

從全文到表格：地方志職官資料的擷取

彭維謙¹、程卉²、陳詩沛³

摘要

本論文介紹一個針對中國地方志開發的電腦輔助全文標記系統，以幫助歷史學者及編輯人員從文本中有效率地蒐集「何人什麼時候任什麼官」的紀錄，將全文轉換成表格，作為大數據分析運算的基礎。中國地方志數量龐大，各地方志在描述「職官」時有大略一致的格式，然而在細節處差異仍大，難以運用電腦以統一的格式擷取各地方志的任官資料。本系統提供簡單易用的人機介面，內建擷取人名、官名、朝代、任職時間、籍貫、入仕等資訊的預設格式，讓使用者能簡單地按一個鍵套用，系統即辨識出文本中大多數的人名、官名、朝代、任職時間、籍貫、入仕等資訊，再供使用者校對。系統更提供使用者能因應文本格式差異、修改擷取格式的功能，以便從不同地方志中均能擷取到正確的任官資訊。

關鍵字：地方志、中國史、數位人文、資訊擷取、正規表達式

¹ 國立台灣大學資訊工程學碩士

² 上海紐約大學學術事務部助理

³ 德國馬克斯普朗克科學史研究所研究員

From Text to Data:

Extracting Posting Data from Chinese Local Gazetteers

Wai-Him Pang⁴, Hui Cheng⁵, Shih-Pei Chen⁶

Abstract

This paper introduces a semi-automatic text tagging interface to help historians efficiently garner posting records from the Chinese Local Gazetteers (*difangzhi* 地方志) with the format of “who, when, which posting”. By turning texts into tabular data forms, this interface aims to lay the basis for analyzing Chinese local gazetteers on a large scale. Although local gazetteers from various locations all follow a general pattern when recording posting data, they still differ in details due to their substantial amount. Therefore, it is unfeasible to ask programmers to extract the posting data using a one-size-fits-all computer program. This tagging interface, on the other hand, provides a simple user interface with built-in patterns to extract the subjects’ names, posting titles, dynasties, posting times, basic addresses and entry methods. This allows users to tag most of data in the text quickly, which can then be proofread by them themselves in order to check the tagging result and to correct mistakes. The interface also enables users to adjust the extraction patterns according to each text in order to accurately extract posting data from local gazetteers with distinct patterns.

Keywords: Local gazetteers, Chinese history, digital humanities, data extraction, regular expressions

⁴ Master of Computer Science, National Taiwan University

⁵ Senior Associate to Academic Affairs, New York University Shanghai

⁶ Digital content curator, Max Planck Institute for the History of Science

一、序論

（一）研究背景

中國歷代人物傳記資料庫（China Biographical Database，簡稱 CBDB⁷）是個專門收錄中國歷史人物傳記的資料庫，此資料庫不同於一般的史料全文資料庫，它並不收錄傳記文的全部文字，而是將傳主的人名、生卒年、籍貫、別名、出身（入仕）、任官、著作、親屬、社會關係人等不同類型的傳記資料，分別收錄在不同的資料表內，再透過各種代碼——人物的代碼、地點的代碼、官職的代碼、親屬關係及社會關係的代碼等，將資料在不同表格間串連起來，形成具有結構性的「關聯式資料庫」（relational database），這種儲存資料的模型允許研究者跳脫傳記文歷來從單一人物切入的視角，轉而能從傳記資料的其他面向維度切入，對資料庫內全部人物的傳記資料進行整合性、大規模的深度查詢，例如可以詢問「南宋所有進士的籍貫」這種不鎖定單一人物的大規模問題，這類問題是全文資料庫所難以回答的問題類型。截至 2013 年 10 月止，CBDB 已收錄從唐代到晚清共三十二萬人的傳記資料。

早期 CBDB 收錄資料的方式，是人工錄入的方式，由研究者或具備知識背景的編輯人員閱讀人物傳記，手動將傳記內容提到的各類資訊分別錄入到不同的資料表去。從 2011 年起，CBDB 開始與資訊學者合作，運用電腦從數位化的傳記全文中自動擷取傳記資訊，再交由編輯人員校訂，此舉大量提升了資料蒐集的速度。能這麼做的原因是，傳記文具有一致的寫作方式，不論是正史內的傳記、墓誌銘、或是當代學者編纂的傳記資料索引，對於 CBDB 收錄的人名、生卒年、籍貫、別名、入仕、任官、著作、親屬、社會關係人等客觀性傳記資料上，都有相當規律的寫法，因此我們能運用「正規表達式 (regular expressions)」⁸來描述上述資料在傳記文中寫法的規律性，再讓電腦按照規則、大規模地

⁷計畫網址位於 <http://isites.harvard.edu/icb/icb.do?keyword=k16229>。中文查詢系統網址位於 <http://db1.ihp.sinica.edu.tw/cbdb/ttsweb?@0:0:1:cbdbkm>。

⁸ 英文稱為 regular expressions，為一種描述文字出現規律性的電腦語言，請詳見 Regular Expressions Info 網站 <http://www.regular-expressions.info/>，或維基百科的 Regular Expression 頁面 http://en.wikipedia.org/wiki/Regular_expression。

到傳記文中辨識並擷取符合規則的資訊，找出傳記文中人名、生卒年、籍貫、別名等各類傳記資料的所在位置。CBDB 運用這套方法處理了宋人、元人、明人傳記資料索引⁹，從中增加了五萬個人物及其相關傳記資訊到 CBDB 的資料庫。至此，CBDB 已經收錄正史中的列傳與傳記資料索引，需要新的資料來源，因此我們將目標轉向了中國地方志。

（二）研究文本：中國地方志

正如傳記一樣，地方志是中國歷史產出的特殊文類，紀錄著地方上（通常是行政區）從建置歷史、地方政府編制、建築古蹟、山水等自然地理、物產、動植物、氣候災害，到官員名冊、歷朝進士錄、人物傳記、藝文等各種類型的知識，是地方官員了解其管轄區域的重要參考書。中國地方志的數量龐大，¹⁰地域分布廣，不論是省、路、州、府、縣、甚至城鎮或軍事營地，只要是擁有了一定人口、或是有治理目的的地域，幾乎都會編纂方志。並且，大多數方志的內容都涵蓋有上述列出的主題，如果想蒐集中國歷史廣大地域上關於特定主題的資料，現存方志便是個很好的資料來源。CBDB 便將目標鎖定到了方志中的「職官志」，這個部分詳列了該地區每個朝代、年號各設置哪些官職，歷任官員的姓名、籍貫、入仕、與任職時間，有時甚至提供官員更多的個人背景資訊，譬如擔任此官職前、後擔任的其他官職，有無傳記等，能夠提供 CBDB 極其豐富的「人物任官」資料，即：「何人、於什麼時候、擔任何種官職」，例如下圖為《惠民縣志》¹¹職官志的首頁。

⁹ 昌彼得、王德毅。1974。《宋人傳記資料索引》。臺北：鼎文書局。王德毅、李榮村、潘柏澄。1979。《元人傳記資料索引》。臺北市：新文豐出版公司。國立中央圖書館。1965。《明人傳記資料索引》：臺北：國立中央圖書館。

¹⁰ 據推測，現存的方志有八千種，請參考「中國大陸各省地方志書目查詢系統」
<http://webgis.sinica.edu.tw/place/>。

¹¹ 光緒年《惠民縣志》，（清）沈世銓修（清）李昺纂，清光緒二十五年柳堂校補刻本。

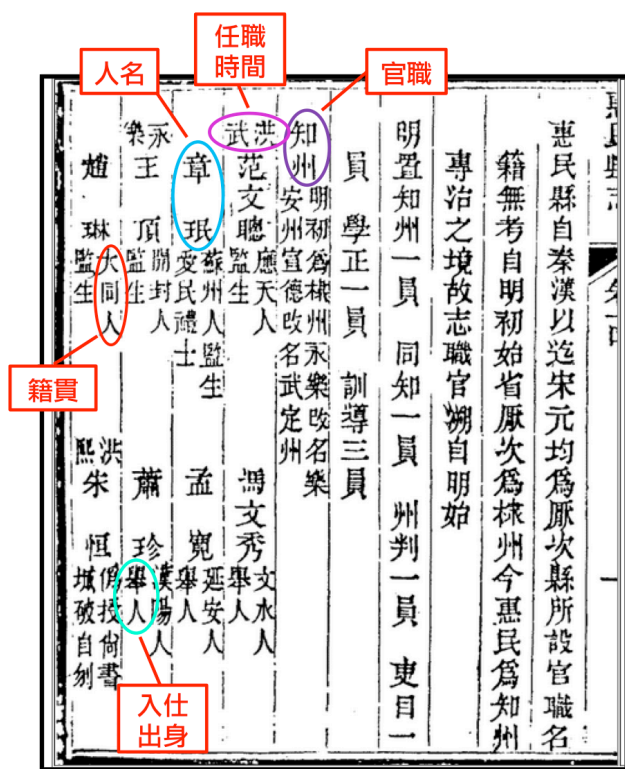


圖 1 《惠民縣志》職官志

從上圖可見，在書寫上，職官志大致是清單的形式，目的為列出所有曾任地方官員的人，但在官職、人員眾多的情況下，會以一定的分類架構、依序將人員列出。以《惠民縣志》來說，其先以官職為軸，各官職底下再按時間列出各朝代、各年號，最後於年號底下列出曾經擔任此官職的人物，記其姓名、籍貫、出身入仕，偶有其他關於此人的事蹟記錄。

職官志的寫作格式簡單，並非長篇敘述性的文字，僅將重點資訊條列出，非常適合資料的蒐集。然而，隨機觀察另一本方志的職官，便會發覺各方志對於職官的寫法，即便都包含類似的資訊量：官職、朝代、年號、姓名、籍貫、出身入仕，然而細節的資訊量、甚至給出資訊的順序，並不規律，以圖 2《新會縣志》¹²為例，雖然同樣先以官職為軸，官職底下按時代條列人員，但並不列出朝代，僅列年號，並且各人物有了詳細任職年份的記載，如「二十三年」，此指「道光二十三年」，在此，年號與年份並未寫在一起，這也是

¹²道光《新會縣志》，（清）林星章修（清）黃培芳纂，清道光二十一年刻本。

自動擷取正確任職時間時需考慮的問題。再以圖 3《濟南府志》¹³為例，其職官志並不先按照官職排，而是以朝代為軸，將各朝曾任官職的人員都列在一起，再於人名底下列其所任官職及任職時間。這種每本方志寫法皆有少許差異的情況，導致了難以針對所有方志設計一組統一的擷取規則（可含多條），讓電腦程式讀取以自動進行抓取。

