Research Tools for the Taiwan History Digital Library

Jieh Hsiang*

Taiwan History Digital Library (THDL) is a full-text digital library designed and built for research in Taiwanese history. It current contains three corpuses: Ming-Qing Archives (*MQA*) (37,831 Imperial Court documents about Taiwan, totally over 70,000,000 words), Dan-Xin Archives (*DXA*) (19,557 prefectural level documents, totally over 11,000,000 words), and old land deeds (*OLD*) (32,152 pre-1900 land and other deeds, totally over 16,000,000 words). They were collected and selected from over 250 sources and span over 250 years, from the 17th to the early 20th century. While *MQA* reflects how Imperial China dealt with events occurred in a remote island of the empire, *DXA* illustrates how Taiwan was governed at the local level and *OLD* shows activities at the grass root. Together, they reflect Taiwanese's political, social, economical, and ethnic history before the 20th century..

THDL was built with two goals in mind:

(1) To build a comprehensive full-text digital library of pre-1900 primary historical documents of Taiwanese history.

(2) To build a research platform for historians to study documents, observe relations, discover new research subjects, and conduct research.

As part of the second goal, we have developed a number of tools, which we shall describe in this presentation. They can be divided into three categories:

(1) **Referential tools**. These are tools that historians refer to when they conduct research. Traditionally they often come as reference books. Currently there are 3 such tools: *Calendar Converter between Chinese, Gregorian, and Japanese*; *Measurement Converter between Suzhou code and numerals*; *Qing Official Graph* that shows the Taiwanese government structure and officials during the Qing Dynasty.

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

(2) **Query analysis and presentation tools**. These are tools that analyze query results and present them in different ways. Currently we offer 4 such tools. They are *Tree Map Analyzer of Land Deeds*, *Appositional Term Analyzer*, *Geographic Locator of Land Deeds*, and *Event Analyzer*.

(3) **Relation Mining Tools**. These are tools that explore hidden relations among documents. These relations are usually very difficult to discover by human. We present 4 tools in this category. They are *Land Transitivity Graph Analyzer*, I*mperial Edict/Memorial Citation Analyzer*, *Land Deed Relationship Analyzer*, and *Dan-Xin Litigation Analyzer*.

臺灣歷史數位圖書館研究工具集

項潔 *

「臺灣歷史數位圖書館」(Taiwan History Digital Library，以下簡稱 THDL)，目前含有超過一億字的全文資料，主要包括「明清臺灣行政檔案」、「古契書」與「淡新檔案」。本系統除提供一般檢索功能外，最重要的特色是設計多種分析工具，包括檢索後分類、年代分佈圖與詞頻分析。利用這些工具，使用者得以在針對大量的檢索結果，進行觀察與研究，進而發現眾多潛在的議題。目前約已經有四百人透過線上機制，申請帳號，使用系統，包括許多臺灣史研究領域中的重要學者。

本文要簡要介紹針對THDL的資料，延伸發展的幾項工具。這幾項工具獨立於THDL，可以提供不同的介面，或更爲複雜的分析功能。但另一方面，他們與原來資料庫之間仍保持連結，研究者可以依照使用需求，選擇路徑，或是相互參照。工具集的頁面位於http://thdl.ntu.edu.tw/tools，亦可從THDL首頁連結。

以下，我們首先要介紹三種通用的「參考工具」，包括中西曆轉換對照查詢、清代臺灣文官官職表、蘇州碼轉換器。其次，則是四種「檢索分析工具」，包括前後綴詞查詢系統、總督府抄錄契書地區分析、總督府抄錄契書地理資訊、與文件年代分佈與事件標註器。最後是四種「關係探勘工具」，包括土地移轉查詢系統、臺灣相關明清檔案引用關係、契約文書買賣角色分析、與淡新檔案分析圖。

