全國漢籍データベース構築覺書

高田時雄
（京都大學人文科學研究所）

はじめに

全國漢籍データベース（日本所藏中文古籍數據庫）は、日本の公私圖書館に所藏されるすべての漢籍の所在情報とその書誌データを総合した全國的・網羅的な目録データベースとして構想された。1999 年、京都大學人文科學研究所附属東洋學文獻センターが、翌 2000 年 4 月に漢字情報研究センターとして擴充改組されることが豫定されていたのに伴い、これまで行ってきた様々な事業に加えて、これを新しい中核的事業の一つとして取り組もうと考えたのである。

もともと東洋學文獻センターは、日本學術會議が打ち出したドキュメンテーション・センター構想に基づき、東洋學に關する文獻・資料を收集・整理し、研究者の共同利用に供することを主目的として、1965 年 4 月に京都大學人文科學研究所に設置されたものであった。翌 1966 年には東京大學東洋文化研究所にも同じ趣旨のもとに同名のセンターが設けられた。以來、この二つの施設は東洋學に關する文獻センターとして、相互に連携を保ちつつも東西に棲み分けるかたちで各々の事業を展開してきた。所藏の漢籍目録を公刊し、國内の漢籍所在調査を行い、また漢籍整理に從事する全國の圖書館職員のために講習會を開催するなどがその例である。

京都のセンターは漢字情報研究センターという名稱に明らかなように、漢字文獻に特化した發展的改組を行ったのに對して、東京のセンターのほうは廣くアジア全域にその對象を擴大すべく東洋學研究情報センターに改組された。二つのセンターは改組によってかなり性格の異なる道を歩み始めたわけであるが、東京の新センターではまた漢籍に關するこれまでの事業の一部を引き繼ぐことが決定されていたので、漢字情報研究センターとしては、新しい全國漢籍データベースの事業も同センターと連携しつつ進めていくことが望ましいと考えた。また日本における全國的な総合圖書目録データベースを運營している學術情報センター（現國立情報學研究所）とも協議する必要があった。

そういった條件を踏まえつつ、全國データベースの實現を模索するための助成金が認められ[1]、具体的な準備作業に入った。この科研費によって行ったことは、以下の三項目である。

(1)全國約 600 の公私圖書館に對するアンケート調査

(2)データベースのフォーマット設計

---

[1] 2000 年度（2000 年 4 月～2001 年 3 月）科學研究費助成金［基盤研究 C（企畫調査）］「全國漢籍データベースの實現にむけて」代表者：高田時雄、分擔者：丘山新、宮嶌博史（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）。

(3)全國漢籍データベース協議會の準備

今、それらにつき逐一簡單な説明を行っておきたい。

アンケート

一體、日本の公私圖書館にはどれだけの數量の漢籍が存在するのか、またその整理状況はどうか。漢籍目録がすでに冊子として公刊されているか、あるいはカード式の目録があるか。さらにデータベースへの取り組みはどうか等、基本的な現状把握を行わなければ、全國漢籍データベースの實現に向けた戰略の立てようがない。そこで、全國約 600 の圖書館に對して以上のような諸項目につきアンケート調査を行ったところ[2]、約三分の一の圖書館から回答が寄せられた。そのうち漢籍所藏目録の有無に對して「有」と回答されたものは 87 館、更にその約半數 43 館がすでに冊子目録を公刊していることが判明した。一方、漢籍目録が「無」の圖書館に對して、「今後の作成豫定」を尋ねたところ、豫定無し或いは無回答が 111 館で、「ある」という回答は 10 館に過ぎなかった。要するに漢籍を所藏している圖書館ではかなりな數がすでに冊子目録乃至カード目録を作成濟みであって、今後新たに目録を作成しようとしている圖書館は稀であるということが分かる。この現況を考えると、すでに作成されて存在する目録を基礎にしてデータベースの構築を行うことが最も効率的であることが豫想された。目録が備わっていない状態から出發するとすれば、先ず調査整理から始めなければならず、これは極めて困難な作業である。さらに目録といってもカードの場合は冊子目録にくらべて取り扱いが厄介である。冊子目録であれば、入手も比較的簡單であり、又たとえ非賣品で入手し難い場合も複寫することは困難ではない。もちろんカード目録の場合も、所藏館の協力があれば冊子目録同樣にコピーを作って、それを入力の基礎とすることは不可能ではないが、少し餘計に手間が掛かると思われる。そんなわけで、大まかなやり方としては、冊子目録の存在するものから手を付けようということになった。アンケートに回答の寄せられた圖書館だけに就いて言っても、43 館が冊子目録を公刊しているのである。さらに樣々な情報源から知り得る日本の公私圖書館所藏の漢籍目録冊子の數はそれよりずっと多いことが豫想される。特に日本の代表的な漢籍所藏機關はほとんど冊子目録を公刊していることが判っており、とりあえずはこれらだけでもデータベースとして総合することが出來れば、その利便性は計り知れないであろうし、それだけでも當面の仕事量を滿たすには充分と考えられた。

