BIBLID 1026-5279 (2001) 90:2 pp. 197-216
國家圖書館館刊　九十年第二期(2001.12)

中文古籍全文資料庫建置比較研究

顧 力 仁

摘　要

中文古籍全文資料庫的建立，包括「逐字輸入」以及利用「光學文字辨識技術」等兩種方式。本文以中央研究院的「漢籍電子文獻資料庫」、中華電子佛典協會的「漢文電子大藏經」、香港迪志文化出版有限公司的「文淵閣四庫全書電子版」以及國家圖書館的「古籍無定型工整手寫文字辨識、檢索與管理系統」等電子資源為例來介紹中文古籍全文資料庫的發展現況，並就建立中文古籍全文資料庫所牽涉到的問題，包括：輸入方式、校對問題、缺字及造字、標誌及標準格式、檢索等問題加以討論，以瞭解圖書館目前以及未來在建立中文古籍全文資料庫此一課題上所面對的問題及發展趨勢。

一、前　言

由於資訊數位化及網路技術的演進，促使數位圖書館的興起，數位圖書館具有資訊保存、組織、展示、利用、教育推廣與研究等功能。也由於社會大眾對文化及歷史資產的重視，圖書館珍藏的歷史文獻遂成為數位化的重要對象，例如：UNESCO Memory of the World、National Digital Library Program 以及國內國科

關鍵詞（Keywords）：古籍；古籍整理；全文資料庫；光學文字辨識；標誌
Digital Library；Organization of Ancient Books；Image Processing；MARC；Metadata
顧力仁：國家圖書館特藏組編輯；E-mail: klj@msg.ncl.edu.tw

會擬進行的「國家典藏數位化計畫」都是重要的圖書館珍藏文獻數位化計畫。

隨著資訊科技對圖書館作業型態的影響，古籍的整理也起了相當大的變化，武亞民論及「數字圖書館與古籍整理的關係」中認為：[1]

1. 數字圖書館。（即「數位圖書館」，另有譯成「數據圖書館」）將各種不同類型的文獻信息有機地結合在一起，因此古籍成為數字圖書館中不可缺少的信息源之一；
2. 中國的數字圖書館在建設的過程中必須考慮和首先解決古籍的中國傳統文化特徵，這是具有中國特色的數字圖書館的關鍵。

由此可知，借諸現代化的手段，古籍的整理、組織與利用逐漸成為圖書館的重要職責。

此外，謝清俊教授更強調在網路環境中，古籍可藉由電子媒體充分將其優越性質表達出來，因為：[2]

1. 古籍的電子版本可無限地複製，是取之不盡、用之不竭的資源，可供全民共享。
2. 透過網路，電子古籍可以瞬息千里，沒有運輸和分配的問題。
3. 電子版本的古籍容易匯集，鉤稽參照後，能產生新的訊息。
4. 電子古籍好儲存，體積小，便於檢索、應用及處理。

所以他認為電子化的古籍是使古籍活出最佳現代風貌、也是唯一的選擇。

圖書館除了應善盡存護古籍的責任外，尚須設法提供使用者有關古籍內文中的「知識」。古籍中知識的獲取有賴先整理出其「線索」，所以古籍的「內容」及「線索」是整理工作的兩個重要對象，而數位圖書館中的古籍整理也以此兩者為鵠的。數位圖書館提供古籍「內容」所用的方式包括影像化及建置全文資料庫，前者重現古籍原貌，以便即時閱覽、傳遞及列印；後者將古籍全文轉為電子本文，以便檢索、儲存及編輯。此外，圖書館也進行網路上古籍資源的組織與檢索，藉以描述並揭示古籍的「線索」。本文以「古籍全文資料庫」為例，藉以瞭解資訊科技對於圖書館收藏中文古籍的影響。

圖書館為什麼要發展「古籍全文資料庫」？這個問題牽涉到兩方面，一方面

[1] 武亞民，〈數字圖書館與古籍整理〉，《圖書館學刊》，1998：2（1998.3），頁 25。

[2] 謝清俊、林晰，〈中央研究院古籍全文資料庫的發展概要〉，1997，頁 2。網址：http://www.sinica.edu.tw/~cdp/paper/1997/19970301_1.htm。

是古籍的影像化有若干限制，包括：1.無法進行檢索、儲存及編輯；2.若原稿的文字甚小或蟲蛀嚴重，則透過影像無法判讀細小的註記文字。而另一方面將古籍的全文轉成電子文本對使用者有其優點，試從中央研究院的「漢籍電子文獻資料庫」可以舉出實例來看古籍電子文本的快速查詢功能及其對學術研究的便利。「裴松之註《三國志》時，徵引了許多目前已經遺佚的典籍，注文中也有許多「臣松之案」或「臣松之以為」之類的案語。學者要利用這些案語，研究當時的史學思想；或利用引注，輯錄散佚的古籍。以往必須先閱讀 86 萬字的《三國志．注》，從中找尋資料，整個過程十分耗神費時。有了「資料庫」後，只要打入「松之」二字，便可在一秒之內，查遍整部《三國志》863,469 字，找到 228 項，259 詞，獲得裴松之意見的初步資料。至於輯佚的工作，透過電腦查詢，更是快捷。例如，若想自《三國志．注》中輯出《魏略》這部書，則只需打入「魏略」一詞，便可以在一秒之內，找出143 項，191 詞。」[3]

