數位人文研究叢書 4
Series on Digital Humanities

數位人文要義：尋找類型與軌跡

項潔　編

Essential Digital Humanities:

Defining Patterns and Paths

臺大出版中心
NATIONAL TAIWAN UNIVERSITY PRESS

數位人文研究叢書 4
Series on Digital Humanities

數位人文要義：尋找類型與軌跡

項潔 編

Essential Digital Humanities:

Defining Patterns and Paths

序

從全球各國數位典藏發展的歷史來看，臺灣開始得不算太早，但也不算晚。暫且不論個別學校、單位或個人的研究和努力，從國科會在1998年以「數位博物館計畫」開始投入國家資源，有系統地發展數位典藏來算，到現在也已將近15年了。在這段不算短的時間裡，我國投入了大量經費，也將大量的文化資產數位化，其中最重要，從2002年開始執行的「數位典藏國家型計畫」，更產生了指標性的作用，讓「數位典藏」在臺灣成為一個大眾語言內的詞彙。

數位典藏國家型計畫的成功，至少有一部分歸功於許多資深且傑出的研究人員，尤其是人文學者，不計報酬地全心投入。但是從大約2005年開始，我們漸漸發覺這些優秀的學者在他們本身的研究中，似乎並沒有充分利用他們花了大量精力數位化的檔案；這令我感到困惑，因為這樣不是事倍功半嗎？為什麼不能將數位化工作和本身的研究結合在一起呢？觀察到這個現象後，我開始去探討它的環節。我發現其中的一個重要原因是數位典藏系統的設計往往沒有考慮到使用者——尤其是研究者——的需求，以致於一直到現在絕大多數的研究者還是認為檢索系統只能幫他們找資料罷了，而不能幫忙整理或分析資料。這是很可惜的，因為許多資訊技術已經十分成熟，如果能夠與數位檔案結合並有效地運用在這些系統中，應對人文研究產生非常大的助益。在更進一步的探討後，我發現這個問題並不是臺灣獨有的，在國外亦有學者思考這個問題，而且已有一個研究社群，那便是「數位人文」。

用最簡單的話來講，「數位人文」就是結合大量數位材料，運用資訊科技，來從事人文研究。顧名思義，這是一項跨領域整合的工作。要達到這個目標，除了要有大量高品質的數位資源可供使用外，更需要人文學者與資訊學者密切的互動與合作。有鑑於此，國立臺灣大學數位典藏研究發展中心從2009年開始舉辦每年一度的「數位典藏與數位人文國際研討會」，希望藉由這個會議，不但能夠介紹世界最新的研究成果，並能提供國內外人文與資訊學者交流的機會，讓臺灣的數位人文研究能夠遍地開花，並與國際接軌。

《數位人文研究叢書》即是這個年度會議產出的一項成品。叢書中的每一篇文章均在研討會中發表過，再經修改及至少兩位審查人的審查通過。在此特別感謝臺大數典中心的蔡炯民博士、陳怡君小姐與全體同仁對本叢書投注的心力。我們希望透過這個系列的叢書，提升國內學界對數位人文的認知，並激發進一步的研究。

項潔

2011年9月於臺大

Preface

Taiwan started its cultural digitization effort in earnest in 1998, when the National Science Council initiated the Digital Museum Program. Since then the government has invested a large amount of money and has digitized a significant portion of Taiwan's cultural heritage. Indeed, the effort has been so successful that 數位典藏, the Chinese translation of "digital archives", has become a household term in Taiwan.

The success of the digitization enterprise, in particular the National Digital Archives Program, is at least partly due to the altruistic devotion of many senior scholars. However, around 2005 I started to notice that many of these scholars do not seem to utilize the outcome of their digitization work in their own research. This phenomenon baffled me. Since an important purpose of digitization is to make materials more accessible, why couldn't they use the fruits of their own hard work to expedite their own research? After some studies, I observed that at least one reason is that the digital archive systems are not designed with the need of the users, especially researchers, in mind. Thus most people still think that a retrieval system can do just that, retrieve, but not to help them organize, observe, and explore what has been retrieved. This is a pity, because combining information technology with massive high quality digital objects should provide tremendous opportunities for humanities research. I then learned that this phenomenon is not unique to Taiwan. Indeed, many scholars in the world have been thinking about this challenge, and the community is called Digital Humanities.

To put it in the simplest term, digital humanities is humanities research with the help of digital materials and information technology, with an emphasis on the type of research that cannot be accomplished otherwise. It is interdisciplinary by nature. In 2009, the Research Center for Digital Humanities of the National Taiwan University started an annual International Conference on Digital Archives and Digital Humanities to serve as a forum for researchers from different disciplines and different parts of the world to showcase their work and exchange ideas. This series, the Series on Digital Humanities, is a product of this effort. All papers in the volumes have been presented in one of the conferences. The final version is then submitted and went through rigorous reviews. We hope that the Series will promote digital humanities and stimulate further research.

Jieh Hsiang
September, 2011
National Taiwan University

目　錄
Contents

Part II 語料庫語言學
Corpus Linguistics

Part III 地理資訊
Geographical Information

導論
Introduction

數位人文的變與不變
The Change and Unchange of Digital Humanities

數位人文的變與不變

項潔*、翁稷安**

摘 要

本文以史學方法中「史無定法」的觀念為例，說明傳統人文學在研究的方法上，本身即保有開放的態度；在數位時代的今日，與資訊技術合作，開展出新的數位人文研究，是符合人文學傳統的進步。

在學科專業化的今日，現階段數位和人文最好的合作方式，可能是打造成一個溝通、開放的團隊。本文認為數位人文是數位時代人文研究的一種新的方式與選項，是一種助益而不是傷害，它分享人文學不變的堅持和執著，也體現著人文學對方法的開放和創新。這變與不變的平衡和掌握，將會是投身於此領域的學者，全力以赴的挑戰。

* 國立臺灣大學資訊工程學系特聘教授，國立臺灣大學數位典藏研究發展中心主任。
** 國立臺灣大學數位典藏研究發展中心碩士後研究員。

The Change and Unchange of Digital Humanities

Jieh Hsiang *, Chi-an Weng **

Abstract

Starting from Yu Ying-shih's thesis that there is no fixed methodology for historiography, this article explains that it is within the very nature of the humanities to explore new research methodologies. In that respect, utilizing digital resources and information technology in humanities research conforms to the progressive nature of humanities scholarship.

In today's world of specialization, the most promising mode of collaboration between digital technology and the humanities might be to build a team of people from different disciplines that is both inclusive and interactive. This Introduction points out that digital humanities is a novel and significant option for humanities research, and does not represent a harmful development in the field. Such collaboration also embodies "the unchange" elements of the humanities—persistence and perseverance—as well as "the change"—openness and innovation. The decision of what should be changed and what should not be changed, then, will be a major challenge for scholars in this field.

* Distinguished Professor, Department of Computer Science and Information Engineering (CSIE), National Taiwan University. Director of Research Center for Digital Humanities, National Taiwan University.

** Research Associate, Research Center for Digital Humanities, National Taiwan University.

一、本書內容簡介

本論文集是「第三屆數位典藏與數位人文國際研討會」的論文集結，這已經是國立臺灣大學數位典藏研究發展中心連續三年舉辦這個會議，離我們希望建立一個每年一度常態會議的目標又更向前邁進了一步。之所以會有這個想法，原因在於數位人文發展的研究方式和型態裡，不斷的對話與交流是最為重要的關鍵；這樣的對話應當是寬廣而多元的，跨越不同國界、不同研究議題、不同技術領域。我們相信唯有經由集思廣義，吸引更多研究者投入，才能彰顯數位人文作為研究方法的價值和意義。在這三屆的研討會中，可以明顯感受到數位人文這個議題，不管是在研究的量或質上面，都有著長足的發展：研究的議題增加，越來越多的人文研究，嘗試應用數位的技術進行處理，建立越來越專精的研究系統或典藏資料庫；研究者的背景也更形多元，許多人文出身的研究者，開始留心甚至主導數位與人文的結合；更多的數位技術被納入數位人文的思考，無論自然語言、3D呈現等新穎的技術，開始被思考和人文議題結合的方式與可能。讀者眼前的這本書，可以說是這蓬勃景象的一個縮影。

本論文集共收錄了九篇文章，可粗略分為兩個部分，代表著現階段數位人文領域發展的兩種主要類型。一類是文字的分析，第一部分和第二部分所收錄的文章，即為此類。在第一部分對檔案的分析中，〈自然語言處理技術於中文史學文獻分析之初步應用〉企圖結合自然語言的資訊技術和傳統史學文獻分析，試圖析理出結合的可能。〈以文本分析呈現臺灣海外史料政治思想輪廓〉則用數位技術，綜觀而量化的去分析過往政治思想史所研究的場域。第二部分將焦點放在具體的議題上，應用數位技術做大規模的語意分析，〈結合漢典古籍虛詞常見字與統計量化分析進行漢譯佛典譯者風格辨別〉一文從大量的佛教經典的字句整理中，去回答譯者風格，這是佛教傳入的重要課題。〈「共現」詞頻分析及其運用——以「華人」觀念起源為例〉提出了新的理論視角，賦與詞頻分析新的詮釋，並以「華人」這個觀念做具體的實驗。〈漢語方言語音資料庫自動擴增補完方法〉則是希望能應用數位科技，去處理語言學中十分困難的方言問題。上述這些研究，所面對的都是十分龐大的文本資料，而它們所努力達成的目標，都是以人力所十分困難達成，需要耗費大量時間和心血的課題。

數位人文研究的第二類關懷則是在呈現上，如何用新技術去展現過往文字所不能負載的成果，本書第三部分即屬此類，並將焦點置於GIS技術的呈現上。空間一向是人文研究所看重，卻又最常被忽略的部分，如今透過GIS技術，學界始逐漸填補這個斷層。〈京都大比例尺地圖（京都市明細圖 *Kyoto-shi meisai-zu*）數位化〉、〈數位典藏應用的社會效益與永續經營——以阪神大地震資料3D視覺化為例〉、〈「太平洋史前Lapita陶器線上數位資料庫」的建立與運用〉，三篇文章的關注點各

不相同，無論是針對底圖、3D視覺化或對具體應用等等不同的開發，但其共同之處便是要將空間帶回人文學的關懷之中，提供新的從地理資訊出發的思考角度。此外，筆者於本書中另有一文，〈多重脈絡——數位檔案之問題與挑戰〉旨在說明數位系統對脈絡探勘所能帶來的協助，同時結合了文字處理和圖像呈現，試圖討論數位人文所能開啟的全新研究視野，是我們長期以來在這個領域中努力的一點體會與想法，就教學界。

在這篇〈導論〉中，則想換個角度去討論數位和人文之間的關係，本文將就人文學界對數位人文的接受與否為主要的焦點，在第一部分將以史學為例，說明在傳統人文學中，對新方法的開放心態；第二部分則承續第一部分的討論，試圖勾勒出數位和人文兩者在現階段的合作可能。如果〈多重脈絡——數位檔案之問題與挑戰〉一文是由專業論述的角度表明數位人文對傳統檔案研究方式所能帶來的改變，那麼本篇短文則希望能用對話、一般性的立論，寬泛討論人文學者對數位人文新方法所應持的開放心態，兩文立意雖略有出入，但仍可併觀。需特別強調，數位和人文兩者之間的對話，應當是一個沒有止境、不斷持續的過程，「數位人文研究叢書」的出版，其內每篇文章都是這樣努力下的嘗試，而非最後的定論。對我們來說，能讓這樣的對話不斷延續，成為引玉之磚，促使相關研究和見解的蓬勃發展，才是我們最終的殷盼。

二、從「史無定法」談起

讓我們先從余英時〈中國史學的現階段：反省與展望〉這篇文章談起，該文寫於1981年，作為《史學評論》這份刊物的發刊辭。文章一開始便指出現代史學雖擁有深厚的學術傳統作為後盾，但在當下已面臨衰落的處境，這是人文或社會科學在20世紀後半中國學界的共同困境，史學的研究本質上即為「綜合貫通之學」，「必須不斷而廣泛地從其他學科中吸取養份」，所以人文社會研究整體的不彰，必然對史學造成很大的影響。作者回顧過去中國史學的發展，可以歸結為兩個主要流派：一是以史料的搜集、整理、考訂、辨偽為核心任務的「史料學派」；另一則是主張從系統、理論的觀點對中國史進行全程通釋的「史觀學派」。兩者在中國近代史學的發展中，呈現對立。作者以司馬遷的「通古今之變，成一家之言」的說法，指出雙方其實各捕捉到史學的不同面向，各有利弊，並對兩者皆提出了深刻的批判，但也指出兩者本身雖有缺陷，但各自均取得了一定的成果，兩者皆顯示了近代史學追求科學化的不同途徑，在反省和調整後，仍是中國近代史學發展的立足點。

那麼如何具體的以前人成果為基點，開展出史學研究的下一階段？余英時提出了幾點觀察和建議：(1)史學作為一種通貫之學，必須不斷從其他相關學科「吸取養料」。這是一個高遠而略顯抽象的理想，從過去的訓詁或語言之學，到今日社會

或自然科學在這標準下皆可和史學相結合。但在學術日趨專業化的情況下，作者指出「綜合貫通」是有一定限制的：「史學家在研究某一時代歷史事象時不但一方面要照顧到該事象在其前後時代中貫時性發展的線索，另一方面則要顧及共時性的橫向關係。這和社會學家去尋求、擴充通則是完全不同的。」所以歷史學家只是要從史學研究的需要去找尋「涉及」的部分即可，是「要看研究範圍的『切已』情況而定」。(2)作者從方法論的角度出發，指出不論近代何種史學派別，都強調自身所持者為最新的科學方法，形成了一種流行的觀念：「史學的進步主要是靠史學方法的進步」。余英時認為所謂的「史學方法」包含兩個不同層次，一是將史學方法等同於一般科學方法，只是應用的研究場域不同，「大膽假設，小心求證」即屬此類。第二層涵義則是應用各種專門學科的具體分析技術去處理歷史問題，作者特別點明：「史學家當然也有他獨特的一套工作程序，如確定證據，建立史實，發現史實與史實之間的關係，解釋變化等等，但是這些工具卻是與其他學術的發展息息相關的。」也因此各個不同研究領域內的「新學術的興起有時開拓了史學家的視野」，令史學研究者可以對「證據」、「史實」有新的了解，很多不被當作「證據」、「史實」的對象，也因之被賦與了重新解釋的空間。作者強調這樣的立論不是在證明史學為一門無方法論的學問，而是「史學的確沒有固定的方法；在技術層面上，史學是在不斷地吸收其他各有關科學的方法以為己用的」。此即「史無定法」的主張，符合了史學綜合通貫的性格。

史學家必須不斷吸收新方法，並且是具批判、不盲從的篩選新方法的運用，誠如余英時所總結的：「史無定法，而任何新方法的使用又隱藏著無數的陷阱；這一事實充分說明在史學研究上沒有捷徑可走的，一切都要靠史學家自己去辛苦而耐心地摸索。」這是史學的困難和吸引力之所在。(3)最後，余英時提出對史學未來發展的期待：「希望史學研究可以逐漸使我們從多方面去認清中國文化的基本型態及其發展的過程；……更希望這種對過去的確切瞭解可以照明我們今天的歷史處境。」更具體的說，即是在肯定文化有型態，以及歷史連續性的前提下，通過歷史研究，去尋找並解釋歷史發展的趨勢。在文章的最後，作者以「大處著眼，小處著手」來勉勵後進，在認清了中國史學研究現階段的任務後，便要對具體、關鍵性的問題反覆分析與綜合，並指出可以借鏡社會學家默頓（Robert K. Merton）的「中距程理論」（the middle-range theory）。（上述兩段整理自余英時，1982）

這篇文章在當時便引起了很大的重視，可說是中國史學研究的定音之作，《史學評論》在數期後還不斷邀請各方學者專家對這篇文章進行回應和討論。余英時先生這份對史學的建言與期許，距今已三四十年的時間，其中的幾個重要觀點，即便在今日看來仍歷久彌新，深具振聾發聵之效。之所以如此重要，因素很多，最根本的關鍵，在於這篇文章回應了中國史學研究的危機，並從宏觀的角度，提出一個具體可行的方向。事實上，史學是個危機感很重的學門，原因之一可能是史學本質的

問題，或許，作為一個探究過往真實的學問，隨著真實的難以測知，史學的研究目標很容易被視為唐吉訶德式的努力，過往的真實就像一個永遠難解的謎團，任何企圖接近的努力似乎最後都僅是徒勞無功的自說自話。另一方面的危機，則是來自更現實層次的挑戰，即在面臨各種學科，諸如社會科學的影響時，史學如何還能保有自身學術規範的獨立和自主。當各種新的社會學方法和解釋紛紛出現時，史學這一學門的價值究竟何在？歷史研究不當只是破碎的考證，也不該只是特定理論架構的附庸，它應當有其自身的價值，而且還是對當代有所幫助的價值。後者成為〈中國史學的現階段：反省與展望〉一文主要針對的目標，並提供了明確而具體的解答，而後者答案的確立相當程度便回答了前者，畢竟要回答對史學的質疑，最實際的方法不是在玄虛的理論之間打轉，而是用實作的方式，以實績去回應所有的懷疑，否則爭論再多只不過是空談而已。是以，當我們一旦確立了史學和其他學門之間的關係，便能在人文科學龐雜、巨大的意義之網內，找到安身立命之處；換句話說，從方法論上所提出的解決方案，有其不容抹殺、跨越時代的重要意義，而這也正是〈中國史學的現階段：反省與展望〉一文最大的價值所在：強調史學「綜合貫通」的性質，並以「切己」為標準，兼容並蓄各種不同的學門。海納百川的結果，不僅回應來自不同領域的挑戰，也彰顯了歷史研究的核心關懷。

這也是我們在討論數位方法和人文研究之間關係時，特別提及這篇文章的主因；數位科技之於人文，簡單來說就是為人文研究提供了方法上的擴充，以史學這個具有悠久傳統的人文學門而言，即是在其綜合貫通的本質之內，提供研究者一個觀察過去的新方式和新選項。方法本身不可能取代學科本身，任何方法也均不是藥到病除的神丹妙藥，使用者和方法之間，不論任何領域，原本就是一個不斷磨合、糾正、適應的過程，唯有透過彼此的溝通和學習，才能達到駕輕就熟的境界。對方法不經檢討的迷信是不可取的，同樣的，對方法直覺或主觀的排斥亦非正確的態度。方法應該配合使用者的需要而不斷修正，不當只是為展現效能而展現；同時方法的開發也應跳脫使用者想像力的限制，幫助使用者看到下階段的可能。要能達成這個目標，使用者就必須用更開闊的態度去面對新方法的出現和改良，促成對話的開展。

對新科技的疑慮與不信任是人類永恆的難題，這或許也是數位人文所背負的原罪，但「君子役物，小人役於物」，至少在研究方法或工具的層次上，數位技術對人文研究所帶來的幫助是明顯的，它讓研究者可以節省更多的時間和心力，專注在課題鑽研和思考，更貼近過往的人事物。方法的變化，會帶來研究視野的變化，也會促進研究者對議題的深化，但並不會影響人文學的本質，再以〈中國史學的現階段：反省與展望〉所言為例，數位的新方法本身並不違反史學，甚至完全符合「綜合貫通」海納百川的研究傳統，而史學作為人文學門的核心關懷絕不會「役於物」

的被方法所取代，事實上兩者是相輔相成的，數位方法的便利能令研究者專注於對史學的核心關懷；史學的核心關懷則能令數位方法脫離破碎的枝節或炫技的展演，發揮其應有的作用。這構成了史學的變與不變，史學如是，人文學亦然，一個優秀的人文研究者，應當在不變之處堅持不懈，而在可變之處與時俱進。

三、數位與人文，在數位環境下的人文研究

讓我們再回頭思考「數位人文」究竟為何的根本問題。誠如前述，這是任何一個投身或關注此領域的研究者，都當不斷思索、面對的問題。我們無意提供一個「權威」的答案，因為這樣的答案無疑會限制數位人文發展的空間和想像，這是我們所不樂見，也是違反數位技術追求日新月異的本質。我們僅能從現階段的成果和嘗試，賦與一個寬泛的定義，並期待不斷的深化和演進。在本系列第一本《從保存到創造：開啟數位人文研究》的〈導論〉中，我們為數位人文提供了精簡的界定：數位人文「指的是那些唯有借助數位科技方能進行的人文研究。反過來講，數位人文的研究，即是企圖尋找在前數位時代中難以觀察的現象、無法想像的議題與無法進行的研究。這樣的定義或仍不免顯得寬泛，但正因如此，它能夠容納比較多的可能性——對於初生的領域而言，與其畫地自限，這樣的可能性應該更為重要。另一方面，這個定義提醒我們，數位人文研究應該專注於真正的突破，並帶來概念上跳躍性的發展（conceptual leap），而不只是小規模的改善。」（項潔、涂豐恩，2011）

這個定義反映了我們對「數位人文」深切的期許與盼望，一定程度上是一個「取法乎上」的高標準，一方面期待它能彌補人力所無法觸及、難以觀察到的角落，並帶來概念上的跳躍；這也是我們一路以來一直努力的目標。但這並不表示我們只能以高標來界定數位人文，會特別以此標準自期，只是對數位人文這樣的一個新興的領域來說，「畫地自限」是嚴重的致命傷，如同古語所云朝「取法乎上」的目標努力，最少能有「僅得乎中」以上的結果，甚或在無數次的嘗試之後，才有開創概念的契機。是以，我們並不是要否定傳統的人文研究方式，要能達到「唯有借助數位科技方能進行的人文研究」，顯然需要以傳統研究方式作為借鏡和學習的對象。從這個角度出發，如果「畫地自限」是數位人文在發展階段最大的致命傷，那麼第二項所該避免的則是「數位」和「人文」的對立。顧名思義，「數位人文」就是「數位」和「人文」兩種不同領域的有機融合，而不是對立，兩者之間不當存有鴻溝。一方面在業已步入數位時代的今日從事人文研究，多數研究者皆或多或少開始應用數位科技或使用相關數位資料庫，數位科技已走入了人們生活的每一個角落，人文研究自無置身於外的可能，頂多僅是程度多寡的差別。另一方面，數位科技和人文兩個不同的學門，雖然有極為不同的思考背景乃至學科訓練，這樣的差異不當被視為是隔閡，它反而是可以彼此刺激、相互影響而激盪出火花的動力。要能達到

這個目的，數位和人文兩造之間就必須進行充分的對話，如同上節所提及的，單只是技術的開發，在沒有人文素養的問題意識為導引下，往往很難碰觸到問題的核心；反過來說，不管是何種領域的人文研究者，為了要觸及自身研究的核心關懷，都應該殫心竭慮，用盡各種方法去進行理解和探勘，不當自囿於先入為主的成見。在雙方都持開放的心態下，持續的溝通，數位科技的研發者可以理解人文學者在處理議題時的思維和限制，人文的研究者可以去理解資訊學者在開發技術時的考量和盲點。一旦能開始這樣的對話，數位和人文兩造之間的有機結合，才能真正落實，不僅只是「高貴的夢想」。

換句話說，前段所描述的數位人文運作，就是一種團隊的運作模式，這很可能是當下最適合的數位人文的運作模式。當然如果人文學者能具有程式能力，或者資訊工程學者可以獻身人文研究，是最理想的狀態，因為親身投入、橫跨兩個面向者，最能掌握彼此的需要及箇中的甘苦，也不會有溝通不良的情況發生，但在學科日益分工、專精的今天，要同時能身兼兩個領域，確實是有現實的難處，資訊專業需要花時間理解，人文素養亦需長期的涵養，我們期盼在不久的將來，隨著數位人文的研究方法和成果日漸受到重視肯定，會有越來越多的學者願意投入時間和精力去學習和思考，進而出現能夠充分掌握兩端的研究者。勒華拉杜里（Le Roy Ladurie）在1970年代那句充滿危機的預言：「明日的史家，為了要生存，都必須要先學會電腦的程式語言。」四十多年過去證明是有些過慮了，也許我們可以用比較正面的角度，將這段話重新詮釋，史學或任何人文學科都有自身的學科規範（discipline），有其不容被輕易取代的獨立性，數位技術帶來的是助益而非挑戰。要能充分的掌握這項技術，擴充自己的學門的研究可能，研究者可能要對資訊技術有初步的認識，如果要期盼有更進一步的需求，最快速而簡單的方式，便是和資訊背景的學者合作，以團隊的方式進行數位人文的探查。讓我們再借用勒華拉杜里的譬喻略加修改，數位技術的加入，其目的不是為了讓人文學能「生存」，而是要讓人文學可以有更好的「生活」。

這並不是我們獨創的觀察，知名的史學研究者黃寬重教授，也從歷史學研究的角度出發，提出類似的見解。在民國100年10月中，在紀念漢學研究中心成立三十週年所舉辦「臺灣漢學新世紀——漢學研究中心三十周年學術論壇」裡，黃教授以〈數位時代人文研究的衝擊與蛻變〉為題發表演講，該次演講後來以文章的型式，刊登於《漢學研究通訊》中（黃寬重，2012），國家圖書館數位影音服務系統亦保有影音的錄像。[1]在該次演講中，黃教授以親身參與的經驗，回顧臺灣史學研究從1984年以來，近三十年的時間進行相關數位化的成果，並提出數位資源對人文研

1 網址見：http://dava.ncl.edu.tw/MetadataInfo.aspx?funtype=0&id=588499&PlayType=1&BLID=5188089（最後確認時間：2012年7月6日）

究所帶來的助益和衝擊。在助益的方面，他將傳統文本數位化的優勢，歸納成下列四點：「首先，典藏原件能夠在不影響公眾利用下，獲得更好的保存，甚至永續庋藏。其次，經數位化後，文物典藏圖像資訊可透過便捷的網際網絡，不限次數送至任何地點，有效突破地域限制。再者，數位資料可透過建置後設資料與聯合目錄，將各類資料彙整分類，便於利用。其四，也是最重要的一點，數位資料能極快速搜尋檢索。」這四項功能，「最終有助資源共享、消弭區域差距，達成不分地域、種族均能公平地利用人類公共文明資產的理想」，是自印刷術發明之後，人類對掌握知識方式的最大革新，「不僅造成資料蒐集、整理、分析等有形研究過程的顯著今昔差異，龐大而易於利用的數位資源更有助研究者整合不同領域，進行跨時段、跨地域的長期觀察」，總之，黃教授認為數位技術對人文研究的涉入，重新定義了傳統所謂「博學」的概念，刺激了各種創新性的研究和說法，臺灣史學界黃一農院士所提出的「E考據」和哈佛大學包弼德教授主持的「中國人物傳記資料索引」相關計畫，都是最好的例子。

然而，水可載舟亦可覆舟，這些優點卻也替人文學者帶來新的衝擊，黃教授指出了下列幾個影響的面向：首先是數位資料庫的品質良窳參差，其原因一是出於商業利益帶來的粗製濫造，加上在建置資料過程中專業知識的不足。其次，資料取得的便利，將增加使用者對資料解讀的門檻，面對能快速獲取的龐大資料，在「缺乏該領域專業訓練與研究資源先備知識」的情況下，反而對研究是種傷害。其三，以數位的方式處理人文知識，會造成新的「知識利用門檻」，即人文學者數位技術能力的缺乏，難以發揮相對應的優勢。作者對資訊技術對文史學者造成門檻的原因，有十分精闢的分析：「資歷較長的學者其學術養成過程並無數位學習環境，與新科技間的疏離自不在話下。年輕學者與資訊科技的隔閡，部分源於國內學科分流過早，人文學學生缺乏自然科學訓練，對於科技陌生疏離，甚至懼於接受。部分也與年長人文教師習於傳統訓練方式，較少利用數位化資源豐富教學有關，年輕學子面對新的學術資源，多僅能自行摸索，不易建立有系統性而專精的應用能力。」這種出於研究習慣和科技恐懼的限制，在作者看來，反而讓人文學者在數位時代進行人文研究時，不如自然或社會科學學者，因為他們本來便長於科技和理論，一旦對人文議題感興趣，轉換跑道的難度會減低很多。也因此，黃教授警告：「因大量數位資料資源的出現，或將有助人文學界開展議題、深化研究質量，但此願景尚須面臨其他優勢學科的競爭；倘使人文研究者無法認知此一情況，未來當自然或社會領域學者逐漸將研究觸角及於人文領域，人文學者的學科優勢則將明顯消退。」

人文學者該如何面對這樣的挑戰呢？黃教授提出了許多建言，包括在學者養成階段改變教學訓練的方式、文史學者積極投身數位工程等等，最關鍵的，還是首重對自身研究領域的專業訓練之強化，「（史學研究者）若企望保有學術競爭優勢，則

更應加強史料來源的掌握與研讀能力，改變過度倚賴詞彙檢索的偏頗風氣。尋找自身關懷且具有時代或環境意義的問題，掌握與之結合的關鍵資料群……將資料還原至原有時空脈絡中，充分瞭解史料內涵與當時社會關切的事物，以顯現具有時代意義的論題。」易言之，人文學者必須要確立人文研究的學科主體性，才不會主客易勢，迷失於各式數位資料庫當中。

不同於紙本，在演講的過程中，黃教授特別突顯資訊學者與人文學者合作的模式，他特別以謝清俊和毛漢光兩位教授的合作，說明中研院的「漢籍全文資料庫」的成功，很大的原因便是在其前身計畫「史籍自動化——〈食貨志〉輸入電腦」執行過程中，謝、毛兩位教授彼此的尊重和互動，讓數位和人文的結合成為可能，之後不管計畫如何擴充，這樣的合作模式提供了堅實、穩定的基礎。事實上，無論在演講或紙本中，在研究習慣的改變上，黃教授都直接點出人文研究者應該跳出過往單打獨鬥的研究習慣，以跨領域、跨學科、跨時段等團隊合作的方式，去因應數位時代人文學研究的挑戰：

> 一旦研究習慣改變，數位資源實有助於研究者推動跨領域的團隊研究。目前人文研究仍以個人發揮創見式的獨立研究為主，這與人文學科強調個人對資料的深入理解與意見闡發的傳統有關。然而，往昔學者可憑藉的史料典籍較少，以個人之力所能涉獵的資料廣度亦有所不足，因此所論多專，難於遍蒐長時期跨越斷代的資料，以議題為中心進行深入的長期觀察，尤難與其他領域學者進行跨領域合作。如今，大型資料庫內容無所不包，所涉領域甚廣，若能合理規劃議題，透過史料研讀，共同學習，當能跨越斷代或學術領域，推動中、大型或跨領域研究。

黃教授的論點和分析，兼及數位和人文兩方面的實務經驗，是十分具有洞見的觀察，對多數投身數位人文嘗試的研究者而言，一定都會贊同黃教授的想法。這種團隊的運作，絕非只是取巧，一個好的數位人文團隊，要能充分、有效的對話，雙方都必須有一定準備，不能僅專注在自己的領域裡，人文學者必須對資訊技術有基礎的理解，資訊學者則對人文議題有基本的認識，更重要的雙方都必須不斷增進自己的專業素養，並保持對另個領域的開放心態。在這樣的一個團隊裡，沒有主從或發號施令的問題，而是一起摸索、一同深入，一旦這樣的運作成立，其實就形成了一個數位人文運作模式的縮影，不斷擴充和發展，數位人文才能有真正發光發熱的一天。

讓我們再回到余英時教授〈中國史學的現階段：反省與展望〉一文中，所提及的「切己」的觀念，團隊合作的模式，其實正呼應了類似的訴求。在學科分流日趨細緻的今日，尤其數位和人文都是十分巨大而困難的領域，要僅憑一人之力將兩者

加以窮盡，無疑是有些困難的。故研究者所能做的，只能先確立自身研究的範疇，以及和此範疇相關的技術。數位人文要能成立亦是如此，兩個領域的學者都必須先對各自領域有足夠的專業知識，然後以開放和學習的心態去學習、理解雙方的交集之處，這「交集之處」即是「切己之處」，也是「史無定法」的具體運用和展現。總而言之，數位人文作為人文學的新風貌，對人文議題的關懷是永遠不變，而如何切入、理解、分析、推論的方法，則是永遠與時俱進、推陳出新不斷在變化之中。人文的價值是不容動搖的堅持，方法的開發是不會止息的演進，這種「變」與「不變」共生並存的狀態，本就是人文學的特質，而在數位人文裡將被更加彰顯。

四、結語

筆者曾經服務過的臺灣大學圖書館，將歷年所有的碩博士論文統一收藏在地下一樓「博碩士論文暨指參專室」裡，這些研究生們的心血，共同集結於一室，可說蔚為大觀。歷史學研究所的碩博士論文置於該房間的最角落，佔了兩面書架，從中隨手便能抽出許多寫成於民國50、60年代的論文，它們或是手寫、或是油印，字紙皆已泛黃，可以明顯看出它們和一同陳列的、寫於近十幾年論文的差別；今日每一篇論文都是用Word或類似軟體打成，清一色的新細明體。不僅於此，如果細究內容，我們相信已有許多研究者開始應用Excel、Access、Endnote、Zotero、ArcGIS、Xmind等軟體對論文資料進行處理，我們早已無法回去那個單靠紙筆完成論文的「手工業時代」了。儘管那時代其實並不遙遠，今日1970年代出身的研究者，多少都還曾經歷過那用稿紙「爬格子」來進行論述、繳交報告的時光。

數位科技的發展，在不知不覺中，已然改變了史學研究——或人文學的研究習慣。這當然是形式上的變化，和數位人文所主張的藉由科技發掘人力所難見的線索仍有所出入。然而，一來那麼龐大數量的形式變化，那麼多數位工具的使用，很難令人相信完全不會有任何一絲「質」的變化。二來至少這現象表明了在這短短的數十年間，資訊技術的發展和普及已成為不可逆轉的趨勢，人文學者與其刻意視而不見，自摒於浪潮之外，倒不如選擇欣然接受，於潮流之中去蕪存菁，為人文研究帶來更多的助益。相當程度上，數位人文的種種方法和訴求，只是上述工具的再進一步，而且還是以人文為起點的前進，對研究方法而言，是有益而無害的。

是以，或許關鍵不在於對數位人文方法的接受與否，識者所該深思的真正課題反而是什麼是人文學最核心、最關鍵的變與不變，於不變處鉤深索隱，精益求精；於可變處廣納百川，與時俱進。總而言之，不管我們將數位人文視為一種新穎的工具，或是一個充滿潛力的新範疇，它都將承繼、分享著人文學的價值和觀念，就如同那並陳於同一書架，年代跨越五六十年的論文一樣，它們或許形式上有明顯的不

同，但對人文世界的關懷卻是一致的。「蛇化為龍，不改其鱗」，數位人文是數位時代人文研究的一種新的方式與選項，是一種助益而不是傷害，它分享人文學不變的堅持和執著，也體現著人文學對方法的開放和創新。這變與不變的掌握，將會是投身於此領域的學者所全力以赴的最大挑戰。

參考文獻

余英時，1982，〈中國史學的現階段：反省與展望〉，《史學與傳統》，頁1-29，臺北：時報文化。

項潔、涂豐恩，2011，〈導論——什麼是數位人文〉，收於項潔編，《從保存到創造：開啟數位人文研究》，頁9-28，臺北：國立臺灣大學出版中心。

黃寬重，2012，〈數位時代人文研究的衝擊與蛻變〉，《漢學研究通訊》，31：1（總121期），頁1-6。

Part I
檔案史料
Archives & Documents

多重脈絡——
數位檔案之問題與挑戰

項潔*、翁稷安**

摘 要

數位化的檔案保存已經成為時代的趨勢，卻也面對質疑的聲音。其中最大的不安，在於原本紙本檔案保存的脈絡，是否會在數位化過程中遭到破壞。為了回應這樣的質問，本文回顧了傳統檔案學的保存原則，即依機構、依時間、依事件的整理，其目的便在於保存所謂檔案的既有脈絡，並指出這些原則本身所隱藏著的矛盾，如依事件的整理本身就會打破檔案依機構和時間所產生的原生脈絡，編纂者的後見主觀和選擇必然會影響檔案。更何況從研究的角度，人文研究本身便是要從史料既有脈絡中，重新找出新的解釋脈絡的過程。檔案數位化保存後，並不一定會破壞原本脈絡，恰恰相反，數位化後的檔案反而較紙本更具調整和變動的可能，產生紙本檔案所無法想像的多元脈絡。要實現這樣的可能，必須建立起一個功能強大的系統，此為檔案數位化的關鍵，數位檔案無法離開系統而單獨存在，系統的好壞將決定檔案的可用度。檔案檢索系統應盡量提供文件的各種脈絡及觀察脈絡的環境。

本文不只是被動的回應質疑，也試圖提出對數位化檔案保存的積極看法，即多重脈絡的挖掘。系統的建構不應再只是檔案數位化過程中的附屬物，而是依據檔案特性所設計的重要環節。一個好的系統所能提供的各種脈絡可能，分別為：（一）原始脈絡的保存；（二）重組原始脈絡產生的多重脈絡；（三）縱觀的脈絡；（四）鳥瞰型的脈絡；（五）文件間統計型的脈絡；（六）文件間隱藏的新脈絡。總結上述，經由檔案的數位化，檔案單一脈絡的限制被打破，檔案整理與檔案研究之間的矛盾隨之減輕；而透過檔案檢索系統的設計，檔案的豐富且多重的脈絡可被發現、被呈現、被觀察。因此，檔案檢索系統是未來在思考檔案數位化時不可或缺的關鍵方法。

* 國立臺灣大學資訊工程學系特聘教授，國立臺灣大學數位典藏研究發展中心主任。

** 國立臺灣大學數位典藏研究發展中心碩士後研究員。

Multiple-contextualization: Problems and Challenges on Digital Archives

Jieh Hsiang*, Chi-an Weng**

Abstract

Although preserving archives through digitization has become the trend of the era, it has also faced criticisms. The primary concern is that the original context of archive might be compromised during the digitization process. In response, this paper reviews the traditional principles of archiving: by institution, by timeline, and by event. Regardless of which one is adopted, its purpose is to preserve the original context of the historical documents. We point out that these principles might in fact be contradictory. For instance, re-organizing an archive according to events is bound to compromise the timeline structure or organization according to agencies. An archivist's personal view and preferences also inadvertently disrupts the original context. Furthermore, humanities research, historical studies in particular, largely involves the discovery of previously un-observed context from among primary documents. This paper argues that digitization will not only preserve the original context of the archive (if there is any), but also yield an enormous possibility of re-contextualization. To realize this possibility, however, it is imperative to design a powerful search system. This is the key to digital archives, since a digital archive can only be used and utilized through a system, not the naked eye. In addition to search and retrieve, such a system should also provide a environment that allows the forming of multiple contexts and the facility for their observation.

Therefore, the software system for a digital archive should not be an after-thought of the digitization procedure. Rather, it should be a key component designed to fully utilize the nature and characteristics of the documents of the archive. The kinds of contextualization that the system should provide include: (a) the original context; (b) re-ordering of the original context according to other metadata attributes; (c) horizontal context; (d) holistic context; (e) statistical context among the documents; (f) hidden (semantic-driven) context. In summary,

* Distinguished Professor, Department of Computer Science and Information Engineering (CSIE), National Taiwan University. Director of Research Center for Digital Humanities, National Taiwan University.

** Research Associate, Research Center for Digital Humanities, National Taiwan University.

through digitization, a digital archive breaks away from the limitation of the single context of the original (paper) archive and reconciles the tension between the archival requirement and the contextualization need of the researchers. Through a carefully designed system, the rich and multiple contexts implicitly existed in an archive can be discovered, displayed, observed, and explored.

一、前言：在傳統與創新之間

隨著資訊科技的進步與發展，數位化已成為影響各領域的時代趨勢，人文研究亦無法置身其外，數位與人文兩個研究領域的結合隱隱然成為蓄勢待發的學術範疇。數位人文所涉及的範圍十分廣闊，非三言兩語所能盡述。[1]在人文研究和數位技術結合的過程中，檔案數位化的典藏保存可視為最基本的第一步；以此為基礎，數位和人文兩者才能擁有共同對話的基準點。事實上，單從檔案保存的角度，將檔案進行數位化整理，其必要已十分明顯，此外，透過網際網路的連接，也促進了檔案取得和使用上的便利。也因此，近年來，無論是國家政府機關或各學術研究單位，檔案資料的數位化工程都成為備受重視的首要之務。

然而，在這股熱潮中，也開始出現質疑的聲音，其中以數位化的形式保存檔案，在收藏上雖有著種種優勢，但是否會割裂切斷檔案既有的脈絡？換句話說，從研究的角度，檔案的數位化在帶來方便的同時，是否反而對史料的原始脈絡造成更深層的傷害？質疑者認為數位後的檔案資料，在性質上必定和紙本保存方式不同，重新數位化便等於徹底的重新編輯，當書本裡一頁頁的資料變成一個個的圖檔或文字檔，原本貫穿書本的編輯邏輯勢必亦將瓦解，失去了原本紙本編排的邏輯後，史料原本的意義脈絡也將面臨了崩壞的危機。

這些質問是檔案數位化所無法規避的問題，它們反映了兩個面向：一是反映了檔案學長期發展後所形成的研究傳統，一種對保障檔案真實原貌的堅持和使命；任何有可能對檔案原始狀態造成傷害的方法，都應該謹慎。在這樣的思路中，數位化很可能是包裹著糖衣的毒藥，保存或搜索的便利很容易讓研究者走入見樹不見林的陷阱裡。另一方面，這些懷疑也反映對數位化最根本的不安全感，數位化不同於紙本，並不是使用者能親手觸摸的實際存在，一旦紙本編成的史料被沒有實物的資訊技術取代，過去的閱讀習慣必將有所改變，這會不會只是造成數位化的「斷爛朝報」？尤有甚者，以紙本方式存在的斷爛朝報人們還有手動整理的可能，而數位化似乎只能依靠資訊技術人員一途了。

在本文接下來的討論中，將試圖證明這樣的恐懼是出於對數位化技術的表面認知，以及面對紙本閱讀習慣改變時的多慮。檔案的數位化並不會違背檔案學長久以來形成的學術規範，正好相反，通過數位化的重新整理，反而可以糾正紙本編輯所造成的「編輯者的盲點」，還原史料原本的脈絡，貼近使用者的研究習慣和需要。此外，檔案數位化不但不會造成斷爛朝報的現象，透過資訊技術和工具的研發，反而可以透過建立數位資料的觀察系統，讓研究者可以更自在的去觀察史料，更自由

1 相關討論可參見項潔、涂豐恩，2011，〈什麼是數位人文〉，《從保存到創造：數位人文研究的發端》，頁9-28。

的進行分析與論述；換句話說，數位化並建立一套可供使用者有效觀察的系統，可使檔案的潛在價值得到完全的解放。數位技術和人文研究應是相輔相成而非相互抵觸的概念，這也才是檔案數位化工作真正的意義。

創新的技術必然會面臨傳統的質疑，但並不表示創新和傳統兩者之間為不可跨越的對立，事實上，所有的創新都一定要去回應傳統，試著和其對話，不斷尋找共識的建立，如此才能讓創新的技術有更加深化的可能，而這將是本文的主要目的。本文將分成兩部分，首先對傳統的檔案文獻保存方式進行概括式的描述，指出其保存的邏輯和重點所在；然後進一步討論傳統檔案保存和檔案研究之間的關係。第二部分，則由數位時代的保存方式出發，指出它和過往的不同之處，以及這樣的特殊性對檔案研究所能帶來的新衝擊與貢獻，尤其在打造一個適合於研究的系統後，將會替數位時代的檔案保存勾勒出新脈絡的可能。

二、傳統檔案保存的脈絡和研究方式

（一）傳統檔案保存的方式及其原則

讓我們試著藉由下面兩位學者的討論為例，來呈現華人世界對檔案保存的看法。首先是大陸前輩學者秦國經，他於1980年代起任職中國第一歷史檔案館，長期鑽研明清檔案學，撰寫過許多相關文章，並於2004年出版了厚達近900頁的鉅著《明清檔案學》，其內容仔細介紹了明清所留下的相關檔案，從檔案的來源、分類、收藏、管理和史料價值等等都有著詳盡的介紹。在該書關於檔案整理的篇章中，他指出明清檔案應按照下面步驟進行整理：首先是區分全宗，即按行政機構歸屬進行約略的區隔。其次則是就全宗內的檔案加以分類，可以依照文件名稱（例如將內閣檔案分詔書、題本、奏啟等類）、機構（如將宗人府檔案分為經歷司、左司、右司、黃檔房、銀庫、玉牒館）、所涉及問題（如將外務部檔案按所涉國家區分）和年代（如會議政務處檔案從光緒27年依時序分至宣統3年）等四種方式加以細分。最後，則是加以組卷和編目的工作。行文中作者不斷強調「文件之間具有緊密的聯繫，不能任意加以分割」，所以要努力保持機關檔案的聯繫和完整；即如按類劃分細目時，亦特別強調「類目的設置應以保持文件的歷史聯繫和便於利用為原則」，其理想是希望「類目要如實反映出檔案的內容」。[2]凡此，皆可見得其對「如實保存」，尤其是按歷史原貌的「如實保存」的堅持。

這樣的堅持也成為中國第一檔案館等相關機構，在處理檔案文件的基本流程。事實上，秦國經所服務的中國一檔館作為一個積極的保存者，除了依文件出處和時

2 參見秦國經，2005，《明清檔案學》，頁761-769。

間的整理外，更一直在進行依事件分類編輯、出版檔案的工作，諸如《明清宮藏中西商貿檔案》、《明清宮藏臺灣檔案彙編》、《中琉歷史關係檔案》等，都是近年該單位出版的重要史料集成。

另一位學者則是政治大學圖書資訊與檔案學研究所的薛理桂，薛教授從事相關教研工作多年，對檔案管理的發展，以及國內各機關的檔案收藏了解極為深入；並出版了《檔案學理論》、《檔案學導論》等書，可說是臺灣檔案學領域中，影響最大的參考書籍。他指出，檔案管理人員較喜歡用「描述」（description）與「編排」（arrangement），取代圖書館界慣用的「分類」（classification），此乃因為檔案的數量和內容，都無法和一般書籍等量／等質齊觀，通常以「文件系列」（series）為建檔的基本單位，即指在全宗之下的構成單位，由同一業務所產生的一組文件。這組文件具有類似的形式、排列順序或主題性質。文件系列可依其性質再細分為若干個「副系列」（subseries）。換句話說，是根據全宗原則的整理。他將對檔案的處理歸納為三點原則：一是尊重檔案的來源原則，按檔案產生的出處，維持其完整，其預設為「檔案產生係隨著機構與個人的業務、生長與活動，呈有機的成長」，不當輕易加以割裂，這樣的處置將會使史學研究者更加了解檔案的來源與目的。第二個原則是尊重檔案的全宗原則，即是來自於同一個行政單位、公司或家庭視為一個全宗，應當組合在一起。第三個原則是尊重檔案的原始順序，此原則強調經由檔案的原始順序，可反應原有產生與使用情況，意即當檔案為現行文書時，有其出於業務需要而生成的特質，如果強迫使用人為的方式，會對檔案的歷史證據價值帶來傷害，所以原始順序對檔案的處理至關重要，不當任意予以改變。[3]

薛理桂的著作充分闡釋了19世紀至20世紀以來檔案學發展的風貌，更可說替檔案學作為一專業學門劃下了明確的學科界線；而他綜合西方各家研究所歸納出的三原則，具體而微地反映了檔案學對文獻保存的理念與態度。秦國經則從長期實作中，順應著傳統既存的檔案結構，歸納出檔案保存應有的準則，依名稱、機構、年代和所涉問題的分類，概括檔案整理多數的需要。兩者可謂殊途同歸。綜合兩位學者的論述，我們可以將檔案學的檔案脈絡大致整理成下列三點：一是以檔案出處為中心的保存脈絡，這可說是一種最安全而妥當的保存方式，作為一份文書資料在仍是現行文書時，本身便有其產生的需要與邏輯，顯示了該機構運作的各個面向，因此檔案的生成機構可說構成了該檔案的天然疆界，以此為範疇，完好保留其運作的規律，提供了日後研究者在觀察和切入時的重要參考座標。第二種整理脈絡則是以時間的順序作為檔案典藏的脈絡，這是最直覺的區隔方式，檔案作為一種史料，蘊藏著進入或理解歷史的鎖鑰，而人類理解歷史最直接的方式便是順著時序去解讀，這也是多數歷史研究者所習慣的方式。以時間先後保存，理論上將能體現事件發生

3 薛理桂，1998，《檔案學導論》（初版），頁181-184。

的先後順序，可說是檔案資料的天然肌理。第三種脈絡則是前述兩者的進階，即依據事件的相關度進行整理，以某個議題或歷史事件為核心，將相關檔案史料打破各種原先的限制，重新加以編排。其優勢在於用事件的角度切入，能將散見於各處的資料集中化，賦與零碎的檔案一個初步的整理，給予一個粗略的「故事軸」，以利日後研究者在觀察時的需要。

總而言之，上述這三種方式符合了研究者最直觀的考量，亦反應著構成歷史的基本要素，機關脈絡代表著和檔案相關人、物的保存，時間脈絡則代表著變化，事件脈絡則代表著將檔案串連的故事，這些都是當我們在談論過去，或從事人文研究時所不能或缺的因子。我們或許可以說，檔案學的三種檔案脈絡，其實便是為日後人文研究進行基本清理、編排的預先準備。

（二）傳統檔案脈絡和研究之間的關係

上述的討論其實已提示了檔案學的保存邏輯，是和歷史研究的發展相契合的，並出於進一步歷史研究的需要所建立起的保存形式。事實上檔案學於19世紀末開始萌芽發展之際，正是和近代史學的生成——蘭克典範（the Rankean paradigm）相輔相成的。19世紀德國史家利奧波德．馮．蘭克（Leopold von Ranke,1795-1886），反對當時流行的浪漫主義（Romanticism）或歷史主義（Historicism），強調歷史學應該是可以客觀、科學的學門，不受制於當代政治與任何哲學系統，亦不為預知未來，而是要探求真實發生的過去。這樣的信念帶來了深遠的影響，成為近代歷史學的主要骨幹。要建立科學的史學，在蘭克看來最重要的關鍵便是對原始史料的直接引用與批判，尤其是對一手檔案的使用，這成為後來蘭克典範的核心之一。[4]事實上，將文獻學的方法引用至近代歷史學研究之中，可被視為蘭克最大的貢獻。鑑定史料的真偽，仰賴「最純正、最直接的文獻」，全面的對相關史料進行檢索，直探檔案，成為近代史學的基本程序；即便在今日，這項重要的突破，仍可被視為歷史學研究和教學的根基。[5]

曾赴德國求學的傅斯年，也受到蘭克史學思想的洗禮，在其於1928年發表的〈歷史語言研究所工作之旨趣〉一文中便直接提出了「近代的歷史學只是史料學」的說法；史家的工作在於使用自然科學的方法，將一切所能觸及的史料進行整理。這種以科學治史的態度，和盡量保存檔案完整脈絡，使其不受後人編排所污染的態度，可說遙相呼應。當然這樣的價值觀，似乎隱隱暗示著在作為史料的檔案中有著

4 此段關於蘭克本人史學著述和思想的大致內容，整理自汪榮祖，〈回顧近代史學之父蘭克的史學〉，收於氏著，2002，《史學九章》，臺北：麥田，頁49-70。這並非說蘭克史學只是一味的主張實證、客觀而已，關於這部分在下文將會提及。這中間涉及許多複雜討論，已為史學史或史學方法領域的重要課題，非本文所能詳述。

5 Richard J. Evans著，潘派泰譯，2002，《為史學辯護》，頁22-28。

「不證自明」的過去藏於其中，史家的工作只是讓檔案史料自己說話。上述蘭克或傅斯年的觀點，有其發展的時代背景，在成為史學研究典範的同時，也必然面對了許多質疑的聲浪，使得近代史學對其典範產生了深刻的反省，對蘭克的論述本身也有了重新認識的機會。事實上，長於檔案使用的蘭克從未認為「史學即史料學」，而是主張「歷史要寫得像過去已發生的事一樣真實」，前者是後者層層錯譯之後的結果。他所謂的客觀、科學的史學，絕不只是檔案的堆砌排比，史料仍需要靠史家主觀的選擇，在不違背對個別史實的客觀研究中，從整體的潮流加以掌握。[6]

總而言之，無論蘭克或傅斯年的主張都有其時代的針對性，不能僅從字面以僵化的教條方式看待，但無論如何，他們對史料求真的尊重是任何歷史研究者都該遵守的基本守則。這樣的原則一樣適用於所謂「後現代情境」下的史學研究，Richard J. Evans在他著名的《為史學辯護》一書中，便指出後現代史學也許可以從理論面對史學傳統帶來一些思考的刺激，但終究只是理論，史學的方法是不會改變的，也就是「由蘭克所設下，並在他之後被人們以多種方式發揚光大的那些鑑定史料的法則」。無論是哪一派別的史學研究者，他們大部分的心力都投注在「確定那些個別事實，重建那些個別事實，根據史料去研究，越穩固越好」這是連後現代史學最熱情的擁戴者都必須遵守的金科玉律。[7]

歷史研究本身便是在一個個不同的檔案和史料之間，找尋新脈絡，藉此進行結構性分析或訴說被時光淹滅的故事。檔案本身並不會說話，而是如同一個個散落的接點，需要靠史家的目光和關懷加以串連編排，並揭示其意義。檔案自身的脈絡是史家的參考座標，是研究的起點而非結束；史學研究所從事的，便是從檔案的既有脈絡中，在不違反歷史事實的原則下，以新的視角建立新的脈絡，挖掘出過去所尚未被世人所知悉的面向。

近代史學對史料的重視，和檔案學對檔案完整的重視是無法分開的；理論上，史家的工作便是在檔案學家所清理好的戰場上努力奮戰。但這並不表示尊重檔案真實的原則，就該被既有結構所僵化和限制，所謂的研究本身即是在尋找、建造出新脈絡的過程，一個好的研究者應該尊重檔案的既有架構，又不當受其限制，如是才能開創出具有新意的研究。此外，紙本的檔案存放原則本身並非沒有矛盾，我們將在下節中進行討論，並進一步指出，數位化正是化解這種矛盾的解答，且數位化的檔案處理將會更貼近研究者的思考方式和需求，不只是清空戰場，更宛如提供了無堅不摧的神兵利器。

---

6 參見汪榮祖，2002，〈回顧近代史學之父蘭克的史學〉，《史學九章》，頁62。

7 參見Richard J. Evans著，潘派泰譯，2002，《為史學辯護》，頁144-146。

三、傳統檔案保存的盲點與問題

（一）何謂檔案

在上節中，我們大致將傳統檔案的保存歸納成依形成單位、依時間、依事件的三個脈絡，並指出在尊重傳統檔案真實的原則下，以研究者的問題意識，重新勾勒出檔案彼此間新的脈絡是所有歷史研究必經程序。接下來我們必須進一步回答，以數位的方式保存檔案，在這樣的既有關係中，能夠帶來什麼新的可能？我們將先指出傳統檔案保存的不足之處，以此為基礎突顯出檔案數位化的優勢，並進一步說明以數位化的方式整理、留存檔案所具有的新潛力。

在開始進行討論之前，我們或許須先思考檔案究竟是什麼？薛理桂在1998年版的《檔案學導論》裡，在綜合各家對檔案的定義後給予「何謂檔案？」的問題一個狹義的答案，即：

> 政府單位在處理公務的過程中所產生的文書，在歷經一段時間後，這些文書已不再使用，經鑑定其具有長期保存價值後，予以保存，以供使用。

他指出在狹義的定義中，包含著產生單位（政府）、時效（非現行文書）、價值（具長期保存的價值）等三個要素；廣義的定義，則是不限政府單位，私人機構、家庭或個人所產生的文書，都可以視為檔案之一種。[8]他更詳細歸納了檔案的文種特性，分別是：唯一性、多元性、公平性、可靠性、自然性和關聯性。[9]我們可以再將這六種特質，粗分成三方面：一是對檔案真實性的信任，特別它是機關在運作過程中自然生成的產物，除少數作假的特例情形，檔案都應該是最真實、可靠的稀有史料。其次則是檔案種類的多元性，尤其在今日，從紙本式的手稿、公文書，到錄音帶、錄影帶、照片、地圖、工程圖、微縮片、電腦磁片、光碟等，檔案的種類可說無所不包。最後則是檔案來源的多源性，因為政府機關內的業務常會有跨部門的情況，有時除了和同機構內不同的單位關聯，甚至還可能會與機構外其他單位結合，形成一份檔案文書多種不同來源的情形。

這樣的定義其實是取其最安全的方式，即以機關為主軸的方式界定檔案，再按時間的方式進行排列，此種作法可以說是最沒有爭議的作法，但仍有其問題存在，首先就如同上段所說的，等於是種「無作為的作為」，一旦數量龐大超過一個限度，人們將無法有信心能夠詳閱其中全部的資料，造成研究者將大部分時間耗費在史料搜集、考證，使得所當關注的研究焦點無法顯現。這也就是前面在討論秦國

8 參見薛理桂，《檔案學導論》，頁3-4。
9 參見薛理桂，《檔案學導論》，頁6-7。

經的檔案脈絡時，作為積極的檔案保存者，會試圖用「事件」去對檔案進行區分的原因。檔案必須出版才能為研究者廣泛的使用，然而一旦經歷出版或編輯的程序，特別再加入「事件」為參考座標後，便很可能（甚或必然）改變了所謂檔案的原有脈絡，無論來源、全宗或原始順序都面臨重新的編排與調整，這便形成了理論和實作間的矛盾。其次則是檔案多元化和多源化的衝擊，各種類型和各種來源的檔案，互相雜混，特別在以「事件」為核心的檔案整理過程中，這樣的問題更加激化。凡此，皆形成了傳統紙本檔案保存方式的盲點所在。

（二）紙本的檔案保存方式的盲點

人們常常不自覺地預設紙本保存方式、尤其是以印刷出版方式的客觀性。這有其歷史脈絡，印刷術的出現對人類的歷史帶來巨大的影響，而對書本的尊重更是自古已然，無論中外。這種對書本的尊重和習慣，會使我們忽略了將檔案以書本的方式編輯出版，本身便是一個重新整理的過程，它對檔案既有脈絡所帶來的殺傷力，並不會小於數位化過程，甚至是更為巨大的。紙本對檔案的整理、出版，不出依機構、時間和事件三種方式，並以依事件分類者為大宗，因為依單位和時間整理，雖保存檔案的原貌，卻無法帶來研究上的便利；一頁頁的翻閱，或能使研究者心安，卻不見得能轉換成研究效率和質量的提升，也因此依事件整理成為最受檔案整理者重視的分類型態。

依事件積極對檔案進行整理的方式，並非今日才出現的思維，在檔案形成之初便有類似的認知。以清代的檔案管理為例，清代所謂的「專案檔」，是有深遠的歷史傳統的。專案檔指的是清代軍機處為了便於查考舊案，依例會將定時經辦的事務按照內容另行抄錄，加以儲存。精研故宮檔案多年的莊吉發，將軍機處的專案檔定義為「是以事為綱，逐日抄繕成冊，其每一種檔冊，僅關一類之事，並不雜載」。他將目前國立故宮博物院現藏的清代專案檔約略分為三類：「一類是清廷用兵隣國整理邊界的檔冊，如緬檔、安南檔、廓爾喀檔等；一類是剿辦秘密宗教及髮捻等內地民變的檔冊，如東案檔、東案口供檔、林案供詞檔、剿捕檔等；一類是平定少數民族之亂的檔冊，如金川檔、苗匪檔、剿滅逆番檔、剿捕逆回檔等。」[10]換句話說，即便在檔案最原初的生成階段，將檔案從機關等原生脈絡中抽出，改以事件為主軸的脈絡，已是檔案彙整的常態現象。這反映了檔案必然的現實，如果不對原生狀態進行加工、「後製」，要對其進行縱向（檔案自身的前後順序性）或橫向（跨機關單位的相關性）串連，都是件十分困難的事，反而無法呈現檔案內所記錄之事件的全貌。

這樣的整理傳統至今日已演變成一種專門的學問，即所謂「檔案文獻編纂學」，徐紹敏於2001年所再版的《檔案文獻編纂學》一書中詳盡列出了檔案編作工作的六

10 莊吉發，1983，《故宮檔案述要》，頁237。

個步驟，分別是：（1）確定編纂題目；（2）編寫選材大綱；（3）查找和挑選檔案；（4）正文的加工和編排；（5）編寫彙編內參考材料；（6）彙編的校審和出版。[11]事實上該書便將每個細節加以說明，從如何選題、編前研究、查找和挑選檔案、內容考訂、到編排的體例等，無不鉅細靡遺地加以介紹。這樣詳細的說明當然顯示了編排檔案時的負責態度，但同樣地，這麼小心翼翼、如履薄冰的態度，也正反應了檔案重新編排時的高度脆弱性和危險性。因為依事件關係替研究者對檔案進行梳理，不得不面對檔案編纂和分類的主觀危險。所謂的史料編輯本身，便是一種具主觀的選擇，最嚴重的情況，還是刻意為特定目的或動機服務的選擇。《檔案文獻編纂學》一書便反覆強調進行彙編的政治性動機和意義。這種為特殊目的編纂方式所帶來的影響，還比較容易避免，只要具有相當警覺心的研究者都應能迴避，更麻煩的則是無心的主觀流入，其動機出於真正對學術需要的整理，但卻在無意中，將編纂者自己理解過程的預設不自覺地帶入檔案，以自己的架構替代了檔案的原始架構。一旦整理完畢編印成冊後，檔案史料的解釋空間很可能便被凝滯於某一點上，窄化了解釋的可能，甚至形成一種誤導。以下即以戴炎輝先生整理的《淡新檔案》為例，說明紙本編輯者分類整理後對檔案所造成的影響。

《淡新檔案》是指清朝乾隆41年（1776）至光緒21年（1895）間淡水廳、臺北府及新竹縣的行政與司法檔案，該批檔案於日本時代由新竹地方法院承接，轉送覆審法院（即高等法院），最後再轉贈臺北帝國大學文政學部，供學術研究之用。日本時期因為該批資料主要是清朝治臺之相關檔案，故命名為《臺灣文書》。戰後移交臺灣大學法學院，由法律系戴炎輝教授主持整理工作，以該批資料所涵蓋之地域改命名為《淡新檔案》。戴炎輝對該檔案的整理功不可沒，在面對如此龐大、無章法的資料，他進行了一系列基於近代法律學架構的整理，將之分成行政、民事、刑事三編，其下再細分成16類，類以下再區隔成102款。[12]這一巨大的工程，對臺灣史的研究帶來深遠的影響。

戴炎輝的分類方式，確實提供了理解《淡新檔案》時切入的便利性，但卻無法除去以「後見之明」的方式解讀史料的必然問題，他引用西方現代法學觀念詮釋清朝臺灣的衙門運作，其侷限十分明顯。諸如分類中所見的「雜事款」（分別出現在行政編總務類、行政編財政類）、「其他款」（出現在刑事編總務類）的分類細目，不明確的題名，正顯示出現代法學觀念無法完全包涵《淡新檔案》的內容。原本依照吏、戶、禮、兵、刑、工的傳統六房分類的《淡新檔案》，其原初脈絡在戴炎輝的分類中喪失，如學者林文凱所指出的「戴炎輝這一新的分類與命名整理，無法完整

11 徐紹敏主編，2001，《檔案文獻編纂學》（第二版），頁8-9。

12 關於《淡新檔案》的詳細資料與介紹，可參見國立臺灣大學圖書館為該檔案所建設的相關網頁，內容十分詳盡：http://www.lib.ntu.edu.tw/cg/events/TH/tanhsin_page_01.html。上列關於《淡新檔案》的簡介即為整理網頁資料而得。

呈現出一個行政與訴訟事務在地方官府的整體統治中原有的體系性」。尤有甚者，這樣的分類方式也易造成研究者在使用《淡新檔案》時的盲點，諸如滋賀秀三、黃宗智與艾馬克等學者，在進行有關淡新檔案的土地法律文化研究時，就都僅以民事門的土地控案為分析，不自覺的接受戴氏分類法中三權分立的體制，忽視了當時土地政策、土地行政與土地訴訟間的真正關聯，難以清楚詮釋出清代土地的法律文化。[13]

當然，需要特別強調，對《淡新檔案》進行重新分類，很可能是戴炎輝不得不然的選擇，因為當戴炎輝挖掘出這批史料之時，保存狀態極差，原有的分類應早無留存。[14]選擇以西洋法的分類進行處理，是在別無他法下的選擇。《淡新檔案》的例子，說明了紙本的檔案保存絕非萬無一失，編輯者的主觀選擇，會直接影響到檔案的整理與出版，更會間接的影響到以後的研究取徑。這形成了一個積極檔案保存者的兩難，一是選擇最安全的方式，以所出機關或時間的方式進行整理，但這樣卻失去了使用者在運用該檔案的便利性；另一則是積極的以事件或較現代的觀點，用使用者比較好理解和操作的方式進行編排，但卻又不得不冒著污染檔案的風險。

然而，這樣的矛盾正突顯了檔案史料數位化的優勢，也說明了數位化的必要。在下文中將進一步說明，紙本檔案本身所存有的脈絡困境，在進入數位化的時代，在看似加劇的同時，卻正提供了解答的良方；此外數位化後的檔案在與一強大的系統搭配後，將為檔案的脈絡帶來新的觀察與研究可能。

四、數位時代檔案研究的新可能

在說明紙本檔案亦有盲點之後，在本節中將指出檔案數位化後，將會帶來新的檔案編排和保存的型態，這種新的型態不僅可以充分回應對數位化檔案脈絡的質疑，並能從更寬廣也更關鍵的角度，體現數位化的方式對檔案進行處理在研究上的優勢。種種質疑之聲，有些是承繼紙本的缺陷而來，有些則是因數位保存的特性而產生的，在檔案數位化後，前者得到顯著的改善，後者在經過強有力系統的處理後，則變成了研究者在數位時代的優勢；在下文中將指出，數位化的檔案可以更進一步解決紙本的缺陷，並完成紙本檔案所無法完成的功能，為人文研究帶來全新的視野。

---

13 林文凱，2004年12月，〈清代晚期臺灣的土地法律文化——淡新檔案內淡新地域漢墾莊在十九世紀四十年代以來土地抗租案件為主的分析〉，發表於臺灣社會學會等主辦「2004 臺灣社會學會年會暨『走過臺灣——世代、歷史、與社會』研討會」。

14 據戴炎輝孫子戴桓青的回憶：「光復之後臺大法學院圖書館的陰暗角落裡有一批殘破的清代檔案，原是臺北帝大的收藏，因蟲蛀水浸無人問津幾乎變成廢紙，祖父看到了卻如獲至寶。……在還沒有影印機與電腦的時代，祖父傾10年之力修補整理……許多蟲蝕水浸闕漏之處，是祖父參酌考證之後再加以褙裱補寫才維持完整。」可以想見當年在處理這批檔案時的艱鉅。見戴桓青，2008，〈四代百年臺大情〉，《臺大校友雙月刊》，第60期，頁24-25。

檔案數位化的優勢不僅只是儲存方式的不同，更重要的是，數位化檔案優異之處在於它是和「系統」相結合的，一套功能強大而完備的資料庫系統，除了能彌補過往檔案處理的不足，並能為人文研究帶來新的觀察面向與研究功能。

（一）數位時代檔案的收集與整理的新樣貌——再論「何謂檔案」

數位化檔案給予研究者最大的不信任感，如前已提及的，一是集中在是否會破壞檔案的原始脈絡上，在前文中我們試著說明所謂「檔案的原始脈絡」本身便是一項迷思，即使是在紙本的年代，為了增加檔案研究的便利，讓使用者不會迷失在史料的浩瀚無垠之中，將檔案從原生的機關、時間脈絡中抽離，一直是檔案使用的常態。之所以為常態，正因為這是符合使用者需要的最佳方式，然而問題便在於要進行這樣的整理工程，需要花費大量的時間、人力和物力，而且經常會有掛一漏萬的風險。二則是檔案從紙本經過數位處理之後本身性質的改變，其新的特性，會讓習於紙本的使用者感到不安，因為一旦數位化，檔案變成了看不見的位元，不再令使用者能親身接觸。以下我們試著以檔案的多元／源化和檔案的原子化兩者為例，[15]加以論證，希期去除掉這類直覺式的恐慌。這些數位化後新的特性，經過精細的系統的處理後，不但不會造成盲點，反而會成為新的研究利基，創造出一種全新的檔案使用邏輯和視野。

首先是檔案的多元／源化，如前述這是在紙本時代即有的難題，過往在思考檔案時往往由印刷的角度去思考，非文字的資料如影音等視聽資料，除非經文字化處理，否則必難以同時保存，進入數化位時代後，這樣的問題便不再存在。[16]多源化則是指檔案來源的不同，這可說是傳統檔案形式所有的形式之延伸，不單只是指跨機關的檔案整理，更涉及了許多原本就不具連貫性的資料文件，如「臺灣歷史數位圖書館（THDL）」[17]所藏的古契書就是最明顯的例子，該資料庫收集了從清代到日本時期臺灣地區的土地契約文件，其內容多半為古地契，記載著某塊土地地權的轉移或交易情況，其總量截至目前為止共有39,255件，[18]每月仍不斷增加中，其來源共有

15 這並不表示說檔案的數位化所面臨的新問題，僅多元／源化和檔案的原子化這兩點而已，除了這兩點之外，檔案內容的「質」的問題，亦即其數位內容和後設資料的正確度，仍是史學研究者在使用上，所感到擔心的。莊樹華在一次研討會中就指出，這種「質」的殘缺，往往比檔案「量」的不齊，帶來更大的風險。莊樹華，〈數位檔案運用的契機與風險——以近代中國外交檔案為例〉，發表於國立政治大學圖書資訊與檔案學研究所主辦，「數位檔案加值與教學應用研討會」（2010年5月27日）。此外，還有檔案Silo化的問題，當數位化如星火燎原般四處擴散時，如何不會各自為政，乃至重複，也成為檔案數位化後的新難題。但這些問題都比較偏向於技術性和整合性，和本文從研究的角度提出的質問略有不同，故無法於此處進一步加以細論。

16 以目前國史館資料庫中「元首文物」的檔案分類而言，蔣中正、蔣經國、李登輝等總統文物，其類型便包含了文件、底片、照片、視聽，同時存在一個資料庫中，供使用者觀察。

17 臺灣歷史數位圖書館（THDL）網址為：http://thdl.ntu.edu.tw/。

18 2012年10月12日。

100多個，除了已出版成冊的書籍外，還有許多原件資料，其數量和延伸性皆非過去紙本時代所能想像。

檔案原子化的問題，在前文已稍涉及，關鍵在於數位化的過程所給予人們的想像，和紙本整理的過程正好相反。過往紙本的整理過程是將一份份分散的檔案，透過抄錄、打字彙集成書，檔案的數位化則是將散落或已編輯成冊的檔案，再一份份散解轉換成數位的檔案；中間更涉及了由實體變成虛擬的過程，加深了人們心中主觀的不信任感。如何在原子化的數位檔案之間，「回復」、「保有」檔案的既有脈絡就成為多數人在思考檔案數位化的焦點。[19]此外，數位檔案的原子化所涉及的另一個層面，對大部分人來說，數位檔案資料庫的使用，多半僅是針對全文或檔案描述進行關鍵字檢索，很容易形成「見林不見樹」的情況。[20]總之，各種類型和複數來源的組合，造成數位化檔案不同於過往的集合型態；由實體轉變成虛擬的過程，也使得檔案數位化的保存予人更形分離的感覺，這種疏離感使人們恐懼無法掌握資料的全貌。這兩種思考的結合，造成了對數位化檔案的不信任。

然而，問題便在於這些對數位資料的擔心，其實是基於既有研究習慣後刻板印象式的恐慌，反而忽略這些憂慮所代表的，乃是一手檔案史料的本質，之前之所以無此困擾，並不是因為這些問題不存在，而是因為紙本印刷的編輯形式，掩蓋甚至割裂了檔案資料的原有特質與風貌，一旦檔案在應用上的可能性被壓縮，自然就不會存有風險，而檔案一旦經由數位化之後，從紙本的限制中獲得解放，無限的可能被釋放，重新回復其本色，是以那些對於數位化檔案的擔憂往往正是其長處所在。譬如多元／源化的問題，這本來就是一手史料的特性，多元化在史學研究多角發展的今日，諸如影視史學、圖像歷史早已成為新的研究重點，研究者原本就不應該自囿於文字的資料，反過來說如何結合文字和影像進行觀察才是困擾研究者的癥結所在，其起因便在於在紙本世界中，文字的邏輯和媒介都難以將影像納入，這樣的困難在數位世界中根本無法立足，因為文字和多媒體的互動，原本就是數位的常態。史料的多源化與其說是研究者的惡夢，或說對檔案脈絡的傷害，正好相反，史料的多源化反而應該是研究者的美夢成真，傅斯年對史學研究者查找資料理應「上窮碧

