當資訊科技碰到史料——
臺灣歷史數位圖書館中的未解問題

涂豐恩 * 、杜協昌 ** 、陳詩沛 *** 、何浩洋 **** 、項潔 *****

摘 要

史料數位化的工作十多年來獲得了長足的進展。以清代臺灣為例，多數重要的材料：官方檔案、民間文書、地方志、文人文集與碑刻等，都可以在不同的資訊庫中瀏覽檢索，對史料蒐集提供極大的方便與助益。但除了檢索與蒐集基本材料外，若能利用已有的資訊技術，如文本探勘（text mining），對這些數位化史料進行分析，很可能會發現一些以人工不易察覺，也因此是此前少有人注意的課題。

本文要以國立臺灣大學數位典藏研究發展中心所建置的資料庫——臺灣歷史數位圖書館為基礎，結合不同的資訊技術，嘗試找出史料中潛藏的問題。依據史料不同的特性，我們將分別從時間、空間、詞頻分析、史料間的關係等幾個面向出發。藉此本文希望說明兩點：第一，資訊技術的應用必須配合著史料的特性；第二，若能進一步探索這些嶄新的、未解的疑問，不僅可以進一步認識史料，也可以增進對於歷史現象的理解。

* 國立臺灣大學數位典藏研究發展中心碩士後研究人員。
** 國立臺灣大學資訊工程學系博士後研究。
*** 國立臺灣大學資訊工程學系博士候選人。
**** 國立臺灣大學資訊工程學系博士候選人。
***** 國立臺灣大學資訊工程學系特聘教授。

Information Technology and Open Problems in the Taiwan History Digital Library (THDL)

Feng-en Tu* , Hsieh-chang Tu** , Szu-pei Chen*** ,
Hou-ieong Ho**** , Jieh Hsiang*****

Abstract

Digitizing historical archives is one of the largest national projects not just in Taiwan but in other countries as well. Along with the growing number of digitalized materials, one exciting new field arises, namely, digital humanities. Utilizing a variety of information technology tools, the study of digital humanities often brings novel perspectives to current historical and humanities studies.

In the current paper, we provide several examples to illustrate how information technology, such as text mining and Geographic Information System (GIS), can assist historians in observing their materials and discovering new historical problems. We focus on the materials and tools in the Taiwan History Digital Library (THDL), a full-text database built by the National Taiwan University Research Center for Digital Humanities. The discussion concentrates on the following topics: time, space, term frequency, and relevance of materials. From these aspects, the current paper recommends further investigations on open problems found in the THDL.

* Research Associate, Research Center for Digital Humanities, National Taiwan University.
** Postdoctoral Research Fellow, Department of Computer Science and Information Engineering, National Taiwan University.
*** Ph.D. Candidate, Department of Computer Science and Information Engineering, National Taiwan University.
**** Ph.D. Candidate, Department of Computer Science and Information Engineering, National Taiwan University.
***** Distinguished Professor, Department of Computer Science and Information Engineering, National Taiwan University.

一、前言

近幾年來史料數位化的風氣，在臺灣或世界各國都相當興盛，無論中文或外文，大量的線上資料庫不斷湧現。人文學界的研究者，大概都能察覺到這股潮流對研究方式產生的或隱或顯的衝擊。對學者而言，當前的問題已經不在於資料庫是否會帶來改變，而是它究竟會帶來什麼改變，又應該如何反應。作為建置資料庫的一方，我們同樣想問：資料庫，或者資訊科技，會對人文研究產生什麼衝擊？還有，如何才能讓資料庫在學術研究的領域中，產生最大的效應，帶動更大的改變？

對歷史研究者而言，當前最明顯的變化，應該是史料的蒐集越來越便利。過去的研究者需要跑遍大小圖書館、檔案館，必須在林林總總的書籍裡頭披沙揀金，才能將相關材料蒐羅齊全。在資料庫的年代，這樣的工作被檢索功能大幅取代。對傳統歷史研究而言，蒐集材料如此重要、如此核心，多數討論因此會將焦點集中在「檢索」上，就不難理解。對研究者而言，取得史料是首要目的，其餘似乎都是次要的。

「檢索」的重要性無須否認。但在此，本文要提出另一種想法：「檢索」或是檢索所意謂的「蒐集史料」，只是歷史研究的一環。踏出檢索的範圍之外，在歷史研究的過程中，其實仍有許多不同形式的工作——史料的排比、問題的挖掘、結果的呈現等等。在這些其他的、史料蒐集以外的工作中，資訊科技是否能，或如何能提供協助？

過去在幾篇文章中，我們曾經略述了相關的構想與一路上的嘗試。[1] 具體作為則可見於「臺灣歷史數位圖書館（THDL）」之中。關於這個系統的內容與功能，第二節還會有更多討論。大體而言，我們將這個資料庫定位在「學術研究」之用。因此，它的整體設計必然與「管理」或是「展覽」所用的資料庫有所區隔；它也與一般的網路搜尋引擎（如 google）不同，除了在內容上有特定性（不同於網路資料的無邊無際）外，我們也預設使用者具有以下特性：他們對史料有基本的認識；他們往往不是尋找「一筆」資料而是「一群」資料；他們重視史料產生的時間與來源。凡此種種。以往我們發展了眾多工具（這些工具可以見於：http://thdl.ntu.edu.tw/tools），都是奠基於這些設想之上。

簡單來說，除了檢索之外，我們還希望提供使用者「觀察」史料的工具。觀察史料的作用有二，一是解決研究者原本的問題。例如，透過檢索結果年代的分析，很快可以知道某些詞彙或概念，何時最早出現，又在哪些年代被大量使用。如果資料庫的涵蓋面夠廣，那麼這種工具很快就能解決歷史學者念茲在茲的「起源」問

1 如 Chen, Hsiang, Tu 與 Wu（2007：49-60），相關論文可至 THDL 首頁（http://thdl.ntu.edu.tw）下載。

題。但除了解決問題外，這些工具還有另一個作用，是「發掘問題」。同樣以年代分布為例，檢索結果的年代分布時常是讓人意外，也因此需要被解釋的。

藉由實際的例子，本文企圖論證，如果使用者與建置者將對於資料庫的重心，從「檢索」轉移到「觀察」，那資訊科技與歷史研究的互動層面，將不僅止於蒐集材料，而將更進一步地深入到「問題挖掘」的部分。本文的第二節，將對臺灣歷史數位圖書館的內容，以及相關的工具略作交代；第三節到第六節的部分，將具體羅列這些工具所能發掘的問題。我們將問題分成四個面向：時間與空間的分布、詞頻的意義、相似文件，以及史料間的關係。這幾個面向對應著系統功能的開發方向。本文以下篇幅所提出的問題，大多尚未得到妥善解答，有待研究者的考究，因此將其稱為「未解問題」（open problem）。但無論如何，我們希望這些問題可以引發歷史研究者的興趣，甚至進一步投入探索。

二、THDL 內容與工具介紹

首先要對本文的基礎：臺灣歷史數位圖書館（Taiwan History Digital Library，以下簡稱 THDL）的內容與工具作一介紹。THDL 是一全文資料庫，內容約有 8,000 萬字，收羅的內容主要是與清代臺灣相關的第一手史料。除全文之外，也提供人工所建置之詮釋資料，以及少部分的檔案原件影像。

THDL 目前開放兩個文獻集，其一是「清代臺灣相關行政檔案」，總數有 37,836 筆，包括從內閣大庫、軍機處檔、月摺檔、起居注、宮中檔等來源，挑選出與臺灣有關的資料，近來則增加了中國第一歷史檔案館所出版之《明清宮藏臺灣檔案匯編》[2]。從類型來看，主要包括了奏摺、題本與上諭，以及其他行政文書，也有一部分資料選自清代來臺文人的文集或地方志。此外，其中也有《清實錄臺灣史資料專輯》與《東華錄》等資料。

另一份資料集則是「古契書」，總數有 32,673 件，包含日治時期臺灣總督府檔案中抄錄的契約文書，臺大圖書館所藏的資料（如岸裡大社文書），以及戰後刊印的古文書專輯等。內容以土地交易契約數量最多，此外也有少數帶有官方性質的材料，如清代官方所頒發的「契尾」，或日治初期土地調查所產生的「理由書」。

這兩類材料在歷史研究中，具有相當特殊且重要的性質。明清檔案反映官方行政體系的運作，同時記載許多社會發展與民間活動，有日常的米穀價格，也有極端的集體叛亂行為。檔案研究早已是清代歷史研究的核心，重要性無庸多言（馮爾康，1993：121-168）。契約文書則反映著清代臺灣的另一個重要觀察角度，即民間