關於古籍全文資料庫的建立，目前有以下兩種方式：

1. 將古籍的文字逐字輸入；
2. 利用「光學文字辨識技術 OCR」，由系統自動作影像處理、文件分析及文字識別。

前者如中央研究院長期開發的「漢籍電子文獻資料庫」以及中華電子佛典協會所製作的「漢文電子大藏經」，後者如香港迪志文化出版有限公司於1998 年推出的「文淵閣四庫全書電子版」以及國家圖書館於民國89 年委託中央研究院資訊科學研究所開發的「古籍無定型工整手寫文字辨識、檢索與管理系統」。本文介紹古籍全文資料庫的發展現況，並就建置古籍全文資料庫所牽涉到的問題加以討論。

二、發展現況

以下針對中央研究院的「漢籍電子文獻資料庫」、中華電子佛典協會的「漢文電子大藏經」、香港迪志文化出版有限公司的「文淵閣四庫全書電子版」以及國家圖書館的「古籍無定型工整手寫文字辨識、檢索與管理系統」等四個「全文資料庫」，分別介紹其概要、流程及特色。

[3] 黃寬重、劉增貴，〈中央研究院人文計算的回顧與前瞻〉，《漢學研究通訊》，17:2（1998.5），頁 146。

(一)中央研究院「漢籍電子文獻資料庫」

中央研究院所推動的漢學研究電子文獻包含全文資料庫、研究參考工具資料庫、主題研究多媒體資料庫、書目及關連式資料庫以及利用影像處理技術進行的檔案與古文書光碟影像資料庫等五種。其中以「全文資料庫（現稱漢籍電子文獻資料庫）」的規模最大，應用最廣。

民國 73 年，中央研究院歷史語言研究所與計算中心，合作開發《廿五史》(食貨志)全文資料庫，第二年起逐漸擴大到《廿五史》全部，接者相繼輸入十三經、佛藏、醫書、政書等資料，截至民國88 年 9 月底，已完成的漢籍電子文獻資料庫累計達1 億 2 千萬字，正在進行校對、標誌的資料約有2 億字，並以每年超過 1 千萬字的速度擴增，堪稱為全球最大的中文文獻資料庫。[4]

「漢籍電子文獻資料庫」的製作程序包括：[5]

1. 資料輸入：將資料交由不同人員分別繕打成兩份原始電子文件檔。
2. 初校：利用文書校對程式比對兩份電子文件檔，找出不同處，進行人工修訂。
3. 二校、三校及標誌：以人工作業直接核對原書，修訂電子文件檔，同時添加標誌，標明文獻的篇章節段組織和版面編排，以便據以建立資料庫裡的資料檔和索引檔。
4. 造字管理：古籍中常出現電腦中文系統所無的中文字，需要一一辨識、彙整，於電腦中文系統中造字。目前造字已累積達4,555 個。

「漢籍電子文獻資料庫」的特色包括：[6]

1. 品質最佳：文獻失誤率低於千分之一。
2. 規模最大：開發及完成的合計超過三億字，每年平均以三千萬字建檔，校對完成者約一千萬字。
3. 資料經過一定規模的組織。
4. 軟體設計最佳。

中央研究院於民國 85 年成立「漢籍電子文獻協調委員會」，未來計畫結合相

[4] 黃寬重，〈數位典藏與人文研究：中央研究院文獻資料數位化工作的回顧與展望〉，載於：《中央研究院第三屆國際漢學會議論文》（2000.6.29-7.1），摘要。

[5] 同註 4，頁 147。

[6] 同註 4，頁 146。

關學術研究機構，共同推動「新四庫全書資料庫」計畫，以擴增全文資料庫的內容。中央研究院「漢籍電子文獻資料庫」的網址為http://www.sinica.edu.tw/ftms-bin/ftmsw3，選例如圖一。

圖一：漢籍電子文獻資料庫選例：以「松之」檢索二十五史中「三國志」，找到 228 段，此為第 1 段

(二)中華電子佛典協會「漢文電子大藏經」

中華電子佛典協會（Chinese Buddhist Electronic Texts Association, CBETA）於民國 87 年成立，其宗旨為建立出符合學術標準的數位化中文佛典。所使用的方法除了收集並校對目前各界已輸入的數位化佛典，並利用光學辨識來進行數位化。該協會選擇日本大正13 年至昭和9 年（1924~1934）東京大正一切經刊行會所編輯出版「大正新脩大藏經」來進行數位化。[7] 所以該工作並非將藏經佛典逐字鍵入，而是將紙本的佛經利用光學辨識方法轉換為數位的佛典。

「漢文電子大藏經」數位化的作業程序包括：[8]

[7] 大藏經刊行會編輯，《大正新修大藏經》(東京：大正一切經刊行會，日本大正13 年至昭和 9 年（1924-1934）；臺北市：新文豐出版公司，1983 年翻印)，修訂本一版。

