古籍風華再現：關於古籍數位人文平台之建置

Revitalizing the Splendidness: The Construction of Digital Humanities Platform on Chinese Ancient Books

林巧敏 Chiao-Min Lin
國立政治大學圖書資訊與檔案學研究所教授
Professor, Graduate Institute of Library, Information and Archival Studies, National Chengchi University
Email: cmlin@nccu.edu.tw

陳志銘 Chih-Ming Chen
國立政治大學圖書資訊與檔案學研究所教授
Professor, Graduate Institute of Library, Information and Archival Studies, National Chengchi University
Email: chencm@nccu.edu.tw

【摘要 Abstract】

數位科技的進步為中國古籍之檢索與運用，提供前所未有的挑戰與發展契機。本文介紹國家圖書館與政治大學合作進行之「國家圖書館古籍數位人文平台建置計畫」執行過程與成果，此一通用型古籍數位人文研究平台，為國圖古籍資料提供創新前瞻之應用，不僅可激發人文學者發掘古籍資料在人文研究上的新面向，並可促進一般社會大眾對於古籍的認識與學習興趣。

The advancement of digitization technologies for the retrieval and use of Chinese ancient books is arising an unprecedented challenge and opportunity. This paper describes the process and results of the "Plan for the Digital Humanities Platform for the National Central Library's Chinese Ancient Books," the National Central Library in cooperation with the National Chengchi University. This universal type of digital humanities research platform on ancient book will provide innovative and forward-looking applications. The results not only can stimulate the humanist to

explore new facets in the humanistic research, but also can promote the public to emerge the learning interest and awareness of ancient books.

【關鍵詞　Keywords】

數位人文、特藏古籍、數位化、漢學研究

digital humanities; old rare books; digitization; Chinese studies

一、前言

臺灣於 2002 年展開「數位典藏國家型科技計畫」，將各種國家級珍貴檔案、文物進行數位化典藏，並建立資料庫，提供國內學者進行人文及社會的研究，經數位典藏保存之資料，多數採數位影像掃描方式，建置資料庫提供查詢使用。然而，資訊技術的快速發展，純粹檢索已無法滿足使用者需求與支援研究所需，古籍的使用需求不再只是停留在對於文本內容大量閱讀的啟發，藉由資料庫功能設計，研究者對於數位文本內容的搜尋、詮釋、比對需求日益殷切（吳明德、黃文琪、陳世娟，2006；廖益賢，2012）。加以自 2012 年行政院國家科學委員會（2014 年改制為科技部）開始推動「數位人文主題研究計畫」，鼓勵國內人文學者進行數位人文學，數位人文研究自此蓬勃發展。

數位人文學者初期嘗試以特定學科領域進行單一面向、獨特議題之技術開發及分析，例如將地理資訊系統（Geographic Information System，簡稱 GIS）、視覺化工具、全文資料標記（如 MARKUS）、社會網路分析、文本探勘、主題分析等工具與特定人文議題結合。經過此階段的技術探索，將數位典藏資源加值與分析運用的發展日趨多元，如何回應使用者需求並將工具整合與平台公開化，將是數位人文學研究接下來發展的里程。

國家圖書館（以下簡稱國圖）長久以來致力於珍藏古籍文獻之維護與整理，典藏善本等級古籍數量逾萬部以上，藏品中明人文集占臺灣各機構典藏總數逾半以上，享譽國際漢學研究。明人文集不僅可提供歷史面向研究，也收錄作者對於哲理諸多問題的見解，能提供中國哲學思想研究的素材（濱島敦俊，2001）。文集所收錄的文體涵蓋明代文學所有的體裁，體裁多樣性也意味著內容的多樣性，明人文集作者身分地位

廣泛，所接觸階層不盡相同，觀察社會角度觀點也不同，使得明人文集內容涵蓋多元面向研究議題，是瞭解明代政治、社會、文化發展最重要的素材（陳寶良，2001）。

惟古籍數量浩繁，過往人文學者往往需要戮力鑽研，浸淫比對各文籍載錄之資訊，始有所得；而對於一般大眾而言，也因古籍內容資訊背景與採用語彙和現今多有出入，實難以一探究竟。有鑑於此，國圖與政治大學社會科學資料中心（簡稱政大社資中心）合作，以建置能提供漢學研究學者與一般大眾均可利用之古籍數位人文平台為目標，開發通用型的古籍數位人文研究平台，為國圖數位古籍資料提供創新前瞻之應用，可激發人文學者發掘國圖古籍資料在人文研究上的新面向，並促進一般大眾對於古籍價值的認識與學習。

「國家圖書館古籍數位人文平台建置計畫」（以下簡稱本計畫），以國圖已完成影像掃描的明人文集數位文本為標的，建立資料全文數位化作業流程，進行全文轉製並建立後設資料（metadata）後，導入數位人文平台系統，於系統中外加資料整理、統計分析、文本閱讀、資訊視覺化、文本自動標註、社會網絡分析等數位工具。本計畫將近年政大社資中心在數位人文工具開發上的成果，嘗試匯整於此平台系統，以激發人文學者發掘新的研究主題，更期許能開啟社會大眾對於古籍內容探查的興趣。

二、數位人文研究平台規劃構想

從人文學到數位人文研究的發展歷程觀察，傳統人文學的研究，多以個人的興趣與專長透過大量的文本閱讀，或專注在史料的蒐集與考證上，經不斷鑽研探討特定主題，以提出新觀點。由於需要以研究者個人思維的角度，進行文獻分析與詮釋，故通常人文學者偏好獨自研究，較少激發出跨領域、跨學門的合作。然而在數位時代，當數位資料量以倍速成長，資訊技術對於處理大量資料的能力越來越快時，開始產生了大數據（big data）的研究議題，在應用資訊技術進行人文研究上，也逐漸出現了跨學科合作的「數位人文（digital humanities）」研究。從大量的文字資料中，進行詞頻計算，或挖掘與探查某些現象及其隱含的關係，或找出複雜的人際網絡關係，甚至直接利用網路蒐集歷史資料所發展出的公眾史學（public history），更有因應現代人對視聽媒體的接受程度高，而利用圖像化、聲光方式展現研究成果等，此皆屬於數位人文研究方法的範疇（項潔、涂豐恩，2011）。此一發展為人文學研究開創了不同面向、不同思考取向的研究議題，也提供新的數位研究方法，得以發掘新的議題。