フォーマット

實際にデータベースを構築するためには、どのようなフィールドをたてるかといった基本設計が必要となる。少なくとも早い時期に試案の形でこれを提示しておかなければ、後で變更を加えることは不可能でないにしても、勞力や經費の面で難しい問題が生じること

---

[2]このアンケートの調査票は http://kanji.zinbun.kyoto-u.ac.jp/kansekikyogikai/enq.pdf、またその結果は http://kanji.zinbun.kyoto-u.ac.jp/kansekikyogikai/enq.html に見える。

は明らかである。また冊子目録を基礎としてデータベース構築を進める以上、これまで公刊されて來た目録の形式から著しく遊離しない配慮も求められる。冊子をそのまま外注に出して入力を進められるという便宜を得られるからである。そこで些か手前味噌の感はあるが、『京都大學人文科學研究所所藏漢籍目録』（昭和 54・55 年）をベースにしてフォーマットの作成に取りかかることとした。その結果、上記科研費のスタッフによる討論を經て、「入力フィールド一覧」[3]及び「入力例」[4]が作成された。

そのフィールドの各項目について以下に簡單に説明を加えながら、注意すべき點を擧げておこう。

まず「編號」(nu)であるが、これは適當な番號を用いればよく、特に制限はない。「部」(fi)「類」(sf)「屬」(tg)「目」(ki)などは、中國の傳統的四部分類の枠組みである。基礎となる目録が傳統的な漢籍目録の體例で作成してあれば、經史子集の順に列んでおり、更にそれぞれ下位の分類に從って排列されているから、發注時に適切な指示を與えることで、間違いなく入力されてくる。

「書名・卷數」を一つのフィールド(ti)に統合してある。古い中國書の卷數というのは、書物の物理的な實態と乖離していることが多く、書名と一體のものとして扱っておいても困らない。またこうすることで册數との混同を避けることも出來る。ただし書名だけで檢索が可能となるように、間をコンマで區切る。從ってこのフィールドには「史記,一百三十卷」と書くことになる。册數のほうはエディションによって様々で、これには別のフィールド「册數」(vi)が用意される。ただし中國の書物には書物の別稱を掲げていたり、附録がついていたりすることが多い。これらの書名も出しておく必要があるが、これは前者を「即」書名(pt)とし、後者を「坿」書名(st)として(ti)とは別に出す。

「撰者名」(au)も、中國の書物には正しい意味での author だけでなく、様々な表示が必要になってくる。これは本文と注釋とが一體になっているような形式があったり、附録が付いていたりして、それらをすべて出しておかねばならないからである。例えば「史記」の或るエディションの場合、「漢,司馬遷,撰」「劉宋,裴¥epsfkanji{99F0.eps},集解」「唐,司馬貞,索隱」「唐,張守節,正義」を書き込み、それぞれ數が増えれば増えるごとに番號をつけて au, au2, au3, au4 のようにしておけばよい。コンマで著者名を圍むのは、それだけ取り出して檢索できるようにするためである。

「刊年」(yr)は元號を用い、「康熙七年」「民國二十三年」のように漢數字を用いて書く。桁數まで正しく表記し、「民國二三」年はような略式は不可である。年次が不明な場合は「嘉靖中」のようにする。西暦を書くことも可能であるが、その場合も「一九六四年」のように漢字 4 桁で表記する。

「出版主體」(pb)は人名であったり、書肆名であったり、地名であったりする。たとえば「康熙二十七年仁和邵氏刊本」と目録に書かれてあるような場合の「仁和邵氏」がこれに

