。



圖 5 信義街地圖，展現出在地居民和特色店家多圍繞歷史景點而生（圖片來源：本研究記錄）



圖 6 信義街地圖，呈現越來越多的新住民聚居於此（圖片來源：本研究記錄）

5. 結論

文化是人類生活的表現，凡文明之國無有不致力於歷史文化之弘揚與發展，而「歷史」是人們對過去事實的認知及其傳達的成果，臺灣過去社會中的教育體制上，缺乏教給下一代臺灣的歷史與人文，造成下一代對自己土地上各種事物的漠不關心，如今要實施本土化教育，卻發現找不到屬於這塊土地的記錄，這幾年來一些文史工作者，投入地方史的建構，才慢慢整理出臺灣的歷史資料(周樑楷，2005)。本研究從歷史街區轉型為「創意街區」之角度，希望能為臺灣老街之未來發展提供一種可參照之模式。

另外，老街具備當地不同時代、不同地域的發展特色，但現今老街發展趨勢多著重要經濟開發與觀光旅遊復甦之角度來考量，造成今日老街開發與產業經營模式的規則化與相似性，使得越來越多的臺灣老街逐漸尚失在地重要的人文資產，而臺灣老街在這幾年的發展過程中，也開始逐漸質變和轉型，過去的歷史文化場域因為吸引越來越多的創意青年進駐，而形成風格各異的創意場所，而本計畫之結果也希望能在這波老街轉型之浪潮下，提供一種可觀察的角度和期許，並希望能臺灣歷史街區之未來發展能有所貢獻。

6. 參考文獻

A. J. Scott, 2006. Creative cities: Conceptual issues and policy questions, *Journal of Urban Affairs* 28(1), pp. 1-17.

C. Landry, 2008. The Creative City: *A Toolkit for Urban Innovators*. London: Comedia, Earthscan.

D. P. Propris and L. Hypponen. 2008. “Creative clusters and governance: the dominance of the Hollywood film cluster, Creative cities, cultural clusters and local economic development, ” Cheltenham: Elgar, pp. 258-286.

R. Florida, 2002. *The Rise of the Creative Class*. New York: Basic Book.

R. Florida, 2005. *Cities and the Creative Class*. New York: Basic Books.

United Nations Educational, Scientific and Cultural Organization, 2017. *What is the Creative Cities Network?* Retrieved February 25, 2017, from http://en.unesco.org/creative-cities/content/about-us.

周樑楷編撰（2005）。《總論》。彰化市：彰化縣文化局。

祁政緯（2012）。〈創意群聚效應對都市區域再生的影響—以中山雙連創意街區為例〉。未出版之碩士論文。

邱詠婷、余倩瑋（2014）。〈文化創意聚落對民眾生活美學影響之研究〉。《國民教育》54(5)，頁 44-56。

張溪南（2007）。《南瀛老街誌。南瀛文化研究叢書：第 11 輯；南瀛地景文化專輯 53》。臺南縣：臺南縣政府。

陳歆怡（2012）。〈用設計改造城市〉,《光華雜誌》，頁 111。

黃秋月建築師事務所（2009）。《台南市第二級古蹟─兌悅門。修復工程工作報告書》。計畫委託機關：台南市政府。黃秋月建築師事務所。

黃婉玲、李孟哲 撰稿（2007）。《追憶城西帆影：台南市五條港歷史區域地方文化館》。臺北市：行政院文化建設委員會。

趙文榮（2006）。《南瀛內海誌。南瀛墾拓春秋專輯》。臺南縣：臺南縣政府。

鄭道聰（2013）。《大臺南的西城故事。大臺南文化叢書 1─地景文化專輯》。臺南縣：臺南縣政府。

賴玘華（2014）。〈製造創意、設計臺北：解構創意城市的都市/文化治理〉。《中華傳播學會研討會 2014 年會論文》。

Paper 3

Title

臺南當代藝術訊息重植之跨世代媒材敘事轉譯研究：《映。他聽》傳統媒材敘事建立

A study on the narrative translation of generation-crossing media in the process of rebuilding contemporary art information in Tainan: reflection, apprehended by the others

台南當代藝術訊息重植之跨世代媒材敘事轉譯研究
-《映。他聽》傳統媒材敘事建立

楊斯嵐[1], 林煌迪[2]
[1]助理教授 南臺科技大學 資訊傳播系
[2]藝術家及兼任講師 成功大學 建築系

摘要

藝術與文化應為一體二面，台灣的藝術文化卻與常民生活有著文化脈絡上的落差，本研究的初期基礎研究旨在探討台灣的文化離根現象，研究之目的有二：(1)前期以台南五條港為研究核心，瞭解台灣的文化離根現象；(2)後期尋找現代傳播技術輔助藝術文化重植的可能性。本計劃共完成六場座談，合計取得 144 人時的目標族群論述，並依此進行十位代表性人物重點訪談，累計 196 小時 5652 段檔案共 2.562TB 第一手數位資料，初步發現台南當代藝術持續存在如研究預期般的官、傳、藝、民的分離狀態，且無收合跡象。本研究為數位人文主題計劃 2016 年「運用數位傳播技術於踏查台灣藝術離根化文化現象之實作評估」與 2017「跨世代媒材敘事轉譯研究：台南當代藝術的信息重植」的綜合成果，其中 2016 年進行田野調查之數位紀錄、2017 建立《映。他聽》之傳統數位媒材型式，同為發展 2018《應。你談》網路社群媒體與 2019《In。我獨白》等次世代傳播技術之研究比較基礎。2017 年起之計劃亦為「府城之數位人文創新-歷史遊戲、老街人文、當代藝術、兒童閱讀」之子計劃，計劃期程三年。2018/3 月完成「城內/城外」紀錄片乙隻，本研究同時輔助台南當代藝工作者建立當代藝術近代的重要年表，完成「台南藝術初探 城內/城外」一書，及以數位影像內容支援台南文化中心「城內/城外」藝展。

目次

1. 前言
2. 文獻探討
2.1 當代藝術
2.2 訊息溝通與傳播
3. 研究方法、步驟及執行進度
3.1 研究方法

3.2 第一年計劃-「映。他聽」研究區塊
4. 研究結果
4.1 資料收集
4.2 第一年《映。他聽》-53min 傳統訊息紀錄片敘事結構
5. 結論與討論
6. 參考文獻

關鍵字

台南當代藝術、文化離根、訊息重植、跨世代媒體、敘事轉譯

1. 前言

依 2015 年度數位人文計劃案-「運用數位傳播技術於踏查台灣藝術離根化文化現象之實作評估」初期所發現之現象再進行後續研究。我們在試圖透過傳播技術連接各個散落的藝術孤島、改善台南當代藝術想像孤立化的狀態時，發現傳統的媒體創作者多半透過傳統創作方式記述藝術文化狀態，製作完成後卻可能馬上面臨被次世代的 AR、VR 媒體丟棄的命運。如此快速翻轉的媒體新世代，製作費用將遠超過非主流的族群所能承受。目前傳播領域缺乏不同媒材間進行演譯製作的研究，沒有有效的轉換模式，各媒材間的轉譯高額漫長且容易失真。本研究的目標即是如何讓同一訊息可以在不同的傳播媒材上進行有效的敘事轉譯，第一期在於產生傳統基礎影片 。

2. 文獻探討

2.1 當代藝術

藝評家謝東山在《台灣當代藝術 1980-2000》一書的序文中提到：『全球化』的思維模式正主導著我們在形形色色的社會生活實踐，包括藝術生產與行銷，它的穿透力早在八〇年代已然出現於台灣。…」(謝東山，2002)一九八〇年代末，第一批留學生的返鄉潮將流行於歐美的當代藝術思潮帶回台灣，同時間非官方與非商業的替代空間的另類體制同時引入台灣的當代藝術環境，形構出了台灣當代藝術發展的主軸。在台北市立美術館、高雄市立美術館與國立台灣美術館(原台灣省立美術館)之外，也在台北、高雄、台中與台南出現了若干的另類藝術空間，和甫成立的國立藝術學院(今國立台北藝術大學)與國立台南藝術學院(今國立台南藝術大學)形成了推波助瀾。台南是台北、台中與高雄以外一個藝文活動聚集區域，由於台南曾經作為台灣的政治經濟中心，在不同的時期皆聚集了藝文人士，並在文學藝術等不同領域中有蓬勃的發展，產生完整的產業鏈。然除少數受過台灣當代藝術教育的專業創作者，仍可與世界當代藝術發展保有較同步的互動關係，多數的民眾對當代藝術仍是陌生，或僅止於表面上的形式閱讀。台灣的當代藝術替代空間歷史來自於一九八〇年代末期，所有九〇年代「伊通公園」與「二號公寓」的種種傳說深植在當時藝術學子的記憶中。1995 年「新樂園」成立初期，組織及制度上大多沿襲「二號公寓」，提高對藝術自主性的要求，也更精緻。這樣的氛圍啟發了台南的「文賢油漆工程行」，也為台南的當代藝術發聲埋下伏筆。

2.2 訊息溝通與傳播

現代所謂的圖像傳播早已非遠古的岩畫所能比擬，媒體創作者與觀眾之間的影像溝通，全是來自己心理狀態，如壓縮(將符號融合)、替換(用一個符號取代另一個)(Freud，1949)。德國傳播學和媒介心理學家 Gerhard Maletzke(Meyen&Löblich，2006)定義大眾傳播須符合幾個特徵：透明的 (受眾不因人際交往關係，而範圍有所侷限)、使用科技發送訊息、間接的 (發送者與接收者之間有空間距離)、面向分散的群體 (群體無階層、行業、群組之區別，且受眾是匿名)、單向的 (發送者與接收者之間不得角色互換)。傳播媒體傳遞訊息隱藏著社會文化的內涵以及科技要素，無論缺少了哪一個，都可能會影響整體傳播媒介的發展。Wright(1974)認為文化決定論和科技決定論(technological determinism)的之間，存在著媒體與社會的因果關係(林日璇、李育豪、王茜穎譯，2014)，例如：進行歷史溝通最好的方式，就是利用大眾媒體來傳遞歷史資訊(張廣智，1998)，McQuail(陳芸芸譯，2002)描述大眾傳播是一種以大規模的方式運行，在某個程度上，幾乎能牽涉及影響社會中每個份子的傳播方式。可以看出文化與科技之間，環環相扣的傳播特性。傳統媒材與傳統訊息管道。電視、電影是一種動態影像的技術呈現，在動態影像的世界裏，電影的發展早於電視。電影在技術上被視為傳播的一支，亦因其特性近於藝術創作，在內容上被視為藝術而非傳播。當代電影之所以可以進一步發展，奠立了理論基礎，如：「精神分析電影理論」、「結構主義電影理論」、「女性主義電影理論」，是因為在電影藝術拓展了新的空間，並吸收運用當代科學技術，並將結構主義、現代主義、後現代主義以及理論話語融入電影製作實踐內(張廣智，1998)。

3 研究方法、步驟及執行進度

3.1 研究方法

敘事理論是一個關於敘事與敘事結構這兩者如何影響我們知覺的理論及研究，常見的戲劇、電影、娛樂等都在敘事理論的研究範圍內，Berger(姚媛譯，2001/2002)之定義：「故事是敘述一段時間所發生的事件，講述的是人、動物所曾發生或正發生的事情，這便是敘事」。就像 Berger(1997)指出敘事是說出某個時間裡，一連串「已發生」或「正發生」的事件，就像我們每天在與人交談時，會敘說出在這個時間或過去時間所發生的事情，這也是敘事的一種。Labov(1972)則強調敘事理論應要有過去經驗的摘述(recapitulating)，並要以發展先後次序的方式表達。(政大傳院媒介寫作教學小組，2009)。敘事的方面也包含了女性主義、巴赫金主義、電影理論…等交流，此主張顯現敘事學的多面性，並也因互相溝通的過程，分支了不同面向的分析研究，如主義敘事學、社會敘事學、電影敘事學與網路敘事學…等(Herman，1999；譚君強，2002)。本計劃進行敘事議題這項內

容時，將三年期分成三大重點，為不同媒材形式開發對應的敘事方式。第一年「映。他聽」針對現有已經存在多年，具有固定型式的傳統媒材，包含傳統的電視、電影、雜誌等，此區塊的特徵為已有理論架構、主要目標在建立第二、三年之比較基期；第二年「應。你談」為社群媒材，即所謂的新媒體，包含 FACEBOOK、BLOG、LINE、YOUTUBE…等，此區塊的特徵在於百家爭鳴，因仍處於高度變化之中，尚未發展出一致性的理論；第三年「IN. 我獨白」為次世代的實境技術，包含擴增實境、虛擬實境、混合實境…等。

3.2 第一年計劃-「映。他聽」研究區塊

本年度研究主要著重在「映。他聽」：府城當代藝術論述在傳統媒材上的敘事表現。主要研究計劃由以下五大區塊所構成。

(1) 藝術議題踏查。在台灣藝術發展的過程，經歷過不同植民及戒嚴時期，使得台灣意識壓迫，造成許多歷史記憶、認知思想散落在各地，我們暫時稱為「歷史碎片化」或「認知碎片化」。在此階段為藝術顧問群根據傳統媒材的特性，提出合適的議題歸類與脈絡，其中包含藝術家的社會角色、對議題的深入程度、文化認知、事件…等所有藝術議題的評估與建議。

(2) 敘事研究及相關影像分析。在傳統媒介所播放的影片大多為長片，本階段以「架構性敘事」的方式為探討傳統媒介播映的影片，此步驟收集多數在傳統媒介傳遞資訊的影片進行劇本分析。傳統的敘事結構有七個部份：序幕（prelude）、開場（set-up）、發展(development)、糾葛(intertwine)、高潮(climax)、結局(ending)、餘波(aftermath)（視覺傳播）。以這些結構依照不同類型的影片去分解出架構性的敘事型態、劇本的事件規劃分配、議題深入程度…等。本年度的實驗影片以傳統媒體作為載具，因此從傳統媒介所播映的影片進行技術上的分析與探討，分析不同類型影片的拍攝手法、分鏡、旁白、剪輯節奏……等技巧的應用方式，瞭解影片類型的差異如何牽動影像製作技術所呈現的樣貌。

(3) 解析統整要素。找出傳統媒介裡，藝術、文化之影片的共通模式。文藝電影著重在人的情感上，音樂優雅、畫面浪漫、節奏緩慢，來表現內心沉靜的情緒，又或者黑色電影(Film noir)強調善惡劃分不清的道德觀作為題材的電影，類型較晦暗、悲觀，由製作技巧來顯現此類風格的影片，像畫面昏暗、拍攝角度較低、煙霧效果…等，利用一些手法傳遞一種氣氛(Raymond, B. & Chaumeton, E.，2002)；本年度以傳統的大眾媒體為傳遞資訊的工具，考量台南當代的藝術歷史及文化客群，將傳統媒體的特性視為考量的因子之一，其中包括：影片長度、風格、劇本、分鏡、旁白、畫面設計、音樂、節奏……等，著重在傳統媒體的播映典型、呈現內容的正確性…等。

(4) 探討模式與原型試作。此階段為以文史小組為本步驟核心，透過轉譯，解析出前段時所獲得的資料，並進行相關型式的建立，產生出多種在傳統媒介所播映的影片模式，進而與顧問群討論，逐步漸構影片可呈現的樣貌，由藝術顧問針對試作影片提供建議，產生出更完整的影片構想，進行整理並修改、調整模式，在此步驟將會不斷來回修正不同的試作原型。為確定影片產出的效益及所需傳達的藝術內容是否對大眾有明顯的影響，這個階段為所有步驟中，最為繁瑣的一步。

(5) 展示發表。本步驟將於藝廊、美術館、替代空間…等作為影片展示之場地，並積極推廣至電視台等傳統媒體平台進行播映，以便獲得符合本年度為傳統媒體播放影像之效益。

4. 研究成果

4.1 資料收集

本計劃共取得 196 小時 5652 段檔案共 2562GB 數位資料，其中座談共累計 92 小時 1381 段檔案 1059GB 數位資料，重點訪談共 103 小時 4271 段檔案 1503GB 數位資料。其相關統計如表 1 所示

表 1 數位資料統計表

	次數	影片長度(hrs)	檔案數	檔案大小(GB)
座談會	6 場(48 人次)	92	1,381	1,059
重點訪談	10	103	4,271	1,503
共計	16	196	5,652	2,562

期中包含六場座談，合計 18 小時。每場各八位與談人，其中一次參與者 8 位、二次參與者 20 位，合計 144 人時。本計劃另進行十次重點訪談，每次一位當代代表性藝術家，並對其藝術工作進行紀錄。其分佈如圖 1 所示。

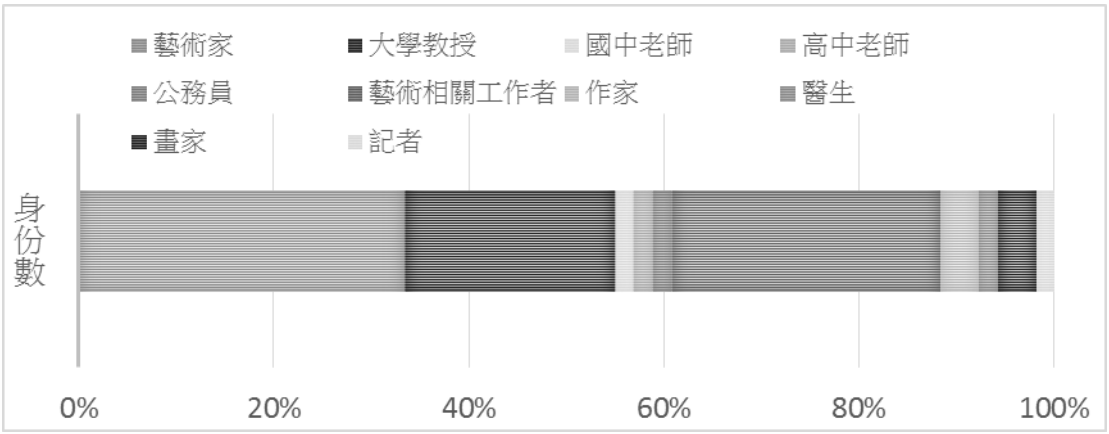


圖 1：與談人及受訪者社會身份分佈圖

4.2 第一年《映。他聽》-53min 傳統訊息紀錄片敘事結構

基礎影片採用台南老中青三個不同世代的十位當代藝術家口述，呈現對台南藝術生態環境的歷史影音紀錄。事件結構從九零年代甫歸國的時空、城內(台南縣市合併前的城區，所謂「府城」)與城外(台南縣市合併之前的「台南縣」)、南台灣新風格、當時的替代空間談起，並且也觸及自身創作的情況。以傳統影片述事手法，透過與談藝術家的回憶，窺見各位藝術家對於整體生態環境的不同角度的經驗，以 53 分鐘時間讓帶領受眾感受台南自解嚴以來的藝術樣貌。

以去年度計劃案與台南城內城外策展團隊合作共同耙梳之重要年表為影片基本時軸，針對十位台南當代代表性藝術家，進行口述採訪，並進行千餘人時之後製、剪輯、調光調色、字幕製作、校稿等，製作成為 53 分鐘紀錄片。劇本內容的專家諮詢累計 38 次，共計 38 小時。

表 2：敘事結構

元素	描述
目的	**藝術空間關係** 1980 年代後興起的「南臺灣新風格」，以及台南藝術大學成立之後，台南城內/城外因此前後興起各式各色的藝術展覽空間、替代空間們，隨之而來也帶動了新的藝術型態發展情形。藝術家們對此發表看法，其中包含了駐村、藝文展覽……等等，藉此紀錄相關這些展覽空間的成立緣起，及其與台南藝術生態環境、文化及歷史之間的關聯性。 **南台灣新風格** 針對 1980 年代間所發展的一種新的藝術創作風氣進行探索，呈現由此引發台南的當代藝術一連串的效應，包含藝術家在台南進行的藝術展覽或活動，包含藝術空間的成長。 **台南環境** 從老、中、青三個世代的藝術家來陳述不論是從外地移居台南，或者本來就是在地台南人的藝術家們口中的台南的歷史經驗。藉此我們得以紀錄並觀察藝術家們的台南經驗，不管是藝術生態、政治環境還是南北的文化差異；對於藝術家而言，台南身為文化古都又是何種看法。我們可從這些口述歷史中，了解到長期在台南創作的藝術

	家，經驗台南從 1980 年代到現在的環境變化、本身的核心思想或者創作脈絡。 **城內城外關係** 李奧・托爾斯泰(Leo Tolstoy)說：「藝術活動就是人類將自己的情感、經驗，藉由文字、色彩、動作…等不同表現形式，來傳遞給其它共同經驗者，以引發共鳴」。因此，處在台南城內/城外的藝術家們，經由各人的感受與經驗，呈現了各人藝術創作風貌，同時也得以窺見台南城內與城外整體的藝術發展。如林鴻文的作品追求的是一種「土味」，這是他對於台南歸仁、關廟的大片竹林顏色感受而來；林書楷認為台南的聚落發展由一個點慢慢堆疊而成，反應到作品的城市意象；黃步青喜愛使用台南撿拾沒有人要的物品，作為他的拼貼創作元素……等。 **藝術家作品** 透過藝術家講述自身創作，可以使觀眾瞭解當代藝術的多樣性、每位藝術家在創作過程的心境以及藝術家居處生活環境和作品的脈絡…等。
事件	**許自貴說台南是文化首都，看古蹟、吃小吃，台南有什麼？** 以此事件來開頭，呈述本片對於「台南」這個地區的藝術歷史文化，因此這句話更像是：「除了看古蹟、吃小吃，台南還有什麼？」 **二號公寓在官方展覽發生火燒** 從此事件可以知道當時二號公寓在官方所展的藝術內容，對於還處在古典系統的台南藝術環境，是比較無法接受的。就此瞭解當時二號公寓、伊通公園的成立，要在當時的藝術環境下生存，是非常艱辛的。 **高雄成立阿普畫廊，許自貴要北上，黃宏德、顏頂生反對** **第一屆新風格展給藝術家八千元的材料費** **從古典系統到 1980 年代末出現南台灣新風格** 從林煌迪論述說道：「1980 年代末那時候，就是有一群人，他做出了有別於你本來很習慣，在台南看到的那個比較古典的系統的一些作品，那這些人就是那個南台灣新風格。」 而南台灣新風格的概念，跟林煌迪在台北唸書的資訊是比較相近的，而這種傾向抽象的繪畫運動延續十年，改變了整個社會樣貌。由此可知南台灣新風格對於台南藝術發展具有重要性。 **繼南台灣新風格後，一群藝術家組成「邊陲文化」** 邊陲文化可能就是台南最早被歸納為替代空間的空間。 **1990 年代的藝術市場曾經大好，開始了一些藝術團體** 在當時藝術市場大好，因此可能標榜不同的藝術形式，因此陸陸續續有一些團體形成，如：「邊陲文化」、「原型藝術」、「文賢油漆行」、台灣新藝」…。
結構	**起》**許自貴的一段話：「有一次我們幾年前聊天，他說什麼台南什麼文化首都，台南有什麼文化？台南只有古蹟跟小吃，美食啊，大家來台南玩就是看古蹟、吃小吃啊，台南有什麼？」。 「台南有什麼？」為起頭，開啟影片最主要核心。 **承》**主要以黃步青、許自貴、楊明迭、林鴻文幾位台南資深的當代藝術家們呈述，敘述當時南台灣新風格、邊陲文化、二號公寓在台南擴展的狀態，從北美館接受二號公寓展覽發生火災、許自貴成立阿普畫廊、第一次舉辦南台灣新風格…，作為影片呈述當代藝術的發展過程，並可看出每位藝術家在台南藝術的歷史洪流中，各自站在不同的位置上，並對於過去的事件有著不同的看法及發展的想像。最後由林煌迪以全觀來論述整體的藝術脈絡。 **轉》**在前段由幾位台南資深的當代藝術家作為呈述，此段主要以林煌迪、方偉文、李承亮、林書楷幾位中世代及新世代的藝術家，講述現今他們對於台南的藝術環境，以及從北部回到台南看到的狀態。如李承亮及林煌迪敘述的台南是較慢的；林書楷認為台南是聚落、堆疊的樣貌；方偉文認為台南是較少發展、資源缺乏的城市。 因臺南藝術大學設立後，台南豐厚的歷史文化所帶來的不同樣貌，也開啟了新的世代在台南發展，如：文賢油漆工程行、台灣新藝…等，並也開始出現各式各色的替代空間。可以從此段看出不同世代對於台南環境的不同感受及看法。 **合》**「你的意識怎麼被轉換出來，它還是有一個社會語言的那個部分，怎麼去轉換，怎麼去談，怎麼去…怎麼去繞，怎麼去跟這個世界產生關係」來綜觀整體的藝術

	環境，而林煌迪認為「當代」通常是出現在有一個流動性的城市，那台南的「當代」，就會出現在城外，會流動的地方。 最後並以許自貴自身多年觀察，暗示藝術生態若能獲得官方助力推動，相信能夠讓藝術家們推向新的高度作為本片為結尾：「連個玉井芒果他都可以推到日本推得那麼好，我們的藝術家還不如芒果。」不僅是暗諷也是說明台南的發展應該有更多的發展性。
時間	依循 2016 年度科技部《府城五條港 - 歷史遊戲、老街旅遊、藝文聚落─運用數位傳播技術於踏查台灣藝術離根化文化現象之實作評估》與台南《城內/城外》策展團隊共同耙梳之台南當代藝術重要事件年表
主體	敘事主體為分別代表老、中、青三個世代的台南十位藝術家。以下為十位藝術家介紹。 <u>資深世代</u> 許自貴：1956 年生於臺灣高雄，定居於臺南。 許自貴創作來自於生活及他輕鬆自得的創作諸脈絡，自然的動作形象，在經過轉化後構成屬於許自貴的神話家族，從他的想像世界延伸，輕鬆的賦予瑰麗色彩、活靈活現的表情。 林鴻文：1961 年出生於臺南仁德車路墘。 創作以平面複合媒材、鋼鐵焊接、環境裝置藝術為主。回溯法則和回歸的思維方式碰觸於現實當下的所有是他所執的方式也是他藝術表現形式中重要元素。 黃宏德：1956 生於臺南關廟，目前工作與生活於臺南。 黃宏德的藝術創作擁有極高的書鳥辨識度，早期的繪畫在空間的安排上，以大比例的放空呈現一種極為宏觀的視覺空間效果。。 黃步青：1948 年出生於臺灣鹿港，定居於臺南。 擅於使用水、木、草、鐵等這些原生材料或已去功能的現成物做為媒材，在畫、立體造型、多媒體藝術、環境藝術等多元創作形式上呈現。 楊明迭：1955 年生於臺灣臺南，定居於臺南。 楊明迭是臺灣重量級版畫藝術家，其人溫文儒雅，避居臺南民風醇樸的街巷中，默默耕耘他澄澈且雋永的版畫風格，在他清雅的藝術作品中看到相同的品質。 <u>中年世代</u> 李昆霖：1965 年生於台南市。2009 年，隱身山林與世長辭。 「孤獨」一直是李昆霖訴說的重要主題，「山賦」是他留給親人朋友們最後、最深刻的誓言。本計劃由其夫人口述。 林煌迪：1971 年生於臺灣臺南，目前工作與生活在臺南。 作品關注所有物件、訊息、符號在生產系統下的資源回收和利用，這樣的回收利用的核心在於如何從失效的聯結性中找回主體意義。 方偉文：1970 年出生於汶萊，1988 年歸籍臺灣。 善於採集、吸納，並為手邊的事物賦予新的生命，無論物件來自手繪，抑或來自工業機器，或者來自自然環境或生活空間，均能被他創造性的融入一個相互應合、對話的整體中。 <u>青年世代</u> 李承亮：1986 年出生於基隆，目前居住與工作於臺南。 創作的媒材多元，關心生活的當下而為生活狀況提供註解，也為生活的下一步找方向，是新生代藝術家中信奉手嘏作身體力行的實踐者。 林書楷：1983 生於臺南市，現居住並創作於臺南。 創作主要以繪畫、裝置為主。2013 年受邀到德國柏林 GlogauAIR 藝術村擔任駐村藝術家，2014 年獲選台灣製造臺北國際藝術博覽會新人推薦特區個展，2015 年受邀到廣島市立美術館參展俯微的世界聯展，2016 年受邀參展臺灣雙年展。

5. 結論與討論

第一年之研究計劃為《映。他聽》，經長達三個月的多方交替採訪、補拍，五個月高達 38 次的劇本與剪輯修正、潤飾後製，已成功以傳統第三方他人為受眾的單向式訊息方式，建立總長 53 分鐘之廣電規格傳播紀錄片，可成為次年《應。你談》及第三年《In。我獨白》等次世代傳播技術的研究基底。相關議題於不同媒材間之轉譯法則將在第二、三年研究討論後提出。

6. 參考文獻

Berger,A.A. 1997. *Narratives in Popular Culture, Media, and Everyday Life Thousand Oask*. CA:Sage.

Freud, S. 1949. *An outline of psychoanalysis*. New York: Norton.

Herman, L. M.. Imitation of self and others by dolphins. In K. Dautenhahn and C. L. Nehaniv(eds.) 1999. *Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts*. Brighton,U.K. Society for the Study of Artiflcial Inteligence and Simulation of Behaviour.

Labov,W. 1972. *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pensylvania Press.

Meyen,M& Löblich,M. 2006. *Klassiker der Kommunikationswissenschaft.Konstanz*: Universitätsverlag Konstanz-UVK.

Raymond,B &Chaumeton,E. 2002. *A Panorama of American Film Noir*. CA: City Light Books.

吳金桃 (2004)。〈台北雙年展應該包含更多的過程及非正式性〉，《典藏--今藝術》147 期，頁 83。

林日璇、李育豪、王茜穎(譯)(2014)。《媒體 ing：認識媒體、文化與科技》。台北市：新加玻商聖智學習。

政大傳院媒介寫作教學小組(2009)。《傳媒類型寫作》。台北市：五南。

姚媛 (2002)。《通俗文化、媒介和日常生活中的敘事》 (原作者：Berger)。南京：南京大學。

陳芸芸譯(2002)。《媒介、傳播與文化(原作者：James Lull) 》。台北：韋伯。

張廣智(1998)。《影視史學》。台北市：揚智文化。

程予誠(2008)。《電影敘事影像美學》。台北：五南。

謝東山(2002)。《台灣當代藝術》。台北：藝術家出版社，頁 6。

Paper 4

Title

不同文化下的動畫角色色彩之研究：以美國迪士尼與日本吉卜力動畫為例

The color research of animated role design in different cultures: Case study of animations in America and Japan.

不同文化下的動畫角色色彩之研究：以美國迪士尼與日本吉卜力動畫為例

徐芳真[1]、楊智傑[2]、陳姿汝[3]
[1]助理教授 南臺科技大學 多媒體與電腦娛樂科學系
[2]副教授 南臺科技大學 多媒體與電腦娛樂科學系
[3]助理教授 南臺科技大學 視覺傳達設計系

摘要

風格通常會隨著時代與地緣的關係而有所不同(Chen & Owen，1997)。角色的風格跟顏色、形體、線條、飾品等息息相關。故事中的角色除了演繹故事情節，更具有該故事風格的代表性。而這些細節都被傳統文化影響著。由於臺灣對於美國和日本的動畫接受度很高，本研究為了瞭解不同文化影響下的角色色彩差異，將以知名的動畫公司迪士尼公司和吉卜力公司在1984年至2013年所推出的動畫作品中的角色為研究對象，探討這兩間動畫公司在角色色彩上的使用差異。本研究以曼賽爾色彩系統紀錄動畫角色的色彩取樣樣本，色彩樣本採取自迪士尼的有704個，吉卜力的有434個。因此在本研究中將色彩樣本依據角色是否為人類、作品出版時間因素分類之後，發現所有的角色類型都經常使用色相“R”、“YR”、“Y”。在“GY”、“G”、“BG”、“PB”、“P”、“RP”等區域，迪士尼和吉卜力兩間公司都很少使用到高彩度的區。其中色相“G”和“BG”較少被迪士尼和吉卜力使用。兩間公司在使用的色彩區域在中彩度、中明度區域呈現重疊的情形。但是迪士尼也會使用高彩度、低明度區域的色彩；吉普力則會再使用低彩度、高明度區域的色彩。從這些結果顯示，不同的文化的確會造成不同國家的動畫角色色彩差異，在本研究中顯示：美國文化偏好鮮豔、飽滿的色彩；日本文化偏好清淡、明亮的色彩。

目次

1. 研究背景
2. 相關文獻
2.1 文化影響下的色彩風格

2.2 日本和美國動畫的差異
2.3 視覺上的色彩刺激
3. 角色設計之色彩樣本取樣與分析
3.1 動畫樣本取樣範圍
3.2 曼賽爾色票
4. 色彩分析結果
4.1 來自動畫角色的色彩樣本
4.2 色彩樣本的分析
5. 討論與結論
6. 參考文獻

關鍵詞

文化，動畫，角色設計，色彩範圍

1. 研究背景

顏色在生活中扮演重要的角色，它不僅僅在美感上影響著人們，更扮演著文化傳承的角色。例如紅色在中國則代表喜慶、長壽。在印度文化中紅色代表著權力、財富、恐懼、純潔、愛、誘惑等等。在南非，紅色則與獨力鬥爭下的犧牲、哀悼有關。在泰國，紅色則與宗教的太陽神有關。顏色在文化中所代表的意義，影響了人們在日常生活中使用顏色的習慣。相同的顏色在不同的文化具有不同的意義，顏色的名稱在很大程度上取決於文化。在作者先前的研究中(Hsu & Hsiang，2018)透過 ANOVA 已經確認不同國家的動畫公司會影響動畫角色的用色。因此可以確認不同國家受到文化影響會有不同的顏色喜好，顏色所表示的意義、情緒不應跨越文化統一闡述。

傳統色彩影響人們在生活中處處可見，最明顯的例子莫過於街道市景。Manav 提出了建築物的顏色確實影響建築環境、城市環境(Manav，2017)。日本在 2004 年公佈並且實施景觀法(Ministry of Justice，2011)，對於都市景觀建築物，包含建築物外觀色彩，有著詳細的規定。有鑒於此，各國所謂的傳統色彩在生活中處處影響著人們的審美觀。臺灣從 1661 年至今，經歷多個政權交替。造成臺灣的文化歷史具有多元文化的特色。這在無形中使的人民對文化認同產生混淆。為了讓台灣人民能認識在地文化、認同在地文化，近年來在政府與民間的推動之下，人民開始思考臺灣在地文化的認同議題。由於美國文化與日本文化對於臺灣的影響深遠，而且動畫是基於傳統與美學產生的綜合成果，因此本研究欲從美國迪士尼公司公司與日本吉卜力公司公司的動畫電影中研究其色彩。將來設計師設計角色時，可以參考日本風格或者美國風格的色彩範圍，依據臺灣的多元文化特色，逐漸發展出適合台灣本身的色彩範圍。

2. 相關文獻

2.1 文化影響下的色彩風格

傳統顏色自古以來就一直使用。通常，獨特的顏色在其使用的每個文化中都具有某種起源或含義，並依此為顏色命名。例如：基於日本國花「櫻花」而命名的櫻花粉色。具體顏色名字可以反映強烈的社會色彩敏感度。在一項顏色名稱的研究中顯示，顏色名稱跟文化具有相關性，而且顏色名稱對於觀看者的情緒具有顯注的影響力(Sutton & Altarriba，2016)。因此不同的國家會有其偏好的使用色彩和不同意義的色彩名稱。為了分析出作品的風格，必須具有至少 4 個特徵重複出現出現在至少 3 個不同的作品中(Chen & Owen，1997)。由於風格具有相互模仿與同化的特性，因此一個單一風格需要數個特徵來建構出其獨特性。在一系列

具有同質性的作品中，形式特質，包含形狀、顏色、排列、大小、質感等，會比作品意義更快被指認出來(Chan，1994)。因此每個分析對象需要具備 8-10 件作品，以提供風格特徵交叉分析所需(Chen & Owen，1997)。

2.2 日本和美國動畫的差異

日本的和色系是日本古代以來一直使用的系統，顏色偏向低彩度、低明度，並不是單純紅色綠色等顏色的使用(張凱翔，2015)，例如：日本的「神隱少女」裡的色彩。此外，日本文化也受到外來文化的影響，並且也呈現在色彩配色和動畫角色創作上。例如：京都寺院裡的五色圍幔，「朱赤」與「萌黃」的配色組合就是受到印度佛教的影響。日本動畫角色「原子小金剛」受到美國迪士尼公司的「小木偶」影響，使單純的木偶人成為具有日本民族風格的機器人。

美國動畫依循好萊塢電影的手法，注重細節的描述且表現較為真實，構圖與用色皆注重於營造劇情內容的氣氛。動畫比電影更具誇張與趣味性，使動畫角色設定具有特色，容易讓觀眾有更深刻的印象。例如，以誇張的身材比例表現人物特質。在角色的色彩上盡量單純、顯色，因此角色的顏色以接近色塊的方式凸顯角色本身(張凱翔，2015)，或使用美國國旗的顏色鮮豔飽滿的紅色和藍色，例如，美國的「超人特攻隊」裡的角色色彩。

2.3 視覺上的色彩刺激

色彩對於情緒有非常大的影響，色彩空間中的顏色可以影響人們的情緒，人們對於喜歡的色彩會感受到正面情緒，而對於不喜歡的顏色則會引發負面情緒(Gong, Wang, Hai, & Shao，2017)。因此，對於讓大部分兒童感到憂慮的牙科，以兒童喜愛的藍色和粉紅色進行環境佈置，可以有效的增加兒童對於牙科的積極態度(Annamary et al.，2016)。不僅心理上受到色彩影響，人體在生理上也受到顏色的影響。在一項熱湯的研究中顯示，湯的顏色會影響食用者的飽足感和身體溫暖度。也就是說顏色的飽和度和亮度證明了對情緒的強烈和一致的影響(Valdez & Mehrabian，1994)。

3. 角色設計之色彩樣本取樣與分析

3.1 動畫樣本取樣範圍

美國與日本的動畫公司因動畫產業發展較早，同一製作公司推出的作品較多。由於分析每一種風格需要至少有 8-10 件作品進行風格分析(Chen & Owen，1997)，本研究以美國與日本動畫數量最多的公司作為研究對象，也就是迪士尼公司和吉卜力公司。吉卜力公司的宮崎駿導演自 1997 年出版第一部動畫作品之後陸續有

作品產出，最後一部已發表的作品是 2013 年的「風起」。考慮到作品的年代對於色彩具有影響力(Hsu & Hsiang， 2018)，為了擷取兩間公司在相同的出版時期內的作品，本研究屏除單一公司更早或更晚的作品，將對兩間動畫公司動畫樣本取樣時間範圍設定在 1997-2013 年之間。除了色彩之外，角色的服裝款式、配件種類等等均不在本次研究範圍之內。

3.2 曼賽爾色票

由於曼賽爾色彩系統具有能直觀的依據人的感官對於色彩濃淡、明暗描述色彩的特色，並使用簡單的記號與數字來紀錄顏色，取代較為曖昧不明的各種色彩名稱，因此被使用在許多色彩課程或研究中。例如在一項色彩情緒與偏好的研究中以曼賽爾色彩系統來記錄顏色(Gong et al.，2017)。本研究採用的電子色票來自於“google play”平台上的“Munsell color chart”(KSGc，2016)。曼賽爾色彩系統將顏色以符號“H”(色相)、“V”(明度)、“C”(彩度)記錄色彩空間。在本研究中“H”共有 10 種，分別為“R”、“YR”、“Y”、“GY”、“G”、“BG”、“B”、“PB”、“P”、“RP”。每一種色相還可細分，並以數字表示細分的情況。本研究採用四等分的色相分級。用來表示色彩的明暗程度的明度“V”則分為 0~10；“V”的數字越大表示色彩越明亮。本研究統計時只有使用到 1~9，因為 V=0 為黑色、V=10 為白色。用來表示色彩的鮮豔程度的彩度分為 0~20 個等級。其中 C=0 為無色彩；本研究統計時只有使用到 1~20。“C”的數字越大表示色彩越鮮豔，在整個色彩空間中離軸心越遠。

4. 色彩分析結果

4.1 來自動畫角色的色彩樣本

角色樣本包含動畫中的主要角色和主要配角，對其頭髮、皮膚、服裝、鞋子、飾品等較大的部位進行色彩取樣。色彩取色工具為“Adobe PhotoShop”的取色功能工具。每個角色可以取出 5-10 個色彩樣本。例如“The Little Mermaid”裡的角色“King Triton”可以取出 5 個色彩樣本(圖 1)。色彩樣本採取自迪士尼公司的有 704 個，吉卜力公司有 434 個。其中，迪士尼公司的人類樣本 258 個，非人類樣本 446 個。吉卜力公司的人類樣本 366 個，非人類樣本 68 個。



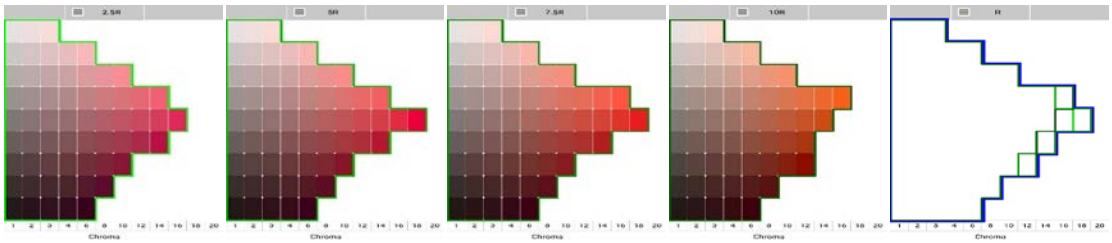
(DVD 資料畫面)

King Triton	H	V	C
皮膚	2.5YR	6	4
鬍鬚	2.5PB	9	2
鱗	2.5BG	8	4
尾巴	2.5PB	5	10
手環飾品	10Y	8	10

圖 1 動畫“The Little Mermaid”中的角色“King Triton”取樣色彩樣本範例

4.2 色彩樣本的分析

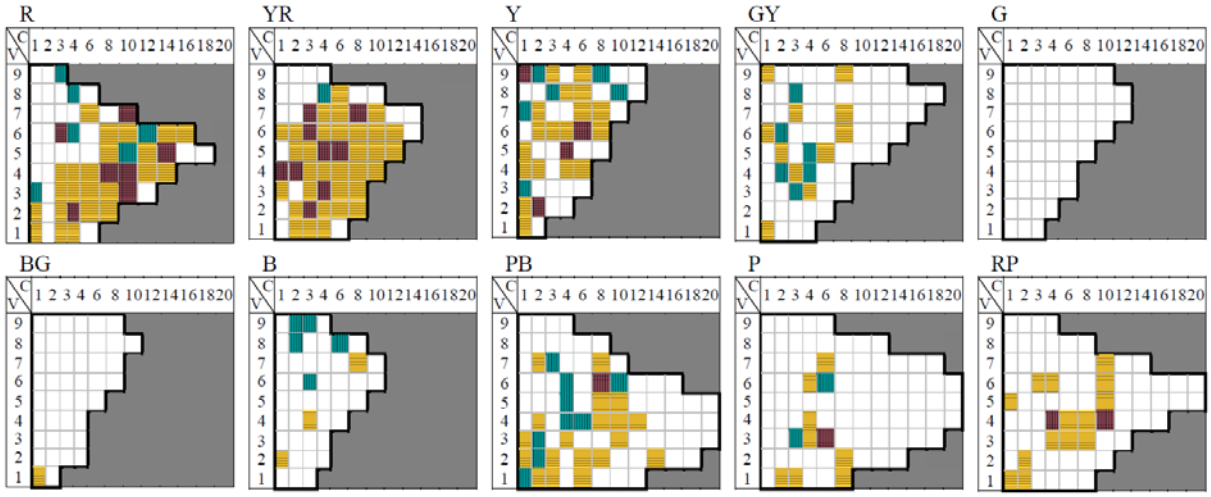
在作者先前的研究中以 ANOVA 分析影響色彩的因素，發現動畫公司別、色彩面積、角色是否為人類、作品出版時間等四因素對角色色彩具有影響性(Hsu & Hsiang，2018)。因此將所有樣本分成三類(圖 3~圖 5)分別進行分析與探討。從動畫中截取角色色彩時，以 10 個色相的 4 個等分色相進行記錄。在統計全部色彩分佈的時候，將同色相 4 個等分色相，依照所佔的明度、彩度位置(如圖 2 (a)-(d)綠色框線)，以聯集方式合併成一個色相範圍(如圖 2 (e)藍色框線)進行色彩分佈說明。



(a) (b) (c) (d) (e)

圖 2 同一色相的 4 個等分色相色票範圍聯集成一個色相範圍

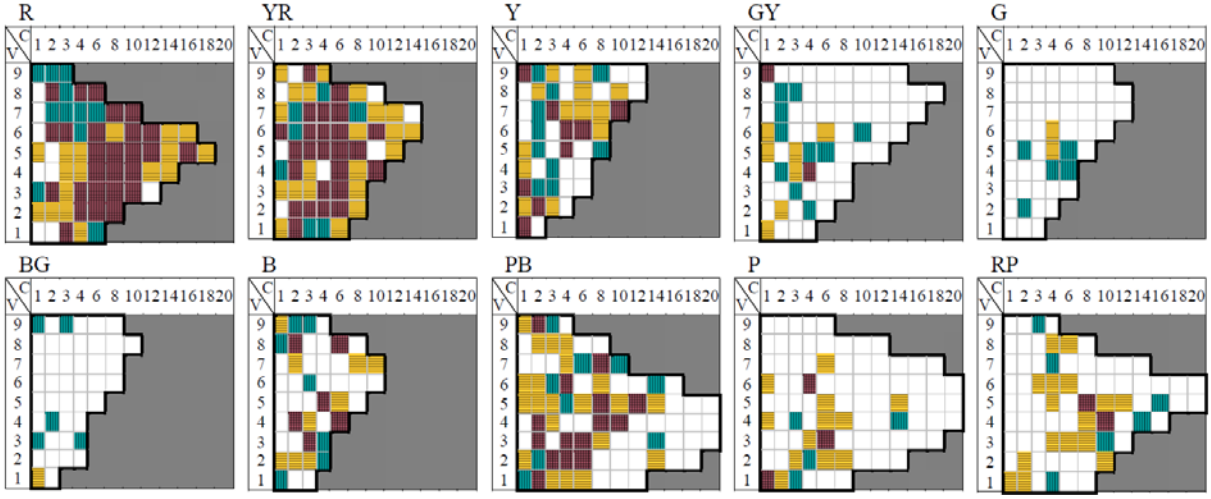
非人類角色的色彩樣本分佈如圖 3。(1)色相“R”、“YR”：被使用的區域很廣。但是低彩度、高明度的區域較少被使用。吉卜力公司比較不使用低明度的“R”和“YR”區域。吉卜力公司不使用高彩度的“YR”區域，但是迪士尼公司會使用到比較多高彩度的“YR”區域。(2)色相“Y”：兩間公司沒有明顯區別。(3)色相“GY”：兩間公司都沒有使用到高彩度的區域。(4)色相“G”和“BG”：兩間公司幾乎不使用“G”和“BG”。(5)色相“B”：吉卜力公司明顯使用高明度、低彩度的區域；迪士尼公司則使用低明度、或高彩度的區域。(6)色相“PB”、“P”、“RP”：兩間公司都沒有使用到高彩度、或高明度的區域。



註：■吉卜力公司使用的色彩。■迪士尼公司使用的色彩。■吉卜力和迪士尼公司共同使用的色彩。

圖 3 樣本角色為非人類的吉卜力公司與迪士尼公司所使用色彩色域

1999 年 12 月以前發表的人類角色色彩樣本分佈如圖 4。(1)色相“R”、“YR”和“Y”：兩間公司使用的顏色範圍幾乎涵蓋整個範圍。(2)色相“GY”、“PB”和“P”：兩間公司都沒有使用高彩度的區域。高彩度、高明度的“GY”和“P”也較少被使用(3)色相“G”和“BG”：這兩個顏色被使用的較少，但吉卜力公司使用的區域稍為多一點，而且都使用低彩度的區域。(4)色相“B”：兩間公司沒有明顯區別。(5)色相“RP”：迪士尼公司使用的色彩偏向中、低彩度。吉卜力公司少見的使用了比迪士尼公司還要高一些的彩度區域。

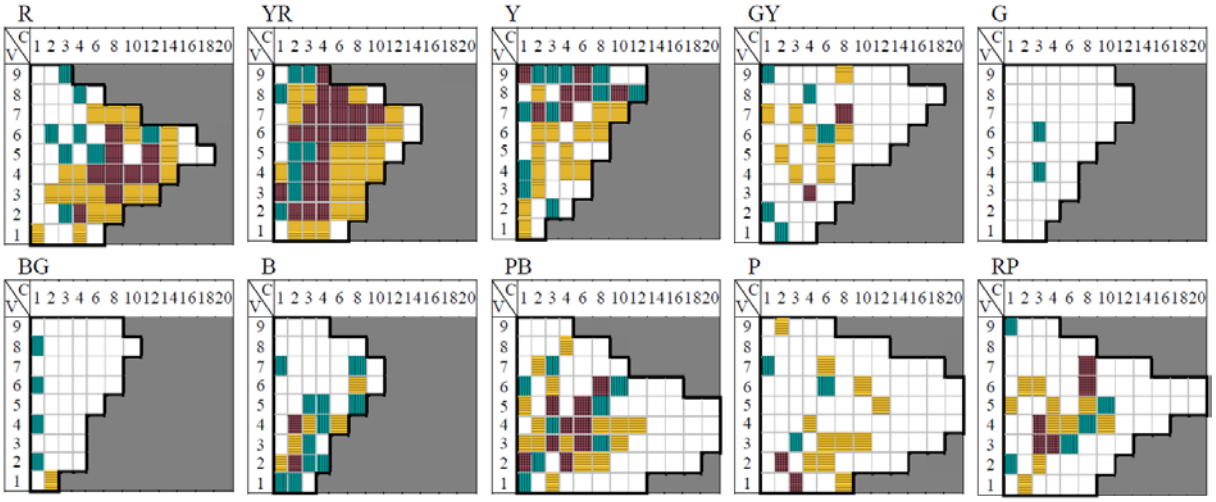


註：■吉卜力公司使用的色彩。■迪士尼公司使用的色彩。■吉卜力和迪士尼公司共同使用的色彩。

圖 4 樣本角色為人類且製作年份在 1999 年 12 月以前吉卜力公司與迪士尼公司所使用色彩色域

2000 年 1 月以前發表的人類角色彩樣本分佈如圖 5。(1)色相“R”：除了中明度、中彩度的重疊區域，吉卜力公司還會使用到低彩度、高明度的區域；迪士

尼公司則偏向高彩度、低明度的區域。(2)色相“YR”：迪士尼公司明顯的使用到較高彩度的區域。(3)色相“Y”：迪士尼公司用到的範圍廣泛，但吉卜力公司沒有使用中明度、中彩度的區域。(4)色相“GY”：兩間公司都沒有使用高彩度的區域。(5)色相“G”、“BG”：這兩個顏色被使用的較少，但吉卜力公司使用的區域稍為多一點，而且都使用低彩度的區域。(6)色相 B”：低彩度、高明度的區域沒有被使用。(7)色相“PB”、“P”、“RP”：高彩度的區域沒有被使用。



註：■吉卜力公司使用的色彩。■迪士尼公司使用的色彩。■吉卜力和迪士尼公司共同使用的色彩。

圖 5 樣本角色為人類且製作年份在 2000 年 1 月以後吉卜力公司與迪士尼公司所使用色彩色域

5. 討論與結論

將色彩樣本依據角色是否為人類、作品出版時間因素分類之後，本研究發現在迪士尼公司和吉卜力公司所有角色裡面，“R”、“YR”、“Y”色域裡面使用的色彩區域最多，幾乎涵蓋全部範圍。在“GY”、“G”、“BG”、“PB”、“P”、“RP”等區域，兩間公司都很少使用到高彩度的區域。“G”和“BG”色域裡面的色彩被使用的區域最少，其中，吉卜力公司在人類角色使用的“G”和“BG”色彩區域比迪士尼公司多。在同一個色相的區域中，除了中彩度、中明度區域比較容易出現兩間公司共同使用之外，迪士尼公司頃向于較會使用到彩度高或者明度低的區域，吉卜力公司則相對使用到低彩度或高明度的區域。

綜合所有的角色類型，“R”、“YR”、“Y”色域都經常被使用；但是迪士尼公司會運用到高彩度區域，這跟我們印象中認為美式色彩鮮豔的觀點一致。在包含藍色和綠色的冷色調中，除了“B”之外，這些冷色調的高彩度被迪士尼公司和吉卜力公司同時的避開使用。而“G”和“BG”的中~低彩度，比較受到日本的青睞。光線的物理特性，色彩的範圍是固定的。但人們受到傳統文化的影響，顏色對不同的人來說意味著不同的事。例如在定義傳統色彩時，對於全部的色彩

都會有相對應的色彩與名稱(Hamada，2013)，但是在實際生活中的運用卻會依照當地的文化而經常使用某些色彩，例如日本奈良縣和山梨市的都市景觀建築物色彩規範(N. C. Government，2009；Y. C. Government，2016)，規定建築物色彩的使用範圍偏向彩度較低的色彩。而“G”和“BG”的中~低彩度色彩正是經常被使用在生活中常建的建築物上。

在近 20 年來臺灣開始著重美學教育。有鑒於台灣過去多個的政權交替，臺灣人民對於所謂的傳統已經不能用單一的漢文化概括，必須要面對每個政權在本地所留下的印記。目前本研究受到科技部計劃的支持，正在進行在地人文故事的媒體設計與製作，希望在角色設計上導入本土的因素，包含角色的色彩、角色的設計，以奠定計劃未來進行本土人文故事的創作基礎。目前本研究的研究成果可以清楚的看到全球知名的動畫公司迪士尼公司與吉卜力公司的用色範圍。未來將基於這個色彩成果，進行本土人文故事的角色設計。

6. 參考文獻

Annamary, Kattakayam, Prathima, Gajula Shivashankarappa, Sajeev, Renganathan, Kayalvizhi, Gurusamy, Ramesh, Venkatesan, & Ezhumalai, Govindasamy. 2016. “Colour Preference to Emotions in Relation to the Anxiety Level among School Children in Puducherry – A Cross-Sectional Study.” *Journal of Clinical and Diagnostic Research : JCDR*, 10(7), pp. 26-30.

Chan, C. S. 1994. “Operational definitions of style.” *Environment and Planning B: Planning and Design*, 21(2), pp. 223-246.

Chen, Kuohsiang, & Owen, Charles L. 1997. “Form language and style description.” *Design Studies*, 18(3), pp. 249-274.

Gong, Rui, Wang, Qing, Hai, Yan, & Shao, Xiaopeng. 2017. “Investigation on factors to influence color emotion and color preference responses.” *Optik - International Journal for Light and Electron Optics*, 136(Supplement C), pp. 71-78.

Government, Nara County. 2009. Nara Prefecture Landscape Regulations and Nara Prefecture Landscape Planning. 2018, from http://www.pref.nara.jp/12775.htm

Government, Yamanashi City. 2016. Yamanashi City Landscape Plan. 2018, from http://www.city.yamanashi.yamanashi.jp/citizen/docs/keikankeikaku.html

Hamada, Nobyoshi. 2013. The traditional colors of Japan: PIE International.

Hsu, Fang-Chen, & Hsiang, Tai-Wei. 2018. “Factors affecting color discrepancies of animated film characters.” *Journal of Interdisciplinary Mathematics*, 21(2), pp. 279-286.

KSGc. 2016. Munsell color chart. 2016, from https://play.google.com/store/apps/details?id=jp.co.kozo.munsellcolorchart

Manav, Banu. 2017. “Color-emotion associations, designing color schemes for urban environment-architectural settings.” *Color Research & Application*, 42(5), pp. 631-640.

Ministry of Justice, Japan. 2011. Japanese Law Translation. from http://www.japaneselawtranslation.go.jp/law/detail/?id=2533&vm=&re=

Sutton, Tina M., & Altarriba, Jeanette. 2016. “Finding the positive in all of the negative: Facilitation for color-related emotion words in a negative priming paradigm.” *Acta* Psychologica, 170, pp. 84-93.

Valdez, Patricia, & Mehrabian, Albert. 1994. “Effects of Color on Emotions.” Journal of Experimental Psychology: General, 123(4), pp. 394-409.

張凱翔 (2015)。〈動畫色彩計畫研究—美、日成功商業動畫長片色彩計畫比較分析〉。《碩，國立臺灣藝術大學，多媒體動畫藝術學系新媒體藝術碩士班》，臺北。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院 Dharma Drum Institute of Liberal Arts (DILA)

Funerary Practices of Taiwanese in the Ryukyu Islands, Muslims in Taiwan and Hong Kong, and Chinese in Southeast Asia

Variations and Invariances of Tombs Epigraphic, Religious and Artistic Craftsmanship

墓葬風俗之流變

琉球群島台灣人、台港穆斯林與東南亞華人於墓葬、碑文銘刻、宗教及民間工藝等風俗的實踐與比對

David Blundell* Oliver Streiter Hanna Yaqing Zhan** Tammy Yiting Liu** Sara Yuting Wang** Mandy Manwai To** Syuanfei Shih** HuiJi Wang*****

National Chengchi University* National University of Kaohsiung Monsoon Asia Cultural Consultancy*****

Funerary Practices of Taiwanese in the Ryukyu Islands, Muslims in Taiwan and Hong Kong, and Chinese in Southeast Asia: Variations and Invariances of Tombs Epigraphic, Religious and Artistic Craftsmanship

墓葬風俗之流變：琉球群島台灣人、台港穆斯林與東南亞華人於墓葬、碑文銘刻、宗教及民間工藝等風俗的實踐與比對

David Blundell
National Chengchi University
Oliver Streiter
Hanna Yaqing Zhan
Tammy Yiting Liu
Sara Yuting Wang
Mandy Manwai To
Syuanfei Shih
National University of Kaohsiung
HuiJi Wang
Monsoon Asia Cultural Consultancy

The 9th International Conference of Digital Archives and Digital Humanities
DADH **2018** 第九屆數位典藏與數位人文國際研討會
Interdisciplinary Research and Humanities Research in Literature Studies, Linguistics, Culture, History, etc., Conducted with Digital Data and Technology
Dharma Drum Institute of Liberal Arts
19th–21st December 2018

Research Domain and Fundamental Research Questions

Our paper is based on variations and invariances of tombs epigraphic, religious and artistic craftsmanship in East and Southeast Asia. Our fieldwork originates in Taiwan as a point of departure from documenting funerary practices of local people in their island environment (**華人於墓葬**) for comparing and analyzing with the people living in other contexts, such in Taiwan, the Ryukyu Islands, or Hong Kong. With this research foundation, we are working collaboratively to bring together data from disparate disciplines and integrate the history of the influence of maritime activity on the development of the major religions and cultures across the region Monsoon Asia.

Archaeological, visual, scientific and textual evidence will be contextualized by time and place providing the opportunity to study a large amount of new and vital material in an integrated manner. This inspires researchers to ask new questions, develop new analysis using readily available new technologies, and develop nuanced interpretations of our past. In addition, the materials will be incorporated into visualization technologies to allow development of new representations and narratives enabling a wide range of users to interact with and re-envision their history.

We are building a geospatial database that can store spatiotemporal information

collected by each individual sub-project.Our software tools to the other sub-projects for converting field data to digital maps with spatiotemporal coordinates such that they can overlay with ancillary map data.Thus we establish a Website to facilitate data sharing among researchers of the integrated project, and publishing of research results on the Internet.

Our research analyzes variations and invariances of tombs epigraphic, religious and artistic craftsmanship in East and Southeast Asia. Observing variations within a cultural domain and yet trying to identify invariances is a regular approach to the study of linguistic, religious, ethnic or cultural communities.



(a.) Iheya, Okinawa



(b.) Xindian, Taiwan

Figure 1. Ink inscriptions in the lid of a bone urn. Different, similar or the same? Which factors can account for similarities, which for differences?

Frequently, practices within a community appear to be invariant when focusing on a specific place or a specific time. However, when analyzing the development of a community through time, or when looking at the same community in different places, the image of an invariant practice gets blurred as more and more variation in practices can be observed. In *Death Ritual in Late Imperial and Modern China*, edited by J. L. Watson and E. S. Rawski, different authors compare death rituals in North and South China, as well as elaborated versus reduced death rituals, of rich and poor people respectively and necessarily observe practices that are not congruent.

Sometimes, the idea of invariance over the cultural domain can be saved, assuming that the observed variances are only a superficial features of a more abstract invariance. One might for example argue, that the reduced practice of poor community members is closer to the abstract, invariant hidden layer of culture (See Figure 2).



(a.) Unten, Okinawa



(b.) Unten, Okinawa



(c.) Yoron, Kagoshima

FIgure 2. Variation of a common theme, representing various degrees of elaboration? What are the differences and how can they be explained?

Yet as the abstract hidden invariant layer can never been exhumed or proven to exist, we might assume that beside the great multitude of surface variances, there are a certain number of hidden layers, which however are not carved in stone, but dialectically interact with the surface forms. After all, these abstract levels are frequently only visible by the trained researcher, communities and its members are frequently left alone when it comes to make sense of their practices. A surface variation is thus not only derived from a hypothesized variant, the variant is also influence, through induction from the surface forms as nothing else but the surface forms are available to community members.

Confronting the emic perspective, that is how communities define and perceive themselves in their search for a social and cultural identity, usually by induction starting from the surface forms, and the etic perspective, that is how researchers perceive communities as being related or unrelated in their cultural practices, usually by comparing communities, is a timely and relevant labor in a period where nationalistic, religious, linguistic or ethnic conflicts, which originate from the elevation of random surface features of cultural practices to sacrosanct features that constitute a social identity, cause thousands of deaths and millions of people to be uprooted, escaping from their homelands in search for a bearable life, no matter which and how many of their social practices they have to give up. This process of liberating oneself from identity defining practices represents the dark, unstudied side of what humanity also means, to give up one's culture in order to find food, shelter, physical integrity and social harmony.

This study will thus look particularly into what we call inverted worlds, worlds in which communities live either in majority or minority positions, as oppressor or oppressed in order to understand which are the cultural practices that are traded for a better life when living in minority position or a the position of the oppressed.

Fieldwork and Data Elaboration

Our fieldwork originates in Taiwan as a point of departure from documenting funerary and epigraphic practices of local people in their island environment for comparing

and analyzing with the people living in other contexts, such as the Ryukyu Islands, Hong Kong or Southeast Asia. Individualized or even psychological accounts on how people manipulate and transform their culture, we called it ‘trading’, can be obtained, e.g. by comparing Taiwanese as a majority in Taiwan to the Taiwanese as a minority in Okinawa, Muslims as a majority in Malaysia to Muslims as a minority in Taiwan or Hong Kong, or so-called ‘Mainlanders’ immigrating from China in northern Taiwan, where they constitute the cultural and economic elite, to those considered ‘Mainlanders’ in southern and eastern Taiwan, where they live comparatively isolated, integrated and assimilated, see Goudin et al. (2011).

To shed light on this and similar questions, Oliver Streiter and Yoann Goudin have created in 2007 the ThakBong Archive of tombs and cemeteries in Taiwan (online at http://thakbong.dyndns.tv). Originally describing cemetery tombs and tombstones of Taiwan, the data set has been enlarged to cover a vast range of spiritual sites across regions of Monsoon Asia (see Figure 3).

With the Asia-Pacific SpatioTemporal Institute (ApSTi, http://apsti.nccu.edu.tw) established in 2014 as Top University Project in Digital Humanities at National Chengchi University in Taipei (Blundell and Jan 2016), and the ThakBong archive merging through collaborative work with data from disparate disciplines, we are able to integrate the developmental history of the major religions and cultures across the region.

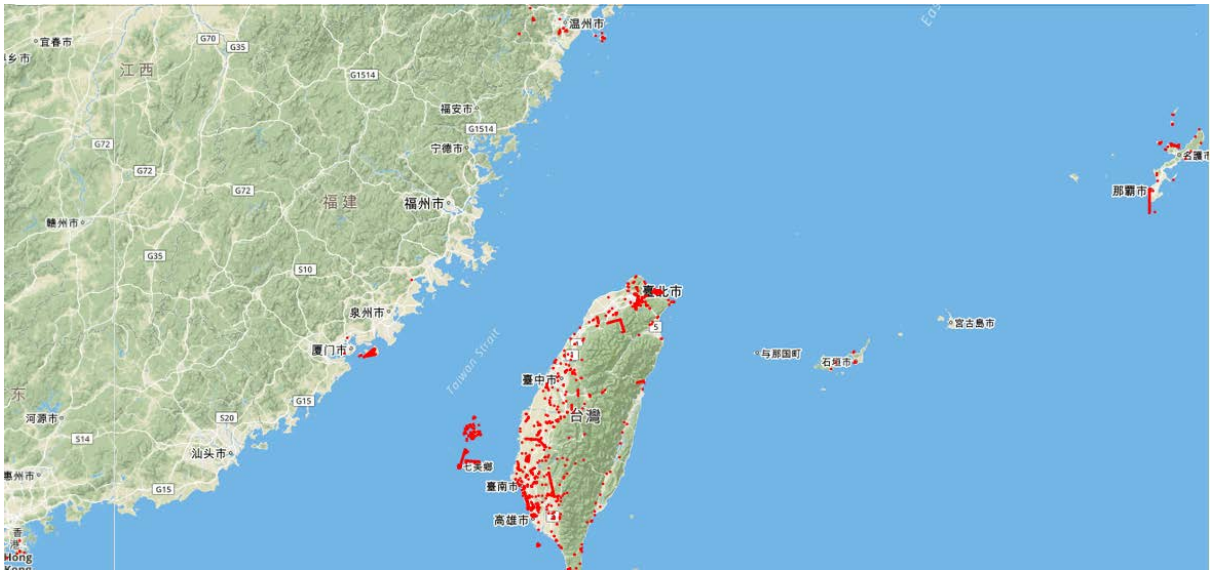


Figure 3. About 1000 burial and spiritual sites between the Pearl River Delta, the Dongtou Islands and Amami have been documented and made accessible in the ThakBong database.

In our extensive fieldwork, archaeological, visual, scientific and textual evidence is searched and digitized through photos, videos, GPS-trackers, audio-recordings etc.

In our laboratory, the digital resources are classified, annotated, transcribed, translated and contextualized by temporal and geo-referential attributes. Time, place, language, material, color, symbols, motifs, place names, family names, etc are only few of the attributes we collect in our database. This creates the opportunity to study a large amount of new and vital material in an integrated manner. Researchers can ask new questions and develop new analysis using readily available up-to-date technologies, and develop nuanced interpretations of our past.

To achieve this integration, we are building a geospatial ARCHES database, that can store spatiotemporal information collected by individual sub-projects, merged into one data model, (see Jan 2016). Our data, currently stored in an outdated idiocratic data model, are gradually re-coded in the CIDOC standard[1]. In addition to the elementary data entities such as, a set of relevant symbols, languages, scripts, motives and colors, more complex linked resources are created within CIDOC: A gazetteer with place names in various languages and scripts, plus their geo-references for various times, a model on artifacts, e.g. texts, tombs, boats, prints, etc. a model on people of the past, their interrelations, their relations to the artifacts, and of all entities, their relations to places and times.

One of the richest resource we find on our field trips, which helps us in establishing the relation between objects, people, space and time are inscriptions on objects, usually referred to as *epigraphy*. A second source for the comparison between communities and cultures derives from aesthetics, as aesthetic features are found in each gesture, each character stroke, each choice of colors and each roof tile.

Epigraphy

Epigraphy, like poetry, refers to a genre of writing, of which each individual writing is referred to as an *epigraph*. Epigraphs are inscriptions carved into an object composed of hard material, such as stone, wood or metal. Objects can be natural, such as rocks or trees, or man-made, such as stone steles or wooden tablets. Epigraphs are one of the main components of our research as they link the awareness of people in the past to real-world objects, representing a rich and valuable cultural heritage, which can be compared over the entire region of East and Southeast Asia.

[1]The CIDOC model is an object-oriented ontology in the domain of cultural heritage and can be implemented in any database system. Its most commonly used representation standard is RDF (Resource Description Framework). In 2006, the International Organization for Standardization (ISO) adopted CIDOC as standard ISO 21127:2014, called "*Information and Documentation: A Reference Ontology for the Interchange of Cultural Heritage Information*".

Similar to poems, epigraphs can be quite stylized and standardized. What seems to be a disadvantage, as we observe less variation compared to freely written texts, ultimately turns out to be a pandora four out structured approach to cultures. Different styles and standards are not randomly distributed and thus become attributes we collect on their own (e.g., one can easily observe how character variants, semantic variants, genre, style and content vary with the community the material and the object). This creates a huge matrix of cultural attributes we can collect. One question is how communities process harder material. Soft materials facilitate the carving of characters or character variants composed of more strokes, while harder or brittle materials might show simplified character variants, e.g. '□' vs. '顯' (xiǎn, an honorific) or semantic variants, e.g. '旦' (dàn) vs. '穀' (gǔ), both in the sense of 'auspicious'.

Also epigraphs refer in generic terms to the object on which they are written such as '墓' (*mù*, tomb) or '神位' (*shénwe*ì, ancestral tablet), we obtain, without much effort, a folk-taxonomy of culturally highly important artifacts and it goes without saying that the folk-taxonomy might vary from time to time and place to place in the same way as the objects themselves. We for example can distinguish communities which indifferently refer to their tombs as '墓' (*mù*, tomb), while others systematically distinguish between '墓' and '壽域' (*shòu yù*, area of longevity), see Goudin and Streiter (2016) or between, as common on Taiwan '墓' , '塋' (yíng, grave) and '佳城' (*jiā chéng, splendid city).* Thus while in Japan, the tomb is usually referred to as '墓' we might wonder how Taiwanese in Okinawa conceptualize and classify their tombs.

