# 「廿五史全文資料庫」與中國歷史的研究

黃 清 連
中央研究院 歷史語言研究所

## 一、「廿五史全文資料庫」簡介

自一九八四年七月至一九九〇年六月，中央研究院歷史語言研究所與計算中心，合作開發了「廿五史全文資料庫」（以下簡稱「資料庫」），目前大致全部完成，並局部對外開外。它可以幫助學者在短短二、三分鐘至十分鐘內，檢索包含約四千萬字的廿五史中的任何字或詞。

這項歷時六年才完成的計劃，前後經過幾位主持人（謝清俊、陳克健、毛漢光、丁邦新、管東貴）的努力，以及上百位工作人員辛苦地輸入、校對、建檔和設計程式，總共花費四、五千萬元台幣，即將成爲輔助中國歷史研究的重要工具了。

「資料庫」輸入的資料，是根據標點本《廿五史》（北京中華書局版及台北鼎文書局翻印版）全文輸入，並經過四次以上的校閱。在校閱過程中偶而發現疑誤，也參考百衲本、武英殿本、汲古閣本等，作必要改正，每項改正，都在「資料庫」的校對說明註出。

「資料庫」包含的內容，自《史記》至《清史稿》（《新元史》除外）共二十五種史籍，全部輸入頁數共86,888頁，約四千萬字

。目前可以查詢約三千四百多萬字的正文，尚有若干「表」雖已鍵入文字，但仍未克服程式設計的困難，無法查詢。

「資料庫」是忠實地按照標點本逐頁輸入，在電腦螢幕或列印螢幕本文資料（詳下）上所見頁碼，與原書完全相同。各標點本正史所有的目錄及一些進表、凡例等，也都涵蓋。例如：標點本《舊五代史》書後所附〈進五代史表〉、〈編定舊五代史凡例〉、〈舊五代史鈔本題跋〉等等，無一遺漏。

中國歷史研究者最有興趣的，自然是想瞭解「資料庫」所提供的功能。就這一點來說，它大致可分四大項目：（一）自由翻閱，即可以根據目錄（如紀、志、表、傳等）、頁次、段落直接查閱本文；（二）自由詞檢索，即可以設定自由詞（free-text，不論單字或二個字以上的詞彙，如人名、地名、官名……或語助詞皆可），並由其組合查詢本文的相關段落；（三）列印，配合普通或系統印表機，可以選擇三種列印方式，即（1）全文列印：印出所查詢字、詞的整個段落，並有目錄標示出處、頁碼；（2）句子列印：查詢結果往往上百、成千筆，爲免久候，可以只列印查詢字、詞前後兩個標點符號［標點符號也可自行設定，如取自由詞前後兩個引號或驚嘆號等等］間的句子。這種列印結果會註出出處、頁碼，查詢者可以取原書核對；（3）製作卡片：查詢者可以利用特殊設計的印表紙［約21.5×10公分］列印出每一張在螢幕上所讀取的資料，包括書名、卷次、頁碼及所查詢字、詞的相關文字；（四）裁文，即利用顯示幕的游標裁切本文，配合卡系、檢索等相關功能使用。有關其他的使用功能，可參閱中央研究院計算中心編的《廿五史全文資料庫使用手册》（1990年7月修訂版）。

使用「資料庫」系統的相關設備，主要包括一部AT&T 3B

系列主機，一部終端機，一部中文字形產生器，及一部印表機。透過多工器的連線，IBM PC/AT，XT或任何相容個人電腦，還可以和3B主機連接使用。

## 二、「資料庫」如何輔助史學研究——若干實例

「資料庫」既然是全文輸入，它所提供的查詢功能，自然強過一般人工編輯的各種人名索引。只要鍵入正確的查詢字、詞，只要標點本正史確實有這項字、詞，查詢結果就一定不會遺漏，並可得出每筆查詢項目在某一種、數種或全部正史中所出現的次數，兼作統計之用。更重要的是，一般的人名索引除了有遺漏的可能外，僅具有人名查詢一項功能；但「資料庫」卻可以提供無數種選擇，讓使用者查詢任何字、詞。

