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

陳詩沛 *、杜協昌 **、項潔 ***

摘　要

「臺灣歷史數位圖書館」（THDL）是臺灣大學近年建置的一個大型全文史料數位圖書館，共含近八萬件的第一手臺灣史文獻，累積的全文字數已達約一億五千萬字。本文會對 THDL 納入史料時進行的資料前置處理方式做說明，並指出其獨特之處。一般數位圖書館在納入資料時，往往不會假設這些資料之間有顯著的關聯，所以在納入分批建置的新資料時，除了重新製作檢索索引（re-index）之外，並不需太多的前置作業。然而，為了讓臺灣史研究者能更有效地利用與分析 THDL 中的大量史料，我們發展了一系列的史料分析工具，這些工具需要對史料作整體性的分析，所以 THDL 對資料的處理方式遠比一般的數位圖書館複雜，尤其在輸入一批新資料的時候，需要與現有資料作整體、精細的前置處理，預先建立全體資料的分析資訊與資料之間的關聯性，讓研究者有能力深入大量的史料，進行分析與研究。這個前置作業程序也反映我們打造 THDL 的原則，即是一個好的數位圖書館不應僅是一個資料倉儲，而應是資料與分析工具之間無縫的結合。

*　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後研究員。
***　國立臺灣大學資訊工程學系特聘教授。

The Document Processing Workflow of THDL, Taiwan History Digital Library

Shih-pei Chen *, Hsieh-chang Tu **, Jieh Hsiang ***

Abstract

In this paper we describe the mechanism for incrementally incorporating new collections into the Taiwan History Digital Library, a digital research environment for research in Taiwanese history. In a system such as the THDL, where documents and tools are fully integrated, adding a new collection involves a lot more work that meets the eye. The challenge mainly comes from the need to provide a global view of the entire document set. Our procedure ensures that a newly added collection is fully integrated with the existing ones, and that a global view of all the documents in the digital library is always preserved. It also reflects our philosophy that a digital library is not merely a warehouse of digital objects but a seamless integration of documents and tools for research in the intended area.

* Ph.D. Candidate, Department of Computer Science and Information Engineering, National Taiwan University.

** Post Doctoral Fellow, 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

過去的十餘年來，全世界有著大量的數位典藏計畫在進行，[1] 運用數位技術將重要文化資產保存，並藉由網站系統的建置、透過網路，將文化資產的數位化版本提供外界使用，許多的數位圖書館與數位博物館便因應而生，這使得文化資產不再受到實體與時空的限制，得以透過網路被眾人取得和使用。

臺灣歷史數位圖書館（Taiwan History Digital Library，簡稱 THDL）也是在這樣的趨勢下所孕育而生的（Chen, Hsiang, Tu, & Wu, 2007）。這是一個專以「明清時期的臺灣歷史」為主題的數位圖書館，收錄了來自臺大圖書館與國立臺中圖書館所全文數位化的第一手臺灣史料，其資料量已達八萬件，全文字數約一億五千萬字。THDL 集合了大量的第一手全文史料於一處，並提供全文檢索，目的是希望這些平日不易取得的珍貴資料，能以數位化的形式，透過網路，提供免費學術使用，進而大幅降低歷史研究人員取得資料的屏障，促進臺灣史的研究。THDL 的最終目標是建立一個數位的研究環境（research environment），在其中提供許多輔助歷史研究的工具，讓歷史研究者可以方便地取得、分析、並探索史料。

在本文中，我們將介紹 THDL 在納入數位化資料時所進行的「資料前置處理程序」。一般數位圖書館所謂的「納入資料」，大致上僅是將史料的數位化檔案放置到系統的存放空間，並重新製作檢索索引（re-index），以確定納入的資料可以正常從系統上檢索、取得、使用得到（search, retrieve, access）。但由於 THDL 並不僅僅以一個單純的檢索系統自居，而是希望打造一個具備多樣化輔助研究工具的數位研究環境，因此，THDL 的資料前置處理程序比一般數位圖書館的作法更為複雜，其過程不僅將資料放入系統的存放空間、re-index、讓研究者可以透過系統檢索取得資料而已；而且得在過程當中建立全體資料的分析資訊與資料之間的關聯性，以供給 THDL 的多項研究工具，使其能基於這些資訊，為研究者提供史料的整體分析，讓研究者有能力深入分析大量的史料，探討史料間的關聯，發現研究議題。

值得一提的是，THDL 現在正以每個月八百件左右的速度，持續收錄新的數位化臺灣史料。分批納入新的數位化資料一事，對於一般的數位圖書館並無特別困難之處，因為一般的數位圖書館系統均當作資料之間是彼此獨立的，因此分批納入資料時，僅需將新的檔案放置到系統的存放空間，再確認新的資料能被系統取用即可，不會產生額外的困難。然而，在 THDL 中，新的數位化史料卻有可能跟已納入系統的舊有史料發生關聯，而這樣的關聯，需得在新進資料的前置處理過程中同步被建立，以確保 THDL 系統的狀態於任何時刻均是一致的（consistent）。這個與一般數

1 可參見 “American Memory” (The Library of Congress, 2000)；〈「數位典藏國家型科技計畫」二期計畫概述〉（數位典藏國家型科技計畫，2008）；“ JISC Digitisation Programm 2004-09” (JISC, 2009)；“Electronic Library Projects” (National Diet Library, 2009)。

位圖書館的不同之處，恰好反映了 THDL 的設計理念：一個好的數位圖書館不應只是一群數位化資料的集合，而應是一個資料與工具能緊密結合、提供真正能輔助研究之工具的「數位研究環境」。