Epigraphs, in addition, epitomize conventionalized forms of interaction that are central to communities such as families, clan associations or temple communities. These interactions can be monetary transactions, the expression of devotion or filial piety, or the re-arrangement of power constellations. Although people and their communities might long have been gone, their interactions can still be reconstructed through epigraphs.

In more general terms, the value of an epigraph derives from its unique location at the intersection of humanities, the science of man-made products, and social sciences, the science of human behavior. An epigraph combines text, material and object, and has been created by specific and frequently known individuals or groups at a specific time, at a specific place, in relation to a specific event, and fulfilling a specific function. Being usually relatively stable, e.g. a wooden tablet attached to the roof of a temple, and inscribed by the year of its creation, space and time, two essential coordinates, are often known. As a consequence, additional data can be attached relatively easily to the social interaction documented in the epigraph.

Each epigraph is thus a node in a network of data that can be linked to historical, sociological, anthropological, geographic, linguistic, or economic data. Each

epigraph has through its unique combination of people, places, times, material and the content of the inscription, the potential to change our perception of the past. They are traces of unique human thinking and behavior. And in some cases an epigraph may convey the only facts we know about a specific person, time or place. Once systematized in an online available database, epigraphs will form a cornerstone of future cultural studies.

Aesthetics

Derived from the Greek *aisthetikos*, aesthetics (æsthetics) is a methodology to deal with the sense perception value of experiences. It identifies traits within cultural workings to make understandings and judgments. Aesthetic experiences form intrinsic attitudes vis-à-vis crafted objects, such as carvings, offerings, or happenings, such as rituals, dance and singing, specifically or generally recognized as worthy of attention. Reflection on the experience in terms of its meaning creates an overall set of tastes or expressive qualities that are harnessed to tell what "tastes good" in the context of a worldview (see Blundell 1996).[3] Aesthetic experience, although more difficult to identify and justify, can, similar to geo-references or character variants, be transformed into attributes which show similarities or potential conflicts between cultures.

Clifford Geertz, for example, introduced 'art as a cultural system' as a way to illustrate guiding factors in defining ethnicity, collective human identity, and the process of community life as traced in the merits of a Moroccan poet reciting the *Quran* (Greetz 1976). Here Geertz writes on the theory of art for understanding a culture as a semiotics [general system of symbolic forms] of art tracing the life of signs in society (*ibid*.:1488). Aesthetics is multi-dimensional for the understanding of cultural expressions in society prevailing for its self-worth, worldview, and human organization.

A. C. Haddon pioneered studies in the evolution of art styles across generations of societies in the region of the Torres Strait. A. C. Haddon, in his *Evolution in Art: As Illustrated by the Life- Histories of Designs*, 1895, expressed the notion of art as an evolutionary human continuum of styles that were extensions of sensual pleasure, information, status, and beliefs.

This is to say that a culture is viewed in terms of a worldview and preferences in life that form an identity of people within a society. The cases we explore come from a heritage shared in cemeteries with their land and history in a network of relationships across time. In terms of the determinants of aesthetics in human life, Rudolf Arnheim (1967) wrote a comprehensive work on "pure" aesthetics that is convincing, yet does not address how people act from the context of their language, heritage, and experience. Styles of civilization (Schapiro 1953:287) were becoming understood as prevailing cultural norms kept by the arts. Jacques Maquet (1986) in his *Aesthetic Experience* further explained universal cultural experience based on aesthetic sensibilities, yet being culturally specific to a contextual origin.

People who have entered the aesthetic process reflect on the values and tastes of their society with openness for speculation and wonder, yet are hemmed in by boundaries that delineate their cultural ethos. Artists and craftsmen are professional specialists who conserve or extend from an ethos or lineage of this circumscribed knowledge. The cultural aesthetic leaders in a society point to the models for primary concepts of structures and forms. Artists and craftsmen transfer aesthetic ideals into forms. Claude Lévi-Strauss explains it this way:

“The painter is always mid-way between design and anecdote, and his genius consists in uniting internal and external knowledge, a ‘being’ and a ‘becoming’, in producing…” art (Lévi-Strauss 1973: 25).

The question that could be asked is how discreet and allusive categories could be sensed in experiencing the aesthetic? This is found in understanding the “golden mean” in Western cultures or the expected tastes and role behaviors in “serving Japanese tea.” The aesthetic qualities are the essence in sharing human information as a “we-ness” or togetherness in the experience to say “ah ha!” The components in the relationship manifest a spatial temporal ethos of otherness-and- self, data sensed, and the aesthetic composition.

This is to say that a culture is viewed in terms of a worldview and preferences in life that form an identity of people within a society. The cases we explore come from a heritage shared in cemeteries with their land and history in a network of relationships across time. Components for this understanding are introduced as a structure and functional process determining the way people cognize a culture based on perceived tastes or sentiments. Anthropology offers the tools for viewing and representing people in their own terms to understand diverse strands of a culture as a holistic notion of beliefs, things, and sense of place.

Vignettes of Our Research

We have case studies based on methodological questions were created on issues of research design and strategy as an empirical science based on spatial humanities, a sub-discipline of digital humanities based on geographic information systems (GIS) and timelines providing an effective integrating and contextualizing function for geo-cultural attributes.

As for information from multiple sources and in multiple formats, they create visual indexes for diverse cultural data. Spatiotemporal interfaces provide new methods of integrating primary source materials into Web-based interactive and 3D visualizations. We are able to chart the extent of specific traits of cultural information via maps using GIS gazetteer style spreadsheets for collecting and curating datasets. The system is based on GIS point locations, routes, and regions linked to enriched attribute information. These are charted and visualized in maps and can be analyzed

with network analysis, creating an innovative digital infrastructure for scholarly collaboration and creation of customizable visualizations. This method gives the researchers an expanse of data in layers of time across space providing new tools to advance humanistic inquiry.

Examples of Religious Networks in Monsoon Asia based on Evidence in Cemeteries

We ask the question to what extent did religious systems circulate in Monsoon Asia facilitated by Austronesian navigation, such as beliefs in the Buddhist/Hindu *dharma*, beginning about 2,500 years ago? This is to say there was a range of influence stemming from Southern Asia across and around the Bay of Bengal to Southeast Asia. The region became receptive to the *dharma* in Myanmar, Thailand, Malaysia, Vietnam, and Indonesia, yet to what extent did the religious system go further northeast across the islands of Taiwan and Japan, and continental China and Korea. How could maritime routes be traced? Was there a limit? And if so, why?

Our supposition is beliefs in the *dharma* as a literary belief system from Southern Asia was carried as far as writing could be traced on palm leaves, metal, and stone. In the 2nd-century CE, the *dharma* moved out by sea travel onboard ships with seasoned mariners, such as Austronesian navigators as mentioned in Southern Asian literature and in stone relief imagery. We further trace the extent of seemingly unrelated cultures intersected, and its periphery (see Blundell 2017).

What is the *dharma*? The meaning depends on context and time of usage. I originated in Southern Asia from prehistory, before the literary traditions, as fundamental constant in the teachings of the ways of nature. As life is born, and nurtured to grow and follow a path, short or long, to its end of the physical form as it returns to the elements of the earth. The essence of life recirculates into a new life depending on the lessons learned, or not, in life path process. These notions extended in to the literary religious traditions of Southern Asia to become known as Buddhism, Jainism, Hinduism, and Sikhism, etc. Theses religious beliefs were based on personal *bhakti* or devotion to a deity or spiritually awaken masters, such as the Buddha, Mahavira, Siva, Vishnu, or Guru Nanak.

The *dharma* spread as literary religious forms across Monsoon Asia with merchants who established sanctuaries of veneration, with such as a *candi*, *stupa*, pagoda, shrine, or temple. With literature came the motifs and symbols associated with the *dharma* traditions, such as lotus, water lily, elephant, lion, deer, etc.

Our research is sourced and measured interactively from Taiwan, where the data seemingly unrelated, yet connected to points based on aesthetic motifs, symbols, and other religious attributes originating from the *dharma* traditions of Southern Asia, extending to Southeast and East Asia. Other traditions in the cemeteries are Daoist, Shinto, Christian, etc. We ask to what extent did these religious systems and related

motifs and symbols spread across ocean island areas of Monsoon Asia? How did these religious influences reach Taiwan and manifest in tombs?

Between Amami and Yaeyama: Transformation of Ryukyu Islands Burial Practices

Earlier this year, Oliver Streiter and Hanna Yaqing Zhan surveyed the Ryukyu Islands between Amami and Yaeyama.They explored tombs of these islands representing a unique looking glass into the transformation of burial practices alongside the transformation of societies from early fishing and agricultural societies into a modern society within a relative short period of a few hundred years. Despite the destruction brought about through War II in the Pacific, the US military installations on the islands and the urbanization on larger islands, examples of even the most archaic burial forms can still be found all over the archipelago through the cultural and ecological protection that prayer-sites, e.g. Utaki (in Yaeyama called 'Ong'), tombs and wells, are placed under Ryukyu cultures.

In addition, old, unique and important tombs have been well documented and researched by local governmental institutions, dating the inscription-free tombs through intensive genealogic research. A factor overlaying the transformation societies is a noticeable cultural Chinese influence during Ming and Qing dynasties, and from the 19th century onwards, a Japanese colonial southwards expansion. As a consequence of this Japanese colonial influence, island after island was subjected to systematic forms of Japanization, such as the Shintoization of local prayer sites as well as the promotion of cremation over the local wind-burials. The research documented revealed characterizes examples of documented tombs according to (a.) time and forms of organization of a society, (b.) their north-south geo-location and (c.) their cultural sphere of influence, e.g. Chinese or Japanese.



(a.) Ishigaki, Okinawa, Toujinbaka/Tangren Mu (唐人墓), a tomb and memorial of Taiwanese who drowned near Ishigaki when trying to



(b.) Izena, Okinawa, turtle-back style tomb. Raw building material, cut into large chunks and perfectly combined. Perfectly integrated in nature, it re-

escape their enslavement by an US-American captain.	creates a natural cave.

FIgure 4. Aesthetic values of people who build tombs as different as these.

Below you will find examples of contrasting aesthetic perceptions epitomized in the Shintō cult, compared to the aesthetic perceptions epitomized. In Okinawan spiritual sites, dominance over nature versus the flow of nature respectively. This also reflects also in the over-stylized, almost geometrical symbolism introduced by the Japanese found on new tombs in Okinawa, compared to the organic motifs on Okinawa funerary urns (see Figure 4). Square-shaped burial plots, tombs and tombstones, as preferred by Japanese and Taiwanese Mainlanders, likewise contrast with round or organic forms preferred for exactly the same objects by Taiwanese and Okinawans. These contrasts are amplified by the preference for light over shade by Japanese and a balanced equilibrium by Okinawans, as well as the preference by Okinawans for pure perceptions, while Chinese perceive the beauty of characters carved on stones and walls.



FIgure 5. The Ryukyus, from Yaeyama to Osumi. Fieldwork has been done so far on Yaeyama, Okinawa and Amami. About 100 sites have been documented so far in this area.

Burial Practices in the Ryukyus

The Ryukyu is an archipelago of the Western Pacific Ocean that has to a large extent been formed by the uplift of coral reefs (see Figure 5). Its landscapes are characterized by limestone formation that are home to thousands of smaller and larger caves. These caves played a central role, among others, as burial sites for at least one millennium.

Yet, after the successive integration of the Ryukyu Islands into the Japanese empire in the 19th century, burial rites, burial sites and the notion of sacredness have transformed, among others, through the statal promotion of cremation on the one hand and the Shintō (神道) belief system on the other. As we will explain below, the introduction of cremation allowed burial sites to be set up as stand-alone structure, independently from the existence of hills and caves, e.g. in flat lowlands or promontories. These new sites much better approximate a Shintō aesthetics of a sacred site, a rigorously curtailed lot of land exposed to the sun, a conception that radically opposes the Ryukyuan conception of sacredness that resides in shady, overgrown, and untouched areas of marked geological formations.

But tombs are only one of the various sacred sites in the Ryukyus that are closely

interrelated with their geological and natural environment (see Figure 6).



(a.) A royal tomb[2]



(b.) A cave with a *yushiri,* a temporary urn



(c.) An *utaki*[3]



(d.) A well associated with a *kaanu-kami,* the deity of the well

Figure 6. Sacred sites on the Ryukyus, merging culture, geology and nature

Of these various sacred sites, the *utaki* is the most intriguing concept. Hard to photograph, they usually consist of nothing one could show except a deep aesthetic sentiment of geology, cosmology, the ocean, nature and human spirituality coming together in a single spot. Interestingly, in their attempt to Japaneseize the Ryukyu Islands, Japanese didn't destroy *utak*i and other sacred sites, but tried to integrate them into the Shintō belief system, usually by successive Shintō layers of temples and gates. There are thus hundreds of utaki which from the outside look like a Shintō temple, and hundreds of Shintō temples, which after some searching show to have an *utaki* as their true origin. This approach to japanization can also be observed with other shrines that have been brought from Fujian to Okinawa, i.e., those dedicated to Tudigong and Mazu (see Figure 7).



(a.) Mount Gusuku, Ie



(b.) Shintō shrine, called



(c.) The *utaki* proper

[2]Photos by the authors on their field trips between June 2016 and September 2018.

[3]Okinawan *utaki,* Japanese *otake,* a prayer site associated with a *nuru* or *nuuru*, Japanese *noro*, a local priestess and shaman. Related words are *taki,* a hill or grove and *utakabi,* meaning a sacred ritual to offer prayers. On Yaeyama, an *utaki* is called *ong*, and we assume that on other islands still other names are used for this kind of sacred site.

Island, Okinawa, Shintō gate (*torii*)	*utaki*	
 (d.) Ishigaki, Okinawa, the Shintō gate (*torii*) of Ama-on Ong	 (e.) The shrine of Ama-on Ong	 (f.) The original gate of Ama-on Ong

Figure 7: Temples and sacred sites of non-Japanese origin encircled by Shintō

As can be observed in cultures of Austronesian speakers, the burial practices of the people on the Ryukyu Islands involved two principal rites, separated by at least seven years. During the primary burial, the corpse was disposed in a half-open coffin in the front part of a ventilated cave to allow for its decomposition (*shiru-hirashi*). The secondary burial, conducted years after the first burial, involved the washing of the bones in ocean water (*shin-kuchi*) and the permanent storage of the bones in the back of the same cave.

Bones have been laid simply on the ground or stored in the ceramic urns (*jiishi-gaami*) or wood chest. Such containers stored the remains of individual persons or married couples or various generations of a certain class or profession. Inscriptions have been added, from the Ming period onwards, on wood sticks placed inside the urn, or with ink the inside or the outside of the urn, where occasionally also carvings are found (see Figure 8).


(a.) Painted in the lid of an urn


(b.) Painted inside an urn


(c.) Carved on the outside of an urn

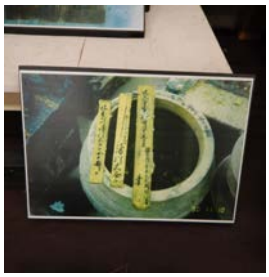

(d.) Painted on wood sticks

Figure 8: Inscriptions in or on urns indicate the name and the date of birth, death, and bone washing.

Motifs, similar to inscriptions, are found mainly on the urns. Only on tombs set up in the 20th century show symbols, inscriptions, and motifs on the outside walls of the

tomb. These motifs and symbols may for a large part be of Japanese origin. On the urns, we we find motifs which predate the Japanese occupation, we can distinguish three types of motifs. First, there are representations of the tomb on the urn with the louvers of the tomb becoming the louvers in the urn. Second, there are conventional *dharma* symbols, and third, motifs of Chinese mythologies (see Figure 9).



(a.) Ie Island, Okinawa: stylized pagoda



(b.) Ie Island, Okinawa: Buddha and lotus



(c.) Noho Island, Okinawa: a dragon



(d.) Ie Island: a stylized tomb structure



(e.) Ie Island: wave-like motif



(f.) Ie Island: zigzag pattern

Figure 9: Urn motifs, styles, and patterns on funerary urns.

Although openness has been a central features of these caves to function as a site for wind burials, these caves have through centuries gradually transformed from open caves over semi-open caves to caves with a front wall with louvers becoming a stylistic element, which as we have seen has been reproduced on the urns. The front wall of the tomb developed into house-like facades, house-like porches and finally stand-alone house-like structures. These house-like structures however have become only possible, as the wind burial has been replaced by cremation, and the burial sites only functioned as storage device of urns. The dry ventilation we find in natural caves, helps to decompose the body of the first burial, while keeping the bones of the second burial dry. In a stand-alone house, ventilation would for at least half of a year be very humid, causing the disintegration of the bones. Therefore, in order not to expose the bones of the second burial to the humidity, most houses are minimally ventilated and thus not suitable for wind-burials. The introduction of cremation as common practice under Japanese thus triggered a transformation of

burial sites from cave sites (*fui-nuchi-baka*) to necropoles (*haka-ji*), which are easy to reach and maintain (see Figure 10).



(a.) Open burial cave



(b.) Semi-open burial cave



(c.) Front wall with louvers



(d.) House-like porch

FIgure 10: The architectural development of burial caves from an open cave to a house-like structure, attached to a hillside.

As these latter house-like structures are no longer dependent on the existence of natural caves, new grave sites have sprung up in sites which previously had not been used for burials, e.g., lowlands, an empty space along a road or beaches. In the northern part of the Ryukyus, administered under the Kagoshima Prefecture, Japanization started earlier and graveyards resemble to different degrees gravesites on Honsho, replacing the notion of house as iconic symbol by the imperial column (see Figure 11).



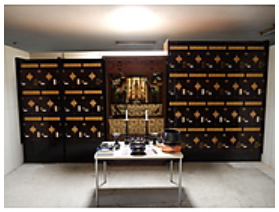
(a.) Ie Island, Okinawa, modern necropolis



(b.) Naha, Okinawa, modern necropolis



(c.) Naha, Okinawa, modern, giant columbarium



(d.) Naminouie Shrine (波上宮, Naminoue-gū,) Naha, Okinawa, a columbarium

FIgure 11: Modern burial forms that developed with the introduction of cremation.

Sometimes original location of the cave-tomb is used to build a new modern family vault, adding the new building just in front of the cave (see Figure 12).



(a.) Ishigaki, Okinawa, the new family vault built in front of the original cave-tomb



(b.) The original cave-tomb behind the modern tomb

FIgure 12: Modern burial style extending the older cave-tomb.

What modern Okinawa tombs with the new invention of tombstones reveal is that some Okinawan families in Naha have beside their Japanese surname also a Chinese surname. These two family name are essentially unrelated and have probably been transferred during the Ming and Qing period by the Okinawa authorities to local families (see Figure 13).



(a.) Ma (麻氏)



(b.) Yang (楊氏)



(c.) Liang (梁氏)



(d.) Fu (傅姓)

Figure 13: Chinese surnames along Japanese surnames on modern funerary epigraphs (photos from Maaji, Naha, Okinawa).

Taiwanese on Yaeyama

Yayama is the most southern archipelago within the larger Ryukyu archipelago. Located about 200 km to the east of Yilan in Taiwan, there has always been an intensive cultural and economic exchange between these two regions. Migration of Taiwanese to Yaeyama was first stimulated by the Japanese authorities when Taiwan was under Japanese control to reclaim arable land and plant pineapples. With the end of WWII and the retreat of Japan from Taiwan, many but not all

Taiwanese left Yaeyama. After WWII, Taiwanese could apply to work on Yaeyama, mainly in the pineapple processing industry.

In order to adapt to the life on Okinawa, many Taiwanese families changed their surname into a Japanese style surname, usually transforming a one-character surname into a two-character surname. Okinawa style family names can also be formed of three or more characters and are frequently derived from place names. A common family name on Ishigaki (石垣) is thus Ishigaki (石垣). The strategy that most Taiwanese followed to change their name was based on the character and not based on the pronunciation. Frequently a second character is added to the Chinese family name. Alternatively, character components are added or removed before a second character is added. The purpose of this strategy is obvious: While Japanese and Okinawan might perceive a common Japanese surname, Taiwanese can within the Japanese surname still identify the Taiwanese surname.

Given names apparently needed less adjustment to become socially acceptable. Table 1. shows only one example where the character '花' has been replaced by the character '代'.

Original Taiwanese surname	**Japanese Surname adopted**	**Example of a living person**
吳 (Wu, Ngo)	吳屋 (Goya)	吳屋吉男 (born in Ishigaki, second generation)
張 (Zhang, TiuN)	張本 (Harimoto)	張光輝 →張本光輝
陳 (Chen, Tan)	東 (Azuma)	陳清龍 →東郷清龍
黃 (Huang, Ng)	横山 (Yokoyama)	黃長發 →横山長発
湯 (Tang, Tng)	湯川 (Yukawa)	湯建興 →湯川建興
林 (Lin, Lim)	林 (Hayashi)	林發→ 林發
王 (Wang, Ong)	玉木 (Tamaki)	王玉花 →玉木玉代
曾 (Zeng, Tsan)	曾根 (Sone)	*unknown*

Table 1: Japanese Surnames adopted by Taiwanese families on Yaeyama.

The Taiwanese Cemetery on Ishigaki Island

Ishigaki is the name of the second-largest and economically most important island of the Yaeyama archipelago. The Taiwanese cemetery in Ishigaki is located near

Nagura, a village where still many Taiwanese reside. It is at the center of the agricultural production on Ishigaki Island.

Founded in 1969 and managed by the Overseas Chinese Association of Yaeyama, the cemetery is now home to family vaults, which most probably contain from recent times ash urns from cremations, and bone urns from personal tombs originally spread over the entire island or even archipelago (see Figure 14). Only one tombstone, put beside a family vault is a personal tombstone and thus maybe a rare instance of a primary tombstone (see Figure 14, j.).

Central to the cemetery is a common tomb for all Taiwanese deceased which don't have a family on Ishigaki to take care of the tomb and to perform the funerary rituals. Common tombs are commonly found among the tombs of Chinese in Southeast Asia, yet to our knowledge, not in Taiwan, Hawaii or the United States generally. As a common tomb bears the risk of mixing up or confounding the bodily remains, in our interpretation, the establishment of common tombs indicates that bones are not intended to be transferred to an overseas hometown.

<table>
<tr>
<td> 

(a.) Common tomb (公墓)</td>
<td> 

(b.) Houtu (后土)</td>
<td> 
(c.) Tudigong (土地公) and Jinlu (金盧)</td>
</tr>
<tr>
<td> 
(d.) Taiwanese-style tombstone, Fulushou (福祿壽), Jintong Yunu (金童玉女)</td>
<td> 
(e.) Okinawa-style tomb with tombstone on the roof</td>
<td> 
(f.) modern lotus or family symbol: Japanese-style</td>
</tr>
</table>



(g.) Lin 林 family tomb



(h.) Sone 曾根 family tomb, originally probably Zeng 曾



(i.) Okinawa folk motif



(j.) First burial tombstone, Taiwanese place name



(k.) Tanghao Longxi (隴西) of Li 李 family



(l.) Okinawa-style front opening to the family vault

Figure 14. Some aspects of the Taiwanese cemetery in Ishigaki.

Although composed of only less than 30 family vaults, the tombs are all but homogeneous. Some tombs have the tombstone before the tomb in central position (Taiwanese style), others have in this place the entrance to the vault and the tombstone on the top (more recent Okinawa style). While some tombs have a Houtu/Tudigong, lions, Fulushou or Jintong Yunu, others don't have any of these representations. Names vary between only Chinese, Chinese and Japanese and only Japanese. Also the epigraphs can be arranged according to Taiwanese conventions, or just mention the family name, as on more recent Okinawa tombs. Overall, there seems to be a transition in time from a Taiwanese style with Taiwanese tiles to a more austere Okinawan style with the entrance to the vault centered in front to the tomb. This might reflect the loss of cultural awareness or even interest in the Taiwanese roots of this community, experienced by younger generations.

The only attributes these tombs probably share is that they all a family vaults that shelter urns that have been built on southwards oriented square lots of land. Interestingly, our informant on Ishigaki, who migrated in his early childhood to

Ishigaki, didn't perceive a noticeable difference between these tombs, nor between the Taiwanese tombs and the Okinawa tombs nearby. Apparently these cultural differences were not relevant to him, similar to the question whether one grabs a glass of water with five or four fingers. Our informant was clearly bicultural and thus apparently not seeing what a monocultural visitor would have observed.

Religious Practices: Tudigong

Each year on August 15th of the lunar calendar, Taiwanese in Ishigaki hold a Tudigong festival (土地公祭) at Nagura Ong (名藏御嶽). The chair of the Overseas Chinese Association of Yaeyama is in charge of this religious practice. In preparation of the event, he informs all members of Overseas Chinese Association about this event, going personally from house to house, recreating the bond between dispersed families on the Yaeyama islands.

In 1930s, Taiwanese immigrants have been too poor to establish their own temple for ritual offering to the gods and goddesses of their hometowns. Therefore, the Taiwanese immigrants borrowed the Ong in Nagura, belonging to the Okinawa community, to worship to Tudigong. The space of the Nagura Ong thus turned into a multi-cultural site: On the access road to the site the entrance to the area is marked by a Shintō gate (torii). The main yard of the site is dedicated to Tudigong and offers ample space for the Taiwanese to display their offerings, such as pigs, gold paper and, moon cakes brought from Taiwan. The back of the Nagura Ong is the sacred prayer site of Okinawan, called *ong* on Ishigaki and *utaki/utake* on other islands of the Ryukyus. Similar to the festival dedicated to Tudigong in Taiwan, people gather to pray for a rich harvest in the upcoming year and to thank Tudigong for the peace experienced this year. The Tudigong festival end with a ceremony that by casting yin yang fortune-telling moon-shaped wood blocks decides who will be the man in charge of worshipping Tudigong throughout the coming year at his home. This person will be called called '爐主' (Lúzhǔ, the master of the furnace). On this day in 2018, about 50 people whose parents and grandparents have migrated from Taiwan gathered at the Nagura Ong. By now, all of them hold the Japanese nationality. Grand- and great grant children don't necessarily understand the rituals performed at the Tudigong festival, but they nevertheless join the activities and respect its traditions in memory of their ancestors.

By celebrating a Tudigong festival on the site of Ong, the Taiwanese connected twice to established practices on Okinawa. With Mazu and Tudigong having been brought from Fujian to Okinawa in the Ming and Qing dynasties, the Taiwanese didn't have to introduce a new deity on the island. They worshipped a known deity on a site that had be designated by the people of Okinawa as a prayer site. The Taiwanese thus renounced in public to their myriade of gods and goddesses and concentrated on one of the two (Tudigong and Mazu) that might help to connect themselves to the Okinawans. They also renounced to their aesthetic conceptions of a prayer side and accepted the shady, partially overgrown clearing in the middle of a

forest, which in the 1930 might not have had a shed, a hut, a stage or shrine. The early immigrants might not have heard of Karl Jaspers, but one of his mottos, “Wahrheit ist, was uns verbindet (truth is what connects us)”, might have been shared by them (see Figure 15). This reflects through the god and goddess they worship, the surnames they adopted and the burial practices they followed. It’s not an ‘either or’ or ‘both’, its the ‘same’, the active attempt of the immigrants to find congruence and harmony with their new social environment by actively abstracting away from the surface forms of cultural features. We now understand the meaning of the common tomb. There is no need to re-bury bones in Taiwan. These people’s home is everywhere.



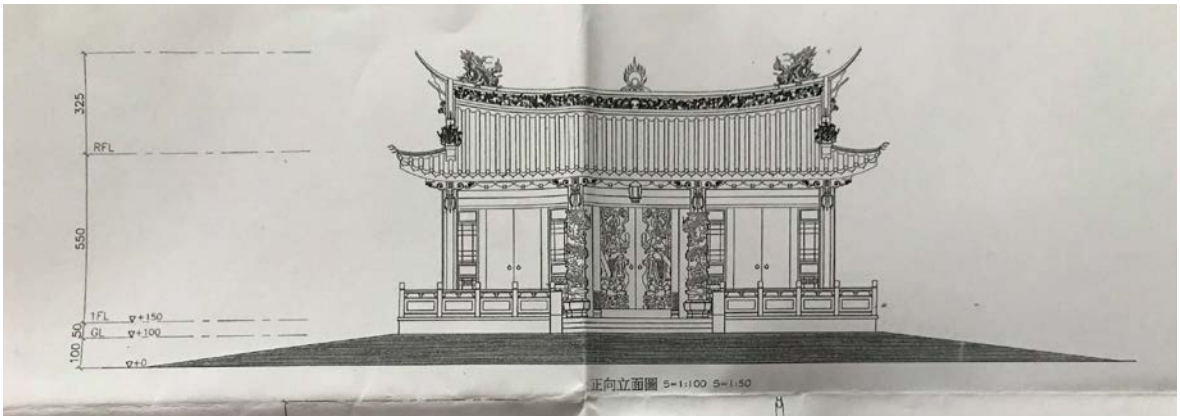
(a.) Nagura, Ishigaki, one of two pigs as part of the sacrificial offerings



(b.) Tudigong festival is held at the Tudigon temple in Nagura Ong, its access is marked by a Shinto gate (*torii*)



(c.) this Tudigong statue has been brought from Taiwan about ten years ago



(d.) The construction plan for a Taiwanese Mazu temple to be built next to the Toujinbaka/Tangren Mu (唐人墓). This temple will also be a new home to Tudigong. Upon its completion, the Tudigong Festival will held here. The plan, probably proposed by a Taiwanese company unfortunately breaks with the aesthetic agreement that Taiwanese and Okinawans have found in the last one hundred years.

Figure 15. Tudigong festival (土地公祭) at Nagura Ong (名藏御嶽) in Ishigaki.

Muslim Tombs in Eastern Asia

Muslim tombs in Eastern Asia are a perfect example to study inverted worlds, e.g. in comparison with ethnic Chinese tombs. While in Taiwan, Kong Kong and Macau, Muslims are a minority in a majoritarian ethnic Chinese community, in other parts of Asia, ethnic Chinese are a minority among Muslims and yet in both cases, both communities have to find ways to find a peaceful coexistence.

Muslim Tombs in Taiwan

Muslims cemeteries in Taiwan have mainly been created for a subgroup of the so-called “Mainlanders”, the Chinese who came after 1945, mostly in 1949 with the KMT, the Nationalist Chinese Party, to Taiwan. The community of Muslims was relatively homogeneous, consisting of the Hui Muslims, the Muslim community in China that shows the highest degree of assimilation of Chinese practices.

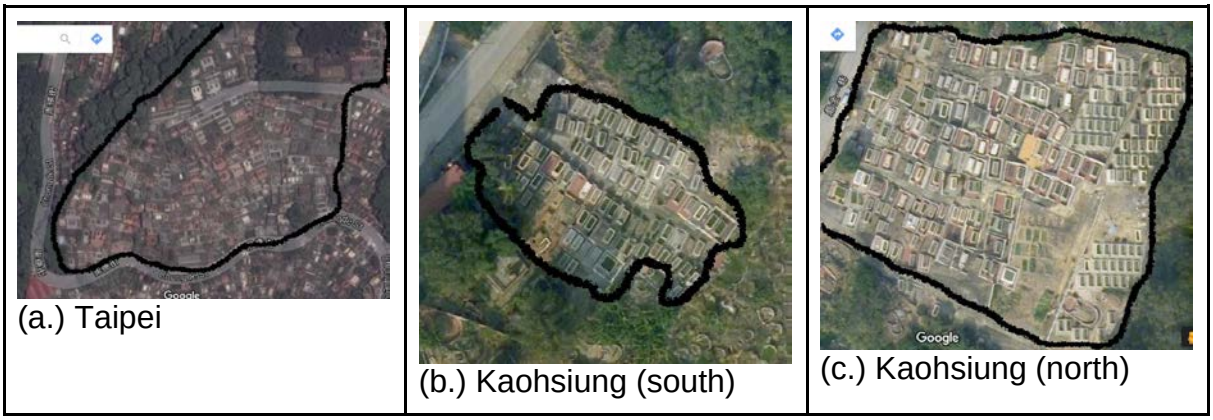
Two Muslim cemeteries have been documented entirely, one in Taipei and one in Kaohsiung, the latter split into two separated sections. Both are managed by mosques in the two cities respectively. Chinese is the main script found on tombstones, reporting names, place names and dates. Although most tombstones also show Arabic text, the function of Arabic on these stones is to evoque the beginning of a prayer as a marker of the religious identity of the deceased (see Figures 16, 17, 18).



FIgure 16. Taipei cemetery of Hui Muslims. Shared with the Mainlanders of Taipei is the preference of rectangular tombs and tombstones.



Figure 17. Kaohsiung graveyard of Hui Muslims.

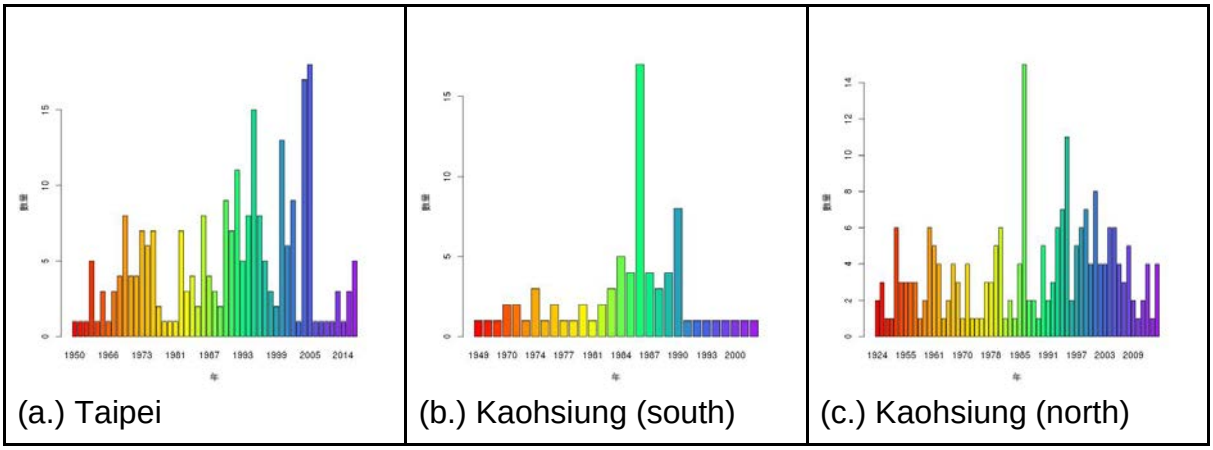


(a.) Taipei

(b.) Kaohsiung (south)

(c.) Kaohsiung (north)

Figure 18. Designated Muslim sections within larger ethnic Chinese graveyards. They all include older, non-Muslim tombs.



(a.) Taipei

(b.) Kaohsiung (south)

(c.) Kaohsiung (north)

Figure 19. The temporal development of the Muslim graveyards after 1945. Since 2018 only the Taipei graveyard is still extant and in usage.

Built on hill slopes, these Muslim graveyards in Taiwan form giant concrete structures with a two-level tomb design. The lower level is an otherwise empty vault where the corpse, wrapped in cloth, is layed down, with the head turned to the right sight, facing Mecca. The empty vault is conceptually similar to the vault where non-Muslim Mainlanders are buried, avoiding the contact with Taiwanese ground, and

hoping to be reburied in the future in their Chinese homelands. In order to follow the Muslim rule according to which a tomb has to be open at the top to let the rain enter, a second level, filled with earth and in most cases covered by grass, is built on the top of the lower vault and rises above ground level. This two-level tomb architecture reflects the double Muslim-Chinese nature we also find in the tombstone inscriptions: A fundamentally Chinese gravestone with a *jiguan* in China, a list of the offspring and few particularities that hint to Islam is usually adorned at the top with the crescent and star, or the onset of a prayer written in Arabic.



Figure 20. The two-level design of Muslim tombs in Taiwan, serving two incompatible conceptions.

Thus very differently from the Taiwanese in Okinawa, how tried to find an intersection between their home culture and the arriving culture, the Hui in Taiwan present themselves as Chinese Mainlanders and as Muslims. Not as ‘either or’, not as ‘the same’ but as ‘both’. To achieve this, also some cultural labor was required, e.g. finding with the two-layer level of the tombs a way to overcome conceptual conflicts, obeying to the Muslim regulations by word, not by their sense, and following Mainlander practices.

This superficial adherence to Islamic burial regulations in order to achieve a closer match to non-Muslim Mainlanders is probably best illustrated by a certain number of Muslim tombs where the grass on top of the tomb is replaced by green tiles, so that if seen from the far, Muslim regulations seem to have been followed.




Figure 21. The visible upper layer of the tombs allow rain to enter, while the lower layers are sealed of. If the entire tombs is sealed, grass is emulated in Kaohsiung. No attempt to emulate a natural surface is made in Taipei, however, see images above.

Muslim Tombs in Hong Kong

Visiting Hong Kong and documenting these tombs help complete a puzzle that the ThakBong team tries to put together. Hong Kong shares with Taiwan a colonial background its population came from different places that brought about the diversity of cultures, languages and religions (see Figure 23). There are however also a number of significant differences.

The Muslim communities in Hong Kong, for example, are much more diverse than in Taiwan. Beside the so-called Hui-Muslims, there are various Islamic groups that have arrived to Hong Kong as part of the Commonwealth military forces and which were not allowed to return in their lifetime to their homelands, e.g. Pakistan, India or Bangladesh. The Muslim community in Hong Kong is thus also much older than in Taiwan, with the oldest tombs dating from the second half of the 18th century.

Two Muslim cemeteries have been documented in Hong Kong, the old Muslim graveyard in Happy Valley, Hong Kong Island and the new Muslim cemeteries in Chai Wan, Hong Kong Island. Both are managed by the Incorporated Trustees of the Islamic Community Fund of Hong Kong,



Figure 21. The oldest part of the Muslim cemetery of Happy Valley in Hong Kong shows a variation of styles and inscriptions.

The latter can be found on foot- and headstones, inside and outside in all possible combinations. English, Chinese and Arabic are usually on different sides of head- or footstones.



Figure 22. Inscriptions in English and Arabic on the Happy Valley Muslim Cemetery.

Particular about this tombstone is its circular arrangement around a hill, as shown in Figure 23. As a consequence, tombs are not uniformly oriented, although believers might have found a way to arrange the bodies so that they can face Mecca.

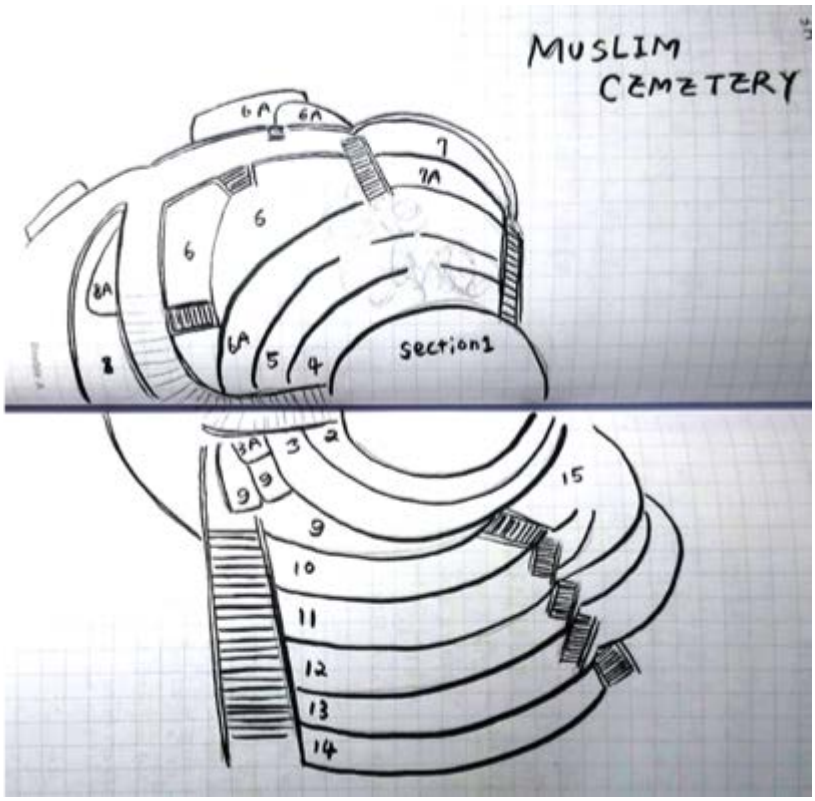


FIgure 23. A map of the old Muslim Cemetery in Happy Valley, Hong Kong.

The new Muslim Cemetery is much more uniform in design and inscription. Plots are uniformly oriented and only headstones are inscribed. Chinese has become the prominent language for inscriptions. Arabic, similar to the usage in Taiwan, is used mainly to evoque a prayer and not as a language of communication, e.g. to transmit a name or the date of birth and death, see Figure 24.



Figure 24. Details of Muslim tombstone in the new Hong Kong Muslim cemetery in Chai Wan.

Muslim Tombs in Southeast Asia

For our comparison of Muslim graveyards in Taiwan and Hong Kong, where Muslims live as a minority with an ethnic Chinese majority, with regions where Muslims are more numerous or form the majority, we also visited Muslim graveyards in Malaysia and Singapore. These graveyards shared a few features with the graveyards we previously visited in Hong Kong and Taiwan. Common features seem to be the absence of incense and a uniform orientation of tombs, with the noticeable exception of the old Muslim graveyard in Hong Kong (see Figure 20). Also more recent fashions, as ornamenting tombs with coloured stones or glass stones are shared between Hong Kong and Southeast Asia.

Yet, there are also a great number of features that set these graveyards apart, showing, how much the Muslims in Hong Kong and Taiwan assimilated languages, aesthetics and rites. Arabic, English, Chinese and Malay can be found, the latter using Jawi script. Generally speaking, and in contrast to Hong Kong and Taiwan, inscriptions are generally reduced, either to a prayer or the name of the deceased. Often, however, no inscription is discernible on the stone or on the tomb. Head- and footstones, which in Hong Kong are common, and might be related to colonial traditions, are also found in Southeast Asia, yet in a completely different design. Shaped like abstract chess figures they show symmetrical forms as if obtained from woodturning wood and stones. Head- and footstone are often indistinguishable in size, material and shape. In Hong Kong, only a few toms in the oldest section share this design, in Taiwan, it is unknown. Sometimes these foot- and headstones are wrapped in white cloths.



(a.) Head and foot stones arranged in the the same direction



(b.) Geometrically styled tombstones in a seemingly abandoned gravesite



(c.) inscription bearing tombstone



(d.) Arabic inscription

Figure 25. Near Masjid Sultan, an old Malay cemetery beside Victoria Lane in Singapore. A majority of the tombs have ornamented foot- and headstones but show no traces of inscriptions. If there have been inscriptions, they may have completely decayed. Tombstone with inscriptions were uncommon. An example is shown in (c.) and (d.) with Inscriptions in Arabic.

Comparing to the above old islamic cemetery of Malay people in Singapore to a modern Muslim cemetery in Singapore, we can observe significant differences. For instance, in the Choa Chu Kang Cemetery Muslim section, where most burials are after year 2000, each designed in an individual style, featuring colourful decorations, with Quran scripture and prayers carved on sides of a tomb. The most common tomb colours are white, green, blue, and brown and decoration materials include marble, wood,cement, tiles, plants, grass, and turf, see Figure 26.





Figure 26. Newer Muslim Cemetery in Choa Chu Kang Cemetery in Singapore. Many headstones and footstones were wrapped with white linen or cotton cloths.



Figure 27. New, tidy, Muslim cemetery with standard-sized burial plots managed by the Singapore National Environment Agency.

Chinese Tombs in Southeast Asia

Similar to the situation of the Taiwanese in Okinawa, many of the graveyards we have visited in Malaysia have been managed by Chinese associations, e.g. temple associations. Landsmannschaft or “huiguan”(會館), among them some more historical and well-established are Associations of Chinese immigrants from Teochew, Hainan, Guangdong and Hokkien. They manage Chinese temples, cemeteries and foster unity among fellow countrymen. Also, common tomb or massive burial are found for burying more than one chinese, without names of every deceased.

Other than Chinese cemeteries in Malaysia, the ThakBong team also visited Chinese cemeteries in Singapore, where a majority of burial lands are managed by the government. The largest cemetery is the Choa Chu Kang Cemetery Complex and one of the earliest existing Chinese cemetery is Bukit Brown Cemetery.

References

Arnhiem, Rudolf. 1974. *Art and Visual Perception: A Psychology of the Creative Eye.* Berkeley: University of California Press.

Blundell, David. 1996. Aesthetic ethos. *Bulletin of the Department of Anthropology,* National Taiwan University, 51:43-58.

_____. 2017. Maritime religious networks: Aesthetic sources of flora and fauna

motifs from India and China on tomb elaborations in Taiwan. *ECAI Workshop: Atlas of Maritime Buddhism, Digital Humanities Support Technology, Cultural Mapping, and Text Translation and Analysis*. International Symposium on Grids and Clouds (ISGC). Academia Sinica, Taiwan. 6-7 March.

Blundell, David, and Jihn-Fa Jan. 2016. Workings of SpatioTemporal research: An international institute in Taiwan. In: *IEEE Proceedings of the 22nd International Conference on Virtual Systems and Multimedia* (*VSMM*). Sunway University, Kuala Lumpur, Malaysia.

Geertz, Clifford. 1976. Art as a cultural system. *Modern Language Note*, 91: 1473-1499.

Haddon, A. C. 1895. *Evolution in Art: As Illustrated by the Life-Histories of Designs*. London: Walter Scott Ltd.

Goudin, Yoann, Streiter, Oliver, Huang, Chun (Jimmy), & Lin, Ann Mei-fang. 2011. Digital Anthropology and the Renewal of Waishengren Studies: From Digitized Tombs to Identity Claims. African and Asian Studies, 15(2), 21–45.

Goudin, Yoann, & Streiter, Oliver. 2016 (May 18–20). Pratiques funéraires en migration : la référence à l'origine au prisme de l'analyse quantitative. In: *FRANCHIR, La Religion des Chinois en France, Colloque international.* Paris, France.

Jan, Jihn Fa. 2016. Digital heritage inventory using open source geospatial software. In: *IEEE Proceedings of the 22nd International Conference on Virtual Systems and Multimedia* (*VSMM*). Sunway University, Kuala Lumpur, Malaysia.

Lévi-Strauss, Claude. 1973. *The Savage Mind*. Chicago: University of Chicago Press.

Maquet, Jacques. 1986. *The Aesthetic Experience: An Anthropologist Looks at the Visual Arts*. New Haven: Yale University Press.

Schapiro, Meyer. 1953. Style. *Anthropology Today*, A. L. Kroeber, ed. Chicago: University of Chicago Press.

Watson, J. L. & Rawski, E. S. eds. 1988. *Death Ritual in Late Imperial and Modern China.* Berkeley: University of California Press.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

Macroscopic trends in Korean based on Newspaper Big data

Ilhwan Kim* **Do-gil Lee****

Professor, Sunshin Univ.* **Professor, Korea Univ****

Macroscopic trends in Korean based on Newspaper Big data

Ilhwan Kim, Do-gil Lee
Professor
Sunshin Univ., Korea Univ

Abstract

The purpose of this study is to construct large amounts of Korean newspapers into big data in order to analyze "macroscopic trends," which are distinct characteristics detected via newspapers that have influence over Korean language, society, and culture, over a long period of time. Our research project processed over 70 years of Korean newspaper big data, utilizing methods such as analyzing semantic prosody, keyword, word frequency, and co-occurrence analysis. The Korean newspaper big data project first began with the creation of the *Trends 21* corpus which incorporated close to a decade's worth of articles from Korea's major daily newspapers. In addition, analysis tools were also developed. However, due to the difficulties faced when capturing and analyzing trends using only 10 years' worth of data at the macro-level, the *Trends 21* corpus project extended its data sets to incorporate relevant newspapers that spanned over 70 years, from 1945 (Korean Independence) until 2014. While applying a number of methodologies toward macroscopic trend analysis, a large amount of emphasis was focused on utilizing word usage patterns which, unlike previously existing methods, does not rely on intrinsic values and intuition; therefore, word usage patterns provided more objectivity and verifiable measures while also capturing greater insight into its processed data. More specifically, macroscopic trend analysis can include other methods such as studying word usage frequency across periods of time, statistical analysis on keyword extraction, co-occurrence network analysis as well as keyword extraction using word co-occurrence, and semantic prosody. By examining the construction of the Korean newspaper big data system, which holds over 70 years' worth of data, and also how it could be used to analyze macroscopic trends through the multidimensional lens of language, culture, and society, our research has demonstrated that identifying trends at the macrolevel is not only quite applicable but also very versatile. Through big data analytics, such as text mining, macroscopic trends analysis can serve as a resourceful tool in a number of research methods as well as across a variety of disciplinary fields.

Keyword

Newspaper Big Data, corpus, Keyword Extraction, Co-occurrence Words, Statistical approach

Contents:

1. INTRODUCTION

2. METHODOLOGY

2.1. Research Method

2.2. Research Scope

3. MACROSCOPIC TRENDS: Analysis on Keywords and Co-Occurring Words

3.1 When Did "Regionalism" Begin in Korea?

3.2 Usage of "Freedom" and "Equality" Over Time

3.3 From "Hope" to "Happiness"

3.4 A Reason for Preferring One Over the Other: "Worker" and "Laborer"

3.5 Different Yet the Same: "Youth" and "Elderly"

4. CONCLUSION

Figures:

Figure 1: The Dong-A Ilbo Newspaper - Year vs. Amount of Text

Figures 2-4: Web-based Corpus Analysis Tool

Figure 5: Relative Word Frequency for "Regionalism"

Figure 6: Relative Word Frequency for "Freedom" and "Equality"

Figure 7: Relative Word Frequency for "Hope" and "Happiness"

Figure 8: Relative Word Frequency for "Laborer" and "Worker"

Figure 9: Relative Word Frequency for "Youth" and "Elderly"

Tables:

Table 1: Size of *Trends 21* Corpus

Table 2: Top 20 Main Co-occurrence Words of "Regionalism"

Table 3: Top 20 Main Co-occurrence Words of "Freedom"

Table 4: Top 20 Main Co-occurrence Words of "Equality"

Table 5: Top 20 Main Co-occurrence Words of "Hope"

Table 6: Top 20 Main Co-occurrence Words of "Happiness"

Table 7: Top 20 Main Co-occurrence Words of "Laborer"

Table 8: Top 20 Main Co-occurrence Words of "Worker"

Table 9: Top 20 Main Co-occurrence Words of "Youth"

Table 10: Top 20 Main Co-occurrence Words of "Elderly"

1. INTRODUCTION

The purpose of this research was to assemble large amounts of Korean newspapers into big data in order to analyze "macroscopic trends" which are distinct characteristics in newspapers that have influenced Korean language, society, and culture, over a long period of time. Our research processed over 70 years' worth of Korean newspaper big data, utilizing methods such as analyzing semantic prosody, keyword, word frequency, and co-occurrence analysis.

The Korean newspaper big data project first began with the creation of the *Trends 21* corpus which incorporated close to a decade's worth of data from Korea's major daily newspapers. In addition, analysis tools were also developed in order to assist in processing data. However, due to difficulties faced in capturing and analyzing trends when using only 10 years' worth of data at the macrolevel, the *Trends 21* corpus project broadened its data sets to incorporate relevant newspapers that spanned over 70 years, from the end of the Korean War in 1945 until 2014.

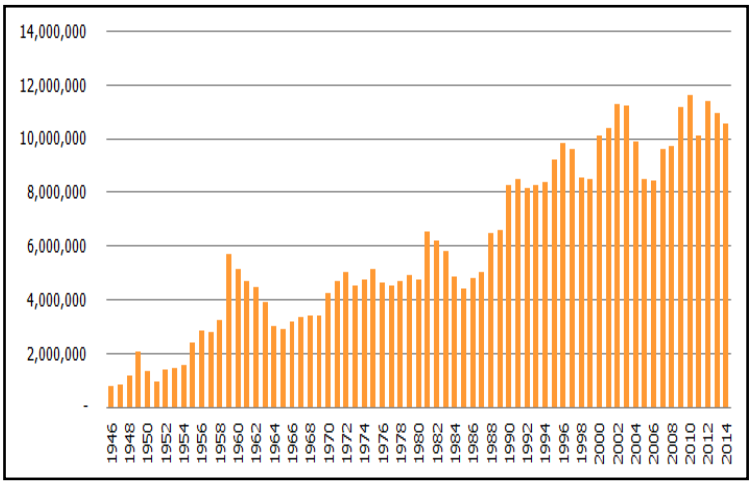
Table 1 represented the *Trends 21* corpus' year-to-year data from all of four daily newspapers - *Chosun Ilbo*, *Dong-a Ilbo*, *JoongAng Daily*, and *Hankyoreh* - while Figure 1 illustrated the extended version of *Trends 21* corpus, which featured the *Dong-a Ilbo*.

Table 1: Size of *Trends 21* Corpus

Year	# of Articles	# of Words
2000	231,361	40,377,975
2001	220,109	39,728,480
2002	214,941	42,298,083
2003	205,635	42,959,317
2004	200,260	40,652,017

2005	180,413	38,062,190
2006	162,055	37,539,133
2007	177,743	42,432,005
2008	176,697	42,322,623
2009	164,314	43,230,845
2010	170,338	48,307,337
2011	157,755	44,425,114
2012	153,227	44,017,316
2013	152,348	45,367,321
Total:	2,567,196	591,719,756

Figure 1: The *Dong-A Ilbo* Newspaper: Year vs. Amount of Text



Due to certain complexities within the Korean language, especially regarding conjugation and lexical semantics, notable keywords were determined and annotated beforehand by using a morphological analysis. The Korean Morphological Analysis Tool (KMAT), which processes large-scale word segmentation, was utilized in the *Trends 21* research.[1]

While applying a number of methodologies to conduct a macroscopic trend analysis, a large amount of emphasis was focused on utilizing word usage patterns which, unlike

[1] Developed by Lee Do-Gil, 2005.

previously existing methods, did not rely on intrinsic values and intuition; therefore, word usage patterns provided more objectivity and verifiable measures while also capturing greater insight into its processed data.

2. METHODOLOGY

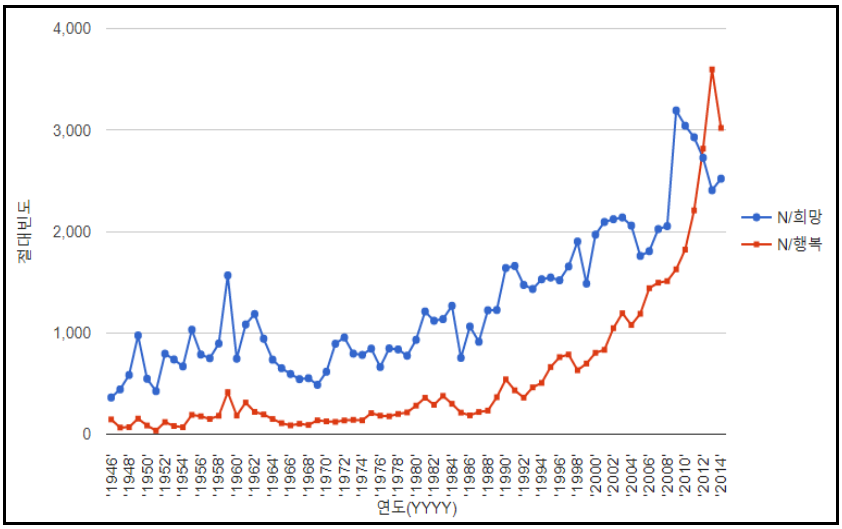
2.1. Research Method

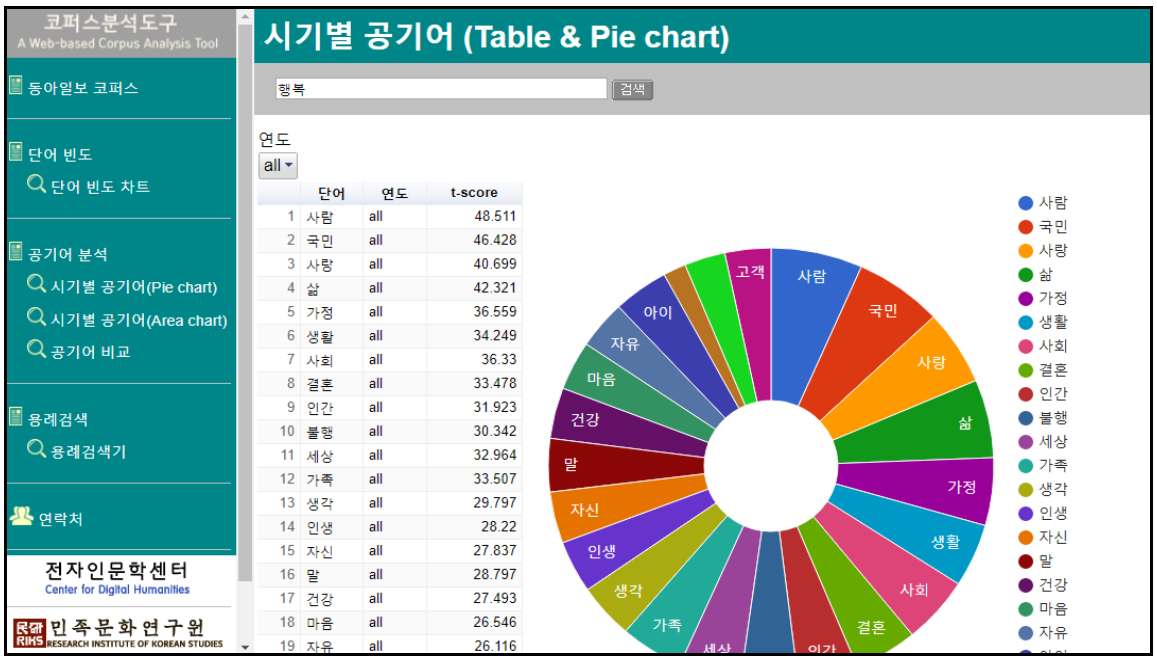
Which Method Would Be Most Appropriate to Capture Macroscopic Trends From 70 Years' Worth of Newspaper Big Data? This research first identified relevant keywords based on time as well as word usage patterns in order to process the data. Afterwards, the selected words went through post-processing in order to capture their implied comprehensive meaning. In particular, three methods were utilized to analyze macroscopic trends: 1) word usage frequency based on a set period of time; 2) statistical analysis on keyword extraction; and 3) co-occurrence network analysis and keyword extraction using word co-occurrence.

In order to evaluate and inspect macroscopic trends in big data, the analytical tools themselves had to efficiently process large amounts of information while also be able to utilize large-scale language resources. The Koreanic language has accumulated upwards to about 600 million words in over the last 70 years, which does not even factor in dynamical factors such as intuitive and qualitative properties of the language itself. In order to analyze macroscopic trends, the Web-based Corpus Analysis Tool (CAT-DongA) was developed for the *Trends 21* corpus project (see Fig. 2-4) .

Figures 2-4: Web-based Corpus Analysis Tool


코퍼스분석도구
A Web-based Corpus Analysis Tool
동아일보 코퍼스
단어 빈도
단어 빈도 차트
공기어 분석
시기별 공기어(Pie chart)
시기별 공기어(Area chart)
공기어 비교
용례검색
용례검색기
연락처
전자인문학센터
Center for Digital Humanities
민족문화연구원
RESEARCH INSTITUTE OF KOREAN STUDIES
웹 기반 코퍼스 분석 도구
웹 기반 코퍼스 분석 도구(CAT-DongA) 소개
코퍼스 분석 도구 개요
웹 기반 코퍼스 분석 도구(A Web-based Corpus Analysis Tool; 이하 코퍼스 분석 도구)는 고려대학교 민족문화연구원 디지털인문학센터에서 동아일보사로부터 제공받은 동아일보 코퍼스를 활용하기 위한 도구이다.
코퍼스 분석 도구는 언어자원의 공유와 확산을 목적으로 하고 있다.
코퍼스 분석 도구의 기능
코퍼스 분석 도구는 크게 다음과 같은 세가지의 도구로 이루어져 있다.
단어 빈도 차트
공기어 분석 도구
용례검색기
코퍼스 분석 도구의 사용에 대해
이 홈페이지에서 제공하는 코퍼스 분석 도구를 사용하는 경우 다음의 내용에 동의 및 인정하는 것으로 간주합니다.
코퍼스 분석 도구는 연구 목적으로 누구나 사용하실 수 있습니다.
다만 연구 목적인 경우라도 인용정보를 참고하여 적절하게 인용을 하시기 바랍니다.
코퍼스 분석 도구를 활용하여 연구 결과를 발표하신 후에 출처를 알려주시기 바랍니다. 또한 저자가 동의하는 경우 이 홈페이지에서 소개하겠습니다.
코퍼스 분석 도구의 결과는 상업 목적으로 사용할 수 없습니다.
코퍼스 분석 도구의 결과나 원자료를 비정상적인 방법으로 수집하려는 행위는 금지합니다.
특히, 용례검색기의 결과로 제공되는 신문 기사의 원문은 확인용으로만 사용해야 합니다.
코퍼스 분석 도구는 신문 자료의 전처리, 데이터 형식의 변환, 자동 언어처리, 색인 및 가공 등의 과정을 거칩니다. 이 과정에서 오류가 포함될 수 있습니다.


4,000
3,000
2,000
1,000
0
N/희망
N/행복
'1946'
'1948'
'1950'
'1952'
'1954'
'1956'
'1958'
'1960'
'1962'
'1964'
'1966'
'1968'
'1970'
'1972'
'1974'
'1976'
'1978'
'1980'
'1982'
'1984'
'1986'
'1988'
'1990'
'1992'
'1994'
'1996'
'1998'
'2000'
'2002'
'2004'
'2006'
'2008'
'2010'
'2012'
'2014'
연도(YYYY)


코퍼스분석도구
A Web-based Corpus Analysis Tool
동아일보 코퍼스
단어 빈도
단어 빈도 차트
공기어 분석
시기별 공기어(Pie chart)
시기별 공기어(Area chart)
공기어 비교
용례검색
용례검색기
연락처
전자인문학센터
Center for Digital Humanities
민족문화연구원
RESEARCH INSTITUTE OF KOREAN STUDIES
시기별 공기어 (Table & Pie chart)
행복
검색
연도
all
단어 연도 t-score
1 사람 all 48.511
2 국민 all 46.428
3 사랑 all 40.699
4 삶 all 42.321
5 가정 all 36.559
6 생활 all 34.249
7 사회 all 36.33
8 결혼 all 33.478
9 인간 all 31.923
10 불행 all 30.342
11 세상 all 32.964
12 가족 all 33.507
13 생각 all 29.797
14 인생 all 28.22
15 자신 all 27.837
16 말 all 28.797
17 건강 all 27.493
18 마음 all 26.546
19 자유 all 26.116
사람
국민
사랑
삶
가정
생활
사회
결혼
인간
불행
세상
가족
생각
인생
자신
말
건강
마음
자유

2.2. Research Scope

The scope and range of this research, which contains 70 years worth of newspaper articles, contained about 600 million words and was processed using CAT-DongA in order to analyze word usage patterns. The top 3,000 key-nouns were extracted based on relative frequency and t-score.

Focusing on word usage frequency based on specific periods of time can serve as a good representative sample, but its limitations lies in the fact that such keywords are latched onto a corresponding timeframe. Therefore, this research has excluded highly time-sensitive keywords and processed keywords using word co-occurrence with high relative frequency that have appeared on a consistent basis over the entire period. In addition, the top 3,000 keywords were organized into a number of different themes, such as: regionalism, freedom vis-à-vis equality, hope vis-à-vis happiness, worker vis-à-vis laborer, and youth vis-à-vis elderly.[2]

3. MACROSCOPIC TRENDS: Analysis on Keywords and Co-Occurring Words

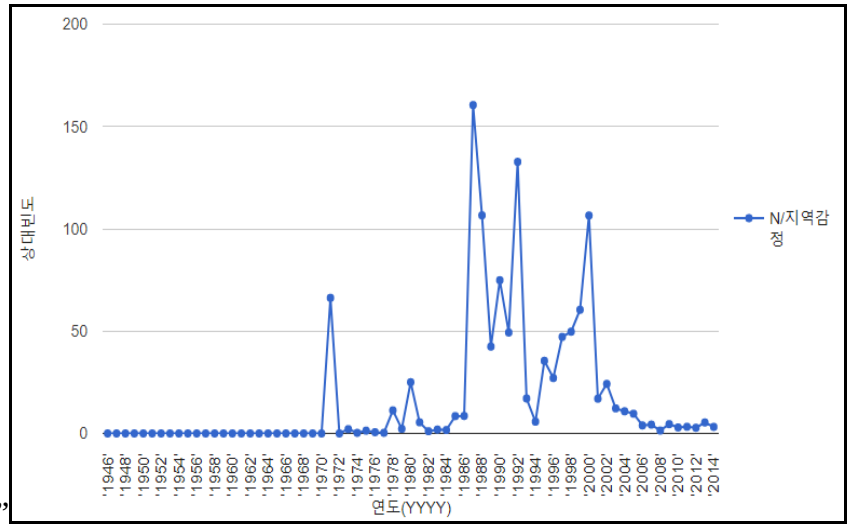
3.1 When Did "Regionalism" Begin in Korea?

Regionalism in Korea has been considered a social phenomena and issue, especially when it pertained to politics. Political preference based on one's home region is considered strong enough to cause a great deal of tension and conflict whenever an important election approaches. However, according to Figure 5, Korean society's understanding about regionalism was somewhat different from actuality.

Figure 5: Relative Word Frequency for

[2] The represented categories take on themes that mostly appeal to an ideology or relate to the populace. These are just a handful from the many organized theme-sets. In order to analyze macroscopic trends, a variety of keywords should be evaluated but this research focused on the ones that were specifically mentioned.

“Regionalism”



Referencing “regionalism” became quite frequent beginning in 1971. Before then, the word “regionalism” itself did not appear in the newspapers. Figure 5 traced sharp intermittent phases with regard to “regionalism” during the late 20th century and it decreased sharply after the 2000s. In order to understand how “regionalism” was used contextually, it was necessary to analyze co-occurrence words in detail.

Table 2: Top 20 Main Co-occurrence Words of “Regionalism”

W1	1970s	t-score	1980s	t-score	1990s	t-score	2000s	t-score	2010s	t-score
지역감정	선거	10.97	해소	23.01	해소	26.05	선거	16.56	선거	6.19
지역감정	후보	7.70	선거	20.08	선거	23.76	조장	15.52	후보	3.96
지역감정	선동	6.53	지역	14.68	지역	22.93	총선	13.64	정치	3.68
지역감정	표	6.03	특위	13.95	후보	21.91	대통령	13.51	조장	3.31
지역감정	P/공화당	5.59	후보	13.38	정치	20.11	정치	13.23	대선	3.22
지역감정	지방	5.56	P/광주	12.31	조장	15.63	지역	12.77	지역	3.19
지역감정	P/호남	5.22	대통령	11.88	P/김	14.89	발언	12.72	P/전라도	2.82
지역감정	해소	5.08	P/김	11.72	국민	14.52	후보	11.42	문제	2.69
지역감정	대통령	5.05	문제	11.38	P/호남	14.21	정치인	11.37	지방	2.65

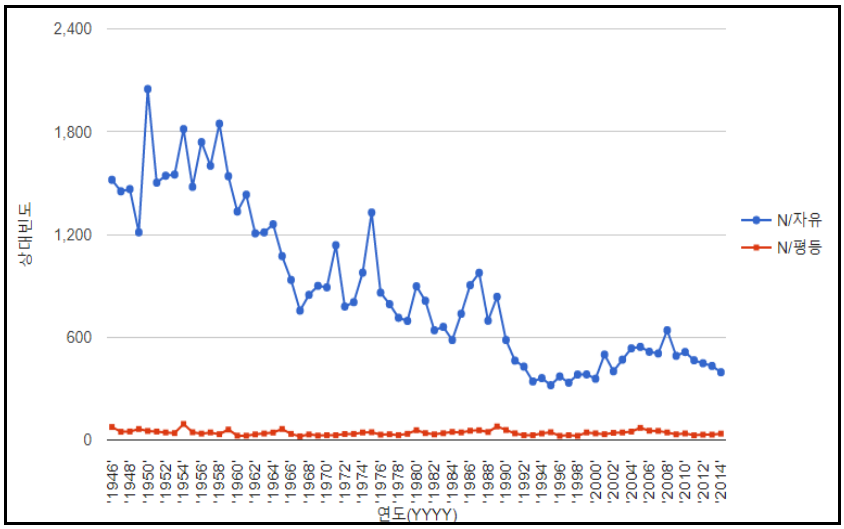
지역감정	유권자	5.01	총재	11.04	망국적	12.79	해소	9.87	P/경상도	2.64
지역감정	민족	4.91	유발	10.12	타파	12.75	P/김	9.82	P/호남	2.60
지역감정	발언	4.85	국민	9.44	발언	12.37	P/한나라당	9.74	국민	2.57
지역감정	유세	4.83	사태	9.32	정당	12.05	총재	9.52	갈등	2.47
지역감정	국민	4.75	정치	9.15	대통령	11.80	자극	9.49	P/김영삼	2.43
지역감정	호소	4.63	폭력	8.63	총재	11.78	정권	9.46	P/부산	2.42
지역감정	P/신민당	4.56	P/호남	8.33	P/김대중	11.59	연대	9.29	자극	2.39
지역감정	유발	4.51	출신	8.21	P/영호남	11.53	P/영남	9.28	기반	2.30
지역감정	정치인	4.38	P/김대중	7.81	P/영남	11.38	P/호남	9.09	정치적	2.29
지역감정	의장	4.35	방안	7.71	자극	11.32	국민	8.82	발언	2.28
지역감정	P/전라도	4.24	P/대구	7.63	주장	11.29	P/민주당	8.79	당선	2.28

Table 2 presented the main co-occurrence words based on the t-score, revealing that many of the co-words were related to politics; however, such relevant words decreased rapidly after the late 2000s. Based on the information provided, especially when newspaper articles debated about upcoming elections, it can be suggested that “regional politics” has embedded itself deeply into “regionalism”, serving as a tool for political incentives rather than simply regarding it as a chronic social issue.