19 史學研究者洪麗完便曾強調「為了保持檔案資料的完整脈絡，入藏時不能隨意拆散（整批入藏），應注意資料本身內在脈絡」；作為一個專業的臺灣史研究者，她認為對龐大古文書的整理「社會、經濟、文化與政治、宗教」等分類，會較以單件契字的性質有意義。洪麗完，2007，〈評論稿〉，收於逢甲大學歷史與文物管理研究所、台灣古文書學會編校，《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，頁331。

20 莊樹華和龐桂芬便為文指出：「檔案的描述雖有助於使用者在大量的資料群中快速擷取資料，但過度依賴關鍵字詞所取得的資訊，往往會切斷了檔案彼此之間的關聯性，以致隱藏於關鍵字詞背後的概念形成，及思考過程之間比較模糊的地帶，恐怕都被湮滅。數位化的便利，對史學研究者是節省了資料收集的時間，可以投入更多的分析與思考，還是讓研究者迷失於資料取得的便捷與豐富，而忽略了歷史研究的深層思考，是值得深思的問題。」具體說明了關鍵字檢索對研究所可能造成的盲點。莊樹華、龐桂芬，2006，〈數位化時代下的檔案館營運——以中研院近史所檔案館為例〉，《檔案季刊》，第5卷4期，頁23。

落下黃泉」的經典名言，其所描繪的理想便是要從各種方面收集史料，讓支持研究的資料能「多源化」。再引前述THDL內的古契書資料為例，在過去要能同時看到那麼多的契書，是件十分困難的事，即便有一「理想的」圖書館能收集所有編印出版的契書史料，全部攤開於研究者眼前，對研究者來說反而是種亂無頭緒的巨大負擔與挫敗。而今日只要透過網路的連結，使用者便能在自家電腦上立刻接觸所有的古契書，不用再舟車勞頓去尋覓那「理想的」圖書館。尤有甚者，數位化不單解除了一本本古契書史料集的紙本限制，同時更開啟了有效重構並觀察史料脈絡的可能。不只是多源，而且是挖掘出潛藏條理枝繁葉茂的多源景象，令使用者可以依據其研究的尺度，對檔案資料進行閱讀，使多源不再是難能可貴的理想，更不是使用者的負擔，而是使用者能輕易使用的利基。

檔案史料原子化的問題，則可分成兩個層面，一是在研究流程上，「見樹不見林」的問題，並非是數位時代所獨有，而是紙本時代即有的現象，研究者僅翻閱可能和研究有關的部分，只選取關鍵字「拿了就跑」的割裂式引用，一直是被批評、教導學生避免的陷阱。反過來說，之所以如此，便在於人有限的精力和時間在面對龐大的史料量時，無法負荷，這正是資訊技術所能輔助之處。是以，割裂式的引用是從紙本時的延續，而有效的資訊技術正是可能的解答之一。在另一個層面上，檔案資料在相當程度上，其本質便理應被原子化的對待和處理，紙本的編輯形式硬生生改變、重組檔案脈絡出版成書，反而對檔案史料帶來難以估計的傷害。前述《淡新檔案》即為一例，重新以後人的眼光編排雖是一種不得不然的安排，但卻也必然對檔案的理解帶來侷限的可能，如果能維持其個別檔案原子化的型態加以保存，並能維持一種可自由調整關聯的串連方法，無疑才是最佳的整理方式，而這是在紙本世界所難以達成的。《淡新檔案》絕非歷史上的孤例，事實上紙本保存的脆弱，讓很多檔案資料在流散後重輯的過程中，都面臨了一樣的難題。如在宋代研究中十分重要的史料《宋會要輯稿》為例，該書即是在《宋會要》亡佚後，從《永樂大典》中輯出，最初的編纂者徐松，便以無法有效回復《宋會要》的原貌，只能自行加以重構，而後又面臨多次的轉手、修正和割裂，今日的《宋會要輯稿》和《宋會要》之間有著再多的考證仍難以填補的差距。[21]然而一旦還原其原子化的風貌，輔以各種脈絡觀察的可能，一來或可析釋後世修刪者的詮釋，二來能讓《宋會要輯稿》的內容依據時間、空間或其他研究需要自由編排，雖然無法恢復原貌，但或許更能切合《宋會要》作為「政書」的原意。

總而言之，一旦當我們重新去深入思考檔案資料的本質，檔案唯有經過數位化後，才能彰顯出檔案的本質。紙本編排方式的穩當，仍是建立在對檔案資料之脈絡和可能性的壓縮上。是以，當檔案數位化重新呈現其風貌時，無論在研究和應用

21 參見陳智超，1995，《解開《宋會要》之謎》。

上都能開發出全新的可能。檔案理應多元、多源和原子化，那些令人擔憂的難題，與其說是針對檔案的還原，倒不如說是在檔案從紙本限制中解放後的駕御。因此，如何駕御數位化的檔案資料，才是數位人文真正的關鍵之處，而要能達到這樣的目的，其解答便是「系統」。

（二）數位化檔案的核心方法論：系統

如前述，數位化的檔案要能發揮其研究價值，最關鍵之處便在於「系統」。傳統檔案整理過程中，檔案的品質和整理分類決定了檔案的使用方式和難易，面對數位檔案則需要非常不同的思維方式。一個非常基本但常常被忽略觀察是數位檔案必須經過中介的媒體才能夠被使用。換句話說，數位化檔案無法單獨存在，使用者一定要透過一個軟體系統才能對檔案進行有組織的研究與使用。這個觀察雖然看起來是老生常談，但卻是數位檔案和傳統檔案最關鍵的差異。也就是說，談數位檔案就不能不談檢索系統，而檢索系統的設計與功能也就決定了數位檔案的可用度，所以我們在探討數位檔案時就不能僅僅從檔案的層面看問題，而是要把檔案的系統當成和檔案一樣重要的一個環節。

既然數位檔案和系統是密不可分的，我們需要花一些篇幅探討一個數位檔案的系統應該，或說能夠，提供什麼。

傳統的檢索系統，包括搜尋引擎、圖書館自動化系統，以及大部分的文件查詢系統等皆以求準率（precision）和求全率（recall）作為評斷系統可用度的指標。直觀的來說求準率指的是檢索結果與使用者的需求的相關度，求全率則是檢索結果包含使用者期望得到的所有文件的比例。既然研究者最關心的是相關的文件有沒有被找到，一般認為一個針對研究者需要所設計的檢索系統需要有高的求全率。高的求全率當然是需要的，但它是不是研究者唯一的需求呢？

當我們進一步檢視檢索系統的設計時，會發現這些傳統系統背後大都有一個假設，就是系統內所收集的文件之間沒有相互的關聯。換句話說，當使用者下了一個關鍵詞，系統將搜尋到的文件用系統內建的排序函數（ranking function）排序後依照系統認為的相關順序條列給使用者。但是因為系統內的排序函數和使用者對相關的認知可能有很大的差異，所以所回傳的文件的順序往往不符合使用者期待，而給使用者一種支離破碎的感覺，這也是為什麼數位檔案的使用系統會給人一種破壞原有檔案脈絡的印象。

但是對一個研究者而言，檔案中的文件是有關聯的；當一個研究者使用一個檔案系統尋找資料時，除非他很清楚自己要找的是某一個特定文件，不然的話，通常都會要研讀所搜尋到的一堆文件（而不是單一的一件）。而在研讀這些文件時，要

找的也就是文件與文件間的脈絡。換句話說，對研究者而言，文件除了本身的意義外，更有文件間經由脈絡的串連所共同產生的意義，這些脈絡往往才是研究者在尋找的，而文件本身即是支持這些脈絡所產生的意義的證據。

所以一個以研究需求為出發的數位檔案檢索系統需要考慮到文件間的脈絡問題，其建置的基本邏輯除了要保持所收錄的檔案本身原有的脈絡外，還需要具有開放性，能夠產生與呈現各種不同連結可能的多元脈絡；一個符合使用者需求的檔案檢索系統應盡量發掘文件間各種關聯，並提供使用者一個觀察這些脈絡的環境。

既然檔案中的文件不是單獨存在的，一個檔案檢索系統理論上來說應該可以產生及呈現檔案中任何一個文件子集合中的重要脈絡，事實上也是如此。但是為了說明的方便起見，我們可以把任一文件子集合想成一個檢索成果的集合（a query return as a sub-collection），而檔案檢索系統的一個重要目標就是提供各種方法讓使用者對檢索成果的集合中的檔案文件間脈絡做觀察與分析。

以上我們指出一個檔案的檢索系統不能僅被視為檔案數位過程中的「附加物品」，而應當成是數位檔案中重要的一環；任何一個系統的建置，都應依據檔案的特性，量身設計出所需要的系統。而這也才是建構數位檔案最核心的方法論，也才能讓數位化檔案的研究潛能真正發揮。

（三）數位化檔案系統所能提供的檔案脈絡

既然一個檔案檢索系統的任務不僅是讓使用者可以檢索到需要的文件，而且更要提供文件間的脈絡，我們就需要進一步分析一個檔案系統應該提供給使用者怎樣的史料脈絡。我們可大致將這些脈絡歸結成六類：（一）檔案原始的脈絡；（二）重組原始脈絡產生的多重脈絡；（三）鳥瞰型的脈絡；（四）子文件集的縱觀脈絡；（五）文件間統計型的脈絡；（六）文件間隱藏的新脈絡。以下將用國立臺灣大學數位典藏研究發展中心目前所發展的檔案檢索系統做例子，[22]說明這六種脈絡的意義及如何達到。

1. 原始脈絡的保存——保持原本脈絡的瀏覽功能

無論數位化檔案的系統功能多麼強大，其基礎仍還是構築在史料保存的真實和完整上，誠如論者已指出的「檔案的使用者絕多為史學研究者。史學研究者對資料的掌握，既要求完整又求正確」，再加上許多單位在檔案數位化後，基於保護史料的原則更不希望使用者閱讀原件，所以資料庫要提供檔案的瀏覽功能，應讓使用者

22 臺灣大學數位典藏研究發展中心的首頁為：http://www.digital.ntu.edu.tw/，除本文所提到的相關資料庫外，該中心網頁還提供了豐富的訊息和資料庫，歡迎讀者前往使用。

可以留有傳統的閱讀習慣。[23]如果從研究的角度來看，即是要保留檔案的原始脈絡，讓研究者在開展自己論述的同時，可以隨時回歸到史料最原初的狀態，掌握檔案的時間、機構、全宗、相關事件等等資料。以「臺灣省議會史料總庫」為例，[24]使用者可以在不輸入任何關鍵字的情況下，直接點入「會期瀏覽」，按會期一頁頁的閱讀。此外如圖1所示，「類別瀏覽」功能則可以讓使用者一目瞭然檔案的歸屬。當使用者在閱讀「三七五地租」的相關檔案時，可以立刻知道其原本是屬於民政事務＞地政＞地權；當在閱讀「菸酒公賣」的資料，可以馬上知道是在「質詢」的類目下。

2. 重組原始脈絡產生的多重脈絡——既有知識結構的重組

紙本印刷的最大限制是只能提供使用者一種脈絡，這個限制使得檔案只能以一個線性的方式發展。這個限制在數位的世界可以被完全打破，而允許使用者將原有的脈絡重組成多重的脈絡。

一般來說，一個整理完好的檔案都有它本身的分類（樹狀）結構，而這個知識結構一旦固定後，檔案中的每份文件也就被這個分類固定住。我們可以舉THDL研究工具集中「清代臺灣官職表」系統為例，[25]該系統以鄭喜夫編纂，國史館臺灣文獻館所出版的《臺灣地理及歷史．官師志》為藍本，這本書是研究清代臺灣官制重要的參考書籍，當我們試圖將這本書由紙本轉換為數位系統時，不只是把資料放上網路，也不僅提供檢索。更重要的，這些資料變成立體的，有結構的，彼此連結的。從原書的目錄即可知，該書是以官職作為分類，按年代先後，依序記錄曾任該職位的人名、就任時間與原委、離職時間與原因。這樣的安排堪稱詳密，但卻無法方便的對一個人仕宦生涯進行觀察，即以人作為出發點，而只能被限定以職位作為唯一的視角，而且還是沒有上下隸屬關係的孤立記載，官職背後所具有的變動或其他觀察可能都被排除了。

但轉化為數位檔案後，即能打破紙本的凝滯，以研究者的需求靈活排列，進行資料組合與調整，並以樹狀結構還原官職原本應有的歷史結構（見圖2）。以譬喻形容的話，前者檔案就像拼圖，後者則像積木，一片片拼圖雖然看似可以分開，但一定還是要合在一起才有意義，而其拼湊的方法也只有一種，就是原來設計者所畫出的圖像。積木則大不相同，它可以依照各種不同的想法和需要，而拼造出各式各樣組合。

---

23 莊樹華、龐桂芬，〈數位化時代下的檔案館營運——以中研院近史所檔案館為例〉，《檔案季刊》，第5卷4期，頁26。

24 「臺灣省議會史料總庫」網址為：http://ndap.tpa.gov.tw/drtpa_now/。更多關於該資料的介紹可參閱項潔、董家兒、蕭屹灵，2008，〈臺灣省議會檔案數位典藏系統之研發與建置〉，《臺灣省諮議會會訊》，16期，頁33-46。

25 該系統網址為：http://140.112.30.230/Career_tb/。

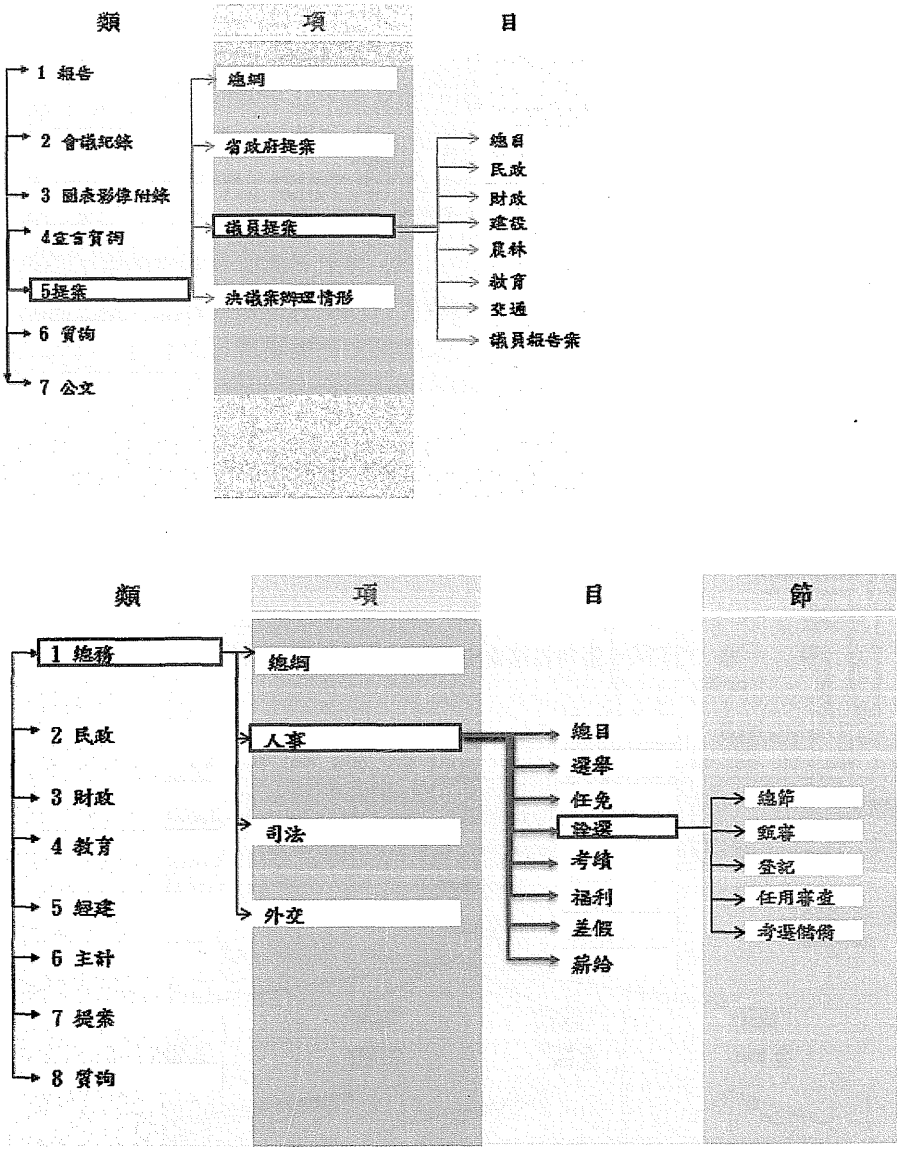


圖 1-1　臺灣省議會史料總庫的原始結構

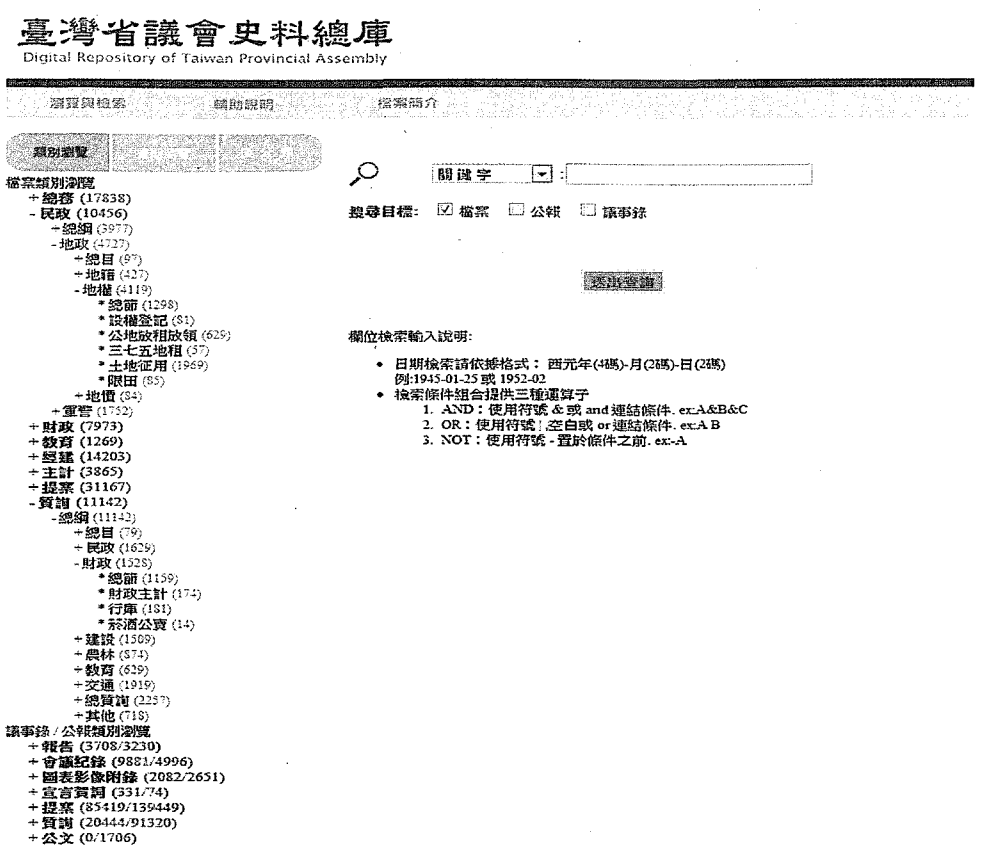


圖 1-2　臺灣省議會史料總庫的原始脈絡保存

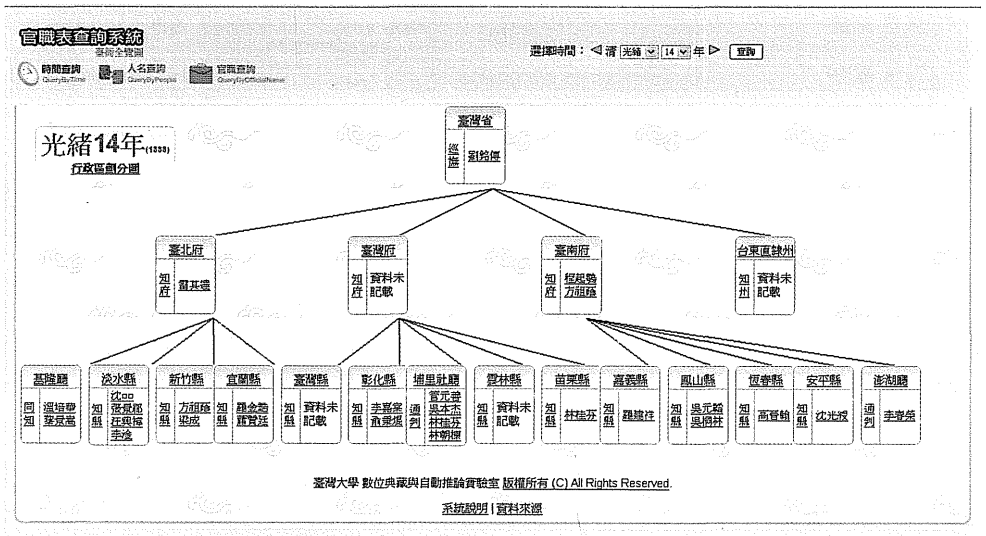


圖2　清代臺灣官職表的樹狀結構呈現

3. 鳥瞰型的脈絡——文件集整體意義的顯現

當我們在面臨大量的檔案資料時，好的檔案系統可以提供使用者立即的鳥瞰脈絡，對檔案的構成有即時而大致的理解，如圖3為例，透過THDL的整理，可以將龐大而多源的明清檔案內所收的文件，立刻於時間軸劃出年代的分布。假設某個年份的文件數量越多，則表示當年某些事件對臺灣政治社會的震撼越大；是以對臺灣影響最大的前六件事件依序為：（1）1787年林爽文事件；（2）1884年中法戰爭相關事務；（3）1895年中日甲午戰爭相關事務；（4）1874年牡丹社事件；（5）1806年蔡牽之亂；（6）1833年張丙事件。然而倘若和圖4明清檔案中由《清實錄臺灣史資料專輯》中的文件分布進行比較，可以明顯發現林爽文事件壓倒性的勝過其他事件，兩相比較，或者可以假設清廷官方對於林爽文事件的重視程度，遠超過中法或甲午等戰役。此外，影響臺灣西部開墾甚為重大，並造成族群劇烈遷移的郭百年事

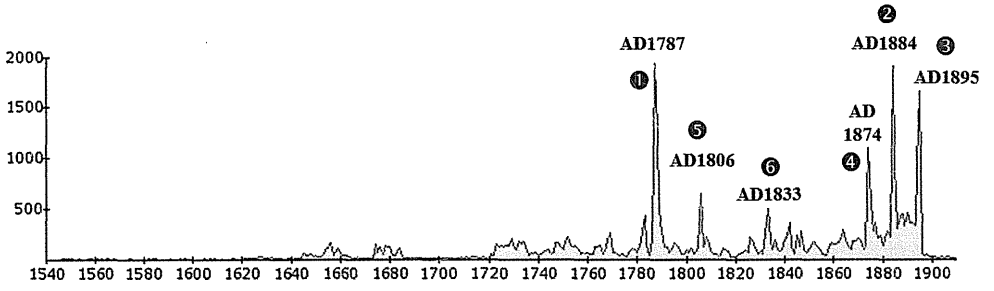


圖3　明清檔案文件在年代上的分布圖

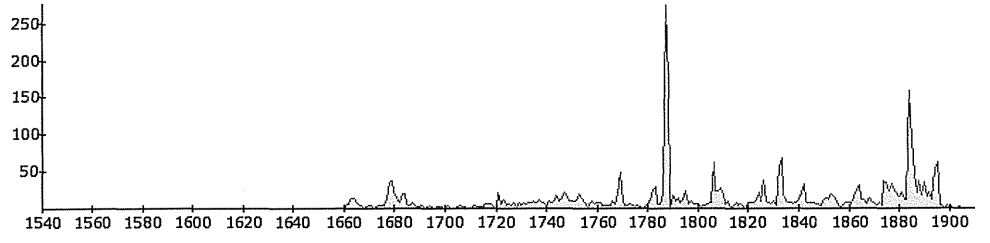


圖4　明清檔案中《清實錄臺灣史資料專輯》文件在年代上的分布圖

件，[26]卻未在明清臺灣行政檔案之中留下任何的紀錄，再度顯示了中央和地方在視野上的差異，而這種差異性是單憑人力一件件閱讀所無法勾勒的。前述的解釋當然還有待研究者進一步的析理，但要強調的，這種鳥瞰性脈絡的觀察方式，是紙本或傳統檢索系統所難以想像的。

4. 子文件集的縱觀脈絡——檔案原有知識結構的多重脈絡呈現

縱觀脈絡是一個檔案檢索系統應該提供的最核心的脈絡，也是我們方法論中最基礎的觀念。大體來說，每個檢索系統都允許使用者用查詢的方式尋找文件，而查詢的結果可能來自檔案知識結構的各個支系。這個查詢結果可以被看成原有文件集的一個子集合，而所謂縱觀脈絡就是指這個子集合在原來知識結構之多重脈絡裡所呈現的各種關係。這些關係中最明顯的是原本詮釋資料（metadata）的欄位中本就蘊含的。舉個例來說，如果年代是一個詮釋資料中的欄位，則年代可以用來作為子集合中文件分類的一個關係。我們稱這種呈現關係的方法為「後分類呈現」的技術。如在THDL中，使用者可以將檢索得到的子集合依年代、出處、作者、類型、地域等欄位做後分類，然後進行更進一步的脈絡觀察。

圖5是THDL中一個檢索成果的後分類呈現。其中左上角是檢索成果的地理分布；（針對古契書我們也做了地理資訊的分析，[27]這些座標資料也是後分類的一個方法）中間是時間分布；左下是詳細的五種後分類，分別是年代、出處、作者、性質分類及地域。圖6顯示這五種後分類的內容。

但我們要指出詮釋資料的欄位並不是唯一的後分類方法。其他資訊（如子集合中出現的人名、地名）也可以用來做後分類。後分類的目的是讓使用者對檢索成果得到的子集合中的文件用各種不同的角度觀察時可以看出的分布情形。這些分布往往是重要脈絡發現的開始。

我們再以「日治法院檔案資料庫」為例，說明後分類的效果。「日治法院檔案資料庫」的主要內容為臺北、新竹、臺中、嘉義等四個地方法院所典藏的日治時期各類司法文書，包括民刑事判決原本、民刑事案件登記簿、非訟事件以及強制執行事件等卷宗、公證書原本、有關法人等的各種登記簿、行政卷宗等，一共 5,645冊，[28]檔案內容分為「條」、「項」、「款」，在詮釋資料欄位另外以「冊」及以「案」

---

26 相關論述與分析可參見陳嵩明，2010，《清代臺灣界外武力拓墾：噶瑪蘭、水沙連、與竹塹東南山區之比較研究》，臺北大學社會學系碩士論文。

27 蔡炯民、歐仲翔、翁稷安、項潔，〈時空資料的地方再現：GIS與數位典藏的交會〉，《國土資訊系統通訊》，78期（2011.6），頁13-20。歐仲翔，〈使用者取向之歷史地理資訊系統：古契書與統計資料呈現〉，國立臺灣大學資訊工程學研究所碩士論文，2011。

28 項潔、蕭屹灵、董家兒，2009，〈日治法院檔案數位典藏系統之研發與建置〉，收於王泰升主編，《跨界的日治法院檔案研究》，頁83-148。

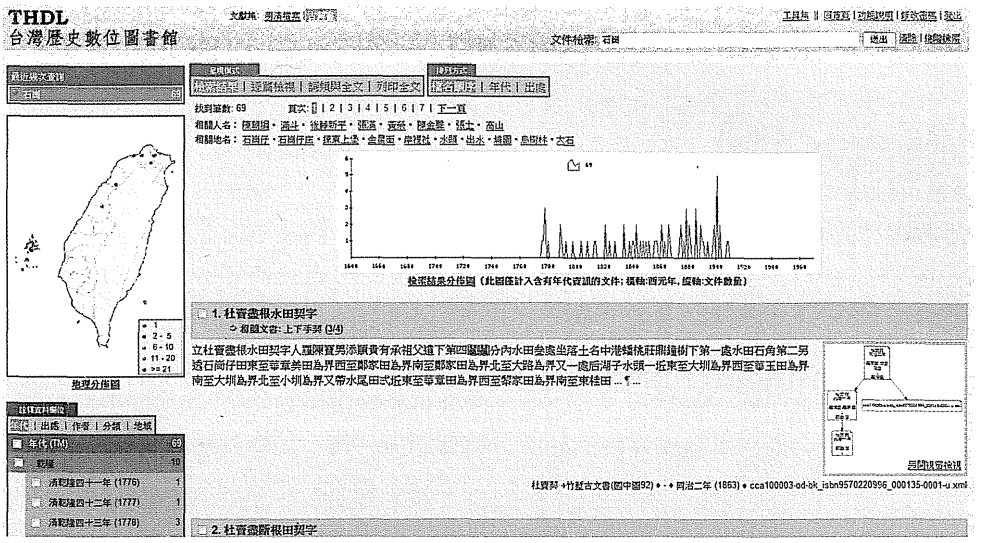


圖5 THDL中以「石岡」為檢索詞的檢索結果

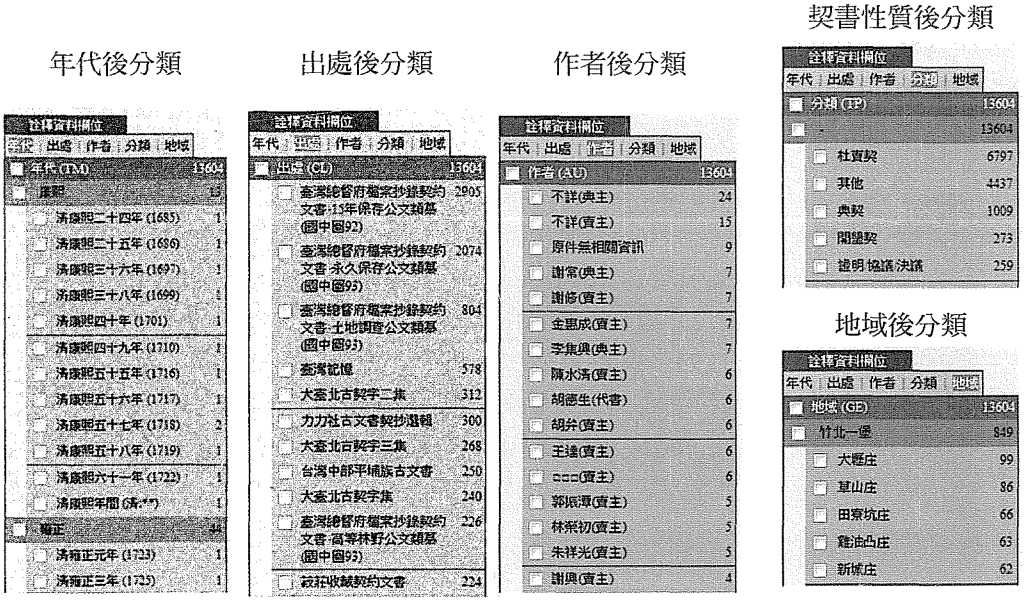


圖6 「石岡」檢索成果的年代、出處、作者、契書性質、及地域的後分類

為單位分別建置的情況下，共有307,139 案的詮釋資料，200 萬餘張影像。這批資料除了數量龐大與保存現況甚差外，內容本身在當初形成的時候也因來自各個法院而有差異，如果要對這批資料做綜合性的研究會十分困難。

「日治法院檔案資料庫」和我們建置的其他資料庫一樣，都用檢索成果後分類作為輔助使用者觀察縱觀脈絡的機制。[29]我們以關鍵詞「離婚」為例。以離婚做關鍵詞檢索可以找到2,740筆資料，從法院所屬後分類可看出新竹、臺中、臺北、嘉義地方法院均有（分別為204、875、1,108、553筆，其中民事判決原文分別為204、875、1,103、553案，其餘六件為公正證書），年代分布從明治38年（1905）的兩案到昭和20年（1945）的五案，每年都有，其中在大正9年（1920）起進入每年超過百案的高峰期。雖然在昭和8年（1933）後案件數量回復兩位數字，但一直到太平洋戰爭開始（1941）才降到每年五十案以下。這種透過檢索得到的子集合中的文件可能散布在檔案的各處，這些文件中的縱觀脈絡透過後分類可以輕易被觀察到，但是不用後分類的機制則不易被察覺。針對這2,740筆資料做進一步的觀察可以發現，其中85%的離婚案件的原告是（臺灣）女性。這個現象透過後分類可以立刻被觀察到。如果沒有後分類的觀念則極為困難。

後分類的觀念和faceted search [30]類似，但基本的出發點有很大的差異。Faceted search後分類的目的是讓使用者在使用不夠精準的關鍵詞檢索的情況下可以比較容易的從大量的檢索結果中找到想要的文件，所以它的設計還是以找到某個特定文件（也就是作為加強求準率的效能）為主。我們的後分類的目的則是提供使用者所檢索到的子集合中的文件之間的關係。因為不知道使用者所想的脈絡是什麼，所以要盡量提供各種不同的後分類及各種呈現的方法，以讓使用者可以對他所需的做觀察。

又或者如「總督府抄錄契書GIS工具」，如前述在面對大量的古契書資料，使用者往往不知如何進入，一旦輔以GIS系統便可以立即掌握契書所在的位置，同時還可以再依文件種類和時間的過濾，對土地開發有概略的理解。以姜秀鑾對新竹一帶的開發為例，圖7「姜秀鑾對新竹開發示意圖」是以道光14年（1834年）12月，淡水廳同知李嗣業任命的金廣福墾隘的粵籍頭人姜秀鑾為關鍵字，分別檢索1835-1836年、1835-1838年、1835-1850年、與1835-1895年這四個時間區段，觀察與姜秀鑾有關之契約文書在「總督府抄錄契書地理資訊」雛型系統上的座落分布，從圖中可以看出這些契約文書的座落分布，最早由今日的新竹縣的竹東市開始，逐漸向北埔、寶山、及峨眉三個鄉拓展。除了讓我們可以依時間序列與地理分布兩者關

29 「日治法院檔案資料庫」的網址為：http://tccra.lib.ntu.edu.tw/tccra_develop/。相關研究範例，可參見項潔、蕭屹灵、董家兒，2009，〈日治法院檔案數位典藏系統之研發與建置〉，收於王泰升主編，《跨界的日治法院檔案研究》，頁83-148。

30 參見Vanda Broughton, 2006. The Need for a Faceted Classification as the Basis of All Methods of Information Retrieval, *Aslib Proceedings*, Vol. 58 Issue1/2, pp. 49-72.

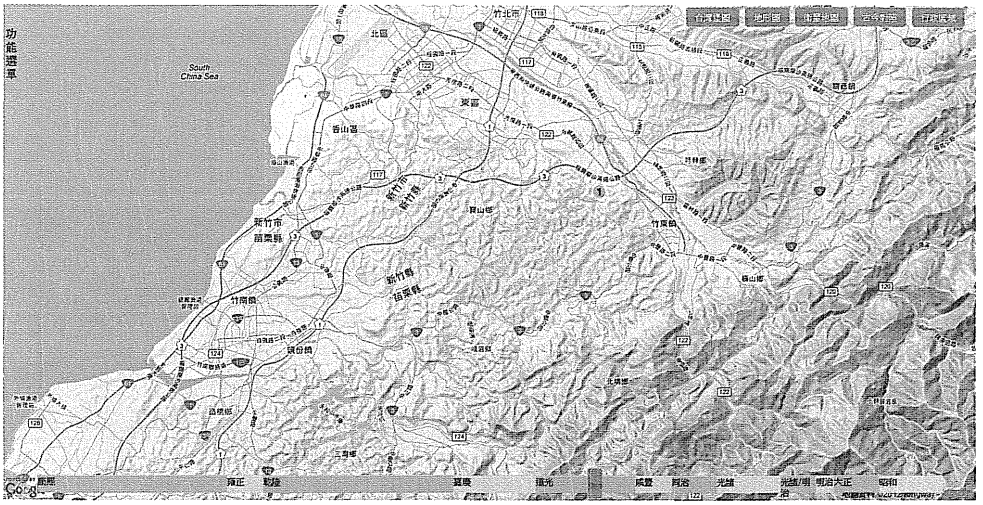


圖7-1　1835年至1836年，金廣福開墾初期

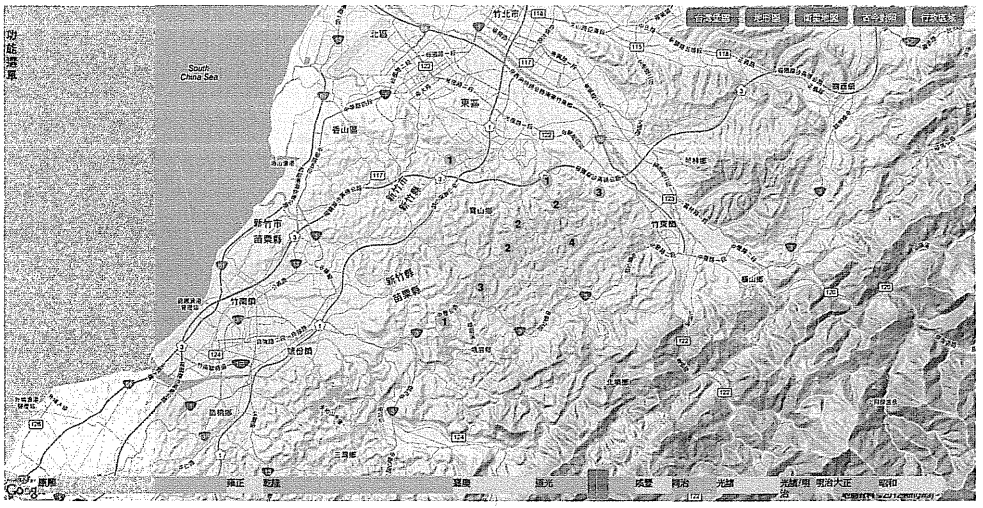


圖7-2　1835-1838年，開始拓展至峨眉鄉

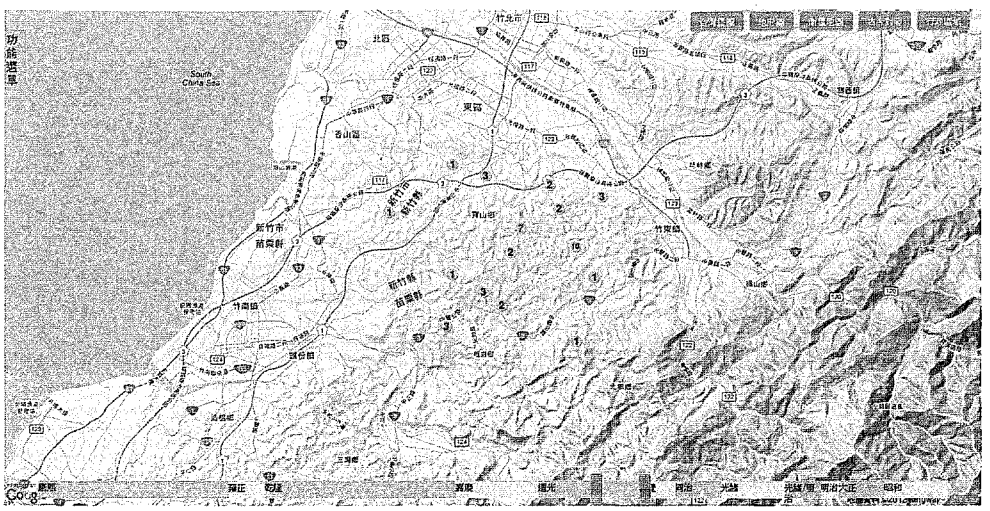


圖7-3　1835年至1850年，以峨嵋鄉為基地，持續開發峨嵋溪流域

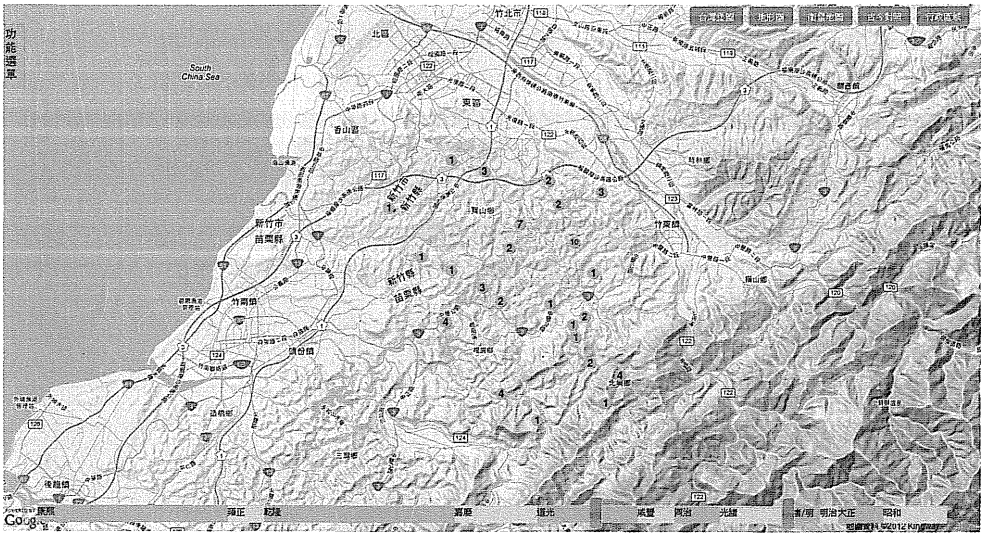


圖7-4　1835年至1895年，可看出金廣福開發範圍

係，約略對照出當年金廣福整合新竹東南山區的官民隘，向新竹東南山區推進的拓殖開發之情形，更可進一步探索其拓展情形是否受其他地理環境或政經社會條件所影響。若依文書種類進行檢索後分類，也可以看出不同文書種類在時間序列上所顯示的差異，尤其1886年（光緒12年）以後，福建臺灣巡撫劉銘傳推行清賦，裁廢墾隘，金廣福正式走入歷史，雖然還有與姜秀鑾相關的契約文書出現，但大概為與分割或轉手相關的種類，而非首墾契書了。[31]

5. 文件間統計型的脈絡——文件子集合中文字及數量的分析

系統如何幫助研究者，在文件中清理、建立出統計型的脈絡，THDL中的詞頻分析，以及「臺灣歷史數位圖書館工具集」中的前後綴詞分析工具是最好的例子。詞頻工具是指詞彙在檢索結果中出現的頻率，利用詞夾子程式，[32]清理、擷取出THDL檔案中的人名、地名和其他專有名詞，再統計出現的頻率，亦會使研究者在檢索之後，能從事更有效與簡單的再觀察對象。如朱一貴在不同年代的高峰，若再加入詞頻進行分析，便會發現和朱一貴最常出現的人名竟不是和事件直接關聯的藍廷珍、施世驃等人，反而是林爽文、常青等，在另一亂事中的人物，在經由對史料內文的解讀後，我們可以初步推論當乾隆在面對林爽文之亂時，其祖父康熙在平定朱一貴之亂的種種戰略與思考，成為他重要的參照點。可能正是因為林爽文事件的存在，使得朱一貴之亂重新被提出、詮釋，構成新的歷史記憶。這樣交插、跨研究對象的比對和討論，是單從紙本閱讀經驗中所難以察覺的。[33]

前後綴詞和詞頻工具相同，亦是應用詞夾子程式，製作出一個簡易的工具介面，讓使用者可以對THDL所藏的檔案文件，進行關鍵字的前綴詞的分析。[34]可以查詢所輸入的關鍵字，其左方（即關鍵字之前）或其右方（即關鍵字之後），所出現的字串，並統計其出現的篇數和次數。在檢索的過程除了可以指定字串的長度外，還可以指定全文需含有的詞彙。比如說我們可以輸入「銀」為欲分析的關鍵詞，觀察其前方三個字數的綴詞，藉以觀察明清檔案中和臺灣有關的銀圓種類，其結果為：佛銀（含佛面銀、佛銀、佛頭銀、番佛銀、佛番銀、清水銀）9,127件、番

31 參見蔡炯民、歐仲翔、翁稷安、項潔，2011，〈時空資料的地方再現：GIS與數位典藏的交會〉，《國土資訊系統通訊》，78期，頁13-20。

32 關於詞夾子程式的設計與應用，可參見謝育平，2010，〈同位詞夾子：主題式分類詞庫萃取演算法〉，《數位人文研究的新視野：基礎與想像》，頁133-162；謝育平、楊龍慶、趙建宏、黃銘立、古馮文、林郁智，2009，〈使用詞夾子建立中文典籍分析加值服務〉，發表於「銘傳大學2009資訊科技與實務研討會」；張尚斌，2006，《詞夾子演算法在專有名詞辨識上的應用——以歷史文件為例》，臺北：國立臺灣大學資訊工程學研究所碩士論文。

33 涂豐恩、杜協昌、陳詩沛、何浩洋、項潔，2010，〈當資訊科技碰到史料：臺灣歷史數位圖書館中的未解問題〉，《數位人文研究的新視野：基礎與想像》，頁21-44。

34 該工具網址為：http://thdl.ntu.edu.tw/SimpleTools/TermPat/TermPatSimpleUI.php。

銀866件、龍銀842件、劍銀210件、花邊銀（含花銀、灼銀）76件、紋銀301件、六八銀157件。可知主要皆為外國銀，分布如圖8。

6. 文件間隱藏的新脈絡——藏於語意（semantics）中的脈絡

發掘研究者所未察覺的新脈絡，可以說是系統最終極、最理想的功能，意即系統不單只是被動的觀察工具，而是能夠更主動的替使用者發現具深刻意義的研究議題。這樣新的脈絡往往藏於語意之中，單憑搜尋、檢索難以發現。同樣收錄在「臺灣歷史數位圖書館工具集」的土地移轉圖和引用關係圖兩樣工具就是最好的例證。土地移轉圖重新將古地契間的脈絡進行重建，[35]利用資訊科技完成這項用人力幾乎不可能完成的任務，清理出THDL中來自100種不同出處超過30,000件的地契文書，析理出土地的身世。它重建了土地的上下手契關係，利用資訊技術自動擷取契書的標題、人名、四至、土地面積、賣價、地號、時間等特徵，在兩兩比對後找出契書間的關聯。接著再把上述一對或一組的上下手契與鬮分契串連起來，就成為某一塊土地在不同地主手中的轉移圖。在三萬多件契約中，出現了2376個土地移轉圖（最大的有103份契約），每一個圖都反映著一塊土地的身世。如圖9這是一塊苗栗永和山地區的土地交易情形，圖中主角為率先開墾此地的廖姓家族。這個圖在社會史、經濟史和家族史所呈現的意義，有待研究。

THDL《明清臺灣行政檔案》引用關係圖利用明清行政檔案當中，臣子的奏事文書、與皇帝諭旨文書之間相互的引用關係，將數量龐大卻零散的行政檔案，連結成反映詳細政治決策過程的「引用關係圖」。[36]它打破了原有的檔案脈絡，即按檔案所屬部門的分類，如它是來自宮中檔，還是來自內閣大庫；或者，它是直通皇帝的奏摺，還是要經過六部的題本；改以奏摺與上諭間彼此間的引用和回應，去重建事件處理過程中官方的往返與運作。其反映了清帝國內訊息的傳遞和流通，以及政策形成的決策過程，更重要的，它也顯現了面對同一事件，不同地區、不同層級的官員，出於自身立場或利益考量的迥異看法，這些新的脈絡都有待研究者去進一步的考察。表1為引用關係所涉檔案最多之前二十件，及其大致內容。

35 土地移轉圖查詢系統網址為：http://140.112.30.230/GRAPH/INDEX/newindex.php。其數位化的方法和過程可參見黃于鳴，2009，《臺灣古地契關係自動重建之研究》，臺北：國立臺灣大學資訊工程學研究所碩士論文。對於古契書的數位整理可參見項潔、陳詩沛、杜協昌，〈台灣古契約文書全文資料庫的建置〉，2009，《第三屆台灣古文書與歷史研究學術研討會》，逢甲大學歷史與文物管理研究所、台灣古文書學會主辦，臺中市：逢甲大學；盧家慶，2008，《台灣古契書自動分類與依分類定義契書角色》，臺北：國立臺灣大學資訊工程學研究所碩士論文。

36 THDL《明清臺灣行政檔案》引用關係圖網址為：http://140.112.30.230/citation/list_mq_graph_8.php。

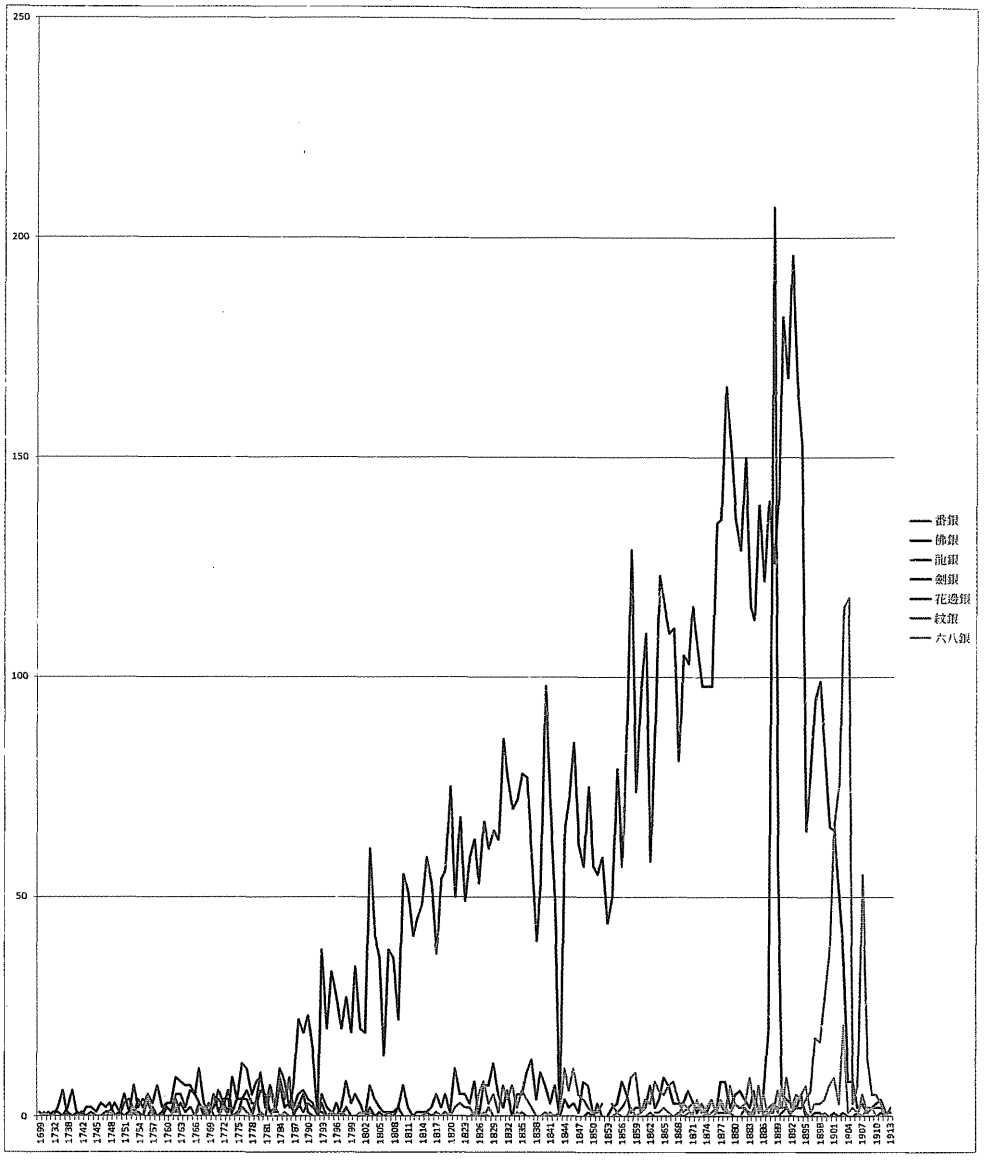


圖8 古契書中主要銀圓類別的年代分布

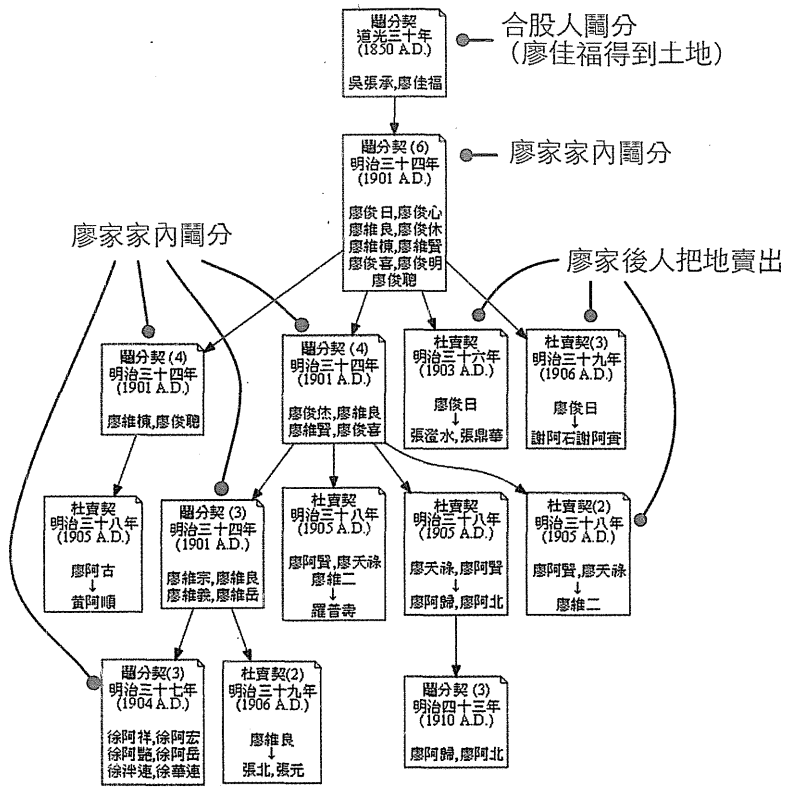


圖9-1　苗栗永和山地區土地交易關係圖

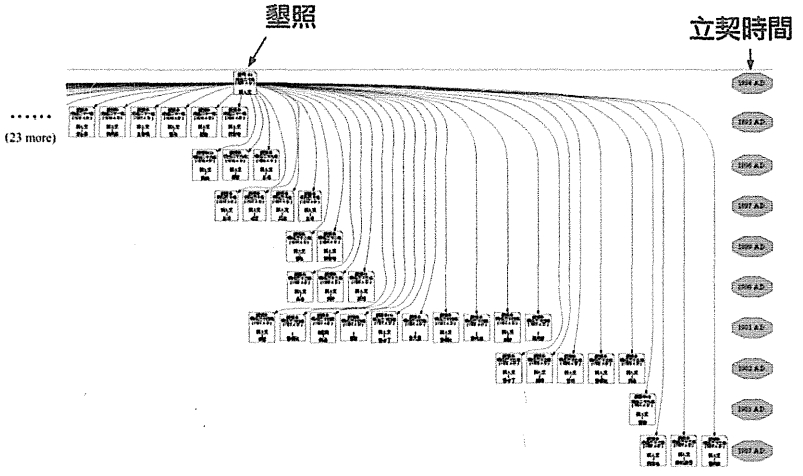


圖9-2　苗栗永和山地區土地交易圖

表1 引用關係所涉檔案最多之前二十件

引用關係圖（檔案數）	主要年代	內容
153, 126, 109	乾隆52年（1787）	林爽文事件產生的奏諭
107	乾隆53年（1788）	林爽文事件的奏諭
95	乾隆52年～同治6年（1787-1867）	林爽文事件引發清治臺灣的政策討論：乾隆下令「嗣後凡遇有台灣道員缺，俱自加按察使銜俾得自行奏事」，日後臺灣道一職補授必定引用此諭旨。
85	嘉慶11年（1806）	蔡牽事件產生的奏諭
83	同治13年（1874）	牡丹社事件產生的奏諭
62	光緒10～11年（1884-85）	中法戰爭期間法軍攻打臺灣的基隆與滬尾產生奏諭
47	同治9年～光緒1年（1870-75）	牡丹社事件後日本兵船盤據龜山一帶
45	乾隆51～52年（1786-87）	林爽文事件剛發生時的奏諭
45	道光21～22年（1841-42）	中英鴉片戰爭期間，兩艘英國船隻出現在淡水與大安，被臺灣道姚瑩與臺灣鎮總兵達洪阿逮捕。
43	光緒11～16年（1885-90）	劉銘傳上任臺灣巡撫後開山撫番及其事後請功的內容
42	光緒11～14年（1885-88）	清廷檢討海防事宜，臺灣身為「南洋門戶」因此也是重點地區之一，包括臺灣建省。
37	同治13年～光緒1年（1874-75）	牡丹社事件後，光緒皇帝點名多位官員提供海防籌備計畫。
35	道光12～13年（1832-33）	張丙事件
30	道光6年（1826）	彰化、大甲、中港、後壠械鬥事件
29	同治4年～光緒19年（1865-93）	臺灣人事升遷的商議文件
29	同治7～8年（1868-69）	同治年間的中外衝突
28	康熙60年～道光12年（1721-1832）	巡臺御史存廢的討論
28	乾隆52～56年（1787-91）	林爽文事件過後恢復事宜的報告

五、結語

傳統的檔案保存方式有其自身的邏輯，是一個經過長期運作，於實務之中不斷改善後所得的學問，是值得重視並尊重的學術典範。然而，當保存方式由紙本演變成數位，印刷文字被二進位位元所取代後，保存的邏輯必然要與時俱進，並非要完全否定傳統檔案保存重要性，而是要在尊重傳統核心的前提下，保留其核心價值，並賦與其數位的全新風貌，衝撞、生成全新的可能，呈現出數位化時代檔案保存的全新價值與脈絡。

檔案本身自然不會說話，必須要透過史家的詮釋才有可能彰顯其意義，紙本的檔案處理並不是完全沒有問題的，尤其是單一脈絡的編輯方式，很容易造成誤導和盲點。理論上，通過學者的勤勞與用心，在檔案間細細爬梳，仍然可以躲過編輯者的分類大刀，重新組合出自己的需求，但終究只是理論上而已，面對汗牛充棟的史料處理，對運氣的需要很可能勝過努力，更遑論過程中的費時費力。

數位化後檔案的新脈絡不但不是要否定檔案和史家之間的既有關聯，更是要加以拉近兩者的距離。檔案經過數位化處理後，不但能保存原有的收藏邏輯，更能賦與其不受篇幅和印刷限制的靈活和彈性；史家仍是史料最終的闡釋和解讀者，但通過資訊的處理，除了創造出一個令研究者更能夠觀察的環境，並析理出埋沒在史料中的新線索，使史家可以將心力投注在對史料的解讀。或許，當史料獲得和整理的問題被資訊技術解決之後，不再是困擾研究者的重擔，如何發展出鞭辟入裡的新論述，可能才是數位化時代的檔案脈絡所給予研究者的真正挑戰。總之，經由檔案的數位化，傳統檔案單一脈絡的限制被打破，檔案整理與檔案研究之間的矛盾亦隨之減輕，關鍵便在於建立一套以研究為主要思考取向的檔案檢索系統，透過檔案檢索系統的設計，檔案的豐富且多重的脈絡可以被發現、被呈現、被觀察。

西方重要史家勞倫斯．史東（Lawrence Stone），曾有一段對於史家使用檔案的鮮明描述，他說：

> 當你在檔案堆裡工作，你就遠離了舒適的家，你會很不耐煩，你焦躁不已，你會像發瘋似的塗塗寫寫。你總會犯錯，我不相信西方世界有任何一個學者做註可以做得十全十美。檔案研究是一項煩人的生計。[37]

我們衷心希望，在數位化的檔案時代，史東所感受到的焦躁可以一掃而空，使用檔案可以變成一個令人感到愉快和刺激的過程，並大大的提升對檔案引用的正確率。就如同愛德華．卡爾那句經典名言：「歷史是歷史學家跟他的事實之間，不斷

37 轉引自Richard J. Evans著，潘派泰譯，2002，《為史學辯護》，頁140。

交互作用的過程，是現在跟過去之間，永無止境的對話！」[38]我們相信數位化工程，將會在史家和史實的交互作用中扮演起重要的對話，它將使史家能更為貼近、理解他所面對的史料，進而開展出一場深刻而愉快的對話。

38 轉引自 Edward H. Carr 著，江政寬譯，2009，《何謂歷史？》，頁 126。

參考文獻

Edward H. Carr著，江政寬譯，2009，《何謂歷史？》，臺北：博雅書屋。

Richard J. Evans著，潘派泰譯，2002，《為史學辯護》，臺北：巨流出版社。

Vanda Broughton. (2006). The need for a faceted classification as the basis of all methods of information retrieval. *Aslib Proceedings*, Vol. 58 Issue1/2. pp. 49-72.

汪榮祖，2002，《史學九章》，臺北：麥田出版社。

林文凱，2004，〈清代晚期臺灣的土地法律文化——淡新檔案內淡新地域漢墾莊在十九世紀四十年代以來土地抗租案件為主的分析〉，發表於臺灣社會學會等主辦「2004臺灣社會學會年會暨『走過臺灣——世代、歷史、與社會』研討會」。

洪麗完，2007，〈評論稿〉，收於逢甲大學歷史與文物管理研究所、台灣古文書學會編校，《第三屆台灣古文書與歷史研究學術研討會論文集》，頁331，臺中：逢甲大學出版社。

徐紹敏主編，2001，《檔案文獻編纂學（第二版）》，頁8-9，杭州：浙江大學出版社。

涂豐恩、杜協昌、陳詩沛、何浩洋、項潔，2010，〈當資訊科技碰到史料：臺灣歷史數位圖書館中的未解問題〉，收於項潔編《數位人文研究的新視野：基礎與想像》，頁21-44，臺北：國立臺灣大學出版中心。

秦國經，2005，《明清檔案學》，北京：學苑出版社。

張尚斌，2006，〈詞夾子演算法在專有名詞辨識上的應用——以歷史文件為例〉，國立臺灣大學資訊工程學研究所碩士論文。

莊吉發，1983，《故宮檔案述要》，臺北：國立故宮博物院。

莊樹華，2010，〈數位檔案運用的契機與風險——以近代中國外交檔案為例〉，發表於國立政治大學圖書資訊與檔案學研究所主辦「數位檔案加值與教學應用研討會」。

莊樹華、龐桂芬，2006，〈數位化時代下的檔案館營運——以中研院近史所檔案館為例〉，《檔案季刊》，第5卷4期，頁23、26。

陳智超，1995，《解開《宋會要》之謎》，北京：社會科學文獻出版社。

陳嵩明，2010，《清代臺灣界外武力拓墾：噶瑪蘭、水沙連、與竹塹東南山區之比較研究》，國立臺北大學社會學系碩士論文。

項潔、涂豐恩，2011〈導論——什麼是數位人文〉，收於項潔編《從保存到創造：開啟數位人文研究》，頁9-28，臺北：國立臺灣大學出版中心。

項潔、陳詩沛、杜協昌，2009，〈台灣古契約文書全文資料庫的建置〉，發表於逢甲大學歷史與文物管理研究所、台灣古文書學會主辦「第三屆台灣古文書與歷史研究學術研討會」。

項潔、董家兒、蕭屹灵，2008，〈臺灣省議會檔案數位典藏系統之研發與建置〉，《臺灣省諮議會會訊》，16期，頁33-46。

項潔、蕭屹灵、董家兒，2009，〈日治法院檔案數位典藏系統之研發與建置〉，收於王泰升主編《跨界的日治法院檔案研究》，頁83-148，臺北：元照出版社。

黃于鳴，2009，《臺灣古地契關係自動重建之研究》，國立臺灣大學資訊工程學研究所碩士論文。

歐仲翔，2011，《使用者取向之歷史地理資訊系統：古契書與統計資料呈現》，國立臺灣大學資訊工程學研究所碩士論文。

蔡炯民、歐仲翔、翁稷安、項潔，2011，〈時空資料的地方再現：GIS與數位典藏的交會〉，《國土資訊系統通訊》，78期，頁13-20。

鄭喜夫纂輯，1980，《臺灣地理及歷史．官師志》，臺北：臺灣省文獻委員會。

盧家慶，2008，《台灣古契書自動分類與依分類定義契書角色》，國立臺灣大學資訊工程學研究所碩士論文。

戴桓青，2008，〈四代百年臺大情〉，《臺大校友雙月刊》第60期，頁24-25。

薛理桂，1998，《檔案學導論》，臺北：漢美出版社。

謝育平，2010，〈同位詞夾子：主題式分類詞庫萃取演算法〉，收於項潔編《數位人文研究的新視野：基礎與想像》，頁133-162，臺北：國立臺灣大學出版中心。

謝育平、楊龍廉、趙建宏、黃銘立、古馮文、林郁智，2009，〈使用詞夾子建立中文典籍分析加值服務〉，發表於「銘傳大學2009資訊科技與實務研討會」。

自然語言處理技術於中文史學文獻分析之初步應用

劉昭麟*、金觀濤**、劉青峰***、邱偉雲****、姚育松*****

摘 要

自然語言處理是計算機科學中具有相當歷史的學科，過去主要應用於分析與處理現代文字語料。文字作為人類溝通與記錄的主要工具，詞意與語法都與時俱進。因此，處理現代文字語料的計算技術，不見得可以立即應用於歷史語料的處理工作。

本文以「中國近現代思想及文學史數據庫」為例，實驗如何利用自然語言處理技術輔助史學研究。我們利用PAT Tree技術從大量史料中，透過專家的協助來擷取與史學研究相關的詞彙，進一步分析詞彙的語境與共現的現象，最終估計個別文件與研究議題相關度，希望藉此輔助學者以比較有效率的方式，覓得相關的史學文件和分析文件內容。

本文所討論的語文分析技術已經實際應用於兩個研究工作。研究者應用計算技術分析了西元1905年到1911年間的《清末籌備立憲檔案史料》，成果發表於〈社會行動的數位人文研究：以清末預備立憲為例〉（金觀濤等，2011）；同時也探索了西元1875年到1911年間的《清季外交史料》，成果同時發表於另一專文〈「共現」詞頻分析及其運用：以「華人」觀念起源為例〉（金觀濤等，2011）。

自然語言處理技術固然不能完全取代史學研究者從事史學研究，但是初步經驗顯示，自然語言處理技術有足夠的潛力為史學研究者提供初步分析的服務，讓史學研究者可以比較有效率的方式處理大量的語料，並且把珍貴的研究時間用於知識層次的分析工作。

* 國立政治大學資訊科學系教授。
** 國立政治大學講座教授。
*** 香港中文大學中國文化研究所名譽研究員，《二十一世紀》創刊編輯。
**** 國立政治大學中國文學系博士班研究生。
***** 國立政治大學歷史學系碩士班研究生。

An Exploration of Analyzing Historical Chinese Documents with Natural Language Processing Techniques

Chao-Lin Liu *, Guantao Jin **, Qingfeng Liu ***,
Wei-yun Chiu ****, Yih-soong Yu *****

Abstract

Natural language processing (NLP) is a well-known research area in computer science, and has been successfully applied to handle and analyze modern text material in the past years. Whether we can extend the applications of current NLP techniques to historical Chinese text is a challenge. Word senses and grammars changed over time, when people of different times assigned different meanings to the same symbols and word patterns and when they used different word patterns.

We explored the applications of NLP techniques to support the study of historical research, based on the text material available at the Database for the Study of Modern Chinese Thoughts and Literature. In recent attempts, we applied the PAT Tree method to extract useful Chinese words from the corpora, with the help of historians to finalize the keyword selection. We also analyzed the occurrences of the keywords and the collocations of the keywords over the years of interest, in order to find a way to rank the historical documents so that historians may find key documents and identify the key sentences more effectively.

In this paper, we report how we employed NLP techniques to support two historical studies. The first is about how the Qing government attempted to convert itself from a monarchy to a constitutional monarchy between 1905 and 1911, using the documents

* Professor, Department of Computer Science, National Chengchi University.
** Chair Professor, National Chengchi University.
*** Research Fellow (Honorary) of the Institute of Chinese Studies, Chinese University of Hong Kong (CUHK). Founding Editor of *the Twenty-First Century* in Hong Kong.
**** Ph.D. Student, Department of Chinese Literature, National Chengchi University.
***** M.A. Student, Department of History, National Chengchi University.

recorded in 清末籌備立憲檔案史料. The second issue is about how the attitude of the Qing government towards the overseas Chinese workers during the late 19th century and the early 20th century, using the documents recorded in 清季外交史料. Details about these historical researches are reported in two other papers in this conference (金觀濤等，2011).

No one may expect that NLP techniques will replace the major roles of historians in the historical studies, but the techniques should be able to work with the historians to make the studies more efficiently and more effectively. Preliminary results reported in this paper and other papers in this conference have suggested this potential of NLP techniques. With the help of computing technologies, historians can delegate some search work to computers and spend more time on higher level thinking than before.