[8] 釋自衍，〈從漢譯大藏經數位化發展現況談紙本資料數位化努力的途徑〉，《中國圖書館學會會報》，63（1999），頁 49-54。

1. 資料輸入：該工作主要採掃瞄書本及使用OCR 技術來產生經文，而不做人工輸入，在進行 OCR 時，利用「去點程式」來去除經文圖檔的「雜點」；另利用「看圖校對程式」來處理OCR 常見且重複規則性的錯誤及誤判。
2. 標誌：即「編輯格式」，除了在經文每一行之前記錄「冊數」、「經號」、「頁碼」、「欄」、「行」等資料，以便日後的查書訂正作業；此外，經文並製作 SGML（現已改用 XML）標記，以便可以跨越各種不同平台使用資料，並產生各種格式檔案。
3. 校對：使用「檔案比對」、「看圖校對」、「人工校對」等多種方法進行校對。
4. 缺字處理：選擇中研院資訊所的作法，對於目前 Big-5 碼所無法輸入的文字，先以「一般組合字」來表示電腦缺字(例如以「口*兄」來表示「咒」)，並建立好「缺字相關資訊」。

「漢文電子大藏經」數位化的特色包括：[9]

1. 提供多種版本，便利需求：版本包括（1）普及版，即一般文字檔；（2）App 版，即「行末句點」格式之文字檔；（3）HTMLhelp 版，具有目錄、索引、全文檢索多功能；（4）HTML 版，可以直接使用網路瀏覽器閱讀。
2. 跨行檢索功能：使用跨行檢索及忽略行首資訊的中文化搜尋程式功能，以避免古籍文獻依紙本格式排列檢索，而碰到換行時，則無法完全檢索到一專門詞彙。
3. 版本比對校對方式，發揮校勘功能及提昇經文品質：在比對不同版本的藏經中，若發現「大正藏」的錯誤，則以紅色顯示經文。
4. 結合電子辭典閱讀藏經：該系統在HTMLhelp 的瀏覽器上，將丁福保「佛學大辭典」電子檔字辭典與經文相連接。

海內外電子佛典的工作進行多年，目前以中華電子佛典協會所做的「大正藏」數位化工作最具代表性，也累積了許多寶貴的實務經驗。「漢文電子大藏經」的網址為 http://ccbs.ntu.edu.tw/cbeta/result/index.htm（「中華電子佛典協會—線上藏經閣」），選例如圖二。

[9] 同註8，頁54-57。

圖二：漢文電子大藏經選例： CBETA 電子佛典光碟版大正新脩大藏經內《大般涅槃經》，完成日期：2001.7.1

(三)香港迪志文化公司「文淵閣四庫全書電子版」

香港迪志文化出版有限公司於1998 年推出「文淵閣四庫全書電子版」，四庫全書」原於清乾隆年間纂修，共收圖書3,469 多種，約 7 億字，3 萬 6 千餘冊，該公司以掃瞄方式將原書圖像數位化，再用光學文字辨識技術，將圖像轉為近7億的電腦編碼文字，並研究出多種檢索功能。

「文淵閣四庫全書電子版」的製作程序包括：[10]

1. 輸入：採取掃瞄方式將原書圖像數位化。
2. 辨識：將清華大學所開發的「非特定人手寫識別OCR 技術」發展為「多特定人準規範手寫 OCR 引擎」，並利用此軟體進行逐頁的「單字切分」，也就是將頁面中的大字／小字、交錯／粘連、局部圖等等規範和非規範的頁面，OCR 識別率平均可達92%。
3. 校對：經過 OCR 處理仍然辨識不出來的文字，再採取「人工輔助糾錯」方

[10] 張軸材，「四庫全書電子版工程與中文信息技術」。上網日期：2001.1.14。網址：http://www.sikuquanshu.com/。

式，專門開發出「校得快」軟體，主要的功能包括：⑴將原文字跡（圖）與識別結果（編碼漢字，文）一一對應，方便進行圖文對照、順序瀏覽校對；（2）按漢字聚類（例如所有的「之」字），瀏覽校對；（3）隱蔽其他內容、突出重點校對；（4）將概率較高的第二候選字與滑鼠一起移動，來置換錯誤的「第一候選字」；（5）點選其餘的九個候選字；（6）利用組合序列方式（例如：@,=,~ 等符號）來描述缺字等等。

4. 檢索：檢索機制包括：（1）傳統的部、類、屬、書目的（樹狀）分類結構；（2）建立作者與作者、作者與書目、書目與書目之間的超連結；（3）編製出卷名、書名、形式標題、邏輯標題、詩詞名、表名、圖名及段落的首句組合等近 200 萬條綱目；（4）全文檢索；（5）建立「漢字關聯」（包括：簡體／繁體關係、正體關係、正字／訛字（形近異義字）、通假／被通假、古今字、新舊字形、中日差異等）以增加古漢語檢索。

「文淵閣四庫全書電子版」的特色包括：

1. 圖文並列：「電子版」分為「原文及標題檢索版」及「原文及全文檢索版」，前者包含原書頁的圖像及標題檢索，後者另增加全文檢索功能。
2. 使用 Unicode 字集：於 Unicode 的基礎上，建立近3 萬中文字集。
3. 跨平台：適用於臺灣（Big5）及大陸（GBK）中文視窗，以及其他如英文及日文等語文平台。