3.2 Usage of “Freedom” and “Equality” Over Time

According to Figure 6, Koreans seemed to have more interest in “freedom” than “equality”.

Figure 6: Relative Word Frequency for “Freedom” and “Equality”



Referencing "freedom" was particularly high during the 1950s and 1960s. The frequent usage of "freedom" was held in juxtaposition to the "anti-communist" environment during those decades which was conjectured after conducting a co-occurrence network analysis.

Table 3: Top 20 Main Co-occurrence Words of "Freedom"

W1	1950s	t-score	1960s	t-score	1970s	t-score	1980s	t-score	1990s	t-score	2000s	t-score	2010s	t-score
자유	언론	66.92	P/중국	70.66	P/중국	103.90	언론	69.44	무역	47.74	언론	68.38	경제	52.56
자유	세계	57.72	언론	52.28	언론	67.41	노조	66.06	언론	44.59	경제	65.97	구역	50.61
자유	국가	54.11	세계	46.33	P/중공	41.46	민주주의	54.33	민주주의	37.95	구역	59.67	민주주의	38.59
자유	국민	52.44	국가	41.18	보장	40.60	P/중국	53.32	보장	35.91	민주주의	50.07	선진	33.99
자유	보장	52.08	P/한국	39.60	민주주의	39.14	보장	49.57	경제	35.43	무역	43.10	P/인천	30.90
자유	P/중국	49.15	보장	39.20	국민	36.27	P/폴란드	49.12	지대	31.31	표현	37.74	표현	29.17
자유	분위기	44.06	국민	38.42	국가	33.94	체제	40.54	국가	30.37	침해	34.98	언론	26.86
자유	진영	43.45	진영	33.17	P/미국	33.17	국민	37.78	체제	28.61	보장	34.87	무역	24.36

자유	민주주의	39.79	P/미국	32.74	P/일본	32.14	국가	34.20	표현	27.74	P/인천	34.37	헌법	22.01
자유	선거	39.45	민주주의	32.51	세계	30.26	민주	34.02	시장	26.99	헌법	34.20	보장	21.70
자유	경제	37.28	신문	28.15	P/한국	28.40	사회	32.28	민주	26.15	선진	33.73	계약	20.99
자유	P/한국	36.76	대북	27.86	대북	28.23	권리	31.90	경선	26.09	도시	29.36	P/북한	19.68
자유	권리	36.52	권리	27.84	말	28.03	표현	30.99	헌법	25.50	지정	28.28	인권	19.20
자유	평화	34.79	P/월남	27.80	P/아시아	27.87	무역	29.50	침해	24.59	국제	27.70	침해	18.23
자유	민주	34.62	경제	26.71	민주	27.84	정부	29.34	예금	24.53	권리	27.62	권리	17.65

According to Table 3, the co-word that drew in the most attention was “the media”, which was used consistently as a concurrence for each decade. Along a broader perspective, the high frequency of “freedom” was clearly visible in the political sphere, as well as extended out into the economical and other fields. On the one hand, the frequency of “equality” was lower than “freedom” which can infer that either there was less interest in “equality” or the ideology of “freedom” weighed more importantly among the populace. The changing trends of main co-words for “equality” over the decades can be seen in Table 4.

Table 4: Top 20 Main Co-occurrence Words of “Equality”

W1	1950s	t-score	1960s	t-score	1970s	t-score	1980s	t-score	1990s	t-score	2000s	t-score	2010s	t-score
평등	자유	20.35	자유	13.67	원칙	16.08	자유	23.62	여성	21.41	자유	22.30	양성	13.81
평등	사회	13.34	원칙	12.74	자유	14.11	사회	17.38	자유	21.11	양성	22.06	여성	13.33
평등	국민	11.89	호혜	10.81	관계	13.43	법	15.74	법	19.20	여성	20.59	교수	13.02
평등	인간	10.76	법	9.74	사회	12.94	여성	15.15	사회	18.50	교육	19.87	자유	12.76

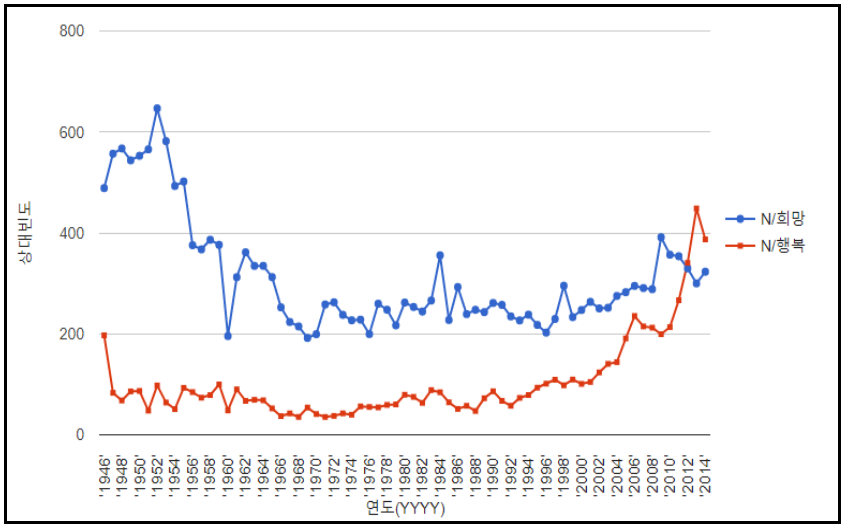
평등	권리	10.25	국민	9.59	호혜	12.59	원칙	13.86	남녀	18.07	사회	19.28	사회	11.32
평등	국가	10.18	국가	8.19	국가	12.21	인간	13.04	고용	17.42	남녀	15.16	교육	10.53
평등	인권	10.15	P/한	7.61	상호	11.20	남녀	12.27	원칙	14.21	원칙	13.48	성	9.91
평등	민주주의	10.01	권리	7.18	평화	10.78	권리	10.90	부부	12.49	정책	13.33	정의	9.63
평등	경제적	9.79	입각	6.91	여성	10.67	호혜	10.84	실현	11.05	성	12.88	기회	8.79
평등	정의	9.68	관계	6.87	존중	10.42	보장	10.68	차별	10.91	법	12.76	남녀	8.54
평등	원칙	9.18	민주주의	6.79	법	10.07	기회	10.58	민주주의	10.88	기회	12.25	원칙	8.21
평등	헌법	9.13	사회	6.65	주권	10.01	정의	10.37	인간	10.77	고용	12.00	법	8.04
평등	법률	8.56	P/일	6.63	입각	8.88	관계	10.21	헌법	10.42	인간	11.38	실현	7.73
평등	보장	8.41	상호	6.51	불간섭	8.58	국가	10.01	정의	10.38	헌법	11.27	정책	7.55
평등	법	7.87	만인	6.40	인간	8.50	민주주의	9.95	기회	9.95	실현	11.21	고용	7.50
평등	인류	7.56	양국	6.30	협력	7.95	이념	9.63	보장	9.23	가치	11.15	민주주의	7.28
평등	사회적	7.53	보장	6.25	권리	7.80	차별	9.55	관계	9.21	차별	11.13	공정	7.23
평등	정치적	7.44	존중	6.22	남녀	7.50	국민	9.41	만인	9.21	강조	10.79	추구	6.85
평등	선언	7.20	인간	6.17	민주주의	6.98	고용	9.08	권리	9.10	분배	10.69	가치	6.83
평등	만민	7.20	경제적	6.04	이념	6.96	평화	9.06	발전	9.04	민주주의	10.38	경제	6.61

In other words, the usage of “equality” during the 1950s and 60s was largely conceptualized through a macroscopic lens and did not necessarily relate itself to actual social circumstances. Beginning in the 1970s, “women” became a main co-occurrence to “equality” while “gender” appeared more frequently than “employment” as a co-word during the 1980s and 90s. Entering into the 2000s, “positive” became a close relative word to “equality”.

3.3 From “Hope” to “Happiness”

Relative frequency of “hope” and “happiness” revealed interesting results (see Fig. 7).

Figure 7: Relative Word Frequency for “Hope” and “Happiness”



According to Figure 7, as the frequent use of “hope” decreased over the years, “happiness” considerably increased beginning in 2000. “Hope” was mostly used within a future context while “happiness” was referenced in relation to the present, revealing a major shifting trend in society. During the 1950s and 60s, “peace” and “unification” were the main co-words for “hope” while “employment” increasingly became the main neighboring word beginning in the 1970 (see Table 5).

Table 5: Top 20 Main Co-occurrence Words of “Hope”

W1	1950s	t-score	1960s	t-score	1970s	t-score	1980s	t-score	1990s	t-score	2000s	t-score	2010s	t-score
희망	회담	27.36	말	18.70	말	18.90	P/한국	20.33	꿈	19.14	꿈	27.29	꿈	25.34

희망	P/한국	26.10	P/한국	17.35	P/미국	17.03	말	18.80	사항	18.10	국민	22.12	미래	21.75
희망	말	25.30	회담	16.67	P/중공	16.94	회담	18.04	국민	18.06	퇴직	21.20	버스	18.95
희망	P/미국	21.73	P/일본	14.43	평화	16.79	관계	17.98	P/북한	17.62	미래	19.88	연대	18.86
희망	P/유엔	19.68	국민	13.69	P/한국	15.86	P/중공	16.65	회담	17.03	신청	17.83	지원	17.65
희망	휴전	19.02	해결	13.52	관계	15.29	대통령	16.62	취업	16.73	절망	17.64	퇴직	17.58
희망	P/소련	18.11	대통령	13.39	회담	14.25	P/북한	15.34	참여	16.24	장래	17.35	사회	16.29
희망	P/일본	18.03	정부	13.26	대통령	13.26	방문	15.32	P/한국	16.02	메시지	16.75	장래	15.37
희망	평화	17.74	P/미국	12.76	P/한반도	12.64	본인	15.07	말	15.89	학생	16.59	홀씨	15.20
희망	송환	17.47	평화	11.85	기대	12.46	양국	14.90	미래	15.05	말	16.54	사업	14.93
희망	해결	16.95	양국	11.27	대화	12.30	취업	14.52	관계	14.03	삶	16.36	절망	14.83
희망	문제	16.88	협상	10.60	P/일본	12.17	국민	14.48	절망	13.90	사랑	16.29	대출	14.46
희망	대통령	16.11	문제	10.54	취업	11.90	도전	14.04	평화	13.63	근로	15.66	국민	14.24
희망	수상	14.78	관계	10.43	업체	11.19	회고록	13.20	장래	13.17	사람	15.10	행복	13.41
희망	포로	13.65	기대	10.18	진학	11.08	협력	13.07	신청	12.86	지원	15.03	참여	13.26
희망	양국	13.61	P/월남	9.92	방문	11.06	망명	12.92	진출	12.74	용기	14.82	일자리	13.19
희망	P/영국	13.45	P/일	9.50	P/소련	10.83	업체	12.64	용기	12.69	사항	14.53	취업	13.16
희망	정부	12.90	표명	9.06	가입	10.64	외교	12.41	정치	12.66	어린	14.42	서민	13.05

											이			
희망	표명	12.73	장래	8.93	P/남북한	10.06	기대	12.35	퇴직	12.53	돼지	14.30	청소년	12.79
희망	외상	12.71	본인	8.88	정부	9.96	평화	11.89	대통령	12.04	취업	14.07	메시지	12.70

Table 6 shows a stark change in the usage of “happiness” beginning in the 2000s.

Table 6: Top 20 Main Co-occurrence Words of “Happiness”

W1	1950s	t-score	1960s	t-score	1970s	t-score	1980s	t-score	1990s	t-score	2000s	t-score	2010s	t-score
행복	자유	13.85	사람	10.06	사람	11.40	사람	16.34	사랑	19.88	사람	29.92	국민	39.16
행복	국민	13.02	생활	9.91	생활	11.29	생활	14.67	사람	19.73	삶	25.60	사회	26.74
행복	생활	12.66	자유	9.33	인간	11.10	결혼	14.41	가정	17.95	세상	23.39	삶	24.39
행복	인류	11.20	인간	9.26	가정	10.45	사랑	14.01	결혼	17.49	사랑	22.42	고객	22.72
행복	인간	11.11	사랑	8.59	결혼	10.28	가정	13.69	삶	16.62	가족	20.77	사람	21.90
행복	평화	10.74	가정	8.29	불행	9.37	인간	11.89	생활	14.43	가정	18.70	가족	19.41
행복	사람	9.29	번영	8.24	사랑	9.05	삶	11.74	가족	13.78	아이	18.25	지수	18.95
행복	사회	9.10	국민	8.20	사회	8.60	추구	10.60	인간	13.77	일	17.84	기금	18.82
행복	생각	9.09	나라	7.80	생각	8.50	불행	10.55	불행	13.74	생각	16.40	사랑	17.71
행복	민족	8.96	생각	7.72	인생	8.25	자유	10.46	자신	13.48	결혼	16.22	공헌	17.66
행복	세계	8.67	여성	7.66	국민	7.76	사회	10.44	집	13.32	자신	16.07	기업	16.75
행복	가정	8.35	불행	7.54	마음	7.61	부부	9.94	부부	12.46	말	15.93	주택	16.50
행복	결혼	8.31	민족	6.88	자신	7.22	권리	9.69	남편	12.44	인생	15.87	경영	15.78
행복	사랑	8.20	사회	6.75	삶	7.17	생각	9.39	밤	12.12	인간	15.78	사회	15.72

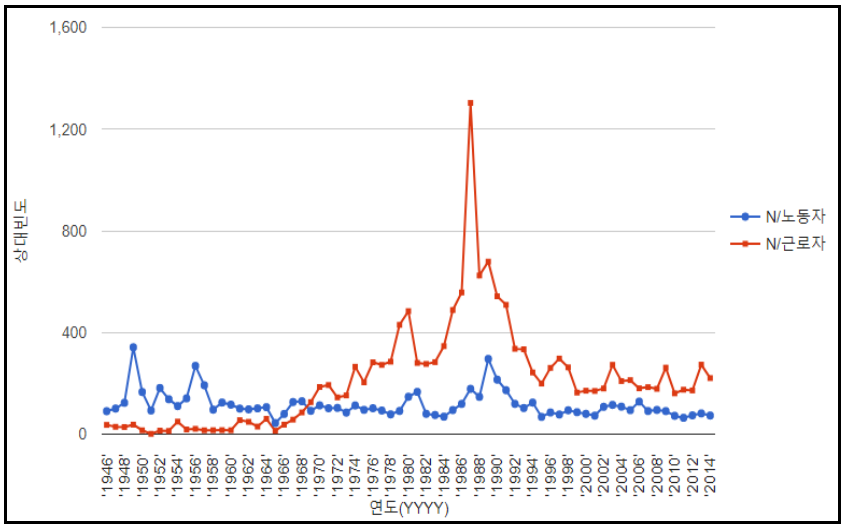
												적	
행복	길	8.15	결혼	6.66	남편	6.97	말	8.97	생각	11.80	사회	15.71	아이	15.65
행복	정치	8.12	말	6.47	추구	6.65	자신	8.95	아이	11.41	마음	15.70	대상	15.45
행복	나라	7.48	인류	6.36	아내	6.62	건강	8.75	인생		책	15.69	건강	15.41
행복	자신	7.41	길	6.26	자유	6.56	인생	8.73	세상	11.24	불행	15.64	세상	15.40
행복	여성	7.34	남편	6.19	부모	6.11	남편	8.73	추구	10.99	생활	15.28	가정	15.24
행복	인생	7.11	마음	5.96	개인	5.97	국민	8.64	마음	10.57	국민	15.08	말	15.16

One of the main co-words of “happiness” during the 2000s was “marriage” but it lost its spot among the top 20 in the following decade. As the correlation between “happiness” and “marriage” decreased over the years, other co-words like “family” and “customer” increased.

3.4 A Reason for Preferring One Over the Other: “Worker” and “Laborer”

According to Figure 8, “laborer” was used more regularly than “worker” up until the 1960s. Overall, both words embody similar meanings but were used a bit differently within certain contexts.

Figure 8: Relative Word Frequency for “Laborer” and “Worker”



The difference in usage patterns between “worker” and “laborer” became more apparent after

conducting a co-occurrence network analysis.

Table 7: Top 20 Main Co-occurrence Words of “Laborer”

W1	1950s	t-score	1960s	t-score	1970s	t-score	1980s	t-score	1990s	t-score	2000s	t-score	2010s	t-score
노동자	노동	21.20	노동	16.25	파업	22.01	파업	32.29	노동	29.82	외국인	37.74	외국인	19.99
노동자	파업	20.65	임금	16.12	임금	20.25	노조	29.88	외국인	24.48	노동	26.63	이주	17.16
노동자	농민	18.92	파업	16.11	노동	17.93	노동	27.91	임금	23.74	노조	24.26	임금	14.01
노동자	공장	17.67	농민	14.87	공장	16.40	P/폴란드	27.47	노조	23.10	임금	22.02	노동	13.19
노동자	임금	15.66	노조	12.77	노조	15.86	농민	23.65	파업	22.02	이주	20.89	공장	12.48
노동자	P/항가리	15.61	공장	11.73	농민	15.30	자유	22.28	여성	18.54	노총	20.42	감정	12.07
노동자	노조	14.53	학생	11.65	요구	12.41	투쟁	20.56	동맹	18.48	비정규직	19.44	노조	12.02
노동자	P/부다페스트	13.39	노임	10.10	인상	12.26	공장	20.13	사회주의	17.47	민주	18.80	일	11.41
노동자	시간	11.90	인상	8.98	데모	12.23	운동	19.84	농민	17.15	파업	17.07	해고	9.41
노동자	평의회	11.83	요구	8.94	학생	11.72	요구	18.51	투쟁	16.77	인권	16.73	인권	9.36

노동자	P/소련	11.69	총파업	8.68	부두	11.38	학생	18.31	계급	16.59	농민	16.43	다문화	9.35
노동자	노동조합	11.01	권익	8.57	정부	10.93	임금	18.19	공장	16.46	고용	16.22	비정규직	9.27
노동자	운동	10.89	P/프랑스	8.44	총파업	10.85	시위	17.59	단체	16.43	투쟁	15.91	농민	9.14
노동자	권익	10.72	복지	8.28	항의	10.36	민중	16.64	운동	15.55	공장	15.07	건설	8.92
노동자	노총	10.68	혁명	8.25	폭동	9.95	여성	16.32	전국	15.35	불법	14.44	출신	8.50
노동자	부두	10.66	노동조합	8.01	P/영국	9.78	P/바르샤바	15.70	해고	15.22	사회	13.82	민주	8.03
노동자	노임	10.54	전국	8.00	P/스페인	9.38	계급	15.48	정부	15.21	시위	13.46	고용	7.98
노동자	요구	10.53	생활	7.88	P/이란	9.16	결성	15.23	시위	14.82	일	12.94	일용직	7.90
노동자	강철	10.46	기업	7.84	혁명	9.14	연합	14.87	고용	14.36	주장	12.58	P/북한	7.69
노동자	P/대한노총	10.41	농업	7.80	권익	9.02	정부	14.61	사회	14.08	운동	12.54	가정	7.65

Table 8: Top 20 Main Co-occurrence Words of “Worker”

W1	1950s	t-score	1960s	t-score	1970s	t-score	1980s	t-score	1990s	t-score	2000s	t-score	2010s	t-score
근로	근로	12.14	임금	16.6	임금	49.83	임금	59.54	임금	53.76	임금	46.68	임금	30.84

자													
근로자	법	9.53	근로	16.3	근로	35.56	노조	51.03	저축	45.02	고용	42.76	고용	27.24
근로자	노동	8.59	노동	16.1	소득	31.64	회사	49.57	주택	41.60	외국인	42.38	외국인	24.80
근로자	보장	8.55	사업장	14.8	노동	29.59	노사	49.29	근로	41.24	비정규직	39.95	기업	24.33
근로자	기준	8.37	노조	13.6	기업	29.43	농성	48.65	고용	38.34	소득	34.09	시간	23.77
근로자	보건	7.63	법	12.5	노동청	27.35	분규	37.58	해고	35.26	근로	32.96	소득	23.72
근로자	노조	7.11	노동청	12.2	노조	27.19	요구	37.27	회사	34.72	기업	31.12	근로	23.58
근로자	사회부	7.04	기준	11.2	사업장	25.23	노동	36.39	기업	34.31	노조	28.11	비정규직	22.00
근로자	사용자	6.99	사용자	10.8	보호	24.95	기업	35.49	노조	34.18	노동	27.64	공단	20.57
근로자	권익	6.97	권익	10.8	복지	24.82	근로	35.36	노동부	33.22	사업장	26.94	근무	19.25
근로자	사업장	6.16	전국	10.5	법	24.46	노동부	35.00	노동	32.43	노동부	25.76	일	18.86
근로자	규정	5.95	보험	10.5	기준	23.33	인상	32.69	소득	31.96	해고	25.44	정규직	18.58

근로자	일	5.78	보건	10.5	노사	23.06	해고	32.52	노사	28.73	파견	25.13	퇴직	17.45
근로자	임금	5.74	소득	10.4	여성	22.70	파업	32.09	사업장	28.50	저축	24.38	시간제	17.28
근로자	생활	5.62	쟁의	10.1	이상	21.55	소득	31.07	외국인	28.43	공장	23.76	공장	17.20
근로자	권리	5.53	실시	10.1	저축	21.55	업체	30.33	복지	28.40	정규직	23.71	중소기업	17.00
근로자	실시	5.38	건강	10.0	보장	21.05	여성	30.03	장기	27.17	공단	22.82	업체	16.04
근로자	향상	5.32	진단	10.0	업체	20.79	공장	27.21	업체	26.30	기간	22.54	사업장	15.86
근로자	시간	5.31	재해	9.3	보험	20.78	사업장	26.73	사용자	25.16	근무	21.89	일자리	15.74
근로자	단체	5.30	가입	8.8	도시	20.60	조업	25.29	가입	24.88	이하	21.65	채용	15.56

According to Tables 7 and 8, “strike” was a co-occurring keyword for “laborer” while “wages” appeared frequently with “worker”. Word co-occurrence analysis illustrated how similar words like “worker” and “laborer” were used dissimilarly in certain contexts. For instance, “farmer” was closely associated with “laborer” and loosely affiliated with “worker”; however, the attached corresponding usage of “farmer” with “laborer” has weakened beginning in the 2000s.

3.5 Different Yet the Same: “Youth” and “Elderly”

There was a gap in the amount of use between “youth” and “elderly” up until the 2000s which decreased significantly since then (See Fig. 9).

Figure 9: Relative Word Frequency for “Youth” and “Elderly”

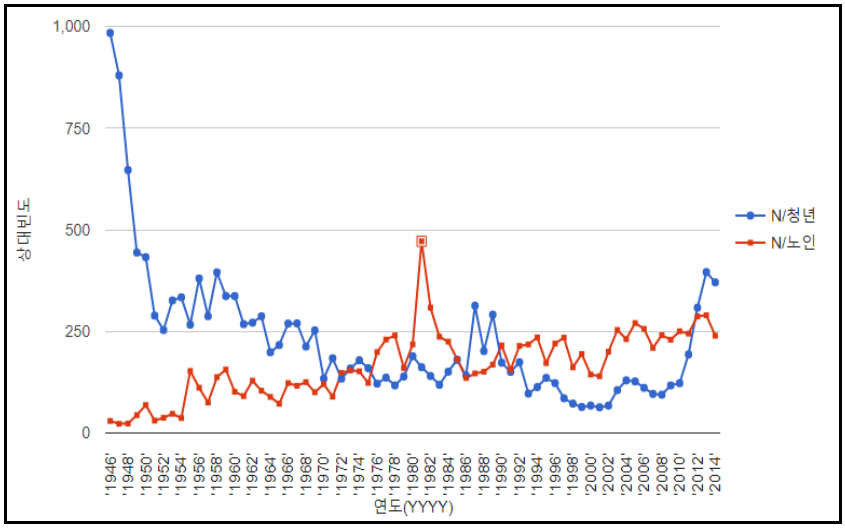


Table 9 and 10 explicates the top main co-occurrence words for “youth” and “elderly”

Table 9: Top 20 Main Co-occurrence Words of “Youth”

W1	1950s	t-score	1960s	t-score	1970s	t-score	1980s	t-score	1990s	t-score	2000s	t-score	2010s	t-score
청년	반공	33.01	경찰	20.90	P/서울	21.90	학생	31.57	학생	23.34	실업	33.04	일자리	41.82
청년	단체	20.40	반공	19.32	경찰	20.50	축전	21.27	조직	21.20	일자리	20.17	창업	41.56
청년	애국	20.04	당원	16.81	집	17.01	조직	20.72	운동	18.34	실업자	19.58	취업	37.99
청년	일	18.69	집	15.68	현금	15.78	운동	20.51	단체	16.24	실업률	19.27	센터	31.41
청년	시내	17.46	조직	15.36	당원	14.76	경찰	19.99	경찰	15.16	취업	16.24	캠프	30.55
청년	경찰	17.27	일	14.02	운전사	13.62	당원	19.73	통일	15.03	고용	14.77	실업	28.24
청년	청년단	15.22	학생	13.59	부인	13.60	협의회	18.74	대학생	13.96	해소	14.49	고용	26.07
청년	집	14.68	운동	12.51	범인	11.92	대학생	18.33	당원	13.88	채용	13.14	지원	25.87
청년	석방	14.27	회원	10.80	위협	11.57	P/평양	17.75	사랑	13.69	시절	12.81	대학	23.87
청년	운동	13.34	봉사회	10.71	학생	11.41	P/서울	17.19	위원장	13.49	사회	11.09	창출	21.01

청년	정체불명	13.17	연행	10.40	차림	11.37	단체	16.87	여성	13.30	단체	10.91	구직자	20.93
청년	군복	13.04	전기	10.20	택시	11.22	후보	16.18	동맹	13.01	대책	10.87	채용	19.38
청년	조직	12.93	마을	10.19	조직	11.20	작가	14.63	연합	12.51	연합	10.85	기업	19.18
청년	학생	12.52	단원	10.11	반공	10.88	연합	14.05	단장	12.38	운동	10.77	실업률	17.97
청년	번지	11.99	시내	10.00	칼	10.25	회원	13.90	시절	12.28	작가	10.64	중소기업	17.77
청년	전기	11.89	범인	9.96	돈	10.14	P/민정당	13.64	협의회	11.95	학생	10.47	멘토링	16.69
청년	사건	11.76	학도	9.96	동맹	9.93	민주	13.51	지구당	11.90	인턴	10.43	위원회	14.68
청년	구타	11.45	사건	9.81	신고	9.68	기독	13.48	활동	11.55	대학생	10.43	공공	14.66
청년	밤	11.25	단체	9.80	차	9.60	P/북한	12.65	작가	11.30	창업	9.86	해소	14.46
청년	P/서울	11.00	부락	9.72	회관	9.56	시위	12.39	회원	10.90	창출	9.78	세대	13.80

Table 10: Top 20 Main Co-occurrence Words of “Elderly”

W1	1950s	t-score	1960s	t-score	1970s	t-score	1980s	t-score	1990s	t-score	2000s	t-score	2010s	t-score
노인	집	13.31	집	12.91	복지	21.73	복지	32.38	복지	44.29	복지	48.95	복지	34.43
노인	사람	12.29	아들	12.57	사회	19.40	문제	30.56	시설	32.17	시설	37.28	홀몸	30.59
노인	일	10.76	여인	11.01	양로원	17.92	사회	28.27	사회	28.05	사회	31.67	시설	25.19
노인	파출소	9.77	경찰	10.11	집	16.51	양로원	21.84	치매	24.78	장애인	30.30	요양	25.02
노인	시내	9.14	사람	9.33	마을	15.34	시설	21.56	무료	24.16	치매	29.41	장애인	24.46
노인	인계	9.03	마을	9.23	학교	14.64	경로	21.04	생활	23.73	건강	29.37	사회	24.10
노인	아들	8.37	어린이	8.86	어린이	14.03	생활	20.08	대상	22.75	인구	28.23	인구	23.91

노인	전기	7.87	시골	8.58	P/서울	13.91	집	19.73	무의탁	22.31	요양	27.10	일자리	21.32
노인	몸	7.80	방	8.43	사람	13.73	인구	19.53	가정	21.81	봉사	26.17	봉사	20.96
노인	돈	7.67	바다	8.40	문제	13.66	가정	18.24	보호	21.62	일자리	24.74	건강	20.95
노인	가족	7.31	P/서울	8.23	가정	13.37	젊은이	17.27	가족	21.39	병원	23.81	연금	20.94
노인	길	7.24	밤	7.46	할머니	13.35	가족	16.38	건강	21.17	무료	23.61	치매	20.02
노인	술	7.04	병원	7.22	생활	13.26	자녀	16.37	장애인	21.03	복지관	23.56	센터	19.17
노인	병원	6.95	돈	7.10	가족	12.00	건강	15.54	병원	20.79	센터	23.38	복지관	18.74
노인	양로원	6.89	딸	7.00	건강	11.61	우대	15.47	문제	20.67	활동	22.84	지원	18.28
노인	경찰	6.82	시내	6.73	불우	11.59	일	15.24	복지관	20.22	생활	22.66	기초	17.96
노인	동정	6.74	자살	6.60	보호	11.57	수용	15.15	인구	19.42	대상	21.88	활동	17.64
노인	사정	6.73	칠순	6.52	시설	11.48	할머니	15.04	어린이	18.01	전문	21.12	병원	17.04
노인	거주	6.70	청년	6.49	할아버지	11.08	어린이	14.68	양로원	16.99	혼자	19.89	생활	16.56
노인	이야기	6.48	양로원	6.43	자녀	10.77	무료	14.51	진료	16.83	보험	18.92	보험	16.09

In particular, the frequency of “youth” was high during the 1940s and 50s, which can better reflect the situation related to that period. For instance, “youth” was closely associated with words like “anti-communism”, “student movement”, and “unemployment”. On the other hand, “welfare” became closely affiliated with “elderly” beginning in the 1970s.

In addition, Table 10 exemplifies that the once intimate relationship between “elderly” and “family” has continually weakened since the 2000s. As the concurrence rate decreased between these two words, one may posit that the issue regarding Korea’s rapid aging society has emerged out of the familial and into the wider social realm.

4. CONCLUSION

The purpose of this paper was to highlight how the *Trends 21* corpus project culled in 70 years' worth of Korean newspaper big data as well as the tools that were developed, such as the Web-based Analysis Tool (CAT-DongA), in order to process and analyze macroscopic trends which can provide a clearer glimpse into Korean culture and society.

This research demonstrated how utilizing big data newspapers, which spans over a period of time, produced a wealth of information that can contribute to a number of studies in the Humanities. If the *Trends 21* corpus can continue devloping language resources, this research project can unceasingly serve as a conduit for developing new and more exciting research.

Works Cited

김일환, 강진웅, 이도길, 배석만 [Kim Ilhwan, Kang Jin-woong, Lee Do-gil, Bae Seok-man]. *키워드, 공기어, 그리고 네트워크 - 신문 빅데이터가 보여주는 것* [*Insights into Keywords, Co-words and Networks*]. Seoul: 소명출판 Somyong Publishing (2017).

김일환, 이도길, 강범모 [Kim Il Hwan, Lee Do Gil, Kang Beom Mo]. "공기 관계 네트워크를 이용한 감정명사의 사용 양상 분석" ["A Study of Emotion Nouns Based on Co-

occurrence Relation Networks"]. *한국어학* [*Korean Linguistics*], vol. 49, 2010, pp. 119-48.

——. "SJ-RIKS Corpus: 세종 형태의미 분석 코퍼스를 넘어서" ["SJ-RIKS Corpus: Beyond 21st Sejong Morph-Sense Tagged Corpus"]. *민족문화연구* [*Korean Cultural Studies*], vol. 52, 2010, pp. 373-403.

김일환 , 정유진, 강범모, 김흥규 [Kim Ilhwan, Chung Eugene, Kang Beom-mo, Kim Heung kyu]. *물결 21 코퍼스의 구축과 활용* [*Handbook of 'Trend 21' Corpus*]. Seoul: 소명출판 Somyong Publishing (2013).

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12 18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

近代韓國「文化」概念的軌跡

宋寅在
翰林大学翰林科学院教授

近代韓國「文化」概念的軌跡

宋寅在
教授
翰林大学翰林科学院

摘要

本文運用近代韓國報刊語料庫宏觀透視「文化」概念的軌跡。近代韓國報刊語料庫收錄從 19 世紀末到 20 世紀 30 年代出版的重要報紙、期刊、學會報的正文, 因此研究者可以根據分析語料庫結果成爲瞭解朝鮮末期到殖民時期韓國文化概念的宏觀趨勢。由於政治不穩定的形式，近代韓國報刊情況也不穩定，特別是，以 20 世紀 10 年殖民化、20 世紀 19 年 3.1 運動為分歧，韓國的言論出版狀況也隨之經歷過變局。除了歷史背景意外，詞匯使用以及語境也表現過幾次變局。近代韓國報刊語料庫可以顯示這樣的語言上的變遷趨勢。本文將運用語料庫，分析「文化」概念的年度次數、共現詞趨勢、元數據信息等，闡明 20 世紀初韓國「文化」概念的傳入、詞匯變化、歷史作用。

雖然，「文化」這一詞在 19 世紀末文獻也出現，該詞語還在 20 年代初被看作一種新術語，并且「文化」頻率在 20 年代初驚爆地上漲，這令人推測在近代韓國「文化」這一詞還算是新術語，1920 年成爲一種轉折點。本文根據這種假設仔細地觀察整個 20 世紀初文化概念的語義、語境變化趨勢。大體上，1910 年之前「文化」主要與發達、增進、程度等詞語在一起出現，成爲跟發展有關的術語。20 世紀 10 年代，一些留日學生把文化看作跟學習、教養、藝術有關的詞語，其發展典範、比較對象是日本、世界。1919 年 3.1 運動結束以後，在韓國文化主義運動進行得很活躍，隨之，「文化」這一次的使用次數也有非常大的增長，界限「文化」的種種論述還頻繁地出現。這一時期「文化」在建設朝鮮文化建設，與中國、世界文化比較的語境下出現了。另外，從 1920 年後期開始，大衆文化的潮流逐漸地成長了，還影響的「文化」這一詞的內涵。這是日常生活、娛樂、休閑方面的內涵比以前突出了。本文將基本上根據預料庫分析，提出上述的「文化」概念變遷，最終顯示近代韓國文化概念的宏觀趨勢。

目次

1. 引言

2. 「文化」的年度頻率趨勢與其語言語境

3. 「文化」的變遷趨勢

3.1. 自内之外的標準轉移

3.2. 作爲政治力學的外物

3.3. 改造與建設的文化

3.4. 穩定與狹隘化

4.結語

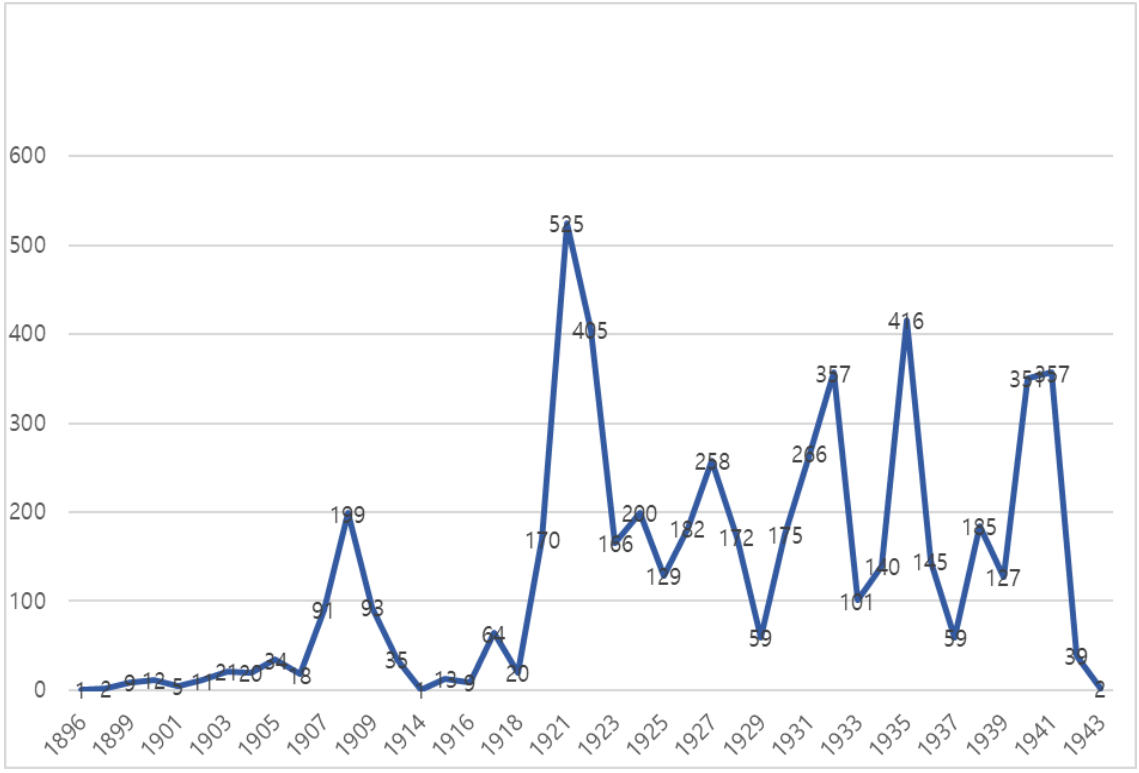
關鍵詞

文化，政治，運動，藝術，生活

1.引言

本文討論韓國傳統社會崩壞的時期到近代轉型期「文化」概念的變化軌跡。當今在歷史敘述上，「文化」一詞被使用於整個時期人類的生活、歷史遺產。可是「文化」這一詞的命運不那麼簡單。現代常用的文化體現多面性，并且「文化」本身不是從中古時期帶著普遍性被使用的，而是在近代形成的概念。比如說，英國的著名文化研究著作《文化與社會》表示，工業革命以後「文化」才成爲指稱社會生活全盤的術語。在這種認識下，本文試圖宏觀透視從 19 世紀末到 20 世紀前半期在韓國「文化」概念的變遷。爲此，本文運用以該時期在韓國發行的期刊、學會報正文構成的《近代韓國報刊語料庫》。這件語料庫收錄在 1896 年到 1942 年出版的 20 種期刊、學會報，1900 年代發行的 1 種報紙。[1] 本論通過分析關鍵詞年度頻率與共現關係，解讀重要正文，試圖宏觀描述「文化」的語義變遷、脈絡以及其內涵。

2.「文化」的年度頻率趨勢與其語言語境



[1] 《近代韓國報刊語料庫》是由翰林大學建置的，當初以 19 九種期刊正文構成還稱之謂《19 種語料庫》或者《翰林概念史語料庫》。現在因爲收錄資料繼續擴充，還包括一部分報紙，需要新名稱，所以臨時使用《近代韓國報刊語料庫》。

图 1。1896-1942 年文化的年度使用次数

本章根據語料庫統計觀察從 19 世紀末到 1940 年代初「文化」概念的出現頻率趨勢，檢出各個時期「文化」的主要共現詞，説明近代韓國「文化」概念的語言語境。這一些分析結果為透視近代韓國「文化」概念的宏觀趨勢提供資料基礎。

首先看從 1890 年代末到 1940 年初期刊、報紙上「文化」的使用次數。〈圖一〉顯示「文化」概念的年度使用次數。在整個時期「文化」一共出現 5644 次。年度使用次數幫助我們劃分幾個時段觀察「文化」的宏觀趨勢。第一、1920 年以前「文化」的使用次數除了 1908 年以外都不到 100 次，使用次數比較少。并且當時期刊上出現的「文化」還是地名，不是概念，因此作爲概念的「文化」的使用次數會更減少。第二、1910 年與 1920 年之間的次數微微了了，這樣的結果跟當時歷史背景有關，1910 年日本的殖民統治正式開始。歷史狀況還影響到出版、言論環境，日本政府全面禁止在韓國使用韓文，隨之韓文資料也全面萎縮了。1919 年 3.1 運動爆發之後日本政府的殖民政策轉換從武斷統治到文化政治，從此才可以重新進行韓文言論活動。因此預料庫只能收錄當時留日韓國學生發行的期刊《學之光》，所以，本文統計上的 1914 年以後 1920 年之前的數據反應在日本的韓國青年們有關文化的論述。第三、在整個統計數據當中，「文化」使用次數的高峰在 1920 年代，特別是 1921 年的 525 次，從了 1929 年下降到 99 次以外其他年度的使用次數都是一百次以上。這裏還有韓國的言論出版狀況背景。20 年代韓國宗教集團發行了《開闢》雜志進行了新文化建設運動 雜志在 1920 年創刊 1926 年停刊了。所以 1920-1926 年有關「文化」的論述大多是在《開闢》雜志進行的。據此，我們可以以 1926 年為境界劃分時段。第四，1926 年《開闢》停刊以後雖然「文化」的使用次數不到高峰，可是使用次數沒有萎縮。1932、1935 年與 1941 年記錄了 300 次以上的使用次數，達到了了第 2、3 次高峰。據此，我們估計這一些年度是「文化」概念演變的關鍵時期。本文將分析例句、共現關係的變化，解讀「文化」在什麼樣麼歷史及語言語境下，體現什麼樣的意義變遷。

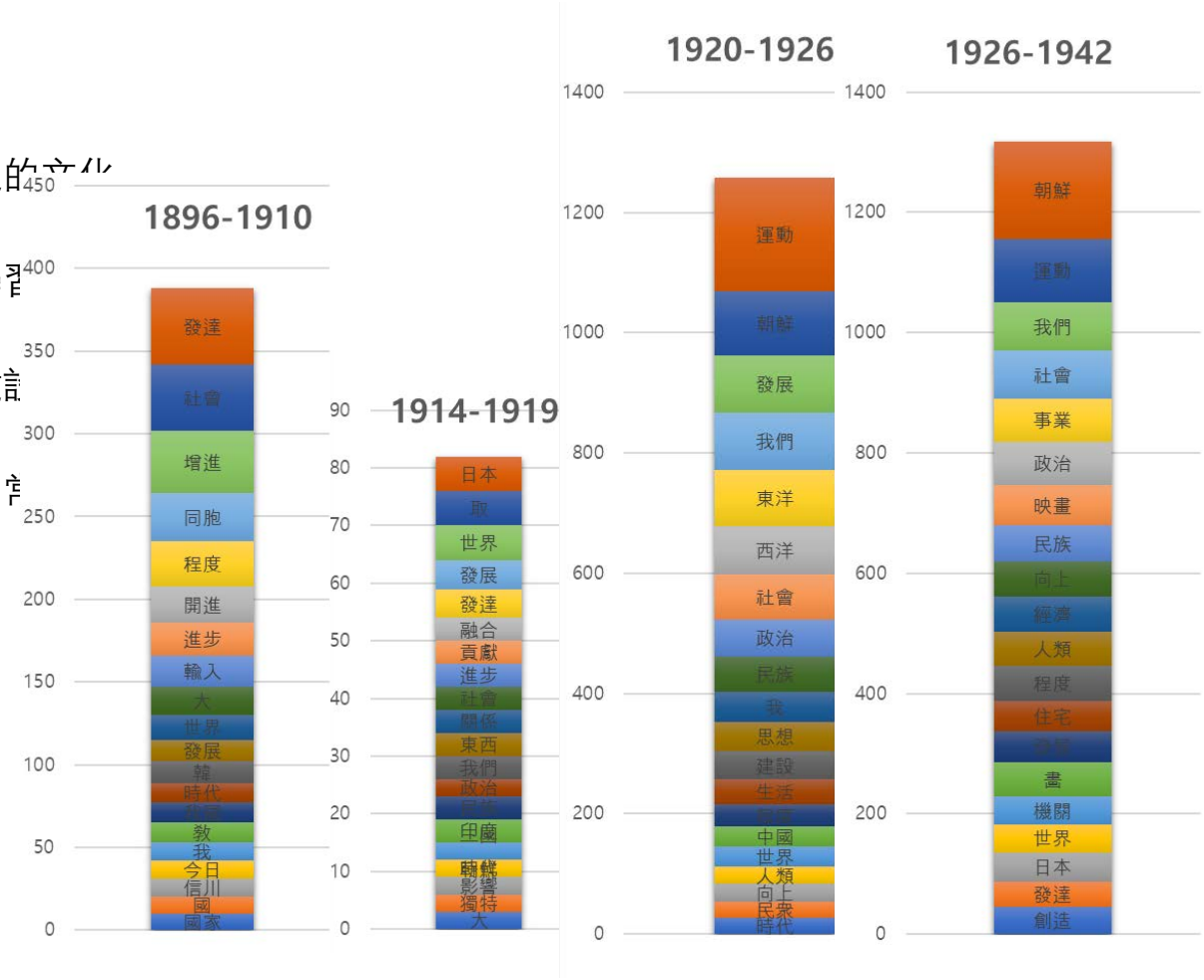
接下來看「文化」的共現關係演變趨勢。本文根據期刊發行日子劃分年代分開分析了「文化」的共現關係。如上述的三個年度劃分 4 個時期，第一、1910 年以前學會報、《皇城新聞》，第二、1910 年代朝鮮留日學生雜志《學之光》，第三、1920-1926 年《開闢》、《學之光》、第四、1926-1942 年《別乾坤》、《三千里》、《東光》、《萬國夫人》。各個時期「文化」的共現關係表現如《圖二》。

3.退的文化

3.學習

4.建議

5.日常

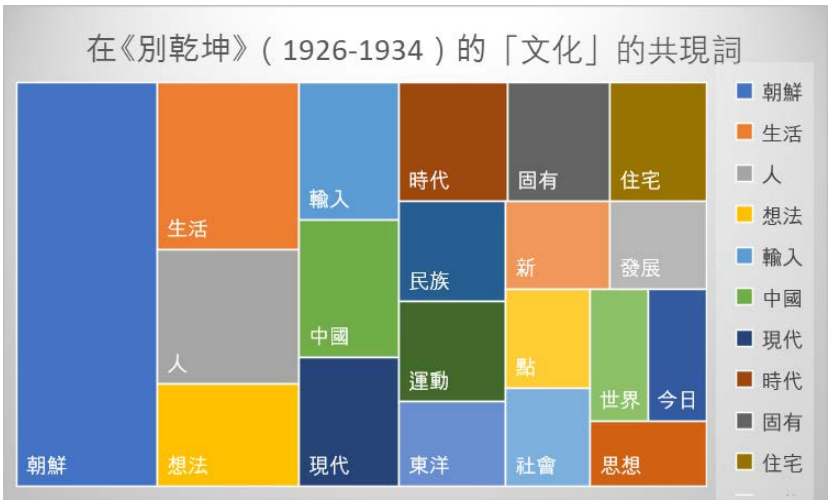


《圖二》顯示出與「文化」共現次數排名第 1-20 位的共現詞。從第一時期到第四時期上位 20 個共現次的出現頻率各自 388 次、82 次、1258 次、1318 次。雖然單純的數據上第四時期的共現關係出現次數最多，可是各個時期期間不一樣，因此如果算出年比共現次數的話，各自一年 22.8 次、13.6 次、179.7 次、77.5 次。這一結果告訴我 1920-1926 年有關「文化」的論述最活躍。在共現關係的持續性方面，「發展」、「社會」、「世界」在整個時期出現為上位 20 位共現詞，表示持續性。在三個時期出現的共現詞是「民族」、「發達」、「朝鮮」、「我們」、「程度」、「政治」。其中「我們」在第一時期沒出現代之「我」出現，因此還算是整個時期出現的共現詞，并且這一些此實際上指示朝鮮，因此構成同一個詞匯群。「發達」也在韓文的語境下是「發展」的類似詞，所以有關發展或發達的論述也跟文化的關係密切。進而説的話，「程度」跟發展、發達會形成一種搭配關係。這一些共現詞構成幾個意義群，「發展」、「發達」、「程度」等跟某一個方面的演變與其狀態有關，「朝鮮」、「我們」、「世界」、「民族」都有關在全球化階段表示某一種因素的單位，此外，「政治」、「社會」是在人類生活中跟「文化」區別還是關聯的方面。這一些信息，推斷出一種假設，就是近代韓國「文化」大體上圍繞

演變、狀態、主題、單位、有關領域進行。更正確而細緻的討論與檢證會通過更微觀的觀察和分析進行。

共現的分期特點也根據《圖二》掌握。第一時期，既是從 1896 年到 1910 年，文化的主要共現詞當中發達、增進、程度、改進、進步、發展等有關演變的詞語比較突出，其次同胞、世界、韓、我國、國、國家等有關單位的詞語還不少出現。這樣、舊韓末時期文化論述的主要論點是圍繞發展、國家、世界形成了。第二時期，既是殖民時期最初十年的主要共現詞也是跟單位、演變、狀態有密切關係。有趣的是「日本」這一詞位於雖然第一名從第三名到七名都是「發展」之類的詞語，但是日本世界在前列，東西、我們、民族、印度等有關國家、民族單位在後列。這一結果表示，日本發行的雜志反應對於日本這樣的空間背景、其他國家、民族、世界的興趣。第三時期，即是 31 運動結束以《開闢》進行朝鮮新文化建設運動的年代，運動一詞名列前茅。接下來朝鮮、我們、東洋、西洋、民族等幾個單位接著運動顯示比較高的共現次數。并且發展、發達、建設、程度等有關演變、狀態的詞語被運動壓倒。這樣的結果告訴我們 1920 年《開闢》創刊之後，運動成爲了文化論述的新語境，於此同時東方、西方這樣的世界區劃方式還在文化論述同行了。這是 20 年代前半期的文化論述跟前兩個時期形成差異性的緣故。第四時期，雖然運動、朝鮮等詞語的共現次數站著上位，但是性格跟以前不同的共現詞的出現令人矚目，比如事業、映畫、住宅、畫等。這一些詞是屬於藝術、生活方面，跟前時期的有關單位、運動的共現詞在性格上截然不同。這一結果提供我們 1926 年以後有關文化的論述在新一方面展開了。

爲了補充還提到另一個數據，《別乾坤》雜志上「文化」的共現詞，這是 1926 年開闢社停刊《開闢》之後創刊的期刊。〈圖三〉顯示比較明顯的共現詞變化。運動的共現次數大幅下降，代之生活、現代、輸入等出現為新共現詞。



《別乾坤》的共現詞情況提供切入 1926 年以後文化論述變化的明顯綫索。

下一章根據本章檢出來的數據，探討在各個時期的文化論述，試圖揭示「文化」的語義變遷以及其特點，相互關係。

3. 「文化」的變遷趨勢

3.1 自內之外的標準轉移

1896 年《大朝鮮獨立協會會報》〈會報本旨〉說，「會報本旨吾人培養忠孝的根本發達職分的事業，故有無相資交換各各聞見及知識，實力上講究自國、自己的本來面目、內外古今的利害損益，進步文化仁壽之域。」[2]這一例句使用的文化意味著一種交換見聞知識、講究真面目以後大達到大的目的。這樣的文化跟知識培養，國家成員具有一定的力量有關，有一定的距離與莫一群人類社會形成的生活方式方面的意義。〈會報本旨〉所提出的「文化」還在前現代意義上的文化，即「人文教化」的意義上。作爲過度時期 1890 年代至 1910 年「文化」一詞在語義上這樣混用傳統「人文教化」與莫一群人類社會固有的生活方式的意義。3 比如，1907、1908 年的記事上「文化」以「教育」的連詞出現，在這樣對脈絡上所謂文化教育的進步是人類的面目，文化教育是人生的目的，文化教育的終極目的是養成具有理性的人生。因此擇定教師，依章程教育促成文化的漸進。這樣傳統意義上的文化，帶著實行教育的有效結果的意義出現。

在古今東西的事業下，比較各國的情況的時候，「文化」帶著各國固有的生活方式及其發展程度。進而說，西方的文化是比東方發達，東方國家應該接受西方文化，從西方文化學習。發達、發展、開進之類的論述大多反映文化等級論、文化比較論的事業。當然這樣的語義適應與描述古代韓國各個國家的文化、周圍國家的文化。這樣的比較視野越近於 1910 年越突出，這還説明當時朝鮮知識分子具有强烈意識，他們要發達自過的文化。傳統意義的文化跟教化、人文密切，這是人文的發達程度有關的。作爲普遍的人類理性、智力的文化，在東西方勢力不均衡的時局開始國家、區域的脈絡。因此發達、進步文化的緣故不再是文化本身的水平，而是高發達國家與相對低級發達的國家之間的差距。當時朝鮮是在世界秩序的弱勢國家，比自己發達的的文化成爲應該學習、輸入的典範。1896 年到 1910 年「文化」的最大語義轉型就判斷、評價的標準從本身的内部轉移到比自己優越的他者。

[2] 〈會報本旨〉,《大朝鮮獨立協會會報》1896 年 12 月。
[3] 另外，文化還是郡級地方的名稱，文化郡、文化郡守等詞語還出現了，可是這只是標記上的同一性，不是語義變遷上的問題。因此本文不涉及地名意義上的文化。

3.2. 作爲政治力學的外物

第二節討論留日學生發行的雜志《學之光》上出現的「文化」的內涵。第二章已經看過這一時期文化大多是在國家之間比較事業進行的，其中「日本」的共現次數最高。加之，「文化」的使用次數最多的文章是李光洙（1892-1950）所寫的〈我們的理想〉，本節特別關注李光洙的文化論述。在文章的開頭李光洙特別提出一種提法，「世界文化上朝鮮族的位置」。李光洙提到文化的範疇問題，他區別文化與政治，表示自己敘述朝鮮的歷史位置的切入點，還提到蒙古帝國、古希臘作爲歷史上在文化、政治上的不平衡的典型。相比之下，羅馬是政治、文化上都占有優越的地位。接下來李光洙提出，文化上的優越比政治上的優越更有價值。其例證是中國與印度對世界的文化貢獻，印刷術、火藥、宗教、哲學等。因此李光洙主張朝鮮應該在文化上做出貢獻，這是世界文化還沒達到絕頂，還有發達、成熟的餘地。据他說，具體來說，物理、化學等自然科學還需要發展，以代議制為代表的政治制度還不完善，此外勞資問題、城市問題、農村問題、男女問題等經濟上的諸如問題要求改革者、創造者、思索者、實行者。文學、藝術更有餘地朝鮮人可以做出貢獻。最終，李光洙朝鮮人的出路在於創建新文化。

李光洙論述的突出點是提出文化的範問題。這樣的提法提供朝鮮人在現實政治力學的絕對劣勢找到生命力、恢復自尊的途徑。其方法是對於世界文化的貢獻，其實踐戰略是介入世界上還需要解決的問題來積極提出答案，牽引世界的發展進步。雖然李光洙還提出政治制度改革，這不算是混淆政治與文化的範疇，更有本質性的是，從現實關係上的優劣獲得獨立性，提高自己的能力，發揮積極作用。這樣活動的主要途徑是學術，因爲實現這種戰略，需要深層研究各個部門的核心課題。總之，李光洙所提出的文化離現實政治力學保持距離，還包括人類社會的諸如領域，通過精通的學術能力實現。

3.3.改造與建設的文化

第三節探討 1920-1926 年《開闢》雜志的文化論述。第二章已經看過這一時期文化使用次數的高峰出現了，文化的年均使用次數也最高，這一些數據都是由於《開闢》雜志進行朝鮮新文化建設運動。《開闢》雜志是天道教領導李敦化在 1920 年創刊的，其發行処是天道教青年會。天道教青年會在 1919 年組織了，其宗旨是研究宣傳天道教教理，發展朝鮮文化。《開闢》的創刊目的是「注視世界改造的動態，改造朝鮮社會。」，這樣《開闢》成爲了文化運動的基地。《開闢》對於文化的關注跟日本的潮流有關，當時

在日本文化主義流行，李敦化在内的《開闢》人士收到了日本的影響。因此，他們的論述上出現一句話，「正在文化流行。」加之，文化的種種界定還在《開闢》的文化論述。其中，有人指出文化是 culture，kultur 的翻譯詞，還説文化有教化、教養、德育等詞典性信息。另一些論述還界定文的範疇、主體。比如，「在人類界不息進化的所有的流動狀態。」，「包括物質、精神雙方面」「共同璞碧娜理想整合體的表現」，「一個民族的種種生活方式」，「在歷史上，與政治史區別的，藝術、風俗、學術等一般人文的進步」最後一句是在李光洙的文化界定的連續綫上的，這樣，告訴我們當時日本知識界對於朝鮮知識界的影響。

《開闢》文化論述的另一個特點是中國的文化論述的影響。北旅東谷（李東谷）投稿的几篇文章是最典型的。特別是，〈批判東西文化談論我們的文化〉是《開闢》記事當中文化的使用次數最多，主要在民族文化、東西文化論的視野切入文化運動。據他説，20 年代朝鮮的文化運動是繼承 31 運動的民族運動，新文化運動是一種民族改造運動，進而説民族改造就是文化建設，文化建設就是文化的復興。這樣的論述上運動、建設、民族、改造等當時「文化」的核心關聯詞都出現。還有令人矚目的是李東谷的文化論述與文化概念界定大大參照了陳獨秀的文化論述，比如説，〈東西民族根本思想的差異〉。李東谷在陳獨秀的影響下全面否定東方文化，於此相反全面承認西方文化的積極性。除了李東谷以外，李敦化、金起田（1894-1948）等還指責以儒家為代表的朝鮮舊文化，加它是朝鮮的根本弊端。

這樣的文化認識還成爲進行新文化建設運動的基本思想。有一個人提出，東學（天道教的前身）的創始精神本身就是融合改造東方的哲學，建立東方的新文化。李敦化也把東學的創始人評價爲打到痼疾，開拓新思想的人物。在這一些思想基礎，李敦化積極推進了新文化運動，希望創造所謂「少年朝鮮」。李敦化提出過新文化運動的六個實踐方案，即知識熱, 教育普及, 農村改良, 都市中心主義, 專門家, 思想統一。實現這一些方案的核心途徑是利用新聞、學校進行教育、傳播新思想。

《開闢》筆者的文化論述的最大特點是運動。他們覺得，改變文化是改造社會的中心環節，改造世界觀有日本思想潮流的影響，親西方文化的文化變革論是受到中國新文化論的影響。進而説，《開闢》集團的文化運動雖然帶著社會改造運動，他們的戰略還在通過出版、教育的啓蒙運動。這樣的格局告訴我們；雖然到 1920 年代「文化」的内涵有變話，「文化」負責人的知識中心性認識以及其啓蒙教化式態度在大範圍上還在持續下來了。

3.4.穩定與狹隘化

1926 年以後「文化」的基本界定，即作爲民族的種種生活方式，已經廣汎地接受、使用了。進而，這一時期出現了「文化」的語義擴展，在另外一的角度來看，一種狹隘的「文化」概念出現了。如果廣義的文化指出整體某一個民族的生活方式，狹義的「文化」的範圍相對縮小了，既是特稱更藝術、教養有關部門。比如說，1926 年以後出現了「文化事業」這不能意味社會成員的生活方式本身，而是指稱其一部分的事業，特別是經濟、政治領域以外，跟藝術有關的事業。還有獨特的新語叫「文化住宅」、「文化映畫」這一些詞指稱人類日常生活、藝術的特點，即水平比較高而且跟教養、高雅親密的。這一些例子。這一時期「文化」的語義在整個時期最豐富，廣義、狹義的意思的出現了。還有事例，文化作爲一種未發達或野蠻狀態的反義詞。在這樣的格局下，文化一邊構成歷史社會性宏觀論述的構成部分，一邊跟藝術有關的特定部分，此外，以知識、教育讓人類生活擺脫一種野蠻、未發達狀態的傳統意識還在持續。這樣，近代韓國「文化」概念最終體現多面性。

4. 結語

從 19 世紀末韓國經歷過激烈的變局。隨之，人的思想觀念發生了變化，而且表示思想的概念也有變化。「文化」概念也不例外。特別是「文化」是已有傳統時期的意思與用法，在轉型時期獲得了新一類意思。本文根據初步分析《近代韓國報刊語料庫》，試圖追蹤「文化」概念變遷的軌跡。「文化」概念變遷的特點從其概念本身的現代語義，即統稱人類生活的整個生活方式。因此「文化」成爲現代民族固有的生活方式，這楊的語義涵蓋了民族的過去、現在、未來。所以在 20 世紀初朝鮮「文化」是應該全面改造、建設的因素。另外，近代社會系統的細分化加深，「文化」還成爲大文化的小一部分，特別是跟藝術有關的種種事業、活動。變化當中還有不變之處。傳統意義上文化具有的以知識、教養爲主的取向，在近代韓國整個時期綿綿持續了。總之，「文化」作爲又變又不變的概念體現近代轉型的軌跡。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

The "Woman" Subject in Modern Korea

An Inquiry into the Concept of the Woman in the Colonial Era through Competing Signifiers

Lee Jeong-Seon

Assistant Professor, Dept. of History&Culture, Chosun University, Korea

The “Woman” Subject in Modern Korea : An Inquiry into the Concept of the Woman in the Colonial Era through Competing Signifiers

Lee Jeong-Seon
Assistant Professor
Dept. of History&Culture, Chosun University, Korea

摘要

This paper examines the concept of “women” in early modern Korea through the usage patterns of female signifiers during the Japanese colonial period. For this purpose, lexical statistics on the frequency of words and their co-occurrences were used. The traditional female signifiers ‘*yeoja*’ (女子), ‘*buin*’ (婦人), ‘*bunyeo*’ (婦女) were considered, and also a new word ‘*yeosung*’ (女性), which is a translation of ‘woman.’ These signifiers had been restricted to the specific meanings implied by their Chinese characters, but they gained new meanings corresponding to the social expectations current at the time of use. In particular, from 1920 to 1933, when various social movements were active, they all became the subjects of such movements. Thus, ‘*yeoja*’ was a mostly an educational or nationalist subject, whereas ‘*yeosung*’ and ‘*buin*’ were used as socialist subjects, and ‘*bunyeo*,’ which retained its traditional moral image, decreased in frequency. Starting somewhat later, in the mid to late 1920s, and continuing into the early 1930s, ‘*yeoja*,’ which implied a young girl, also became a primary target of sexual interest and criticism. As the colonial authorities repressed the socialist movement, during the 1930s, the use of ‘*yeosung*’ and ‘*buin*’ in this context tailed off. During the wartime period, ‘*buin*’ and ‘*yeosung*’ again became prominent, now as the subject of Japanese mobilization efforts, whereas ‘*yeoja*’ was deployed as a Korean nationalist subject. The usage of ‘*buin*’ emphasized the role of a woman as hostess for her family, and ‘*yeosung*’ was positioned to describe women considered to possess traditional oriental virtues, from which the aspect of sexuality had been removed. In sum, these female signifiers expressed different visions of the present and of the future, so that the analysis of their usage patterns is a convenient way to examine the formation and transformation of the modern concept of “women,” and also the formation process for female subjects in colonial Korea.

目次

1. Introduction
2. Signifiers of “Woman” in Modern Korea
3. Competing Signifiers of “Woman” during the Colonial Era
  3.1. The Movement Era (1920–1933)
    3.1.1. “*Buin*” and “*Yeosung*” as Subjects of the Socialist Movement
    3.1.2. “*Yeoja*” as the Subject of Enlightenment and Nationalism, and Sexualization
  3.2. The Age of Mobilization (1934–1942): The Desexualized Subject
4. Conclusion

關鍵詞

female signifier, *yeosung* (女性), *yeoja* (女子), *buin* (婦人), *bunyeo* (婦女)

1. Introduction

Modern society is purported to be made up of free and equal individuals who enjoy universal human rights. However, history reveals that, even in France, where the 1789 Declaration of the Rights of Man was set forth, women were still not accepted as equal "people" or "citizens." It goes without saying that the political rights of colonial subjects or people of color were repressed. If we are to understand that, in fact, it was through the distinction and exclusion of such non-people or non-citizens that the individual was able to become the subject of modernity, the concept of individual must necessarily be a western, imperialist, and masculine subject. In this study, we are motivated by this critical viewpoint to focus on the "female" subject in modern Korea, owing to the belief that the means for deconstructing/reconstituting extant concepts can be found in the experiences of the non-western, colonial, and feminine. Here, the discourse on the modern woman is notable in that it began to consider women, in general, as a collective subject regardless of class or caste. It is also worth noting that multiple expressions were used to refer to women at the time. This viewpoint follows the approach taken in the study by Tani E. Barlow (1996), which analyzed modern and contemporary China. According to Barlow, the signifiers referring to women, in general, shifted along with the changes in women's discourse. Therefore, in this study, we aim to approach the concept of "woman" in modern Korea by analyzing the use of signifiers that referred to women, in general. The signifiers selected for analysis were "*yeosung* (여성, 女性)," "*bunyeo* (부녀, 婦女)," "*yeoja* (여자, 女子)," and "*buin* (□□, 婦人)." Additionally, to ensure that the concepts here were close to the social norms of the time, we employed lexicostatistical methods to analyze the frequency of the words, as well as their co-occurring words. The data analyzed in this study were taken from the corpus (March 2015 version) of 19 modern magazines compiled by the Hallym Academy of Sciences, Hallym University, the original source being the Korean History Database at the National Institute of Korean History.

2. Signifiers of "Woman" in Modern Korea

The various signifiers of "woman" were not identical in terms of meaning. The meanings were determined by the single-syllable characters included in the words, such as "*yeo/nyeo* (女)" and "*bu* (婦)." "*Yeo/Nyeo*" referred to a daughter, while "*bu*" referred to a married

woman such as wife or daughter-in-law. Thus, "*yeoja* (女子)," which was formed by conjoining "*yeo*" and "*ja* (子)," meaning child, strongly denoted the sense of daughter. Likewise, while "*yeoja*" meant an unmarried woman, "*buin*" and "*bunyeo*" that include "*bu*" usually referred to married women. One notable aspect of such lexicographic definitions is that women were typically categorized according to their age or marital status. This stems from the norms of the times, wherein women were expected to perform familial roles, either as daughter, wife, or daughter-in-law. This does not differ significantly from the usages found in pre-modern Korean sources such as the *Veritable Records of the Joseon Dynasty*. In terms of total frequency of occurrence, apart from "*yeosung*" (8), those of "*yeoja*" (727), "*buin*" (721), and "*bunyeo*" (854) are similar. However, any occurrences of "*yeosung*" in the *Veritable Records* were not signifiers for "woman," as "*yeosung*" was a newly minted term that arose in Japan during the late 1880s as a translation of the western terms "woman" or "womankind."

The unique senses of the signifiers of "woman" continued into the 1920s. Figure 1 indicates the frequencies of occurrence of each signifier, by year, as obtained by analyzing the Hallym Academy of Sciences corpus using WordSmith6 software. During the period from 1896 to 1942, "*yeoja*" occurred a total of 15,736 times, followed by "*yeosung*" (3,733), "*buin*" (3,327), and "*bunyeo*" (550). The growth of "*yeoja*" and "*yeosung*" is marked, as is the decline in the use of "*bunyeo*." In particular, "*yeoja*" and "*buin*" become prominent circa 1906,

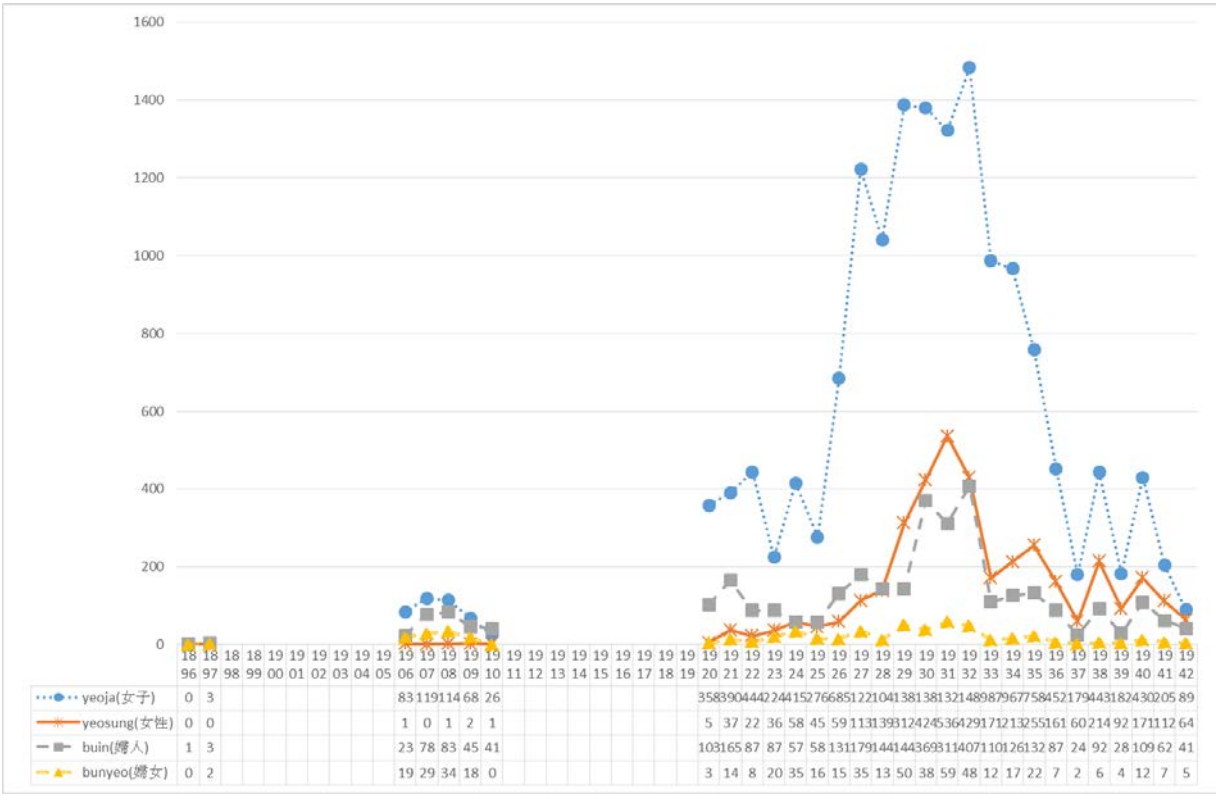


Figure 1 The absolute frequency of the signifiers of "woman" (1896-1942)

while "*yeoja*" becomes dominant following the 1920s, along with the growth of the neologism "*yeosung*" to compete with "*buin*" for the second-highest frequency.