一、引言

自然語言是資訊科學中用以泛指自然界中生物所直接、間接使用的語言，大多時候用以代表人類所用的語言，例如中文、日文與英文。自然語言處理技術，則是指資訊科學研究者利用計算機的軟硬體設備處理自然語言的相關技術。由於近十幾年來計算機硬體生產技術的突飛猛進，人們可以極低的價格獲得高速運算能量與大量儲存空間，近幾年來許多軟體科技也隨著有很大的進步。除了計算機的軟硬體設備之外，自然語言處理技術的研究者研究的是語料，當然是要有適當的語料才能讓技術派上用場。近二十年來網際網路的迅速普及，使得基礎的語文資料的獲得變得十分容易，也使得自然語言處理的研究範圍迅速擴增，不再像過去只能侷限於現代且學術性的文字資料，而能夠擴及多種層面的語文資料。

自然語言處理技術雖然尚且未臻完美，但是已經在許多領域展現相當成果與潛力。自然語言處理技術是在網際網路上提供資訊檢索服務的重要基礎，透過適當技術處理網頁資料的內容與建立索引資料，搜尋引擎讓人們得以在前所未有的便利環境，透過網際網路找尋有用的資訊。自然語言處理技術也讓我們可以處理特定領域的資料，研究者利用語文對話模型，可以找尋心理疾病的相關文獻資料（Yu et al., 2009）。同樣的，基於相關的財務管理知識，搭配自然語言處理技術，研究者也可以深入分析財務報表的內容（Chen et al., 2011）。近年來，學者也深入研究表面文字對於焦點議題所意涵的正面意見與負面意見（古倫維，2009）。

學者也應用自然語言處理技術來分析非現代漢語之文字資料，例如，早在1997年即有學者建立《紅樓夢》的資料中心，提供學者研究該書的內容（羅鳳珠等，1997）。本研究群於近年則致力於蒐集中國近現代的思想史與文學史相關文獻，以人工將文獻資料轉換為數據資料，作為研究近現代史之基石（金觀濤、劉青峰，2011）。

透過該數據庫的內容，我們可以分析詞彙來審視歷史事件的發展。圖1的折線圖顯示《清末籌備立憲檔案史料》（故宮博物院，1979）的關鍵詞彙分析。透過計算機軟體的協助，我們可以計算該史料中的總字數與各年份文本的年份總字數。然後以各年份的總字數除以全部史料的總字數，以得到年份比例。以c_{1905}、c_{1906}、c_{1907}、c_{1908}、c_{1909}、c_{1910}和c_{1911}分別代表《清末籌備立憲檔案史料》在西元1905、1906、1907、1908、1909、1910和1911年的資料的字數，我們再將各年份的字數加總得到文獻的總字數t（如下方公式所示）。

$$t = c_{1905} + c_{1906} + c_{1907} + c_{1908} + c_{1909} + c_{1910} + c_{1911}$$

再將c_{1905}、c_{1906}、c_{1907}、c_{1908}、c_{1909}、c_{1910}和c_{1911}分別除以t，這樣就可以年份i作

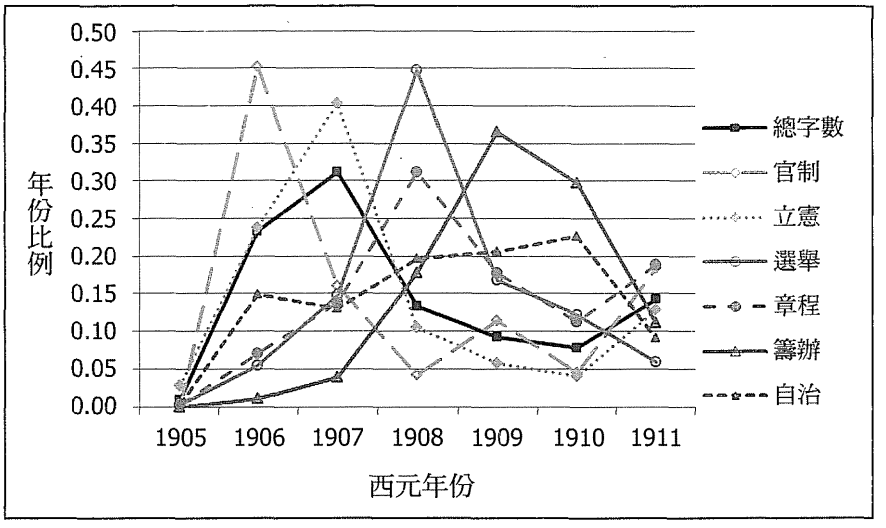


圖1 《清末籌備立憲檔案史料》的關鍵詞彙分年分析（僅示部分資料）

為橫軸、$\frac{c_i}{t}$作為縱軸，以i為1905、1906、1907、1908、1909、1910和1911年來繪製圖1中關於「總字數」的折線圖。其他6個關鍵詞彙的折線圖，也是以同樣的方法獲得。例如，我們從史料中計算每一年「官制」出現的次數，並且計算總次數，然後把每一年份的出現次數除以總次數，即可以繪製圖中關於「官制」的折線圖。

如果個別詞彙的某一年度比例高於總字數的同一年度比例甚多，則意味著該年度有一些事件與該詞彙有相當密切的關係。這樣的分析角度，可以讓我們看到個別關鍵詞彙最重要的年份，而比較不會礙於個別關鍵詞彙出現的總次數。如果有一個關鍵詞彙幾乎集中在某一、兩個年份出現，似乎就暗示著與該詞彙語意相關的事件在那一、兩年的重要性。以圖1為例，「官制」與「立憲」出現的高峰分別是西元1906年和1907年 ，並且這兩詞彙的年份出現比例明顯高於所有文獻資料總字數的年份比例 ，暗示著這兩年是討論立憲與否議題的重要時期。「選舉」與「章程」在1908年的高峰則暗示著立憲行動的濫觴，而「籌備」與「自治」在1909與1910兩年的高峰則意味著立憲行動的逐步落實 。

關鍵詞彙出現頻率和比例的變化，固然可以用以輔助分析事件的發展，但是有時單一關鍵詞彙並不一定能夠反映事件的全部資訊。如果可以考慮更多的語境資訊，則可能可以更加精確的掌握事件的面貌。以「君主立憲」為例，分析「君主立憲」出現的頻率，固然反映不同時期的文獻中討論相關議題的熱烈程度，但是沒能細分反對者與贊成者的消長。因此，考慮相關關鍵詞彙一起出現的頻率，是精確化詞彙語意的重要方法。

一起出現的關鍵詞彙的重要性在自然語言處理技術中早已經為人所熟知（例如，Manning & Schütze, 1999），collocation所代表的技術就是討論何謂詞彙共同出現和討論如何計算這類事件的頻率的相關議題。數位人文學者當然也有類似的觀察，Wang和Inaba（2009）所稱的co-word其實也就是自然語言處理所稱的collocation。以下我們將以「共現詞組」來指涉兩個一起出現的關鍵詞彙。舉例來說，分析「籌備」與「自治」在30個漢字[1]的範圍內一起出現的次數，就是對於｛籌備、自治｝[2]這一共現詞組的分析。

這一篇論文的主要內容乃關於我們如何以自然語言處理技術，分析「中國近現代思想及文學史專業數據庫」中的《清季外交史料》（王彥威、王亮，1934）和《清末籌備立憲檔案史料》的關鍵詞彙和共現詞組的出現頻率，並且進而提供深入分析史料文獻時的有用建議。在這裡我們著重於所應用的計算技術，而這一些計算技術對於人文研究的功能與相關之史學觀察則分述於〈「共現」詞頻分析及其運用：以「華人」觀念起源為例〉（金觀濤等，2011）和〈社會行動的數位人文研究：以清末預備立憲為例〉（金觀濤等，2011）兩篇論文。

二、文獻相關數據、關鍵詞彙與齊夫定律

本文所分析的資料有兩筆，兩者均來自現正在國立政治大學與香港中文大學建構之「中國近現代思想及文學史專業數據庫」。《清季外交史料》擷取西元1875年

---

1 以30個漢字作為定義共同出現的基礎，是一個任意的選擇。在資訊科學的文獻中，這一字數所框出的文字範圍稱為「共現窗口」（collocation window）。以關鍵詞為中心，考慮關鍵詞前後各30個漢字作為共現窗口，因此我們的共現窗口的大小（collocation window size）是60。研究者在運用計算機軟體分析語文資料的共現資料時，可以自行訂定這一數量的大小。如果採用數量比較少的漢字，則兩個詞彙就會比較不容易被認定為共同出現，因此那將是比較保守的選擇。在一些初步的實驗中，我們曾經嘗試不同數量的漢字，不同的選擇當然影響了所選擇的部分共現詞組的排序，但是一些主要的共現詞組仍然經常出現。經過初步的比較，史學研究者選擇採用60個漢字的共現窗口。
本文評審對於共現窗口的選擇特別關注，有下列重要評述，對於窗口大小的選擇具有重要意義，因此引用如下：

> ……但是作者亦提及有其他研究者已經標注標點符號，因此若是研究者可以掌握更具有訊息的語料文本，應該善用這樣的訊息。通常視窗概念用於具有結構的單位，如句子、段落、文章，不同的研究工作會使用不同的單位，以跨語詞典的建構為例，通常使用句子為單位，以作者另一篇討論「華人」的論文為例，則是以段落為單位較為適宜。……例如，若是決定以句子為單位，但是手中的文本並沒有標點符號，便可以人工標注少量的語料，估算該文本句子的平均長度，以此作為視窗的大小。

本文作者完全同意評審對於窗口大小的選擇的評述，窗口大小的選擇應該從文字資料的語意出發，如果語意資訊無法明確，則可以考慮語法訊息，例如句子。而中文句子又不盡然以一般所認知的句號作為終結，因此這一問題牽涉到中文語意的分析問題。在古文斷句的計算技術方面，可以參考《以序列標記方法解決古漢語斷句問題》（黃瀚萱，2008）。

2 以下我們將以大括號以及其內部以頓號分隔的兩個關鍵詞彙，來表示由這兩個關鍵詞彙所構成的共現詞組。

表1　相關文獻之基本統計數字

文獻	準詞彙個數	總字數	相異字數	文件份數
《清末籌備立憲檔案史料》	3,288	713,131	4,097	399
《清季外交史料》	29,315	2,875,032	5,225	5,758
《民報》	7,784	1,450,623	6,230	325
《海國圖志》	2,649	679,410	4,916	160
《新民叢報》	33,378	5,259,590	6,647	1,524

到1909年之間的文件，包含5,758份檔案；不包含檔案標題、作者等基本資訊，僅文件內容合計2,900,938個漢字。這一份資料本身尚且沒有標點符號，如果再去掉段落前後之空白符號，則總字數減為2,875,032字。《清末籌備立憲檔案史料》出版時間介於1905年到1911年之間，包含399份檔案；不包含檔案標題、作者等基本資訊，僅文件內容部分合計720,498個漢字、空白和標點符號。如果去掉段落前後之空白符號，則總字數減為713,131字。

在2011年第三屆數位典藏與數位人文國際研討會」中，本研究群主要提報關於《清末籌備立憲檔案史料》的研究與《清季外交史料》中的華工相關問題的研究。除了這兩筆文獻資料之外，我們也分析了《民報》、《海國圖志》與《新民叢報》的內容，這幾批文獻的基本統計數字請參見表1。[3]

（一）關鍵詞彙

中文分析的首要工作通常是中文斷詞。[4]在分析現代中文資料時，絕大多數的研究者都是依賴電子詞典並且搭配其他計算技術來進行這一工作。不過中文斷詞迄今還無法作得盡善盡美。以「教廷駐臺大使館」為例，這一連續的字串可以解讀為「教廷　駐臺　大使館」或者「教廷　駐臺大　使館」。實際上，前者才是正確的斷詞結果，但是在文法上後者並無不可，例如我們可能說「教廷　駐高雄　使館」。後者之不正確，是基於事實之不成立；而除非計算機中有相關的知識否則難以判斷真假。例如「教廷　駐臺北　使館」就可為「教廷駐臺北使館」之正確斷詞。不同的斷詞結果，將影響計算機軟體如何回報「臺大」這一詞彙的頻率。因此斷詞的正

3　表1的數據全部是基於「中國近現代思想及文學史數據庫」的資料中目前已經收錄的資料所得的統計結果。不同版本的文獻，可能有不同數量的字數，例如吳通福博士所標註的《清季外交史料》超過四百萬字（包含個別文件標題、附件與標點符號）。又如《海國圖志》也有內容不盡相同的版本。

4　或稱「中文分詞」，亦或「中文切詞」。

確與否，將影響我們計算文獻中相關詞彙的詞頻的可靠度。

在處理現代漢語資料時，我們可以找到關於現代漢語的電子詞典，可是現在尚且沒有適用於「中國近現代思想及文學史專業數據庫」中文獻資料的電子詞典。因此，在處理非現代中文資料時，我們所面對的問題還不只是中文斷詞的正確與否，而是目前沒有立即可用的中文斷詞器。

缺乏對於所欲分析的文獻的時代的詞典，不只是使得中文斷詞難以全面自動化，同時也意味著在從事研究工作時，我們不容易確定該透過哪一些關鍵詞彙來探索研究議題。如果我們不確定與研究議題相關的關鍵詞彙，也就不容易利用計算機軟體技術來協助我們的研究工作，就連尋找相關參考文獻都變得不容易。

缺乏關鍵詞彙的問題對於專業的研究者或許不是明顯的問題，專業研究者本應該對於所研究之領域具備相當知識，所以總該可以列舉一些重要的關鍵詞彙。然而，既然有計算機軟體的協助，我們是否有機會找到更多的關鍵詞彙？或者從另一個角度來看，即使後代史學者能夠推想當年的史學家「應該」是用哪一些詞彙來討論某一些議題，但是這樣的猜測是否確實成立，卻也可能是需要再多作驗證的。

為了確立研究議題的關鍵詞彙詞集，我們利用計算機軟體技術中的PAT Tree（Chien, 1999）來記錄與分析文獻中任意字串的頻率。

這一方法是記錄出現在文獻中所有可能的字串的頻率，當我們的語料量足夠大時，真實有用的詞彙的頻率自然會比無意義字串的頻率要高。以尋找兩個漢字所構成的二字詞為例，在處理「數位人文研討會」這一假想文本資料時，我們會記錄文獻中出現一次的「數位」、「位人」、「人文」、「文研」、「研討」和「討會」。如果再處理「人文風氣」這一假想資料則會記錄一次的「人文」、「文風」和「風氣」。這就會使得「人文」的總次數變為2。我們寄望這樣的程序，在處理數量夠多的語料之後，透過累積的效應，有用的詞彙的頻率會自然的比無意義字串的頻率高。

這樣的程序可以用於尋找由三個字所構成的三字詞，以及更多字所構成的詞彙。因此在處理「數位人文研討會」時，我們的程式也會記錄出現一次的「數位人」、「位人文」、「人文研」、「文研討」和「研討會」；其中只有「數位人」和「研討會」比較可能常出現在現代的中文文字資料之中。

記錄和分析了任意字串的詞頻之後，我們可以請專家以其專業知識找出實際上有利於歷史分析的關鍵詞彙。由於上述方法實際上是漫無目的的計算所有字串的詞頻，甚至也包含內含標點符號的字串，所以我們會找到數量極為龐大的字串。以下，我們以準詞彙來稱呼這一些以程式找到的高頻率字串。

以《清季外交史料》為例，即使我們忽略出現次數偏少的字串，只考慮出現次數超過3次的字串，也得要過濾超過七百五十六萬個準詞彙。實際上我們無法請任

何專家從這樣大量的準詞彙中篩選有用的詞彙。因此，如果是要透過人工過濾關鍵詞彙的話，篩選的門檻通常會因準詞彙的數量與參與的專家人力而定。除非我們能夠以人力檢驗所有字串，否則設定任何預先篩選的門檻值，都預示了我們可能遺漏一些有用的罕用詞彙的風險。表1所列的「準詞彙」的認定標準是在個別的文獻中出現次數不低於10次的字串。[5]

除了借重專家的專業知識來過濾詞彙之外，我們也可利用一些經驗法則來自動篩選重要或者有用的詞彙。在本研究群另一篇專門討論清末立憲問題的論文中（金觀濤等，2011），我們探討如何利用齊夫定律（Zipf, 1949）來篩選詞彙。結果發現與Luhn（1958）所提出的詞彙篩選程序類似的篩選原則。在〈社會行動的數位人文研究：以清末預備立憲為例〉（金觀濤等，2011）這一篇論文中，我們先以人工篩選有用的詞彙。另外再模仿檢驗齊夫定律的一些作法，也就是將詞彙依照出現的詞彙排名（詞頻高者排名編號較低），然後把詞彙排名乘上自己的詞頻所得的乘積作為繪圖的縱軸，橫軸則是詞彙的排名，依此我們可以得到一條曲線。以主觀的判斷將這一曲線分成三個區段：分別對應到詞彙詞頻的高段、中段和低段。然後，依據這三個區段的詞彙分別進行本文後面節次所述的文件推薦工作；同時我們也以人工所篩選的所有詞彙進行文件推薦工作。最後將依據三個區段的詞彙所推薦的文件與依據全部人工篩選的詞彙所推薦的文件相比較，檢驗推薦內容的異同；結果發現使用中段的詞彙與使用全部詞彙時的推薦文件最相吻合。這符合Luhn所推測的經驗法則：最高頻和最低頻區域的詞彙都不是用來檢索文件的最好選擇。這一經驗提供了一個全自動選擇關鍵詞彙的契機。

（二）齊夫定律的一些觀察

以文字的出現頻率來說，齊夫定律指出，如果我們將某大量文獻中的詞彙依照詞頻予以排序，讓最高頻的詞彙為序號一號的詞彙，第二高頻者序號為二，則在這一文獻中所出現的詞彙的頻率將與其序號成一個反比的關係。齊夫定律雖然是由語言學家所提出，但是這一個定律所描述的現象，可以在許多不同的情境得到支持（Adamic & Huberman, 2002）。例如，如果以全世界的城市的人口來給城市排序，則城市人口的多寡與其排序的序號也遵循齊夫定律（Ioannides & Overman, 2003）。

以f代表詞彙在某一批文獻中出現的頻率（frequency），以r代表該詞彙排序的序號（rank），則齊夫定律可以用下列的公式表示：

$$\text{公式（1）} \quad f \propto \frac{1}{r}$$

5 在簡立峰的程式中，我們將minfreq、freqthd、kindthd和kindfreqthd分別設定為10、0.5、10和1。

因此，依據這一定律，f 和 r 的乘積是某一個常數（constant）。令這一個常數為 c，則公式（1）可以改寫為 $f \times r = c$。如果我們把這一個等式的兩邊同時取對數（logarithm），則這一個等式就變成下列公式：

$$\text{公式（2）} \qquad \log(f) + \log(r) = \log(c)$$

雖然我們可以直接以 f 和 r 的關係繪製圖形，來觀察個別文獻的詞彙是否遵循齊夫定律，但是個別文獻的詞彙分布經常有相當大的差異，所以圖形容易擠在一個小小區域，不容易在同一張圖上面直接繪製容易閱讀的 f 和 r 關係圖。如果改以 log（f）和 log（r）分別作為縱軸和橫軸，則兩者的關係會呈現一種約略的線性關係，圖形會比較清楚。我們將會在下面的繪圖工作中用到公式（2）。

我們可以利用上一小節所提到的PAT Tree演算法，為表1所列的五批文獻資料中，分析出現超過10次的準詞彙是否符合齊夫定律。由於表1所列的文獻合計包含超過一千萬個字，且出現次數超過10次的任意字串也超過七萬六千個，所以我們並沒有全部逐一去檢驗這一些詞彙是否屬於有意義的詞彙。為了圖形的可讀性，我們是以準詞彙的頻率的對數作為縱軸，以準詞彙的序號的對數作為橫軸作圖。因為取過了對數的關係，依照公式（2）的說明，齊夫定律預期橫軸與縱軸數據的關係是由左上向右下走的直線關係。

圖2是以表1中五份文獻的資料的準詞彙所繪製的圖形。我們分別以「立憲史料」和「外交史料」代表《清末籌備立憲檔案史料》和《清季外交史料》。我們可以看到，雖然個別文獻，可能因為相異字數和總字數而造成一些可見的差異，但是不同文獻幾乎都遵循了齊夫定律所預測的直線關係。[6]因為序號愈低者是出現頻率較高的詞彙，所以圖中的曲線都是由左上往右下的走向。因為我們所取的高頻率詞彙中最低頻的詞彙出現10次，而且圖中縱軸都是以10為底數，所以縱軸的最小值就是1。如果我們允許頻率低於10次的詞彙也是高頻率詞彙的話，圖中的縱軸的最小值就會小於1。

圖中五條曲線幾乎彼此上下堆疊，由上而下的順序正好對應到個別文獻的總字數的順序（參閱表1的統計數字）。雖然總字數的差異並不保證這一些曲線絕對不會交叉，但是以大趨勢而言，總字數愈多，則同一序號的詞彙出現的次數就容易偏高，因此針對不同文獻中同一序號的詞彙的詞頻高低也大致隨著總字數的提高而變

6 很多準詞彙實為半詞彙，它是有意義的，只是這種意義並沒有固化，這和其他語種完全不同。準詞彙符合齊夫定律，是一件有趣的發現。一般來說，以文字詞頻來進行齊夫定律的相關分析時，應該是要採取有意義的詞彙，不應該採用沒有意義的詞彙。這也是我們在上一小節提及金觀濤等（2011）的工作中，先以人工篩選高頻的詞彙，才以所得的詞彙集來分析文件的主要原因。

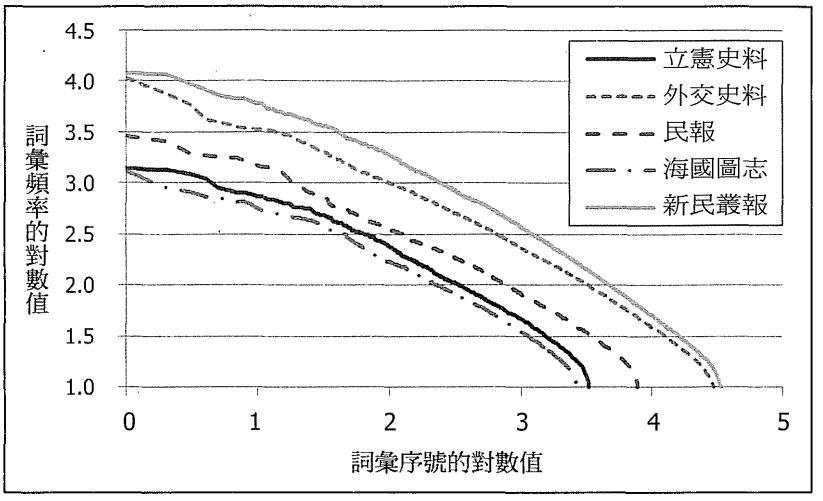


圖2 文獻中準詞彙遵循齊夫定律

高，所以我們會看到圖2中這樣相當整齊的排列。

為了消弭文獻總字數對於圖2中個別曲線的影響，我們可以把詞彙頻率先行除以研究中的文獻的總字數，之後再取其對數值作為縱軸。以N代表某一文獻的總字數，我們改以 $\log \frac{f}{N}$ 作為縱軸，但是仍然維持以$\log(r)$作為橫軸來作圖。這裡要注意的是，N是表1中所記錄的文獻總字數，所以不是一個常數。由於任何詞彙的頻率必然小於詞彙所屬文獻的總字數，所以 $\log \frac{f}{N}$ 是小於零的負數。依照這樣的作法，我們可以為表1所列的五份文獻作圖，得到圖3的折線圖。這一個圖形更加突顯了目前所觀察的五份文獻的準詞彙幾乎是以一樣的模式遵循齊夫定律。

圖2所繪製的是沒有經過專業篩選的準詞彙的趨勢。我們進一步從《清末籌備立憲檔案史料》和《清季外交史料》所篩選出來的準詞彙中挑選關於所欲研究的立憲與華工議題相關的詞彙，然後以這一些詞彙的詞頻以及它們在這一些詞彙集裡面的排序來作圖，則可以得到圖4。我們從《清末籌備立憲檔案史料》挑選了539個關於立憲議題的詞彙，從《清季外交史料》挑選了175個關於華工議題的詞彙。為了方便對照，除了為這一些挑選過的詞彙作圖之外，我們複製了圖2中《清末籌備立憲檔案史料》和《清季外交史料》的圖形到圖4之中。

在圖4之中，我們分別以「立憲（選）」和「華工（選）」標示從《清末籌備立憲檔案史料》和《清季外交史料》所篩選出來的詞彙。這一些經過人工挑選的詞彙的曲線的趨勢雖然不再像圖2中的五條曲線那樣的符合理想狀況，但是仍然大致遵

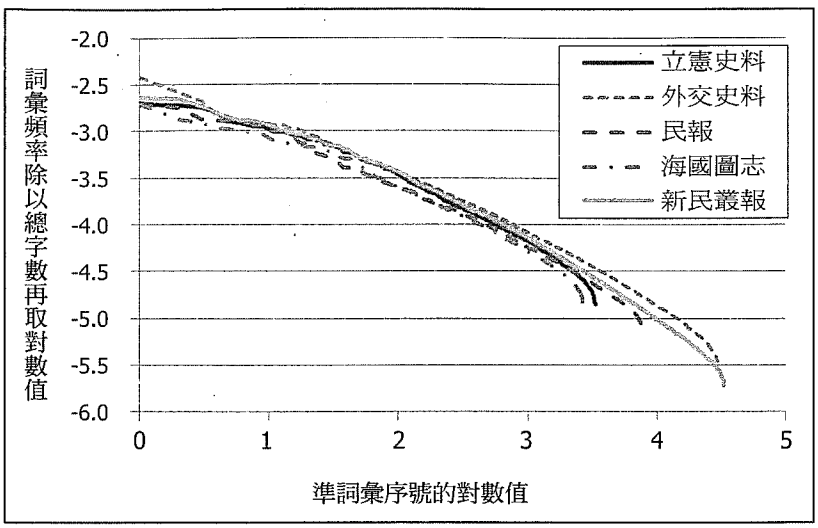


圖3 文獻中準詞彙遵循齊夫定律（經文獻總字數之調整）

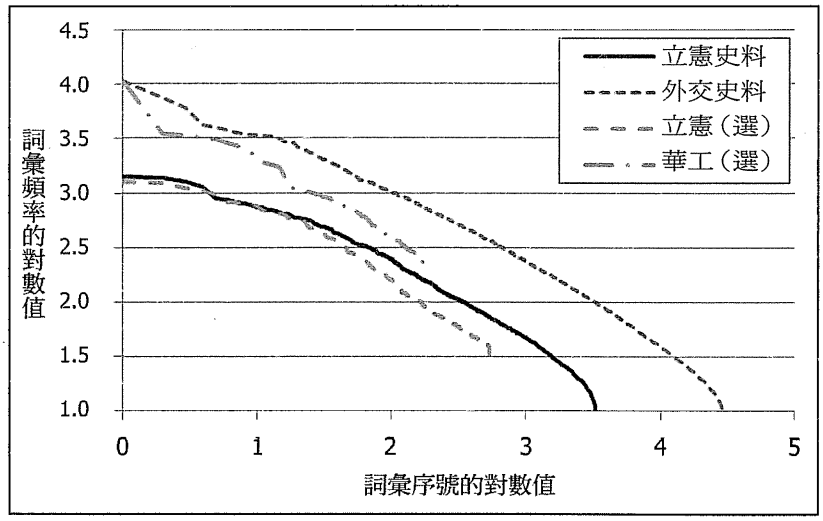


圖4 篩選過的詞彙的詞頻與序號也相當接近齊夫定律的預測

循了齊夫定律的預測。「立憲（選）」不過是由五百多個詞彙的數據所繪製，而「華工（選）」則更只是由不到兩百個詞彙的數據所繪製；如果我們挑選更多詞彙來繪圖的話，則曲線走勢將更接近齊夫定律的預測。目前圖中的曲線所呈現的趨勢，已經說明了這一些非現代的中文也相當遵守齊夫定律的預測。

三、關鍵詞彙之分年詞頻分析

確定關鍵詞彙詞集之後，我們可以據以計算這一些關鍵詞彙在各年份文件中的出現次數。得到每一關鍵詞彙個別年份的出現次數，也就可以計算該關鍵詞彙在所研究之文獻中出現的總次數。

我們計算關鍵詞頻率的方法，只是去尋找文獻之中有沒有相同的連續字串。如果有的話就會記錄該關鍵詞彙出現一次。我們需要提醒的是，基於第二節中關於中文斷詞的說明，我們現在計算詞彙的詞頻並不絕對精準。例如，即使「人文風氣」與「文風鼎盛」中的「文風」並不相同，但是在計算「文風」的詞頻時，我們的程式會認定「人文風氣」與「文風鼎盛」都包含了「文風」這一關鍵詞彙。儘管如此，經驗顯示這樣的問題雖然使得我們所呈報的數據不是絕對完美，但是當文獻資料量夠大時，程式所提供的數據仍具備研究上的價值。

得到關鍵詞彙在各年份資料的出現次數與全文獻的總出現次數，我們就可以把各年份出現的次數除以出現總次數，得到關鍵詞彙在各年份出現的比例。這就是我們得到圖1中各個關鍵詞彙的年份比例的方法。

在獲得各關鍵詞彙出現的年份比例之後，我們試圖找出各年份文本中的重要關鍵詞彙。關鍵詞彙是否重要，最可靠的判斷者應當是專家學者。但是計算技術是否能夠提供一些幫助呢？在分析研究文獻時，我們可以計算各年份文本的字數，進而計算整個文獻的總字數。以《清末籌備立憲檔案史料》為例，從西元1905年到1911年的字數分別是6,331、168,121、224,482、95,738、66,513、56,122和103,191字，合計720,498字。因此，個別年份的字數比例是0.9%、23.3%、31.2%、13.3%、9.2%、7.8%和14.3%。

關鍵詞彙出現在個別年度的比例是我們定義重要關鍵詞彙的主要依據。「官制」在整個文獻出現總次數是667次，但是其中的302次出現在1906年的文件中，佔了所有出現次數的45.2%，這一數字遠高於1906年文件的字數佔所有文件總字數的比例（也就是23.3%）。「立憲」在整個文獻出現總次數是958次，但是其中的386次出現在1907年的文件中，佔了所有出現次數的40.3%，這一數字遠高於1907年文件的字數佔所有文件總字數的比例（也就是31.2%）。如果「官制」和「立憲」是一般的關鍵詞彙，則它們在這一些年出現的比例或許應該跟該年份在全部文獻所佔的比

表2《清末籌備立憲檔案史料》各年份的10個重要關鍵詞彙

1905	1906	1907	1908	1909	1910	1911
立憲	**官制**	中國	議員	**籌辦**	廳州縣	內閣
大臣	各國	**立憲**	**選舉**	**自治**	**籌辦**	大臣
各國	行政	學堂	諮議局	審判	**自治**	審判
憲法	大臣	滿漢	議長	行政	州縣	經費
日本	各部	天下	章程	地方	各屬	預算
臣等	臣等	各國	資政院	憲政	成立	審判廳
朝廷	中國	法律	議事	**章程**	巡警	顧問
衙門	考察	今日	督撫	刑律	地方	衙門
天下	裁判	日本	各省	諮議局	省城	資政院
政治	日本	人心	議事會	籌辦處	憲政	地方

例相似。因此，這樣的現象暗示了1906年和1907年分別是討論「官制」和「立憲」的重要年份。所以，我們將這兩個詞彙分別定義為各該年份的重要關鍵詞彙。

上面的敘述只是闡明了質性思考的方向，但是我們在程式之中還是得清楚定義「遠高於」的認定標準。這一個標準顯然是研究者需要經過思考才能決定的，在目前的研究工作中，我們暫時以詞彙的年份出現比例超過該年份佔總字數比例的1.1倍作為認定標準。以前面「官制」的例子來說，「官制」出現在1906年的比例是45.2%，而該年度的文件字數佔全部文件總字數的23.3%，而45.2除以23.3大於1.1，因此我們認定「官制」是1906年的重要關鍵詞彙。

基於以上的分析程序，我們可以建構不同文獻資料在各年份資料的重要關鍵詞彙。表2是《清末籌備立憲檔案史料》各年份的10個重要關鍵詞彙。表2中以粗體字特別標示出來的就是圖1中有較詳細資料的重要關鍵詞彙。

四、共現詞組之分年詞頻分析

共現詞組的分年頻率分析與前一節討論關鍵詞彙的分年分析的技術其實沒有很大的差異。兩者都是計算所關切的文字訊息在不同年份的文件中的出現次數。比較主要的差異在於共現詞組的數量可能很大。如果關鍵詞彙有一百個的話，則這一百個關鍵詞彙最多可以組成一萬個共現詞組。如果是要把未經過濾的任意準詞彙拿來當作計算共現詞組的基礎的話，則所需要的計算能量更會超過合理的範圍。我們在

第二節提到過，我們可以用PAT Tree技術從《清季外交史料》擷取到超過七百五十六萬個出現次數超過3次的準詞彙。如果再企圖計算這七百五十六萬個任意字串所組成的所有共現詞組的話，則最多會有超過五十七兆個共現詞組。因此進行共現詞組的分年分析之前，還是只能考慮合理數量的關鍵詞彙。以下我們以《清季外交史料》中特別與「華工」相關的資料來說明進行共現詞組的分年分析的方法。

擇定一組關鍵詞彙之後，我們可以利用程式去分析文獻之中，所有關鍵詞彙組合在文獻之中一起出現的次數。我們在第一節就已經提到，所謂「一起出現」是一個需要精確定義的概念。兩個詞彙可以一起在同一份文件、同一段落或者同一語句一起出現。如果兩個關鍵詞只是在同一份文件中一起出現就當作一起出現，似乎標準太過於寬鬆。如果必須在同一個句子一起出現，則比較合理。然而，所謂一個句子，在現代中文中或許是以標點符號的逗號和句號所隔離的陳述來認定；而古文卻未必都有標點符號可以依賴。而即使有標點符號可以依賴，詞義共同出現的關聯性是否真的不能跨越標點符號，也尚有爭議。因此，我們就以詞彙之間的距離是否不超過30個漢字或者標點符號，作為認定兩個關鍵詞彙是否曾經一起出現的標準。

選定認定標準之後，我們就可以讓計算機軟體去計算關鍵詞彙的共現頻率。以30個字作為認定標準，即使我們有175個關鍵詞彙，實際上只有將近2,100個出現次數超過20次的共現詞組。所以，即使我們所處理的《清季外交史料》包含超過兩百八十七萬個漢字，要計算這2,100個共現詞組在個別年份出現的頻率，在一般的個人電腦上，也不需要超過一個小時的時間。[7]

得到個別共現詞組在各年份出現的次數之後，我們就可以模仿關鍵詞彙的分析程序，先行計算個別共現詞組在各年份出現的比例，再以各年份的共現詞組數目與全部文獻的總次數的比例相比較，就可以找出個別年份的重要共現詞組。跟認定重要關鍵詞彙的問題一樣，我們得要訂一個何謂「重要」的標準。跟第三節的說明一樣，我們是以1.1倍作為認定標準。表3所列是一些年份之中的重要共現詞組。

圖5和圖6則是類似圖1的折線圖，只是圖1是關於重要關鍵詞彙的年份分析。而圖5和圖6是關於一些重要共現詞組分年分析的折線圖。因為所謂「重要」共現詞組是以全文獻的年份分析為基礎，所以圖5和圖6都有全部文獻的年份分析的折線圖。由於這兩個圖所包含的曲線數目較多，且彼此有時接近，因此每一線條都盡量以線條的樣式和所附帶的標記符號來區分。這兩個圖中的共現詞組的折線圖各有自

7 跟以30個漢字當作一起出現與否的認定標準類似，這裡以20次作為是否進一步處理共現詞組的標準，也只是主觀上的選擇。（參考表1）以包含5,758份檔案，總計包含超過兩百八十七萬字的《清季外交史料》來說，出現20次的共現詞組算是出現次數相當少的。如果研究上需要，我們當然可以調低這一篩選門檻。

表3《清季外交史料》部分年份與華工議題相關的重要共現詞組

1875	1878	1879	1884	1886	1887
{大臣、條約}	{日國、使臣}	{領事、中國}	{祕魯、華民}	{華商、華工}	{國民、領事}
{祕魯、使臣}	{衙門、使臣}	{華人、美國}	{華民、華工}	{金山、美國}	{通商、領事}
{祕魯、華工}	{出洋、華人}	{領事、華民}	{祕魯、美國}	{美國、華工}	{香港、中國}
1888	**1889**	**1890**	**1893**	**1894**	**1896**
{小呂宋、領事}	{美國、華工}	{日國、使臣}	{領事、華民}	{註冊、美國}	{機器、華商}
{華人、華工}	{華人、美國}	{華民、華工}	{華民、美國}	{註冊、中國}	{華民、華商}
{領事、中國}	{華商、華工}	{金山、領事}	{古巴、華民}	{註冊、華工}	{製造、機器}
1899	**1902**	**1905**	**1906**	**1908**	**1909**
{領事、中國}	{古巴、領事}	{公司、中國}	{日本、華人}	{華僑、商務}	{合同、工人}
{衙門、使臣}	{國家、華人}	{鐵路、中國}	{土人、華人}	{商務、南洋}	{上海、大臣}
{古巴、領事}	{領事、中國}	{日本、大臣}	{華僑、中國}	{貿易、華人}	{大臣、政府}

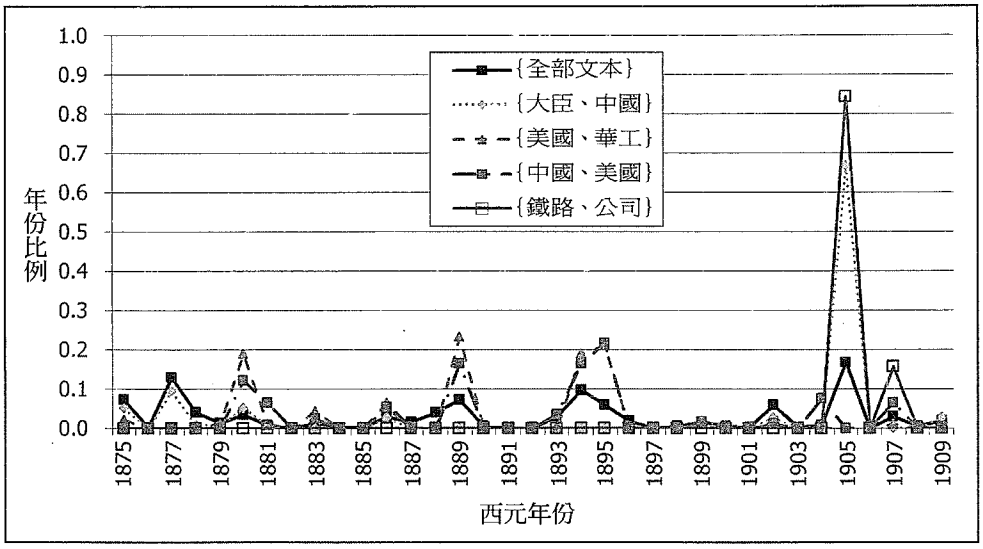


圖5《清季外交史料》的共現詞組分年分析（僅示部分資料，第一組）

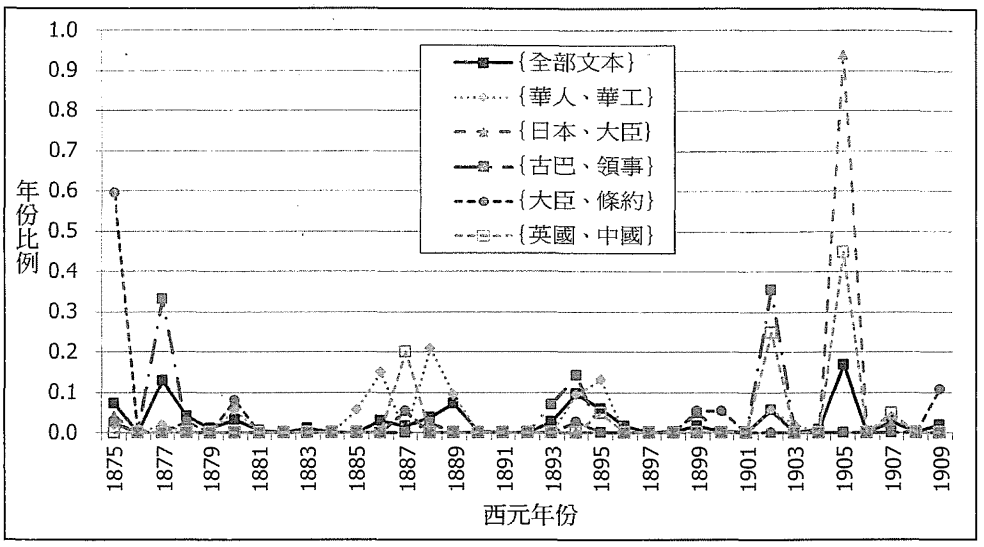


圖6 《清季外交史料》的共現詞組分年分析（僅示部分資料，第二組）

已的趨勢，顯示與不同共現詞組相關的議題在不同時刻成為研究華工相關問題的重點；一個共現詞組的重要性是隨著歷史的推演而改變的。

五、關鍵詞彙、共現詞組與文件權重

透過第三節與第四節所說明的技術，我們可以為每一年份選擇重要的關鍵詞彙和共現詞組，因此可以藉此來檢驗個別文件之中是否用到這一些重要的關鍵詞彙和關鍵共現詞組。如果關鍵詞彙是與某一些觀念密切相關或者某一些共現詞組是與某一些歷史事件密切相關，則一份文件包含很多重要的詞彙或者詞組，就可能是當年度比較關鍵性的文件。基於這樣的理念，我們以文件包含重要關鍵詞彙和重要共現詞組的總數，來當作個別文件的相對權重。當然，這樣的量化機制並非完美，因此慎重的審視文本內容仍是最重要的研究工作。

依照第二節所說明的，如果研究者一開始是依照自己的研究興趣來挑選關鍵詞彙，則經過第三節與第四節的程序，我們希望最後所得的這一些文件權重，能夠讓比較相關的文件獲得比較高的權重，也使得研究者能夠比較容易找到相關的文件。個別文件應當有其歷史意義，量化的機制通常難以面面俱到，因此不能只看文件權重來挑選文件。

表4《清末籌備立憲檔案史料》的文件權重排序

1905		
文件權重	作者	篇名
65	載澤等	出使各國考察政治大臣載澤等奏請以五年為期改行立憲政體摺
57	劉汝驥	御史劉汝驥奏請張君權摺
30	載澤等	出使各國考察政治大臣載澤等奏出洋考察政治請調員隨同差委摺
1906		
文件權重	作者	篇名
420	戴鴻慈	出使各國考察政治大臣戴鴻慈等奏請改定全國官制以為立憲預備摺
312	楊晟	出使德國大臣楊晟條陳官制大綱摺
122	殷濟	內閣校簽中書殷濟為預備立憲條陳籌經費建海軍等二十四條呈
1907		
文件權重	作者	篇名
352	吳劍豐	候選道吳劍豐條陳改良財政言路吏治學務陸海軍警察等六事呈
299	程淯	分省補用道程淯條陳開民智興實業裕財政等項呈
197	劉寶和	學部主事劉寶和條陳立憲預備施行大綱以通上下之情明上下之權呈
1908		
文件權重	作者	篇名
871	憲政編查館	憲政編查館奏核議城鎮鄉地方自治章程並另擬選舉章程摺〔附清單〕
784	憲政編查館等	憲政編查館等奏擬訂各省諮議局並議員選舉章程摺〔附清單〕
632	憲政編查館資政院	憲政編查館資政院會奏憲法大綱暨議院法選舉法要領及逐年籌備事宜摺〔附清單二〕

表4表列《清末籌備立憲檔案史料》中，西元1905年到1908年排名前三名的文件。舉例來說，依據表4，我們知道〈御史劉汝驥奏請張君權摺〉這一份文件用到了57次，被認定是《清末籌備立憲檔案史料》中1905年資料的重要關鍵詞彙。我們以文件包含關鍵詞彙的多寡為文件排序。這一些排名只是建議性質，權重之間的些微差異不能被過度詮釋。同一年份的文件的權重是該文件包含當年度關鍵詞彙或者關鍵共現詞組的總數，權重較高者，代表可能談到較多的關鍵概念，因此可能比較具有參考價值。所以權重65的文件可能比權重只有5的文件具關聯性，但是權重65的文件是否絕對比權重57的文件要值得閱讀，則是有高度爭議。儘管是經過篩選的關鍵詞彙和共現詞組，這一些詞彙與詞組是在怎樣的語境中被提及，當然深深影響這一些文字的重要性，因此文字的意涵仍要靠專業來判讀。猶有甚者，不同年份文

件的權重更不可相互比較，例如，儘管1906年的〈出使德國大臣楊晟條陳官制大綱摺〉的權重是312，而1907年的〈分省補用道程清條陳開民智興實業裕財政等項呈〉的權重是299；但是我們不知道這兩份文件的相對重要性。這一些權重的計算都是基於個別年度的關鍵詞彙，而不同年度的關鍵詞彙集合是不一樣的。因此312和299這兩個數字是基於不同基準所得的權重，直接比較當然是不甚恰當。

我們當然可以應用其他的技術，設計一個適用於全文獻的基準，讓我們可以有一個方法來給不同年份的文件設定相對權重，但是目前我們並沒有採用這樣的設計。

六、討論與未來展望

雖然許多的重要詮釋與判斷終究還是要靠史學專家來拍板認定，但是自然語言處理技術可以很有效率的處理與分析極大量的文字資料，兩者能夠相輔相成。史學文獻固然已經有既定的內容，計算技術也可以幫助研究者抽取相關的內容，然而在史學文獻內容的詮釋工作上，不僅需要專業的知識，且研究者可能採取不同的角度詮釋相同的史料內容。因此專業的判讀與辯證仍為史學研究的最重要環節。[8]這一研究報告說明我們如何利用關鍵詞彙與共現詞組的相關資訊來協助人文學者進行研究，比較深入的史料分析成果，請參看「第三屆數位典藏與數位人文國際研討會」金觀濤等（2011）的其他兩份報告的內容。雖然我們的技術還沒有能夠達到盡善盡美的階段，但是目前的經驗已經讓我們看到數位技術能夠貢獻於人文研究的初步證據。

文字分析技術在數位人文領域的潛力，不只是在國內；其實就連2011年在史丹佛大學舉行的「數位人文國際學術研討會」中，七個教學課程（tutorials）中就有三個是以介紹文字分析相關軟體或者技術為主要的題材。[9]這樣的趨勢或許可以用以回應，許多中文史學家在批評中文數位人文學時，所質疑數位人文是否只是過去的計量史學的另一說法的問題。北京師範大學的方維規教授於2011年的「近代東亞的觀念變遷與認同形塑」國際學術研討會中討論到此一議題。金觀濤先前在不同場合也有所討論與回應；透過分析關鍵詞彙的使用，包含出現的頻率分析與相關意義，我們可以審視關鍵詞彙所代表的觀念的相關歷史。[10]項潔與翁稷安也討論到這一問題，認為以現在計算機大量運算的能力去分析大量的史料，提供研究者以前所未有的角

8 國立臺灣大學資訊工程學系項潔教授於2011年「第三屆數位典藏與數位人文國際學術研討會」中對於本研究之評述。

9 Tutorial 4: Introduction to Text Analysis With Voyeur Tools (https://dh2011.stanford.edu/?page_id=517); Tutorial 5: Gabmap: A Web Application for Analyzing Linguistic Variation (https://dh2011.stanford.edu/?page_id=521); Tutorial 6: Natural Language Processing Tools for the Digital Humanities (https://dh2011.stanford.edu/?page_id=525).

10 金觀濤，2011，〈數位人文研究的理論基礎〉，《數位人文研究的新視野：基礎與想像》，頁45-61。

度審視史料所攜帶的資訊，其固然有量化的部分表象，但是實質意義卻不只侷限於量化。[11]

幫助研究者自大量文字資料中提取重要的關鍵詞彙是一件非常重要的工作。在這一方面，本文約略提及〈社會行動的數位人文研究：以清末預備立憲為例〉（金觀濤等，2011）利用齊夫定律的作法，是一個相當新穎的嘗試。關鍵詞彙的自動提取不只是在數位人文學中能夠感受到它的重要性，在其他自然語言處理的技術領域，例如資訊檢索（information retrieval），其實還有不少其他前人的經驗可以借鏡，例如latent semantic analysis（Deerwester et al., 1990），也是一種提取關鍵詞彙的相關技術。目前比較關鍵的瓶頸是以PAT Tree技術所找到的準詞彙，還是需要專業人力來篩選關鍵詞彙，當整體文獻資料量非常大的時候，這樣的篩選工作幾乎就接近了編輯詞典的工作。編輯不同時代的常用詞彙對於利用資訊技術處理不同時代的中文語料，具有極重要的貢獻，這類的工作已經有漢學研究者進行中，例如德國海德堡大學的華格納教授的「清末民初詞彙數據庫」（Wagner, 2011）。如果分析文獻的最終目標是推薦文件的話，則〈社會行動的數位人文研究：以清末預備立憲為例〉（金觀濤等，2011）所嘗試的方法與結果是相當值得參考的。儘管如此，如何在缺乏完整的詞典的情形中，可以像處理現代中文文件那樣有效率的處理歷史文件，應該是一個有挑戰且非常重要的研究議題。

誌謝

本研究承蒙教育部補助國立政治大學之邁向頂尖大學計畫100H51與國家科學委員會NSC-99-2221-E-004-007及NSC-100-2221-E-004-014之部分補助。本文與先前研討會論文的評審提供了許多寶貴的建議，協助本文盡量兼顧資訊與人文背景讀者的資訊需求，是本文作者非常感激的。本研究所使用之PAT Tree程式源於谷歌（Google）臺灣分公司總經理簡立峰博士。謹此一併致謝。

11 項潔、翁稷安，2011，〈導論——關於數位人文的思考：理論與方法〉，《數位人文研究的新視野：基礎與想像》，頁9-18。

參考文獻

王彥威、王亮編，1934，《清季外交史料》。

古倫維，2009，《意見分析之研究與應用》，國立臺灣大學資訊工程學系博士論文。

金觀濤、邱偉雲、劉昭麟，2011，〈「共現」詞頻分析及其運用：以「華人」觀念起源為例〉，《第三屆數位典藏與數位人文國際研討會論文集》，臺北：國立臺灣大學，頁199-223。（本篇論文的增修版本收錄於本書之中。）

金觀濤、姚育松、劉昭麟，2011，〈社會行動的數位人文研究：以清末預備立憲為例〉，《第三屆數位典藏與數位人文國際研討會論文集》，頁309-310，臺北：國立臺灣大學。

金觀濤、劉青峰，2011，中國近現代思想及文學史專業數據庫，國立政治大學與香港中文大學合作建構。（尚未開放）

故宮博物院明清檔案部編，1979，《清末籌備立憲檔案史料》（上、下冊），北京：中華書局。

黃瀚萱，2008，《以序列標記方法解決古漢語斷句問題》，國立交通大學資訊科學與工程所碩士論文。

羅鳳珠、張如瑩、江姿瑩、彭瑜璇，1997，〈以「互動觀念」建立「紅樓夢網路資料中心」對紅學發展之影響〉，1997年「北京國際紅樓夢學術研討會」發表之論文。

Adamic, Lada A. & Huberman, Bernardo A. (2002). Zipf's Law and the Internet. *Glottmometrics, 3*, 143-150.

Chen, Chien-Liang, Liu, Chao-Lin, Chang, Yuan-Chen & Tsai, Hsiangping. (2011). Exploring the Relationships between Annual Earnings and Subjective Expressions in US Financial Statements, *Proc. IEEE International Conference on e-Business Engineering 2011.*

Chien Lee-Feng. (1999). PAT-tree-based Adaptive Keyphrase Extraction for Intelligent Chinese Information Retrieval. *Information Processing and Management, 35*(4), 501-521.

Deerwester, Scott, Dumais, Susan T., Furnas, George W., Kandauer, Thomas, K. & Harshman, Richard. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science and Technology, 41*(6), 391- 407.

Ioannides, Yannis M. & Overman, Henry G.. (2003). Zipf's Law for Cities: An Empirical Examination. *Regional Science and Urban Economics*, *33*(2), 127-137.

Luhn, Hans Peter. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, *2*(2), 159-165.

Manning, Christopher D. & Schütze, Hinrich. (1999). *Foundations of Statistical Natural Language Processing*, Cambridge, MA: The MIT Press.

Wagner, Rudolf G.，2011，「近代東亞的觀念變遷與認同形塑」國際學術研討會所提及之德國海德堡大學之「清末民初詞彙數據庫」。網址：http://www.zo.uni-heidelberg.de/sinologie/institute/staff/wagner.rudolf/cv.html

Wang, Xiaoguang & Inaba, Mitsuyuki. (2009). Structures and Evolution of Digital Humanities: An Empirical Research Based on Correspondence Analysis and Co-word Analysis, *Proc. International Conference of Digital Archives and Digital Humanities 2009.*

Yu, Liang-Chih, Wu, Chung-Hsien & Jang, Fong-Lin. (2009). Psychiatric Document Retrieval Using a Discourse-aware Model, *Artificial Intelligence, 173* (7-8), 817-829.

Zipf, George Kingsley. (1949). *Human Behavior and the Principle of Least Effort*. Boston: Addison-Wesley.

以文本分析呈現臺灣海外史料政治思想輪廓

劉吉軒*、柯雲娥**、張惠真***、譚修雯****、
黃瑞期*****、甯格致******

摘 要

政治制度的發展決定國家利益與人民福祉，臺灣從專制到民主的演變過程中，海外臺灣人民在僑居地資訊充分與言論自由的條件下，曾經對臺灣政治制度的改革扮演著重要的角色，包括思考、辯論、發聲、啟迪，這一段歷史的縮影記錄於當時海外臺灣人民所發行的政論性刊物。然而這些刊物大部分為非正式出版，發行量少、資料散佚、國內罕見，隨著時空環境的轉變，已經面臨逐漸流失的嚴重危機。因此，這批刊物的徵集、整理、數位化、全文化，不僅幫助保存這一段歷史紀錄，作為國內外學界進行臺灣歷史、政治、社會等面向的重要研究資產，也提供了數位人文研究的珍貴素材。本研究以左派刊物為例，嘗試探討史料分析的二個面向的議題：第一個議題為人工關鍵詞與自動關鍵詞的比較，通常人工關鍵詞以單篇內容為範圍、呈現人事時地物等面向的核心資訊，而自動關鍵詞則以文字符號組成單元的計算方式，以全文文本中詞語出現次數與使用重複性的比較為基礎，挑選出代表性詞語。本研究探討二者間的差異及其意涵。第二個議題為社會網路模型的應用，本研究以社會網路模型中之共現網路建構史料文本中關鍵詞之關聯情形，透過社會網路的視覺化工具，觀察關鍵詞的角色及彼此的關聯，提供一種新型態的內容解讀與分析結果。整體而言，本研究透過數位人文的研究方法，呈現出海外臺灣人民部分的政治思想輪廓，並希望能促進後續更全面、更深入的研究成果。

* 國立政治大學資訊科學系教授兼圖書館館長。
** 國立政治大學圖書館編審。
*** 國立政治大學圖書館組員。
**** 國立政治大學圖書館組長。
***** 國立政治大學圖書館專案助理。
****** 國立政治大學資訊科學研究所博士研究生。

Text Analysis on Overseas Taiwanese Journals for Political Thought Profiling

Jyi-shane Liu *, Yun-er Ke **, Hui-chen Chang ***,
Hsiu-wen Tan ****, Ruei-chi Huang *****, Ke-chih Ning ******

Abstract

A nation's political system has a profound influence on its growth and its people's welfare. Under the favorable conditions of abundant information and free speech in foreign residency, many overseas Taiwanese intellectuals are enthusiastic in deliberating, debating, advocating, and inspiring political thought, thus, playing an important role in the process of Taiwan's political transformation from anarchy to democracy. A part of this history was captured and reflected in political journals published overseas. These journals were mostly informally published with limited circulation, scattered and lost, and rarely seen in Taiwan. There is a serious threat of forever losing these journals. Therefore, the work of collecting, organizing, and digitizing these journals will help preserve important historical records and provide valuable materials for academic studies of Taiwan's history, as well as its political and social evolution.

The current research attempts to investigate two approaches to text analysis using left-wing journals as an example. The first approach compares human annotated keywords with computer-extracted keywords. Human annotated keywords usually represent core ideas of an article interpreted by human readers, whereas computer extracted keywords are generated by statistical selection of frequently used words. Differences between the two sets of keywords and their implications are analyzed and discussed. The second approach involves the use of social network modeling, and visualization in representing and observing entity co-occurrence

* Professor, Department of Computer Science, National Chengchi University. University Librarian of National Chengchi University Libraries.
** Executive Officer, Digital Preservation Section, National Chengchi University Libraries.
*** Officer, Digital Preservation Section, National Chengchi University Libraries.
**** Section Chief, Digital Preservation Section, National Chengchi University Libraries.
***** Project Assistant, Digital Preservation Section, National Chengchi University Libraries.
****** Ph.D. Student, Department of Computer Science, National Chengchi University.

relations in text. Overall, as an empirical study in the paradigm of digital humanity, the current research helps illustrate the profile of political thought expressed by overseas Taiwanese intellectuals and hopes to facilitate further investigation for in-depth results.

一、前言

臺灣是一個海島移民社會，數百年來向海外發展一直都是臺灣歷史非常重要的一部分。19世紀中期開始，臺灣人的足跡已出現在南洋、中國、日本等地，戰後則以美國為主要根據地，並漸次擴展到世界其他地區（藍適齊，2002）。二次世界大戰以後，臺灣海外活動的推動者除二二八事件後逃亡到海外的部分反對人士外，最主要是因1950年代開放出國留學，至60年代後蔚為風潮，臺灣的留學生幾乎遍及世界主要的文明先進國家。

留學的目的最初只是為了學習先進國家的技術，以便學成返國能一展所學，對自己的鄉土有所貢獻。但留學生到了先進國家，受到不同文化之衝擊，進而比較與反省自己國家社會的種種問題，當獲知事情的真相之後，自然會發聲批判不合理的體制，並提出相關解決方案（鍾才，1994）。但當時臺灣島內因威權統治下的言論限制，任何對當權者的不滿和批評，在島內幾乎完全被消音，於是海外則成為知識份子抒發理想、集會結社的唯一自由空間。

發行刊物，最初也只是為了覓得一個共同發表言論的地方，作為彼此研究、見聞的交流園地（侯榮邦，2005：577）。同時，因經濟因素，不論撰稿者或是雜誌編輯者，大都是義務性質，純為理想出發。但隨著島內政治情勢之演變，刊物內容逐漸轉為對社會現實之批判性言論，且為了向國際社會告發獨裁政權統治下的臺灣現狀以及發出臺灣人民真正的心聲，喚起國際社會輿論的注視與關心，於是紛紛出版政治性刊物，進行政治思想的啟蒙與宣傳工作，亦代表著臺灣島內的沉默大眾，向國際社會傳達另一種不同的聲音。而宣傳刊物依訴求對象的不同，而出現不同的語文版本，有中文、日文、英文等，形成百家爭鳴的局面。

這些海外刊物，通常因資金、人力、訂戶量等種種因素考量，故發行量稀少；另一方面則因空間地域之隔閡及當時島內政治社會環境之限制，言論不自由，並不允許公開發行，多以祕密方式傳布。因此，島內人民無從得知其真正內容，自然對事件的來龍去脈不熟悉，甚至漠視之。但這些刊物與島內的黨外雜誌相互唱和，提供知識份子養分，與臺灣民主運動二者間關係密切，間接影響著臺灣的政治與社會發展，是珍貴的臺灣研究史料。

傳統人文研究大多是通讀文獻後進行理解，由研究者主觀分析擬構，以全文大意的方式去分析文本，主張通讀全文理解之文意才是作者欲表達的真正意義（邱偉雲，2010）。資訊科技的應用改變人文學者與材料間的關係，也為人文學術研究提供了新的工具。研究歷史，除了對歷史文獻細密而又富想像力的解讀，資訊科技與解釋或意義的發掘並不互斥，甚至還可以尋找到一個新的接榫點（王汎森，2004）。資訊科技除了可以彌補傳統人文學術研究方法之不足，同時也能創新（洪

一梅，2009），換言之，資訊科技的應用，不僅會改變學科原有的生態，甚至引起學科的典範轉移，讓人文與社會科學有新的內涵及發展方向。

數位人文研究即是數位典藏、資訊科技與人文研究者三者互動的新興領域，不但擴大人文學科研究的素材範圍，也因數位典藏成果與資訊科技的結合而產生新的應用。透過數位化資訊，能發掘出新的研究主題及素材，使其後續有更深度的研究意義與更寬廣的成果空間。Todd Presner（2010）指出第一波的數位人文浪潮是在1990年代晚期到2000年代初期，著重在大規模的數位化工作與基礎架構的建立。而目前第二波的浪潮則是重心轉移至數位人文2.0（Digital Humanities 2.0），著重在原生數位資源的產製、流通和互動所需的數位人文研究的環境與工具開發，開創新興研究場域、研究方法、出版模式，樹立跨領域研究的新典範。

本研究即是利用國立政治大學圖書館建置的「臺灣政治與社會發展海外史料資料庫」所收錄的海外刊物為研究對象，透過文本分析與詞語關聯的共現網路模型，進行多面向之比較觀察，發掘其中所隱藏的資訊，藉以探知當時海外知識份子的政治思想輪廓，不僅直接顯現當時社會狀況與知識份子的關注焦點，也是重現歷史的直接線索，對於探討臺灣政治社會之階段性發展有不可抹滅之意義。

二、文獻回顧

（一）詞頻分析

隨著大量數位化史料的聚集，以及資訊科技的發展，數位人文的發展，自人文計算（humanities computing）轉化而來，並立基於數位典藏工作而產生的大量數位史料。應用資訊技術來協助人文研究，使得人文學者對於史料大規模的進行觀察與檢索成為可能。大量的史料適合做整體性的觀察，但史料的產出背景需大概一致，並具有一定程度的完整性，才能一致的進行詮釋，並支持整體觀察的詮釋（陳詩沛，2011）。

詞頻分析即是以統計方法計算詞語在文本中出現的頻率，透過特定詞語出現頻率計算，一方面觀察詞語的強度所反映的文本主題或詞語演變，另一方面也可觀察詞語的分布與彼此之間的關係。共詞分析目的在建構領域關鍵知識的連結性，藉由詞頻的統計與詞語的共現關係，呈現出領域知識的群集結構及演進趨勢。相關學者已使用共詞分析進行各個學域的研究，進而發掘該領域的熱門議題與發展焦點。而關聯法則分析是資料探勘技術最常被使用的方法，藉由支持度與信心水準兩個指標找出資料集合中某些項目間的關聯性；由於關聯法則的表現明確易懂，因此被廣泛的運用於不同的領域（如商業、網路及醫學等），然而卻較少於主題領域（如知識

管理）研究議題中進行探討（陳良駒、張正宏、陳日鑫，2007）。

（二）主題分析

主題分析是辨識某作品所包含之知識內容（intellectual content）的過程；係依據文獻顯著的特性加以解析，並以數字、符號、名詞、形容詞與名詞的組合或片語標示出文獻中所述及的主題，以作為資料查詢的檢索點。主題分析的結果可能會以二種方式呈現在目錄或書目之中：一是數字符號，如分類系統；一是語言詞語，如標題或索引詞（Chan, 1994）。

主題分析常使用的標題法，是一種以經過規範化處理的標題詞為文獻主題概念呈現的主題法，其優點為主題用語明確、詞語使用一致、解決語意問題、可表達主題之階層附屬關係（陳明來，2002），但傳統的控制詞語索引除了回現率不如自然語言索引外，最為人詬病的是不符合經濟效益（陳光華、伍健廷，1998）。而隨著電腦檢索的應用，關鍵詞描述法出現在美國1950年代中期，是直接引用文獻所使用的主題詞語，用以代表文獻主題概念，並作為檢索用詞，包括題名、作者，乃至全文等出現的原始詞語。然而關鍵詞屬於自然語言，通常來自文獻本身內容，如題名、摘要、目次等，詞語間並沒有事先建立參照及層次附屬關係，因此，相對於使用控制詞語的主題概念呈現方式，同一主題款目的文獻，常會因為作者用詞的不同而分散（陳明來，2002）。

（三）社會網路與共現網路

社會網路是將個體之間的關聯結構以圖像的方式呈現，由節點與其間的連結所組成。節點可以是任何的個體，連結則代表著個體與個體間的關係。社會網路分析（social network analysis, SNA）是一種研究社會結構、組織系統、人際關係、團體互動的概念與方法，是在社會計量學（sociometry）基礎上所發展起來的分析方法。社會網路分析在許多不同的學術領域被視為一重要的工具，如社會學、人類學、社會心理學、經濟學、物理學、資訊科學等，更由於網際網路的發展，成為熱門的研究方法（Scott, 2000）。社會網路分析主要是從網路中找出顯著的樣本與觀察其配對之間的資訊流動的關係。其基本觀察要件有點、線、集中性、密度、位置、居間中心性、距離及結構洞等項（Wasserman & Faust, 1994）。社會網路憑藉不同於傳統分析的視點，可提供較大規模的觀察面向；應用資訊科技的輔助，能夠從大量的資料片段中，浮現概念或實體間的關係與輪廓（Otte & Rousseau, 2002)。

共現網路是以個體的共同出現情形為主要的資訊目的，個體是網路中的節點，而兩個個體在某一定義下的共同出現事件，則表示為兩個節點之間的連結。因此，共現網路其實是一種特殊的社會網路，其主要應用則為大量文本的分析與資料挖

掘，包括俗民分類的結構分析（Cattuto et al., 2007）、新聞報導的內容探討（Özgür et al., 2008）、學術文獻中的知識結構（Su & Lee, 2010）。字詞共現網路的建置，可以視覺化的方式呈現網路圖形，並進一步觀察不同網路圖形的節點、連結線數、網路密度和節點間的距離，另外也可利用社會網路分析中相關的指標量化計算方法，分析網路中節點或節點之間的特徵，進而提供許多文本分析之豐富線索與有利資訊。

三、研究方法

（一）研究架構

本研究旨在利用統計工具與社會網路模型方法，分別比較人工關鍵詞與自動關鍵詞在內容解讀與資訊呈現兩個面向之差異。前者在於觀察專家註解之詞語與作者用詞間的異同；後者觀察文本中重要詞語之間的關聯性所呈現之左派刊物關注之議題與政治思想。其中，人工關鍵詞是由圖書資訊專業人員直接檢視文本後給予，自動關鍵詞則利用中研院CKIP中文斷詞工具並經過條件篩選後得之。

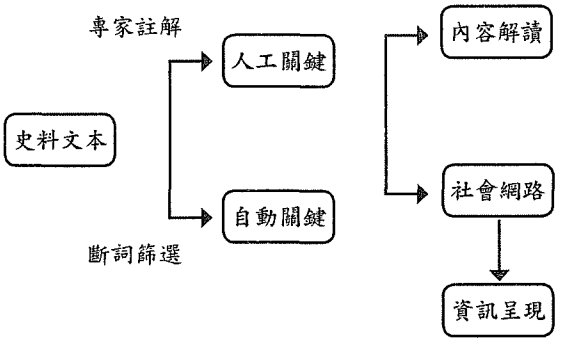


圖1　研究架構

（二）資料庫介紹

「臺灣政治與社會發展海外史料資料庫」是將1950-1990年代，臺灣人於海外所發行的政論性刊物數位化，其中包含各種不同意識型態的言論內容，作為以後國內外學界進行臺灣相關研究很重要的史料，一方面能方便需求者易於近用，另一方面也幫助社會大眾進一步了解海外臺灣民主運動的原因與動機及其演變過程。

本資料庫共收錄85種、2,247期海外刊物，以地區分，有歐洲、日本、及美國；以屬性分，可粗略分為左派、右派及中間派等，各派都有其發行的代表性刊物。如：右派的《美麗島》週刊；左派的《台灣人民》、《台灣革命》；革新保臺中間派

的《波士頓通訊》。這些雜誌可看出雖然立場不一樣、主張不一樣，可是它們關心臺灣，要奉獻臺灣民主前途的出發點是相同的。本研究初期無法以全部刊物為研究對象，乃立意取樣以資料庫中左派刊物為例。

林照真（2008）認為左派是國共內戰後的臺灣禁忌，白色恐怖更將左派思想斬草除根，臺灣內部沒有左派的土壤，也極難看到左派成形的組織。左代表對社會的激烈反省與改革，右則代表保守。在臺灣極少有人對馬克思理論深入探討，臺灣人不談左，左是一種負面象徵。因此，本研究以海外的左派刊物為研究對象，以關鍵詞詞頻及社會網路分析，觀察刊物內容所呈現的思想層面與趨向，並比較各刊間之同異，彼此間相互之關係，一窺當時左派刊物之言論重點及思想演變，作為研究1950-1990年海外臺灣人左派運動之橋樑。

（三）資料樣本與背景資訊

本研究所選取的左派海外刊物計有八種，為《台灣人民》、《台灣革命》、《台灣時代》、《台灣思潮》、《台灣解放》、《台灣天地》、《海外政論》及《建台》，如表1。各刊物之基本背景簡述於下。

● 台灣人民

左雄、許登源等，1972年10月於北美洲加拿大所發行之刊物。此刊物為臺灣社會主義者於北美洲集結，表達思想與付諸實踐的重要指標，反映出臺灣社會主義者對「臺灣問題」從辯證到認識的過程。整個海外臺灣左派運動，左雄路線近乎唯

表1 本研究涵蓋之左派刊物

刊物名稱	發行年代	總發行期數	收錄卷期	發行地
台灣人民	1972/10-1975/2	10	1-10	Halifax, Canada
台灣革命	1975/9-1976/12	5	1-5	Toronto, Ont., Canada
台灣時代	1977/1-1982/4	13	2-13	Canada
台灣思潮	1981/5-1984/4	8	1-8	USA
台灣解放	1987/2-1987/6	2	1-2	Darien, IL, USA
海外政論	1981/2-1984/12	16	1-16	New York, USA
台灣天地	1985/10-1987/6	15	1-15	Summit, NJ, USA
建台	[1970]*	不詳	12（1972/7）	Seattle, WA, USA

*本資料庫僅收錄第12期，以西雅圖華盛頓大學圖書館館藏推估其創刊號於1970年發行。

一，倡導社會主義臺灣獨立運動，屬左派思維之啟蒙期。至1975年2月止，共發行10期。

● 台灣革命

左雄於1975年創刊發行，標幟了臺灣左派運動民族、民主革命的性質與民族、反帝的運動綱領，同時提出了列寧主義的組織路線原則。是主張臺灣獨立走社會主義路線的刊物，可算是開創左派政治組織路線內容的充實確立階段。於1976年12月停刊，共發行了5期。

● 台灣時代

左雄於1977年所發行之刊物。是屬組織路線上的具體發展鞏固期。整個海外臺灣左派運動，自《台灣人民》、《台灣革命》至《台灣時代》，可說圍繞在左雄的身上為其運動發展過程的中心焦點。至1982年4月止，共發行了13期。

● 台灣思潮

1981年5月於洛杉磯創刊，由離開《美麗島週報》的左派人士所發行。主要是從社會、勞工群眾等層面切入，而非以政治的、菁英式的立論點來探討國家的位置，其統派立場認為臺灣資產階級的民主要求是不徹底的，只有臺灣無產階級形成階級的力量和組織所領導的臺灣人民自我解放的革命才是究竟。至1984年4月止，共發行8期。

● 台灣解放

1987年2月創刊，一反先前以左雄為中心思想的路線，對左雄的路線的運動學理主義進行嚴厲的批判。主要針對國際時事評論對臺灣民族解放運動的影響，及臺灣民族解放運動中各個階級的性質與傾向之探討。由刊物內容可看出左派思潮之轉變。至1987年6月為止，共發行了2期。

● 台灣天地

張金策1985年10月以「新台灣同志會」的名義於休士頓發行。此刊物是張金策力行實踐左派的革命宣傳工作，傳播社會改造的理念。其探討臺灣獨立的意義，是主張獨立的力量來自於無產階級的觀點，認為必須透過群眾運動來實踐獨立，因此鼓吹社會運動。至1987年6月為止，共發行了15期。

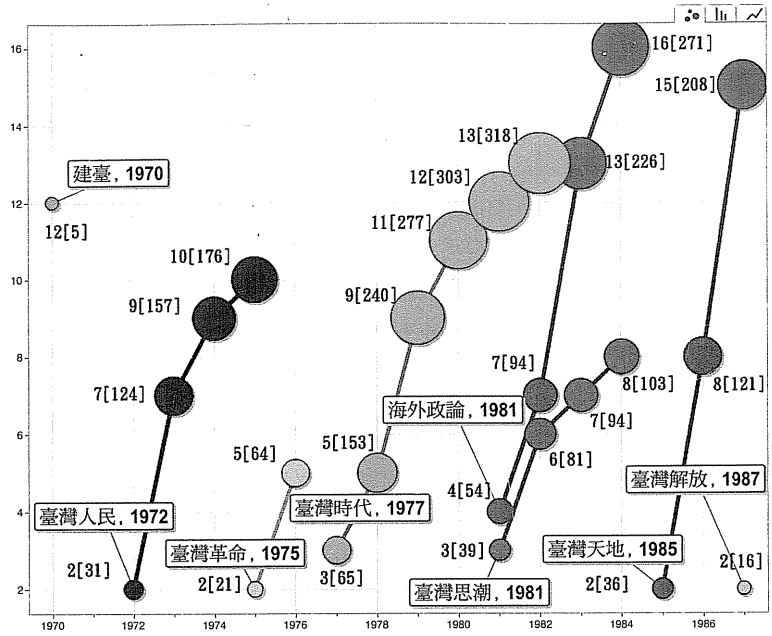


圖2　本研究樣本刊期與時間分布圖

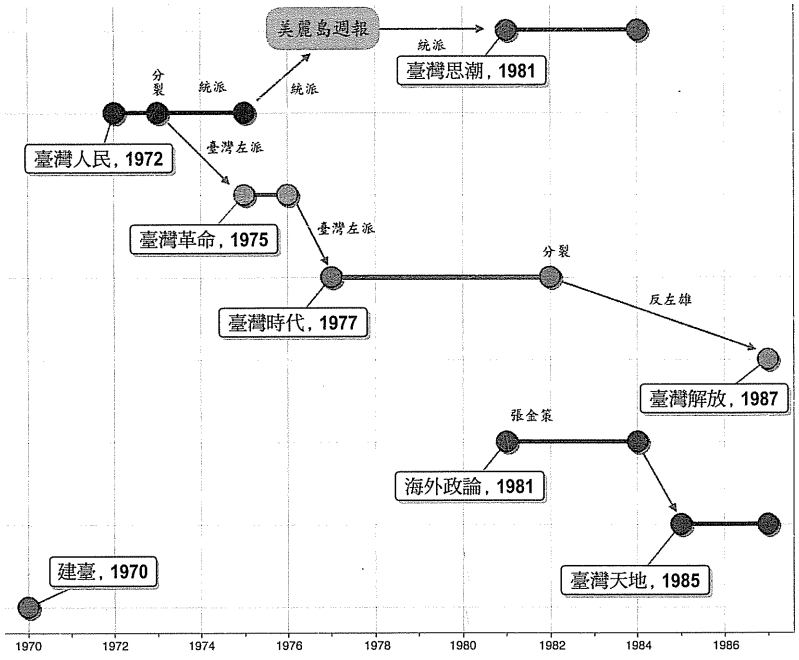


圖3　左派刊物發展脈絡

● 海外政論

高成炎、鍾維達、張金策等人於1981年在美國休士頓發行的刊物。除關懷弱勢、宣導各項抗爭議題及傳播社會改造的理念外，也曾與《台灣時代》一起挑戰「臺獨聯盟」的霸權地位，對其親美國帝國主義的態度一直有著極為嚴厲的批判，由此可看見黨外運動中的不同聲音，以及對另一種帝國霸權的批判。第8期起改為定期雙月刊，每逢雙月發行。至1984年12月止，共發行16期。

