「大數據時代的古籍進化論研討會」心得報告

流通組 謝鶯興

一、前言

「大數據時代的古籍進化論研討會」是希望透過邀請學者，提供個人運用大數據時代的「海量古籍資料庫內容，用傳統無法達到的研究方法，讓證據說話，並擴大研究者的空間與視野。」以研究者的「基本素養，加上大數據時代的古籍知識礦場探勘應用，佐以自己的創新思維」。黃一農院士「認為大數據的時代足以為人文社會科學領域開創一個學術黃金期。」(以上引「會議緣起」語)

基於會議「緣起」所揭櫫的概念，身為館藏古籍整理的工作者，引發莫大的興趣，而參加這個研討會。

二、議程及各場次內容概述

研討會議程

■ 時間：2017 年11 月2 日（星期四）8:30 - 16:45
■ 地點：中央研究院學術活動中心二樓 第一會議室
（臺北市南港區研究院路2段128號）

時 間	講 題	講 者
08:30 - 09:00	領取資料及入座	
09:00 - 09:15	致歡迎詞	羅志承 總經理 漢珍數位圖書公司
09:15 - 10:15	"隱藏版"的史實： 資料庫的使用與侷限	衣若蘭 副教授 國立臺灣大學 歷史學系
10:15 - 10:30	茶敘	
10:30 - 11:40	e考據與文史研究的新機遇	黃一農 教授 中央研究院院士/清華大學教授
11:40 - 12:00	數位古籍資料庫展示	楊學斌 經理 漢珍數位圖書公司
12:00 - 13:30	中午餐敘	
13:30~14:20	「漢達文庫」之構建與應用	王利 博士 香港中文大學 中國文化研究所
14:20~15:00	古籍資料庫應用經驗分享	劉衛君 小姐 國立臺灣大學 中國文學研究所博士生
15:00 - 15:15	茶敘	
15:15 - 16:15	城市商號研究的新工具： 廣告資料庫的介紹與應用	連玲玲 副研究員 中央研究院 近代史研究所
16:15 - 16:45	綜合討論	劉毓欣 協理 漢珍數位圖書公司

本議程共邀請五位學者，分就他們個人使用資料庫的經驗以及認知，提出他們的見解或觀點。

1.衣若蘭教授「『隱藏版』的史實：資料庫的使用與侷限」，首先揭開在資料庫使用的歷程，及其發現的侷限。以她曾研究的課題：「明清旌表節婦的變化」為例，從關鍵字的取得與搜尋，感受到資料庫的侷限，認為研究者需配合自身對於文本的閱讀與史料的翻覽，才能搭配資料庫的檢索機制取得相關

資訊，強調進行研究時，切莫以資料庫的檢索結果即輕易地認為是唯一的答案的基本觀念。

2.黃一農教授「e 考據與文史研究的新機遇：以查索古代詩文之用典為例」，說明他嘗試深入學習該如何融通治學方式與數位研究工具，進而歸納出幾種有效且快速的操作程序，在研究實例中，從：「『蜍志』之用典」、「『呂袋』之用典」為例，提出 E 考據的主要精神並非只著重在搜尋，而是幫助文史工作者直接且迅速地與大量原典對話，以從事度學習，以及還要懂得隨時善用大數據的環境去充自我，在覓得重要材料時，要再去瀏覽該文獻的其它內容，以積累並擴展從點狀到面狀的知識；提出在大數據時代我們所缺乏的不是資料，而是對資料的敏感度、解析力與整合力。這種對「e 考據的反思」，對於初學者具有濃厚的指導原則與進行研究的基本觀念，相當值得我們省思。

3.王利博士「『漢達文庫』之構建與應用」，概述「漢達文庫」構建的緣起，現有的內容：傳世文獻資料庫，包括先秦兩漢資料庫、魏晉南北朝資料庫；中國傳統類書資料庫；詞彙資料庫；出土文獻資料庫，包括甲骨文資料庫、金文資料庫、竹簡帛書資料庫。並介紹他們利用這些資料庫進行的研究成果，有逐字索引，漢達古籍研究叢書，出土文獻專書等。在介紹幾種資料庫中同時說明這些資料庫的應用，亦即如何利用布林邏輯的檢索方式，取得所要的資料。最後應主辦單位的要求，附帶提出他個人對於「中國基本古籍庫」使用的看法與建議。

4.劉俐君博士「古籍資料庫應用經驗分享」，是從自身使用的經驗進行分享，並論及各資料庫的優缺點。首先是「中文古籍資料庫資源述略」，概分為古籍全文資料庫和古籍書目資料庫兩大類，劃分古籍資料庫的建置者類型，介紹「公開取用古籍資料庫」，再者談各大學採購商業型的古籍資料庫現況。第二單元是對資料庫應用的經驗分享，分：常見的使用目的，資料探勘：傳統與科技的互補互利，研究實例，古籍資料庫常見版本問題，關於資料庫主機：本地版與在線版的資料落差等敘述。在經驗分享中特別提出「檢索注意事項」：著錄方式影響檢索結果，正體字、簡體字影響檢索結果、空格斷行影響檢索結果、二次檢索的局限、關於界面設計的建議(廠商設定畫面呈現字數涉及檢索的結果)等。版本問題方面提出：消失的序跋、目錄、責任人、附錄等；全本？節本？未曾聲明的內容刪削；電子全文與「原據版本」不合；未必理想的底本等四項，分享其見到的問題。

