第 22 回日中業務交流
業　務　交　流
平成 14 年 11 月 20 日

国家図書館書誌データベースとデジタル化資源の構築現状と展望

中国国家図書館善本特蔵部主任
張　志清

※当日は、劉康寧副処長の発表とあわせて 1 つにし、「中国国家図書館電子文献提供サービスとデジタル化資源の構築の現状と計画」と題して劉副処長が発表した。

中国国家図書館は全国の書誌作成センターであり、書誌データの構築を重視し、多量の経費と人員を投入してきている。この数年は、文献のデジタル化及びデータベースの構築にも成果を上げている。

一、書誌データベースの構築

当館の中国語図書書誌データは 1988 年から作成が開始された。十数年間の経験を経て、系統的な中国語図書の全国書誌（中国国家書目）が構築され、その他の資料の遡及書誌データベース構築も基本的に完成されている。現在、日常業務における書誌データの作成はすでに軌道に乗っており、中国語図書のデータだけでも年に 12 万件増加している。

データの類型は、書誌データ（目次情報を含む)、典拠データ、マルチメディアデータ及び全文データなどに分けられている。2001 年末まで完成したデータの類型と数量は以下の通りである。

1．書誌データ約 250 万件。中国語・外国語図書、中国語・外国語逐次刊行物、国内修士・博士論文、音声映像電子出版物、古典籍資料、地方志、石刻拓本等の書誌。

2．目次情報約 70 万件。中国の年鑑、中国、外国の法律・法規、敦煌学に関する論文・論著、地方志に関する論文・論著、在外華人の新聞コレクション、地名資料のうち、地図等の目次データ。

3．典拠データ約 50 万件。中国語件名典拠、名称典拠、分類法典拠等。

4．マルチメディアデータ約 6,000 時間、CD 約 22 万曲。所蔵の音声映像資料を変換したデジタル資源と貴重映像資料等。

5．中国語図書全文画像データベース 6,370 万ページ。中華民国期の中国語逐次刊行物のマイクロフィルムをスキャンした画像データ 180 万コマ。金石拓本画像データ 5,000 件。

二、近年の所蔵資料デジタル化及びデータベース構築

中国国家図書館には豊富な紙媒体の資料が所蔵されているが、現在の情報化社会においては、デジタル化及びネットワーク情報に対する要求がますます高まっており、図書館は紙媒体の資料管理とサービスを特徴とする伝統的な図書館から、デジタル資料の管理とサービスを特徴とする電子図書館へと移行しつつある。デジタル資料提供サービスは情報サービスの主導的な地位を占めていくであろう。

所蔵資料のデジタル化及びデータベース構築には統一的な標準と典拠が必要である。書誌データと典拠コントロールとを結合し、索引の科学性と実用性に配慮し、電子図書館建設に対応すると同時に、利用者の文献に対するさまざまなニーズに応えなければならない。

一般的に、優先してデジタル化すべきなのは以下の文献である：(1)不安定メディア上の文献、(2)価値が高い文献、(3)利用率が高い文献、(4)修復が必要な文献。また、データベース化計画を選定する際には、その資料自体の価値、文化的価値、芸術的特色を重視し、より価値のある研究資料の提供ができるようにしなければならない。以上の原則と所蔵資料の状況を鑑み、当館では以下のデジタル化計画を実施している：善本書（古典籍)、甲骨及び甲骨拓本、石刻拓本、敦煌文献、西夏文献、地方志、民国時期逐次刊行物、博士論文等。その内容は当館の歴代の職員が九十年以上の努力を結集し、心血を注いで収集し保存してきた、中華文化の典籍の精華である。