The evolution of the changing occurrence frequencies of the signifiers of "woman" is attributable to the fact that women had come to engage with the public sphere in a different manner, such as being expected to perform social responsibilities or forming organizations of their own. First, the dramatic growth of "*yeoja*" is associated with education. Looking at the co-occurring words of "*yeoja*" from 1896 to 1910, "education (**教育**)" is by far the most frequent, occurring 88 times in 28 articles. This is because, as the Japanese empire encroached upon Korean sovereignty, Koreans looked to the education of women as a means of implementing new education to self-sufficiently realize independence. "*Yeoja*" was tied to education probably because younger and unmarried women were amenable to schooling. "*Yeoja*" also frequently occurred as part of proper nouns, as it was included in school names.

"*Buin*" is another signifier whose usage grew along with the patriotic enlightenment movement. Following the Meiji restoration, proponents of the Japanese enlightenment movement called for equality/equal rights between men and women, eschewing the derogatory "*onna* (女)" and replacing it with "*buin*," a term on equal standing with man. In Korea, as well, "*buin*" was used along with "society (**社會**)" (most frequent, 15 times), "Korea (**韓國**)" (8 times), "academic society (**學會**)" (5 times), and "our nation (**我國**)" (5 times) prior to 1910. It occurred particularly in the names of organizations formed by women and was used alongside "Korean (**大韓**)" and "patriotic (**愛國**)" to indicate that they worked for patriotic purposes.

Meanwhile, "*buin*" and "*bunyeo*" strongly tended to refer to married or adult women. Also appreciable is the division between "*buin*" as mother and "*bunyeo*" as a protected entity. Since the married "*buin*" had sexual knowledge, she was distinguished from "maiden (**處女**)." To this day, terms associated with childbirth or reproduction, such as those for "obstetrics (**産婦人科**)" or "gynecopathy (**婦人病**)" are still formed in conjunction with "*buin*."

Furthermore, in view of the persistence of meaning, let us mention the characteristics of "*yeosung*," although it rarely occurs prior to 1920. "*Yeosung*" began to be used in Japan within the context of the equal rights movement. Even in the first occurrence of "*yeosung*" at 1906 in the Hallym Academy of Sciences corpus, "*yeosung*" denotes someone on equal standing vis-a-vis the man. Additionally, in that it includes the character "*sung* (性)," which is a pointer meaning "quality" or "trait," "*yeosung*" has also been used to refer to femininity since the 1920s. Compound words that are still in use today, such as "feminine (女性的)" and "feminine beauty (女性美)," are prominent examples.

3. Competing Signifiers of “Woman” during the Colonial Era

The usage of the signifiers of “woman” followed a different trend after the 1920s, in the wake of the annexation and colonization of Korea by Japan. Based on the occurrence frequencies reported in Figure 1, Figure 2 is a graph of the percentage shares of the relative frequency of each signifier.

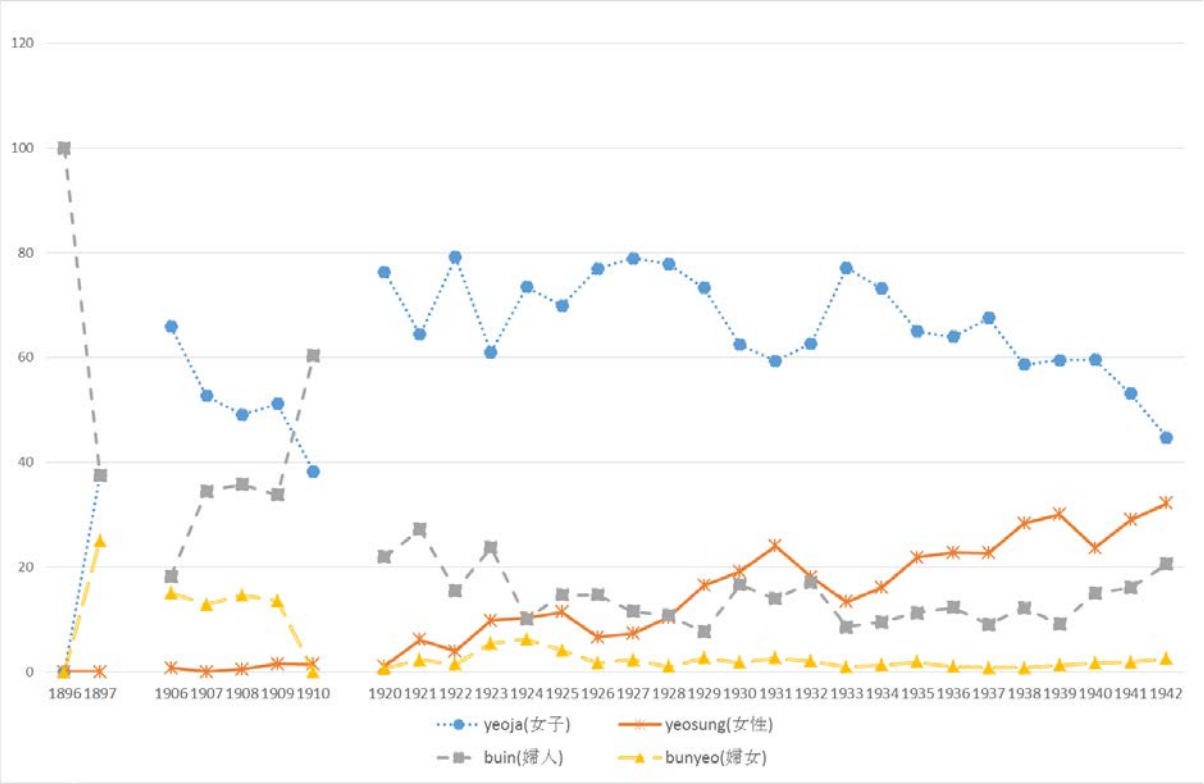


Figure 2 The percentage shares of the signifiers of “woman” (1896-1942)

Here, we examine the state of women’s discourse during the colonial era by analyzing the co-occurring words of the signifiers of “woman.” In this study, we defined a co-occurring word as a morpheme that occurs at least five times within a range spanning the 10th left (L10) to the 10th right (R10) of a signifier of “woman.” Due to the ambiguity in meaning of single-syllable words, only words with at least two syllables were considered. While nouns, primarily, were considered for co-occurring words, adjectives that describe feminine traits or characteristics were also partly included. After discarding redundancies, there were a total of 516 co-occurring words for all signifiers of “woman.” Thereafter, these co-occurring words were categorized into six groups.

Let us consider the co-occurring words shared by the signifiers of "woman." Excluding "*bunyeo*," which has a markedly smaller number of co-occurring words, the remaining signifiers share 35 co-occurring words. The most frequent category that includes these words is "movement-organization." Even after including "*bunyeo*" among the signifiers, the six shared co-occurring words were "man (男子)," "society (社會)," "life (生活)," "Joseon (朝鮮, Korea)," "status (地位)," and "liberation (解放)." Again, the dominant context is that of the movement. Women's discourse, which had risen to prominence following 1906 via debates regarding the national/social responsibilities of women, unfolded during the colonial era centered on the movement for "liberation."

3.1. The Movement Era (1920–1933)

3.1.1. "*Buin*" and "*Yeosung*" as Subjects of the Socialist Movement

Although "movement" is the overarching context for women's discourse, the changes in the types of co-occurring words show a discontinuity before and after the year 1932. In view of this, we categorized the period from 1920 to 1933, during which the "movement" context is most marked, as the era of movement. Figures 3 and 4 report the changes in type of the co-occurring words "*buin*" and "*yeosung*."

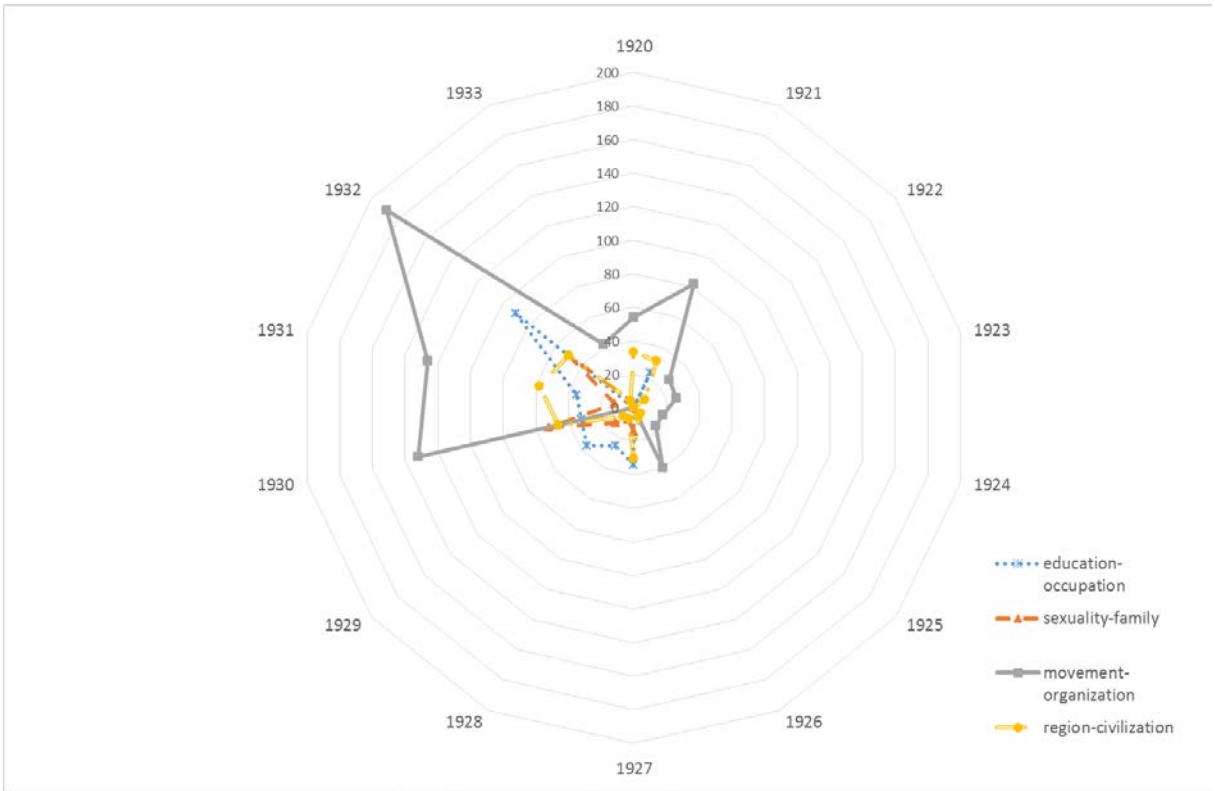


Figure 3 The changes in type of the co-occurring words "*buin*" (1920-1933)

In the case of "*buin*," co-occurring words of the "movement-organization" type were predominant. As women's discourse unfolded within the context of the movement, "*buin*"—now redefined as a social entity—was strongly associated with the movement. This may also be understood as the reason that "*buin*" became a signifier referring to women, in general, during the colonial era, as evidenced by expressions such as "the *buin* issue (婦人問題)," "the *buin* movement (婦人運動)," and "*buin*'s liberation (婦人解放)." Furthermore, words co-occurring with "*buin*" from 1924 to 1925 have strong socialist connotations. As the socialist faction gained prominence within the nationalist movement, the association of *buin* with the movement morphed into an association with socialism. While "*gukje* (國際, international)" was categorized under the "region-civilization" type, it refers to the "international socialist" here.

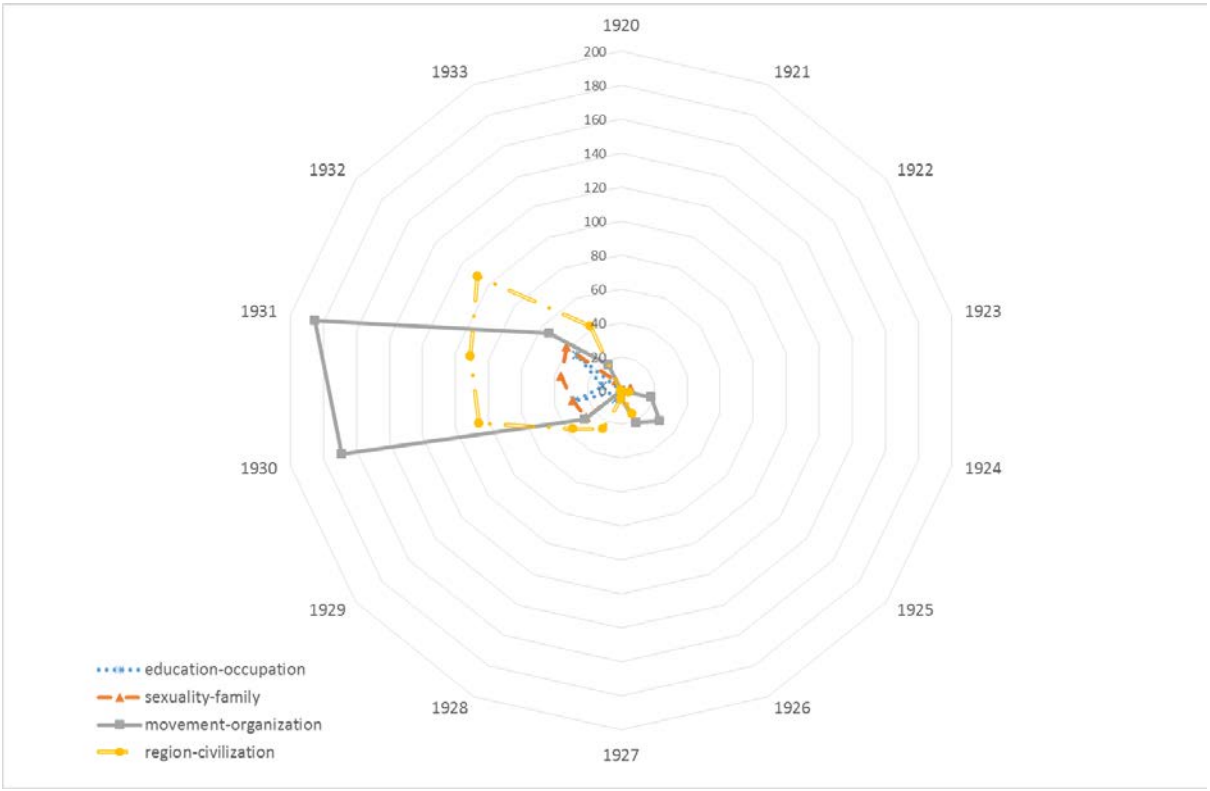


Figure 4 The changes in type of the co-occurring words "*yeosung*" (1920-1933)

The strengthening of socialist connotations is also evident in the case of "*yeosung*" during these time periods. In 1921, "*yeosung*" first appeared as an entity that contrasted with, or was the amorous counterpart of, the man. However, sudden shifts toward an association with movement became apparent in 1924 as it co-occurred with words such as "association (同友會)," "organization (組織, 團體)," and "Joseon (朝鮮)." This trend was most

pronounced from 1925 to 1926, when it became the signifier most strongly associated with movements, based on the establishment of the *Joseon Association of Women* (朝鮮女性同友會), Korea's first women's organization to call for women's liberation, in terms of class struggle.

In the cases of both "*buin*" and "*yeosung*," the connotations of movement receded substantially from 1927 to 1929. It is notable that this change coincides with the activities of the *Shingan Association* (**新幹會**, 1927), which was formed by nationalists and socialists. "*Yeosung*" retained some of its movement-related connotations in that it saw increased co-occurrence with "We/Our" and "Joseon" under the "region-civilization" type in 1926 and co-occurred with "liberation" from 1927 to 1929. However, "movement-organization"-type co-occurring words completely vanished in the case of "*buin*." The dissociation of "*buin*" from the core context of women's discourse at the time (movement) is the backdrop against which "*yeosung*" rose to take the place of "*buin*" as the second-most frequently used signifier of "woman" in 1928.

The temporary rise of "*bunyeo*" from 1923 to 1925 is also attributable to the context of the movement. The only co-occurring words of "*bunyeo*" at the time (1924) were "Joseon (**朝鮮**)" and "liberation (**解放**)." "*Bunyeo*'s liberation (**婦女解放**)" was an expression used among socialist figures who were active or familiar with the state of affairs in China. Looking at words that are closely associated, albeit infrequently, with "*bunyeo*" from 1920 to 1942 via mutual information (MI3) returns terms such as "respectable family (良家)," followed by "respectable person (良人)." "*Bunyeo*," as used alongside "respectable family," referred to chaste women who were worthy of protection. Thus, unlike in China, where old moral norms were still reflected in the use of "*bunyeo*," in Korea, the "*buin*" and "*yeosung*" were named the subjects of the socialist movement.

3.1.2. "*Yeoja*" as the Subject of Enlightenment and Nationalism, and Sexualization

While "*buin*" and "*yeosung*" were socialist subjects, "*yeoja*" was the subject of enlightenment and nationalism. This characteristic is evident in Figure 5. In the case of "*yeoja*," the marked type of co-occurring words from 1920 to 1922 was "movement-organization," with "liberation (**解放**)" (99 times) in 1920 being the most frequent. Also included are other terms associated with enlightenment, such as "personality (**人格**)" and "reconstruction (**改造**)." This is associated with the characteristic of "*yeoja*" as the object of education. Figure 5 also shows that the "education-occupation" type accounts for

a substantial share.

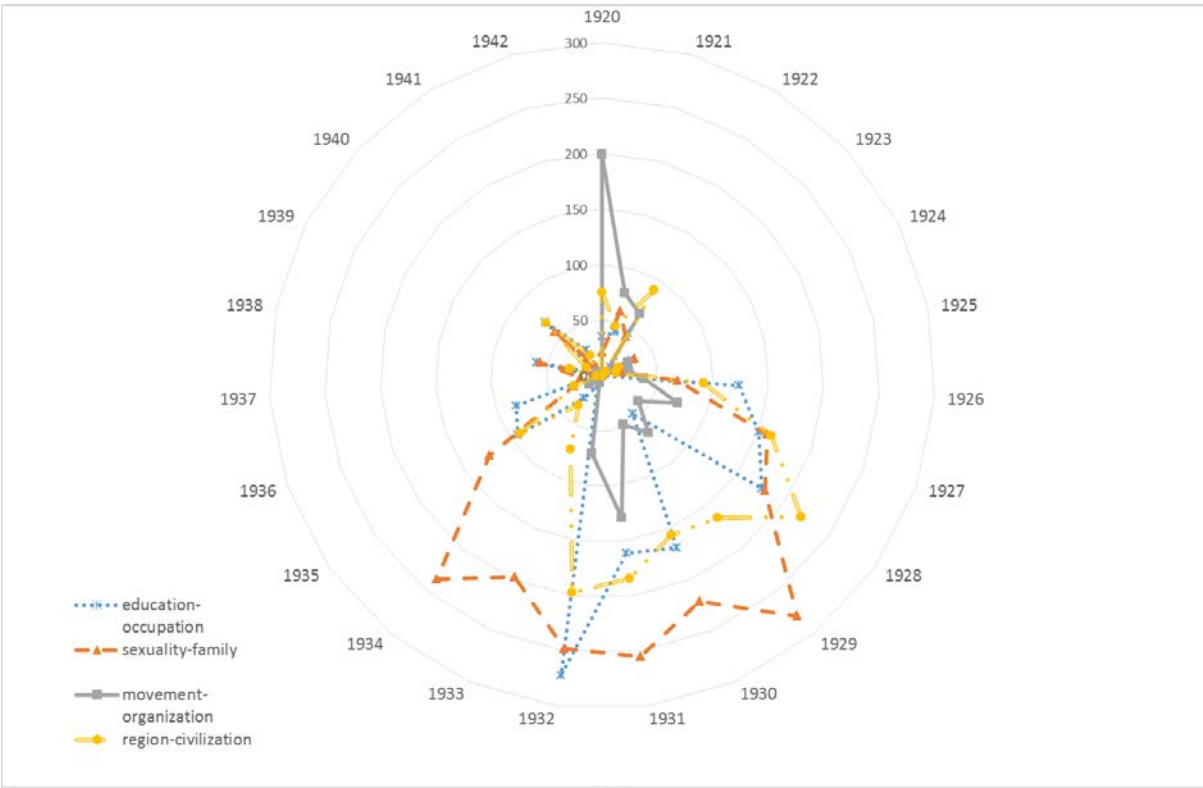


Figure 5 The changes in type of the co-occurring words "*yeoja*" (1920-1942)

"*Yeoja*" has a lower share of "movement-organization"-type co-occurring words, with higher shares seen from 1921 to 1923 and from 1927 to 1932. This reflects the fact that "*yeoja*" was the subject of the movement before the socialist movement branched off from the nationalist movement or during periods when the two factions were in cooperation. The "region-civilization" type also has a substantial share; "Joseon (朝鮮)" co-occurred most frequently, indicating that "*yeoja*" was the subject of nationalism.

Why, then, did the share of "*yeoja*" decline as of 1933? First, one might consider the decreased interest in women's education due to the economic downturn. This was accompanied by a sharp decrease in co-occurring words of the "education-occupation" type for "*yeoja*." Furthermore, "*yeoja*" also became dissociated from the context of the movement after a brief intensification. The pattern that arises from 1928 to 1933 shows that it is "*yeosung*" that is associated with "movement (運動)" and "liberation (解放)." This indicates that "*yeosung*"—a person on equal standing with a man—had grown to become the subject of movements encompassing nationalism, socialism, and feminism.

A comparison of Figures 3, 4, and 5 shows that words co-occurring with "*yeoja*" under

the “sexuality-family” type increased from the late 1920s to the early 1930s. Thereafter, the “sexuality-family” type becomes the most frequent type for “*yeoja*.” Co-occurring words of this type also include expressions that comment on appearances, in addition to various words that express objection to women’s unbridled sexuality, such as “concubine (妓生),” “pregnancy (娠娠),” “cohabitation (同居),” “illegitimate child (私生子),” “corruption (墮落),” and “harem (하렘).” When the “new woman (shin*yeosung*)” became the target of umbrage as the personification of excess and decadence, it was the “*yeoja*” student who was mainly denounced.

3.2. The Age of Mobilization (1934–1942): The Desexualized Subject

Following 1934, the rebelliousness that sought liberation was diminished in women’s discourse. This vacuum was instead filled by denunciation/enforcement/control regarding the chastity of female students. “*Buin*” co-occurs with “female student (女學生)” only from 1935 to 1936, when the conservatism of older “*bunyeo*” and “*buin*” were upheld as virtues that contrasted with sexual indulgence.

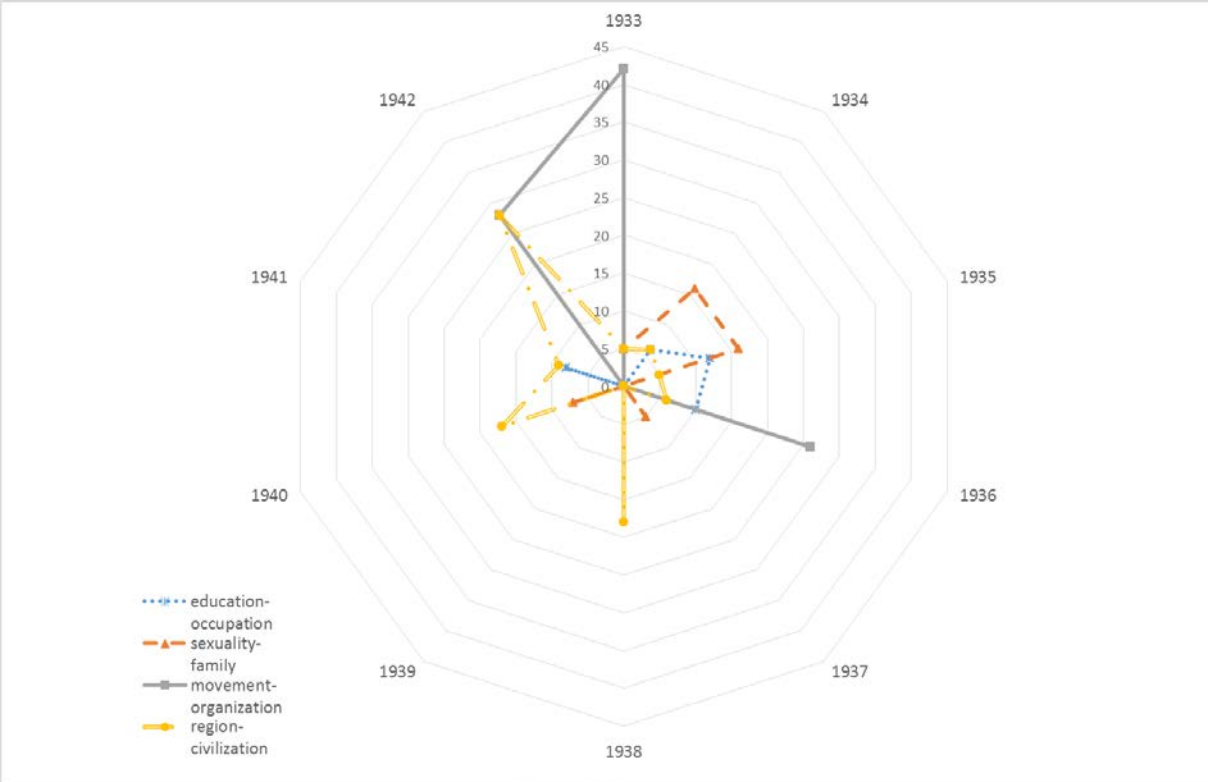


Figure 6 The changes in type of the co-occurring words “*buin*” (1933-1942)

As the Japanese Empire prepared for war, it was the "*buin*" that became the focus of mobilization. The "*buin*," which was a social entity, was not sexually dissolute, and embodied motherhood was deemed the ideal subject to take responsibility for the rear guard. After 1942, especially, when the conflict had escalated into the Pacific War, a multitude of war-related terms such as "rear guard (銃後)," "national defense (報國)," and "decisive battle (決戰)" co-occurred with "*buin*." The increased share of "movement-organization"-type words in Figure 6 reflects this change.

Meanwhile, Figure 2 reveals that the 1940s saw a rise in "*yeosung*," accompanied by a sharp decrease in "*yeoja*." One reason for this might be that "*yeoja*" was a nationalistic subject that frequently occurred alongside "Joseon (朝鮮)." In Figures 6 and 7, there is an increase in "region-civilization"-type words that co-occur with "*buin*" and "*yeosung*" during the 1940s, with the new additions being "the peninsula (半島)" and "the Orient (東洋)." Japanese authorities at the time preferred the term "peninsula" to "Joseon," which had strong nationalist overtones. Likewise, they replaced "Joseon person (朝鮮人)" with "peninsula person (半島人)," "peninsula countrymen (半島同胞)," and "Japanese in the peninsula (半島日本人)." However, "Joseon" remained a frequent co-occurring word with "*yeoja*," while "peninsula" did not co-occur either to the left or to the right of "*yeoja*." On the other hand, "*buin*" and "*yeosung*" were included in patterns of expressions such as "our *buin* of the peninsula" and "our 10 million (or 12 million) *yeosung* of the peninsula." While the use of "*buin*" and "*yeosung*" instead of "*yeoja*" was better suited to mobilizing Korean women for

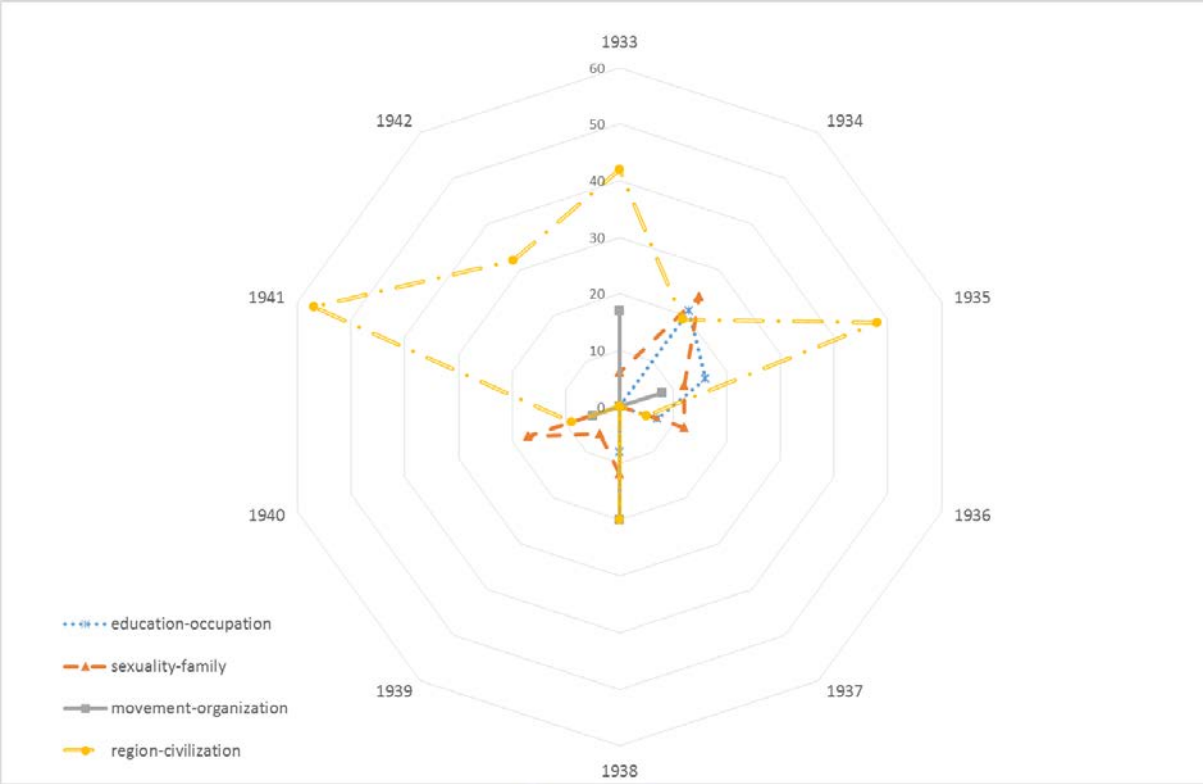


Figure 7 The changes in type of the co-occurring words "*yeosung*" (1933-1942)

Japan's war, there was also the need to refer to "*buin*" and "*yeosung*" as belonging to a province (i.e., the peninsula) of Japan.

"The Orient (東洋)" co-occurs with "*yeosung*" from 1941 to 1942, due to the emphasis placed on the character/traits of the oriental woman. As used in conjunction with the oriental under an anti-Western mood, "*yeosung*" was closer in image to the virtuous "*bunyeo*" than a sexual object. Here, "oriental *yeosung* (東洋女性)" does not refer solely to Korean women: After the Pacific War, there was a greater effort to associate women with "the Orient," with reference to the so-called "Greater East Asia Co-Prosperity Sphere (大東亞共榮圈)."

In short, the Japanese authorities at the governorship-general of Joseon sought to convert Korean women to Japanese people by choosing the term "*buin*," which had less social or sexual connotations. At the same time, it tried to include Asian women as the Orient "*yeosung*," unlike Western women. This was the context in which "*buin*" and "*yeosung*" became prominent, while "*yeoja*"—both a nationalistic subject and a sexual object—declined.

4. Conclusion

To conclude, we briefly summarize the contents of this study and outline directions for future research, as follows.

First, the concept of "woman" should not be defined biologically but through historical analysis. Traditional signifiers, especially, such as "*yeoja*," "*buin*," and "*bunyeo*," in addition to the translated term "*yeosung*," not only have their own unique meanings but were imbued with varying contexts and meanings in different historical periods, in accordance with the prevailing social expectations of the time. Such changes are the outcome of a struggle to mobilize women who identify as "*yeoja*," "*buin*," or "*yeosung*" to the enunciator's movement. At the same time, these changes were also efforts to imbue the signifiers with the enunciator's ideology/beliefs, so that they could be spread to women, in general.

Second, having been converted into the subjects of nationalist or class-based movements, women became desexualized entities. This point holds, regardless of whether the enunciator was Korean or a colonial ruler. Within the dichotomy of sainthood/whoredom, women had to efface their sexual aspect to belong as a "national" or "countrywoman." Contemporary history after liberation from Japan also shows how desexualized the concepts of "nation" or "citizen" were in Korea. Therefore, a re-conceptualization of "individual" or "citizen" must be undertaken, through gender consciousness or consideration for diverse sexual orientations.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

Emotions of colonial Joseon and its History

LEE, Jura
HK Research Professor
Hallym Academy of Sciences at Hallym University

Emotions of colonial Joseon and its History

LEE, Jura

HK Research Professor

Hallym Academy of Sciences at Hallym University

Abstract

In this study, I want to analyze the historical changes of the emotion as a concept, to extract the dominant emotions of early modern Korean society, finally to find features of Korean modern culture. Emotions can be constituted and learned socially and culturally. It is possible to form some discourses about specific emotion by social intentions. So a society can emphasize and exclude an emotion. As a result, emotions can reflect social features and historical changes. Therefore, this study will analyze the historical changes of the specific emotion, and will find out distinctive sensibility and desire in Korean history.

Most studies on emotions have been done in a way that analyzes specific individual emotions in a single text. In contrast, this study will examine quantitative analysis of the process and the features in the course of modernization of the concept of emotion itself. At the same time, I will extract the major emotions that have gained attention in the discourse of modern Korea and draw pictures of the changing map of emotions. For this analysis, I will extract and analyze lexical collocations and frequencies of emotions, based on Hallym Corpus, the main newspapers and magazines corpus of early modern Korea. First of all, this study will analyze emotional features of Korea. This will reveal the differences in cultural sensitivity among the Northeast Asian countries. Understanding the differences would be the

basis for mutual understanding and communication.

Table of contents

1. Introduction

2. Colonial Joseon and the Agitation of Emotions

3. Changes in History in the Early 20th Century and People's Responses

3.1. Establishment of commercial culture and discovery of sensual senses

3.2. The Crisis of War and the Rise of Anxiety

3.3. The Realization of War and the Back of Optimism

4. Conclusion

Keywords

emtion, feeling, anxiety, optimism, popular culture, colonial period, <Hallym Corpus>, <Trend 21>

1. Introduction

Emotion is usually used as a vocabulary to indicate a person's inner state of mind. And emotional state indicates a state of passion or excitement. It is often linked to the state of being love or dating. This emotional or emotional state has a close relationship with romantic or passionate situations.

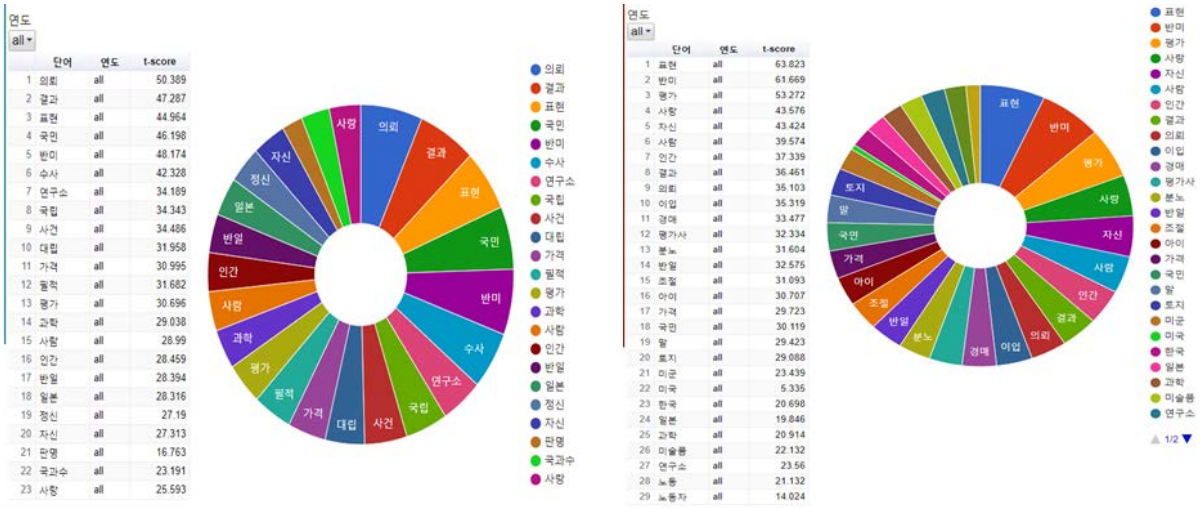


chart 1 co-words of the late 20th and early 21st century

However, the social function of the word of ‘emotion’ is very different from the meaning used in everyday life. The chart1 above is a presentation of the co-words of ‘emotion’ that appeared in Korean newspapers in the late 20th and early 21st centuries. In both periods, the most closely used co-words of ‘emotion’ are words such as ‘anti-American’, ‘nation(國民)’ and ‘anti-Japanese’. When emotions are used in social discourse, unlike the usual usage, emotional subject is set as collective subject such as nation(國民), not individuals. It also reflects historical conflicts between the two countries. That is, the romanticism of emotion disappears and negative emotion is revealed. In such a social discourse, emotions reveal the collective sensibility or atmosphere of the society, which can be called national sentiment. The emotions of a collective subject are often expressed as anger, reflecting immediately the problem situation of a society. Such feelings directly show the point where irrational and unrestricted energy is emitted at the national or institutional level, enabling us to sensitively capture the problematic situation and the atmosphere of the times.

It was after the beginning of modern society that the word ‘emotion’ gained social

meaning. In a Confucian-centered society, emotions were nothing but objects of thought and moderation. Emotional expression and eruption become possible in modern society. In fact, the word of ‘emotion’ in Korea has emerged from social discourse in large quantities since the 1920s. The frequency of the word ‘emotion’ in the 1920s magazine is only six. So, how was the word emotion used in early modern Korea? The purpose of this study is to examine the role of emotion in Korea through comparison of the co-words of emotion used in the early 20th century when Korean modern society began, and take a look the historical changes and the public's response in Korean society.

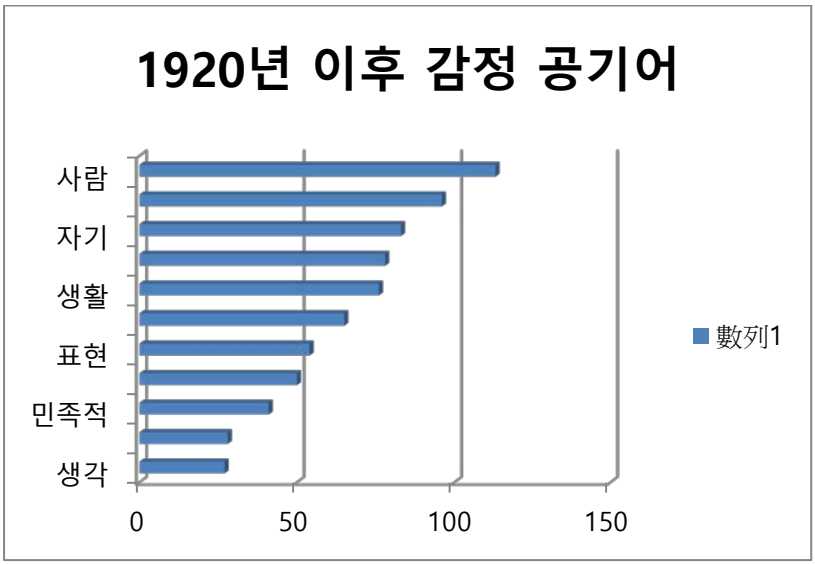


chart 2 co-words in early 20th contury

Even in the early modern times, the word of ‘emotion’ in social discourse showed the state of the collective subject. However, our people(民族), not the nation(國民), represented the collective subjects. This reflects the idea of separating the Japanese Empire and the Korean people in the colonial period. On the other hand, in the early period of the modern era, specific emotional conditions such as love and excitement are presented in detail. The interest in love

reveals the social atmosphere of early colonial Joseon, also known as sentimental period. In the early modern era, Korea believed that it could form an individual subject by experiencing the feeling of love and thought that the discovery of this modern individual was the beginning of modernization achievement.

Korea, which experienced colonies at the beginning of modern times in the early 20th century, expressed emotional understanding of modernity and colonial situation through the modern vocabulary of emotion. Therefore, this study will find out what emotions meant to the Korean people during the colonial period and what roles emotions played. Furthermore, by looking at the political and social changes of the colonial Joseon, which is represented by the words of 'emotion', this study will find out where the emotions of Joseon people that lived in contemporary society erupted and what were the collective desires of the colonial Joseon people through them.

2. Colonial Joseon and the Agitation of Emotions

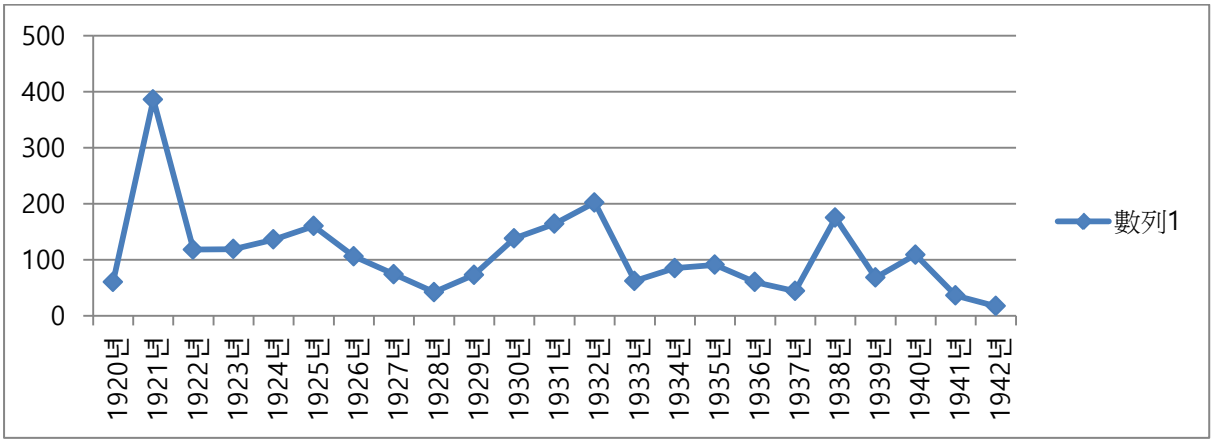


chart 3 frequecy of the word 'emotion'

The use of emotional vocabulary in the colonial Joseon begins to explode

immediately after the March First Independence movement. Until 1926, the use of the word ‘emotion’ is quite high, reflecting the 1920s as a sentimental age. Since then, the frequency of the word of ‘emotion’ increases around 1932 and 1938. What happened at this time? And did the word emotion really only express the sorrow of the public in the colonial period? Then, why did the word of ‘emotion’ increase in 1932 and 1938 when the Manchuria Incident and the Sino-Japanese War broke out, after the 1920s as a sentimental era? To find the answer, this chapter focuses on the 1920s, when the word of ‘emotion’ was the most used. By analyzing the so-words of ‘emotion’ used in the mid-1920s, we will find out what roles and functions emotion played in colonial Joseon society.

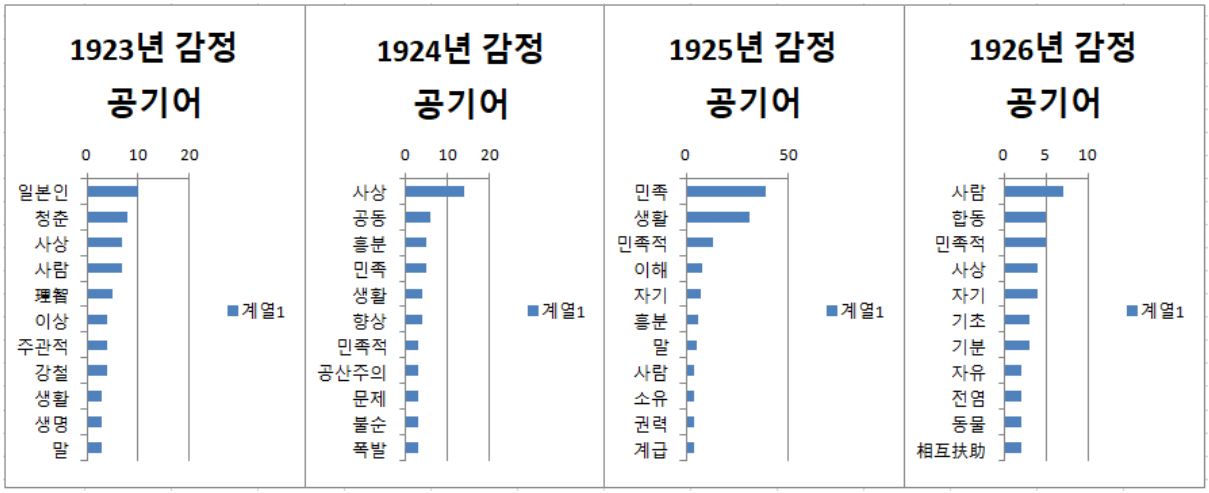


chart 4 co-words of emotion from 1923 to 1926

In 1921, when the word "emotion" was used as a trend, there were a comprehensive set of the uses of the word ‘emotion’. So while it may be possible to confirm that there was a growing social interest in emotions in the early 1920s, it is difficult to summarize the specific uses of the word ‘emotion’ at that time. After 1921, the co-words of ‘emotion’ began to reflect the atmosphere of the times. In the mid 1920s, the co-words of ‘emotion’ appeared in words such as ‘people(民族)’,

‘community’, ‘class’, ‘thought’, ‘ideal’, and ‘improvement’. This period was the time when various national movements were developed after the March First Independece Movement, and socialism, a critical discourse of modern society, was settled in Joseon society. At this time, I can guess that the emotion was used like this. Emotion and its co-words conveyed these meanings, such as urging people to express thoughts and feelings of ethnicity or common ideology or class, aiming to pursue ideal society by improving ideas and emotions.

Until now, the 1920s were analyzed as a sentimental age. But here, emotion did not convey sentimental sentiments such as sorrows or emotions related to love. Emotion played a role in promoting national movements and spreading new ideas through emotional uplift. Emotions did not remain sentimental in the personal sphere. Emotion played a role in driving the movement of ethnic groups(民族). Even socialism that emphasizes the cold analysis of reality, uses emotional stimulation to propagate ideas and actions of the public. Emotions are not just personal feelings. Emotion can serve as the driving force for social thought and practical movement through collective transmission and diffusion.

3. Changes in History in the Early 20th Century and People's Responses

Emotions reflect the collective movement of people and society. In other words, it shows how collective emotions move according to social changes. In this chapter, I will analyze how the word ‘emotion’ is used at certain times through the change of the co-words of ‘emotion’. I will also refer to the co-words of the word ‘feel’ since the 1920s to see exactly what the public felt at each time. Hopefully, this will help to grasp ‘what emotions they felt’ during the early modern era.

3.1. Establishment of commercial culture and discovery of sensual senses

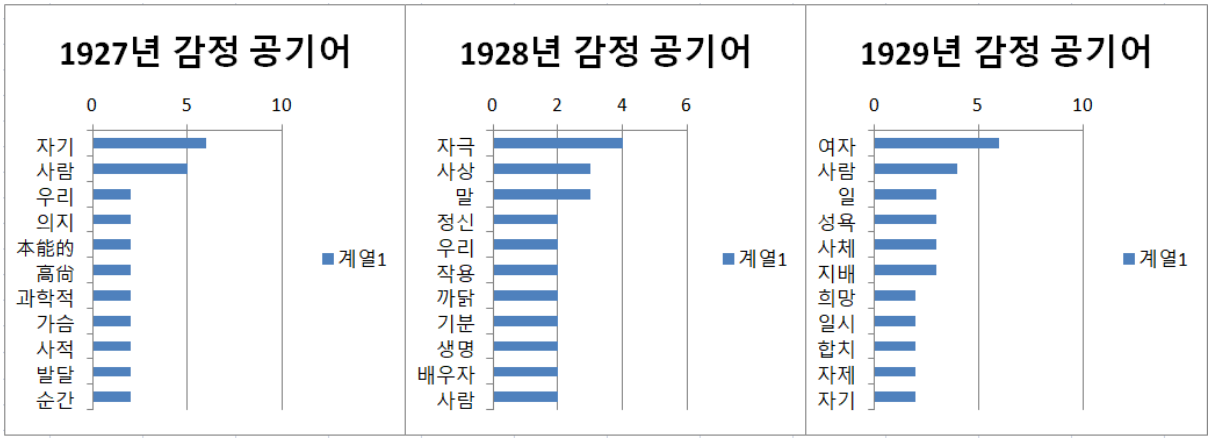


chart 5 co-words of emtions (1927-1929)

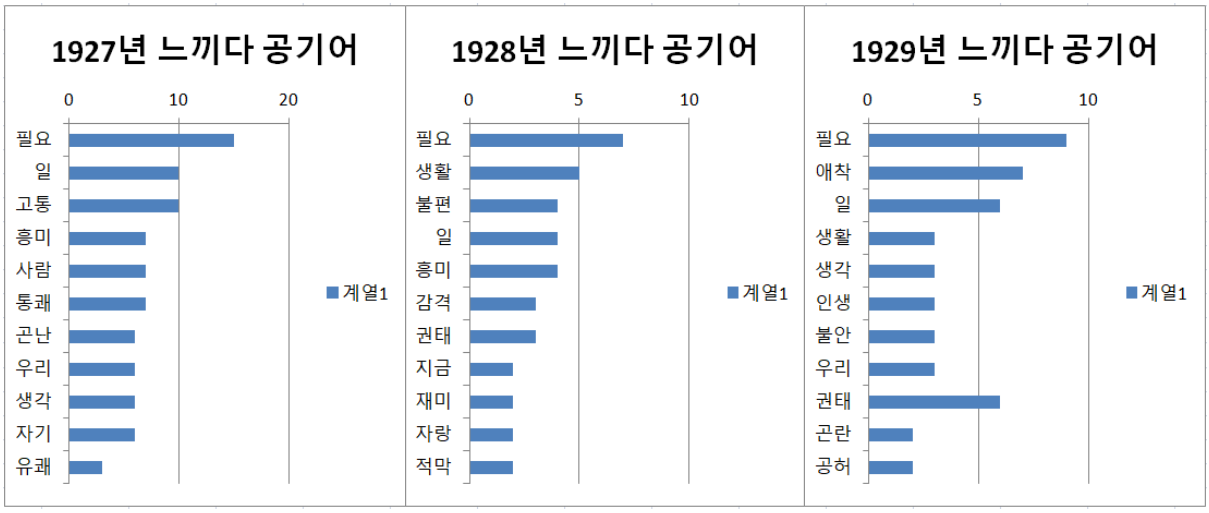


chart 6 co-words of feel (1927-1929)

Between 1927 and 1929, the co-words of ‘emotion’ differ from previous periods. In the mid - 1920s, the co-words of ‘emotion’ were interested in enchancing the ideas and ideals of the Korean people, the collective subject. On the other hand, words such as ‘self’, ‘private’, ‘instinctive’, ‘sexual desire’ appear as the co-words of emotion in the late 1920s. The words ‘self’ and ‘private’ are a vocabulary that refers to an individual subject. ‘Instinct’ or ‘sexual desire’ is a vocabulary that expresses a sensual senses felt by an individual subject. This

reflects the stimulation of discourse about the instinctive pleasure of the individual at this time. It shows that personal emotion is consumed in the area of sensual pleasure.

This change is also evident in the co-words of the word ‘feel’. The co-words of ‘feel’ specifically describes 'what individuals felt' in the late 1920s. At this time, the words on the list of words are ‘interesting’, ‘fun’, ‘excitement’, and ‘pleasant’. It is a vocabulary similar to the co-words of ‘emotion’. It can be seen that individuals pursued the same interest and fun as instinctive pleasure. And the ‘interesting’ and ‘fun’ they think are also exciting and pleasant. Unlike the depressed atmosphere, which was regarded as general emotion of this period, there was a society that pursued funny entertainment.

The late 1920s was a time when modern popular culture settled and developed. Under the influence of the Taisho Democracy of Japan, a commercial culture for the public has developed. In addition, socialists have paid attention to the way they mobilize the public to spread socialist ideas. The interest in the public created within the Joseon intellectual society, the establishment of capitalism based on the stabilization of colonial rule, and the spread of the commercial culture that came in through Japan made popular culture flourish in colonial Joseon in the late 1920s. The creation of *Byeolgeongon(別乾坤, Another World)*, a magazine for popular taste and fun, in November 1926 reflects this social atmosphere. At this time, popular culture was formed mainly by the code of ‘ero(tic), gro(tesque), nonsense’. As the commercial culture began to set in, the Korean public was interested in individual senses and instinctive pleasures, and pursued a pleasant laugh out of the sadness of the previous era.

3.2. The Crisis of War and the Rise of Anxiety

1930년-1935년 감정 공기어											
1930		1931		1932		1933		1934		1935	
자기	10	민족	10	사람	11	사람	4	사람	7	때	5
사상	9	사상	9	민족적	8	지배	2	경이	6	정서	4
사람	9	표현	6	理智	7	자기	2	거즛	6	표현	3
국민적	8	우리	6	우리	6	작자	2	자기	5	사람	3
불쾌	5	의지	4	표현	5	기분	2	우리	4	사랑	3
정서	4	사람	4	자기	5	같이	2	정서	4	일시	3
감염	4	악화	4	사상	5	악화	2	극도	4	자기	3
생활	3	고급	3	말	4	양국	2	녀자	4	말	3
까닭	3	이상	3	악화	4	무엇	2	완화	3	충돌	3
예민	3	생활	3	예술	4	비판	2	순실	3	융화	3
평생	3	불쾌	3	지배	4	선생님	2	강렬	3	모순	3

chart 7 co-words of emotion (1930-1935)

1930년-1935년 느끼다 공기어											
1930		1931		1932		1933		1934		1935	
때	11	필요	31	필요	32	때	13	때	12	필요	12
필요	9	우리	10	권태	11	흙	8	필요	12	소리	12
불안	8	공포	9	흥미	11	흥미	6	공포	9	마음	12
사람	8	비애	8	공포	10	필요	5	불안	8	공포	10
우리	7	불안	8	흙	9	사람	3	생각	5	불안	9
비애	5	사람	7	일	8	이상	3	쾌감	5	사람	8
모욕	5	흥미	7	불만	8	시작	3	소리	4	흙	6
부족	4	권태	6	불안	7	힘	3	우리	4	우리	6
문제	4	실망	5	소리	6	충동	2	권태	4	憎惡	5
공포	4	만족	4	불쾌	6	허무	2	환멸	4	몸	4
孤寂	4	부족	4	만족	6	환멸	2	불쾌	3	쾌감	4

chart 8 co-words of feel (1930-1935)

The word ‘emotion’ that had dropped in frequency in 1928 began to increase again over the 1930s. As the usage of emotional vocabulary increased, they concentrated on emotional problems. However, all of the emotion related vocabulary in this period appears as negative words. For example, words such as ‘displeased’, ‘anxiety’, ‘aggravation’, and ‘explosion’.

The words ‘people(民族)’, ‘ethnic(民族的)’, and ‘national(國民的)’, which refer to collective subjects, also appear again. As the negative emotions increase, Joseon society became worse and ready to explode.

Here we need to pay particular attention to the emergence of the word ‘national(國民的)’. Unlike the word ‘people(民族)’ which refers to the ethnic group, the state and nation are words based on the acceptance of the rule of the Japanese empire. The fact that national feelings, not ethnic emotions, have been used prominently means that the conflict of two countries has been raised as a social issue, and that the political discourse of the Japanese Empire has begun to exert its full force through the media. In fact, since Manchurian incidents took place in 1931 and Manchuria was built in 1932, the relations between Japan and China deteriorated. As the tensions between the two countries increased, the crisis of war began.

This led to anxiety and fear in the Korean society, which had been under Japanese rule. When we look at the co-words of ‘feel’, we can see that the specific emotions felt in the Joseon society during this period were ‘anxiety’, ‘fear’, ‘disillusionment’, ‘hatred’. In previous periods, feelings of anxiety were sometimes revealed, but at this time anxiety is combined with fear such as horror, accompanied by feelings of hatred for something. Negative emotions are rising in the society due to the conflicts between the two countries, and it can be seen that this is expressed as a violent expression such as hatred toward the opponent.

Meanwhile, during the period of anxiety and fear, the Joseon public had a desire to escape from tense sentiment. In addition to ‘anxiety’ and ‘fear’, words such as ‘disappointment’, ‘weariness’, ‘futility’, ‘disillusionment’, and ‘desolation’ in the list. It can also be interpreted that there is a sense of skepticism about the rising tensions in society due to the

crisis of war. In times of national and social conflict, violent emotions such as dislike and hatred are revealed, but at the same time disappointment and disillusionment over the governance methods that make daily lives uncomfortable are revealed. These feelings, of course, are not a form of direct resistance to the political situation that stresses the one's daily life, but they properly express public discontent with social anxiety and fear.

3.3. The Realization of War and the Back of Optimism

1936년-1941년 감정 공기어											
1936년		1937년		1938년		1939년		1940년		1941년	
우리	9	연애	12	감각	9	결합	4	생활	5	생활	7
사상	5	자기	5	사람	9	사상	3	인간	4	우리	3
표현	3	사상	4	연애	8	표현	3	자기	4	진실	2
조선	2	인간	4	인간	5	成熟	3	사상	3	사람	2
회계	2	조선	2	우리	5	단순	3	말	3		
지금	2	우리	2	지성	5	적	3	적	3		
때	2	결혼	2	말	4	국민적	3	우리	3		
말	2	남편	2	문학	4	연애	2	청춘	2		
사람	2	발로	2	자기	4	정열	2	처음	2		
모든	2	마음속	2	생각	3	인간	2	취미	2		
민족적	2	동물	2	표현	3	의지	2	향수	2		

chart 9 co-words of emtion (1936-1941)

1936년-1941년 느끼다 공기어											
1936년		1937년		1938년		1939년		1940년		1941년	
필요	7	사랑	3	필요	13	우리	4	때	22	때	9
마음	5	때	3	사랑	10	말	4	향수	14	피로	6
생각	5	님	2	때	8	사랑	3	필요	13	필요	6
소리	5	흥미	2	자기	7	朝鮮	3	매력	8	우리	6
사람	4	청년	2	깊이	5	필요	3	우리	7	불안	5
사랑	4	앞	2	책임	4	때	3	행복	7	공포	4
悲哀	4	외로움	2	시작	4	어머니	3	소설	6	切實	4
흥미	3			불쾌	4	행복	3	권태	5	생각	4
적막	3			고통	4	책임감	3	불안	5	感激	3
행복	3			가슴	4	오늘	2	마음	4	어머니	3
환멸	3			의무	3	충동	2	情懷	4	행복	3

chart 10 co-words of feel (1936-1941)

The Sino-Japanese War, which began in 1937, was extended to the Pacific War in 1942. Paradoxically, the emotions that were emphasized during the war were related to vocabularies such as ‘love’, ‘life’, ‘taste’, and ‘nostalgia’. This is same aspect with the co-word of ‘feel’ at the same time. The co-words of ‘feel’ were also the words such as ‘love’, ‘happiness’ and ‘nostalgia’, which express the stability and satisfaction of one’s personal life.

However, it can be seen that analogy of the context in which ‘love’ and ‘happiness’, which appeared as emotional expressions at this time, is not simply a universal meaning of ‘love’ and ‘happiness’. In the late 1930s, love and happiness emerged with the words ‘responsibility’, ‘duty’ and ‘responsible feeling’. In this context, we can see that ‘love’ and ‘happiness’ at this time are not the full emotions that individuals feel in concrete experiences. Rather, ‘love’ and ‘happiness’ are used as an abstract concept that a collective subject must obligatorily pursue in the social discourse. In fact, ‘love’, which appeared in 1939, appears with

‘Joseon’ and ‘mother’. It can be inferred that it means the love of Joseon's mother. At that time, the love of the mother of Joseon was related with the ideology of Japanese empire. The Japanese Empire emphasized the love and sacrifice of the mother who keeps the daily life instead of her husband in battlefield and the mother who raises the soldiers of the future. The emotional expressions that appeared in the late 1930s were used to spread national ideology rather than reveal the personal inner side.

In 1938, the national discourse reinforced the propaganda to ensure a stable war. Ideology for war does not consist only of extreme violence. It should also be emphasized that the optimistic prospect that the present sacrifice will lead to future victories. Paradoxically, during the war, the discourse of optimism is revealed on the surface of society. Through the media, the offensive of national ideology is intensified. The offensive of national ideology consists of an emotional approach as evidenced by the co-words of ‘emotion’ and ‘feel’ in the late 1930s. Ideology spreads through emotional sharing rather than through rational persuasion. The frequency of the word ‘emotion’ increases at the time of national ideology attacks, as in 1938.

Unlike ostensible optimism, however, the public anxiety that experiences war directly does not disappear. As shown in the co-words of the 1941, ‘fatigue’, ‘anxiety’, and ‘fear’ of war do not disappear at the time of transition to the Pacific War. Of course, in 1942, as the word ‘pride’ and ‘boasting’ show, the ideology of war sutures individual anxiety and fear. But behind the happiness and love driven by national ideology, that is optimism about the future, there has always been anxiety and fear. Social discourse tries to spread political ideology through emotion, but individuals who live in that society express the true feelings they experience through emotion. Social discourse tries to spread political ideology through emotions, but the individuals living in that society express their true feelings through emotions. These emotions

immediately reflect the problematic point of society that the state and system cannot control. Only when such feelings can be read will the public's desire within the historical changes of society be read.

4. Conclusion

This study examined the dominant emotions of Korean society in the early 20th century through the social role and the changing pattern of the concept of emotion. Emotions can be constituted and learned socially and culturally. It is possible to form some discourses about specific emotion by social intentions. So a society can emphasize and exclude an emotion. As a result, emotions can reflect social features and historical changes. In this study, emotions with this historical context were the key words. This allowed us to identify the social functions of emotions, and to sort out what collective subject felt and how they reacted in the historical changes in Korean society of the early 20th century.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

中古僧人的人名辨識

以《高僧傳》、《續高僧傳》、《出三藏記集》為例

Name Recognition of Medieval Chinese Monk Names

A Conditional Random Field Approach for Name-Extraction based on Biographies of Eminent Monks (高僧傳), Continued Biographies of Eminent Monks (續高僧傳) and Collection of Records Concerning the Chinese Buddhist Canon (出三藏記集)

謝薇娜*　陳雅琳**　林品傑*
中央研究院*　國立中正大學**

中古僧人的人名辨識──以《高僧傳》、《續高僧傳》、《出三藏記集》為例

Name Recognition of Medieval Chinese Monk Names: A Conditional Random Field Approach for Name-Extraction based on *Biographies of Eminent Monks* (高僧傳), *Continued Biographies of Eminent Monks* (續高僧傳) and *Collection of Records Concerning the Chinese Buddhist Canon* (出三藏記集)

摘要

本文擬透過 Conditional Random Field（CRFs）的方式，訓練出 CRFs 序列標記（Sequence Labeling）的辨識模型，搭配建立關鍵詞庫（Dictionary）與特徵值詞組，進行《僧傳》的人名擷取，期能修正長度較長的人名與別名辨識模型，以增加下一階段遊方事件擷取的準確度。實驗將以兩種外國僧人最活躍的中古僧傳文本為主要處理材料，並以現存最早的完整佛教經錄《出三藏記集》為參照，擷取並標記僧人傳記文本所收錄的眾多僧人人名、簡稱與別名，挖掘出對傳記文本更具辨識強度的人名特徵，並在此基礎上，完成建立人名辭典的工作流程，進而分析僧名與時代、地域、科別的相關性，凸顯中古佛教的特性。

數位人文研究領域中，雖已有針對古籍文獻中的人名進行辨識，如建置人物傳記資料庫的 CBDB，即透過《清代人物生卒年表》、《宋人傳記資料索引》等具有歸整資料的工具書，以正規表達式（Regular Expression）提取其中的人名資訊；或如以《清實錄》和《資治通鑑》為對象的人名擷取，透過字詞之間的機率進行斷詞，進而挖掘人名。但這些用以進行人名辨識的文本，多集中在有固定書寫體例的史料與完整的人名資訊，對於容納大量人物資料與訊息的傳記文本，卻因包含太多文學性的敘述手法，不易進行人名辨識與資料擷取，尚有待進一步開發技術。換言之，人物傳記在敘述傳主生平事蹟時，往往會使用人名子集（Subset），或是透過人稱代名詞來指涉傳主，用以保持文本敘述的流暢性。如僧人竺法蘭的傳記中，便透過人名子集「法蘭」與「蘭」，或是人稱代名詞「之」、「其」等，作為指涉傳主的人物簡稱與別名。而文句中的行動者又是判斷事件的重要線索，如何辨識文本中的人名，以及其所包含的人名子集與人稱代名詞，便成為進行下一步事件擷取前不可逃避的程序。

一、前言

現代文本的人名實體辨識方式，往往需要透過百家姓字典作為特徵標記，而能夠處理的人名實體字元長度也限縮在二到四個字元的範疇，倘若需要處理字元長度較高的人名，便會因為複姓與詞組過度合併的風險而表現不佳。由於中古僧傳文本中記錄許多外國僧人，其人名多經翻譯之故，時常會出現超過三個字元的僧名，甚至是長達六到八字元的人名，如：釋若那跋陀羅（Jñānabhadra）、波羅頗迦羅蜜多羅（Prabhākaramitra）等，此種過長的人名，在目前現有的辨識方法中，便會出現難以挖掘的困境，亦無法透過百家姓字典作為辨識特徵。在此基礎上，本次所進行的人名辨識實驗，便以〔梁〕釋慧皎所撰的《高僧傳》、〔唐〕道宣《續高僧傳》為實驗對象，將重點放在此類具有特殊性質的外國僧人姓名之上，期能處理字元長度更長、以及組成更為複雜的人名實體，並從中建構出一套辨識人名實體的模型，復以〔梁〕釋僧祐《出三藏記集》為測試文本，覆核此套模型的準確性。

本次實驗將以《高僧傳》、《續高僧傳》、《出三藏記集》的 TEI 標記文本，做為 Conditional Random Field（CRFs）的辨識模組之訓練文本。並結合法鼓文理學院的人名規範資料庫（Buddhist Studies Person Authority Databases），藉由其中收錄的人名規範資料做為辭典與規則庫，進行序列標記。亦透過 TEI 標記文本中的人名前、後詞綴，提取出僧人人名之特徵詞組。不同於具有歸整書寫形式的工具書，僧傳文本中蘊含大量僧人人名、人名子集、別名，以及翻譯後的外國僧人人名。首先，將利用 CRFs 模型針對二到四個字元的僧人人名與翻譯僧人名字元過長之人名進行初步辨識。配合人名字串切割、字頻統計選出僧人別名，標記出僧人在文本中的簡稱與別名，進而探討 CRFs 在僧傳文本中的效能。

二、文本資料

釋慧皎(497-554)《高僧傳》與釋道宣(596-667)《續高僧傳》蒐羅超過一千名僧人傳記，是中古時期重要且較為完整的佛教傳記資料。同時，僧傳文本所具有的固定敘述格式，也有利於進行命名實體辨識。《高僧傳》與《續高僧傳》不僅是中古時期佛教歷史文化的寶庫，更是探索中古時期中外僧人姓名的重要資料來源。釋僧祐(445-518)所撰之《出三藏記集》記錄中土所翻譯的經律論三藏，並以簿錄體形式編纂，僧祐之前雖有經目問世，然而此書卻是現存最早的佛教目錄，為早期佛教研究提供寶貴的素材。除了緣記、名錄、經序等內容外，此書更收錄不少人物列傳，雖非現存最早的僧人傳記，然可用於檢核僧傳的演變與差異。[1]

（一）資料來源與型態

本文以南朝梁至唐代《出三藏記集》、《高僧傳》、《續高僧傳》三本書為實驗資料。《高僧傳》、《續高僧傳》於法鼓文理學院執行佛教傳記文學計畫時，已為二書標記人名、地名與時間，這些人工標注正好能作為本文模型學習與測試評估之使用。[2]在本次使用的三部僧傳資料，總計有 1098 個僧人傳記和後序。每篇僧人傳記採用文本編碼規範（Text Encoding Initiative）進行人名、地名、時間實體的人工標記[3]，並使用 XML （eXtensible Markup Language）標記語言作為檔案形式。具體言，TEI 規範的標籤名稱以 persName、placeName、date 代表人名、地名與時間。並以前後標籤<persName>人名</persName>、<placeName>地名</placeName>、<date>時間</date>將目標文字夾住。如下圖呈現。

[1] 《高僧傳》、《續高僧傳》以及《出三藏記集》的內容、版本、歷史價值等研究，參看陳垣：《中國佛教史籍概論》，收入《陳垣全集》第十七冊（合肥：安徽大學，2009 年），頁 497-500、517-532。

[2] 佛教傳記文學計畫共標記從南朝梁至明朝的《梁高僧傳》、《續高僧傳》、《宋高僧傳》、《明高僧傳》、《比丘尼傳》、《出三藏記集》、《明僧傳抄》、《補續高僧傳》八部僧傳的人名、地名、時間，並連結至相應的規範資料庫。
佛教傳記文學地理資訊系統 http://buddhistinformatics.ddbc.edu.tw/biographies/gis/interface/。

《佛教傳記文學視覺化平台》標記工作手冊
http://wiki.dila.edu.tw/pages/%E3%80%8A%E4%BD%9B%E6%95%99%E5%82%B3%E8%A8%98%E6%96%87%E5%AD%B8%E8%A6%96%E8%A6%BA%E5%8C%96%E5%B9%B3%E5%8F%B0%E3%80%8B%E6%A8%99%E8%A8%98%E5%B7%A5%E4%BD%9C%E6%89%8B%E5%86%8A

[3] （引：杜正民，佛學數位資源的建置與開展）



然而，標記檔案仍有許多文本以外的訊息，如校勘碼、行號、段落以及關係點（linkGrp）標籤等。在進行人名辨識模型的訓練與測試的實驗中，我們必須對資料集進行前置處理，將其轉換成程式理解的格式，並且加上字符的特徵讓模型學習。

（二）僧傳中的人名

進行辨識前，必須要釐清一下僧傳中標記資料對於人名的定義。本次實驗中，我們對於人名實體採取相對廣泛的定義。其定義為，指涉單一人物之詞彙即視為人名。在大多人名辨識的數位研究中，主要都以完整人名代表為人名挖掘的對象。我們的人名涵蓋了如下表五種有關人名的表述方式。

完整人名	複合人名	官職、封號	人名子集	代名詞
竺法蘭	陳太守	河東郡王	法蘭	汝
釋道宣	蘭師	太后	蘭	其

（三）前處理

文本處理目的是將文本處理成模型能夠學習格式，並讓設計者在該形式上靈活的增加特徵值標註。在文本前置處理中，主要流程有三步驟；第一，去除掉人名、地名、時間以外的標籤，僅留下文字與作為訓練的正確標籤。第二，將文本切成以句子為單筆的序列資料。第三，將文字轉為一行一個字，並給予單一的字相對應的所屬標籤。

文本 TEI 標記檔案中，人名、地名、時間的標籤即是模型所想要辨識出的目標。第一步處理中，我們想要其他代表文本其他資訊的標籤清理掉，僅留下辨識目標。接著，CRFs 是根據句子為一筆訓練資料之單位，所以便需要再將文本做分句。在三部僧傳的資料量下，總共得到 64, 451 句子。



安	B-PER	N	N	B-PER
世	I-PER	N	N	I-PER
高	E-PER	N	N	E-PER
傳	N	N	Y	O
第	N	N	Y	O
一	N	N	N	O
安	B-PER	N	N	B-PER
清	E-PER	N	N	E-PER
，	N	Y	Y	O
字	N	Y	Y	O
世	B-PER	N	N	B-PER
高	E-PER	Y	N	E-PER
，	N	Y	N	O
安	N	N	N	B-PLA
息	N	N	N	I-PLA
國	B-PER	Y	N	E-PLA
王	E-PER	Y	N	O
政	N	N	N	O
后	N	N	N	O
之	N	N	N	O
太	B-PER	Y	Y	O
子	E-PER	Y	Y	O
也	N	Y	Y	O
。	N	Y	Y	O

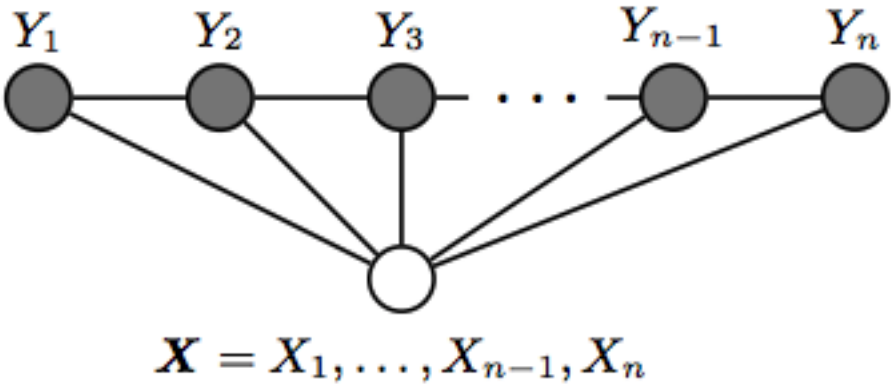
前處理後的文本

三、 條件隨機域（Conditional Random Fields）

（一）CRFs 簡介

本文使用條件隨機域（Conditional Random Fields, CRFs）來解決中古傳記文學人名辨識問題。人名辨識的任務即屬於一種序列標記（Sequence Labeling）問題，在每個字符上皆有一個相對應的標籤。以語言處理的例子說明，句子由為數不一的單字詞所組成，其組成可以視為一串連續單字的序列。CRFs 由 John Lafferty 等人提出後[4]，便經常被使用在解決序列標記的問題上。序列中每個字對應到一個標籤，形成一組由標籤組成的標籤序列，CRFs 是便是透過機率最大化找出一組最佳的序列標籤。

[4] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In International Conference on Machine Learning, 2001.



