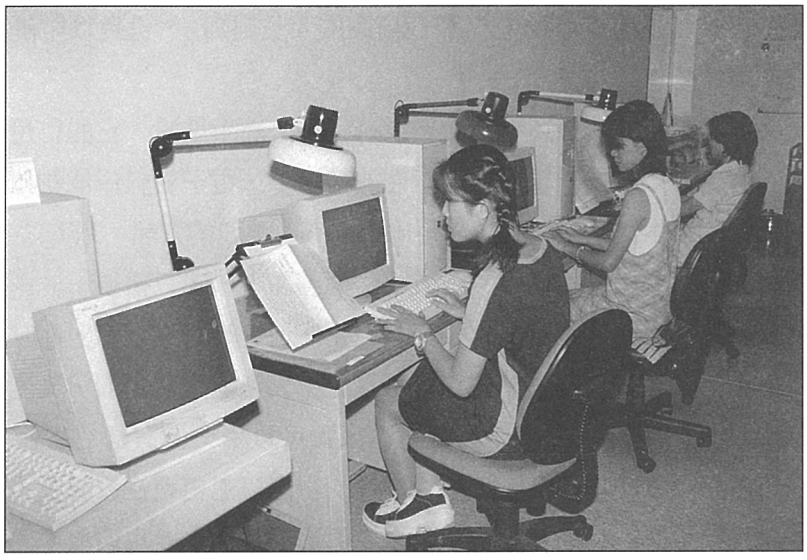
試談電子文獻資料庫在歷史研究上的應用



歷史語言研究所「漢籍全文資料庫」工作室

陳弱水（中央研究院歷史語言研究所）

在這篇文章，個人想從一個研究者的觀點談談電子文獻資料庫的應用問題，目的是在對作爲一種新型研究工具的電子文獻資料庫的價值作一個評估。中央研究院從一九八四年開始發展漢文古籍資料庫，至今已有十四年的歷史，院外的個人和機構也製作了許多同類的資料庫，現在應該是考察這個問題的適當時機。這個討論，或許也有助於我們對資料庫發展前景的考慮。個人在研究工作上，對資料庫的使用並不頻繁，對院外製作的資料庫的了解也很有限。我之所以不避淺陋，撰寫這篇文字，是因爲個人曾經主持過三年史語所漢籍自動化計劃的工作，對相關問題曾略有考慮，也許還有一些值得大家參考的意見。

這篇文章在結構上分爲三個部分。第一部分是要對電子文獻資料庫在歷史研究上應用的現況作一個描述，並討論有關製作研究用資料庫的若干問題。其次，我要對資料庫在歷史研究上的價值作一個評估，特別希望揭示資料庫與其他種類的研究工具的關係。最後，個人將根據以上兩部分的討論，對電子文獻資料庫在本院與台灣未來的發展，提出一些看法。

現在，我要設法對資料庫在史學研究上的應用情況作一個大體的描述。這個描述，基本上是根據個人的經驗。至於我所謂的「資料庫」，也是以中研院史語所開發的全文資料庫爲藍本。

電子文獻資料庫的功能，當然是在找尋資料。簡單地說，它的特色有以下幾點：任意檢索、全文檢索和快速檢索。在理想的情境下，電子資料庫能夠以任何詞語爲查詢單位，無遺漏、極快速地找到並調集出研究者所需要的資訊。以「二十五史資料庫」爲例，在這個約有四千萬字的資料庫查詢一個詞語，一般只需要一、兩秒鐘。和傳統的工作方式相比，電子文獻資料庫不但能大幅度地提高研究效率，而且可以帶來許多新的發現——因爲電子計算機的檢索既可以避免人工閱讀的遺漏，又能讓我們隨時進行以前幾乎無從著手的查詢工作。在這裡，我可以舉一個有趣的實例。一九九七年末，我在日本東京大學作研究時，一位年輕的中國思想史教授告訴我，他的一個學生在舉行學士論文口試的時候，一位年長教授對他的論文中表現的博學非常驚訝，他不知道，這些博學的表現其實都是從電子文獻資料庫查來的。由此可見，資料庫的威力驚人，它可以使一位異國的大學部學生輕易掌握廣泛的中國古典資訊。