本計畫目的在於引進數位科技技術，以擴展國內文史研究之視野，建構古籍全文資料環境，作為未來發展人文資料分析之大數據。「古籍數位人文平台」之建置，是在國圖原有古籍典藏與查詢系統的基礎上，運用科技研究成果與使用者訪談蒐集需求資訊，建置能符合漢學研究需求的數位人文研究環境，因而提出具備開放性、易用性與展示性的數位人文研究平台。整體計畫架構如圖 1。

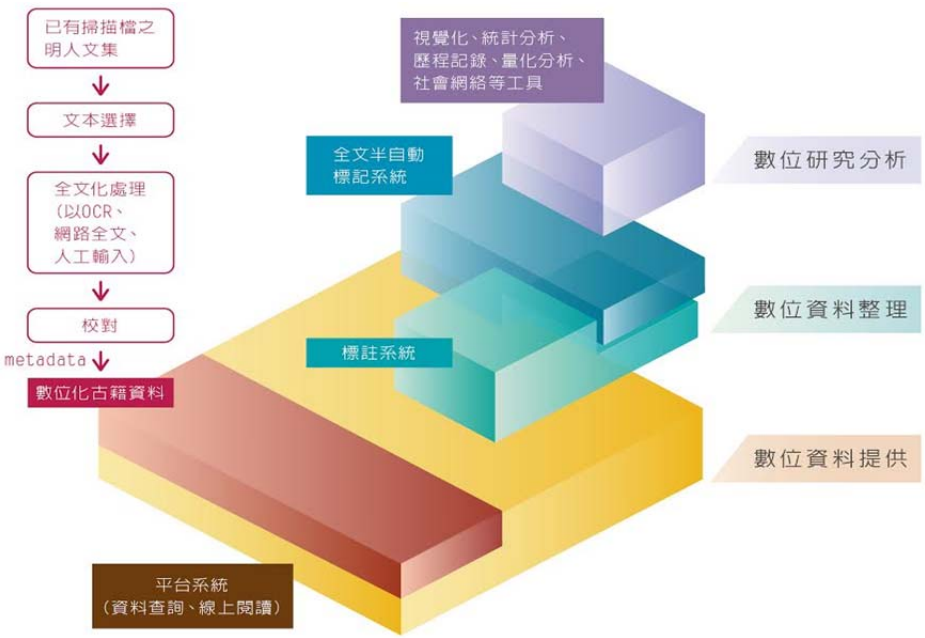

圖 1　本計畫整體架構圖

整體計畫架構是由「數位人文平台系統功能建置」以及「全文資料轉製與導入」兩大部分組成，而數位人文平台建置的部分，除了最底層的全文資料庫平台系統，提供數位資料的典藏、提供資料查詢與線上閱讀功能外，亦包含數位資料整理及數位研究分析等二層次的工具開發。本計畫兩項核心工作理念，分述如下。

（一）數位人文平台系統建置：建置與開發具備通用特性之古籍數位人文平台系統，整合全文資料庫檢索功能與數位人文研究工具。

政大社資中心進行此數位人文平台之技術開發，首先需先建立具全文資料庫之數位典藏系統，使其具備承載全文及後設資料功能，提供完善的資料瀏覽、檢索、檢索後分類等查詢功能，以及全文文本與掃描影像的對照閱讀環境。

接續於系統之上外加開發通用性之數位人文研究數位工具，在數位資料整理層面，以公開的人名、地名、官職名、年號等詞庫，進行文本中對應詞彙的標記與調整，而成為全文半自動標記系統，同時可透過 API 連結查詢解釋專有名詞等外部參考資源，如「中國歷代人物傳記資料庫」（China Biographical Database Project，簡稱 CBDB）、「中國歷史地理信息系統」（China Historical GIS, CHGIS）等公開辭彙工具。此外，亦加入合作標註系統之概念，讓人文學者對於文本能加註個人的解釋及補充資料，並可進行合作資料解讀與討論辨證。而在數位研究分析層面上，則提供統計分析、量化計算、視覺化呈現及社會網絡等數位人文計算與呈現工具，並進一步規劃納入使用者之歷程紀錄，可分析使用者在系統中探索使用資料之歷程。

（二）全文資料轉製與導入：探索國圖明人文集主題特色挑選核心文本，同步規劃發展數位影像轉製全文之作業流程與規範。

本計畫於系統建置之際，同時展開全文資料轉製與後設資料導入工作。國圖典藏明人文集數量豐富，基於全文轉製之成本與效益考量，優先選擇具有研究需求與使用族群之文本，進行全文轉製。本計畫根據過往明人文集著作之研究主題與內容取向，找出優先轉製全文之核心文本，並比對國圖現有之掃描影像檔，擇選本計畫此階段需要轉製之 40 種文集內容。

光學文字辨識（Optical Character Recognition，簡稱 OCR）技術雖可協助全文轉製，但影響 OCR 軟體辨識精確度因素，包括軟體本身、語言複雜程度以及辨識文件的清晰度（Zhu, Tan, & Wang, 2001；Balk, & Ploeger, 2009；Holley, 2009；Zhou, 2010；Cojocaru, Colesnicov, Malahov, & Bumbu, 2016）。由於古籍全文辨識問題複雜度高，涉及因素不只單一，需要綜合各項因素，故將兼採字元辨識、人工輸入校正等彈性交叉運用方式。本計畫針對明人文集文本特性，經討論評估先採 OCR 辨識後，再輔以全文比對與人工校正，基於實測作業經驗，發展出最有效率之全文建置過程。

本計畫為期 10 個月，計畫內容是將原本僅有掃描影像的明人文集進行全文數位化轉製，並建置與掃描影像能對照瀏覽的資料庫系統，同時結合使用者需求，於平台中研發各種數位工具，提供使用者針對古籍全文進行分析與探索。完成之系統平台，

未來可不斷新增全文數位文本內容，並透過與學者的合作，發掘古籍內容價值，擴展古籍研究成果。此外，亦可結合政大社資中心近年的數位策展經驗與技術，進行視覺化與數位展示，向社會大眾呈現整體計畫的成果，促進終身學習的理念。

三、數位內容建置與平台功能設計

（一）古籍內容全文轉製

根據《國家圖書館善本書志初稿》收錄之明人文集，其中別集、總集兩類合計為1,497 種，收錄種類豐富，基於全文轉製成本與效益考量，優先選擇具有研究需求與使用族群之文本。本計畫選擇文本的原則，係綜合下述層面考量：（一）分析已發表之明人研究論著內容，找出過往研究探討之文集名稱、研究議題；（二）探詢明人研究學者認知重要的主題與文本，加入使用者推薦優先轉製之全文標的；（三）網路或資料庫已有載錄部分或完整全文的明人文集，如已有前人貢獻全文，意味有使用需求存在，亦為優先考量因素。