古典籍コレクションのデジタル化及びデータベース構築は、近年の当館のデータベース構築における重点である。中国国家図書館の古典籍コレクションの総量は 280 万点である。そのうち、善本書 27 万点には 1,600 部の宋元善本、「四大専蔵」と言われる敦煌遺書、『趙城金蔵』、『永楽大典』及び『四庫全書』が含まれている。普通古籍（民国時代の線装本を含む）は 180 万点にものぼる。その他、1911 年以降出版された革命文献を中心とする新善本、各時代の代表的な印刷と装丁の特徴を備えた上製本、西洋、日本、ロシアの善本も数多く所蔵している。著名人の手稿は約 3,000 種あり、特に近現代文学史上の代表作品が豊富である。また、総出土数の四分の一を占める 35,000 点の甲骨片、3 万余種 20 万件の金石拓本、15 万枚の国内外地図、１万枚の貴重な雷一族建築図面（訳注：雷氏は清代の建築家一族で円明園、頤和園等を設計した)、７万枚の古写真、5,000 枚の初期年画、11 万冊の貴重な各少数民族古籍文献がある。これらの古典籍コレクションは中華文明の最も貴重な文化財である。当館はその中でも最も価値のある部分を優先的にデジタル化にしたいと考え

ている。

現在、当館のデジタル化資料データベースの構築はすでにある程度の成果を生んでいる。主なプロジェクトとその内容は以下の通りである。

1．甲骨画像・拓本データベース：当館が所蔵する甲骨片は合計 35,651 点で、重要な学術価値を有している。甲骨画像データベースが完成すれば、甲骨片の実物に触れることなく、ただその拓本のみによって解釈や記録を行なうという長年の研究手段を変えることができ、甲骨片のもう一方の部分――背面にうがたれた穴の形態研究という新たな分野も開拓されるだろう（※訳注：甲骨とは動物の骨に穴を開けて焼き、ひびわれの入り具合で吉凶を占ったもの）。そして、拓本（模写、写真）によって甲骨を再構成するという昔ながらの方法も改められる。このデータベースは 2、3 年間で完成する予定であり、正式なテストデータベースは来年にはオンラインで提供される。甲骨データベースのレコードには、当館の登録番号、出所番号、卜占人、時代、出土場所、甲骨片の材質、大きさ、卜占の内容類別、拓本の出所、拓本の大きさ、再構成の状況、記録等である。撮影には高解像度撮影（400dpi）を採用し、甲骨片に刻まれた文字（特に小さな文字）及び背面の穴もよくわかるようにした。甲骨片の特殊な部位、例えば文字のある関節部は、多角度撮影に加えて 3D バーチャル技術を採用し、画像を回転して各面から見えるようにする。甲骨データベースはさらに参考資料データベースとリンクする予定である。例えば、『甲骨文合集』の出所表及び釈文部分、『甲骨文字典』、『金文字典』等である。

2．中文石刻拓本画像データベース：中国国家図書館が組織する大型の館際協力プロジェクトである。2000 年 6 月以来、『中文拓本目録規則』、『中文拓本 MARC フォーマット使用マニュアル』、『中文拓本画像加工規則』、『デジタル資源の整合規則』、『デジタル資源統合規則』などの規則を制定し出版した。『中文拓本 MARC フォーマット使用マニュアル』に基づいた MARC フォーマット拓本の MARC データを作成し、その MARC フォーマットのデータをダブリンコアを用いた MARC データに変換する。インターネットで提供する画像データは 150 dpi、72 dpi、サムネイル方式で提供する。現在までに作成されたデータは 2 万件近くになり、9,700 件はインターネットで試験的に閲覧できる。館際協力プロジェクトは現在進行中で、2005 年の末までには全てのデータベースを完成できるように努力している。

3．国際敦煌学（IDP）プロジェクト：2001 年 3 月、中国国家図書館と英国図書館は五ヵ年の協力プロジェクトを締結し、国際的な協力を通じて敦煌写本の研究と保存を促進する国際敦煌学プロジェクトに参加した。敦煌写本の安全を確保することを前提に資源共有を図り、貴重なコレクションを公開するというものである。このデータベースは国際敦煌学プロジェクトが提供した専用の 4D データベースを使用し、最も精密な PHASE1 というデジタルスキャナーで敦煌写本を一枚ずつ高解像度の画像にする。画像は写本の全ての内容―