「文淵閣四庫全書電子版」的簡介網址為http://www.sikuquanshu.com/，全文選例如圖三。

(四)國家圖書館「古籍無定型工整手寫文字辨識、檢索與管理系統」

中央研究院資訊研究所曾開發出「影像文件辨識、檢索與管理系統」，兼可製作影像及全文資料庫，使用的技術包括：影像處理、文件分析、文字識別、容錯性檢索等，其特點為：1. 原文件的影像能以完全不失真的方式重現；2. 文件的全文皆可檢索；3. 將文件掃瞄輸入電腦後，系統自動從事各項分析、識別、檢索與建構資料庫的工作，使用簡易。[11] 然而包括此系統只能處理以鉛字、活字版或電腦印刷的文件，而對於手寫或木刻的古籍尚無法辨有效辨識。所以國家圖書館於民國 89 年委託中央研究院開發「古籍無定型工整手寫文字辨識、檢索與管

[11] 張復，「影像文件辨識、檢索與管理系統」。網址：http://140.109.19.199/Osweb/Osweb.asp

理系統」，針對各種不同「版本」及「字體」的古籍；如：刻本、抄本、明體、楷體等）為對象進行實驗，希望將來能改良出一套最適合製作古籍影像及全文資料庫的理想方案。

圖三：文淵閣四庫全書電子版選例：史部，別史類，東都事略，卷一。

該系統的作業程序包括：[12]

1. 掃瞄、影像處理：針對典藏書籍紙質及文字特性，設計合適的掃瞄及二色化或三色化方法。
2. 文件分析：根據典藏書籍文字分佈的特性，設計合適的文件分析及文字切割法。
3. 文字識別：文字識別的步驟包括大部比對、文字大分類及細部比對，並歸納文字特徵，調整分析方法。

該系統的特色包括：

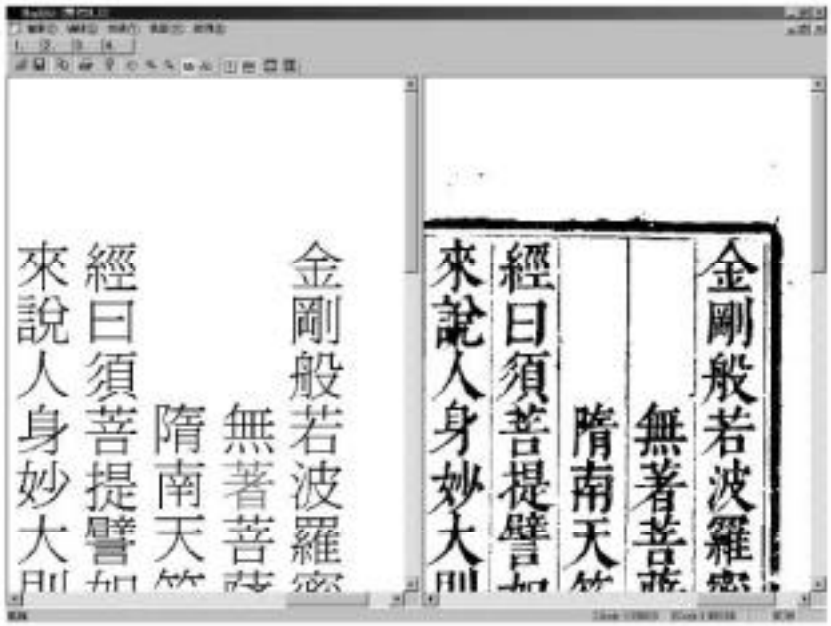
1. 開發出新的文字辨識技術：本計畫利用已開發的「影像文件辨識、檢索與管理系統」內的「文字辨識模組」，過濾出「無發辨識的文字」，再進一步分

[12] 張復（研究計畫主持人），「古籍無定型工整手寫文字辨識、檢索與管理系統計畫執行構想報告書」，影印本（中央研究院資訊研究所文件分析與辨識實驗室, 2000），頁 63-64。

析、比對文字的特徵，歸納初期規律性，轉而提昇上述文字辨識模組的辨識能力。該計畫在辨別兩字是否相同所開發的技術包括：掃瞄線文字細線化、線比對、接點比對、部位比對、投影峰比對。[13]

2. 容錯檢索：該系統所利用的文字資訊不僅是文字檔裡的文字，而且考慮到各種形狀相似的文字，所以其檢索功能具有「容錯性」，且其正確率高於文字的辨識率。[14]

該系統的將重點是在「文字辨識率」的提升，所以選擇文字清晰、工整且沒有眉批及校點的古籍作為初步實驗的樣本，目前所選擇兩部館藏善本分別是「嘉興楞嚴寺方冊藏經」中的《摩訶般若波羅蜜經》及《金剛般若波羅蜜經論》。[15]「古籍無定型工整手寫文字辨識、檢索與管理系統」的網址為 http://192.83.186.77/legacyweb/，全文選例如圖四。