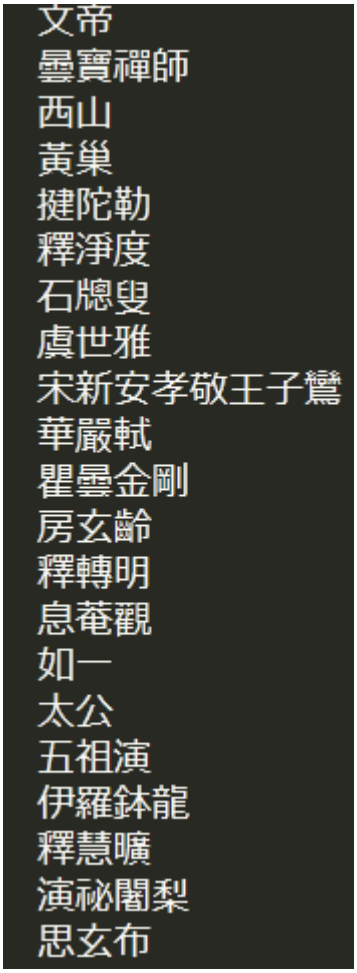
訓練上，CRFs 是一種機率架構的無向圖模型（Undirected Graphical model），透過最大概似估計法（maximum likelihood estimation）來訓練出模型的參數。本文使用 CRFs++ 作為建立模型的工具。此開源工具是目許多人認為最有效率運用 CRFs 工具之一。其允許使用者自行定義特徵集，在運算上也提供（quasi-newton algorithm）和減少記憶體占用之設計，提供更好運算效率。

（二） 特徵值

1、特徵值擷取

訓練資料準備的過程中，特徵值設計是影響辨識效果相當重要的因素。一般而言，在判斷是否為特定命名實體時，會仰賴命名實體的外部特徵（Outside Feature）與內部特徵（Inside Feature）。以人名為例，位於人名位置左右的前綴（Prefix）詞和後綴（Suffix）詞，即是屬於人名實體之外的外部特徵。反之，人名本身個別的字便是命名實體的內部特徵。

本文依照辨識人名實體的方向，設計三組特徵詞組；分別為僧人人名、人名前 2 字以及人名後 2 字。我們藉由八部僧傳，收集所有被標記為人名的詞做為僧傳人名詞庫。僧傳中的人名，其實不單只有僧人與一般常人之名子。在傳記文學的書寫中，作者慣於用僧人名當中的部分名子做為僧。例如使用「蘭」代替「竺法蘭」，以「什」來代替「鳩摩羅什」。又或是字元長度為一的代名詞（吾、汝、公、王）來指稱人物。



在對字符進行特徵值標註時，為確保能更準確標記可能為人名的字詞。我們將選用長度為一以上的人名作為僧傳人名詞庫，總計 13,422 僧人或人物名稱。值得注意的是，人名詞庫中實際上不僅僅是只有僧人的名子，還涵蓋所謂部分人名與人名加上稱號與代名詞。例如，向居士、誠法師、文悟大師、高座、超公、八龍王、安侯道人、其母。本文是基於法鼓文理學院所標記資料的研究，採取較定義寬定的人名。遂亦把各式指稱為人名的詞也包含在內。除了僧人人名詞庫外，我們對三部僧傳中的人名進行外部特徵的抽取。將傳記文本中人名長度為二的前綴詞與後綴詞作為前綴詞組與後綴詞組。

復有	來問	初隨	弟子

2、 IOBES 標記法

句子是由各式的文字所組構而成，其文字可以是命名實體辨識目標或是非辨識目標。我們藉由特定的標記形式，來表示個別的單字是否為辨識目標。舉例說明， IOB 標記型式（Inside, Outside, Begin）即是用來表示命名實體的開始（Begin，簡稱 B）、實體中剩下部分（Inside，簡稱 I）以及非辨識目標（Other，簡稱 O），是以簡稱為 IOBES 標記法，藉此幫助我們表示出一種以命名實體為中心的序列標籤表示法。

依據 TEI 標記檔案所標記的人名、地名、時間標籤，我們定義以 PER 、PLA 與 DATE 是模型想要辨識的目標標籤[5]。

根據命名實體的長度，標籤形式也會有些微不同。例如句子「後遊長安，從什公受業。」，「長安」和「什公」分別為地名與人名實體。我們除了給予每個字相對應的 PER、 PLA 標籤外，還加上 IOB 標記法表示每個字元在地名與人名實體中特定的位置，進而來作為一種更詳細的觀測符號。按照每個中文字元與標籤符號，對應如下表。

觀測符號	後	遊	長	安	，	從	什	公	受	業	。
狀態符號	O	O	B-PLA	I-PLA	O	O	B-PER	I-PER	O	O	O

而在本文中，我們採用更精準表示實體位置的 IOBES 標記法（Ratinov and Roth. 2009.），來表示每個字符自己所屬的標籤。此種規則延伸出結束（End，簡稱 E）、單一實體（Single，簡稱 S），在 Dai, H.-J., Lai 等人的研究中顯示有對訓練模型有更佳的表現[6]。

（三）人名辨識結果

本文設計以《高僧傳》、《續高僧傳》作為條件隨機域（Conditional Random Fields）模型的訓練資料，並以《出三藏記集》作為測試資料。在全部 64, 451 句子中，按照以上劃分。正好切成 90 % 訓練資料、10 % 測試資料。測試《出三藏記集》上達到 86.67 %準確率（precision）、90.17 % 召回率（recall）以及 88.39 % F1-Score。

accuracy	**precision**	**recall**	**FB1**
97.1	86.67	90.17	88.39

在僧傳中，有大量的僧人人名常以人名子集的部分指稱。以安世高傳記為例，全篇「安世高」僅出現一次，「世高」代替全名使用有 26 次，佔該傳人名出現次數最多。另外仍有「安公、安侯、高」的使用敘述。CRFs 的模型辨識過程中，不僅能完整人名辨識出來，其人名子集與部分人名皆能辨識出。

值得注意的是，CRFs 除了長度在二到四之間的人名，如「悉達」、「陳壽」、「馬遷」、「智輪」、「釋法濟」皆能辨識出來外。外國僧名中亦得到不錯的效果。例如來自東印度烏茶

[5] 本文欲討論在中古傳記文本的人名辨識，遂將僅討論人名辨識的成效以及分析原因。

[6] Dai, H.-J., Lai, P.-T., Chang, Y.-C., & Tsai, R. T.-H. (2015). Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. Journal of cheminformatics, 7(S1), S14.

國人的僧人「善無畏」，其原名「善無畏」以及別稱「戍婆揭羅僧訶」、「輸波迦羅」正確地找出來。又或是菩薩、沙門僧之漢譯梵文名「七俱胝尊」、「達摩掬多」、「迦毘羅神」、「般剌剌若」也正確辨識出來。

在 CRFs 模型測試《出三藏記集》之後，我們另外使用《宋高僧傳》做為模型的對照組。測試成果為 %準確率（precision）、 % 召回率（recall）以及 % F1-Score。

accuracy	precision	recall	FB1
94.21	80.59	69.7	74.75

分析兩組實驗表現有所差異，究其因素有二。第一，人名前二字、後二字的特徵值抽取的範圍原本就涵蓋《出三藏記集》。在進行特徵值標記時，人名前後二字的特徵，更能相較正確地標記人名前後地位置。

第二，資料集大小為《出三藏記集》十倍大的 《宋高僧傳》有更多未知人名。在未辨識到的人名當中，以兩個字人名最為大宗。例如「大雄」、「不空」、「痴人」、「大覺」、「曇一」、「靈一」、「小威」。模型在訓練時，這些字符都比較接近非人名常用字。

訓練資料集中包含了作為人名的人稱代名詞，而在《出三藏記集》卻未有人名標記。例如「我當往廣州畢宿世之對」、「不在吾後」、「真得汝矣」、「賊遂殺之」中之代名詞，模型皆有辨識出。但在原始標記資料上並未有人名標籤，以致有此誤差。

最後，在部分的案例例如人名「太子德王」，「太子」與「王」同時皆為特徵詞組內中的特徵值，模型判斷「太子德王」的名子時，容易認為是人名以外的特徵。或是特殊的超長人名「應運統天叡文英武大聖至明廣孝皇帝」。這樣特殊的稱號，在模型中幾乎未見此類人名而無法辨識。

四、中古僧人的人名意義與特徵

探索中古時期的僧傳，不僅可以從中國僧人的法名思考其對印度佛教的接受概況，更可觀察出其對宗教系譜的重要性，是探索中國中古時期佛教歷史的重要面向之一。與此同時，佛教用語也出現在許多六朝人物的姓名之中。[7]若以僧人姓名作為觀察的切入點，僧名往往與僧人的出生地或活動區域相關，是以可以從中探索人物與地方之間的關係，以及展示人物的流動趨向。同時，由於僧名通常會借用某些佛教核心觀念，是以可以從中探索早期佛教被群眾接受或排斥的觀念，以及佛教在中國流傳的特色，與不同區域對某些核心宗教觀念的接受概況。

人名不僅是一種語言現象，更標誌著人物在群體中的社會認同，以及個體生命的變化歷程，是以人名便成為一個值得探索的的重要元素。[8]傳記文學中，傳主的姓名具有錨定文本歸屬的

[7] 宮川尚志，<六朝人名に現はれたる佛教語>，《東洋史研究》，第三卷第六號（1938），頁 503，第四卷第一、二（1938），頁 71、180 以及第六號（1939），頁 538。

[8] Elwys De Stephani, “Names and Discourse”, in The Oxford Handbook of Names and Naming, ed. by Carole Hough, with assistance from Daria Izdebska (Oxford: Oxford University Press, 2016), pp. 53 - 54.

效用，如何識別藏身在大量文字訊息中的傳主姓名，便成為一個重要的事情。由於傳記本身所具有的文學性，致使人名作為一種元素，往往會被許多訊息所掩藏，同時行文過程又會受到書寫筆法影響，出現許多指涉同一人物的別名或簡稱。

實際進入慧皎撰寫的《高僧傳》進行觀察，可以發現：帶有「竺」字稱號者，一般來自天竺或月支，如卷一竺曇摩羅剎；帶有「釋」字稱號者一律為中國人，如卷三釋曇無竭、釋智嚴；姓「康」者，通常出生在天竺或泛稱「西域」之地，偶爾也指涉中國僧人；若名前有「僧加」或「佛陀」者，往往是罽賓人，如卷一僧伽跋澄、僧伽提婆、卷二佛陀耶舍等，但也有來自天竺者，如卷三僧伽跋摩。此外，許多來自天竺、罽賓或西域者，其名字多為翻譯而來，如求那毗地，因此尋求其姓或稱號，便需考慮到原文所具有的特徵。中古時期正處於佛教東傳初始階段，是以此時期的僧人以外國僧人為多，展現於僧名上的特徵，則以長度較長的音譯詞彙為主流。

其後，進入唐宋時期，此時期的僧名組成有三個慣例：其一，僧人的法名可以簡稱，並在名前加「僧」、「釋」，或在名後加「禪師」等稱號。其二，僧人的表字必須全稱，不得簡稱，在表字後可以加稱號。其三，僧人可以連名帶字一起稱呼，名取簡稱，字取全稱，名在前，字在後。[9]據考察，晚明僧名多由四字組成，前兩字為「字」，後兩字為「名」,「名」中的第一個字通常為派輩用字，可以省略。一個僧人的全稱順序是地名/寺廟名+號/字+名。[10]

五、結論

人名辨識命名實體辨識即是針對每個字元給予正確的標籤，以 CRFs 為實作對人名辨識的成果雖受人名未知詞影響。但比較基於人名權威詞庫的語料方式來標記人名，機器學習模型的方法更有可能辨識出因為音譯不同導致有不同音譯的外國僧人名。而若是與模型訓練資料相似的傳記文學，也能藉由此模型得到不錯的結果。未來將針對人名未知詞著手收納更多僧傳做為訓練資料，並且嘗試較具普遍性的特徵詞組，以交叉驗證（cross validation）得到更接近 CRFs 穩定表現之數據。

總的來說，運用 CRFs 所訓練出的模型，能縮短人工標記所耗費的大量時間。與權威詞檔案進行標記來比較，更能彌補對人名子集與代名詞的標記不足。數位人文的標記任務迫需大量且精準的標記工作。我們透過以機器學習的模型挖掘中古傳記文學中的人名，希望能為減少人工標記的渴求，並經過未來調整後擬能運用在更多中古古籍文本。

[9] 如周裕鍇，〈略談唐宋僧人的法名與表字〉，《佛學研究中心學報》，第九期（2004 年），頁 122。
[10] 張雪松，〈晚明以來僧人名號及譜系研究〉，《玄奘佛學研究》，第十五期 2011 年 3 月，頁 249。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

中國古代詩歌自動斷詞、文本標記及其應用

羅珮瑄*　王昱鈞**
政治大學中國文學系博士候選人*
法鼓文理學院佛教學系助理教授**

羅珮瑄（政治大學中國文學系博士候選人）
王昱鈞（法鼓文理學院佛教學系助理教授）

本文旨在探索中文斷詞技術在古代漢語詩歌研究上的應用，從詩歌史的研究議題出發，設計一套適用於中國古代詩歌斷詞與選音的自動化流程，並與其他中古音韻、地理資訊、人物傳記、典故與辭書等資料庫結合，拓展應用的層面，無論是詩歌研究裡專門的古代格律發展問題、抑或基礎研究如詩歌典故、詩歌詞彙、句法與詩法的比較、註解等，期皆能有所助益。

一、中文自動斷詞技術

漢語作為詩歌的載體，相對於印歐拼音語系，其單音獨體的孤立語性質，使漢詩容易形成整齊的節奏，及同音相和的呼應效果；又因漢語具有聲、韻、調的語音元素，其中獨特的聲調叶韻，更易使漢語詩歌產生一種特殊的音樂美感。其構詞不以屈折形態區分詞類，不以固定的語法範疇限定詞序，而是由單字、辭彙、詞序、句構相互間的關係，共同扮演生發語意的功能，因而有極大的靈活性與多元的組合彈性，甚至結合格律規範來變換語序，與散文所呈現出來的樣貌極為不同。

在資訊科技的領域裡，中文斷詞在中文的自然語言處理上，是一項基礎而重要的工作，它是發展其他進階技術——諸如文本標記、事件擷取、進階檢索、問答系統、機器翻譯、語音辨識等等——必須先走過的前置處理。而另一方面，在語言學的領域裡，分詞、建置語料庫、標記語法等工作，也是一項具有漫長傳統的工作，台灣在上世紀 90 年代開始，隨著政府與學校、圖書館、民間基金會推動數位典藏時代的研究計畫，除了現代中文與方言之外，古代漢語也進入數位化的研究視野當中，其中以中央研究院語言學研究所「古代及現代漢語機讀文獻資料之語言分析——文獻語言學之奠基研究(1991.7-1993.6)」和「上古漢語詞彙之蒐集與詞彙庫之建構(1994.7-1996.6)」兩項計畫下所建置的上古（先秦到西漢）、中古（東漢六朝）、近代（唐代以後）漢語語料庫為最大，受限於當時的研究經費、古代文獻數位化的數量和實況、以及不同文獻所需要的文本清理工作的難度，這個大型的漢語斷代語料庫以上古與中古為主，而近代語料庫的完成度則最低[1]。

對於建置語料庫的工作而言，古代漢語的處理方式也與現代漢語不同，古代漢語語料庫選擇從「語法」的角度來進行，是為了斷代的需求；1994 年開始進行的「中央研究院現代漢語平衡語料庫」建置計畫，則從「詞彙」的角度，而有所

[1] 魏培泉、譚樸森、劉承慧、黃居仁、孫朝奮：〈建構一個以共時與歷時語言研究為導向的歷史語料庫〉，《中文計算語言學期刊》第 2 卷 1 期（1997 年 2 月），頁 131-145。

謂分詞規範的研擬。分詞是基於數位典藏的目的，概念切分不同會導致解讀的歧異與混亂，標準一致才能定義並處理知識本體，從而提高搜尋引擎檢索、錯別字更正等應用技術上的準確率。整個討論從 1991 年開始，中華民國計算語言學會（ROCLING）初步訂定學會共用的分詞原則；1995 年及 1997 年 ROCLING 接受中央標準局委託，進行分詞規範的研擬，並由中央研究院執行；1998 年舉辦分詞規範公聽會；1999 年正式通過為國家標準，編號 CNS14366，名稱「資訊處理用中文分詞規範」。其主要訴求為：將具有獨立意義，且扮演固定詞類的字串視為一分詞單位，從而將中文詞彙區分為信、達、雅三級，同時在自動分詞的技術難度上也是遞增的；信級可提供基本資料交換，達級可以進行一般自然語言處理，如建立詞雙連語言模型等，雅級則可作語法和語意上的抽取，如語音合成、語意分析、機器翻譯等[2]。

綜言之，台灣數位人文學發展早期的數位典藏時代，基於建置語料庫的需求而進行有關中文斷詞、分詞規範的探索，技術上則以辭書比對、構詞式來擷取辭彙，再輔以人工校正。縱使在分詞規範上提出了信達雅三級制度，並形成相當程度共識，卻仍然有很大程度是仰賴人為的預設定義，對於其他研究領域的應用而言，仍有一定的限制。

需要制定分詞規範的理由，源於漢字一字一音孤立語的特質，如何處理詞彙邊界與歧義性，是第一個門坎。2000 年左右，當台灣的數位人文學發展進入了所謂數位人文時代，在中文斷詞技術上，開始了更多嘗試，2003 年 M. Li 等人(Li, et al., 2003)提出一種非監督式(unsupervised)訓練，藉由訓練 Naïve Bayes 分類器，來解決中文斷詞的交集型歧義(overlapping ambiguity)問題，實驗結果可達到 94.13% 的準確率；而 1999 年 J. H. Zheng 等人(Zheng & Wu, 1999)則以規則式(rule-based method)來處理組合型歧義(covering ambiguity)，並達到 85 %的準確率；2002 年 X. Luo 等人(Luo, et al., 2002)使用類似於自然語言處理領域中解決「詞義消歧」(word sense disambiguation)的方法，使用 TF.IDF 權重計算的公式，重新定義新的 TF 與 IDF 公式，來解決組合型歧義問題，達到 96.58 % 的準確率。

歧義問題是在已知詞彙的基礎上，探索詞彙邊界時所遭遇的困難；中文斷詞技術的另一難題，是如何挖掘未知的詞彙。中央研究院資訊科學研究所陳克健等人於 1997 年提出關於解決未知詞問題的研究(Chen & Bai, 1997; Chen & Ma, 2002; Ma & Chen, 2003)，初期透過統計與人工判斷並用，從斷詞語料庫偵測單一字元之已知詞，並擷取規則以合併這些單一字詞而成為未知詞；後期則將所有種類未知詞的構詞方式以上下文無關文法(context free grammar)表示出來，並搭配疊代式合併排序法(bottom-up merging algorithm)來解決大部分統計特性低的未知詞擷取問題。其他如 2002 年 Zhang 等人(Zhang, et al., 2002)則使用類似詞性標示(part-of-speech tagging)的作法，稱為「角色標示」(roles tagging)，角色指的是在未知詞的組成成分、上下文以及句子中的其他部分，並且依據句子的角色序列來辨識出

[2] 黃居仁：「CNS14366 中文分詞標準與分詞的實際操作」，http://linganchor.sinica.edu.tw/data/file/LC030909LC03.ppt (2018.11.23.)

未知詞。實驗部分針對中國人名以及外國翻譯名等未知詞做測試，並且達到不錯的準確率以及召回率[3]。

近年來的研究主要趨向於機器學習式的方法來處理中文斷詞，均可見企圖擺脫人為干預所帶來的主觀限制，而更訴求統計與演算來呈現詞彙分布，並進而挖掘出更多的未知詞彙。但基於大規模的訓練語料與商業應用的需求，中文斷詞技術在現代漢語的側重遠大於古代漢語，而出於研究需求所進行的古代漢語研究，則又重散文而略詩歌（或韻文），在斷詞技術上均未考慮過漢語詩歌體式和聲韻上獨特的結構，無法回應中國古代詩歌研究的需求；另一方面，將漢語詞彙視為一個整體，而缺乏時空發展的脈絡，也不易觀察詞彙生成的過程，作為漢語當中一塊特殊的詩歌語彙，本文嘗試運用既有技術，以詩歌史的研究課題出發，為中國古代詩歌重新設計一套斷詞流程。

二、漢詩斷詞分析工具

「漢詩斷詞分析工具」的研發動機，源自於研發團隊的另一項目「漢詩文獻分析系統」與「漢詩格律分析工具」對漢語詩歌聲韻處理的程序，該項目依據現今流傳文獻中關於永明沈約與初唐元兢詩學之記載，統理出其中具有規範性質的格律內容，結合「小學堂漢字古今音資料庫」之中古聲韻材料，建構沈約四病、元兢四病與元兢調聲三術、劉滔二四字不同、近世平仄譜、第四字黏對、五言略頌、新五言略頌等 8 種格律模型、21 種格律規範條件的分析系統，以呈現格律發展史中從永明至初唐的承變關係[4]。面對漢語一字多音與詩歌詞彙的搭配問題，目前這套工具的演算法採取了無罪推定的原則，暫時忽略詞彙邊界和語義的問題，在電腦自動計算過程中，出現二讀以上的情況時，以不與格律規範條件衝突者為優先，來觀察漢詩格律的發展。

當研究推進到可以運用各種交叉運算與統計模型來分析細節的時候，我們意識到在無罪推定原則下所造成明顯的選音錯誤問題，越來越影響我們對於特定格律規範條件生成的時代與作品的論斷，細節上的錯誤已無法用大數據的容錯能力來忽略時，必須尋求其他解決的途徑。研發「漢詩斷詞分析工具」來建置漢語詩歌語料庫，在進行選音演算以前，先確立特定語彙的音韻資料，成為一種可能方案。

開發的初期階段，首先針對五言句的基本句式（2-3）與七言句的基本句式（4-3），將漢語詩歌切分為 2 字串、3 字串與 4 字串，再搭配兩種演算模型，進行詩歌斷詞。兩種演算模型分別為：

[3] 以上技術介紹，參見林千翔、張嘉惠、陳貞伶：〈結合長詞優先與序列標記之中文斷詞研究〉，《中文計算語言學期刊》15 卷 3-4 期（2010 年 9 月），頁 162-165。

[4] 研究成果參見蔡瑜：〈初唐格律發展史觀——以詩格、詩選、詩作交互探索〉，《臺大中文學報》第 59 期（2017 年 12 月），頁 1-54。

1. 以正、逆向的長詞優先法搭配辭書比對，確認漢詩固有辭彙及其選音模型，最後再進行漢詩衍生辭彙的消歧，從而建置具有時代特徵的漢語詩歌詞庫。

辭書部分，我們選用清代康熙年間官方敕編、歷時八年而成的《御定佩文韻府》，根據御制序曰：「嘗謂《韻府群玉》、《五車韻瑞》諸書，事繫於字，字統於韻，稽古者近而取之，約而能博。諸書簡而不詳，略而不備，且引據多誤，朕每致意焉。欲博稽眾籍，著為全書。爰于康熙四十三年夏六月，朕與內直翰林諸臣親加考訂，證其訛舛，增其脫漏，或有某經某史所載，某字某事未備者，朕復時時面諭，一一增錄，漸次成帙。」可見該書乃是集此前詩韻系統韻書之大成，詳加考訂，體制以韻為綱，其下統字、詞、典故、詩例，又以官方之力為之，從而具有詩人創作之際，實務操作上的功能；數位文本方面，付費資料庫如四庫全書電子版、中國古籍資料庫、以及開放資源如中國哲學書電子計劃（Ctext）等皆有全部或部分文本，雖然仍需進行相當多的文本清理、格式轉換和人工校對，但這是現階段較為適合開發「漢詩斷詞分析工具」的辭書基礎。

另外一部為《漢語大辭典》，該辭典由漢語大詞典編輯委員會組織中國四百多位專家參與編寫，從 1975 年開始到 1986 年第 1 卷出版，歷時 11 年，到 12 卷出齊，歷時 18 年。《漢語大詞典》共分為 12 卷，收詞目 375,000 餘條，包括 22,000 餘個單字，350,000 餘條典故，是一全面性的辭書。

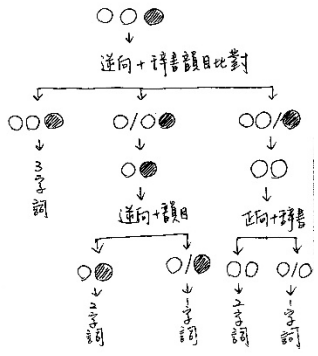
作為提供詩歌創作參考用的《佩文韻府》，在斷詞的成效上，必然優於《漢語大辭典》，且以韻為綱，亦可提供部分的選音建議；然而也正因其以韻為綱的編輯方式，無法全面照顧到所有的詩歌詞彙。《漢語大辭典》雖可補《佩文韻府》之不足，但因為不是針對詩歌創作而編輯的辭書，在詞彙邊界上以詞素為基本單位，往往導致切分過細，不便於提供選音建議，但對於其他基礎的文本研究工作而言，有可能是合適的。

至於模型方面，五言句與七言句的基本句式，包含 2 字串、3 字串與 4 字串，運用正、逆向長詞優先法，其流程分別為：

(1) 2 字串：採用正向長詞優先法與全辭書比對，可斷出 2-0、1-1 兩種模式。

〇〇→正向長詞優先法+全辭書比對→2-1 或 1-1

(2) 3 字串：考慮 3 字串皆在五言句與七言句的末尾，有韻腳的限制，比對流程較為複雜。在長詞優先法上，遇到包含韻腳的字串時，必須採用逆向長詞優先法，不涉及韻腳的字串，則同於 2 字串的處理方式；在辭書的比對上，也必須區分為全辭書與單一韻目兩種，遇到包含韻腳的字串時，必須先判定韻目，再從單一韻目中比對詞彙，不涉及韻腳的字串，則依舊採用全辭書比對。補充說明，這個流程僅限於採用《佩文韻府》作為辭書語料的時候成立。斷詞模型圖示如下：



可斷出 3-0、1-2、2-1、1-1-1 四種模式。

(3) 4 字串：採用正向長詞優先法與全辭書比對，可斷出 4-0、3-1、2-2、2-1-1、1-3、1-2-1、1-1-2、1-1-1-1 八種模式。

故五言句方面，可以切分出 8 種句式；七言句方面，可以切分出 32 種句式。

2. 非監督式斷詞

我們參考 Magistry and Sagot 於 2012 年所提出之非監督式中文斷詞方法[5]，進行詩歌的斷詞。非監督式斷詞方法之基本概念為藉由文字在未標注之語料中統計其結合之良度（Goodness）來衡量字串成詞的可能性。而 Magistry 之非監督式中文斷詞方法採用基於標準化分支熵變化量（Normalized Variation of Branching Entropy, nVBE）作為良度之計算方式。該方法為計算一 n 元詞組從左側與右側減少一字時的訊息熵之變化量以作為良度之估算方式來決定詞彙的邊界。我們在此也同樣設計了兩種模式：

(1) 先將詩集中的五言句與七言句，先進行 2-3 與 4-3 之基本節律的切割後，其再作為 nVBE 斷詞方式之訓練資料。

(2) 完全不做任何前置的切割程序，直接透過學習出之 nVBE 斷詞模型進行非監督式斷詞。

非監督式斷詞模型的基本精神在於，盡量減少人為的干預，透過機器學習的方法嘗試找出適用於詩歌研究議題的斷詞模型。

本文選取《先秦漢魏晉南北朝詩》、《全唐詩》兩部詩歌總集，以及一部晚清文人的詩歌別集《葉德輝詩集》作為研究案例，考慮的理由有二：(1)「漢詩文獻分析系統」上依據逯欽立輯校《先秦漢魏晉南北朝詩》（北京：中華書局，1983

[5] Magistry, P., & Sagot, B. (2012, July). Unsupervized word segmentation: the case for mandarin Chinese. In Proceedings of the 50th Annual Meeting of the ACL: Short Papers-Volume 2 (pp. 383-387).

年）[6]與《全唐詩》（北京：中華書局，1960 年）[7]建置了精校的中國古代詩歌文本，並與「漢詩格律分析工具」架接，來進行中國早期詩歌格律發展更為複雜與細節的研究，故「漢詩斷詞分析工具」在此數位文本基礎上，首先以這兩部詩歌總集為語料，測試兩套斷詞模型的適用性。(2)《葉德輝詩集》為晚清文人葉德輝（字煥彬，號郋園，1864-1927）的詩歌別集，包含《消夏百一詩》、《觀畫百詠》、《古泉雜詠》、《曲中九友詩》、《和金檜門觀劇絕句》、《崑崙皕詠集》、《觀古堂詩集》等七種，凡 27 卷，其中《觀古堂詩集》又分為《南遊集》、《朱亭集》、《歲寒集》、《書空集》、《漢上級》、《于京集》、《還吳集》、《北征集》、《浮湘集》等九種。無論在時代上、性質上、篇幅上、內容上，這部晚清文人詩集皆與《先秦漢魏晉南北朝詩》和《全唐詩》十分不同，它使用了更成熟、更規律的律體詩歌，並以七言為主，同時亦是晚於《佩文韻府》的詩歌別集，在詩歌詞彙的用法上，呈現出更多現代性。

通過兩種斷詞模式，搭配不同性質的辭書和詩歌語料庫，本文進而比較「漢詩斷詞分析工具」的實驗結果，並反思這套斷詞流程的設計背後，必須回應的數位人文學課題，包含「小數據」的可能、詩歌語彙的邊界、以及對詩歌史與格律發展史研究的具體貢獻。

[6] 《先秦漢魏晉南北朝詩》為逯欽立（1910-1973）在明人馮惟訥（1513-1572）《詩紀》與近人丁福保（1874-1952）《全漢三國晉南北朝詩》的基礎上纂輯而成。全書凡 135 卷，囊括詩歌篇什跨越逾千年，作為一部詩歌總集，《先秦漢魏晉南北朝詩》可視為《全唐詩》的前接，為目前蒐集、考證唐前詩歌較為完整且可信的版本。

[7] 《全唐詩》一書為清康熙年間彭定求（1645-1719）、楊中訥（1649-1719）等，於明人胡震亨（1569-1645）《唐音統籤》與清初季振宜（1630-1674）《唐詩》之基礎上，「又旁採殘碑、斷碣、稗史、雜書之所載，補苴所遺」編纂完成。全書共 900 卷，收錄唐人詩作凡四萬八千九百餘首，作者二千二百餘人，是現存唐詩彙輯本中蒐羅詩人作品最為完整的版本。中華書局本《全唐詩》依據揚州詩局刻本進行校點，並錄有日人毛河世寧（1739-1820）《全唐詩逸》（知不足齋叢書本），輯錄《全唐詩》未收詩作共 3 卷，蒐羅唐人詩作更廣，是目前最為周延完整的版本。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

基於史籍所載氣象紀錄之事件分類與時空資訊整合研究

Development of Event Classification and Spatio-temporal Information Integration System based on Meteorological Records in Historical Texts

吳承翰*　吳尚芸**　白璧玲***　蔡融易****
黃詩芸*****　蔡宗翰******　范毅軍*******

國立中央大學資訊工程系碩士研究生*
國立中央大學大氣科學學系大氣組學士**
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***
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&
國立中央大學資訊工程系碩士****
國立中央大學資訊工程系碩士研究生*****
國立中央大學資訊工程系教授&
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******
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*******

基於史籍所載氣象紀錄之事件分類與時空資訊整合研究

Development of Event Classification and Spatio-temporal Information Integration System based on Meteorological Records in Historical Texts

吳承翰[1]、吳尚芸[2]、白璧玲[3]、蔡融易[4]、黃詩芸[5]、蔡宗翰[6*]、范毅軍[7]

Cheng-han Wu, Shang-yun Wu, Pi-ling Pai, Jung-yi Tsai, Shi-yun Huang, Richard Tzong-han Tsai, I-chun Fan

摘要

在異常氣候現象趨於頻繁的今日，氣候變遷帶來的衝擊已逐漸顯著，如何洞悉其因果以尋求因應之道，成為重要的研究議題。欲追溯近代以來氣候災害的發生與影響，在中國早期史料所留下的豐富紀錄中，可發現許多線索，而《中國三千年氣象紀錄總集》(2004)[8]的出版，摘引 7 千多種史籍，並分地區、日期加以彙集，更有助於據以進行歷史時期氣候現象的時空特性分析。由於本團隊近年來致力於研發非監督式文本事件分群技術，將之應用於《明實錄》衛所事件分類上，得到很好的效果，尤其去年在數位人文頂尖國際研討會 Digital Humanities 2017 也順利發表文本事件分類技術的具體應用範例(Tsai et al., 2017)，進一步則思考將此技術應用於其他文本的可行性。因此，依據近年來中央研究院已參照《中國三千年氣象紀錄總集》所建置的「中國歷史氣候資料庫」(P.-K. Wang et al., 2018)，本研究嘗試就地方志與正史等史料有關清代前期(1644-1795)的氣象紀錄內容，建立事件分類處理架構，並開發時空資訊應用系統，提供利於研究者使用之圖文檢索操作介面，期藉由實際研究應用與回饋，拓展文本分析方法於巨量歷史資料分析的效用。

本研究建立氣象分群系統架構如下：

*通訊作者

[1] 國立中央大學資訊工程系碩士研究生 電話: (03)4227151#35203 email: tfssai@g.ncu.edu.com

[2] 國立中央大學大氣科學學系大氣組學士

[3] 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

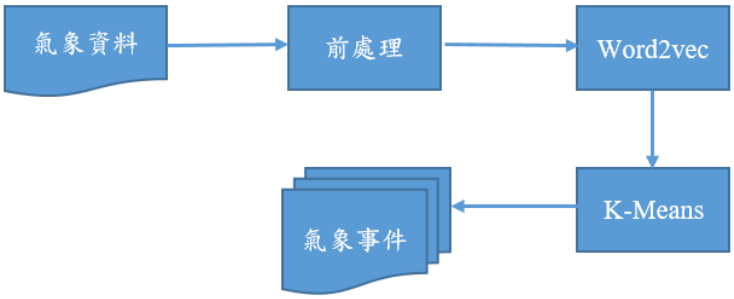
[4] 中央研究院人文社會科學研究中心地理資訊科學研究專題中心、國立中央大學資訊工程系碩士

[5] 國立中央大學資訊工程系碩士研究生

[6] 國立中央大學資訊工程系教授、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@g.ncu.edu.tw

[7] 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

[8] 張德二主編，《中國三千年氣象紀錄總集》，南京：鳳凰出版社，江蘇教育出版社，2004。



圖一 氣象分群系統架構

氣象資料主要是從地方志等史料以人工方式整理出來的資料表，欄位分別為:ID、年份、省、縣市、文本內容及出處(見圖二)，總共 36,123 筆。

	A	B	C	D	E	F
1	ID	年份	省	縣市	文本內容	出處
2	1643-01		北京市		正月庚寅，大風霾。乙卯，是日大風霾，登城西望，埃塵漲天。二月丁卯，大風霾，五色遞變，暗室照之赤如血。三月丙申，大風霾，晝晦，風腥不可觸。丙午，大雷電雨雹。	《明實錄・附錄・崇禎》卷十七
3	1643-02		北京市	昌平縣	春正月朔，大風霾。三月，大風霾，晝晦。	康熙《昌平州志》卷二十六紀事
4	1643-03		北京市	延慶縣	春，大疫。	乾隆《延慶州志》卷一災祥
5	1643-04		北京市	密雲縣	元旦,大風霾，晝晦。三月，大風霾，晝晦。	雍正《密雲縣志》卷一災祥
6	1643-05		天津市		人染異病，十喪八九，親友不敢相吊，俗傳為探頭病。	康熙《天津衛志》卷三災變
7	1643-06		河北省	正定縣	春正月初一日寅刻，赤氣亙天。二月初二日，白虹貫日。二月二十日巳刻，西方有聲，如雷，山出 青氣。	順治《真定縣志》卷四祥異
8	1643-07		河北省	靈壽縣	夏旱。	康熙《靈壽縣志》卷三災祥
9	1643-08		河北省	辛集市	四月十六日未時，有風自西北來，黑氣如墨，相對不辨人物形，移時微雨，乃漸清明。六月間，夜雨如注，南城樓旗杆頂上有火光如拳。	康熙《保定府祁州東鹿縣志》卷九災祥
10	1643-09		河北省	新樂縣	有年。	乾隆《新樂縣志》卷二十災祥
11	1643-10		河北省	深澤縣	三月二十日辰時，風行空際，黑沙蔽天，陰霾至午，日見無光。	康熙《深澤縣志》祥異

圖二 氣象紀錄資料表建檔格式

氣象資料首先經過前處理，將多餘的雜訊濾掉或將同類型的文字替換成統一的符號，這樣可以使得氣象資料更加的適合機器學習語意以及氣象事件的分類。像是我們會將有關天干地支、年分、月份和數字分別替代成 g、y、m、n，以及把地名還有標點符號給去掉，例如□春正月庚寅，以廣東旱，發倉穀七萬石賑之。□經過前處理後會變成□春 mg 以旱發倉穀 n 石賑之□。舉例來說□免田租十分之三□和□免田租十分之七□都是相同的事件，而這兩筆資料經過前處理都會變成□免田租 s□，如此一來就能將這兩筆資料分在同一群。

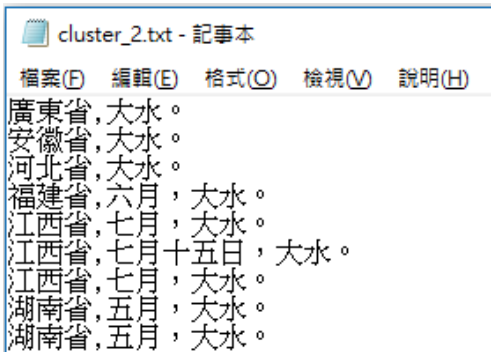
接著我們將氣象資料透過 word2vec 演算法讓機器去將每個文字轉換成 200 維度的向量，而 word2vec 可以計算出向量空間上的相似度，來表示文本語義上的相似度。之後，透過程式利用 word2vec 計算出每筆氣象資料的向量後，再將資料倒入 K-means 做分群的動作。K-means 是一種非監督式的分群演算法，對給定的樣本集按照樣本之間的距離大小，將樣本劃分為 K 個群。K-means 的優點在於原理簡單，收斂的速度快，且分群的效果較優，並可以簡單地依照使用者需求去調整 K 值。若其他文本需要利用此項方法做事件分類，資料格式需比照氣象

資料將一個事件斷成一筆資料，另外，前處理的部分也需要針對不同的文本進行不同的前處理才能達到後續分群的最佳效果。

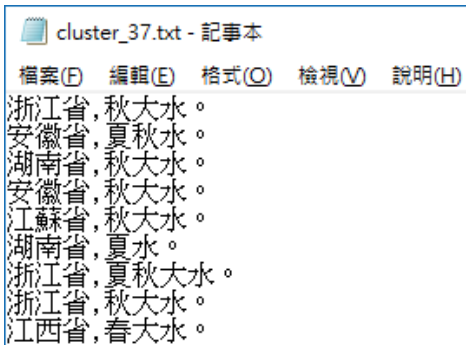
本項分類結果的評估方法，是利用已標註的人工段落共 9,530 筆去計算分群的效果如何，並可以得知此次實驗準確率在 82%，而此次實驗我們將所有資料分成 45 群，詳見表一。由表一可以得知許多群的分類都是相同的，像是以第 2(圖三)、29(圖四)、37(圖五)群為例，這三群皆為水患，但是觀察群內資料可以知道這幾群的特徵都不同，第 37 群比第 2 群多了季節的資訊，而第 29 群又比另外兩群多了有關建築損毀的事件，且由□提下口決□該筆能順利分類成水患可以得知我們 word2vec 的訓練效果不錯。

Cluster_0:農業豐欠	Cluster_15:動物異類	Cluster_30:社會現象
Cluster_1:風類	Cluster_16:乾旱	Cluster_31:降水
Cluster_2:水患	Cluster_17:社會現象	Cluster_32:飢荒
Cluster_3:農業豐欠	Cluster_18:農業豐欠	Cluster_33:水患
Cluster_4:病蟲害	Cluster_19:農業豐欠	Cluster_34:農業豐欠
Cluster_5:降水	Cluster_20:乾旱	Cluster_35:農業豐欠
Cluster_6:水患	Cluster_21:未判定	Cluster_36:降水
Cluster_7:飢荒	Cluster_22:社會現象	Cluster_37:水患
Cluster_8:社會現象	Cluster_23:飢荒	Cluster_38:未判定
Cluster_9:社會現象	Cluster_24:降水	Cluster_39:水患
Cluster_10:水患	Cluster_25:疾病	Cluster_40:乾旱
Cluster_11:水患	Cluster_26:農業豐欠	Cluster_41:降水
Cluster_12:水患	Cluster_27:病蟲害	Cluster_42:農業豐欠
Cluster_13:農業豐欠	Cluster_28:風類	Cluster_43:降水
Cluster_14:社會現象	Cluster_29:水患	Cluster_44:農業豐欠

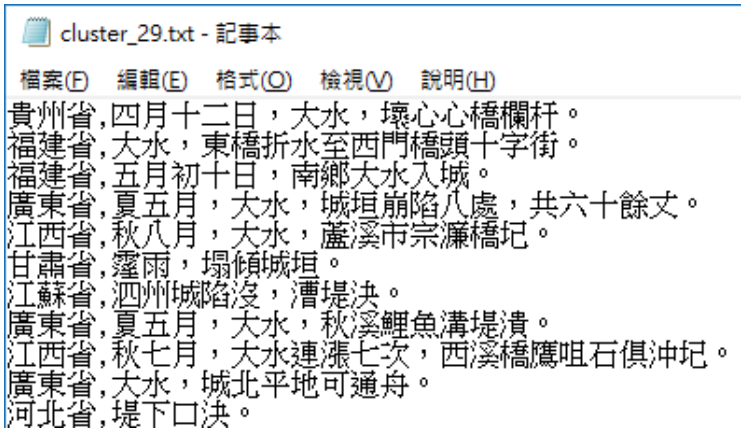
表一 實驗分群結果



圖三 Cluster_2 資料



圖五 Cluster_37 資料



圖四 Cluster_29 資料

依據歷史氣象紀錄的事件發生年份與地點資訊，我們運用「中華文明之時空基礎架構」(CCTS)地名時空對位應用服務，進行坐標定位，使得各類氣象事件可透過地圖介面呈現，進而開發時空資訊應用系統，提供利於研究者使用之圖文檢索操作介面(見圖六)。主要功能包括設置可拉動的時間軸，以及可彈出式的條件篩選視窗；當使用者篩選完條件之後地圖會及時響應，把符合該條件的氣象文本統計出來，並顯示在地圖上。另外，將滑鼠移至圖面地點上，即可看到該地點的詳細資訊，像是氣象類型、次數、詳細文本等等。



圖六 氣象事件時空資訊應用系統介面

自 1550 年至 1770 年間，全球溫度出現下降的現象，氣候學家稱之為「小冰期」（Little Ice Age）。小冰期帶來的影響，除了氣溫下降外，還使得植物生長季節變短，土壤降溫，使糧食作物產量變少，穀物價格上升，造成全球各地頻繁出現飢荒與瘟疫。也因為死亡率上升含，這使全球人口成長率在這段時間減緩。

中國歷史悠久，具有豐富的歷史文獻記載，近一千年來具連續性的巨量歷史文獻，包含地方誌、政府的歷史報告，主要分佈在東部，其中有直接歷史氣候記

載像乾旱、洪水、嚴寒、大雪等，以及間接的氣候環境的記載像人口、飢荒病蟲害等。由於中國位於亞洲東部，具西高東低的梯形貌階梯，也是亞洲氣候的盛行區。青藏高原對北半球大氣環流的動力和熱力作用、西風急流的南北分支在東部輻合，東亞季風和南亞季風環流變化的時空特點，形成較為複雜的氣候環境格局，且目前氣後紀錄中無論是冰蕊、樹輪、湖泊沈積皆尚未有明確「小冰期」發生時間點定義，故本案例將以故本案例將以廣義清初小冰期間 1650 年至 1700 年為探討。

在廣義的中國清初小冰期間 1650 年至 1700 年間，中國歷史氣候文本中究竟如何記錄氣候現象？「小冰期」對當時中國氣候環境如何造成的直接及間接影響？本案例將以直接的氣候環境記載（嚴寒、大雪、乾旱、湖凍、大旱）；以及間接氣候環境記載（嚴寒大雪造成的社會災害、饑荒、人口下降）做為重點討論。

為詳細了解歷史氣候文本記載「小冰期」氣候相關內容，我們以定義資料中「溫度」類別作為篩選標準，時間定義自 1650 年至 1700 年間，共在系統中搜索出 263 筆「溫度」紀錄。在 263 筆紀錄中，共有 68 筆「大寒」、「奇寒」、「劇寒」等低溫類別。由地圖中可看出，發生區域由中國熱帶區延伸至溫帶區域，可了解「小冰河」時期全球嚴寒氣候不僅影響中國中高緯區域，更影響了低緯熱帶區域，臺灣也出現降雪歷史文本紀錄。



圖七 1650 年至 1700 年間氣溫出現低溫的區域

小冰期帶來的影響主要包含嚴寒的氣候，冬天奇寒無比。由歷史氣候文本記錄內容分析中，可以了解在小冰期間嚴寒氣候造成的自然現象以及社會影響。以分群中「降水」類別作為篩選標準，自 1650 年至 1700 年的 50 年間，共有 3,014 筆「降水」紀錄，其中共搜索出 379 筆「大雪」相關資料，文本中直接氣候環境描述不光包含「大雪」現象，更可以觀察同時大雪奇寒現象、連續三日以上大雪

現象；以及間接氣候環境描述，包含大雪造成樹木凍絕現象以及大雪造成鳥獸人畜凍死現象。下表統計出歷史氣候地圖關於「大雪」直接與間接氣候現象，其中共有 379 筆大雪資料中共有 46 筆連續三日以上大雪歷史氣候文本資料；8 筆草木凍絕資料以及 17 筆人畜凍死歷史文本資料。相較於非小冰期（1745-1795 年）50 年間，共 1,618 筆資料，其中共有 119 筆「大雪」的歷史的氣候文本資料，10 筆連續三日以上「大雪」資料；2 筆樹木凍絕/草木萎死；6 筆鳥獸凍死/人畜凍死資料。可以看出現有資料中，1650 年至 1700 年的小冰期大雪造成嚴寒氣候現象與災害頻率高於非小冰期（1745-1795 年）期間。

現象	小冰期（1650-1700）	非冰期(1745-1795 年)
降水資料	3,014	1,618
大雪	379	119
大雪，旬餘/月餘/連旬/連三日以上	46	5
大雪，樹木凍絕/草木萎死	8	2
大雪，鳥獸凍死/人畜凍死	17	6

表二 「降水」類別中小冰期與非小冰期間「大雪」的直接與間接的氣候環境歷史文本筆數

「小冰河」時期全球嚴寒氣候不僅影響造成連續數日大雪現象，我們更發現臺灣歷史氣候文本中出現罕見的降雪紀錄。臺灣位於副熱帶與熱帶氣候區之間，所處緯度較低，高海拔山區，例如玉山、雪山、合歡山比較常發生降雪現象，於 1683 年文本中記載的降雪地點竟包含嘉義縣及臺南市，此氣候現象實屬罕見，除發生在小冰河時期外，其他氣候原因則有待更深入探討。

1683	臺灣省	嘉義縣	降水、溫度	夏五月，大雨水，時霪雨連月，鄭氏土田多沖陷，有“高岸為谷”之歎。冬十一月，始雨雪，冰堅厚寸餘。諸羅有霜無雪，是歲甫入版圖，地氣自北而南，信有矣。

1683	臺灣省	臺南市	降水、溫度	春，鯽魚潭涸。夏五月，大雨水，田園多沖陷。六月，澎湖潮水漲四尺。秋八月壬子，鹿耳門潮水漲。冬十有一月，雨雪冰。是臺地氣暖，從無霜雪，是歲八月甫入版圖，冬遂雨雪，冰堅寸許，地氣自北而南，運屬一統故也。

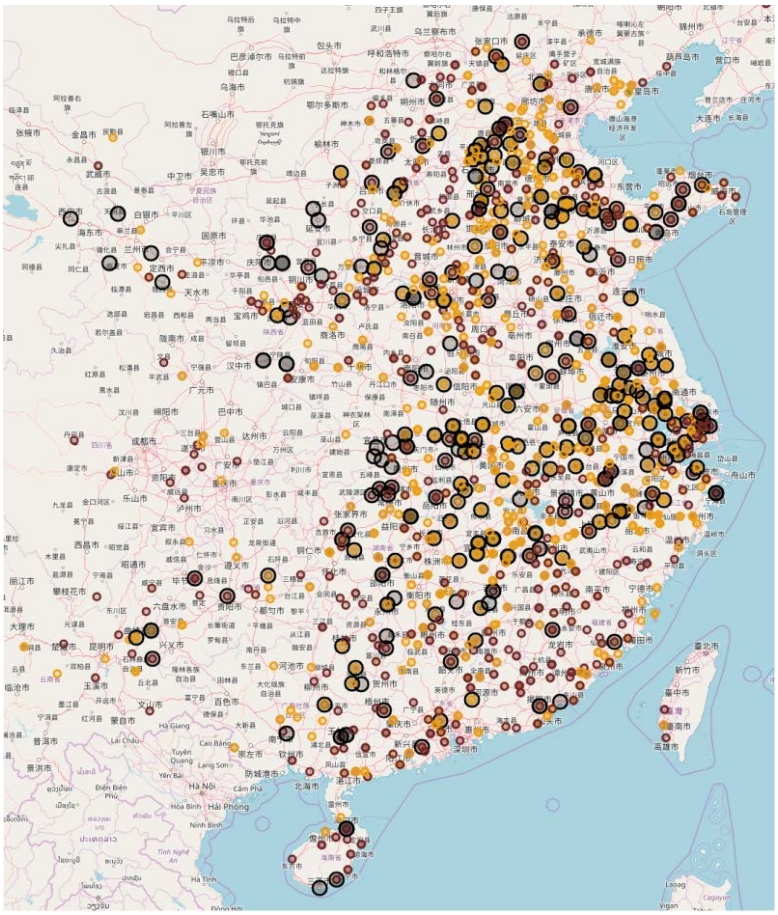
表三　罕見的台灣降雪記錄

在降水紀錄中，除了觀察到「大雪」氣候現象外，在冬季寒冷的區域，河川湖泊可能封凍，故觀察歷史文本河川湖泊相關紀錄，不失為約略窺知氣候冷暖依據。在歷史氣候文本「降水」類別中，共有 2 筆「湖凍」相關資料，可以知道 1690-1692 年間安徽省巢湖市的巢湖；湖北省的廣濟縣湖凍皆長達一個月時間。巢湖市的氣候屬於北亞熱帶濕潤季風氣候，具有雨量適中、光照充足、四季分明的特點，北亞熱帶季風氣候中湖凍現象少見，於 1690 年間發生長達一個月湖凍時間。

1690	安徽省	巢湖市	降水	大旱。十二月初三日湖凍，至正月十五日方解。
1692	湖北省	廣濟縣	降水	大雪，平地三四尺，湖凍彌月不開。

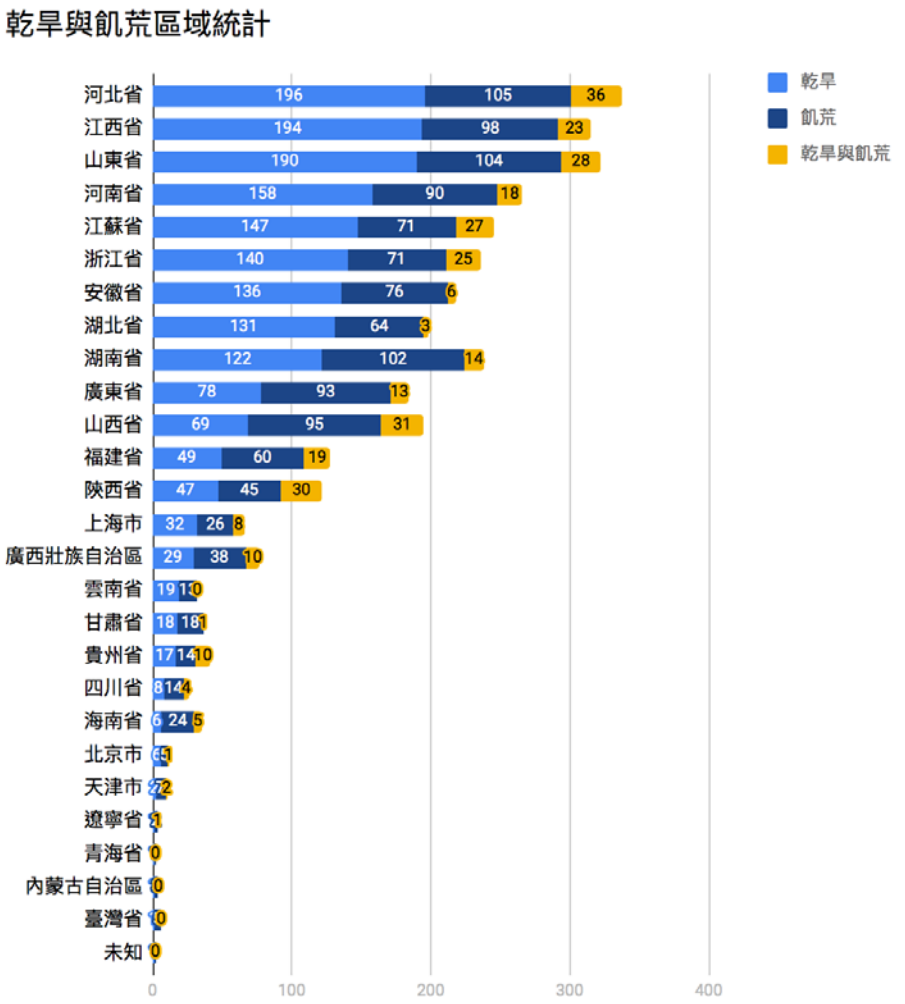
表四　「降水」類別中湖凍相關記錄

全球氣溫下降，導致水氣減少，小冰期間全球乾旱不曾減少，乾旱現象與溫度下降也造成植物生長期縮短，農業萎縮以及飢荒。自 1650 年至 1700 年 50 年間，共有 1,769 筆乾旱紀錄以及 1,241 筆飢荒現象紀錄。由歷史氣候地圖中選取乾旱與飢荒現象，可了解發生區域如圖八：黃色記號為乾旱現象紀錄區域；咖啡色為飢荒紀錄區域；黑色為兩者同時發生紀錄區域。根據地圖所顯示，無論是乾旱紀錄區域、飢荒紀錄區域遍佈中國東部。為了解當時乾旱造成飢荒詳細紀錄，我們篩選出同時符合「乾旱」、「飢荒」兩分類的紀錄共 411 筆。



圖八　1650 年至 1700 年間出現乾旱與飢荒的區域

黃點為乾旱，咖啡點為飢荒，黑點為皆有



表五　「乾旱」與「飢荒」分類區域統計

根據統計結果顯示（表五），以「河北省」發生「乾旱」並造成「飢荒」頻率較高。河北省為中國農業大省，是小麥和玉米的重要產地，因氣候變化造成乾旱、飢荒更甚是食物短缺，可能增加當時社會動亂現象。部分歷史研究也認為，中國明末清初時期社會人口下降、動亂現象，其可能原因為小冰期導致的氣候現象間接造成的社會影響，詳細原因則有待氣候學家及歷史學家共同探討。

現代科技發達，天氣現象依舊難以掌控；長時間的氣候現象則更需要多樣氣候的證據證明，更況是歷史氣候案例探討。在尚未發展大氣科學、氣候觀測技術，尚未重視天氣數據的中國古代社會中，為探討歷史氣候現象，除以冰蕊、樹輪、湖泊沈積了解當時溫度、溫室氣體含量後重建氣候模型外，歷史氣候文本紀錄也是探討歷史氣候的有力證據之一。

藉由將歷史氣候文本紀錄透過 word2vec 演算法學習語意之後，再以 K-means 作分群，將事件做分類，運用「中華文明之時空基礎架構」(CCTS)地名時空對位應用服務，進行坐標定位，建立歷史氣候時空資訊整合檢索系統、製作歷史氣候地圖。儘管目前氣候資料年份僅涵蓋 1647 至 1795 年，無論是資料量、或是分群準確度仍有開發空間，本計畫期待未來建立更完善歷史氣候時空資訊檢索系統，提供相關研究者快速搜尋歷史氣候紀錄，實際將檢索系統應用於其研究中。

References

Chinea-Rios, M., Sanchis-Trilles, G., & Casacuberta, F. (2015). *Sentence clustering using continuous vector space representation.* Paper presented at the Iberian Conference on Pattern Recognition and Image Analysis.

Hammouda, K. M., & Kamel, M. S. (2004). Efficient phrase-based document indexing for Web document clustering. *IEEE Transactions on Knowledge and Data Engineering, 16*(10), 1279-1296. doi:10.1109/TKDE.2004.58

Kotlerman, L., Dagan, I., Gorodetsky, M., & Daya, E. (2012). *Sentence clustering via projection over term clusters*. Paper presented at the Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, Montréal, Canada.

MacQueen, J. (1967, 1967). *Some methods for classification and analysis of multivariate observations.* Paper presented at the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, Berkeley, Calif.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Qian, G., Sural, S., Gu, Y., & Pramanik, S. (2004). *Similarity between Euclidean and cosine angle distance for nearest neighbor queries*. Paper presented at the Proceedings of the 2004 ACM symposium on Applied computing, Nicosia, Cyprus.

Solomon, S., Qin, D., Manning, M., Averyt, K., & Marquis, M. (2007). *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC* (Vol. 4): Cambridge university press.

Tsai, R. T.-H., Lai, Y.-T., Pai, P.-L., Wang, Y.-C., Huang, S. H.-M., & Fan, I.-C. (2017). *WeisoEvent: A Ming-Weiso Event Analytics Tool with Named Entity Markup and Spatial-Temporal Information Linking.* Paper presented at the DH.

Wang, D., Li, T., Zhu, S., & Ding, C. (2008). *Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization*. Paper presented at the Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, Singapore.

Wang, P.-K., Lin, K., Liao, Y. C., Liao, H. M., Lin, Y. S., Hsu, C. T., . . . Ting, T. T. (2018). Construction of the REACHES Climate Database Based on Historical Documents of China. *Scientific Data*, in press.

王苏民, 刘健, & 周静. (2003). 我国小冰期盛期的气候环境. *湖泊科学, 15*(4), 369-376.

张娴, 邵晓华, & 王涛. (2013). 中国小冰期气候研究综述. *南京信息工程大学学报: 自然科学版, 5*(4), 317-325.

竺可楨. (1972). 中國近五千年來氣候變遷的初步研究. *考古學報, 1*, 21.

Facing
the Era of AI + DH

第九屆 數位典藏 與 2018 DADH

9th International Conference of Digital Archives and Digital Humanities

數位人文國際研討會



平行會議 WORKSHOP



數位人文跨域共授課程之群組討論行為模型分析

Analysis of Behavioral Patterns on Group Discussion of Digital Humanities Interdisciplinary Co-teaching Courses

朱志明

國立宜蘭大學資訊工程學系

數位人文跨域共授課程之群組討論行為模型分析

Analysis of Behavioral Patterns on Group Discussion of Digital Humanities Interdisciplinary Co-teaching Courses

朱志明

國立宜蘭大學 資訊工程學系

Email：cmchu@niu.edu.tw

摘要

近年來強調以跨領域學習的創新課程愈來愈多，授課教師在課程設計上，使用分組討論的方式也頗為常見，本研究為了解參與跨域共授課程學生，在群組討論的品質與行為模型，以台灣某大學之通識選修課程「紅樓有夢x科技有愛」跨域共授課程為實驗，該課程為單一學期2學分，共有27位分別來自電機資訊學院、生物資源學院及工學院的學生修課，本研究以行動研究方式進行，將學生分成六組，採異質性分組，每組4至5人，由任課教師指定課後時間及主題，要求學生上網進行群組同步討論，每次1小時共進行3次，將各組的討論內容以量化內容分析（QCA, Quantitative Content Analysis）和滯後序列分析（LSA, Lag Sequential Analysis）等工具進行分析。實驗結果顯示學生在跨域共授課程的群組討論品質良好，各組討論行為皆在本研究的六項分類中表現顯著。

關鍵字：跨域共授統整課程，群組討論，討論品質，行為模型分析

壹、 前言

一、文獻探討

1. 跨域共授統整課程

統整課程的類型或模式有很多種（林怡秀，1999；黃政傑，1997；黃譯瑩，1998；Fogarty, 1991；Jacobs, 1989），就學科而言，可分成單一學科、 跨學科與科際整合，就方法而言，可分成主題及學科統合，各種統整模式中又以 Fogarty 及 Jacobs 二位學者提出之統整課程分類模式較為完整和清楚。Fogarty（1991）提出十種課程統整方式，包含單一學科統整、兩個以上學科間的統整、以及學習者的統整等三類，其中兩個以上學科間的統整有並列式、共有式、結網式、線串式及整合式等五種模式，著重於教材組織的方式；Jacobs（1989）則提出科技整合單元模式（interdisciplinary units model）的課程統整模式，強調應用不同學科的方法及語言，讓學生由主題探究中了解學科間的關係，另外 Beane（1997）的課程統整模式更是超越學科界線，重視增進學生對自我和社會意義的建構，涵化其民主的涵養，因此課程統整模式除了考量單元間與學科間之外，也考量學習者經驗的統整。Beane （1997）特別強調課程統整與知識學科不是敵對的關係，也不能缺乏學科知識的基礎，但是他對課程統整採嚴格的認定，其提出「倘若課程統整不是以主題為學習焦點，而是以學科知識為學習焦點，不論其形式為何，皆不算是統整（integration）」之論述。Erich Jantsch 對「跨學科」的相關概念提出系統性論述，他將學科之間的關係依照跨科際的程度分為五個層次，分別是 Multi-, Pluri-, Cross-, Inter-, Trans- Disciplinarity。Jantsch 根據學科之間需要相互協調統整的程度，由低至高提出這五個概念，從不同學科領域間單純的並列關係，直到跨科際(Transdisciplinarity)代表許多不同學科因為複雜的合作關係而發展成一個多層次的學科知識體系，而學科體系協調研究、創新以及教育三個環節(Jantsch, 1972)。

2. 線上討論與內容分析

線上討論雖然可以和問題導向式學習和專題式學習等各種不同的教學策略結合（Gilbert & Dabbagh, 2005; Kanuka, Rourke, & Laflamme, 2007; Sung, Chang, Chiou, & Hou, 2005），但在實際進行線上討論時，因其內容是動態而複雜且不斷更新又難以掌握，讓教師對於學生在討論時會發生甚麼狀況很難預測和掌握，對於如何方能提升學生的線上討論品質也覺得很吃力（Mazzolini & Maddison, 2007），如果太強調要老師做好學生線上討論品質的把關工作，可能會減弱教師使用線上

討論平台作為提升學生學習成效的動機，而透過分析學生在線上討論的過程、內容和行為模型，可以清楚理解學生在社群討論時，所出現的知識建構情形和同儕協作之過程（Hou, Chang, & Sung, 2007; Hou, 2010），如此的分析結果是在學習科技上的重要發現，還可以為教師提供更佳的教學策略以提升學生的線上討論品質（Hou & Wu, 2011）。有研究指出線上討論時若經常被打斷或插話也會影響討論品質（Hewitt, 2005），也有許多研究使用不同的教學非同步討論策略來探討學習者的行為模型（Hou et al., 2007, 2008），這些研究結合量化內容分析（QCA, Quantitative Content Analysis）的編碼（Coding）和量化分析以及滯後序列分析（LSA, Lag Sequential Analysis）來探討其討論品質（Bakeman & Quera, 1995; Hou, Sung, & Chang, 2009a; Hou, 2010），這種方法可以讓研究者了解學生在線上討論時的知識建構內容，也可以使用統計方式了解學生在固定時間區段的各種討論行為，而序列分析則可以看到視覺化的行為序列模式，如：學生的討論內容在各個編碼的顯著與否（Wang & Woo, 2007）。

二、研究動機

近年來強調以跨領域學習的創新課程愈來愈多，授課教師在課程設計上，使用分組討論的方式也頗為常見，本研究動機為了解參與跨域共授課程學生，在群組討論的品質與行為模型。

三、研究目的

為了解參與跨域共授課程學生，在群組討論的品質與行為，本人利用「紅樓有夢 x 科技有愛」跨域共授課程進行教學實驗，希望能達到以下研究目的：

(一) 探究學生在跨域共授課程群組討論中之討論品質。

(二) 探究學生在跨域共授課程群組討論中之行為模型。

貳、研究方法

在台灣某大學通識選修課程開設「紅樓有夢 x 科技有愛」跨域共授課程，共有 27 位分別來自電機資訊學院、生物資源學院及工學院的學生修課，大學四年級學生有 23 人，佔修課人數 85%，本研究以行動研究方式進行，將學生分成六組，採異質性分組，每組 4 至 5 人，由任課教師指定課後時間及主題，要求學生上網進行群組同步討論，每次 1 小時共進行 3 次，再將各組的討論內容以量化內容分析（QCA, Quantitative Content Analysis）和滯後序列分析（LSA, Lag Sequential

Analysis）等工具進行分析，本研究使用 Bloom 在 1960 年提出認知領域(Cognitive Domain)的教學目標類別予以分類，分別為知識(Knowledge)、理解(Comprehension)、應用(Application)、分析(Analysis)、綜合(Synthesis)與評鑑(Evaluation)等六類，另外再加上其他(Other)共七類，做為量化內容分析編碼之用，以分析學生的討論品質和行為模型。

參、 初步結果

表 1 是將各組課後線上討論內容總計 2,829 則，由受過心理學訓練的研究人員，使用量化內容分析（QCA）工具分類後的結果，另外尚有量化內容分佈圖及行為模型待續。

表 1. 各組之群組討論編碼量化一覽表

組別	E1 知識	E2 理解	E3 應用	E4 分析	E5 綜合	E6 評鑑	E7 其他	小計
1	32	40	12	118	5	3	189	399
2	35	50	38	132	24	11	325	615
3	13	36	23	55	8	6	110	251
4	71	44	38	194	30	28	126	531
5	54	46	45	199	19	24	86	473
6	77	41	48	226	40	32	95	560
合計	282	257	204	924	126	104	932	2829

肆、 討論

學生在跨域共授課程的群組討論品質良好，各組討論行為皆在本研究的六項分類中表現顯著。

伍、 結論

學生在線上討論時，雖有學生擔任助教加入，但為求實驗客觀與公平性，皆無干涉其討論內容，建議未來在實施線上討論時，教師或助教也能加入討論，扮演適時發言和適切引導的角色，相信對於討論品質會有一定程度的幫助。另外因為課程以實做為主，而教學現場有 27 位學生（分成 6 組），但是只有一位教師和助教，導致當有問題的學生較多時教師便無法應付，影響學生學習成效。建議每組搭配一位助教，以隨時處理有問題的學生，而且研究者從現場觀課中發現，對於比較沉悶

但很認真學習的小組，若有助教介入引導，會有較好的學習成效。

陸、 參考文獻

中文文獻

林怡秀（1999）。**國民小學課程統整模式之研究**。國立花蓮師範學院國民教育研究所碩士論文。花蓮，未出版。

黃政傑（1997）。**課程改革的理念與實踐**。台北：漢文。

黃譯瑩(1998)。課程統整之意義探究與模式建構。**人文及社會科學**，**8**(4)，616-633。

英文文獻

Bakeman, R., & Quera, V. (1995). *Analyzing Interaction: Sequential analysis with SDIS and GSEQ*. New York: Cambridge University Press.