人名	官職	任職時間	籍貫	入仕出身
許寶珩	知縣	前任言良鈺以上具載林志	湖南善化人	八二年署
聶爾康	知縣	江西萍鄉人舉人二十三年署	湖南衡山人	院庶吉士九年署十年補
帖臨藻	知縣	山東臨縣人進士二年任有傳	山西應州人拔貢六年署	
陳應聘	知縣	山東臨縣人進士二年任有傳	山東臨縣人	
胡湘	知縣	人監生二十九年代理	湖南人	
邱才穎	知縣	福建光澤人舉人二十七年任	福建光澤人	
陳壽頤	知縣	山人進士二十五年任	山人	
彭邦晦	知縣	山人監生二十三年署	山人	
文晟	知縣	江西萍鄉人舉人二十三年署	江西萍鄉人	

圖 2 《新會縣志》職官志

¹³道光《濟南府志》，（清）王贈芳修（清）成瓘纂，清道光二十年刻本。

濟南府志 卷二十二 隋 秩官略備左

而載之其未備者以新唐書世系表益之而有隋三十八年之

人名	官職	任職時間	籍貫	別名
張喬	刺史	齊	齊州	張喬
元寂	刺史	齊	齊州	元寂
裴頌	刺史	齊	齊州	裴頌
王頌	刺史	齊	齊州	王頌
公孫景茂	刺史	齊	齊州	公孫景茂
王伽	刺史	齊	齊州	王伽
乞伏慧	刺史	齊	齊州	乞伏慧
韋藝	刺史	齊	齊州	韋藝
樊子蓋	刺史	齊	齊州	樊子蓋
趙軌	刺史	齊	齊州	趙軌
劉宏	刺史	齊	齊州	劉宏
盧賁	刺史	齊	齊州	盧賁
辛遵	刺史	齊	齊州	辛遵
張齋	刺史	齊	齊州	張齋
張須陁	刺史	齊	齊州	張須陁
房恭懿	刺史	齊	齊州	房恭懿

圖 3 《濟南府志》秩官志

另一個欲以程式自動擷取的困難是，職官志條列人員時通常會將重複的資訊省略不列，像樹狀分類一樣，譬如圖 1 中的「知州」、「洪武」、「永樂」、「洪熙」均僅列第一次，但是電腦需要明確的規則才能判斷哪些人員屬於相同的分類——哪些人都是「知州」、哪些人是「洪武」、「永樂」、「洪熙」年間任職；若採取撰寫電腦程式來自動擷取資訊的作法，則需先由程式員觀察規則，將規則寫進電腦程式內，電腦再根據程式中的規則去擷取資料。然而，我們發現，幾乎每本方志的寫法規律均不盡相同，例外的情況繁多，甚至同一本方志內職官的前後規則都不同，欲透過撰寫電腦程式，根據預先觀察蒐集的、能涵蓋大部分內容的一組規則，讓程式到多本地方志內自動擷取資料，並不可行。因此，我們興起了建立一套人機互動系統的念頭，將「觀察文本規律」與「編寫規則」都交由欲對文本進行標記的人（可能是歷史學者或沒有計算機科學背景的標校人員）來做，程式設計師應做的是提供一個足夠彈性、能輔助這種工作流程的系統。

（三）研究目的

我們希望設計一個人機介面，允許沒有受過編寫電腦程式訓練的使用者也能一邊閱讀方志、觀察任官記錄的書寫規律，一邊以簡單直覺的使用介面來編寫擷取規則，編寫完成後只需要電腦套用，系統便會根據規則自動於全文標記出符合的字符串，供人審訂。人更能因應不同方志的寫法變異，修改先前建立的規則，再套用在不同的方志上，以因應各方志的不同狀況來擷取資訊。也就是說，為了因應各方志的寫法變異，我們需將原本全程交由程式員負責的「閱讀全文觀察書寫規律」、「蒐集規則」、「讓程式根據規則在全文中擷取資訊」的過程，改為前兩個步驟由任何未經程式寫作訓練的使用者直接操作，使其能在「文本」、「規則」、「套用」三者間直接互動的過程，將程式員能彈性指揮電腦根據規則擷取資訊的能力，傳遞給不寫程式的使用者。

從系統設計的觀點來看，需要知道何為此系統的「輸入」與「輸出」（input and output）。這套人機介面的 input 是地方志某個章節的全文，最終的 output 是易於轉成 CBDB 關連性資料庫的表格形式，表格的每一行是一筆地方志中人物任官的紀錄，即「何人於什麼時候任過什麼官」，一筆記錄中至少應包含人名、任職時間（可能為朝代、年號、或年份）、官職名稱，但也可能有任官者的額外資訊，如：籍貫、入仕方法（進士、舉人、監生等，可能亦載入仕年份與入仕地點）、之前官職與之後官職、及其他描述。

在輔助使用者擷取任官資訊的角色上，此人機介面應提供適當的標記工具，以助使用者：

- (a)能根據文本內容與欲擷取的資訊類型，設定想標記哪些類型的資訊
- (b)能簡易直覺地編寫正規表達式，系統便根據規則自動標記出符合的資訊
- (c)能立即看到套用的結果，一邊閱讀一邊審訂標記結果
- (d)有標記錯誤的地方（標錯或者漏標），能快速修正
- (e)能根據文本的書寫規律之不同，修改規則再套用
- (f)若文本的文字有錯誤，也能直接修改文字，儲存
- (g)能將結果匯出成具有上述欄位的輸出表格

這套系統的開發有以下的挑戰。第一，正規表達式對於未受過電腦程式訓練的人來說，並不容易理解，如何讓未學過程式寫作的人也能了解正規表達式，並能把從文本觀察到的書寫規律、按照正規表達式的概念編寫成擷取規則，是需要設計的。第二個挑戰是系統需設計成把人「觀察文本規律」、「編寫規則」、「電腦擷取」的工作流程抽象化，僅將流程背後確定的步驟寫成電腦程式，允許使用者足夠的彈性。

（四）現有相關工具

1. CBDB RegexMachine

CBDBRegexMachine¹⁴是哈佛大學資訊系博士生 Elif Yamangil 為 CBDB 發展的一個工具，其目標是將 CBDB 用於傳記資料擷取的技術普及給其他人文研究者，這個工具是以 Java 程式語言開發，使用者能用此工具載入欲標記的全文檔，以視覺化的方式編寫正規表達式（此工具內建了擷取中國史料日期的幾種不同規則），讓機器自動找出符合規則的字符、並以顏色於全文中標示（見圖 4），能將標記結果輸出為 XML 檔，結果能立即以 Excel 打開，即轉為表格。

CBDBRegexMachine 提供使用者以直覺性編寫正規表達式的功能，它將每個 regex 的編寫解構成不同部件，譬如想編寫能辨識「道光二十三年」這類日期的正規表達式，需要三個部件「年號+數字+年」以此順序組成，而各部件分別有不同文字規則，「年號」可從中國歷史帝王年號列表中任選其一，「數字」應為中文數字一到十的排列組合，「年」則為「年」這個字元。CBDBRegexMachine 讓使用者編寫 regex 時能自己為部件取名，稱為 handle，編寫部件規則時能從事先上傳的清單中（年號、中文數字清單）來組合，如此，使用者不需要知道正規表達式的複雜語法，也能編寫規則，並因為有 handle 的設計，能直覺了解編好的規則，如圖 5 所示。

但 CBDBRegexMachine 設計時對於輸入的全文檔有較高的要求，不適合載入地方志全文。並且，介面中不允許使用者改正機器標錯或漏標的地方，也無法修改數位化過程中

¹⁴關於更多 CBDB Regex Machine 的資訊，請詳見

<http://isites.harvard.edu/icb/icb.do?keyword=k16229&pageid=icb.page515758> 與 Bol, P., Chen, S-P., Yamangil, E. (2012, Oct. 5-6). A RegEx Machine. Paper presented at New Directions in Analyzing Text as Data, Cambridge, MA.

打錯的文字，兩者需等匯出 XML 檔之後才能修改，如此使用者得將「編寫規則」、「套用」、「修改」的過程從中分段，不符合一般人的標記過程，因此我們決定不採用此工具標記地方志職官。