在以下的章節中，首先我們簡單地展示 THDL 現有的整體分析工具，讓讀者對於 THDL 的資料前置處理程序將產生的成果有所概念。接下來，我們詳細說明新進資料前置處理程序的過程，並說明其如何搭配 THDL 的整體分析工具，為研究者提供有用的資料分析。最後，我們對前述的論述做一總結，帶出 THDL 的遠程目標。

二、簡介「臺灣歷史數位圖書館」的整體分析工具

THDL 是一個專以「明清時期的臺灣歷史」為主題的數位圖書館，收錄了明清時期各種第一手臺灣史文獻的數位化全文，目前總計已蒐集近八萬件的史料，全文字數共達約一億五千萬字。當中最主要的文獻集有二——「明清臺灣行政檔案」與「古契約文書」，前者是從明清兩朝中央政府的行政檔案中，挑選出與臺灣相關的史料（吳密察、翁佳音、李文良、林欣宜，2004a：73-99；邱婉容，2006：128-129），而後者是臺灣民間訂定的各種契約文書，以土地相關契約居多（吳密察、翁佳音、李文良、林欣宜，2004b：101-135）。這兩類文獻因為產生的背景不同，分別反映了明清中央政府看待臺灣的角度，以及臺灣基層人民的生活樣態。

面對如此大量、豐富的第一手臺灣史文獻，僅僅提供檢索，並不足以發揮這批史料的潛在價值，應該要進一步提供「能幫助歷史研究者有效利用大量史料、進行研究」的工具。為此，THDL 發展了一系列基於「群體」概念的輔助研究工具：歷史資料，對比於圖書、網頁、商品而言，彼此之間蘊含了更多的關聯性，而經由同一檢索條件所回傳的多篇史料，其彼此之間的關聯性應更為強烈。為了彰顯史料之間隱含的種種關聯，THDL 發展了「檢索後分析」的概念——我們認為，系統應於檢索之後，主動地為研究者分析回傳史料之「整體」所共同呈現的特徵，例如：史料主要分布的年代，當中常出現的人物與地點，以及該人物、地點與回傳史料之間的相依性等。

但除了被動地為研究者分析檢索結果，我們認為系統還應該要主動、預先的對整個文獻集進行整體分析，將整個文獻集中彼此有關的史料連結在一起，例如：把同一塊土地的上、下手契約連結在一起；或是藉由對整個文獻集進行整體的年代、人物分析，將具有相同特徵的史料連結，以幫助研究者能用新的角度探索 THDL 的大量史料。這些整體分析的結果，一方面可以幫助研究者快速地掌握史料整體的重要特徵，以及史料間彼此關聯的狀況；另一方面，它們更提供了研究者得以用「不同於閱讀」的「整體視角」來觀察史料，或許能協助研究者在大量的史料之中，發

現以人力閱讀時未能發現的議題。

THDL 的整體分析工具有檢索後分類、檢索結果年代分布圖、詞頻分析、相關文書、內容相似文件等，分述如下：

（一）檢索後分類

以年代、出處來源、作者、性質分類等四種面向，呈現回傳史料於每一面向的整體分布狀況，其畫面如圖 1。以「年代後分類」為例，我們以年號、年分兩階層的樹狀分類做為分類主軸，再將回傳史料依據其成文日期，分類到年號、年分所形成的子類別之中，並統計每個子類別含有多少篇此次回傳的史料。

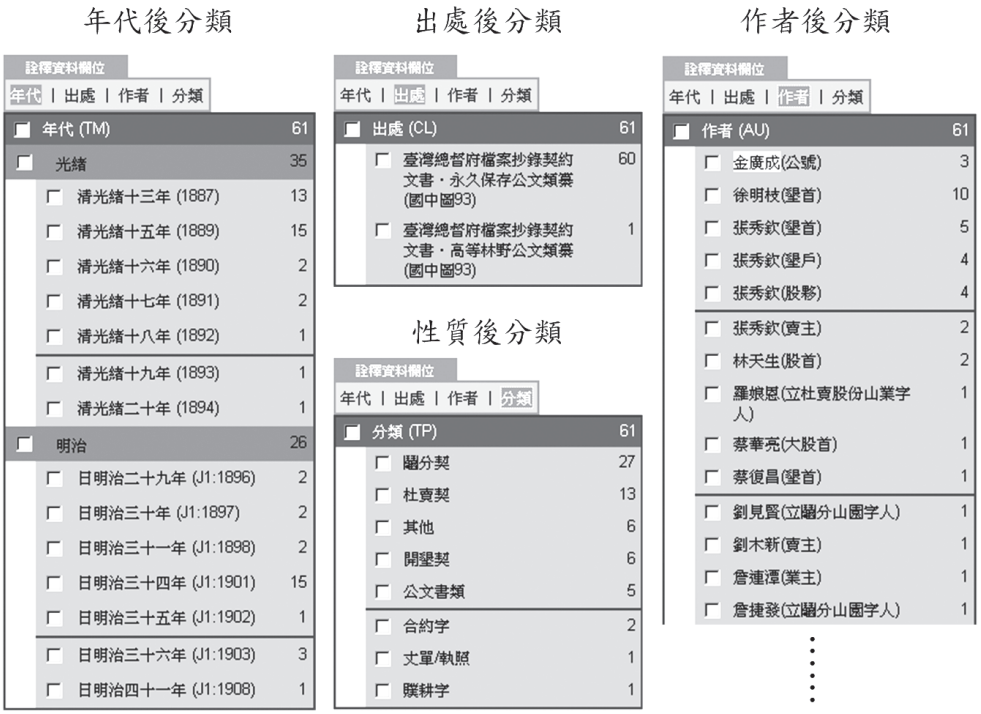
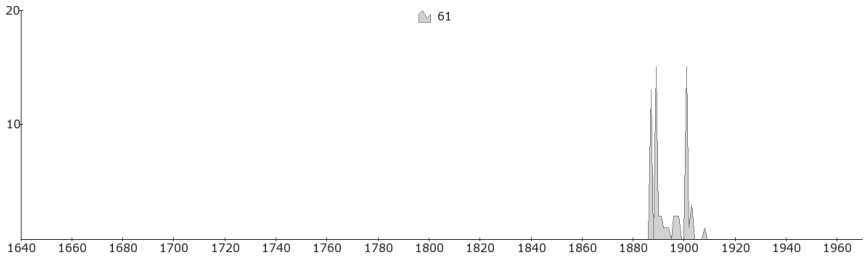