2 其中一大部分與臺灣所藏的明清檔案重複，因此並未重複打字。

的社會經濟活動。不同種類的契書，呈現出不同的歷史面向。如墾照代表漢人開墾勢力的移動和擴張；土地買賣契約，記錄各地經濟的活動與地權的轉移；研究賸耕字，則可以挖掘土地開發過程之中，地主與佃戶的租佃關係；鬮分契記錄著分家的過程和決策，也是家族史重要的課題；尤其值得注意的是，契約文書還存留了不少原住民的土地交易活動（王世慶，2004：198-211），「岸裡大社文書」就是最為知名的例子，近年來幾本關於清代臺灣族群關係的重要著作，都相當倚重契約文書所留下的線索（陳秋坤，1994；柯志明，2003）。

THDL 的建置是針對歷史研究，因此在功能的設計上，也盡可能地貼近傳統文獻的特性，以及歷史研究者的使用需求。就前者而言，雖然許多文獻本身有進行斷句標注，以便閱讀，但在檢索時系統會略過標點符號，即以未標點的形式進行檢索。就後者而言，THDL 則提供檢索之外眾多的分析工具，例如能讓使用者將所需文件加以整理的「自訂文件集」，或是能對檢索結果加以觀察的「後分類」、「詞頻」以及「檢索結果圖」。關於 THDL 中眾多功能的使用方式與意義，讀者可參閱其他相關文章，或網路上的使用說明（項潔、陳詩沛、杜協昌，2009），本文在此不一一介紹。

不過，以下行文會反覆提及「後分類」、「詞頻」，這兩個詞彙對一般讀者而言，或許有些陌生，因此在此需要稍作解釋。「後分類」指的是在檢索結果後，將查詢的結果進行分類。其中，因「明清檔案」與「古契書」的性質不同，兩者的分類範疇也有些許差異。明清檔案的分類範疇包括「年代」、「出處」（即檔案出處，如《宮中檔乾隆朝奏摺》）、「作者」和「分類」（即檔案性質，如上諭、移會或奏摺）；古契書則在上述四項之外，另增加了「地域」，即契書的地理資訊。此一功能設計的出發點，是在一般資料庫中，使用者往往會查詢到相當大量的資料，透過後分類的功能，使用者可以對結果有一個大略的瞭解，同時也可以根據後分類進一步縮小觀察範圍。

「詞頻」，顧名思義，則是指詞彙在檢索結果中出現的頻率。我們利用詞夾子程式，將各篇文件中的人名、地名與其他專有名詞擷取出來，再統計它們出現的頻率（謝育平等人，2009；陳詩沛、杜協昌、項潔，2009）。此一功能同樣可以提供使用者在檢索後，快速而簡單的觀察查詢結果。另一方面，我們在接下來的討論中也會發現，研究者往往可以從中發掘到一些意外的現象，進而引發進一步的問題。

首先，我們要從時間和空間兩個角度觀察 THDL 中的材料。時間，是歷史研究中最核心也最基本的概念；空間，則是近年來歷史研究中日興的研究視角。從這兩個看似簡單的觀念出發，仍會從史料中發現許多意想不到的課題。

三、時間與空間的分布

（一）明清檔案

圖 1 是 THDL 中所有明清檔案的年代分布圖。其中的幾個高峰很容易會引起關注。熟悉臺灣史的讀者大概可以推想每個高峰時期在臺灣發生的事件。如最高峰的 1787（乾隆五十二）年是林爽文事件；次高峰的 1884（光緒十）年，則是中法戰爭進行到基隆、淡水的年代；第三個高峰則是 1894-1895 年甲午戰爭爆發和馬關條約的簽訂。其次的幾個高峰，則包括了 1874（同治十三）年的牡丹社事件、1806（嘉慶十一）年的蔡牽事件、1833（道光十三）年的張丙事件等。

乍看之下，這幾個時間點，似乎都反映著清朝中央政府對臺灣關注程度的陡然升高。其原因，若非所謂的民眾叛亂，就是清朝與其他國家勢力在臺灣的交鋒。另一方面，我們也會注意到大約從 1685 到 1720 年之間，文件的數量非常少，甚至在 1700（康熙三十九）年，只有一件檔案來自《東華續錄》，內容是關於福建巡撫調任的紀錄，與臺灣的關係可以說是有些遙遠的。以往對清代初期對臺政策的評價往往指稱為「消極治理」。這樣的印象或論斷也許一部分也來自檔案的如斯分布。

清代臺灣檔案的分布，可以與《清實錄臺灣史資料專輯》（以下簡稱《專輯》）的時間分布（圖 2），作一比較。若我們基本上接受《專輯》的內容是足夠完整的，那麼《專輯》的年代分布就應該有其代表性和獨特性：它同樣反映官方視角，而且是經過官方整理、檢查過的視角。兩相比較之下，在《專輯》中，林爽文事件同樣一枝獨秀，數量遠遠超過了 19 世紀晚期的幾個大事件。尤其 1874 年的牡丹社事件，數量變得很少，甚至不能構成一個高峰。為何會出現這樣的差異，值得思考。

其次，圖 1 是將 THDL 的明清檔案文獻集中，所有文件皆予以標示。但我們可以進一步將圖 1 細緻化，聚焦在上諭、奏摺兩類文件之上。圖 3 即是將 1718-1813 年之間，臺灣相關的上諭與奏摺的數量，以直條圖加以比較。其中有幾個值得注意的現象。第一，在 1718 年到 1766 年之間，上諭的數量非常之少。換言之，就算清代初期對臺治理並非單純的「消極」政策，對於臺灣統治的討論也鮮少是由皇帝所發動的。但從 1766 到 1813 年之間，情況就有明顯的不同，不僅奏摺出現了幾個明顯的高峰，同樣的，上諭的數量也大幅增加。尤其明顯的，仍是乾隆年間的林爽文事件。其中上諭的數量，甚至幾乎要接近奏摺。若我們想到上諭是因皇帝一人意志而生（儘管有人代筆），奏摺則來自眾多地方官員之手，必然會對這種現象感到有些訝異。