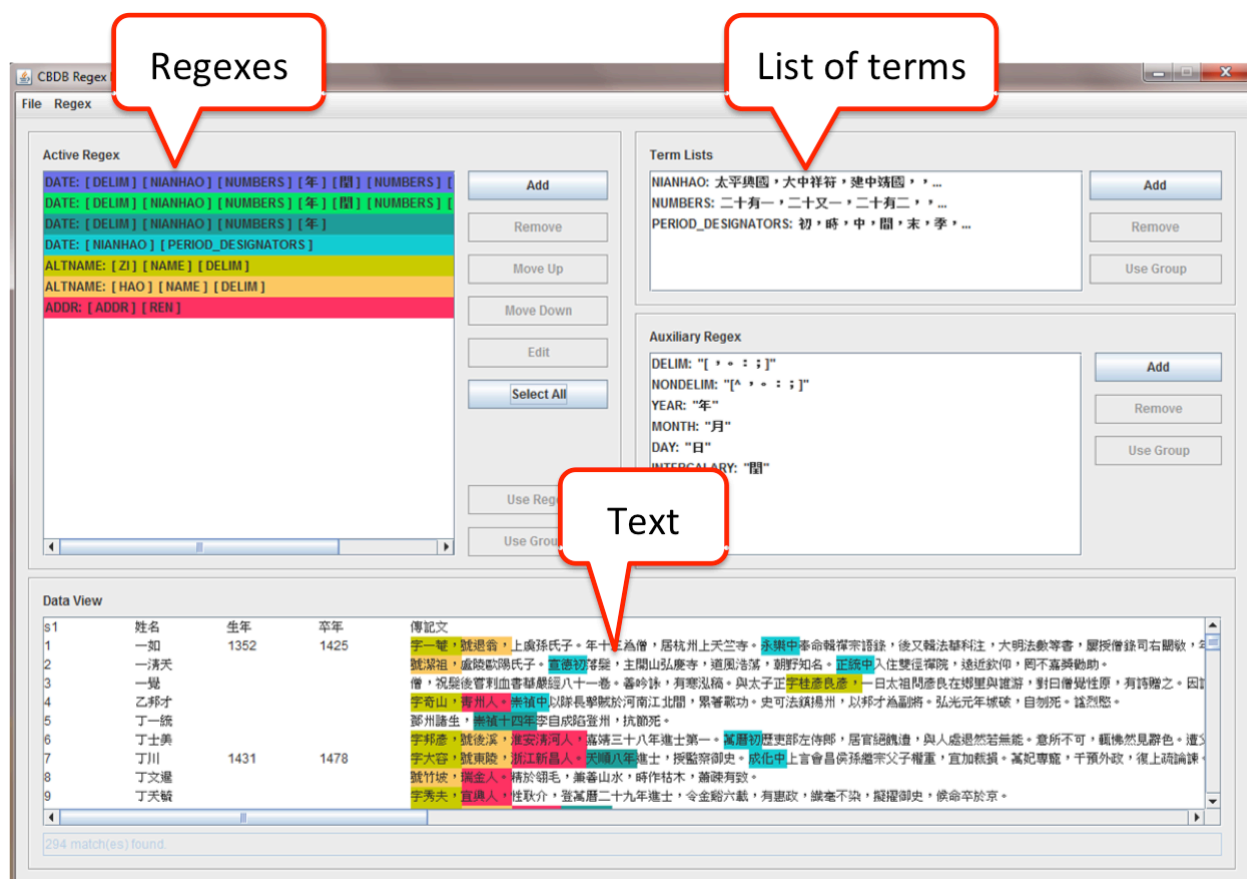


圖 4 CBDBRegexMachine 使用介面

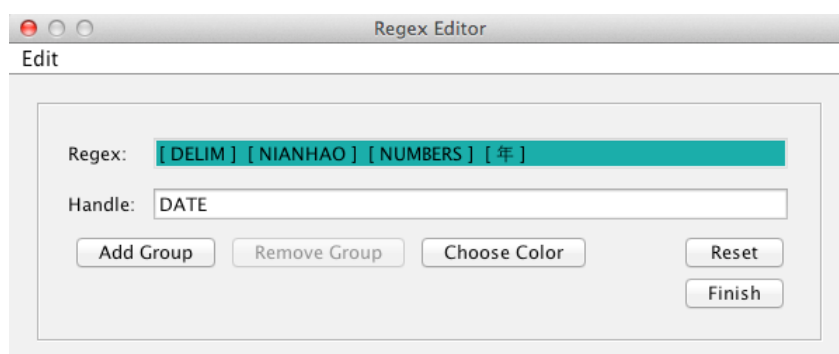


圖 5 使用者能自己命名 regex 的各個部件

2. Import.io¹⁵

讓人能在全文中標記資訊的工具很多，各種 XML 編輯器都可以做此用途，但想要有將標記過程自動化的工具，尤其用在中文資料的工具，則極少見。Import.io 的目標跟我們一樣，希望透過好用的使用介面，幫助不懂寫程式的使用者，將有規律的文本變成表格。Import.io 鎖定的文本類型是網頁，標榜使用者不需要寫程式，也能透過該工具半自動地將網頁上有規律性的資料（例如，購物網站上的商品清單）轉變成表格。Import.io 的技術是利用了網頁文字背後的 HTML 結構來轉換，透過使用者的點擊，系統辨識被點擊的文字背後的 HTML 結構後，自動建構擷取規則。

然而，在我們的職官志計畫裡，想自動擷取的是沒有任何背景結構的純文字，對電腦來說僅是一連串有順序的符號，僅有文字本身符號性的結構可供識別，尤其純文字材料不若網頁材料有 HTML 標籤的邊界可讓系統自動辨識使用者點擊的範圍，在純文字材料中不易採取相同作法得到高正確率的自動規則，反之，因為地方志職官的寫法規律，欄位類型固定，不如直接讓使用者編寫規律，更為有效。

¹⁵ 網址位於 <https://import.io/>，它們網站號稱“Web Data Extraction Made Easy. Turn any website into a table of data or an API in seconds – no coding required.”

二、系統結構

（一）系統需求及結構

這系統的結構主要分為三大部分：輸入、標記工具、輸出。輸入是任一本方志中的職官志區段全文，輸出是 HTML 表格，以易於轉成別種表格格式，譬如 Excel、CSV（comma separated values）、關聯式資料庫表格。

接下來是系統使用的技術，根據開發 CBDB Regex Machine 的經驗，本系統決定從使用 Java 程式語言轉到基於瀏覽器的 Web based 網頁工具，本系統的資料儲存在 MySQL，使用 PHP 作為 Server side 程式語言，而標記工具則是以 JavaScript 來編寫，方便使用者在瀏覽器上使用並得到即時回應。以下為開發本系統時的目標：

- (a)能根據文本內容和欲擷取的資訊類型，設定要標記哪些類型的資訊，使用者可以自行定義標籤（Tags）及其顏色。
- (b)能讓使用者根據文本的書寫規律之不同，簡易直覺地編寫正規表達式，系統會根據規則自動於全文標記出符合的資訊。在本系統中稱之為 SmartRegex，使用者自行定義一個或多個元件組成一個 Regex，元件可以是特定文字、或者一堆文字的列表，在系統介面能讓使用者新增、定義及編輯文字列表。通過使用者自行定義的 Regex，系統可以自動將其轉換成正規表達式並運行，達到自動標記的效果。
- (c)能立即看到運行自動標記的結果，當使用者運行 SmartRegex 或套用一些標記功能後，系統會在全文中以顏色標示來顯示該文字已被標籤，讓使用者可以一邊閱讀一邊審訂標記結果。
- (d)能讓使用者在發現標記有標錯或者漏標時，快速修正，如：標記文字、移除標籤、批次標記、與批次移除等功能。
- (e)能讓使用者在發現文本文字有錯誤時，直接修改文字。而且系統具備文字編輯器的一些進階功能，如複製、貼上、復原、儲存等，系統也根據每次儲存記錄下不同

文字版本（即 **versioning**），以便使用者發現問題後可以調回舊版。另外系統會用 XML 的格式儲存標記，所以在標記後使用者亦可編輯文字。

- (f) 能將結果匯出成如上述的輸出表格，因為要讓系統知道那些文字標記是同一條記錄，所以在標記時按照人名分行，輸出時就會是一行為一個人的資料。在標記的過程中，使用者可以使用帶換行的標記按鈕在標記時分行，更可利用系統內建中以百家姓來判斷人名自動分行。每一個記錄需要有人名才會被匯出，若記錄中有其他的標籤，就會變成表格中不同欄位，這種表格是介於文本與 **CBDB relational database** 結構之間的形式，容易之後整理匯到 **CBDB** 資料庫。

（二）標記流程及工具

1. 輸入資料

系統要求把文本分割成一定長度的段落以便處理，如果段落太長，在標記處理上每一個計算的等待時間會較長，而且如果段落中有不同類形或編寫方法的文字，會導致自動標記錯誤率提高；相反如果段落太短，使用者必須在每個不同的段落重覆一些標記的步驟，反而降低了效率，所以建議把類形相近、資料形態比較相似的文字放在同一個段落。在本研究中，我們把每本不同的方志先按其卷數切開，之後擷取每一卷的主題（例如：職官志、人物志、地理志等），如果連續兩卷的主題一樣則合併為一個段落，相反如果一卷中有不同的主題則分拆為兩個段落。

在文本分段條目都準備好後，就可以進行文本的符點空白一致化，把所有不影響標記資料標點符號和空白都替換成一個文本沒有的全形符號，在這個實驗中的文本幾乎沒有標點符號，我們只把空白替換成「空白圓圈」符號，即「○」（Unicode 編號 U+25CB），因為在文本中空白代表了原來書寫的格式，所以不同的空白數量（即替換後的空白圓圈數）有其格式意義，所以在標記過程中可以使用空白圓圈來做一些標記條件。

2. 系統基本功能簡介

當資料前置處理完成及匯入系統後，可以在系統中以不同段落為一頁進行標記，圖 6 就是系統的擷圖，左邊為正文部份，右邊為系統功能鍵。

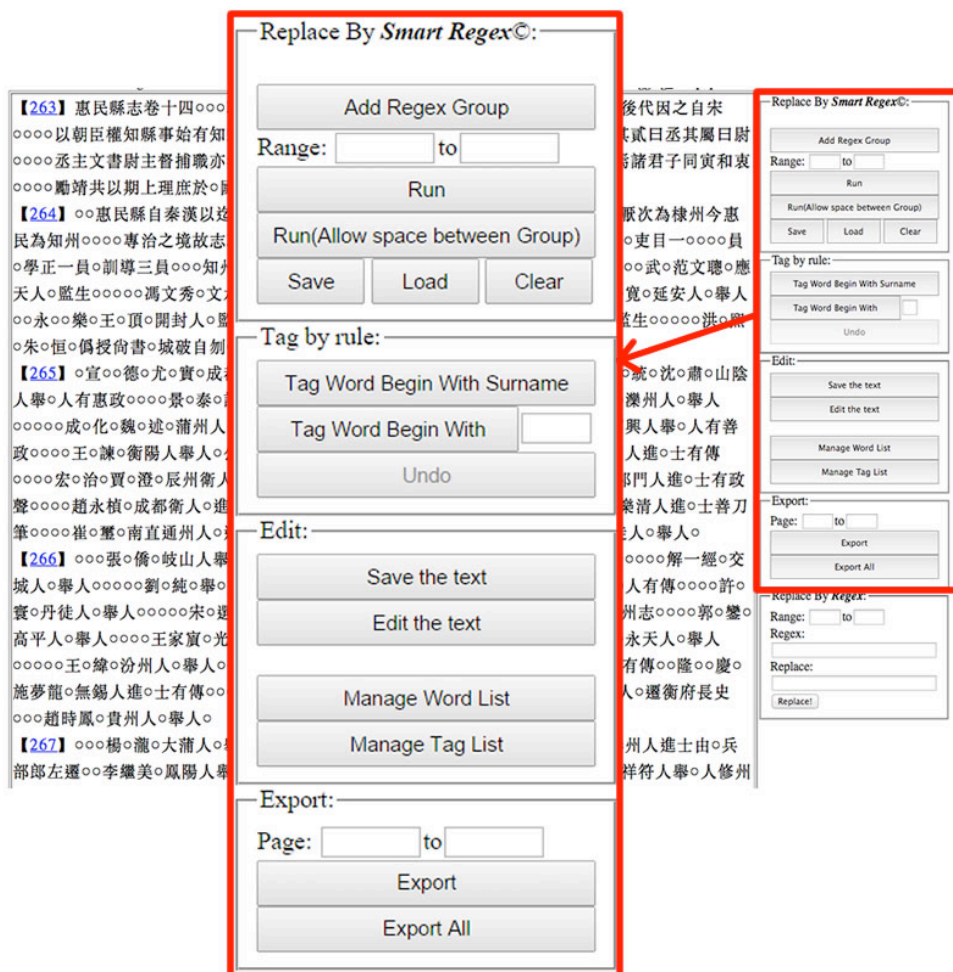


圖 6 標記系統擷圖

在使用此系統標記前，要先定義需要的標籤（Tags），即需要標記的不同類型資訊。在右方系統功能鍵中有“Manage Tag List”按鍵，進入後可新增標籤“Add New Tag”，編輯已有的標籤——包括其標籤名稱、英文簡寫、標記顏色，以及刪除不需要的標籤。使用者