綜合上述三項因素分析，訂出本計畫進行全文轉製或下載修正的數位化處理文本清單。並進行下述階段的全文轉製過程處理：

1. 文字辨識處理

首先評估分析 OCR 軟體處理古籍之識別程度，以尋求影像轉製全文作業流程的最佳處理方式。由於影響 OCR 軟體辨識精確度的因素，包括軟體本身、語言複雜程度以及辨識文件的清晰度。儘管目前 OCR 辨識率尚未達到百分之百的精確，然而如果能達到理想的 OCR 辨識率，將可節省人力以及時間的耗費。

本計畫挑選之 40 種國圖典藏的明人文集，是分布於萬曆至嘉靖年間刊刻的文本，將每種文集隨機抽樣挑選 5 頁（1 頁線裝書含 2 頁影像），合計 40 種共 400 頁的古籍影像，進行 OCR 軟體的辨識比較。影像經 OCR 辨識結果，各文集的辨識率差異極大，辨識率較佳者可達 0.858，但辨識率不佳者僅有 0.325，40 種文集之平均辨識率僅達 0.588（參閱表 1）。

表 1

明人文集 OCR 辨識率統計表

次序	文集名稱	行段數	行字數	辨識率
01	誠意伯文集	11	21	0.381
02	王文成全書	9	19	0.591
03	蟣蠓集	10	19	0.524
04	鳳池吟稿	9	19	0.540
05	空同集	11	20	0.458
06	槎翁詩集	10	20	0.567
07	四溟集	10	20	0.653
08	白沙集	9	18	0.728
09	備忘集	9	19	0.495
10	存家詩稾	9	18	0.570
11	覆瓿集	9	20	0.744
12	具茨集	10	20	0.772
13	西隱集	9	20	0.327
14	薛文清集	10	20	0.410
15	考功集	9	18	0.733
16	練中丞集	9	20	0.518
17	西菴集	8	18	0.520
18	柏齋集	10	22	0.692
19	方簡肅文集	9	18	0.813
20	芻蕘集	10	22	0.534
21	滄溟集	10	20	0.620
22	讀書後	8	18	0.858
23	皇甫少元集	10	18	0.671
24	迪功集	9	16	0.770
25	遜志齋集	10	20	0.527
26	大復集	10	20	0.577
27	類博稾	10	20	0.473
28	荊川集	10	20	0.659
29	羅圭峯文集	11	22	0.492
30	康齋文集	10	21	0.439
31	陸子餘集	10	18	0.753
32	竹澗集	12	20	0.745
33	望雲集	10	20	0.691
34	楓山集	10	20	0.542
35	重編瓊臺會稾	11	24	0.395
36	華泉集	10	22	0.543
37	甫田集	11	21	0.740
38	蘇門集	10	16	0.325
39	震澤集	11	20	0.645
40	謙齋文錄	10	20	0.477

資料來源：本研究整理。

由於古籍版本會依朝代、印刷、版刻的不同，而有不同的字體、版式、字型呈現（駱偉，2004）。古籍文本不同於現今圖書，會有更多影響 OCR 辨識率的因素，其中尤以文字字體、文字字型、文字大小、文本狀況等因素對 OCR 辨識率影響極大。說明各項因素對於 OCR 辨識的限制與影響如下：

（1）文字字體

文字在歷經演變的過程，會產生不同的字體，這些字體因型態的差異，造成 OCR 辨識的困難。以中文為例，中文字體包含隸書、楷書、草書、行書等，即使同樣字體在跨朝歷代之後，也可能產生型態的差異，但文字形態的差異會造成字形辨識的誤差。如果將同文字的不同字體，皆儲存至字辭庫中，除辨識時間會拉長外，系統儲存空間也需要相當的負載。因此，最好的解決方式，是針對辨識文本建立專門的字體庫，或運用演算法辨識文件內有差異的字體，不僅可減少系統空間負擔，也能提升文字辨識速度（曾逸鴻、林裕淵，2007）。

Cojocaru, Colesnicov, Malahov 及 Bumbu（2016）在掃描辨識 18 至 20 世紀羅馬的印刷書過程中，發現書中並存各時期字體，而有不同字體穿插出現，勢必影響 OCR 辨識的成效。為解決此字體差異的問題，則建立不同字體模板及字體對照表，透過字體交叉比對，輔以軟體字辭庫資料，將不同字體辨識出來，並在輸出時轉製成相同字體。藉由字體對照表的建立，能方便不同字體的字母比對，協助辨識涵蓋多種字體內容的文本。

（2）文字字型

OCR 軟體不擅長進行手寫辨識，原因在於手寫字體不同於印刷字體的型態固定，由於個人書寫習慣的差異，容易造成文字字型沒有規律性，尤其是手寫文字，在字與字之間的空間很不一致。OCR 是運用空白區間以辨識文字最小單元，但過於靠近的文字會造成辨識困難，使得辨識結果混亂。藉由智能字元辨識（intelligent character recognition）的技術，運用神經網絡（neural network）的分割手法，將過於靠近的文字，自動調整切割區塊，能協助辨識不同的手寫文字（Mariner, 2010）。

（3）文字大小

文字前處理需將文字與背景相互分離，此時，字體的區間與大小是影響分離的關鍵。文字區間的空白可協助文字分割，但字體大小也同時影響字與字之間空白區的大小，亦即越大的文字，區間空白也越大。顯然較大文字更方便文字與背景的分離；相反地，越小的文字，區間空白越小，必

須靠高解析掃描或文字線條粗化的輔助，才能確保文字與背景順利分離。文字辨識前處理，影響後續特徵抽取的成效，文字大小對於辨識結果的正確性會有影響（Zhu, Tan, & Wang, 2001）。

（4）文本狀況

年代久遠的文本，可能因保存不佳等因素，造成紙質變形、文字扭曲等現象，或因印刷技術不純良，導致印刷品質不佳、掃描透頁、字跡暈開等現象，更可能會因印刷本年代差異，產生用字過時、字體過小或版面不規律等狀況，此皆會造成 OCR 辨識正確率問題（Balk, & Ploeger, 2009）。余顯強（2005）進行《北平世界日報》數位典藏計畫時，曾採數位化方式協助報紙保存，並建置全文資料庫供檢索使用。民國初年的世界日報，不僅年代久遠，又經歷戰亂時期，使得保存狀態不完好，加上報紙用字的差異、排版的不規律、印刷紙質不佳等因素，在諸多因素考量下，最終捨文字辨識而決定改採逐字輸入的方式建立全文資料庫。