● 建台

本資料庫僅收錄第12期（1972年7月），無法探知詳細資訊，只知由建台會原在加拿大溫哥華發行，1972年3月起改於西雅圖發行。認為在臺灣人的各種自救運動中，因對社會主義的不了解而產生恐懼的狀況，因此詳細介紹了社會主義與革命的關係或其外交政策、經濟制度等事項，以對社會主義有進一步的認識。

綜上所述，本研究所包含之範圍如圖2所示，依發行的年度排列；各刊之圓圈大小依年度累積文章篇數而變化，並標示其刊期於括弧前，文章累計篇數則標示於括弧中。

另外，本研究在研究人員對於各刊資料內容的充分掌握及其他研究結果的部分佐證之下（張清水，1987；許維德，2001），整理了左派刊物更為扼要清楚的發展脈絡（如圖3所示），作為後續文本解讀與研判的背景資訊。本研究之八種刊物，雖然都屬左派刊物，但刊物之發行有分分合合之現象，大約可分成以下四種不同脈絡。一是以《台灣人民》及《台灣思潮》為範圍；二是以《台灣革命》、《台灣時代》及《台灣解放》為範圍；三是以《海外政論》及《台灣天地》為範圍；四則是《建台》，如圖3所示。

（四）研究工具與流程

1.文本處理

首先將研究所需之刊物，剔除刊物中所有的廣告、漫畫等資訊後，將有意義之論述文章全文繕打，並移除文章之格式、作者、文章標題等，給予單篇標記與段落標記後，存成純文字（Big5）檔案。

2.專家註解

每篇文章由圖書資訊專業人員直接檢視全文後，進行主題分析，依文章之內容主題與形式，給予不等個數的人工關鍵詞，代表該文之重要意涵。

3. 中文斷詞

本研究採用中研院中文詞知識庫小組（Chinese Knowledge and Information Processing，簡稱CKIP）提供之中文斷詞工具，進行全文斷詞以及詞性標注（中文詞知識庫小組，n.d.）。

4. 斷詞詞性篩選

首先設定單詞詞性篩選條件為名詞、動詞及非謂形容詞；複合詞詞性篩選條件則參考人工關鍵詞之詞性組合。

5. 詞頻統計

針對全文斷詞及詞性篩選後之詞語進行頻率統計，剔除較無意義的詞語，得出每刊的單詞詞頻統計及複合詞的詞頻統計。另外，也對人工關鍵詞進行詞頻統計。

6. 重要自動關鍵詞及人工關鍵詞篩選

自動關鍵詞之篩選門檻為總詞頻之1%；人工關鍵詞之篩選門檻為總詞頻之0.5%及絕對詞頻2；大於門檻值之詞語則列為重要關鍵詞。門檻值之設定以得到合適數量之重要關鍵詞為目的，以全文斷詞總詞頻之1%為篩選門檻，八個刊物共得到91個重要自動關鍵詞；而以各刊人工關鍵詞總詞頻之0.5%及絕對詞頻2為篩選門檻，則八個刊物共得到161個重要人工關鍵詞。本研究認為90-160之詞語節點數量為合適觀察分析的網路大小，因此，個別門檻值的選擇應為恰當。雖然重要自動關鍵詞之數量少於重要人工關鍵詞，但自動關鍵詞之產生完全以詞頻為依據，若詞頻過低（小於1%），則將使詞語之代表性不足。在人工關鍵詞部分，本研究認為重要關鍵詞應該至少出現兩次以上，另外，有些刊物的文章數量要遠高於其他刊物，因此，必須再加上總詞頻比率之門檻，以避免過度寬鬆之篩選。詳細之篩選結果請參見表4及表8。

7. 共現網路模型

本研究以共現網路為詞語關聯分析的基本工具，圖4為共現網路建置的流程。

字詞共現網路的建置過程說明如下：

第一步：完全連結（Permutation），以單位範圍（如全篇或是段落）內文本提取出來的詞語集合 $g_j = \{o_1, o_2, ..., o_i\}$（重要人工關鍵詞或是重要自動關鍵詞），建立

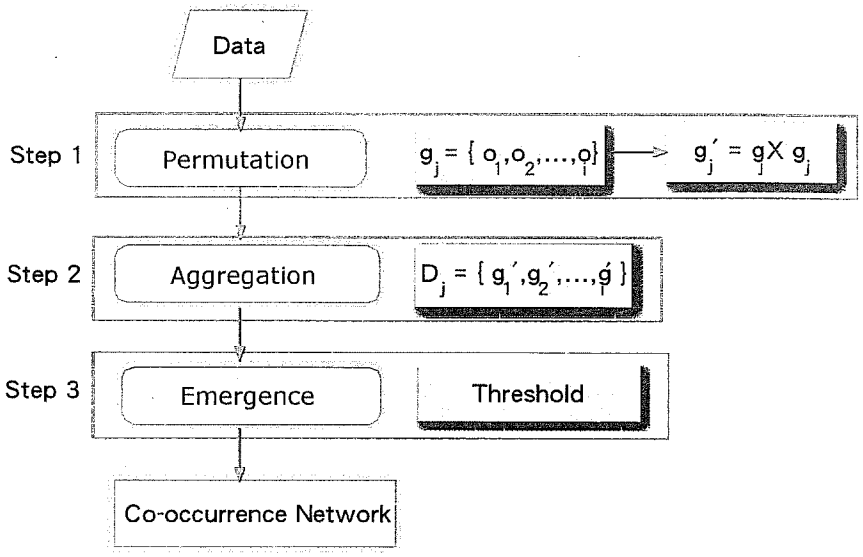


圖4 詞語共現網路建置流程

出相鄰矩陣（adjacency matrix）。同一個詞語集合中之各詞語元素可視為網路中之節點，將這些節點依照彼此之相鄰關係連結起來，便可建立出屬於該詞語集合之詞語子網路g_j'。

第二步：聚合（Aggregation），將上一步驟所產生的各詞語子網路集合g_j'，依據指定之聚合範圍（例如以一篇文章、一整刊，或數刊），將各個子網路加以疊合，產生一個整體的聚合網路 $D_j = \{g_1', g_2', \ldots, g_i\}$。逐步聚合的過程中，我們須同時考慮各個子網路中相同的節點或連結並加以疊合（疊合次數的累計視同加強該節點或連結的重要性），以及將不同節點與現有節點加以建立連結（產生關聯），之後便可依照節點或連結的重要性多寡來篩選相對重要的詞語群組。

第三步：浮現（Emergence），以各節點重要性數據排列，依照設定的域值門檻，過濾相對較不重要的節點，而篩選出較具顯著連結的網路。

建構共現網路的基本選擇為共現範圍的設定，本研究在自動關鍵詞的關聯分析上，以文章全文的段落為關聯範圍；在人工關鍵詞方面，則以同一篇文章為關聯範圍。最後，研究過程中的相關工具參數整理如表2。

表2 本研究之相關工具參數

重要詞語	來源	篩選門檻	共現範圍
人工關鍵詞	專家註解	總詞頻之0.5%以上 ＋絕對詞頻 ≧ 2	一篇文章
自動關鍵詞	全文斷詞	詞性＋ 總詞頻之1%以上	文章段落

另外，本研究採用由美國史丹福大學研究人員開發之社會網路視覺化工具Vizster（Heer & Boyd, 2005）。該軟體工具提供了便利的網路結構與節點連結觀察方式，透過選取及拉曳，使用者可彈性的檢視網路的整體或局部。Vizster網路另一特殊之處在於一個根節點（root node）的設計，作為網路中所有節點的共同連結。通常，根節點的存在與顯示，並不影響大部分的觀察與分析目的。

四、關鍵詞分析

本研究以左派刊物為例，針對大量史料內容，以關鍵詞分析及共現網路之觀察，探討當時左派刊物的意識型態及言論重點。關鍵詞是呈現文件主題意義的最小單位，因此，透過關鍵詞可呈現出刊物的論述核心。

（一）人工關鍵詞

為完整呈現刊物之全貌，故資料庫之著錄是以篇為內容單元，除了有意義的論述文章外，尚包括刊物中所有的廣告、漫畫、啟事等資訊。但本研究只擷取有意義之論述文章，剔除較無論述性質的篇章，各刊之文章篇數如表3所示。

由表3得知，本研究採用之八種刊物中，以《台灣時代》中的論述文章篇數佔總篇數之比例最高，《台灣革命》次之，《台灣思潮》最低。人工關鍵詞共2,075個詞語，去除重複詞，則總詞語數共1,445個，各刊關鍵詞數如表4所示。以《台灣時代》最多，《台灣天地》次之，《台灣解放》的人工關鍵詞數最少。各刊人工關鍵詞數量會受收錄刊物的期數多寡，及專家解讀的主觀因素影響，但推論另一個重要的原因是該刊物內容之論述焦點，如果議題集中的話，人工關鍵詞個數較少但頻率高；反之，若議題分散，則關鍵詞個數會增加，而頻率降低。為比較各刊中重要的關鍵詞語之異同，本研究依表2之篩選標準，八刊共得161個關鍵詞，如表5所示。

表3 各刊收錄之文章篇數表

刊物名稱	論述性文章篇數（%）	資料庫中總篇數
台灣人民	142（80.68%）	176
台灣革命	55（85.94%）	64
台灣時代	279（87.74%）	318
台灣思潮	67（65.05%）	103
台灣解放	12（75.00%）	16
海外政論	227（83.76%）	271
台灣天地	163（78.37%）	208
建台	10（66.67%）	15

表4 各刊重要人工關鍵詞語篩選情形

刊物名稱	不重複關鍵詞數	關鍵詞總頻率	最高詞頻次數	最高詞頻佔總詞頻之百分比	選取頻率百分比	選取頻率次數	關鍵詞選取個數
台灣人民	124	160	9	5.6%	1.3%	≧2	13
台灣革命	178	272	13	4.8%	0.7%	≧2	40
台灣時代	978	1984	68	3.4%	0.5%	≧10	22
台灣思潮	79	104	4	3.8%	1.9%	≧2	18
台灣解放	45	67	8	11.9%	3.0%	≧2	11
海外政論	263	421	20	4.8%	0.7%	≧3	30
台灣天地	350	504	11	2.2%	0.8%	≧4	20
建台	58	69	3	4.3%	2.9%	≧2	7

依表5所列，觀察重要人工關鍵詞於各刊的分布情形，同一詞同時出現在二種刊物以上者，共有29個關鍵詞，佔所有重要人工關鍵詞的18%。其分布如表6所示。顯示本研究範圍的刊物中，最普遍被討論的主題是「革命運動」，出現於六種刊物中；其次是出現於五種刊物中的「社會主義」及「帝國主義」；出現於四種刊物中的「臺灣」、「臺灣民族」、「政治路線」、「國民黨」、「資產階級」等5詞語；出現於三種刊物中的「臺灣左派」、「臺灣革命」、「政治運動」、「革命」、「許信良」、「無產階級」及「解放運動」等6個詞語。

表5 各刊重要人工關鍵詞語

序號	台灣人民	台灣革命	台灣時代	台灣思潮	台灣解放	海外政論	台灣天地	建台
1	社會主義	臺灣革命	國民黨	經濟	臺灣左派	許信良	許信良	革命
2	革命運動	政治路線	革命運動	社會階級	政治路線	臺獨聯盟	全美臺灣同鄉會	社會主義
3	階級鬥爭	無產階級	革命	就業人口	左雄	國民黨	民進黨	階級
4	資產階級	國民黨	許信良	編者文字	學生運動	讀者投書	政治運動	臺灣
5	帝國主義	臺灣左派	帝國主義	革命科學	馬克思主義	臺灣民族	會長選舉	蔣幫
6	資本主義	帝國主義	臺灣革命	製造業	資產階級	海外臺灣人	留學生	辜寬敏
7	勞動人民	民族壓迫	社會主義	經濟發展	無產階級	臺灣革命	革命運動	資產階級
8	臺灣人民社會主義同盟	資產階級	美國	社會發展	社會主義	統治思想	楊黃美幸	
9	右派	臺灣問題	臺灣左派	資本	臺灣民族	世臺會	左派	
10	兩岸關係	機會主義	臺灣民族	庸屬發展	解放運動	海外臺灣運動	國民黨	
11	民族革命	馬列主義者	中國	無產階級	革命運動	黨外運動	同鄉會	
12	啟事	小資產階級	臺灣時代	革命		康寧祥	帝國主義	
13	民族民主革命	臺灣政論	讀者投書	批評		統治哲學	革命黨	
14		馬列主義	民主運動	臺灣		臺灣人	臺灣民族	
15		社會主義	郭雨新	政治發展		郭雨新	臺灣人	
16		解放運動	臺獨聯盟	政治經濟學		政治路線	康寧祥	
17		美國	選舉	階級理論		封建思想	黨外運動	
18		階級立場	政治運動	書評		政治討論	臺灣獨立	
19		階級鬥爭	資本主義			陳鼓應	臺灣	
20		獨立聯盟	政治路線			軍國帝國主義	政治討論	
21		蘇聯	左派			政治活動		
22		機會主義者	臺灣獨立			革命運動		
23		國際主義				文化		
24		愛國主義				政治運動		
25		葡萄牙				政治遊說		
26		革命理論				FAPA		
27		貢薩維斯				解放運動		
28		越南				同鄉會		
29		工人運動				帝國主義		
30		工人政黨				臺灣		
31		革命組織						
32		人民革命黨						
33		民主革命						
34		和平解放						
35		革命運動						
36		革新保臺論						
37		安哥拉						
38		新臺灣人民派						
39		左派聯合						
40		馬克思主義						

表6 重要人工關鍵詞於各刊之分布

刊名 / 關鍵詞	台灣人民	台灣革命	台灣時代	台灣思潮	台灣解放	海外政論	台灣天地	建台
臺獨聯盟			◎			◎		
臺灣				◎		◎	◎	◎
臺灣人						◎	◎	
臺灣左派		◎	◎		◎			
臺灣民族			◎		◎	◎	◎	
臺灣革命		◎	◎			◎		
臺灣獨立			◎				◎	
左派			◎				◎	
同鄉會						◎	◎	
社會主義	◎	◎	◎		◎			◎
帝國主義	◎	◎	◎			◎	◎	
政治討論						◎	◎	
政治路線		◎	◎		◎	◎		
政治運動			◎			◎	◎	
美國		◎	◎					
革命			◎	◎				◎
革命運動	◎	◎	◎		◎	◎	◎	
馬克思主義		◎			◎			
國民黨		◎	◎			◎	◎	
康寧祥						◎	◎	
許信良			◎			◎	◎	
郭雨新			◎			◎		
無產階級		◎		◎	◎			
階級鬥爭	◎	◎						
解放運動		◎			◎	◎		
資本主義	◎		◎					
資產階級	◎	◎			◎			◎
黨外運動						◎	◎	
讀者投書			◎			◎		
合計	6	13	18	3	9	18	14	4

（二）自動關鍵詞

自動關鍵詞的目的是透過資訊科技的協助，從文本中自動擷取篩選出能代表內容意義的詞語，大幅減少專家解讀的人力與時間成本。然而中文文本的處理，要面對中文斷詞、詞性判斷、未知詞辨識、複詞組合等操作面的困難與障礙，因此，技術層次及文本內容的特性都會影響到自動關鍵詞產出的品質與適當性。本研究在各階段的資料產出說明如下：

如表7所示，《台灣解放》每篇文章的平均字數最多，是長篇論述型的刊物；而《台灣天地》的平均字數最少，推論該刊物以短篇文章為主。斷詞及詞性標注之處理結果如表8所示。總斷詞數為使用CKIP斷詞工具後之詞語個數，單詞為本研究根據詞性條件篩選所得之詞語個數，複合詞則是參照專家註解之人工關鍵詞的詞性組合篩選而得之詞語個數，詳細詞性列表請見附錄。

表7 各刊內容之全文統計

刊物名稱	期數	文章篇數	總字數	平均字數／篇
台灣人民	10	142	477,269	3,361
台灣革命	5	55	239,383	4,352
台灣時代	12	279	938,369	3,363
台灣思潮	8	67	553,441	8,260
台灣解放	2	12	119,635	9,970
海外政論	16	227	613,856	2,704
台灣天地	15	163	259,577	1,592
建台	1	10	25,926	2,593

表8 各刊內容斷詞篩選統計

刊物名稱	總字數	總斷詞數	篩選後單詞個數	篩選後複合詞個數
台灣人民	477,269	239,600	19,120	18,620
台灣革命	239,383	104,718	8,503	7,890
台灣時代	938,369	570,100	25,986	31,115
台灣思潮	553,441	330,558	21,091	18,363
台灣解放	119,635	22,378	6,025	5,292
海外政論	613,856	364,272	20,482	18,156
台灣天地	259,577	165,865	11,593	9,076
建台	25,926	10,624	3,272	1,260

本研究文本經過全文斷詞及詞語篩選之後，結果共得91個詞語，如表9所示。但其中有5個無意義的詞語，「這」、「一」、「個」、「有」及「上」重複出現於各刊中，若剔除無意義及重複詞語，最後只剩下29個重要自動關鍵詞。依此將各刊重要自動關鍵詞交叉統計，觀察重要自動關鍵詞於各刊的分布，而同一詞卻同時出現在二種刊物以上者，共有9個詞語，如表10所示。顯示本研究文本中出現頻率最高的概念是「臺灣」，於八種刊物中皆是重要自動關鍵詞；其次是出現於六種刊物中的「運動」；出現於五種刊物的「革命」；出現於四種刊物的「臺灣民族」、「帝國主義」、「階級」、「資產階級」等；出現於三種刊物的是「臺灣人民」。

表9 各刊重要自動關鍵詞

序號	台灣人民	台灣革命	台灣時代	台灣思潮	台灣解放	海外政論	台灣天地	建台
1	臺灣	臺灣	臺灣	一	運動	臺灣	臺灣	階級
2	一	階級	一	階級	臺灣	這	運動	革命
3	階級	這	這	這	階級	一	一	臺灣
4	這	革命	個	個	個	有	有	我們
5	革命	一	革命	有	一	革命	個	一
6	個	個	運動	生產	社會	個	這	有
7	人民	主義	帝國主義	臺灣	左派	運動	民主運動	社會主義
8	資產階級	人民	臺灣人民	資產階級	發展	臺灣民族	臺灣民主	社會主義社會
9	臺灣人民	運動	臺灣民族		路線		臺灣民主運動	資產階級
10	帝國主義	臺灣人民			這		帝國主義	臺灣人民
11		資產階級			政治		臺灣民族	帝國主義
12		無產階級			上		臺灣獨立	
13		帝國主義			左派運動		國民黨政權	
14					資產階級		臺灣社會	
15					臺灣左派			
16					臺灣社會			
17					臺灣左派運動			
18					臺灣時代社			

表10 重要自動關鍵詞於各刊之分布

刊名 / 詞語	台灣人民	台灣革命	台灣時代	台灣思潮	台灣解放	海外政論	台灣天地	建台
臺灣	◎	◎	◎	◎	◎	◎	◎	◎
臺灣人民		◎	◎					◎
臺灣民族	◎		◎			◎	◎	
臺灣社會					◎		◎	
帝國主義		◎	◎				◎	◎
革命	◎	◎	◎			◎		◎
階級		◎		◎	◎			◎
資產階級		◎		◎	◎			◎
運動	◎	◎	◎		◎	◎	◎	

（三）比較分析

人工關鍵詞與自動關鍵詞各有其產生及使用上的限制。一般來說，人工關鍵詞的產生成本較高，其品質則取決於人力素質。自動關鍵詞在產生成本上具有優勢，尤其在大量文本或不同主題內容上更為明顯，但在詞語選擇的適切性上，需要更進階的資訊技術，才能有效提升品質。本研究試圖從幾個面向提出兩者之間比較的初步觀察：

1. 詞語數量及代表性

八種刊物中，人工關鍵詞共有約2,000個，去除重複後，計有約1,500個。而全文自動處理部分，總斷詞數約有一百八十萬個，符合詞性篩選條件的詞語數量則為數十萬個。二者間以數量觀，其差異性頗大。因此，簡單的篩選條件所自動產生的詞語尚無法成為關鍵詞。若以表2的篩選條件，並去除重複詞語，人工關鍵詞共得114個重要關鍵詞，而自動關鍵詞則只得29個重要關鍵詞，若將門檻值降低為0.5%，則得70個重要關鍵詞。由此可知，人工關鍵詞相對於簡單技術產生的自動關鍵詞，較能適切的呈現文本中的顯著資訊而反映出核心內容。單以詞頻作為篩選之依據有其侷限性，難以充分展現該文本內容的各個面向，不若專家解讀後給予之關鍵詞，較能夠精準的詮釋該文本的內容主題。

2.核心詞語

政論性刊物主要是針對當時發生的事件報導或批評，甚至宣揚政治理念，因此，人工關鍵詞及自動關鍵詞二者同時出現的詞語，表示經專業人員及以全文分析產生的詞語，應是左派刊物的核心詞語，代表當時主要的思想概念，或是關切的焦點。二者皆出現的核心詞語，計有：臺灣、臺灣左派、臺灣民族、臺灣獨立、左派、民主運動、社會主義、帝國主義、革命、無產階級、階級及資產階級等11個詞語。由此即可浮現，當時左派刊物所關注的焦點是以臺灣為軸心，探討臺灣當時所面臨的情勢及左派刊物所關注的政治傾向。

3.意同詞異

有些詞語會因用法及斷詞之差異而無法併為同一詞，如：人工關鍵詞有「國民黨」，而全文詞語有「國民黨政權」，但此二者所指之意義應相同，即是指當時的執政者。人工關鍵詞的取法是以概念為取向，而自動關鍵詞則是呈現作者實際之用詞，故會有詞語用法歧異的情形出現。

4.專有名詞

中文自動斷詞不似英文，可以空格作為斷詞區隔，通常必須依賴內建之詞庫為依據。因此，詞庫之樣本大小及性質會影響到斷詞之正確性，尤其是專有名詞常無法正常顯現於詞語中。而本研究觀察到，樣本中常出現的人名，人工關鍵詞上有，但全文處理會被斷開成二詞，必須再以人為方式補救，才能顯示出正確詞語，這是中文自動斷詞上須特別留意的地方。

5.選詞之差異

人工關鍵詞是經專家解讀內容後產生的詞語，是以整篇文章之內容主題為選詞依據。但自動關鍵詞則是以文本上出現的文字為選詞依據，所以，就會出現人工關鍵詞於自動關鍵詞中找不到的情形。如「讀者投書」一詞，人工關鍵詞中出現頻率高，但於自動關鍵詞中卻沒發現，因為該詞語在文本中所佔的比例太低，未能達到篩選之門檻。

6.長尾效應

一般文本中使用的詞語出現的次數，通常是有幾個詞語次數特別多，但多數的詞語卻僅出現一次，即有所謂的長尾效應。本研究所探討的人工關鍵詞及自動關鍵

詞，都同時有這種現象，甚至是每一刊物皆有此現象。以《台灣人民》為例，關鍵詞詞頻的曲線如圖5及圖6所示；詞頻次數最高的前25個關鍵詞之詞頻曲線如圖7及圖8所示。

為了便於整體觀察歸納，本研究將上述幾個面向之比較分析重點整理為表11。

表11　人工關鍵詞及自動關鍵詞之重點比較

	詞語數量	詞語代表性	核心詞語	意同詞異	專有名詞	選詞差異	長尾效應
人工關鍵詞	少	佳	共同（11/114）	較少	可涵蓋	佳	有
自動關鍵詞	彈性（依篩選條件）	有限	共同（11/29）	較多	困難	技術限制	有（更顯著）

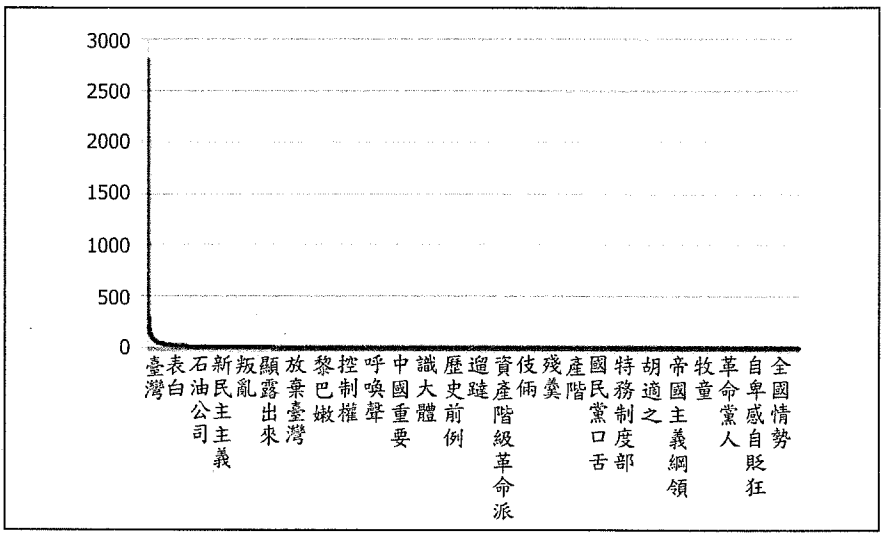


圖5　《台灣人民》自動關鍵詞詞頻分布

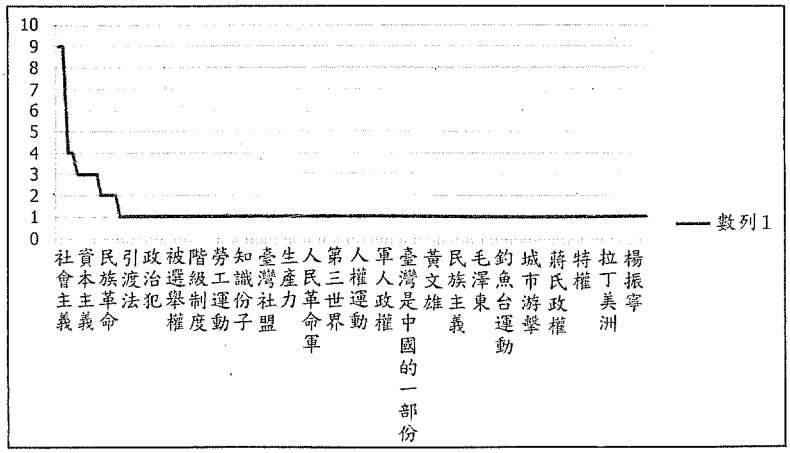


圖6 《台灣人民》人工關鍵詞詞頻分布

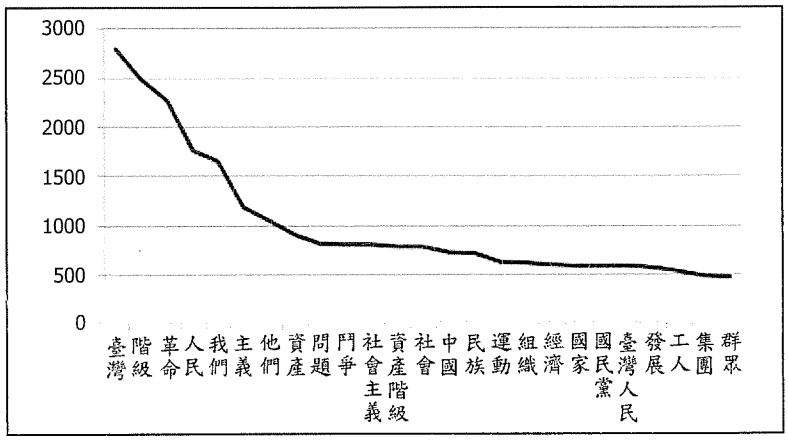


圖7 《台灣人民》自動關鍵詞前25名詞詞頻分布

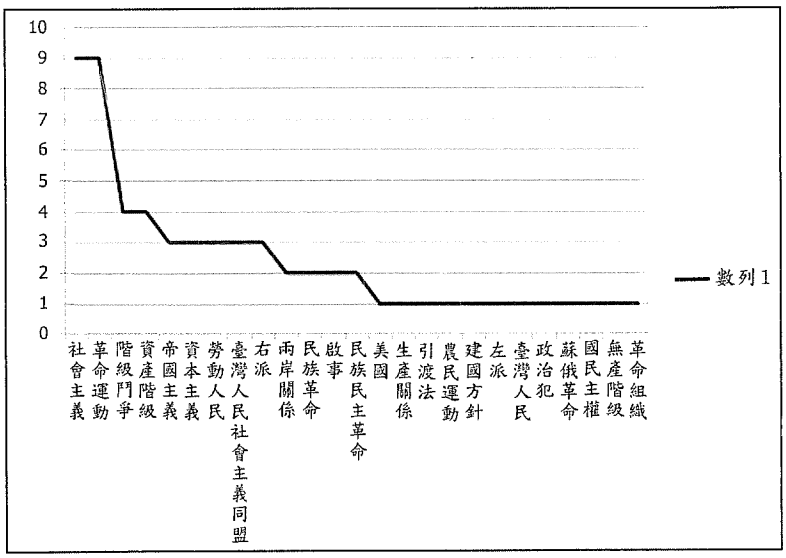


圖8 《台灣人民》人工關鍵詞前25名詞詞頻分布

五、社會網路分析

從上面研究，透過文本詞語分析，雖然可了解刊物內容所呈現出來的詞語使用狀況，及刊物彼此間的關係程度，但卻是平面的展現，對於一般人而言，缺乏更清楚之面貌。於是本研究進一步嘗試以社會網路分析方法，觀察個別刊物之詞語共現關係，並呈現局部與全體之結構資訊。本研究利用Vizster社會網路視覺化軟體工具，對於關鍵詞共現網路的結構觀察，進行探索性的檢視分析，並印證研究人員對內容的領域知識。

（一）人工關鍵詞共現網路

重要人工關鍵詞共有114個詞語，其共現網路如圖9所示，從圖中明顯可看出四個不同區塊。檢視所群聚的詞語，對照表5之各刊物重要人工關鍵詞語，由順時鐘方向依序檢視群聚的詞語。右上角以「小資產階級」為中心的區塊中的詞語大多屬《台灣革命》所有；右下角與「右派」相連的詞語區塊，則為《台灣人民》的重要詞語；中間偏右下方向者與「臺灣左派」、「左雄」相連的詞語為《台灣解放》；中間夾在「同鄉會」、「革命運動」、「國民黨」之間的區塊，則顯示《海外政論》、《台灣天地》、《台灣時代》；左下角獨立成球狀者為《台灣思潮》。上述人工關鍵詞的詞語群聚分布，可以視覺化方式清楚顯示《海外政論》與《台灣天地》所關注的議題與概念相近。而《台灣時代》則與《台灣革命》部分相似，另一方面也與《海外政論》與《台灣天地》接近，對照左派的刊物演變脈絡，可推論《台灣時代》的關注焦點的確從社會主義的討論逐漸轉移至海外臺灣人運動組織相關時事。

（二）自動關鍵詞共現網路

本研究篩選連結強度為前一百名之自動關鍵詞，建置其共現網路如圖10所示，由於此自動關鍵詞單以詞語出現的頻率為計算基準，因此出現在共現網路中的詞語必定是詞語配對反覆出現在同一段落之中，明顯與人工關鍵詞不同。再深入檢視該詞語性質，可發現能夠納入自動關鍵詞共現網路的詞語，其性質多為寫作行文時的針對性用詞，例如「反霸」、「反修」、「反帝」、「反左」等詞，若非熟知該主題領域用語或情境（context）的研究者，難以立刻知道這些詞語是「反對霸權主義」、「反對蘇聯修正主義」、「反對帝國主義」、「反對左派運動」。換言之，若非該領域之研究者較難藉此共現網路觀察出該文本之主題涵蓋哪些面向，不若人工關鍵詞較能表現各刊文章主題意涵。

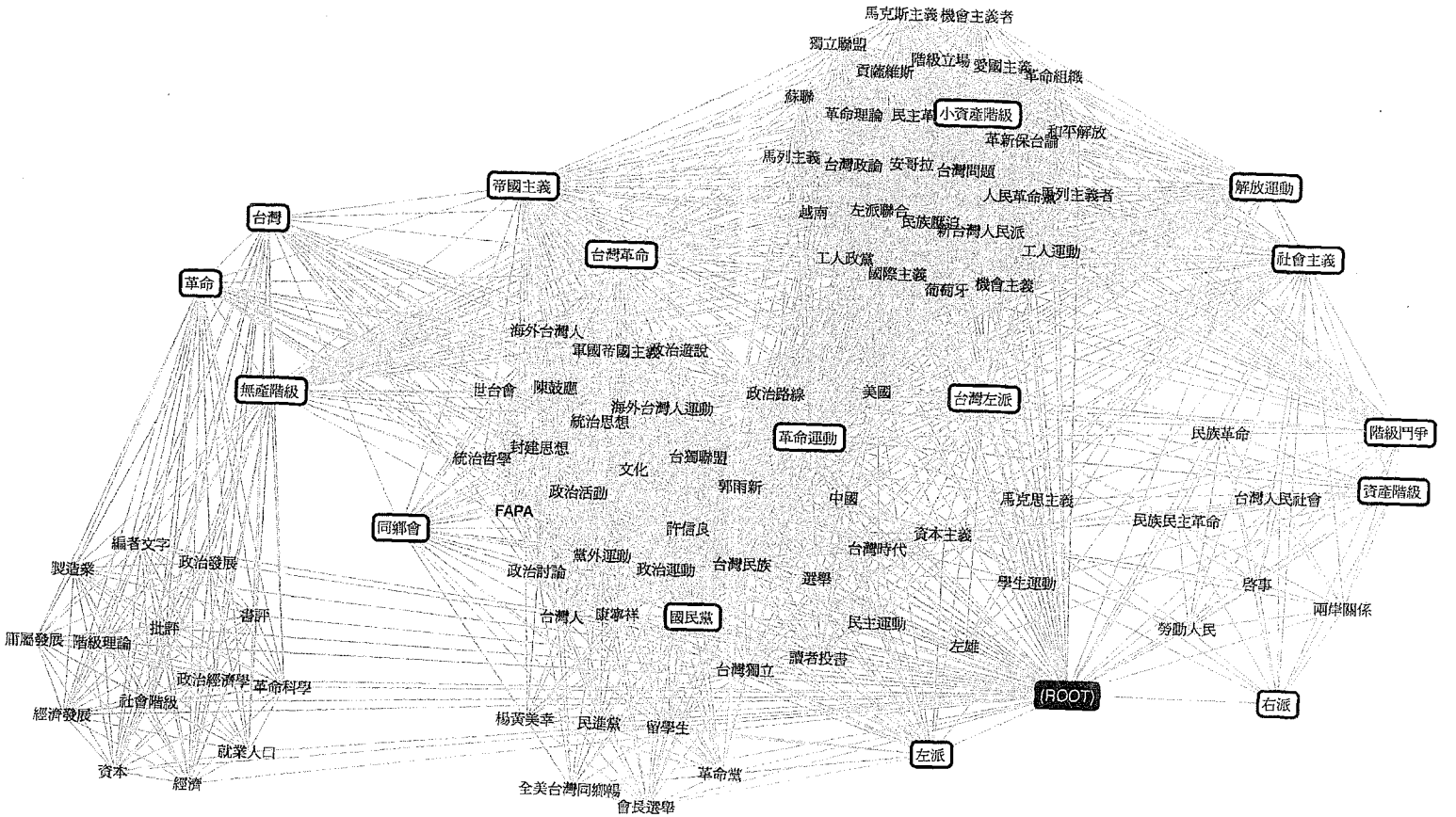


圖9　重要人工關鍵詞共現網路

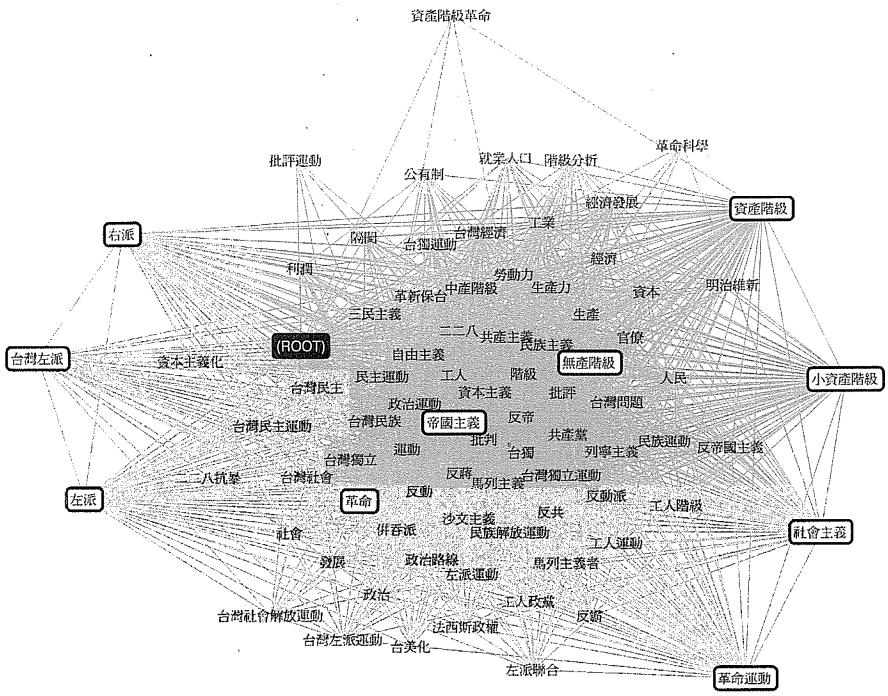


圖10 重要自動關鍵詞共現網路

（三）兩刊之共現網路比較

本研究進一步以《海外政論》與《台灣時代》互相比較，觀察兩者的重要自動關鍵詞之詞語分布，兩刊的共現網路如圖11、圖12所示，確實可呈現其刊物本身用詞的歧異，例如《海外政論》的共現網路即呈現放射狀，各關鍵詞彼此之間的連結較少，顯示其刊登之文章主題較為龐雜；而《台灣時代》之共現網路則呈現球狀，關鍵詞之間的關聯性較強，顯示其文章主題較為集中。

進一步檢視兩刊共現網路中的詞語，可清楚看出兩刊所關注的議題明顯不同。《海外政論》浮現較多的海外與島內的黨外人士和組織，或時事人物；《台灣時代》則較偏重社會主義的實踐與其他國家解放運動。對照兩刊的編輯群背景，《海外政論》是由臺灣出走的前礁溪鄉長張金策主導，較關注海外臺灣人組織以及臺灣島內現況；而《台灣時代》則由海外左派領導人物左雄為主，著重在社會主義的實踐。兩刊內容主題與重點之差異，雖然亦可由表5、表6、表9、表10的各刊重要關鍵詞語提供部分資訊，但透過網路視覺化工具對兩刊各自的共現網路的呈現，研究者可觀察到特定關鍵詞之間的顯著關聯，或關鍵詞關聯性之整體分布，而有助於對局部與全貌之情境資訊之掌握。

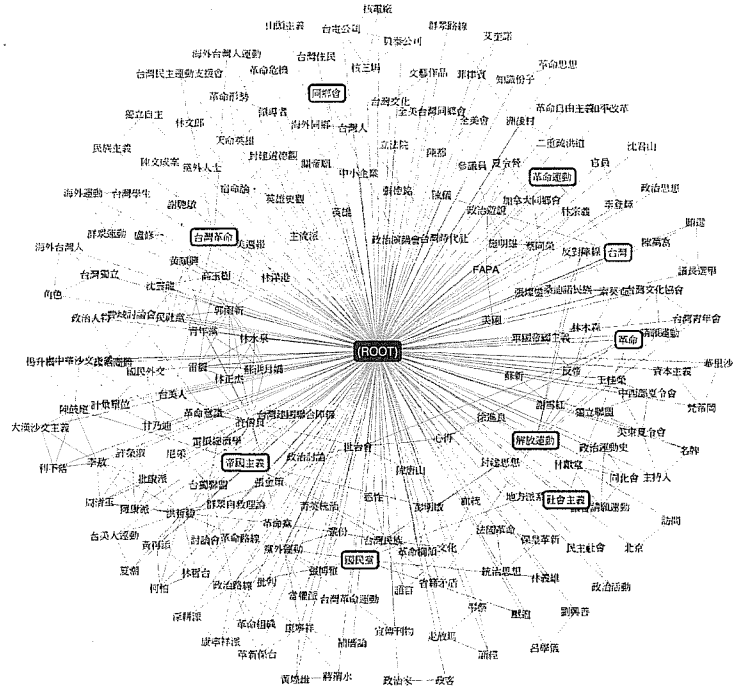


圖 11 《海外政論》之共現網路

圖 12 《台灣時代》之共現網路

六、結論

本研究透過詞頻統計與共現網路，分析《台灣人民》、《台灣革命》、《台灣時代》、《台灣思潮》、《台灣解放》、《台灣天地》、《海外政論》及《建台》八種左派海外刊物之內容，其結果可歸納以下幾點說明：

1.人工關鍵詞較能精準的呈現文本內容的核心資訊，也較能涵蓋完整的內容面向。相對而言，簡單技術（詞頻統計）下產生的自動關鍵詞所能反映的內容資訊較為狹隘，單以字詞頻率為依據，在真正的字義與資訊的完整性上皆有所限制，代表性不足，並且容易疏漏。因此，自動關鍵詞的使用仍有相當大的改進空間，需要更精良的詞語代表性判斷方法。

2.本研究範圍的左派刊物所關注的焦點是以臺灣為軸心，探討臺灣當時所面臨的情勢及左派刊物所關注的政治傾向，而經過詞語交叉比對後顯示「臺灣」、「臺灣左派」、「臺灣民族」、「臺灣獨立」、「左派」、「民主運動」、「社會主義」、「帝國主義」、「革命」、「無產階級」、「階級」及「資產階級」等11個詞語為核心詞語。

3.不論從詞語交叉比對或從共現網路圖所呈現的詞語分布結果，均顯示《台灣思潮》的主題最為集中，使用詞語與其他刊物差異最大，其使用詞語多為以革命科學的角度，進一步論述社會主義的內涵。

4.重要人工關鍵詞的共現網路可約略分成五個區塊：（1）《台灣人民》；（2）《台灣革命》與《台灣時代》；（3）《台灣思潮》；（4）《台灣解放》；（5）《台灣天地》與《海外政論》；而《建台》的詞語群聚則較不明顯，《台灣時代》則有部分詞語向區塊（5）靠攏。此一結果與刊物發展脈絡相符，顯示分析刊物文章的用詞，確實可反映其編輯群的政治理念。

5.各刊的共現網路的圖形，確實可呈現其刊物本身用詞的歧異，例如《海外政論》刊登文章主題較為龐雜，其共現網路即呈現放射狀；而《台灣時代》之共現網路則呈現球狀，顯示詞語之間的關聯性較強。

6.關鍵詞分析比較及共現網路視覺化呈現，兩者所提供之資訊雖有部分重疊，但仍有極高的互補性，而可互相印證支持。尤其是許多網路視覺化工具提供了相當具有彈性的檢視功能，可依照使用者的需求或興趣，對網路面貌或組成，提供各種角度與結構的觀察方式。因此，建議未來的文本分析研究，仍應善用共現網路的研究工具，擴大文本內容解讀與發掘的空間。

臺灣海外政治運動的意識型態與認識程度，直接表現於海外組織刊物中，從詞頻分析各刊的重要詞語相關性與社會網路模型所呈現的詞語分布及群聚現象，直接

印證研究團隊從文獻中得知之左派刊物演變脈絡，有助於研究者迅速的及直觀的對於左派刊物關注的重要議題或概念，有概括性的了解，並有機會進一步分析其時代意義及所代表的思想層面與趨向，有助於體會臺灣獨立運動的原因動機，及其演變過程的主觀要素和路線發展。

除了利用統計工具進行分析外，數位文本也根據其主題或性質分割成微型顆粒資料，配合適切的演算法與篩選機制，有別於在傳統線型資料的呈現，重新建構資料間的關係，能清楚看到各資料節點間的彼此關係，由視覺化的工具輔佐提供理解數位文本的新視角。以往人文學者必須窮其一生自行蒐集與解析史料，但資訊科技的進步，不僅可以協助整理與分析，而數位化的史料結合新的數位人文分析工具，能夠有效浮現過去因資料分散而隱沒在字裡行間的關聯。本研究在共現網路及社會網路分析的技術方法應用上，只完成了初階的嘗試。在共現網路的視覺化工具使用上，尚未充分嘗試各種檢視的可能，也因此只能得到有限的觀察結果。另外，在社會網路分析中常用到的節點指標與結構分析，本研究也未及完成。這些都將是本研究下一階段的延伸方向與努力目標。

附錄 本研究採用之詞性

精簡詞類	簡化標記	對應的CKIP詞類標記	
A	A	A	/*非謂形容詞*/
N	Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
N	Nb	Nba, Nbc	/*專有名稱*/
N	Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
N	Ncd	Ncda, Ncdb	/*位置詞*/
N	Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
N	Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
Nv	Nv	Nv1,Nv2,Nv3,Nv4	/*名物化動詞*/
Vi	VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
Vt	VAC	VA2	/*動作使動動詞*/
Vi	VB	VB11,12,VB2	/*動作類及物動詞*/
Vt	VC	VC2, VC31,32,33	/*動作及物動詞*/
Vt	VCL	VC1	/*動作接地方賓語動詞*/
Vt	VD	VD1, VD2	/*雙賓動詞*/
Vt	VE	VE11, VE12, VE2	/*動作句賓動詞*/
Vt	VF	VF1, VF2	/*動作謂賓動詞*/
Vt	VG	VG1, VG2	/*分類動詞*/
Vi	VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
Vt	VHC	VH16, VH22	/*狀態使動動詞/
Vi	VI	VI1,2,3	/*狀態類及物動詞*/
Vt	VJ	VJ1,2,3	/*狀態及物動詞*/
Vt	VK	VK1,2	/*狀態句賓動詞*/
Vt	VL	VL1,2,3,4	/*狀態謂賓動詞*/
Vt	V_2	V_2	/*有*/
Vt	SHI	/*是*/	

參考文獻

中文詞知識庫小組，n.d.，中文斷詞系統，上網日期：2011年8月26日。網址：http://godel.iis.sinica.edu.tw/CKIP/onlinesystem.htm#t2。

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。

林照真，2008，〈探尋臺灣左眼的世界〉，上網日期：2011年8月26日。網址：http://www.wretch.cc/blog/powerslide/11060594。

邱偉雲，2010，〈關鍵詞叢與文本意義挖掘的嚐試：以《清季外交史料》為例〉，項潔等主編，《數位典藏與數位人文國際研討會論文集》，臺北：國立臺灣大學數位典藏研究發展中心。

侯榮邦，2005，〈台獨聯盟日本本部發行之刊物〉，張炎憲、曾秋美、陳朝海編，《自覺與認同：1950-1990年海外台灣人運動專輯》，臺北：吳三連臺灣史料基金會，頁575-586。

洪一梅，2009，〈人文學術研究的數位新時代——史語所的思維與作為〉，《古今論衡》，20，頁133-154。

張清水，1987，〈左雄路線運動學理主義的回顧與批判：台灣左派運動問題檢討之一〉，《台灣解放》，1，頁2-15、38。

陳光華、伍健廷，1998，〈控制詞語之自動索引〉，《中國圖書館學會會報》，61，頁81-102。

陳良駒、張正宏、陳日鑫，2007，〈以特徵詞共現特性探討知識管理研究議題相關性——使用共詞與關聯法則分析〉，《資訊管理學報》，17（4），頁31-60。

陳明來，2002，〈主題分析在圖書資訊組織之角色探討〉，《圖書與資訊學刊》，42，頁69-84。

陳詩沛，2011，《資訊技術與歷史文獻分析》，國立臺灣大學資訊工程學研究所博士論文。

許維德，2001，〈發自異域的另類聲響——戰後海外台獨運動相關刊物初探〉，《台灣史料研究》，17，頁99-155。

鍾才，1994，〈戰後台灣留日學生的獨立建國運動史〉，《台灣史料研究》，4，頁107-118。

藍適齊，2002，〈再探討戰後海外台獨運動相關刊物及「海外台灣人史」〉，《台灣史料研究》，18，頁99-109。

Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V. D. P., Loreto, V., Hotho, A., Grahl, M., & Stumme, G. (2007). Network Properties of Folksonomies. *AI Communications, 20*(4), 245-262.

Chan, L. M. (1994). *Cataloging and Classification: An Introduction* (2nd ed.). New York: McGraw-Hill.

Heer, J. and Boyd, D. (2005). Vizster: Visualizing Online Social Networks. *Proc. IEEE Symposium on Information Visualization*, 32-39.

Presner, T. (2010). *Digital Humanities 2.0: A Report on Knowledge.* Retrieved September 6, 2010, from http://cnx.org/content/m34246/1.6

Otte, E. & Rousseau, R. (2002). Social Network Analysis: A Powerful Strategy, Also for the Information Sciences. *Journal of Information Science, 28*(6), 441-453.

Özgür, A., Cetin, B. & Bingol, H. (2008). Co-occurrence Network of Reuters News. *International Journal of Modern Physics C, 19*(5), 689-702.

Scott, J. (2000). *Social Network Analysis: A Handbook* (2nd ed.). London: Sage.

Su, H. N. & Lee, P. C. (2010). Mapping Knowledge Structure by Keyword Co-occurrence: A First Look at Journal Papers in Technology Foresight. *Scientometrics, 85*, 65-79.

Wasserman, S. & Faust, K. (1994). *Social Network Analysis: Methods and Applications.* New York: Cambridge University Press.

Part 2
語料庫語言學
Corpus Linguistics

結合漢典古籍虛詞常見字與統計量化分析進行漢譯佛典譯者風格辨別

謝承恩*、洪振洲**、馬德偉***

摘 要

佛教於東漢由印度傳入漢地之時，記載佛教義理的佛經主要仍為印度的梵語或是中亞語言所撰寫。由東漢至唐中葉的數百年間，佛教盛行於中國，因而有了大規模的佛經翻譯活動。透過這些佛經翻譯的活動，不僅產生了巨量的漢譯佛典，也進一步影響了中國的文化，甚至是語言的發展。但是，早期的佛經翻譯由於受到傳抄、戰亂、偽託等現象影響，使得譯者紀錄出現許多錯誤，這也連帶造成相關研究者的困擾。為找出正確的佛典翻譯者，許多佛學研究者使用傳統文獻學之方式，提出新的證據與看法。然而傳統文獻學之研究方式十分倚賴人工判斷與處理，不僅耗時費工，且無法處理大量文獻資料。

在現今資訊科學的幫助下，以數位化資料及統計量化方法進行資料的比對分析，已是現今人文資訊學的一大趨勢。此類方法不僅可以進行大量資料比對，也能夠找出譯者風格的潛在因子，這是傳統文獻研究方式難以達成的工作。而此方面的研究，目前仍以英文文獻為主要研究對象，以類似方法運用於古代中文文獻的效果仍未被正確的評估。因此本研究採用常見的古代中文的虛詞作為風格特徵之評斷指標，並搭配多變量統計分析手法中的主元素分析法，進行譯者之風格分析。為評估本方法之效用，本研究設計多種不同的實驗情境。根據我們的實驗結果，此種方式能成功的區分出作譯者的翻譯風格，協助研究者進行譯者紀錄的判斷。

* 法鼓佛教學院佛學資訊組研究生。
** 法鼓佛教學院圖書資訊館館長。
*** 美國天普大學助理教授。

Authorship Attribution of Early Chinese Buddhist Translations: Using Principal Component Analysis with Commonly Used Ancient Chinese Empty Words

Cheng-en Xie *, Jen-jou Hung **, Marcus Bingenheimer ***

Abstract

The Taishō edition of the Chinese Buddhist canon (1924-1932) contains ca. 1000 Indian texts translated into Chinese between the 2nd and the 11th century CE. Up to 153 of these texts are marked as "失譯", indicating that the name(s) of the translator(s) are unknown. Consequently, translator attributions from the beginning of the Tang dynasty, i.e., from the early 7th century onwards, are relatively reliable, but those for texts translated between the 2nd and the late 6th century have uncertain, problematic, or wrong attributions. We are reasonably confident that the 49 sutras ascribed to Kumārajīva (344-413 CE) were actually produced by him and his team (although this needs further research), but the attributions for many of the other texts translated before the late 6th century likely wrong.

Over the years, Buddhist scholars have leveraged traditional text-critical methods to corroborate or dispute traditional attributions. Although these methods can produce high quality results, they often rely heavily on the intuition of a single scholar honed over many years of research.

Information technology offers an alternative dimension of inquiry that aims to complement rather than supersede the more traditional approaches. To accomplish this, statistical methods, quantitative methods, and artificial intelligence algorithms were adopted to

* Graduate Research Assistant, Dharma Drum Buddhist College.
** Assistant Professor, Director of Library and Information Center, Dharma Drum Buddhist College.
*** Assistant Professor, Temple University, U.S.

analyze ancient Buddhist texts translated into Chinese to discover new evidence to address the problem of translator attribution.

The major advantage of stylometrics and quantitative authorship attribution is the capability to discover hidden patterns that cannot be discerned through traditional approaches. In the past four decades, considerable attention has been focused on the quantitative authorship attribution of literature in Western languages. However, few attempts have focused on texts written in classical Chinese, much less to the "Indian Buddhist Chinese" of early-translated texts. In the present paper, the focus is on the empty word (*xuci* 虛詞) widely used in classical Chinese to express grammatical relations. After measuring their occurrence in Indian Buddhist Chinese, principle component analysis is employed to determine how their use reflects the authorship of some selected sutras, especially the three sutras attributed to Zhu Fahu (竺法護) (trad. 231-308) the *Xuzhen tianzi jing* (須真天子經) (T. 588), a version of the famous Lotus Sutra the *Zhengfahua jing* (正法華經) (T. 263), and the *Puyao jing* (普曜經) (T. 186). We have developed an algorithm to help describe and distinguish the translation style of different translators. The analysis explores different scenarios that need to be accounted for, such as the changes in translation style during the course of a translator's career, an understanding of the commonalities between contemporaneous translations, and quantification the differences between different translations of the same sutra.

一、簡介

佛經自印度傳來漢地，所使用的文字原為印度的梵語，或是當地的方言。對於漢地的人民來說，都是未曾見聞的外語。為使漢地的人民能理解佛經的義理，譯師所進行的佛經翻譯工作便顯得相當重要。早期的佛經翻譯紀錄，在受到傳抄、戰亂、偽託等情形影響之下，使得許多經錄的記載有所出入，例如：失譯及誤判譯者的現象。此外，在同經異譯本之間，作者可能大量互相參考翻譯成果的現象，也都成為現代進行譯者風格研究學者的一大困擾。因此，許多文獻學者以此研究方向為其志業，透過使用文獻學、語言學的方法，提出許多嚴謹且精美的研究成果，成效可說是相當卓越。然而，傳統研究方法所需的時間漫長，不易廣泛的針對不同對象進行研究，一直是傳統研究方法的一項缺憾。

現今，在資訊科學與有心人士的幫助下，大量的佛經資料已經完成數位化，並且提供免費使用。因此，我們可以輕易的取得大量的數位佛經內容，並以程式取代人工進行統計比對分析，所能處理的資料量遠超過傳統文獻方法，處理的速度更是人工作業的數百倍。簡言之，使用量化分析方法的分析具有下列好處：一、更快速且正確的計算：過去如果要計算某一詞語在《大藏經》中的出現頻率，以人工而言幾乎是不可能的任務，現今透過數位化的資料與電腦程式，我們可以快速的得到正確的結果。二、找出決定譯者風格的潛在因子：透過比對大量資料，再結合學者的研究方法，設計出適合的演算法，可以獲取譯者遺留於作品中的線索。

近年來，部分歐美地區的學者在作譯者風格的相關研究中，採用與文本情境內容無關的詞作為作者風格特徵。這類的詞，稱之為非語境詞（non-contextual nouns），如英文中的：a、be、do、for、if等詞（Zhao & Zobel, 2007）。也就是說，當作者使用此類型單詞時，大多是依據個人寫作習慣，而非取決於文本所記載的內容。因此，非語境詞能夠忠實的反應出作譯者的個人寫作特色。同樣的，中文字詞裡也存在著非語境詞，這類的詞語通常並不具有明顯的內容意涵，而幾乎都是詞義已虛化的代名詞、助詞。此一現象在古籍中更為明顯，這是由於漢文古籍並沒有現代標點符號的概念，所以在句子上必須仰賴此類字詞作為斷句。

因此，本研究採用漢典古籍虛詞的常用字作為作譯者風格的特徵值，透過主元素分析法（Principal Components Analysis，簡稱PCA）（Jolliffe, 2002），實作漢譯佛典古籍的譯者風格區分模型。我們的模型使用自訂的公式，計算出該實驗中各部經文的古籍虛詞出現頻率的差值，再以此差值降冪排序，取得前30個出現頻率具有最大變化的古籍虛詞，作為實驗所使用的特徵衡量詞。接著，依此特徵衡量詞對樣本進行統計，計算出各樣本內特徵衡量詞所佔百分比，投入主元素分析運算。最後，將透過主元素分析法運算之結果投影繪圖，並對結果進行判讀，以確認本研究模型的效用。

後續論文章節之規劃如下，在第二節中，我們將回顧近年來相關領域的研究成果，範圍包括了資訊、人文與跨領域合作之研究。第三節說明經文處理與資料分析的步驟，包括選擇經文的標準，與資料來源的介紹。第四節則對模型設計數個實驗，實際運作與分析，驗證模型的可行性。第五節則對本研究做出整體性的結論。

二、文獻回顧

語言學的研究方法中，經常使用文本中的詞彙、語法與句型來進行語言風格研究。詞彙方面，研究者通常選定某一種詞性的詞彙下手，如：林昭君（1998）在《東漢佛典之介詞研究》中，分析了東漢著名的佛典譯者們使用介詞的情況，比對他們在介詞使用的異同、頻率與用法，總結出東漢佛典介詞的共時特色。而汪禕（2008）的《中古佛典量詞研究》，則是利用佛典裡所記載的量詞，考證出中古時期量詞的使用情況，進一步了解佛典在運用量詞時與漢語之間的關係為何。此外，劉芳薇（1995）的《〈維摩詰所說經〉語言風格研究》，則跳脫詞彙層次，而從句型分析上討論其翻譯風格特點，並以虛詞的使用探討該經的口語化特色。

近年來，國內的語言學研究在學者竺家寧、萬金川等人的推動下，有愈來愈多的研究採用佛教文本作為材料，形成一股「佛經語言學」的研究趨勢。此類研究不限於文本本身的內容，更延伸到文本的歷史背景、語言演化等。如：高婉瑜（2006）的《漢文佛典後綴的語法化現象》，根據漢文佛典後綴詞的分析，得出漢語詞綴與派生構詞的意義。蔡佳玲（2007）的《漢地佛經翻譯論述的建構及其轉型》，透過整理譯者僧傳與經序，尋找出譯經時代背景與譯經間的關聯。廖宜君（2008）的《東漢佛典之雙音節動詞研究》中，將東漢佛典中的雙音節動詞，以詞義與構詞兩方面進行分析，了解其結構特色與在漢語語法史上的意義。另外，也有從語音角度切入的研究，如：張嘉慧（2008）的《法雲〈翻譯名義集〉的語言研究——以音寫語段的分析為中心》一文，使用「語音學」和「詞彙學」方法，對《翻譯名義集》進行音理與構詞的分析，得到梵語轉為漢語的音寫用字規則。

在國外的部分，使用統計量化方式，進行文章內容分析，分析文章語言風格，進而推論出較可能的作譯者方式，也是近年來的熱門研究方向之一。實際上早在20世紀末期，Mosteller與Wallace（1984: 158-164）就已經嘗試使用貝式推論分析（Bayesian inference）的方式，來判斷12篇具有爭議性著作權的Federalist Papers文章。近年來由於資訊技術的進步，數位化文獻大量出現。在大量的材料輔助之下，以電腦輔助執行量化譯者分析的研究也開始蓬勃發展，並且受到各界的矚目。作譯者量化分析的研究，多半可以理解成將一個未知作者的文獻，歸類到幾個預先設定好的作者分群之一的分類問題。在實務上，於實際進行分析之前，我們多半可以利用具公信力的外在證據，例如在相關的文獻紀錄中，鎖定幾個可能的作者。藉由蒐集並

分析這些候選作者具有代表性的作品，將未知的對象與這些確認過的作品相比較，以判斷較為可能的作譯者，此即量化譯者分析的主要模型。換句話說，一個有效的、可量化比較兩個文獻是否相近的計算方式，成為此類研究的重點。

綜觀以上眾多研究者所提出的相關研究方法，大多包含以下兩階段的處理方式：首先處理數位文獻內容，並萃取出文章的基本特徵值，然後將這些特徵值利用統計或人工智慧的方式，來判斷出較為可能的作譯者。決定文章特徵值的第一考量，就是希望能夠擷取出與文章內容無關，但卻是作譯者遺留在文章中的線索。為達此目的，許多研究使用與內容無關的特殊功能單字，或是在文章中最常使用的n個字的出現頻率為風格衡量值（Holmes & Crofts, 2010; Hoover & Hess, 2009），此種方式十分直觀、簡單。但若處理不佳，可能讓分類結果偏向於內容分類而非作者風格比較。而另一種常見的作法，就是使用文章內容的字彙豐富度與句子的長度（Grieve, 2007）等隱含訊息來作為文章特徵值的計算方式。由於作者在寫作過程中，較少會意識到使用的詞彙量與語句長短的變化，因此以上兩者皆可視為具有相當效果的寫作風格衡量值。除上述所使用簡單文字的統計量化之外，在許多研究中還會將文章內容預先進行自然語言處理（Nature Language Processing），以取得文章在文法或詞性的變化，並利用這些較為高階的語法資料來進行後續分析的材料，以求得更精準的分析結果（Zhao & Zobel, 2007; Stamatatos, Fakotakis & Kokkinakis, 2001）。另一種較特別的文章特徵抽取方法，是使用n-gram的方式來切割文章（Houvardars & Stamatatos, 2006; Liu, Allison, Guthrie & Guthrie, 2007; Peng, Schuurmans, Keselj & Wang, 2003; Hung, Bingenheimer & Wiles, 2009）。n-gram 的演算法是忽略文字在語法上的組合，而將文章視為一個十分綿長的字串，並依序切割為長度為n的小塊，由其中取出相關的統計值以便進行分析。這種完全忽略文字結構的切割方式，看似沒有道理，但卻可以解決在缺乏直觀字詞邊界的東方語言（如中日韓等語言）中詞彙不易被正確切割的問題。

用來比對特徵值的統計方法也有許多選擇，基本上可以簡單歸類為傳統統計式以及人工智慧演算法等兩大類。常見使用的統計方法，有：貝式分析（Bayer Classifier）（Bozkurt, Ba ho lu & Uyar, 2007）、多變異分析（Stamatatos, Fakotakis & Kokkinakis, 1999）、主元素分析法（Labbe, 2007）、分辨分析（Tambouratzis & Vassiliou, 2007）等方式。而人工智慧學門中經常用來進行分類的演算法，包括：支撐向量機（Support Vector Machine）（Zhang & Lee, 2006）、類神經網路（Kjell, 1994; Tearle, Taylor & Demuth, 2007），及NSC（nearest shrunken centroid）（Jockers, Witten & Criddle, 2008; Schaalje, Fields, Roper & Snow, 2011）等方式，也常被應用到量化作譯者分析的問題之中。

一般而言，以量化方式進行作譯者分析的研究成果，通常因為與傳統語言學的分析方式有很大的不同，而導致人文學者對於量化分析的結果仍抱持著存疑的態

度。因此相關研究成果，不一定會被人文研究引用成為重要的參考證據。這也是以量化方式進行作譯者分析研究所面臨的困難之一，然而此一現象近年來開始有相當程度的改變。Hoover 與 Hess（2009）所發表的研究成果之中，作者嘗試去解決一個因為線索太少而無法使用傳統文獻學推論方式得到確切結果的問題，因此作者轉向使用統計量化分析方式如：PCA、T-testing與Delta（Burrows, 1992）分析，來試圖找出可能的真實作者。根據研究結果，量化分析方法所建議的作者，與一般認為最有可能的候選作者相符，並且，量化分析結果也呈現了許多傳統作法忽略的證據。而Holmes與Crofts（2010），則同時使用傳統文獻的方式與量化分析的手法來處理同一個作譯者分析的問題。雖然兩種分析手法差異頗大，但結果都指向同一個可能作者。透過這些研究的實例，也讓我們更進一步確定量化作譯者分析方式的可靠度。

三、經文處理與資料分析

為取得具有高準確性的譯經風格分析結果，首先我們必須蒐集具有可靠譯者資訊的文本來源作為樣本，以建立可靠的比較對象。再由這些樣本中擷取出譯者的風格，並利用適當的統計方法分析，以便找出可能性最高的譯經者。因此，在本研究中，我們分析的過程分為下列三步驟進行：一、依據現有的文獻研究成果，選定適合的實驗對象，並採用具有可靠譯者紀錄的文本來源作為比較樣本。二、以漢文古籍中常見的虛詞，對各個樣本進行統計，建立該組實驗所需的特徵值。三、最後套用主元素分析法，產生分析結果，將運算結果投影繪圖，並輔以現有的文獻研究成果進行分析與解釋，以驗證分辨模型的效度。以下為各步驟的細節。

（一）文本來源與取樣

研究的第一步，我們必須先針對研究的目標譯者，收集其翻譯作品，建立出一個可靠的譯者翻譯作品集合。然而，此集合建立的過程是十分問題導向的，也就是說，要選取哪些經典作為風格比較對象，完全取決於所要研究的對象為何。因此，我們不容易給定一個絕對的標準。但一般來說，在此利用前人於文獻學上的心血結晶是一個不錯的選擇。本研究中，我們所挑選的實驗對象都是佛教譯經史上早期的譯者，如：東漢、西晉、三國等。而針對該時期的研究中，由呂澂（1991）、俞理明（1993）、許理和（1987）、Lewis Lancaster（2008）等四位學者的經錄研究成果最為廣泛且豐富，也常被近代學者引用於相關研究之中，因此我們使用這些文獻作為選擇目標時排除疑義的基準。本研究中對於經文的選擇，採用從嚴認定的方式。我們比對四位學者的研究成果，對於同一部經典的譯者紀錄，只要有一位學者持不同意見，我們就將該經剔除於實驗集合之外。根據上述標準，我們在第四節的每組實驗中，分別選用了三到五部的經文，作為分析比較之用。各實驗所選用的經文在

第四節將有詳細說明，在此先不贅述。

當用於比對的經典清單決定之後，我們使用中華電子佛典協會所製作的「CBETA 2010電子佛典集成光碟」[1]作為數位文本來源。CBETA自1998年成立以來，一向致力於佛典的電子化工作。CBETA電子佛典最初所採用的底本為日本《大正藏》，經十多年來，文獻人員不斷的校訂與標記，比對現存不同藏經版本，勘正多處錯誤，並依文本結構、意義進行標記，是品質相當良好的數位藏經資料來源。

在取得可靠的文本資料後，為符合統計方法的量化需求，我們必須將經文適當的進行分割，形成數個長度較短的樣本。切分為多個樣本後，便能藉由觀察同一樣本群的落點進行判讀，有驗證風格一致性與偵測錯誤之附加價值。在此研究中，我們將每部經依字數平均分為20個樣本，使同一樣本群裡的各樣本數目接近，減少因數目尺寸所產生的差異，如此一來，各樣本之間的差異更有可能是來自於內容，而非取樣所造成，也能避免取樣差異對實驗結果的影響。

（二）特徵選取

完成取樣後，為衡量各樣本之間的差異，我們需要設定一組特徵值，作為各個樣本之間比較的基準，以利進行後續分析。在此我們採用學界常使用的非語境詞頻率，亦即文本中的各個「虛詞」所出現的頻率作為特徵值。對於「虛詞」的定義，由於在詞類的劃分上有其難度，所以在傳統語言學的研究中仍尚未有統一的定義（戴綉娟，2010），在此我們不打算討論虛詞本身的爭議，而是直接參考王叔岷所著《古籍虛字廣義》所列231個古籍常見虛字，扣除破音字後，得到228字作為初步海選的虛詞清單。

為使分析結果更顯著，減低過多衡量值所形成的干擾，因此我們進一步由這些虛詞中，選擇文本出現頻率變化較大的30個詞彙，作為譯經風格的衡量詞。首先，我們計算出這228個虛字於各經中的出現頻率。我們利用S表示所有選出的經典的集合，而 s_i 表示單一經典，因此 s_i 可以表示為：

$$S = \{s_i \in S | i = 1,2, \ldots n\} \qquad (1)$$

我們以 F_j^i 表示虛字 j 在 s_i 中的出現頻率，因此，F_j^i 可以表示為：

$$F_j^i = \frac{WC_j^i}{L_i} \qquad (2)$$

1 Chinese Electronic Tripi aka Collection CD Version 2010 (Chinese Buddhist Electronic Text Association, CD-ROM, 2010 release). http://www.cbeta.org/

其中，WC_j^i表示虛字 j 於 s_i中出現次數，L_i表示 s_i的總字數。因此虛字 j 於所有經典集合 S 中出現的頻率平均值$\bar{F}_j$為：

$$\bar{F}_j = \frac{1}{n}\sum_{i=1}^{n} F_j^i \qquad (3)$$

最後，我們計算虛字 j 於所有經典中，其出現頻率F_j^i與平均值$\bar{F}_j$的差值總和 $diff_j$ 如下式所述：

$$diff_j = \sum_{i=1}^{n} \left| F_j^i - \bar{F}_j \right| \qquad (4)$$

最後，我們選用 $diff_j$ 值最大的30個虛詞，作為該組實驗的特徵衡量詞。這些特徵詞代表了樣本中差異最大的部分，同時也保留各樣本本身的特徵。之後，我們將計算在各文本樣本中，這些特徵衡量詞所出現的頻率，用以進行下一步驟中主元素分析法的運算。

（三）套用主元素分析法與繪圖觀測

取得特徵值後，必須經由下一步的分析才能觀察出其差異。在此我們採用主元素分析法。主元素分析法是一種非監督式學習（unsupervised learning）的分析方式，在演算過程中，主元素分析法不需要資料預先分群的資訊。而其主要的目的，是將原本人工無法處理的高維度資料，降維（dimension reducing）成為可理解、具代表性的低維度資料。在本研究中，我們藉由主要特徵投影過後的資料之中，選取產出的第一與第二主元素的值，作為特徵依據，並將計算出的樣本投影點描繪至二維空間中，進行繪圖分析與比對。由於主元素分析法可將高維度資料降低至低維度空間，因此經由主元素分析法處理過的資料，所輸出的繪圖具有直觀、容易被解讀的特性，更利於人工判讀。此外，也因為主元素分析法本身沒有預先分群的概念，因此我們可以從繪圖中樣本的落點位置觀測出許多樣本本身的現象（Wold, Esbensen & Geladi, 1987），例如：樣本落點的接近程度、群聚現象、距離等，並用以辨別樣本之間的相似程度。