―正面、裏面さらに文字のない部分まで――を含む予定である。画像は実際のサイズより大きく、解像度も現物を見るのと同じである。中国語のウェブサイト（訳注：http://idp.nlc.gov.cn/）は 11 月 11 日に正式に開設され、以下の内容である：一、高解像度の大型画像データベース：現在、写本の目録情報約 8,000 件と写本約 500 巻が入力されている。二、中国国内の各地方に所蔵されている敦煌文献総合目録。三、研究論著目録資料。四、シルクロード地名典拠データベース。五、敦煌吐魯番学研究記録データベース。六、IDP 中国語通信。七、「シルクロード」特定分野データベース：甘粛、敦煌、新疆、寧夏等の地域の文化資源を含む。八、オンライン展示。九、特定テーマの学術講座：主に「敦煌とシルクロード文化学術講座」の内容である。十、敦煌学会議：主に近年に中国で行われた敦煌に関する会議、特に国家図書館での会議を紹介する。

4．西夏文献及び西夏学研究資料データベース：1917 年に寧夏の霊武県で工事中に出土した二つの大箱に納められた西夏語文献の大部分は 1929 年に中国国家図書館の所蔵となった。大部分が西夏、元代の孤本で、中でも西夏語の『大方広仏華厳経』は明確に活字印刷の特性を有している。合計百数点あり、数量では国内最多、世界的にもロシアのサンクトペテルブルグ東方学研究所に次ぐコレクションである。特定分野データベースを構築するとともに、西夏文献の復元作業を進め、また、デジタル化した復元記録文書も作成している。研究者データベースと論文データベースを構築し、西夏文献の修復も進めている。修復作業は 2003 年 6 月までに完成する予定で、修復完了にあわせて完全な修復記録文書も作成し、研究者にインターネットで提供する。

5．中国デジタル地方志データベース（清代）：このデータベースは全文画像データベース、OCR データベースと 8 つの特定分野サブデータベースからなっている。全文画像データベースは、中国国家図書館分館が所蔵する 1949 年以前に編纂された約 6,000 種の地方志の全文をスキャンし、デジタル化処理をする予定である。利用者は原文を読めるだけではなく、複数の画像を用いての比較、すなわち版本の校勘もできるようになる（最大 4 枚の画像を同時に見ることができる）。OCR データベースは地方志の画像が見られるだけでなく、全文検索も出来る。この OCR データベースでは利用者は自分の必要とする内容に記号や注釈をつけて選択し、編集したりコピーしたり、異なる版本の画像を多画面での比較もできる。8 つの特定分野サブデータベースは地名データベース、人物データベース、事象データベース、作品データベース、挿絵データベース、情景データベース、目次データベース及び関連文献データベースからなっている。これによって、もともと散逸したり不完全な地方志の情報をまとめ、地方志情報ネットワークを形成する。現時点では、全文画像データベースの構築に着手しており、2002 年内には 330 万ページの地方志のスキャンが完了する予定である。来年からはこれに付随する OCR データベースと 8 つの特定分野サブデータベースの構築に着手する。

6．全国善本書総合目録データベース（部分的に原文の画像も含む）：すでに所蔵の古籍善本の書誌データ 4 万件が完成しており、三年間かけて全国善本書総合目録データベース及び画像部分とのリンクを完成する予定である。

7．国内博士論文データベース：博士論文の全文デジタル化には先進的なスキャン技術を採用し、OCR によって論文の概要等のキーになる内容を見分けてからテキストファイルに変換して、使いやすいインターフェイスと強力な機能を持つ検索プラットフォームを構築し、博士学位論文の世界的な共有を目指す。

8．民国時期中国語逐次刊行物マイクロフィルム変換デジタル資源データベース：このプロジェクトは 1999 年の後半からすでに始まっており、180 万コマの処理が完了している。当館の所蔵資料を利用して制作されたマイクロフィルムは 600 万コマある。マイクロフィルムのデジタル化はフィルムのスキャン、画像処理、論文名索引付与、ネットワークフォーマットへの変換という四つの部分からなっている。

9. 音声映像資料変換デジタル資源データベース：このプロジェクトは 2000 年から始まり、二年間の間に、大量の資料のデジタル化を行なっている：MPEG-1 規格で 3,000 時間、MPEG-2 規格で 3,000 時間、MP3 規格で 20 万曲がデジタル化されている。音声映像資料変換デジタル資源データベースの構築は、多様かつ全面的に当館の情報資源を公開すること、また、現存する大量の磁気テープ等の貴重な資料に対し保護措置を講ずることでもある。完成後は電子図書館の基礎データベース群の一部分になる。