根據前述研究分析，本計畫著手進行全文化的古籍，是明朝文人文士所著述的作品，經自身或是後人加以彙編整理，刊刻成為今日所見之古籍集冊。明代雖是中國歷代印刷事業興盛的朝代，但此批文集是由不同彙編者所收錄整理，文集版本可能由於印刷坊刻字的數量或是彙編者加以刪修內容的結果，造成同一文集會因版本差異而內容略有不同。由於本計畫以 OCR 測試此批明人文集影像轉換全文的辨識率不高，加上影響古籍全文辨識率因素絕非單一因素，不僅有字體與文字大小問題，加上原有影像檔品質不清晰，實難以逐一排除問題。因此，改採全文比對、人工輸入校正等交叉運用方式。

2. 全文比對與校正

經查詢比對國圖所藏的明人文集，絕大部分收錄在清乾隆時期編撰的《四庫全書》中，由於《四庫全書》為中國歷史上最大規模的編撰叢書，故優先比對已在網路可提供查詢之四庫全書資料庫，陳述全文與人工校正之作業程序如下。

（1）比對全文資料庫使用

《四庫全書》是中國經典叢書，經兩岸資料庫建置者貢獻，在網路上已有各種資料庫提供不同程度的全文內容。本計畫經評估重要古籍資料庫後，衡量各資料庫優劣，兼採三種全文資料庫，進行文集全文的比對與參考，其為「中國哲學書電子化計劃」（http://ctext.org/zh）、「Kanseki

Repository」（https://www.kanripo.org）以及「文淵閣四庫全書電子版」（3.0 版）。

前兩者為網路開放的免費資源，可提供交互參考，以彌補單一全文資料庫內容的缺漏；後者是需付費使用的全文資料庫，其全文量與內容較為完整。此兩類全文資料庫使用上各有專擅，在網路開放的免費全文僅單純收錄古籍文字，內容缺漏的情形較為明顯；至於付費使用的全文資料庫，具有模擬古籍版面原貌的功能，並能以卷冊標示全文內容，因此在進行文集全文校對上，是以「文淵閣四庫全書電子版」為主，輔以兩個免費資料庫相互參考補正，以減少內容闕漏的問題。

（2）全文校正步驟流程

全文校正流程分為六個步驟：選擇適當全文資料庫、取得全文文字檔、複製貼至文字檔、調整文字檔格式、校正全文內容、上傳至系統。上述六個步驟以流程圖示（圖 2）呈現，並逐一說明操作過程如下：

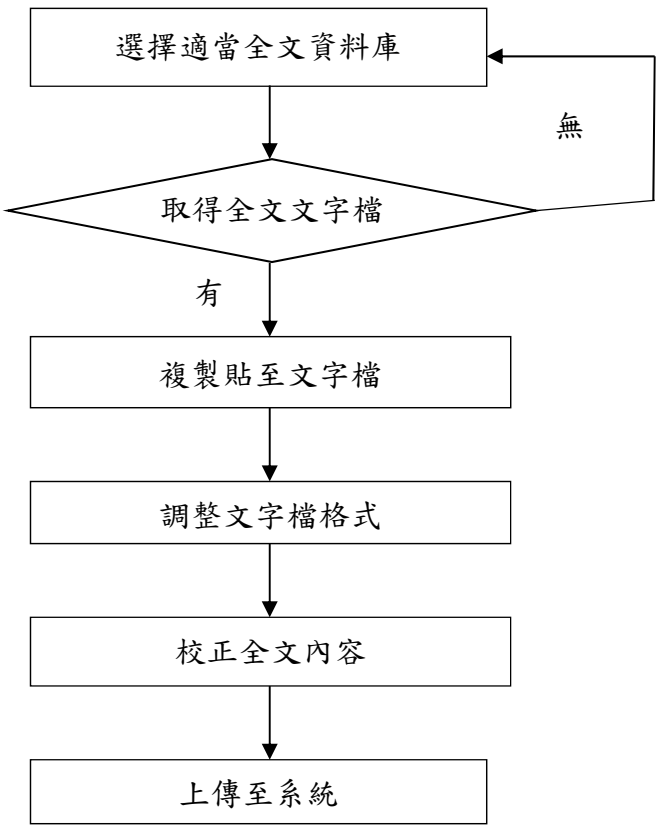

圖 2 校正步驟流程圖

①選擇適當全文資料庫：主要使用前述三種全文資料庫作為文集內容文字比對的參考，三者所收錄內容雖有全文數量差異，但經比對相同文本的內容文字，差異不大。

②取得全文文字檔：在文集內容比對校正時，若遇到缺漏或內容版本差異較大時，先比較不同全文資料庫收錄文本內容的差異，以最為接近國圖影像內容的版本為複製依據。

③複製貼至文字檔：複製儲存之檔案格式為 WORD 檔，執行過程中曾嘗試使用 WORD 檔與 TXT 檔相互比較，結果發現，WORD 檔能模擬古籍由右至左、由上至下的書寫方式，而 TXT 檔無法模擬，僅能以現代化由左至右的書寫方式；字體方面由於古籍許多字體是古體字或異體字，TXT 檔無法支援造字所產生的文字檔，在表現上會以空白方式呈現，經此比較後，選擇使用 WORD 檔為檔案儲存形式。

④調整文字檔格式：為了方便快速校正文字檔內容，須先調整 WORD 檔格式。先將 WORD 設定成直書模式，路徑為「版面配置」／「直書／橫書」／「垂直」；接著，調整 WORD 版面，路徑為「版面配置」／「方向」／「橫向」，讓 WORD 檔可以模擬古籍文字內容走向。對於字體的選擇以「標楷體」呈現，採用標楷體的原因是文集文字較接近標楷體；在字體大小上，為了便利校對流程的順暢，盡可能採接近古籍版面行字數設定字體大小，讓影像檔與複製貼上之文字行字數一致，較方便在電腦螢幕上直觀校對。此批明人文集不同文本的版面行字數不同，經多次比對設定後，找出行字數與應設定之字體級數大小對應如表 2。