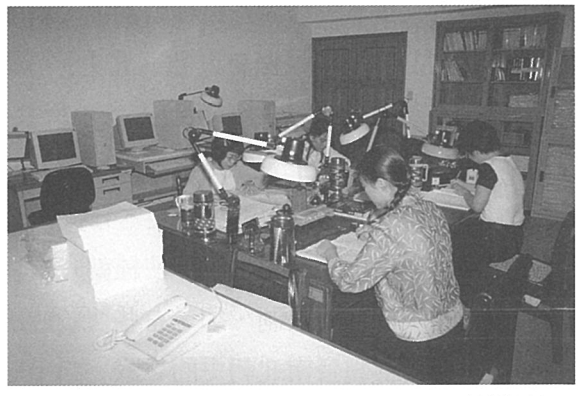
不過，要使電子文獻資料庫發揮上述的理想功能，並不是必然的，它需要某些條件。這就涉及了資料庫製作的問題，現在稍微談一下。首先，資料庫必須根據比較好、錯誤少的版本，這也要是通行的版本，讓研究者能夠容易找到原書，進行比對。在研究的原則上，如果沒有特殊困難，學者都盡量使用好的版本（可以減少錯誤、增加閱讀速度）。如果資料庫的版本選擇不當，就會給研究者帶來困擾，他可能需要用不同的版本來核對，如果卷數、頁數都不符合，翻查費時，資料庫快速查詢的優點就大打折扣了。

即使資料庫依據好的版本，也無法取代書籍，我們寫註的時候，必須引出版物，而不能引資料庫。何以如此的原因有好一些，譬如，資料庫有缺字、校對的問題。使用資料庫常須參考原書，還有一個實際的考慮。中研院的資料庫有瀏覽、閱讀的功

能，但在監視器上，並不方便作大量的閱讀，所以使用資料庫時，視野經常不夠廣，無法讓研究者充分掌握查詢到的資訊的涵義。（也許將來歷史學者可以直接引用資料庫，但現在顯然沒有達到這個地步。）總之，我相信，大多數歷史學者查詢資料庫後，都再拿書籍作對照。在這種情況下，資料庫使用的版本如果不普及，也是不行的。個人在主持史語所漢籍全文自動化計畫時，有一次要把一本重要的傳統醫學文獻製作成資料庫，但發現中研院圖書館內的版本都不理想，好的版本只有所內一、兩位研究人員私人擁有，而且這個版本已經絕版，只好暫時放棄製作。還有一次，一本書的唯一好版本是簡體字，因爲簡轉繁有輸入上的困難，加以校對不易，也放棄製作。

另外值得提出的一點看法是，資料庫如果用現代重排本製作，效果遠比傳統版本好。原因是，重排本有標點，容易在電腦監視器上閱讀。更重要的是，重排本有分段，在使用資料庫時，研究者只須先閱讀查詢字詞所在的段落，容易掌握資訊。傳統版本是以「卷」或文章爲單位，不分段，查詢個別的字詞，常常帶出好幾頁的資料，使用上頗有困難。《二十五史》是中研院最早開發、也是規模最大的資料庫，這個資料庫奠定了中研院漢籍資料庫的基礎與聲譽。依照個人的看法，「二十五史資料庫」的成功在相當程度上要歸功於北京中華書局的版本，沒有這麼精緻的版本作爲底本，這個資料庫恐怕會遜色很多。

除了以上所說的條件，資料庫要發揮好的研究功能，製作的技術和程序也很重要。譬如，檢索的軟體工具、版面設計、校對工作都是關鍵因素。我在這裡想要表達的一個看法是，中研院開發的漢籍資料庫普獲好評，一個重要的原因是，這些資料庫大體符合以上的標準。今後我們繼續開發資料庫，也不能不注意這些條件，否則資料庫的效益可能就會減低。現在再舉一個例子說明這個看法。個人曾經在網路上嘗試使用台灣開發的佛經資料庫，發現大多數在研究的利用上有相當的困難。其中一個原因就是版本使用不當或不明，查對不易。（其實，中研院的資料庫也常常沒有標示版本，都是由使用者憑常識判斷。目前資料庫的數量逐漸增加，如果沒有著作權問題的顧慮，應該明確說明版本。）