圖 1　THDL 的檢索後分類工具

（二）檢索結果年代分布圖

以視覺化的折線圖重新呈現「年代後分類」的結果，讓史家更能感受到檢索結果在時間軸上的分布狀況，如圖 2。



（僅計入含有年代資訊的文件；橫軸：西元年，縱軸：文件數量）

圖 2　檢索結果年代分布圖

（三）詞頻分析

進一步為研究者分析回傳史料的內容中所出現的人物、地點，或其他關鍵詞彙。分析結果以表格的方式呈現，每一橫列代表一個詞彙 t，每一欄位則代表該詞彙的統計數據，例如圖 3 中的 df 與 t → q 兩欄。df（document frequency）是統計詞彙在這次的檢索結果中、出現於幾篇史料，例如查詢「福康安」回傳的 1,550 篇史料中，「林爽文」一詞的 df 是 582，代表「林爽文」出現在其中的 582 篇。因而 df 愈高的詞彙，代表其愈常出現在此次回傳的史料之中，THDL 將 df 愈高的詞彙排列在愈上方，以便研究者觀察具有這樣現象的詞彙。而 t → q（t implies q）是計算詞彙 t 在整個文獻集出現的總篇數中，有多少比例是出現於此次的檢索結果，例如前例中「林爽文」一詞的 t → q 值，是將「林爽文」在此次檢索結果中出現的篇數 582，除以「林爽文」在整個文獻集中出現的總篇數 1,523，便得到 582/1523 = 0.38。藉由 t → q 一值，使用者得以觀察詞彙是否「唯獨地」（exclusively）出現於此次檢索結果之中：t → q 值愈大、愈接近 1 的詞彙，代表愈「唯獨地」出現於此次檢索結果。這可以幫助研究者了解到，究竟 t 是一個高度依存於此次回傳史料的詞彙、還是一個亦廣泛出現於其他史料中的詞彙。由上可知，t → q 一值的計算，也是參考了詞彙 t 在整個文獻集中出現的總篇數，將文獻集視為一個整體，才能知道這樣的資訊。

人名

Term t		df	t→q	回報
福康安	生平	1543	1.000	
李侍堯	-	689	0.427	■■■
林爽文	生平	582	0.382	
常青	生平	523	0.377	
柴大紀	生平	507	0.476	
莊大田	生平	356	0.687	
徐嗣曾	-	339	0.406	■■■
海蘭察	生平	319	0.982	
恒瑞	-	310	0.797	■■■
普吉保	生平	294	0.599	

地名

Term t	df	t→q	回報
臺灣	1187	0.058	■■■
閩	844	0.045	■■■
浙	661	0.051	■■■
福建	500	0.025	■■■
諸羅	425	0.285	■■■
粵	359	0.081	■■■
大里杙	310	0.529	■■■
鹿仔港	303	0.341	■■■
廣東	291	0.075	■■■
廈門	285	0.075	■■■

df：詞彙 t 於此次檢索結果中被提及的篇數　t → q：詞彙 t 僅出現於此次檢索結果的比例

圖 3　人名與地名的詞頻分析

（四）相關文書

歷史資料之間可能蘊含了豐富的關係，其中有些關係是史料產生的當下就已經存在的。例如契書之間可能存在有上下手契的關係，代表同一塊土地歷年的移轉、典賣、鬮分等不同的交易行為，是研究土地開發及地權轉移過程與模式的重要依據。再例如明清行政檔案之中，臣子的奏摺可能是為了回覆皇帝於上諭檔中下達的命令，而皇帝下達的命令也可能起因於另一位臣子所奏報的情事。將奏摺與上諭檔之間的關係釐清，或許可以追溯事件的發展與處理經過，了解當中的決策依據。然而，這些彼此相關的史料常因存放於不同的機構或家族，經過多年的流傳與轉手，最終落在不同的典藏單位手中，若要靠人力找出史料間散失的關係，需要花費相當大的力氣。因此，我們嘗試以自動的方式重建史料之間的關係，目前已經成功重建了契書之間的「上下手契」、「原契與契尾」、「鬮分契多份」的關係（黃于鳴，2009），並利用這些關係，將關於同一塊土地的契書串連起來，形成土地的完整交易歷史，如圖 4。