圖四：古籍文字辨識、檢索與管理系統選例：系統辨識結果的對照，左半部是辨識後的文字，右半部是原文重現的部份。

[13] 同註 12，頁 57-62。

[14] 同註 12，頁 55。

[15] 「古籍文字辨識檢索與管理系統」。上網日期：2001.1.14。網址：http://192.83.186.77/legacy-web/。

三、相關問題與檢討

綜觀上述所進行的幾項與古籍有關的「全文資料庫」，可以瞭解到建立古籍全文資料庫的方式包括「逐字輸入」以及「光學文字辨識」等兩種，「逐字輸入」需要耗費無數的人力、時間以及財力，而「光學辨識」則牽涉到辨識的對象（例如：鉛字、電腦印刷文字、刻板文字或手寫體等）以及辨識率的高低；此外，如：校對的問題、缺字及造字、標誌技術及標準格式、檢索率都是相關的問題，逐項檢討如下。（以下「漢籍電子文獻資料庫」，簡稱「漢籍資料庫」；「文淵閣四庫全書電子版」，簡稱「四庫電子版」；「漢文電子大藏經」，簡稱「大正藏」；「古籍無定型工整手寫文字辨識、檢索與管理系統」（因以館藏「嘉興楞嚴寺方冊藏經」為實驗樣本，簡稱「嘉興藏」）

(一)輸入方式

上述各資料庫的輸入方式列表如下：

表一：各全文資料庫輸入方式一覽表

方式／資料庫	書頁字體	輸入方式	辨識率
漢籍資料庫	印刷體	逐字輸入	
四庫電子版	手寫楷體	光學辨識	92%
大正藏	印刷體	光學辨識	95%
嘉興藏	木刻本明體	光學辨識	90%

古籍全文資料庫的輸入方式包括「逐字輸入」（如「漢籍資料庫」）以及「光學文字辨識」（如「四庫電子版」）此兩種主要方式。比較兩種輸入方式，「逐字輸入」需要耗費無數的人力、時間以及財力，即使「漢籍資料庫」已經輸入了上億字的古籍，但在圖書館內仍有數以百萬計的古籍有待製作，所以就成本效益的角度來看，「光學文字辨識」應較「逐字輸入」為經濟可行之法。然而光學辨識也有其侷限性。就目前的光學辨識軟體的辨識率而言，以「四庫電子版」為例，其辨識率為 92%，辨識不出來的文本仍要靠在大陸引入數百人力來校對及人工鍵入，這在勞力薪資高的臺灣實難辦到，所以目前「光學辨識軟體」的「辨識技術」是一個技術瓶頸。「辨識技術」的高低（亦即辨識率的百分比）牽涉到「辨識對

象」（鉛字、電腦印刷文字、刻板文字或手寫體等）以及「辨識方法」等問題。目前對於印刷書籍的辨識能力雖然有效，然而由於雕版及手抄的古籍文本字型不規範，而且書寫風格有相當大的差異，所以「雕版以及手抄方式的古籍文本文字辨識技術」值得開發研究。

辨識需靠系統不斷地分析、辨識並且學習，「嘉興藏」是針對木刻本的明體字來進行光學辨識，目前已超過90％的辨識率，將來會繼續不斷地提升。

(二)校對問題

上述各資料庫的校對方式列如表二：

表二：各全文資料庫校對方式一覽表

方式 / 資料庫	輸入方式	校對方式
漢籍資料庫	人工繕打	利用程式作檔案比對
四庫電子版	光學辨識	利用程式人工修改
大正藏	人工繕打	檔案比對、看圖校對
	光學辨識	線上校對、字串取代
嘉興藏	光學辨識	看圖校對

「漢籍資料庫」採用檔案比對法，其校對分為初校及複校，初校係利用程式作檔案比對，並進行人工修訂，而二、三校則以人工核對原書校對，如此，所需人力相當可觀。「四庫電子版」的光學辨識率及人工輸入校對方法如表三：[16]

表三：「四庫電子版」的光學辨識及人工輸入校對方法比率表

方法		百分比
OCR	正確識別/無需輸入	89-91%
人工輸入	光標跟隨/二選字輸入	1-2%
	點擊其餘候選字輸入	5-6%
	鍵盤輸入/四庫流行碼	1-2%
	鍵盤輸入外字組合串	<0.1%

[16] 同註10，頁13。

以上顯示「四庫電子版」仍須近 10% 的人工輸入，所需的人力也相當可觀。「大正藏」依據輸入方式的不同而有不同的檔案校對方式，其中針對曾經收集過各方已建檔經文的資料，係採用「檔案比對」及「看圖校對」兩法，每份經文至少有三個輸入版本，再利用程式，或一次比對三個，或分兩次兩兩比對，最後針對比對的結果，翻原書以人工訂正。而比「檔案比對」更快的是利用「看圖校對」程式，一邊看大正藏掃瞄圖檔，一邊做文字訂正。[17] 此外，針對光學辨識所產生的檔案，則採取「線上校對」及「字串取代」等方法，線上校對係利用程式進行「看圖校對」，同時可顯示字串圖形與校對文字，另「字串取代」則是先建立的「OCR 常犯錯誤字串取代表」，再以 OPEN98 所提供的程式自行取代。[18]「嘉興藏」係利用影像、文字對應編輯器來對錯誤的文字進行修正，也屬於「看圖校對」。

即使是使用光學辨識，校對所需的人力仍相當可觀，所以各資料庫都不斷開發新的校對方式，其中「看圖校對」是最普遍的作法，而「字串取代」則直接利用 Microsoft Word 內的「取代」功能，不必另行開發程式，經濟而簡便，未來在校對上應發展人工智慧，加強校對工具的自動學習功能。

