第 23 回中日業務交流サブテーマ

2003 年 11 月 26 日

ネットワーク情報資源の収集、組織化と利用

（中国国家図書館逐次刊行物部副主任　王 志庚）

この 1 年は中国国家図書館にとって非常に重要な 1 年だった。全館の図書館業務管理システムが Aleph500 システムに順調に切り換えられ、同時に、資料の収集・整理・閲覧・保存業務の新しい流れに合わせるため組織改編が行われた。これらは、技術と組織の両面において、国家図書館がインターネット時代に新たなサービスを展開していくための基礎を築くものとなった。蔵書構築の面では、「中華人民共和国録音録画製品管理条例」の規定に基づき、国家図書館は全国の音声映像資料の納本を網羅的に受け入れている。ネットワーク情報資源収集の面では、当館はネットワーク情報資源整理組織化研究班を発足させ、テストプロジェクトを通じて、ネットワーク出版物の収集・組織化・保存に関する技術的・法律的問題を研究することにしている。

本日は、国家図書館が現在取り組んでいる、ネットワーク情報資源の収集・組織化・サービスのテストプロジェクトに関する状況について報告する。

Ⅰ ネットワーク情報資源の定義、分類、特徴

1. 定義（「国家図書館収集条例」による）

「コンピュータネットワークを通じて公開・伝送・蓄積される各種の文献情報資源を総合したもの、ネットワーク文献ともいう」

2. 分類

データフォーマット別：テキスト（doc、pdf、txt）、図表、画像、音声、映像、双方向マルチメディアデータ等

表示形式別：文字、画像、音声、動画、マルチメディア等

伝達方式別：ネットワーク出版物（ネットワーク雑誌、ネットワーク新聞、電子図書、ネットワークデータベース等）及び各種のネットワーク情報（Email、BBS、Forum）

3. 特徴

a）情報量が多く、伝達範囲が広い。

b）情報の増加が速い。毎日新たに増えるホームページ情報は 700 万。専門家の試算によれば、サイトとホームページコンテンツの増加速度は毎年 2 倍となり、幾何級数的に増えている。

c）寿命が短い。ネットワーク情報資源の消滅は速く、ホームページの平均寿命はわずか 44 日である。

d）情報の公開が自由で、出所の範囲が広く、内容が雑多で、質も一様でない。

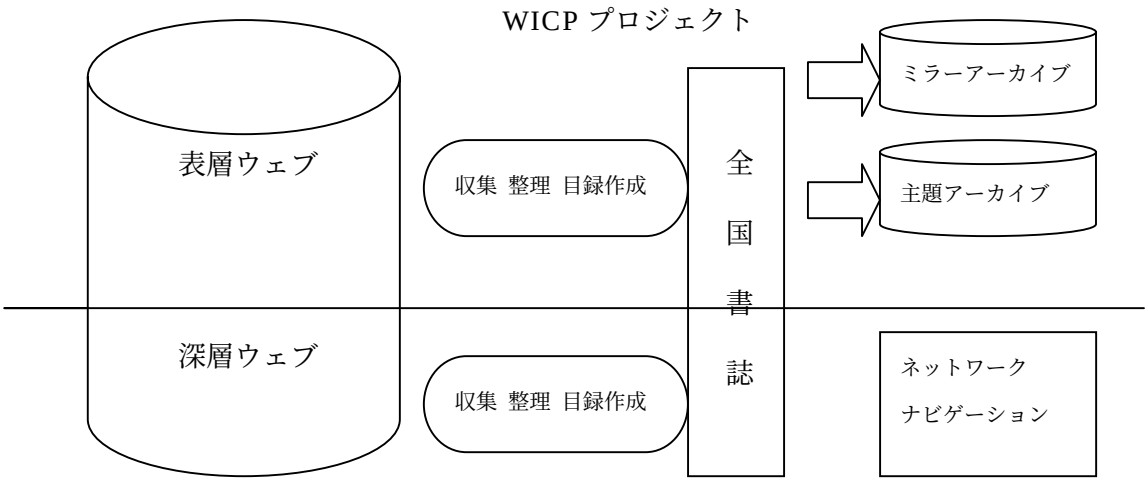
II ネットワーク情報資源の収集と組織化

ネットワーク情報資源は中華文明の成果であり中華デジタル文化遺産の一部であるから、適切に保存・保護されなければならず、ネットワーク情報資源は国家図書館の蔵書構築とサービスにとって戦略的意義を持つものであるから、国家図書館は伝統的な図書資料の収集と同じように、各種ネットワーク情報資源を網羅的に収集しなければならない、と国家図書館は考えている。

ネットワーク情報資源の収集は決して容易ではない。まず、ネットワーク情報資源は数が膨大で、全国のホームページの総数は 1 億 5709 万 1220 に上る。その公開は既に伝統的な出版行為を超えており、政府機関・大学等のネットワーク情報は比較的収集しやすいが、企業や個人等民間のネットワーク情報の収集には大きな困難が伴うだろう。