表 2

古籍行字數與字體級數大小對應參考表

行字數	16	18	19	20	21	22
字體大小	25	22	21	20	19	18.5

⑤校正全文內容：校正時逐頁對應古籍影像進行內容文字的校對，為順暢校正的作業速度，經前述調整文字檔字體大小與影像檔一致後，依照古籍影像版面進行文字檔版面調整。若文字位置為空兩個字為開頭，則文字檔也依樣調整成空兩個字，讓文字檔與影像檔版面呈現一致，如遇影

像檔出現空頁或空行情形，文字檔也如實仿照。但遇古籍影像字體模糊以及無法判斷文字時，先交叉比對不同全文資料庫的內容，協助分辨該文字為何，若經比對其他全文資料庫仍無法判斷該文字時，則以輸入全形的方框（□）暫為替代處理，有待日後學科專家建議再行補正內容。

⑥上傳至系統：將所完成的古籍全文文字檔以卷為單位，存成一個 WORD 檔，檔案命名方式：第一層命名規則為「文集編號-00000」，文集編號係與國圖原影像檔編號一致；第二層將序、卷次、別集、附錄分開命名，以便未來使用者查詢介面可分集、分冊各別呈現。

（3）全文內容校正要求

全文校正作業過程中發現諸多文字判別問題，由於明代至今約 500 年，文字字形也歷經 500 年演變成為現今使用的文字樣貌，500 年前明人文集使用的版刻字樣，部分文字與現今常用字體不太相同。

針對字體樣貌的差異，原先考慮採忠於古籍影像呈現之原字體文字，亦即根據影像文字樣貌輸入外型相同的文字。但處理後發現，許多文字需要另行造字，才能出現完全符合樣貌的文字，而造字所產出的文字無法以常用的輸入法再現。因此，針對古籍所含古體字及異體字的輸入，建立以下處理原則：

①優先輸入可出現的字體

中文輸入法有許多種，本計畫作業執行人員擅長使用的輸入法為「新注音輸入法」，在忠實呈現古籍影像文字為優先前提情形下，部分古體字或異體字可以藉由「新注音輸入法」顯示，例如「将」、「从」、「剥」、「吴」（亦即為今之「將」、「久」、「剝」、「吳」）等字，這些字體雖然並非現今使用的文字樣貌，但由於使用輸入法可以輸入出現古體文字，因此當現今輸入法可顯示古體字時，則直接輸入使用。

②無法藉由輸入法呈現則輸入現今對應字體

當古體字非現今所使用的文字，而且「新注音輸入法」也無法輸入呈現相同字體時，則利用「國際電腦漢字及異體字知識庫」（http://chardb.iis.sinica.edu.tw/）進行古體字查詢，通常使用部件或是相似字的檢索模式，可以找尋到相對應的現今字；另一途徑是使用「教育部異體字字典」（http://dict2.variants.moe.edu.tw/variants/rbt/home.do），

而「教育部異體字字典」找到的異體字比「國際電腦漢字及異體字知識庫」更多，但仍須兩者交互使用。由於異體字無法藉由輸入法輸入顯示，為了方便日後研究比對，本計畫將兩項資料庫所收錄的異體字影像下載，對應現今文字整理完成紀錄，預計可提供研究者未來探索古籍文字差異之比較分析，摘錄部分古體刻字與現今字體之比較例示如表 3。

表 3

古體刻字與現今字體之比較舉隅

可顯示字碼		不可顯示字碼			
古體刻字	現今字體	古體刻字	現今字體	古體刻字	現今字體
隱	隱	覩	睹	旨	旨
乆	久	壤	壤	俾	俾
剥	剝	嘗	嘗	講	講
将	將	歸	歸	歸	歸
徃	往	踈	疏	髙	高
吳	吳	履	履	說	說
吳	吳	鼏	鼎	若	若
悱	悲	負	負	収	收
九	凡	恠	怪	纔	纔

③完全無法找到對應字且無法正確判定時以方框代替

當上述兩種狀況皆無法有效處理時，因無法正確判斷文字，則以方框（□）代替文字。在校正過程如果實在無法判定文字為何，絕不妄加臆測給字，與其錯字誤引，不如以方框替代，反而日後可藉由學者研究內容，協助判斷所缺文字為何，必要時，可用方框找尋更新補正文字。

除了異體字處理上需要制定統一作業標準外，尚有不同版本內容多

寡的差異，小則有些許段落內容文字不同，大則發生收錄詩、詞、文段整個不同，轉錄文字採行與國圖影像版本一致為基本原則，但針對國圖版本與網路全文版本內容的差異，本計畫為求審慎處理，特針對校對後缺少或多餘的文段留下比對紀錄，試圖為日後研究明人文集版本內容差異之研究者，提供基本資料參考。所有完成之文字檔將陸續匯入平台系統，提供系統功能測試與未來開放查詢運用。

（二）系統平台功能開發

本計畫以開放原始碼的數位典藏軟體 DSpace 為基礎發展系統功能，除了作為數位資料儲存平台外，尚具備後設資料設計、數位典藏、身分認證與權限控管機制等基礎功能，並具備高度客製化內容層級架構、工作流程設計、支援多種國際標準資料交換 OAI-PMH 與 OpenURL 等通訊協定之優勢。數位人文研究平台之技術開發，需先建立具全文資料庫之數位典藏內容為基礎，除了需具備承載全文及後設資料之功能，並提供完善的資料瀏覽、檢索、檢索後分類等查詢功能外，尚需提供全文文本與掃描影像的對照閱讀環境，並針對使用者需求重新設計整體網站介面，以提升使用者介面之優使性。

本計畫於系統之上再外加通用性之數位人文研究數位工具，在數位資料整理層面，以公開的人名、地名、官職名、年號等詞庫，進行文本中對應詞彙的標記與調整，而成為全文半自動標記系統，同時可透過 API 連結查詢，能解釋專有名詞的其他外部參考資源或其他免費公開之辭典。此外，加入合作標註系統之概念，讓人文學者對於文本能加註個人的解釋及補充資料，並進行合作資料的解讀與討論辨證。再者，在數位研究分析層面，則提供統計分析、量化計算、視覺化呈現及社會網絡分析等數位人文計算與呈現工具，並進一步規劃納入使用者之歷程紀錄，未來可分析使用者在系統中探索使用資料之歷程。平台各功能的介紹將依層次順序詳細介紹，分述如後：

1. 數位資料提供

（1）後設資料

後設資料除可以協助藏品的儲存、控制、管理、散布與交換數位資源外，還可協助使用者搜尋、辨識、選擇、詮釋、獲取與使用數位資源。本計畫所典藏的文集為明代古籍，出版與典藏環境經過數百年的變遷已經大幅改變，為了縮短使用者瀏覽時與文集的差距，除基本的題名、作者、版本、出版資訊與外部形體特徵的稽核項外，尚有序跋、落款、印記、版本