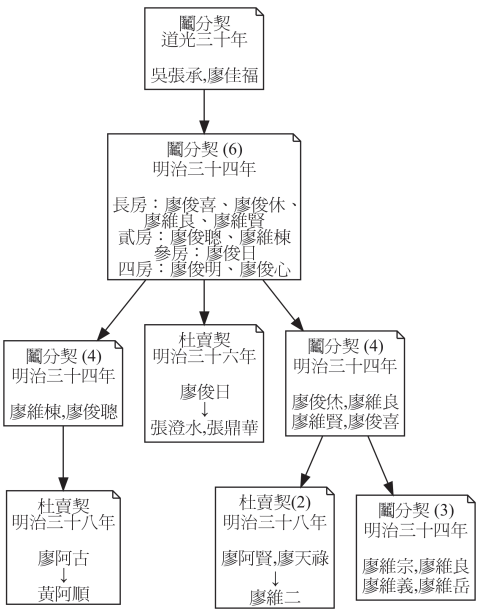


圖 4　利用契書間「上下手契」關係串連而成的契書群組

（五）內容相似文件

內容極為相似的文件，常是具有固定格式的史料，例如明清行政檔案中各省需例行奏報的「雨水糧價情形」與「監生銀數」，再例如契約文書中的「丈單」、「契尾」等官方制定格式的證明文書。THDL 為收錄的史料進行兩兩比對，找出了內容相似

的文件，如圖 5 左側的「為奏報閩省收捐監生銀數事附片」一文，竟有其他 148 件與之內容相似的文件（圖中框線處），皆是奏報監生銀數的文獻。再仔細對照其內容文句，僅在日期與金額等處有不同（圖中虛線所標示之處）。若非有這樣的工具輔助，人力或許很難發現史料間的這層關聯，更不會對其產生研究的興趣。而如今，藉由相似文件工具的幫助，研究者可以很容易地從大量的史料中發現同一格式的史料，或許便能對其進行更深入的研究，從中發現有趣的議題。

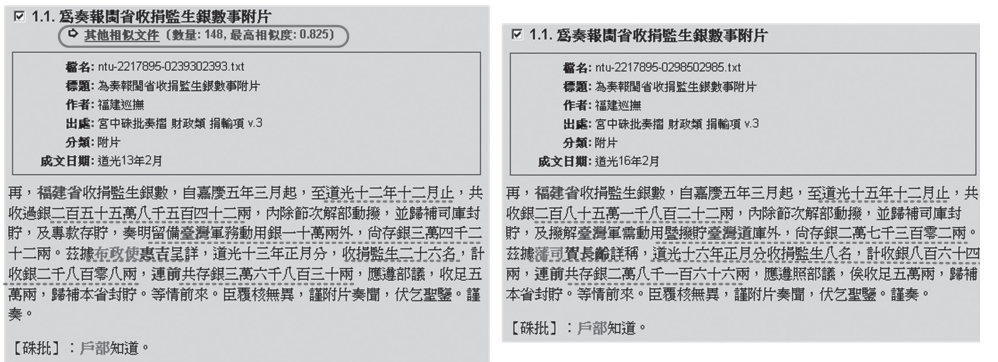


圖 5　透過 THDL 相似文件工具所找到的兩篇相同格式的史料

以上介紹的輔助研究工具，都是 THDL 欲幫助歷史研究者有效地利用大量史料、進行研究的嘗試，我們希望能協助史家，從整體的角度、以多樣化的面向，對史料進行觀察、分析。

三、THDL 的「資料前置處理程序」

THDL 之所以能提供如此多樣化的整體分析工具，全有賴我們的「資料前置處理程序」，對史料的數位檔案進行了細緻的整理。當史料的數位檔案建置完成、準備被納入系統之時，它們需要經過一連串的處理程序，從原本打字全文（full text）與詮釋資料記錄（metadata records）的數位檔案形式，轉化成為 THDL 系統工具後續能夠利用的形式。因此，THDL 的「資料前置處理程序」並不像一般數位圖書館的作法，僅是把史料的數位化檔案納入系統而已。THDL 除了納入史料的數位檔，更重要的是能匯集全體史料的資訊，進行統計、運算，最終建立資料的整體分析資訊，並建立資料之間的關聯性，兩者皆是不存在於個別史料的新資訊。

THDL 的「資料前置處理程序」可分為兩個主要的階段：第一階段是從個別史料的數位檔案——打字全文與詮釋資料——之中，辨識、擷取可進一步利用的內容特徵。第二個階段則是匯集所有史料於第一階段所得的內容特徵，加以統整、運算，得出文獻集的整體資訊，並於此時比對史料的特徵，據此重構史料之間的關係。此

階段所得到的新資訊，會連同史料的數位檔案一同儲存在 THDL 的後端資料庫，以供給 THDL 系統上的各式工具做進一步的運用，將個別的史料內容與來自史料整體的資訊，於適當的時機，以合適的搭配方式，呈現給研究者。

（一）第一階段：蒐集個體資訊

此階段的主要目的是整合史料的打字全文與詮釋資料，並且補充史料的詮釋資訊，以使分批建置而導致詮釋資訊不一致的各批史料，均能擁有豐富性一致、品質也一致的詮釋資訊，以供往後的處理程序所用。

第一步要做的是，將原本分散的全文（full text）與詮釋資料（metadata）合併。當一份史料完成數位化之後，通常會得到兩部分的檔案成品：打字全文與詮釋資料。以THDL所取得的數位化史料為例，這兩部分通常分別存放於不同的數位檔案之中。打字全文通常存放於一個文件檔之中（純文字檔、Microsoft Word 檔，或 PDF 檔）；而詮釋資料通常以整批史料集體的方式，存放在同一個表格檔之中（例如 Microsoft Excel 檔）。因此，第一步是將一件史料的全文與詮釋資料合併存放於同一個數位檔案之中，THDL 選擇以 XML（eXtensible Markup Language）作為此匯集全文與詮釋資料的載體（請見圖 6 與圖 7）。

