

全國漢籍資料庫構建紀要

高田時雄

(京都大學人文科學研究所)

序言

全國漢籍資料庫計劃收錄日本公私圖書館所藏全部漢籍的收藏信息及其書籍數據，成為全國綜合目錄資料庫。1999年，在預定將京都大學人文科學研究所附屬東洋學文獻中心于翌年4月擴充改組成為漢字資訊研究中心（漢字情報研究センター）的同時，也將這項工作加入到迄今實施的各種工作當中，並計劃做為新的核心工作之一努力推進。

最初，東洋學文獻中心是基于日本學術會議提出的文獻中心的構想，以收集、整理東洋學相關文獻、資料，供研究者共同利用為主要目的，于1965年4月在京都大學人文科學研究所成立的機構。翌年，東京大學東洋文化研究所基于同樣的旨趣，設立了同名的機構。之後，這兩個機構做為東洋學相關的文獻中心，在保持互相合作的同時，也分居東西，開展各自的事業。例如，刊布收藏漢籍目錄、調查國內漢籍收藏情況、為全國圖書館從事漢籍整理的職員開辦學習班，等等。

京都的中心，正如漢字資訊中心這一名稱所表明的，進行了以漢語文獻為專門對象的改組。與此相對，東京的中心則將對象擴大到亞洲全體，改組成為東洋學研究資訊中心。基于改組，兩個中心踏上了性質相當不同的發展道路。但因為東京的新中心決定繼續進行與漢籍相關的部分工作，做為漢字資訊中心，也希望能與之合作，推進新的全國漢籍資料庫的工作。此外，與運營全國總合圖書目錄資料庫的學術資訊中心（學術情報センター，現國立資訊學研究所）進行協商也是必要的。

基于上述條件的全國資料庫施行研究的補助金獲得批准¹，已于2000年度進入具體的準備工作。用此科研費實施的工作，有以下三項：

- (1)對全國約600家公私圖書館進行問卷調查。
- (2)資料庫結構設計。
- (3)全國漢籍資料庫協商會的準備。

現在，逐一進行簡單的說明。

問卷調查

究竟日本公私圖書館中藏有多少漢籍？整理情況如何？漢籍目錄是否已經以冊子的形式刊布？或者有沒有卡片式目錄？資料庫的工作進展如何等等基本現狀如果不能把握，實現全國漢籍資料庫的戰略就無法成功。因此，對全國約600家圖書館進行了以上各項目的問卷調查²，約三分之一的圖書館發回了回答。其中，對“有無所藏漢籍的目錄”的問題回答“有”的圖書館有87家，而且其中半數（43家）已經以冊子的形式刊布了目錄。另一方面，沒有漢籍目錄的圖書館對“今後是否準備制作目錄”的問題，回答“沒有計劃”或沒有回答的有111家，回答“有計劃”的不過10家。總之，收藏漢籍的圖書館有不少已經完成了冊子目錄或卡片目錄，沒有目錄而今後準備制作目錄的圖書館則很少。鑒于此，我們認為，以已經作成的目錄做為基礎構建資料庫是最有效率的。如果是在沒有目錄的情況下開始工作，就不得

¹ 2000年度（2000年4月～2001年3月）科學研究費補助金〔基盤研究C（企畫調查）〕「展望全國漢籍資料庫的實現」代表者：高田時雄；分擔者：丘山新、宮島博史（以上東京大學）、花登正宏（東北大學）、井波陵一、安岡孝一、梶浦晉（以上京都大學）。

² 問卷見 <http://kanji.zinbun.kyoto-u.ac.jp/kanseikiyogikai/enq.pdf>；結果見 <http://kanji.zinbun.kyoto-u.ac.jp/kanseikiyogikai/enq.html>。

不先展開調查整理，這是非常困難的。就目錄而言，卡片目錄相對於冊子目錄也是比較難于處理的。如果是冊子目錄，比較容易得到，即如非賣品，入手不易，但復印並不困難。當然，如果得到所藏館的合作，卡片目錄也可以和冊子目錄一樣復制，做為錄入的基礎並非不可能，但恐怕會稍顯麻煩。因此，大致準備從冊子目錄著手。單就回答問卷的圖書館而言，已經有43家刊布了冊子目錄。進而通過各種資訊得知的日本公私圖書館所藏漢籍的冊子目錄的數量會更多。特別是日本有代表性的漢籍收藏機構幾乎都刊布了冊子目錄，如果暫且可以將這些總括成為資料庫，其便利性即不可估量。且僅此項工作，也已佔滿當前的工作量。