行款等詳細記載古籍特徵的後設資料，提供研究古籍資料的專業學者充足的研究資訊，以縮短研究時程。

（2）圖文顯示介面

為模擬使用者在使用平台瀏覽文集時，彷彿閱讀真正的明人古籍一般，因此開發圖文顯示介面。系統在呈現文集內容時，採用掃描圖檔與文集全文共同顯示的圖文介面，提供同時閱覽。使用者可搭配字體大小與圖片大小的調整功能，隨著顯示螢幕大小與使用需求自行調整字體大小與圖片大小，以利瀏覽閱讀。此外，此圖文顯示介面亦擁有文集目錄功能，可讓使用者瀏覽文集的整體架構，並且搭配超連結功能可迅速瀏覽點選的章節進行閱讀。使用者除可以透過掃描圖檔瀏覽文集排版、字型等原本的樣貌，並搭配平台系統所提供的刻工、印記、牌記、序跋與裝訂等後設資料相互參照，研究版本學外，亦可搭配以現代字體呈現的全文，加速瞭解文集內容且搭配全文檢索功能迅速查找所需內容，以利更有效的利用明人文集進行研究。

（3）全文檢索

近年來數位典藏為提供一站購足的使用者體驗，以典藏全文為平台目標，但本計畫平台典藏的明人古籍為非結構化的文本資料，共 1 萬 4 千多頁的文本內容，使用者難以在短時間內用傳統閱覽的方式找到欲查找的內容。為了節省查找所需資料的時間，在計畫平台中提供全文檢索功能，使用者可自行輸入欲查找的關鍵字，讓系統幫忙檢索出符合關鍵字的典藏內容，以提高典藏資料的使用效率。

（4）檢索後分類

若使用單純的檢索功能，可能會因為檢索結果眾多，致使需要花費更多額外的時間找尋所需資源。因此，如果系統能依據檢索結果的資料特色進行初步分類，即可幫助使用者快速瞭解檢索結果的資料分佈概況。本計畫平台提供檢索後分類功能，可將查詢結果以預設之不同特性進行分類，目前系統以文集的後設資料為基礎，分別進行作者與年代的後分類，幫助使用者在眾多的檢索結果中迅速找到欲查找的資料。

2. 數位資料整理

（1）新詞探勘

現今的自動化文本處理技術大多仰賴詞庫進行斷詞，以便後續對文本進行更進一步的處理。在利用詞庫方式進行中文斷詞時，文本若出現未在詞庫內建置的新詞，將會降低斷詞的正確率。因此，本計畫所開發的數位

人文研究平台會利用新詞偵測演算法找出新詞後，由人工協助判斷其正確性方式將新詞加入詞庫，藉此增強斷詞之正確性。

換言之，本計畫所採用的是半自動化的方式產生新詞，當在閱讀文本的過程中，若發現未自動標註但為新詞的詞彙，使用者可以自行將詞彙標註為新詞，此時系統會自動將詞彙新增至詞庫中，進而改善斷詞與自動標註之正確性。

（2）標註系統

本計畫平台所典藏的內容為明人古籍，文本內容非現代慣用語，為幫助使用者在短時間內理解典藏內容，本計畫發展自動標註系統，藉由實作 linked data 匯聚來自各個資料庫資源並加以整合，進而替文本進行自動註解，以方便使用者能即時參照其他資料庫資源，在短時間內瞭解文本內容，並設計友善標註閱讀介面以利資料解讀。圖 3 為自動標註系統架構圖，詳細說明如後。

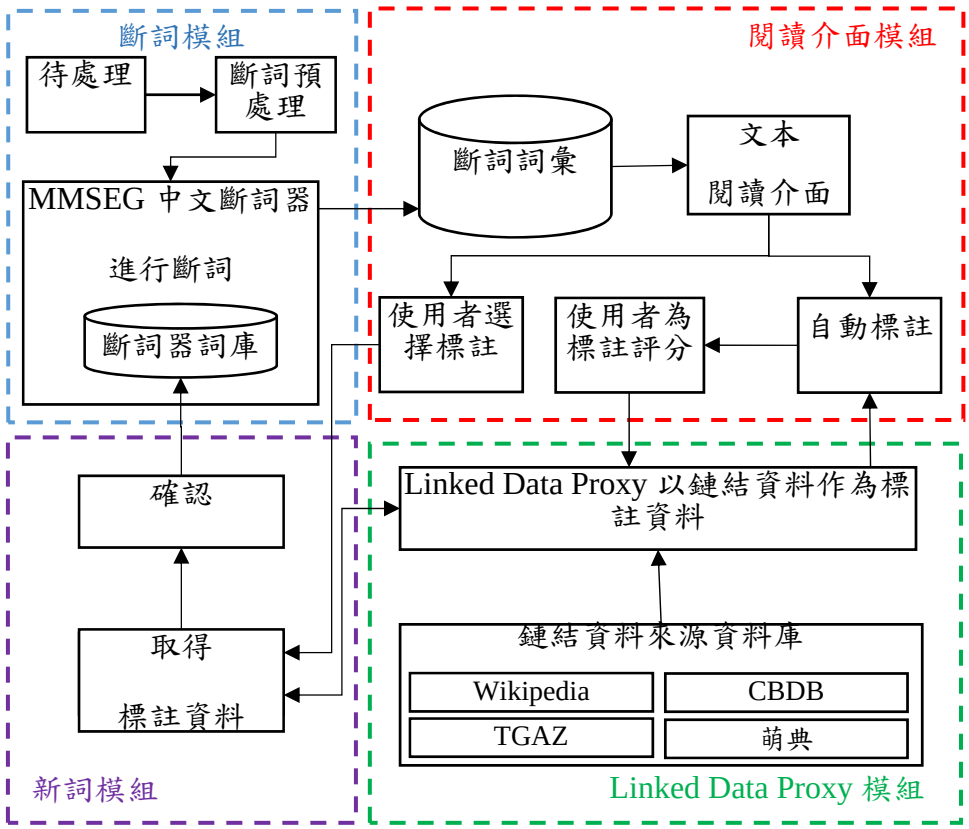

圖 3 自動標註系統架構圖

自動標註系統架構依據其功能，主要可區分為斷詞模組、閱讀介面模組、新詞模組及 Linked Data Proxy 四個模組，以下針對不同模組的特性，依序詳細說明。

①斷詞模組

為使系統順利處理自然語言，必須進行斷詞，將正確的詞切分出來。由於中文結構的關係，在詞與詞之間並無明顯的分界符號，不同於英文在字與字之間存在空格，因此需在斷詞前進行預處理將句子間隔開，再輸入斷詞器中進行斷詞。標註系統所使用的斷詞器是基於最大匹配法且以詞庫為基礎的 MMSEG 中文斷詞器，並將經過斷詞處理的詞彙存入中介資料庫中，以方便系統後續利用。