四、實驗與結果分析

本研究的目的在於利用虛詞與主元素分析法建立一個譯者風格辨析模型，用以辨別不同的作譯者風格，為確認本方法的效果，我們實驗了以下三種可能的分析情況，分別為：

1. 單一譯者的前期作品與後期作品

由於譯師的譯經時間相當長，其寫譯風格多少會受到時間、本身學養能力及當時譯經風氣影響而改變。而某些特定作者，又因為更多的特殊原因，如：學習了新的語言，因此造成關於單一譯者前、後期風格有十分明顯的變化。這類的問題，文獻學者已有相當多的研究成果，這些成果正好能夠作為本論文所提之模型的驗證對象。

2. 相近時期不同譯者的作品

分辨不同譯者作品是本研究最大的目的，因此本實驗就是用以驗證此模型於分辨不同譯者作品方面之能力。但因為譯經作品是以文字記錄，因此，譯經作品不僅受到作者影響，也受到譯經年代當時語言用字習慣的影響，同時期的作品可能會偏好使用相同的虛詞。為驗證本模型在年代相近的情況下，是否可區分出不同譯者，因此我們採相近時期中不同譯者的作品進行實驗。

3. 同經異譯的作品

同樣主題的作品具有相近的內容，在文字的使用上也頗為相近。而譯者在處理同經異譯的作品時，也經常會參考前人作品，有時甚至只進行小規模改寫，這也使得譯經風格分析更加的困難。因此本實驗中，我們將測試在翻譯標的物相同的情況下，本模型是否能夠表現出譯經風格的差異性。

根據上述條件選擇適合的實驗對象後，接著收集文本資料，投入模型進行實驗。以下三個小節，我們將詳細描述各實驗的結果。

（一）單一譯者的前期作品與後期作品

在判別單一譯者前後期風格的實驗中，我們選擇竺法護作為實驗對象。竺法護為西晉著名的譯經家，據現存文獻記載，其譯經始於太始2年（C.E. 266），終於永嘉2年（C.E. 308），歷時至少43年。在此期間，竺法護本身的語言能力有十分顯著的變化。根據《出三藏記集》中所述，竺法護世居敦煌，祖先為月支人，相當具有語言天分，[2]一生所會的語言更多達36種。[3]而在譯經的晚期，竺法護對漢語的理解更是已達精通的程度（梅迺文，1996）。如同《出三藏記集》中另一處所記載，竺法護翻譯時：「手執胡本，口宣晉言。」可見他對於漢語的理解已相當純熟，才能

2 《出三藏記集》卷13：「竺法護，其先月支人也，世居燉煌郡。年八歲出家，事外國沙門高座為師。誦經日萬言，過目則能。」（CBETA, T55, no. 2145, p. 97, c20-22）

3 《出三藏記集》卷13：「……護乃慨然發憤志弘大道。遂隨師至西域，遊歷諸國。外國異言三十有六，書亦如之，護皆遍學貫綜古訓音義字體無不備曉。遂大齎胡本，還歸中夏。自燉煌至長安，沿路傳譯，寫以晉文所獲。」（CBETA, T55, no. 2145, p. 97, c26-p. 98, a1）

以類似現代即席口譯的方式進行翻譯。綜合上述，由於其翻譯歷時長久，加上語言能力的進步，故其前、後期的風格有相當大的不同（湯用彤，1997）。

因此，我們分別選擇了竺法護於前、後期具有代表性的作品。在前期的部分，我們選擇了《須真天子經》，據經錄記載，該經譯於太始2年（C.E. 266），是目前可見竺法護最早所譯的一部作品。後期部分則選擇其於太康7年（C.E. 286）所譯的《正法華經》，此經譯出的時間距離《須真天子經》譯出有20年，並且竺法護於當時已學會晉言，因此可視為後期的代表作品。為了實驗模型在區分前、後期的效果，我們將更晚期的《普曜經》（C.E. 308-309）加入實驗中，雖然《普曜經》的年代比起《正法華經》更晚20年，但兩經同屬於竺法護精通漢語後的作品，很可能已經具有相同風格，因此加入實驗作為對照之用。

依據3.2的特徵選取方法，我們將《須真天子經》、《正法華經》及《普曜經》，以228個古籍常見虛字依公式進行計算，得到差值總和 $diff_j$ 最大的前30字，如表1所示。

從表1中我們可以初步觀察出，竺法護在早期《須真天子經》中，使用到表中所列虛字的比例佔了該經的32.47%，比起後期二者的18.57%與15.53%高出許多，幾近兩倍。其中「是」、「於」、「得」、「為」等字，在《須真天子經》使用機率相當高，但是這些字在後期的《正法華經》與《普曜經》中卻很少出現，是相當值得注意的地方。另一個特別的點在於《正法華經》中使用「斯」字的出現頻率有0.4158%，然而在《須真天子經》卻是幾近為零，在《普曜經》中的頻率也不到0.1%。我們可以發現，這三部經在簡單的公式計算下，似乎已能看出差別。

接著，為了進行主元素分析法的運算，我們先將三部經進行取樣，將每經依該經字數平均切割為20個樣本。然後將這三個樣本群分別配對實驗，以表1所列的30個虛字作為特徵值，投入主元素分析法運算並將結果繪圖。

1.《須真天子經》與《正法華經》

首先，我們先以竺法護早期的《須真天子經》（C.E. 266），與後期作品《正法華經》（C.E. 286）兩部經進行比較，結果如圖1所示。

圖1中《須真天子經》的落點偏向於圖的左半部，大部分的落點皆聚集在一起，只有少數幾個落點距離較遠。《正法華經》則集中於圖的右半部，落點的分布顯得相當集中，有著明顯的聚集情形。圖中比較特別的是《須真天子經》的第14號樣本，它落在座標空間中左上角（-4.7001, 7.4357）的位置，距離其他樣本相當遙遠。在該樣本中，大部分的虛字頻率都與其他樣本相近，但是在「是」、「行」二字的使用上，頻率卻高出其他樣本許多，其中「是」字的頻率佔全經「是」字總比率

表1 表列《須真天子經》、《正法華經》及《普曜經》中虛詞頻率差異最大30字，降冪排序

虛字	F_j^i			$\bar{F}_j$	$diff_j$
	T15n0588 須真天子經	T09n0263 正法華經	T03n0186 普曜經		
是	**2.2778%**	**0.8081%**	**0.7590%**	**1.2816%**	**1.9923%**
於	**2.1286%**	**0.8221%**	**0.6661%**	**1.2056%**	**1.8460%**
得	**1.9487%**	**0.6836%**	**0.5422%**	**1.0581%**	**1.7811%**
為	**2.0207%**	**0.8914%**	**0.6428%**	**1.1850%**	**1.6714%**
何	1.3985%	0.2091%	0.2664%	0.6247%	1.5477%
所	2.3292%	1.4035%	1.0100%	1.5809%	1.4966%
者	1.7379%	0.7153%	0.5329%	0.9953%	1.4851%
不	2.2623%	1.2109%	1.4143%	1.6292%	1.2663%
復	1.1414%	0.3065%	0.2556%	0.5679%	1.1472%
故	1.0180%	0.2314%	0.2277%	0.4924%	1.0513%
無	2.3960%	1.5268%	1.7132%	1.8787%	1.0346%
言	1.1466%	0.2349%	0.5360%	0.6391%	1.0149%
而	1.1312%	0.6084%	0.3516%	0.6971%	0.8682%
一	1.3008%	0.7482%	0.7002%	0.9164%	0.7689%
亦	0.7455%	0.2560%	0.1828%	0.3948%	0.7015%
行	1.1209%	0.5755%	0.6599%	0.7854%	0.6709%
云	0.5090%	0.0376%	0.0341%	0.1936%	0.6309%
已	0.6684%	0.1997%	0.1952%	0.3544%	0.6280%
事	0.5861%	0.1139%	0.1487%	0.2829%	0.6064%
來	0.1954%	0.7517%	0.4384%	0.4618%	0.5797%
諸	0.9615%	1.4352%	1.0983%	1.1650%	0.5405%
以	1.0232%	0.6248%	0.6243%	0.7574%	0.5315%
當	0.1337%	0.6331%	0.3780%	0.3816%	0.5030%
斯	**0.0154%**	**0.4158%**	**0.0991%**	**0.1768%**	**0.4780%**
如	0.9512%	1.2661%	0.8984%	1.0386%	0.4550%
便	0.4936%	0.2267%	0.1100%	0.2768%	0.4337%
若	0.2262%	0.5779%	0.2850%	0.3630%	0.4296%
今	0.0206%	0.3030%	0.3532%	0.2256%	0.4100%
則	0.4730%	0.2419%	0.0898%	0.2683%	0.4095%
正	0.1080%	0.5074%	0.3160%	0.3104%	0.4050%
合計	**32.4662%**	**18.5662%**	**15.5287%**		

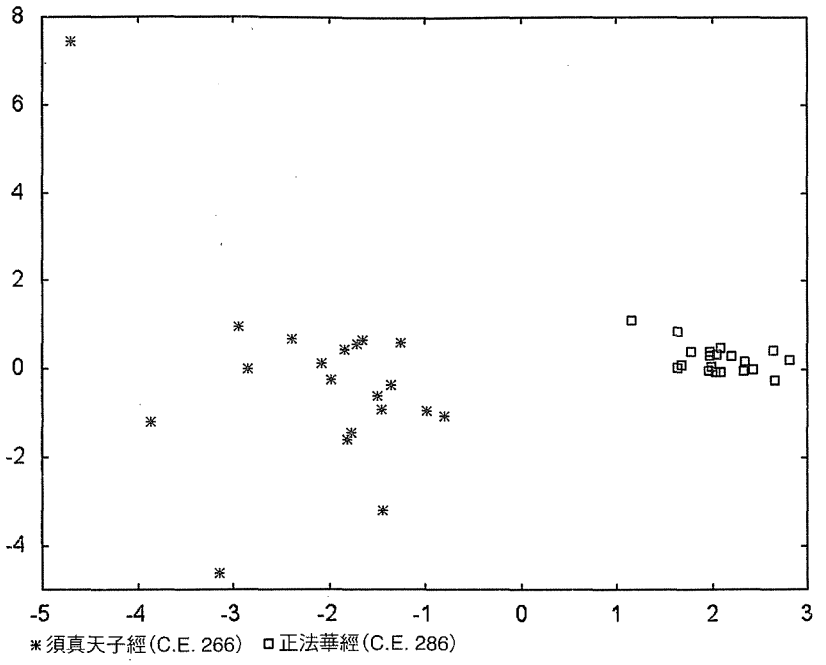


圖1 竺法護早期作品《須真天子經》與晚期作品《正法華經》

的12.8%，而「行」字更高達27.5%。我們推論這可能是此樣本離群的原因，當我們採用人工檢視第14號樣本文字內容時，發現該樣本的大量「行」字，大多數是帶有實際的意義在，而非作為虛詞使用。因此，我們推測該樣本離群現象疑似來自於內容，而與作者風格變化較無關係，然而，這部分的推論仍需未來進行更進一步的詳細驗證。

整體而言，從圖1中我們可見實驗的兩個樣本群並無交集，且兩群樣本的落點具有相當距離，可以簡單的用圖的左右部分將兩者良好區分開來。由於主元素分析法本身並沒有分群的概念，因此能得到此分群的結果，可見模型的效果相當顯著。

2.《須真天子經》、《正法華經》與《普曜經》

在此實驗結果的基礎上，我們將更晚期的《普曜經》加入實驗，觀察它的落點與原有兩經間的關係，如圖2。

圖2仍然保有《須真天子經》與《正法華經》的明顯區隔，然而《普曜經》與《正法華經》卻出現重疊的現象。這種現象顯示出兩種可能：一是如同我們之前的假設，《普曜經》與《正法華經》的確具有相同的語言風格；二是由於《須真天子經》與其他兩經差異過大所造成的錯誤。

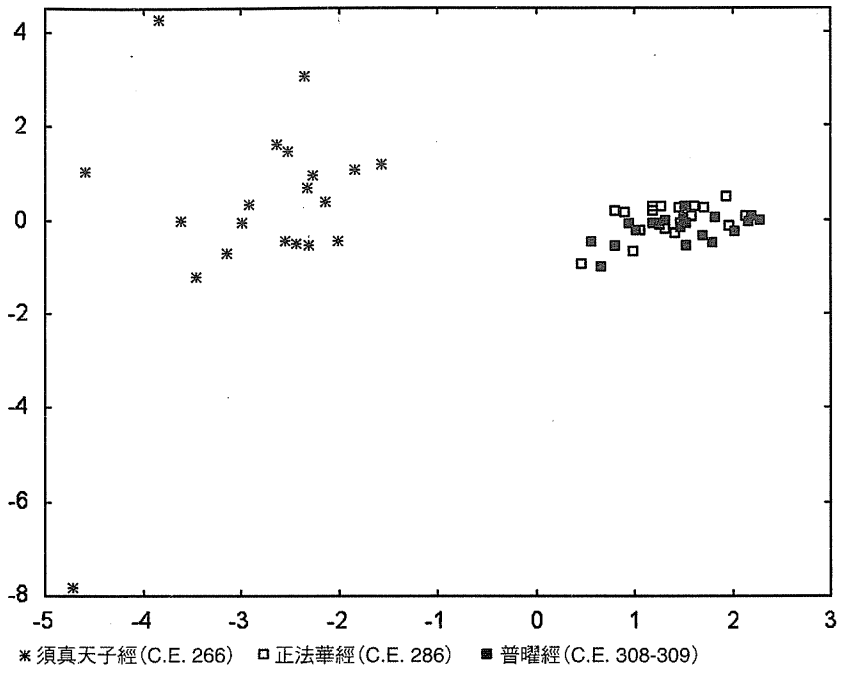


圖2 竺法護《須真天子經》與其他兩經間明顯分群，
而《正法華經》與《普曜經》具有重疊現象

為排除第二種可能，我們將《須真天子經》排除在外，單就《普曜經》與《正法華經》進行運算並繪圖如圖3。

從圖3中仍然可見《普曜經》與《正法華經》的樣本落點有著明顯聚集，但仍能以x軸區分出兩經，大部分的落點還是十分相近，因此可證明圖2的落點區隔並非來自單一樣本差異，而是模型真正能夠表現出風格間的接近程度。因此，我們認為本模型在區分竺法護前、後期譯經風格的效果算是十分良好。不過，上述實驗三個結果中，也提醒我們另一個明顯的事實——各經典雖然已經切分成20個樣本點，但同一經典的不同樣本點間的群聚效應仍然十分明顯。我們相信這是因為一個經典的翻譯為一次完整的作業，因此同一譯者在同一經典翻譯過程中，具有相同的語氣，是十分合理的現象。

（二）相近時期不同譯者的作品

為進行相近時期不同譯者間的作品比對，我們選擇竺法護與支謙作為比較對象。選擇竺法護的原因，一方面是因為我們在上述實驗中，已經進行過廣泛測試，

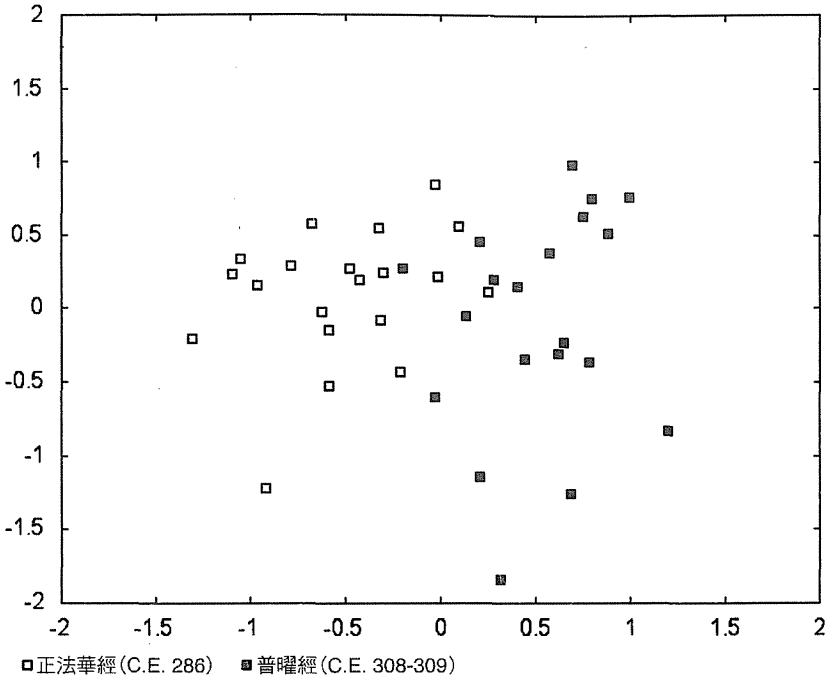


圖3 竺法護《正法華經》與《普曜經》落點有明顯的重疊現象

使用竺法護作為比較對象，可以省下許多實驗時間。另一方面，根據經錄的記載，我們可以肯定支謙為月支人，[4]而竺法護則是居住於敦煌的月支僑民，兩人為同一民族的後代，在語言文化上必定有其相近的地方。其次，據現有記載推測出支謙的生年約在西元192至221年間，卒年約在254至283年間，與竺法護相較之下，竺法護的生年雖不可考，但以其譯出《須真天子經》時值西元266年來說，兩人生活的年代相去不遠，可視為相近時期。此外，支謙與竺法護皆是大譯經師，名氣遠播，所譯出的作品數量眾多，為譯經史上的重要人物。綜合上述，兩位譯者在民族、時代與名氣皆十分相近，故符合實驗的需求。

在此我們對兩位作者各取兩部經作為實驗，在支謙所譯經中，我們先選擇了較著名的《太子瑞應本起經》與《佛說阿彌陀三耶三佛薩樓佛檀過度人道經》，再與竺法護本身風格已趨穩定的《正法華經》、《普曜經》進行實驗，如前步驟取得虛字清單30字，繪圖如圖4。

4 DDBC人名規範資料庫：http://dev.ddbc.edu.tw/authority/person/，規範碼A000164。

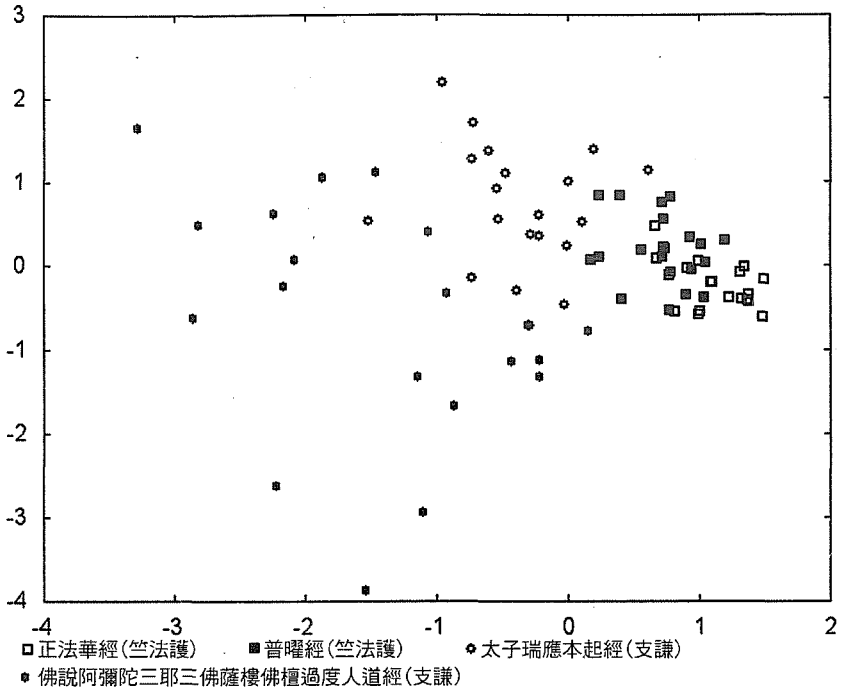


圖4 竺法護所譯《正法華經》、《普曜經》落點皆於畫面右半部，落點間仍有重疊現象，與支謙所譯經典在左半部有明顯區別

從畫面中顯示，以x軸可劃分出竺法護所譯《正法華經》、《普曜經》落點均在畫面右半部，而支謙所譯的兩部經，除了零星的3個落點較為接近竺法護外，其餘皆落在左半部。另外，屬於支謙的兩群落點之間也有部分重疊，顯示風格有所接近。而竺法護的樣本落點則是相當集中，呈現重疊群聚的現象。為了進一步觀察支謙譯經是否具有相近的風格，我們增加支謙的譯經，將《佛說義足經》加入實驗中，得到新特徵值，繪圖如圖5。

從圖5可見，新加入的《佛說義足經》與原有的《太子瑞應本起經》、《佛說阿彌陀三耶三佛薩樓佛檀過度人道經》均有部分重疊之處，三部經大部分的落點仍落在畫面的左半部。其中《太子瑞應本起經》與《佛說阿彌陀三耶三佛薩樓佛檀過度人道經》的重疊程度也較上個實驗提高，顯示此兩經在新特徵值使用上，頻率較上個實驗的特徵值來得接近。而竺法護的兩群落點仍在右半部，具有明顯的聚集，不受新特徵值的影響。此外，我們也發現，在《佛說義足經》加入分析之後，原本沒有交界的《佛說阿彌陀三耶三佛薩樓佛檀過度人道經》與竺法護的兩個經典樣本點群，開始出現一些交疊的現象。我們相信這是因為加入分析的樣本點已經太多（100

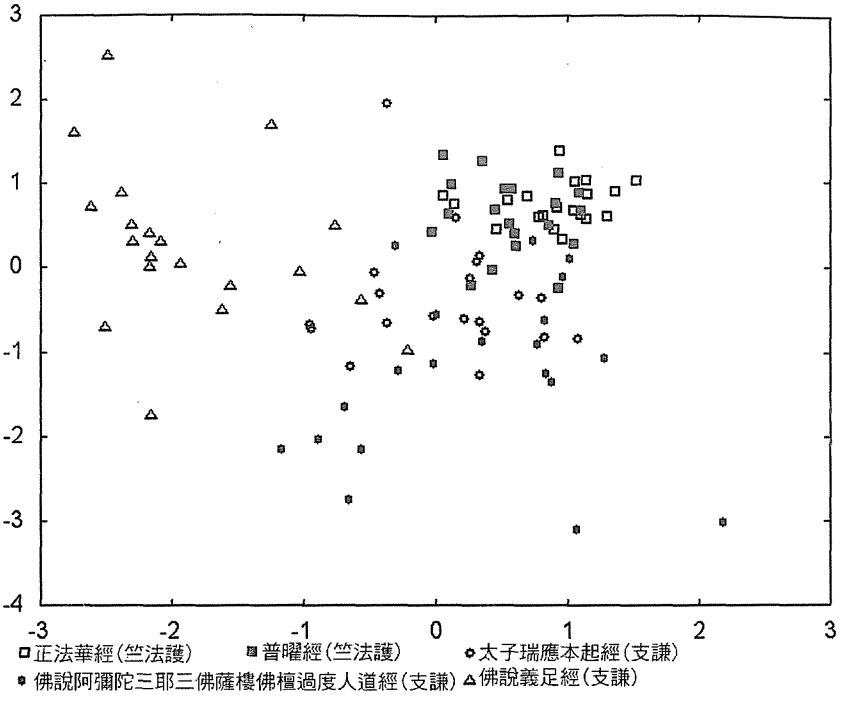


圖5 新加入支謙的《佛說義足經》後，原有《太子瑞應本起經》與《佛說阿彌陀三耶三佛薩樓佛檀過度人道經》落點顯得更為接近

個樣本點），僅使用產生於第一與第二主元素的投影結果，已經無法顯示出所有群間的差異。雖然如此，其中的界線仍勉強可以分辨。因此，我們仍相信採用本研究模型能區分不同譯者之作品。

（三）同經異譯的作品

以經文的內容而言，同經異譯的作品其本質會較為相近，再加上譯經師於進行譯經工作時，經常有參考前人作品改寫的情形，因此，同經異譯的作品，在風格上於同經異譯經典的風格分辨能力，可能出現彼此接近的現象。為測驗本模型於同經異譯經典上的表現，我們選用具有多個異譯版本的《無量壽經》作為實驗。現今可取得的《無量壽經》漢譯版本有五種，分別為：後漢支婁迦讖譯《佛說無量清淨平等覺經》、吳支謙譯《佛說阿彌陀三耶三佛薩樓佛檀過度人道經》、曹魏康僧鎧譯《無量壽經》、唐菩提流志譯《大寶積經無量壽如來會》以及趙宋法賢譯《大乘無量壽莊嚴經》（周睦修，2005）。由於前三個版本是屬於譯經史上早期的作品，後兩個版本則是較晚期的，因此我們先粗略劃分，將兩個時期各挑一部經進行比較。早期

部分，我們選擇了支謙譯《佛說阿彌陀三耶三佛薩樓佛檀過度人道經》，晚期則選擇了法賢的《大乘無量壽莊嚴經》。兩經繪圖如圖6。

從圖6中，我們可以發現兩經的落點有所區隔，但仍可說相近，且範圍有部分重疊，這顯示出兩經的虛詞使用頻率相當接近，符合我們先前對於同經異譯作品風格可能相近的假設。雖然《大乘無量壽莊嚴經》有兩個樣本呈現離群，但對結果的影響不大，樣本的落點並未遭到排擠。其中離群的樣本，分別是第1號與第2號樣本。第1號樣本在「有」、「復」、「已」三字有較高的使用頻率，這是由於此經一開始大量的宣說佛名，而每宣說一尊佛名即以「次復有佛」作為開端，再以「已過去」作為結尾，造成此樣本的特殊性。第2號樣本則因為「得是願乃作佛。不得是願終不作佛。」的重覆出現，顯得在「是」、「作」二字有較高頻率。此外，該樣本在「有」字的使用頻率也遠低於其他樣本，因此呈現較大差異。

另外在《無量壽經》的五個版本中，曾有日本文獻學者認為支謙的《佛說阿彌陀三耶三佛薩樓佛檀過度人道經》一經，可能並非為支謙初譯，而是他參考了支婁迦讖的版本所改譯；也有學者認為該經其實應為支婁迦讖的作品（釋慧嚴，2001）。雖然此一爭議在學界至今尚未有定論，但從學者們的討論中，不難看出此兩經具有

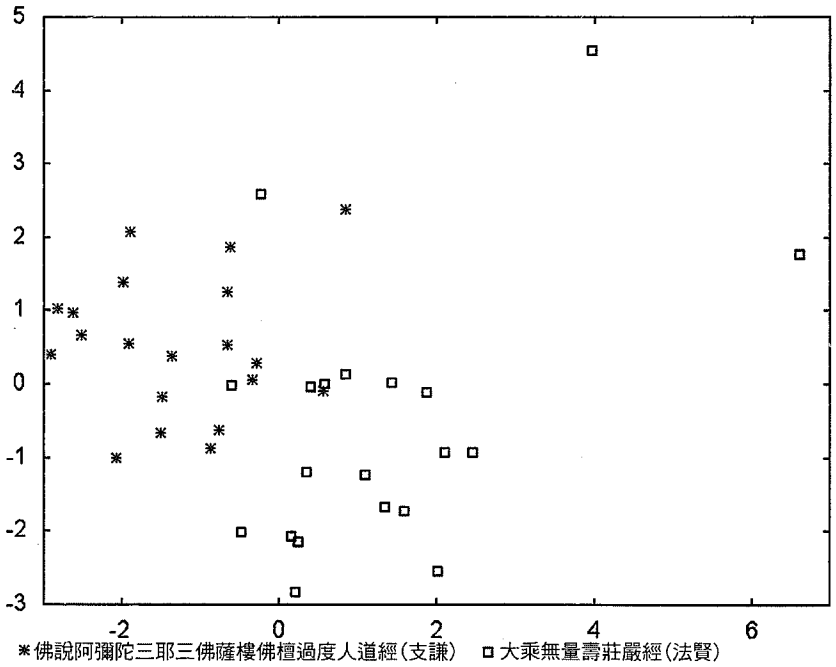


圖6　支謙與法賢的同經異譯版本，在畫面上的落點相當接近，並有部分重疊

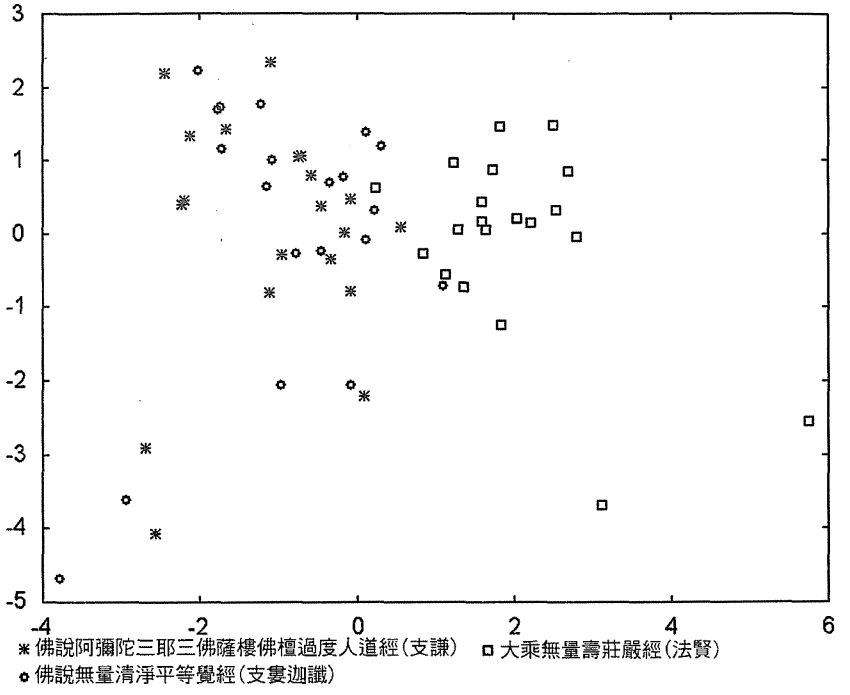


圖7 支謙與支婁迦讖的同經異譯版本，落點重疊程度高，與法賢版本區別明顯

相當密切的關係。加上歷史上記載著支謙與支婁迦讖之間有著師承的關係，[5]因此，我們可以相信這兩人的《無量壽經》版本風格，應該比起其他人的版本更為相近。所以，接著我們將支婁迦讖所譯《佛說無量清淨平等覺經》加入實驗，取得新的特徵值計算，繪圖如圖7。

由畫面可以看到支婁迦讖的《佛說無量清淨平等覺經》與支謙的《佛說阿彌陀三耶三佛薩樓佛檀過度人道經》有著明顯的重疊，且無法切分。而與法賢《大乘無量壽莊嚴經》的區分則比上個實驗來得更加明顯，應證了學者對於支謙、支婁迦讖版本相近的推論。因此，本模型在驗證相近的作品風格上，具有相當明顯的效果。

五、結論

本研究利用古籍常見虛字與主元素分析法，建立出譯者風格判別模型，以協助文獻研究者進行譯者判斷。在實驗中，我們將樣本切割為數量相等的數個小樣本，

5 《佛祖統紀》卷35：「月氏國優婆塞支謙來雒陽。謙受業於支亮，亮受業於支讖，世稱天下博知無出三支。」(CBETA, T49, no. 2035, p. 331, b26-28)

並使用自訂的公式選取較為顯著的風格特徵量值進行分析。實驗證明了模型對於不同作譯者風格有著相當良好的區分效果，在相近的作譯者風格上，也確實能表現出其相近的特質。

雖然以古籍常見虛字結合主元素分析法在譯者風格實驗初步有效，但是本研究還有許多可以細部調校的部分，舉例如下：一、改良取樣方式：基於PCA的特性，我們得知樣本的切分是實驗的基礎，因此，如何選擇適當的樣本數，以及如何使樣本的個別差異減到最低、同時保留最多作譯者特徵，是未來研究的重要課題之一。二、特徵值的篩選：特徵值的選擇與數量，對於最後的分析結果有著重要影響，本研究以30字作為初步實驗，已得到不錯的實驗結果。未來可研擬更詳細之訂定方式，例如：採用公式決定特徵值數量，以及結合現有的佛學詞彙研究成果，過濾出更為符合虛字定義的特徵值。三、模型功能的擴展：我們希望能繼續發展此一模型，利用數學公式以建立非人工的自動判別功能，並可延伸為「判斷譯者未知的文件之實際譯者」模型。以上所述皆為本模型可以繼續改進討論的地方。此外，公開化流通也是重要的考量方向之一，我們希望能將研究成果製成網路工具公開分享，提供給相關領域的學者們作為研究方向的參考。我們希望模型的判斷結果，不但可作為研究者的佐證，支持其論點；更能進一步挖掘出隱藏在資料中的新課題，等待研究者的發掘。

參考文獻

王叔岷，2007，《古籍虛字廣義》（再版），北京：中華書局，（初版：1900年）。

呂澂，1991，《新編漢文大藏經目錄》（再版），中國濟南：齊魯出版社，（初版：1980年）。

汪禕，2008，《中古佛典量詞研究》，南京師範大學博士論文。

林昭君，1998，《東漢佛典之介詞研究》，國立中正大學中國文學研究所碩士論文。

周睦修，2005，《〈無量壽經〉譯者考——以佛經語言學為研究主軸》，南華大學宗教學研究所碩士論文。

俞理明，1993，《佛經文獻語言》，四川：巴蜀書社。

高婉瑜，2006，《漢文佛典後綴的語法化現象》，國立中正大學中國文學研究所博士論文。

梅迺文，1996，〈竺法護的翻譯初探〉，《中華佛學學報》，9，頁49-64。

許理和（Erik Zürcher），1987，《最早的佛經譯文中的東漢口語成分》，蔣紹愚譯，《語言學論叢》第14輯，北京：商務印書館。

張嘉慧，2008，《法雲〈翻譯名義集〉的語言研究——以音寫語段的分析為中心》，國立中央大學中國文學研究所碩士論文。

湯用彤，1997，《漢魏兩晉南北朝佛教史》（再版），北京：北京大學。

廖宜君，2008，《東漢佛典之雙音節動詞研究》，國立政治大學中國文學研究所碩士論文。

劉芳薇，1995，《〈維摩詰所說經〉語言風格研究》，國立中正大學中國文學研究所碩士論文。

蔡佳玲，2007，《漢地佛經翻譯論述的建構及其轉型》，國立中央大學中國文學研究所碩士論文。

戴綉娟，2010，《〈維摩詰所說經〉虛詞研究》，南華大學宗教學研究所碩士論文。

釋慧嚴，2001，〈彌陀淨土信仰對漢儒內心世界的影響〉，《華梵大學第五次儒佛會通學術研討會論文集》，頁121-135。

Bozkurt, I. N., Ba ho lu, O. & Uyar, E. (2007). *Authorship Attribution: Performance of Various Features and Classification Methods.* Proceedings of the 22nd International Symposium on Computer and Information Sciences, 2007. doi: 10.1109/

ISCIS.2007.4456854

Burrows, J. (1992). Not Unless You Ask Nicely: The Interpretative Nexus between Analysis and Information. *Literary and Linguistic Computing*, *7*(2), 91-109.

Grieve, J. (2007). Quantitative Authorship Attribution, an Evaluation of Techniques. *Literary and Linguistic Computing*, *22*(3), 251-270.

Holmes, D. & Crofts, D. W. (2010). The Diary of a Public Man: A Case Study in Traditional and Non-traditional Authorship Attribution. *Literary and Linguistic Computing*, *25*(2), 179-197.

Hoover, D. L. & Hess, S. (2009). An Exercise: In Non-ideal Authorship Attribution: The Mysterious Maria Ward. *Literary and Linguistic Computing*, *24*(4), 467-489.

Houvardars, J. & Stamatatos, E. (2006). *N-Gram Feature Selection for Authorship Identification*. Proceedings of 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications 2006, LNAI4183, 77-86.

Hung , J., Bingenheimer, M. & Wiles, S. (2009). Quantitative Evidence for a Hypothesis Regarding the Attribution of Early Buddhist Translations. *Literary and Linguistic Computing*, *25*(1), 119-134.

Jockers, M., Witten, D. & Criddle, C. (2008). Reassessing Authorship of Book of Mormon Using Delta and Nearest Shrunken Centroid Classification. *Literary and Linguistic Computing*, *23*(4), 465-491.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). USA: Springer.

Kjell, B. (1994). Authorship Determination Using Letter Pair Frequency Features with Neural Network Classifiers. *Literary and Linguistic Computing*, *9*(2), 119 -124. doi:10.1093/llc/9.2.119

Labbe, D. (2007). Experiment on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*, *14*(1), 33-80.

Lancaster L. (2008).*Catalogues in the Electronic Era: CBETA and The Korean Buddhist Canon: A Descriptive Catalogue*. CBETA, Taipei, 2008 (electronic publication). Retrieved from http://jinglu.cbeta.org/lancaster.htm

Liu, W., Allison, B., Guthrie, D. & Guthrie, L. (2007). *Chinese Text Classification without Automatic Word Segmentation*. Proceedings of ALPIT 2007, 45-50.

Mosteller, F. & Wallace, D. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer Verlag.

Peng, F., Schuurmans, D., Keselj, V. & Wang, S. (2003). *Language Independent Authorship Attribution Using Character Level Language Models*. Proceedings of the 10th Conference on EACL, 267-274.

Schaalje, B., Fields, P. Roper, M. & Snow, G. (2011). Extend Nearest Shrunken Centroid Classification: A New Method of Open-set Authorship Attribution of Text of Varying Sizes. *Literary and Linguistic Computing*, *26*(1), 465-491.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. (1999). *Automatic Authorship Attribution*. Proceedings of the 9th Conference on EACL, 158-164.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. (2001). Computer-based Authorship Attribution without Lexical Measures. *Computers and Humanities*, *35*(2), 193-214.

Tambouratzis, G. & Vassiliou, M. (2007). Employing Thematic Variables for Enhancing Classification Accuracy within Author Discrimination Experiments. *Literary and Linguistic Computing*, *22*(2), 207-255.

Tearle, M., Taylor, K. & Demuth, H. (2007). An Algorithm for Automated Authorship Attribution Using Neural Networks. *Literary and Linguistic Computing*, *23*(4), 425-442.

Wold, S., Esbensen, K. & Geladi, P. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1-3), 37-52.

Zhang, D. & Lee, W. S. (2006). *Extracting Key-substring Group Features for Text Classification*. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 474-483.

Zhao, Y. & Zobel, J. (2007). *Searching with Style: Authorship Attribution in Classic Literature*. Proceedings of the 13th Australasian Conference on Computer Science, 2007, 59-68.

「共現」詞頻分析及其運用
——以「華人」觀念起源為例*

金觀濤** 、邱偉雲*** 、劉昭麟****

摘 要

本文研究目的旨在透過《清季外交史料》(1875-1911)當中對「華工」議題的歷時性描述，形構出當時官方對於「華工事件」的認知轉移過程。再從「華工事件」中諸多「關鍵詞」內涵的歷時性變化中，進一步觀察「華人」觀念是如何隨著「華工」事件認知焦點的轉移而逐漸形成之脈絡。本文之研究範圍以四百多萬字之《清季外交史料》為底本，挑選曾出現過「華工」此一關鍵詞之文獻共有109篇，總字數為118,899字，旁及目前近現代中國「華工議題」與「華人意識」之相關研究。本文主要以「共現詞頻分析法」為數位輔助程式，透過此程式，可協助研究者快速在所欲研究的龐大文本中，切割出眾多的待選詞，再由研究者確定重要關鍵詞後，進一步將關鍵詞兩兩配對成共現詞組，觀察各共現詞組的共現頻度及歷時性分布。而研究者即可從高共現頻度之共現詞組中，架構出以「事件」為核心之「重要關鍵詞叢」，進而觀察「事件」與「觀念」之間的互動過程。期待透過本文的小試，能將數位人文學推廣給更多人文學界研究者使用，藉由數位方法能讓人文研究者得以處理龐大的文獻底本，進行更為宏觀的觀察與研究，此為本文結合數位方法與人文研究之推廣前景所在。

---

* 本文所使用之共現詞頻分年分析法，請參見論文：劉昭麟、金觀濤、劉青峰、邱偉雲、姚育松，〈自然語言處理技術於中文史學文獻分析之初步應用〉，「2011第三屆數位典藏與數位人文國際研討會」發表之論文，國立臺灣大學。

** 國立政治大學講座教授。

*** 國立政治大學中國文學系博士研究生。

**** 國立政治大學資訊科學系教授。

Frequency Analysis and Application of "Co-occurrence" Phrases: The Origin of the Concept "Hua-ren" as an Example

Guantao Jin *, Wei-yun Chiu **, Chao-Lin Liu ***

Abstract

This article aims to demonstrate the official conceptual change of the "Hua-gong Event" (華工事件 the Chinese worker event) with the chronological records in *Qingji waijiao shiliao*(清季外交史料)(1875-1911) and to illustrate the context of the concept "Hua-ren" (華人 the Chinese) developed with the attention shift during the events and the chronological change of the keywords content. The primary source is the all-text data of *Qingji waijiao shiliao*, which contains over four million characters of official records from 1875 to 1911. There are 109 articles (the word counts overall is 118,899) include the phrase "Hua-gong" (華工), regarding to the relevant research on the Hua-gong Event and Hua-ren consciousness. This research utilize the "analysis of co-occurrence frequency" (共現詞頻分析法) developed by Dr. Liu as its chief digital approach. With the program, various candidate phrases could be selected among an enormous quantity of text. After the researcher chose a collection of the keywords from the candidates, the program could compare every one with each other and chronically show their relations with frequency of co-occurrence(共現頻度). Those of high frequency of coexistence with each other establish a keywords cluster which demonstrates the attention of the events and thus one can observe the interaction of concepts and events. Hopefully the experiment of the research could encourage more humanists to apply the approach of digital humanities. With digital approach, it is more likely to conduct humanities researches on immense amount of data and accordingly depict a panorama of the past, which is the outlook of this research.

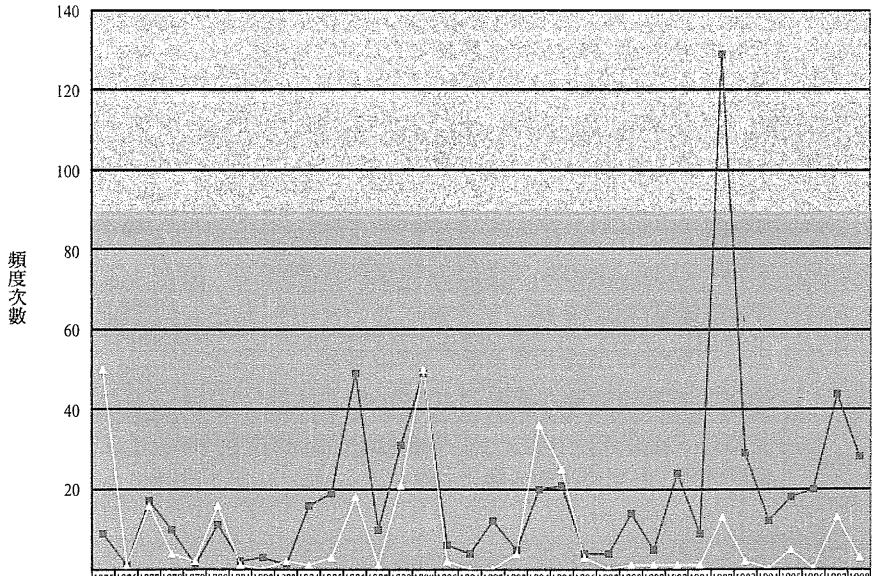
* Chair Professor, National Chengchi University.
** Ph.D. Student, Department of Chinese Literature, National Chengchi University.
*** Professor, Department of Computer Science, National Chengchi University.

一、「華人」觀念的起源和共現詞頻分析法

「華人」觀念是如何起源的？幾年前吳通福博士在標點《清季外交史料》時告訴作者，該文獻中長時間地頻頻提到華工問題，我們意識到這對解決該問題的重要性。於是對《清季外交史料》全部文獻中「華人」和「華工」這兩個關鍵詞進行檢索，發現包含「華人」關鍵詞的文獻早期幾乎全部被蘊含在包含「華工」關鍵詞的文獻之中，如圖1所示。

從圖1可以明顯看出，從1882年開始，《清季外交史料》中「華人」關鍵詞頻度才開始凌駕「華工」之上。再從1882年前兩個關鍵詞的文章分布圖來看，「華人」一詞在早期1882年之前是隱含於「華工」文本之中出現的，如表1所示。

我們立即用一億兩千萬字的「中國近現代思想史專業數據庫」對這一想法進行檢驗，在1882年前，「華人」關鍵詞出現的23篇文章51次使用中，共有12篇文章34次「華人」之用例出現於「華工」議題討論中，比例高達66.666%，故包含「華人」關鍵詞的文獻早期幾乎大多數被蘊含在以討論「華工」議題為主的文獻中，證明該觀點成立。這樣，可以用《清季外交史料》中提及「華工」議題的文獻，作為研究「華人」觀念起源的基礎文本。其方法是分析「華人」這個詞是如何與「華工」一詞「共現」的，因為「共現」過程的意義結構中一定包含了「華人」觀念的起源。



	1875	1876	1877	1878	1879	1880	1881	1882	1883	1884	1885	1886	1887	1888	1889	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908
—●— 華人頻度	9	1	17	10	1	11	2	3	1	16	19	49	10	31	49	6	4	12	5	20	21	4	4	14	5	24	9	129	29	12	18	20	44	28
—△— 華工頻度	50	1	16	4	2	16	1	0	2	1	3	18	1	21	50	2	0	0	4	36	25	3	0	1	1	1	1	13	2	0	5	0	13	3

資料來源：「中國近代思想史專業數據庫（1830-1930）」

圖1 《清季外交史料》1875-1908年「華人」與「華工」歷年分布頻度表

表1 「華人」一詞於1875-1882年間與華工文獻重出對照表

華工文本	年代	《清季外交史料》中「華人」一詞於1875-1882年間出現總目	作者
	1875	滇督岑毓英奏英員馬嘉理被戕一案派員查辦摺〔附上諭〕	岑毓英
	1875	總署奏英員馬嘉理被戕一案英使詞意叵測請加意邊防海防摺〔附上諭〕	總署
1	1875	直督李鴻章等奏請設駐祕魯使臣保護華工片	李鴻章
2	1875	直督李鴻章等奏請保護祕魯華工謹防誘拐片〔附上諭〕	李鴻章
3	1875	直督李鴻章奏請派丁日昌互換祕魯條約片〔附上諭〕	李鴻章
	1876	附總署　覆日本國使臣森有禮節略	總署
4	1877	照錄古巴華工條款	清季外交史料
5	1877	總署奏與西班牙公使訂定古巴華工條款摺〔附條款〕	總署
	1877	使英郭嵩燾等奏保薦伍廷芳摺	郭嵩燾
	1877	江督沈葆楨奏美國旗昌公司願並歸招商局摺〔附上諭〕	沈葆楨
6	1878	總署致西使赴古巴華工如有臨時不往者借墊款項由關道取保俾得有著照會	總署
7	1878	總署奏與西使通融互換古巴條款片〔附條款正文〕	總署
8	1878	西使覆總署華工赴古巴如有船主墊款由關道取保於條款易於得手照會	西使
	1878	總署奏新加坡設總領事經費薪俸辦法摺	總署
	1878	使美日祕陳蘭彬等奏應派駐美中國領事以資保護僑民片	陳蘭彬
9	1879	總署奏檀香山擬設商董由駐美公使發給諭帖摺	總署
10	1880	總署奏美國修約提出限制華工條款摺	總署
11	1880	中美條約〔其二〕	清季外交史料
	1880	總署奏定內港江河行船免碰及救護賠償審斷專章	總署
	1880	中德續修條約善後章程	清季外交史料
	1880	總署奏美國修約使臣來華請派大員與之商議片	總署
12	1881	直督李鴻章奏巴西修約情形摺〔附條約及節略〕	李鴻章
	1881	直督李鴻章奏朝鮮陪臣金允植密陳該國王議商外交情形相機開導摺〔附朝鮮密書跿錄及上諭〕	李鴻章
	1882	總署奏新疆開埠中俄一律免稅摺	總署
	1882	北洋大臣李鴻章奏妥議朝鮮通商章程摺〔附章程〕	李鴻章
	1882	總署奏議覆朝鮮通商章程摺	總署

資料來源：「中國近代思想史專業數據庫（1830-1930）」

所謂共現（co-occurrence）是指：「特徵項描述的資訊共同出現的現象。通過對共現現象的定量分析，可以揭示資訊的內容關聯和特徵項所隱含的知識。」（楊立英，2006），它用於本文研究就是分析共現詞彙集。該方法曾被用於研究教育評鑑發展趨勢（曾元顯、林瑜一，2011）、探勘資訊傳播學領域的研究主題與關係（林頌堅，2010）、中文醫學概念空間（李軍蓮、李丹亞、黃利輝、孫海霞、冀玉靜、王鈐，2010）、中國大陸信息作戰領域發展（陳良駒、傅振華、楊誌瑋，2010）等。方法學研究則包括對於詞共現的文本相似度計算（曹恬、周麗、張國煊，2007）、提出新的共現辭彙演算法——FDC（陳鐘、彭波，2005）、基於辭彙吸引與排斥模型的共現詞提取（郭鋒、李紹滋、周昌樂、林穎、李勝睿，2004）、共現分析的文本知識挖掘方法研究（王曰芬、宋爽、熊銘輝，2007）等等。

目前該方法已被運用到人文歷史研究中。如曾元顯提到日本學者村田忠禧（2002）：

> 以1949年到2002年人民日報每年的元旦社論為材料，先自訂待觀測的關鍵詞彙，然後統計這些詞彙出現的篇數與年代。依此資料分析每年焦點詞彙的變化，並觀看某年焦點詞彙在其他年份出現的狀況，以瞭解該年份與其他年份的連動關係。村田忠禧認為這些詞彙的變化可以宏觀地觀察歷史變動，且此種分析結果不是在推翻過去對歷史的解釋，而是做進一步地補充（亦即讓資料自己說話），他認為這正是其詞彙頻率統計分析有效性之所在。（曾元顯、林瑜一，2011）

但這種方法碰到最大困難是不知選擇哪些關鍵詞進行共現分析才是適當的。正如這方面研究者所說：「其最重要的起始步驟，需自訂『待觀測的關鍵詞彙』，是此方法最耗費專業知識、人力與時間的地方。」（曾元顯、林瑜一，2011）

換言之，要用「關鍵詞的共現」來探討某一觀念的起源，原有基於數據庫中一個和數個關鍵詞檢索的統計方法就不夠用了。除了眾多關鍵詞之分析研究非人力所能及之外，最大的困難在於我們完全不知道分析過程中還需要涉及哪些關鍵詞。例如本文欲研究「華人」觀念如何從「華工」議題中轉化出來，必須找到兩者發生關係的事件和論述，它們涉及哪些關鍵詞呢？要在四百多萬字的《清季外交史料》中將其找出絕不容易，為此必須尋找新的數位方法。

我們發現，可以運用對齊夫定律的最大偏離即是該文本最重要的關鍵詞叢的預設，從有關文本中找出最重要的關鍵詞叢（金觀濤、姚育松、劉昭麟，2011）。然後再去建立研究兩個不同關鍵詞「共現」現象的分析方法。

目前有關於詞「共現」判定演算法有如下三種：「1.以整篇文檔作為視窗單

元，計算在每個給定文檔中每對辭彙同時出現的次數；2.計算辭彙對在同一個句子、同一段文章或章節中出現的次數；3.計算辭彙對在文檔中與在文檔集中共現的相對頻率。」（李軍蓮、李丹亞、黃利輝、孫海霞、冀玉靜、王鈐，2010）而本文則是採取不超過三十個漢字之距離作為視窗單元，來討論兩個關鍵詞是否共現。我們認為，這個距離雖然可任意選擇，但若能根據研究者就文本之書寫風格屬性閱讀下，以確定最適合之視窗單元，將更能添加人文研究者之專業判斷，亦將比任意選擇更能貼近歷史文本語境。[1]

二、齊夫定律與關鍵詞叢

1949年哈佛大學的語言學家George Kingsley Zipf提出「齊夫定律」。它通常表達成公式（1）：

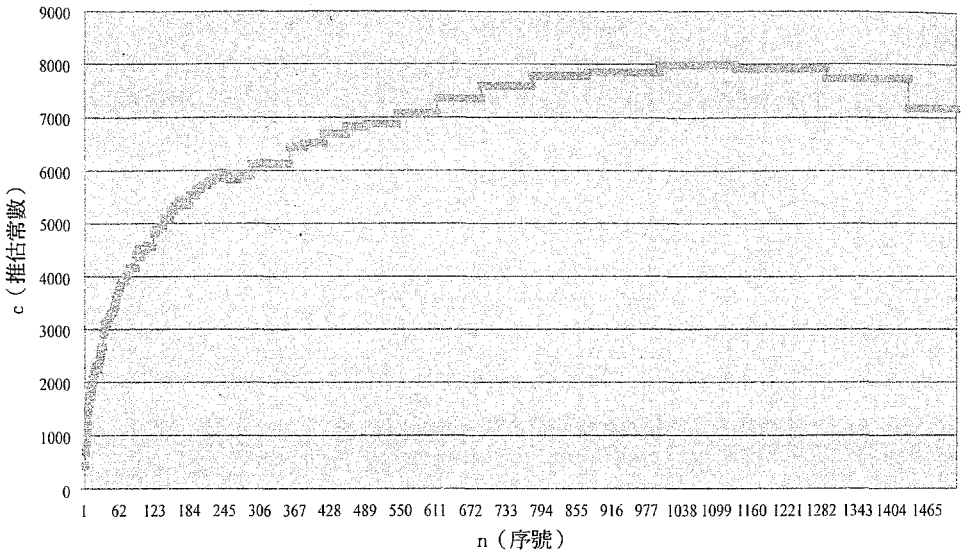
公式（1）　$f \times n = c$

其中f為某詞總次數，n為該詞在次數遞減排列中的序號，c為常數。[2]

作者金觀濤等曾提出根據齊夫定律選擇特定文本重要關鍵詞的想法，即齊夫定律作為語言學法則和用語言表達特定內容無關。這樣對齊夫定律的偏離可以分為「大偏離」和「一般（線性）偏離」，「大偏離」即為該文本要表達的最主要的意

1 作者邱偉雲（2011：173）曾指出以研究者關注之核心關鍵詞前後10字，作為視窗單元最接近一個話語句子長度，故將前後10字作為核心關鍵詞的指涉範圍較為適切，也符合《清季外交史料》奏摺文本的「段落語意長度」。我們根據不同研究方法觀察文本較為適切的視窗單元字數，若以某一核心關鍵詞出發，則其前後10字的20字文本，最接近該核心關鍵詞，也最能作為該核心關鍵詞的「定義語境」。透過核心關鍵詞前後10字文本，共20字的視窗單元，較適合提取「關鍵詞叢」與進行「語境式定義」。而本文則不同前者從「某一核心關鍵詞」出發，而是以「兩詞共現」角度出發，故兩個詞彙皆同樣為研究者觀察重點，視窗單元便需增加以擴大語境，方能呈現出較為整全之「共現語境」。然這樣的視窗單元仍非絕對值而應具有開放性，如作者劉昭麟等（2011：155）曾指出以30個漢字作為共現視窗，是一個任意選擇，研究者可自行訂定數量大小，若採用較少漢字，則兩詞彙便不容易被認定為共同出現，因此將是比較保守的選擇。而綜合人文學者與數位學者對於「視窗單元」之看法，基本上對於「視窗單元」的多寡應可於進行視窗單元文本量設定前，關注不同文本屬性而加以變化，例如奏摺類文本屬性語意段落較短，而一般文集則語意段落較長，故設定視窗單元可搭配研究文本的屬性來觀察與判斷較為適切，因此並未具有一個絕對標準，應依人文研究者對文獻閱讀之語感作視窗單元之判斷，呈現出一種開放性。

2 《維基百科．齊夫定律》：「在自然語言的語料庫裡，一個單詞出現的頻率與它在頻率表裡的排名成反比。所以，頻率最高的單詞出現的頻率大約是出現頻率第二位的單詞的2倍，而出現頻率第二位的單詞則是出現頻率第四位的單詞的2倍。這個定律被作為任何與power law probability distributions有關的事物的參考。比如，在 Brown 語料庫中，『the』是最常見的單詞，它在這個語料庫中出現了大約7%（100萬單詞中出現69,971次）。正如齊夫定律中所描述的一樣，出現次數為第二位的單詞『of』占了整個語料庫中的3.5%（36,411次），之後的是『and』（28,852次）。僅僅135個字彙就占了Brown 語料庫的一半。」（《維基百科．齊夫定律》，上網日期：100年9月30日。網址：http://zh.wikipedia.org/wiki/齊夫定律）。另可參考蔡明月，1999，〈齊夫（Zipf）定律〉，《教育資料與圖書館學》，37（2），頁165-183。



資料來源：「中國近代思想史專業數據庫（1830-1930）」

圖2 《清季外交史料》華工資料準詞彙對數值

義。它規定代表該文本意義結構的關鍵詞叢（金觀濤、姚育松、劉昭麟，2011）。

目前在一億兩千萬字的「中國近現代思想史專業數據庫」中，《清季外交史料》共導入了1875-1909年之檔案資料，包含5,758份檔案；不包含檔案標題、作者等基本資訊，僅文件內容合計2,900,938個漢字。然本文由於需對整部《清季外交史料》進行歷時性觀察，故另採用了吳通福博士的繁體標點版來作為「華工」議題史料的基礎文本，較「中國近現代思想史專業數據庫」中《清季外交史料》多了1909-1911年的檔案，包含檔案標題總字數為4,025,966字。本文即從1875-1911年《清季外交史料》中，擷取出曾出現過「華工」此一關鍵詞之文獻共有109篇，總字數為118,899字。我們使用數位技術對《清季外交史料》中「華工」109篇史料進行斷詞，以出現5次以上之組合稱為「準詞彙」。其數目共有1,515個。照公式（1）可繪出準詞彙在該文本中的齊夫曲線（圖2）。顯而易見其中水平段為常數，即c等於8000。從8000至6874為近似斜線，即「一般偏離」，共有587個準詞彙；而小於6874者為「大偏離」，有543個準詞彙。這樣它規定了與《清季外交史料》華工議題關係最為密切的關鍵詞叢，經過分析篩選共有168個具有研究意義的詞彙，可列為表2。

表2給出文本的關鍵詞叢，它對於人文研究者深具價值。這張關鍵詞頻表，可使研究者了解在「華工」事件發展史中，有哪些關鍵詞與「華工」事件並生。我們

表2《清季外交史料》華工議題具研究意義詞叢表

具研究意義詞彙	詞頻	具研究意義詞彙	詞頻	具研究意義詞彙	詞頻	具研究意義詞彙	詞頻
中國	432	利益	49	租界	28	中國政府	18
華工	348	總領事	49	開辦	28	美境	17
大臣	278	執照	46	禁華工	28	干預	17
領事	263	南洋	46	土人	27	閩省	17
華人	256	總統	44	新例	27	粵督	17
美國	251	總督	44	教民	27	遵旨	17
華民	198	英國	41	公使	27	澳門	16
公司	195	工人	41	修約	26	入境	16
衙門	192	商人	40	美外部	26	諭旨	16
保護	151	出口	40	總公司	26	紐約	16
外部	146	祕魯	39	中國全權大臣	26	參贊	16
條約	127	限制	39	請旨	25	電報	16
使臣	126	商務	38	教堂	25	利息	16
鐵路	118	機器	38	優待	25	倫敦	16
日本	115	商民	38	稅務司	24	交涉官	16
照會	104	國家	37	外務部	24	邦交	15
古巴	98	洋人	37	限禁	23	免稅	15
合同	98	貿易	36	巡撫	23	納稅	15
章程	92	華僑	36	小呂宋	23	專約	15
國人	90	中美	36	日國使臣	23	墨使	15
本國	89	奴才	35	賠償	22	教案	15
領事官	89	鄭藻如	35	續約	22	警察	15
日國	88	美廷	34	傳教	22	苛例	15
華商	87	貴大臣	34	德國	22	自主	14
地方官	87	洋商	33	外洋	22	苛待	14
政府	86	新約	33	祕國	22	滿洲	14
總理衙門	78	土貨	33	朝廷	21	釐金	14
大西洋國	75	會議	33	民生	21	中華	14
金山	74	美使	32	外國	21	駐美	14
通商	69	進口	31	甲必丹	20	美民	14
大清國	66	關道	31	保護華工	20	管轄	14
總署	65	海關	30	舊約	19	督撫	14
條款	65	禁止	30	約章	19	流寓	14
內地	64	漢文	30	墨國	19	護照	14
互換	62	口岸	30	俄人	19	護路兵	14
教士	61	李鴻章	30	上海	19	古巴華工	14
出使	60	張蔭桓	30	中國人	19		
出洋	56	專條	29	直隸臨城礦務局	19		
註冊	55	各埠	28	日本國全權大臣	19		
聲明	55	貨物	28	貴國	18		
借款	54	香港	28	日人	18		
換約	52	立約	28	洛士丙冷	18		
招工	50	他國	28	蘆漢公司	18		
議院	50	駐紮	28	中國國家	18		

資料來源：「中國近代思想史專業數據庫（1830-1930）」

透過數位方法協助，能夠找到與「華工」相關之重要關鍵詞，而不致於受限於研究者主觀想像中的幾個觀念。例如若以傳統方式來處理「華工」事件，可能會思考到的是「古巴」、「美國」、「墨西哥」等等曾出現「華工受虐」問題的地點，或者如「保護華工」觀念；但從詞表中，卻可看到與「華工」共現的如「華人」、「華民」、「華商」、「華僑」等一組詞叢，這時即可刺激人文研究者去思考「華工」事件與「華人」意識之間的關係為何，又與「華民」、「華商」、「華僑」之間有何種互動關係。

當研究者從詞表中看到了「華工」事件與「華人」意識之間可能的關係時，以往人文學者會開始找尋曾經一起討論「華工」與「華人」的相關文獻加以閱讀，進而分析並連綴成文，形成一條從「華工」事件引起「華人」意識興起的論述脈絡。這樣的傳統論述雖然可行，然而卻未能夠舉出有利的證據去證明「華工」事件確實引起「華人」意識興起。而今處理這個問題，因數位人文學的出現而有新的進路。關鍵詞共現現象的數位人文研究表明，「華工」事件因「華人」意識的覺醒被賦予新的意義，甚至「華工」事件之所以被清廷所重視，從原本的「天朝棄民」轉而成為清廷所極力保護的對象，是由於「華人」意識之興起。

三、共現詞頻分析

首先，我們可以透過「詞彙比例」的計算，來確定「華工」史料中，「華人」關鍵詞的歷年重要性。先計算《清季外交史料》中從1875-1911年論及「華工」議題的史料總字數，並各自除以每一年的史料字數，得出每一年「華工」文本的重要性比例；接著以《清季外交史料》中從1875-1911年「華工」議題史料中「華人」一詞出現的總詞頻254次，各自除以每一年「華人」一詞出現詞頻數，得出每一年「華人」一詞出現於「華工」文本中的重要性比例，再將兩者搭配來看，若是「華人」關鍵詞重要性比例，超過該年「華工」的文本重要性比例，則可推斷出該年「華人」一詞於「華工」文本中具有重要性，為「關鍵詞高比例年份」，也可判斷那是「華人」意識重要的年份，如表3。

從圖表中可以看見粗體字為以超過比例為1.1倍作為依據，可以初步看出幾個重點時間。若再以2倍作為依據，則可更明確看出在《清季外交史料》「華工」史料中「華人意識」被討論的幾個關鍵年代，如表4所示。

研究者從表4即可快速且明確的判斷出在「華工」論述語境下「華人」意識的重點共現年份，而研究者即可就這些年份加以判斷與分析，為何這些年代「華人」意識會與「華工」論述呈現高度相關。

表3 《清季外交史料》華工文獻中「華人」意識重要年份分布表（1.1倍）

年份	1875	1876	1877	**1878**	1879	1880	1881	1882	1883	1884	**1885**	**1886**	1887	**1888**
華工史料歷年總字數／華工史料該年字數	0.057	0.009	0.054	**0.016**	0.007	0.013	0.005	0.000	0.004	0.003	**0.003**	**0.051**	0.045	**0.047**
華人一詞歷年總詞頻／華人一詞該年詞頻	0.028	0.000	0.059	**0.031**	0.004	0.012	0.000	0.000	0.000	0.000	**0.020**	**0.094**	0.000	**0.098**
年份	**1889**	**1890**	1891	1892	1893	1894	1895	1896	1897	1898	1899	**1900**	1901	**1902**
華工史料歷年總字數／華工史料該年字數	**0.062**	**0.003**	0.000	0.000	0.037	0.064	0.055	0.041	0.000	0.003	0.009	**0.009**	0.022	**0.081**
華人一詞歷年總詞頻／華人一詞該年詞頻	**0.142**	**0.008**	0.000	0.000	0.020	0.047	0.047	0.000	0.000	0.000	0.000	**0.024**	0.004	**0.161**
年份	1903	1904	1905	1906	**1907**	**1908**	1909	1910	**1911**					
華工史料歷年總字數／華工史料該年字數	0.011	0.021	0.092	0.000	**0.077**	**0.020**	0.036	0.032	**0.010**					
華人一詞歷年總詞頻／華人一詞該年詞頻	0.008	0.000	0.000	0.000	**0.114**	**0.031**	0.024	0.012	**0.012**					

資料來源：劉昭麟博士統計，邱偉雲製表

表4 《清季外交史料》華工文獻中「華人」意識重要年份分布表（2倍）

年份	1875	1876	1877	1878	1879	1880	1881	1882	1883	1884	**1885**	1886	1887	**1888**
華工史料歷年總字數／華工史料該年字數	0.057	0.009	0.054	0.016	0.007	0.013	0.005	0.000	0.004	0.003	**0.003**	0.051	0.045	**0.047**
華人一詞歷年總詞頻／華人一詞該年詞頻	0.028	0.000	0.059	0.031	0.004	0.012	0.000	0.000	0.000	0.000	**0.020**	0.094	0.000	**0.098**
年份	**1889**	**1890**	1891	1892	1893	1894	1895	1896	1897	1898	1899	**1900**	1901	1902
華工史料歷年總字數／華工史料該年字數	**0.062**	**0.003**	0.000	0.000	0.037	0.064	0.055	0.041	0.000	0.003	0.009	**0.009**	0.022	0.081
華人一詞歷年總詞頻／華人一詞該年詞頻	**0.142**	**0.008**	0.000	0.000	0.020	0.047	0.047	0.000	0.000	0.000	0.000	**0.024**	0.004	0.161
年份	1903	1904	1905	1906	1907	1908	1909	1910	1911					
華工史料歷年總字數／華工史料該年字數	0.011	0.021	0.092	0.000	0.077	0.020	0.036	0.032	0.010					
華人一詞歷年總詞頻／華人一詞該年詞頻	0.008	0.000	0.000	0.000	0.114	0.031	0.024	0.012	0.012					

資料來源：劉昭麟博士統計，邱偉雲製表

而我們還可以再以「華工」與「華人」這兩個關鍵詞的共現現象來進行二度觀察，確認在「華工」論述語境下「華人」意識的重點共現年份。首先我們使用「共現詞頻分析法」，同樣以《清季外交史料》中109篇「華工」史料為底本，切出自然關鍵詞，經過研究者篩選，去除非詞彙後餘下有意義詞彙，形成詞表，再由詞表進行兩兩配對，並統計配對共現詞組的詞頻，以「歷年共現詞組總組數」除以「該年共現詞組總組數」，得出「每一年共現詞組重要性比例」。再以所欲研究的「共現詞組」出發，以該共現詞組（例如「華工與華人」這一組）的「歷年共現詞組總次數」，除以「該年共現詞組總次數」，得出「每一年該共現詞組（如「華工與華人」）的共現重要性比例」。透過兩者比較，即可了解「共現詞組」在歷時性上的重要分布年代，而且透過數位方法的兩兩排列，可以得出超越研究者可設想範圍之共現詞組現象，這是數位人文學之長處所在。以「華工」為觀察對象的共現詞組分布，見表5；以「華人」為主要觀察對象的共現詞組分布，見表6；其餘還有「華民」、「華商」共現詞組分布見表7。

上述這些自然配組的詞叢，可以向研究者展示關鍵詞間彼此交流的過程，並可提供給研究者新的思考方向。而本文主要處理的是華工事件與華人意識之間的問題，故將焦點鎖定回兩個關鍵詞共現的歷年分布狀況如表8。

表5 《清季外交史料》華工文獻中1875-1911年「華工」共現詞組表

華工	領事	華工	續約	華工	雇用	華工	駐紮
華工	華民	華工	註冊	華工	擬請	華工	擬定
華工	華商	華工	新約	華工	新章	華工	總統
華工	衙門	華工	貿易	華工	會議	華工	議定
華工	禁止	華工	華工赴美	華工	照約	華工	議約
華工	照會	華工	貴國	華工	照覆	華工	鐵路
華工	議院	華工	領事官	華工	禁止華工		
華工	貴大臣	華工	辦法	華工	稟報		
華工	禁之	華工	總理衙門	華工	弊端		
華工	辦理	華工	薩島	華工	德國		

資料來源：劉昭麟博士統計，邱偉雲製表

表6《清季外交史料》華工文獻中1875-1911年「華人」共現詞組表

華人	華工	華人	照會	華人	德人
華人	領事	華人	衙門	華人	澳門
華人	華民	華人	進口	華人	聲明
華人	關道	華人	歐美日本人	華人	驅逐
華人	華商	華人	續約		
華人	新例	華人	禁止		
華人	貿易	華人	遊歷		
華人	辦理	華人	領事官		
華人	總統	華人	總理衙門		
華人	註冊	華人	議院		

資料來源：劉昭麟博士統計，邱偉雲製表

表7《清季外交史料》華工文獻中1875-1911年「華民」、「華商」共現詞組表

華民	領事	華民	貿易
華民	領事官	華民	禁之
華民	總領事	華民	經商
華民	衙門	華民	照會
華民	華商	華民	稟報
華民	註冊	華民	墨境
華民	總統	華民	稽查
華民	傭工	華商	領事
華民	辦理	華商	機器
華民	總理衙門	華商	衙門

資料來源：劉昭麟博士統計，邱偉雲製表

表8《清季外交史料》華工文獻中1875-1911年
「華人」與「華工」共現詞頻年度分布表

年份	1875	1876	1877	1878	1879	**1880**	1881	1882	1883	1884	**1885**
歷年共現詞組總組數／該年共現詞組總組數	0.090	0.003	0.109	0.037	0.011	**0.024**	0.008	0.000	0.008	0.002	**0.001**
歷年華工與華人共現總次數／該年華工與華人共現總次數	0.019		0.019			**0.057**					**0.057**
年份	**1886**	1887	**1888**	1889	1890	1891	1892	1893	1894	1895	1896
歷年共現詞組總組數／該年共現詞組總組數	**0.026**	0.061	**0.034**	0.056	0.001	0.000	0.000	0.025	0.078	0.056	0.029
歷年華工與華人共現總次數／該年華工與華人共現總次數	**0.151**		**0.208**	0.094					0.094	0.132	
年份	1897	1898	1899	1900	1901	1902	**1903**	1904	1905	1906	1907
歷年共現詞組總組數／該年共現詞組總組數	0.000	0.004	0.013	0.021	0.006	0.051	**0.005**	0.009	0.162	0.000	0.027
歷年華工與華人共現總次數／該年華工與華人共現總次數						0.057	**0.019**				0.038
年份	1908	1909	**1910**	**1911**							
歷年共現詞組總組數／該年共現詞組總組數	0.004	0.025	**0.011**	**0.004**							
歷年華工與華人共現總次數／該年華工與華人共現總次數			**0.038**	**0.019**							

資料來源：劉昭麟博士統計，邱偉雲製表

由上可知，「華工事件」與「華人意識」之間密切相關之年代為1885年、1886年、1888年這三年，可以看見這三年乃是「華工」事件與「華人」意識結合的關鍵年代，而從「共現詞頻分析法」中，透過數位人文方法的便捷，我們可以快速羅列出歷年「共現詞組」的分布篇章，如表9、表10、表11所示可以看見，「華工」事件真正與「華人」意識明顯共現，是從美國對華工之虐待事件中所產生。而1885年的華人與華工高峰，正呼應《近代中國史事日誌》1885年9月2日所載美國窩民（Wyoming）州慘殺華工，焚燒房屋，死19人（一說28人），被逐出洛士丙冷（Rock Springs）者約六百人之洛士丙冷慘案（郭廷以編著，1987：785）。