資料庫結構

為了實際構建資料庫，欄位的基本設計是必要的。如果不能盡早提出欄位設計的試行方案，其後的變更即便不是不可能，也顯然會在勞力和經費等方面造成困難。此外，既然是以冊子目錄為基礎推進資料庫的構建，那麼也要力求與迄今刊布的目錄的形式不至有太大的差別，因為這樣將冊子目錄委托外部公司輸入的時候會比較方便。因此即便稍顯自滿，我們還是以『京都大學人文科學研究所所藏漢籍目錄』（1979—1980年）為基礎制作了資料庫的結構。經上記科研費的成員討論，已經制成「輸入欄位一覽」³及「輸入例」⁴。

以下對各欄位進行簡單的說明，並舉出應注意的地方。

首先是「編號」(nu)，用適當的號碼即可，沒有特別的限制。

「部」(fi)「類」(sf)「屬」(tg)「目」(ki)等，是中國傳統的四部分類的框架。做為基礎的目錄如果是按照傳統的漢籍目錄的體例作成的，就會按照經史子集的順序排列，並進而按照各自下面的子類排列，委托輸入的時候給予適當的指示，應可以正確錄入。

「書名 卷數」組合成一個欄位 (ti)。中國古書的卷數與書籍的物理實態乖離的情況很多，和書名合在一起，也不會混亂。這樣也可避免和冊數混同。但為了僅用書名也能檢索，書名和卷數之間用逗號分開。這樣，在這個欄位中，就會記入「史記，一百三十卷」這樣的數據。冊數因版本不同而不同，所以設置了單獨的欄位「冊數」(vi)。中國書籍多有別稱和附錄，也需標示。前者做為「即」書名(pt)欄位，後者做為「附」書名(st)欄位，與(ti)分別標示。

「撰者名」(au)也是如此。中國書籍不僅要標示真正意義上的撰者，其他各種表示也是必要的。因為有的書籍採用本文和注釋合為一體的夾注形式，有的則附有附錄，這些部分的撰者也不能不標示出來。比如「史記」的某個版本，著錄有「漢，司馬遷，撰」，「劉宋，裴駰，集解」，「唐，司馬貞，索隱」，「唐，張守節，正義」，隨著著錄項目的增加，應給每個增加的欄位加上號碼，設置成 au, au2, au3, au4 這樣的欄位。用逗號隔開著者名，是為了可以單獨抽出著者名進行檢索。

「刊年」(yr)採用年號紀年，象「康熙七年」、「民國二十三年」這樣用漢字數字表示。位數也要正確標記，不能用「民國二三」年這樣的略式。年次不明的情況下，以「嘉靖中」這樣的形式表示。以西歷紀年的情況也可能有，此時也要象「一九六四年」這樣用漢字四位數字表示。

「出版主體」(pb)可能是人名、出版社名或者是地名。如「康熙二十七年仁和邵氏刊本」中的「任何邵氏」。如果是「嘉靖中福建刊本」，就應著錄地名「福建」。

「版本」(ed)多數情況下就是「刊本」，但「抄本」、「排印本」、「景照」、「景印」等也可能有。附錄書名的(yr)(pb)(ed)各項應在(syr)(apb)(sed)欄位中予以著錄。

「藏板」(sd)不一定都有，象「康熙三十一年自序刊本未學齋藏板」這樣的情況，應著錄「未學齋」。

³ <http://kanji.zinbun.kyoto-u.ac.jp/kansekiyogikai/field.pdf>

⁴ <http://kanji.zinbun.kyoto-u.ac.jp/kansekiyogikai/example.pdf>

漢籍的特徵之一，是叢書、叢刻甚多。叢書中還包含有叢書的情況也不少。為了把叢書的層級構造在資料庫中表現出來，有必要設置表示“父子關係”的欄位。如將某書籍設為「子」，含有該書籍的叢書即為「父」。例如『十三經注疏』中收有「周易兼義九卷」，登錄「周易兼義」的數據的時候，父編號(oy)欄位中應記入『十三經注疏』的「編號」(nu)。相反，『十三經注疏』的數據中，子編號(ko)欄位中應記入「周易兼義」的「編號」(nu)。這樣明確標記，兩者的父子關係在資料庫中即可正確反映。當然『十三經注疏』中不只有「周易兼義」，還有很多其他書籍，因此(ko)欄位也必須象 ko2, ko3, ko4... 這樣增加。此外，某叢書、叢刻包含在別的大叢書的情況下，該叢書的數據中父編號和子編號都是必要的。儘管有些麻煩，以這種結構可以表現無窮的層級結構，還是有利的。這也可以說是其他資料庫未見的特色吧。

