古籍數字化：現狀、問題與趨勢

——從一個使用者的角度看

吳宣德

數字圖書館（電子圖書館、虛擬圖書館）和大型電子文獻資料庫的開發和建設，近年來業已成爲學術界和相關技術部門的一個熱門話題。在這個話題中，古籍的數字化也得到了高度重視。並且，伴隨著一些重要產品的發行，數字化的古籍在學術研究中的利用率也在不斷提高。而古籍數字化的價值，也正是在這樣的使用中一點一點體現出來。

然而，僅就筆者個人的專業應用來看，古籍數字化在技術處理和用戶的使用需求之間仍然有一些距離。本文即擬對此談一點粗淺的看法，供有興趣者參考。因筆者對中國大陸以外地區的相關產品了解很少，取例有所偏重，尚祈讀者諒之。

一、現狀一瞥

目前古籍的數字化可謂“繁花似錦”，除專門的製作公司外，一些單位和個人也投放了不少注意力，並且也確實在各自的工作領域取得了一些重要成果。茲據筆者涉獵所及，將相關情況略述如下：

（一）豐富多樣的文件格式

從文件格式上來看，目前的數字化古籍除常見的 txt、doc、html 格式外，還有 exe、pdf、wdl、pdg、ebk、edb 等。

1．exe 格式。所見有秦昌榮（秦堤居士）的“家庭百寶箱”系列（含諸子百家、二十五史、資治通鑑三種）（中華文史軟件網 http://www.jtbbx.com/）。另外，北京天安億友公司（http://www.eyousoft.com/html/index.htm）出品的《二十五史》、《資治通鑑》也包含了這種格式，詳見 2688 阿里巴巴網站（http://www.2688.com/product/zmkmdefault.asp）的“芝麻開門”圖書系列。

2．pdf 格式。所見有紅旗出版社（http://5051.peoplespace.net/）的"家庭藏書集錦"所包含的部分古籍。博庫（http://www.bookoo.com.cn/default.asp）圖書也採用了這種格式。需要 pdf 格式文件專用瀏覽器，如 Adobe 公司的 Adobe Reader、北大方正的 Apabi Reader。

3．wdl 格式。北京華康信息技術有限公司（http://www.dynalab.com.cn）開發的電子讀物

文件格式。所見有博庫圖書採用了這種格式。它保留了原來的版面設計，可以在線閱讀，也可以將下載到本地機閱讀。需要該公司的專用閱讀器 DynaDoc Free Reader。

4．ebk 格式。所見有深圳市百博電子商務網絡技術有限公司的“百博書城”（http://www.bbook.net/gb/default.asp）圖書使用。書籍可在線閱讀，也可下載到本地。需要百博專用的圖書閱讀器。

5．edb 格式。北京天安億友公司開發的圖書格式（Eyousoft Digital Book）。

6．oeb 格式（OpeneBook）。所見有遼寧省出版集團開發的掌上書房（http://www.cnbook.com.cn/bottom/aboutus.htm）使用，有專用閱讀器，也可使用北大方正的 Apabi Reader 閱讀器閱讀。

7．sep、ifr 格式。由書生之家（http://www.21dmedia.net.cn/zhtsw.asp）開發的電子圖書格式，需要其專用閱讀器來閱讀。

8．xeb 格式。以 oeb（Opene Book）爲基礎，以 xml(eXtensible Markup Language,可擴展標記語言)技術爲核心的中文電子圖書格式。北大方正的 Apabi Reader 等瀏覽器支持閱讀。

9．pdg 格式。超星公司（http://www.ssreader.com/）開發的圖書格式。需通過超星專用的圖書瀏覽器閱讀。

10．nlc 格式。中國數字圖書館(http://www.d-library.com.cn/index.php)
採用的圖書格式，需要其專用瀏覽器 NLC Reader 閱讀。

11．基於 unicode（通用多於 8 位編碼字符集）、倉頡碼、Big5 碼或其他字符集，以 html、xml、sgml 等爲核心的全方位電子文獻全文檢索格式。這是目前勢頭最盛、也顯然是更有發展前途、更具實用價值的一種古籍數字化形式。臺灣中央研究院的《漢籍電子文獻資料庫》（http://www.sinica.edu.tw/ftms-bin/ftmsw3）、香港迪志文化出版公司與上海人民出版社等合作開發的《四庫全書》標題檢索版和全文檢索版（http://www.sikuquanshu.com/ 、http://www.skqs.com ）、北京書同文數字化技術有限公司（http://www.unihan.com.cn/html/index.htm）開發的《四部叢刊》等全文檢索版均屬於這種形式。漢文化聯盟開發的《漢文化資料庫》和目前北京大學正在開發的《中國古籍基本庫》也應屬於這種形式。

此外，筆者還見到有位圖形式和多媒體格式的數字化古籍。前者如濟南開發區匯文科技開發中心研製、武漢大學出版社出版發行的《四庫全書》原文電子版。後者如方圓電子出版社出版的《中國古典文學精品書庫》（見 2688 阿里巴巴網站 http://www.2688.com/product/zmkmdefault.asp 的"芝麻開門"圖書系列），包含了《紅樓夢》、《三國演義》、《西遊記》、《初刻拍案驚奇》、《二刻拍案驚奇》、《儒林外史》、《醒世名言》、《喻世通言》及《警世恆言》10 部古典名著全文，還包含由北京廣播電臺的播音員配音的唐詩三百首