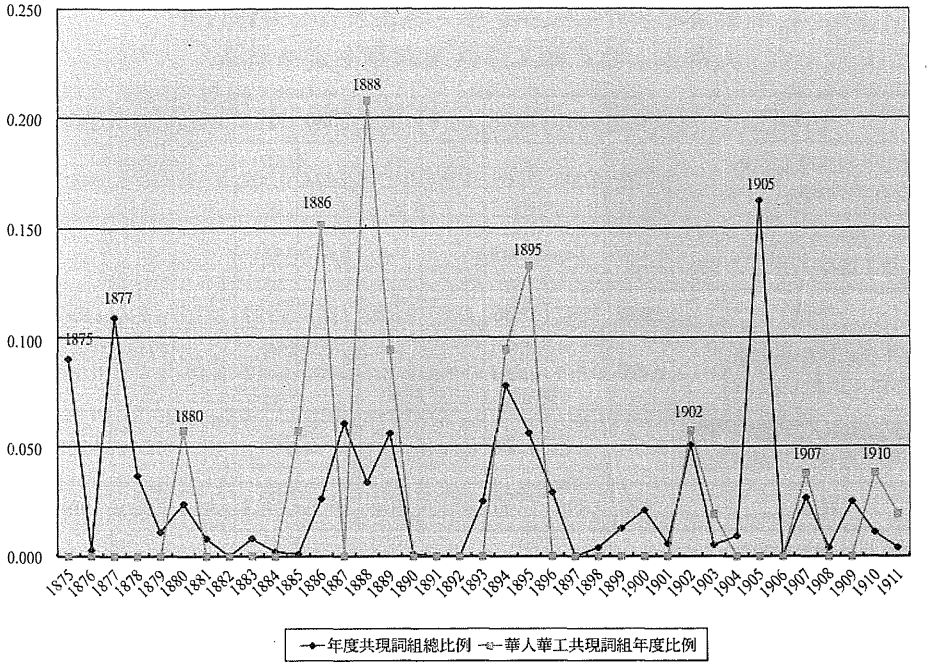


圖3 《清季外交史料》華工文獻中1875-1911年「華人」與「華工」詞頻年度比例分布圖

表9 《清季外交史料》華工文獻中1885年「華人」與「華工」共現詞組重要篇章分布表

1885年檔案內容分析		共現詞組權重	1.8	0.17	0.6
		共現詞組年度比例	0.6	0.057	0.2
		共現詞組年度頻度	3	3	3
		詞彙一	洋官	華人	土人
		詞彙二	華人	華工	華人
文件權重	作者	篇名			
12	鄭藻如	使美鄭藻如致總署洛士丙冷案田使若來剖辨乞嚴詞拒之冀美廷速辦電	0	3	0
7	鄭藻如	使美鄭藻如致總署報美煤工焚斃華工電	3	0	3

資料來源：劉昭麟博士統計，邱偉雲製表

表10 《清季外交史料》華工文獻中1886年
「華人」與「華工」共現詞組重要篇章分布表

1886年檔案內容分析		共現詞組權重	2.79	1.29	1.21	0.54	0.06	0.04
		共現詞組年度比例	0.31	0.429	0.151	0.109	0.031	0.042
		共現詞組年度頻度	9	3	8	5	2	1
		詞彙一	華工	保護	華人	保護	保護	保護
		詞彙二	華商	華商	華工	華民	華工	華人
文件權重	作者	篇名						
241	粵督張之洞	粵督張之洞奏舊金山華民被害請催美國懲辦摺	1	2	1	4	2	0
101	張蔭桓	使美張蔭桓奏陳舊金山戕害華工案辦理情形片	8	0	5	0	0	0
54	舊金山中華會	舊金山中華會董致總署土人殺害華人求請美廷保護電	0	1	0	0	0	0
40	粵督張之洞	粵督張之洞致總署美使來粵晤談洛士丙冷案允電外部電〔附旨〕	0	0	0	1	0	0
23	李鴻章	直督李鴻章致總署鄭使電美土人戕害華工電	0	0	2	0	0	0
23	粵督張之洞	粵督張之洞致樞垣請敕電鄭使商美廷保護華工電	0	0	0	0	0	0
18	粵督張之洞	粵督張之洞致總署美害華工領事來牘道謝西人因懼生訛乞商禁電	0	0	0	0	0	1
17	旨	旨寄張之洞外傳該督因美國華工被害將圖報復有無此議著電聞電	0	0	0	0	0	0

資料來源：劉昭麟博士統計，邱偉雲製表

表11《清季外交史料》華工文獻中1888年
「華人」與「華工」共現詞組重要篇章分布表

1888年檔案內容分析		共現詞組權重	10.29	2.28	2.5
		共現詞組年度比例	0.857	0.208	0.5
		共現詞組年度頻度	12	11	5
		詞彙一	小呂宋	華人	入境
		詞彙二	領事	華工	華人
文件權重	作者	篇名			
352	兩廣總督張之洞	粵督張之洞奏訪查南洋華民情形擬設小呂宋總領事以資保護摺	12	4	0
141	張蔭桓	使美日祕張蔭桓奏與美外部議訂華僑善後事宜摺	0	3	0
45	張蔭桓	使美日祕張蔭桓致總署中美立約係舊約不用寶電〔二件〕	0	3	4
34	張蔭桓	使美日祕張蔭桓奏陳兼使祕魯所至情形摺	0	1	0
24	張蔭桓	使美日祕張蔭桓致總署聞美廷擬禁止華工再來電	0	0	0
24	李鴻章	直督李鴻章致總署美新約禁華工往請緩訂電	0	0	0
23	張蔭桓	總署致張蔭桓商華工新約美使已電美外部電	0	0	0
12	張蔭桓	使美張蔭桓致總署美議院議定華工須領新照電	0	0	0
8	張蔭桓	使美日祕張蔭桓致總署美禁華工案請乘機結束電	0	0	1
8	李鴻章	直督李鴻章致總署美立苛約禁華工進境請設法拯救電	0	0	0
2	張蔭桓	使美日祕張蔭桓致總署美議院議禁華工新約請定準駁電	0	0	0
0	張蔭桓	總署致張蔭桓美定華工新例請力辨電	0	0	0

資料來源：劉昭麟博士統計，邱偉雲製表

而1888年的高峰，正呼應1888年3月13日出使美國大臣張蔭桓與美國訂立限禁華工赴美條約六款。1888年10月1日美國總統批准限禁華工入境案。1888年12月26日李鴻章再電總署，美國完全不准華工入境，自粵到美輪船，均被迫將華人載回。（郭廷以編著，1987：818-824）這裡可以明顯看出清人已將華工事件與華人意識結合，以往是分開的，而如今美國除了禁止華工之外，更擴大禁止華人入境，禁止華商去經商，中國內部開始重視海外華人所帶來的外匯經濟，所以至此保護重點已不再是傳統記憶中的天朝棄民，而是保護華人和華商。從華工與華人共現詞頻高峰的1888年可知，華工事件成為華人意識抬頭的導火線，因為美國藉著華工問題而限禁華人，這也導致中國為了維護華人權益，自然應當從保護華工入手，因此華工事件從1888年以後，已經不單純是華工問題，而成了中國在國際上地位升降的比武場，故而為了維護中國的形象，清廷自然得保護華人，保護華人自然以保護華工為主，故而與以往保護華工出於自認為自己是天朝上國心態不同，1888年之後的保護華工，是為自己的顏面留下最後一絲尊嚴，故華工事件雖一，但由於加入華人意識，故而展現了不同的華工內涵，從這裡也可看出事件對於觀念之影響所在。

由上述諸表格可以看出「共現詞頻分析」的長處，可以減少人文學者對於資料擷取的時間，可透過數位協助，快速的進入文本之中進行閱讀與分析。

四、從數據到關鍵詞意義結構

提到「華工」一詞，很多人會聯想到晚清所謂的「賣豬仔」，認為所有華工都是被賣到國外的奴隸，這個說法目前已有學者提出反駁，認為「豬仔」與自由移民美國的華工不能混為一談（蔡石山，1971：198-199）[3]。而本文所談到的「華工」內涵，則跨越了從被賣到古巴、祕魯的「豬仔」到自由移民往美國的「華人」。而從「豬仔」到「華人」的過程，隱含著是中國對於海外子民的認同心態轉移，因此從清廷對於華工事件的處理態度，亦可看出中國對外的心態轉變史。

從歷史文獻中可知，中國對於海外子民的態度，從明清以來即是以坐視不管、放任自生的態度去面對：「『逃民』、『罪民』和『潛在的漢奸』這些形象，與對商賈的傳統偏見一起，構成了明代對海外華人敵視政策的基礎。凡返回中國的海外華

3 參見蔡石山：「一八五〇年代至一八七〇年代之間，一種慘無人道的販賣人口的勾當猖獗於我國的東南沿海。這一種人口販賣當時的人稱之為『豬仔』販運，洋人稱之為 coolie trade（苦力貿易）。因為『豬仔』販運盛行時恰好華工大量地湧入美國，以致很多人誤會移往美國的華工即是類似奴隸般的『豬仔』，這是大錯特錯的誤解！美國西部工人在十九世紀末葉排華暴動時，動輒宣傳華工都是『豬仔』，他們謊稱所有華工都是受役於六大會館（美國人稱The Six Companies）的奴隸，所以華工都是不自由的，受人宰割的。為了解放華工，使他們重獲自由，美國工人才會有排華的運動。這些冠冕堂皇的『理由』是美國反華份子造出來辯護他們犯罪行為的護身符。」（蔡石山，1971，〈華工與中美外交〉，《美國研究》，3，頁198-199。）

人，都受到拘捕和懲處。在明代晚期，當海禁最終解除時，朝廷對海外華人的敵視只不過是降到漠不關心的程度而已，對於他們身處海外，並不予以保護。」（顏清湟著，粟明鮮、賀躍夫譯，1990：14）從這裡可以看出華工為何被視為天朝棄民的傳統記憶。

自明清以來被視為天朝棄民的華工，是如何成為晚清極力保護的對象？即是一個必須要梳理的脈絡，且若是要理解晚清清廷政府的外交心態轉移，也必定要從清廷對於華工事件態度的轉變來著手。如蔡石山在〈華工與中美外交〉一文中提到：「崔國因在光緒十九年九月初四奏摺中云：『……美日祕三國交涉之事以美國為繁，美國交涉之事以保護華民為要。』我們可以說，一部早期中國對美外交史絕對脫離不了移民與護僑的問題。」（蔡石山，1971：197）從上述中，可以了解，中國對外關係之態度改變，可以從對待華工事件的處理方式來觀察。即沿襲以往對於明清以來忽視華工之態度，轉而成為保護華工之態度。在這轉變之間，看似從一「不保護」到「保護」之簡單過程，但其中牽涉了整個中國的「自我認同」，以及在國際上的「自我定位」。換言之，以往中國強盛為「天下中心」時，華人並不會受到欺壓，但隨著中國在「他者」心中形象漸漸的呈現敗象，各國也就開始欺壓華人，這時面臨從「天下觀」轉變到「萬國觀」之清廷，為了護住天朝上國的尊嚴，自然會對任何可以強化自我認同之機會加以把握，而在以「華工」受虐為核心的事件中，正是一個清廷不斷對外進行談判、協商、退讓的自我宣示，也是一個自我認同的轉換過程。透過整個華工事件長達三十年的處理過程，中國在這段期間也逐步從「天下觀」轉變到「萬國觀」，而華工事件真正得到解決，也是等到中國真正的接受萬國觀念，捨棄天朝心態並引《萬國公法》自我保護下，才真正的透過國際公法完成了保護華工之工作。（金觀濤、劉青峰，2008：231）而其實華工事件只是引子，清廷真正要保護的乃是大批的「華人」，而這批華人背後所象徵與隱喻的，是中國的「尊嚴」、「破敗」與「不堪」。「尊嚴」指的是維持那「名存實亡」的天朝上國「薄面」；「破敗」是指保護華工目的在於不讓華工歸國，以避免殘破不堪的中國經濟更加危殆；「不堪」是指不僅不讓華工回國，甚至還要希望他們能在國外賺取外匯，補助中國。因此從「華工事件」出發，形成「保護華工」與「保護華人」的兩個面向，基本上就是在上述三個考量之下所形構而成。

而目前研究華工問題者，多聚焦於美國如何迫害華工、清廷對華工所持的態度來進行說明研究。說明了清廷從不保護華工到保護華工的心態改變、保護華工的理由、保護華工的協商過程、保護華工的結果。上述這些論述前人多已說明，但基本上乃是建立於少數文獻的引證，而本文即試圖通過數位人文學的方法，以可重複驗證之數據，來建立前人研究結果的正當性，或是補充其不足。更希望透過這樣的嘗

試，能夠提供人文研究者一種新的研究方法與工具，是為本文之期盼，因而以下即以透過數位人文方法所得出之結果，來與前人研究對話，觀察數位人文研究引入文史研究中的輔助性與可行性所在。

五、保護對象的改變

前人對於華工研究中，皆將重點關注於保護的行動與過程，透過這樣的論述，讀者可以擬構出一條清廷對於華工的保護心態發展譜系，但這條譜系雖然道出保護華工的發展過程，卻未觀察到「保護華工」背後的「不在場訊息」，即「保護」對象的轉移。

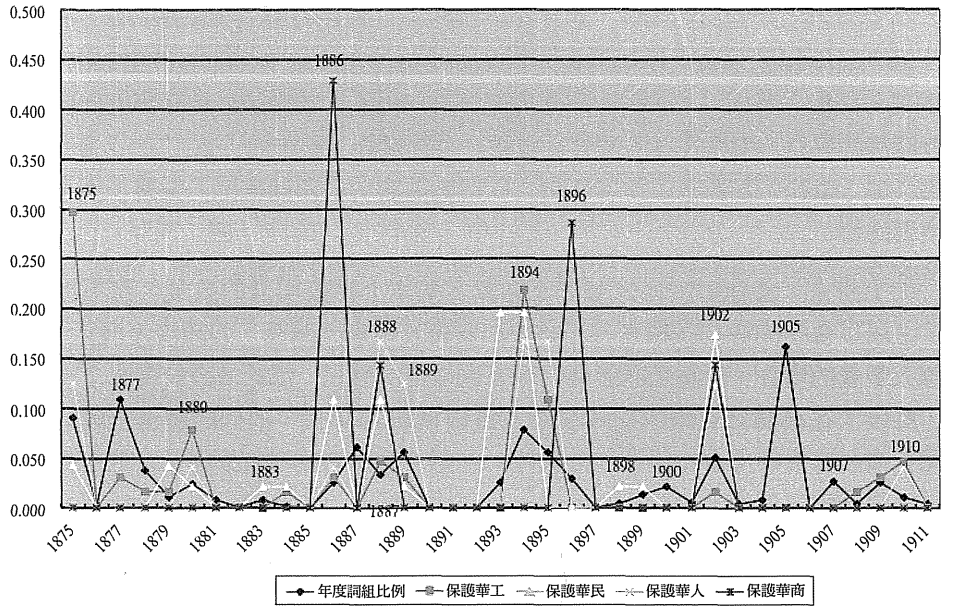
本文從數據中觀察出來，華工事件確實引發了清廷保護海外華人的心態，但這樣的保護觀念，卻有一條從保護「華工」向保護「華人」移動的過程，即透過「華工事件」引發「華人意識」的發展趨勢。這樣的趨勢，極難從人文研究者以傳統閱讀綜理分析中梳理出來，由於時間跨幅長，且文獻量多，除非浸淫日久，才有辦法產生這樣論述的構圖，而如今透過數位協助，可以使人文研究者快速的繪出觀念發展圖形，並且從圖形中很快的觀察出一些殊異現象，而從這些殊異現象中客觀的去閱讀史料與進行歷史解釋。故由上可知，數位人文學並非以取代人文學研究為宗旨，而是以輔助人文研究者快速觀察出歷史殊異點，協助研究者快速進入研究焦點核心為依歸，這是數位人文學發展的殷切期盼。

以下我們即從透過「共現詞彙分析法」所得出之分析表中，擷取出以「保護」為主軸，搭配「華工」、「華民」、「華人」、「華商」等詞彙之共現詞組比例，進行歷時性的排列，觀察在《清季外交史料》中「保護」觀念與「華工」、「華民」、「華人」、「華商」等觀念共現密切度之歷時性展演，並藉此得出「保護」觀念對象的演變過程。

如上所言，清廷對於海外華工之態度，基本上延續明清的政策，而不加以重視，之所以會在晚清有所轉變，乃是由於會傷及中國之顏面，如顏清湟提到：「海外華人數量的增長，無論是通過契約勞工還是自由移民，都給清政府帶來了一些問題。許多人在異鄉受到虐待或被殺死，到19世紀70年代，他們的遭遇已成為國際醜聞。清政府對華工本來並無什麼感情，然而，這些醜聞不僅損害了它的尊嚴，也有損於它的國際地位。這個新的現實迫使清政府採取一些保護措施，以保護海外華人。」（顏清湟著，粟明鮮、賀躍夫譯，1990：12）而清廷保護對象也是有轉移與擴大過程的：「總理衙門的照會與新的章程，也許是清政府對有關移民問題所曾頒佈的文件中最重要的文件。這自然是清政府按照自己的主張來管理招工，並用它認為

合適的方式來保護移民的第一次嘗試。這也是政府首次視保護海外移民為對其臣民的應盡之責。這種責任感後來被擴大對所有海外華人的保護。」（顏清湟著，粟明鮮、賀躍夫譯，1990：110）既然前人已看出保護傘擴大現象，但還可以追問的是這個擴大程序是如何發動？又這擴大的主要發動者為誰？即可從數位人文圖示中看出。

從圖4中可以看到，保護觀念的擴大階段，從1875-1885年「華工」與「保護」觀念共現頻度有增強趨勢；而1886年「華商」與「保護」觀念共現頻度也突顯出來；而1888年「華人」與「保護」觀念共現頻度也有明顯的高點；1893年「華民」與「保護」觀念共現頻度則為最高峰；而1894年「華工」與「保護」共現頻度產生最高峰，而同時也伴隨著「保護華民」與「保護華人」共現頻度的最高峰。而這顯示 1894年「華工」、「華人」、「華民」三個觀念在該年都有高度使用。而1896年「華商」與「保護」觀念繼1886年後又再次出現高峰；而在1902年後可以看見，「保護」與「華工」的共現現象已經大幅降低，而「保護」與「華民」、「華商」、「華人」觀念共現頻度則增高。由上可以看出的是清廷官方的保護對象，具有一條保護意識的擴張變化，即「保護華工」的共現觀念從1895年後就未出現高峰，而「保護華民」、「保護華商」、「保護華人」則從早期零星出現到1895年後皆各自呈現高峰，從這可以看出「甲午戰敗」後，中國自信心大受打擊，而西人更加輕蔑，故不僅只壓迫華工，於1895年後西人更擴大壓迫對象至華人、華民與華商，故這裡即



資料來源：劉昭麟博士統計，邱偉雲製圖

圖4 《清季外交史料》華工文獻中1875-1911年「保護」觀念共現詞頻年度比例分布圖

可看出清廷與西方的競逐焦點，有從早期針對「華工」至晚期針對「華人」、「華民」、「華商」的趨勢。[4]

而這裡可以看出華工事件所引發的一個重要觀念轉移，即對於華商的重視，更進一步的說，是由華商主導了整個中國對於海外華人的重視，這個關鍵即在於中國對於商人意識的改變。這個改變可以從郭嵩燾對於海外華商貿易之重要性強調中看出來，郭嵩燾認為：「由於貿易為一國之本，西方政府要求開放口岸，並派駐領事，以保護商人和他們的商業活動。與此相反，中國對貿易卻不感興趣，它既不想博得商人的支持，也不想去保護他們。為了補救以往的失誤，中國應效法西方，保護它的那些散居海外並已數代定居於當地的商人。」從這個觀點出發，郭嵩燾便要中國外交官在保護華工之外，還要發揮另一種重要作用，即保護海外華商，以求獲得他們對中國經濟近代化的支持。（顏清湟著，粟明鮮、賀躍夫譯，1990：152）

由此可知，清廷對於華商觀念的改變，也成了華人意識興起的一個重點。而從「保護」與「華商」觀念的結合圖來看，這個結合是從1886年才開始在清廷官方史料中形成論述，故由此可知，即使郭嵩燾在1877年即已看到華商對於中國經濟的重要性，但整個中國內部卻要在十年之後才真正開始重視商民，而1886年這時間點，即是洛士丙冷案發生的時間，故可說「保護」觀念與「華商」、「華民」、「華人」觀念之合流，乃是由於洛士丙冷慘案這事件所引起的觀念內涵轉移現象，如顏清湟所言：「洛士丙冷大慘案是中國改變它對保護美國僑民態度的一個重要里程碑。此案引起總理衙門和一些高級官員的很大注意，迫使他們在這一問題上採取更強硬的立場。其結果是中國獲得一次小小的外交勝利——美國答應賠償大屠殺中的受害者。這次小勝利主要應歸功於中國駐華盛頓公使鄭藻如和兩廣總督張之洞的努力。」

4 共現觀念詞組間重要性的比較，應當以單一共現觀念的重要性論述為主，如說明「保護」與「華工」共現詞組有一個從高到低的觀念發展過程，而「保護」與「華人」共現詞組有一個從低到高的觀念發展過程，但不可將兩個共現詞組進行比較，認為這即是一種保護華人共現觀念取代了保護華工的共現觀念，因為比例高可能其背後數據的共現詞頻低，而比例低也可能實為高共現詞頻，若驟然進行不同共現詞組比例之間的比較，恐會有所歧出，故應謹慎觀察文本，觀察不同共現詞組升降起伏的交叉點，以交叉點的文本閱讀作為根據，觀察其中是否確實有共現詞組之間的取代關係，如此論述將更為可靠與確證。換言之，即是在數位人文方法的協助下，再次經過人文研究者的判讀與檢證，即可免除誤讀之可能。如1896年保護華工雖然次數多，但從文本中可以看見當年華工次數多乃是因為排華法案將商人與學生納入排斥身分中，故而當時是以華商為討論核心，而華工的使用出現乃是出現於被當作附錄之條約中，僅是一種附錄引用，並非真正關注的核心。換言之，詞頻次數多，有可能是被作為一種次要的語彙，無任何深意的引用，然而真正關心的詞彙正因開始萌芽故而次數雖少，但卻可見比例躍升的過程。從此看來，詞彙次數不是關鍵，因為可能是一種「無目的性詞彙」，但卻不能忽略，如「我的」一詞詞頻必然高，但卻不是一個重要且有目的性的關鍵詞彙，而是從比例中去進一步觀察觀念間的起伏升降才是關鍵。因此本文認為所謂詞彙，應有一個從「焦點性關鍵詞彙」轉向一個「非焦點性關鍵詞彙」的過程。如華工一詞雖然次數到了1896年仍多，但已非焦點性詞彙，僅是討論華商被美國列入排擠行列問題時帶到的詞彙，故而即使次數多，也不能將其視為比華商觀念於當時更為重要。反而是在1896年甫現之「華商」一詞，雖然剛出現次數很少，卻每次都是「焦點性關鍵詞彙」的展演，故我們得知其重要性。我們日後也許觀察所謂觀念之間的取代過程，或許不能只以詞頻，而應以比例，因為比例中可以看出所謂從「焦點性關鍵詞彙」轉向「非焦點性關鍵詞彙」之過程，此點有待討論。

（顏清湟著，粟明鮮、賀躍夫譯，1990：238-239）而從上述的圖示與歷時性敘述中可知，清廷對於海外華人的保護具有一擴大的趨勢，也很清楚可以看見發動的過程。[5]在這保護對象的轉移之中，可以看見的是「華人意識」乃是從清廷處理「華工事件」中，不斷的發展所產生，此即是事件引發觀念形成的一條發展譜系。

六、限禁對象的改變

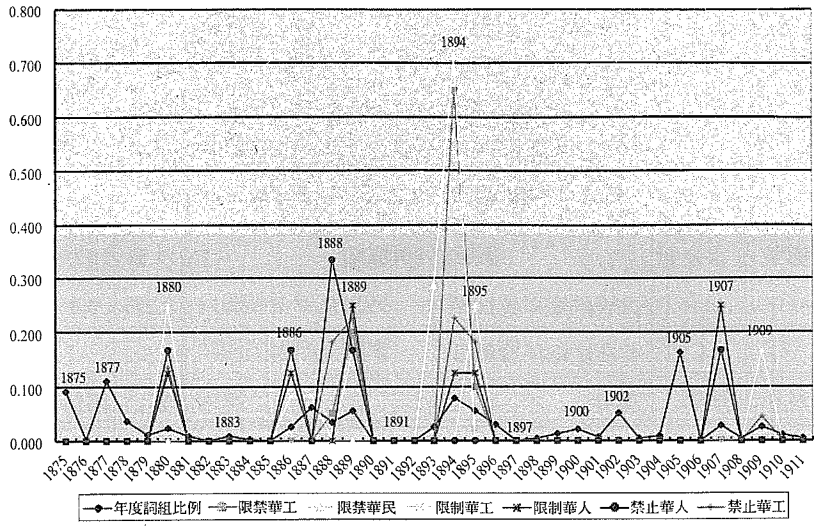
美國從1878年開始在加利福尼亞州提出排華法案，法案內容包括禁止華人擁有財產，禁止華人經營貿易，不准華人加入美籍，不准華人攜帶武器，不准華人在涉及白人的案件中到法庭作證等。並且添加這些違反美國聯邦憲法的法條進入該州新憲法，提供合法性的依據，更禁止移入華人擔任公共工程之工人，否則該公司會受罰。由此可以看見，美國人在當時已將所有移入美國的「華人」都視為「華工」，故而美國的「排華」並非限禁「華工」，而是擴大到整個「華人」圈，因此美國對於「排華」所排對象的擴大，也影響了中國將保護的對象從華工擴大到所有華人之上，這是美國對於限禁對象的擴大造成中國保護對象擴大的一個脈絡所在。故從圖5可以看見這個限禁狀況的發展。

從圖5中，可以看出1878年之後，限禁的共現詞頻就有明顯的起伏，與「限禁」觀念共現者有「華工」與「華人」，而此時「限禁」與「華工」共現呈現高頻度現象，可知這組觀念在當時被加以重視。「華人」與「限禁」共現頻度在當時也呈現高頻度現象，可見當時，美國也已經有「限禁」、「華人」與「華工」之觀念是並起的。這樣的現象直到1886年仍然相同，兩組共現詞頻都維持差不多的頻度，圖示中可看見，「限禁」、「華人」之共現頻度與「限禁」、「華工」共同出現，但這樣的共同出現仍是在萌芽階段，可以說仍未達到大量運用的自覺階段，僅於文章中零星出現，見表12。

而在1889年限禁「華工」與「華人」這兩組共現觀念篇章散於三文，這表示限禁「華工」與「華人」的共現觀念，已經在各個討論中都被作為一個組合議題處理，唯有該共現觀念開始被自覺後，才會開始呈現普遍討論之分布，見表13。

而在1889年後相關的觀念討論又消失，直到1894年才真正大量出現限禁觀念頻度的躍升。故由上可知1886-1889年是華工與華人觀念開始混合階段，故而限禁華工與限禁華人的觀念開始被置放在一起討論。然而這一個現象直到1894年即有了突

5 如顏清湟所言：「在與西班牙和祕魯就有關華工保護問題的整個漫長談判中，清政府都試圖為其受壓迫的子民獲得盡可能好的條款，但他的嘗試卻因西方列強公使的干涉而受到限制。儘管如此，1876年中國—祕魯條約與1878年中國—西班牙條約的締結，都充分表明了清政府保護華工的決心。這種保護精神後來擴大到了對世界上其他地方華僑的保護。」（顏清湟著，粟明鮮、賀躍夫譯，《出國華工與清朝官員：晚清時期中國對海外華人的保護　1851-1911年》，頁130。）



資料來源：劉昭麟博士統計，邱偉雲製圖

圖5 《清季外交史料》華工文獻中1875-1911年「限禁」觀念共現詞頻年度比例分布圖

表12 《清季外交史料》華工文獻中1886年「限禁」共現詞組重要篇章分布表

1886年檔案內容分析		共現詞組權重	0.17	0.08	0.13
		共現詞組年度比例	0.167	0.083	0.125
		共現詞組年度頻度	1	1	1
		詞彙一	禁止	限制	限制
		詞彙二	華人	華工	華人
文件權重	作者	篇名			
241	粵督張之洞	粵督張之洞奏舊金山華民被害請催美國懲辦摺	0	0	0
101	張蔭桓	使美張蔭桓奏陳舊金山戕害華工案辦理情形片	0	1	1
54	舊金山中華會	舊金山中華會董致總署土人殺害華人求請美廷保護電	0	0	0
40	粵督張之洞	粵督張之洞致總署美使來粵晤談洛士丙冷案允電外部電〔附旨〕	0	0	0
23	李鴻章	直督李鴻章致總署鄭使電美土人戕害華工電	1	0	0
23	粵督張之洞	粵督張之洞致樞垣請敕電鄭使商美廷保護華工電	0	0	0
18	粵督張之洞	粵督張之洞致總署美害華工領事來牘道謝西人因懼生訛乞商禁電	0	0	0
17	旨	旨寄張之洞外傳該督因美國華工被害將圖報復有無此議著電聞電	0	0	0

資料來源：劉昭麟博士統計，邱偉雲製表

表13《清季外交史料》華工文獻中1889年
「限禁」共現詞組重要篇章分布表

1889年檔案內容分析		共現詞組權重	1.14	0.8	0.33	0.5
		共現詞組年度比例	0.227	0.2	0.167	0.25
		共現詞組年度頻度	5	4	2	2
		詞彙一	禁止	限禁	限制	限制
		詞彙二	華工	華工	華工	華人
文件權重	作者	篇名				
379	鄭藻如	謹將前使臣鄭藻如咨總署自禁華工來美節略恭呈御覽	2	0	0	1
294	鄭藻如	謹將前使臣鄭藻如咨總署草約未成約稿恭呈御覽	2	3	1	1
257	張蔭桓	使美張蔭桓奏美約中輟請設法補救並述前使草約及美國新約摺	1	1	1	0
118	張蔭桓	使美張蔭桓奏美國積案清償完結摺	0	0	0	0
27	恭呈御覽	謹將美國現行新例恭呈御覽	0	0	0	0
2	李鴻章	直督李鴻章致總署何天爵使華恐難駕馭請酌辦電	0	0	0	0

資料來源：劉昭麟博士統計，邱偉雲製表

破，而這個突破則是由於「華工」與「華民」觀念已被混一認同。

此時清廷已經了解到美國不僅是要針對華工進行排擠，而是對整個海外華人都有意圖的進行排擠動作。當美國只是排斥華工之時，基本上中國仍會認為只是單一個案予以訂法條與進行領事館設置予以保護即可，但當意識到美國是針對整個華人圈都予以排擠之時，中國所欲保護的除了華人之外，就是中國在國際上的地位與尊嚴了，更包含了要維護國內財政穩定與保持外匯收入等經濟因素，在這狀況下，中國內部即自覺應當更積極的與美國進行談判與交涉，在這排華問題之上。因此可以看到1894年之時，整個討論串產生了高峰，這一年的「限禁」、「華民」與「華工」共現高峰討論，如表14。

楊儒〈使美楊儒奏與美外部重訂限禁華工保護華民約款摺〉中提到：「出使美日祕國大臣楊儒奏：為遵旨與美國外部重訂限禁華工保護華民約款事。竊奴才於上年七月接任之始，時值美國迫行華工，註冊新例，美西各邦紛紛拘人，方擬遣送回

表14《清季外交史料》華工文獻中1894年
「限禁」共現詞組重要篇章分布表

1894年檔案內容分析		共現詞組權重	8.45	3.57	1.14
		共現詞組年度比例	0.65	0.714	0.227
		共現詞組年度頻度	13	5	5
		詞彙一	限禁	限禁	禁止
		詞彙二	華工	華民	華工
文件權重	作者	篇名			
568	楊儒	使美楊儒奏與美外部重訂限禁華工保護華民約款摺	6	3	0
355	慶親王奕劻	總署奏重訂中美約款保護寓美華工摺	5	1	4
243	楊儒	使美楊儒奏通籌寓美華民善後事宜並派員赴墨西哥察看情形摺	2	1	0
134	奕劻	總署奏請飭楊儒妥議中墨約章請旨遵行片	0	0	1
114	慶親王奕劻	總署奏重訂中美保護華工約本請批准摺	0	0	0
35	奕劻	總署奏交犯專約應由楊儒照會美外部訂期畫押片	0	0	0
34	楊儒	使美楊儒奏派美日祕三國參贊領事摺	0	0	0

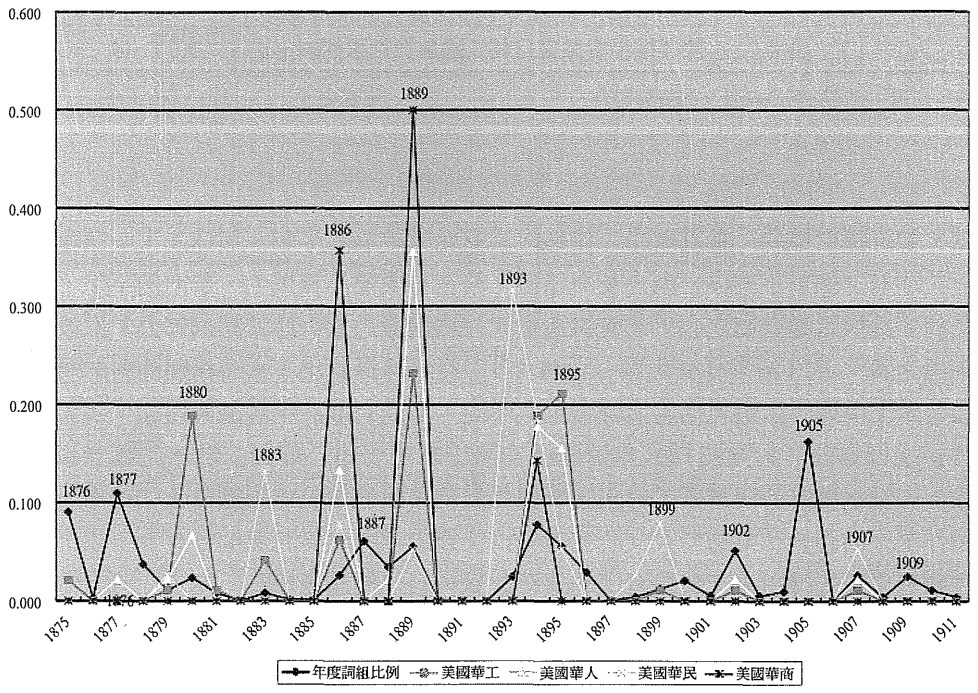
資料來源：劉昭麟博士統計，邱偉雲製表

華，當經援引條約，嚴詞駁詰，美外部始商允議院展限半年，被拘工人一律釋放，而於註冊之例堅不改移。竊念銜命遠使，務在尊國體而奠民生，若前例不除，工商交病，既無以盡保衛之心，而使職所重尤在結兩國之歡，亦當開誠佈公，庶彼此無不達之隱。」(《清季外交史料　1894年》) 從這可以看出，當時楊儒面對的問題不僅是華工問題，而還包括著海外華工與華商，並且有關於中國國家體面，故而才與美國進行協商。可見在1894年清廷已有將「華工」與「華民」混一之觀念產生，故而方派楊儒簽訂限禁華工與保護華民的條約。故而從上述由數位人文技術所協助製作出來的圖表，有助於人文研究者可以快速的掌握關鍵訊息進行分析，減少了以往研究者可能得用人工耗時費工的進行大量資料之閱讀後摘要、寫卡片等工作，直接進入重要文本分析，且透過數位人文技術的協助，可以指出歷史中的某些關鍵時刻以及重要觀念詞組，使研究者可以快速的在大量文獻資料中掌握關鍵議題而進行處理，又以一種可驗證之數據來加以佐證，可以說是大大的加速與便利了人文研究者，是為數位人文技術方法的優點所在。

誠如上述所言，中國「保護」的對象，是從「華工」到「華商」到「華人」與「華民」之過程。而從「華工」到「華人」的發展過程，其對象他者之共現，亦為可以觀察之發展圖像。如圖6所示。

而這個歷時性的觀念轉變過程，與中國相對最為重要的他者乃是美國，故而此處我們即要透過數位技術方法來觀察與「美國」這一關鍵詞共現的關鍵詞詞頻，來說明當時與美國他者共現之觀念為何。從圖6可以看見，與美國他者共現之自我指涉具有一個從「華工」→「華商」→「華民」之過程。如1897年時期：「伍延芳集中抨擊限制應受豁免一類的人員入境問題和在美華商受到虐待問題。根據條約權利，中國學生和商人可以自由進入美國。但是這種權利並未得到鮑得利及其支持者的尊重，他們有意識地、積極地策畫完全排斥華人。伍廷芳迅即揭露他們的目的，維護中國學生和商人的權利。」（顏清湟著，粟明鮮、賀躍夫譯，1990：318）

在這過程中，本文觀察出來，華工事件之所以作為促進華人意識的關鍵事件，乃是由於美國以排斥華工為由，一併的想要完全排除華人圈，而這樣的態度，也使一直以來受到保護之既得利益華商權益受損，美國這樣的排擠華商政策，比排擠華工對於中國的影響更大，因為排擠「華商」不僅損害到國家顏面，也會實際損害到



資料來源：劉昭麟博士統計，邱偉雲製圖

圖6 《清季外交史料》華工文獻中1875-1911年「美國」一詞共現詞頻年度比例分布圖

國內經濟穩定。（顏清湟著，粟明鮮、賀躍夫譯，1990：161）故而在此關鍵時刻之下，由於華商對於中國內部經濟的穩定具有一定的重要性，故而中國之外交策略不能僅保護華工，而因應美國對於整體中國人民的排擠，自然從保護華工擴大到保護華人，而擴大之原因，即是由於美國對於華商的排擠所導致。

從上述由數位人文技術所分析與分析出的歷時性觀念共現詞頻圖的協助，研究者可以很快的掌握觀念的個別與共現發展，並且可以透過可驗證的數據資料，去加以佐證相關的歷史論述，並且以一可驗證性資料去重新架構歷史觀念發展的系譜，有別於傳統利用少數文獻加以論述之片面，透過大量資料的分析，當更可強化人文研究者將來進行歷史研究中的科學性與可驗證性，並且可透過這樣的方法去重新檢證前人之歷史研究論述，加以重新確認、肯定或是重新商榷與討論有關的歷史論述，賦予歷史新的討論空間，也是數位人文學發展的一個方向所在。

七、結語

「華人」觀念起源於「華工」，但其本質卻是中華帝國被納入全球化民族國家組成的世界體系的結果。無論清廷對海外保護對象的變化，還是西方對中國抵制對象的擴大，都顯示出華人觀念取代華工的必然性。

本文的結論是用數位方法對關鍵詞意義結構統計分析得到的結果。它一方面表明數據庫和統計方法的運用比Reinhart Koselleck的概念史更進一步；另一方面亦顯示了數位方法對人文研究的意義。也就是說，這兩者之結合並不會改變人文研究的本質，但顯示了數位方法的巨大力量。

首先，整個分析離不開關鍵詞意義分析和文本解讀，其核心是純粹的人文研究。但是，《清季外交史料》共有四百萬字，以傳統人文學研究法，欲研究華工事件，必得先行通讀四百萬字文獻，進行挑選出有華工事件曾經出現之文獻奏摺，這個工作量雖人力可完成，但恐耗時費工，而且旁岔極多。華人觀念起源過程極可能被淹沒。而這時透過數位方法，可以快速的擷取出四百萬字中有華工相關討論之篇章，省去研究者通讀其他較不相關文獻之時間，直接可針對主題進行分析與研讀。

在研讀資料過程中，數位方法的運用使我們可以用關鍵詞叢來抓住蘊含在龐大的資料中的主要觀念之關聯。如究竟華工議題中，又以哪些觀念最為重要？而又是與哪些國家對話？又與哪些議題相關？這些疑問，傳統人文研究就必須在所收集來的華工史料中，再進行第二步的閱讀、摘要與寫筆記。在這個過程之後，才可能提出上述問題意識的初步觀察，且這個觀察仍然是「印象式的觀察」而非「可驗證式的觀察」。因此很清楚的從傳統人文研究方法切入，著實非常耗費研究者之心力在初步的資料整理上，若如今因為數位方法的協助，能夠減少資料檢索與整理筆記的

心力，而直接投入於分析與研究步驟，想必人文研究將會有事半功倍的效益產生。

正是基於這樣的理念，我們可以從人文角度出發，思考數位方法能提供什麼樣的數位協助，這也是數位人文研究的真正價值所在。

誌謝

本項研究中的「華工」詞彙的資料，取自於「中國近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研究開發，劉青峰主編）；現由臺灣政治大學「中國近現代思想與文學史專業數據庫」計畫辦公室提供檢索服務，謹致謝意。目前正由兩校共同完善、開發數據庫。英文摘要由張靖怡小姐協助翻譯，謹此一併致謝。

參考文獻

王曰芬、宋爽、熊銘輝，2007，〈基於共現分析的文本知識挖掘方法研究〉，《圖書情報工作（北京）》，51（4），頁66-70。

李軍蓮、李丹亞、黃利輝、孫海霞、冀玉靜、王鈐，2010，〈基於詞共現的中文醫學概念空間研究〉，《現代圖書情報技術》，26（11），頁59-63。

村田忠禧，2002，〈愛国主义と国際主义について『人民日報』社說を素材にした分析〉，《日中相互理解とメディアの役割》，日本僑報社、日中コミュニゥーション研究会編，頁74-76。

林頌堅，2010，〈以詞語共現網絡分析探勘資訊傳播學領域的研究主題與關係〉，《圖書資訊學研究》，4（2），頁123-148。

邱偉雲，2011，〈關鍵詞叢與文本意義挖掘的嘗試：以《清季外交史料》為例〉，項潔等主編，《數位人文在歷史學研究的應用》，頁159-188，臺北：臺大出版中心。

金觀濤、姚育松、劉昭麟，2011，〈社會行動的數位人文研究：以清末預備立憲為例〉，《第三屆數位典藏與數位人文國際研討會論文集》，臺北：國立臺灣大學。

金觀濤、劉青峰，2008，〈從「天下」、「萬國」到「世界」——兼談中國民族主義的起源〉，《觀念史研究：中國現代重要政治術語的形成》，香港：香港中文大學。

曹恬、周麗、張國煊，2007，〈一種基於詞共現的文本相似度計算〉，《電腦工程與科學》，29（3），頁52-53、73。

郭廷以編，1987，《近代中國史事日誌》，北京：中華書局。

郭鋒、李紹滋、周昌樂、林穎、李勝睿，2004，〈基於辭彙吸引與排斥模型的共現詞提取〉，《中文信息學報》，18（6），頁16-22。

陳良駒、傅振華、楊誌瑋，2010，〈詞彙共現分析在中國大陸信息作戰領域發展之實證研究〉，《中國大陸研究》，53（2），頁111-145。

陳鐘、彭波，2005，〈一種辭彙共現演算法及共現詞對檢索系統排序的影響〉，《清華大學學報（自然科學版）》，45（S1），頁1857-1860。

曾元顯、林瑜一，2011，〈內容探勘技術在教育評鑑研究發展趨勢分析之應用〉，《教育科學研究期刊》，56（1），頁129-166。

楊立英，2006，〈「共現」現象研究〉，《科學觀察》，1（5），頁10。

劉昭麟、金觀濤、劉青峰、邱偉雲、姚育松，2011，〈自然語言處理技術於中文史學文獻分析之初步應用〉，《第三屆數位典藏與數位人文國際研討會論文集》，臺北：國立臺灣大學。

蔡石山，1971，〈華工與中美外交〉，《美國研究》，3，頁198-199。

蔡明月，1999，〈齊夫（Zipf）定律〉，《教育資料與圖書館學》，37（2），頁165-183。

顏清湟著，粟明鮮、賀躍夫譯，1990，《出國華工與清朝官員：晚清時期中國對海外華人的保護　1851-1911年》，北京：中國友誼出版公司。

漢語方言語音資料庫自動擴增補完方法

林居正*、王昱鈞**、蔡宗翰***

摘 要

在歷史語言學研究之中，以語言田野調查收集語音資料為其研究方法重要之依據，故針對漢語的聲韻學研究領域，漢字於現代各種漢語方言中之發音資訊是探討漢語語音之歷史演變，及漢語方言間之親屬關係的重要研究材料。為利於資料的保存與處理，並結合資訊技術之應用，有部分漢語方言語音已進行數位化資料庫的建置。然而漢語分為七至八種方言群，因文化傳播與使用人口等因素，各漢語方言語音資料的完整度和數位化程度各不相同，許多漢語方言的語音資料庫仍未臻齊備，諸多漢字尚未有語音資訊。因此我們提出一個新的基於機器學習方法的模型，用以自動擴增補全漢語方言語音資料庫中缺少的漢字發音資訊，該模型利用既有的方言語音內容，以及從中古漢語韻書中找出各方言中語音對應的模式，藉以預測未知的漢字發音。我們以「漢語方音字彙」資料庫作為評估資料集，以漢字語音的音韻特徵如聲母、韻母、聲調等來評估擴增結果之準確率。藉由我們的方法，不僅能擴增漢語方言資料庫內容，其預測結果能進一步協助研究者發現具研究價值的語音現象，從而發展更多研究議題的可能性。

* 國立臺灣大學資訊工程學研究所碩士。
** 國立臺灣大學資訊工程學研究所博士生。
*** 元智大學資訊工程學系副教授。

An Automatic Augmentation Method for Chinese Dialect Pronunciation Databases

Chu-cheng Lin *, Yu-chun Wang **,
Richard Tzong-Han Tsai ***

Abstract

Field phonetic data collection is an important basis of historical linguistics' methodology. In Chinese historical phonology, phonetic qualities of Chinese characters in modern dialects are valuable materials. These are crucial for investigation into diachronic Chinese phonology, and phylogenetic relationship among Chinese dialects. In order to employ information technologies for preservation and further processing, work has been done to build dialectal pronunciation databases. However, Chinese, a large language family, can be divided into 7-8 dialect groups. As population and issues of culture propagation differ from a dialect to another, collection of their pronunciation also differ greatly in terms of completeness and degree of digitalization. Many dialects still lack satisfactory databases. And many character pronunciations are yet to be recorded. Therefore we propose a new machine learning based model to augment Chinese dialectal pronunciation databases automatically, filling out missing character pronunciations. This model uses existing dialectal pronunciations and medieval rime books to find patterns across dialects, in order to predict unknown character pronunciations. The Project DOC dataset is used for evaluation purpose. Phonological features, such as initial, rhyme, tone, are used to evaluate the accuracy of our results. Our model augments the pronunciation database. Moreover, the predictions made by our model may facilitate researchers discover interesting phenomena.

* Master, Department of Computer Science and Information Engineering, National Taiwan University.
** Ph.D. Student, Department of Computer Science and Information Engineering, National Taiwan University.
*** Associate Professor, Department of Computer Science and Engineering, Yuan Ze University.

一、簡介

字音資料庫是語音處理應用，如語音辨識與合成的重要資源。對於國家的官方語言而言，這樣的資料庫不虞匱乏。舉例而言，英語有「CMU語音資料庫」（Carnegie Mellon University, 1998），而標準漢語有「Unihan資料庫」（Jenkins & Cook, 2009）。

但對於其他語言而言，這樣的數位發音資源並不充足。在中國，這樣的情形特別明顯。一份2004年關於漢語方言的普查顯示，超過86%的中國人口可以非普通話的方言交談。而僅有53%能以普通話交談（Tong, 2006）。然而，非普通話方言的數位字音資料庫可謂非常缺乏。這樣的情況阻礙了這些資源匱乏方言的語音處理科技與應用的發展。因為收集這樣的資源需要大量人工，我們期望能開發工具，自動化諸方言之字音預測。在數位人文研究方面，若有了這樣的完整資料庫，將可以進行許多有趣的分析，例如以語言歧異程度分析文化傳播的速率等等，都仰賴同源語言的字表（Grey et al., 2009）。

目前大部分的方言字音資料庫皆由獨立研究者所建，其完整度往往差異頗大。若已有某方言其相關方言之完整字音資料庫，我們即可用既有之監督式學習方法預測某漢字於某方言之字音。

然如前所言，絕大部分漢語方言的字音資料庫都仍未完備。因此，我們提出一個新的生成模型，利用現有之方言發音資料與中古韻書，發掘跨方言的規律。與先前文獻不同之處在於我們的模型並沒有語言樹狀演化的假設，而僅要求相關方言的漢字字音有跨語言的規律。我們提出的模型可以以其他方言字音表（即便不完整）與中古韻書為基礎，填補某特定方言的方言字音。填補完成後，即可建構一個以分類器為基礎之字音預測系統。

二、漢語語音背景

（一）漢語方言間相互溝通度

漢語雖然依通見分為各種方言，然而各方言之間多無法相互溝通。依據Tang與van Heuven（2009）的研究，漢語南方之方言相互之間句子溝通度平均小於30%，與之相較，葡萄牙語與西班牙語的相互溝通度卻大約有60%（Jensen, 1989），顯之漢語方言間彼此幾乎無法相互溝通。

雖然漢語方言之間相互溝通度很低，但由於各方言受限於源於古代漢語的語音演變規則的制約，漢字之發音屬性於不同方言之間仍大多保持其一致性。如漢字「肝」與「寒」於閩南語及華語之間發音相差甚大，然而該二漢字之韻母於閩南語及華語中仍保持其對應。

（二）韻書

由於漢字的文字系統本身並非完整的拼音文字，對於實際的語音表示上具有其侷限性。於魏晉南北朝時期逐步發展出「反切」的標音方式，用以標註漢字於當時的實際發音。反切之原理，其以二個漢字來標記單一漢字的發音，第一個字稱作反切上字，以反切上字本身之聲母用以標記該漢字的聲母；第二個字稱作反切下字，以反切下字本身之韻母及聲調來標記該漢字的韻母及聲調。以現代華語為例，漢字「東」之反切可為「德工」，反切上字「德」表示其聲母為d，而反切下字「工」表示其韻母為ong且聲調為陰平聲。藉由反切，表音不便的漢字便有了得以較準確標音的方法。

反切的標音方式出現之後，以反切記錄漢字發音的語音字典便隨之而生，即所謂之韻書。現存可考的最早之韻書為隋代開皇年間由陸法言等人所編撰之《切韻》，其所記錄的發音為中古時期之漢語。《切韻》全書早已亡佚，只有部分內容見於唐代時的敦煌殘卷，然而《切韻》一書所揭櫫之中古漢語之語音體系，在而後北宋時期的韻書《廣韻》之中得以繼續保存下來。《廣韻》全名《大宋重修廣韻》，為北宋真宗年間（1008年）由陳彭年等人奉詔編修的韻書，為中國歷史上第一部由官方所編修之韻書，其參考了更早的韻書如《切韻》等書進行編撰，其按聲調分為五卷，按同樣的韻母將漢字進行分類，同韻母的漢字歸屬於同一個韻部，共分為206個韻部。《廣韻》為中古漢語時期最重要的韻書，而後的韻書體例亦多按《廣韻》之架構編排。自宋代以後，歷代私人與官方皆陸續編撰韻書以記錄當時的漢語發音。如元代熊忠編彙的《古今韻會舉要》、周德清編撰的《中原音韻》、明代的《洪武正韻》與清代的《佩文詩韻》等，記錄了各個時期的漢語之實際語音。

雖然反切的出現解決了漢字標音的問題，但是標註同一發音之反切的上下字的可能性相當多，如「東」字於現代華語之中之反切可以為「德工」或「大中」等等。故在南宋時期的《韻鏡》一書開始，另一類被稱作「韻圖」的韻書開始出現。韻圖將漢字依照聲母和韻母排列成矩陣的表格圖形，透過行與列的組合便能夠查詢漢字的發音。韻圖之中對於每一個漢字標註其六種屬性，分別為：聲母、韻、攝、聲調、呼……等。聲母多以36個漢字所組成之36字母進行表記，以代表當時漢語的36種聲母，韻則是以《廣韻》類的韻書所分的206韻為基礎。由於韻母分作二百多種，在分析上過多過細，因此韻圖中提出了「攝」的概念，攝是比韻更寬的分類，將主要元音與末尾子音相同的韻歸併在一起，不論其聲調和介音之差別，一般韻圖多將《廣韻》之韻母整併歸納為十六個攝，如「寒」韻（an）和「先」韻（ian）都歸納於「山」攝（an）之中。聲調部分則是傳統的漢語四聲分為四類：平、上、去、入。「呼」的概念主要分為合口與開口二類，基本上是介音[-u-]的有無之差別，有介音[-u-]者為合口，無者則為開口。韻圖在每一個聲母和韻母組合之中，又

細分為四類，分別為一至四「等」，對於韻圖中的「等」之解釋，歷代聲韻學家看法分歧，一二等與三四等之間的差異可能為介音[-i-]之有無，而一二等之間與三四等之間內部的差異則為母音發音部分的前後與開口度之差異。

韻圖對於每個漢字的六種語音屬性，在現代的諸多漢語方言之中多仍保持其對應關係。如「含」（han）與「站」（zhan）二字，其皆歸於「咸」攝，為押韻字，而該二字於現代華語、閩南語、廣東話之中皆為押韻字，儘管該二字的實際發音於華語、閩南語、廣東話中皆不相同，但其於韻圖中所記錄之屬性，於該三種漢語方言中仍予保持。是故韻圖的語音屬性，是決定漢字於各漢語方言中實際發音的寶貴參考資料。

三、相關研究文獻

已有許多當代字典使用音標標示特定方言的字音，如《粵音韻彙》。1962年第一部綜合的跨方言字彙，《漢語方音字彙》（以下簡稱《字彙》）出版了。原始的《字彙》包含十七個現代漢語方言，大約兩千五百個字音的IPA標註。此外還包含了一本韻書《韻鏡》的分類。在出版不久之後，《字彙》就在Project DOC下數位化了（Streeter, 1972）。《字彙》對於歷史音韻學研究是相當具有價值的。然而，還是有許多方言未收進《字彙》。另一個問題是《字彙》僅有大約兩千五百個音讀；這遠小於漢字總數（超過五萬）。這兩個問題使得《字彙》不太適合當成方言字典。我們提出可以用《字彙》裡的方言字音增補先前未見的方言字音。

以已知資訊增補未知資料並不是個新想法，在Nigam、McCallum、Thrun與Mitchell（2000）和 Lu、Zheng、Atulya與Zhai（2006）已經可以見到。資料增補通常是透過引入隱藏變數描述training data達成（van Dyk & Meng, 2001）。在我們的問題裡，我們必須要描述方言發音資料。Bouchard-Côté、Liang、Griffiths與 Klein（2007）提出一個羅曼語的發音模型，能產生現代語言與重建的古代方言的發音。他們以系統發生樹描述古典拉丁語、通俗拉丁語、西班牙語以及義大利語的演化關係。在此樹裡，古典拉丁語為樹根，通俗拉丁語為其子，而西班牙語以及義大利語是通俗拉丁語的後嗣。在他們的模型裡，樹根語言的發音必須為已知。

但是對於漢語方言而言，這種樹狀模型是否能適用是有爭議的。Ben Hamed與Wang（2006）指出以網路描述漢語方言的發展也許較為合適。即便以Bouchard-Côté等提出的模型描述漢語方言而將中古漢語，大多數漢語方言的祖語，當成根，我們仍會遇到下面的問題。由於古典拉丁語的音韻已大致明瞭，Allen（1978）實際的發音可以從拼字簡單的推測出來。不像古典拉丁語，中古漢語的音韻以及字音並不是完全明瞭。舉例而言，我們對實際的調值仍一無所知。現在的重建相當倚賴中古韻

書，而中古韻書已知是至少兩個中古漢語方言的綜合體。丁邦新（1995）要推導出正確的系統發生樹，我們勢必得分別這些中古方言（根據丁邦新的說法至少有兩個）並指派給正確的後嗣。然而目前的研究顯示有些吳方言至少有兩個substrata，其一來自北方中古漢語，其一來自南方。梅祖麟（2001）指出這直接違反了樹狀假設。對一個語言 l 來說，若沒有 l 祖語的實際發音，我們無法以Bouchard-Côté等的模型預測 l 的字音。

有些研究者試圖利用其他語言的資源幫助資源匱乏的語言。Snyder、Naseem、Eisenstein、Barzilay（2009）指出加入多語言未標註的文字可以增進非監督式詞性標記的效能。Stüker、Metze、Schultz、Waibel（2003）以多語聲學資料，在語言間共用音節特徵，增進了一個新見語言的辨識效能。這些研究假設訓練時的語言資料具有某些規律，可以用於新見語言。但我們的研究僅假設漢語方言彼此之間與中古漢語都具有一致的音韻對應。

四、模型

（一）問題定義

我們的目標是要填補漢語方言的語音資料庫。就每筆紀錄而言，我們的字音資料庫表列出這二十一個方言內既有的方言字音。也就是說，有些紀錄可能是不完整的。這些字音以音素組合表示（我們將稱之音韻特徵）。我們的填補模型不只利用既有的發音，也利用了韻書資料。

令 c 為一筆紀錄的漢字；令其韻書類別特徵為 Y_c。舉例而言，漢字「含」的韻書類別特徵為〔匣, 覃, 咸, 平, 開, 一〕。多類別向量 Y_c 接著可以將其每個分量「攤平」，串接成一個二值向量 F_c。例如一個有三個可能值的可以「攤平」成三維二值向量。因韻書類別特徵 Y_c 有六個分量，F_c 為一二值向量，維度為 $\sum_{i=1}^{6}\dim(Y_{c,i})$。令 $L_1 \ldots L_N$ 為 N 個現代方言，每個方言的音韻特徵數目都固定。舉漢字「含」為例，其廈門方言之音韻特徵為〔'12', '43', /h/, ϕ, /a/, ϕ, false, /m/〕（見表1）。

則問題可如此闡述：假設所有方言有共 K 個音韻特徵 $t_1 \ldots t_k$，另有漢字的二值韻書特徵 $F_{t1} \ldots F_{tG}$，及一維度 $K \times G$ 之部分填滿之音韻特徵表，我們的目標等同於把此表填滿，如圖1。

（二）模型定義的一些考量

如第二節所言，絕大多數漢語方言音韻皆與韻書內描述之類別特徵及其他方言高度相關。舉例而言，韻書特徵「深攝」、粵語的/am/、閩南語的/im/皆有高度相

表1 DOC dataset 音韻特徵編碼

音韻特徵	例值
調類	“上”
調值	55
聲母	/b/
介音	/i/
元音	/o/
雙元音後半（若有）	/e/
是否有鼻化現象	是
韻尾	/t/

…	…	…	…	…	?	…
…	…	…	…	…	…	…
〔見,屑,山,入,開,四〕	‘kiat’	‘tɕie’	…	‘kit’	…	…
〔匣,覃,咸,平,開,一〕	‘ham’	‘xan’	‘xæ˜’	‘ham’	‘ɦø’	‘xaŋ’
…	…	…	…	…	…	…

圖1 輸入格式。共有G個字音，其韻書特徵皆已知。若干音韻特徵可能缺漏。我們的目標是將缺漏的特徵值補滿，輸出為一完整表格。

關。縱使單憑韻書即能得到對這些方言音韻的許多洞見，方言字音並不總是依循韻書分類，但方言間仍有規律，這些現象都必須列入考慮。此外，在資料收集方面，韻書資料較沒有缺漏的問題；而在資料集中方面，卻不見得每個方言都有收齊所有的漢字音。

我們提出了一個利用所謂的superlingual rhymes（SLRs）的隱藏變數，同時考慮方言間音韻規律以及韻書特徵的生成模型。我們的模型將每筆紀錄分成兩個部分：一部分是韻書特徵，另一部分是方言音韻特徵。我們想將方言音韻特徵這部分補齊。我們知道韻書特徵與漢語方言音韻特徵高度相關。因此，我們利用韻書特徵估計方言音韻特徵。另一方面，我們也利用其他方言的音韻特徵。我們的想法大致上

是引入superlingual rhyme作為韻書特徵與方言音韻特徵的中介，如此每個字的音韻特徵可視為所有superlingual rhyme的混合。因每個漢字對superlingual rhyme的機率及superlingual rhyme對音韻特徵的機率都可求得，我們即可填補遺缺的音韻特徵。

事實上，我們亦可以主題模型（topic model）的角度來看。在我們的模型之中，每個漢字可以視為主題模型中的文件（document），而SLR可以視為主題（topic），每個音韻特徵可以視為主題模型中的單詞（word）。在主題模型之中，每個文件皆為許多主題的混合（mixture），而每個主題又為許多單詞的混合。就我們的應用而言，我們期望得到單詞的後驗分布，以預測新方言的漢字音。

（三）模型描述

圖2是我們提出模型的plate diagram。令觀察值o為一二元組(p, l)，p為一觀察到的音韻特徵，而l為p出現的方言，每個漢字c的觀察值$o_{c,i} \in \{o_{c,1}...o_{c,n(c)}\}$都有一個隱藏的superlingual rhyme $s_{c,i}$；而漢字c是superlingual rhyme的混合。為了方便描述，我們在敘述中假設每個方言都僅有一個音韻特徵t。在現實裡每個觀察值$o_{c,i}$對於每個方言$l_{c,i}$都有多個音韻特徵，但我們所做的簡化敘述並不影響推導。實驗時我們會同時採用多個音韻特徵。

以下描述我們的模型。在我們的模型裡，二元韻書特徵的長度為|F|，有G個漢字、H個superlingual rhyme $SLR_1...SLR_H$，而每個漢字音c有n(c)個讀音，每個$s \in \{SLR_1..SLR_H\}$都有N個音韻特徵的多項分布$\theta_{s,t,L_1}...\theta_{s,t,L_N}$；以及方言$L_1...L_N$的多項分布$\phi_s$。$\theta$和$\phi$都有均勻的先驗分布Dirichlet($\beta$)與Dirichlet($\gamma$)。在我們的實驗裡，$\beta$與$\gamma$的每個分量都設為$10^{-3}$，讓先驗機率較為疏落。

再重複一次漢字c的二值韻書特徵為F_c。我們令漢字c的superlingual rhyme $SLR_1...SLR_H$的分布為一參數為ψ_c的多項分布，並令ψ_c的先驗分布為F_c的函數，權重向量為$\lambda_{SLR_1}...\lambda_{SLR_H}$，維度皆與$F_c$同，而$\psi_c$的先驗分布即為：

$$\text{Dirichlet}(\alpha_c) = \text{Dirichlet}(e^{F_c^T \cdot \lambda_{SLR_1}},...,e^{F_c^T \cdot \lambda_{SLR_H}})。即 \alpha_c[i] = e^{F_c^T \cdot \lambda_{SLR_i}}。$$

換言之，superlingual rhyme s的先驗機率與$e^{F_c^T \cdot \lambda_s}$成比例。

以下我們將描述生成過程。圖2有此模型之plate diagram。

1. 對於每個superlingual rhyme s，

甲、$\lambda_s \sim N(\mu,\sigma^2 I)$，其中$\mu$為零向量。

乙、$\phi_s \sim$ Dirichlet(γ)

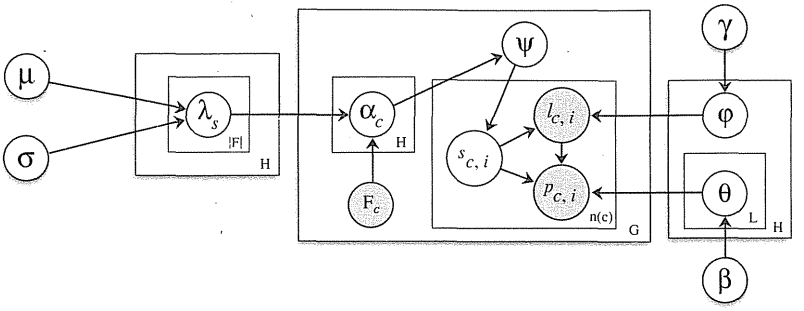


圖2 我們提出生成模型之plate diagram。陰影節點為觀察值

2.對每個方言l，$\theta_{s,t,l} \sim \text{Dirichlet}(\beta)$

甲、對於每個漢字c與其二值音韻特徵F_c，

i. 對於每個superlingual rhyme s，$\alpha_{c,s} = e^{F_c^T \cdot \lambda_s}$

ii. $\psi_c = \text{Dirichlet}(\alpha_c)$

iii. 對每個$o_{c,i}$，

① $s_{c,i} \sim \text{Multinomial}(\psi_c)$

② $l_{c,i} \sim \text{Multinomial}(\phi_{s_{c,i}})$

③ $p_{c,t,l_{c,i}} \sim \text{Multinomial}(\theta_{s_{c,i},t,l_{c,i}})$

我們描述了一個完整的生成模型。貝氏統計的相關文獻對於此等模型的參數估計已多有描述。

（四）推論

因為音韻特徵為類別資料，很自然的可以利用多項分布描述。就跟所有的貝氏模型一樣，我們給這些多項分布加上了先驗分布。如許多文獻所指，如Blei、Ng與Jordan（2003）和Goldwater與Griffiths（2007），我們選用Dirichlet分布，因為如此一來後驗分布就有解析解，並且也是Dirichlet分布。漢字對superlingual rhyme的機率是遵循Dirichlet分布；而這Dirichlet分布的參數是由這漢字的韻書特徵的對數線性函數決定。這方法又叫Logit模型。因為多項分布與Dirichlet分布具有共軛關係，一個漢字的superlingual rhyme後驗分布很容易就可以得到。Heinrich（2008）、Resnik與Hardisty（2010）如此混合生成模型與Logit模型，非常類似Berg-Kirkpatrick、

Bouchard-Côté、DeNero與Klein（2010）提倡的範式。同樣的，利用多項——Dirichlet共軛，我們可以得到superlingual rhyme對於音韻特徵的後驗分布。我們如此以馬克夫鏈蒙地卡羅法得到變數$p_{c,t,l}$的後驗分布，並用以填補缺失的音韻特徵。而在實作方面，為了加速訓練過程，λ_s在生成模型部分視為定值。儘管我們給它常態先驗分布，但我們在馬克夫鏈蒙地卡羅法過程中並沒有更新它的值，而是以L-BFGS法尋找能讓生成模型似然函數最大化的值。實作時我們間歇的以Gibbs Sampling法取樣每個音韻特徵的superlingual rhyme，而後固定某一取樣，並最大化生成模型似然函數。這有點類似於著名的Expectation Maximization法。

五、資料與評估指標

（一）資料

我們的實驗皆利用第三節所述之DOC dataset 進行。在此dataset中，每筆紀錄皆對應至一個漢字音。舉例而言，多音字「正」有兩個普通話 φ 發音（zheng1與zheng4）。如此一字多音的現象，在《廣韻》裡就可以見到。每筆紀錄都列出在21個方言裡的讀音。原本的格式是以IPA音標轉寫。Cheng（1997）以8個音韻特徵表示IPA轉寫，如表1所示。因為共有21個方言、8個音韻特徵，每筆紀錄最多可能包含168個音韻特徵，但並不是每筆紀錄都包含這完整的168個音韻特徵。把每個多音字的讀音都單獨記錄後，我們得到5,403筆紀錄。

（二）評估指標

我們依兩個指標評估：分別音韻特徵準確率（Individual Phonological Feature Accuracy, IPFA）與整體音韻特徵準確率（Overall Phonological Feature Accuracy, OPFA）。IPFA是指預測正確的音韻特徵佔測試集音韻特徵總數的比例。而OPFA是指整組字音的音韻特徵皆預測正確的字音佔整個測試集字音總數的比例。

（三）評估方法

評估方言d的預測準確率時，我們將所有方言d的音韻特徵都當成ground truth label。有些非d的方言可能會遺漏一些音韻特徵。隨著設定不同，我們利用提出的模型或是baseline分類器填補這些遺漏的音韻特徵。

如四（三）節所述，我們焦點之一是要填補遺漏的音韻特徵。在填補的實驗中，我們隨機將除d以外所有方言內的音韻特徵隨機移除。詳細程序如下：首先我們建立有10%或20%欄位缺漏的dataset子集。接著我們利用先前所述的方法填補這些欄

位。方言d的音韻特徵並沒有被用來預測其他方言的音韻特徵。填補完成後，所有的紀錄都沒有遺漏的音韻特徵。在評估時，我們以在資料集中有記載的、經由田野調查得來的方言音韻特徵評估音韻特徵準確率。

我們進行了t檢驗以確定統計顯著性差異，重複了以下的步驟30次。我們隨機將dataset裡的紀錄以2:1分入training和test兩組。因為每筆紀錄都有多個label，我們利用multiclass SVM分別學習這些標籤。提供給SVM分類器的特徵有二值韻書特徵向量(F_c)以及除了方言d外所有方言的音韻特徵。對應的label是方言d的音韻特徵。因此分類器的輸出值也是方言d音韻特徵的預測值。

（四）t檢定

我們利用雙樣本t檢定檢驗是否某個配置比另一個好，並且具有顯著性差異。

我們選用雙樣本t檢定，因為我們假設樣本彼此獨立。在樣本數大且樣本標準差已知時，便可利用下列的雙樣本t統計：

$$t = \frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

在此μ是平均準確率、σ^2是準確率變異數，而n是樣本數。（在我們的實驗裡，$n = 30$）。自由度為29，如果得到的t值小於或等於1.67，就接受虛無假設，否則拒絕。

六、實驗

我們設計了三個關於潮州方言字音的實驗。潮州方言屬於閩南語，主要在廣東東部使用。我們想評估以下因素造成的影響：

（一）方言資料對標準分類器的影響

傳統聲韻學家對漢語方言的處理態度是去尋找韻書中韻部等分類與現代方言字音的對應；這通常需要大量人工。然而兩者之間的清楚對應並不一定存在。舉吳方言為例，夬韻與佳韻無法清楚辨別，有時以「-ua」出現，有時以「-uo」出現。我們認為加入方言資料（即其他方言的音韻特徵）可能會對辨別某些方言的發音有幫助。

表2 加入方言資料對預測準確率的影響

設定	OPFA
R	41.5%
R+F	62.8%

在我們使用的資料集中有正確的、經田野調查收集而來的方言資料。我們用以訓練了SVM分類器並評估預測潮州方言的漢字發音的準確率。如前所述，我們實驗了兩種設定：

僅有韻書（R）：這種設定僅包含韻書特徵，就是「聲母」、「韻」、「攝」、「聲調」、「呼」以及「等」。

韻書與全部方言資料（R+F）：除了韻書資料，我們利用了所有的方言資料。如果有缺漏的音韻特徵，我們就補上隨機猜的值。

結果列於表2。很明顯的，加上方言資料，我們大幅提升了預測效能。

（二）鄰近方言的影響

Snyder、Naseem、Eisenstein、Barzilay（2008）表示POS tagging 的效能可以透過納入更多語言提升，尤其是關係密切的語言。我們進行了實驗，想知道是不是利用韻書特徵（R）加上關係密切的方言（+C）比加上關係疏遠的方言（+D）更有效。

我們比較了西安方言與潮州方言的OPFA。這兩個方言分別屬於官話與閩語系統。我們納入的官話方言有濟南、太原以及北京；而閩語我們使用了廈門、福州以及建甌。每個方言我們都有兩個設定，利用同方言群的方言以及利用不同方言群的方言兩種。為了使比較有意義，我們每次實驗都會隨機拿掉音韻特徵，使遺漏的音韻特徵數目都相等。接著我們利用我們提出的模型，填補這些遺漏的音韻特徵。我們重複30次，得到的平均OPFA列於表3。結果顯示無論是西安方言或是潮州方言。R+C的準確率都大幅超越R+D，並具有統計上的顯著性。

（三）資料填補的效果

如第一節所述，許多漢語方言的資料都十分稀少。我們的模型是設計來填補遺缺的音韻資訊。倘若我們的填補模型有效，將可以讓我們利用多個資源匱乏方言填補另一個方言的語音資料庫。至於資料填補，我們用五（三）節的步驟將潮州方言缺漏的音韻特徵填補完整。在我們的實驗裡，以（-A）標示我們的模型。

表3 不同方言群對準確率的影響

設定	方言	30次平均OPFA	t
關係密切方言	西安	80.5%	40.8537
關係疏遠方言	西安	67.9%	
關係密切方言	潮州	51.2%	25.8430
關係疏遠方言	潮州	42.8%	

我們另外用了三個方法填補遺漏的資訊當做baseline來比較：

Logistic Regression（-L）：使用韻書特徵F_c訓練一個Logistic Regression模型以預測缺漏的音韻特徵值。Logistic Regression為社會科學中處理離散變項的一常見方法。我們以各韻書特徵作為特徵值，並以隨機梯度下降法訓練參數。

Naïve Bayes（-N）：與Logistic Regression模型相似，訓練一個生成式模型以預測缺漏的音韻特徵值。在使用Naïve Bayes分類器時我們並不用訓練參數，而為了平滑預測結果，我們把每項特徵值的計數都加上1。

Random（-R）：隨機猜測遺缺的音韻特徵值。我們隨機補上一個可能的猜測值。

在我們的實驗裡，我們測試了兩種不同比例的遺缺值，10%以及20%。接下來所有的SVM classifier都使用RBF kernel，參數$c = 512$，$\gamma = 2^{-7}$。Superlingual rhyme的個數在我們的實驗裡設為200。

實驗結果與相對應的t值都列於表4中。使用我們的填補模型可以相當一致的增進OPFA。有趣的是，對-10%以及-20%這兩個dataset，使用關係密切方言資料都比關係疏遠方言來的有效，這也與六（二）節的結果一致。

七、分析與討論

我們對於training data方言的選擇對於個別音韻特徵預測的影響有興趣。表5列出我們的模型對於baseline隨機填補法的IPFA進步幅度，以百分比表示。韻書特徵加上關係密切方言（R+C run）在填補後除鼻音化以外所有的音韻特徵中都有進步，原因仍不明朗。

而韻書加上關係疏遠方言，聲調、聲母，韻尾的準確率在填補後都變差了。我們認為考慮我們模型的假設可以解釋這個現象。我們的模型假設方言之間有音韻特

表4 資料填補對於關係密切與關係疏遠方言資料的影響

設定	移除的百分比	30次平均OPFA	t
R+C-A	10%	54.6%	
R+C-L	10%	51.4% (+3.2%)	9.5379
R+C-N	10%	51.5% (+3.1%)	9.1315
R+C-R	10%	51.2% (+3.4%)	8.9327
R+C-A	20%	54.0%	
R+C-L	20%	48.1% (+5.9%)	15.6103
R+C-N	20%	48.3% (+5.7%)	14.0192
R+C-R	20%	46.4% (+7.6%)	23.1526
R+D-A	10%	44.5%	
R+D-L	10%	43.0% (+1.5%)	3.4887
R+D-N	10%	42.6% (+1.9%)	5.0235
R+D-R	10%	42.8% (+1.7%)	4.0456
R+D-A	20%	42.3%	
R+D-L	20%	40.6% (+1.7%)	3.1019
R+D-N	20%	41.1% (+1.2%)	1.8283
R+D-R	20%	39.5% (+2.8%)	9.8251
R	10%	41.5%	
R+F	10%	62.8%	

表5 IPFA表

資料	R+C	R+D
調類	1.33%	-1.68%
調值	1.33%	-1.68%
聲母	3.99%	-0.09%
介音	2.57%	1.51%
元音	4.7%	0.53%
雙元音後半（若有）	2.22%	0.27%
是否鼻化	-0.27%	0.00%
韻尾	1.77%	-0.62%

徵的對應關係。也就是說，跨方言對應的音韻特徵會被歸到同一個 superlingual rhyme 下。因此若方言間缺乏這樣的對應，填補的音韻特徵可能不準確。而的確，聲調、聲母、韻尾並沒有很好的跨方言區對應。吳瑞文（2005）近期的研究提示閩語的聲調可能與原始吳閩方言的一個創新有關，而官話並沒有共享這個創新。至於聲母，官話和閩語對輕重唇音的區別也截然不同。洪惟仁（1999）韻尾也沒有很明確的對應：閩語保留了中古漢語大部分的塞音韻尾，而在官話方言許多脫落了。因此利用官話方言預測閩語的韻尾相當困難，而反之亦然。

對於關係密切與關係疏遠的方言，IPFA指標似乎都反映了方言間對應的程度。透過比較每個IPFA進步幅度，可能可以用以探討方言間的親疏關係，我們認為這可能是有趣的研究方向。

八、結論

我們提出了一個新的生成模型，同時利用現有的方言字音資料以及中古韻書發掘跨方言之音韻規律。在我們的模型中稱這樣的音韻規律為superlingual rhyme。我們提出的模型能夠利用現有的方言字音表（即使不完全）及韻書來預測另一個方言的字音。我們評估的指標包括音韻特徵的準確率。就每筆紀錄，我們評估整體音韻準確率（OPFA）。我們的第一項實驗顯示在只有利用韻書資料的baseline SVM分類器加入方言發音資料即可大幅度提升OPFA。第二項實驗裡我們比較了使用關係密切方言與關係疏遠方言音韻特徵在SVM分類器上準確率的影響。在第三個實驗我們指出使用我們提出的資料增補模型來填入遺缺的資料可以增進SVM模型的OPFA達7.6%。我們也發現進步幅度在使用關係密切方言資料時較大。

參考文獻

丁邦新，1995，〈重建漢語中古音系的一些想法〉，《中國語文》，249（6），頁414-419。

吳瑞文，2005，《吳閩方言音韻比較研究》，國立政治大學中國文學系博士論文。

洪惟仁，1999，〈漢語送氣音與鼻音衍化的動機與類型〉，發表於「第六屆國際暨第十七屆中華民國聲韻學學術研討會」，臺北：國立臺灣大學。

梅祖麟，2001，〈現代吳語和「支脂魚虞，共為不韻」〉，《中國語文》，280（1），頁1-15。

Allen, S. (1978). Vox *Latina: A Guide to the Pronunciation of Classical Latin*. New York: Cambridge University Press.