最後是「所藏機關」(or)。應所藏機關錄入各種各樣編號的需要，設置了「登錄號」(rn)、「請求號」(si)、「文庫名稱」(se)、「文庫代碼」(sn)等欄位。當然，也設置了「冊數」(vi)。(rn)以下各項只在必要的情況下使用，不是必須登錄的項目。此外，為以上欄位無法著錄又必須說明的事項，設置了「注記」(no)欄位。在此，可記錄卷數原來的缺失、缺某卷、補配、補抄、識語、圖記等各種情況。這些會全部成為關鍵字檢索的對象。

全國漢籍資料庫協商會

建立資料庫的主要工作當然是由新漢字資訊研究中心承擔的，但為構建全國規模的資料庫有必要進行周密的聯合調整。因此設置了做為聯合組織的全國漢籍資料庫協商會，其第一次總會，于 2001 年 3 月 9 日，在學術綜合中心的一橋紀年講堂舉行。

事務機關，目前，漢字資訊研究中心以外，還有東京大學東洋文化研究所附屬東洋學研究情報中心及國立資訊學研究所。事務局設置在漢字資訊研究中心。如前所述，東大的東洋學研究情報中心的前身和京大的漢字資訊研究中心同樣是東洋學文獻中心，在漢籍的搜集和整理等方面，共同起到核心作用。從漢籍的收藏量上說，兩者均為國內最大規模的圖書館，想必會成為資料庫的中核。此外，資訊學研究所（前學術資訊中心）是日本當前學術資訊網絡建構的中心，運營著大規模的全國總合目錄資料庫（WEBCAT）。全國漢籍資料庫如果不與之合作，當然無法有效進行。因此，資訊學研究所做為協商會的事務機關參加本項目，具有重大的意義。

協商會每年 3 月召開總會，迄今已經舉行三次⁵。通過這一機會，與報告資料庫進展情況同時，也計劃交換與漢籍資料庫有關的各種各樣的資訊。國際合作也做為今後的課題加以考慮，今年 3 月召開的第三次總會即有韓國漢城大學奎章閣的客人參加。

協商會不採用會員制，以自由參加總會的形式運作。資料庫的加入，則基于事務局（即漢字資訊研究中心）與各個圖書館之間的個別協商另外進行。因此，總會與資料庫的加入未必有直接關係，而是希望通過總會獲得相關的資訊。總會的參加者是全國的圖書館館員以及關心漢籍的研究者，歷年參加人數平均 100 人左右。

進展情況

經過上述準備，終於于 2001 年度開始資料庫的實際構建工作。因為幸運地得到了科學研究費成果公開費的補助，所以可以從京大人文科學研究所和東大東洋文化研究所的漢籍開始著手工作。東洋文化研究所的漢籍已經先行資料庫化，但從此著手，可以比較容易和低成本地完成工作的期望未能實現。因為做為東大資料庫基礎的『東京大學東洋文化研究所漢籍分類目錄』（1973-75）是沿襲『京都大學人文科學研究所漢籍分類目錄』（1963-65）體例的分類目錄，其目錄中列記的書名沒有反映出書籍的物理的情況，而將叢書、叢刻的子目置于四

⁵協商會的主頁（<http://kanji.zinbun.kyoto-u.ac.jp/kansekiyogikai/>）載有會議記錄，請參考。

部分類中的適當位置。因此以此目錄做為基礎，就很難在資料庫中記錄書籍的層級關係。將父子關係逐一插入原來數據中的工作難度超出預想，儘管第一年度未能完成，成為今後長期的工作，但現在的情況已大有改善。

資料庫的編制順序大致分成兩個階段。首先是錄入的工作，這主要委托公司完成⁶，然後獲得經過三次校正的數據。委托時在目錄或其復印件中附以錄入說明書是必要的。因為各目錄的記錄方法未必相同，什麼部分錄入什麼欄位必須逐一說明。在第二階段對得到的數據由包括編制委員會的成員⁷在內的人員進行更嚴格的校正。在這個階段會發現很多想不到的錯誤。此外，科學研究費的章程規定各年度未必須公開該年度的研究成果，此時也往往會出現未經過充分校正的數據。這些數據還會繼續進行訂正，望予理解。

資料庫設計的技術工作全部由漢字資訊研究中心的安岡孝一承擔。檢索引擎採用 Open Text，可以實現高速檢索。詳情請參同氏的「全國漢籍資料庫的設計和 WWW 的運用」⁸。技術設計完成後，全國漢籍資料庫做為 2001 年度科學研究費的成果，終於于 2002 年 3 月進入公開運作⁹。