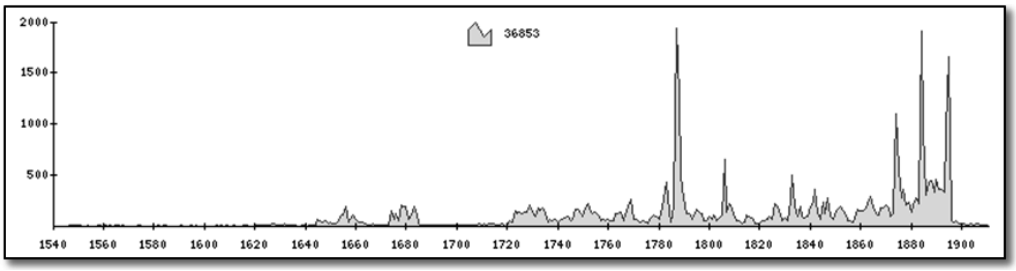

圖 1　THDL 中所有明清檔案的年代分布圖

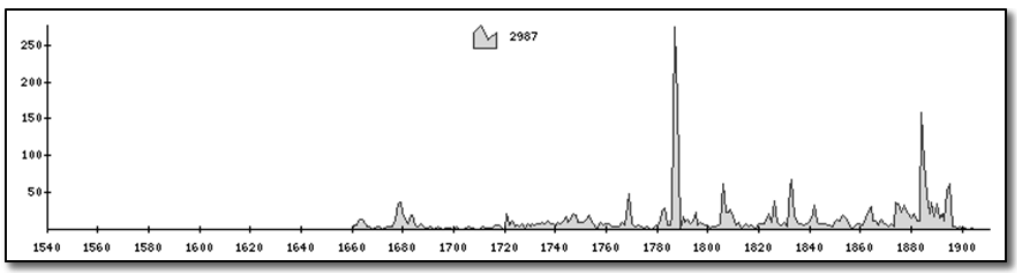

圖 2《清實錄臺灣史資料專輯》年代分布圖

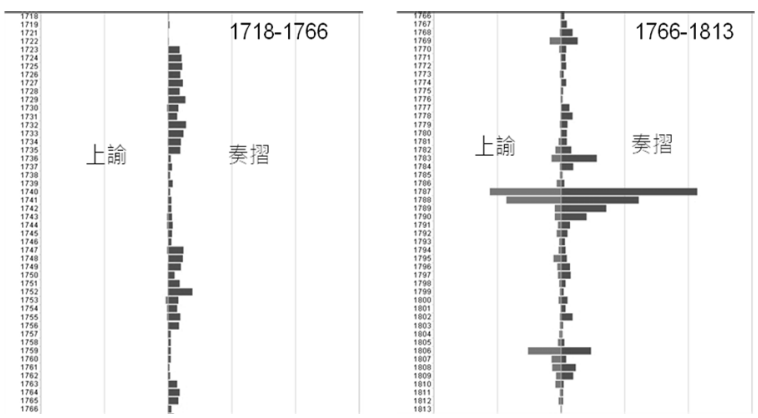

圖 3　上諭與奏摺數量之比較

圖 3　1718-1813 年上諭與奏摺數量之比較

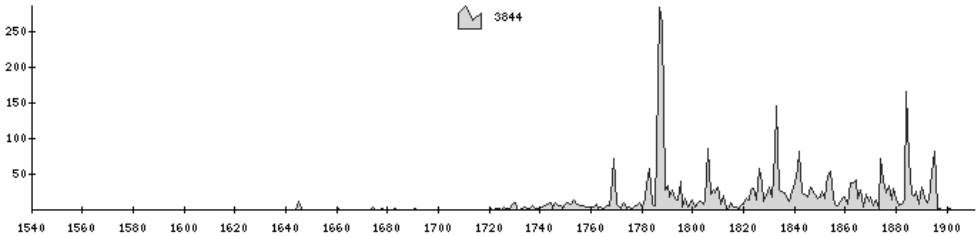

圖 4 THDL 中上諭的年代分布

圖 3 提醒了一件事：儘管我們將這些不同的文獻統稱為「明清檔案」，但其內部其實包含著不同種類的文件，將它們合而觀之，自然有其意義；將其分而觀之，也能提供不同的視野。

圖 4 則延伸自圖 3，是單獨以「上諭」為本所繪製的年代分布圖。將它與圖 1、圖 2 相比較，可以明顯感覺到一些差異。除了林爽文事件外，在其他幾個不同的時期也出現了高峰，尤其是張丙事件與牡丹社事件特別突出。此外，發生於林爽文事件之前的黃教事件（1769，乾隆三十五）也成為了另一個高峰。

從這些不同的年代分布圖中，我們看到的是，同樣一段清代臺灣史，同樣透過所謂的明清檔案加以觀察，卻可能產生許多不同的視角。奏摺、上諭代表了從事件發生當下，不同位置的觀察者對事件本身的感知；《清實錄》則是事件過後的統整與紀錄。這可以解釋幾張圖表何以會出現差異，不過，若要更深入地詮釋這些差異，還必須有更多的研究。

而在前面的討論中，我們大多強調年代分布圖上的高峰。但我們也可以反向思考另一些問題，比如，為何某些似乎重大的事件並未反映在上述分布圖中。舉例而言，一般臺灣史教科書中習慣將「朱一貴」、「林爽文」與「戴潮春」並列為「清代臺灣三大民亂」，然而，我們在前文中卻會發現，朱一貴事件和戴潮春事件在檔案數量上，並不算特別多，尤其是與林爽文事件相比，甚至還不及比方說張丙事件。隨之而來的問題是，在什麼基礎上他們可以並列「三大民亂」？又是什麼時候開始他們成為「三大民亂」？關於這個問題，下一節中還會觸及。

（二）古契書

接下來討論契約文書，我們同樣從年代分布開始。圖 5 是 THDL 中目前所收錄之契約文書的年代分布圖，與上一小節相比，圖 5 的曲線一直到 19 世紀晚期之前，都顯得相當平緩，不像明清檔案的劇烈震盪。其中三個高峰，第一個是 1888（光

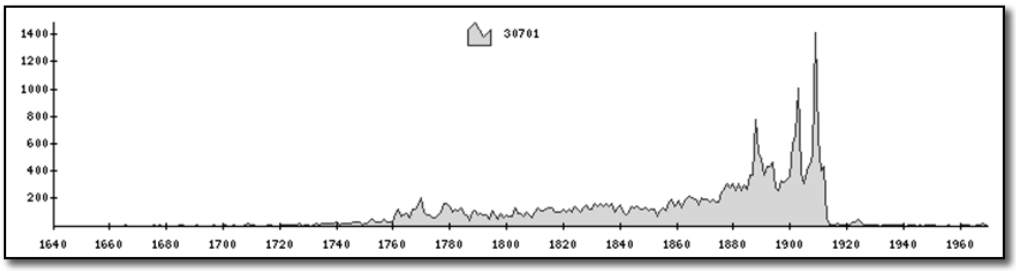

圖 5 THDL 中目前所收錄之契約文書的年代分布圖

緒十四）年劉銘傳發動臺灣土地清丈，當年頒布了許多「丈單」，在 THDL 中至少就有 424 件。而後兩個高峰則與日治時期總督府的土地調查事業有密切關係（李文良，2004）。換言之，從這張圖表中，首先吸引我們注意的，其實與一般人民的土地交易關係較低，反倒先呈現了官方對土地控制的行動。

若暫時不去注意這三個高峰，那麼整張分布圖似乎反映了社會歷史相對靜態、轉變緩慢的一面。在明清檔案中十分關鍵的年代，如包括林爽文事件在內的各個叛亂活動或是外國勢力的入侵，在圖 5 中都不見蹤影。這是否表示，人們生活中的經濟活動，受這些突發事件的影響並不多。這讓人想起上個世紀法國年鑑學派對歷史時間的觀點：那些看似驚心動魄的政治或軍事事件，放在長時段的社會史中，好像不過是水面上旋起旋落的浪花，轉瞬即逝，不留痕跡（Braudel, 1976: 21）。

當然，明清檔案與古契書的文件集本身，存在著重要的差異。THDL 中的明清檔案，儘管不能說沒有遺漏，但整體而言，還有一定的完整性與代表性。古契書則不然。儘管 THDL 可稱得上這方面目前收錄數量最為龐大的資料庫，但三萬多件古契書究竟占清代臺灣契約文書的多少比例，還很難斷言。因此，前一段的斷言，目前還只能暫時當作假說，未來如果能擁有更多的資料，不妨再加以驗證。

在古契書的資料集中，來自「臺灣總督府檔案抄錄契約文書」的資料，大約占了一半，將近 16,000 件。而其他各地館藏與出版的資料，則占了另外一半，約有 17,000 件。如果我們將兩者分開觀察，也會浮現一些有趣的問題。

圖 6 是所有「總督府抄錄契約文書」的年代分布圖；圖 7 則是「總督府抄錄契約文書」以外的年代分布圖。兩相比較，差異非常明顯。圖 6 中的高峰集中在 20 世紀的土地調查事業；圖 7 則精彩得多，除了 1888 年前後的高峰以外，我們終於可以看見其中出現比較多的高低起伏。比如在 1760 年前後，突然出現了上揚的趨勢，但到了 1770 年之後，又突然下降，出現了一個突兀的凹谷。另一個凹谷則出現在 1895 年左右。這兩個凹谷要如何解釋？或者，它們是可以被解釋的嗎？它們可以反

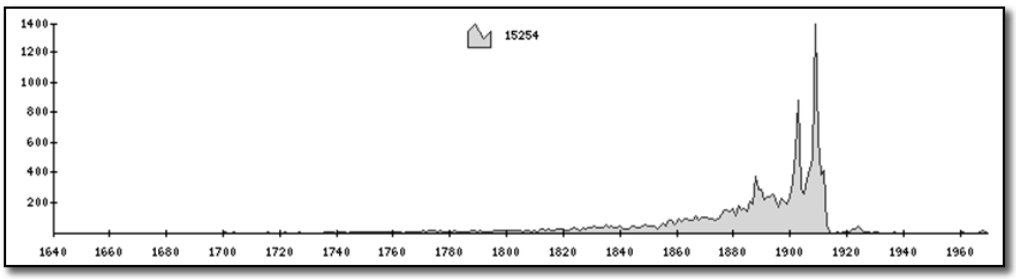

圖 6　THDL「總督府抄錄契約文書」的年代分布圖

映當時臺灣社會經濟的發展趨勢嗎？這個趨勢有代表性嗎？

其次，「臺灣總督府檔案抄錄契約文書」包括了四個部分：「15 年保存公文類纂」、「永久保存公文類纂」、「土地調查公文類纂」、「高等林野公文類纂」。其中「高等林野公文類纂」數量較少，只有 468 件，但其餘三個部分都有數千件。這三者中所包括的抄錄契約文書有何不同？以下三張年代分布圖，圖 8 是數量最多的「15 年保存公文類纂」、圖 9 是「永久保存公文類纂」、圖 10 則是「土地調查公文類纂」。