---

[3] http://kanji.zinbun.kyoto-u.ac.jp/kansekikyogikai/field.pdf
[4] http://kanji.zinbun.kyoto-u.ac.jp/kansekikyogikai/example.pdf

當たる。「嘉靖中福建刊本」とあれば、地名の「福建」を入れておくことになる。

「エディション」(ed)というのは、多くの場合「刊本」であるが、「鈔本」「排印本」「景照」「景印」などもあり得る。(yr)(pb)(ed)については、附録書名についても(syr)(apb)(sed)として同様に記述しておく必要がある。

「藏板」(sd)は必ずしもあるとは限らないが、例えば「康熙三十一年自序刊本未學齋藏板」とあった場合、「未學齋」と書いておくことになる。

さて漢籍に特徴的なことの一つは、叢書や叢刻の多いことである。叢書の中に更に叢書が含まれるといったことも珍しくない。そこで叢書の階層構造をデータベースで表現するために、親子關係を記したフィールドを立てる必要が起こってくる。ある書物が「子」とすると、その書物が含まれる叢書を「親」と考えるわけである。例えば『十三經注疏』にはまず「周易兼義九卷」が收められているが、「周易兼義」のデータを採録する場合、親番號(oy)フィールドには『十三經注疏』の「編號」(nu)を記入しておくわけである。反對に『十三經注疏』のデータには、子番號(ko)フィールドに「周易兼義」の「編號」(nu)を記入する。こうしておけば兩者の親子關係がはっきりと明記されているので、データベース上に正しく反映させることが出來る。もちろん『十三經注疏』には「周易兼義」のみならず多數の書物が含まれているので、(ko)フィールドは ko2,ko3,ko4...という風にその數だけ增やさねばならない。またある叢書や叢刻が別の大きな叢書の中に含まれているような場合には、當該叢書のデータには親番號、子番號の兩方が必要になる。多少面倒くさいようだが、この仕組みによって、何段でも階層構造を表現できる利點がある。これは他のデータベースには見られない特色と言えよう。

最後に「所藏機關」(or)のフィールドがある。さらに所藏機關によって様々な番號を入れておきたいという要望に答えるため、「登録番號」(rn)、「請求番號」(si)、「文庫名稱」(se)、「文庫代碼」(sn)を用意した。當然「册數」(vi)も用意してある。(rn)以下は必要な場合のみ使用するもので、必須のものではない。また以上のフィールドのどこにも書き込めない情報で、是非とも採録しておきたいという事項については、「注記」(no)のフィールドを設けた。ここには、卷數の原闕や闕〇卷、補配、補鈔、識語、圖記といった様々な事柄を書いていただけばよい。すべてキーワード檢索の對象となる。

全國漢籍データベース協議會

データベース作成の實務を主として新しい漢字情報研究センターが擔うことは成り行き上當然として、全國規模のデータベースの構築には周到な連絡調整が必要となることが豫想された。そこで連絡組織としての全國漢籍データベース協議會を設置することとし、その第一回總會を、2001 年 3 月 9 日、學術総合センターの一橋記念講堂で開催した。

幹事機關は當面、漢字情報研究センターのほか、東京大學東洋文化研究所附属東洋學研究情報センター及び國立情報學研究所がこれに任じ、事務局を漢字情報研究センターに置くこととした。前述のように、東大の東洋學研究情報センターの前身は京大の漢字情報研

究センターと同じく、東洋學文獻センターであって、ともに漢籍の蒐集・整理に關して國内の中心的役割を果たしてきたという經緯がある。また漢籍收藏量からいっても兩者がともに國内最大規模の圖書館に數えられることは明らかで、データベースの中核を形成することも想定されていた。また情報學研究所（舊學術情報センター）は現時點に於いて日本の學術情報ネットワーク構築の中心的存在であり、大規模な全國總合目録データベース（WEBCAT）を運用している。全國漢籍データベースもこれと連携することなしには効果的な運用は難しいことは自明である。したがって協議會の幹事機關として情報學研究所の參加を得たことは大きな意味を有するものであった。

協議會は毎年一度 3 月に總會を行っており、これまで三回が開催された[5]。この機會を通じて、データベースの進捗状況の報告を行うとともに、漢籍データベースに關連するさまざまな情報交換が圖られている。國際的な連携も今後の課題として視野に入れており、今年 3 月に開かれた第三回總會では韓國ソウル大學の奎章閣からのゲスト出席もあった。