(三)缺字及造字

上述各資料庫對「缺字及造字」的處理方式列如表四：

表四：各全文資料庫對「缺字及造字」處理方式一覽表

方式／資料庫	缺字表現及造字方式	中文字庫
漢籍資料庫	漢字組合法、組字規則	中文字形資料庫
大正藏	漢字組合法、組字規則，同時顯示該字的圖形檔	OPEN98 漢字庫
四庫電子版	國際碼（Unicode）	

「漢籍資料庫」係使用漢字組合法及組字規則來表示電子古籍中的缺字處理，並以網路上所建立的「中文字形資料庫」提供給使用者使用。[19]「大正藏」

[17] 吳寶原，〈從實務經驗談電子佛典初步工程之演進〉，《佛教圖書館館訊》，14（1998.6），頁 29-30。

[18] 同註 17，頁 27-28。

[19] 莊德明、謝清俊、林晰，「中央研究院古籍全文資料庫解決缺字問題的方法」。網址：http://www.sinica.edu.tw/~cdp/paper/1998/19990511_1.htm。

採取以下步驟來處理「缺字造字」，包括：1. 臺灣網路界通用的組字法做校對及普及版的發行；2. 轉成 SGML 碼以處理缺字問題；3. 採用目前國際上造字最多的日本「今昔文字竟」的字型檔以呈現所缺的字。[20]「四庫電子版」則採用國際碼（Unicode），為了兼顧長遠字量增長以及跨平台的需求，「四庫電子版」係從 GBK（CJK 字彙）開始，再過渡到 Unicode／CJK＋－為基礎的平台上，如此大約多了 1 萬餘個字符。以「四庫全書」經部前19 冊使用 CJK＋前後（四校前後）缺字量的變化為例，說明如表五：

表五：「四庫全書」經部使用CJK＋前後缺字量變化表

代碼 Code	語料（字次）	外字				
		出現率（字次）	出現率（萬分之）	模糊不清	不肯定的異體代換	可用組合串表示
GBK	8,769,710	4,887	5.57	1,830	2,158	899
CJK+	8,771,787	1,520	1.73	1,345	39	75

除此外，另「模湖不清」的字可以在光標所指「方框（即□圖形）」之處顯示原文字跡。[21] 比較使用「CJK+」前後，顯示缺字出現的頻率大幅降低（約 3 倍）。

全文資料庫不論是「檢索」或「顯示」，皆需要足夠的漢字，類似「中研院」所使用的專用字庫只是一個暫時的權宜作法，而且僅能通用於臺灣地區，不利於資料的交換。若為長遠計，應有一套既能適應各種平台、又可作為國際通用的字集，目前國際所頒佈的ISO10646／Unicode，應為可以考慮的方案。

(四)標誌及標準格式

古籍若要維持原始文件的「版面訊息（如頁碼、行次）」以及「文件結構（如：標題、篇、章、節、小節、段落、註解等）」，則要考慮所謂的標誌（Markup）工作，而標誌所採用的方式需符合一定的標準格式，例如：SGML、XML 等。上述全文資料庫中只有「漢籍資料庫」及「大正藏」採用「標誌」。

[20] 杜正民，〈佛教藏經的文字問題與解決方案〉，載於：中央研究院漢籍電子文獻協調委員會，《電子古籍中的文字問題研討會》（會議資料）（臺北市：中央研究院，1999.6.14-16），頁 43。

[21] 同註 10，頁 5-8。

「漢籍資料庫」係以人工或程式自動進行描述性標誌（Descriptive Markup），再轉換為WWW的HTML。[22]「大正藏」除了標誌「冊數」、「經號」、「頁碼」、「欄」、「行」等資料外，並製作經文的SGML 標誌，在「CBETA」手冊中詳定SGML格式目前所用的標記符號，以便處理不分卷、卷終、註解的跨行、一個註解中包含另一個註解等情形。[23]

古籍全文資料庫的電子檔建構在開放的標準格式上有兩個作用，第一個是「有助於資料的互通與合作」，劉文卿認為「標準格式」對制訂漢文電子佛典的重要性包括：1. 跨越各種平台，共享資源；2. 方便系統開發，適用所有的數位資料；3. 方便網路傳輸；4. 建立新的索引格式，作為查詢索引的標準界面；5. 具有長期而穩定的環境。[24] 第二個是「可以將標誌過的古籍轉成『超文件Hypertext全文資料庫』，以利內容的檢索」。古籍全文資料庫雖然解決了篇章及字詞的檢索問題，但對於古籍各版本之間的對應（如：存世《史記》各種版本之間的關係）、注疏和正文間之間的參照（如：《尚書》漢孔安國、唐陸德明等各家傳注與正文的關係）仍無法處理，針對這一點，中央研究院已開發了「中文文獻處理系統」，可以做「古籍超文件處理」，並且已有若干先導實驗解決古籍各類資訊之多版本連結導行之問題。[25]

---

[22] 同註2，頁6。

[23] 杜正民、維習安，〈佛學網路資料庫的建構與開發〉，載於：中央研究院漢籍電子文獻協調委員會、中央研究院計算中心，《漢學研究網路環境的開發座談會》（會議資料）(臺北市：中央研究院，1999.3.29），附「電子佛典手冊」頁27-30。