此三者的高峰都在 1900 年以後，這不令人訝異。不過比較之下，永久保存公文類纂與土地調查公文類纂二者，尤其集中。反倒是 15 年保存公文類纂是逐漸上升的曲線。這個現象又該如何解釋？答案或許牽涉到土地調查展開的歷史過程，以及永久保存和 15 年保存的不同性質。

除去單純的年代分布外，我們也可以循著前小一節明清檔案的模式追問：不同類型的契約文書，分布又是如何？在考慮這個問題之前，我們同樣碰到了契約文書與明清檔案的差異。作為官方運作的一環，明清檔案的產生有相對明確的框架，比如題本與奏摺的分別、傳諭與字寄的不同等。儘管這些分別偶爾也會混淆，但大體而言，制度性的框架是清楚的。契約文書則缺乏了這樣的制度性框架，因此，契約文書的分類經常是浮動的，隨著研究者的視角而改變。在此我們不處理這個棘手的問題，只挑選其中兩個範疇為例：「開墾契」與「鬮分契」。

先看鬮分契。在 THDL 中的「鬮分契」一類項下，共有 4,572 件契約。它們的年代分布如圖 11，從中可以明確感覺到上升趨勢。大體來說，19 世紀下半葉開始，鬮分契的數量要遠遠超過此前的數目。當然，可以想見，不管分家或是財產分管，都是土地開發過程中比較晚期的行為，上升的趨勢因此不讓人特別驚訝。但我們仍然可以探問，何以自 1880 年左右開始，數量開始顯著地上升。在此同時，我們也必須注意，鬮分契往往是一式多份，因此目前的統計數字還有加以精緻化的必要。[3]

3　最近有研究者對類似的題目進行考察，見李朝凱（2010：143-149）。

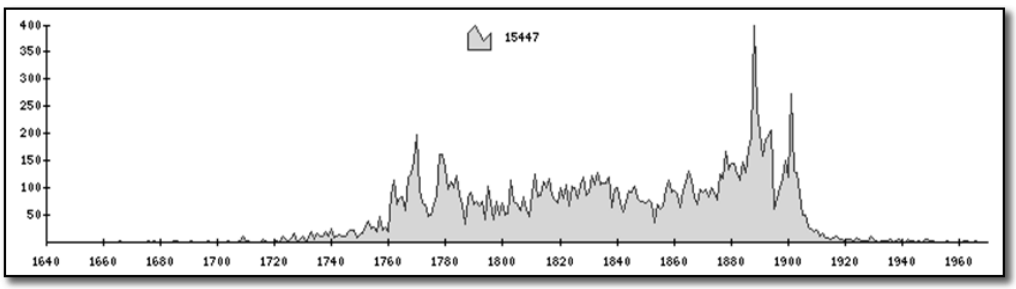

圖 7　THDL「總督府抄錄契約文書」以外的年代分布圖

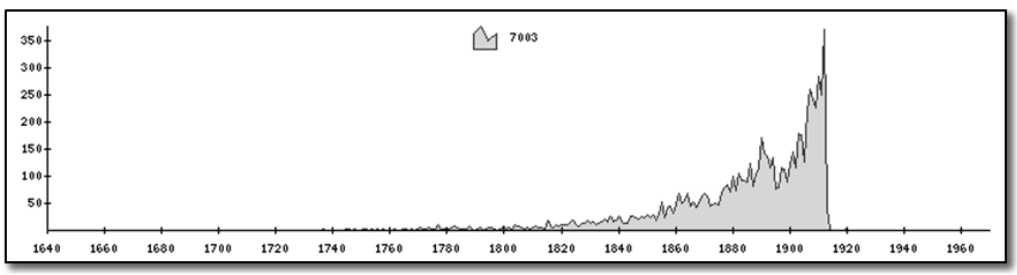

圖 8「15 年保存公文類纂」年代分布

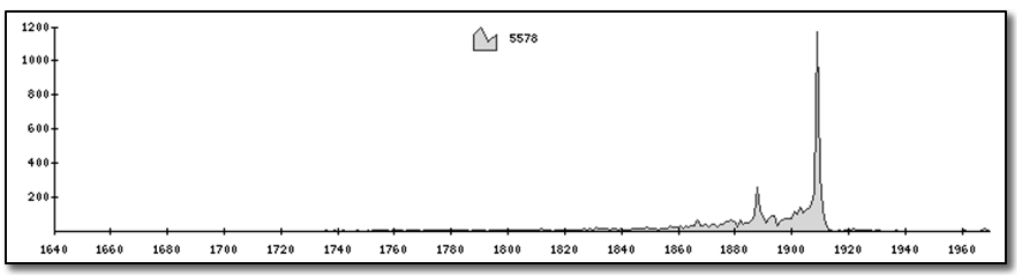

圖 9「永久保存公文類纂」年代分布

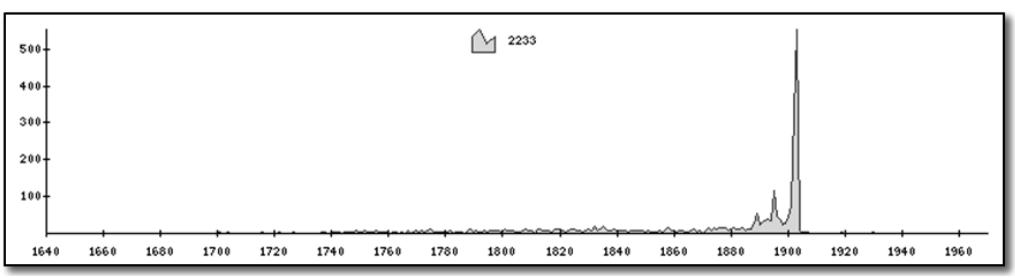

圖 10「土地調查公文類纂」年代分布

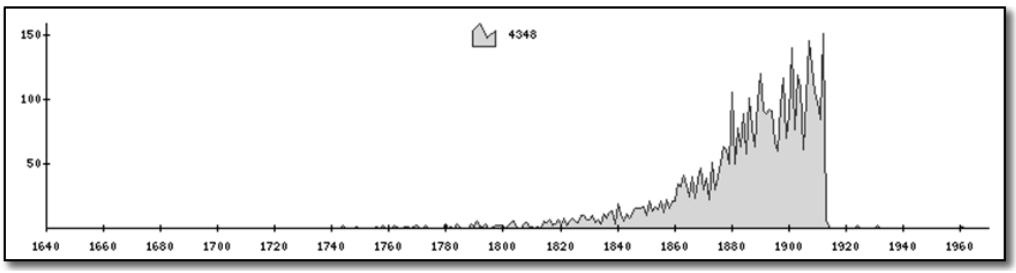

圖 11　THDL 中「鬮分契」之年代分布

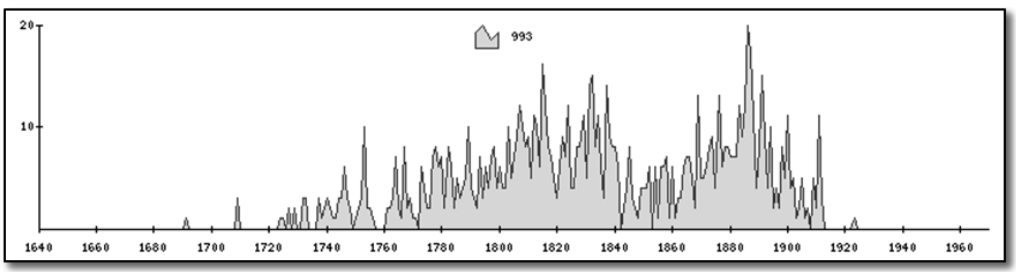

圖 12　THDL 中「開墾契」之年代分布

「開墾契」（圖 12）的年代分布圖，則與「鬮分契」截然不同。與此前幾張年代分布圖也有差別。看起來時高時低，起伏不定，不太容易掌握到什麼規律。我們能說，高峰處是臺灣開墾史上發展比較快速的部分嗎？反過來，數量較少的又代表了什麼？又或者，這張圖只反映了現存史料的年代分布，而與開墾史的實像無關？