ネットワーク情報資源の収集に関して、国家図書館は法的環境整備に積極的に取り組んでいる。現在制定途上の「中国図書館法」(意見募集稿、文化省起草) には、国家図書館への電子出版物とネットワーク出版物の納本について明文規定がなかった。国家図書館はこれに対して修正意見を提出し、国家図書館にネットワーク情報資源の納本を受ける権利を与え、ネットワーク情報資源の複製・編集・長期保存・公共サービスを行う権利を保障すべきだと提案した。現在この法律はまだ意見募集の段階にある。

国家図書館は、ネットワーク情報資源収集に関する技術と基準の研究にも積極的に取り組んでいる。2003 年初頭、国家図書館はネットワーク情報資源整理組織化研究班を発足させ、テスト用ソフト・ハード環境を構築し、関連の研究とテストを開始した。実験では、表層ウェブと深層ウェブに対し、異なる整理組織化戦術を取っている。即ち、ウェブ情報資源収集保存テストプロジェクト（Web Information Collection and Preservation）とオンラインデータベースナビゲーションプロジェクト（Online Database Navigation）である（下図参照)。



テストを通じてネットワーク情報資源の収集・整理・目録作成・保存・サービスに存在する問題点を見つけ、問題解決策を提案する。テストで定められた保存対象は政府情報、学術情報、公共情報である。テスト段階では、WICP プロジェクトはサイトユニットとページユニットに分けて、ネットワーク情報の収集・目録作成・保存を行い、それぞれミラーアーカイブと主題アーカイブを形成する。

現在、国家図書館にはこのプロジェクト専門の担当部署はまだ設置されておらず、プロジェクト班は、英語・コンピュータ・図書館学・情報学等を専攻するレファレンス・目録作成・ネットワーク管理部門の当館職員で構成されている。テストプロジェクトであるため、国家図書館の LAN においてのみサービス提供している。

(1) ウェブ情報資源収集保存テストプロジェクト（WICP）

1. ミラーアーカイブ

ウェブロボット Wget を用い、あるサイトのトップページからダウンロードしていき、ダウンロードしたデータは元の目録構造を維持し、一つの情報ユニットとして保存する。ネットワーク情報は頻繁に更新されるため、異なる時点で同一対象を重複ダウンロード（スナップショット）する必要があり、このようにして複数の情報ユニットが作られる。それと同時に、ダブリン・コアや MARC によってこれらの情報ユニットの目録作成を行い（あるサイトのアーカイビングされた情報ユニットの総和に対し目録作成するものであり、一つ一つの情報ユニットについての目録作成ではない)、書誌データは全国書誌（NLC-OPAC）に収載する。

1.1 収集対象：政府機関サイト 100、電子ジャーナルサイト 100、大学サイト 100、企業サイト 100、その他（ポータルサイト、メディアサイト―新聞・ラジオ・テレビ等）100

1.2 収集コンテンツ：政府情報、学術情報、公共情報（広告等は削除）

1.3 収集ツール：Wget　バージョン：1.8.2

1.4 メタデータ：「国家図書館ネットワーク情報資源メタデータ規則」は現在制定途上にあり、ミラーアーカイブは MARC21 やダブリン・コアによって保存対象に対し目録の作成を行い、目録データは Aleph システムとテストシステムにそれぞれ収載する。主な目録記述要素は、サイト名、著作権者、発行者、開通日、分類、件名、リソース類型、URL である。

1.5 技術的課題

a）ロボットの性能：HTML 文法の誤り、PDF ファイル内リンク、拡張子のないホームページ、大容量ファイルのダウンロード

b）長期保存技術：コンピュータ技術の発展によりハード・ソフトウェアがバージョンアップした後、アーカイブデータの可読性をいかに保証するか

c）原本同一性：アーカイブデータの原本性、同一性の保障。例えば MD5 アルゴリズムの利用など

d) 一意的な識別子：信頼性の高いデータ識別体系が必要。DOI、URN 等

e) 蓄積空間：データの圧縮・解凍、データ転送、大容量データの保存技術

2. 主題アーカイブ

「国家図書館収集条例」第 45 条は、「ネットワーク情報資源の収集はいくつかの主題別に行うことができる。国際及び我が国の政治、経済、文化、科学技術、スポーツ等に関わる重大事件については、主題に基づく収集を行わなければならない。」と規定している。テスト段階で我々は、「オリンピック」「SARS」「有人宇宙飛行」について主題別収集を行った。