② Linked Data Proxy 模組

為使用者在閱讀明代文集時，可以即時參考不同網站的資源，建立資料與資料間的連結，因此利用斷詞處理過的詞彙會分別從「維基百科（Wikipedia）」、「中國歷代人物傳記資料（China Biographical Database Project，簡稱 CBDB）」、「Temporal Gazetteer，簡稱 TGAZ」、「萌典」四個資料庫中獲取參照資料，並將資料轉換為資源描述框架（Resource Description Framework，簡稱 RDF），並設計可以針對不同來源資料庫標註給予評分的機制，以作為較為適切標註之顯示排序依據。

③閱讀介面模組

將斷詞過的文本輸出至文本閱讀介面上並顯示全文後，系統就會取用 Linked Data Proxy 模組的資料將全文進行自動標註。待使用者利用滑鼠點擊帶有自動標註的詞彙後，會自動顯示不同來源資料庫資料，此時可針對不同來源資料的適切性進行評分，做為資源顯示排序依據，如圖 4 所示。

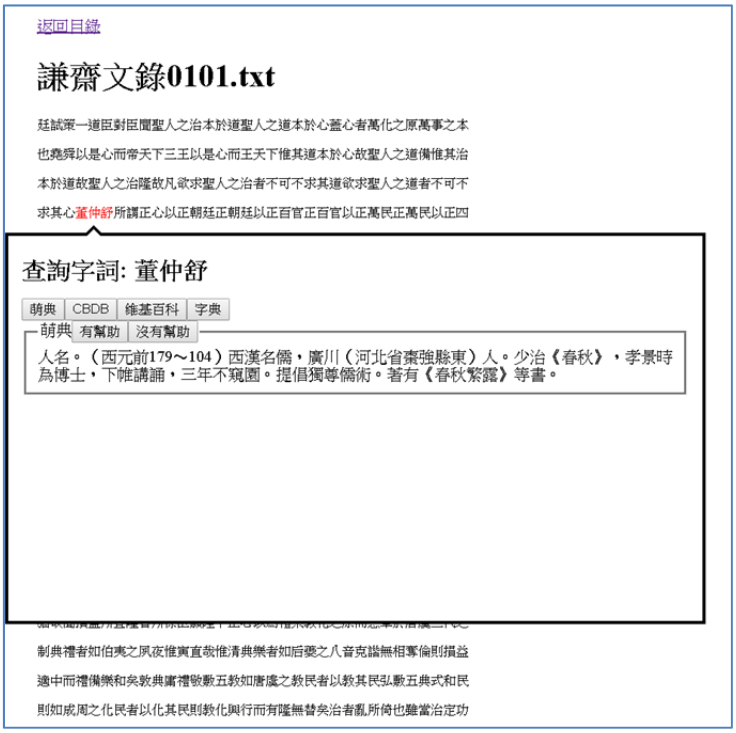

圖 4 “董仲舒”透過 Linked Data 之自動標註結果

④新詞模組

使用者標註未知詞為新詞時，系統會將詞彙傳送至 Linked Data Proxy 模組中搜尋是否有可作為註解的資料並回傳，作為是否要將該詞彙添加為新詞的依據。經由使用者確認為新詞後，該詞彙新增至斷詞器的詞庫中，提升斷詞結果正確性，如圖 5 所示。

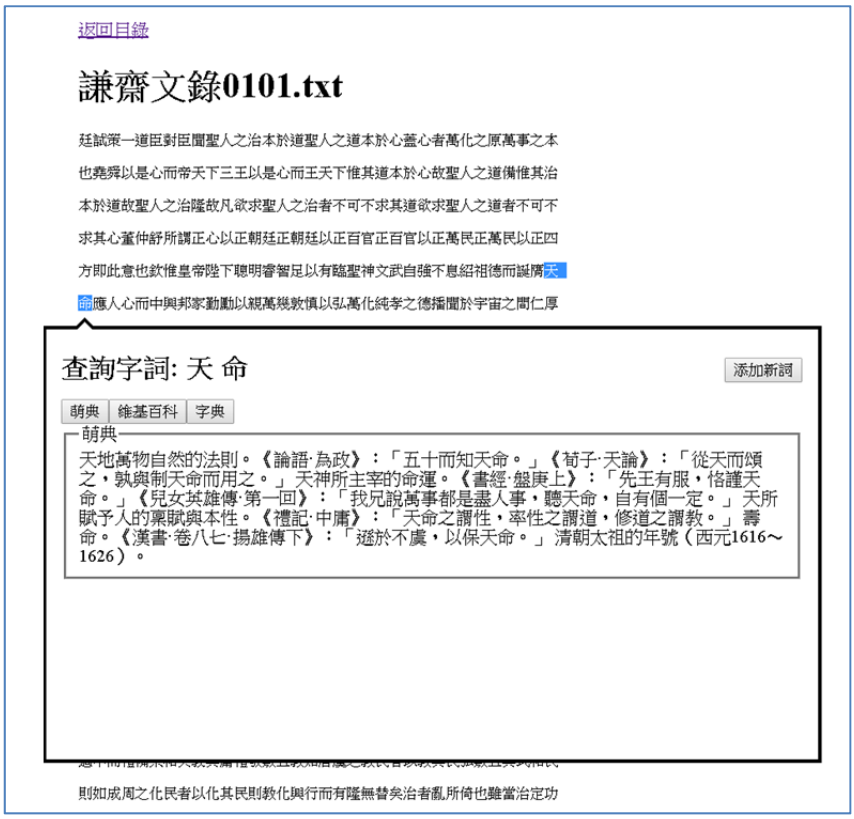

圖 5 由人工輔助判斷「天命」為一個有意義的新詞

3. 數位研究分析

本計畫發展的數位人文研究平台除了提供基本的數位典藏資料的後設資料、明人古籍掃描檔與對照全文、全文檢索與後分類、自動文本標註與新詞探勘等數位資料整理與閱讀介面工具外，未來亦將提供資訊視覺化與使用歷程紀錄等研究分析工具，詳細說明如後：