以上的討論，都是以文件的時間分布作為觀察基點。但契約文書還有另一個特性，它的空間資訊，包括位置、大小等等，都相當關鍵。因此，本小節的最後，也要從空間分布的角度提出一個問題。

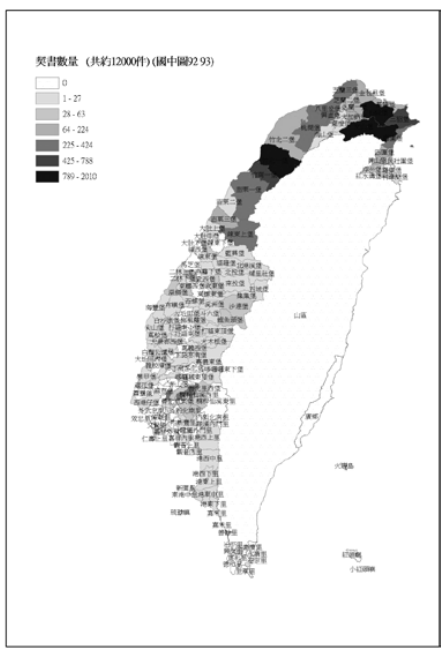

圖 13　總督府抄錄契約文書之地理分布

圖 13 是從「總督府抄錄契約文書」中 15,000 多張契約，找出約 12,000 張具有精確地理資訊的契約，繪製而成的空間

分布圖。其底圖是 1904 年殖民政府所繪製的臺灣堡圖。同樣地，我們很快會問：何以它呈現如斯分布？

初步觀察看來，契約特別集中在北部，特別是竹北一堡與文山堡兩個地區。中部以下相對較少，甚至有地區完全是空白。這又是為什麼？當然，若我們把剩下的約 3,000 張契約，再標示到圖上，可能會與目前所見的分布有些差異。不過問題依舊，契約的分布仍需要進一步的詮釋。

四、詞頻的意義

（一）明清檔案

前一節曾經提到，從臺灣相關明清檔案的整體數量看來，林爽文事件從各種角度看來，似乎都高居第一，與它並稱的「朱一貴事件」與「戴潮春事件」，相對而言就不那麼突出。此處要以「朱一貴事件」為例，對這個問題作進一步的討論。

如果以「朱一貴」為關鍵字檢索，在 THDL 中會找到 246 件檔案。其年代分布圖大致如下。這張分布圖的兩個高峰是有些奇怪的。尤其啟人疑竇的是，朱一貴事件發生在康熙六十（1721）年，但圖上的最高峰，卻是乾隆五十二（1787）年，比前者還要高出一些。何以如此？

從詞頻上來看更有意思。在人名一項的前十名中，與朱一貴事件直接相關的著名人物，如藍廷珍、施世驃，當然排在其中（如表 1）。但讓人意外的是，其中有不少人名不但與朱一貴無關，甚至生平年代與朱一貴事件都頗有差距。而我們自然會注意到，高居第二的是林爽文，與林爽文息息相關的常青、福康安、莊大田等人，也出現在詞頻之上。這與上述年代分布圖的狀況顯然是一致的。但我們仍要問：為何如此？

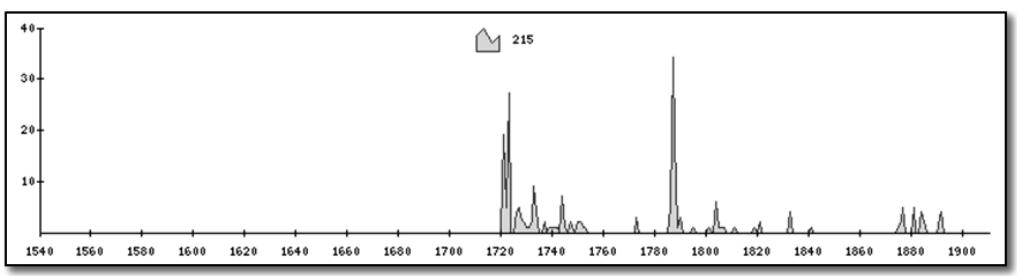

圖 14　檢索「朱一貴」之年代分布

表 1 檢索「朱一貴」之詞頻

人名				
Term t		df	t→q	回報
朱一貴	生平	245	1.000	
林爽文	生平	88	0.058	
常青	生平	53	0.038	
藍廷珍	生平	52	0.338	
滿保	生平	44	0.141	
施世驃	生平	39	0.639	
福康安	生平	36	0.023	
黃仕簡	生平	35	0.040	
任承恩	生平	30	0.076	
莊大田	生平	30	0.058	

這個問題還值得深入探究，這裡僅提出一些初步的觀察。如果回到檔案中，在林爽文事件發生當下，中央與地方之間頻頻往來的通訊中，朱一貴頻頻被提起，乃是作為臺灣將官平亂無力的對照。對乾隆而言，在他祖父康熙的時代，類似的亂事如此輕易平定，何以現下臺灣官員竟會放任讓林爽文事件恣意擴大，竟至難以收拾。一篇字寄中就寫到：

> 從前康熙年間，奸民朱一貴聚眾滋擾，至於全臺失陷，經官兵剿捕，不及一月，即已收復蕆功。今此等么麼賊匪，不過占據縣城，較之從前更易辦理。俟大兵到齊，同時並力夾攻，自可即日蕩平（中國第一歷史檔案館，1980：203-205）。

不只如此，乾隆更想參考朱一貴事件時官方的戰略，甚至找出曾經參與其事的藍鼎元之著作，他說到：

> 朕披閱藍鼎元所著東征集，係康熙年間，臺灣逆匪朱一貴滋事，官兵攻剿時，伊在其兄藍廷珍幕中，所論臺灣形勢，及經理事宜，其言大有可採（中國第一歷史檔案館，1980：250-252）。

換言之，朱一貴事件發生後數十年，之所以又被往事重提，是因為它與當下的事件結合在一起。他成為歷史當事人（尤其是乾隆）思考和行動的重要參照點。循此，我們可以問的是民眾叛亂的歷史記憶：朱一貴事件在什麼時間點，又是如何在歷史中被記錄、被詮釋。從上述簡短的討論看來，或許正是因為林爽文事件方才賦予了朱一貴事件一個新的歷史位置。

（二）古契書

我們同樣可以從詞頻的角度觀察古契書。在全部的古契書中，出現的人名前十名如下表：敦仔、中村是公、林人文、張公藝、後藤新平、潘明慈、小松吉久、潘士萬、潘輝光、楊獅。這十個名字中，有些我們並不陌生。如敦仔、潘明慈、潘士萬和潘輝光，都是岸裡社的成員；至於中村是公、後藤新平和小松吉久，則是日本殖民政府的官員，中村是公時任臨時土地調查局的局長、小松吉久曾任宜蘭廳長，至於大名鼎鼎的後藤新平，更無須介紹。這樣的分配反映出 THDL 中古契書文件集的特性：它與岸裡大社和土地調查二者的關係特別密切。

表 2　古契書詞頻分析列表的前十名

人名			
Term t	df	t→q	回報
敦仔	509	1.000	■■■
中村是公	483	1.000	■■■
林人文	472	1.000	
張公藝	347	1.000	■■■
後藤新平	305	1.000	
潘明慈	298	1.000	■■■
小松吉久	260	1.000	■■■
潘士萬	196	1.000	■■■
潘輝光	177	1.000	■■■
楊獅	176	1.000	■■■

上述七人外，剩下的三個名字林人文、張公藝和楊獅，看來就有些陌生。關於林人文和楊獅，在下一節中會有更多的討論。在此只聚焦於張公藝一人。張公藝是誰？如果思考古契書所代表的歷史意義，可能會猜想此人是清代的大地主，因為出現在古契書的人名，多數都是名不見經傳的小人物，出現的次數越多，理論上意謂著他曾經參與過大量的土地交易行為。不過，一旦進一步考索，就會知道這樣的猜測完全是錯誤的。張公藝並非臺灣人，甚至也不是清朝人，而是生活在唐代，以「九世同居」著稱。《舊唐書》中有一段對他的記載：

> 鄆州壽張人張公藝，九代同居。北齊時，東安王高永樂詣宅慰撫旌表焉。隋開皇中，大使、邵陽公梁子恭亦親慰撫，重表其門。貞觀中，特敕吏加旌表。麟德中，高宗有事泰山，路過鄆州，親幸其宅，問其義由。其人請紙筆，但書百餘「忍」字。高宗為之流涕，賜以縑帛（《舊唐書．列傳》，卷一百八十八）。