和中小學語文教材中的古詩文，帶聯機字典及語音。另外，人民郵電出版社出版（http://www.ptpress.com.cn）的《二十五史多媒體全文檢索閱讀系統》也利用了多媒體形式，該產品除收錄百衲本二十四史和關外二次本《清史稿》的全部內容外，還配置了簡繁字異體字對照表、古代年表等，附錄了《歷代輿地圖》近 2000 幅大比例歷史地圖和張元濟校輯百衲本二十四史時所撰《校史隨筆》以及《左傳》、《戰國策》等多部史學名著、清至當代學者的部分輯佚與校勘成果。

顯而易見，如此眾多的圖書格式，顯示出中國古籍數字化的途徑非常多樣。而且，也正是通過這些途徑，數字化了的古籍才快速地進入了人們的視野，並爲廣大愛好者和專業學術研究人員所注意。

（二）便捷靈活的傳播途徑

從傳播方式上看，除通過磁盤或光盤等介質進行傳播外，最值得一提的是網絡方式和正在引起人們關注的專用電子書閱讀器。

1．網絡傳播

數字化的古籍大範圍進入研究者的視域，應該歸功於網絡。網上書庫、網絡全文檢索、網上圖書館和相對專業的個人網站相得益彰，爲專業研究者提供了古籍電子文獻的不同的使用路徑。

網上書庫多數屬於個人網站。從內容不限於古籍的黃金書屋（現地址 http://goldnets.myrice.com/）、新語絲網上書庫（http://www.xys.org/library.html）、陳清書齋（http://www.chenqin.com/）、亦凡公益圖書館（http://sousuo.shuku.net/）、中華古籍（http://pastbook.myrice.com/）等等到相對專門的國學網站（http://www.guoxue.com/），一大批中國文化的愛好者將一些常用古籍搬上了網絡，對滿足普通讀者的閱讀和使用需求發揮了極其重要的作用。此外，一些專業研究者、學人集體和研究機構也利用網頁提供了一些專門的書籍。比如簡帛研究（http://www.bamboosilk.org/index.htm）上提供的部分簡帛資料、孔子 2000（http://www.confucius2000.com/）上所提供的部分儒學原著、《象牙塔》（http://www.ssdph.com.cn/sailing/book/index5.htm）提供的幾種史籍資料、中華佛典寶庫（http://ccbs.ntu.edu.tw/DBLM/cindex.htm）提供的《大藏經》的免費閱讀和下載，等等。這些資料因多數爲研究者個人積累或專門機構特別提供，所以常常起到補現有一些大型數字化圖書之不足的作用。

與網上書庫、個人網站的單純網頁瀏覽不同，網絡全文檢索提供了古籍資源利用的更爲便捷的手段。除近年來陸續開發的《四庫全書》、《四部叢刊》等大型數字化產品均在單機板之外開發有網絡版外，網上目前已有的讓讀者免費或有條件檢索的相關資源大體集中在臺灣。這方

面，臺灣中央研究院的《漢籍電子文獻》系列(http://www.sinica.edu.tw/ftms-bin/ftmsw3)、臺灣故宮寒泉檢索系統（http://libnt.npm.gov.tw/s25/index.htm）、臺灣元智大學工學院的“網絡展書讀”（http://cls.admin.yzu.edu.tw）、臺灣中華電子佛典協會“線上藏經閣”（http://www.cbeta.org/result/index.htm）《大正新修大藏經》全文檢索等可爲代表。相反，在中國大陸，目前可提供網上全文檢索的古籍資料寥寥無幾，所見僅有北京大學中文系的《全唐詩線上全文檢索系統》（注冊後使用，http://chinese.pku.edu.cn/cgi-bin/tanglibrary.exe）。此外，北京中醫藥大學開發的中藥方劑數據庫（收錄公元 960 年至今的 24 種古籍中的全部方劑）也可以通過網絡進行有償檢索（http://wall.cintcm.ac.cn/webdkrh1/）。

在中國大陸，目前利用率最高的當屬一些網上圖書館。這方面，超星數字圖書館和中國國家圖書館特別值得提起。尤其是超星數字圖書館，是目前中國最大的數字化圖書館。它在 2000 年 1 月由北京時代超星公司與廣東中山圖書館合作開通，現已成爲一個由全國各大圖書館支持的龐大數字圖書展示推廣平臺，並已將其數字化方案成功應用於中央檔案館、中山圖書館、深圳圖書館、解放軍醫學圖書館、溫州圖書館、美國加州大學聖地亞哥分校圖書館等國內外 500 多家單位。其“館藏”的古籍，集中於“專題圖書館”中的“北大圖書館古籍”，以及“古代文獻圖書館”、“國家檔案文獻庫”、“地方志圖書館”等圖書館。而其數量之多，已經可以在很大程度上滿足專業研究者的閱讀需求。此外，超星公司目前還在開發自己的 e_Book。