協議會は會員制を採っておらず、總會への自由參加のかたちで運營されている。一方、データベースへの參加は、箇々の圖書館と事務局（すなわち漢字情報研究センター）との個別の折衝に基づいて別途行われる。したがって總會は必ずしもデータベースへの參加とは直接の關係を持たないといってもよいが、總會の場を通じて關連情報獲得の機會にしていただきたいと考えている。ちなみに出席者は全國の圖書館職員および漢籍に關心を有する研究者で、出席者數は例年平均して 100 名程度である。

進捗状況

上で述べてきたような準備を經て、いよいよ實際のデータベース構築に取りかかったのは、2001 年度であった。幸いにして科學研究費成果公開費の助成を得ることが出來たので、京大人文科學研究所と東大東洋文化研究所の漢籍から取りかかった。東洋文化研究所の漢籍については、すでに先行してデータベース化が進んでおり、そのデータに手を入れれば、比較的容易かつ安價に仕上がると考えていた目算は見事に當てがはずれることとなる。東大のデータベースが基礎とした『東京大學東洋文化研究所漢籍分類目録』(1973-75) は、『京都大學人文科學研究所漢籍分類目録』（1963-65）の體例を踏襲した分類目録であって、その目録に列記された書名は物理的な書物のありようを反映していないのである。つまり叢書や叢刻の子目を抜き出して四部分類の然るべき場所に置いてあるために、これを基礎にするとデータベース中で各書物の階層關係を記述することが難しい。元のデータに對して親子關係を一々挿入していく仕事は豫想外に難航し、初年度ではとても完成せず、かなり長期にわたって尾を引くこととなったが、現在ではかなり改善されている。

さてデータベースの作成手順は大きく二段階に分かれる。先ず入力作業であるが、これ

---

[5] 協議會のホームページは http://kanji.zinbun.kyoto-u.ac.jp/kansekikyogikai/。そこに會議の記録が載せられているので參照されたい。

は民間企業に委託しており[6]、そこから三回の校正を經たデータが納入される。委託時には目録の現物（或いはコピー）に入力指示書を付ける必要がある。各目録の記述の仕方は必ずしも同じでないために、どの部分をどのフィールドに入れるかを一々指示しなければならないのである。第二段階では納入されたデータに對して、作成委員會[7]のスタッフを含めた人員によってさらに本格的な校正を行っている。この段階でも思いがけない誤りの見付かる場合がかなりある。また科學研究費の規程により各年度末には當該年度の成果を公開しなければならないために、充分な校正を經ないままのデータが反映されていることも間々あるが、これは繼續して訂正を行っているので、ご理解を得たいと思う。

データベース設計の技術的側面はすべて漢字情報研究センターの安岡孝一が擔當した。檢索エンジンとして、OpenText を採用し、高速な檢索が實現されている。その詳細は同氏の「全國漢籍データベースの設計と WWW での運用」[8]に述べられているので參照されたい。結果、全國漢籍データベースは 2001 年度科學研究費の成果として、2002 年 3 月遂に公開の運びとなった[9]。

本年度（2003 年度）、科學研究費によるデータベース構築は三年目を迎えているが[10]、現時點で全國漢籍データベースは以下の目録データを採録している。

京都大學人文科學研究所
東京大學東洋文化研究所
京都産業大學（小川文庫）
鹿兒島大學（玉里文庫、岩元文庫、松本文庫）
立命館大學（各部局、及び別に詞學文庫、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
神戸市外國語大學
廣島大學（斯波文庫）
實踐女子大學（山岸文庫）

---

6 日本國内ではコストもさることながら、正字體の入力に問題があるため、台灣の某社に依頼している。幸い有能なスタッフに惠まれたことと、經驗量の增大によって、現状ではほぼ滿足すべき状況にある。

7 漢字情報研究センターに設置されたデータベース作成の實行部隊で、科學研究費の申請主體でもある。

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 これまでの科學研究費による補助は以下の通り。平成 13 年度(2001)科學研究費補助金（研究成果公開促進費）、課題番號 138128、補助金額 13,500,000 圓；平成 14 年度(2002)科學研究費補助金（研究成果公開促進費）、課題番號 148046、補助金額 21,900,000 圓；平成 15 年度(2001)科學研究費補助金（研究成果公開促進費）、課題番號 158033、補助金額 20,500,000 圓。