那麼，張公藝為何在古契書的詞頻中高居第四？若我們以「張公藝」為關鍵字，觀察其檢索結果的後分類，可以看到大多數均是「鬮分契」。原來，在清代臺灣的鬮分契中，時常會提到這樣的句子：「竊聞張公藝九世同居，此風足慕」或「盡聞九世同居張公藝忍堂著美」。他們之所以提及張公藝之名，並非真的是要追隨九世同居的生活模式；正好相反，他們儘管高舉這樣的理想，卻往往在契約接下來的內容，大嘆無奈，不得已而分家析產（涂豐恩，2010：11-20）。

儘管我們已經知道張公藝的背景，有一些問題還有待討論。首先，我們似乎還不清楚為什麼是張公藝。綜觀清代鬮書內的修辭，他們偶爾還會提及其他一些性質類似的名字，如「七世同財范稚春」，可是數量上遠遠不如張公藝。此外，這是否為臺灣地區鬮分契的特色？比較清代中國其他地方，如徽州、福建等地的鬮分書中，似乎就很少見到類似的修辭模式（張傳璽，1995；陳支平，2009；福建師範大學歷史系，1997）。[4]

五、史料間的關係：土地移轉圖

接下來要談的問題是土地移轉圖。土地移轉圖是奠基在另一項研究「臺灣契約文書交易關係自動重建」之上，從這個研究中，我們利用自動技術找出現存契約文書中的「上下手契」等關係（黃于鳴，2009）。若將這些單一的關係再組合成更為複雜的關係，比如連續的上下手契，或是鬮分後再杜賣、杜賣後再鬮分等，就有可能找出相對豐富的土地交易關係。比如在 3,287 組上下手契與鬮分契的關係中，大約有超過 150 個歷經兩次交易以上的契約關係。在這些較為複雜的契書關係中，似乎有些故事隱然成型。將這些關係繪製成圖後，就能以視覺化的方式，表現一塊土地轉手復轉手的身世。我們將這個成果稱為「土地移轉圖」。

每一張移轉圖都隱含了一則故事，或大或小。在此，我們僅舉其中兩個比較複雜而有趣的案例，作為說明：

圖 15 是本研究牽涉最多契書的一組關係，因為數量過於龐大，在紙本上不容易找到較為適當的呈現方式。這張圖的主角是臺南地方士紳林人文，也就是我們前一

4 但這幾本書所收的鬮書數量並不多。

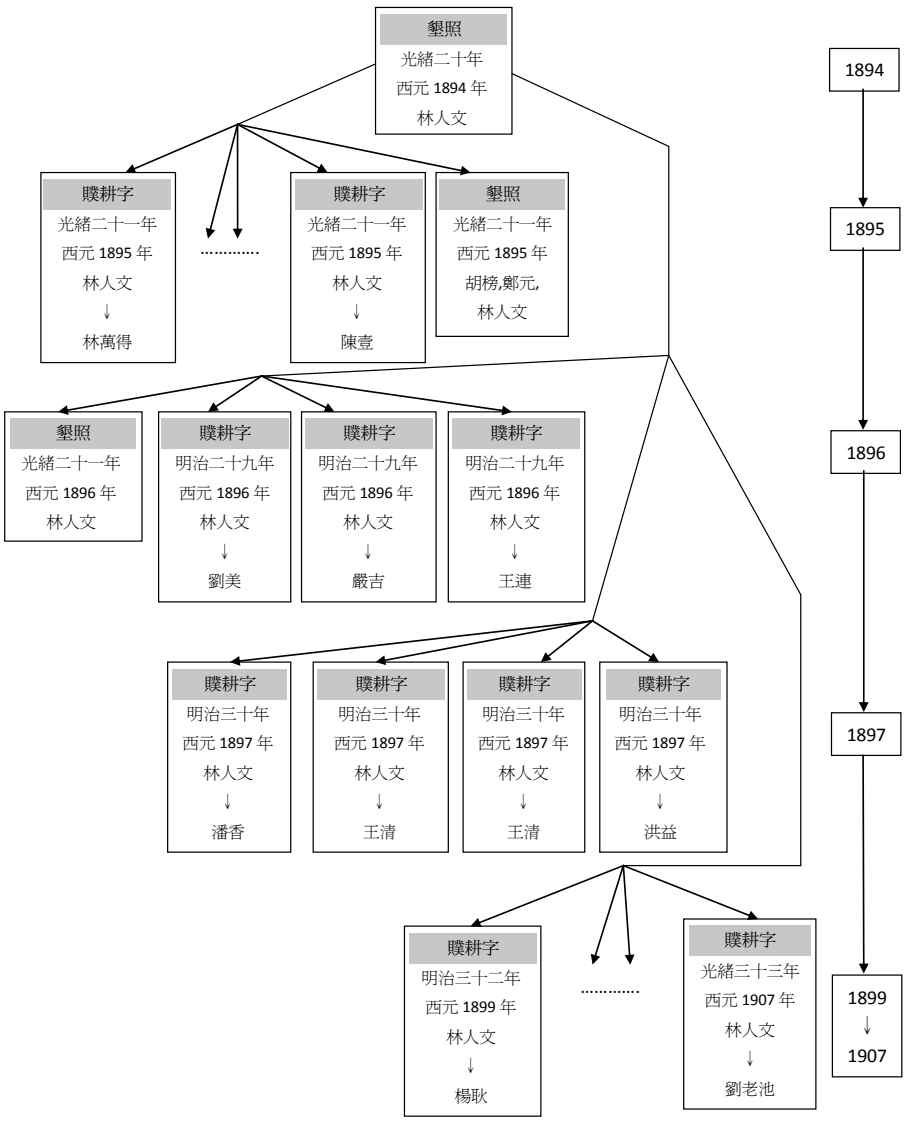
墾照
光緒二十年
西元 1894 年
林人文
1894
1895
1896
1897
1899
↓
1907
賸耕字
光緒二十一年
西元 1895 年
林人文
↓
林萬得
……
賸耕字
光緒二十一年
西元 1895 年
林人文
↓
陳壹
墾照
光緒二十一年
西元 1895 年
胡榜,鄭元,
林人文
墾照
光緒二十一年
西元 1896 年
林人文
賸耕字
明治二十九年
西元 1896 年
林人文
↓
劉美
賸耕字
明治二十九年
西元 1896 年
林人文
↓
嚴吉
賸耕字
明治二十九年
西元 1896 年
林人文
↓
王連
賸耕字
明治三十年
西元 1897 年
林人文
↓
潘香
賸耕字
明治三十年
西元 1897 年
林人文
↓
王清
賸耕字
明治三十年
西元 1897 年
林人文
↓
王清
賸耕字
明治三十年
西元 1897 年
林人文
↓
洪益
賸耕字
明治三十二年
西元 1899 年
林人文
↓
楊耿
……
賸耕字
光緒三十三年
西元 1907 年
林人文
↓
劉老池

圖 15 土地移轉圖之一

節在詞頻分析中曾看過的人名。他在光緒二十（1894）年向政府申請執照，計畫開墾善化地區約五十餘甲的土地，其後分別與小租戶訂立合約字。其中大多簽訂於光緒二十一（1895）年，也就是林人文拿到墾照的後一年，但最晚的合約也遲至明治三十五（1902）年方才簽訂。[5]

圖 15 以非常具象的方式，呈現大租戶與小租戶之間的關係。若稍加閱讀契書內文，則可知林人文所開墾的是曾文溪旁的浮復地。為此，在日治初期，他與善化地方大族的楊家，也就是前一節提及的楊獅，還曾起過地權糾紛。楊家宣稱該地原由他們墾成，也應該具有所有權。為此日本政府介入調查，因而在總督府檔案中留下大量的理由書。

圖 16 是另一個有趣的例子，它與前者有個不同，因為圖中涵納的契書不只是關於同一土地，也關於一個家族。這些契書來自永和山庄地區，最早是道光三十年廖佳福與吳張承等人，立約均分竹南一堡地方的土地。到了明治三十四（1901）年 6 月，廖家的四房子孫，包括長房廖俊喜、廖俊烋、廖維良、廖維賢，二房廖俊聰、廖維棟，三房廖俊日，四房廖俊明、廖俊心等人，因為「家口之紛紜，保無私心之自用，欲常安靜，殊覺維艱」[6]，因此簽訂鬮分書，將祖父廖佳福所留下的土地一分為四。