2．專用電子書閱讀器

專用電子書閱讀器（Electronic Reader，簡稱 eReader）是一種手持離線閱讀電子書的專用設備，它的開發也是近年一個值得關注的事情。目前已見到一些產品面世。如朱邦復先生主導推出的“文昌電子書”Ⅰ號和Ⅱ號，臺灣廣杰科技股份有限公司開發的“EB-2100”，遼寧出版集團推出的“掌上書房”，天津南開津科公司(http://www.jinke.com.cn/ebook/ebook.asp)開發的翰林Ⅰ、Ⅱ、Ⅲ號，北京青創投資管理有限公司(http://www.qingchuang.com.cn/）出品的“金博覽掌上讀”（電子書下載地址 http://www.qingchuang.com.cn/jbl818/docc/libery.htm），等等。

專用電子書閱讀器雖然並非專爲古籍數字化開發，但因其體積小、容量大以及攜帶方便、無特殊閱讀條件的限制，而相應的電子書籍中又包括不少古籍，其對一些古代經典著作的傳播無疑會發生積極作用。

儘管在這些傳播方式中，有些並不需要太高的技術（如網上書庫、個人網站的建設），但它們在推進中國古籍數字化上的作用還是不應被忽視。就廣大的中國文化愛好者甚至是專業研究者而言，其對古籍數字化之價值的認識，並非靠專門製作的大型數字化系列圖書，而恰恰來自他們簡單的網頁瀏覽。

（三）恢宏闊大的開發規模

從上面的羅列中可以看出，中國古籍的數字化主要依賴於學術研究機構（如臺灣中央研究院）和學校（如北京大學），專門技術公司（如超星、北京書同文）和出版社（如迪志、漢文化聯盟），圖書館（如北京圖書館、上海圖書館），並且已經有了不少優秀的成果問世。此外，就筆者瀏覽所及，目前正在開發的古籍數字化相關工程還有：

1．北京大學劉俊文先生主持的“中國基本古籍庫”光盤工程（http://www.cn-classics.com/default.htm）。這是中國目前最大的一個古籍數字化工程了。該工程 1998 年正式啓動，全套光盤庫共 500 張，分哲學、史地、藝文、綜合 4 個子庫 20 個大類，範圍涉及先秦至民國的重要典籍 1 萬餘種並提供多重檢索功能，每種典籍有 1 個通行版本的全文信息，另附 1－2 個珍貴版本的圖像數據，預計全文 20 億字，版本圖像 2 千萬頁。應該說，這個項目具有很好的前瞻性，一旦開發成功，將基本上可以滿足文史和其他方面研究者的研究需求。

2．漢文化聯盟的“漢文化資料庫”（http://www.hanculture.com/），內容涵蓋歷代重要古籍，按照經、史、子、集進行分類，精選底本勘校而成，與單純的全文檢索不同。資料庫中包含《龍藏》（部分）、《歷代書法全集》（圖、文）和“漢文化考證檢索系統”，極富特色。

3．中國中醫藥文獻數字化。國家科技部基礎工作項目。該項目由中國中醫研究院中國醫史文獻研究所(http://www.cintcm.ac.cn/catcm/ys/yssjj.htm）承擔，山東中醫藥大學、南京中醫藥大學、上海中醫藥大學以及浙江省中醫研究院、天津中醫研究院等十四家中醫藥教育、科研單位協作。

4．《歷代石刻史料匯編》、《永樂大典》全文檢索。北京書同文數字化有限公司開發，預計不久即可見成果。該公司亦將地方志的數字化列入了規劃。

5．中國數字圖書館工程（http://www.nlc.gov.cn/dloff/）。該工程由中國國家圖書館倡議於 1998 年，2001 年 10 月正式國務院批准立項。目前已成立了“中國數字圖書館聯盟”，聯絡了 41 家公共圖書館、30 家高校圖書館、8 家科研機構圖書館、7 家其他類型的圖書館和 14 家技術公司。該工程的總體目標是在寬帶 IP 網上形成超大規模的、高質量的中文資源群，支持國家整體創新體系的形成與發展，通過國家骨幹通信網絡系統向全國乃至全球提供服務。其工程的重要內容之一，就是發掘歷史遺產。

6．“中國教育科技數字圖書館”（China Education and Research Digital Library，簡稱 CERDLIB）工程。該項目由美國卡內基－梅隆大學教授 Dr. Raj Reddy 和西蒙學院陳劉欽智博士、中國科學院研究生院常務副院長高文博士共同發起，旨在建設面向教育和科研的包含 100 萬冊（中、英文書籍各 50 萬冊）圖書的數字圖書館，由教育部“211”工程建設辦公室歸口管理，浙江大學和中國科學院研究生院等單位共同承擔建設任務和運行管理。計劃通過 4 年的時間，建成 2 個數字圖書館技術中心(浙江大學，中國科學院研究生院)和 12 個數字資源中心（北京大學、清華大學、吉林大學、武漢大學、西安交通大學、中國科學院研究生院、浙江大學、

復旦大學、南京大學、中山大學、四川大學、上海交通大學），開發出 30 個左右的大型特色數字資源庫。項目中包含了古籍數字化的內容。

7．臺灣典藏數位化計劃。2001 年啓動，參加機構有臺灣中央研究院、臺灣故宮博物院、臺灣歷史博物館、臺灣大學等。

8．中國歷史地理信息系統（CHGIS）。復旦大學歷史地理研究中心（http://yugong.fudan.edu.cn/chinesegb.asp#）、美國哈佛大學、哈佛燕京學社、澳大利亞格林菲斯大學亞洲空間數據中心等機構合作開發，主要資助單位是美國羅斯基金。項目目標是建立中國歷史時期基礎地理信息系統數據庫，使其成爲中國歷史 GIS 數據的基礎平臺。這是一種研究性、工具實用性都很強的開發項目，與單純的把古籍轉化成圖片或檢索文本的數字化不同，也顯示出古籍數字化的另外的發展方向。