異なる主題について、収集対象とするポータルサイトとメディアサイト（新聞・ラジオ・テレビ等のサイト）を決め、ロボットを用いて関係するウェブページの検索とダウンロードを行い、ダウンロードしたページはソフトウェアを用いて分類の自動付与と索引の作成を行い、また、ページのスナップショットをソフトウェアで自動保存する。それと同時に、HTML ページに対してコンテンツの抽出を行い、ページ中のテキストと画像（表などのフォーマットファイルは抽出不能）を抽出してデータベースに保存する。

2.1 ツール：31/CGRS/TRS

2.2 メタデータ：ウェブページの主題、公開日、公開時間、オリジナル URL、分類、件名、作者、序列番号、本文、スナップショット、内容説明等

2.3 技術的問題：自動分類・索引や内容説明自動作成の技術、自動重複調査技術、コンテンツフィルタリング技術等

(2) オンラインデータベースナビゲーションプロジェクト（ODBN）

カタロガーがサーチエンジン検索でデータベースを発見し、データベースの目録作成、分類、索引を行い、書誌データを実験システムに収載し、リンク技術を基にネットワークナビゲーションを行う。このプロジェクトは異なるデータベース間の横断検索サービスは行わない。

メタデータ：データベース名、著作権者、開通日、分類、件名、サーバーアドレス、サービス方式、課金の有無、データ量、検索方式、更新頻度等

技術的問題：

a）データベースの発見：現在は手作業でサーチエンジン検索によりデータベースを発見し、目録を作成しているが、サーチエンジンでは検索できないデータベースも多く、また、新しく開かれたデータベースは通常一定時間経過しないとサーチエンジンに収録されないので、すぐに収集するのが難しい。

b）データベースのフォローアップ：データベースのコンテンツ更新状況等は、年度毎あるいはもっと短い間隔でフォローアップしなければならない。

c）ナビゲーションの有効性：サーバーアドレスの変更状況、サービスの有効性等のフォローアップ

III ネットワーク情報資源のサービス

WICP と ODBN はいずれも実験プロジェクトなので国家図書館の LAN においてのみ提

供され、情報量が少ないため、対外的な検索サービスは基本的にまだ行っていない。次に、国家図書館が自ら構築したデジタル資源のサービス状況と、購入しているネットワーク出版物のサービス状況を紹介する。

IDP プロジェクト

敦煌文献は中国近代文化史における四大発見の一つである。現在世界に敦煌文献は約 5 万点余り存在し、主に中国、英国、フランス、ロシアに収蔵されている。1994 年、中国国家図書館と英国図書館、フランス国立図書館、ベルリン国立図書館等が国際敦煌プロジェクト（International Dunhuang Project）を正式に発足させ、事務局を英国図書館に置いた。その目的は、敦煌及びシルクロードのその他の遺跡から出土した 11 世紀以前の文書・芸術品の研究と保護を促進することである。IDPの初期の事業は修復と目録作成に集中し、近年になってデジタル化事業が増加してきた。中英両館は 2001 年から 2006 年にかけて共同で敦煌文献のデジタル化を行い、両館の所蔵するすべての敦煌自筆稿本の目録をデータベース化し、インターネットで全世界の学者に無料で提供することにしている。英語サイトは 1998 年にすでに開通している。中国語サイトは 2002 年 11 月 11 日に開通した (http://idp.nlc.gov.cn/)。中国語サイトに収載されている文献目録データは 3 万件、文献は 50 余点、画像情報は 4 万余件である。最終的には敦煌・シルクロードにおけるすべての収集品が、インターネット上で無料で閲覧できることになる。

ネットワーク版電子ジャーナルのサービス

国家図書館は資料購入費を用いて、ネットワーク電子ジャーナルデータベースの出版事業者と取り決めを結び（単独購入とコンソーシアムによる購入の 2 種類）、ネットワーク電子ジャーナルの使用権を購入している。図書館 LAN 内のコンピュータでは、国家図書館ミラーサイトに保存されている電子ジャーナルを利用することができ、ユーザー名とパスワードを登録したリモートサーバーでも必要な文献を検索することができる。来館利用者は電子閲覧室で電子ジャーナルを検索し、ダウンロード、プリントアウトを行うことができる。利用者サービスは有料である。現在利用できる主なネットワーク版電子ジャーナルは、英語リソースとして Ei Village、Elsevier Science、IEEE/IEE、IOP 英国王立物理学会電子ジャーナル、OCLC ECO 全文ジャーナル、Springer 電子ジャーナル、EBSCO 全文データベース、Swets Blackwell、中国語リソースとして中国ジャーナルネット、維普中国語科学技術雑誌全文データベース、万方デジタルジャーナルがある。

以上、国家図書館のネットワーク情報資源の収集、組織化、サービスについて紹介した。国家図書館はこの方面についてはスタートが遅かったので、日本国立国会図書館の手法を参考にさせていただいたところが多い。貴館の専門家各位のご意見ご提案をいただくと共に、ネットワーク情報資源の収集・保存の面で一層協力を深めていきたいと願っている。