不必在標記之前就定義好全部的標籤，如在標記途中發現需要新的標籤，可隨時進入此頁面新增標籤。在本實驗中，我們的標籤有：

(a)個人基本信息

人名、朝代、號、籍貫、別名、字

(b)入仕信息

入仕方法：進士、貢士、監生等表示的其如何通過科舉考試進入仕途的內容入仕

地點：與入仕方法相連的地點

入仕時間：與入仕方法相連的時間

(c)任官信息

官職：具體擔任的官名名稱

任官時間：擔任該官職的時間

任官地點：擔任該官職的地點

任官種類：“任”、“兼任”、“復任”等表示任官種類的詞

前官：先前擔任的官職，關鍵詞有“由[前官]陞”等

後官：之後擔任的官職，關鍵詞有“後遷[後官]”等

(d)其他信息

親屬關係、出處、及無意義（使用於自動化標記中，在本文稍後章節會有解釋）

Add New Tag					
ID	Name	Tag	Color	Save	Delete
15	朝代	dy		Save	Delete
24	入仕地點	entry_addr		Save	Delete
19	入仕時間	entry_time		Save	Delete
20	入仕方法	entry_way		Save	Delete
23	號	hao		Save	Delete
18	籍貫	jiguan		Save	Delete
25	親屬關係	kins		Save	Delete
30	後官	next_office		Save	Delete
14	年號	nianhao		Save	Delete

圖 7 自定義標籤頁面擷圖

在定義好標籤後就可以開始進行標記，使用者可以在右邊選單中點選“Edit the text”來編輯文本，包括增減文字、改變順序等，但最常使用到的應該是換行，因為在此系統中把每一行視為一筆記錄，所以使用者應該在每一記錄開始處按一下“Enter”為文字換行。之後使用者可以利用滑鼠圈選文字再按右鍵，在彈出來的選單中選擇該詞的標籤。為方便使用者標記，可以為經常作為記錄起始的標籤加上換行屬性，當使用者按下“Tag as: 人名”（快捷鍵為 Alt+1）時會自動換行，如果不須換行則選擇“Tag as: 人名（No BR）”。另外亦可選擇“Tag as: 標籤（ALL）”，則在整個文本中有該文字的都會在同一時間標記被起來，起到批次標記作用。

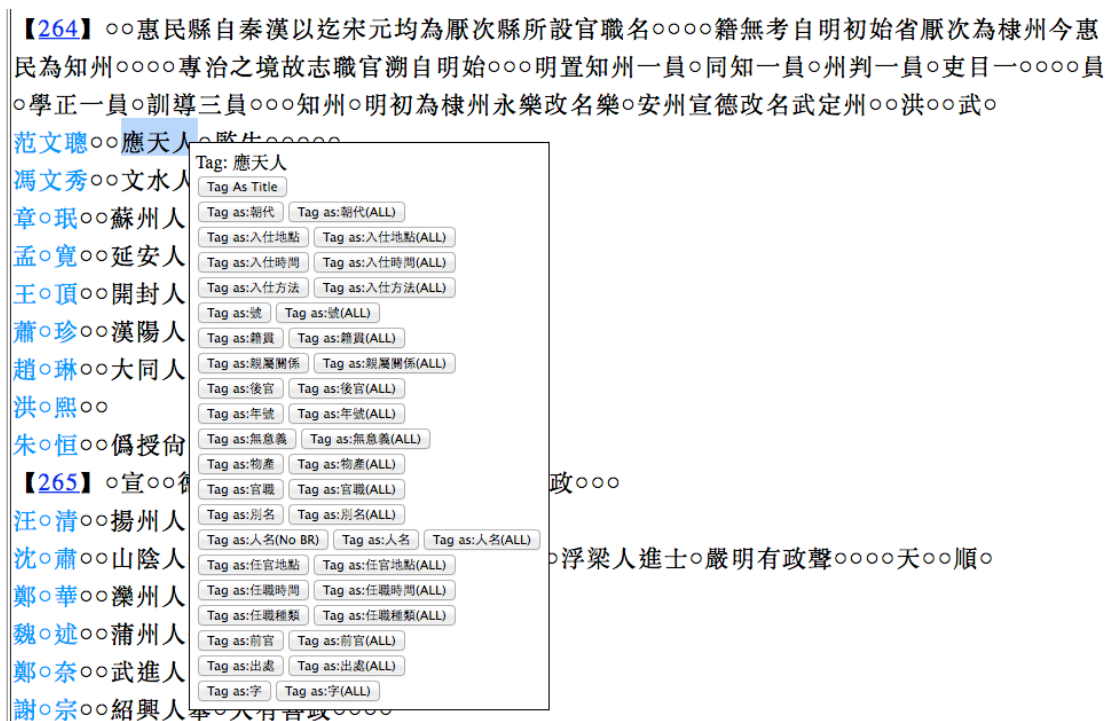


圖 8 手動標籤文字功能表擷圖

如果發現標記錯誤時，可在已標記的文字上點一下左鍵，選擇“Remove this”取消此一標記，或“Remove all”取消全文中標記為該標籤的同一文字。



圖 9 手動移除標籤功能表擷圖

在標記地方志時我們還發現了一個記錄資料方式，在這一本地方志中，它會先提到朝代“明置知州一員 同知一員……”，然後出現官職“知州”，然後是年號“洪武”、“永樂”。這就表示，在出現下一個朝代之前，這一批人物都是明代的；在出現下一個官職

前，這一批人物都是知州，其中，范文聰、馮文秀、章珉、孟寬在洪武間任官。此處的朝代“明”、官職“知州”、任官時間“洪武”都是屬於多行資料共享的信息。

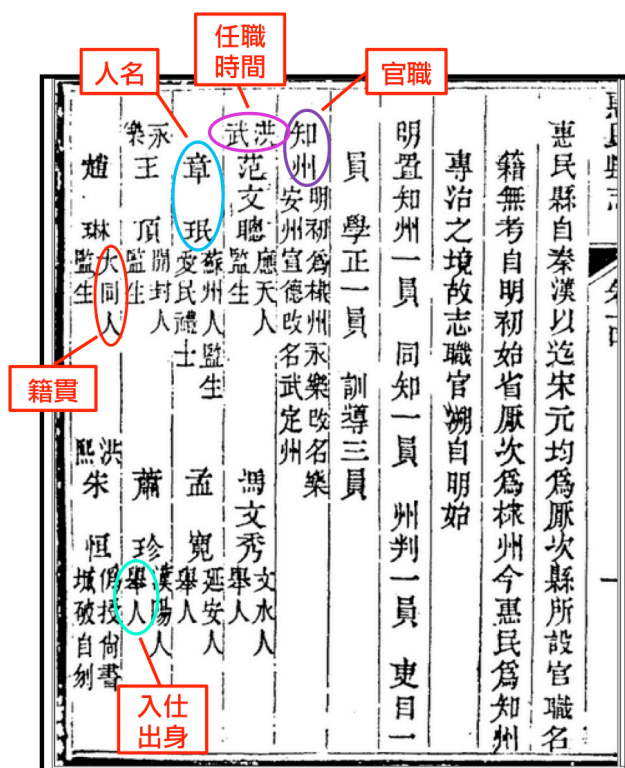


圖 10 原始文本中的編寫格式

因為這種情況在古代文獻資料中經常出現，所以本系統有一個功能來簡化這些標記，在此我們把它們稱作“標題”（title）——選中要標記為 Title 的內容，再選擇“Tag as Title”（快捷鍵為 Alt+2）。

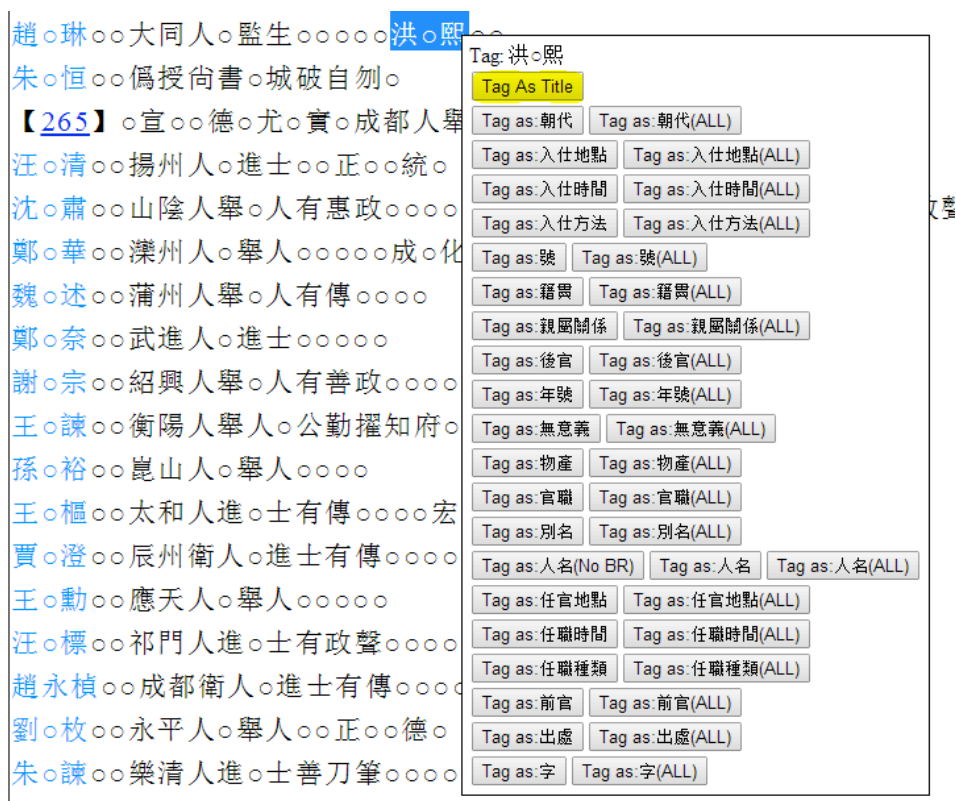


圖 11 手動標籤標題功能表擷圖

因為在輸出時，每一行的資料輸出成表格中的一行，其內容會按照標籤的種類放置在不同欄位。而剛才被標成“Title”的內容，是多行資料所共享的內容，所以要一次把這些共享信息加到每一行資料裡去。從 Title 內容的括號開始，一直選中到想要加上此“Title”內容的最末行為止。Title 中的內容會自動出現在第一行右邊空格處，在左邊的下拉菜單中選擇該 Title 的標籤類型，再點擊“Add Title Tag To Each Line”。