9．“國際敦煌學項目”（The International Dunhuang Project，簡稱 IDP）（http://idp.bl.uk）。英國圖書館開發，開始於 1993 年。項目中包括了英國（並計劃擴展到世界範圍內）的敦煌文獻數字化。目前可線上查看英國圖書館收藏的 3 萬餘件中亞寫本和印本文件，以及 15000 餘件殘片的高質量彩色圖片（http://idp.bl.uk/IDP/idpdatabase.html）。

10.“古藤堡計劃”（Project Gutenberg，簡稱 PG）（http://promo.net/pg/），美國伊林諾斯大學文理學院米歇爾·哈特（Michael S. Hart　http://promo.net/hart/）倡始於 1971 年。計劃對世界上的一些經典古籍進行數字化，其中包含了中國經典、文學、歷史和地圖等方面書籍和資料的數字化內容（主要爲英文），如《論語》、《桃花源記》、《三國演義》等。現可在網絡上進行目錄檢索（可使用中文，http://promo.net/cgi-promo/pg/t9.cgi）。相關信息可通過在北京大學哲學系網站下載（ftp://elib.phil.pku.edu.cn/pub/gutenberg/），另有它還在中國設立了一個限中國境內訪問的鏡像站（ftp://ftpbook.dhs.org/mirrors/gutenberg/）。

此外，像超星數字化圖書館和地方上一些公共圖書館等也在其數字化圖書中不斷增添新的古籍，一些古籍研究機構和個人也在工作中將相關文獻變成電子文本，也是古籍數字化隊伍中的不可小視的力量。

合現在已有的成果而觀之，中國古籍數字化規模之宏，形式之夥，數量之多，範圍之廣，都標示出其良好的發展路向。這些產品的開發，不僅更加有效地保護了古籍，而且在豐富人們的學習資源、提高學術研究的質量以及擴大中國文化的影響上，無疑也會發揮積極的作用。

二、存在的問題

毫無疑問，古籍數字化的最終目的，是更大範圍、更便捷、更有效地使這些文獻爲人們所利用。而由於用戶群體及其需求的差異，這個“用”也顯示出不同的層次。因此，滿足這些用

戶群體的需求，也應該是古籍數字化所需要考慮的問題。從這方面看，目前的古籍數字化產品還有不少可以完善的地方。

（一）選題內容重複，分散了數字化資源開發的力量

選題的好壞直接關乎一個產品的成功與否。在這方面，《四庫全書》的原文電子版（圖像，濟南）、《四庫全書》的全文檢索版、《四部叢刊》全文檢索版，以及超星圖書館中的北大古籍等“館藏”（pdg 圖像格式），等等，從解決用戶的迫切需求上而言，都是成功的。但也無庸諱言，在今天的各種數字化產品中，內容的大量重複也是一個不爭的事實。這種重複不僅存在於一些小規模帶有普及性的電子書製作上，而且也同樣存在於一些大型的數字化工程中。最明顯的例子是《二十五史》，幾乎現在稱得上數字化項目的產品都全部或部分包含了它們。這種重複在人力和財力上都是一種浪費，它提高了開發成本，相應減少了對其他可開發項目的投入。

造成這種重複的原因，除了有些文獻本身屬於基本的典籍，製作單位需要利用它們以獲得基本的用戶外，資源共享上的隔膜恐怕也是一個重要原因。特別是在電子文獻資源、技術資源和信息資源上的共享上，這種隔膜對更多更好項目的開發是相當不利的。這方面的例子可以列出許多，其大而又大者如《四庫全書》原文圖像版就有兩種，超星圖書館和國家圖書館及其他公共圖書館在圖書內容上的大量重複。特別是對後者擁有的數字化資源在互利互惠的前提下進行合理合法的利用，使致力於 OCR、全文檢索引擎等數字化技術和實用工具開發的公司將更多的精力投向相關技術開發，對產品質量的提高無疑是有重要作用的。

（二）文件格式繁多，造成了閱讀和資料提取的麻煩

數字化圖書格式之繁多，從上文的相關羅列中就可以看出。由於不同的圖書閱讀器互相難以兼容，因而給讀者帶來了許多麻煩。頗爲有趣的是，有一個數字圖書館採取 pdf 文件存儲格式，卻只能使用 acrobat reader 去閱讀，連 adobe acrobat 也不支持。事實上，沒有一個讀者會願意在自己的電腦上安裝七八個不同的閱讀器去閱讀圖書。以筆者個人爲例，筆者曾經是許多網上圖書館或書庫的註冊用戶，但因爲煩不勝煩，現在常常光顧的只剩下超星數字圖書館了。