Beane, J. A. (1997). *Curriculum integration*. New York, USA: Teachers College, Columbia University.

Fogarty, R.(1991). *How to integrate the curricula*. Arlington Heights. Illinois: IRI/Skylight training and publishing.

Gilbert, P. K., & Dabbagh, N. (2005). How to structure online discussions for meaningful discourse: a case study. *British Journal of Educational Technology, 36*(1), 5–18.

Hou, H. T. (2010). Exploring the behavioural patterns in project-based learning with online discussion: quantitative content analysis and progressive sequential analysis. *Turkish Online Journal of Educational Technology, 9*(3), 52–60.

Hou, H. T. (2011). A case study of online instructional collaborative discussion activities for problem solving using situated scenarios: an examination of content and behavior cluster analysis. *Computers & Education, 56*(3), 712–719.

Hou, H. T., Chang, K. E., & Sung, Y. T. (2007). An analysis of peer assessment online discussions within a course that uses project-based learning. *Interactive Learning Environments, 15*(3), 237–251.

Hou, H. T., Chang, K. E., & Sung, Y. T. (2009b). Using blogs as a professional development tool for teachers: analysis of interaction behavioral patterns. *Interactive Learning Environments, 17*(4), 325–340.

Hou, H. T. & Wu, S. Y. (2011). Analyzing the social knowledge construction behavioral patterns of an online synchronous collaborative discussion instructional activity using an instant messaging tool: A case study. *Computers & Education, 57*(2), 1459-1468.

Jacobs, H. H.(1989). *Interdisciplinary curriculum: Design and implementation*. Alexand ria: Association for Supervision and Curriculum Development.

Jantsch, E. (1972), "Towards Interdisciplinarity and Transdisciplinarity in Education and Innovation," in Innovationfor Educational Research and Center (ed.), Interdisciplinarity. *Problems of Teaching and Research in Universities*, 97-121, Paris, French: OECD.

Kanuka, H., Rourke, L., & Laflamme, E. (2007). The influence of instructional methods on the quality of online discussion. *British Journal of Educational Technology, 38*(2), 260–271.

Martin, S. (2009). Learning to teach science. In K. Tobin & W.-M. Roth (Eds.), World of science education: North America, 567-586. The Netherlands: Sense Publishers.

Mazzolini, M., & Maddison, S. (2007). When to jump in: the role of the instructor in onl ine discussion forums. *Computers & Education, 49*(2), 193–213.

Sung, Y. T., Chang, K. E., Chiou, S. K., & Hou, H. T. (2005). The design and application of a Web-based self and peer-assessment system. *Computers and Education, 45*(2), 187– 202.

Wang, Q. Y., & Woo, H. L. (2007). Comparing asynchronous online discussions and face -to-face discussions in a classroom setting. *British Journal of Educational Technolo gy,38*(2), 272–286.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

Digital Political Science Learning Strategies

Yun-Cheng Tsai* Da-Chi Liao**

Associate Professor, Center for General Education of National Taiwan University*

Professor, Institute of Political Science of National Sun Yat-sen University**

Digital Political Science Learning Strategies

Yun-Cheng Tsai[1] and Da-Chi Liao[2]
Associate Professor[1], Professor[2]
Center for General Education of National Taiwan University[1]
Institute of Political Science of National Sun Yat-sen University[2]

Abstract

Digital Political Science (DPS) is interdisciplinary research in Political Science through the application of Computer Science algorithms and methods, which include Computational Text Analysis, Data Mining, Natural Language Processing, Association Analysis, and Social Network Analysis. Programming is one of the most valuable skills in the DPS. However, the traditional Political Science training is lack computational thinking and programming learning environment. The following third steps help us think about how to build up the learning environment. The first, we need to help students figure out why they need to learn to code. The second, how to choose the suitable programming language. The third, we need to build a strong mentorship relationship. We wish the programming learning environment is full of interdisciplinary people who are willing to help the next generation of programmers.

We provide learning strategies self-taught at home and collaborative learning in the class for the DPS to inspire students to obtain interdisciplinary problem solving and programming skills. During the implementation process, we are mentors and lead students to teach themselves to code R programming language through problem-based learning. Because too many various programming learning resources online to learn, we help students set the online learning path for the DPS questions. Students participate in coding at home and ask questions in the classroom. Learning to code through problem-based learning in the DPS, we design short-term and long-term goals. In the short-term goal, the students can code to collect the Facebook fan page of Members of the Legislative Yuan. Through the goal is achieved, the students learn how to program the R programming language. And for the long-term goal, we wish the students can explore legislation topics of the Legislative Yuan through the application of Computer Science algorithms and methods. Besides, Prof. San-Yi Huang helps us build up a database which provides the dictionary of word analysis used by the exclusive Legislative Yuan, develop relevant keywords, and part of speech and word segmentation methods. In the future, we can scale up the problem-oriented exploration in the DPS and help the students trans to computation thinking.

1 Current Works and Results

We wish the programming learning environment is full of interdisciplinary people who are willing to help the next generation of programmers. However, the traditional Political Science training is lack computational thinking and programming learning environment.

Digital Political Science (DPS) is interdisciplinary research in Political Science through the application of Computer Science algorithms and methods, which include Computational Text Analysis, Data Mining, Natural Language Processing, Association Analysis, and Social Network Analysis. Programming is one of the most valuable skills in the DPS. We provide learning strategies self-taught at home and collaborative learning in the class for the DPS to inspire students to obtain interdisciplinary problem solving and programming skills.

1.1 Current Works

The following third steps help us think about how to build up the learning environment. The first, we need to help students figure out why they need to learn to code. The second, how to choose the suitable programming language. The third, we need to build a strong mentorship relationship.

During the implementation process, we are mentors and lead students to teach themselves to code R programming language through problem-based learning. Because too many various programming learning resources online to learn, we help students set the online learning path for the DPS questions. Figures 1 and 2 are our R programming language self-taught learning guides.



Figure 1: The URL of online learning path for R programming language.



Figure 2: The online learning resources for R programming language.

The students participate in coding at home and ask questions in the class. We use an online free discussion platform (https://discordapp.com/) to record students' questions and encourage students to reply to others questions. Figures 3 is shown the interaction records on the platform. Discord designs for communities, which specializes in the text, image, video and audio communication between users in a chat channel. Discord runs on Windows, macOS, Android, iOS, Linux, and in web browsers. Figure 4 is shown the discussion platform (Discord). The platform helps us establish the collaborative practice in the class quickly.



Figure 3: The interaction records on the online free discussion platform.

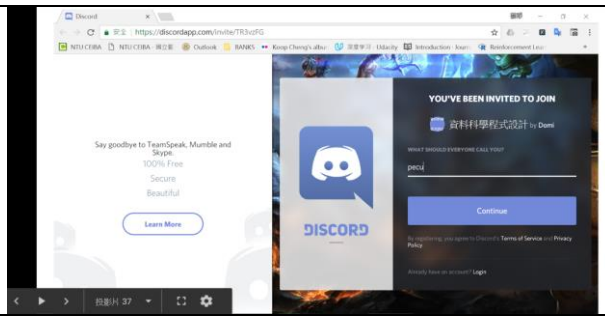


Figure 4: Discord - Free Voice and Text Chat to build up a community quickly.

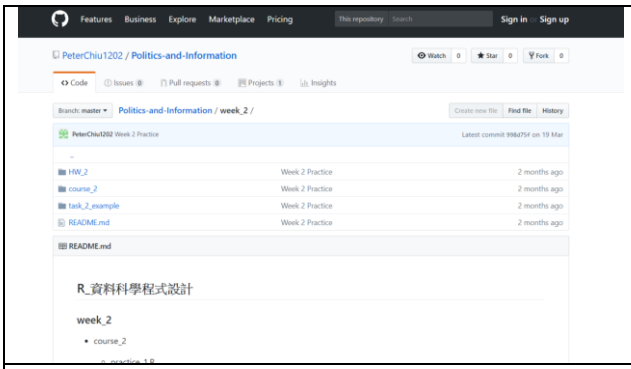


Figure 5: The example codes on GitHub are from Prof. Yun-Cheng Tsai.

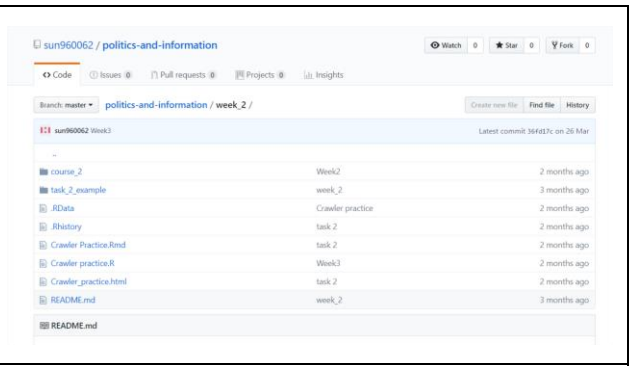


Figure 6: The practice codes on GitHub are from the student.

As Learning to code through problem-based learning in the DPS, we design short-term and long-term goals. Because all learning begins with imitation and learning programming are no exception, we use GitHub (https://github.com/) to share Prof. Yun-Cheng Tsai's example codes to help the students learn by doing. Figure 5 is shown the codes on GitHub (https://github.com/NTU-CSX-DataScience/106-2RSampleCode). Figure 6 is shown the shared practice codes on GitHub from the students. GitHub is a web-based hosting service for version control. It offers all of the distributed version control and codes management. It also provides access control and several collaboration

features such as bug tracking, feature requests, task management, and wikis for every project.

1.2 Results

This course is interdisciplinary learning combining Political Science and Computer Science. The mentors of the class are including the professors from the Institute of Political Science, from the Department of Information Management and Computer Science experts. The mentees are including the Political Science students and the Information Management students. They contribute to the construction of different research fields through understanding and communication. Figure 7 is shown the results of literature reading and sharing in the class.

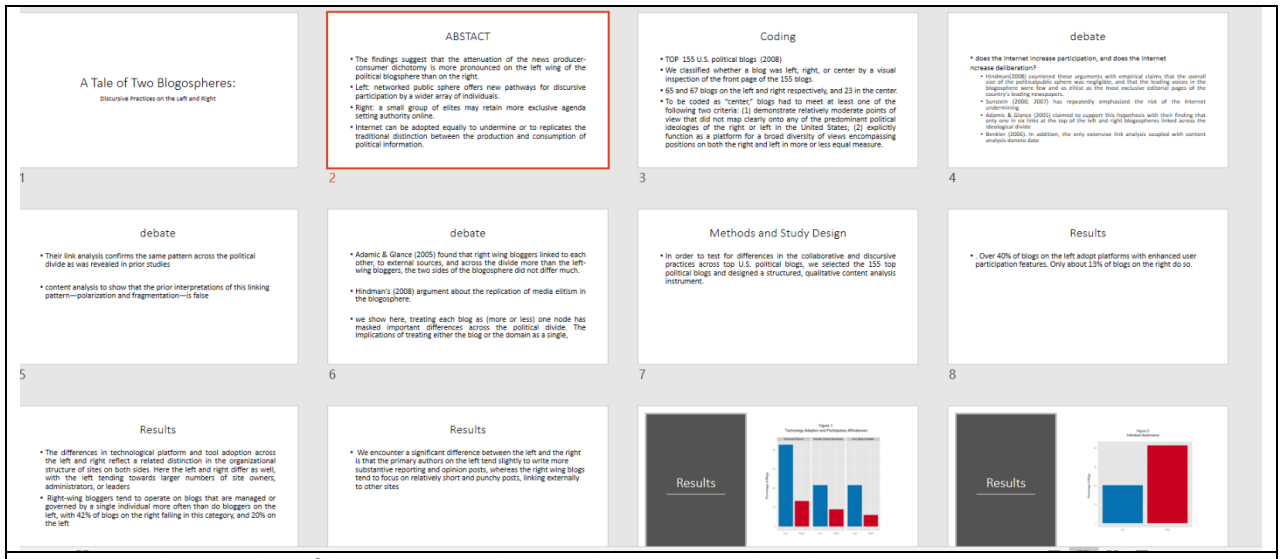


Figure 7: The results of literature reading and sharing in the class.



Figure 8: The students collected the Facebook fan page of Members of the Legislative Yuan and made the data into a word cloud.



Figure 9: The students analyzed keywords the behavior of Members of the Legislative Yuan from the Facebook fan page.

In the short-term goal, the students can code to collect the Facebook fan page of Members of the Legislative Yuan. Figure 8 is shown that the students collected the

Facebook fan page of Members of the Legislative Yuan and made the data into a word cloud. Through the goal is achieved, the students have learned how to program the R programming language.

The students should be familiar with the relevant literature on online politics, especially research and theory involving online study and the use of information technology to analyze political issues. Figure 9 is shown that the students analyzed keywords the behavior of Members of the Legislative Yuan from the Facebook fan page. Furthermore, the students should have the ability to capture information from the social media and initially analyze political texts, including Facebook and PTT, and write papers on preliminary analysis. Therefore, the students can conduct textual analysis of the Legislative Yuan's unstructured data and expect to build a problem-oriented query data and produce information. Figure 10 is shown that the students analyzed the behavior of Members of the Legislative Yuan from the Facebook fan page and produced explainable information.



Figure 10: The students analyzed the behavior of Members of the Legislative Yuan from the Facebook fan page and produced explainable information.

For the long-term goal, we wish the students can explore legislation topics of the Legislative Yuan through the application of Computer Science algorithms and methods.

2 Future Works and Adjustment

Prof. San-Yi Huang leads a research team in College of Management, National Sun Yat-sen University to build up an analysis platform (http://cm.nsysu.edu.tw/~msrc/wp/). The platform provides the dictionary of word analysis used by the exclusive Legislative Yuan, develop relevant keywords, and part of speech and word segmentation methods. Figures 14 and 15 are shown the analysis platform how to management the corpus and the primary text processing results.



Figure 14: The corpus management on the analysis platform.



Figure 15: The primary text processing results on the analysis platform.

We can use the platform to scale up the problem-oriented exploration in the DPS and help the students trans to computation thinking. In the future, we will develop a series of courses to make in-depth knowledge of exciting and valuable political research topics. Hence the students will be constructed to allow students to progress from basic to

intermediate classes. The course group will deepen the course content. The team will propose a group application plan for the DPS program for the new year. Figure 16 is shown the Direct Mail Advertising (DM) of a series of courses to make in-depth knowledge of the DPS program.



Figure 16: The Direct Mail Advertising (DM) of our series courses.

3 Appendix

3.1 The URLs of teaching materials:
https://1drv.ms/f/s!Ah2A7TwrBl04htcOplAXcPjPCDjzbw
https://1drv.ms/f/s!Ah2A7TwrBl04htcPL-dhmXopDgzAkA

3.2 The URL of collaborative learning in the class:
https://1drv.ms/f/s!Ah2A7TwrBl04htcRcFQunRK3UBykhw

3.3 The URL of the class live video:
https://1drv.ms/f/s!Ah2A7TwrBl04htcrVZ9Rg9t5FQwgAw

3.4 The URL of the students' midterm reports:
https://1drv.ms/f/s!Ah2A7TwrBl04htcNsIj5l2U96osQhQ

3.5 The URL of the students' course feedbacks:
https://1drv.ms/f/s!Ah2A7TwrBl04htMO6gindO_agRht2w

Facing
the Era of AI + DH

第九屆 數位典藏與 2018 DADH

9th International Conference of Digital Archives and Digital Humanities

數位典藏與人文國際研討會



海報展示 POSTERS



Research on the Key Components of Construction of Historical Village Archives Digital Repository

Lin-Tao LU Yong-Jun XU

School of Information Resource Management

Renmin University of China, Beijing

Research on the Key Components of Construction of Historical Village Archives Digital Repository

Lin-Tao LU Yong-Jun XU

School of Information Resource Management

Renmin University of China，Beijing

This poster is one of the research of a significant project of the National Social Science Fund of China“Digital Protection and Inheritance of Historical Villages: Theory, Method and Application” (which is hosted by Professor Hui-Ling Feng)

Poster Abstract:

In many countries especially developing countries, under the enormous pressure of rapid urbanization and the basic economic and social development background of rural areas quo, historical villages are obviously eliminating. The historical villages scattered all over the world carry rich historical information on every change in the long history. They are important and natural spiritual homeland for the villagers and even the general public. They have extraordinary significance for shaping collective memory, promoting identification, and building community with shared cultures and history. How to deal with the current embarrassment of the protection and management of historical villages, and to defend the cultural

homeland in the context of the demise of the entity? Archives can provide intertextuality for the village for the social memory at the most. Living preservation of archives is closely related to the construction of social memory. The historical village archives that is comprehensively and intrinsically linked to the village will be the essence of solving the current rescue problems in historical villages. They recorded the history and they reconstruct the villages' memory actively. For a long time, protection and utilization of historical village archives have been at a low level of development- traditional quantitative protection and basic utilization, which has restricted the release of their archival memory attribute and social values. Construct a historical village archives digital repository, rely on digital archival resources, integrate and share with digital means, truly record the historical evolution and development context of historical villages, and properly restore the memory of the villages and villages to promote the preservation and utilization of historical village archives, and protection of historical villages.

It is a systematic project to build a historical village archives digital repository. The current focus on macroscopic ideas and micro-applications lacks serial integration: macro-construction to determine the basic framework for the construction of digital repository, and micro-focus on the implementation of specific aspects of digital repository construction. From the perspective of the system engineering elements, this article fully

examines the complex characteristics and unique needs of the digital archives and digital repository of historical villages. The components are relatively independent deployable units, and can be built through the assembly of cohesive systems engineering. Carding components are the basic elements of the analysis system engineering as well as assembling components is the basic operating mode for building system engineering Conducting research of components can not only clarify the basic content of the progress of the system engineering construction, but also emphasize the coordination and interoperability among the elements of the unit so that the independence and cohesion are fully unified, and the overall process of the macro work is fully integrated on the basis of the micro content. We analyze the key components of the historical village digital archives repository-guidance, strategy, platform, and operation. The key components propose a key component propulsion model that consists of a thrust construction group composed of guidance and strategy components, and a tolerance construction group composed of platform and operation components.

The guidance component is the macro component of the construction of digital repository. The construction of the digital repository is clearly based on the construction goals of the integrated digital resource custody and utilization platform and prominently guided by principles of multiple construction, comprehensive integration, and open application.

The strategy components are connected with each other, and they inherit the macro-direction framework established by the guidance components. Under the framework, the relevant contents of the preliminary preparation of the digital repository construction are detailed, and the platform construction and business management of the application operation layer are followed. Strategy components consist of two types of elements: planning and regulatory. The planning elements focus on the clarity of the preliminary preparations for the construction of digital repository, which involve feasibility studies, planning and design, etc. The regulatory elements concern the institutional construction of digital repository, including the promotion of the layout of the three level-national-provincial-county (city), the joint conference mode of the construction of the project organization system and personnel, technology, funding support detailed requirements.

The platform component is the core component of the “output” of the key component model, and it focuses on the establishment and presentation of the open platform for the sharing of digital archives resources. The first is the content organization, which clearly conveys the sources of digital archives content and defines the content generation of different digital tangible cultural and intangible cultural archives. The second is the technical framework. We focus on building a digital repository technology architecture of “One Database One Port”: a database refers to the core of

the platform of the digital repository—the back-end database. It consists of a digital archive resource database, a metadata database, an index database and a retrieval database which is the basic resource database of the digital repository realizing the storage and specification information output of the historical village digital archives. In addition to the database, a pre-stored digital archives database will be set up, and data cleaning and indexing will be performed on the original historical village archives resources which are created in digital environment or re-processed by digital means like electronic text coding and animation scene recovery to form a standardized digital archive resources. One port refers to the user side of the digital repository which means the web page of the PC and the HTML5 mobile terminal planned to be developed. The front end web page is a resource outlet for the public, and is an access page for the database to implement retrieval services, and implements user interface retrieval input and information output and presents and personalize content descriptions. In the whole work process, we will apply many advanced technologies in platform's building and operation such as 3D modeling, associative aggregation and mobile communications.

The network management of historical village digital archives has become more complex than traditional archives due to the complexity of management objects. The standard specification system for data quality (object data description and metadata description), long-term preservation

(OAIS Model) and intellectual property of digital archive resources (statutory permission principle) is a basic operation component of construction of digital repository. It is an important guideline for standardization specific operations of digital repository. The maintenance management components including the user management system, service terms, and use of accessories like directory, index and electronic reference are also necessary components for ensuring the daily operation of the digital repository. Utilization is an important direction for the construction of digital repository. The utilization of digital archive resources is naturally one of its important business components. The utilization of digital repository mainly includes the use of digital archive resources directly, covering the basic use of information retrieval and serving for digital humanities research especially in the field of sociology and anthropology, providing external interfaces including APIs for professional developer, and general interface for public to self-descript, setting up social media platforms and developing social cooperation.

Comprehensively sorting out the guide, strategy, platform, and operation components involved in the construction of digital repository of historical village archives will be more profound and scientifically promote the pre-planning, mid-term construction, and late-stage operation and maintenance of digital repository with module operations. What's more, it is better to promote the preservation and utilization of historical village

archives resources, to retain the most precious memory landscape of the township land area in China and all over the world.

In this research, we use typical research methods to make the conclusions more reasonable and scientific. On one hand, we comprehensively reviewed related literatures about archival memory views, rural archives and digital archives, digital humanities so that we can analyst research status fully and in-deep. On the other hand, we focused on the "benchmark" of digital archives programs all over the world like American Memory, Venice Time Machine to know what we should do and how we can follow their layout and detailed technologies.

Author's Information

Name: Lin-Tao LU

Address: No.59, Zhongguancun Street, Haidian Distract, Beijing, P.R.C

E-mail: 1203151540@qq.com/ lintao.lu@yahoo.com

Tel-phone: (086)13708263898

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

人物檔案資源數據化研究

以上海交通大學錢學森圖書館館藏為例

孙 逊* 于英香** 张现民***
上海交通大学钱学森图书馆馆员、博士研究生*
上海大学图书情报档案系教授、博士生导师**
上海交通大学钱学森图书馆研究馆员***

人物档案资源数据化研究

——以上海交通大学钱学森图书馆馆藏为例

孙 逊[1] 于英香[2] 张现民[3]

1 上海交通大学钱学森图书馆 馆员 、博士研究生

2 上海大学图书情报档案系 教授 、博士生导师

3 上海交通大学钱学森图书馆 研究馆员

摘要 人物档案因其收藏场所各异，具备档案、文献、文物的多重属性。本文从“资源”层面重新界定其内涵，以上海交通大学钱学森图书馆藏为对象，分析“多级著录”在特藏“629 袋”中的应用，提出人物档案资源数据化的重要性，进一步地选择“钱学森往来书信”这类重要馆藏，阐释如何对其数据化分析乃至“共现图谱”形成的过程，以期实现人物档案资源数字化向数据化过渡的路径。

目次

引言

1.人物档案资源的界定及属性分析

1.1 人物档案资源的界定

1.2 人物档案资源的属性分析

2.人物档案资源著录方法——“多级著录”

2.1 人物档案资源著录存在的问题

2.2 人物档案资源“多级著录”的必要性

2.3 “多级著录”在钱学森图书馆特藏“629 袋”中的应用

3.人物档案资源数据化路径

3.1 人物档案资源数据化的重要性

3.2 实证范例：“钱学森往来书信”

3.2.1 “钱学森往来书信”概况及价值概述

3.2.2 数据化分析与采集

3.2.3 “共现图谱”的生成

关键词 人物档案资源，多级著录，数据化，共现图谱

引言

一直以来，有关人物档案的探讨主要集中其管理方面。具体地讲，包括征集、建档等。随着档案信息化建设的发展，档案数字化工作有了较大幅度的提升，档案信息检索在一定范围内满足了人们的利用需求。然而，大数据时代下，人们对于档案文化知识的需求日益凸显，迫使我们开始关注人物档案的数据化问题。

钱学森是享誉海内外的杰出科学家、中国航天事业的奠基人。2002 年起，上海交通大学档案馆开始征集钱学森相关文献、实物及声像材料。2011 年 12 月 11 日，在钱学森诞辰 100 周年之际，钱学森图书馆正式建成开馆。现在的钱学森图书馆，收藏着钱学森大量的纸质文献，总量约 6 万余件，占全部馆藏的 95%以上。其中，有一类极为珍贵的馆藏，便是钱学森生前亲自剪切、黏贴的剪报，并连同其他多种类型的资料（如，钱学森手稿、与他人的来往通信、会议资料等）所形成的“合集”。钱学森将这些资料，按照社会经济发展涉及的不同主题进行整理、分类，并装进了不同的牛皮纸袋中，形成了其一生中留存世人的最为宝贵的“私人档案”，这样的牛皮纸袋共有 629 个，我们将其俗称为“629 袋”。鉴于“629 袋”是经由钱学森本人收集、整理、分类的原始材料，因此，这批档案是人物档案中的一个典型。本文将紧密结合特色馆藏，重新界定人物档案的内涵，分析人物档案的属性，重点探讨人物档案在包括著录方法在内的数据化路径，以期与同仁交流。

1.人物档案资源的界定及属性分析

1.1 人物档案资源的界定

人物档案的称法不一。在图书馆，有人将其称作“名人手稿”；在博物馆，则将其视为名人遗物或者文物藏品。早在上世纪 90 年代，徐锦[1]曾撰文对“名人档案”、“名人全宗”、“名人档案资料”、“人物档案”等概念作了分析。实则，至今尚未有正式而统一的称谓和定义。2012 年，王利伟[2]对“人物档案基本问题”进行了思考与总结，指出人物档案真实记录了一个人从出生、成长、乃至成才、成功的全过程，因而包含了丰富的知识信息，具有较高的历史与经济价值。对于人物档案收集途径方面，提到建立人物多媒体资料库建设。蒋友钧[3]梳理了人物档案整理的步骤和方法，提出采用“人名-问题”方法进行分类、组卷。不难发现，作为一类专门档案，人物档案的研究主要集中于收集、整理、管理等方面。然而，随着档案管理信息化水平的提高以及人们对档案信息利用需求的不断增长，对于人物档案信息的利用也变得更为迫切。特别是一些人物纪念馆，其承担着展现某一重要人物的辉煌一生以及留存给世人宝贵的精神和物质财富，并弘扬这一人物崇高价值和精神财富的重大责任。例如，美国肯尼迪总统图书馆

[1] 徐锦，<“名人档案”研究综述>，《档案》3，页 31-33。
[2] 王利伟，<关于人物档案基本问题的思考>，《山西档案》3，页 65-68。
[3] 蒋友钧， <人物全宗档案的管理>，《浙江档案》11，页 13-15。

（*https://www.jfklibrary.org/*），实则兼顾档案馆、博物馆的属性，其所藏资源相当丰富，更主要的是它们被详尽而系统地呈现在网络上，供全世界民众访问并使用。这与其著录工作的发展水平不无关系。

肖明[4]认为，资源的一般定义可以表述为，在一定的科学技术条件下，能够在人类社会经济活动中用来创造物质财富和精神财富并达到一定量的客观存在形态。那么，对于信息资源，众多学者倾向于“广义”的信息资源，即信息和它的生产者以及信息技术的集合。人物档案作为一类珍贵的被留存下来的历史见证物，更应从“资源”层面上去理解和研究它。这也是本文研究选择“人物档案资源”的缘由之一。

正如前文所述，人物档案因为在不同的收藏保管机构，其称谓不尽相同。对于图书馆、档案馆，甚至博物馆中的人物档案资源，我们更应该打破图书馆、档案馆、博物馆之间的壁垒，站在更为兼容、宽广的视角对其加以审视。因此，本文认为，所谓的“人物档案资源”是指网络信息时代下，被 LAM（LIBRARY ARCHIVE MUSEUM）收入的对时代发展有较大影响力的人物或者人物群体有关的历史见证物信息以及相关信息技术。当然，这一界定并不是一成不变的，它也会随着认识和实践的不断深入逐渐完善，但是基于跨 LAM 层面之上对人物档案资源进行界定值得参考。

1.2 人物档案资源的属性分析

如上文所述，人物档案资源的内涵正是基于跨 LAM 这一前提下界定的。本文认为，人物档案资源兼具档案、文献、文物等多重属性特征。

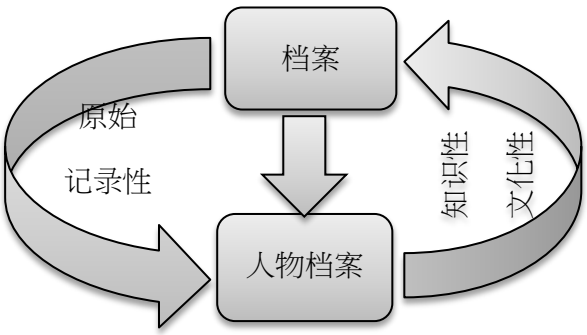


图 1 档案与人物档案的属性关系

学界一般认为，“原始记录性”是档案的本质特性之一，是档案区别于其他事物尤其是相邻事物的独一无二的本质规定性所在。然而，冯惠玲、张辑哲[5]在《档案学概论》中指出，这一本质特性在现实中和许多复杂事物的本质特性一样，并不是表现得很固化，而是具有明显的相对性、动态性特点，是与相关事物及相关因素条件的对照、比较中表现出来的，如文稿、书稿等对象。这将对档案与人物档案的异同提供了比较

[4] 肖明，《信息资源管理》，页 17。
[5] 冯惠玲、张辑哲，《档案学概论》，页 7。

基础。我们发现，人物档案除了具备“原始记录性”之外，还具备文献通常具备的“知识性”。同时，由于国内外人物类纪念馆的存在，亦将人物档案资源也视为文物，甚至对其进行专业鉴定、定级，他们认为其收藏的正是对历史发展的见证物，因此，人物档案也具有“文化性”。综上，我们认为，人物档案资源实则具备多重属性，如图 1 所示。

正是其具备的“知识性”、“文化性”等特性，在对待人物档案资源的著录、挖掘等方面既要遵循保持原貌的基本原则，又要进一步提炼、挖掘其蕴含的数据及知识层面上的内涵。以上海交通大学钱学森图书馆的馆藏为例，从定性上看，我们认为其正是“人物档案资源”的典型。本文将紧密结合馆藏特色，应用专门分析特藏的著录方法，选择“钱学森往来书信”作为实证分析对象，阐明人物档案资源数据化的思路和路径。

2 . 人物档案资源著录方法——“多级著录”

2.1 人物档案资源著录存在的问题

人物档案是某个人在其一生的社会活动中所形成的原始记录。不管如何称谓，其表现形式和内部结构应当保持其原貌。这一原则正是类似于档案学中的“来源原则”。人物档案的研究大多集中于如何征集、管理方面。其实，人物档案整理是人物档案资源开发、利用（资料库检索、展览展示）以及价值发现的基础。整理的过程中，信息著录是对人物档案进行规范控制的方式。2004 年，上海图书馆刘炜[6]等人曾对该馆名人手稿的元数据方案进行过深入探索。2016 年，上海交通大学图书馆张洁[7]等人曾撰文分析科学手稿如何组织的问题。两者的共同点在于，他们均对“资源”的类型及定义作了较好的界定，并且对元数据定义模型等作了探讨，其成果给业界带来了较大的参考价值。然而，不足的是，尚未系统地揭示有关人物的资源整体，未能揭示所谓“森林”的模样。这种现象在国内档案界较为普遍。那么，目前人物档案著录中到底存在哪些棘手的问题？经过分析与研究，本文认为，大致有如下几点：

首先，人物档案的分类界限模糊，存在交叉。论述人物档案分类的文章不少，主要问题在于，将内容分类和形式分类进行交叉、混合分类的现象较为常见。以上海图书馆名人手稿馆为例，他们在其研究中将资源进行了严格分类。当然，这大致是基于上海图书馆所收集到的资源自身情况基础之上的。对于每个机构而言，仍需要结合自身的馆藏实际进行区分。由此可见，针对人物档案的分类问题，我们必须要按照某一标准严格区分，不能有所交叉。

其次，人物档案查全率和查准率低下。目前，国内档案检索的普遍现象是，基于案卷级和文件级的检索并呈现，不能详细检索到其他层级的文件。甚至，因为信息揭示的不足，导致查询结果严重缺失的情况屡见不鲜。当然，这里与档案开放、获取的规则不无关系。这也是国内档案信息检索中存在的普遍问题。

[6] 刘炜，《基于本体的数字图书馆语义互操作》。
[7] 张洁, 彭佳, 郑巧英，<图书馆学与档案学双重视角下的科学手稿组织方法研究>，《图书馆杂志》3，页 74-79。

第三，人物档案描述层级简单。通俗地讲，查询的目标对象“只见树木，不见森林”。人物档案的检索本身具有现实局限性。很多门户上，人物档案的查询和展示，呈现单一化、一维性的特点。我们无法全面、系统地知晓该资源在整个体系中的位置及与其他资源间的关系。当然，有图书情报工作者采用了“资源集合”的概念，甚至会跟关联数据（Linked data）相联系。然而，作为档案本身，其具有自身的结构，应在层级上进行科学而客观地剖析。国际档案著录标准理事会认为，著录信息应能反映档案整理的各个层次，精确表述档案信息的收集、鉴定、分析和组织，以便于揭示档案材料所形成的活动全貌，对其进行信息控制和提供利用。

人物档案的整理类似于普通档案整理，又不完全相同。人物档案通常以全宗形式固定。通常，以大类进行区分。具体的细分类不尽相同。这些划分的标准因馆各异，仁者见仁、智者见智。从管理层级上看，通常仅包括两层，即大多数局限在案卷级和文件级的著录。

2.2 人物档案资源“多级著录”的必要性

多级著录的概念和实践均来源于国外。多级著录思想最早由 Oliver W. Holmes 提出[8]，他提出了多级著录的 5 个层级，分别是存储处（depository）、文件组（records group）、系列（series）、案卷（file unit）、文档（document），依据从上至下的先后顺序进行著录，Olive 的分级整理思想为档案多级著录理论奠定了基础。随着信息技术发展和理论研究的深入，多级著录逐渐由思想发展为理论。在现代档案著录工作中，要达到对档案控制的规范、可靠及有效，就要遵循档案著录的基本原则，即来源原则、尊重全宗原则、反映管理级次原则，这三个方面的内容又被称为档案著录的“马德里原则”。

档案多级著录概念的提出建立在现代档案著录概念的基础之上。就社会利用环境角度而言，在数字化条件下，现代档案利用需求出现了由传统手工向计算机、网络利用扩展，由本地利用向远程利用扩展，由单级、单方面利用向多级、多层面的梯度利用过渡与转变的新特点和趋势[9]。档案多级著录是按照“全宗整理级别结构模式”实行的，它反映了档案全宗是一个不可分散体系的档案的本质与特点，因而档案的多级著录是对全宗理论的具体体现[10]。实现人物档案的多级著录，其必要性有如下三点：

1）实现人物档案的多级著录，可以提高人物档案的检索效率。传统的单级著录，不能体现由全宗到类别，再到案卷，最后到文件之间的逻辑关系。实行人物档案的多

[8] Holmes O W, “Archival Arrangement—Five Different Operations at FiveDifferent Levels”， American Archivist, 27，p. 21-42.
[9] 段荣婷，<论数字时代的档案多级著录>,《档案学研究》s1，页 159。
[10] 李和平, 李超, 刘波，<档案检索系统中的多级著录研究>，《兰台世界》5，页 11-12。

级著录，就可以建立起各著录级别之间的逻辑联系，这样只通过一次检索，就能获得由文件到全宗的全部信息，从而提高了检索速度，节约了检索成本。

2）实现人物档案的多级著录，是保留人物档案原貌的途径。人物档案不同于一般档案，很多档案的形成来自于个人，他们留存下来的原始档案有其自身的逻辑与思考。通常具有明显的多层级关系。对其进行层级分析及多级著录，有助于揭示人物档案本来的结构和信息。

3）实现人物档案的多级著录，是人物档案资源开发与利用的基础。人物档案资源开发与利用，是在详细描述人物档案信息的基础上进行的深度组织和挖掘。

综上，本文认为，人物档案多级著录可以多维、立体、完整地体现人物档案的特点，不仅能揭示其形式特征和内容特征，还包括人物档案的背景信息以及与之相关的关联信息，丰富著录对象的信息内容，为用户直观判断检索结果是否符合检索需求提供依据，同时为用户对其他关联信息的发现创造条件，为进一步扩检、缩减、改检提供了可能。下文将以钱学森的“629 袋”私人档案作为案例，探讨多级著录思想与方法的应用。

2.3 “多级著录”在钱学森图书馆特藏“629 袋”中的应用

“629 袋”时间跨度为 20 世纪 50 年代至 2004 年前夕，是钱学森看书读报、与人学术交往过程中形成的原始记录。它不仅反映了钱学森科学思想的形成轨迹，而且是钱学森留给世人的一笔文化与智力财富。我们对其的整理，即遵循多级整理原则。

首先，“子全宗”的确立。将钱学森捐赠的“629 袋”作为“钱学森全宗”的“子全宗”看待。因为，这批材料是钱学森晚年读书看报的真实写照，呈现的是钱学森与各界学人学术交流的记录。最重要的是，钱学森对其进行系统地收集、整理并分类。仔细审视其中的材料，我们发现它存在着复杂的文件之间的关系。如图 2 所示。

其次，各级的确立，包括“物”层级。简单地说，钱学森捐赠的“629 袋”具备以下几层结构：子全宗级、案卷级、文件级、子文件级。鉴于博物馆的缘故，我们整理时将每袋中的文件视为“物”来揭示。例如，钱学森用于放置相关主题文件的资料袋，在档案著录中，通常将其视为案卷级信息，实际操作中，我们将其也视为“物”，与其中的文件平级，是 collection-level 层级的著录。

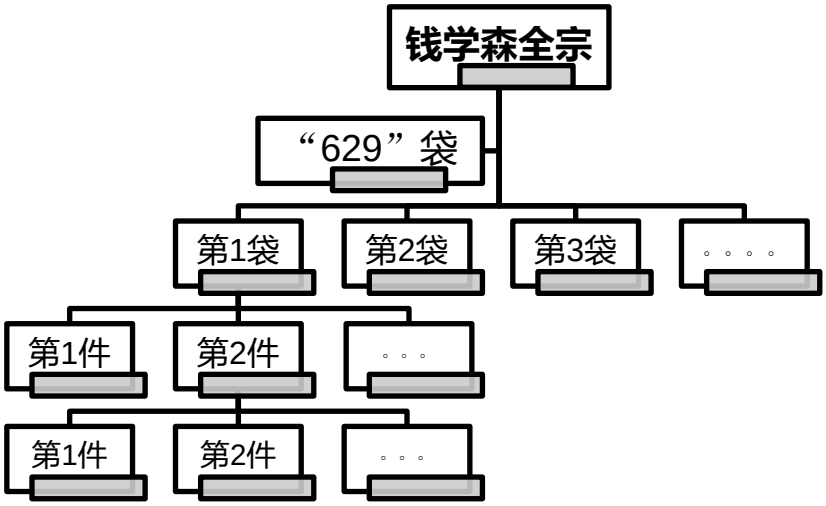


图 2 特藏“629”组织结构示意图

本文将着重阐释著录项目，分为全宗级、子全宗级、文件级、子文件级，以示多级著录在我国人物档案著录中的应用，如图 3 所示。

1）全宗级著录项目的确定。包括：名称、数量、背后故事等。以钱学森全宗为例，单位可为卷，也可为长度，也可为件/套。背后故事与整个馆藏，甚至馆的总体定位、宗旨等密不可分。该层较为宏观。

2）子全宗级著录项目的确定。包括：名称、数量、时间、征集经过、流传经过、背后故事等。以钱学森“629 袋”为例，时间跨度为“629 袋”中最早文献的时间至最晚文献的时间。对其的征集、流传经过的描述是对整个子全宗级的描述，避免对每件文献重复的描述。

3）案卷级著录项目的确定。包括：名称、主题、数量、时间、关联等。以钱学森“629 袋”中每袋为例，我们需要描述其主题、数量，时间跨度以及与其他相关案卷的关联。在“629 袋”中，存在着围绕同一主题的若干案卷组合的情况。

4）文件级著录项目的确定。包括：定名、数量、尺寸、完残状况、关联信息等。以钱学森“629 袋”每袋中的每件文献作为描述对象。需要注意的是，这里的“定名”要参照博物馆文物藏品的定名规范来进行，将为后期的文物定级工作打下基础。对尺寸、数量、完残状况进行一一描述。同样地，文献间存在着大量而复杂的关联关系。这里，需要用关联信息加以描述。照此层级，也能较好地处理档案与文物在整理、管理、描述、利用等方面的矛盾。这里需要指出的是，关于此层的描述，应当将其视为“物”层面作以理解，这与国际著录标准与实践中所强调的 collection-level 一致，也与博物馆一般围绕“物”本身的管理及研究的做法不谋而合。



图 3 特藏“629”元数据描述层级示意图

3 .人物档案资源数据化路径

3.1 人物档案资源数据化的重要性

在大数据概念出现之前，“数据化”并不单独作为一个术语出现，而是以数据化管理、数据化运营等形式出现。美国大数据学家维克托·迈尔-舍恩伯格在他的《大数据时代》一书中提出“一切皆可量化”的数据化概念，即指一种把现象转变为可制表分析的量化形式的过程，这里数据代表着对某件事物的描述，数据可以记录、分析和重组它[11]。另外，一些计算机文献中，常将数据化定义为将数据结构化后存入数据库中，可以利用数据库管理系统对这些数据进行管理和利用。

一些档案馆的数字化工作是通过扫描的形式把一些非数字化信息变成“0”和“1”的数字化形式以便于计算机阅读，但是扫描的数字化的内容是以图像的形式储存，不能通过检索词进行检索，也就是数字文本没有数据化。同时，也因为没有做好数字化档案的著录、标引等工作，因此无法完成主题词、关键词与全文数据库的检索。然而，大数据时代提出的一个新目标是价值服务，因为“拥有大量的数据本身并不会增加任何价值，数据的核心是价值，而驾驭数据的核心是分析。[12]”因此，大数据时代下，我们应当首先重视档案数据化工作，而不仅仅是数字化扫描工作，只有这样，我们才能积累更多“活”的数据，才能“从数据积累的量变过程转化为‘数据智能’的质变过程[13]”。

图书馆在探索智慧博物馆道路上更是重视数据化、数据分析的重要性。从数字人文研究视野看，数据的分析、利用已成为图书馆提升其服务能力的一个重要表现形式，如基于数据的研究热点、趋势分析等。同时，在数字人文研究中，对研究对象的计量化及深入全文的本体化处理，也将生成大量数据，通过对其充分地分析与挖掘以满足用户在内容层级的数据服务需求[14]。人物档案既作为档案的一类分支，加之，前文分析的人物档案具备的“知识性”属性，也理应当更加注重其数据化，以最终实现向知识

[11] 维克托·迈尔-舍恩伯格，肯尼思·库克耶，《大数据时代：生活、工作与思维的大变革》，页 104。
[12] Bill Franks，《驾驭大数据》，页 1-5。
[13] 佚名，<百度李彦宏：大数据已走到技术变革的临界点>，《上海经济》，页 58。
[14] 刘炜，<数字人文的技术体系与理论结构探讨>，《中国图书馆学报》43（5），页 32-41。

化的过渡。数据化的意义是将利用文献的方式从“读”转变为“分析”，其核心是重组文献内容，置入使用者所建立的新的文本或数据结构中，也即文献的结构化[15]。

本文所探讨的“人物档案资源数据化”，不但包括前文讨论的著录方法这一基础内容外，更是以某类具体资源为分析对象，采用数据分析工具对其进行数据提取、数据分析。这里需要说明的是，本研究所结构化的对象并非常规、一般意义上的文献，而是关于某位名人与他人的往来信札。

3.2 实证范例：“钱学森往来书信”

3.2.1 “钱学森往来书信”概况及价值概述

书信是钱学森进行学术交流的重要方式，与钱学森以通信的方式来求教、讨论学术问题的情况非常普遍，且这些通信人主要在科学界。钱老在书信中经常与他们交换看法、探讨学问，其中不乏钱老对各学科的指点、评论，以及与学科发展紧密相关的意见和建议等。2007 年，《钱学森书信》（十卷本）正式出版，共计收入 3331 封。2012 年，《钱学森书信补编》（五卷本）正式出版，共计收入 1980 封。他们是记录钱学森为实现中华民族伟大复兴而奋斗的光辉历程的重要载体，是钱学森科学精神、科学思想和科学方法的重要组成部分。

按照钱老本人对科学的分类体系来看，书信涉及自然科学、社会科学、数学科学、系统科学、思维科学、军事科学、地理科学、建筑科学和文艺科学等 11 个科学大部类。于景元[16]曾撰文再现了钱学森从系统思想出发而提出的现代科学技术的矩阵式结构，从横向上看有这 11 个科学技术部门，从纵向上有三个层次的结构。如图 4 所示：

谈及钱学森书信的价值，研究者们[17]认为，从历史长时段看，钱学森书信不仅属于其个人，同时还属于其所处的时代，具有鲜明的时代特征。通过阅读书信不难看到，钱学森书信还可以作为研究 1970 年代后期至 1990 年代后期中国社会、政治、经济和思想等的重要文献资料。汪长明指出，在某种意义上，钱学森通过书信动态地反映出自己在晚年构建以马克思主义哲学为引领的“现代科学技术思想体系”的具体过程，以及关心青年学者成长、国家后备人才和繁荣科学事业的心路历程[18]。

[15] 赵思渊，<地方历史文献的数字化、数据化与文本挖掘：以<中国地方历史文献数据库>为例>，《清史研究》11，页 27。
[16] 于景元，<钱学森的现代科学技术体系与综合集成方法论>，《中国工程科学》11，页 12。
[17] 钱学森书信研究课题组，<钱学森未刊书信>导言，《钱学森研究》3，页 31。
[18] 汪长明，<钱学森手稿的学术发现与当代价值——国家社科基金重点项目<钱学森手稿整理与研究（1955-2009>结题报告>，《山西档案》9，页 36。

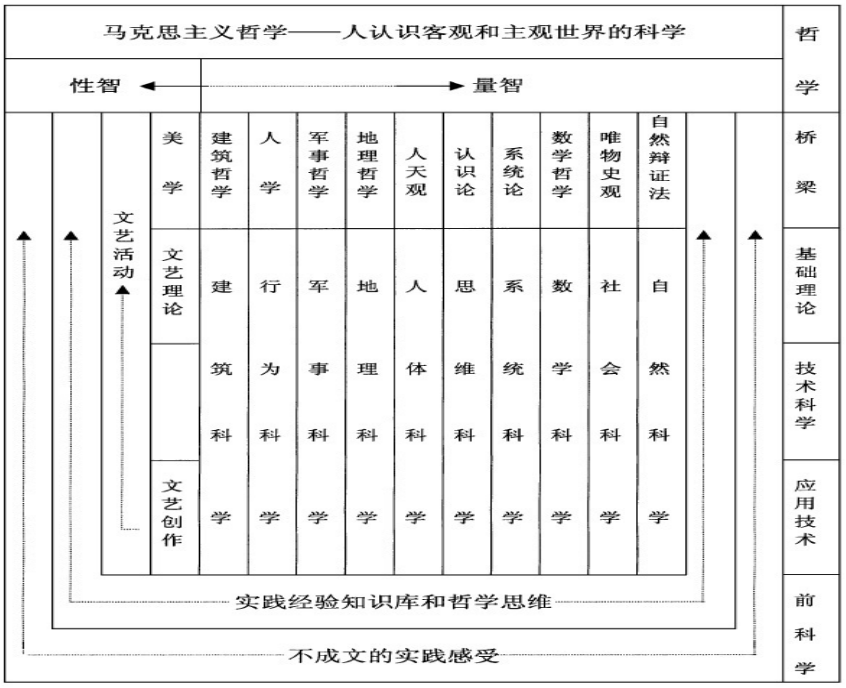


图 4 “现代科学技术”的矩阵式结构

3.2.2 数据化分析与采集

从前文提及的“现代科学技术体系”以及诸多学者探讨的钱学森系统科学思想的学术成果上可见，系统科学思想贯穿整个现代科学技术体系之中。鉴于书信总量的巨大及本研究开展研究的范围所限，本研究拟选定在系统科学这一领域，主要围绕钱学森与其为主导的“七人小组”成员间的书信作为分析对象。简单地以“Y”、“W”、“D”、“S”、“Q”、“T”代称。

我们选择书信往来较为密集的“1970 年代末到 1990 年代末的 20 年间”作为我们分析的时间段。分析对象上，我们选取 “七人小组”中的“Y”，共计约 235 封书信。数据化过程中，我们参考书信现有成果中归纳的关键词索引，结合每封书信的具体内容，手工提取了“发信人、收信人、时间、关键词、相关人物”等信息项。

发	时间	收	关键词								相关人物	
钱学森	1984/10/21	于景元	系统工程								邹瑜；郑拓彬	
钱学森	1984/1/21	于景元	人口问题								陈社英；冯瑞；王寿云	
钱学森	1987/5/5	于景元	复杂系统								方继乾	
于景元	1987/5/19	钱学森	灰色系统理论								方继乾；周政	2219/5/1
钱学森	1987/7/16	于景元									郑应平	
钱学森	1988/4/20	于景元	复杂巨系统	系统学	社会系统工程							
于景元	1988/4/21	钱学森	系统科学	系统学								
钱学森	1988/4/23	于景元	复杂巨系统	还原论	整体论							
钱学森	1988/5/5	于景元	复杂巨系统								周政；薛嘉桥；张劲夫	
钱学森	1988/6/18	于景元	社会经济系统	复杂巨系统							胡传机；马老（马炎）；	
钱学森	1988/6/22	于景元	社会行为学	复杂巨系统	人体学							
于景元	1988/6/27	钱学森										
钱学森	1988/7/4	于景元	复杂巨系统	社会系统	子系统	系统动力学	社会心理学	社会行为子系统			Forrester；沙蓝鲁	
钱学森	1988/8/29	于景元	系统科学								马洪；郑必坚；王寿云	
钱学森	1988/10/8	于景元	社会系统工程	复杂巨系统	社会系统	简单巨系统	协同学				王兆强	
钱学森	1988/10/17	于景元	社会系统	社会经济系统	地理生态系统	人体系统					梁惠民	
钱学森	1988/10/24	于景元									马炎；王存璐；严馨友；王天恩	
钱学森	1988/11/7	于景元									胡传机；马老（马炎）；乔培新	
于景元	1989/1/4	钱学森										
钱学森	1989/1/6	于景元									李家镐	
钱学森	1989/2/10	于景元										
钱学森	1989/3/1	于景元									孙凯飞；马老（马炎）	《社会主义文明的协调发展需要社会主义政治文明建设》
钱学森	1989/3/18	于景元									孙凯飞；马老（马炎）；黄硕风	《社会主义文明的协调发展需要社会主义政治文明建设》
钱学森	1989/5/8	于景元	开放的复杂巨系统	社会系统	系统综合	社会系统学					朱照章	
钱学森	1989/7/17	于景元	简单巨系统	复杂巨系统	协同学						D. L. Stein	
于景元	1989/7/27	钱学森									周政	
钱学森	1989/7/31	于景元	涨落	自组织	开放的复杂巨系	社会系统					吴学谦；M. Eigen	
钱学森	1989/8/2	于景元	系统科学	控制论	系统学	运筹学						
于景元	1989/8/5	钱学森	系统学									

图 5 书信数据化采集信息图

3.2.3 “共现图谱”的生成

以 Y 为分析对象，通过建立矩阵后，在 NODEXCEL 中用 CLAUSET-NEWMAN-MOORE 方法聚类之后分成八类，过滤掉“中心性小于 1”的点，得到如下图 6 所示：

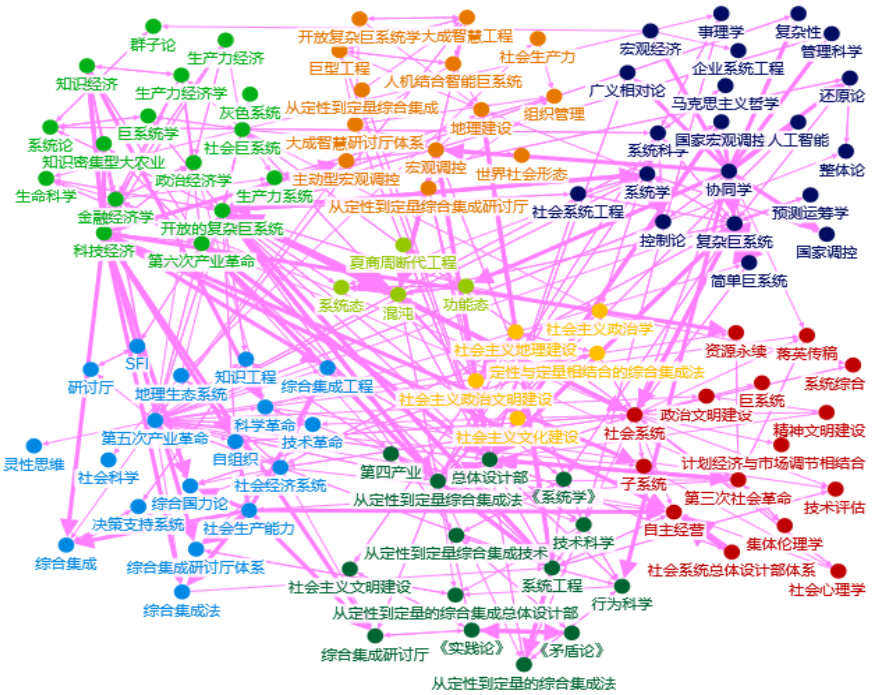


图 6 “共现图谱”

关于此图的分析，需要进一步结合系统科学相关的文献加以结合分析。

结论

人物档案资源具备档案、文献、文物等多重属性。大数据时代下，人物档案资源的数据化将成为必然的趋势，这与人物档案资源自身所蕴藏的属性密切相关。本研究提出人物档案资源“多级著录”方法，旨在引导同行关注档案内部多层次结构，维护人物档案资源间的内部逻辑。如何挖掘其中的内涵与知识，并将其网络化、可视化等将会是人物档案资源数据化发展下一步努力的方向。

参考书目

1 马文峰、杜小勇（2007）。《数字资源整合:理论.方法与应用》。北京：北京图书馆。

2 苏新宁（2014）。《面向知识服务的知识组织理论与方法》。北京：科学出版社。

3 王知津、 李培、李颖（2009）。《知识组织理论与方法》。北京：知识产权出版社。

4 陈燮君、盛巽昌（2005）《二十世纪图书馆与文化名人》。上海：上海社会科学院出版社。

5Marshall P D. 2014.*Celebrity and Power:Fame in Contemporary Culture*. Minnesota ：University of Minnesota Press.

6Tammaro A M, Madrid M, Casarosa V. 2013.*Digital Curators' Education: Professional Identity vs. Convergence of LAM (Libraries, Archives, Museums)*. Berlin：Springer Berlin Heidelberg.

7 肖希明, 田蓉（2014）。< 国外公共数字文化资源整合的现状与发展趋势>，《国家图书馆学刊》23，页 48-56.

8 刘立冬, 刘华银（2007）。<基于对象人物显性知识关联的知识库构建研究——以地方历史文化名人“卞和”文献资源开发为例>，《图书馆学研究》13，页 65-67。

9 周耀林, 章珞佳, 常大伟（2017）。<名人档案信息化建设进展、问题与发展趋势>，《中国档案》1，页 76-78。

10Wendy Davis, Katherine Howard.2013.“Cultural policy and Australia's national cultural heritage: issues and challenges in the GLAM landscape,” Australian Library Journal 62,pp.15-26.

11Evens T, Hauttekeete L.2011.” Challenges of Digital Preservation for Cultural Heritage Institutions,” Journal of Librarianship & Information Science 43,pp.157-165.

12 王巧玲, 孙爱萍, 李希（2017）<私人档案资源建设行为与意识调查研究——以北京市普通民众为调查对象>，《档案学通讯》6，页 53-57。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 12/21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

《1960》

在衝突的年代討論自由與生命的意義

1960

Investigate the meaning of freedom and life in a conflicting era

林愛*　吳苡瑄*　程靖庭*　施保旭**
世新大學數位多媒體設計學系研究助理*
世新大學數位多媒體設計學系副教授**

《1960》：在衝突的年代討論自由與生命的意義

1960: Investigate the meaning of freedom and life in a conflicting era

林愛　吳苡瑄　程靖庭

Ai Lin, Yi-Xuan Wu, Ching-Ting Cheng

世新大學數位多媒體設計學系研究助理

Research Assistant, Dept. of Digital Multimedia Arts, Shih-hsin University

施保旭

Pao-hsu Shih

世新大學數位多媒體設計學系副教授

Associate Professor, Dept. of Digital Multimedia Arts, Shih-hsin University

*通訊作者：施保旭，地址：台北市 116 文山區木柵路一段 17 巷一號 世新大學數位多媒體設計學系，電話：（02）22368225x3299，E-mail:sposh@mail.shu.edu.tw。

摘　要

隨著科技愈來愈進步，人們也愈來愈依賴科技，從前我們可以用紙筆或行動力解決的，現在幾乎都離開不了科技的幫助，也因此這幾年探討科技對人類影響的作品也愈來愈多，而人工智慧，更是一個大家都好奇的題材。

這款遊戲選用的背景是 60 年代，用這個年代的背景來講述一個人工智慧的故事。60 年代是一個暴風雨前寧靜的年代，它正在進行、也即將發生一項大事件，這起事件是一個階級的反轉、文化的斷層，這樣的大事件或許與一個檯面下的人工智慧逃離人類沒有太大的關係，但卻又極其的相似，這是一款藉由一位機器人反抗人類的行為中，去呈現出一個追求自由的過程、去探討一個生命存在的價值與意義的遊戲。

我們的氣氛從開始畫面就開始去營造了，開始畫面會跟著玩家的鼠標移動，看出去的是商店街的一角，並且有著科技面板的介面……諸如此類，使玩家有種以機器人的身分開始遊戲的感覺。而整體的美術風格是走懷舊風加上科技感的融合，操作介面上大量使用綠色與黑色來營造古老的科技感，畫面上也在打光上下了一點功夫，試圖營造懷舊昏黃的感覺。

玩法採取探索解謎與打怪為主要的通關方式，加了探索最主要是為了能使玩家了解到世界觀，也可以有一個通關的目標，而不是盲目打怪。戰鬥方面則採取 ARPG 融合一點 TPS，玩家可以根據自己的喜好來選擇武器，有武器更換的系統。入戰方式則是與傳統遊戲遇怪進入戰鬥畫面類似，但我們在這之上做了一點創新。另外還有升級系統，可以使玩家更有陪著主角成長的感覺。

仿生人到底會不會夢見電子羊呢？我相信這是一個無解的結論，但是我們可以一起玩這款遊戲，一起去看著主角慢慢地找到自己價值、反極權的故事。

關鍵字：科幻、遊戲性、歷史

Abstract

Along with the progress of science and technology, people will become more and more rely on them. We can solve the problems encountered by pen or action before, but now almost without the help of science and technology. Therefore, study on the impact of technology on human work is gathering the attention of researchers in the recent years. And artificial intelligence then more is a subject of everyone’s interest.

The background of this game is set up in the 1960s. It tells a story about artificial intelligence. The 1960s, is an era with calm before the storm. It is going to change and a big incident approaching. This incident brought about a class of inversion, the fault of culture. Such a big incident seemingly have nothing to do with a mesa under artificial intelligence fleeing from the human beings, but actually extremely similar. Our design is a game borrowed from the human behavior by having a robot to pursuit for freedom, to explore the value and significance of the existence of life.

The atmosphere of game was taken into consideration from the beginning of the game. The game screen will move with the player's mouse to see the corners of the street, and it has the interface of the technology-styled panel. It gives the player a sense of starting the game as a robot. And the art style of whole fusion is go nostalgia and a sense of science and technology. Extensive use of green and black on the interface was utilized to create the ancient science and technology style. We also put in a little bit of effort on lighting, trying to create a nostalgic feeling of yellow.

The gameplay of the game is to explore the puzzles and kill monster encountered. Exploring the puzzles is intended to make players to learn about the worldview and to also have a clear picture of the goal, rather than blindly killing the monsters. Fighting has taken a TPS and ARPG style. Players can choose weapons based on his favor. We also have the weapons replacement system. The way of going into fighting is like the traditional games, encountering monster will be brought to the battle scene. But we have done a little more on this innovation. There is also an upgrade system to make players feel more like they are growing up with the main character.

Will robots dream of electronic sheep or not? I believe this is an unanswerable question, but we can play this game together and watch the story of the protagonist slowly finding his own value and resisting the power.