歷史語言研究所「漢籍全文資料庫」工作人員進行校對的情況

現在要換一個話題，討論一下資料庫在研究工作上有哪些具體的功用。依照個人的體會，資料庫在以下三個時機最有用：研究開始、研究結束以及處理具體難題時。研究開始，或考慮進行某一研究的時候，必須搜集資料，建立基本資訊。這個時候，如果有適當的資料庫可以利用，學者可以快速、有效率地開展工作。而且，由於研究伊始，這些資料的邊際效益都很高。研究進行中，當碰到特定的難題，找不到適當的工具書或既有研究來解決時，查詢資料庫是脫離困境的一條捷徑。另外，在研究將近結束，要考慮是否還有重大缺漏時，利用資料庫查詢各種與研究內容有關的關鍵詞，是拾遺補闕、自我檢查的好辦法。

以上談的三個時機，都是研究過程的重要部分。資料庫在這些地方能發揮重大功能，也可以看出資料庫對史學研究具有重要價值。然而，純就查詢資料而言，電子文獻資料庫仍然有它的限制，關於這方面的問題，本所同仁李貞德先生曾經有所討論❶。在這裡，我只想談一個——但我認爲是最重要的——限制。這就是，電子資料庫無法告訴我們要查什麼，或者有什麼可查。對於這樣的評論，有人也許會說，這本來就是研究者的事，與資料庫無關。我提出問題，並不是爲了引起這樣的反應。因此，現在想舉兩個例子，把意思說得清楚些。

第一個例子是個人研究的經驗。五、六年前，當時我還在加拿大工作，有一次作一個關於唐代初期政治與女性的研究，主要處理的人物有武則天、韋后、安樂公主、太平公主等。由於我探討的主要是心理、觀念層面的問題，需要的材料很細膩，瑣碎的記載也不能放過。當時我就用上海古籍出版社出版的《新舊唐書人名索引》，來找出兩唐書中關於這些人的所有資料，查得很辛苦，非常耗時。當時我就想，如果能利用中研院開發的「二十五史資料庫」，就方便多了。

過了一年，我回台灣，到史語所任職，上面所說的研究已經完成，論文也即將發表，但我還忘不了那個煩瑣的翻查資料工作。我就想像，我如果是利用中研院的資料庫，會是怎樣的情況。結果，出乎意料，我發現，我必須使用人名索引。如果我使用資料庫而不參照人名索引，會遺漏許多資料，研究的結果可能受到嚴重影響。

這裡有兩個問題。第一，我如果不用人名索引，根本不可能掌握應該查詢的詞語。以武則天爲例。在兩唐書中，武則天以二十一種稱呼出現過：武氏、武曌、阿武、武媚、武才人、武昭儀、武宸妃、武惠妃、武后、天后、聖母神皇、聖神皇帝、大聖天后、則天皇后、則天順聖皇后、則天大聖皇后、天皇聖帝、則天大聖皇帝、武太后、則天皇太后、則天大聖皇太后。純從研讀資料的過程，很難發現這麼多稱呼，從而一一檢索。其次，在史料中，有時人物根本不以完整的稱呼出現。最常見的情況是，當一個人和他的家人的傳記資料一起出現時，史書經常只提這個人的名字，而不提姓。但古人單名很多，在這種情況下，利用資料庫根本無從發現只提單名的傳記資料。最極端的例子是《南史》和《北史》。這兩部書通常把同一家族人的傳記放在一卷，除了該卷起首的人物，其他人就不稱姓。我們利用資料庫查詢南、北史中的人物，可能什麼都查到了，就是沒發現這個人的本傳。但另一方面，編輯嚴謹的「人名索引」等於是以人物爲關鍵詞的文本分析，很少有重大的疏漏。