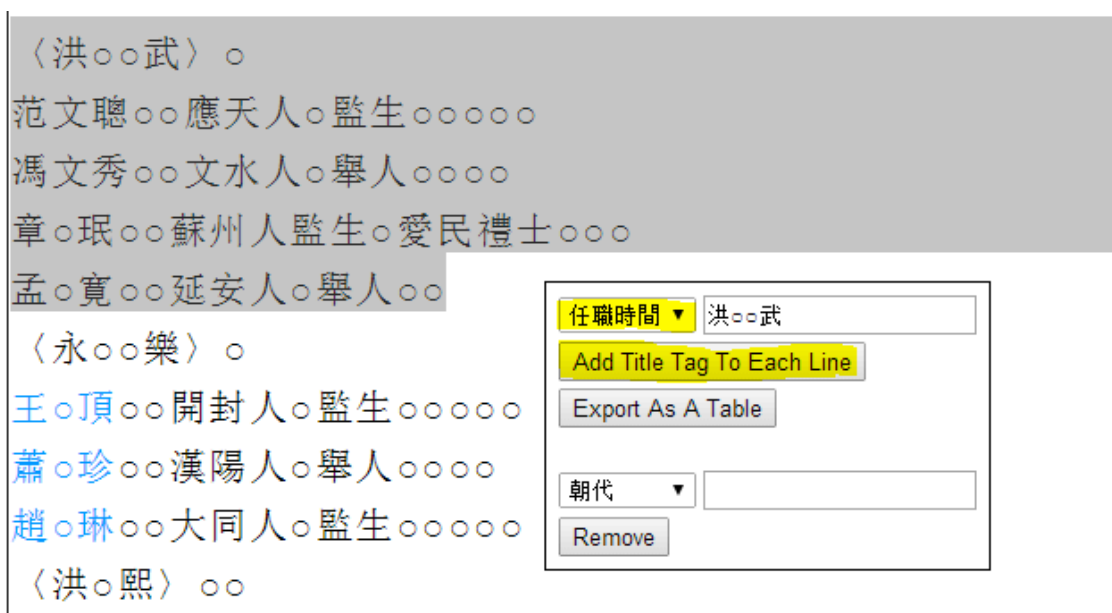


圖 12 把標題標記到內文功能表擷圖

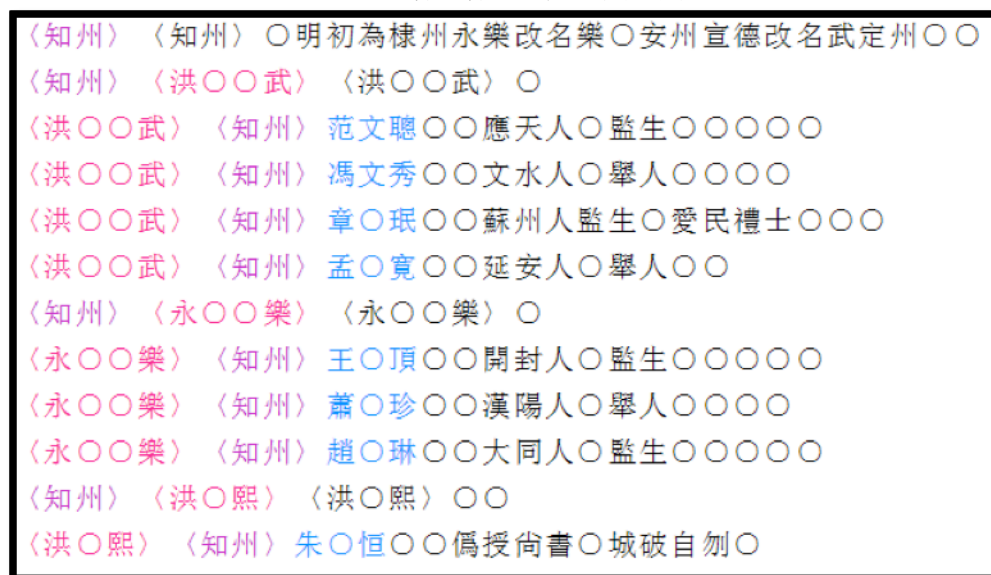


圖 13 把標題標記到內文效果擷圖

3. 利用系統 SmartRegex 建立標記規則

經過上述步驟的整理之後，可以明顯看得出來每一行記錄中有想要標出來的資料，如別名（字、號）、籍貫、入仕信息（時間、地點、方法）、任官信息（時間、種類）等，如果每一個資料都圈選再標記是十分耗時的，所以本系統提供一套 SmartRegex 工具幫助

使用者直覺地編寫正規表達式（regular expression，簡稱 regex），然後一次快速地把符合這個規律的文字全部標記出來。在這邊舉一個例子：

- (a)[1 到 6 個字]後跟上一個“人”字→籍貫（紅色框中部分）
- (b)[符合“入仕方法”詞庫的文本]→入仕方法（綠色框中部分）
- (c)[數字]“年”“任”→任職時間（粉色框中部分）

〈雍○正〉 〈雍○正〉 ○
 〈雍○正〉 蔡○□○江南泰州人監○生十二年任○
 〈乾○隆〉 〈乾○隆〉 ○
 〈乾○隆〉 尹○任○雲南浪穹人○舉人三年任○○
 〈乾○隆〉 艾○深○隴黃旗人監○生八年任○○○
 〈乾○隆〉 徐繩甲○浙江烏程人○舉人九年任○○
 〈乾○隆〉 張維垣○甘肅固原人○監生十年任○○○
 〈乾○隆〉 朱容極○湖北漢陽人舉○人十三年任○
 〈乾○隆〉 王允深○江西鄱陽人舉○人十三年任○○○
 〈乾○隆〉 鄧○瑛○順天宛平人舉○人十七年仕○
 〈乾○隆〉 王錫蕃○山西榆次人舉○人二十一年任○○○
 〈乾○隆〉 蕭○樑○貴州都勻人舉○人二十三年任○
 〈乾○隆〉 陳名儉○山西鳳臺人舉○人二十四年任○

圖 14 地方志中人物資料格式

在正規表示式不單單要用一個字來指定規則（例如 a 中的“人”），更可以利用一個字詞列表（word list），或稱詞庫，來建造規則（如 b 中的“入仕方法”，該詞列表包含貢生、官生、副榜、解元、進士、歲貢、貢生、監生、舉人、吏員、拔貢、廩貢、廩生及生員等詞）。在系統右側中部“Edit”框中，點擊“Manage Word List”，可以管理所有的字詞列表。

在觀察出文本中的規則後，就可以利用系統的功能來自定義 Regex，對於如何使用 Regex 在本文中不作詳細說明。在右側選單點擊“Add Regex Group”，可添加需要的 Regex 組，並可將經常使用的儲存起來，以供以後使用。這個系統的設計是在標記初期，可以讓資訊專業的技術人員協助分析文本，並定下 Regex 然後儲存起來，在以後的標記

過程中只要讀取出已設定好的 **Regex** 即可使用，亦可針對性作出修改以符合個別特例使用，更可把修改後的 **Regex** 另存成新的規則，在這樣的流程下可以省去很多重覆的步驟進而有效地標記文本。接下來舉一些本實驗的 **Regex** 規則及其使用流程，以供參考。

各個 **Regular Expression** 的功能如下：

Zi：標出該人的字 **字 字名**

Hao：標出該人的號 **號 號名**

Basic_Address：標出籍貫內容 **籍貫 人**

Entry_method_student：標出入仕方法，由三個部分組成 **入仕地點 干支 入仕方法**

Post_year_month_day_type：標出精確到日期的任官時間，**年號 年份 年 月份 月 日子 日 任**

Post_year_type：標出精確到年份的任官時間，**年號 年份 年 任**

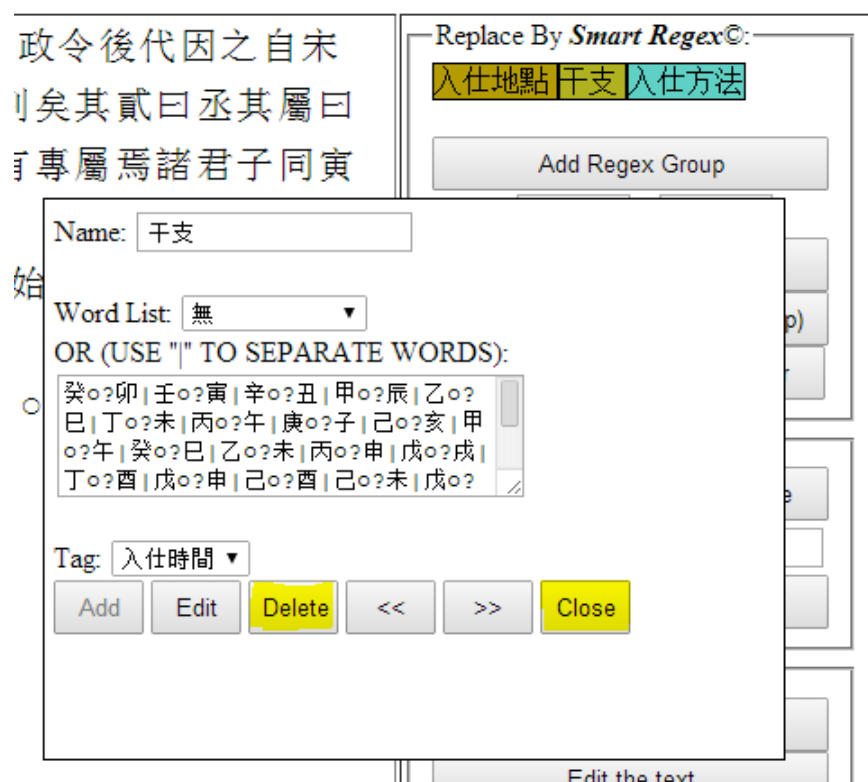


圖 15 系統中 **Regex** 工具擷圖

經驗所得，CBDB 標記職官志工作時的 Regex 操作順序是 Zi→Hao→ Basic_Address→ Entry_method_student→ Post_year_month_day_type→ Post_year_type。其中要注意的是，在使用“Entry_method_student”**入仕地點 干支 入仕方法**後，由於地方志中不一定會出現「入仕地點+時間+方法」的組合，所以為了標出「入仕地點+入仕方法」及僅有「入仕方法」的組合，我們需要把這一支 Regex 中的「干支」組刪去，然後再跑一遍。最後，我們再把「入仕地點」組刪去，再跑一次，確保所有的入仕資料都被標出。對於“Post_year_month_day_type”**年號 年份 年月份 月 日子 日 任**這一支 Regex 也一樣要對其進行微調。在加載並運行完之後，我們要把「日子」和「日」兩個組刪除，再運行一次，這樣，我們就可以標出精確到月的任官時間了。



圖 16 系統中 Regex 工具擷圖

在每執行完一支 Regex 後，使用者可以在地方志全文中稍微看一下標出的內容是否是自己想要的內容，執行完所有 Regular Expressions 後，再對全文進行仔細檢查。另外針對進階使用者可能需要使用一些系統未內建的 Regex，所以預留兩個欄位讓使用者以 Regex 原始語法來編寫規則，如下圖。

Export:

Page: to

Export

Export All

Replace By *Regex*:

Range: to

Regex:

Replace:

Replace!