Key words: science fiction, gameplay, history

壹、緒論

首先，我們先對本作品之創作動機與背景以及相關進行方法作一個簡單的說明。接著，則會說明本研究進行的方法與過程。

一、創作動機

最開始在構想這款遊戲時，是以開發團隊成員的興趣融合在一起而創造出來的，一個組員想創作修仙，一個組員想創作民國風格或著類似年代懷舊的東西，又想創作機器人的故事，點子就這麼蹦的出來了。後來再詳細圓了這個故事時，我們突然覺得這樣的故事很適合去融入一個歷史背景……把時間設定從民國初年推延至 1960，那是個文革前蠢蠢欲動的年代，是個社會看似和平卻動盪不安的年代，尤其是那年代的上海，那樣熱鬧的，充滿了繁華的假象，於是我們決定將故事的舞台設定在了上海。而主角身為機器人所要做的事是反抗束縛他的人類，在熱鬧繁華的上海角落裡，不為人知的默默在進行著反抗這件事，就像真實世界裡即將到來的爆發性革命一般，這件不起眼的、默默的事情就這樣在沒有人知道的狀況下開始與結束，並且以開放式的結局來使玩家能自行解讀這件事——甚至是那場即將到來的革命——的意義與否。

這整個故事走向恐怕不是有趣的，我們想表達的是一個反極權、反烏托邦的故事，但會用有趣的遊戲手法來讓玩家融入這個過程，玩家可以選擇不去看沉重的劇情而是去享受遊戲玩法，享受多變的連擊技巧，以及需要思考的解謎流程。另外我們選擇融入科幻的考量，其一是為了使遊戲風格更加多元特別，其二是為了與真實世界區分的更明顯，雖然有意帶入那個時期的歷史背景，但大致上還是希望能架空的；最重要的一點是，我們希望藉由代表新事物的科技、與舊事物的背景的相撞與融合來表現出衝突感。

我們將更注重在於美術氛圍的表達，在遊玩的過程中使玩家可以更加融入這個我們添加了科幻色彩的懷舊年代。

二、創作背景

近年來，市場趨勢走向畫面精美、玩法容易歸類但又有一些小創新的遊戲。2017 年 TGA 遊戲排名上大多是如此類型的大作，例如在小細節精雕細琢給人驚喜的《開放世界 AVG 薩爾達傳說》、給予玩家一個封閉舞台刺激性體驗的《TPS 絕地求生》，又或者是畫面以及風格上別出心裁的傳統《JRPG 女神異聞錄 5》。在多數大作中只有少數一兩款玩法特殊風格獨特的遊戲，而那是特別開出獨立遊戲製作的獎項。（TangBao 2017）

於是我們決定也做一款如此的遊戲，最先是把目標放在 RPG 上頭，再詳細討論後決定在 RPG 這塊做點取捨以融入 AVG，使遊戲更加有趣；畫面精美也是一開始就決定好的，畫面能帶給玩家太多東西，這是我們無法捨棄的。

而技術發展已經到了 VR 漸漸成為各大廠投資的現在，也是有考慮過做 VR 遊戲的可能性。VR 能帶給玩家更融入氣氛的體驗，融入氣氛是我們所追尋的，但是考慮到現今團隊的工作量與對於相關技術的掌握度，還沒有十足把握，所以暫時先把此想法放入下一個階段。

三、研究方法與過程

雖然電腦遊戲的發展相對於其他類型的創作歷史顯得相當短促，但在快速的產業週期壓縮下，遊戲的設計與開發已逐步分工為如下的幾個主要工作：

- 企劃工作：關於整體遊戲的類型、風格、時代背景、故事、角色，乃至關卡挑戰的設計等等。
- 美術工作：包括美術風格的呈現與設計、2D 貼圖的繪製、3D 角色的建模、動作調整、場景的建構，乃至於執行效率的考量等等。
- 程式工作：角色能力的賦予、人工智慧、物件互動機制、系統效能，以及整體的整合。

而隨著開發工作的進展，前述的分工往往是重疊而重複的進行，傳統軟體工程之技術與工具在此領域應用得尚屬十分有限。圖一所示的是這些工作間的大略整合概念圖，圖二則是本作品開發時程表的最上一個階層。

本論文後續的部分將從遊戲類型及作品主題兩個不同的角度進行探討與介紹，然後再針對設計與製作得細節進行說明。

貳、類型探討與相關類型作品比較

本章將從遊戲類型的角度對於本作品做一個定位，並與同類型的產品做一比較探討。

一、類型定義

遊戲的發展在不同的創意下有不同的玩法與挑戰，在探討或介紹一項作品前，先說明他的屬性屬於哪一個類型將可以省下許多的「定位」時間。然而，對於遊戲軟體的分類目前並沒有一個大家公認的標準，同時，隨著技術與遊戲創意的演進，這種分類更是有許多的演變。在此，我們綜合「互動百科」(2016）的遊戲類型分類概述以及 Andrew Rollings and Ernest Adams（2003）在其名著 On Game Design 一書中對於不同類型遊戲設計的詳細分析，將各種遊戲類型的定義以及設計重點簡述如下：

- ACT 動作遊戲：講究打擊的爽快感和流暢的遊戲感覺；
- AVG 冒險遊戲：玩家的主要任務是體驗其故事情節。亦有融合動作遊戲要素的冒險遊戲（A·AVG）新類型；
- FPS/TPS 第一／第三人稱視點射擊遊戲：本類型獨特魅力是給予玩家極其強烈的浸入感。隨著 3D 技術的不斷發展，FPS/TPS 也向著更逼真的畫面效果不斷前進。
- FTG 格鬥遊戲：FTG 誕生於街機，每局對決時間很短，是動作遊戲的戰鬥部分的進一步昇華。透過玩家對戰中的各種判定，搖動搖桿後按下相應的按鍵，使遊戲角色作出對應的動作。
- MUG 音樂遊戲：MUG 遊戲的系統說起來相對簡單，就是玩家在準確的時間內做出指定的輸入，結束後給出玩家對節奏把握的程度的量化評分。這類游戲的主要賣點在於各種音樂的流行程度。
- PUZ 益智類遊戲：PUZ 遊戲多需要玩家對遊戲規則進行思考、判斷。系統表現相當多樣化，主要依遊戲規則制定。由於對遊戲操作不需要太高要求，是現在受眾面最廣的遊戲類型之一。
- RAC 賽車遊戲：RAC 以體驗駕駛樂趣為遊戲訴求。常見的 RAC 系統就是在系統給定的路線內，根據玩家的速度值控制背景畫面的捲動速度，讓玩家在躲避各種障礙的過程中，於限定的時間內，趕到終點。
- RPG 角色扮演遊戲：RPG 遊戲是最能引起玩家共鳴的遊戲類型。能把遊戲製作者的世界完整的展現給玩家。架構一個或虛幻，或現實的世界，讓玩家在裡面盡情的冒險、遊玩、成長、感受製作者想傳達給玩家的觀念。這是一種故事性可以很強的遊戲類型。而進一步將動作遊戲、角色扮演遊戲和冒險遊戲的要素合併的作品則稱為動作角色扮演遊戲，簡稱為 A-RPG 或 ARPG。
- RTS 即時戰略遊戲：RTS 遊戲是戰略遊戲發展的最終形態。玩家在遊戲中為了取得戰爭的勝利，必須不停的進行操作，因為「敵人」也在同時進行著類似的操作。
- SLG 模擬/戰棋式戰略遊戲：SLG 遊戲有多種含義，其一是指對飛機、坦克、機器人……，一般玩家接觸不到的設備的虛擬體驗，以及對某種現實生活的體

驗；另一種則是專門指戰棋式戰略遊戲。

- SPG 體育運動遊戲：SPG 遊戲就是現實中各種運動競技的模擬。有靠玩家點擊頻率，節奏取得勝利的，也有像動作遊戲一樣要求玩家精確操縱的。
- STG 射擊遊戲：現在一般指的是捲軸式射擊遊戲，遊戲性相對簡單許多。
- TAB 桌面遊戲：一般是現實生活中的玩具的模擬，也可以是桌遊的電子版延伸。一般以大富翁類遊戲最具有代表性。
- ETC 其他類型遊戲：無法歸入上述幾種類型的遊戲、玩家互動內容較少的遊戲類型、或作品類型不明確的遊戲類型，皆可歸於本類。
- GAL 戀愛養成遊戲：典型 GAL 遊戲所具備的特性是：以文字（語言）為表達主體、具有一定劇情基礎、以刻畫人物為主要手法、結合多種藝術表現形式給玩家以體驗感的數位載體。

二、本作品類型探討

就戰鬥系統來說，本遊戲是偏向 ACT 與 TPS 融合的遊戲，玩家隨著不同的武器可以切換戰鬥方法，近戰武器則是有連擊系統的 ACT，遠程武器則偏向 TPS。

故事方面屬於 AVG 以及 RPG 的融合，硬要分類的話更偏向所謂的 ARPG，角色的成長系統我們使用了成長樹替代，並且希望玩家能扮演主角，在那個我們精心布置的世界中感受到共鳴。

三、相關類型作品

在創作過程中，我們參考研究了許多現有的作品，本節將舉其中二例做一分析。重點在於我們在其中學到的東西以及在設計上所受到的影響。

（一）、萬代南夢宮遊戲公司的《時空幻境：緋夜傳奇》

萬代南夢宮遊戲公司於 2016 年 8 月 18 日出版的《時空幻境：緋夜傳奇》獲得了 2016 法米通遊戲大獎優秀獎，在此遊戲中玩家扮演一位曾經被最敬愛的親人背叛、導致家破人亡甚至變成半人類的女主角，她的任務將是尋找那位親人並向其復仇。（幻影神月，2017 及金手錶，2017）

《1960》主要是一款 RPG 遊戲，因此我們參考了許多《時空幻境：緋夜傳奇》的系統以及玩法，他是一款相當傳統的日式風格 RPG，卡通式的畫面、精美的人設以及豐富的劇情，帶給喜歡日式 RPG 的玩家一種親切的感覺（參考圖三）。另外，他的戰鬥系統特別別出心裁，打擊感非常爽快，技能招式也有許多玩法及變化（參考圖四）；而故事屬於引導式的敘事方式，讓玩家不會不知道自己該做什麼。

然而，其野外某些場景地圖太大、在劇情與劇情間必須度過的地圖遼闊卻又沒有什麼能讓玩家有驚喜的探索地，有的只有散布在路上的怪物以及不必要的零星寶箱或探索點，使得移動這件事成了一個枯燥乏味的事情。於是我們在設計地圖時便盡量限制在主要場景發生的範圍，設計一個剛剛好的空間並且把解謎要素均勻散佈，希望不會造成玩家的乏味。

（二）、Quantic Dream 的遊戲《Detroit: Become Human》

另外一款是由法國公司 Quantic Dream 於 2018 年 5 月 25 日推出在 PS4 平台的遊戲《Detroit: Become Human》，它以交叉敘述的方式講述了三位仿生人分別在底特律這座城市展開的故事，因此這遊戲有三位主角。（維基百科，2018a 及 TangBao 2018）

這是一款互動式電影的遊戲，畫面非常精美，在人物呈現上也下了很多功夫。故事以三位主角為主軸去敘述一個仿生人們尋找自身存在、追尋自由的故事，並且此遊戲中包含了大量的分支以及支線（參考圖五），每一個決定都會影響到結局，因此，這遊戲有非常多種結局，每個人玩出來都會有屬於自己的故事（參考圖六）。儘管互動式電影的遊戲無法有太多遊戲性，但它卻做到了許多遊戲所追尋的不同體驗感，這也是這款遊戲極大的魅力之一。

我們認為所謂的「不同體驗感」不僅僅只是不同的打法、不同的走法等這種淺顯的東西，而是使玩家能更發自內心的有種「這才是我的個性、符合我、只有我才能玩出來的故事」，一般來說只有開放世界較能達到這種境界。雖然這類型的互動式電影遊戲總是被批評遊戲性差、玩過就不想再玩，但就是這款沒什麼遊戲性的遊戲卻達到了給予玩家不同體驗感這件事，使我們不禁開始思索起遊戲性的定義到底是如何。我們認為我們的遊戲可以在此多下點功夫，因此武器更換系統以及升級系統等都是為了拓展不同道路和玩法而創造出來的。

叁、主題闡述與文獻探討

在這一章，我們將從作品主題的角度介紹本遊戲，並與類似定位的其他類型的創作做一個比較探討。

一、主題闡述

本作品以「反極權」當作此作品的關鍵字，並且進行一系列的延伸。那些在主角以前的其他機器人，所代表的是一個被權力壓榨的角色，同時也是人類為了達到他們的目標所利用的工具，他們短暫的一生至死亡皆是被當作道具一般利用，他們遵守著人類賦予他們的身分度過了一生，但最終還是因失去價值而被拋棄；而主角做為一個機器人卻沒有踏上同樣的道路，在 AI 的自主學習後，他思考起了自身存在的意義，並且決定脫離人類的束縛，他的反抗正是一種「反極權」、「反烏托邦」的象徵。機器人們是否應該在看似安逸的、由人類塑造的「美好校園」內生活，就這樣做著人類給予他們的工作度過一生，還是應該去更廣闊的世界冒險，即使可能機體壞掉後無人維修？

而人類設定為修仙人士除了是想要在科幻上增加玄幻設定，更是為了讓表面光鮮亮麗的形象更加凸顯，在歷史的洪流來看，權力愈大常常使得人愈容易腐敗，遊戲中的修仙人士為了參透成仙的奧秘，製造出了機器人供他們做實驗，即使在了解到機器人們能自主思考後依然沒有給予平等的對待，與宗教神聖不可接觸的形象形成一種反差，使得階級矛盾更加明顯。

另外 1966 年以前的學校、上海也象徵著看似和平，實則暗潮洶湧的社會表象，顛覆階級並不是一件容易的事情，因此主角必須從這裡脫離也必須經過重重考驗以及阻礙。主角在脫離被營造出和平假象的校園後，進入到了象徵社會的上海街頭，在那裡一切都是他在學校中沒有看過的事物，人類也會在街頭發現他的蹤跡並且追捕他，「校園並不是一切的終點，而是一切的開始。」

最後結局或許主角永遠不可能成為人類，但是他最後能以人類的身分活在了這個世界上。

二、文獻探討

我們想表達一種「反烏托邦」的概念，「反烏托邦是烏托邦（utopia）的反義語，希臘語字面意思是「不好的地方」（not-good place），它是一種不得人心、令人恐懼的假想社群或社會，是與理想社會相反的，一種極端惡劣的社會最終形態。」（維基百科，2018b）

反烏托邦主要反映表面上看似公平、和諧的理想社會，實際上卻是極端惡劣的型態，充斥著各種諸如貧窮、極權、階級矛盾等弊病，是一種反人性、倫理及道德的社會，有些故事描述物質文明氾濫使精神受控於物質，因此反烏托邦社會尤其常見於未來設定的作品中，人類在科技發展形成的高水準生活下是虛弱空洞的精神。

反烏托邦基本是出現於虛構的故事中，以警醒人們注意現實世界中有可能出現反烏托邦狀況的議題。

三、相關主題作品探討

不同型態的創作對於主題的呈現方式均有所不同，主要原因來自不同的創作法法所能使用的工具不同所致。文字創作可以詳細而細緻的描述各個角色的內心世界，乃至場景的聲光色變化。影像創作將場景或角色的外觀帶到了閱聽人面前，但也無法將角色的內心世界完整呈現。動畫或電影帶進了「動」的要素，但終究苦於難以將關鍵的畫面深刻的停駐在觀眾心中。遊戲更是難上加難。我們必須將要呈現的主題融入背景故事以及角色設定中，還需兼顧遊戲本身的「遊戲性」。而後者，才是遊戲有別於其他娛樂的最重要關鍵。

本節介紹我們從主題角度所取經的兩部作品。

（一）、阿道斯·雷歐那德·赫胥黎《美麗新世界》

英國作家阿道斯·雷歐那德·赫胥黎於 1931 年創作 1932 年發表的反烏托邦作品《美麗新世界》（Brave New World），是一部以「反烏托邦」為主題的名著，講的是在一個極端科學主義的集權控制下，「野蠻人」主角反抗擔任管理者「文明人」的故事。這本書由一位與社會格格不入的文明人開始，到野蠻人的自殺結束，之中探討了許多關於人性、文化、自由等相關的內容。它與《一九八四》和《我們》並列為世界三大反烏托邦小說。（維基百科，2018c）

《美麗新世界》是影響世界巨大的「反烏托邦」經典名著，裏頭對於自由的詮釋除了文明人們所認定的好的正面意義外，也附帶了野蠻人所追求的痛苦與不確定的不安，他沒有否認被極權統治的安定社會，也沒有推崇為自由付出的野蠻人，最後的自殺以及安於現狀的社會或許都是對兩者的嘲笑以及認同。而此遊戲中，我們結局定為開放式結局，主角逃離了人類後卻又融入了人類生活，這本身就是不能道明哪方是更好的原因，於是未來會發生什麼事將簡單交代，給予玩家想像空間。

（二）、尤金·薩米爾欽《我們》

俄國作家尤金·薩米爾欽於 1920 年完成，1988 年出版的作品《我們》（俄語：Мы），是一部日記體反烏托邦諷刺小說，一般被認為是這類小說的始祖。該作品講述了一個透明的、美麗又蒼涼的世界，所有人的生活都一絲不苟，所有人都以代號為名，所有人都生活在聯眾國規定的幸福之中。而主角原本是個對聯眾國滿懷忠誠的標準代號，但在他遇見了 I-330 之後，開始慢慢有了轉變，整部書我們將陪伴著主角的視野看一個人、一群人從「我們」變成「我」的故事。（維基百科，2018）

《我們》雖然名氣不比並稱的另外兩位高，但卻是最早帶出「反烏托邦」思想的始祖。以代號稱呼、被數字以及「無所不能者」所規範的人類，就好像機器人一般，在美麗透明的玻璃方框裡被監視著走完人生，一出了差錯（有了感性），便可能遭受到被廢棄的命運。這正是我們以機器人為主角的一大原因，機器人一詞無限放大了一個被控制的、沒有感性的人類的概念，而隨著遊戲的進行，機器人也會慢慢地找尋到了自己的意義以及感性。

肆、遊戲結構、場景設計與角色設計

本章將針對作品的設計與實作細節進行論述。

一、故事大綱

那是發生在 20 世紀中的一個不為人們所知的小故事，一群對於修仙的狂熱分子在取得了製作機器人的技術後，便構想了一個瘋狂的計畫——科技化修仙，在看似繁榮熱鬧的上海街頭，這項計畫便這麼在人們看不見的地方悄悄進行著。作為目前最好的實驗品的機器人六十一號——我們的主角陸壹，在日復一日的任務與實驗中不禁開始思索了自身存在的意義，直到遇見了大學新來的教授——黎昕邀請他離開學校，他才開始了解到世界是很多采多姿的，於是他決定答應黎昕的邀請離開大學。

然而實際上黎昕就是製造主角的那群組織的重要人物，這所大學本來就是組織的根據地之一。會慫恿主角逃脫完全是出自於想研究機器人的私心，他帶著他離開大學，並且偷偷的引誘他前往倉庫，使他與前一號機器人見面以測試他最終的實驗——機器人是否具有人性，一切的一切，都是在這位科學家的研究中偷偷的開始，並且也一直在這位科學家的計畫掌握之中。

二、場景設計

場景主要分成四關，對應著四個階段的挑戰。設計參考了一些 60 年代的街頭風格，以及加入一些我們對於機械及宗教的幻想。

商店街-開始有共存的怪物 假熱鬧

倉庫-打敗自己

（一）、場景一：實驗室

此關是屬於新手教學的關卡，教導玩家如何使用基本遊戲操作的場景。我們沒有把這關設定的很大，僅僅只是一間小小的房子，並且也把關卡流程設計的很簡單，除了方便操作教學以外，它的小也象徵了一種封閉性，主角的思想、人類的禁錮，在打開那一刻象徵了一種開始、一種逃離。

作為本遊戲第一個場景，為了使玩家對這遊戲有更加強烈的印象，我們把這間房間的風格做的非常強烈——充滿舊式風格的家具、散落的老式機器、以及牆上貼滿著宗教或科學的海報，我們希望藉由這些充斥著我們作品的軸心風格的裝飾，能使玩家更能在遊戲一開始便感受到這遊戲的氛圍。

美術設定圖稿請見圖七。

（二）、場景二：大學校園

大學是莘莘學子們吸收知識的地方，也是能改變國家未來棟樑思想的地方，由實驗室一出來就是大學內部，代表著所謂的「思想」已經被極權所掌握在手中。另外，主角第一個任務就是打倒大學中的怪物，也是一種剷除異己的象徵。

場景上為了不使這所大學看起來像現實存在的大學，我們在設計這關時只用了很多對於大學的「印象」，佈滿道路的樹與修剪整齊的草、石磚鋪成的小路、圖書館與教室、紅磚屋頂與大片的窗戶等等，將這些要素拼湊起來，成為一個校園該有的樣子。

（三）、場景三：商店街

繁華的商店街，形形色色的人與你擦身而過，60 年代的上海早已是一個空有繁華假象的空殼，有身分地位的人哪個還留在上海？主角一離開和平的空殼，隨即接觸的便是輕飄飄的繁華，而在暗處依然有許多不可告人的事。

商店街的設計參考了許多 60 年代上海的街景，也參考了許多老街的排列方式。最後選擇呈現一個略微繁華的街道，街道上會有一些巴士、一些較現代的車子，以及用雜亂的招牌呈現出的繁華感。另外此關主要目標為潛行，因此設計了許多小巷子供玩家躲避。美術設定圖稿請見圖八。

（四）、場景四：廢棄倉庫

這是最後一個場景，也是主角與大 BOSS 決戰的地方，廢棄倉庫內，破敗的屋頂有些漏洞，陽光打下來的光束給玩家一種希望的感覺，象徵著結局即將迎向那未知的外頭光芒。

三、角色設計

角色設計主要參考了 50-60 年代的服飾，但終究這是一款遊戲，在設計上我們無法太拘泥於實際的材料，必須兼顧趣味性以及操作細節。因此，我們也加入了一些自身喜愛的要素。

- 正面主角－陸壹

  為了體現出留學回來的有錢人家子弟的服飾風格（組織給機器人們的身分），我們特地參考了那個年代歸國子弟的服飾，走了一點洋風。美術設定圖稿請見圖九。

- 負面主角－黎昕

  參考了那個年代教授們的服裝，為了增加知識感以及穩重感，設定了風衣以及圓眼鏡，使這個角色看起來更無害點。

- 配角－其餘路人

  男生我們設定了那個年代大多數人常穿的中山裝；而女生我們則是在那年代常

見的工作服、軍裝以及碎花彩裙之中選擇了碎花彩裙，主要是想體現出大學生年輕的活力，以及上海街頭繁華的氛圍。

伍、創新說明與檢討

一、故事與企劃方面

在市面上對於懷舊故事，大多是走民國風，然而此遊戲卻採取了較少見的 60 年代為背景，為的就是想以這年代特有的歷史地位來詮釋一些我們想表達的理念——反烏托邦。

風格上加入科幻以及玄幻的設定進去亦是因為如此。這樣的風格融合是一種創新的挑戰，為了將這三種風格完美的融合，我們在美術上下了一些功夫，而這樣的風格融合除了是挑戰更是一種必要的敘事方式，融合了背景故事的敘事方式。

在玩法上，為了增加可玩性，這款遊戲除了探索解謎以外還融合了 ARPG 以及 TPS 要素，使玩家可以因應挑戰的不同而更換武器。我們在戰鬥上也下了一點功夫，入戰時設定成在原地生成範圍，並且清空周遭所有路人，造成一個使玩家進入戰鬥場景的感覺，此設定不僅為了融合劇情解釋，更是為了使玩家在探索時不至於與太多怪物相撞。

二、技術與程式方面

以下將以對話系統的程式碼改進為例，以說明程式設計上遭遇的一些實務問題。常見的傳統寫法是很直覺的，但在實作時便可以發現它的不方便。原始寫法如下：點到某物便把物件布林值打開，開啟後出現txt檔第幾行的對話，在Unity介面中設定點擊對話框執行呼叫AddLine方法。但當物件多起來，要宣告許多物件布林，並把所有該跳第幾行寫出來，腳本就會變得又臭又長，如果換一個場景就要寫一個新的腳本，許多重複的程式碼就代表是可以縮減的，這讓我們開始思考，要如何寫才能變得更好。

```
public class DialogueManagerRoom : MonoBehaviour
{
    //對話管理
    ……
    void Update(){
        //點擊後，偵測是否有點擊到tag為book的物件，若有布林book=true，
        //  並將currentLine=0
        Ray ray = new Ray(Camera.main.transform.position, Camera.main.transform.forward);
        RaycastHit hit;
        if (Input.GetMouseButton(0) && Physics.Raycast(ray, out hit)) {
            if (hit.collider.tag == "book"){
                book = true;
                currentLine=0;}
        }
        //布林book==true時，開啟對話，文字讀取行數為currentLine，
```

```
            //設定可跳行數的最大值
            if (book == true) {
                  theText.text = textLine[currentLine];
                  MaxLine = 2;
                  AllStatic.openTalk = true; }
                  ……
          }
         //增加currentLine的方法
         public void AddLine()
         {
               if (currentLine < MaxLine) {
                     currentLine++;}
         }
}
```

改良的寫法如下，將一個腳本拆成兩個腳本，對話管理 DialogueManager 為一個腳本，物件 DialogueObject 為另一個。每個場景對話腳本只會有一個，而物件腳本可以有多個。不同的物件，就在 Unity 介面中設定不同的公開變數如文檔、行數等等。當點擊到物件後，呼叫對話管理的公開方法 ClikObject，更新對話管理腳本文檔、現在的文字行數與最大行數的變數，就可以成功的與不同物件對話。

當我們這樣寫會便利許多，可以重複使用這兩個腳本，並且在每個場景都能普遍使用，改善了一個場景就要多寫一個腳本的問題，也成功改善多一個物件，就要多一個物件變數，用物件變數去控制文字這樣寫死的方法。

```
public class DialogueManager : MonoBehaviour
{
       //對話管理
       ……
      void Update(){
            ……}
     //更新此腳本txt文檔、現在的文字行數與最大行數
      public void ClikObject(TextAsset otextFile,int omin,int omax){
            textFile = otextFile;
            currentLine = omin;
            MaxLine = omax;}
}
public class DialogueObject : MonoBehaviour {
     //點擊
     ……
     void Update () {
          //點擊後，偵測是點擊到的物件上是否有DialogueObject腳本
```

```
        ……
        if (Input.GetMouseButton(0) && Physics.Raycast(ray, out hit) &&
            AllStatic.openTalk == false) {
                DialogueObject _object = hit.transform.GetComponent<DialogueObject>();
                //若有DialogueObject腳本_object，開啟對話，呼叫dialogueManager的方法
                //ClikObject，參數為_object的txt文檔、現在的文字行數與最大行數
                if ( _object!=null && _object.enabled==true) {
                        dialogue.ClikObject(_object.textFile, _object.MinLine,
                        _object.MaxLine);
                        AllStatic.openTalk = true;
                }
            }
    }
}
```

三、美術與音效方面

為了使畫面看起來有老舊感，在 Unity 中我們調整了懷舊的濾鏡特效，畫面有些微的泛黃、雜訊。界面設計方面設計了懷舊科技感的風格，大量使用綠色和黑色為底色，以及像素風格的文字。

音效上因懷舊風的在無版權上較難找，所以只找了聽起來有些許老舊感的音樂風格。尤其是開頭畫面及教學關卡的背景音樂，還帶了點淡淡的留聲機的雜音，使玩家能在一打開遊戲就融入了這個 1960 年代。

四、創作過程相關檢討

創作過程中，對時間的掌握度還非常的不足，容易在超出時間才完成東西，這點在之後的任何專案都還需改進。為了在設定的時程內完成本作品，許多最初始的設計到了實作階段都只能放棄。這些擱置的設計將在下一章簡要地加以敘述。另外，在製作作品的途中感受到自身能力的不足，還需要花時間去精進，團隊整體的能力才會上升。

在強調「好玩」為主的遊戲作品中加入「主題」要素是我們的大膽嘗試，也遇到了許多不易掌握的技巧問題。在遊戲型態趨於雷同的現代遊戲市場，這是一條必須摸索的道路。相信，對於其他類型的作品有更多的研討，將可提升這一方向的能力。

陸、未來創作建議

本作品只是一個起點，後續可以加入設計的還有很多。例如，我們建議可以有多結局的分支、好感度系統、裝備商店系統以及參數值的優化。雖然我們已有一個從頭講述到尾的故事了，但若是想完善整個世界觀、使玩家更能感受到當下歷史背景的氣氛，可以從多結局多分支入手，增加一點小支線事件、或是一些小彩蛋等，若是時間允許，還可以在我們設定成開放式結局的後面加點主角逃離的過程與關卡。

另外雖然我們有更換武器的系統，但卻因為受制於遊戲的大小以及必要性，因而刪減了商店街中一些本來可以購買物品的功能商店，於是武器大多是一種類型只有一把，並且少數才從商店街取得，若是可以把遊戲做的更大更豐富，想把武器類型也多設定些，使玩家可以挑選自己順手的武器。除了武器，也可以增加一些禮物商店或是書店等，原本我們還有設定禮物可以刷重要ＮＰＣ好感的功能，但因為太瑣碎而也刪除掉了。以上提到刪除掉以及建議增加的東西，都將使遊戲可以更加豐富並且使玩家有更多東西可以探索，甚至允許的話，可以考慮做個半開放世界。

參考文獻

互動百科（2016.01.01），〈遊戲類型〉，取自：http://baike.baidu.com/view/18461.htm。

百度百科（2018），〈我們〉，取自：https://zh.wikipedia.org/wiki/%E6%88%91%E4%BB%AC_(%E5%B0%8F%E8%AF%B4)。

幻影神月（2017.04.17），〈《時空幻境 緋夜傳奇》堅持自我的緋色復仇劇〉，取自：https://forum.gamer.com.tw/C.php?bsn=544&snA=38168，參閱於 2018 年 4 月 07 日。

金手錶（2017.08.14），〈時空幻境 緋夜傳奇〉，取自：https://home.gamer.com.tw/creationDetail.php?sn=3682746，參閱於 2018 年 4 月 07 日。

維基百科（2018a），〈底特律：變人〉，取自：https://zh.wikipedia.org/wiki/%E5%BA%95%E7%89%B9%E5%BE%8B%EF%BC%9A%E8%AE%8A%E4%BA%BA。

維基百科（2018b），〈反烏托邦〉，取自：https://zh.wikipedia.org/wiki/%E5%8F%8D%E4%B9%8C%E6%89%98%E9%82%A6。

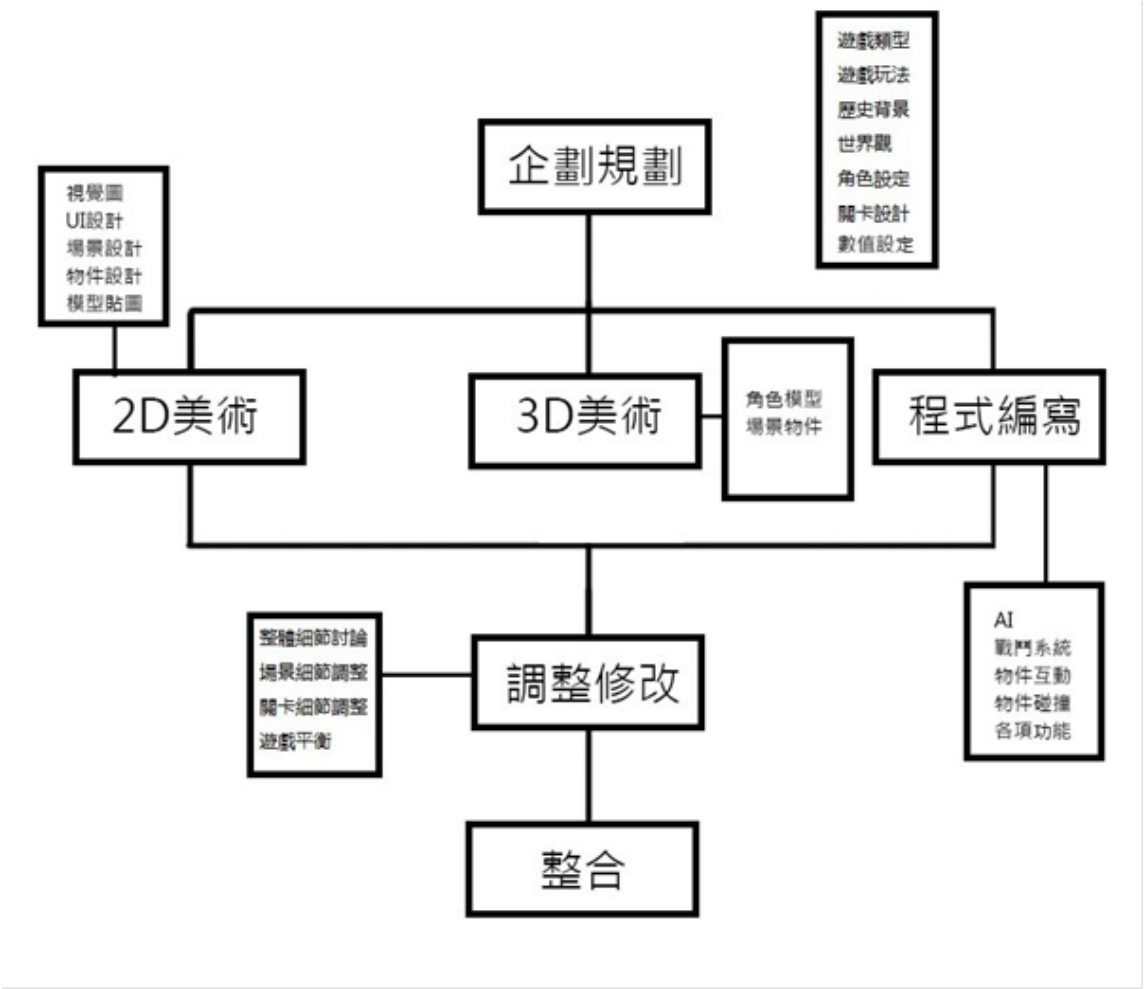
維基百科（2018c），〈美麗新世界〉，取自：https://zh.wikipedia.org/wiki/%E7%BE%8E%E9%BA%97%E6%96%B0%E4%B8%96%E7%95%8C。

Andrew Rollings and Ernest Adams, On Game Design, 中文譯本：《大師談遊戲設計》，上奇：台北市，2003。

TangBao（2017.12.08），〈TGA 2017 遊戲大獎完整得獎名單，《薩爾達傳說：曠野之息》與《地獄之刃：賽奴雅的獻祭》抱三獎大豐收〉， 4GAMERS 網站，取自：https://www.4gamers.com.tw/news/detail/33851/the-game-awards-2017-winner-full-list，參閱於 2018 年 4 月 07 日。

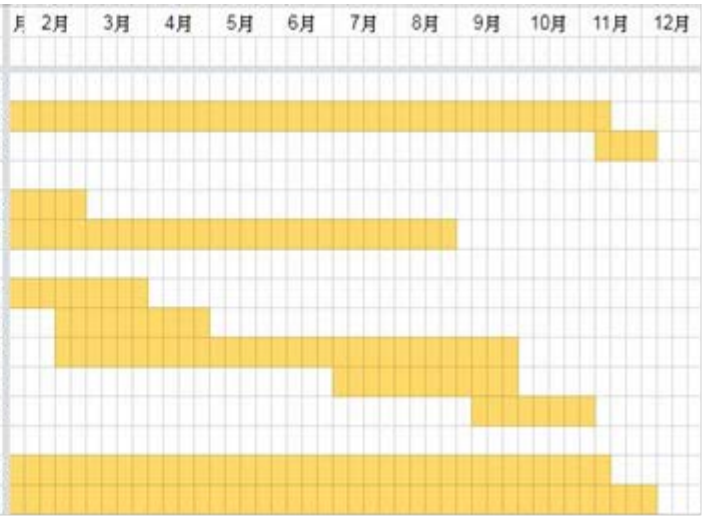
TangBao（2018.05.24），〈QTE 最高傑作！《底特律：變人》上市前全破評測，革命不是報復是求生〉，4GAMERS 網站，取自：https://www.4gamers.com.tw/news/detail/34999，參閱於 2018 年 6 月 12 日。

圖一、遊戲開發工作整合架構



圖二、本作品開發時程表（最上階層）

	事項	DAY	STAR	END
遊戲整體			2018/1/4	2018/12/9
	遊戲完成	301	2018/1/21	2018/11/18
	遊戲測試與調整	28	2018/11/11	2018/12/9
企劃				
	企畫書完成	40	2018/1/19	2018/2/28
	腳本完成	223	2018/1/19	2018/8/30
美術				
	2D設計完成	71	2018/1/19	2018/3/31
	3D建模	69	2018/2/20	2018/4/30
	3D建模完成	201	2018/2/20	2018/9/9
	3D材質貼圖	208	2018/3/20	2018/10/14
	特效製作	61	2018/9/9	2018/11/9
程式				
	程式完成	303	2018/1/19	2018/11/18
	程式測試	324	2018/1/19	2018/12/9



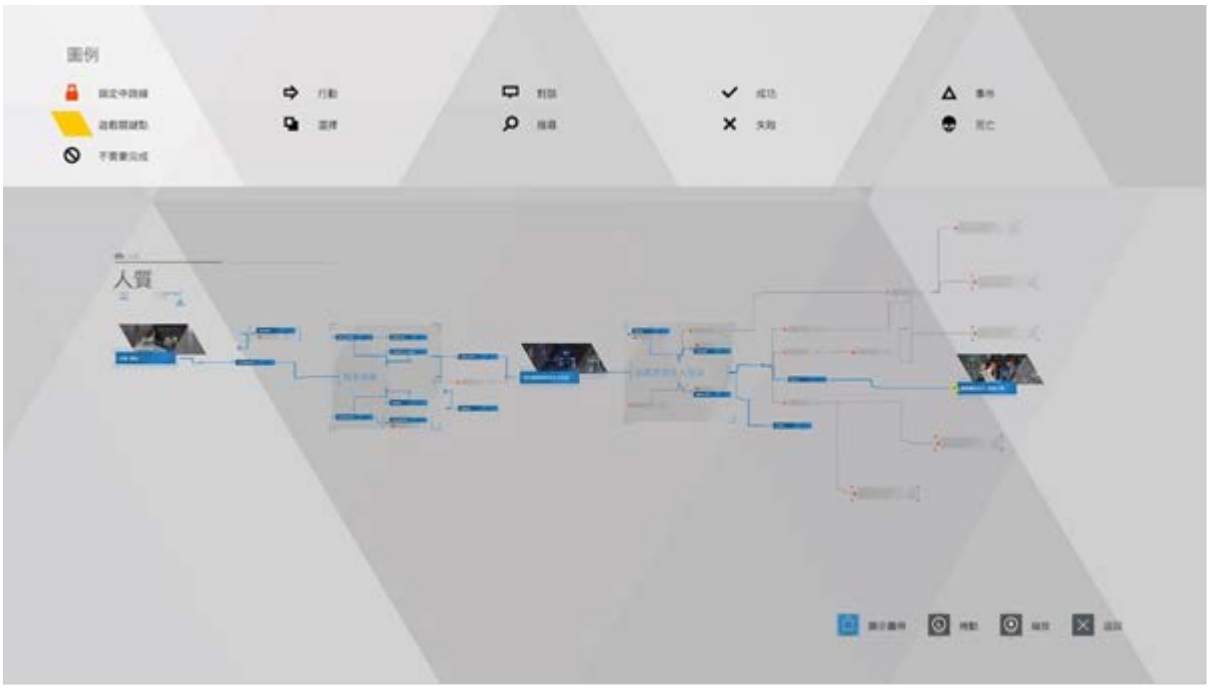
圖三、遊戲精彩畫面（資料來源：
https://home.gamer.com.tw/creationDetail.php?sn=3682746）



圖四 、爽快的戰鬥方式（資料來源：
https://home.gamer.com.tw/creationDetail.php?sn=3682746）



圖五、充滿分支的章節結局，每個分支都會影響到下一幕（資料來源：https://www.4gamers.com.tw/news/detail/34999）



圖六、玩家可自行選擇自己所要傳達的訊息（資料來源：
https://www.4gamers.com.tw/news/detail/34999）



圖七、場景一實驗室的美術風格設定



圖八、場景二校園的美術風格設定



圖九、正面主角（陸壹）的美術風格設定



第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

以何「半部」《論語》治天下？

基於文檔向量相似性的論語篇章結構分析

楊浩* 楊明儀** 劉千慧** 王軍***
北京大學哲學系助理教授*
北京大學資訊管理系本科生**
北京大學資訊管理系教授***

以何「半部」《論語》治天下？

——基於文檔向量相似性的論語篇章結構分析

楊浩[1]楊明儀[2]劉千慧[3]王軍[4]
1:助理教授 2,3:本科生 4:教授
1:北京大學哲學系 2,3,4:北京大學資訊管理系

摘要

《論語》為中國古代最有代表性的語錄體著作，其有各條語錄組成二十個篇章，這些篇章到底是隨意的排列，還是有一定的規律，歷代學者眾說紛紜。有著名的「半部」《論語》的典故，本文採用了計算文檔向量相似性的方法，來探索此語中「半部」的說法是否成立，以及是以何種形式成立。我們得到的結果為：與整部相比，上半部與下半部並不存在明顯差異，而且上下部也沒有明顯的匹配對應關係。從與整部相似度的比較來說，「半部」也可能是重點篇章組成的占半部篇幅的章節。本文為研究《論語》篇章結構引入了電腦化的方法，相較于傳統的分析方法，得到的結果更客觀、可驗證。

目次

1. 「半部《論語》治天下」與《論語》的篇章結構
2. 研究思路
3. 實現原理、步驟與方法
3. 1. 實現原理
3. 2. 資料準備與資料清理
3. 3. 相似度匹配
4. 結果對比與分析

關鍵詞

半部《論語》，篇章結構分析，文檔向量，相似度

《論語》是中國的傳統典籍，千百年來深刻地影響著一代又一代中國人的思想。其

作為文體，也成為中國語錄體著作的濫觴。絕大多數學者都會承認，《論語》各篇、各章之間是有著一定的有意安排，瞭解這樣的安排對於深入理解《論語》是不無助益的。

據傳宋人趙普有「半部《論語》治天下」一說，對於研究《論語》篇章結構可以說是一個有益的突破口。此處「半部」的說法，到底是前半部分、後半部分的「半部」？還是概指《論語》中的一部分呢？如果用傳統的研究方法進行分析，必定是見仁見智的，不會有較為公認的答案。

近年來，數位技術的發展為人文研究帶來了一些方法，雖然還不能代替學者對文本細緻的分析，但是也可以從宏觀角度量化地理解《論語》篇章之間的聯繫，至少從方法論角度提供了一種更為客觀的分析結果。

1.「半部《論語》治天下」與《論語》的篇章結構

鑒於本文討論的主題，此處有必要對此一典故的核心要點做一簡單梳理。「半部論語」的故事的主角趙普是北宋開國的重臣，據《宋史·趙普傳》記載：「普少習吏事，寡學術，及為相，太祖常勸以讀書。晚年手不釋卷，每歸私第，闔戶啟篋取書，讀之竟日。及次日臨政，處決如流。既薨，家人發篋視之，則《論語》二十篇也。」[1]這個記載有幾個要點：一，趙普年輕的時候讀書很少；二，當了宰相之後，宋太祖趙匡胤經常勸他讀書；三，晚年經常讀書，死了之後，家人發現所讀為《論語》。據學者們分析「半部論語」的故事可能與真實的趙普沒有什麼關係，但無疑趙普確實與《論語》有關係。也許後人正是依據趙普少不讀書，老常讀《論語》的原型演繹成後來「半部論語」的故事。

這則故事產生之後，後人對此故事所持態度不一，關注重點也不一。有信用的，也有貶斥的，也有質疑的。陸敏珍〈故事與發明故事：「半部論語治天下」考〉一文對「半部論語治天下」在歷史上對此典故的各種態度做了很好的梳理。追溯了「半部論語治天下」的故事形成的過程，肯定最遲在宋末元初，「論語半部」已經成為典故詞藻。

應當看到，從語義上來說，典故中的「半部」的說法，並非虛指。那麼，「半部」到底是前半部？還是後半部？還是某些章節組成的半部？甚或是從《論語》中選取一半數量的條目作為半部？2

朱熹在《朱子語類》中就表達出《論語》後半部內容與編排上不如前半部的意思。朱熹說：「大抵《論語》後數篇間不類以前諸篇」3，「大抵後十篇不似前十篇」4。甚至

[1]林駉撰：《古今源流至論》前集卷八，臺北：臺灣商務印書館影印文淵閣《四庫全書》本，第 942 冊，第 113 頁。

[2]日本僧人瑞溪周鳳（1391—1473）曾說：「趙普以《論語》半部輔宋太祖，恐非自一至五又五至十之義，於一部之中以可半部之事佐耳。」更有意思的是，民國時期，重慶藝新圖書社曾出版《半部論語與政治》一書（初刊於 1936、再版於 1943），該書託名為趙普所作，原作者實為趙正平，該書共彙集《論語》語錄 166 條，並逐條解讀。

[3]黎靖德編，王星賢點校：《朱子語類・卷第四十七　論語二十九・陽貨篇・子曰由也章》，中華書局，1986 年 3 月，第 1 版，第 1185 頁。

[4]黎靖德編，王星賢點校：《朱子語類・卷第四十四　論語二十六・憲問篇・子貢曰管仲非仁章》，中華書局，1986 年 3 月，第 1 版，第 1129 頁。

《易經》經文與《繫辭》也有下半部分不如上半部分的情況。5這個問題應當說與《論語》的總體的篇章結構有緊密的關係。

關於《論語》的篇章結構，古今學者的觀點有這樣一些要點：第一，《論語》各篇內部具有一定的主題。正是有了這些《論語》注釋者作為前驅，絕大多數《論語》注釋者都認為，《論語》的篇章組合不是任意的。第二，《論語》整體篇次安排具有一定邏輯。例如皇侃《論語集解義疏》力圖將《論語》篇次、篇名、篇義關聯起來，構築一個完整的理解體系。第三，《論語》總體篇章分佈並不均勻，比如上半部與下半部有一定的差異。不少學者將《論語》看成由上下兩編構成。伊藤仁齋《論語古義》認為：「《論語》二十篇，相傳分上下，猶後世所謂正續集之類乎」。日本太宰春台《論語古訓外傳》認為《論語》「上論文章簡潔而下論文章詳密」。日本市村瓚次郎《支那研究史》認為《論語》「上論十篇輯成最早，可視作《論語》的正編。」錢穆《論語新解》認為自〈鄉黨〉「以前十篇為上論，終之以〈鄉黨〉篇，為第一次之結集，下論十篇為續編。」[6]這些觀點雖不免猜測，但是也確實反映了《論語》上下部確實有一定的不同。

2.研究思路

本文採用向量空間模型分析《論語》的整部、上下半部、每一篇之間的相似度。分析思路主要從如下四個方面展開：

第一，上半部、下半部與整部的相似度比較。這個比較旨在探索上下半部是否存在較大的差異。如果相似度差異較大，則說明上下部從文本內容上來說確實有一定的不同。

第二，每一篇與整部的相似度比較。這個比較可以顯示出哪些篇章更具有代表性

第三，上半部與下半部篇章之間的相似度匹配。旨在探索上下半部在內容上是否有一定的對應關係，即上半部能不能在一定程度上代表下半部，反之亦然。

第四，每一篇與其他篇之間相似度的匹配。可以說明哪些篇章在內容上是「明星」章節，在一定程度上可以成為《論語》的代表。

3.實現方案

3.1.實現原理

向量空間模型是一種將非結構化文本資訊表示為數學向量的方法，在這一模型中，對於整個系統文檔集合 D 中的任一文檔 $d_j \in D$，可以把它表示為以下 t 維向量的形式（t 為概念集中概念的個數）：

$$d_j = (W_{1j}, W_{2j}, \ldots\ldots, W_{tj})$$

其中，向量分量 W_{ij} 表示第 i 個概念 k_i 在文檔 d_j 中所具有的權值，$W_{ij} \geqslant 0$。概念權值 W_{ij}

5「六十四卦，只是《上經》說得齊整，《下經》便亂董董地。《繫辭》也如此，只是《上系》好看，《下系》便沒理會。《論語》後十篇亦然。《孟子》末後卻剗地好。」〔宋〕黎靖德編，王星賢點校：《朱子語類・卷第六十七易三・綱領下・上下經上下系》，中華書局，1986 年 3 月，第 1 版，第 1672 頁。

6本段轉引自湯洪：《<論語>篇章結構研究述評》，載《延邊大學學報(社會科學版)》2018 年第 1 期。

的大小我們考慮局部權值和全域權值，決定使用 TF-IDF 進行衡量。

在文檔向量化表示的基礎上，文檔之間的相似度就可以通過各自向量夾角的余弦函數來計算，即兩個文檔 $d_j = (W_{1j}, W_{2j}, \ldots\ldots, W_{tj})$和 $d_p = (W_{1p}, W_{2p}, \ldots\ldots, W_{tp})$的相似度

$$sim(d_j, d_p) = (d_j \cdot d_p) / (|d_j| * |d_p|)$$

$0 \leqslant sim(d_j, d_p) \leqslant 1$，值越大，說明兩篇文檔越相似。

3.2.資料準備與資料清理

我們選取朱熹《四書章句集注》中的《論語》文本，根據其章節進行二級標號。如第一章學而的第一節編號為 1.1，依此類推，獲得了此次進行文本分析的原始語料。

為了從多種角度去探索《論語》篇章之間的相似度，我們建立了兩大類，共 6 種概念集。概念集的建立均先排除人名的內容。

第一類基於詞頻的概念集選取所有單字作為概念。當單字的頻次分別大於 3、5、10 時，獲得 3 種單字概念集。

第二類基於核心概念的概念集是引入人為判斷建立起來的，也有三種概念集。第一種採用 ngram 切詞方法，然後由兩位研究人員判斷並保留其中包含有效語義的部分，兩位研究人員存在分歧的判斷由第三位研究人員進行最終的確認，共得到 221 個核心概念。第二種概念集是根據《論語》核心內容人為選定的 57 個核心概念。第三種概念集是《論語》中的「五常」「仁義禮智信」。

3.3.相似度匹配方法

為了檢驗《論語》中上下部的對應關係，針對《論語》前 10 篇，在下半部的中找出與其最相似的篇章，記為下半部中與前 10 篇最佳匹配結果。反之，記為上半部中與後 10 篇最佳匹配結果。

除此之外，拋卻上下半部的限制，對《論語》中的每一篇匹配與其最相似的篇章。第一種匹配是針對每一篇，在其他 19 篇中找出與其最相似的章節。由於某一章有可能既是 A 章，也是 B 章的最相似章節，這樣的匹配對中章節存在重複情況，因此記為章節之間重複匹配結果。第二種匹配是在 20 個章節中，找到 10 個匹配對，這樣的匹配對中章節不存在重複情況，因此記為章節之間非重複匹配結果。

4.小結

我們初步的分析結論是：

第一，上半部、下半部與整部的相似度比較。與整部相比，上半部與下半部並不存在明顯差異，而且上下部也沒有明顯的匹配對應關係。

	單字（>3）	單字（>5）	單字（>10）	221 個核心概念	57 個核心概念	仁義禮智信
上半部 v.s.整部	0.990	0.991	0.993	0.961	0.975	0.996
下半部 v.s.整部	0.991	0.992	0.994	0.976	0.985	0.995

第二，每一篇與整部的相似度比較。其中第 14、15、6、12 篇比較突出，可以說是《論語》篇章中的代言篇章。

	單字（>3）	單字（>5）	單字（>10）	221 個核心概念	57 個核心概念	仁義禮智信
與整部最相似的前三篇	14	14	14	15	14	15
	15	15	15	14	15	17
	6	6	6	12	12	13

第三，上半部與下半部篇章之間的相似度匹配。上部主要集中在第 6、7 等章，而下半部則主要是 14、15 兩章。從匹配度上沒有形成一定的分散性，也就是說上部不能代替下部，下部也不能代替上部。

	單字（>3）	單字（>5）	單字（>10）	221 個核心概念	57 個核心概念	仁義禮智信
下半部中與前 10 章前三匹配對	（6,15）	（6,15）	（6,14）	（1,15）	（7,15）	（7,15）
	（7,14）	（7,14）	（4,15）	（6,15）	（6,15）	（4,17）
	（4,15）	（4,15）	（7,15）	（7,15）	（5,15）	（6,17）
上半部中與後 10 章前三匹配對	（15,6）	（15,6）	（14,6）	（15,1）	（15,7）	（15,7）
	（14,6）	（14,6）	（15,6）	（14,7）	（14,7）	（14,7）
	（13,6）	（13,6）	（17,6）	（17,5）	（19,6）	（17,4）

第四，每一篇與其他篇之間相似度的匹配。比較結果顯示，從總體上，《論語》中最能夠代表《論語》整體內容的是第 14、15 章。也就是《論語》的重心應該在下部，而不是上部。

	單字(>3)	單字(>5)	單字（>10）	221 個核心概念	57 個核心概念	仁義禮智信
篇章重複匹配對前三	（15,14）	（15,14）	（15,14）	（15,14）	（15,14）	（15,7）
	（14,15）	（14,15）	（14,15）	（14,15）	（14,15）	（7,15）
	（6,15）	（6,15）	（6,14）	（13,12）	（19,15）	（6,4）
				（12,13）		（4,6）
篇章非重複匹配對前三	（14,15）	（14,15）	（14,15）	（14,15）	（14,15）	（7,15）
	（7,9）	（7,9）	（7,9）	（12,13）	（6,7）	（4,6）
	（6,13）	（6,13）	（6,17）	（5,17）	（16,19）	（3,10）

但是從篇與篇之間相似度來說，上半部的篇章概念分佈比較分散，沒有形成特別突出的篇章，而下半部第 14、15 篇則比較突出。因此如果從概念分散程度來說，半部可以是「前半部」。而最能代表《論語》，則是應該是有 14、15 章的下部。

當然，也可以換一種思路，那就是從與整部相似度的比較來說，「半部」也可能是重點篇章組成的占半部篇幅的章節。那麼以第 14、15、6、12 篇為代表篇章可以領銜。

至於從《論語》三百條中選取一半來代表《論語》則不是本文所採用方法能夠解決了的，有待通過更好的演算法進行研究。

可能令讀者沮喪的是，本文並不可能通過這一種演算法最終確定「半部」到底是哪半部，但是我們認為至少通過這樣的辦法，拓展一種宏觀研究《論語》篇章結構的一種方法。

參考文獻

林駉《古今源流至論》(1983)。《四庫全書》。臺北：臺灣商務印書館。
黎靖德編，王星賢點校（1986）。《朱子語類》，北京：中華書局。
湯洪（2018）。《<論語>篇章結構研究述評》，《延邊大學學報(社會科學版)》1。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

華語文領域學者資料定義與資料使用的詮釋探討

Data in the Humanities

Exploring Humanists' Perception and Usage of Research Data in the Field of Chinese Language

Chieh-Yun Lin Wei Jeng

華語文領域學者資料定義與資料使用的詮釋探討

Data in the Humanities: Exploring Humanists' Perception and Usage of Research Data in the Field of Chinese Language

摘要

豐富的人文資料有助於領域研究的創新與進步，同時也是數位人文發展的基礎。本研究聚焦於資料的產製者與詮釋者，取徑於深度訪談，探索華語文相關學者對於人文資料的詮釋與定義，以及研究他者再利用（reuse）自身之研究資料的認知與實踐。研究結果顯示，華語文學者在與研究材料互動時，對於原始資料（raw data）與次級資料(secondary data) 之間的分野較為模糊；然而，在經由學者詮釋過後的的資料中，受訪者的分享意願較低，因為此等資料已灌注研究者自身的問題意識，進而轉換為著作權的意識存在，資料的價值幾乎與最後產出的研究成果同等重要。本研究期於經由發現華語文領域學者與人文資料互動之認知與經驗，提供未來數位人文領域或是記憶型機構如學術圖書館若欲建置典藏庫時之參考。設計方或館方可藉此研究參考相關學術領域的資料使用方式，設計出符合使用者意向的系統、工作流程與使用機制。

1. 導言 (Introduction)

大量的資料產出是數位人文發展的基礎，其有助於研究的創新與進步（項潔，涂豐恩，2011）。然而，大量未有效管理、未遵循標準釋出的數位資料，除了造成正在進行研究的研究團隊造成困擾，也使得後續再利用資料的學者不便觸及與利用，形成資料氾濫（data deluge）情形 （Borgman, 2007）。因此，研究資料應該如何整理與組織才能夠弭平上述挑戰，並使其發揮最大價值，是資訊學領域逐漸提升關注的焦點。

然而，在眾多探討研究資料整理與組織的議題中，有別於自然與應用科學領域，人文領域所產製的資料有兩大特性：一）多為質性資料，以及 二）資料週期的模糊性。質性資料是人文領域學者經常主要使用的一項研究資料基礎架構，藉由了解主觀與客觀的詮釋以利找出其研究的價值。對於質性資料的定義各方解釋皆有所不一，其中根據美國聯邦政府 DOE Policy for Digital Research Data Management 定義研究資料，是為科學界普遍接受的用於驗證研究結果必要之「事實性紀錄材料（recorded factual materials）」。美國國家學術人文基金會（National Endowment for the Humanities）認為人文學的研究資料是在進行研究過程中產生或蒐集的資料，例如：引文、數據庫、數位工具、文件與論文等；但卻不包含：初步分析資料、論文草稿、未來研究計畫書，以及會導致侵害個人隱私之資料。在廣泛的資料類型中，人文學者所產製的資料中多屬質性資料，文章脈絡對於建構資料至關重要。即便是相同文本，運用在不同情況和不同的研究問題時，更需學者進行處置、再詮釋。

	基本資料	學位背景	目前研究領域	近期研究主題
P01	A校碩士三年級	英國文學 美國 MBA	清代文學	清代詩歌研究
P02	B校碩士三年級	中國文學	現當代文學	伊斯蘭教議題小說研究
P03	A校碩士一年級	中國文學 教育學程	民間、敦煌文學、神話學	無聲戲研究
P04	C校國文系助理教授	中國文學 公共行政	儒學	荀學史研究
P05	D校碩士一年級	中國文學	文字學	清末民初報刊研究
P06	A校博士二年級	中國文學	唐宋文學	唐宋古文相關研究
P07	A校中文系教授	中國文學	中國近現代思想與文學	近現代報刊研究
P08	E校中語系教授	中國文學 英美文學	辭賦	晚清民國相關拓本與辭賦文化
P09	B校中文系助理教授	中國文學	易學	易學相關研究
P10	B校台文所助理教授	台灣文學	臺灣電影 臺灣文學	戰後小說相關研究
P11	F校中文系副教授	中國文學	魏晉文學	南北朝相關文學研究
P12	G校中文系副教授	中國文學	唐宋詩學	詩學相關研究

表一、受訪者背景一覽（N=12）

承接著人文學者再詮釋的特性，在該領域中原始資料 (raw data)、二次資料 (secondary data) 與學者加值之後所產製的資料 (processed data)，界線顯得模糊。再者，由於人文領域分析方法的適用性，時常只能根據特定的研究問題進行評估：人文領域中資料產出者與再用者之間的應用皆存在差異。以我國中國文學、華語文相關系所之研究內容為例，領域學者多注重主觀意識、詮釋者內在觀點的探析與延伸、以及在自然情境脈絡下感受的經驗。因此，不同於量化研究，華語文領域學者的研究典範側重於對社會性事實進行分析，並納入各方與自身的經驗與詮釋：經過蒐集許多豐富的描述性資料，研究結果始而清晰。Bishop（2007）結合對質性資料之原始資料和二次資料分析（secondary data analysis, SDA），探討他們二者之間的差異、相似性和關係的反思進行探究。結果發現許多學者多半不願提供他們認為自己仍未完整經消化、磨光的資料。因此不願將自己的錯誤攤下陽光下，意即不願分享自己經消化的資料再供他人二次使用。儘管所有質性資料都是在上下文中所建構出，但重複使用資料會在不同時間和不同研究的問題上，重新為文本添加上不同的面向。且倘若學者對於原始資料與二次資料分析的界定趨於模糊，則資料可能經過無數次地詮釋，爾後再將資料接連分享，資料的有限性將會變的更明顯，也就是資料開始愈來愈有專一性與特定性，資料的再使用範圍漸趨縮小。

總的來說，質性資料所為人限制之再使用(reuse)上的困難度在於「描述性」不一，隨著時間與背景的改變，質性資料的內容與價值也可能跟著改變，抑或不符合當今的觀點，更重要在於它使用上的學術倫理考量，意即「何謂正當使用途徑？」在歷史久遠的人文學院之中文系對於資料的定義又是為何，何種類型的形式是屬於研究當中的「資料」，以及在他們的研究中如何使用這些資料，都是本研究預計探討的面向。然而，現今與資料實踐相關的既有文獻，多以針對自然科學、生醫領域學者等量化資料探討（如 Piwowar et al., 2008; Hey, A.J.G. et al, 2003）。少數調查中，有涉及到質性資料的實徵研究中，也多以社會科學領域或歷史學者進行訪談與調查（如 Jeng et al., 2016; Jeng 2017），較少針對純人文領域學者進行探究研究資料的相關議題。因此，本研究得以發現缺口以進行探索。

本研究以資料的產製者與詮釋者——人文學者的角度出發，探討國內華語文相關學者對於資料的定義與研究他者使用自身之研究資料的認知與實踐。本研究針對以下研究問題進行探討：

(1) 華語文領域學者如何定義第一手與第二手研究資料？

(2) 華語文領域學者對於他者使用自身之研究資料，進行二次資料分析 (secondary data analysis) 的意向為何？華語文領域學者對於他者再利用自身研究資料 (data reuse) 之意向為何？

(3) 華語文領域學者對於使用研究資料度用典藏庫 (data curation repository) 系統的意向為何？

2. 研究設計 (Research Design)

本研究旨在探討人文學華語文學者對於資料的定義，以及對資料的定義與詮釋之意向。針對研究型大學的華文領域共 12 位學者進行半結構式訪談：受訪者分別為兩位碩士一年級生、兩位碩士三年級生、一位博士二年級生、三位助理教授、兩位副教授以及兩位教授；其中 8 位學者已完成受訪，4 位學者已經確認邀約，預計在 2018 年八月悉數完成訪問，期可在全文截稿前（十月）順利完成分析與撰寫。

訪綱內容除了為符合本研究而設計的問題之外，亦改編自由 Purdue University Libraries 與 University of Illinois Urbana- Champaign 的圖書資訊學研究生共同執行的 Data Curation Profiles Toolkit（DCP）（Witt et al., 2009）。 為了進一步解釋訪談結果，本研究以相關文獻回顧的輔助製作此半結構式開放性問卷。開放性問題尤針對以下三項主題：(1) 中文系學者對於研究資料的定義與分類 (2) 研究資料分享的程度與劃分等級 (3) 質性資料典藏庫(repository)建置意向。爾後以分類方式分析與觀察資料呈現方向；此外，也發現了一些新出現的研究結果，如此亦擴大了未來研究的框架。

3. 早期研究結果 (Preliminary Results)

1) 對「研究資料」的定義與研究歷程中的挑戰與發現

針對目前已詢問之華文系所受訪者，當問及「研究資料」的定義時，學者所認知的「研究資料」大多為一完整文本或期刊、論文，然而這也是他們所認定的「原始資料」形式。結合其研究主題可以看出，中文系學者的研究方法幾乎是針對經典古籍或具有年代性的文本進行分析與詮釋，資料的形式與類別相對比較單純；但因文本詮釋的專業性極高、獨特性強，更是研究者的心血產製，因此對於文本詮釋的資料分享意願與分享程度皆低。學者認為，受自己詮釋加值過後的資料已有自己的著作擁有權的意識存在，它的價值幾乎同等最後產出的研究成果重要。

另外，在研究歷程的發現中，華語文領域學者雖然主以文本進行研究，形式上紙本電子檔皆常使用，尤也能了解針對研究性質所符合的數位資料庫並善用之；例如大成老舊期刊資料庫。縱使有著廣泛的資料可供研究使用，但其中卻有一半以上受訪者表示他們在分類資料上的窒礙，尤以龐雜的資料令他們在研究上產生相當的困擾程度。

2) 學者對於他人進行自身「二次資料分析」與「資料再用」的見解

受訪者 P02 表示，文本詮釋資料並非是一般資料形式，它包含了生產者的個人主觀想法，若供他人進行二次使用並分析會有許多應考慮的使用規範。縱然如此，學者在「應如何規範」的做法上卻保有很大的空間。舉例來說，P04, 05 認為若使用文本資料的詮釋可以規範再用者(reuser)的方法也應只有標註引用，推測可能有關乎典藏庫設置的疑慮問題與必要性。延伸至華語文領域學者定義之研究資料主為文本，另有學者提出典藏庫與資料庫的差異與使用方式應該加以妥善區分。

對於資料再用的分享看法上，大部分學者認為，自己的資料至多僅可給予相關研究者參考與使用，對於分享給校內甚至是校外相關領域學者的意願性不高。應可推測為，華語文學者對於自身的資料有高度的擁有權與保護權，並且對於應如何保護智慧財產權的相關做法了解有限。

3) 典藏庫建置意向

以往傳統上的資料庫與研究資料典藏庫（repository）的差別是華語文領域學者極其困惑的地方。P06 學者認為，兩者系統裡面所儲存的內容應該都大同小異，往後若在建置典藏庫前應須加以審視存放的資料屬性。此處討論的原因在於中文系學者本身對於研究資料的認定：意即欲針對要進行分析的文本，學者認為大都已可在現有的資料庫或是圖書館中找到；因此對於典藏庫的定義須再加以審視、思量與區分。

4. 小結與未來研究規劃

質化資料相較於量化資料存在許多待解決的矛盾與衝突，包含資料再使用的規範、學者分享意願低落，以及相關的學術倫理問題，都是本研究發現的問題。藉此，資訊領域專家或可思索未來建置典藏庫(repository)的注意事項，是否典藏庫需要針對質化與量化的資料進行個別的管理，甚至建造單一質化資料的典藏庫，尤針對以分析大量且雜的文本詮釋與再詮釋的學者。

鑑於研究結果發現華語文領域學者對於資料的定義與意向，未來學術圖書館若欲建置典藏庫時，可以藉此研究參考相關學術領域的資料使用方式，設計出符合使用者意向的模型，更能加以省思是否不同領域的專業知識，其對於研究資料的看法與用途有不同之處。因此，希冀在未來能更深入了解華語文領域學者對於質性資料的應用，本研究預計將再訪問四位學者期望更能全面性的捕捉與探究，並將完整結果提交至十月之會議論文。

5. 參考資料

Bishop, L. (2007). A Reflexive Account of Reusing Qualitative Data: Beyond Primary/Secondary Dualism. *Sociological Research Online* ,12(3). Advance online publication.

Borgman, C.L., Wallis, J.C., Mayernik, M.S. and Pepe, A. (2007). Drowning in Data: Digital Library Architecture to Support Scientific Use of Embedded Sensor Networks. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, *ACM*, pp. 269-277.

Hey, A.J.G. and Trefethen, A.E. (2003). The Data Deluge: An e-Science Perspective. In, Berman, F, Fox, G C and Hey, A J G (eds.) *Grid Computing - Making the Global Infrastructure a Reality.* Wiley and Sons, pp. 809-824.

Jeng, W., Mattern, E., He, D., & Lyon, L. (2016). Unpacking the "Black Box": A Preliminary Study of Visualizing Humanists and Social Science Scholars' Data and Research Processes. Proceedings of iConference 2016.

Jeng, W. (2017). Qualitative data sharing practices in social sciences. Unpublished doctoral dissertation. University of Pittsburgh.

National Endowment for the Humanities (2018). Data Management Plans for NEH Office of Digital Humanities Proposals and Awards. National Endowment for the Humanities. Last accessed 2018 at https://www.neh.gov/files/grants/data_management_plans_2018.pdf

Piwowar, H.A., Becich M.J., Bilofsky H., Crowley R.S., on behalf of the caBIG Data Sharing and Intellectual Capital Workspace (2008) Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers. PLoS Med 5(9): e183. https://doi.org/10.1371/journal.pmed.0050183

U.S. Department of Energy (2018). DOE Policy for Digital Research Data Management: Glossary. Retrieved from:https://www.energy.gov/datamanagement/doe-policy-digital-research-data-management-glossary#Digital%20Research%20Data

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation*, 4(3), 93-103.

項潔、涂豐恩 (2011) 導論—什麼是數位人文。項潔編《從保存到創造:開后交婁文位人文研究》,頁 9-28 。臺北：國立臺灣大學出版中心。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

Wikidata, a Low-tech Solution to Leverage Semantic Technologies?

Fudie Zhao
University College London

Wikidata, a Low-tech Solution to Leverage Semantic Technologies?

Digital Humanities and GLAM (Galleries, Libraries, Archives, and Museums) are actively applying semantic technologies to their projects. They have constructed ontologies, published semantically-enriched data, linked with external sources and imported semantic data to enrich their own. However, the high technical requirements are always a barrier for these institutes to fully leverage semantic resources.

Wikidata is a free and open knowledge base that acts as central storage for the structured data of its Wikimedia sister projects, like Wikipedia. Featuring a WikiProject ontology[1], a live SPARQL endpoint for data query[2], regular RDF dumps[3], and interlinks to other open data sets[4], it is Wikimedia's response to semantic technologies. Beyond the scope of Wikimedia, it also provides support to other sites and services, like Google Knowledge Graph (Tanon et al., 2016).

Originated from Wikipedia, it maintains a simple, user-friendly, collaboration-oriented editing interface (fig.1)[5]. Following its tutorials, even novice users are able to create and publish semantically-rich structured data conforming to Wikidata's ontology[6]. Its strong community has developed various applications and tools to support data publication and integration which further lower the technical barrier[7]. It has already linked to many databases from GLAM-DH domains, like VIAF (Virtual International Authority File)[8] and LCNAF (Library of Congress Name Authority File)[9].

[1] https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology
[2] https://query.wikidata.org
[3] https://www.wikidata.org/wiki/Wikidata:Database_download
[4] https://tools.wmflabs.org/mix-n-match/#/
[5] https://www.wikidata.org/wiki/Wikidata:Introduction
[6] https://www.wikidata.org/wiki/Wikidata:Tours
[7] https://www.wikidata.org/wiki/Wikidata:Tools
[8] https://www.wikidata.org/wiki/Property:P214
[9] https://www.wikidata.org/wiki/Property:P244

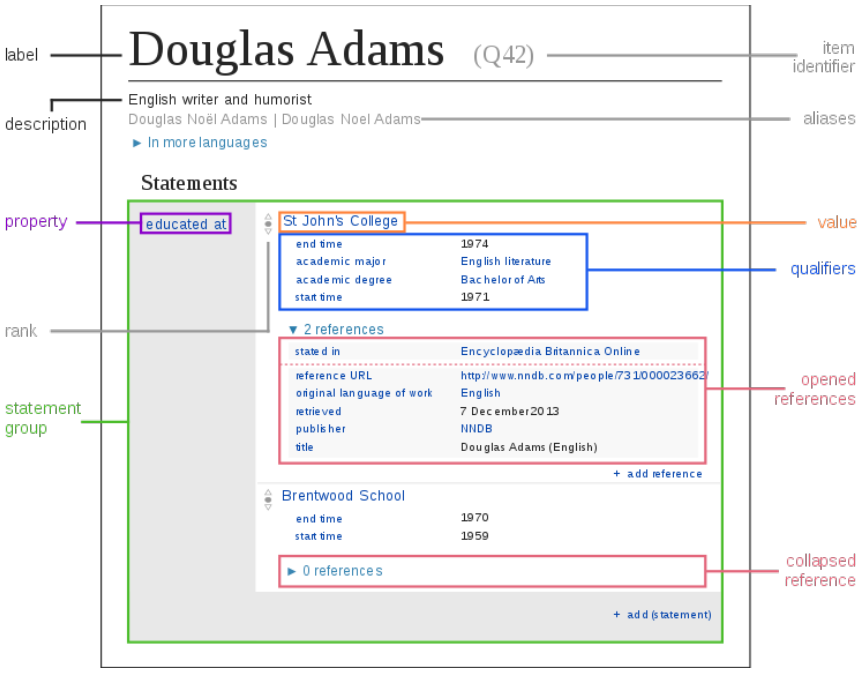


fig.1

Wikidata is a powerful resource for GLAM-DH to access the semantic technologies. Some institutes from the GLAM-DH domain have been embracing Wikidata for their projects. For example, Europeana, the EU digital platform for cultural heritage, has taken Wikidata into its semantic strategy to enrich its multilingual, heterogeneous data sets[10]. OCLC (Online Computer Library Center), the maintainer of WorldCat, is introducing Wikidata to libraries[11]. A well-known Digital Humanities project for Chinese historical figures, CBDB (China Biographical Database Project), has reconciled 83.4% of its data with Wikidata[12].

Despite the numerous empirical works conducted so far, GLAM-DH's cautious attitude towards the quality of the crowd-sourced data make them hesitate to cooperate with it (Davidson, 2012). It is therefore crucial to have a systematic review of these empirical projects to identify the potentials and challenges in cooperation with Wikidata so that GLAM-DH can critically leverage this powerful source of semantic technologies while avoiding its pitfalls.

The poster will demonstrate the process and results of the systematic review, and suggest possible approaches for GLAM-DH to utilise Wikidata. The poster will present three potential ways for GLAM-DH to use Wikidata to benefit themselves from semantic technologies:

[10] https://www.slideshare.net/antoineisaac/glamwiki-2015
[11] https://www.oclc.org/research/events/2018/06-12.html
[12] https://tools.wmflabs.org/mix-n-match/#/catalog/158

1) Dissemination of Information – Play Safe: Access the Wider Landscape of the Semantic Web through Wikidata.
2) Consumption of Information – Embrace the Breadth and Fuzziness: Enrich Datasets with Wikidata
3) Production of Information - From Open Data to Open-sourced Metadata

The first approach is the least controversial to collaborate with Wikidata as it only takes Wikidata as a platform to disseminate information generated by GLAM-DH. The second is a bolder step as the GLAM-DH become consumers of the crowd-sourced data. As the quality of the data is not necessarily guaranteed, GLAM-DH needs to be more careful. The third approach is the most advanced, in which GLAM-DH give away the right of curating its datasets to the public on Wikidata rather than holding it for its own experts. The poster will then address the challenges of these approaches, including the assessment of resources, selection of tools and methods, and evaluation of results.

References

Davidson, Cathy N., Humanities 2.0: Promise, Perils, Predictions. In *Debates in the Digital Humanities*. Minneapolis, MN: U of Minnesota P, pp. 476–489.

Tanon, T.P.P. et al., 2016. From freebase to wikidata: The great migration. *25th International World Wide Web Conference, WWW 2016*, pp.1419–1428.