[24] 劉文卿，〈制訂漢文電子佛典標準格式的重要性〉，《佛教圖書館館訊》，15（1998.9），頁42。

[25] 相關文獻包括：

(1) 周亞明，〈使用SGML 與物件導向資料庫轉換古書為Hypertext〉，中國中文信息學會、國家古籍整理出版規劃小組辦公室、北京語言學會，《海峽兩岸中國古籍整理研究現代化技術研討會論文集》(北京市：中國中文信息學會，國家古籍整理出版規劃小組辦公室，1993），頁 164-178。係將古書及其注解轉換為Hypertext 系統，並以「尚書」為例，系統可顯示各家傳注（如：漢孔安國、唐陸德明等）對正文的說明。

(2) 謝清俊、莊德明，〈古籍校讀工具「中文文獻處理系統」的設計〉，中國中文信息學會、國家古籍整理出版規劃小組辦公室、北京語言學會，《海峽兩岸中國古籍整理研究現代化技術研討會論文集》（北京市：中國中文信息學會，國家古籍整理出版規劃小組辦公室，1993），頁 1-11。該系統為超文件系統，包含文件本身、文件間相連關係及檢索和瀏覽功能，可處理古籍各版本間之對應、注疏和相關資料與原文間之參照，以及原文內容之標誌等。

(3) 陳昭珍，「古籍超文件全文資料庫模式之探討」（臺北市：國立臺灣大學圖書館學研究所

(五)檢索問題

「漢籍資料庫」係採用在 UNIX 系統下所開發的「中文全文檢索系統 CTP/FTMS」（現改稱「瀚典全文檢索系統」，1997 年 11 月，1.3 版），其文件擷取方式（Access Method）所用的技術是字串比對（String Matching）。在 SUN 工作站中，每秒可處理 70 萬字以上。[26]「四庫電子版」開發了一個以 Unicode/CJK+ 為基礎的全文檢索引擎，可以檢索到篇章／字位。針對1 億 400 萬條語料來實施檢索，其效果如表六所示：

表六：「四庫電子版」語料檢索效果表

硬 體 設 備	檢索語料	所需時間
PII/233MHz 128M RAM,Windows NT 4.0 Server	李白	0.470 秒
	商鞅變法	1.271 秒
PII/166MHz 32M RAM,Windows NT 4.0 Wordstation	李白	1.482 秒
	商鞅變法	1.487 秒

此外，「四庫電子版」還提供「漢字關聯」（包括：簡體／繁體關係、正體關係、正字／訛字（形近異義字）、通假／被通假、古今字、新舊字形、中日差異等），以增加古漢語檢索。[27]「大正藏」所研製的全文檢索的引擎之功能並不強，尤其全文資料尚未作好權威控制（提供使用同義詞作為查詢之關鍵詞），使檢索時呈現高回收率，低精確度的情形，浪費使用者的時間。[28]「嘉興藏」的全文檢索子系統具有「容錯性」，也就在做全文檢索時，只要輸入的字是被搜尋字的相似字，例如：菩薩、善薩，就會被系統偵測出來。

古籍全文資料庫的「檢索問題」不僅與檢索系統的功能以及所使用的硬體有關，也牽涉到權威資料的製作。例如「四庫全書電子版」提供出若干不同對映關係的漢字關聯字，以增加檢索的效果。此外，為了提高古籍全文檢索的「查全率」

博士論文，1994.12）。分析古籍之超文件性質，並實際以文心雕龍為例，利用前述「中文文獻處理系統」，解決古籍各類資訊之多版本連結導行之問題。

[26] 同註 2，頁 4。

[27] 同註 10，頁 16。

[28] 同註 8，頁 57。

和「查準率」，也有建議應提供古籍的規範詞表，如：人名、室名別號對照詞表、地名沿革詞表、職官沿革詞表、同義詞對照表等，以利使用者輸入任一同義詞，系統即可將含有全部同義詞的句、段同時提供。[29]

四、結　語

藉諸資訊科技來整理中國古籍，是對古籍組織與利用的的一個推廣和延伸，使得古籍不僅是被動的等待讀者取用，也能主動的將古籍呈現在使用者之前，以充分發揮圖書館服務的功能。

電腦對於圖書館整理古籍究竟有什麼影響？首先在「內容」方面，將古籍的內容透過影像處理，可以實現數位化典藏的功用，更可在網路上傳輸，以方便流通；再就全文檢索而言，古籍的原文轉化為電子文本後，除了方便儲存、傳輸以及查詢外，還有「勾稽參照」的功能，也就是古籍的電子文本若加以標誌（Markup）化後將可轉成「超文件 Hypertext 全文資料庫」，對於古籍內容的開發和其中所包含知識的貫串整理甚有幫助；此外，目前網路上有關圖書館所收藏中國古籍的資源有限，而不同圖書館所使用的書目及資料庫的查詢系統又不盡一致，讀者需要分別認識並且逐一查詢，既耗時又費力。長久以來圖書館就希望能讓讀者跨越不同的平台進行檢索，在這個理想的期待下，「網路資源的組織與檢索」應運而生，透過這個作法可以將網路上的資源有效的加以組織與檢索，以實現資源共享的目標。