不僅如此，由於許多圖書採用了圖片格式，也給使用者提取資料帶來了許多困難。這當中，超星因其巨大的影響而常常成爲批評的對象。筆者無意於否認超星保護版權之努力的必要性和合法性，但因其相關技術開發不同步，其本應發揮更大作用的大量圖書資源並未得到很好的利用。限制打印頁數（每次 10 頁）、瀏覽器所帶的截圖工具每次只能截取頁面的 1/4（最新版本已經改成可整頁截圖）徒自給合法用戶（讀書卡註冊用戶）增添麻煩（因爲其瀏覽器的破解版本完全可以破除這些限制），而且其所提供的 OCR 工具，對當代標準印刷的直排繁體圖書的文字識別已經無能爲力，更談不上識別古代寫、刻、鈔、稿本的文字了。在這一點上，超星以及

類似的數字圖書館似乎僅僅把自己定位在爲讀者提供普通的圖書閱讀，而對專業研究者所需要的快速檢索、並將檢索結果直接轉換成編輯文本方面還關注不夠（超星提供的全文檢索工具頗差）。

（三）隊伍組織、項目規劃單調，導致了產品開發缺乏連續性

在隊伍組織方面，多數項目的開發以計算機技術和圖書館人員爲主，技術公司和圖書館之間的合作，更多的是因爲圖書館擁有原始文獻資源。而圖書館本身對館藏圖書的數字化，又常常拘泥於傳統圖書館的圖書借閱形式，而只是將過去的人工手段轉化成計算機通訊（這可能也是現在的數字圖書館多數採取圖片存儲格式的一個原因）。在這一點上，現在的一些項目的開發與用戶的需求之間還有相當的距離。

應該說，在隊伍組織和項目規劃方面，目前並非沒有比較成功的範例可以借鑒。臺灣中央研究院的系列電子化項目、元智大學的“網路展書讀”、漢文化聯盟的《漢文化資料庫》採取的都是專業研究專家與技術人員、圖書館三方合作的方式，使得開發的項目與用戶（尤其是專業研究者）的實際需求相切合。北京大學的《中國基本古籍庫》在設計思路上也採取了這種方式。而超星數字圖書館目前在其瀏覽器設計中已加入了可由用戶編輯專題的虛擬圖書館，使資源開發者與用戶之間建立起一種動態合作關係，也顯示出一種可喜的變化。

在項目的整體規劃方面，也有不少可以挑剔的地方。按照我個人的理解，一個項目的開發至少應該包含這樣的一些步驟：

開發者對自身開發能力的合理評估和發展目標的合理定位－根據前者選擇選題方式、進行市場調查和確定選題－選題可開發內容的信息搜集－確定開發的具體目標（主產品和副產品）－採樣－技術處理過程－測試－修改與完善－發布－市場反饋－補丁

可以看出，從選題開始，項目開發就是一種多方互動的活動。從選題方式上看，單純依賴開發者自己的想像，或者依賴文獻資源擁有者的倡導，或者依賴部分學者的評議，甚至依賴權威的一兩句斷言，顯然都是不合適的。最根本的一點，就是項目開發應該根據“甚麼最需要”而非“是否有價值”（如果考慮到開發公司自身的生存問題，還應該加上“是否能夠盈利或具有盈利的希望”）的原則去選定，而專家學者的看法通常著眼於“價值”，這種價值認定又往往因其專業限制而難免有局限。

選題確定以後，對選題可開發內容的信息搜集直接關乎項目開發的連續性。尤其是一些具有“原創”性質的開發項目，其本身所具有的系列開發內容就很豐富。僅我個人所思及，就有這樣一些方面：

其一，版本信息（版本類型、年代、版式、字體、刊刻地點、刻工姓名等）。這部分內容可以通過掃描而得到影像資料而進一步開發。

其二，全部書籍的詳細目錄匯總（用以滿足不能購買整套軟件但希望掌握相關信息的用戶的需求）。

其三，最常用或極具價值的書籍資料（單行或選編進一類專用書籍中，以滿足一般用戶的需求）。

其四，項目中所包含的各類專題資料（用以滿足專題研究者需求，或引導一般用戶進行相關查詢）。

反觀現在的一些開發項目（特別是全文檢索項目），除了所謂“單機板”、“網絡版”之類的“系列”外，在其他方面的開發幾乎爲零。項目開發者常常抱怨開發出的好產品沒有更多的用戶使用，卻往往忽略了另外一個問題：爲甚麼不利用已有的開發成果，而再開發出能適應不同用戶群體需求的產品來?

除以上三方面外，技術處理上的缺陷、成果推廣上的遲緩、開發成本過高導致產品價格過高等，也都對成果的更大範圍的應用有著一些影響。

三、“我”需要甚麼：《四庫》全文檢索案例分析

選擇《四庫全書》全文檢索版作爲案例，是因爲它是目前古籍數字化的一個非常突出的代表。而對它進行分析，只是想根據我自己的使用感受回答這樣一個問題：用戶究竟需要甚麼?