同月，長房的廖俊喜、廖俊烋、廖維良、廖維賢四人，又分別簽訂分家鬮書，將土地一分為四。[7] 分到十六分之一土地的廖維良，又在同年與兄弟廖維義、廖維岳簽訂鬮書，再將永和山楊梅凸地方的土地，再次均分為三。[8] 至於二房的叔姪廖俊聰與廖維棟，也在同年 6 月，將所得之土地析分為二，從該鬮書中還可見到，隔年廖維棟即把一處銃櫃坪的土地賣與黃錫璋。[9] 換言之，光是在明治三十四年，廖家就簽訂了四份鬮分契書，而廖佳福原先所開墾的土地也就在這樣的過程中，被切割析分。

但故事還沒完，在這波鬮分高潮後，分到土地的廖家子孫遂紛將土地出售，如明治三十九（1906）年，廖維良就將名下土地賣給同在永和山庄的張比、張元兩人；其餘如黃阿順、羅普壽等人，也陸續分別向廖家收購土地。[10] 廖家對土地的所有權就這麼逐漸轉讓與外人之手。

上述對契約關係的解讀仍嫌粗淺，僅僅初步勾勒契約文書所串連的故事，尚未

5 臺灣歷史數位圖書館，檔名：cca100003-od-ta_04411_000005-0001 。
6 臺灣歷史數位圖書館，檔名：cca100003-od-ta_01841_000312-0001。
7 臺灣歷史數位圖書館，檔名：cca100003-od-ta_01841_000304-0001、cca100003-od-ta_01841_000335-0001、cca100003-od-ta_05568_000111-0001、cca100003-od-ta_05568_000160-0001。
8 臺灣歷史數位圖書館，檔名：cca100003-od-ta_05568_000115-0001。
9 臺灣歷史數位圖書館，檔名：cca100003-od-ta_01841_000308-0001。
10 臺灣歷史數位圖書館，檔名：cca100003-od-ta_01841_000318-0001、cca100003-od-ta_05568_000113-0001。

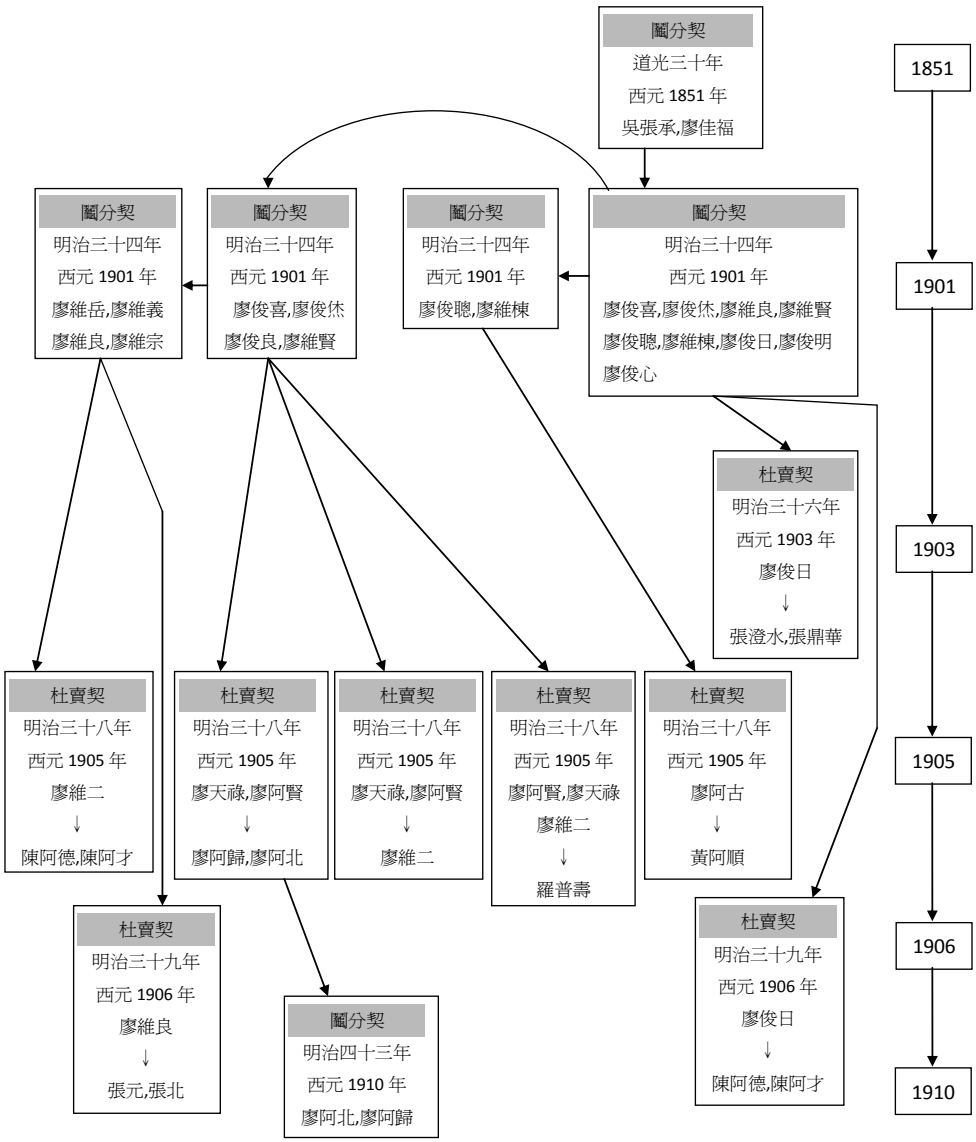
1851
1901
1903
1905
1906
1910
圖分契
道光三十年
西元 1851 年
吳張承,廖佳福
圖分契
明治三十四年
西元 1901 年
廖維岳,廖維義
廖維良,廖維宗
圖分契
明治三十四年
西元 1901 年
廖俊喜,廖俊然
廖俊良,廖維賢
圖分契
明治三十四年
西元 1901 年
廖俊聰,廖維棟
圖分契
明治三十四年
西元 1901 年
廖俊喜,廖俊然,廖維良,廖維賢
廖俊聰,廖維棟,廖俊日,廖俊明
廖俊心
杜賣契
明治三十六年
西元 1903 年
廖俊日
↓
張澄水,張鼎華
杜賣契
明治三十八年
西元 1905 年
廖維二
↓
陳阿德,陳阿才
杜賣契
明治三十八年
西元 1905 年
廖天祿,廖阿賢
↓
廖阿歸,廖阿北
杜賣契
明治三十八年
西元 1905 年
廖天祿,廖阿賢
↓
廖維二
杜賣契
明治三十八年
西元 1905 年
廖阿賢,廖天祿
廖維二
↓
羅普壽
杜賣契
明治三十八年
西元 1905 年
廖阿古
↓
黃阿順
杜賣契
明治三十九年
西元 1906 年
廖維良
↓
張元,張北
圖分契
明治四十三年
西元 1910 年
廖阿北,廖阿歸
杜賣契
明治三十九年
西元 1906 年
廖俊日
↓
陳阿德,陳阿才

圖 16 土地移轉圖之二

深入挖掘其中隱含的豐富資訊。若要進一步梳理，或可配合幾項工作。其一是將其他相關但未被納入圖中的契書一併找出，因為目前圖中所有的關係都是著眼於同一塊土地，但其中牽涉的人物則可能擁有不同土地，這些人物可能不只出現在契約文書中，包括地方志或文集中的資料都可能存在相關資訊，如圖 15 的林人文就曾編寫《三字經》，也曾擔任地方上的教師。這些材料，即便是旁證，仍可提供立體而多面的歷史圖象。

六、相似文件

電腦另一個重大的功能，是能夠對文字自動作比對。儘管它無法認知到文句中的意義，但可以輕易而快速地比對大量文字，並判斷它們彼此的相似性。將文件比較的功能應用在不同的文件上，也會發現一些不同的問題。

（一）明清檔案

以明清檔案而言，透過相似文件的功能，可以對所有的文件作一全面的整理，進而發現有幾種的資料會因為內容大同小異而集結成群。比如，有關雨水糧價的奏摺，或是每年的監生銀數。這幾種資料，因具有大致固定的報告格式，而被電腦判定為相似文件。我們將監生銀數所出現的年份，以及其中記載的監生人數製成圖 17。其中與文監生相關的文件，大約自道光十二年便開始出現，而武監生則自道光二十三年之後，短暫存在，其後數量變得相當稀少。這樣的變化，應與臺灣當時整體的財政和對外關係，都有所牽連。[11]