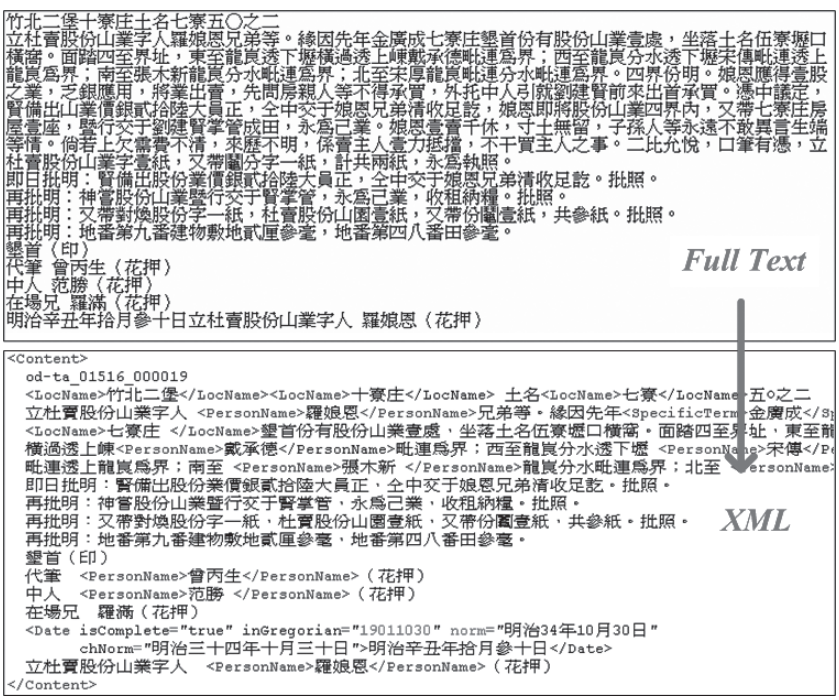


圖 6　將全文轉換為 XML

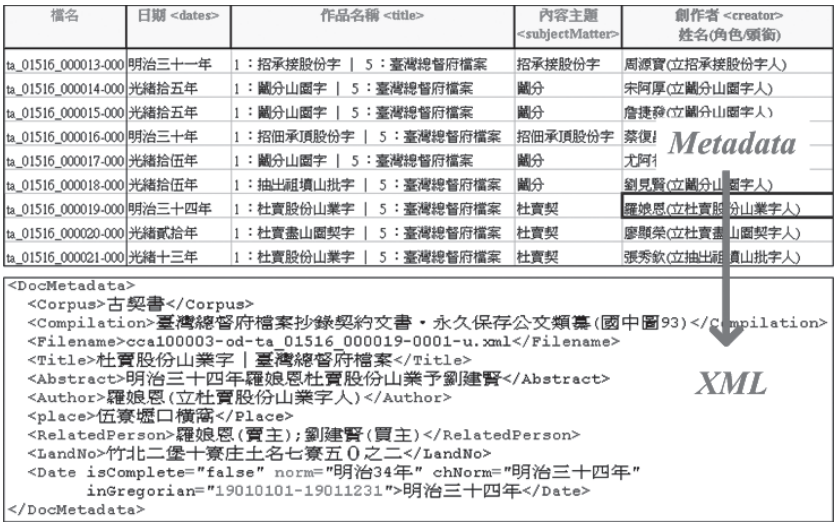


圖 7　將詮釋資料紀錄轉換為 XML

由於 THDL 的史料是由多批的數位典藏成果所彙集的，在分批建置的狀況下，各建置單位所使用的建置規範並不盡相同，因此建置出來的詮釋資料難免會有不一致的狀況，像是記錄的格式不一致（例如日期有多種書寫格式）；又如欄位名稱相同、但各數位化單位對該欄位的內容意義之定義不同（例如契書分類沒有一致的分類法則）；又如詮釋資料欄位的多寡不同，造成某批資料的欄位較其他資料為少（例如 THDL 的部分資料集沒有建置人名、地名欄位）。在這樣的情況下，為了對 THDL 收錄的史料皆有一定的掌握，我們決定盡量為所有的資料集，補充缺乏的資訊，並統一記錄的格式，俾使所有資料集都能有一致同等豐富的資訊。

1. 日期書寫格式的正規化

我們對所有的日期，進行了格式的正規化。史料中日期的書寫方式頗為多變，「雍正七年閏七月三十日」可能寫成「雍正柒年閏柒月參十日」——年分不以數字記年、而以干支記年，而中文數字的寫法亦有簡體、繁體之可能。為此，我們先是發展了辨識各種日期寫法的方法，能夠辨識日期中的年號、年、月、日各個區段，然後將年月日部分一律轉換為中式簡體的數字，[2] 以及阿拉伯數字兩種約定的標準格式。我們並將辨識日期的方法應用在史料的內文之中，擷取所有內文中提到的日期，並且也同樣轉換成為約定的標準格式。此外，為了較妥善地處理臺灣因從清治時期跨越到日治時期，所產生農曆日期、陽曆日期轉換的問題，我們也為所有中曆及日本曆的日期找出其對應的西元日期，以作為比較的一致標準。以上對日期的各