（一）信息容量

包含 3400 餘種書的《四庫全書》全文檢索，無疑是現在容量最大的一個古籍數字化工程了。雖然《四庫》本身因編纂、版本等方面的問題而爲學者所詬病，但因其文獻集中，而影印本又可以很快解決掃描底本問題，選擇它進行數字化在目前無疑是非常正確的。比較一下它和《中國基本古籍庫》的工作進度，就可以發現《四庫》全文檢索在解決用戶的最迫切需求上厥功甚偉。《中國古籍基本庫》自 1998 年啓動，至今將近四年，尚未見到成型的產品面世。而《四庫》前後僅三年就完全開發成功。甚至當初極力反對這項工程的學者，現在也成爲它的積極的使用者，這本身就說明了這項工程的價值所在。

《四庫》全文檢索之受到歡迎的一個重要原因，就是其信息量的巨大。這種情形，也反映出另外一個問題：倘若在版本與信息容量上不能兼顧，是選擇版本好但容量少，還是選擇版本稍差但信息容量大的圖書進行數字化? 據筆者本身以及所知的一些情況來看，恐怕多數人還是選擇後者。舉例說來，筆者的一位學友想搜集歷史上蝗災的資料，每日前往圖書館翻閱圖書（逐頁翻查，苦不堪言），猶恐遺漏，後通過筆者檢索《四庫》“蝗”字，即刻得 4535 卷、11329 個匹配。由此把節省的大量時間轉入資料的考訂和搜集《四庫》所無之書中的資料，較之其先

前的工作方式，優劣判然。

也因爲如此，筆者深感已經大大超越同類數字化工程容量的《四庫》全文檢索，在容量上仍然不能滿足要求。比如筆者目前正在進行《明儒學案》的文獻學研究，想查證其中的傳記資料和黃宗羲摘編的學術資料的原始來源，《四庫》全文檢索對多數人物無能爲力。《四庫》中宗教類資料很少，明代著作未收者頗多，而清代因修書時代限制幾乎無法利用，這些缺憾都還需要其他數字化項目來補充。

（二）顯示模式

《四庫》電子版採取了檢索結果、原文圖像、全文閱讀三種顯示模式並可快速切換，亦屬獨創。這種顯示模式的確有它的好處。原文圖像和全文閱讀的精確對應，在兩者之間建立起了直接的勘校關係，可以解決全文閱讀時的部分文字錯誤。特別是對古今字、異體字、避諱字等的關聯檢索沒有達到完善匹配的時候，原文圖像在補字和校正錯字方面就有重要作用。

但是，這種方式也存在諸多問題，匯總如下。

1．檢索結果方面

1）單機版的檢索結果能打印但不能複製（網絡版可以通過網頁拷貝方式複製），有卷數、書名而無其他可顯示該條資料的內容，使得在缺乏隨身攜帶全文檢索的情況下無法與其他的書籍內容進行比對。（這一點在書同文《四部叢刊》的開發中已經得到修正。）

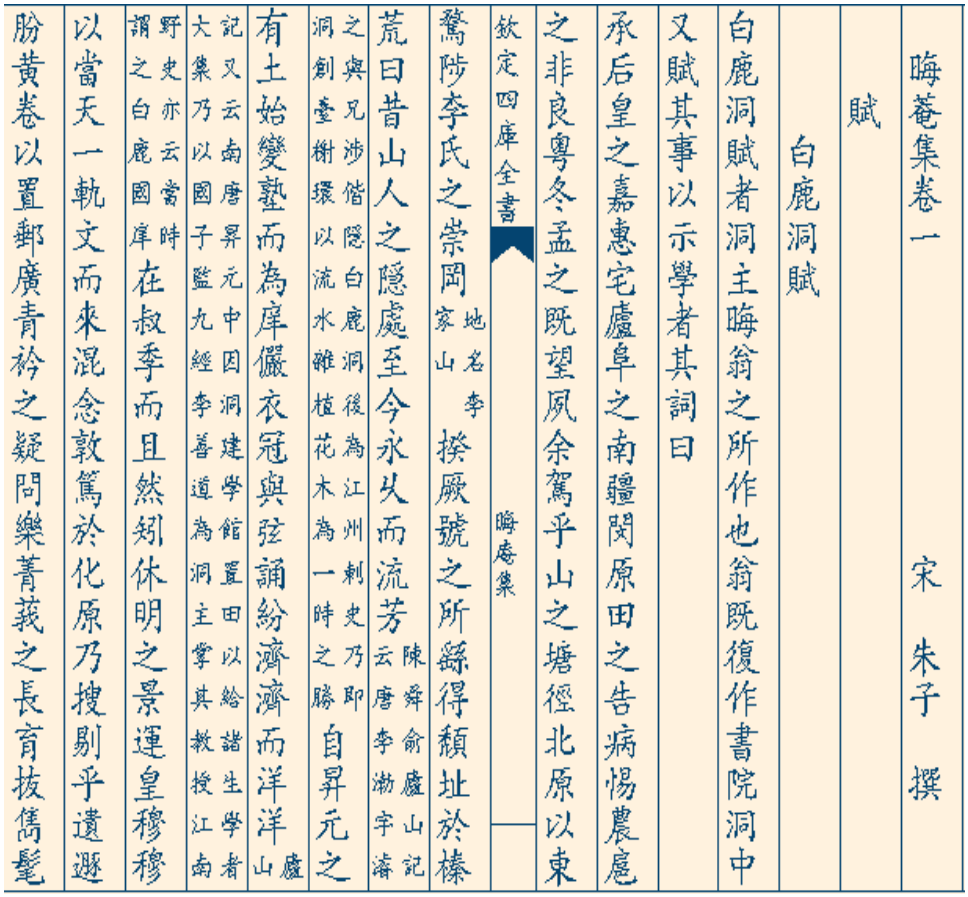
2）檢索結果必須通過閱讀原文才能知曉具體內容，不能集中顯示，也給用戶使用帶來了一些麻煩。比如“朱子”的檢索結果就高達 9133 卷、37910 個匹配，若將此外的“文公”（13180 卷，33153 個匹配）、“晦庵”（1665、3293）、“朱熹”（1839、4508）加在一起，計有 25817 卷、78864 個匹配。至於“孔子”，更是高達 23757 卷、111641 個匹配。假定每個匹配的閱讀時間平均爲 1 分鐘（加上複製相關資料、標點，實際一條資料的處理時間遠遠超過 1 分鐘），每天八小時不間斷地閱讀，則“朱子”等條資料需要花 164 天、“孔子”需要花 233 天才能閱讀完畢。如果是通過網絡閱讀，其麻煩會更大些。（臺灣中央研究院的《漢籍電子文獻》、陳郁夫先生的“寒泉”檢索系統都採用了可以分段顯示的方法，甚是便利。）