本年度（2003 年度），由科學研究費支持的資料庫構建進入第三年¹⁰，目前全國漢籍資料庫採用了以下目錄：

京都大學人文科學研究所
東京大學東洋文化研究所
京都產業大學（小川文庫）
鹿兒島大學（玉裡文庫、岩元文庫、松本文庫）
立命館大學（各部局、及詞學文庫、高木文庫）
滋賀大學（教育學部）
高知大學（小島文庫）
東北大學
新潟大學
神戶市外國語大學
廣島大學（斯波文庫）
實踐女子大學（山岸文庫）
千葉縣立中央圖書館
三康圖書館
宮內廳書陵部
大阪府立中之島圖書館
東洋文庫

此外，現在正在錄入的目錄如下（本年度或今後一兩年中加入資料庫）：

九州大學（附屬圖書館及六本松分館）
一橋大學

⁶因日本國內既有成本問題也有字體的錄入問題，所以現在委托臺灣某公司進行。工作組富於能力，隨著經驗的增多，現在的工作令人滿意。

⁷漢字資訊研究中心的數據庫編制組也是科學研究經費的申請主體。

⁸ <http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/publications/2002-11-19.pdf>

⁹ <http://www.kanji.zinbun.kyoto-u.ac.jp/kanseki/>

¹⁰此前科學研究費的補助情況如下：平成 13 年度（2001）科學研究費補助金（研究成果公開促進費）課題號：138128，補助金額：13,500,000 日圓；平成 14 年度（2002）科學研究費補助金（研究成果公開促進費），課題號：148046，補助金額：21,900,000 日圓；平成 15 年度（2003）科學研究費補助金（研究成果公開促進費），課題號：158033，補助金額：20,500,000 日圓。

東京都立中央圖書館
關西大學（內藤文庫、增田文庫、泊園文庫）

並且計劃明年開始錄入國立國會圖書館所藏漢籍的資料庫，現正在協商。

資料庫構建開始以來近三年的發展，工作組成員的努力自不待言，也是全國收藏漢籍圖書館積極合作的結果。萬分感謝，今後還希望得到更多的支持。另外，在這裡還想強調，此項工作能夠這樣迅速地進展，與全日本圖書館界和學術界對漢籍的強烈關心是分不開的。

今後的課題

目前的錄入從刊行的冊子目錄開始的大方針不變，但也要盡快考慮如何處理卡片目錄和未採用漢籍目錄體例的目錄（例如按十進制排列的目錄）卡片和冊子相比，作成錄入用的原稿比較麻煩，之後的工作則大致相同。現在，已經開始嘗試錄入少數卡片目錄。

卡片目錄往往未以四部分類。按十進制分類法或其他分類法排列的目錄也有同樣的問題。將這些目錄所載漢籍逐一加上四部分類從勞力上說是幾乎不可能的。因而這些目錄（包括卡片目錄）錄入的時候不得不缺記四部分類。現在已經有這樣的數據。這雖與漢籍資料庫本來的旨趣相反，但如果只考慮檢索的方便，還是可以充分利用的。此後補入分類也並非不可能。如何處理這些數據是今後的課題。

此外，根據情況，將完全未加整理的漢籍從頭開始錄入數據也不是不可能的。此時，根據既定的欄位錄入，用 UTF8 編碼的文本文件保存的話，可以立刻組合進資料庫。如果資料庫中已有同樣的書籍的數據存在，給錄入的數據加入必要的變更也是可以做到的。但是最簡單的還是使用編制委員會開發的免費軟體「漢籍編輯器」¹¹。這個軟體運行在 Windows2000 以上版本的操作系統上，將必要的項目逐一錄入，即可成為遵循全國漢籍資料庫格式的數據。將數據送交編制委員會，即可在資料庫中顯示這些數據。

這個軟體已經用于漢籍管理職員學習班（文部科學省和京都大學人文科學研究所附屬漢字資訊中心共同主辦）的實習。

為了進行書誌學的研究，能夠看到各個版本的書影當然是最理想的。為此，已開始將人文科學研究所所藏圖書中對於書誌學比較重要的書籍的書影圖像漸次揭載。全部數據都加入圖像是不可行的，但力爭在既定的條件下盡可能多的揭載圖像。但是圖像數據的作成沒有保存機關的全面合作是不行的，今後能在多大程度上擴大規模還是未知數。

國立資訊學研究所運作的 WEBCAT 做為全國書籍資料庫，利用的頻度是很高的。全國漢籍資料庫和 WEBCAT 之間如果可以以某些形式互相切入，就會更加便利。技術可能性的探討，做為資訊學研究所和漢字資訊研究中心間的共同研究已于 2002 年進行。基本的問題點大致明確，得出了很大部分可以互相切入的結論。但付諸實施，似乎還要再花些時間。

（2003 年 9 月）

¹¹ "kanseki edi tor1.01"。可從資料庫協會的主頁下載，附有安裝說明。