國外電子圖書館所典藏的電子文件中，不乏古籍文獻在內，以美國維吉尼亞大學圖書館的電子文件中心為例，美國維吉尼亞大學圖書館於1992 年成立電子文件中心（Electronic Text Center），旨在建立一個以SGML 編碼的全文及影像館藏，並可透過網路進行查詢。該中心已建立數以千計用SMGL 及 XML 處理過的電子全文及影像檔案，並將這些檔案與圖書館所提供的軟硬體服務結合在一起。該中心所收錄線上以及離線的人文學科電子文件約有5 萬種，內容涵蓋史學、文學、哲學、宗教等各學科；另有 35 萬幅影像，包含插圖、封面、手稿、善本古籍書影等。所處理的電子文件有 12 種語文，包括：英文、法文、德文、拉丁語

[29] 董燊、邢素麗，〈中國古籍電子出版物製作技術淺探〉，《現代圖書館情報技術》，1999：3（1999.5），頁51。

文、中文、日文、俄文、希臘文、希伯來文、西班牙文、藏文、冰島文、義大利文、葡萄牙文等等。大部份的電子文件都可在網路上以同一介面查詢，每一筆電子文件都含有被建檔文件及其紙本來源的書目標頭（Bibliographic Header）[30]，的確是一個將「全文資料庫」、「影像資料庫」以及「網路資源組織與檢索」三者結合在一起的好例子。

圖書館所收藏的古籍在資訊科技的衝擊下，顯而易見有幾個目標，包括：如何利用資訊技術將著錄的內容作更精確的存取？如何提供更深化的連結？如何表達出更親合的呈現？這些都是圖書館的古籍面臨資訊科技衝擊的問題。吳政上認為近年來網際網路盛行，網路資源的詮釋、管理、呈現，成為熱門的話題，他預測此一趨勢將會改變古籍編目（善本圖籍、金石拓片影像資料）的內涵，同時也提出一個未來努力的方向，即是設法將「詮釋資料」、「聲音影像資料」以及「全文資料」作深入且有效的連結。[31] 此一目標也已經在中央研究院歷史語言研究所傅斯年圖書館所進行的「國家典藏數位化計畫」中開始在實施中。[32]

國家圖書館於民國 90 年開始進行國科會所推動的的「國家典藏數位化計畫」，並且研擬「古籍文獻典藏數位化計畫」，預定在三至五年內將館藏重要善本古籍約 6 千部加以數位化，主要進行的工作要項包括：典藏品的數位化以及檢索系統的開發，後者相當於 Metadata 書目資料庫的建置，此書目資料庫係依據國家圖書館 Metadata 研究小組編撰之中文詮釋（Metadata）格式彙編」初版一書中「古籍善本詮釋資料及著錄範例」為基礎，加以修訂，來訂定Metadata 所需的欄位。在規劃欄位時，除了將古籍的篇目加入外，並且特別將有關古籍的「全文資料」包含在內，所謂古籍的全文資料，不僅是古籍的「內容全文」，並且還將此一古籍特殊版本的包含在 Metadata 的欄位在內。由於國家圖書館過去曾將館藏古籍的「序跋」以及「題跋」分別加以標點排印，所以很方便轉為「全文資料庫」，此外，也由於前述國家圖書館所開發的「古籍無定型工整手寫文字辨識、

[30] Electronic Text Center, University of Virginia. Retrieved October 14, 2000, form http://etext.lib.virginia.edu/.

[31] 吳政上，「古籍著錄過程中若干問題及處理方式」，《古籍聯合目錄資料庫合作建置研討會》（論文）（臺北市：國家圖書館，2001.4.19-20），頁 48（46-48）。

[32] 傅斯年圖書館善本書目全文檢索資料庫。上網日期：2001.10.10。網址：from http://ultra.ihp.sinica.edu.tw/~fsnlib/frames7.htm。

檢索與管理系統」，已具備了將特定字體的古籍轉為「全文資料庫」的功能。未來也可實現將「全文資料庫」、「影像資料庫」以及「網路資源組織與檢索」三者結合在一起的目標。

圖書館將其所典藏的珍貴文獻，建置為全文資料庫，甚至包括影音資料庫，而所製作的電子文件都加以適當的描述與組織，這是未來圖書館開發古籍資源，提供知識利用的良法和的趨勢。傳統館藏藉著進步的資訊技術，將可實現圖書館提供知識、普及利用以及弘揚文化的職志。

BIBLID 1026-5279 (2001) 90:2 pp. 197-216

A Comparative Study on the Establishment of the Chinese Ancient Books Full-text Database

Li-jen Ku

Abstract

This paper takes several full-text databases of Chinese ancient books as an example to discuss the impact of information technology on the Chinese ancient books collected in the library. There are two methods to establish a full-text database. One of them is transcribe and proofreading, the other one is using optical character recognition technology. However, both of them face different kinds of problems when processing. The author first introduces several examples of full-text databases and then analyzes the problems encountered in this process. In conclusion, possible future developments are discussed.

Keywords（關鍵詞）：Digital Library；Organization of Ancient Books；Image Processing；MARC；Metadata
古籍；古籍整理；全文資料庫；光學文字辨識；標誌

Li-jen Ku： Editor, Department of Special Collection, National Central Library；E-mail: klj@msg.ncl.edu.tw