2．原文閱讀方面

1）原文顯示上區分正文和注文，一方面將有些並非注文的小字誤作注文，另一方面在拷貝時將注文置於頁末，頗爲不便。（後者在《四部叢刊》全文檢索中已經得到糾正。）

2）未妥善解決異體字等關聯問題，導致有些文章在拷貝後必須補充大量空缺的文字（四庫自帶了方正楷體大字庫，但並非所用使用者的本地機上都有這種字庫，因而在無此字庫支持的電腦上閱讀和編輯都不方便）。

此兩者可舉《晦庵集》中一條爲例。原文截圖如下：


晦菴集卷一　宋　朱子　撰

賦

白鹿洞賦

白鹿洞賦者洞主晦翁之所作也翁既復作書院洞中又賦其事以示學者其詞曰

承后皇之嘉惠宅廬阜之南疆閔原田之告病惕農扈之非良粤冬孟之既望夙余駕乎山之塘徑北原以東驚陟李氏之崇岡（地名李家山）揆厥號之所繇得頹址於榛荒曰昔山人之隱處至今永久而流芳（陳舜俞廬山記云唐李渤字濬）自昇元之有土始變塾而為庠儼衣冠與　誦紛濟濟而洋洋（廬山）在叔季而且然矧休明之景運皇穆穆以當天一軌文而來混念敦篤於化原乃搜剔乎遺遞胐黃卷以置郵廣青衿之疑問樂菁莪之長育拔雋髦

欽定四庫全書　晦庵集

拷貝到相關編輯器中的結果如下（截圖）：

晦菴集卷一宋朱子撰賦白鹿洞賦白鹿洞賦者洞主晦翁之所作也翁既復作書院洞中又賦其事以示學者其詞曰承后皇之嘉惠宅廬阜之南疆閔原田之告病惕農扈之非良粤冬孟之既望夙余駕乎山之塘徑北原以東驚陟李氏之崇岡揆厥號之所繇得頹址於榛荒曰昔山人之隱處至今永久而流芳自昇元之有土始變塾而為庠儼衣冠與　誦紛濟濟而洋洋在叔季而且然矧休明之景運皇穆穆以當天一軌文而來混念敦篤於化原乃搜剔乎遺遞胐黃卷以置郵廣青衿之疑問樂菁莪之長育拔雋髦地名李家山陳舜俞廬山記云唐李渤字濬之與兄涉偕隱白鹿洞後為江州刺史乃即洞創臺榭環以流水雜植花木為一時之勝廬山記又云南唐昇元中因洞建學館置田以給諸生學者大集乃以國子監九經李善道為洞主掌其教授江南野史亦云當時謂之白鹿國庠↵

3．原文圖像方面

筆者無意否認附帶原文圖像的價值，但這種做法事實上造成了產品難以在更大範圍推廣。《四庫》的全文檢索安裝盤僅 16 張光盤，而圖像盤高達 167 張，這無疑也提高了製作成本，從而相應帶來了價格的提高。國內不少用戶對《四庫》全文檢索心嚮往之，卻最終沒有去購買，高昂的價格大概是使他們望而卻步的一個重要原因。

（三）檢索模式

《四庫》提供了全文檢索、分類檢索、書名檢索、著者檢索以及“開啓當前檢索條件”用以修正當前檢索的內容。全文檢索還可分部、分書或分著者進行，並支持複合檢索。而在全文閱讀狀態下，還可以通過選擇當前閱讀頁面中的文字進行再檢索。分類、書名、著者檢索提供簡單、具體、詳細三種顯示方式並提供了相關鏈接以進行切換。此外，在全文檢索、書名檢索、著者檢索中還增添了一些輔助功能設計，分類檢索中則包含了部、類、書、目錄的層級搜索方式。這些檢索方式，可以滿足用戶不同的檢索要求，使用也很方便，非常值得贊賞。而附加的聯機字典、添加筆記、放大鏡等工具也頗具實用價值（其聯機字典猶具價值，惜釋義稍簡）。

感覺不方便的地方是：原文的卷次顯示於頁面的底端，位置不當。有些著作的卷次顯示尤有問題。比如別集類明代的一些著作，標成“集部，別集，洪武至崇禎，？…”（？爲書名的第一個字），顯示了前面一堆無用的信息，而關鍵的卷次信息卻被省略。