2 指「一二三四五六七八九十」字。

種轉換，都以 XML tag attribute 的方式，連同史料中日期的原始寫法，一併記錄在 <Date> 這個 XML tag 之內。[3]

2. 契書分類一致化

各單位所建置的契約文書，雖都有分類相關的欄位，但對該欄位的內容意義卻無一致的認定準則，造成 THDL 雖然蒐集了數萬件的契約文書，卻難以對其進一步處理、運用。因此我們參考了專家學者的建議，根據契約文書的性質，初步訂定了簡單的十四個分類，並發展一套自動分類的方法，將 THDL 的契書自動分類到這些類別之中（盧家慶，2008），便得以針對其中的幾類契約文書，進行更細緻的特徵擷取。像是對於買賣性質的「杜賣契」，便可進一步擷取買方、賣方的姓名；而對土地相關的契書，便可擷取土地座落、四至界址、面積、大小、價格等的土地特徵，這些特徵是用於建構契書關係的重要資訊，因此，亦將之記錄於 XML 之中（如圖 8）。

```
<LandFeatures>
  <DeedType>杜賣契</DeedType>
  <Seller>羅;娘;恩</Seller>
  <Buyer>劉;建;賢</Buyer>
  <Place>伍;寮;堘;口;橫;窩;十;七</Place>
  <Placenum>五〇之二;九</Placenum>
  <Time day="19010101" month="190101" year="1901" dynasty="1868"
        timelevel="year">明治三十四年</Time>
  <IsInherited>False</IsInherited>
  <LandPrice number="26" unit="大員">銀貳拾陸大員正</LandPrice>
  <Size>0.023;0.003;0.026</Size>
  <Edge_E>龍崑</Edge_E>
  <Edge_W>龍崑</Edge_W>
  <Edge_S>張木</Edge_S>
  <Edge_N>宋厚</Edge_N>
</LandFeatures>
```

圖 8　為土地相關契約文書擷取出更細緻的內容特徵

3. 人地名欄位之補充

因有許多資料集的詮釋資料皆未建置人名、地名相關欄位，我們也發展了一套方法，自動從史料的全文之中，擷取了 90,948 個人名與 3,496 個地名（張尚斌，2006）。

經過上述的資料彙集、整理、格式轉換、特徵擷取等種種處理之後，所有的史料便都已被轉化為一致的 XML 形式了，在每篇史料的 XML 檔案中，除包含史料的全文與原始詮釋資料，更含有從全文自動擷取出的重要內容特徵，並且已正規化為一致的標準格式。即便部分詮釋資料欄位有不足的狀況，我們也於此階段中將缺乏的資訊補上。到此，THDL 所有的史料都具有了一致且同等豐富的資訊，以便下一

3 我們在 XML 中記錄日期的方式如下：<Date isComplete=" true" inGregorian=" 17290922" norm=" 雍正 7 年閏 7 月 30 日" chNorm=" 雍正七年閏七月三十日" > 雍正柒年閏柒月參十日 </Date>

階段進行的群體處理。

（二）第二階段：匯集全部史料，建立整體資訊

這個階段是 THDL 之所以有能力提供整體分析工具的關鍵。在這個階段中，我們匯集所有史料，進行群體的分析運算，最終得到整個文獻集的整體資訊。更利用上一階段於史料全文中擷取的細緻特徵，進行特徵的比對，據此建構史料之間的關係。

THDL 的多項整體分析工具均是基於同樣的「群體」概念：其所利用的資訊並非僅來自於單一史料，而是集合了全體史料才能得到的。例如「詞頻分析」工具當中，各個詞彙的 df [4] 是匯集了所有檢索結果中的史料才計算出來的，而為了進一步計算 t → q，[5] 甚至得匯集整個文獻集，得到詞彙 t 於整個文獻集出現的總篇數後，才能計算。

其中，「檢索後分析」的作法是，利用第一階段所擷取、標示於 XML 中的史料特徵，依據不同的檢索後分析工具所要分析的特徵，匯集所有檢索結果中的史料，取出該特徵的 XML tag，記錄每一種特徵值出現於哪些史料，供給前端的「檢索後分析」工具取用，以便其能快速地整理出對研究者有幫助的分析結果。以人名之「詞頻分析」為例，我們匯集所有 XML 中的人名標籤之後，依據人名之不同，記錄其出現於哪些史料之中，並計算其出現的史料數量，最後便得到所有 XML 中人名的統計數量表。而 THDL 的詞頻分析工具便是根據這個表，繼續計算得到 t → q（請參見圖 9）。

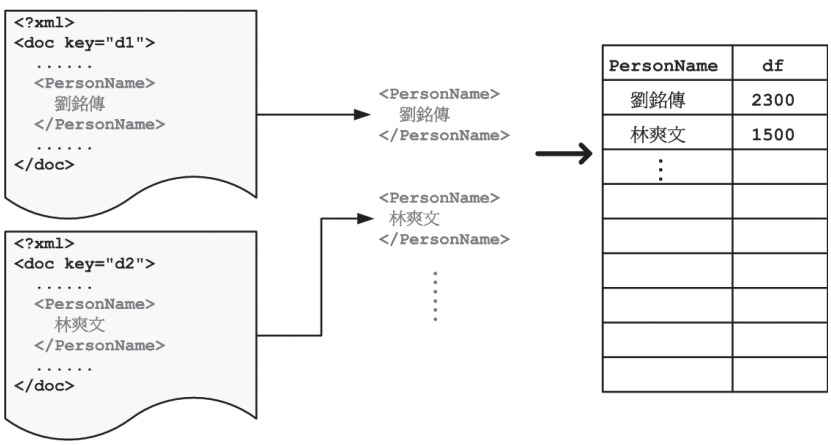


圖 9　匯集全部 XML 中的人名標籤後，便可統計各人名出現的史料篇數

4 指該詞彙於這次的檢索結果中，出現於多少篇史料之中。請參見前文「詞頻分析」所舉的例子。
5 此值代表詞彙 t 有多麼「唯獨地」出現於這次的檢索結果之中。請參見前文「詞頻分析」中的說明。