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

“一花開五葉”
禪宗傳承視覺化平臺構建

楊明儀*　王軍**　邱勇***
北京大學資訊管理系本科生*
北京大學資訊管理系教授**
微軟（中國）有限公司產品經理***

“一花開五葉”——禪宗傳承視覺化平臺構建

楊明儀[1]　王 軍[2]　邱 勇[3]
1：本科生　2：教授　3：產品經理
1,2：北京大學資訊管理系　3：微軟（中國）有限公司

摘要

本研究基於臺灣法鼓佛教學院佛學規範資料庫中的人名規範資料庫，建立了禪宗人物師承關係、禪宗人物傳承地理分佈和人物資訊的展示平臺，為探索禪宗人物的傳承關係和演變歷史提供了直觀易用的工具。

研究禪宗的師承關係是一個宏大的課題。儘管如《五燈會元》、《續傳燈錄》等佛教典籍對禪宗人物的師承關係進行過梳理，但一來典籍之間詳略不同，且不免存在錯漏之處，二來對傳承關係的展示不夠直觀，臺灣法鼓佛教學院佛學規範資料庫對這些資料進行彙集整理，形成了較為全面可靠的資料，我們在此基礎上，構建禪宗人物師承關係的展現平臺，通過圖形化介面和交互設計將師承關係的展現變得簡便、易於挖掘和探索，供研究人員及感興趣的用戶使用。

禪宗人物師徒傳承的過程也是各宗派傳播的過程，基於 GIS 對禪宗人物師徒關係的地理資料進行視覺化展現，可以直觀地反映出“五家七宗”在傳播過程中的歷史流變和地理分佈，以促進大眾在互聯網環境下的對禪文化的瞭解。

目次

1. 引言
2. 基本架構
3. 實現方案
3. 1. 獲取、清洗人物和關係資料
3. 2. 法脈傳承樹狀圖的呈現
3. 3. 傳播分佈圖分析與呈現
4. 小結

關鍵詞

禪宗，視覺化，法脈傳承，傳播分佈

1.引言

禪，發源于印度，興盛於中國，在東亞的文化中生根發芽，並傳至世界各處。禪對於西方人而言是充滿了神秘的東方色彩的詞彙；而絕大多數中國的人在提起禪時，常常會仿佛有所體悟，卻不知從何談起。我們希望討論和研究的問題是：在互聯網環境下，我們能否借助互動式的、視覺化的呈現手段幫助大眾直觀地瞭解禪宗的源流與傳播。

就佛學而言，目前有 CBETA 電子佛典、慧海佛教百科資源庫等提供佛典的線上閱讀和搜索，佛學規範資料庫提供以人物和地點為中心的相關資訊的展示。這些研究都是以資料庫、資源庫的形式存在，是將文本的資訊數位化後的成果。本文基於這些已有的數位化資源，提供一個面向用戶的、互動式的視覺化產品。

就時空範圍歷史資料的視覺化而言，Schich M 等人[1]使用人物的出生死亡地點描述人類兩千年來的智力流動，對於文化歷史做出了宏觀的描述。本文借鑒此研究思路來呈現禪文化在時間和空間的傳播，以此幫助用戶對於禪宗的傳播歷史有一個整體、宏觀的感受和理解。

2.基本架構

本系統所面向的使用者有兩類，一類是對禪宗感興趣的大眾，另一類是相關的佛教史研究人員。基於此，平臺總體分為四個部分，第一部分為視覺化效果呈現，也就是系統的主幹部分，這一部分的內容見實現方案之第二小節；第二部分對系統的設計思路、資料來源進行簡單介紹；第三部分是對研究團隊——KVision 實驗室的介紹頁面；第四部分為語言切換，支援簡繁切換。

總體而言，禪宗的流派是在不斷地演變和發展的，並不局限於這五家，如在五家之前有青原系和南嶽系，臨濟宗又分出楊岐派和黃龍派。僅以這五家論，出現和存在的時間也有先後早晚之分。因此首頁是對禪宗的整體歷史視覺化頁面：以達摩為根，對於從達摩到五家宗派的創始人之間的傳承路徑進行了描繪使用者可以從首頁跳轉到整體傳播圖頁面，查看禪宗整體的時空傳播圖景。

對於五家宗派，用戶可以從首頁跳轉到五家宗派的介紹的子頁面。在介紹頁面，可以跳轉至法脈傳承圖和該宗的傳播分佈圖，同一宗的法脈傳承圖與傳播分佈圖可以自由切換。

[1] Schich, M., Song, C., Ahn, Y. Y., Mirsky, A., Martino, M., & Barabási, A., et al. (2014). A network framework of cultural history. Science, 345(6196), 558-562.

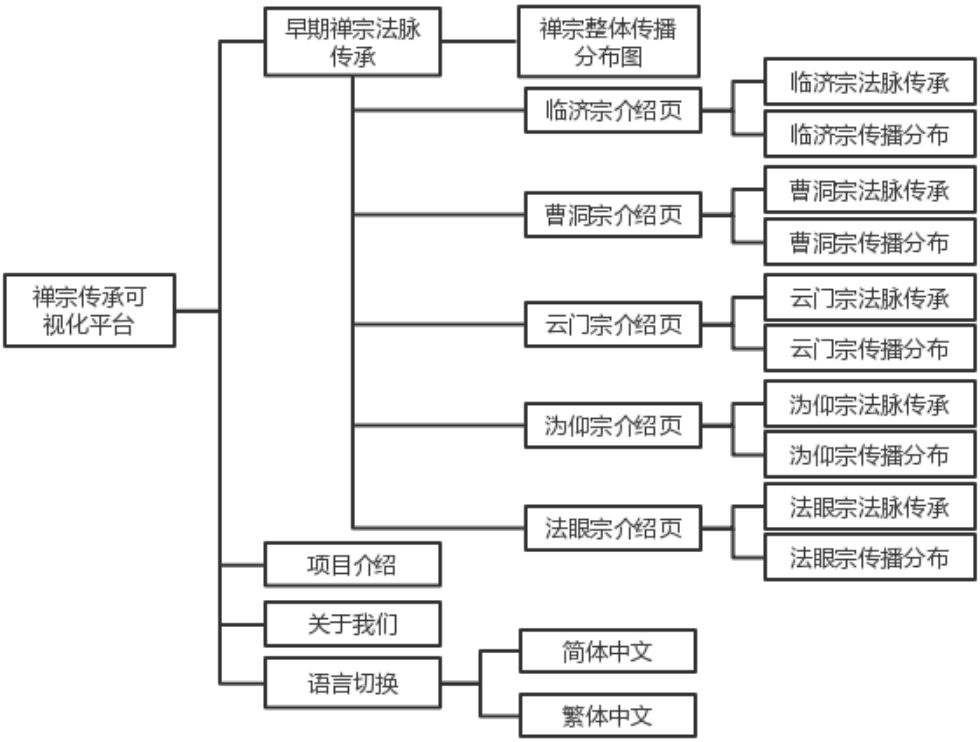


圖 1 禪宗傳承視覺化平臺架構

3.實現方案

3.1.獲取、清洗人物和關係資料

本專案的資料來源於法鼓佛教學院的佛學規範資料庫。此資料庫的目的為整合已完成與進行中之各項目的人物與地點資料，並建立歷史對照年表，以利於日後數字佛學專案之資源分享與未來可能之跨項目搜尋應用，因此建立時間、地點、人物與佛經目錄等四個規範資料庫。本研究主要採用了其中的人物與地點規範資料庫。

人名規範資料庫提供佛典相關人名規範資料查詢，截至 2018 年 8 月底，共收錄 41739 條人物資訊。對於收錄的人物，資料庫提供姓名、別名、生年、卒年、朝代、籍貫、圓寂地、性別、作譯資料、出現於何典籍之中等資訊及注解。

地名規範資料庫提供佛典相關地名名規範資料查詢，截至 2018 年 8 月底，共收錄 58859 條地名資訊。對於收錄的地點，資料庫提供地名、別名、經緯度、所屬朝代、出現於和典籍之中等資訊及注解。

基於法鼓文理學院的佛教規範資料庫，我們獲得了 xml 格式的人物資料，採用 beautifulsoup 為工具提取所有人物以及師承關係資訊，保存為 json 格式。由於資料中沒有對禪宗以及各宗派資料進行標注，我們採取的假定是：若一個人物根據其師徒關係能追溯到達摩，則判定其為禪宗人物 ，若能追溯到五家宗派的創始人，則判斷其為對應

宗派的人物。

在我們獲取到的資料中，共有禪宗人物 16561 人，各宗人數如下表所示：

表 1 原始資料中各宗人數

宗派	人數
臨濟	13800
曹洞	4768
雲門	723
溈仰	106
法眼	300
其他	1385
總計	16561[2]

3.2.法脈傳承樹狀圖的呈現

將師徒傳承關係以樹狀圖呈現方法與傳統法脈傳承圖一致，符合用戶的習慣。以可交互的方式呈現出來，使用者可以自由選擇查看哪一分支，隱藏無關的人物節點。

我們採取 echarts 的樹狀圖對師徒法脈傳承進行視覺化，具體呈現效果如下圖所示。其中空心節點代表其為葉節點或其子節點已展開，實心節點代表其有未展開的子節點。用戶可以通過按一下節點實現展開或隱藏的動作。

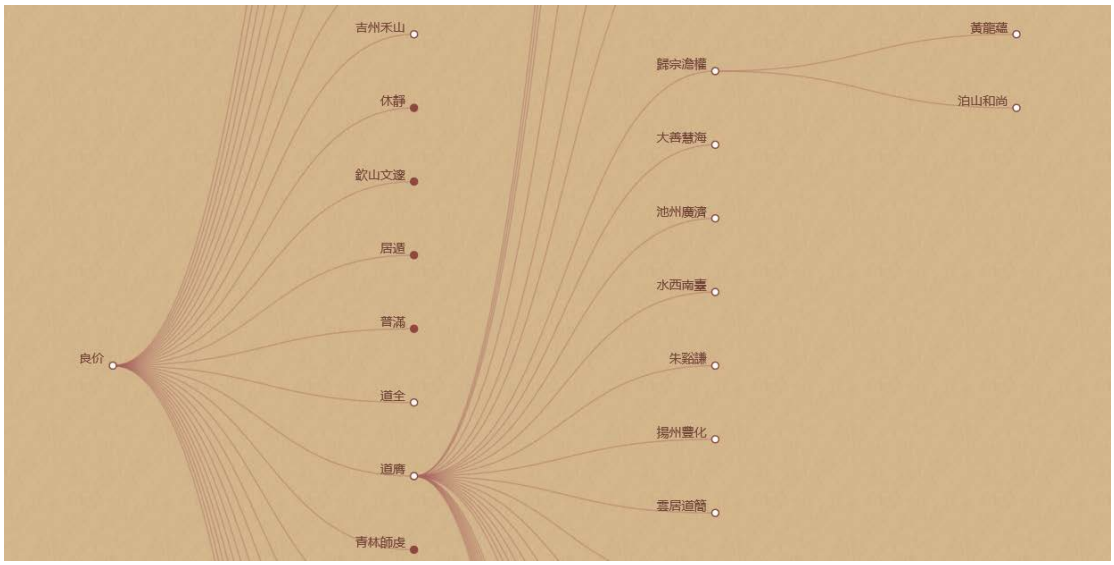


圖 2 曹洞宗法脈傳承圖頁面截圖

在對師徒關係進行視覺化時，樹狀圖層級清晰，但無法呈現一個弟子拜多人為師的情況。這種情況需要網狀圖才能完整呈現，但考慮到呈現為網狀圖後複雜度會提高、清晰度會降低，且這這種情況較為特殊，我們規定仍以樹狀圖呈現，弟子會多次出現，即凡是他拜過的師父的弟子中，都會出現他的名字。

[2] 存在一人同属于多个宗派的情况。

此外，對於法脈傳承圖的一大挑戰是顯示問題：師徒傳承代數多，且一位師父的弟子數有可能極多。考慮到使用者使用習慣一般為上下滾動頁面，我們採用橫向作為代際的延伸，初始顯示四代弟子，若用戶點開新的一代，則根據頁面寬度調整兩代之間的距離；縱向作為一代弟子的展開，對於一人可能有很多弟子，系統會根據弟子數量計算頁面長度，弟子越多則頁面縱向的長度越長，用戶可以通過上下滾動實現流覽全部弟子。

3.3.傳播分佈圖分析與呈現

在禪宗人物的資料中，其中有出生年記載的 1955 人，有籍貫記載的 2205 人，有出生年和地點記載的人物重疊部分為 1163 人，我們選取這部分人物進行禪宗傳播的視覺化。資料庫缺少時間和地點的情況有多種可能，總的來說不外乎兩種可能：一是本身書籍中就沒有記載，二是在資料庫錄入的過程中還沒有加入。對於第一種情況，對人物記載較少的說明這個人物的影響力相對而言是不夠大的；但是也可能存在一些年代久遠不可考的人物，如果影響力較大的話，後人基本也會對其出生年和籍貫有所推測，這也在資料庫中有所記錄。對於第二種情況，隨著資料庫的進一步完善，我們也可以將新錄入的資料補充進來。

各宗派有出生年、地人數及其所占比例如下表所示：

表 2 有出生年、地的宗派人數及比例

宗派	有出生年、地的人數	占該類總人數的比例
臨濟	822	5.96%
曹洞	157	3.30%
雲門	52	7.19%
溈仰	9	8.49%
法眼	29	9.67%
其他	240	17.33%
總計	1163	7.02%

臨濟和曹洞的比例相對較少，意味著在傳播分佈圖中，臨濟和曹洞宗的熱度是被低估的；而在五家之外的人物，包括早期禪宗人物、北宗，以及五宗之外的其他宗派，在傳播分佈圖上會存在被大大高估的情況。這一情況也對系統的顯示效果造成了影響，但是排除掉早期禪宗及北宗（藍色表示）後，相對而言五家宗派的相對比例還是在可接受的範圍內的。

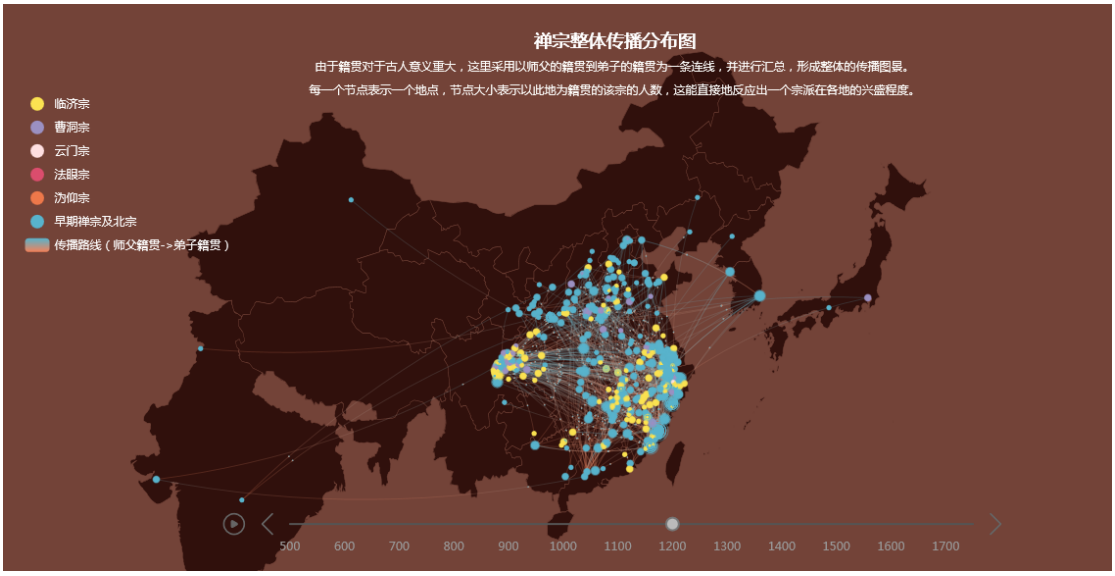


圖 3 禪宗整體傳播分佈圖截圖

對於禪宗在時空範圍內的傳播，我們借鑒了 Schich M 等人在文化歷史方面的視覺化研究，但與之不同的是，我們的資料中人物之間是有聯繫的。我們將每一對師徒的籍貫連線，從師父的籍貫指向徒弟的籍貫（或出生地）。古人的地點相對穩定，即使發生變動，一個人的籍貫也基本可以反映一個人從哪裡來這一問題，因此徒弟的來源地也基本可以反映師父的聲名傳到了何處，也就是我們想要呈現的傳播的情況。

總體資料中人物出生時間跨度為從 300 年到 1800 年，由於 300-500 年間、1700-1800 年間人數較少，我們最終選取了 500-1700 年進行呈現，並將 300-500 年間的人物作為已有資訊呈現在 500 的節點上。各個宗派呈現的時間跨度不同，與其興衰時間有關，具體如下表所示：

表 3 各宗派傳播分佈圖起止年

宗派	開始時間	結束時間
臨濟	900	1700
曹洞	800	1700
雲門	850	1300
溈仰	750	900
法眼	850	1100

在各宗派的傳播之中，我們將有重大意義的事件、人物以及對某一時期該宗派的評論與傳播圖景共同展示，幫助用戶除了瞭解傳播到了哪裡，還能瞭解導致這一現象出現的原因是什麼，或者說是在現實中究竟發生了什麼與禪宗傳播有關的事情。我們主要參考了《中國佛教史》、《中國禪學思想史》 、《增訂本中國禪思想史》、《中國禪宗史》。

例如，在曹洞宗傳播圖中的西元 1200 年，出現了連接到日本的一條線，與此相伴的解釋為：西元 1200 年，希元道元出生於日本京都。他入宋求法，師從天童如淨，回到日本後開創了日本曹洞宗。



圖 4 曹洞宗傳播分佈圖截圖

4.小結

本平臺實現了對禪宗人物師承關係和禪宗傳播情況的大規模資料視覺化實踐，為梳理禪宗發展脈絡、研究禪宗演變歷史提供了新的角度。禪宗歷史傳承悠久，資料不易收集全面系統。未來我們收集更多的資料，以期全面反映禪文化的傳承和流變。

未來我們將在以下幾個方面提升和改進系統：

（1）師徒傳承的樹狀圖不只是宗派的創始人有，從任何有記載的人物開始都可以，因此此處可以支援搜索功能，並對同名人物進行消歧處理。

（2）目前實現的是整體的宏觀的傳播歷程，僅使用人物籍貫的資訊過於粗略，因此下一步可以收集更豐富的資訊，將時間精度調整為以年為單位，呈現出關鍵人物的行跡。

（3）考慮到相較於禪宗的歷史，禪宗的義理與公案可能對於大眾更有吸引力，未來將基於禪話、禪師、地點及其相互關係製作本體並提供更高級、更多維度的流覽和檢索功能。

參考文獻

Schich, M., Song, C., Ahn, Y. Y., Mirsky, A., Martino, M., & Barabási, A., et al. (2014). A network framework of cultural history. *Science,345*(6196), 558-562.

蔣維喬. (2007). 中國佛教史. 上海：上海古籍出版社。

忽滑穀快天. (2002). 中國禪學思想史.上. 上海：上海古籍出版社。

忽滑穀快天. (2002). 中國禪學思想史.下. 上海：上海古籍出版社。

葛兆光. (2008). 增訂本中國禪思想史. 上海：上海古籍出版社。

顧毓琇. (2017). 中國禪宗思想史. 北京：外語教學與研究出版社。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

基於文本挖掘的佛經人物畫像研究

張曉冬 馮國明
北京科技大學經濟管理學院

基于文本挖掘的佛经人物画像研究

1 选题背景及意义

1.1 选题背景

佛经是佛陀说过的话的汇编，是佛教教义的基本依据。其传说部分也许是为了展现佛教的神奇之处；而其哲学与修行部分是值得学习与深思的道理；其将一世因果扩至三世因果无论是否存在都可以合理解释一些科学中无法解释的现象，达到导人向善的作用。

三藏十二部经典，内容可谓浩如烟海。对于佛经阅读者来说，从繁多的经文中快速的定位到自己想了解的人物的相关文章或者定位到人物的简介是非常困难的，这就需要对佛经进行整理和结构化处理。目前佛经的整理和结构化工作很大程度上依赖于人工，海量的经文在进行文本处理工作时会耗费大量的人工成本和时间成本，而且人工处理存在标准不一致和失误的现象。

随着自然语言处理技术和机器学习技术的发展，中文文本处理的自动化技术已经应用到了很多领域，例如自动翻译、摘要生成和人物画像等。其中，人物画像技术在电子商务领域得到了广泛的应用，利用人物画像进行个性化推荐和用户群识别得到了非常好的效果。然而由于经文的词语构成、语法结构以及用字方面的特殊性，经文的文本处理手段不同于现代文，目前大部分自然语言处理技术都只是应用在了现代文，在佛经领域，自然语言处理技术的应用还鲜有研究，将人物画像技术应用于佛经领域，能填补自然语言处理技术在经文中的研究空白，推动经文的自动化研究的进展。对佛经的知识库构建有一定意义，同时能有效提高佛经的阅读方便程度。

1.2 选题意义

在理论方面，对于古文，尤其是藏经的研究甚少。由于经文的词语构成、语法结构以及用字方面的特殊性，经文的文本处理手段不同于现代文，目前大部分自然语言处理技术都只是应用在了现代文而未在经文中进行应用研究。将一些自然语言处理技术针对经文进行改进并加以应用能够填补这个研究空白，推动自然语言处理技术在古文或特殊文本中的研究进程。

在现实意义方面，文本结构化是文本统计分析、快速索引等应用的前提和基础，将文本进行结构化处理能使其便于在计算机中合理存储、统计分析和快速索引。随着佛经网络化、数字化的发展，大量经文被收集和整理，但繁多的经文目前均是以非结构化的形式存储，进行统计分析时或者实际阅读过程中有很大的不便。将自然语言处理技术应用于佛经领域，针对佛经进行人物画像的抽取可以为佛经的知识库构建提供一定的贡献，为其他工作提供基础，同时能有效提高佛经的阅读方便程度。

2 研究内容及研究方法

2.1 研究内容

本文计划以佛经为研究对象，以结合字典与 BI-LSTM-CRF 的 DBLC 模型为基础针对佛经进行优化完成自动分词和实体识别的任务。人工总结术语模板，得到术语初始代表词集，依据初始代表词集利用正则表达式对语料中的句子进行匹配分类，通过人工校验确保分类正确以获得初始分类集。以初始分类集作为语料，基于本人提出的改进的 TF-IDF 方法提取类别关键词扩充术语代表词集，进一步

扩充分类集。最后在分类集的基础中基于规则提取实体的标签值。主要研究内容如下：

（1） 实验语料的处理。

a. 获取佛经原始语料和词典；

b. 语料的清洗（去除特殊字符和无用字符，统一格式）和粗分词（使用 DBLC 模型）；

c. 人工矫正粗分词结果，得到准确的分词结果；

d. 在准确的分词结果中抽取新词扩充佛经词典，将准确的分词结果转化为序列标注结果；

（2） 使用经过标注的语料训练 DBLC 模型，并使用训练好的模型对所有语料进行序列标注，根据标注结果提取实体；

（3） 人工总结术语模板，总结术语关键词，得到术语初始代表词集；

（4） 依据初始代表词集，利用正则表达式对语料中的句子进行匹配分类，并对句子进行添加主语和代词替换处理。通过人工校验确保分类正确以获得初始分类集；

（5） 以初始分类集作为语料，使用本人提出的改进的 TF-IDF 方法对各类别下的词在该类别中的代表程度进行评分，选取前 n 个作为类别关键词扩充术语代表词集，进一步扩充分类集。

（6） 在分类集中基于术语代表词集和实体与术语的关系规则提取实体的标签值，以获得人物画像；

（7） 基于人物画像的聚类分析，使用 k-means 方法；

（8） 实验结果评价与分析。

2.2 研究方法

本次研究中，计划分别采用下述研究方法：

（1）文献研究法：通过收集、鉴别、整理文献获得资料，从而全面地了解中文文本挖掘和佛经文本处理的发展和研究现状。此外，通过研读关于实体识别、人物画像技术的主要文献，在整体上把握已有研究状况，对确定研究思路和方法提供依据。

（2）实证研究法：在序列标注方面，通过对基于字序列标注的相关文献的对比分析，选定目前适用于领域特殊文本序列标注的较好的分词模型——DBLC 模；在关键词提取，选用改进的 TF-IDF 方法；在短文本分类中，由于已有类别关键词，故选用基于规则的正则表达式匹配方法；在聚类分析方面，使用经典的 k-means 方法。上述方法均已经过验证在各自的实验数据上获得了较好的效果。

（3）软件辅助分析法：本次研究计划使用 python 编程对上述研究方法和内容进行实现和实验，使用 echarts 进行结果的可视化与分析。

2.3 技术路线图

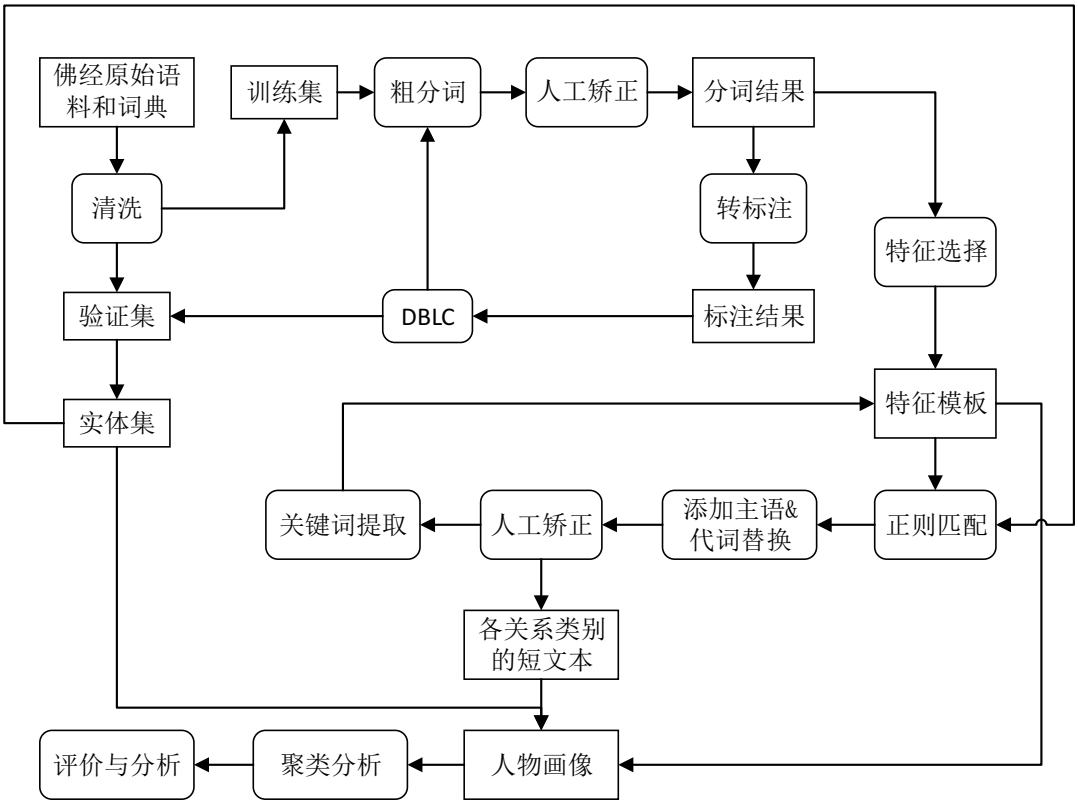


图 2-1 技术路线图

3 研究的创新点和难点

本文的创新点在于：

（1） 将人物画像的研究方法应用于佛经；

（2） 提出了基于 TF-IDF 改进的关键词提取方法；

（3） 提出针对佛经文本改进的 DBLC 模型；

（4） 使用模式匹配与半监督的方式抽取人物画像。

本文的难点在于：

（1） 研究过程含有较多人工处理过程；

（2） 佛经有专业知识门槛；

（3） 实验所涉及的 DBLC 模型需要针对佛经进行改进；

（4） 半监督的学习结果不易评价。

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) - 21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

聖經經節引用的擷取與應用

Extraction and Application of Bible Verse Citations

陳俊良* 林清峰**
長榮大學數位內容設計學系及台南市立安南醫院*
長榮大學數位內容設計學系**

聖經經節引用的擷取與應用[1]
Extraction and Application of Bible Verse Citations

陳俊良[2]、林清峰[3]

摘要

基督教會的主日禮拜都會引用聖經經節。被引用的經節經常可以反應出該教會該週關心的議題。過去不靠資訊技術，很難收集各教會主日禮拜引用的經節，更難對這些資料進行搜尋、探勘。本研究的目的在於收集教會週報、擷取經節引用、然後加值人文利用。

本文是本研究的初步成果，比如我們可以回答過去幾年某間教會最常引用的是聖經的哪一節。據此，我們判斷這是一個可行的數位人文研究方向。

關鍵詞：大數據、資料視覺化、聖經研究、聖經經節引用。

前言

聖經是猶太教、天主教、東正教、基督教(新教)等的共同經典，是有史以來印行最多本的書。內容不僅直接影響這些宗教的信徒，也間接影響到未信仰這些宗教的人們。

教會的例行活動是每週的主日禮拜，會分發教會週報。每個教會的週報涵蓋內容不一，有主日禮拜程序、金句、詩歌、講道稿、代禱、奉獻、報告事項、…等，絕大部分都有主日禮拜時牧師講道引用的經文。

講道是一個單向的活動，牧師講，會眾聽。每週全台灣幾十萬人坐在教堂接收牧師透過選擇的聖經經節傳遞的訊息。主日禮拜是區域性的活動，影響數十人、數百人、頂多千人。這是小眾傳播。主日禮拜是週期性的活動，一週一次；是持續性的，延續幾十年、百年。

假如我們可以收集到很多教會、一段時間的週報，彙整引用的聖經經節，可以試想一些問題。哪些是熱門的經節，被高度引用？是否有區域性？是否南北有別？是否城鄉有別？是否教派有別？

要分析社會輿情，長時間的觀察是必要的。報紙、雜誌、電視、廣播等大眾傳播都是長期經營的，不過他們呈現的輿情是全國性的。相對的，透過幾十年來各地教會週報引用的聖經經節，不知不覺中我們已經在台灣佈建了好幾千個感應器 (sensor)，每週偵測當地輿情。這是一個不插電的人文物聯網 (IoT, Internet of Things)。

有鑑於此，我們擬回收此一人文物聯網的訊息，建立數位平台，然後加值人文應用。

聖經經節引用格式

聖經經節的引用會標註出某卷、某章、某節。英語世界學術界常用的引用格

[1] 本文是科技部計畫 MOST-106-2420-H-309-005-MY3 (整合型計畫：聖經章節數位平台建置與人文應用，子計畫：以聖經章節為定義域之資料視覺化研究) 的部分成果。

[2] 長榮大學數位內容設計學系 及 台南市立安南醫院

[3] 長榮大學數位內容設計學系

式，如 PAP [1], MLA [2], Chicago [3]，都有專節討論如何標註聖經的引用。專門的 Society of Biblical Literature 也有訂定其格式 [4]。英文的經節引用標註方式差異不大。「教養孩童，使他走當行的道，就是到老他也不偏離。」是 Proverbs 22:6。經卷也可用縮寫，如 Pr 22:6、Prov 22:6，有些會有縮寫的點，如 Pr. 22:6。章節的編號都用阿拉伯數字。章節之間普遍用冒號，但是 MLA 用點，如 Prov. 22.6。相對於英文，華語世界有些神學院也訂定了要求其學生在學期間撰寫報告的經節引用格式 [5][6][7][8]，但是差異性比英文的大。神學院學生進入教會服務後，在教會週報的標註方式更五花八門。底下都是可能的中文寫法：箴言第二十二章第六節、箴言 22 章 6 節、箴言２２章６節、箴 22:6、箴二二 6、…。甚至還有箴言廿二章第六節；請留意 六 與 六 的不同。

聖經經節引用擷取

我們的作法分成兩階段，首先快速地找出下一筆可能的經卷名稱，其次再解析跟隨的經章和經節編號。

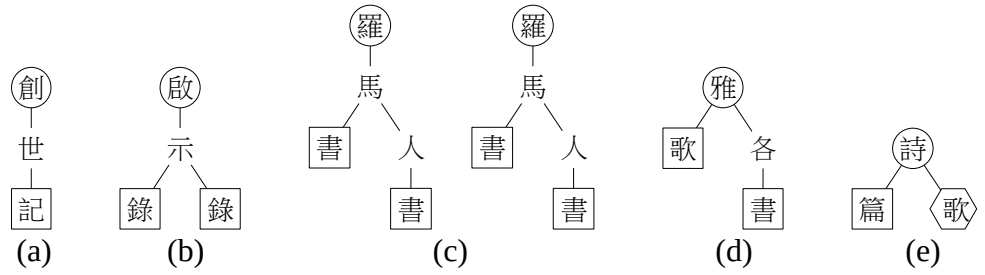
經卷名稱的擷取

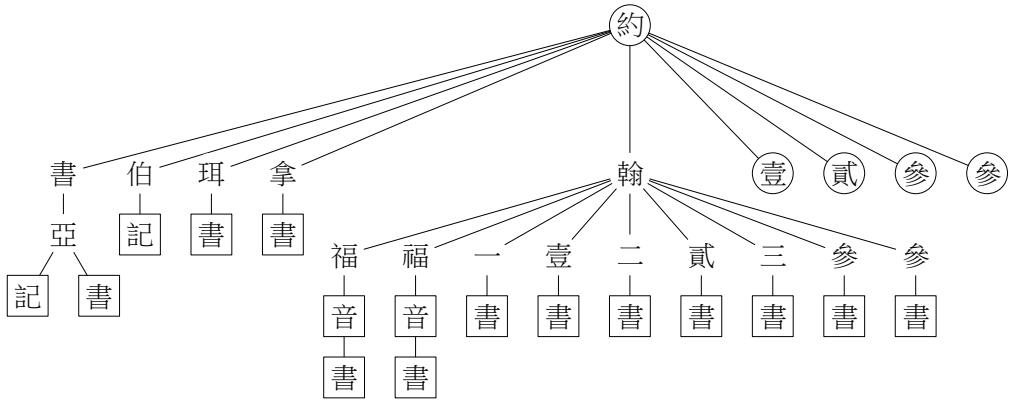
中文經卷名稱的判讀有幾點必須留意：

- 經卷名稱可能是全名 (如：羅馬書) 也可能是簡名 (如：羅)。
- 經卷的全名或簡名可能不唯一 (如：羅馬人書)。
- 經卷名稱採用的中文碼可能不唯一 (如：羅馬書)。請留意 羅 與 羅 的不同。前一個 羅 的 Unicode 7F85 屬於 CJK Unified Ideographs [9]，這是常用字的中文碼區段 (4E00-9FFF)。後一個 羅 的 Unicode F90F 屬於 CJK Compatibility Ideographs [10]，這是相容的中文碼區段 (F900-FAFF)。

為應付這些課題，我們有一個前置作業，輸入所有可能的中文經卷全名、簡名，同時輸入相容中文字碼對應常用中文字碼對照表，輸出術語樹 (Term Tree)。術語樹是 trie 或是稱作 prefix tree 的變體。

圖 1 是術語樹的示意圖。圓形標記和方形標記分別代表經卷簡名和經卷全名，而沒有框框的則是未完整的名稱。圖 1(a) 呈現 創 是一個簡名，創世記 是一個全名，而 創世 只是一個判讀的過程不是一個完全的名稱。碰到相容字時，如 啟示錄 與 啟示錄，可以利用圖 1(b) 樹枝分叉的方式呈現不同的ㄉㄨㄟˋ。圖 1(c) 表現出了 Romans 2 種可能的簡名以及 4 種可能的全名。一棵術語樹可以同時呈現有相同第一字的不同經卷，如圖 1(d) 同時代表了 雅歌 的簡名和全名以及 雅各書 的全名。圖 1(f) 是最大的一棵術語樹，呈現了第一字是 約 的 21 個卷名。





(f)

圖 1：術語樹

我們擷取經卷名稱的方式是，每一棵術語樹配置一個指標指到文章的最前頭，如圖 2 的上半段。然後從頭開始找到對應的第一個字，如圖 2 的中段。排序這些指標，最左邊的就是下一筆可能的經節引用的經卷名稱。接著利用術語樹來決定對應的經卷簡名或全名，如圖 2 的中段將擷取出 約翰福音。當判斷完經節引用後，把 約 的指標移動到下一個 約，如圖 2 的下半段。重複上述流程就可以依序找出經卷名稱。

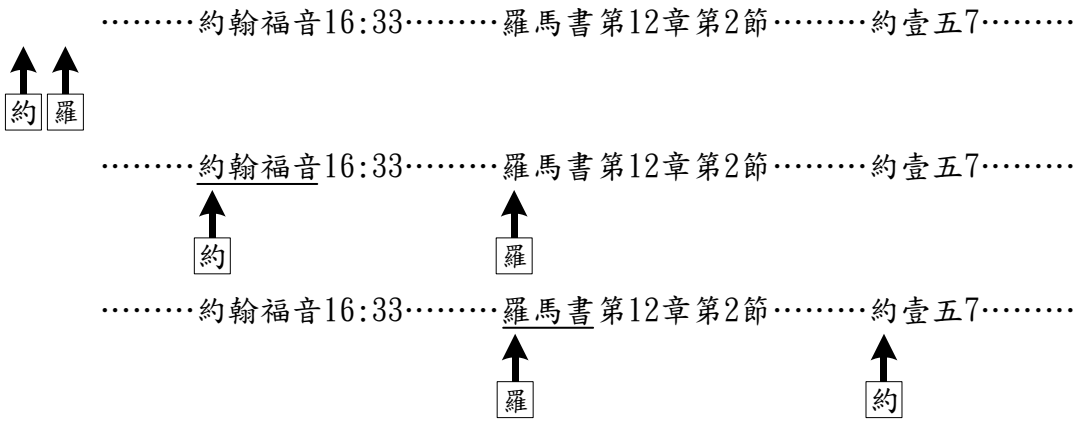


圖 2：經卷名稱擷取過程

經章、經節編號的擷取

得到經卷名稱後，就可以解析跟隨的經章和經節編號。下面是經節引用表示法的文法。

```
BookChapVerse   →  Book
                |  Book  ChapRangeSeq
                |  Book  ChapRangeSeq  VerseRangeSeq
ChapRangeSeq    →  ChapRange  { AndOperator  ChapRange }
ChapRange       →  Chap  [ ToOperator  Chap ]
Chap            →  [ OrdinalPrefix ]  Number  [ ChapPostfix ]
VerseRangeSeq   →  VerseRange  { AndOperator  VerseRange }
VerseRange      →  Verse  [ ToOperator  Verse ]
Verse           →  [ OrdinalPrefix ]  Number  [ SubVerse ]
[ VersePostfix ]
                |  [ OrdinalPrefix ]  Number  [ VersePostfix ]
[ SubVerse ]
```

Number	→	中文數目 ｜ 半形阿拉伯數目 ｜全形阿拉伯數目
OrdinalPrefix	→	第
ChapPostfix	→	章 ｜ 篇 ｜ ： ｜ :
SubVerse	→	a ｜ b ｜ c ｜ d ｜ a ｜ b ｜ c ｜ d
VersePostfix	→	節
ToOperator	→	到 ｜ 至 ｜ ～ ｜ ~ ｜ — ｜ -
AndOperator	→	和 ｜ 及 ｜ 、 ｜ ， ｜ , ｜ ； ｜ ;
Book	→	經卷全名 ｜ 經卷簡名

其中的 ChapRange 以及 VerseRange 是用來描述編號區段。比如，徒八 7-12 代表了使徒行傳第 8 章第 7 節到第 12 節。ChapRangeSeq 以及 VerseRangeSeq 是用來描述編號區段序列。比如，徒八 7-12,22-24 還要加上第 22 到 24 節。這種區段以及區段序列經節引用是很常見的，也加大了經節引用擷取的難度。

特殊的擷取策略

實作過程中，碰到了很多系統分析階段沒有規劃到的情境；這是可以預期的，因為我們的對象是人與人溝通的自然語言，不是嚴謹的程式語言或數學。底下略舉一二。

- 中文經節引用中的編號區段以及編號區段序列超乎我們想像的不規則，前述的文法沒有描述完整。
- 聖經中的俄巴底亞書、腓利門書、約翰二書、約翰三書、猶大書等都只有 1 章，有時經章編號會被省略只寫出經節編號，如猶大書第 1 章第 21 節寫成 猶 21。
- 單字的經卷簡名會產生許多問題。
  - 「保守大家一路平安」將被擷取出路加福音的 路。很明顯這是多餘的，我們的策略是忽略沒有跟隨經章經節編號的經卷簡名。
  - 教會週報很常見地址。「台南市歸仁區長大路一號」將擷取出路加福音第 1 章的 路一。我們的策略是多判斷編號後面是不是量詞 (排除章、篇與節)，假如是其他量詞就不視為跟隨的編號。此一策略可以濾掉很多類似的情形，如「開啟一扇門」的 啟一、「提出三點建議」的 出三。
  - 經卷的單字簡名幾乎都是很常用的字，所以是很多雙字詞或多字詞的部分，如：歡迎參加的 加、詩歌教唱的 歌。建立術語樹時，我們多提供了迴避詞，比如參加、詩歌，如此就可以避免誤判讀 加、歌。圖 1(e) 的六邊形代表 詩歌 是一個迴避詞。

擷取結果可信賴分數

有再多的策略也不可能把經節引用擷取做到百分之百正確，因為面對的終究是自然語言。我們對每一個擷取出的經節引用給了可信賴分數。假如格式正確、編號在範圍內等等條件都滿足，給予高分，未來擬直接匯入資料庫中立即使用。假如很明顯的可以排除，比如迴避詞，給予低分，未來擬直接捨棄。假如模擬兩可，那麼給予中等分數，未來擬讓人工介入判讀。

系統架構

圖 3 是目前的經節引用擷取器的操作流程。

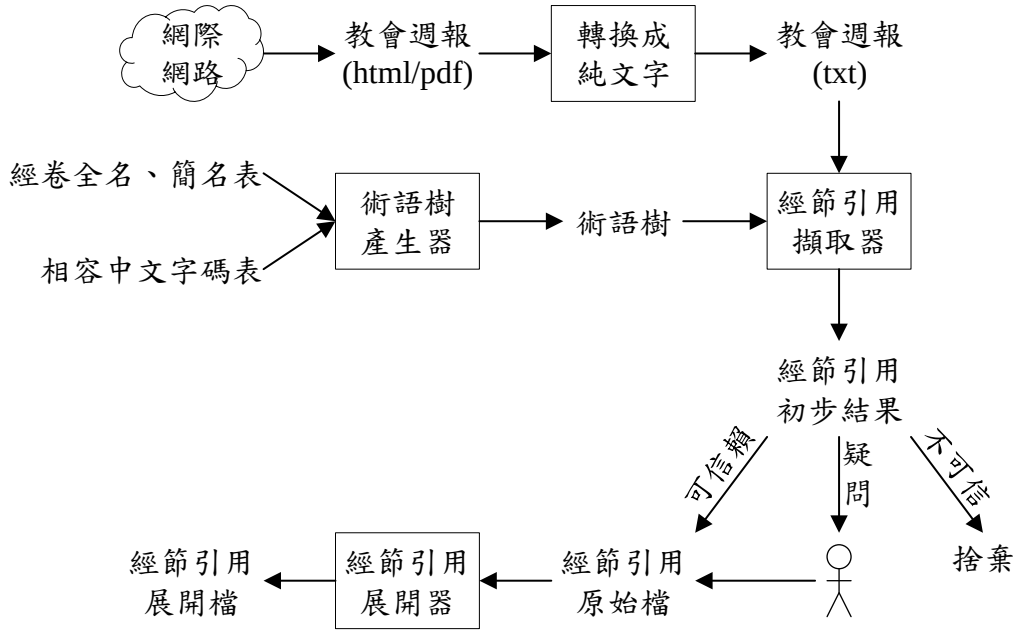


圖 3：經節引用擷取器系統架構

聖經經節引用應用

我們先從網路上抓取了台灣基督長老教會太平境教會 (自 1997/11 起)、真耶穌教會台北教會 (自 2005/06 起) 的教會週報當成測試資料。我們開發了初步的聖經經節引用分析器。底下是一些得到的結果。

最常用的聖經經節

相信絕大多數的教會或是其信徒沒有辦法精確說出過去若干年該教會哪一經節引用最多次。表 1 是台灣基督長老太平境教會和真耶穌台北教會最常引用的聖經經節列表。

可以觀察到兩個教會的引用有別。唯一兩教會都是前 20 名的只有羅馬書第 12 章第 2 節；而太平境教會第 17 名以賽亞書第 55 章第 7 節以及真耶穌台北教會第 12 名約翰一書第 5 章第 7 節居然在另一教會一次的引用都找不到。

表 1：最常引用的聖經經節

聖經經節	台灣基督長老太平境教會 次數 (排名)	真耶穌教會台北教會 次數 (排名)
Acts 1:8	32 (1)	11 (99)
Prov 3:6	30 (2)	9 (164)
Prov 3:5	29 (3)	8 (204)
Heb 4:16	28 (4)	7 (309)
Prov 22:6	27 (5)	16 (34)
Rom 12:2	26 (6)	25 (2)
Eph 4:16	26 (6)	8 (204)
2Cor 5:17	26 (6)	6 (463)
Mic 6:8	25 (9)	11 (99)
Isa 55:6	25 (9)	2 (2480)

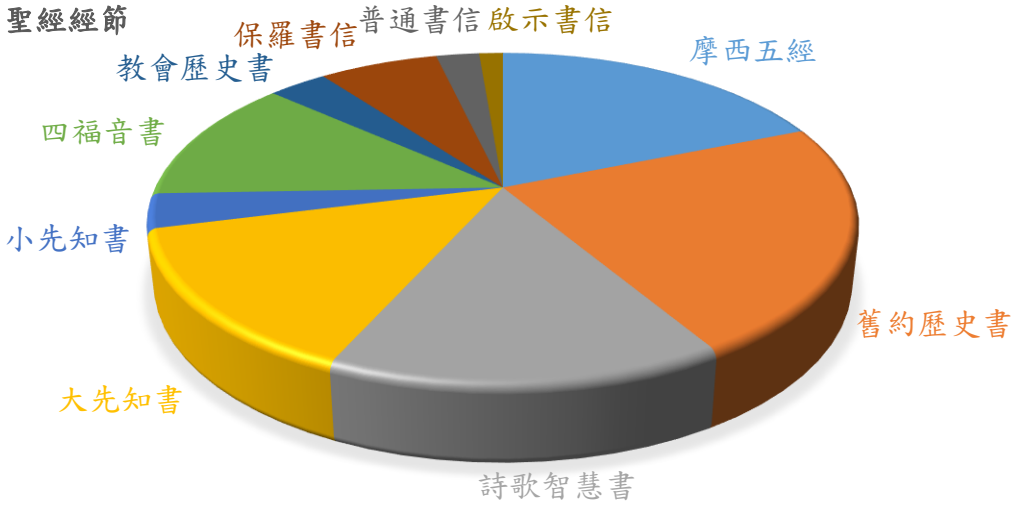
聖經經節	真耶穌教會台北教會 次數 (排名)	台灣基督長老太平境教會 次數 (排名)
Phil 4:6	26 (1)	10 (312)
Rom 12:2	25 (2)	26 (6)
Phil 4:7	25 (2)	11 (218)
John 16:33	25 (2)	8 (587)
Gal 5:22	22 (5)	12 (164)
John 3:5	21 (6)	13 (124)
Acts 2:38	21 (6)	5 (1798)
Gal 3:27	21 (6)	4 (2522)
Eph 4:24	21 (6)	3 (3779)
John 14:16	20 (10)	11 (218)

Matt 28:19	24 (11)	12 (76)
Rom 12:1	23 (12)	14 (51)
Eph 4:12	23 (12)	6 (463)
Isa 6:8	23 (12)	6 (463)
Heb 4:14	23 (12)	5 (707)
Pss 51:10	22 (16)	5 (707)
Eph 4:2	21 (17)	6 (463)
Eph 4:11	21 (17)	2 (2480)
Pss 105:1	21 (17)	2 (2480)
Isa 55:7	21 (17)	0 (8303)

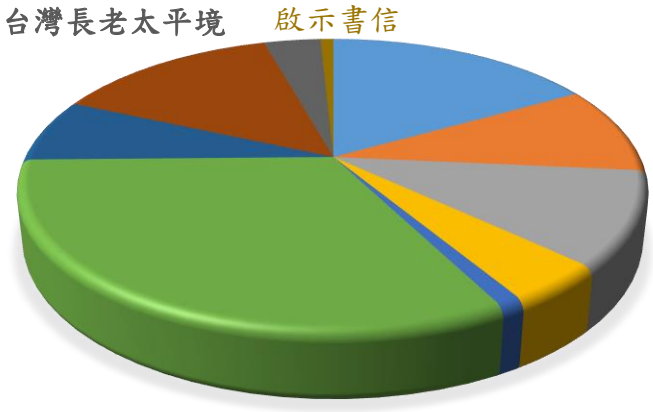
1Cor 15:58	20 (10)	6 (1165)
John 14:6	19 (12)	16 (62)
Acts 2:4	19 (12)	14 (104)
John 1:1	19 (12)	11 (218)
Pss 46:1	19 (12)	10 (312)
Eph 4:22	19 (12)	2 (5594)
1John 5:7	19 (12)	0 (18943)
John 3:16	18 (18)	16 (62)
Acts 2:3	18 (18)	15 (86)
Rom 8:28	18 (18)	15 (86)

不同教會引用經節的比較

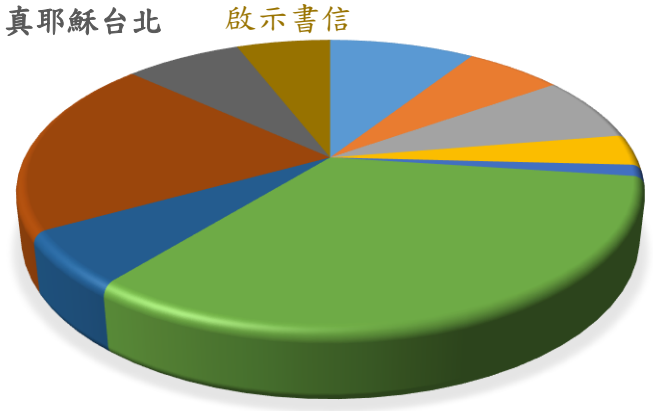
不同的宗派不同地區的教會引用的經節是否有別？圖 4 圓餅圖是台灣基督長老太平境教會和真耶穌台北教會引用經節的比較，採用的分類把舊約聖經分成摩西五經、舊約歷史書、詩歌智慧書、大先知書、小先知書，把新約經分成四福音書、教會歷史書、保羅書信、普通書信、啟示書信，共 10 類。圖 6(a) 是聖經總共 31,103 經節的數量比例，舊約聖經佔了約四分之三，新約聖經約四分之一。圖 6(b) 和 6(c) 分別是兩教會引用經節歸屬分類的比例。可以觀察到兩教會引用的經節均偏重於新約，太平境教會約六成，新耶穌台北教會約四分之三。偏重新約是目前台灣基督教共有的現象，並不令人意外。另外還可以觀察到，真耶穌台北教會引用的啟示書信比例顯著的多於台灣長老太平境教會。啟示書信事實上只有一經卷，即聖經最後一卷的啟示錄。這是宗派的區別嗎？



(a)



(b)



(c)

圖 4：不同教會引用經節比較圓餅圖

啟示錄歸屬於啟示文學的範疇，舊約的但以理書也是。為進一步理解教會引用但以理書的情形，我們把經節引用歸屬到聖經的 66 經卷。切成 66 份的圓餅圖將不易閱讀，所以我們用圖 5 的文字雲來呈現，一個詞單一字母的面積約略就是該分類的比例。可以看到，真耶穌台北教會引用的啟示錄 (Rev) 大於台灣基督長老太平境教會的；這是跟圖 4 吻合的。不過，兩間教會引用但以理書 (Dan) 的比例差異不大；有賴神學與人文研究者進一步的解釋了。



圖 5：不同教會引用經節比較文字雲[4]

[4] 此一文字雲是利用 WordClouds 線上工具產生的。https://www.wordclouds.com/

未來努力方向

本文所述乃為了驗證此一數位人文研究方向的可行性。為擴大研究成果，我們尚須在下列事項努力：

- 系統性，更全面收集教會週報，再擴及其他基督教文件。
- 提高經節引用擷取的準確度。並加註引用類型，如：講道經節、金句、宣召、會務報告、...等。
- 除經節引用外，自動擷取其他有用資訊，如：講道題目、講道者、(與社會事件有關)代禱事項、...等。
- 了解人文學者分析研究的途徑，並建置人文學者易操作之介面。
- 自動化產出資訊圖。
- 導入機器學習、深度學習，以探勘更多值得研究的議題。
- 建置公開平台，提供成果大家共享。

圖 6 是整體的構想，其中斜體部分是尚待努力的。

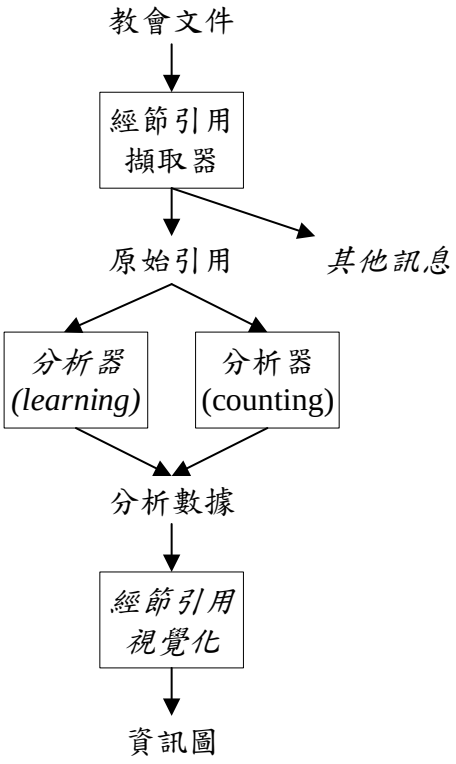


圖 6：整體構想

結語

本文展現了一個藉由數位手段擷取並分析聖經經節引用，進而協助人文與神學學者從事進一步研究的途徑。參看圖 6，只靠分析數據或是資訊圖，就看圖說故事然後下結論，是下策，不該是數位人文的正途。依靠分析數據或是資訊圖，加入有質量的神學和人文的解釋，是中策。

上策應該是，先觀察分析數據或是資訊圖，然後回頭看原始引用甚至是原本的週報，以便做進一步的人文研究。終究數位人文合在一起講時，“數位”的角色是協助“人文”更快的切入研究的重點，不該喧賓奪主。

參考文獻

[1] *The Publication Manual of the American Psychological Association*, 6th ed., Washington, DC: American Psychological Association, 2011.

[2] *The MLA Handbook for Writers of Research Papers*, 8th ed., Modern Language Association of America, 2016.

[3] *The Chicago Manual of Style*, 17th ed., Chicago, IL: University of Chicago Press, 2017.

[4] Society of Biblical Literature, *The SBL Handbook of Style*, 2nd ed., SBL Press, 2014.

[5] 台灣神學研究學院教務處, *台灣神學研究學院研究報告格式手冊*, 2017. http://www.taitheo.org.tw/bin/downloadfile.php?file=WVhSMFlXTm9MemN2Y0hSaFh6RTROakJmTVRVd05UTTBNMTg1TlRBNE5DNXdaR1k9

[6] 聖經網, *Manual of Style 文字格式*, http://www.aboutbible.net/Ab/A.L.08.ManualOfStyle.html

[7] 天主教輔仁大學神學院研究所, *註腳暨書目格式* http://theology.catholic.org.tw/alantung/Gbookform.doc

[8] 衛道神學研究院教務處, *專論寫作格式指引*, 2009. https://uwgi-edu.org/seminary/index.php/2014-12-31-09-55-23?download=33:pdf

[9] *CJK Unified Ideographs*, https://unicode.org/charts/PDF/U4E00.pdf

[10] *CJK Compatibility Ideographs*, https://unicode.org/charts/PDF/UF900.pdf

第九屆 DADH 2018
數位典藏與數位人文國際研討會
9th International Conference of Digital Archives and Digital Humanities
12/18 (W2) -21 (W5)
Venue 地點 / 法鼓文理學院
Dharma Drum Institute of Liberal Arts (DILA)

「曹錕賄選」與知識份子群體的政治選擇

以胡適日記（1923）為中心的觀察

蔡明遠
國立臺灣大學

「曹錕賄選」與知識份子群體的政治選擇：以胡適日記（1923）為中心的觀察

主旨：

「1928 年的中國展望一片光明，雖然內部還是面對許多問題，且尚無力解決，但是在當時國民黨的軍事力量以及時年 41 歲的蔣介石強力領導之下，已經少有中國人懷疑國民黨對於管理公共事務的能力，而其本身早也藉由歷史證明瞭其為了公眾的目的，有著動員以及引導民眾力量的組織。」[1]

《劍橋中國史》這段話說明國民政府於北伐（1926-28）成功後，地位大幅上升。後續學人的研究中，亦認為國民黨政權因此獲得統治正當性。[2]然而，筆者認為民國時期，在關注知識份子的視角下，尚有以下三個待追問的課題。

首先，北伐論述強調國民黨作為辛亥革命以及北伐的主要行動者，用以表示敵對政權被革命推翻的必要性，[3]然而當時位居北方的知識份子會是如此認知嗎？[4]其次，位居北方的知識份子，這群內部組合多元、非同質性的群體，於五四運動（1919）爆發後，以北京政府為當時中國合法政權的現實出發，他們表現出的政治選擇或是政權認同為何？最後，做為一個特別的階層，北方知識份子群體內部的相互人際關係為何？他們是以何種方式形成連結？這種連結有怎麼樣的特性？又以何種方式散發出影響力？

本文將集中討論「曹錕賄選」這一個案，嘗試藉由社會網絡分析（Social Network Analysis）觀察在「曹錕賄選」事件前後，不同知識份子群體的政治態度與反應為何？討論將集中在上述所提及的三個帶追問課題。

筆者希望藉由新方法與視角，重新觀察「曹錕賄選」除了帶來政權更替之外，是否也帶給當時知識份子群體一些學界過往較少注意到的影響。

方法：

本文將不採取過往的人群分類方式，如省籍、學派、政黨、文人期刊等，筆者想從史料與事件出發，藉由數位人文方式凸顯民國初年的人群網絡，[5]並試圖展示這組人群網絡及其背後微妙、隱約影響到表象行為的複雜性。基於前述三個有待追問的課題，特別是第二點，筆者欲以人際關係為基礎，進一步審視「知識份子群體」與「政治選擇」兩者之間的連結。

[1] 費正清(編)，《劍橋中國史》民國篇(上)（台北：南天出版社，1999），頁 884。

[2] 如羅志田，《亂世潛流：民族主義與民國政治》（北京：中國人民大學出版社，2013），原序頁 4。

[3] 這由「北洋政府」一詞當時所帶有的兩種意涵便可得知：一、地方性私人集團竊奪的國家政權；二、與南方國民政府的革命政權為敵的軍閥政權。桑兵，《歷史的本色：晚清民國的政治、社會與文化》（廣西：廣西師範大學出版社，2016），頁 245。

[4] 若是，則這群知識分子是如何認知管理自身的北洋政府？形成他們地域及政治上認同歧異的原因為何？若否，他們如何評斷更像是入侵者的南方國民黨政權？換言之，筆者認為這觸及知識份子如何認知國家、政權兩者，以及其與自身之間的關係。

[5] 此處主要將採取人際網絡分析的方式，方法上的討論詳見後文。

以人際關係（特別是朋友群）為重心，[6]重新去連接這群知識分子的原因在於：我們可以更容易觀察到跨越傳統分類的交流，也能夠以新的方式去解釋這些交流存在的原因及可能性。若再進一步考慮到民國初期，在空間與時間上都十分集中在沿岸大學、出版社等據點散發影響力的知識圈與學人，那麼藉由人際關係去觀察知識份子們行為將是值得嘗試的。

本文選擇以胡適為重心討論的原因在於其 1917 年回國後，很快成為當時意見領袖，並藉由不同刊物擴散其言論影響力。藉由胡適所串連起的人際關係，雖然其相關人物所能提供的文本材料量十分龐雜，要完全閱讀乃至於建成資料庫難度極高。但正因為這些數量龐大的材料，才讓我們得以觀察民國初年複雜且開放的人際網絡。

本文雖以胡適所存材料為基準出發，但這不代表要將胡適視為這個「群體」的唯一領導人進行討論，或是只將討論的範圍侷限於胡適及其社群。筆者更希望呈現的是胡適及其社群如何認知及面對其他群體，換言之，希望觸及群體與群體間的互動與往來。

本文以「政治選擇」這一「行為」作為觀察視角的原因在於：民國初年位於北方的知識份子不斷地面臨政權可能的轉移，[7]且身處於各式政治風波、學潮，乃至於西方政治思想的輸入、宣傳等。這代表此時的北方知識份子，不可迴避的要一次次回答「政治選擇」的問題，[8]其中一個例子正是「曹錕賄選」。

最後，筆者認為需要更進一步回答的則是知識份子的互動是基於何種理念或核心價值？這種核心價值能否幫助知識份子超越政治立場的隔閡？[9]是故，筆者關心的是知識份子內部解析，而不全然聚焦於知識份子及其輿論與政府的互動關係。筆者嘗試藉由回答這樣的問題，重新理解當時知識份子的行為模式。

[6] 實際上，人際關係並不是單純的只有友誼的選項，它也包含負面的相處方式。以中國歷代人物傳記資料(CBDB)為例，裡面所記錄的人際關係有相當多種，如恩主、推薦、攻訐、彈劾等。在這邊以友誼為主的原因在於史料的性質，日記及書信通常多是以友人的相關紀錄為多。而有關於「友誼」相關的研究，中國近現代史領域幾乎沒看見專著，目前筆者所見較相關者為 Hu Ying, “Burying Autumn: Poetry, friendship, and Loss,” 但值得注意的是這樣的研究傾向在上古及中晚明已經累積了相當的成果。此部分可詳見詹前倬，〈中晚明儒者友誼研究〉一文中的討論。

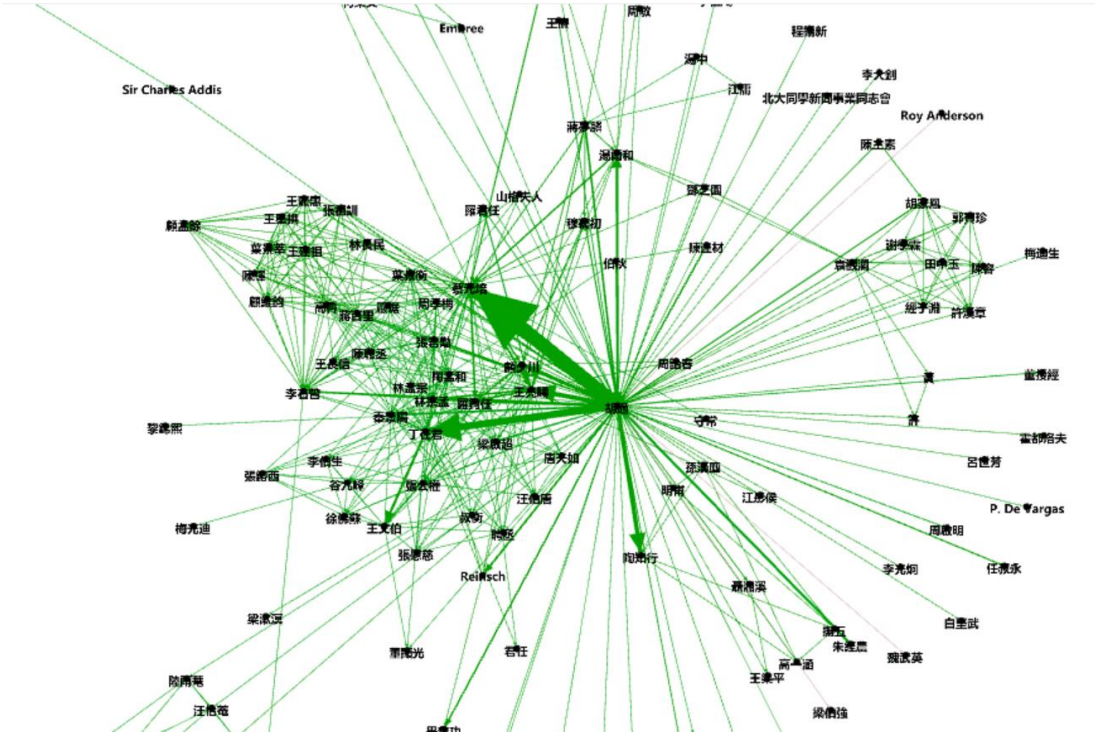
[7] 從 1916 年袁世凱取消帝制開始，一直到 1928 年國民黨政權統一中國，中間 12 年歷經七人出任過總統或國家元首八次；四個攝政內閣；一次復辟。另外尚有二十四屆內閣；五次國民議會或國會；至少四部憲法或基本法。詳見《劍橋中國史》民國篇（上），頁 378-379。

[8] 我們必須理解，知識份子是一群面對政治、參與政治的階層，筆者認為以這樣一種對知識份子最基礎的定義出發，去關注他們在政治上的選擇與表態是一個合適的選擇。

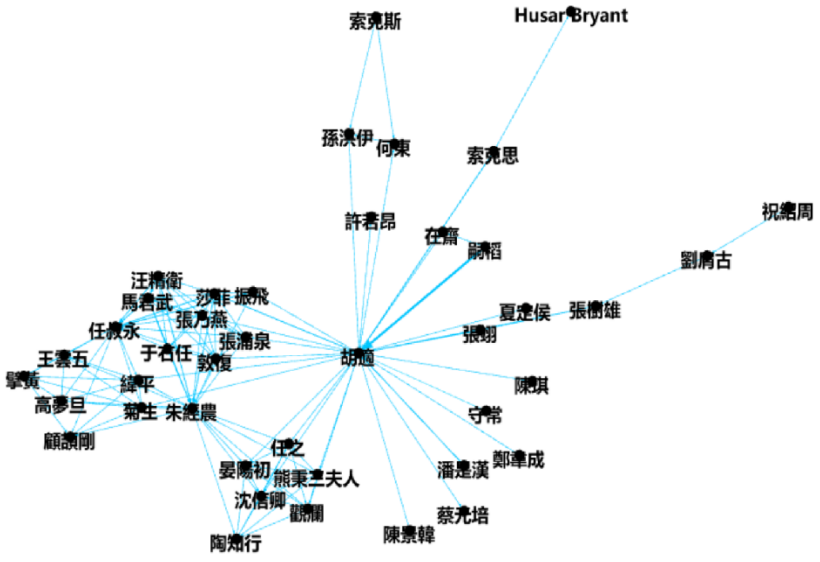
[9] 在這裡，筆者的猜想，「菁英認同」是這群知識份子結合的核心，換言之，要在這個群體中得到發言和被重視的權力，他們必須要有相對應的知識水平。有關民國初年「菁英」相關的討論，例如李金銓(等)，《報人報國：中國新聞史的另一種讀法》(香港：中文大學出版社，2013)，頁 3、5、8-9、33；章清，《胡適派學人群與現代中國自由主義》（上海：上海古籍出版社，2004），Ch1-2；章清，《清季民國時期的「思想界」》（上海：社會科學文獻出版社，2014），Ch1-3。較常被徵引的材料有丁文江，〈少數人的責任〉、胡適，〈我們的政治主張〉等。然而筆者認為還有繼續發展和細緻說明的空間。

試作：

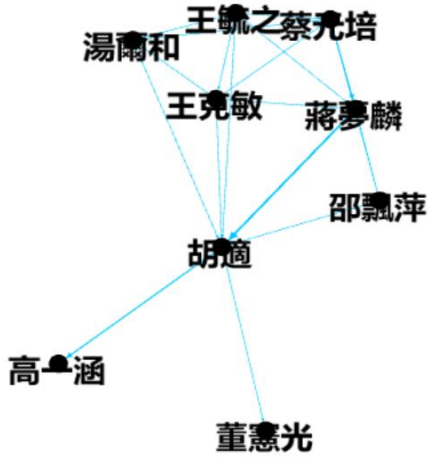
以下分別為 1922-1924 年胡適的政治人際交流。



(1922 年，可以觀察到集中在蔡元培、丁在君、陶知行三人身上。)



(1923，可觀察到重要的節點為任叔永、朱經農兩人。)



(1924 年，政治相關的討論幾乎消失。)

從這三張圖我們可以很清楚的發現，1922-24 年，胡適對於政治討論的熱忱快速下滑，這除了「好人政府」的失敗之外，1923 年的「曹錕賄選」是否會是壓倒駱駝的最後一根稻草呢？

若進一步觀察會胡適好友：任叔永與高一涵，他們對於曹錕的賄選都持反對態度，[10]且對於當時國會南遷一事沒有抱持多大的信心。特別是高一涵更進一步認為議員們這樣的行為無外乎是以金錢做最後的考量，若回歸到《努力周報》對於政治以及政府的主張，則無一及格。[11]仔細觀察胡適於 1923 年末的文章書信，也可發現他此時常露出對於政治的絕望。若與 1922 年積極參與政治的態度相比，此時對政治問題的意氣消沉不難看出，這樣的意見更可能影響到了他的朋友。[12]

[10] 任叔永，〈國會南遷的第一幕〉，《努力周報》第 62 卷 （19230722）、高一涵，〈這一週：答 KC 君〉，《努力周報》第 63 卷（19230729）。
[11] 詳見涵，〈這一週：答 KC 君〉，《努力周報》第 63 卷（19230729）。
[12] 胡適，〈一年半的回顧〉，《努力周報》第 75 期（19231021）。同文亦收錄於《胡適來往書信選（上）》（北京：中華書局，1979），頁 156-157。