一、參考工具

首先介紹通用型的參考工具。這類工具是將傳統研究者時常查閱的參考工具書，加以數位化，並設計便於查索的介面。使用線上的參考工具，比起翻檢厚重的參考書，往往更有效率。

(一)中西曆轉換對照查詢

研究前近代東亞歷史的學者，最常碰到的問題，就是不同曆法之間的轉換。不少相關參考工具便應運而生。而本系統同樣提供中西曆轉換對照，但內容與設計則針對臺灣史研究。因此資料範圍和功能，涵蓋臺灣歷史中常用的年代，包括明、清、日本和中華

* 國立臺灣大學資訊工程學系特聘教授暨數位典藏研究發展中心主任。

民國。此外，也納入南明、太平天國與滿洲國等年號。上述幾種政權在某些時刻，便重疊在一起。舉例來說，西元 1868 這一年，同時是日本明治元年、清同治七年和太平天國十八年。[1]透過本系統西曆轉中曆的功能，可以輕易地對照不同政權間的年號。

（二）清代臺灣文官官職表

本系統可查詢清代臺灣文官官職的相關資料。使用者可以透過時間、人名、官名，三種方式進行檢索。與紙本參考書不同，本系統提供多維度的觀察工具，比如檢索特定人物時，同時呈現該人物的官職之間的上下屬關係，或是同一職位的前後任官員；也可顯示被查詢者所擔任過的官職列表。此外，本系統亦連結故宮人名權威檔，作爲參考。除了與臺灣相關之官職之外，並會列出此人在中國擔任的其他職位。凡此種種，都可讓使用者在查找時，得以觸類旁通。

（三）蘇州碼轉換器

蘇州碼是中國民間常用的計數方式，經常使用在契約文書的書寫中，如文中若提到土地、銀兩／元、重量、長度，其計數方式便會使用蘇州碼。現在一般人對蘇州碼大多已經難以辨識。此工具可將蘇州碼轉換爲阿拉伯數字，並幫助使用者解讀與蘇州碼一同出現的單位。如輸入蘇州碼「〒又三乂十」，便可直接轉換爲「6.9847」。

二、檢索分析工具

第二部份的工具，與 THDL 的資料有直接關係，許多功能在 THDL 中也可使用。但由於 THDL 功能眾多，部份進階功能操作上相對複雜，部份功能則不易放入現行版面中。我們因此將其中四項功能獨立出來。這四項工具皆是針對 THDL 資料進行檢索，並將提供便於觀察的呈現工具。

（一）前後綴詞查詢系統

在 THDL 查詢欄中輸入特定語法，可以檢索連綴在關鍵字前後的詞彙。比如，以明清臺灣行政檔案爲例，若想要了解這群文件中出現過多少種「某某主義」，以人力一一查閱，勢必大費周章。但電腦可以很快檢索出包括「帝國主義」、「民族主義」、「社會主義」、「均勢主義」、「國家主義」等 47 種主義，同時能把每個詞彙出現的次數加以統計。但由於操作上過於複雜，我們提供另一個簡單的工具介面，讓使用者進行同樣的分析。

---

[1] 因爲曆法的不同，清朝與太平天國與西元年的轉換會稍有差異，如清同治六年 12 月 17 日，便進入西元 1868 年 1 月 1 日。

（二）文件年代分佈與事件標註器

THDL 另一個方便的功能是文件年代分佈圖，使用者可以從分佈圖中，快速掌握檢索結果的概貌，而且往往能得到一些意外的線索或啓發。不過在 THDL 中的分佈圖較小，我們因此提供另一個年代分析工具，以便研究者進行比較細部的觀察。同時，在 THDL 中只能呈現一個關鍵字的年代分佈，或兩個關鍵字的年代比較。在此獨立的工具中，則可以比較三個以上關鍵字的年代分佈情形。

（三）總督府抄錄契書地區分析

接下來兩種工具則與地理資訊有關，其中涵蓋資料較少，只包括「古契書」中的「總督府抄錄契約文書」。這是因爲總督府在抄錄契書時，皆會註明該地契所指涉的地點（即地號），因此有著比較準確清楚的空間資訊。我們將這個這些空間資訊自動擷取出後，建置了兩個分析工具。在本工具中，使用者可以輸入特定詞彙，結果則利用 TreeMap 呈現。在 TreeMap 階層圖中，外層的區塊爲『堡』，內層的區塊爲『庄』，區塊越大的堡、庄，代表隸屬於此地的契書數量越多，同時呈現『庄隸屬於堡』的階層關係，與契書分佈的情形。