而欲建構史料之間的關係，則更需要基於「群體」的概念，以整個文獻集作為搜尋比對的基礎，才能放大找到史料關係的可能性。然而，史料間關係的建構，需要遠比統計、加總更為複雜的比對與運算。例如，史料之間的「內容相似」關係，乃是利用字串比對的方式，兩兩比對史料內容中文句的書寫順序，計算出兩篇史料內容的相似程度，再據此數值建構史料間的「內容相似」關係。而契書之間的「上下手契」關係，又需要更為複雜的比對方法，我們利用第一階段所擷取的、土地契約文書之各種特徵——契書性質分類、人物及角色、立契時間、土地座落、四至界址、大小、價格等，兩兩比對兩份契書的各種特徵是否相符，據以判斷兩者是否為同一塊土地的上下手契，或是一群人分割土地的鬮分契約（黃于鳴，2009）。建構出來的史料關係會先預存起來，供系統前端的「內容相似文件」工具與「相關文件」工具運用。

當所有的史料皆通過 THDL 的資料前置處理程序之後，會產生關於全體資料的分析資訊，以及史料間的關係，這些資訊都是 THDL 在資料前置處理程序中，集合了所有史料，自行整理、產生的重要知識，不是用個別的史料便可以得到的。這些資訊，連同史料的全文、詮釋資料，一起儲存於 THDL 的後端資料庫之中，以供 THDL 系統前端的整體分析工具根據這些材料、進行加值運算，為研究者呈現有幫助的分析結果，其示意圖如圖 10。例如將人名的統計資訊用於詞頻分析，可幫助研

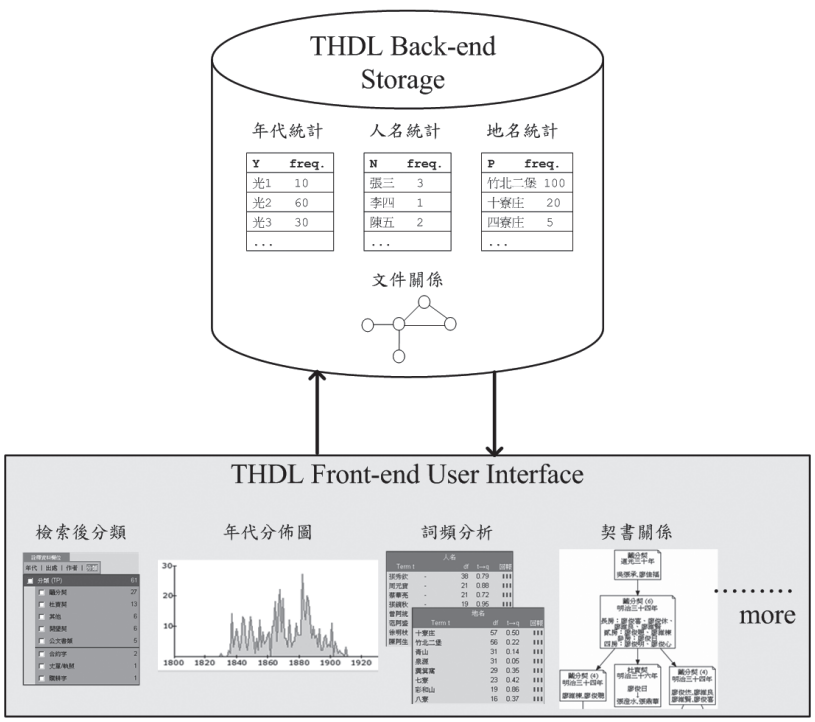


圖 10　將預先運算好的整體資訊儲存下來，以供 THDL 整體分析工具運用

究者了解該人物與此次檢索結果的相關性與相依性；又如建構出的兩兩契書關係，可據以將同一塊土地的相關契書全部串連起來，形成契書關係圖，幫助研究者了解該塊土地歷年的交易全貌。

THDL 仍在持續地收錄更多數位化臺灣史料，目前大約每個月都會有八百筆左右的新進史料加入 THDL。然而，新納入的史料可能跟已經存在於系統的其他史料發生關聯，而新資料的加入也改變了 THDL 的「整體」資料，因此，我們透過「資料前置處理程序」的控管，在新進史料納入系統之前，重新建立 THDL 的整體資訊，也重建新舊資料之間的關聯，以確保 THDL 系統的整體分析在任何時刻均是 consistent 的。

新進史料的處理流程如下：首先，新進史料的數位檔案需跟其他史料一般，通過資料前置處理程序的第一階段，確保這批史料含有跟其他史料一致豐富的資訊，才能進入第二階段繼續處理。而在第二階段中，我們集合所有的新、舊資料，重新運算、建立 THDL 的整體資訊——包括統計分析、內容相似關係、契書關係等，最後得到的「整體資訊」必然會與先前不同，應該更為豐富。也就是說，THDL 每納入一批史料，就會產生更加豐富的整體資訊，整個資料庫蘊含的資訊是會一直不斷演進的。