5.連玲玲教授「城市商號研究的新工具：廣告資料庫的介紹與應用」，介

紹利用中研院近史所自建的：近代婦女期刊資料庫，申報資料庫、Chinese Women's Magazines 資料庫，中國商業廣告資料庫，近代城市小報資料庫的經驗，以及如何利用這些資料庫研究，如女性形象比較，城市間的商品廣告比較等，並提出資料庫建置架構、檢索功能、統計功能等皆將影響使用者的研究成果，雖然可以藉由圖表看出統計數量，但不可逕認為是唯一的，正確的研究成果。

三、心得與感想

聆聽五場次的講演之後，發現現今對於數位化，似乎還是以掃描文件，再轉換為文字，提供關鍵詞或全文檢索就足夠的觀念進行，不知道是否有注意到使用者的需求，是否注意到使用者的最終目的在於研究需求，以及如何提高或強化資料檢索的功能。

以第一場次「『隱藏版』的史實：資料庫的使用與侷限」及第二場次「e考據與文史研究的新機遇」，講者都強調回歸文本的翻閱史料，要對資料有敏感度、解析力與整合力。亦極在大數據時代中的運用資料庫的檢索功能，仍要回歸傳統的基本訓練與閱讀文本的培養功夫。言之諄諄，令人感佩他們治學的嚴謹。

聯想到自己平日在館內處理讀者查詢資料的問題，只要讀者提出為何找不到他所要的資料，第一個反應是問他使用何種方式、哪些字串查詢。如何他所用的字串太長而找不到，就會建議少幾個字，也就是不要用太精確、完整的字句檢索；如果用簡單幾字檢索還是找不到時，就建議試用「同義詞」、「同義字」，或是衍生詞、擴大義等方式，因為中國字的字義，「同義多詞」的現象很多，可以考慮利用。

因為讀者所面臨的問題，與「隱藏版」史實的講者所提資料庫的侷限頗為相似，她建議不要僅單層的檢索，要有第二層、第三層的根據所得結果再進一步檢索的功能。個人認為或可從圖書館界的「權威檔」模式思考，將「同義多詞」的字詞以及擴大義的詞彙，整合為一個詞，若檢索出來的資料繁多，則在第二層出現的「多詞」的字串，再在「檢索結果」中搜尋；甚至可以在第三層再進行搜尋，或可得到想要的資料。

從第四場次「中文古籍資料庫應用經驗分享」所提的「檢索注意事項」，告訴我們：「著錄方式影響檢索結果」、「空格、斷行影響檢索結果」的事實，也經常在館內讀者檢索過程，特別是被收入叢書之書最為常見。以「著錄方式」言，館藏檔的建置是按照「中國機讀格式」，只要建入固定欄位之書籍，

絕對可以檢索到，但因叢書僅出現該冊數種書籍的第一本書名，較少使用的讀者就會誤以為沒有他所要的書籍。至於資料庫的全文檢索，雖有「空格」或「斷行」的因素而影響到結果，可以善加運用「布林邏輯」的進階檢索功能，還是可以克服，如黃教授在第二場次中，針對「中國基本古籍庫」沒有詳細說明資料庫的使用方法，但從他個人使用經驗中所揭示的「進階檢索符號」的搭配便用，或如第四場次劉博士的應用經驗分享，特別是在進階檢索時，相同「詞串」的檢索位置改變，得到的結果會有差異經驗，提供我們解決一些資料庫的侷限。

綜合此次的研討，以及個人對於資料庫的建置看法，提出下列的幾個面向的思考；

一是資料庫的建置架構，能以使用者角度思考，從資料庫的建置目的與使用者的需求為目標，瞭解數位化的最終目的是否僅是文件掃描的另一種保存方式而已，還是要提供檢索使用的；提供關鍵詞或全文檢索為終極目標而已，還是能提供研究者更深入的研究；研究者可以經由資料庫的運用，得到有如第三場次「漢達文庫」建置後的副產品：出版各種相關研究的出版品。

二是古籍資料庫內容的正確性，與底本的選定和校訂工夫是息息相關的。但檢索方法的設定，更會決定這個資料庫的適用與否的關鍵。如進行同義詞、類似詞、相關詞的整合，建立人性化的檢索，應可降底檢索之不足，特別是關鍵詞的考量與選定，更是一個關鍵。如果能延請學者參預，不僅是「顧問」，而是「團隊」的加入，相信對資料庫與學者研究，應是相輔相成，互蒙其利的。

三是傳統古籍的內容，對於年代的記載多是天干地支的紀年，雖然已有「中西曆對照表」可以檢索換算，終究是要再經一道手續。如「文淵閣四庫全書電子版」附有「古今紀年換算」、「干支/公元年換算」，即使可以在同一畫面輸入，似乎可以思考直接點選該詞串(紀年)，即可換算其公元年，是可以節省一道動作。