Ben Hamed, M. & Wang, F. (2006). Stuck in the Forest: Trees, Networks and Chinese Dialects. *Diachronica, 23*(1), 29-60.

Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., & Klein, D. (2010). *Painless Unsupervised Learning with Features*. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 582-590.

Blei, D., Ng, A., & Jordan M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Bouchard-Côté, A., Liang, P., Griffiths, T. L. & Klein, D. (2007). *A Probabilistic Approach to Diachronic Phonology*. Paper Presented at the Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic.

Carnegie Mellon University. (1998). CMU Pronouncing Dictionary. Retrieved from http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Cheng, C.-C. (1997). Measuring Relationship among Dialects: DOC and Related Resources. *Computational Linguistics, 2*(1), 41-72.

Gray, R. D., Drummond, A. J. & Greenhill, S. J. (2009). Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science, 323*(5913), 479-483.

Goldwater, S. & Griffiths, T. L. (2007). *A Fully Bayesian Approach to Unsupervised Part-of-speech Tagging*. 45th Annual Meeting of the Association of Computational Linguistics, 744-751.

Heinrich, G. (2008). Parameter Estimation for Text Analysis. Germany: University of Leipzig.

Jenkins, J. & Cook, R. (2009). Unicode Han Database. The Unicode Consortium.

Jensen, J. (1989). On the Mutual Intelligibility of Spanish and Portuguese. *Hispania*, *72*(4), 848-852.

Lu, X., Zheng, B., Atulya, V. & XiangZhai, C. (2006). Enhancing Text Categorization with Semantic-enriched Representation and Training Data Augmentation. *Journal of the American Medical Informatics Association*, *13*(5), 526-535.

Nigam, K., McCallum, A., Thrun, S. & Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, *39*(2-3), 103-134.

Resnik, P. & Hardisty, E. (2010). Gibbs Sampling for the Uninitiated. University of Maryland.

Snyder, B., Naseem, T., Eisenstein, J. & Barzilay, R. (2008). *Unsupervised Multilingual Learning for POS Tagging*. EMNLP'08 Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1041-1050.

Snyder, B., Naseem, T., Eisenstein, J. & Barzilay, R. (2009). *Adding More Languages Improves Unsupervised Multilingual Part-of-speech Tagging: A Bayesian Non-parametric Approach*. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 83-91.

Stüker, S., Metze, F., Schultz, T. & Waibel, A. (2003). *Integrating Multilingual Articulatory Features into Speech Recognition*. Paper presented at the Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech -2003), Genf, Schweiz.

Streeter, M. (1972). DOC, 1971: A Chinese Dialect Dictionary on Computer. *Computers and the Humanities*, *6*(5), 259-270.

Tang, C. & van Heuven, V. (2009). Mutual Intelligibility of Chinese Dialects Experimentally tested. *Lingua*, *119*(5), 709-732.

Tong, L.-Q. (2006). Survey on the Usage of Chinese Languages and Script. Retrieved October 7, 2011, from http://www.china-language.gov.cn/LSF/LSFrame.aspx

van Dyk, D. A. & Meng, X. L. (2001). The Art of Data Augmentation. *Journal of Computational and Graphical Statistics*, *10*(1), 1-50.

Part III
地理資訊
Geographical Information

Digitalization and Utilization of the "Large-scale Maps of Kyoto City (*Kyoto-shi meisai-zu*)"

Naomi Akaishi *, Toshikazu Seto **, Yukihiro Fukushimav***, Keiji Yano ****

Abstract

This paper discusses the "Large-scale Maps of Kyoto City (*Kyoto-shi meisai-zu*)," which consists of 291 maps, produced between 1927 and 1951 as well as the related digitalization project. Made for fire insurance purposes, these maps contain various kinds of building-related information necessary to prevent fire. The information comprises not only lot numbers of buildings, but the number of stories and usage of each building as well. The color-coded information, which was added to the maps by hand, can significantly contribute to the development of historical research of Kyoto City. For this reason, the Historical GIS Research Group of the Global COE Program "Digital Humanities Center for Japanese Arts and Culture" of Ritsumeikan University is conducting a digitalization project for these maps. The digitalization will help to reconstruct the urban landscape of Kyoto between the 1920s and the 1950s, and to analyze the descriptions on the maps. The map database is constructed using ArcGIS™. The following steps are undertaken to digitalize the maps. First, the maps are traced and placed into the computer. Subsequently, the maps are overlaid on, and adjusted to match Kyoto City's Digital Maps for city planning in geographic information system (GIS). The procedure, which is called geometric correction, facilitates the acquisition of positional information. The procedure allows the comparison between the former and present landscapes using GIS. Second, blank spaces are clipped on the rectified maps using GIS. A total of 286 clipped maps are joined to achieve an integrated display. The procedure reveals the features of the city after its merger with neighboring towns in 1918. Third, polygon data about each building are derived through digitizing the vertexes on GIS. In addition, a database of

* Postdoctoral Fellow, Department of Geography, Ritsumeikan University, Japan.
** Ph.D. student, Department of Geography, Ritsumeikan University, Japan.
*** Curator, Kyoto Prefectural Library and Archives, Japan.
**** Professor, Department of Geography, Ritsumeikan University, Japan.

materials, which includes the number of stories and usage, is created for each group of polygon data. Furthermore, rectified image data is converted to the KML format, which makes it possible to display maps in Google Earth™. As a result, "Large-scale Maps of Kyoto City" can be easily compared with the present city. The comparison reveals dramatic changes in the city's landscape. For example, Kyo-machiya, Kyoto's traditional wooden houses, were changed to Western-style buildings, and major streets such as Oike, Horikawa, and Gojo have become wider. The "Large-scale Maps of Kyoto City" are digitalized, thus helping reconstruct the landscape of modern Kyoto City and reveal historical changes.

京都大比例尺地圖（京都市明細圖 *Kyoto-shi meisai-zu*）數位化

赤石直美*、瀨戶壽一**、福島幸弘***、矢野桂司****

摘 要

本文探討京都大比例尺地圖（京都市明細圖 *Kyoto-shi meisai-zu*）及其數位化專案的進行過程。這些地圖涵蓋了1927年至1951年間繪製的291份地圖，比例尺為1:1200，大小約為38×54公分。這些地圖的繪製目的是消防安全之用，因此包含了建物相關的各種防火必要資訊，例如各建物之地號、樓層數與使用目的。用色彩編碼對建物相關各類資料加以區分，並陸續以人工方式新增到地圖中。日本立命館大學「日本藝術與文化數位人文中心」乃日本文部科學省設置之全球卓越計畫（COE Program）研究中心，基於京都市明細圖所載之豐富資訊對研究京都市的歷史發展有極大貢獻，其所屬之歷史地理資訊系統（HGIS）研究群，特別針對此批地圖執行數位化計畫，期能藉此重建1920至1950年代京都的都市景觀，並對地圖所呈現之資料進行分析。此計畫運用ArcGIS™軟體建置地圖資料庫，並透過以下方式進行數位化。首先，這些大比例尺的歷史地圖可以利用GIS地理資訊系統軟體，將其轉繪成數位檔案，然後與京都的都市計畫數位地圖相疊合，並加以調整，讓兩者相互對應。這些地圖經幾何校正，取得定位資訊，便能用以比對古今地景。接著，將286張已校正的地圖，裁切掉空白區域，組合成為一整份地圖，藉此我們可以理解此份地圖所呈現1918年時京都市及其鄰近市鎮完整市容。將建物各個端點數位化後，每棟建物便能建立一個多邊形（polygon）資料，連結其所屬之物質特性資料庫，例如樓層數與用途等。我們也將校正後的影像資料轉換為KML格式，地圖便能在Google Earth™上顯示，方便我們將京都市明細圖與現代的京都市容作比較，看出京都市景觀所產生的巨大變化，例如，京都傳統木造建築京町屋已被西式洋房取代，御池通、堀川通與五条通等主要道路也拓寬了。顯見京都市明細圖的數位化對重建現代京都市的歷史景觀與觀察歷史變化有極大助益。

* 日本京都立命館大學地理系博士後研究員。
** 日本京都立命館大學地理系博士生。
*** 日本京都府立綜合資料館研究員。
**** 日本京都立命館大學地理系教授。

1. Introduction

This paper describes the digitalization and construction of a geographic information system (GIS) database for the "Large-scale Maps of Kyoto City" (*Kyoto-shi meisai-zu*) and discusses several research topics using the maps in the context of "historical GIS." The discussion is presented to demonstrate the usefulness of GIS databases in investigating landscape restoration and land use in modern Kyoto.

Landscape restoration calls for past geospatial information. For that purpose, public maps and data, such as picture maps, geographical books, and statistical information, as well as white monochrome photos and private maps are used. New insights can be derived from these sources because private maps contain some information not shared by official maps. However, issues such as uncertain geospatial information and other data are associated with private maps (Ushigaki, 2005). The use and analysis of private maps require checking out of these issues.

At present, a significant amount of information from the past is undergoing digitalization. In 2003, the Library of the University of California at Berkeley entered into a partnership with David Rumsey to implement a digitalization project of Japanese historical maps (Ishimatsu, 2003). This project elicited relevant response not only among researchers but from the mass media as well. In most cases, old maps are precious data, and pictures of them cannot be taken or copied because of preservation reasons. In the past, time and effort had to be invested to research old maps. However, these maps have been digitalized, and the site is open to the public. Thus, visitors to the website can save the maps as part of their collections. Moreover, the visitors can analyze, rotate, enlarge, crop, and compare the maps with their modern counterparts. These features highlight the value of digitalization as an aid to scholarship. The Berkeley website is accessed by a wide range of visitors. The site continues to register numerous hits.

In addition, a considerable number of historians and historical geographers use GIS for cartography (Gregory & Healey, 2007). GIS has numerous functions, and mapping is only one of them. GIS is a special form of database, because each item of data, be it a row of statistics, a string of text, an image, or a movie, is linked to a coordinate-based representation of the location that the data refer to. Thus, GIS is better regarded as database technology, with which constructing various databases of historical information becomes possible. This new field is called "historical GIS." In less than a decade, historical GIS emerged as an important field for historians and historical geographers.

Another important use of GIS is the introduction of a spatial viewpoint using GIS to the

digital humanities, which was derived from the traditional humanities and computer science disciplines (Yano, Nakaya & Isoda, 2007; Yano, Nakaya, Kawasumi & Tanaka, 2011). This development increased the awareness of the importance of geography among humanities scholars who previously would have little interest in geography. In short, digitalization of past information, construction of databases of digitized data, and historical GIS are expanding the field of the humanities.

Two Center of Excellence (COE) Programs at Ritsumeikan University, namely, the 21st Century COE Program "Kyoto Art Entertainment," and the Global COE Program "Digital Humanities Center for Japanese Arts and Cultures" are conducting research and education on the digital humanities. The Virtual Kyoto Project, which aims to reconstruct the historical landscape of Kyoto using historical GIS, is a part of these programs.

The Virtual Kyoto Project comprises four phases: a) archiving geo-referenced materials such as current digital maps, old topographic maps, cadastral maps, aerial photos, picture maps, street photos, landscape paintings, archaeological site data, and historical documents; b) creating a database of all existing buildings, including early modern buildings and structures, such as shrines and temples that are historically and culturally significant; c) creating 3D virtual reality models of the buildings and structures mentioned above; and d) estimating and simulating land use and landscape changes over the study period using the aforementioned materials.

In this process, paper maps and statistical information, which include interim 1:20,000 maps (*kaseizu*) and old aerial photographs of Kyoto taken in 1928 (the Japanese year Showa 3), are scanned and digitalized. In particular, large-scale maps of Kyoto made from the late Meiji era to the Taisho era are valuable resources in the restoration of the landscape of the city of Kyoto. These maps include the "Kyoto Cadastral Map" (1:1,200-1:2,000 scale) of 1912 (Taisho 1) and the "Urban Planning Map of Kyoto City" of 1922 (1:3,000 scale).

In addition, databases of historical architectural forms, such as kyo-machiya (Kyoto-specific wooden townhouses), which are results of a field survey jointly conducted by the Kyoto municipal government and the geography department at Ritsumeikan University are being converted to GIS data. The project is building and compiling a GIS database about Kyoto.

By using these GIS databases, the distribution of land values in the Meiji-Taisho era, and the social geography of Kyoto are clarified. For example, land value distribution in the modern period is discerned as different from the present land value distribution in Kyoto (Inoue,

2007). Kyoto's daily rhythm is clarified using the 1937 "Traffic Census on City Trams and Buses" (Inoue, 2011). Another study utilizes these databases to analyze the social geography of Kyoto. These endeavors provide new insights about modern Kyoto. Digitalizing and constructing a GIS database makes overlaying several maps and information, which clarifies ambiguous parts of documents or maps, possible. These approaches are believed to lead to new areas in the research field. Digitalizing and constructing a GIS database can overcome problems in existing private maps, transforming them into useful documents for study.

This project deals with the digitalization and construction of a database of "Large-scale Maps of Kyoto City (*Kyoto-shi meisai-zu*)." Recently, historical geography and urban history scholars became interested in fire insurance maps. Seemingly, these maps were drawn for calculating fire insurance rates (Ushigaki, 2005; Kokaji, Uchida, Shimizu & Fuse, 2006). These large-scale maps contain various kinds of information related to buildings that can contribute to spatial and social analyses of the targeted urban area. Recently, the development of urban historical research with fire insurance maps has drawn the attention of Japanese geographers. Moreover, the "American Treasures" exhibition of the US Library of Congress has a collection of fire insurance maps that are employed in a wide variety of research.

"Large-scale Maps of Kyoto City" is a fire insurance map, which became available to the public in the Kyoto Prefectural Library and Archives two years ago. These maps contain rich information about historical buildings that are valuable for historical research on Kyoto. This paper describes the process of constructing a GIS database for the maps and discusses research topics related to historical GIS. The characteristics of "Large-scale Maps of Kyoto City" are explained. Subsequently, the process of digitalization and construction of the maps is discussed. This study aims to demonstrate how the GIS databases can be useful in the investigation of landscape restoration and land use in modern Kyoto.

2. "Large-scale Maps of Kyoto City (*Kyoto-shi meisai-zu*)"

"Large-scale Maps of Kyoto City (*Kyoto-shi meisai-zu*)," which is owned by the Kyoto Prefectural Library and Archives, consists of 291 maps, produced between 1927 (Showa 2) and 1951 (Showa 26). The scale of these maps is 1:1,200, and the size of each map is approximately 38 cm by 54 cm. Made for fire insurance purposes, these maps contain various kinds of building-related information necessary to prevent fires. Building lot numbers are included, as well as the number of stories and usage of each building. These various kinds of information related to the buildings were color-coded and added to the maps by hand over time. For example, red represents stores or finance-related buildings, while green denotes

residential houses. Blue represents factories, while yellow denotes temples or shrines. This rich information significantly can contribute to the development of historical research on Kyoto.

However, who added and when exactly the information was added to "Large-scale Maps of Kyoto City" after 1927 remain unclear. To find out how "Large-scale Maps of Kyoto City" were made, an efficient and comprehensive analysis needs to be conducted. The digitalization and construction of a GIS database will help reconstruct the urban landscape of Kyoto between the 1920s and the 1950s, and to analyze the descriptions on the "Large-scale Maps of Kyoto City."

In Japan, fire insurance maps were produced before and after World War II in Tokyo as well as other cities, such as Yokohama and Osaka (Ushigaki, 2005). Before the war, fire insurance maps were produced for the 28 districts of Tokyo and the surrounding cities of Kiryu, Kawasaki, Atami, and Okaya, which are located in the Kanto region. In the Kansai region, maps were produced after World War II for cities like Kyoto, Osaka, and Nishinomiya. Previous studies suggest that the oldest fire insurance maps in Japan are the maps produced in 1928 for Kyobashi Ward, Tokyo. However, "Large-scale Maps of Kyoto City" was produced in 1927, and it is considered that at present, the maps are the oldest insurance maps. Likewise, this means that fire insurance maps were produced ahead of those of Tokyo. Moreover, although these maps were crafted at the same time, they were produced by different companies. Thus, it is safe to assume that there were various processes for making fire insurance maps.

Globally, it is believed that the oldest fire insurance maps were produced in London between 1792 and 1799 (Arlitsch, 2002). Elsewhere, numerous fire insurance maps were published in the late 19th century. Notable are the Sanborn fire insurance maps that document the rise of American cities from 1867 through 1970. The University of Utah Digital Sanborn™ Project website states that the Sanborn fire insurance maps were designed in 1866 by surveyor D. A. Sanborn to assist fire insurance agents in determining the risk associated with insuring a particular property. The D. A. Sanborn Co. was the first to offer insurance maps on a national scale in response to the growth of urban communities after 1850. The company's surveyors meticulously documented the structural evidence of urbanization - building by building, block by block, and community by community (http://content.lib.utah.edu/cdm4/az_details.php?id=0).

The Sanborn fire insurance maps, which have a scale of 1:600 or 1:2,000, provide a block-by-block inventory of the buildings in the built-up or congested parts of towns. The outline or footprint of each building is indicated, and the buildings are color-coded to show

the construction material (pink for brick, yellow for wood, and brown for adobe). Numbers inside the lower right corner of each building indicate the number of stories the building had, while the numbers outside the building, on the street front, refer to the street addresses. This information allows researchers to correlate these locations with census records and city directories. Individual dwellings are marked with "D" or "Dwg," but the residents or owners are not identified. Factories, businesses (such as hotels, saloons, and liveries), churches, schools, and other public buildings (city halls, assay offices, and libraries) are labeled by name (http://www.loc.gov/exhibits/treasures/trr016.html).

These characteristics of the Sanborn fire insurance maps are similar to those of the Japanese maps. However, their color-coding method is different from the "Large-scale Maps of Kyoto City." The former method intended to show construction materials and the latter focused on the usage of each building.

In 2001, ProQuest Information and Learning unveiled the digitalized Sanborn fire insurance maps. Aside from the intended audience of insurance companies, the maps are valuable to genealogists, demographers, environmentalists, urban planners, historians, and laypersons. Thus, it is not surprising that attempts were made in recent years to digitize them.

In Japan, digitalization of fire insurance maps is not popular. Studies on the use of fire insurance maps are few. The presence of numerous sheets and miscellaneous data in fire insurance maps deters analysis. In this case, digitalization and construction of a GIS database for fire insurance maps will aid an efficient and comprehensive analysis of the maps. This approach is the same as the digitalization of old maps and documents. Japanese fire insurance maps have the same value as the Sanborn fire insurance maps. For this reason, this paper discusses the significance of fire insurance maps and historical GIS from the perspective of the digitalization and construction of "Large-scale Maps of Kyoto City."

3. Digitalization and Construction of a GIS Database of the "Large-scale Maps of Kyoto City"

The database of maps is constructed using ArcGIS™. The following steps are taken to digitalize and construct a GIS database of the "Large-scale Maps of Kyoto City."

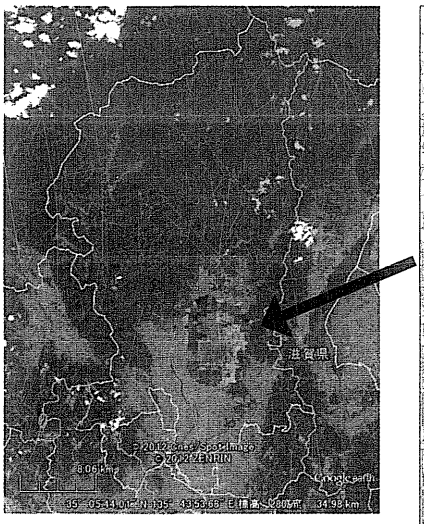
3.1 Scanning of the Maps

The maps are traced. To do so, a big scanner, which enables tracing of an A2-size paper (420mm by 594mm), is used. A big scanner is brought to the Kyoto Prefectural Library and

Archives. Scanning the maps, one by one, took four days to complete. There were tears in several maps, necessitating carefulness. The resolution is set to 500 dpi; the data size of a map image is 300 MB. The total data size is approximately 90 GB.

3.2 Geometric Correction

The maps are traced, overlaid on, and adjusted to match Kyoto's digital city planning maps. ArcGIS™ was used. The procedure, which is called geometric correction, allows the maps to acquire positional information. By using GIS, the procedure permits the comparison of the former and present landscapes. Subsequently, blank spaces are clipped onto the rectified maps using GIS, and 286 sheets of clipped maps are joined to display them as an integrated whole (Fig. 1). In its entirety, the integrated map shows the features of the city after its merger with neighboring towns in 1918.



Map of the present day Kyoto City and the "Large-scale Maps of Kyoto City"

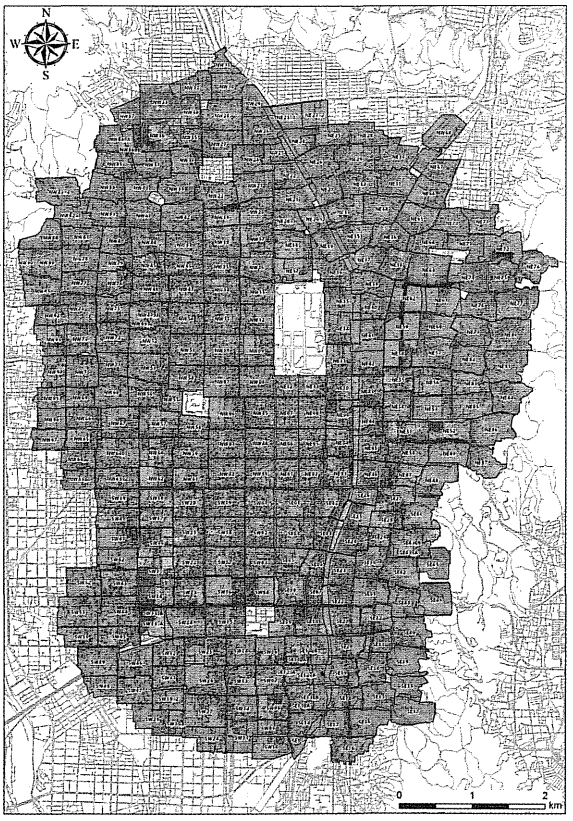


Figure 1 Whole view of the "Large-scale Maps of Kyoto City"

3.3 Making Vector Formats

The information about the buildings includes the outline or footprint of each building, the number of stories, color codes, and usage. Vector formats are formulated for each piece of information, such as polygon data for building outlines and point data for color-coding. These formats provide a structure that makes answering queries, not only about what features are in the database, but where they are located as well, possible.

The objective is to create one figure that holds a significant amount of information. The figure is managed on GIS. This approach is similar to peeling the layers of history. The significant implication of the construction of this GIS database is that the damaged original maps will not be used in the analysis.

3.4 Taking Photos of the "Large-scale Maps of Kyoto City" with an Infrared Camera

Photos of the "Large-scale Maps of Kyoto City" were taken with an infrared camera. While the maps were produced between 1927 and 1951, sheets with new information pasted particularly on the large urbanizing areas were found. To know what was previously drawn on these sheets, the paper was peeled off. Subsequently, photos of the maps were taken. This approach makes the analysis of the original map beneath the new sheets possible. The photographed maps consisted of 154 pieces. Through these maps, the appearance of these regions, which were rapidly urbanizing in 1927, was identified to belong to the beginning of the Showa era (which began in 1926) to the postwar era. For example, in the area south of Kyoto Station and the Shimogamo district, a river covered with a conduit as part of urbanization stands out in relief. Thus, the detailed block layout of the town that is not drawn on city planning maps is revealed.

3.5 Publishing the "Large-scale Maps of Kyoto City" Online

Currently, a database on the individual buildings shown in "Large-scale Maps of Kyoto City" is being constructed for the Virtual Kyoto Project. These image data will be available online from the website of the Kyoto Prefectural Library and Archives starting July 2011.

The site is called the "Library of Kyoto's Memory" with the address of http://kyoto-shiryokan.jp/kyoto-memory/index.php.

The position of the images on the maps will be displayed on Google Maps, and the website, which provides the links to an image data published on "Library of Kyoto's Memory,"

will be constructed ("Virtual Kyoto Project" and "Large-scale Maps of Kyoto City" http://www.geo.lt.ritsumei.ac.jp/meisaizu/meisaizu.html).

At the Art Research Center of Ritsumeikan University, the construction of a database of modern documents is ongoing. Part of the database has been published (http://www.arc.ritsumei.ac.jp/dbroot/top.htm). This approach enables the use of this material not only for study but also by the general public.

4. Applications of the GIS Database of the "Large-scale Maps of Kyoto City"

In this section, the "Large-Scale Maps of Kyoto City" is compared with another GIS database of modern Kyoto. The Virtual Kyoto Project aims to reconstruct the historical landscape of Kyoto using historical GIS. Various GIS databases of modern Kyoto, such as the "Kyoto Cadastral Map" (1:1,200-1:2,000 scale) of 1912 (Taisho 1) and the "Urban Planning Map of Kyoto City" of 1922 (1:3,000 scale), are being established. Comparing the databases will help in determining how the maps were made. Moreover, the GIS database will allow the reconstruction of the urban landscape and urban society of Kyoto between the 1920s and the 1950s.

4.1 Displaying the Maps onto Google Earth™

In many cases, historical GIS is used to explore the process of geographical change. For this purpose, rectified image data of the "Large-scale Maps of Kyoto City" are converted into the KML format to facilitate display of the maps onto Google Earth™. As a result, the maps easily can be compared with present-day Kyoto, thus revealing dramatic changes in the city's landscape.

The comparison shows that the kyo-machiya, Kyoto's traditional wooden houses, were replaced by Western-style buildings. In addition, major streets, such as Horikawa, Oike, and Gojo, were widened (Fig. 2). The widening of the streets resulted from "building evacuation" to prevent fire from spreading during the World War II. At present, numerous buildings are situated around Kitayama and Nishi-oji Streets. In the past, only a few buildings existed in the area, as revealed by the "Large-scale Maps of Kyoto City" (Fig. 3). However, several places hardly changed over time. For instance, a section of the block and the streets in the Kamigamo area have remained the same for the last 80 years (Fig. 4). The investigation of the changes or their absence will reveal how Kyoto expanded during the modern period.

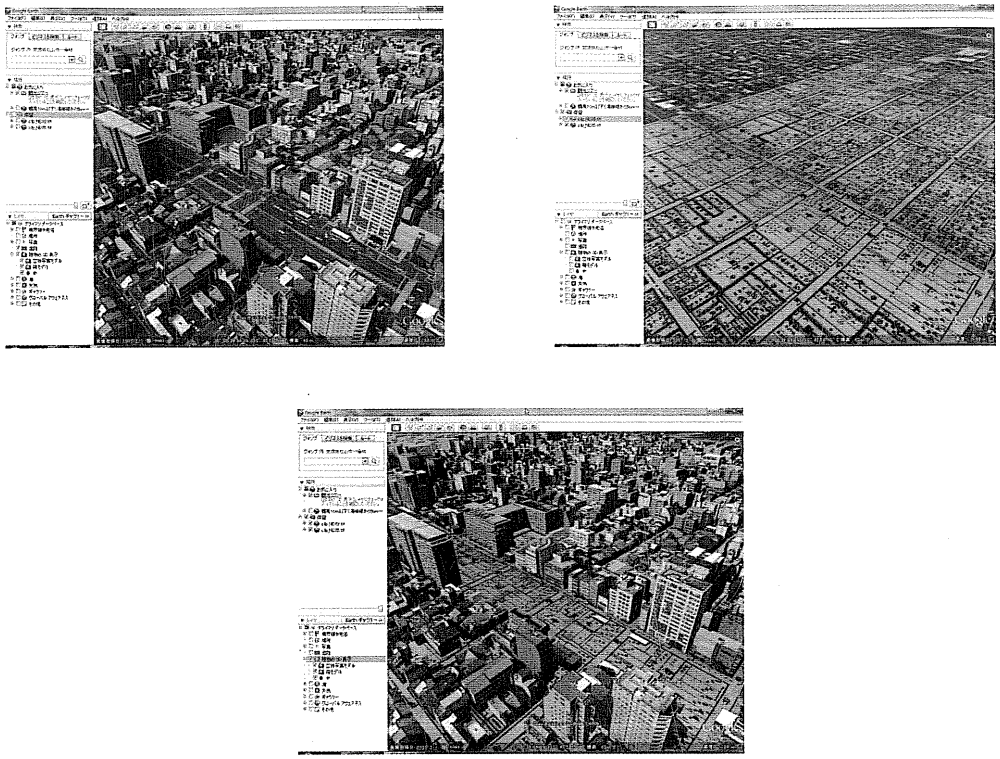


Figure 2 Maps overlaid onto Google Earth™: Oike area

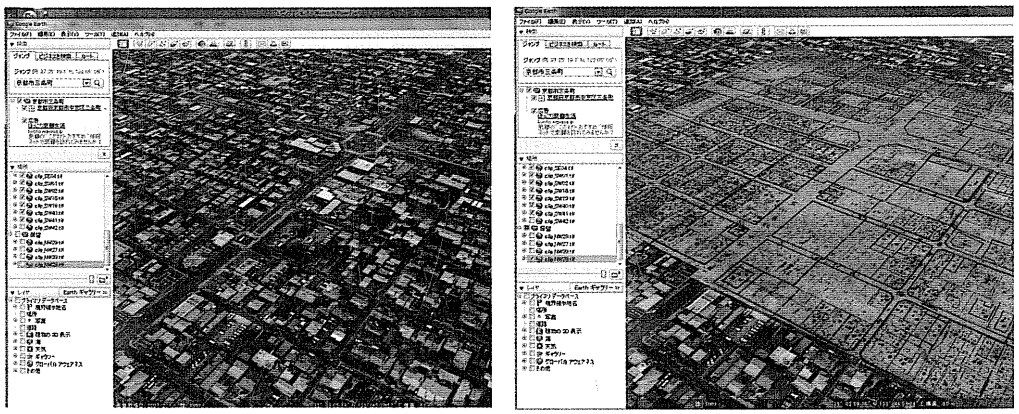


Figure 3 Maps overlaid onto Google Earth™: Kitayama area

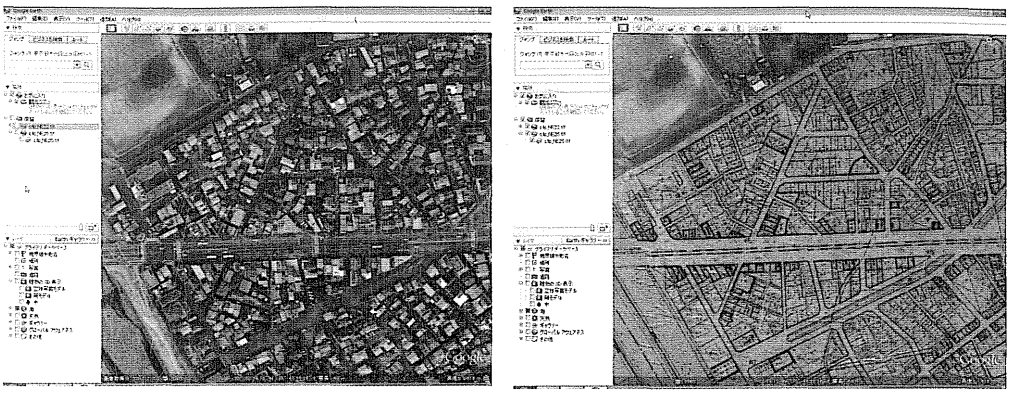


Figure 4 Maps overlaid onto Google Earth™: Kamigamo area

The administrative maps and documents covering "building evacuation" are similar to the style of the "Large-scale Maps of Kyoto City." These maps can be used to identify the details of buildings that were chosen for "building evacuation." The "Large-scale Maps of Kyoto City" can be used not only for fire insurance but for city planning as well.

4.2 Overlaying with "Kyoto Cadastral Map"

The Virtual Kyoto Project has accumulated various GIS databases for modern Kyoto, allowing comparison of the "Large-scale Maps of Kyoto City" with other databases, such as the "Kyoto Cadastral Maps" (Fig. 5). Published in 1912 (Taisho 1), the "Kyoto Cadastral Maps" contain information about the landowner, land price, and land use parcel by parcel. Cadastral maps are used in historical geography and urban history to restore landscapes and analyze the differences between landowners. However, distribution of space in cadastral maps is described based only on distinctions between landowners only. While the distinction between cultivated land and residential land from cadastral maps can be determined, the shape or usage of the buildings cannot be clarified. Comparing the "Kyoto Cadastral Maps" with the "Large-scale Maps of Kyoto City" will help in analyzing the relationship between landowner and land use.

Figure 6 shows the "Large-scale Maps of Kyoto City" and the "Kyoto Cadastral Map" overlay. In the city center of Nishiki-koji, one building occupies a whole parcel of land, which indicates that the "building owner" is essentially the "landowner." In contrast, numerous rented houses existed in the periphery of the city, such as Senbon-Imadegawa (Fig. 7). Previous

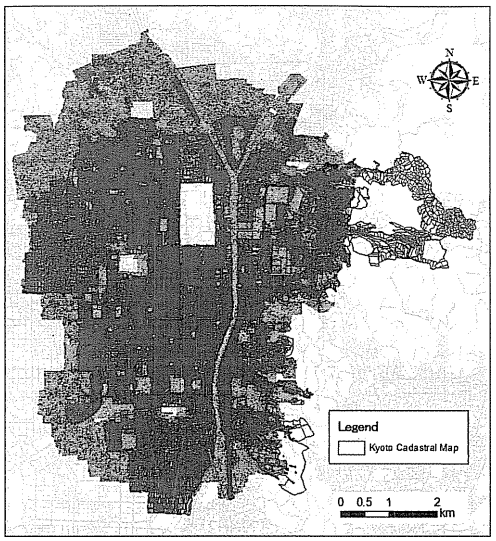


Figure 5 The "Large-scale Maps of Kyoto City" and "Kyoto Cadastral Map" overlaid

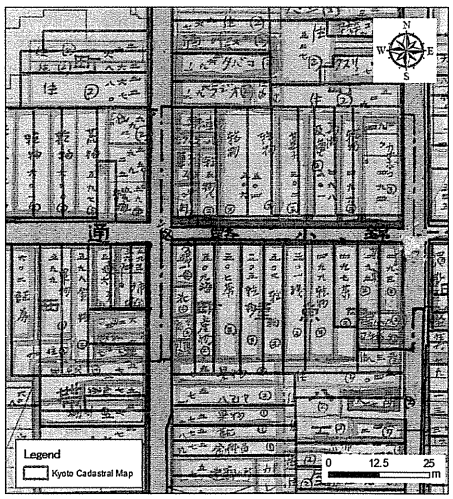


Figure 6 Example of an area around Nishikikoji

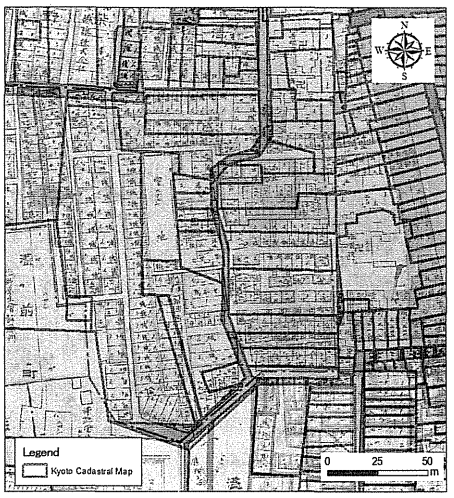


Figure 7 Example of around Senbon-Imadegawa

studies suggest that a limited number of landowners held the majority of the land in the Nishijin area, which was famous for its textile industry. Thus, houses were rented by craftsmen (Mizushima, 2003). Because the lands were divided at the beginning of the Showa era, the urban structure must have changed during the modern period.

Thus, the relations between building and landowners in the inner city and peripheral areas were not the same. The comparison between the "Large-scale Maps of Kyoto City" and the "Kyoto Cadastral Map" enabled the analysis of the urban structure and urban society of Kyoto.

4.3 Distribution of Yuzen Silk Dyeing and Nishijin Brocade Weaving Manufacturers in 1951

In this section, land use in Kyoto in 1951 is analyzed using the GIS database of the "Large-scale Maps of Kyoto City." The spatial distribution of traditional industry in Kyoto immediately after the World War II is revealed. The "Large-scale Maps of Kyoto City" include considerable information about land use, such as the merchandise that a shop was selling, the products of a factory, the illnesses a medical clinic treated, and so on. This study focuses on the location of the buildings related to Yuzen silk and Nishijin brocade industries. Silk and brocade are traditional handicrafts of Kyoto, thus a number of manufacturers for these products existed in the city.

Point data on buildings used for Yuzen silk and Nishijin brocade were created. Figure 8 shows a description of the workshops. Buildings related to Yuzen silk dyeing can be seen around Gojo-Horikawa (the city center) and Takano (the northeastern area of Kyoto). The locations suggest that craftsmen used the Horikawa, Takase, and Takano Rivers (the eastern area) to wash dyed textiles. Moreover, in Takano, a big spinning mill was established in Meiji 41 (1908), with numerous small factories surrounding the mill. The use of the Horikawa River for industrial purposes became unpopular because of increasing pollution. The figure likewise shows a transitional period for the dyeing factories, characterized by the move from the center of the city to the periphery.

The industry-related area of the Nishijin brocade weaving can be seen between Senbon-Imadegawa, in the Nishijin area and north of Kitaoji. As of 1951, only a few buildings existed north of Kitaoji. A number of yarn dyers live in Nishijin. In the "Large-scale Maps of Kyoto City," particular buildings were marked as used for twisted yarn and yarn dyeing. Various craftsmen were involved in producing the Nishijin brocade, as seen from the "Large-scale Maps of Kyoto City." Numerous small factories sprung up in the kyo-machiya in this area, which were distinct from the ordinary houses and houses of merchants elsewhere.

The investigation of late-Meiji housing structures in the center of the city revealed the

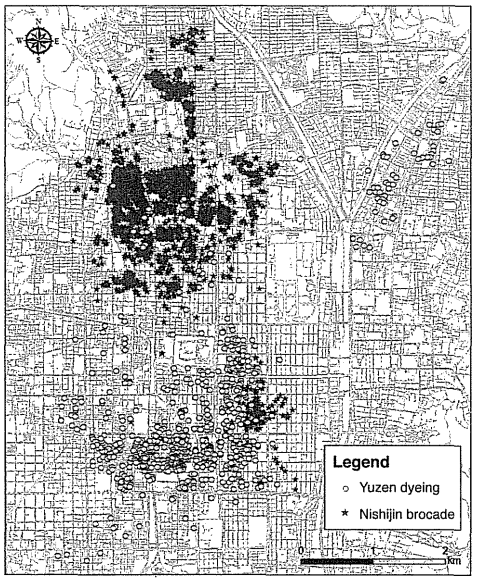


Figure 8 Distribution of Yuzen dyeing and Nishijin brocade establishments

existence of large mercantile establishments and manufacturers in the late Edo period. The "Large-scale Maps of Kyoto City" demonstrate these characteristics of Kyoto's urban structure.

4.4 A Study of Occupied Kyoto and the "Large-scale Maps of Kyoto City"

In this section, the characteristics of the "Large-scale Maps of Kyoto City" are discussed using GIS and other data. When the maps were produced, between 1927 (Showa 2) and 1951 (Showa 26), various information about buildings were added to the maps. In particular, building color-coding and usage information were included after World War II. Between 1940 and 1955 (prewar and postwar), the social situation dramatically changed in Japan, thus, the background of the production of the maps has to be analyzed. Between 1945 and 1952, Japan was occupied, and the situation of the occupied Kyoto has to be considered.

Administrative documents about the occupied Kyoto owned by the Kyoto Prefectural Library and Archives were exhibited in the spring of 2010. One of the documents details the accidents that the Occupation Army caused in Kyoto. Another document lists the buildings confiscated by the Occupation Army. These documents contain positional information. Thus, previous research on the mapping of these data by hand is important in learning about the occupied Kyoto (Nishikawa, 2010). Below, the contents of the maps are examined based on the results of the study on the occupied Kyoto.

Among the administrative documents stored in the Kyoto Prefectural Library and Archives are documents about accidents and confiscated buildings. The occupation army confiscated several buildings. In this paper, the buildings used for businesslike operation by the Occupation Army are called "confiscated establishments," and buildings used for dwellings by the Occupation Army are called "confiscated houses." Vector data for each confiscated building were constructed based on the positional information to improve the precision of previous research and facilitate the comparison of maps. In addition, confiscated building data were converted to the KML format. Thus, the maps can be accessed through Google Earth™.

In the results, confiscated establishments stand out around Shijo-Karasuma and Kawaramachi-Sanjo, near the Kyoto Station. On the other hand, confiscated houses were distributed east of Horikawa, especially around Shimogamo, Shirakawa, Okazaki, Higashiyama, and the Old Imperial Palace. Previous studies pointed out the need to refurbish Japanese kyo-machiya to Western-style house for the convenience of the Occupation Army. Therefore, it is assumed that the so-called modern architecture or the large residences that could be refurbished were chosen as the objects of confiscation. Therefore, the narrow kyo-

machiya at the center of the city, and around Shijo-Karasuma, was not a target of confiscation. In contrast, the accidents that the Occupation Army caused in Kyoto were distributed in Sanjo Street, Shijo Street, and Karasuma Street. The Occupation Army moved by jeep and truck, which the Japanese were unused to. Thus, damage to buildings and fatal accidents were inevitable particularly on the main streets. In Sanjo Street, the accident sites were in the neighborhood of prefectural border with Otsu city. In addition, accidents occurred in the entertainment districts and in the areas near the confiscated buildings.

The "Large-scale Maps of Kyoto City" were compared with present-day Kyoto on Google Earth™ based on the distribution of confiscated establishments, confiscated houses, and accidents related to the Occupation Army. As a result, the meanings of the Roman letters recorded on the maps were understood and the kind of building that was confiscated was determined.

The "Large-scale Maps of Kyoto City" was examined in the context of the historical knowledge gained about the postwar period. Construction of a GIS database is useful to distinguish the timeframes and facilitate analysis. Overlaying a GIS database of modern Kyoto and past information define the direction of the study on occupied Kyoto and present-day Kyoto, as well as enhance the understanding of the content of the "Large-scale Maps of Kyoto City."

5. Conclusions

Based on the discussion on the digitalization of the "Large-scale Maps of Kyoto City (*Kyoto-shi meisai-zu*)" and on the research topics based on the context of historical GIS, the following conclusions are presented.

(1) Conversion of the rectified image data of the "Large-scale Maps of Kyoto City" to the KML format allowed the display of the maps onto Google Earth™. As a result, the maps can easily be compared with present-day Kyoto. The dramatic changes in the city's landscape are revealed.

(2) The "Large-scale Maps of Kyoto City" demonstrate the characteristics of Kyoto's industrial structure in relation to the information on building usage for activities like Yuzen silk dying and Nishijin brocade production.

(3) Overlaying between "Large-scale Maps of Kyoto City" and "Kyoto Cadastral Map" enabled analysis of the urban structure and urban society of Kyoto. The relations between buildings and landowners were different between the inner city and the

peripheral areas.

(4) Integration with the study of occupied Kyoto can assist in the identification of the information of building usage found in the "Large-scale Maps of Kyoto City."

The construction of a historical GIS database of the maps helps in the reconstruction of the landscape of modern Kyoto and reveals historical changes. Moreover, the construction enables the examination of the complicated information given in the "Large-scale Maps of Kyoto City."

A major factor, which slowed down the development of the historical GIS, was the length of time required to build databases. Nevertheless, the maps are full of information that is useful for better understanding of the history of modern Kyoto. Thus, the maps should be efficiently used. For a long time, the field of urban history of Kyoto has focused on urban structure, such as the changes in land ownership and land use (Mikura, 2008; Iwamoto, 2011). In this field, the study area tends to be small, often at the street level. Owing to this, analyzing Kyoto more comprehensively by looking at the city from a broader perspective is necessary. In addition, focusing not only on the historical change of each small area, but on the contemporary regional differences within Kyoto City as well is necessary (Kirimura, 2009). For these reasons, digitalization and construction of a GIS database of the "Large-scale Maps of Kyoto City" are crucial. In these processes, identification of the process of and people involved in the survey and production of the "Large-scale Maps of Kyoto City" in the early Showa era was important. The analysis provided by this study will aid in the development of urban map study using fire insurance maps and large-scale maps.

On the other hand, problems on digitalization of old maps exist (Ishimatsu, 2009). One of these problems is copyright, which has become evident in recent years. Another problem is related to the content and format of a digitized image. The correct content and format facilitate the display of the maps onto Google Earth™, where people can pinpoint the position of communities that were discriminated against on the old map. Thus, recognizing these discriminated communities is important. In this project, a digitalized image and GIS database of the "Large-scale Maps of Kyoto City" is published on the web, bringing about unexpected problems. Nevertheless, digitalization and construction of a GIS database of old maps and documents will be effective in challenging historical orthodoxies, answering questions that had previously been difficult to address and posing entirely new questions. As stated above, it is important that these databases be widely opened and used under certain rules.

The Virtual Kyoto Project has generated a vast quantity of geospatial GIS data including

a wide variety of old maps, old aerial photos, cadastral maps and wide-ranging registrations, and directories on kyo-machiya. The next challenge involves restoring the past urban landscape of Kyoto during the modern period on the basis of the historical GIS database of Kyoto.

Historical GIS has been accepted by scholars in the fields of historical geography and urban history in terms of quantitative and qualitative studies. Moreover, historical GIS has continued to develop. Construction of GIS databases is the foundation of historical GIS. Therefore, further exploration of the process through which historical GIS can be applied to historical study may be done.

References

Gregory, I. N. & Healey, R. G. (2007). Historical GIS: Structuring, Mapping and Analyzing Geographies of the Past. *Progress in Human Geography, 31*(5), 638-653.

Inoue, M. (2007). Distribution of Land Values in the Meiji-Taisho Era. *Virtual Kyoto*. Kyoto, Japan: Nakanishiya, 62-65.

Inoue, M.(2011). Kyoto's Daily Rhythm as Seen from Traffic Flow in the Prewar Showa Period. *Historical GIS of Kyoto, Kyoto*, Japan: Nakanishiya, 263-272.

Ishimatsu, H. (2009). Mapping the Past: The Digitalization of the Japanese Maps at University of California at Berkeley (<Special feature> Applications of the geographic information systems). *Information Science and Technology*, 59-11, 557-667.

Ito, O. (2006). Sanborn Fire Insurance Maps in Urban Cities. *Kansai University Library Forum, 11*, 33-34.

Iwamoto (Mikura), Y. (2011). Cho (Township) and Landownership in Modern Kyoto: The Case of Kitanogomon-cho. *The Urban History Annual, 18*, 65-82.

Kenning, A. (2002). Digitizing Sanborn Fire Insurance Maps™ for a Full Color, Publicly Accessible Collection. *D-Lib Magazine*, 8. http://www.dlib.org/dlib/july02/arlitsch/07arlitsch.html

Kirimura, T. (2009). Changes in Residential Structure in 20th-century Kyoto City. *Japanese Journal of Human Geography, 61*(6), 528-547.

Kokaji, M., Uchida, Y., Shimizu, H. & Fuse, T. (2006). Applicability of Fire Insurance Maps in the Urban History: For the Street Research during Pre- and Postwar Periods in Tokyo. *Annual Meeting of the Japan Society of Photogrammetry and Remote Sensing*, 47-50.

Matsui, K. (1979). The Structure of Nishijin Industrial Region: Study of the Small-scale Wearing Industry in Kyoto Metropolitan Area. *Japanese Journal of Human Geography, 31*, 117-136.

Mikura, Y. (2008). Land Owners Trend and Urban Space in Modern Shinkyogoku. *Journal of Architecture, Planning and Environmental Engineering (Transactions of AIJ), 629*, 1651-1656.

Mizushima, A. (2003). A Study on Trend of Land Employment of Landowners in Early 20th Century in Nishijin Area, Kyoto. *Journal of Architecture, Planning and*

Environmental Engineering (Transactions of AIJ), *565*, 373-378.

Nishikawa, Yuko. (2010). Occupied Kyoto: Introduction to a Study of Occupied Kyoto. *Arena, 10, 143-162.*

Ushigaki, Y. (2005). Characteristic and Use of Fire Insurance Maps as Large-scale Maps in the Showa Period. *The Historical Geography*, *226*, 1-16.

Yano, K., Nakaya, T. & Isoda, Y. (Eds.). (2007). *Virtual Kyoto*. Kyoto, Japan: Nakanishiya.

Yano, K., Nakaya, T., Kawasumi, T. & Tanaka, S. (Eds.). (2011). *Historical GIS of Kyoto*. Kyoto, Japan: Nakanishiya.

Towards Social Application and Sustainability of Digital Archives: The Case Study of 3D Visualization of Large-scale Documents of the Great Hanshin-Awaji Earthquake

Akinobu Nameda *, Kosuke Wakabayashi **,
Takuya Nakatsuma ***, Tomomi Hatano ****, Shinya Saito *****,
Mitsuyuki Inaba ******, Tatsuya Sato *******

Abstract

The present study concerns possible methods to increase the sustainability of digital archives in cultural heritage and social institutions. Advancements in technology over the past decade enabled digital archiving of large amounts of data. Although digital archives deliver a vast amount of information, accessing key information remains difficult. Therefore, the present study offers improvements in how key information is accessed, in order to increase the sustainability and social merit of digital archives. Our system is expected to be applied in visualizing key information related to the Great East Japan Earthquake that happened on March 11, 2011.

We visualized the database of the Great Hanshin-Awaji Earthquake that occurred on January 17, 1995 which was created and published by the government of Japan. The database included textual information, such as what happened in towns and local places during the

* Ph.D. student, Center for Law and Psychology, Ritsumeikan University, Japan.
** Research Assistant, Center for Law and Psychology, Ritsumeikan University, Japan.
*** Ph.D. student, Center for Law and Psychology, Ritsumeikan University, Japan.
**** Lecturer, Center for Law and Psychology, Ritsumeikan University, Japan.
***** Lecturer, Digital Humanities Center for Japanese Arts and Cultures, Ritsumeikan University, Japan.
****** Professor, Digital Humanities Center for Japanese Arts and Cultures, Ritsumeikan University, Japan.
******* Professor, Center for Law and Psychology, Ritsumeikan University, Japan.

earthquake, and the timeline of events in each social sector. For the visualization, we utilized KACHINA CUBE (KC) system, a web-based platform that allows the storage, plotting, and display of information in three-dimensional space. After inputting data from the government database into the KC system, we designed the KC output: a 2D geographical space map and a time line on the vertical axis, with segmented articles in the database plotted into three-dimensional space.

With the KC containing and displaying database textual information on the earthquake, we analyzed the mapped patterns of segmented articles in each time period. In other words, articles on different locations, time Periods, and topics were visualized. Articles were grouped according to their particular place, time, and topic; for some of these classifications, not many articles existed. The possibilities of visualizing using the KC system to improve the accessibility of key information in archives are discussed in this study.

數位典藏應用的社會效益與永續經營——以阪神大地震資料3D視覺化為例

滑田明暢*、若林宏輔**、中妻拓也***、破田野智己****、
齋藤進也*****、稻葉光行******、佐藤達哉*******

摘 要

本研究旨在探討如何針對文化遺產保存與社會機構等，找到應用數位典藏與提升數位典藏永續性的可行方法。過去十年，由於科技發展，大量的資料得以進行數位典藏。然而，數位典藏雖然產出豐富的資料，但許多重要資訊仍然不易使用。本研究希望能開發出重要資訊的方法，以提高數位典藏的永續性與社會價值。我們希望未來能利用這套系統，將2011年3月11日東日本大震災的重要資料作視覺化呈現。

本研究以日本政府建置公布的1997年阪神大地震資料庫，作視覺化平台的試驗，未來將可作為其他震災資料視覺化的參考。此資料庫除記錄地震期間市鎮與地方的狀況，也以時間軸記載當時的社會事件。我們利用KACHINA CUBE（KC）網路平台，在立體空間中儲存、模擬與顯示資料，我們將政府建置的阪神大地震資料庫輸入，利用所設計的輸出界面，結合2D地圖與時間縱軸，將一個個時間片段繪製在3D空間，使之視覺化呈現。

透過KC系統所展示地震庫的資料，我們可以進一步分析每一個時間片段的資料。由於每筆資訊的地點、時間與類型等都可以直接觀察到，我們可以發現，在某些特定地點、時間或主題上，資料特別集中；但對有些地點、時間與主題而言，資料卻很稀少。

---

* 日本京都立命館大學法律心理中心博士生。
** 日本京都立命館大學法律心理中心研究助理。
*** 日本京都立命館大學法律心理中心博士生。
**** 日本京都立命館大學法律心理中心講師。
***** 日本京都立命館大學日本藝術與文化數位人文中心講師。
****** 日本京都立命館大學日本藝術與文化數位人文中心教授。
******* 日本京都立命館大學法律心理中心教授。

1. Introduction

Individuals are surrounded by large amounts of information in modern society. Such abundant information has been shaped and supported by digital technology, thus we are able to share and exchange much information fast. This sharing and exchanging would be benefited by the development of technology that allows the storage of vast information in digital space.

A large number of people and organizations are known to use digitally stored information. For instance, the government retains the information from the surveys and from meetings regarding administrative matters, and makes this available on the Internet. Many citizens search for information from government websites. Various other fields retain information digitally. As archiving information in digital space becomes more common, the need to develop technology that supports this increases. In fact, the creation of the following archives of narratives and visual information on the Great East Japan Earthquake has started: the Digital Archive of Great East Japan Earthquake 2011 by Reischauer Institute of Japanese Studies, 2011; and the Archive project of Great East Japan Earthquake by Tohoku University, 2011. Society generally collects considerable information on disasters, such as earthquakes, and utilizes these archives when a similar disaster occurs.

Although digital archives deliver tremendous amounts of information, accessing key information remains difficult. In other words, archiving not only involves storing vast information, but also providing access to such information. The required information may be maintained in archives, but is useless unless found, retrieved, and used. Information from archives is typically accessed in order to contribute to society. To use an analogy, a child in the first year of primary school, who is given a dictionary with words at the adult level, is not likely to read and understand the text, thus gaining no information regarding what he is seeking from the dictionary. Another example is when the proceedings of a government meeting are transcribed then posted on a national website. On one hand, we would be able to access raw data pertaining to what transpired during the meeting. On the other hand, ordinary citizens may find difficulty in reading the transcription and understanding the main topics, or the conclusions and implications of the discussions of the meeting. Without clear guidance, the main points discussed fail to reach the audience. Thus, although the information is stored and retrieved, when it is too complex to understand or time consuming to extract main points, then it cannot be considered accessed but merely scanned. Such situation where in the information fails to reach citizens is regarded unfair, referencing the notion of "informational justice" discussed by Professor Ibusuki (2011) during an 12th Annual Conference of Japanese Society for Law and Psychology held in Japan. The information in archives should be

organized and easily accessible, except in private archives which are closed to the public.

A well-organized archive facilitates its usage by citizens thus benefiting them; the information in archives should be clearly presented. Therefore, the present study explores ways to improve accessibility to key information in digital archives in order to increase their sustainability and social merit. We will develop a concrete system that visualizes the information database of the Great Hanshin-Awaji Earthquake that struck central Japan in 1995. Such system will eventually be applied to visualize key information related to the Great East Japan Earthquake of 2011.

2. Visualizing the Information of the Great Hanshin-Awaji Earthquake Using KACHINA CUBE System

2.1 The Tool, Concept, and Methodology for Visualization

KACHINA CUBE (KC) system was utilized in visualization; it is a web-based platform that allows one to store, plot, and display information in three-dimensional space (Saito, Ohno & Inaba, 2009; Ohno, Saito & Inaba, 2010). After inputting data on the Great Hanshin-Awaji Earthquake from the government database into the KC system, we designed the KC output which contained a 2D geographical map of the Hanshin-Awaji area, and a time line on the vertical axis (the Great Hanshin-Awaji Earthquake is described later in the next section), and the segmented textual information from the archive was plotted into three-dimensional space. A visual image of the KC output of the present study is presented in Figures 1 and 2.

The KC system in the present study was used to clearly display the phenomenon to viewers. KC was generally designed to allow viewers to simultaneously view, in detail and in whole, the articles contained in the information database (e.g., Saito, Ohno & Inaba, 2010). Using the KC system, we focused on exploring the meanings of articles, rather than pursuing a quantitative analysis of the phenomenon, in order to allow viewers to gain new findings.

To promote a clear understanding of the phenomenon, the KC system operated on three main functions. First was to access the original articles in the archive. Sentences from the articles could be viewed by clicking information fragments plotted as segmented information. Second was to allow the cube in the system to rotate so that viewers could locate each information fragment on the map according to time period, at different angles. Finally, the system was equipped with a search function. Readers could thus retrieve only information that interested them, or which they wanted to analyze.

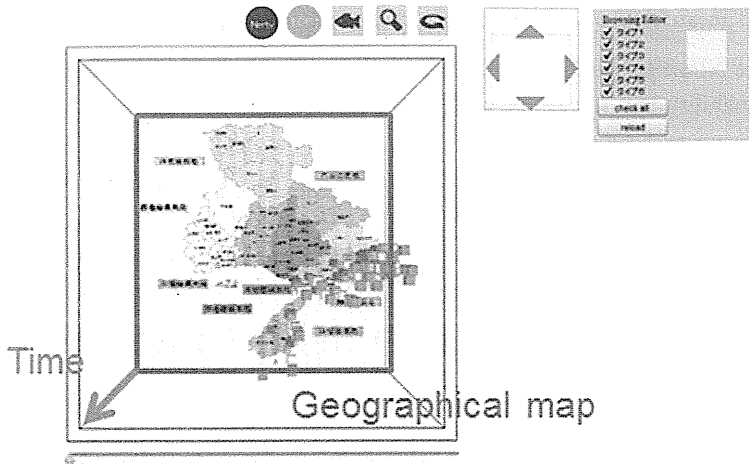


Figure 1 Image of the KACHINA CUBE output in the present study, as seen from the top

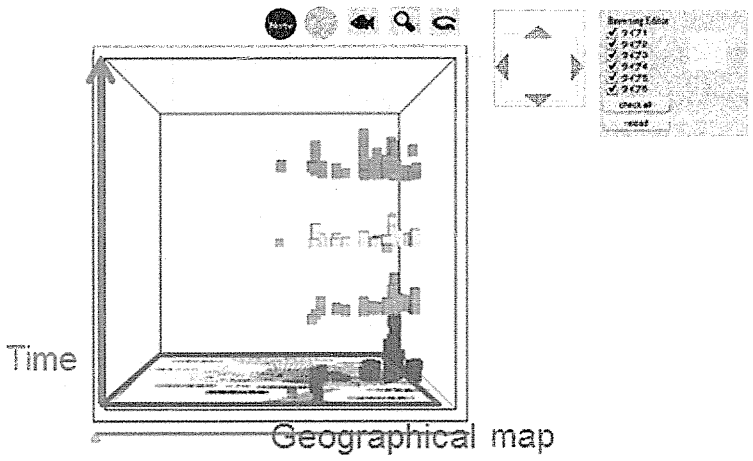


Figure 2 Image of the KACHINA CUBE output from the present study, as seen from the front

By arranging textual information in three-dimensional space using KC, viewers can gain a visual and spatial image of the textual information in the archive. Knowledge and an intuitive understanding of the textual information are therefore obtained. In the present study, the articles in the database were arranged on the geographical map and time line. As the result, the place and time under which each input or textual information belongs becomes clear. Moreover, by accessing the original articles, textual information is found. In summary, the use of the KC allows viewers to gain knowledge with integrating spatial and temporal perspectives similar to that in a map showing the weather forecast.

The spatial and temporal visualization of the articles in the archive also results in better understanding of the relevance of articles. For example, the search function identifies information fragments with the same expressions. Connecting these information fragments shows how these are related to each other. Viewing the relevance of the information fragments in different places and time Periods depicts the connections among these. Such depiction creates knowledge, in the reader, about the textual information.

The present study aims to demonstrate and suggest a visualization technique using the KC system described above. The first step in visualization is when the viewer gains a visual image of the articles in the information database, by mapping these as information fragments in 3D space. The second step considers the quantitative differences among information fragments across different places and time. In this step, the viewer discovers the volume of articles available in particular locations, and whether such volume varies across different places and time. The third step is when the viewer explores the textual information while making inferences as to why such differences in volume of articles exist. Finally, in exploring the answer to their inferences, viewers locate key information and acquire knowledge from the archive by accessing the original articles displayed in the KC, or by using the search function.

2.2 Target of the Visualization

To provide a way to access key information in archives, we visualized the information database of the Great Hanshin-Awaji Earthquake of 1995, which was created and published by the government of Japan (Cabinet Office, Government of Japan, 2006), for the case of the other great earthquakes in the future.

The Great Hanshin-Awaji Earthquake struck the Hanshin area and Awaji island which are located in west central Japan. A magnitude of 7.2 was recorded at the epicenter, and 6,400 fatalities. Tremors caused by the earthquake were felt in several Japanese islands from the west such as Kyushu area, to the east such as Kanto area which includes Tokyo, and even the north

eastern part of Japan (Cabinet Office, Government of Japan, 2006).

The purpose of creating the database is to establish and offer systematic textual information useful for improving disaster prevention measures in the future (Cabinet Office, Government of Japan, 2006). Textual information was collected from a broad range of documented resources such as published literature and reports from public institutions, and from experts and journalists concerned about the reactions and reconstruction after the earthquake, and people who suffered from the disaster. In creating the database, the textual information was organized and documented in terms of three viewpoints, as follows: facts on the events following the earthquake, actions taken, and action plans that still needed to be addressed. The original information database was structured and described in three levels, namely: lessons learned, abstract version of textual information, and original articles from the resources. In our visualization, only the descriptions of the original articles from the resources were used.

The database included relevant articles on what happened in towns and local places within the Hanshin-Awaji area 10 years after the earthquake, or from 1995 to 2005. These articles were classified into four periods. Period 1 covered the actual earthquake up to 72 hours later. Period 2 was from Day 4 to three weeks after. Period 3 was from 4 weeks to 6 months after the earthquake. Finally, Period 4 was the period after 6 months. Articles included in each period were structured, and consisted of various aspects such as the series of events relevant to the earthquake, and damages and reconstruction of infrastructure. Period 1 includes articles on what happened and how people, organizations, and the government reacted to what happened during the earthquake. In other words, what was the extent of damage and how were these handled. Period 2 describes the reactions of the government and other institutions; although quite similar to Period 1, the topics of the articles were different as these happened in a different period. In Period 3, the recovery of affected areas was the main subject for organizing articles. Finally, Period 4 included articles on action taken during reconstruction. Details on these topics are presented in Table 1.

2.3 Procedures for Putting Information Database into the KACHINA CUBE System

In inputting database information into the KC system, the textual information was first segmented into items containing several sentences from the original article that describe one aspect of the situation or events related to the earthquake. Second, these items were classified according to municipality. To avoid the complicated mapping of items within areas of the

Table 1 Content structure of articles in the information database of the Great Hanshin-Awaji Earthquake

(from Cabinet Office, Government of Japan, 2006)
(Note: The authors translated from Japanese to English.)

Period 1 (from the actual earthquake to 72 hours after)	**Period 2 (from Day 4 to 3 weeks after the earthquake)**	**Period 3 (4 weeks to 6 months after the earthquake)**	**Period 4 (6 months after the earthquake)**
1-01.Actual earthquake	2-01.Operating evacuation centers	3-01.Building emergency dwelling	4-01.Reconstruction of life
1-02.Initial reactions	2-02.Supporting life in disaster area	3-02.Rebuilding residences and life	4-02.Revival of the industry and cities
1-03.People's behavior	2-03.Determining the situation	3-03.Planning reconstruction	
1-04.Rescue and emergency medical treatment	2-04.Volunteers	3-04.Demolishing damaged buildings	
1-05.Fire handling	2-05.Reestablishing urban infrastructure	3-05.Industry recovery	
1-06.Emergency transportation			
1-07.Food and supplies			
1-08.Health and hygiene			
1-09.Lifeline			
1-10.Reaction among companies			
1-11.Preventing a second disaster			

municipality, we assigned particular coordinates to each area then manually plotted each item as an information fragment on the 3D space in the KC system. Thus, each fragment of information that mentioned the name of a particular municipality was marked in a particular space within the municipality on the geographical map; these information fragments were then grouped according to time period. The visual output image of the KC with information fragments plotted is shown in Figures 3 and 4. A total of 1,000 fragments were created in the KC system, in contrast to the thousands of information fragments in the original information database.

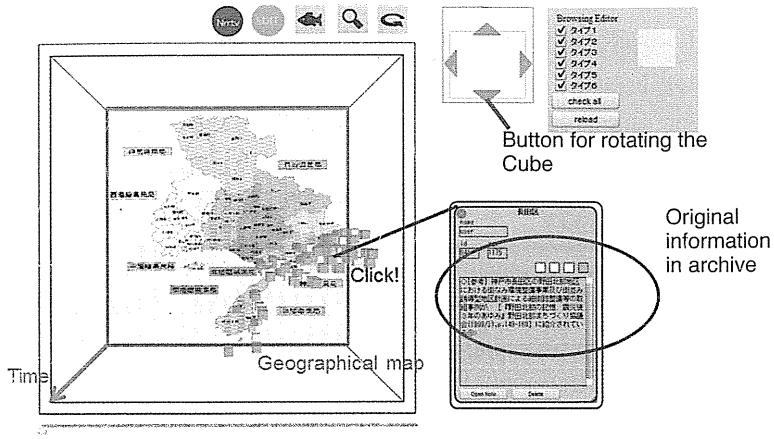


Figure 3 Visual image of the KACHINA CUBE output (as seen from the top) and its functions of rotating and accessing original information of a fragment

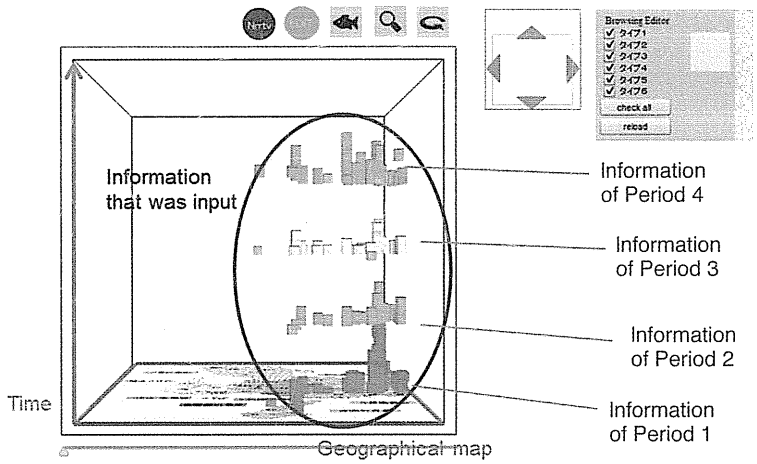


Figure 4 Visual image of the KACHINA CUBE output, as seen from the front

3. Results and Discussion: Findings from Visualization Using the KACHINA CUBE

3.1 Advantages of Visualization Using a KACHINA CUBE with a 2D Geographical Space Map and Timeline on the Vertical Axis

The information database of the Great Hanshin-Awaji Earthquake was visualized by inputting information into the KC system; such system stores, plots, and displays the textual information in three-dimensional space, and contains a 2D geographical space map and timeline on the vertical axis. As the result, the textual information was visually organized according to place, time period, and topic. The KC system allowed us to search and use the textual information from the large documented database, with reference to space and time.