舉例來說，裴松之爲《三國志》作注，徵引大批目前已亡佚的史料，他在注文引用史料之後，往往有「臣松之案」、「臣松之以爲」之類的案語，或對史料、史實考據，或對史事評述，值得作一番考訂、研究。如果要用人工遍查《三國志・注》，當然要花很長的時間，但在「資料庫」的自由詞檢索查詢欄上，只要鍵下「松之」兩字，並下指令在《三國志》上查詢，則僅花31秒時間，就可檢索863,469字，得出228項，259詞。這項查詢結果，對有關裴「松之」部份，都不會遺漏。得出結果後，螢幕上會列出這228項自由詞「松之」在《三國志》上的頁數，使用者可以直接在螢幕上逐項閱讀本文，或以上述三種方式之一列印結果。閱讀本文時，螢幕上出現的每個自由詞會以反白顯示，有助於立即辨認出其相關位置。

使用同一自由詞，也可同時檢索幾種史籍。譬如：要檢討「匈奴」與兩漢的關係，可以鍵入自由詞「匈奴」二字，先對《史記》、《漢書》、《後漢書》的記載，作一些基本資料的搜集，然後根據那些資料再抽絲剝繭，繼續研究。這項查詢，共費時261秒，檢索三種史籍共4,904,334字，得出1192項，2207詞。使用者可以隨各人喜好，或者在電腦上閱讀，或者將結果列印，將來再仔細過濾這批數量不少的材料。

三、五個或數十個自由詞，在「資料庫」檢索系統中，可以一次同時查詢。例如：使用者若想檢索唐代時期東北地區「靺鞨」、「渤海」、「奚」、「契丹」、「室韋」五個民族在兩《唐書》中的資料，可將上述五個自由詞同時輸入，在《舊唐書》與《新唐書》共4,675,833字（少數「表」未計）的資料中，費時235秒，得出869項，1141詞。這項查詢結果，是按照《舊唐書》頁1-5407、《新唐書》頁1-6472的順序排列，因此上述五項自由詞會在目錄中交叉出現。使用者如想讓每一項自由詞的查詢結果一起放置（在本例中，即指「渤海」一詞的資料全部依頁序排列，其他各詞亦同），可以按照鍵盤「改排序」指示，按下X鍵，瞬間即完成改排動作。

以上三個例子，在於說明使用「資料庫」檢索的幾種主要方式。其實，運用之妙，存乎一心。使用者如果能在蒐集論文資料之前，對一些關鍵字、詞有所瞭解，更可將它們在電腦中建檔，然後再利用檔案中的無數個詞彙，一次或分批去檢索「資料庫」。這樣，的確可以收到事半功倍的效果。

## 三、餘言

無庸置疑，「廿五史全文資料庫」提供給史學工作者的，是

一項便捷、功能強大的查詢工具。這項工具如何促進中國歷史研究的發展，端視使用者如何對它善加利用。使用者當然不可能只靠它的查詢結果，就寫出一篇精彩的論文。但論文撰述者在蒐集材料的過程中，卻可以藉「資料庫」之助，迅速查到相關資料，節省許多寶貴時間。

繼「廿五史全文資料庫」之後，中央研究院歷史語言研究所持續推動「史籍自動化計劃」，準備逐步建立先秦兩漢史籍、《十三經注疏》、《十通》等大型資料庫。至於一些小型資料庫，如漢簡、漢墓、「歷代名臣奏議索引」及「永樂大典所收典籍名稱及著者索引」等，有的已發展完成，有的正在建檔中。這種建立資料庫的趨勢，在日本和香港也有令人欣喜的發展，如：日本京都大學人文科學研究所的「東洋學研究資料庫」計劃、香港中文大學「漢及以前全部傳世文獻電腦化資料庫」計劃，目前都在進行中。

電腦硬體與軟體的發展日新月異，人類與電腦的關係也日益密切。在可預見的將來，人類利用電腦大量貯存資料與迅速分析資料的能力，勢必將學術研究推進嶄新的世紀。也許，在電腦影響、甚至主導一切知識的時代中，諸如「廿五史資料庫」以及其他許許多多可能發展出來的資料庫，必將提供史學工作者更大、更廣的發展空間。在這一個新時代中，應該是蘊育「新史學」的最佳契機。