圖 17 系統中 Regex 工具擷圖

4. 輸出資料

在輸出前，我們建議使用者以人工的方法掃看一次整段文字，如果文本結構比較工整的話，若是還有標記遺漏、錯誤，在彩色的文字中是十分容易被發現的。如果發現標記不完整，可以使用圈選後“Tag as: 標記（ALL）”的方法來把整段中同一樣的文字標記，或是有標記錯誤，則可點選後取消，那些範圍的標題標記亦可在系統選取整段後，一次全部取消。

在檢查完標記的內容後，在系統右下方點擊“Export All”，將所有資料輸成一張表格，此時可以對輸出的資料進行二次檢查。先觀察表格中“其他”一欄，如下圖，此欄中系統輸出的是在原文同一行記錄中沒有被標記的文字，如果發現此欄出現應該要被標記的內容，則可回到系統中改正並重新輸出。除了檢查漏標的內容之外，還要注意在輸出的表格中有沒有包含引號“<”和“>”。如有則可能同一文字內容被包在了多個記標中而造成錯誤，同樣也要回到原文中，對其進行修正。

#	Book ID	Section ID	方志	部	頁數	朝代	入仕地點	入仕時間	入仕方法	號	籍貫	親屬關係	後官	年號	無意義	物產	官職	別名	人名	任官地點	任職時間	任職種類	前官	出處	字	其他
1	35213	98798	臨澧縣志	職官志		明					肥鄉						縣令		蔣鳳		成化年間					□□
2	35213	98798	臨澧縣志	職官志		明											縣令		高顯		成化年間					
3	35213	98798	臨澧縣志	職官志		明	江夏		舉人								縣令		余本政		成化年間					
4	35213	98798	臨澧縣志	職官志		明	漢陽		舉人								縣令		李仲良		成化年間					
5	35213	98798	臨澧縣志	職官志		明	靈寶		舉人								縣令		侯敬		弘治年間					
6	35213	98798	臨澧縣志	職官志		明	開州		舉人								縣令		王世臣		弘治年間					
7	35213	98798	臨澧縣志	職官志		明	鄞縣		進士								縣令		周致		弘治年間					
8	35213	98798	臨澧縣志	職官志		明	柏鄉		舉人								縣令		趙璧		弘治年間					

圖 18 輸出資料表格

系統所輸出的表格可以直接進行全選、複製，貼到 Excel 中作進一步的處理。另外在標記過程，按下系統右邊的“Save the text”可把文本以 XML 的方式保存在系統中，再次進入這本地方志時，會把保存的檔案開啟，以供使用者繼續標記，使用者亦可把 XML 檔案匯出到其他資訊系統利用。

三、結果與分析

(一) 標記結果

自系統於 2013 年 10 月開發完成以來，位於哈佛大學的 CBDB 工作組在 10 月至 12 月完成了部分地方志中職官志的標記工作。在 2014 年 3 月開始，位於北京大學的工作組繼續地方志的標記工作。截至論文成稿時，相關的數據如下：

表 1 使用本系統完成標記的地方志數據

	標記地方志本數	依照目錄切分而成的篇數	資料筆數	地方志頁數	花費時間（小時）
CBDB	119	143	101,593	9,196	152
北大	176	215	148,663	15,053	288
總計	291	358	250,256	24,249	420

其中，標記地方志工作所花費的單位為小時，且花費時間則依據 CBDB 與北大工作小組估算。依據上表，共計在 420 小時中（約 52.5 個工作天），完成了 291 本地方志的標記，獲取約 25 萬筆資料，耗時約 420 小時。這樣大規模的資料擷取，是用傳統的人工輸入方法所不能想像的。工作組成功地使用系統對地方志進行標記，也從一個側面上驗證了系統對於實際使用者，在提高資料擷取效率上有一定幫助。不過，為了更好地闡釋本系統在標記文本、擷取資料上所做的獨特設計，以及本系統可以為人文學者提供的幫助，本章將通過實驗及數據分析來探究並嘗試回答兩個主要的研究問題，即：

- (a) 利用系統標記地方志文本中的資料的效能。
- (b) 與傳統的手動標記方式相比，本系統的優勢與劣勢之所在。

由於時間與人力所限，本文中的實驗採取了小範圍的定量研究，並主要討論系統標記的效率、準確性等方面，同時將其與傳統的人工手動標記方法相比較。

（二）實驗設計

1. 實驗參與者

本次實驗選用了 4 名實驗參與者，於 2014 年 9、10 兩月，分三次進行小範圍實驗。實驗參與者包括具有地方志文本相關背景知識的學者、學生，以及擁有一般地方志文本閱讀能力的編輯人員。這樣設計是為了驗證本系統是否不論對專業的人文學者者，還是非專業的資料擷取者，都有較高的效能。同時，實驗者學術背景的多樣性也與本系統潛在的使用者構成較為吻合，使得我們的實驗結果具有較強的說服力。而在選取實驗參與者時，除了保證他們學術專業上有所不同之外，我們也盡可能地包括多種年齡層、來自多個地域的實驗參與者。參與實驗者的年齡分佈為 24—52 歲，來自北京、南京及澳門。這樣一來，我們所獲得的實驗數據可以較客觀地反映系統標識地方志文本的效用及準確性。

2. 實驗方法

本文中的實驗共選用三篇地方志文本選段，由四位實驗參與者，使用手動標記與系統標記兩種方法，完成了 16 組實驗。本節將從實驗文本、實驗設計、數據收集三個部分進行詳細闡述。

2.1 實驗文本

本文中選用的實驗文本為在《中國方志庫》¹⁶中抽取的三篇地方志文本中的官職片段。三篇地方志分別來自《偏關志》（職官）、¹⁷《德州志》（卷八，職官）、¹⁸《濟南府志》（卷三十一，秩官九）。¹⁹在原文中的頁數為 10 頁、30 頁及 80 頁，相對應的純文本長度分別約為 1600 字，5200 字及 14000 字。抽選三種長度的地方志文本是為了驗證系統在標記不同長度的地方志文本時的效用的一致性。為保證實驗文本具有代表性，三篇地方志樣本均為原文結構較為清晰、包含人名——職官對應的職官志內容。

¹⁶ 《中國方志庫》為愛如生大型資料庫的一種，收錄從漢魏到民國共計一萬種方志，每種皆具數位全文與原版影像。

¹⁷ 道光《偏關職官志》，（清）盧承業原編（清）馬振文增修，清道光間刊民國四年鉛印本。

¹⁸ 乾隆《德州志》，（清）王道亨修（清）張慶源纂，清乾隆五十三年刻本。

¹⁹ 見註腳 13。

為了行文方便，在之後的討論中將這三篇地方志文本標記作文本一（《偏關志》職官）、文本二（《德州志》卷八，職官）與文本三（《濟南府志》卷三十一，秩官九）。

2.2 實驗設計

實驗前：採集實驗參與者基本資訊，包括年齡、專業背景、所在地。對於沒有接觸過系統的實驗參與者，統一對其進行系統使用培訓。²⁰

實驗中：每一位實驗參與者都會得到以網頁形式呈現的實驗文本，並在依照先手動標記，後用系統標記的順序，對實驗文本進行處理。實驗文本可細分為兩個部分，即主要的、已經數位化的文本界面及原文圖像供實驗參與者參考。

在進行手動標記地方志文本時，我們要求實驗參與者建立一個 Excel 表格，並按照“每一行只放置單一人的資料、不同類型資料分不同欄位輸入”的原則進行標記。其中，欄位的設置以實驗參與者習慣為準，並不要求表格欄位遵循樣表的順序。為實驗參與者提供的樣表如下：

表 2 手動標記地方志文本之樣表，實驗文本一

官名	人名	籍貫	入仕方法	入仕時間	字、號	入仕年號	前後官信息	其他
長山知縣	周懋臣	直隸密雲	貢生	二年		順治		
長山知縣	朱天寧	江西新喻	進士	四年				
長山知縣	張國維	浙江縉雲	貢生	五年				
長山知縣	趙國海	奉天		六年				
長山知縣	趙一心	山西汾陽	進士	七年				
長山知縣	王來舉	奉天		九年	字彥升			
長山知縣	馮隆運	直隸昌黎	生員	十三年				
長山知縣	郭振郛	山西平陽	舉人	十五年				
長山知縣	賈復元	山西崞縣	進士	十六年	字臨溪			
長山知縣	謝龍	奉天錦州	舉人	二年		康熙	升知州	

²⁰ 培訓材料為「地方志標記系統操作手冊」。

手動輸入標記時，我們不提供額外的檢查時間，一但實驗參與者完成全篇地方志文本的標記，即算做實驗結束。

在使用系統進行標記時，我們要求實驗參與者自行完成地方志文本的標記，不得尋求其他幫助。同樣，在全篇地方志文本標記完成後，也不提供額外的檢查時間。實驗參與者必須即刻輸出表格，同時實驗計時結束。