One chief advantage of using the KC system above is that the whole information can be seen at a glance. A visual view of the database is likewise provided. The database originally contained a large scale of documents hence getting a glimpse of the information on these documents was difficult. The KC system, however, allows users to visually understand and depict the textual information based on a particular time and space. Moreover, the original articles could be viewed by clicking the information fragments on the system. As a result, an overview of the original articles could be obtained, prior to accessing it.

Another advantage is that users of the KC system can view and understand the quantitative distribution of articles in the geographical map, in each time period. In other words, the KC system in the present study allows viewers to determine which places or time Periods contained no articles on the issue. For example, the visualization in the present study shows that a large amount of textual information belongs to the central part of the Hanshin-Awaji area, and is sparsely distributed in other places. In this regard, the user is likely to visually concentrate on the area with many articles. Few may want to view and explore other areas with less information, and will find no article in areas where no information exists. Therefore, the information database of the Great Hanshin-Awaji Earthquake, in fact, collected quantitatively unbalanced textual information, in terms of geographical distribution.

Another finding can be inferred by paying attention to this so-called quantitative distribution of articles in the geographical map. Comparing the number of articles across particular places and time Periods reveals how textual information on the Hanshin-Awaji Earthquake is accumulated. Few information fragments were related to the Nagata area in Periods 1, 2, and 3 (Figure 5). During these Periods, the Nagata area experienced the most

damage, with the large number of houses destroyed in Kobe city (City of Kobe, 2010); however, the Nagata area did not have as much information as Nishinomiya city or Ashiya city. Thus, we can hypothesize that gaining information from severely damaged places is more difficult.

With regard to the change in number of articles across time Periods, another finding is revealed. There were relatively more articles on the Nagata area in Period 4 (Figure 5). In checking the content of the articles within the KC system, we find that the various articles and investigation reports on the Nagata area relate to various aspects of reconstruction, including the population, development of towns, and the people's lives in general after the earthquake. In sum, studying the changes in volume of articles over time in particular places allows the users of KC to be familiar with the different aspects and events related to the earthquake. Quantitative changes in information fragments in particular places over time, as seen through the KC system, lead to new perspectives and further exploration.

The use of the search function reveals more characteristics of the earthquake. For example, in searching information fragments with the word "volunteer" (the searched word was in Japanese), numerous results were obtained in different periods (Figure 6); however, Japan, in fact, had the most number of volunteers in Period 2. In reviewing the contents of the information fragments that contained the word "volunteer" in the KC system, we find that the volunteers mainly provided support in Period 2, which was four weeks after the earthquake. Articles from Period 4, covering six months to 10 years after the earthquake, described some volunteers as having continued their support, while some volunteer groups stopped working. Although we need to consider that the articles in the database were limited to those collected by the editors, authors, and creators of the database on the said topic, we could confirm that the volunteers mainly worked four weeks after the earthquake and that some of them have stopped providing support, as gleaned from the content of the original articles.

Whereas articles on the volunteers were contained in the database, articles of Nonprofit Organizations (NPO) were barely seen. Searching information fragments with the word "NPO" only yields a few results (Figure 7). This can be attributed to NPOs' being legally certified by the government a couple of years after the earthquake. The NPO law that granted organizations involved in volunteer activity the title "NPO" was enforced in 1998 (Cabinet office, Government of Japan). The term NPO is now widely used by in Japanese society, and citizens are familiar with their activities, unlike before. Searching for textual information on the KC system causes viewers to realize new facts thereby encouraging them to appreciate the phenomenon more.

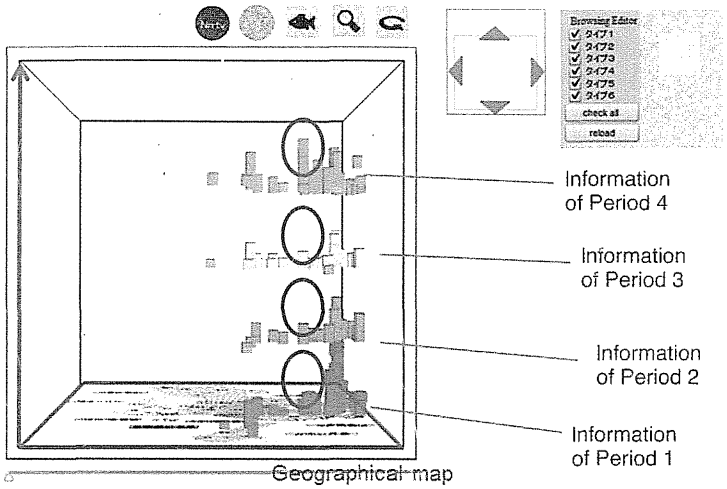


Figure 5 The quantitative change in information fragments related to the Nagata area from Periods 1 to 4. The fragments in red circles are the articles on Nagata area in each period (the KACHINA CUBE system as seen from the front)

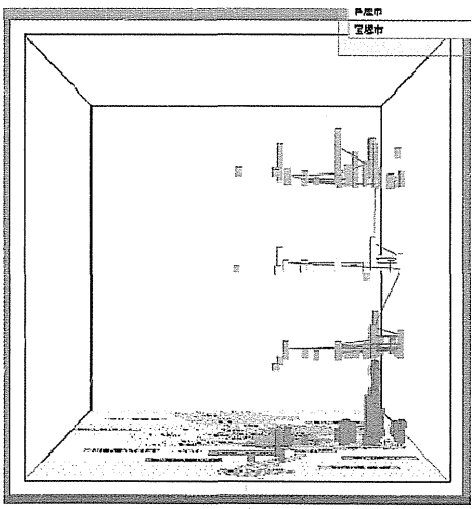


Figure 6 The result of searching information fragments containing the word "volunteer" (The information fragments with "volunteer" are connected with lines in the Figure)

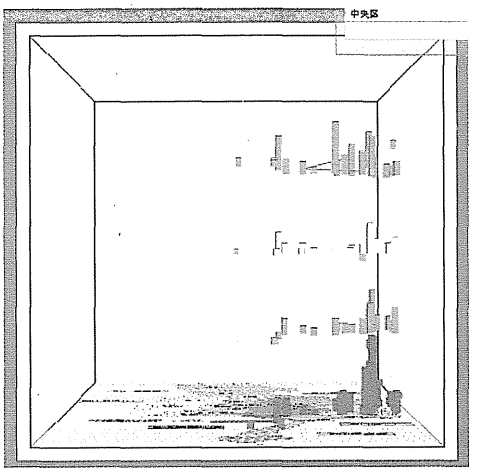


Figure 7 The result of searching information fragments containing the word "NPO" (The information fragments with "NPO" are connected with lines in the Figure)

3.2 Problems and Limitations

In order to utilize the KC system more efficiently, certain problems should be resolved. One is that the contents of information fragments are not provided in one view. Although users can view how the information fragments were distributed on the map over time, at a glance, the types of articles included in each information fragment are not visible until the users click the information fragments to see the original articles. Presenting a screen that shows only the distribution of information fragments on the map over time reduce the efficiencies of visualization in the KC system; finding an efficient way to view the contents of the articles at a glance would make the system more efficient.

Another problem is that some articles were excluded when the textual information was entered into the KC system. Data input was carried out by uploading articles that included mentions of municipalities such as cities, towns, and villages in the Hanshin-Awaji area. Thus, articles that did not mention such expressions did not appear in the KC system and were not included in the analysis. Volumes of articles, in fact, mentioned the names of buildings and natural attractions, such as highways, the city hall, or mountains but did not include the names of cities, towns, and villages. This type of textual information is also an important reference in addition to those with the names of municipalities.

4. Discussions for Future Research: Towards an Archiving System that is Sustainably Beneficial for Society

4.1 Pursuing the Improvement of the Visualization and Archiving with the KACHINA CUBE

The present study presented the information database on the Great Hanshin-Awaji Earthquake in three-dimensional space using the KC system, showing prospects of visualizing textual information and offering a way to access key information. Concretely, the textual information of the database was organized according to timeline and geographical space. The visualization of textual information in 3D space can still improve with regard to accessing key information.

Three main aspects of the advantages of visualizing the information database using the KC system were discussed above. First, the system provided viewers a visual picture of the large amount of textual information. Viewers can easily obtain the textual information at a glance, on the geographical map, over different time Periods, in contrast to the more cumbersome use of piles of documents contained in the information database.

Visualizing images on 3D space also presents the quantitative changes over time, in relevant articles found in particular places. As a result, a new hypothesis or perspective was gained with regard to the phenomenon. Finally, the functions to rotate and search articles in the KC, and to access original articles in each fragment, are useful in exploring further the inputs from the information database. Therefore, viewers of the KC are given more opportunities to appreciate the phenomenon resulting from the earthquake.

Whereas the system used in this study showed positive prospects in terms of visualizing textual information, certain problems need to be addressed. One critical problem is that the topics of the articles included in each information fragment are not provided in one view. The KC system used in the current study only presented viewers with the number of articles regarding a particular place in each period on the 3D space. Viewers needed to click in order to explore each information fragment, although the system was equipped with the functions of searching and accessing original sentences in the information fragments. One way to solve the problem is by coding the contents of the information fragments and presenting these in color. Colors were used in this study to classify information fragments according to time period. Similarly, if the content of the information fragments were classified and presented through colors, then the types of information available can also be identified at a glance.

Also a critical problem is the exclusion of a large amount of articles during the process of data input. These articles may have contained significant information excluded in the analysis of the present study. Such articles mentioned the names of places but not the actual names of municipalities such as cities, towns, and villages; instead, more general, abstracted or local expressions were used, such as the names of prefectures, areas, buildings, or city hall. For visualization to be more systematic and comprehensive, a method of recognizing expressions of places other than municipality names is required. Thus, the invention of a system that translates the expressions of places into geographical information would be ideal.

As regards the fact that not all the information in the database was entered into the KC system in the present study, a positive perspective for effectively establishing archives is achieved. To solve the problem above, we recommend that the textual information that will be archived and placed in the database should include particular expressions, such as the name of municipalities. The type of expressions to be included can be finalized after discussions among experts or other persons involved in particular areas. For further studies on the use of the KC system at least, using the names of municipalities is recommended. If all the articles contained the name of municipalities, then no textual information would be excluded in the process of uploading the data into the KC system. Accordingly, all the textual information stored in archives or databases can be placed on a geographical map.

In line with the suggestion above, we propose the use of more concrete expressions of time in articles for archiving. The database used in the present study classified articles into four periods only. Placing the event described in each article on a continuous time line, or organizing articles by year of publication would be beneficial. As a result, viewers can conduct a more detailed analysis through the visualization.

4.2 For Future Studies

An archiving system that is limited to collecting and storing information is not enough. To attain sustainably of such system, the information contained in archives should be classified and organized in a manner that is clear to viewers. However, classifying and organizing textual information manually is time consuming. Automatic and systematic input and classification of data is preferred. In future studies, such automatic and systematic method for archiving vast information can be achieved through digital technology which requires less manpower to input, classify, and organize information in the database.

Improving the visualization of vast textual information using the KC system, which is the goal of this present study, benefits citizens in general. Considering the possibility of providing

users with a new perspective on a particular phenomenon through visualization using the KC system, academic scholars, experts, and officers analyzing the Great East Japan Earthquake of 2011 can employ the same system. Such can be used in conjunction with findings from other studies employing qualitative GIS (e.g., Kwan and Ding, 2008) and other relevant methods.

In addition to the discussions above, how viewers use the KC system and how they acquire or gain knowledge from it should be analyzed. How instructions on the use of the KC system will be given is likewise necessary. Therefore, one of the next issues that we intend to address involves testing whether users clearly understand how to access key information from the KC system, thus exploring effective ways for them to use the system.

References

Archive project of Great East Japan Earthquake by Tohoku University. Retrieved October 5, 2011, from http://www.tohoku.ac.jp/japanese/newimg/pressimg/tohokuuniv-press20110912.pdf

Cabinet Office, Government of Japan. (2006). Information Database of the Great Hanshin-Awaji Earthquake. Retrieved October 5, 2011, from http://www.bousai.go.jp/1info/kyoukun/hanshin_awaji/index.html

Cabinet Office, Government of Japan. Web Page of Non-Profit Organization. Retrieved 10, 2012, from https://www.npo-homepage.go.jp/index.html

City of Kobe. (2010). *Comprehensive Strategy for Recovery from the Great Hanshin-Awaji Earthquake*. Digital Archive of Great East Japan Earthquake 2011 by Reischauer Institute of Japanese Studies. Retrieved October 5, 2011, from http://www.jdarchive.org/?la=en

Ibusuki, M. (2011). *Informational Justice*. Paper Presented at the 12th Annual Conference of Japanese Association of Law and Psychology, Nagoya: Japan.

Kwan, M. & Ding, G. (2008). Geo-narrative: Extending Geographic Information Systems for Narrative Analysis in Qualitative and Mixed-method Research. *The Professional Geographer*, *60*, 443-465.

Ohno, S., Saito, S. & Inaba, M. (2010). A Platform for Mining and Visualizing Regional Collective Culture. In T. Ishida (Ed.), *Culture and Computing*, LNCS (*Lecture Notes in Computer Science*) *6259* (189-199). Berlin: Springer.

Saito, S., Ohno, S. & Inaba, M. (2009). A Platform for Visualizing and Sharing Collective Cultural Information. Paper Presented at the Proceedings of International Conference Digital Archives and Digital Humanities, Dec. 1-2, 2009, National Taiwan University, Taipei, Taiwan.

Saito, S., Ohno, S. & Inaba, M. (2010). A Platform for Cultural Information Visualization Using Schematic Expressions of Cube. *Digital Humanities Society, the Proceedings of Digital Humanities 2010 Conference*, 365-368.

「太平洋史前Lapita陶器線上數位資料庫」的建立與運用

邱斯嘉*、郭潔**、蘇郁尹***

摘 要

經過六年來的努力，「太平洋史前Lapita陶器線上數位資料庫」已然成為當今擁有研究太平洋史前Lapita陶器最完整資料來源的資料庫，是全球第一個建立線上數位資料庫以供各國研究人員互相比較各地之Lapita陶器收藏，以統計方法來互相對照、比較遺址之間的研究工具，其中的大筆收藏也將成為日後各國研究人員在研發相關資料庫時必須參考的基準點。此計畫已奠定臺灣在南島語族史前Lapita物質文化研究的數位典藏上的重要地位，日後也將吸引更多國外考古研究人員與臺灣考古學界合作，更進一步促成及推動跨國與跨領域的太平洋史前Lapita陶器研究。本文將著重於介紹本資料庫的建立理念、跨國合作成果，以及資料庫功能，並以資料庫利用GIS呈現Lapita陶器紋飾在各島群的分布狀況為例，簡單說明此一資料庫的應用方式及其對後續研究工作的助益。

* 中央研究院人文社會科學研究中心考古學研究專題中心副研究員。
** 中央研究院人文社會科學研究中心考古學研究專題中心約聘助理。
*** 中央研究院人文社會科學研究中心考古學研究專題中心勞務承攬人員。

Establishment and Research Applications of the Online Database for the Study of Lapita Pottery

Scarlett Chiu *, Chiech Kuo **, Yu-yin Su ***

Abstract

The Lapita Pottery Online Database is now the leading source of information for researchers and the general population to learn more about this particular pottery tradition. It is the first database established with the vision of providing online storage management and research tools that will help investigators locate and compare Lapita pottery excavated from seven countries in the Pacific. It also contains standardized recode themes that serve as platforms for further comparisons. This database has established itself as a great source for the research of Lapita pottery, a means to establish and maintain international collaborations among scholars, and a useful e-learning tool for Pacific islanders.

This paper will focus on describing the research collaborations achieved so far for establishing this Lapita Pottery Online Database, its basic functions, and future plan for promoting the usage and research applications. A brief summary of what has been accomplished in the past six years will be presented in the second section, while future research questions will be proposed in the last section.

* Associate Research Fellow, Center for Archaeological Studies, Research Center for Humanities and Social Sciences, Academia Sinica.

** Research Assistant, Center for Archaeological Studies, Research Center for Humanities and Social Sciences, Academia Sinica.

*** Contract employee, Center for Archaeological Studies, Research Center for Humanities and Social Sciences, Academia Sinica.

一、引言

自歐洲大航海時代以來，大洋洲的南島文化傳統是如何在不同的島嶼上產生與演變的，人與陸地資源稀少的自然環境之間是如何互相適應的，人群之間如何能夠同時保有其特色、又維持長時間及長距離的交換系統來獲取在島嶼上生存所需的物資及婚配伴侶，整個大洋洲裡為何會出現從平權到高度階級化等種種不同的社會組織，都是大洋洲考古及人類學所專注的議題。在短短三百年間，史前Lapita文化從巴布亞新幾內亞的俾斯麥群島，急速擴散到涵蓋巴布亞新幾內亞、索羅門群島、萬那度、法屬新喀里多尼亞、斐濟、東加及薩摩亞一帶的廣大範圍，至今已有超過200個遺址出土（Anderson et al., 2001; Bedford & Sand, 2007）。其中帶有繁複紋飾的梳點壓印紋陶，長久以來便是大洋洲的考古學者賴以建立相對年代關係、社群分類過程的重要項目之一（例見：Allen, Gosden & White, 1989; Anson, 1986; Bellwood, 1987; Best, 1984; Burley, Storey & Witt, 2002; Chiu, 2007; Gifford & Shutler, 1956; Golson, 1971, 1972b; Green, 1978, 1979, 1990, 2003; Hunt, 1988; Kirch, 1984, 1997; Mead, Birk, Birks & Shaw, 1975; Sand, 2000, 2001; Sharp, 1988; Specht, 2007; Spriggs, 1984, 1990; Summerhayes, 2000a, 2001a, 2001b）。由於Lapita文化與史前南島語族擴散過程緊密相關，探討太平洋史前Lapita文化與臺灣及東南亞史前文化間的關聯性也成為過去五十年來的研究重點（Bellwood, 1979; Chiu & Sand, 2007; Golson, 1972a; Hung et al., 2011; Pawley, 2007; Shutler & Marck, 1975; Spriggs, 2007），而這當中許多的資訊都必須倚賴從考古遺址當中所出土的物質文化遺留來提供線索。透過了具有特殊裝飾風格的Lapita陶器在技術、風格、形狀等方面的演變，搭配時空的資訊，便成為討論西南大洋洲史前文化流傳模式上的重要方法之一（邱斯嘉，2011）。

二、Lapita線上資料庫的建置過程

「太平洋史前Lapita陶器線上數位資料庫」主要是為協助考古學家處理來自西南太平洋不同國家、為數龐大的考古標本所建置的（邱斯嘉，2011）。所處理的材料，並非只針對博物館的展示精品，而是為保存考古資訊、提供線上學習資料，及開發未來研究議題而設立的。針對太平洋史前Lapita文化叢的陶片及相關遺址文物，製作同時便利專家「研究」以及一般大眾「學習」的線上資料庫，不僅可將資源更方便的開放給學者專家，也能增進民眾對於世界不同地區文化的相互比較與理解。因而在選取材料上便不能夠只選取少量精品來進行數位化的工作，而是著重於詳實且精確的描述與記錄考古標本及其出土脈絡資訊，在累積一定數量後方能提供學者做進一步的研究。透過與各國學者的密切合作，逐步新增更多的資料供參與的學者共同使用，就可以在擴大資料基礎的同時，也使資料庫能夠在各國研究人員的齊力貢獻下，永續經營下去。因此前幾年的首要目標都是放在透過國際合作蒐集重

要遺址資料，並針對陶器的各項特徵將資料標準化，且訂定標準化作業流程來掌控資料本身的精確度，輔以資料庫系統的設計與逐步改善其使用功能，用意是在建置一完善的系統，並提供可信度極高的資料，供各國研究者評估此一資料庫的可行性與可用性，進而吸引更多人投入合作建置（見圖1）。

為詳細記錄現今研究陶器傳統時所常用的陶器特徵，以及提供使用者文字與視覺上的資訊，在陶片資料的處理程序當中，首先是實地復原陶器並測量陶片的尺寸、重量等，以取得確切的器型。接下來在工作室裡將照片檔案標準化，加上浮水印，以便網站搜尋使用。第三部分的工作，則是以人工分析陶片上的紋飾，以23個特徵加以分類，註記到對應陶器器壁區塊，且必須要再新增新紋飾的定義及繪製復原圖等，以供查詢與比對之用。第四部分的資料，則是來自分析陶器的岩象與化學分析數據。最後，整批資料必須比對回遺址的發掘文獻資料，然後輔以統計方法來探究不同的研究議題，才能重建時空下的變化，成為後續研究的基礎材料。其中也包含了校準地圖、確認遺址所在位置、建構器型與紋飾分類標準等步驟，在多位學者的共同努力下才建置出這套資料庫的系統（邱斯嘉，2011）。

自2006年起，本資料庫便將出土於新喀里多尼亞（New Caledonia）、巴布亞新幾內亞（Papua New Guinea）和索羅門群島（Solomon Island）等地區的陶片，以分期分群的方式進行整理與登錄的工作。將上述島群中的10個重要大型遺址（Vatcha、

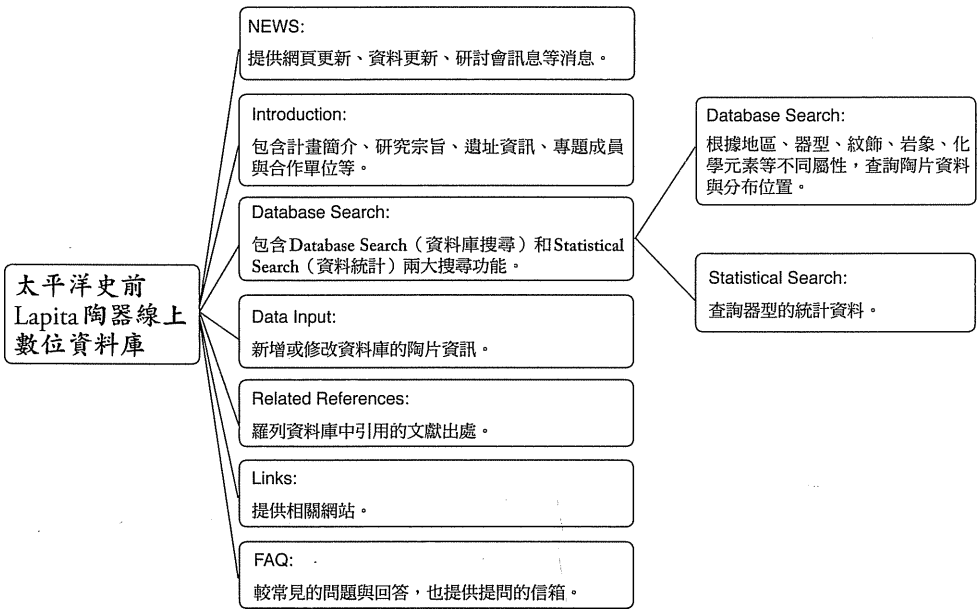


圖1 資料庫架構圖

13A、Nessadiou、Vavouto、Goro、Kamgot、RL2、RL6、SZ8和SZ10）所挖掘的陶片及其遺址相關資訊匯入「太平洋史前Lapita陶器線上數位資料庫」中，並根據材料製作出相關的63個後設資料定義檔。目前，本資料庫已是全世界擁有最多Lapita陶器資料的典藏及搜尋系統，其中囊括了來自四大地理區219個遺址之資訊，32,558張陶片標本的照片，並且制定出26大類76種陶器類型（Vessel Type）和3,916個紋飾（motif）圖樣的分類標準，持續架構出太平洋史前Lapita文化叢的面貌（邱斯嘉，2011）。

由於時間、人力，以及經費上等等的限制，在蒐集資料的考量上，共分為以下兩種方式來處理：

（一）出自大洋洲七個國家裡的重要考古遺址發掘，目前典藏於各地博物館、大學之實驗室的陶器實物

此類標本因出自重要遺址，具有指標性質，能夠吸引世界各地研究人員一同來參與往後的資料建置，因此本資料庫在剛起步時，與各國合作學者商量進行此一資料庫的建置過程中，便首先針對16個重要遺址進行資料的蒐集與整理。此類標本需要透過國際合作取得研究許可之後，再經過研究人員的整理、進行陶器復原等措施，才能提供詳細的標本數量。然而實際上大多典藏單位並未能掌握詳細數字，大多數的典藏品也尚待修復，且由於每張原始相片當中會包含數量不等的陶片，在在影響到本資料庫將蒐集的預估數字與實際處理的數字。經過六年來的努力，現已整理完其中10個遺址共40,732張數位化圖片，及其所對應的遺址資訊、陶片的形式、測量、部分紋飾分析等資訊。目前尚有6個遺址約61,243片陶片尚待處理。其中萬那度（Vanuatu）的三個遺址（Makué、Teouma、Vao）共計2,376張原始照片，是預定在今年要進行整理的部分。另外三個遺址估計約有30,000張原始照片（約58,867片陶片）的資料，則希望可在明後年的計畫中完成。在各國研究者的文章發表過後，這些資料也將逐步開放給大眾使用，提供電子化學習的內容（見表1）。

表1中的第七位，加拿大Simon Fraser University的David Burley教授及其研究團隊，是現在第一位公開表示將會在其論文發表後，直接把資料上傳給本資料庫的合作學者。因此這部分的資料將來只需要後設資料組工作人員確認已標準化，並轉換其分類代碼即可公開上線。未來本資料庫希望盡量吸引這類的合作對象，在他們各自的資料登錄過程中自行進行標準化，在發表了遺址相關論文之後，便將其中陶片資料上傳給本資料庫，然後就可公開上線以利研究資訊之迅速傳播及線上教學使用。

表1 資料庫現有之資料與資料輸入情況

合作學者	已取得之遺址資料		資料輸入情況		
	島群	遺址	目錄屬性	照片處理	紋飾登錄
Prof. Roger Green, University of Auckland, New Zealand	Solomon Islands	RL2	●	●	◎
		RL6	●	●	◎
		SZ8	●	●	◎
		SZ10	●	●	◎
Prof. Glenn Summerhayes, University of Otago, New Zealand	Papua New Guinea	Kamgot	●	●	◎
Dr. Christophe Sand, Institute of Archaeology of New Caledonia and the Pacific, New Caledonia	New Caledonia	13A	●	●	◎
		Vavouto	●	●	○
		Goro	●	●	○
		Nessadiou	●	●	○
		Vatcha	●	●	○
		Kurin	●	●	○
Prof. Patrick Kirch, University of California at Berkeley, U.S.A.	Papua New Guinea	ECA	◎	◎	◎
		ECB	○	◎	○
Prof. Matthew Spriggs and Dr. Stuart Bedford, Australian National University, Australia	Vanuatu	Teouma	○	◎	○
		Vao	○	◎	○
Prof. Jim Specht, Australian National Museum, Australia	Papua New Guinea	Watom	○	◎	○
Prof. David Burley, Simon Fraser University, Canada	Fiji	Vorovoro	○	○	○

●已完成；◎部分完成；○待製作

（二）尚未被收納入資料庫的其他遺址材料

由於本資料庫是當今世界上針對Lapita陶器研究最完善的資料庫，其中的資料又包含了所有重要指標性遺址，資料的分類與建置也一併標準化，提供未來大規模對比研究的紮實基礎。日後所有相關研究勢必需要引用本資料庫的資訊做地域性的對比與演繹，才能確保其研究成果不致失誤。這些未被列入重要遺址所出土的陶片資料，將來也會盡量以合作方式透過各國的研究者自行整理上傳，以換取運用本資料庫內容的特許權利。往後本資料庫的工作性質也將會隨之轉變為提供更優良的後設資料服務，確認並新增各地傳來的新器型或紋飾資料，控管資料上傳與下載等特許功能給不同性質的使用者。

只有透過提供重要研究相關資訊及服務，才能同時吸引相關研究族群成為忠實的使用者，也能夠逐步形成口碑且成為此一資料庫的推廣者，甚至是提供新資料的合作者。在形成良性回饋機制之後，本資料庫便能夠及時提供更多有用的資訊，成為研究大洋洲史前史不可或缺的工具，達到本身就可以永續經營並推廣南島文化資產的目的。而線上學習，特別是對教育及交通資源稀少的大洋洲島民而言，是珍貴的學習與取得研究材料的機會，也是促進臺灣與此區人民彼此文化交流的方式。在表2列出自2010年7月資料庫上線以來的使用者記錄，可以看出排名前十名的使用者所在國家，大多為與大洋洲相關研究領域關係密切的國家，而同一使用者造訪50次以上的數據，則顯示出長期使用本資料庫的習慣。

由於本資料庫的性質特殊，其中所蒐集的資料，主要是透過與各國學者的密切合作，在各國研究人員的齊力貢獻下才能永續經營下去。而唯有透過資料本身的可信度及廣度、搭配資料庫及網頁的使用流暢程度，才能持續吸引更多潛在的合作者投入此一資料庫的建置當中。因此，為了增強資料庫內容的廣度及使用的深度，在接下來的工作中，除了在國際學術會議上的推動，也已規劃透過提供資料庫中英文操作手冊、影片說明、圖像呈現等方式來吸引閱覽人數，增強社會教育與文化交流方面的功用。同時也規劃透過社群網站以及電子報等功能，加強與使用者之間的對話，激發潛在的使用者，逐步凝聚研究社群。

三、Lapita線上資料庫的功能簡介

在逐年進行相關整理研究擴增及更新資料庫內容與系統之後，本資料庫的功能可以概分為幾大類。第一是妥善保存太平洋史前文化Lapita陶器各項相關文獻與標本數位記錄資料；第二是將以上的資料搭配GIS及Google Map功能，利用地理位置

表2 使用者記錄表

此訪客的造訪次數（包含當次造訪）	該訪客的第N次造訪	佔造訪總數的百分比
1-50次	3,015	70.16%
51-100次	494	11.49%
101-200次	405	9.42%
201+次	384	8.93%

排名	國家／領域	造訪	單次造訪	平均網站停留時間	%新造訪
1	Taiwan	3,739	3.77	00:05:39	19.60%
2	United States	141	4.11	00:02:53	62.41%
3	New Zealand	88	3.80	00:02:13	61.36%
4	Australia	53	3.64	00:02:28	79.25%
5	Canada	29	5.97	00:09:45	62.07%
6	China	25	2.84	00:01:21	92.00%
7	France	21	3.95	00:03:09	100.00%
8	United Kingdom	20	3.80	00:01:16	70.00%
9	Germany	17	4.12	00:02:16	100.00%
10	Japan	14	5.93	00:04:16	71.43%

來查詢及檢視陶器分布資訊（見圖2）；第三則是設計出兼具裝飾主題模型搜尋與裝飾排列之比較功能，以方便研究者同時檢視多種陶器特徵來詳細對比各島群所出土的考古資料的異同之處。同時，也提供各項數據之統計結果的網上呈現與下載功能，有效管理並開放加值利用數位化資料，以作為相關學界的學術交流橋樑。

此資料庫已於2010年7月29日正式開放使用者註冊，使用者可藉由基礎或是進階表單，以遺址、器型與陶器裝飾母題等屬性，或是配合資料庫地圖視窗進行Lapita陶器搜尋。其中，為讓使用者更容易藉由裝飾母題複雜之分類進行陶片檢索，陶器裝飾母題檢索頁面規劃設計為選單勾選介面，選單內容則含括裝飾母題定義、規則與裝飾母題種類等主要分類，設計出不同單複選選項供使用者依據其研究主題來自由運用。

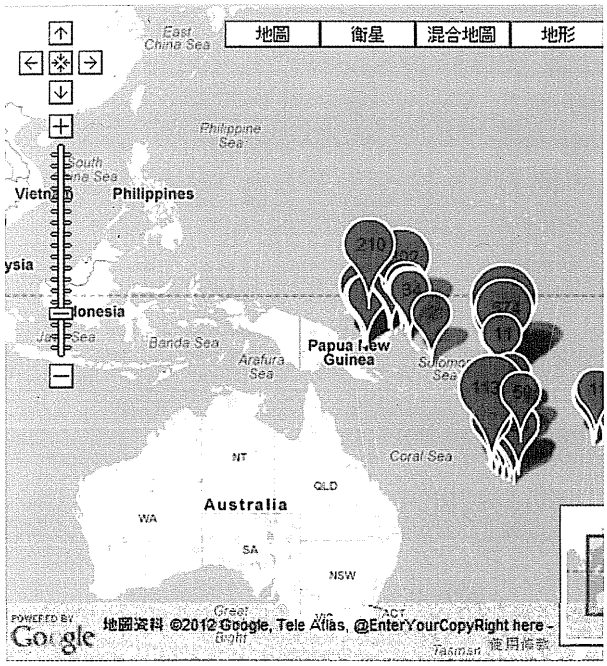


圖2 利用結合於資料庫中的Google Map顯示出現有的資料量及資料來源分布

四、資料庫應用範例——Lapita紋飾分布之分析

以下的資料出處為從實際考古資料以及文獻當中所整理出來的，共有4,195種不同的紋飾登錄在資料庫中，其中包括了15個巴布亞新幾內亞的遺址、11個索羅門群島的遺址、5個萬那度的遺址、10個新喀里多尼亞的遺址、5個斐濟的遺址、9個東加的遺址，以及1個薩摩亞的遺址（Anson, 1983, 2000; Bedford, 2006; Bedford & Spriggs, 2007; Bedford et al., 2007; Best, 1984, 2002; Birks & Birks, 1975; Doherty, 2009; Donovan, 1973; Frimigacci, 1977; Galipaud, 1988; Gifford & Shutler, 1956; Gorecki, Head & Bassett, 1991; Green, 1976, 1990; Hedrick, N.D.; Kay, 1984; Kirch, 1988; Lilley, 1991; Mead, 1975; Parker, 1981; Poulsen, 1987; Sand, 1996, 1997, 2001, 2007, 2010; Sand, Bolé & Ouetcho, 2002; Shaw, 1975; Jim Specht, 1991; Specht & Attenbrow, 2007; Spriggs, 1990; Summerhayes, 2000b; Wickler, 2001）。其中的570種分布在一個以上的遺址內，最受歡迎的紋飾是A1這種簡單幾何紋的圖樣（見圖3），在56個遺址中出現過24次（見表3）。

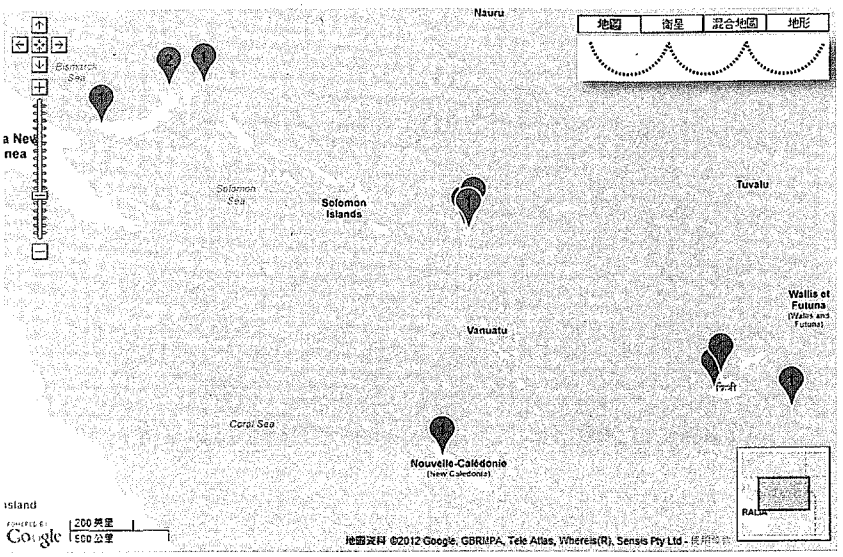


圖3 以A1裝飾紋路的分布圖為例

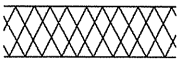
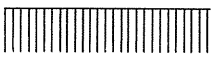
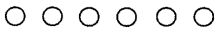
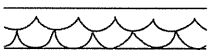
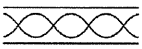
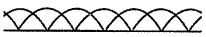
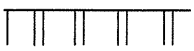
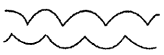
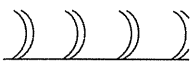
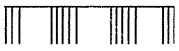
表3列出出現過10次以上的紋飾圖案。然而這些受歡迎的紋飾，在Lapita文化叢最晚期，也是最東邊的薩摩亞遺址中一個都沒有出現過。薩摩亞遺址中只有兩個紋飾是與其他遺址相同的，而且都集中在Lapita東區的斐濟與東加的遺址當中。這說明了此兩種紋飾屬於晚期在Lapita東區自行發展出來的，而且很可能屬於在Lapita東區已經與西邊的其他區域斷絕往來後才發展出來的地方性紋飾。

在所有出現過5次以上的紋飾中，A275（8次）、A167（6次）以及A120、A241、A306、A439（各5次）都只出現在Lapita遠西區巴布亞新幾內亞的遺址中，A5（5次）則只出現在Lapita西區索羅門群島的遺址中（見表4）。

從技術方面來說，這4,195種紋飾，絕大多數都利用到精細梳點壓印紋（3,170種，佔75.57％強），刻劃紋次之（695種，佔16.57％強），而圓管狀壓印紋再次之（552種，佔13.16％強），晚期較為多見的粗糙梳點壓印紋則只佔7.2％強（見圖4）。其餘十數種裝飾技術，例如透雕、貝印紋等，在整個技術層面來說最多佔0.2％強，表明了當時的陶匠在裝飾陶器方面具有強烈的使用梳點壓印紋的取向，但是也在極少數的例子當中使用新的技巧（見表5）。

從定義方面來說，4,195種紋飾中有3,714種可以看出它們分別由哪些基本裝飾單位架構而成。這108種不同的裝飾單位中，利用弧線原型（original）所組成的紋飾數量最多（約25.90％），頂點在上的三角形（triangle point up）次之（約7.03％）（見表6）。由於利用弧形這樣的梳點壓印紋工具在島嶼東南亞的遺址中尚未出現，

表3 在56個遺址中出現過10次以上的紋飾

Motif_ID	紋飾	出現次數
A1		24
A237		23
A448		22
A417		20
A18		19
A37		19
A497		19
A35		18
A441		18
A2		16
A421		16
A53		16
A435		15
A496		15
A206		14
A444		14

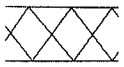
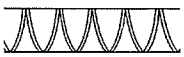
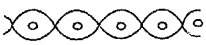
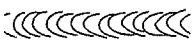
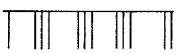
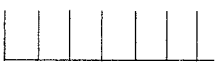
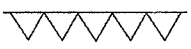
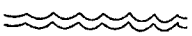
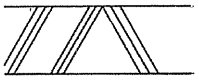
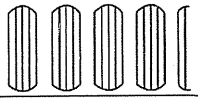
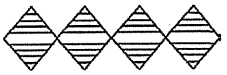
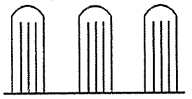
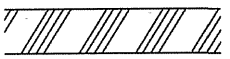
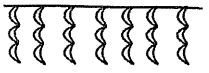
Motif_ID	紋飾	出現次數
A231		13
A436		13
A6		13
A207-2		12
A333-2		12
A44		12
A499		12
A260		11
A442		11
A73		11
M77-1-1		11
A162		10
A207-1		10
A3		10
A55		10

表4 只在特定區域中流行的紋飾

Motif_ID	紋飾	地區
A275		巴布亞新幾內亞
A167		
A120		
A241		
A306		
A439		
A5		索羅門群島

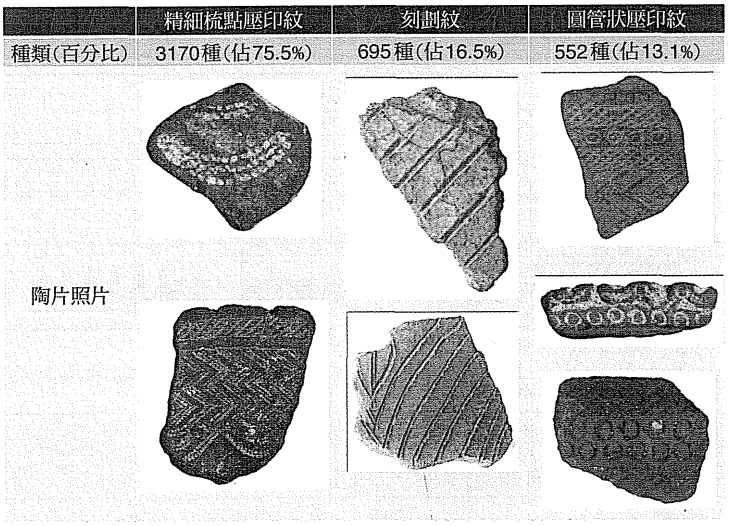


圖4 前三大類的裝飾技術

表5 製作紋飾的技術種類及使用次數

編號	類型	使用數量	百分比
1	fine dentate-stamping 精細梳點壓印紋	3,170	75.57%
3	incising 刻劃紋	695	16.57%
24	impressed with a circular tube 圓管狀壓印紋	552	13.16%
2	coarse dentate-stamping 粗糙梳點壓印紋	302	7.20%
23	carving 浮雕	86	2.05%
5	shell impressed 貝印紋	57	1.36%
21	rockening fine-dentate 精細梳點滾壓紋	30	0.72%
9	applique 貼花	23	0.55%
4	cut-outs 透雕	17	0.41%
17	impressed dots 壓印點紋	16	0.38%
20	rockening 滾壓紋	15	0.36%
12	notching（V-shape）V型凹痕	13	0.31%
8	paddle-impressed 拍打棒的拍打紋	12	0.29%
7	notching（U-shape）U型凹痕	11	0.26%
10	pierced 刺穿	9	0.21%
11	finger nail impressed 指甲印紋	7	0.17%
22	rockening shell 貝殼滾壓紋	7	0.17%
15	line impressed 壓印實心線	5	0.12%
29	rockening coarse-dentate 粗糙梳點滾壓紋	4	0.10%
13	crenate（finger pinched）手指捏夾紋	2	0.05%
27	rockening with sharp objectv 以尖物滾壓	2	0.05%
16	triangle impressed 壓印實心三角型	1	0.02%
28	impressed small circle 壓印圈紋	1	0.02%
25	rectangle impressed 壓印正方形	1	0.02%

註：百分比是以使用量除以可分析的技術總數（4,195）所計算出。

表6 裝飾單位（Design Unit）的種類及使用次數

編號	類型	使用數量	百分比
14	original 弧線原型	962	25.90%
8	triangle point up 頂點在上的三角形	261	7.03%
6	quadangle 四邊形	213	5.74%
24	circle 圓	163	4.39%
9	triangle point down 頂點在下的三角形	162	4.36%
59	spear head 單箭頭狀	158	4.25%
1	ㄇ-shape ㄇ形	150	4.04%
60	thigh 大腿骨狀	150	4.04%
15	one point attached 一點相交弧線	112	3.02%
80	jelly fish 水母狀	89	2.40%

註：百分比是以使用量除以可分析的裝飾單位總數（3,714）所計算出。

因此這種在Lapita陶器中佔絕大多數的基本裝飾單位代表了Lapita陶匠在工具與設計上的創新（Ambrose, 2007; Bedford & Sand, 2007: 3; Sand, 2007）。

從製作紋飾的規則方面來說，這4,195種紋飾中有3,160種可以看出它們利用了哪些基本走向來架構紋飾。19種基本走向中，單方向平行移動（horizontal translation）最受歡迎（約70.25%），單方向垂直移動（vertical translation）次之（約4.24%）（見表7）。這種高度專注於某一特殊基本走向的性質，反映出即使是處於不同島嶼的陶匠，彼此之間確實遵行著一定的心理模式規律來製作紋飾圖形，且此一後天習得之規律性很少受到地理區域或是時間先後的影響，因此Lapita陶器的製作的確能反映出文化偏好。這樣的結果，將與Sharp（1988）所觀察到的現象做更深入的比對研究。

在4,195種紋飾中，共有1,683種屬於43大類的複合圖樣，其餘則是屬於簡單幾何紋路。複合圖樣中簡化人臉面（simple face）佔最多數（約5.35%），長鼻臉面（long-nose face with headdress）次之（約3.74%）（見表8）。圖5則是所有簡單臉面紋飾的地理分布圖。

從以上的資料綜論來說，未來研究的方向大致上可以從不同地區的偏好做更為細緻的區分，利用統計的群集分析等方法，來辨識各個遺址所獨有的紋飾製作特性開始著手進行研究。從現有的基礎資料當中已經可以找出在不同島嶼廣為流傳的幾何形紋飾，也可以看出不同地區各自所偏好或是獨有的紋飾，同時印證了梳點壓

表7 基本走向（Basic Direction）的種類及使用次數

編號	類型	使用數量	百分比
1	horizontal translation 單方向平行移動	2,220	70.25%
3	vertical translation 單方向垂直移動	134	4.24%
13	superimpose/overlapping 重疊	125	3.96%
5	horizonal reflection 水平鏡射	115	3.42%
9	horizontal glide translation 平行位移	108	3.64%
7	tilted 任意角度傾斜	103	3.26%
8	half turn 反轉	74	2.34%
19	interspersion 散布	67	2.12%
18	fold 摺	53	1.68%
16	expansion 擴散	35	1.11%
6	vertical reflection 垂直鏡射	28	0.89%
12	interlock 連鎖	16	0.89%
14	90 degree rotation 90度旋轉	15	0.66%
17	radiaton 放射	16	0.51%
15	linked 連結	13	0.47%
2	opposite directions-horizontal expansion 反方向平行移動	13	0.41%
11	concentric circle 同心圓輻射	3	0.09%
4	opposite directions-vertical expansion 反方向垂直移動	1	0.03%
10	vertical glide translation 垂直位移	1	0.03%

註：百分比是以使用量除以可分析的基本走向總數（3,160）所計算出。

表8　複合圖樣（Complex Motif）的十大種類及使用次數

編號	類型	使用數量	百分比
2	simple face　簡化臉面	90	5.35%
5	long-nose face with headdress　長鼻臉面	63	3.74%
40	straight band　直線帶狀	57	3.39%
3	triangular face　三角臉面	36	2.14%
45	vertical zigzag　垂直閃電狀	26	1.54%
26	curvature band　彎曲帶狀	19	1.13%
1	indeterminate　未命名	13	0.77%
6	odd face　怪臉面	10	0.59%
20	single undulated　單幅波浪狀	10	0.59%
27	double spear-heads　雙箭頭	10	0.59%

註：百分比是以使用量除以可分析的複合圖樣總數（1,683）所計算出。

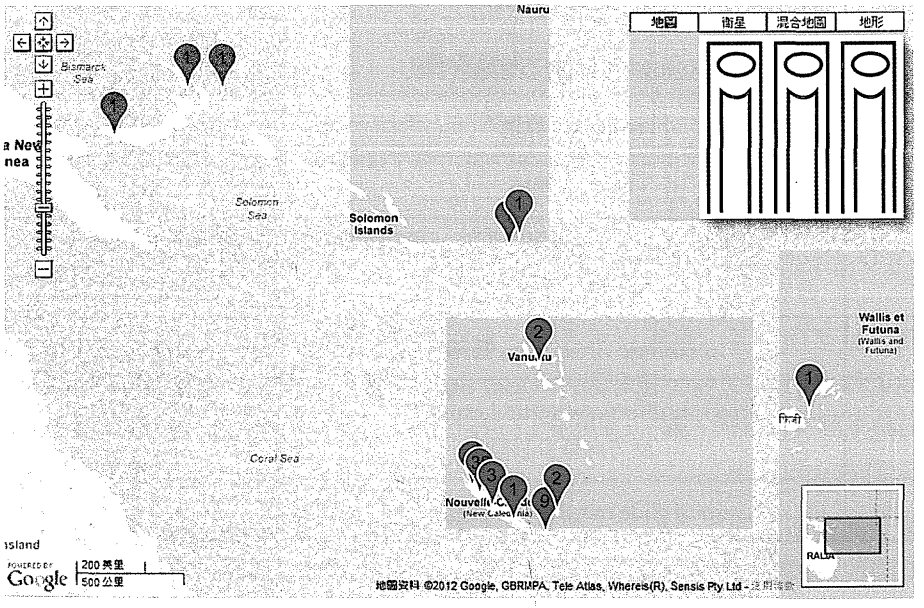


圖5　簡單臉面紋飾（simple face）分布圖

印紋、刻劃紋與圓管壓印紋是組成絕大部分Lapita紋飾的重要裝飾元素。這三者間的互相搭配原則是否會因地域或時間的不同而有所差別，用這些技術製作出來的基本裝飾單位（如直線或是弧形的梳點壓印紋）在不同地區出現的時間早晚與所佔比例，也是值得進一步探究的線索。而絕大多數的紋飾是以單方向平行移動法來建構的，其他紋飾建構方式是否可以反映出不同地區的偏好，甚至指出不同陶器的流傳方向，則有賴進一步的調查。

以往無法統整資料的弊病現在藉由文字與影像的雙重呈現方式得以解決，日後可以開始著手重新檢視原有的Lapita文化分區標準，也能夠更為細緻地處理各文化分區本身的陶器傳統演變史，尋找出各分區之間的互動過程（邱斯嘉，2011）。而Lapita陶器傳統與密克羅尼西亞以及島嶼東南亞之間的關聯到底有多密切（Hung et al., 2011），也可以經由大範圍的搜尋而取得可信度更高的證據。往後再搭配上相關遺址其他出土的器物種類與密集度，便能進一步討論不同紋飾的分布狀況是否與遺址本身的生業經濟能力強弱、與其他遺址之間的往來交通便利性，或是其他社會因素有關（Bedford & Sand, 2007: 5；邱斯嘉、郭潔、蘇郁尹，2011）。

透過利用資料庫所提供的器型與紋飾結合搜尋功能，未來也可以開始著手進行研究在同一時間內、不同島嶼出現的不同類型陶器是否會施加相類似紋飾，或是配合陶器在遺址出土脈絡資訊，來討論不同功用的陶器是否會呈現出不一樣的紋飾與器型特徵。這些特徵，也可以進一步透過陶器的岩象與化學成分分析結果來探究陶器的流傳或交換模式。這些資訊都能夠促進研究者對於Lapita文化叢本身的理解，以及其與周遭不同地區間的互動過程（Chiu, 2011: 56-58；邱斯嘉，2011：142）。

五、結語

運用線上資料庫可以容納大量數位典藏材料的便利性及即時性，來協助考古學家處理龐大複雜的考古標本，使得發掘出土的考古標本，連同發掘記錄、遺址地理資訊、相關影音影像記錄等，得以統整，並提供大眾線上學習與參考的資料；也能幫助處於不同地區的考古學家們可更迅速的分享及交流重要發現，使典藏的考古資料得以活化起來，重複被使用，以促進學術研究的發展，以及作為相關各國在列冊保護重要遺址時的重要依據。

由於本資料庫幾乎已經包含了所有Lapita文化的重要指標性遺址，日後相關研究勢必需要引用本資料庫的資訊做地域性的對比與演繹，才能確保其研究成果不致失誤。其中所開發出來的陶器器型紋飾多重選擇功能，在多場國際會議當中已經獲得眾多稱讚，並逐漸建立起在學術圈內的聲譽。往後本資料庫的性質也將隨之轉變為提供更優良的後設資料服務，確認並新增各地傳來的新器型或紋飾資料，控管資

料上傳與下載等特許功能給不同性質的使用者等。然由於「太平洋史前Lapita陶器線上數位資料庫」仍在建置與改善之中，現在大多數應用本資料庫的學者多是以查考資料的文獻出處、取得影像，以及利用資料庫內建的Google Map來查看某種陶器特徵的分布位置為主，而據此開發的深入研究課題則尚未成形。因此本文主要的目的在於介紹此資料庫給國際及國內相關研究者，提供他們在建置資料庫或是進行相關研究時的參考。

誌謝

中央研究院人文社會科學研究中心地理資訊中心廖泫銘研究助技師、李玉亭小姐、林農堯先生，本院計算中心林晰科長、林宗茂先生與本院資訊科學研究所方俞喬先生等也提供許多幫助，特此予以感謝。中央研究院人文社會科學研究中心考古學專題研究中心的王仁君小姐和陳天慧小姐幫忙製作圖表與排版，也一併致謝。感謝創意引晴有限公司在資料庫系統建置上的合作。本文的研究結果是來自於中央研究院人文社會科學研究中心考古學研究專題中心的「利用數位典藏改善學術研究環境計畫」及國家科學委員會（NSC96-2628-H-001-043）的經費支持。

參考文獻

邱斯嘉，2011，〈太平洋史前Lapita陶器：線上數位資料庫的建置與其對後期深化研究的效益〉（Research Impacts of Establishing an Online Database for the Study of Lapita Pottery），《臺大考古人類學刊》（*Journal of Archaeology and Anthropology*），75，頁123-158。

邱斯嘉、郭潔、蘇郁尹，2011，〈應用地理空間資訊系統分析臉面紋飾在太平洋史前Lapita文化叢不同文化區塊之分布情形〉（Analysing Face Motif Distribution Patterns among Different Cultural Provinces of the Lapita Culture Complex by Geographic Information System），賴東進主編，《2011數位典藏地理資訊研討會》，臺北：國立臺灣大學地理環境資源學系，頁39-58。

Allen, J., Gosden, C. & White, J. P. (1989). Human Pleistocene Adaptations in the Tropical Island Pacific: Recent Evidence from New Ireland, a Greater Australian Outlier. *Antiquity, 63*, 548-561.

Ambrose, W. (2007). The Implements of Lapita Ceramic Stamped Ornamentation. In S. Bedford, C. Sand & S. P. Connaughton (Eds.), *Oceanic Explorations: Lapita and Western Pacific Settlement* (pp. 213-222). Canberra: Australian National University.

Anderson, A., Bedford, S., Clark, G., Lilley, I., Sand, C., Summerhayes, G. & Torrence, R. (2001). An Inventory of Lapita Sites Containing Dentate-Stamped Pottery. In G. R. Clark, A. J. Anderson & T. Vunidilo (Eds.), *The Archaeology of Lapita Dispersal in Oceania: Papers from the Fourth Lapita Conference, June 2000, Canberra, Australia* (pp. 1-13). Canberra: Pandanus Books. Research School of Pacific and Asian Studies, Australian National University.

Anson, D. (1983). *Lapita Pottery of the Bismarck Archipelago and Its Affinities*. PhD thesis, University of Sydney, Sydney.

Anson, D. (1986). Lapita Pottery of the Bismarck Archipelago and Its Affinities. *Archaeology in Oceania, 21*, 157-165.

Anson, D. (2000). Reber-Rakival Dentate-stamped Motifs: Documentation and Comparative Implications. *New Zealand Journal of Archaeology, 20*(1998), 119-135.

Bedford, S. (2006). The Pacific's Earliest Painted Pottery: An Added Layer of Intrigue to the Lapita Debate and Beyond. *Antiquity, 80*(309), 544-557.

Bedford, S. & Sand, C. (2007). Lapita and Western Pacific Settlement: Progress, Prospects and

Persistent Problems. In S. Bedford, C. Sand & S. P. Connaughton (Eds.), *Oceanic Explorations: Lapita and Western Pacific Settlement* (pp. 1-16). Canberra: Australian National University.

Bedford, S. & Spriggs, M. (2007). Birds on the Rim: A Unique Lapita Carinated Vessel in Its Wider Context. *Archaeology in Oceania, 42*, 12-21.

Bedford, S., Spriggs, M., Regenvanu, R., Macgregor, C., Kuautonga, T. & Sietz, M. (2007). The Excavation, Conservation and Reconstruction of Lapita Burial Pots from the Teouma Site, Efate, Central Vanuatu. In S. Bedford, C. Sand & S. P. Connaughton (Eds.), *Oceanic Explorations: Lapita and Western Pacific Settlement* (pp. 223-240). Canberra: Australian National University.

Bellwood, P. S. (1979). *Man's Conquest of the Pacific: The Prehistory of Southeast Asia and Oceania*. New York: Oxford University Press.

Bellwood, P. S. (1987). *The Polynesians: Prehistory of an Island People* (Rev. ed.). London: Thames & Hudson.

Best, S. B. (1984). *Lakeba: The Prehistory of a Fijian Island*. PhD thesis, University of Auckland, Auckland.

Best, S. B. (2002). *Lapita: A View From The East*. Auckland: New Zealand Archaeological Association.

Birks, L. & Birks, H. (1975). Dentate-stamped Pottery from Sigatoka, Fiji. In S. M. Mead, L. Birks, H. Birks & E. Shaw (Eds.), *The Lapita Pottery Style of Fiji and Its Associations* (pp. 6-18). Wellington.

Burley, D. V., Storey, A. & Witt, J. (2002). On the Definition and Implications of Eastern Lapita Ceramics in Tonga. In S. Bedford, C. Sand & D. Burley (Eds.), *Fifty Years in the Field. Essays in Honour and Celebration of Richard Shutler Jr's Archaeological Career* (pp. 213-225). Auckland: New Zealand Archaeological Association.

Chiu, S. (2007). Detailed Analysis of Lapita Face Motifs: Case Studies from Reef/Santa Cruz Lapita Sites and New Caledonia Lapita Site 13A. In S. Bedford, C. Sand & S. P. Connaughton (Eds.), *Oceanic Explorations: Lapita and Western Pacific Settlement* (pp. 241-264). Canberra: ANU EPress.

Chiu, S. (2011). Lapita-scape: Research Possibilities Using the Digital Database of Lapita Pottery. *People and Culture in Oceania, 27*, 39-63.

Chiu, S. & Sand, C. (2007). From Southeast Asia to the Pacific: Historical and Theoretical Background to Austronesian Origins and to the Lapita Dispersal in Western Oceania. In S. Chiu & C. Sand (Eds.), *From Southeast Asia to the Pacific. Archaeological Perspectives on the Austronesian Expansion and the Lapita Cultural Complex* (東南亞到太平洋：從考古學證據看南島語族擴散與Lapita文化之間的關係) (pp. 27-48). Taipei (臺北): Center for Archaeological Studies, Research Center for Humanities and Social Sciences, Academia Sinica (中央研究院人文社會科學研究中心考古學研究專題中心).

Doherty, M. (2009). Post-Lapita Developments in the Reef-Santa Cruz Islands, Southeast Solomon Islands. In P. J. Sheppard, T. Thomas & G. R. Summerhayes (Eds.), *Lapita: Ancestors and Descendants* (pp. 181-213). Auckland: New Zealand Archaeological Association.

Donovan, L. J. (1973). *A Study of the Decorative System of the Lapita Potters in Reefs and Santa Cruz Islands.* Unpublished MA thesis, University of Auckland, Auckland.

Frimigacci, D. (1977). *Les Céramiques de Nouvelle-Calédonie,* 7(D. E. C. 101 p.). Nouméa: D.E.C., Bureau Psycho-Pédagogique.

Galipaud, J.-C. (1988). *La Poterie Prehistorique Neo-Caledonienne: Et Ses Implications Dans L'etude du Processus de Peuplement du Pacifique Occidental.* PhD thesis, Université de Paris I, Panthéon-Sorbonne, Paris.

Gifford, E. W. & Shutler, R., Jr. (1956). *Archaeological Excavations in New Caledonia, 18*(1). Berkeley and Los Angeles: University of California Press.

Golson, J. (1971). Lapita Ware and Its Transformations. In R. C. Green & M. Kelly (Eds.), *Studies in Oceanic Culture History, 2* (pp. 67-76). Honolulu: Bernice P. Bishop Museum of Polynesian Ethnology and Natural History, Department of Anthropology.

Golson, J. (1972a). Both Sides of the Wallace Line: New Guinea, Australia, Island Melanesia and Asian Prehistory. In N. Barnard (Ed.), *Early Chinese Art and Its Possible Influence in the Pacific Basin, 3* (pp. 533-595). New York: Intercultural Arts Press.

Golson, J. (1972b). The Pacific Islands and Their Prehistoric Inhabitants. In R. G. Ward (Ed.), *Man in the Pacific Islands: Essays on Geographical Change in the Pacific Islands* (pp. 5-33). Oxford: Clarendon Press.

Gorecki, P., Head, J. & Bassett, S. (1991). A Lapita Site at Lamau, New Ireland Mainland. In J. Allen & C. Gosden (Eds.), *Report of the Lapita Homeland Project* (pp. 217-221). Canberra: Australian National University.

Green, R. C. (1976). Lapita Sites in the Santa Cruz Group. In R. C. Green & M. M. Cresswell (Eds.), *Southeast Solomon Islands Cultural History: A Preliminary Survey* (pp. 245-265). Wellington: The Royal Society of New Zealand.

Green, R. C. (1978). *New Sites with Lapita Pottery and Their Implications for Understanding the Settlement of the Western Pacific, 51*. Auckland: University of Auckland.

Green, R. C. (1979). Lapita. In J. Jennings (Ed.), *The Prehistory of Polynesia* (pp. 27-60). Cambridge: Harvard University Press.

Green, R. C. (1990). Lapita Design Analysis: The Mead System and Its Use, A Potted History. In M. Spriggs (Ed.), *Lapita Design, Form and Composition* (pp. 33-52). Canberra: Department of Prehistory, Australian National University.

Green, R. C. (2003). The Lapita Horizon and Traditions - Signature for One Set of Oceanic Migrations. In C. Sand (Ed.), *Pacific Archaeology: Assessments and Prospects: Proceedings of the International Conference for the 50th Anniversary of the First Lapita Excavation. Koné-Nouméa 2002* (pp. 95-120). Nouméa, New Caledonia: Département Archéologie, Service des Musées et du Patrimoine de Nouvelle-Calédonie.

Hedrick, J. (N.D.). *Archaeological Investigation of Malo Prehistory: Lapita Settlement Strategy in the Northern New Hebrides*. Manuscript draft of PhD thesis, University of Pennsylvania, Philadelphia.

Hung, H.-C., Carson, M. T., Bellwood, P., Campos, F. Z., Piper, P. J., Dizon, E. & Zhang, C. (2011). The First Settlement of Remote Oceania: The Philippines to the Marianas. *Antiquity, 85*(329), 909-926.

Hunt, T. L. (1988). Lapita Ceramic Technological and Composition Studies: A Critical Review. In P. V. Kirch & T. L. Hunt (Eds.), *Archaeology of the Lapita Cultural Complex: A Critical Review, 5* (pp. 49-60). Seattle: The Burke Museum.

Kay, R. (1984). *Analysis of Archaeological Material from Naigani*. Unpublished MA thesis, University of Auckland, New Zealand, Grantor.

Kirch, P. V. (1984). Ancestral Polynesia. In P. V. Kirch (Ed.), *The Evolution of the Polynesian Chiefdoms* (pp. 41-69). Cambridge: Cambridge University Press.

Kirch, P. V. (1988). *Niuatoputapu: The Prehistory of a Polynesian Chiefdom.* Seattle: Burke Museum.

Kirch, P. V. (1997). *The Lapita Peoples: Ancestors of the Oceanic World.* Cambridge, Mass.: Blackwell Publishers.

Lilley, I. (1991). Lapita Sties in the Duke of York Islands. In J. Allen & C. Gosden (Eds.), *Report of the Lapita Homeland Project* (pp. 164-169). Canberra: Department of Prehistory, Australian National University.

Mead, S., Birk, L., Birks, H. & Shaw, E. (1975). *The Lapita Pottery Style of Fiji and Its Associations.* Wellington: Journal of the Polynesian Society.

Mead, S. M. (1975). The Decorative System of the Lapita Potters of Sigatoka, Fiji. In S. M. Mead, L. Birk, H. Birks & E. Shaw (Eds.), *The Lapita Pottery Style of Fiji and Its Associations* (pp. 19-43). Wellington: Polynesian Society.

Parker, V. N. M. (1981). *Vessel Forms of the Reef Island SE-RF-2 Site and Their Relationships to Vessel Forms in Other Western Lapita Sites of the Reef/Santa Cruz and Island Melanesian Area.* Unpublished MA thesis, University of Auckland, Auckland.

Pawley, A. (2007). The Origins of Early Lapita Culture: The Testimony of Historical Linguistics. In S. Bedford, C. Sand & S. P. Connaughton (Eds.), *Oceanic Explorations: Lapita and Western Pacific Settlement* (pp. 17-50). Canberra: Australian National University.

Poulsen, J. (1987). *Early Tongan Prehistory: The Lapita Period on Tongatapu and Its Relationships.* Canberra: Dept. of Prehistory, Research School of Pacific Studies, Australian National University.

Sand, C. (1996). *Le Début du Peuplement Austronésien de la Nouvelle-Calédonie: Données Archéologiques Recéntes.* Nouméa: Département Archéologie, Service des Musées et du Patrimoine de Nouvelle-Calédonie.

Sand, C. (1997). The Chronology of Lapita ware in New Caledonia. *Antiquity, 71*(273), 539-547.

Sand, C. (2000). The Specificities of the 'Southern Lapita Province': The New Caledonian Case. *Archaeology in Oceania, 35*(1), 20-33.

Sand, C. (2001). Evolutions in the Lapita Cultural Complex: A View from the Southern Lapita Province. *Archaeology in Oceania, 36*(2), 65-76.

Sand, C. (2007). Looking at the Big Motifs: A Typology of the Central Band Decorations of the Lapita Ceramic Tradition of New Caledonia (Southern Melanesia) and Preliminary Regional Comparisons. In S. Bedford, C. Sand & S. P. Connaughton (Eds.), *Oceanic Explorations: Lapita and Western Pacific Settlement* (pp. 265-288). Canberra: Australian National University.

Sand, C. (2010). *Lapita Calédonien: Archéologie d'un Premier Peuplement Insulaire Océanien.* Paris: Société des Océanistes.

Sand, C., Bolé, J. & Ouetcho, A. (2002). Site LPO023 of Kurin: Characteristics of a Lapita Settlement in the Loyalty Islands (New Caledonia). *Asian Perspectives, 41*(1), 129-147.

Sharp, N. D. (1988). Style and Substance: A Reconstruction of the Lapita Decorative System. In P. V. Kirch & T. L. Hunt (Eds.), *Archaeology of the Lapita Cultural Complex: A Critical Review* (pp. 61-82). Seattle: The Burke Museum.

Shaw, E. (1975). The Decorative System of Natunuku, Fiji. In S. M. Mead, L. Birk, H. Birks & E. Shaw (Eds.), *The Lapita Style of Fiji and Its Associations* (pp. 44-55). Wellington: Polynesian Society.

Shutler, R., Jr. & Marck, J. C. (1975). On the Dispersal of the Austronesian Horticulturalists. *Archaeology and Physical Anthropology in Oceania, 10*(2), 81-113.

Specht, J. (1991). Kreslo: A Lapita Pottery Site in Southwest New Britain, Papua New Guinea. In J. Allen & C. Gosden (Eds.), *Report of the Lapita Homeland Project* (pp. 189-204). Canberra: Australian National University.

Specht, J. (2007). Small Islands in the Big Picture: The Formative Period of Lapita in the Bismarck Archipelago. In S. Bedford, C. Sand & S. P. Connaughton (Eds.), *Oceanic Explorations: Lapita and Western Pacific Settlement* (pp. 51-70). Canberra: Australian National University.

Specht, J. & Attenbrow, V. (Eds.). (2007). *Archaeological Studies of the Middle and Late Holocene,* Papua New Guinea: Australian Museum.

Spriggs, M. (1984). The Lapita Cultural Complex: Origins, Distribution, Contemporaries and Successors. *The Journal of Pacific History, 19*(3-4), 202-223.

Spriggs, M. (1990). The Changing Face of Lapita: Transformation of a Design. In M. Spriggs (Ed.), *Lapita Design, Form and Composition: Proceedings of the Lapita Design Workshop, Canberra, Australia* (pp. 83-122). Canberra: Department of Prehistory,

Australian National University.

Spriggs, M. (2007). The Neolithic and Austronesian Expansion within Island Southeast Asia and into the Pacific. In S. Chiu & C. Sand (Eds.), *From Southeast Asia to the Pacific: Archaeological Perspectives on the Austronesian Expansion and the Lapita Cultural Complex* (東南亞到太平洋：從考古學證據看南島語族擴散與Lapita文化之間的關係) (pp. 104-140). Taipei (臺北): Center for Archaeological Studies, Research Center for Humanities and Social Sciences, Academia Sinica (中央研究院人文社會科學研究中心考古學研究專題中心).

Summerhayes, G. R. (2000a). Far Western, Western, and Eastern Lapita: A Re-evaluation. *Asian Perspectives, 39*(1-2), 109-138.

Summerhayes, G. R. (2000b). *Lapita Interaction*. Canberra: Pandanus Books. Research School of Pacific and Asian Studies, Australian National University.

Summerhayes, G. R. (2001a). Defining the Chronology of Lapita in the Bismarck Archipelago. In G. R. Clark, A. J. Anderson & T. Vunidilo (Eds.), *The Archaeology of Lapita Dispersal in Oceania: Papers from the Fourth Lapita Conference, June 2000, Canberra, Australia* (pp. 25-38). Canberra: Pandanus Books. Research School of Pacific and Asian Studies, Australian National University.

Summerhayes, G. R. (2001b). Lapita in the Far West: Recent Developments. *Archaeology in Oceania, 36*, 53-64.

Wickler, S. (2001). *The Prehistory of Buka: A Stepping Stone Island in the Northern Solomons*. Canberra: Department of Archaeology and Natural History and Centre for Archaeological Research, Research School of Pacific and Asian Studies, Australian National University.

國家圖書館出版品預行編目資料

數位人文要義：尋找類型與軌跡／項潔編. -- 初版. --
臺北市：臺大出版中心出版：臺大發行, 2012.11
面；　公分. --(數位人文研究叢書；4)

ISBN 978-986-03-4236-9 (精裝)

1.人文學 2.文獻數位化 3.數位科技 4.文集

119.029　　101021943

數位人文研究叢書4
Series on Digital Humanities

數位人文要義：尋找類型與軌跡

Essential Digital Humanities: Defining Patterns and Paths

策　劃　國立臺灣大學數位人文研究中心
叢書主編　項　潔
叢書編輯　陳怡君　蔡炯民

總　監　項　潔
責任編輯　游紫玲
編輯協力　方　誼　林必修　許楚君
封面設計　盧　耽
內頁編排　極翔企業有限公司

發 行 人　李嗣涔
發 行 所　國立臺灣大學
出 版 者　國立臺灣大學出版中心
法律顧問　賴文智律師
印　刷　中康彩色事業股份有限公司
出版年月　2012年11月
版　次　初版
定　價　新臺幣400元整

展 售 處　國立臺灣大學出版中心
臺北市10617羅斯福路四段1 號
電話：(02)2365-9286　傳真：(02)2363-6905
臺北市10087思源街18號澄思樓1 樓
電話：(02)3366-3991~3 轉18　傳真：(02)3366-9986
http://www.press.ntu.edu.tw E-mail: ntuprs@ntu.edu.tw
國家書店松江門市　電話：(02)2518-0207
臺北市10485松江路209號1樓
國家網路書店 http://www.govbooks.com.tw

GPN：1010102515
ISBN：978-986-03-4236-9

數位人文研究叢書 4
Series on Digital Humanities

Essential Digital Humanities:

Defining Patterns and Paths

現今人文研究不僅有鉅量的數位化資料可作為研究材料，更為重要的是，運用數位科技、並考量資料特性與使用者研究需求而建立的資料觀察系統與研究環境，讓研究者可以更自由地挖掘潛藏在浩瀚文本中的多重脈絡；或是進行大規模的詞頻語意分析，或是結合地理資訊呈現人文研究的空間面向，更具開放性地連結各種資料並發掘其關聯性和脈絡意義，為數位時代的人文研究帶來新的觀點與研究方法。

ISBN 978-986-03-4236-9
9 789860 342369 00400
臺大出版中心
NATIONAL TAIWAN UNIVERSITY PRESS
GPN：1010102515 定價：新臺幣400元