千葉縣立中央圖書館
三康圖書館
宮内廳書陵部
大阪府立中之島圖書館
東洋文庫

また現在入力作業中のものに以下があり、今年度内或いは一兩年中に総合が豫定されている。

九州大學（附属圖書館及び六本松分館）
一橋大學
東京都立中央圖書館
關西大學（内藤文庫、増田文庫、泊園文庫）

さらに來年度は國立國會圖書館所藏漢籍のデータ入力を開始すべく、現在折衝中である。

データベース構築開始後、三年未滿のうちにこれだけの發展を見たことは、スタッフの努力は言わずもがな、全國の漢籍所藏圖書館の積極的な協力があってのことである。實に感謝に堪えない。今後一層のご支援をお願いしたい。また一方で、これほどの速やかな進展を得られたのには、廣く日本の圖書館界及び學術界に漢籍に對する竝々ならぬ關心が存在するためと考えられる。あわせてこの點を是非強調しておきたいと思う。

今後の課題

當面の入力は册子目録の刊行されているものから始めるという大方針は變わらないとして、一方でカード式目録の場合や、また漢籍目録の體例を採っていない目録の場合、例えば十進分類法で排列された目録をどうするかという問題もやがて考えなければならない。カードは册子に比べて入力用原稿の作成に手間がかかるが、それ以後の作業は基本的に同じである。現在、少數ではあるが、一部試行的にカード目録からのデータ入力を始めている。

ただカードの場合は往々にしてそうであるが、四部分類がされていない。十進分類法あるいはその他の分類で排列された目録についても同様であるが、これらの目録に採録された漢籍に一々四部分類を付け加えて入力することは勞力の上から言ってほとんど不可能である。したがって（カード目録をふくめ）これらの目録中の書物をデータベースに取り込む場合は四部分類を缺いたままで総合せざるを得ない。現在既にそういったデータが存在する。これは漢籍データベースの本來の趣旨に反するが、單なる檢索の便宜だけを考えれば、充分に使用に耐え得る。分類は後から補うことも不可能ではない。それらをどうするかは今後の課題である。

また全く整理がなされていない漢籍を一からデータ入力することも場合によってはあり得るであろう。その時は既定のフォーマットにしたがって入力し、ユニコード（UTF-8）のテキストファイルで保存すれば、データベースへの総合が即時可能である。すでにデー

タベース中に同一の書物が存在する場合は、そのデータを用いて必要な變更を加えるという方法もあり得る。しかし最も簡便なのは作成委員會で開發したフリーウェアの「漢籍エディタ」を用いることである[11]。これは、Windows2000 以上の OS で動き、必要な項目を逐一入力していけば、そのまま全國漢籍データベースのフォーマットに従ったデータとなる。そのデータを作成委員会に送っていただければ、そのまま全國漢籍データベースに反映される。

このソフトは、既に漢籍擔當職員講習會（文部科學省・京都大學人文科學研究所附屬漢字情報研究センター共同開催）における實習にも用いている。

書誌學的な研究のためには、それぞれの版本の書影が見られることが理想的であるのは言うまでもない。そのために人文科學研究所の所藏圖書のうち書誌學的に重要と思われるものから順次、書影畫像を掲載していくことを既に始めている。すべてのデータに畫像を加えることは不可能だが、與えられた條件の下で出來る限り多數の畫像を載せていきたい。しかし畫像データの作成には所藏機關の全面的協力がなくては出來ないので、今後どの程度まで擴大していけるかは未知數である。

國立情報學研究所の運用している WEBCAT は、全國書誌データベースとして利用頻度の非常に高いものである。全國漢籍データベースと WEBCAT との間に何らかのかたちでの相互乘り入れが可能となれば、利便性は一層向上する。技術的にそれが可能かどうかを探る試みが、情報學研究所と漢字情報研究センターとの間の共同研究として 2002 年度に行われた。基本的な問題點がほぼ明らかとなり、かなりな部分について乘り入れが出來そうだという結論に達した。ただ實用化にはもう少し時間がかかるようである。

（2003 年 9 月）

11　“kansekieditor1.01”。データベース協議會のホームページからダウンロード出來、マニュアルも附いている。