進行兩次標記之間，我們保證實驗者有時間休息，以取得最佳的實驗效果。每個實驗者都使用了兩種標記方式，最少標記兩篇地方志文本，在標記途中均由實驗控制人員計時。

實驗後：將同一位實驗參與者的標記結果保存在一個 Excel 表格文件中，按照實驗文本的不同和標記方式的不同，分別保存成多張表格。統一收錄，以便接下來進行數據分析。

2.3 數據收集

在所有實驗參與者都完成實驗後，我們將 Excel 表格與計時結果匯總如下表：

表 3 實驗的數據收集結果基本情況

	實驗文本			總計
	文本一	文本二	文本三	
實驗人數	4	3	1	3
實驗總時間	284	522	453	1259
實驗結果	8 組	6 組	2 組	16 組

對每一篇地方志文本的標記結果，我們都對照地方志原文進行了檢查工作，為接下來討論手動標記與系統標記結果的完整性、準確性提供數據依據。除了本次實驗所收集的數據之外，我們還將結合 2013 年 10 月，系統開發完成後，哈佛大學、北京大學工作小組使用系統標記地方志所獲得的數據進行進一步分析與討論。

（三）分析與討論

為了使讀者對系統的效能有一個較直觀的了解，我們將在此節中討論使用系統標記地方志時，工作的效率與結果的準確性。同時，我們也會將相對應的數據與傳統手動標記方法產生的結果相對比，以探究系統標記的優勢與劣勢。

1. 效率

本小節中討論的地方志文本標記效率，主要的分析指標是手動標記及用系統標記所用的時間。而標記的完整性與準確性將放在另一部分進行討論。

表 4 手動標記與系統標記地方志文本實驗平均時間

		實驗文本		
		文本一	文本二	文本三
實驗人數		4	3	1
實驗平均時間（分鐘）	手動標記	58.5	121.3	374
	系統標記	12.5	52.7	79

由上表可知，對於三種不同長度的地方志文本，使用系統標記的時間都遠遠少於人手標記的時間。以時間為縱坐標、地方志文本頁數為橫坐標可畫出時間—頁數線性圖。

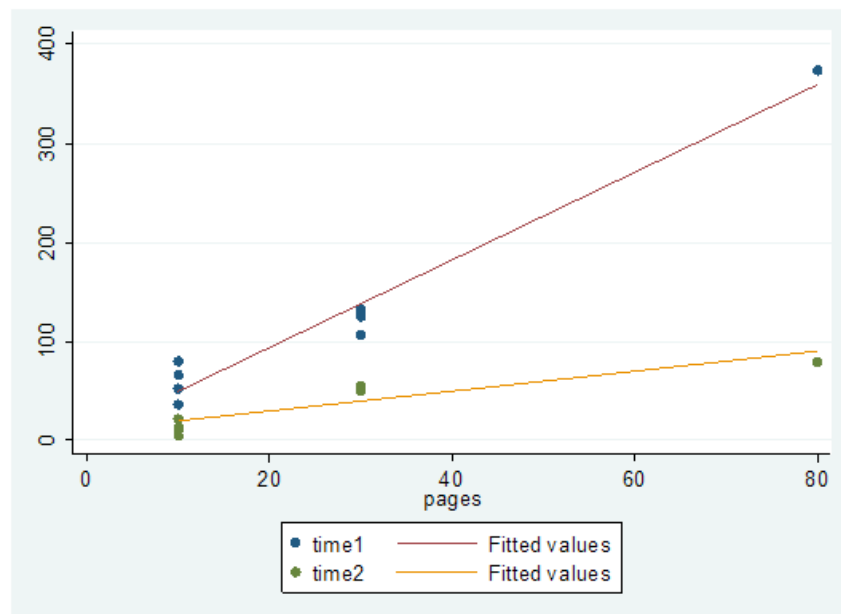


圖 21 手動標計與系統標計，時間——頁數線性圖

表 5 手動標計與系統標計，時間－頁數線性回歸數據

. regress time1 pages					
time1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pages	4.430094	0.3437083	12.89	0.00	3.58907, 5.271118
_cons	5.210031	11.84423	0.44	0.68	-23.77175, 34.19181
. regress time2 pages					
time2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pages	1.009718	0.2035729	4.96	0.00	0.5115928, 1.507843
_cons	9.369906	7.015148	1.34	0.23	-7.795543, 26.53536

若將手動標記地方志文本的時間記作 T1，系統標計的時間記作 T2，則上由圖，可知標記時間 T（分鐘）與文本長度 P（頁數）的關係約為：

$$T1 = 4.43p + 5.21$$

$$T2 = p + 9.37$$

可推算，當地方志文本約為 1.21 頁時，手動標記於系統標記所花的時間的基本相同，當地方志文本長於 1.21 頁時，使用系統標記的時間便小於手動標記時間。若按照假設，標記時間 T（分鐘）與文本長度 P 滿足線性方程，則手動標記時間 T1 大約為系統標記時間 T2 的 4 倍。

對於一篇長度為 100 頁的地方志文本，系統標記時間 T2=109.37（min），而手動標記的時間 T1=448.21（min）。對於平均長度為 70.9 頁的中國方志庫中按章節切分的文本，想要大量處理，顯然使用系統標記的效率是遠遠高於傳統方法的。並且，地方志文本越長，使用系統所能節省的時間便越多。

2. 資料完整度

本處的資料完整度僅考量實驗中參與者在標記地方志文本時，是否將原文中的有效資料²¹，即與人物相關的職官資料內容，完整輸出至 Excel 表格之中。此處我們暫時不考慮

²¹在本文中，我們討論的有效資料為與人物相關的資料。又因為本文中例用地方志中的“職官”部分舉例說明，故些處的有效資料為職官資料。任何與某一人物無關的職官資料，例如該官職何時設立、何時被廢，都不會被計在有效資料中。

資料輸出的準確性，僅統計實驗地方志文本中應有的資料內容，並與實際實驗者標記輸出的內容進行對比。在統計時，我們使用兩個量來對完整度進行衡量。

i. 資料筆數：

在地方志中一個人的資料算做一筆，也就是在文本中包括多少個人的資料，就應該輸出多少筆資料。實驗中在不提供額外檢查時間的情況下，對於文本一，系統標記平均得到 145 筆資料，完整度為 100%，手動標記平均得到 144.5 筆資料，完整度為 99.7%；文本二，手動標記平均得到 413 筆資料，遺漏了 2 筆資料；而在文本三中，手動標記得到 1016 人，原文中應有資料為 1023 筆，遺失了 7 筆資料。

表 6 手動標記與系統標記結果統計

	手動標記	系統標記	原文有效資料數	手動標記資料完整度	系統標記資料完整度
文本一	144.5	145	145	99.7%	100%
文本二	413	415	415	99.5%	100%
文本三	1016	1023	1023	99.3%	100%

ii. 欄位資料個數：

文本每一筆任官記錄中，還包括不同種類的資料，這裡稱為欄位，此處考察的是，在不同欄位下的單元格中的資料是否完整。實驗中兩種標記方式在單元格內資料上都有缺失。其中，手動標記主要缺失內容出現的地方較為隨機，無法總結出規律。而系統標記中產生的遺漏，主要是由於地方志原文文本的格式出現變化，不再能被預先設置好的正則表達式抓取而導致的。在使用系統標記時，對系統較為熟悉的實驗者就會有意識地注意格式產生變化的文本，從而有效地減少該類型的資料缺失。另外，在使用系統標記時，其實並不存在絕對意義上的單元格資料缺少，因為系統在輸出資料時，會將未被 tag 標記，即未放在一般資料欄位下的內容放在“其他”欄位下。這可被視為資料欄位錯誤，故此類型的資料遺漏，我們會在下一節，資料的“準確性”中，進行進一步討論。

在檢查時，我們除了統計缺少的資料之外，還觀察了對於手動標記輸入出的內容，以及系統標記輸出的內容，檢查的難度及檢查後修正的步驟的複雜程度。

檢查完整度的難易：

- 手動標記：較困難——需要逐字對應原文檢查，所花時間長。
- 系統標記：簡單，只需要去系統界面觀察哪些文字沒有被標記上顏色——即未被輸出，再判定這些文本是否是有效資料即可。

檢查到有資料缺失後的解決方法：

- 手動標記：需要手動輸入，將資料補充完整。
- 系統標記：選中資料，將其標上正確的 tag 即可。若錯誤性質相同且符合同一規律，還可以一次批量完成修改。

在經過統計與觀察使用不同方法標記的地方志文本後，可得出三點結論：

- (a) 文本長度越長，資料完整度越低。
- (b) 對於同等長度的文本，手動標記的資料完整度不及系統標記的完整度。
- (c) 在標記完成後，對系統輸出結果進一步檢查與修改較為容易且花費時間短，而檢查與修改手動輸出的的結果較困難且花費時間長。

完整度對於人文學者及資料擷取者來說都十分重要。在標記地方志文時遺漏資料，無疑會給相關學者進行分析與研究時帶來不便。在手動標記時，由於需要大量處理人名，故稍不細心，就有可能遺漏人名，從而造成資料不完整。而在用系統標記時，由於已經有一套機制可以自動識別並標記人名，並且在標記過程中給用戶提供了一個初步檢查的過程，所以可以更好地保存資料完整性。從數據對比來看，利用系統標記可以有效規避手動標記時人名遺漏這一問題，從而在資料完整性上佔有優勢。

3. 資料精準度

此處的精準度定義是，輸出的資料內容與地方志原文相比，資料內容是否準確，資料類型是否正確。在檢查過程中，發現存在的錯誤類型主要有：

i. 欄位放置錯誤：

例：

表 7 手動標記欄位放置錯誤樣例

人名	籍貫	入仕方法	入仕時間	字、號	入仕年號	前後官信息	其他
吳階玉			四十六年			由新泰知縣署任	
下咸和	江蘇江都	進士	五十年	字岷山			
孫衍	浙江嘉善	五十二		字宰工			
沈聖願	江南	舉人	五十七				

此處，“五十二”應該置於“入仕時間”一欄，卻被錯誤地放在“入仕方法”中。

表 8 系統標記欄位放置錯誤樣例

入仕方法	號	籍貫	官職		人名	任職時間	任職種類	後官	其他
監生		漢軍鑲黃旗	新城知縣	張印宗	張印宗	乾隆;六十年	任		調冠縣
歲貢		萊陽	濟陽訓導	崔巍	趙爾汲	順治			陞州同