四、結語

在此篇論文中，我們介紹了 THDL 史料整體分析工具的幕後推手——「資料前置處理程序」。每一批納入 THDL 的史料，都需經過這樣的處理程序，以確保先、後納入的所有史料，均含有一致且豐富的資訊量。也正是基於這個處理程序，匯集了來自所有史料的大量資訊，預先進行了分析、運算以得出關於整個文獻集的整體資訊，使 THDL 的分析工具能夠與全體史料緊密結合，提供快速且穩定的史料整體分析，讓研究者得以深入大量的史料，探討史料間的關聯，甚至發現新的研究議題。

THDL 強調歷史資料之間的種種關聯性，因此設計了許多能反映、彰顯史料關聯的研究工具，供研究者利用。這跟一般數位圖書館將館藏均視為彼此獨立個體的觀點大相逕庭。一般的數位圖書館，僅以提供館藏數位資料的檢索、取得、使用（search, retrieve, access）為最主要的目的；而 THDL 雖名為數位圖書館，其真正的目標卻是為歷史研究者建造一個「數位研究環境」，在其中提供各種輔助研究的工具，包括檢索工具與各式整體分析工具，才能有效地幫助研究者利用大量的資料進行研究。我們認為：一個好的數位圖書館不應只是一個集合了數位化資料、提供研究者查找資料的地方，而應是一個資料與研究工具能緊密地結合、提供真正能輔助研究的研究環境。

值得一提的是，在 THDL 中的史料整體分析工具，均是將「明清臺灣行政檔案」與「古契約文書」兩文獻集分開、各自進行分析統計的。舉例來說，圖 3 查詢「福康安」一詞所得的詞頻分析結果，是僅從「明清檔案」文獻集統計出的數據，並不是對 THDL 的全部史料進行統計。這樣做的原因，需回到 THDL 收錄史料的理念來說：我們為 THDL 選擇史料的原則，是希望藉由與臺灣相關的第一手史料之匯集，勾勒出臺灣史的全貌。「明清臺灣行政檔案」涵蓋的是明清中央政府的行政檔案當中與臺灣有關的資料，這些檔案具有同質性，整體勾勒出的，是從官方視角出發的臺灣歷史；而「古契約文書」涵蓋的則是臺灣人民簽訂的契約及相關文書（主要與土地相關），整體勾勒出的，是臺灣的土地移轉的歷史。這兩類史料的性質與意義均截然不同，若將兩類文獻混和、一起進行整體分析，將使得分析的結果難以被解讀、甚至根本不具意義。例如：在明清檔案中查某關鍵詞，所得結果反映的是官方對此的態度，但若在古契約文書查得同樣的關鍵詞，反映的卻是其如何影響臺灣人民訂定契約的行為，兩種結果若混在一起，恐將漏失各自的整體意義。

THDL 的目標，是希望收錄足以勾勒出臺灣史全貌的臺灣相關史料。可惜此一目標還有很多需要努力的地方——從時代的廣度來說，THDL 目前只涵蓋了臺灣明清時期的史料（主要為清代，部分的古契書跨越到日治時期），並未收錄明鄭以前荷蘭人留下的史料，亦缺乏日治時代豐富的臺灣史料；再從史料種類的廣度來說，在這兩個文獻集之外，也仍有許多常被臺灣史研究者參考利用的材料，未被 THDL 收錄；若再提到更大的歷史觀察架構，例如把臺灣史放在整個中國的脈絡來看，甚至是放在全球大航海時代的脈絡下來看，那麼 THDL 現有的材料更是不足。但我們仍希望，透過 THDL 目前收錄的史料，加上多種的史料整體分析工具，或許已能幫助歷史學家掌握兩個部分的臺灣史內在脈絡了。至於臺灣史全貌的展現，仰賴於更廣大臺灣相關史料的蒐集、彙整，與數位化，則是我們今後努力的方向。

參考文獻

吳密察、翁佳音、李文良、林欣宜，2004a，〈明清臺灣檔案〉，《臺灣史料集成提要》，臺北：文化建設委員會。

吳密察、翁佳音、李文良、林欣宜，2004b，〈清代契約文書〉，《臺灣史料集成提要》，臺北：文化建設委員會。

邱婉容，2006，〈明清臺灣行政檔案全文建檔及標注計畫〉，《中華民國圖書館年鑑》，臺北：國家圖書館。

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。

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。

盧家慶，2008，《臺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。

數位典藏國家型科技計畫，2008，「數位典藏國家型科技計畫」二期計畫概述。網址：http://www.ndap.org.tw/1_intro/intro.php

Chen, S. P., Hsiang, J., Tu, H. C., Wu, M. (2007). On Building a Full-Text Digital Library of Historical Documents. In D. H. L. Goh, T. H. Cao, I. Sølvberg & E. M. Rasmussen (Eds.), *Lecture Notes in Computer Science: Vol. 4822. Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, 10th International Conference on Asian Digital Libraries, ICADL 2007, Hanoi, Vietnam, December 10-13, 2007, Proceedings* (pp. 49-60). Berlin, Germany: Springer.

JISC. (2009). *JISC Digitisation Programme 2004-09*. From http://www.jisc.ac.uk/whatwedo/programmes/digitisation/projects.aspx

National Diet Library. (2009). *Electronic Library Projects*. From http://www.ndl.go.jp/en/aboutus/elib-project.html

The Library of Congress. (2000). *American Memory*. From http://memory.loc.gov/ammem/index.html