（1）資訊視覺化

多數的數位人文平台僅提供數位化掃描檔，並提供後設資料瀏覽與檢索，少部分平台則提供全文瀏覽檢索以及進行基本的詞頻分析統計功能。但如何幫助使用者在短時間內掌握資料彼此之間的關聯，藉此激發出新的研究方向是發展數位人文平台需要關注的議題。因此，本計畫發展之數位人文研究平台未來將提供資訊視覺化分析，將抽象的資料運用視覺技術予以多面向呈現，以有效縮短使用者理解資料的時間。

（2）歷程記錄

大部分數位人文平台難以精準得知每位使用者使用平台的操作行為與歷程，大多事後透過問卷調查或訪談方式瞭解使用經驗，但由於是請其以回憶的方式進行問卷填寫或訪談，問卷內容往往無法完全真正反映出實際的使用歷程及行為。本計畫平台將嵌入使用者操作歷程監控模組，藉此精確且真實地記錄每位使用者的操作微歷程（micro behavior）。平台預計採用 ADL（Advanced Distributed Learning）所提出的 xAPI（Experience Application Programming Interface）進行使用歷程記錄，此一技術除可紀錄任何時間與地點的操作歷程外，還具備跨瀏覽器、跨平台、支援行動載具與高安全性等特性。xAPI 使用 JSON 格式紀錄將使用者操作微歷程區分成「主詞（Actor）」、「動詞（Verb）」和「受詞（Object）」三個部分進行詳細記錄，其中主詞（Actor）為動作執行者（例如：使用者帳號）；動詞（Verb）為主詞所進行動作類別（例如：搜尋）；受詞（Object）為與主詞互動的受詞（例如：關鍵字），並將這些微歷程紀錄結果傳送到學習歷程紀錄資料庫（learning record store，簡稱 LRS）中儲存，以作為後續分析使用者歷程之基礎。如此可藉由使用者歷程充分了解數位人文學者的瀏覽資料行為、查詢資料方法，以及資料解讀行為等細節，除有助於分析使用者行為作為協助改善平台介面操作功能，增加平台的優使性（usability）外，更有助於瞭解數位人文學者解讀資料的思維、方法與歷程，對於發展更符合人文學者使用之數位人文研究平台具有相當高的助益。

四、平台建置成果與發展

本計畫透過與國圖的合作，建置國圖特藏資料之古籍數位人文平台，具體成果包括：

（一）建立創新服務模式與核心服務內容：透過政大社資中心數位人文研發技術與國圖特藏資源結合，發展可有效協助漢學研究學者創造更多元研究面向與議題的數位研究環境與工具。

（二）發展通用型國圖古籍數位人文研究平台：提供便利、易用且能兼顧研究學者與一般大眾查詢與分析使用需求之數位人文平台系統，並期望未來此平台之建立，可持續推廣到其他圖書館所特藏之數位人文研究服務上。

（三）建構數位人文研究實踐、數位典藏加值、學習知識網路：未來可與人文學者合作，透過數位人文之研究，發掘國圖古籍資料的新應用模式，並普及化推廣讓全民認識國圖古籍之知識運用價值。

（四）建立古籍影像轉製為全文之有效作業模式與流程：可做為數位典藏文件影像轉製全文文本之工作依據，以奠定數位人文研究內容探勘分析之基礎。

（五）促進大學研發能量導入社教機構，並培育人文社會科學創新服務人才：透過政大社資中心技術研發單位與系所共同參與計畫，使政治大學人文社會科學研究群師生與國圖密切互動與合作，強化系所教學及研究能量，並有效連結社教館所第一線服務民眾取用資訊之需求。過程中亦同時培育團隊成員，提供未來社會所需之數位人文創新研究與服務人才。

本文為教育部 105 年「大學以社教機構為基地之數位人文計畫」研究成果。

參考文獻

余顯強（2005）。北平「世界日報」：民初歷史性新聞報紙數位化之研究。*圖書與資訊學刊*，*54*，84-95。

吳明德、黃文琪、陳世娟（2006）。人文學者使用中文古籍全文資料庫之研究。*圖書資訊學刊*，*4*（1/2），1-15。

陳寶良（2001）。明人文集之學政史料及其價值。在張哲郎（主編），*明人文集與明代研究*，（頁 339-358）。臺北市：樂學書局。

曾逸鴻、林裕淵（2007）。中文文件影像中之特殊字體偵測。*科學與工程技術期刊*，*3*（4），29-39。

項潔、涂豐恩（2011）。導論—什麼是數位人文。在項潔（主編），*從保存到創造：開啟數位人文研究*，（頁 9-28）。臺北市：國立臺灣大學出版中心。

廖益賢（2012）。電子文獻全文檢索資料庫管窺—以「中央研究院漢籍電子文獻」資料庫所收《文心雕龍》為例。*中國文化大學中文學報*，*25*，285-303。

駱偉（2004）。*簡明古籍整理與版本學*。澳門：澳門圖書館暨資訊管理協會。

濱島教俊（2001）。日本靜嘉堂所藏《朱文肅公文集》與朱國楨。在張哲郎（主編），*明人文集與明代研究*，（頁 13-28）。臺北市：樂學書局。

Balk, H., & Ploeger, L. (2009). IMPACT: working together to address the challenges involving mass digitization of historical printed text. *OCLC Systems & Services: International digital library perspectives*, *25*(4), 233-248.

Cojocaru, S., Colesnicov, A., Malahov, L., Bumbu, T. (2016). Optical character recognition applied to Romanian printed texts of the 18th-20th century. *Computer Science Journal of Moldova*, *24*, 106-117.

Holley, R. (2009). How good can it get? Analyzing and improving OCR accuracy in large scale historic newspaper digitization programs. *D-Lib Magazine*, *15*(3/4), Retrieved from http://www.dlib.org/dlib/march09/holley/03holley.html#1

Mariner, M. C. (2010). Optical Character Recognition (OCR). In Bates, M. J. & Maack, M. N. (Eds.), *Encyclopedia of library and Information Sciences,* (3rd ed., pp. 4037-4044). Boca Raton, Fla: CRC Press.

Zhou, Y. (2010). Are your digital documents web friendly? Making scanned documents web accessible. *Information Technology and Libraries*, *29*(3), 151-160.

Zhu, Y., Tan, T., & Wang, Y. (2001). Font recognition based on global texture analysis. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, *23*(10), 1192-1200.