此處，“調冠縣”與“陞州同”均為相應人名的“後官”資料，卻被錯誤地置於“其他”欄位中。

ii. 單字錯誤：

單字錯誤，即輸入的標記結果中的資料與原文資料相比，雖然資料類型放置正確，但是文字上有誤。此種錯誤類型在手動標記中較為常見，錯字率約為 0.3%。因為手動輸入時，需要處理輸入的文字量非常大。稍有不注意，就會產生打字錯誤。而在使用系統標記時，錯字一般不存在，因為資料是直接從原文中標記。但是，不排除系統標計的電子文本本身就有錯誤——即從地方志原文圖像數位化成電子文本時存在的錯誤。

例：原文電子化時“萬曆”被錯誤數位化為“若曆”

〈明〉〈西樞廡〉胡○恪○○河南萬曆○三年任○○○○○○○○○
〈明〉〈西樞廡〉鄒應芳○○陝西若曆○十二年任○○○○
〈明〉〈西樞廡〉張嘉績○○陝西萬曆○二十年任○○○○○○○○○
〈明〉〈西樞廡〉張○鑒○○陝互萬曆二○十五年任○○○○
〈明〉〈西樞廡〉福文明○○山東萬曆三○十一年任○○○○○○○
〈明〉〈西樞廡〉蔡夢齊○○直隸萬曆三○十五年任○○○○
〈明〉〈西樞廡〉楊若予○○雲南萬曆三○十八年任○○○○○○○○
〈明〉〈西樞廡〉馮國貞○○湖廣萬曆○四十年任○○

圖 22 原文數位化時的錯誤²²

但不可否認的是，手動標記地方志文本中的資料時，更能注意到地方志原文中可能存在的問題。在這一例中，三個實驗參與者都注意到了這個小錯誤。但這種情況一般較少見，且涉及圖像資料的數位化問題，不是本文主要討論的內容，故不在此作深入探究。

根據統計，兩種錯誤類型在使用不同標記方法輸出的結果中出現情況如下：

表 9 準確度錯誤在不同標記方法輸出結果中的出現頻率

		平均欄位放置錯誤（個）	平均單字錯誤（個）
文本一	手動標記	2	2
	系統標記	0.5	0
文本二	手動標記	13	16.7
	系統標記	2.6	7
文本三	手動標記	18	32
	系統標記	11	0

同樣，在檢查時，我們除了統計錯誤的資料之外，也同時觀察了對於手動標記得到的，以及系統標記輸出的內容，檢查的難度及檢查後修正的步驟的複雜程度。

檢查準確度的難易：

- 手動標記：較困難，因為需要逐字對應原文檢查，所花時間長。
- 系統標記：一般困難，首先是關注輸出表格中的“其他”欄位，因為此欄位中會包含未被標記上任何 tag，即未被放置於任何一般資料欄位中的資料。如果在“其他”欄

²²資料來源：乾隆《德州志》，（清）王道亨修（清）張慶源纂，清乾隆五十三年刻本

位中發現了有效資料，就說明此處有資料未被準確標記；其次則是在觀察系統界面中，如果某資料本身的顏色與其屬於的資料類型不符，則說明發出欄位放置錯誤。檢查到資料有誤後的解決方法：

- 手動標記：需要手動修改人，將錯誤的資料重新輸入。
- 系統標記：選中資料，將其標上正確的 tag 即可。若錯誤性質相同且符合同一規律，還可以一次批量完成修改。

經過觀察使用不同方法標記的地方志文本後，可得出三點結論：

- (a) 文本長度越長，資料準確度越低。
- (b) 對於同等長度的文本，手動標記的資料準確度不及系統標記的準確度，尤其在單字準確率上。
- (c) 在標記完成後，對系統輸出結果進行檢查與修改較為容易，而檢查與修改手動輸出的結果較困難且花費時間長。故在錯誤率相同的情況下，檢查並修改系統標記出的內容都更快、更容易。

準確度是我們擷取資料的重要標準。如果資料本身的内容不正確，無論是使用傳統方法對其進行研讀，還是使用程式對其作大規模的分析，都會產生負面影響。與資料的完整性類似，在手動標記時，由於需要大量資料，故較易造成資料不準確。而在用系統標記時，由於直接處理的是數位化後的文本，所以錯誤率會大大降低。同時，系統本身的設計讓熟悉系統的人可以快速甄別出哪些錯誤放放在了錯誤的欄位（依據顏色迅速判斷），所以利用系統標記可以得到更加準確的資料，並且在發現資料錯誤時，可以較容易地進行修正。所以說，系統在資料準確性上同樣佔有優勢。

（四）實驗的潛在問題

由實驗結果及其他輔助數據可知，系統半自動化的設計，讓用戶能夠及時按照文本內容規律，快速標記資料。同時，系統其實提供了一個“邊標記邊檢查”的機制，保證效率的

同時還增強了準確度。使用系統標記地主志文本不僅可以大量節省人力，大幅增強效率，而且能夠提升準確度。不過，由於本次實驗的時間限制與人力所限，在實驗設計與數據收集與分析上難免存在一些潛在問題。

- (a) 實驗參與人數少，實驗樣本的代表性有待增加。
- (b) 對系統的熟練程度這個變量無法控制。本次實驗參與者對系統的熟悉程度並不相同，而在分析實驗結果時很難量化這個數據。對於潛在的更廣大的用戶，還需要研究系統優勢是否還會維持這麼明顯。從另一個側面來說，對系統越熟練，使用系統標記地方志的用時越短，且完整度與準確度越高，所以系統的優點在本文數據中有可能還沒有完全體現，還需要更多實驗數據支持。
- (c) 系統教學時間可以控制得更加嚴格。在本實驗中，沒有對系統的教學時間沒有討論。
- (d) 沒有討論實驗文本的難度差異。由於此次實驗樣本內容相對規整，信息相對簡單，系統標注的速度優勢會很明顯。而對於規律相對不明顯、資料內容複雜的的文本，系統的優勢之所在還需要更多實驗進行分析。

針對以上問題，我們已經設計了對應的解決方案，希望可以在不久的未來進行進一步實驗，收集更多的數據，不斷優化系統。

- (a) 尋找更多的實驗參與者，獲取更多實驗數據。
- (b) 增加實驗文難度這個變量。選取難度不同的文本進行實驗：考慮不同地方志的結構、排版、包含的資料內容數量差異，分析系統所具有的優點與限制的不同體現。
- (c) 將系統教學時間量化，作為一個考量系統效用的因子。
- (d) 將對系統的程度量化，作為一個考量系統效用的因子。
- (e) 採取多重回歸模型而非線性回歸模型，分析多個因子對於時間—頁數關係的影響，得出一個更加準確的回歸方程，從而可以更精確地分析系統的效率。
- (f) 增加定性調查，獲取實驗對象及用戶對於系統的評價，進一步完善系統。

四、結論

本論文介紹了為了從地方志擷取職官資料而發展出的半自動電腦輔助標記系統，讓未曾受程式編寫訓練的人文學者能自行於觀察文本編寫規律後，透過直覺的 **SmartRegex** 介面來製作基於正規表達式的標記規則，再讓電腦根據編寫的規則自動在全文標記出各類資訊。系統內建了一些判斷人名的規則，並提供批次修改標記的工具，讓使用者能在編寫規則、套用到全文、閱覽結果、修正錯誤的步驟之間，順暢且快速地對長篇文本進行標記。標記的結果除了能匯出成表格，系統也將標籤連同原文儲存為 **XML**，並保留多次修改版本，供使用者撤回修改之用。

這個工具使得從地方志快速擷取職官資料變得可能，即使是地方志各個文本存在著許多書寫變異，透過將觀察規律、編寫擷取規則的工作從程式員手中交給使用者，允許使用者能根據各地方志的不同規律性，來客製擷取規則，成功從不同地方志中擷取職官資料。截至論文成稿時，此系統幫助 **CBDB** 工作團隊於 **420** 小時的工作時間內，蒐集到 **25** 萬筆地方志職官資料。透過我們進行的小規模實驗，也顯示此系統的效能明顯優於傳統手動擷取職官資料的作法。

透過將地方志職官從全文轉變成表格，我們將原本對於電腦只是一連串無特定意義文字的職官志，依據資訊類型分開儲存在表格的不同欄位之內，這相當是告訴了電腦該串文字的意義——哪裡是人名、官職、任職時間、入仕方法，等等。這類「結構性資料」對比「全文資料庫」有著能分按語意儲存資訊的格式，人便可以運用電腦對具有語意結構的文字大規模地做更多操作，譬如可以將職官資料按照年代來分析、調出某年號間全中國各地載於地方志的官員等，跳脫原本地方志架構的侷限性，更能將擷取的職官記錄定位到地圖上呈現。

我們希望這套系統也能用在地方志其他內容的擷取，將地方志記載的豐富資料——各地寺廟、穀物作物、動植物、物產、天災、稅收等等，從全文擷取出來，變成表格，形成中國史地方志主題資料庫，如寺廟資料庫、作物資料庫、天災資料庫等，便能開啟更多使用地方志的途徑，尤其是大數據分析的方式。

從數位人文的角度來說，我們希望將資訊科學領域已發展成熟的技術，例如本論文採用的正規表達式，透過建置簡單直覺的工具，推廣給人文學者，讓他們也能享受此技術帶來的便利性，加速人文學者標記文本的工作。許多處理文字資料的資訊技術已經發展多年，非常成熟，但並未被廣泛用在人文領域，如何將資訊技術與人文學者的需求做連結，讓人文學者即便在不寫程式語言的狀況下，也能享受資訊領域發展的成果，幫助他們運用資訊方法進行人文研究，是我們將繼續探討的課題。

誌謝

感謝哈佛大學包弼德教授對此系統開發工作的經費投資。感謝本文撰寫過程中從 CBDB 計畫經理王宏魁先生處得到的諸多幫助。感謝北京大學尹航與吳淑敏對本次實驗的貢獻。

參考文獻

- 昌彼得、王德毅。1974。《宋人傳記資料索引》。臺北：鼎文書局。
- 王德毅、李榮村、潘柏澄。1979。《元人傳記資料索引》。臺北市：新文豐出版公司。
- 國立中央圖書館。1965。《明人傳記資料索引》：台北：國立中央圖書館。
- （清）沈世銓修（清）李勗纂。光緒年。《惠民縣志》。清光緒二十五年柳堂校補刻本。
- （清）林星章修（清）黃培芳纂。道光年。《新會縣志》。清道光二十一年刻本。
- （清）王贈芳修（清）成瓘纂。道光年。《濟南府志》。清道光二十年刻本。
- Bol, P., Chen, S-P., Yamangil, E. (2012, Oct. 5-6). A RegEx Machine. Paper presented at New Directions in Analyzing Text as Data, Cambridge, MA.