（四）總督府抄錄契書地理資訊

除了 TreeMap 外，我們也將地理資訊與 WebGIS 結合。此系統同樣可以用來分析臺灣總督府抄錄契約文書之地域分布，同時結合了日治與現代的地圖。使用者可以按照或特定時間或特定的契書種類，觀察該群契約的分佈；也可以空間範圍查找該地區所有的契約。

三、關係探勘工具

最後一個部份是關係探勘的工具。我們利用自動的方式，將明清臺灣行政檔案，或是古契書中的關係重建出來。

（一）土地移轉查詢系統

THDL 中有三萬多件古契書，包括一百多個不同的來源。其中有不少上下手契、鬮分契、原契與契尾、和契書內容相同的文件。我們利用自動的方法，重建上述的幾種關係，包括上下手契 2409 對、原契與契尾 92 對、鬮分契多份 878 組、契書內容相同 531 組。再將這些關係加以串連後，組合成「土地移轉圖」，數量超過 2,400 張。每一張土地移轉圖代表了一塊土地的身世，也就是該土地如何在不同的人手中流轉。有些土地移

轉圖包括數十張，甚至上百張的契約文書。本系統提供了一個方便檢視土地移轉圖的工具，並可以針對契約的關鍵字加以檢索。同時，呈現時亦提供人物、時間和地理資訊等相關資訊的提示，並可查閱契書原文。

（二）契約文書買賣角色分析

我們已將各契書的角色（如買方或賣方）及人名，以自動方式擷取出來。本系統即提供的工具，查詢個別人名在契書中擔任過的角色。檢索後會列出所有與此人相關的契書之標題、年份、時間、地點，和出現在同一張契書中的其他人物及其角色。同樣可以查閱契書原文。

（三）臺灣相關明清檔案引用關係

明清的行政檔案常有一定的書寫規範，如大臣的奏事文書中，爲了說明上奏的原因，會引述先前收到的諭旨命令作爲引言；而皇帝的諭旨文書中，爲了讓受命臣子瞭解命令的背景緣由，也會簡要摘錄此前的奏事內容，作爲引言。我們將上述奏事文書、與諭旨文書之間相互的引用關係，利用自動方式加以串連。在 37,817 件明清臺灣行政檔案中，共找出了 7,782 組「奏摺引用上諭」，以及 1,732 對 「上諭引用奏摺」，共連結出 1,172 張『引用關係圖』。本系統提供檢視這些引用關係圖的工具。系統預設已列出了 1,172 張引用關係圖，並以圖中所含檔案數量的多寡順序排列，呈現出引用關係圖的縮圖，與圖中涵蓋檔案的簡要資訊，包括年代、作者、人物、地點、出處。使用者也能輸入想查詢的人物、官名、地名、年代、關鍵詞，檢索出感興趣的關係圖。

（四）淡新檔案分析圖

戴炎輝教授曾對淡新檔案進行分類，將總計 1,164 案、19,281 件的淡新檔案，共分爲 3 編、16 類、102 款。本系統所提供的功能，是對淡新檔案的全文進行檢索，並將檢索結果上述分類之間的交互關係，以眼球圖加以呈現。如檢索「方祖蔭」一詞，可得到 167 案。其中 19 案出現在霸佔和鹽務兩類，有 14 案出現在隘務中。這個系統也提供多重詞彙的檢索，如可以同時輸入方祖蔭、劉銘傳、徐錫祉，比較他們各自出現在何種案類中。

結語

以上簡單介紹了總共十一個工具，全部可以從臺灣歷史數位圖書館工具集的首頁連結使用（http://thdl.ntu.edu.tw/tools）。我們相信這些工具對研究者會有許多助益。尤其在查找、觀察和比較方面，可以節省相當的時間成本。未來我們會繼續研發不同面向的工具，更歡迎學者將研究時的需求告訴我們，讓工具的開發更貼近使用者。