這裡所要提出的看法是，即使純就查詢資料而言，電子文獻資料庫有它明顯的限制，而它的一個主要限制，卻是有些傳統的研究工具可以完全避免的。我再舉一個例子。日本有一群魏晉南北朝史的專家，二十七、八年前開始，在前東京大學教授西島定生的領導下，開始製作《魏書》的綜合索引。這個工作最近才告完成，還沒有出版，成果放在東京大學供學者使用。在《魏書》索引編纂的漫長時間裡，上海古籍出版社出版了《魏書人名索引》，中研院開發了「二十五史資料庫」。我去年有機會和

❶李貞德、陳弱水，〈中研院史語所漢籍全文資料庫介紹〉，《中國圖書館學會會訊》，四卷三期，頁4-10。（第二節，李貞德撰寫部分。）

《魏書》索引的主要編纂者，御茶水女子大學教授窪添慶文先生見面。他告訴我，他有些懷疑這部索引還有沒有出版的價值。他跟我講這個話，是因爲知道我曾經負責史語所漢籍自動化的工作。我告訴他，我認爲他們的成果有出版的價值。

我說的不是客氣話。據我所知，日本編纂的《魏書》索引，關鍵詞有十六萬個左右。扣掉與上海古籍出版社人名引得重複的部分，還是一部內容極豐富、極有價值的工具書。這樣的工具書，豈不是告訴我們該查什麼、有什麼可查的好指引嗎？

我舉上面兩個例子，想要表達一個看法。就是，對歷史研究而言，電子文獻資料庫是一個新的研究工具。但它並沒有取代傳統的研究工具。在很大的程度上，兩者的關係是互補的。這就是我對電子文獻資料庫的價值的基本看法。資料庫是一種威力強大的研究工具，但我們也不應過於誇大它的功用。我們應該了解它的限制，突破這些限制，使計算機的技術能爲人文研究作出更大的貢獻。

再來，個人想提出一些對電子文獻資料庫在本院與台灣發展前景的看法。就歷史學的觀點看來，資料庫是台灣學術界在發展研究工具上的一項重大貢獻，可以比喻成一個關鍵性的學術基本建設，引起了全世界漢學界的注意。不過，這似乎也是台灣學界在研究工具發展上的唯一主要貢獻。台灣沒有從事人文研究基礎建設的傳統。舉凡古籍的點校、註釋，文獻目錄、索引、年表、專門性辭書的編製，都很少有人從事。唯有書目的編纂，近年來比較多些。事實上，這方面的工作完全沒有得到制度上的重視，很難得到經費補助。在一九八〇年代以前，漢學工具書的製作，以日本爲大宗。八〇年代以後，中國大陸上的成果逐漸增加。台灣近年來發展電子文獻資料庫，可以說彌補了過往忽視研究工具的缺憾。無論從本土學術發展或爲國際學術社群盡責任的觀點看來，電子文獻資料庫都是很有價值，應當繼續從事的。

但是，在加速開發資料庫的同時，我們仍須注意品質的維持。資料庫要能爲學術研究作良好的服務，有其必要的條件。滿足這些條件，是資料庫開發過程的重要一環。

最後，如果個人的看法不錯，資料庫和傳統研究工具的關係確實主要是互補的，中研院資訊科學研究所謝清俊先生提出的漢學工作站的構想就很有意義。這個問題應該由謝先生自己來談，我不宜越俎代庖。不過簡單地說，我們如果能把重要的工具書——如類書、辭典、年表、文獻或資料索引——數位化，將之與文獻資料庫連結起來，就能在研究工具的發展上，得到進一步的突破。我們不但能很快查得資訊，而且能知道要查什麼，查到了什麼。個人是計算機技術的外行，不過最近有使用電子辭典的經驗，覺得電子辭典對學習語文的最大益處，還不在於查閱的速度，而是在資訊能夠連結、轉移，使學習的效率和品質得到極大的提昇。不過，個人手頭的一部日文電子辭典並不理想，主要原因是辭典本身品質欠佳。看來，任何資料庫的製作，內容和軟體工具都是同等重要的。

附記：本文原爲作者參加一九九八年五月一日本院舉辦的「人文計算研討會」的演講稿，現稍作文字修訂發表。