（四）檢準率

從整體上看，《四庫》全文檢索的命中率應該是很高的。筆者利用它考證一些概念的演變、人物的生平、古籍整理上引文的查核以及校勘等，都取得了滿意的結果。甚至偷閑的時候胡亂檢索一些字詞，也常常得到意想不到的結果。比如“愛情”兩字，檢索得 159 卷 163 個匹配，雖然其中許多都是兩字碰巧排在一起，但也確實有兩字連用者。如《禮記集說》“若愛情在心，則聲和柔”、《續資治通鑑長編》卷一六九“伏望陛下斷以大義，稍割愛情”、《清河畫舫錄》卷一二上“自亦不堪屬目，以徇愛情而已”之類，雖意思與現代的男女情愛邈不相關，亦頗見古今詞義之變化。尤爲有趣者，現代人通常把歷史上的禮制想像得非常可怕，而且似乎歷久不變，然檢索“離婚”兩字，得 172 卷 192 個匹配，其事件可追於春秋，而《晉書》即屢屢見“離婚”二字，若輔以其他記載，諸資料內容頗有可判今人認識之誤者。

要求《四庫》全文檢索在命中率上達到完全無誤，顯然是一種太過苛刻的要求。但從完善產品的角度言，《四庫》在檢索的準確率上的確還有待提高。由於版本不同，筆者無法用其他的全文檢索產品來進行對比，在此僅列舉一個事例以作說明。

檢索“講會”二字資料，得 36 卷、37 條。然通過個人所知者複核，發覺《四庫》脫漏甚

多。比如《法苑珠林》一條，另載於《廣博物志》卷五、《太平廣記》卷九九兩條未檢出；《東都事略》卷一一四、《宋名臣言行錄外集》卷三所載“赴講，會”條，另載於《續資治通鑑長編卷》四〇四、《太平治跡統類》卷二五、《伊洛淵源錄》卷四、《二程遺書》附錄、《近思錄集注》附說、《御纂朱子全書》卷五三、《晦庵集》卷九八諸條未檢出。《晦庵集》檢索得一條，但筆者所知另一條《白鹿講會次卜丈韻》詩（卷七）未能檢出，而此詩又載於《江西通志》卷一五四、《性理大全書》卷七〇、《御纂朱子全書》卷六六、《宋詩鈔》卷六〇，亦未檢出。此外，史部正史類無一條檢出，而《舊唐書・蕭俛》附蕭仿傳，《明史》沈懋學、史孟麟、呂維祺傳及顧憲成等傳贊、儒林傳之陳時芳傳中均有“講會”字。一條檢索出現如此多的失誤，對一個成熟的產品來說是不應該的。

此外，《四庫》全文檢索中還有不少錯字，也影響了它的質量。(《四部叢刊》全文檢索版允許用戶在本地機改正錯字，值得提倡。)

應該說，《四庫全書》全文檢索版的開發，爲中國古籍的大規模數字化提供了一個成功的先例。雖然從用戶的實際需求方面言，這個產品還存在一些問題，但是，不能要求一種產品完成用戶所想做的所有事情，也應該是合理對待這類產品的態度。而正因爲如此，《四庫》以及類似數字化產品依然爲未來的數字化工作留下了很大空間。

四、趨勢

關於古籍數字化的發展趨勢，臺灣元智大學羅鳳珠先生（“網路展書讀”的開發者）《臺灣地區中國古籍文獻資料數字化的過程與未來的發展方向》（http://cls.admin.yzu.edu.tw/present/tarcf.htm）一文言之甚詳，已無需筆者班門弄斧。唯個人覺得：大型綜合性可提供全文檢索等功能的、建立者與用戶動態合作的網絡數據庫（或網絡圖書館）的建設，以及投資商、技術開發公司、圖書館和專業研究機構合作開發更多、更專門的數字化產品，或許是發展的主要趨勢。此外，謀求同國外一些機構的合作，以多種方式將數字化產品推廣出去，借以傳播中國優秀之文化，亦當在考慮之列。

後記：本文行文當中，偶然自網路檢得羅先生此文，細讀之後，甚爲敬佩，若再寫“趨勢”一節，徒爲蛇足，故此略過。然羅先生所建之“網路展書讀”及所知臺灣其他一些可供全文檢索或閱讀的網路資源，雖屢經嘗試而網頁終無法打開，故行文中只能略略道及。另於香港中文大學中國文化研究所的“華夏文庫”及“古文獻資料庫”http://www.chant.org/scripts/main.asp所知甚少，不敢妄及。此外，凡文中所引資料，均來自相關網站，因嫌重複，未注明來源，尚祈諒鑒。而匆匆成文，言詞觀點定多謬誤，亦祈讀者斧正之。

特別鳴謝：漢文化聯盟朱邦復、欒貴明先生提供此難得學習機會；北京書同文數字化技術公司張軸材先生先前提供試用《四庫全書》網絡版及新近提供《四部叢刊》試用版，上海迪志文信息科技發展有限公司由開發公司授權提供《四庫全書》全文檢索個人用戶版。

作者簡介

吳宣德

筆名蘊之，安徽大學中文系文學學士，華東師範大學古籍研究所古代文學碩士，教育學博士學位。

著有《清代前期教育論著選》、《中國教育大系・歷代教育制度考・明代編》、《中國教育大系・歷代教育論著選評・明清編》、《中國教育思想通史》第四卷、《中國教育思想史》第二卷、《中國教育制度通史》第四卷《明代》、《中國教育史話》、《中國區域教育發展概論》，並整理《朱熹集》、《朱子語類》。