（二）古契書

從相似文件來看古契書，乍看之下不甚稀奇。因為古契書的文字內容，時常極為類似，往往只在一些關鍵處，如人名、地名、價錢等等，有所更動，其餘的部分則大同小異。而這些大同小異的部分也往往被一般研究者所忽略。但從另一個角度想，這些在各地契書中不斷抄錄、重複的文字，卻可能代表著文化上的特殊意義，比如前文曾經討論過的張公藝。更明確地講，藉由相似文件，或可以開始探索契書的格式、形制和修辭，以及其所彰顯的意義。

舉例而言，若以「找洗」二字作為關鍵字，會在 THDL 中找到大量的契約。不過，這些契約並非單純的找洗契，反而是以杜賣契居多。而且，其中之所以提到找洗，大多是「不得找洗」，或「永不言找洗」等與找洗正好逆反的話語。以往對找洗

11　關於清代監生的早期研究，見許大齡（1984）。本書寫成於 1950 年代。

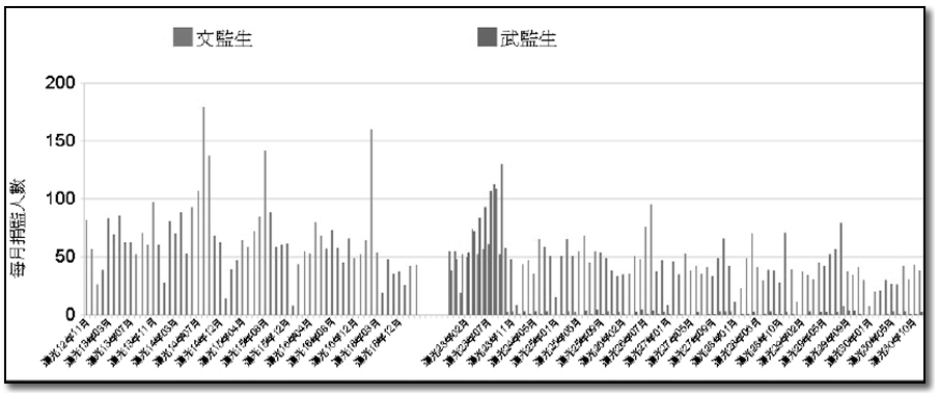

圖 17　文監生與武監生的比較

意義有興趣的學者，大多著重在找洗本身的行為，並以此證明傳統中國社會獨特的地權觀念。但買賣契約中大量對找洗持反對態度的文字，或許可以提供另一種觀察或思考的視角。一如人們在鬮分契中強調對張公藝九世不分家的欣羨，「永不言找洗」是另一種經濟社會生活中，概念與實際悖反分離的有趣現象。我們不應該理所當然將這些文字視為空洞的、無意義的套句，反而可以追問：為何它們會出現？為何會成為這種形式？

相似文件有時也會引發一些更龐大的問題。比如我們以電腦自動比對 THDL 中所有的相似文件，在「總督府抄錄契約文書」的 15,000 多件文件中，與戰後所蒐集、出版的古契書，竟然少有重複。這顯然是個不尋常的現象。理論上，若當時總督府土地調查過後，有將契約文書物歸原主，這些契約應該或多或少會出現在戰後幾波民間契約蒐集工作中。既然沒有，那就不能不讓人懷疑，這些契約最後何去何從？他們仍留在民間，沒有被發現？或是因為失去土地證明效力而被輕易的毀棄？或者，當初總督府抄錄過後，並未發還原地主？若是最後一項，那問題是，他們是否有被任何機構所接管？若非透過相似文件的功能，這樣的問題可能很難被發現。

七、結語

本文提出利用資訊技術，針對數位化的臺灣史料進行分析，提出一些有待深究、有待討論的問題。這些問題，部分是歷史本身的問題，部分則是史料的問題。這兩個面向有時是一體兩面的：哪些史料被存留下來，往往跟歷史的發展過程有緊密關係。

上述眾多問題可以分成四個部分。其一是從年代與空間分布的提問，尤其是前者。利用年代分析圖可以快速掌握一群資料的時間概況，但其分布狀況經常與預設有些差異，其間的落差因此有待進一步解釋；其二是詞頻的分析，我們舉了朱一貴與張公藝兩個不同文件集的例子。有時值得研究的問題不見得是在詞頻上位居前面的名字，反而是應該出現，而沒有出現的狀況。如嘉慶年間的郭百年事件，雖然對中部平埔族的發展造成巨大影響，但郭百年的名字卻罕見於明清檔案之中。換言之，除了注意史料中「所有」之外，我們也應該對史料中「所無」保持敏感度。

第三種形式的問題來自史料間的關係。這裡指的是，先利用電腦進行一些繁瑣的比對，進而重建出史料間的關係。我們從結果中，往往可以發現一些以人工難以發現的現象，如本文提出的土地移轉圖；最後一類則是從相似文件出發的問題。與前一項類似，相似文件也是利用電腦對文件進行大量的、重複的比對，這樣的工作交由人工，十分辛苦，對電腦而言卻是輕而易舉。在此我們看到電腦與人腦的差異所在。但另一方面，當電腦取代了重複卻單純的工作時，更複雜的問題唯有靠能夠深思的研究者才能處理。

透過上列疑問的提出，本文希望說明，在數位人文的研究中，資訊技術可以協助歷史學者發掘一些意外的問題。這些問題往往無法用簡單的方式回答，有時也許根本不可能回答。但這些新鮮的疑問可以引領我們進入不同的視野。如果我們接受這些提問有其意義，那麼，接下來我們也許可以開始探索，資訊科技有沒有可能幫忙解答這些問題？或者，該怎麼做呢？

參考文獻

王世慶，2004，〈臺灣民間和田野所存清代史料及其價值〉，《臺灣史料論文集》，臺北：稻鄉。

中國第一歷史檔案館，1980，《天地會》，冊 1，北京：中國人民大學。

中國第一歷史檔案館，1980，《天地會》，冊 2，北京：中國人民大學。

李文良，2004，〈土地行政與契約文書：臺灣總督府檔案抄存契約文書解題〉，《臺灣史研究》，11，頁 221-240。

李朝凱，2010，〈清代臺灣女性的家庭地位——以女性立鬮書為探討核心〉，《臺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。

柯志明，2003，《番頭家：清代臺灣族群政治與熟番地權》，臺北：中央研究院社會學研究所。

涂豐恩，2010，〈張公藝九世不分家——談臺灣分家鬮書中的修辭〉，《臺灣古文書學會會刊》，7，頁 11-20。

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。

許大齡，1984，《清代捐納制度》，臺北：文海。

陳支平，2009，《民間文書與明清東南族商研究》，北京：中華書局。

陳秋坤，1994，《清代臺灣土著地權：官僚、漢佃與岸裏社人的土地變遷 1700-1895》，臺北：中央研究院近代史研究所。

陳詩沛、杜協昌、項潔，2009 年 12 月，〈史料整體分析工具之幕後——介紹臺灣歷史數位圖書館的資料前置處理程序〉，「第一屆數位典藏與數位人文國際會議」發表之論文，國立臺灣大學。

張傳璽編，1995，《中國歷代契約會編考釋》，北京：北京大學出版社。

馮爾康，1993，《清史史料學》，臺北：臺灣商務。

項潔、陳詩沛、杜協昌，2009 年 3 月，〈臺灣古契約文書全文資料庫的建置〉，「第三屆臺灣古文書與歷史研究學術研討會」發表之論文，逢甲大學。

福建師範大學歷史系編，1997，《明清福建經濟契約文書選輯》，北京：人民出版社。

謝育平、楊龍廉、趙建宏、黃銘立、古馮文、林郁智，2009 年 3 月，〈使用詞夾子建立中文典籍分析加值服務〉，「銘傳大學 2009 資訊科技與實務研討會」發

表之論文，銘傳大學。

Braudel, F. (1976). *The Mediterranean and the Mediterranean World in the Age of Philip II* (Sian Reynolds, Trans.). New York: Harper.

Chen, S.-P., Hsiang, J., Tu, H.-C., & Wu, M.-C. (2007). On Building a Full-text Digital Library of Historical Documents. *Lecture Notes in Computer Science, 4822. Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers* (pp. 49-60). Berlin, Germany: Springer-Verlag. doi: 10.1007/978-3-540-77094-7-11