第 27 回日中業務交流
テ ー マ 報 告 Ⅱ
2007 年 11 月 7 日

中国国家図書館新聞デジタルリポジトリー（DiNeR）プロジェクト

中国国家図書館
報刊資料部主任補佐　李春明

2005 年、私は貴館の招聘を受け IFLA のプレコンファレンスに参加した。その際には貴館の親切な接待を受け、同業のみなさまと交流し、貴館の文献資源の保存と保護の進展について理解する機会をいただいたことに、ここで改めて感謝の意を表したい。前回の会議において中国新聞のマイクロ化とデジタル化の現状について紹介したが、今回の報告は当館の新聞の保存と利用に関する業務の進捗についてである。これも当館のデジタル図書館設立の重要な部分を占めており、私の説明によって、当館の新聞デジタル化事業に関する理解を深めていただければと思う。

1.プロジェクトの背景

新聞は一次情報を大量に含んでおり、ある国や地域、また特定の歴史的時期の社会状況、方針、政策、社会風俗、民俗を研究するための重要なよりどころで、極めて高い史料的価値をもつものである。デジタル化、ネットワーク化の波が押し寄せる中で、伝統的な新聞業界はしだいにデジタル化ネットワーク化へと向かって発展してきている。2006 年 2 月 1 日、武漢で創刊され 80 年近い歴史をもっていた台湾の『中央日報』が停刊した。2006 年 12 月 29 日には、1645 年に創刊し、世界新聞協会によって現在も発行されている最も古い新聞とされていた、スウェーデンの首都ストックホルムの『ポスト・オック・インリッケス・ティドニンガー』（Post-och Inrikes Tidningar）も、2007 年 1 月 1 日から正式に印刷版の発行を停止すると発表した。これらは過去 1 年に起きた二つの典型的な事件である。ここ数年、中国の各新聞社は自社の情報力を高めるために、次々にサイトを立ち上げて、ユーザーにデジタル版新聞閲覧サービスを提供しており、デジタル版新聞はネットユーザーにとって情報入手の新たな場となっている。しかし情報の組織化の点からは、各紙の間で統一された情報スキーマやルールが欠けており、ユーザーインターフェース、検索コマンド、データフォーマットが異なっていることもある。ユーザー側から見ると、ユーザーは情報を調べる際に多くのサイトを行き来しなければならず、時間と労力を費やすことになるだけでなく、検索性と情報の実用性は十分ではない。長期保存の観点から見ると、ネ

ット新聞は寿命が短い。さらに、出版者の多くが商業目的のためにリアルタイムの情報を配信しているが、遡及情報の検索サービスは提供していない。ほかにも、デジタル版新聞の出版者が商業競争に敗退し倒産するなどの理由から、多くのデジタル新聞がそのサイトの閉鎖に伴って消失し、閲覧できなくなっている。

中国国家図書館は、国の総書庫であり、デジタル資源の長期保存戦略構想に基づいて、2005 年に新聞デジタルリポジトリプロジェクト（DiNeR：Digital Newspaper Repository. 以下、DiNeR)）を立ち上げた。このプロジェクトは当館デジタル図書館のコンテンツ構築の重要な部分であり、一般利用者にデジタル新聞のブラウジング・検索サービスを提供するだけでなく、専門主題情報の加工の基礎となる。同時に、新聞社にとっては、データの長期保存と閲覧用のプラットフォームを整備することになる。2 年間の取組みの中で、データはある程度の蓄積ができたが、同時にいくつかの課題が明らかになった。一つはデータ収集が主に人手によっており、業務量が膨大であることである。もう一つは、一部の PDF フォーマットの内容を変換するときに識別ミスが起こることがあり、全文検索の実現が難しいことである。2007 年からは、中国国家図書館は継続して PDF 版新聞の収集を行うとともに、国が定めた電子出版物の納本政策と連動して、出版者に対して出版データの納入を促す取組みを行い、出版元からデータを取得するようにした。同時に、北京大学方正集団公司と共同、同社の出版業者としての長い経験とデータ収集面の優れた実績を活用して、デジタル新聞の保存と利用に関する検討や実験を行った。また、国家図書館のデータ統合とマイニング面での強みを活かし、DiNeR を基にして各種のデジタルサービスを展開することも計画している。こうした事業によって、印刷出版業界や図書館における中国国家図書館の中心的な権威としての価値をさらに高めていきたい。

2. DiNeR システム設計

2.1 システム設計の基本的考え

DiNeR のシステム設計は、図書、雑誌などの資料に対して相対的に特殊性をもっている。その主要な理由は新聞に含まれる内容、組版形式などの特性による。新聞一部には多くの紙面があり、厚いものでは百紙面以上になることもある。また、1 紙面には図、文字、表など様々な情報を含む多くの記事があるほか、複数紙面にわたるものや連載など様々な形式が存在する。その結果、記述形式及び構造形式を含む新聞のメタデータの加工は複雑なものになるので、システム設計時には、OAIS, PRIMS, METS, PREMIS のような関係する国際標準や J2EE, XML, Unicode, Web Service などを参考にしなければならない。完全性、将来展望性、継続性、拡張性を考慮する必要があるだけでなく、ユーザビリティ、安定性、完成度、融通性と開放性も満たし、さらには安全性、拡張性、管理のしやすいユーザーフレンドリーなインターフェース、高性能であることなどの特徴を具現化しなければならない。データの情報検索とブラウンジングができるだけでなく、データマイニング及びデー

タ再組織化の必要性、すなわち、現在のニーズと長期にわたる保存と利用というニーズも考慮しなければならない。

2.2 システム構造

DiNeR のプラットフォームは B/S 構造のシステムであり、J2EE の 3 層構造で開発された。以下の図に示したとおりである。

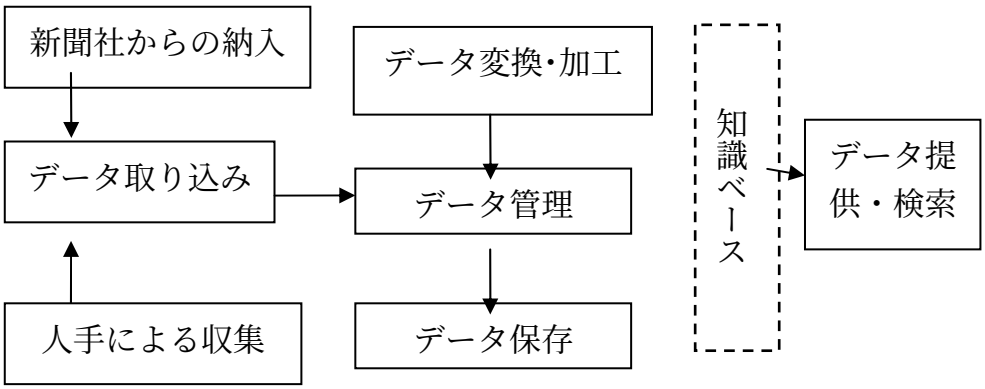

図 1　OAIS に基づいた DiNeR 保存・提供の流れ

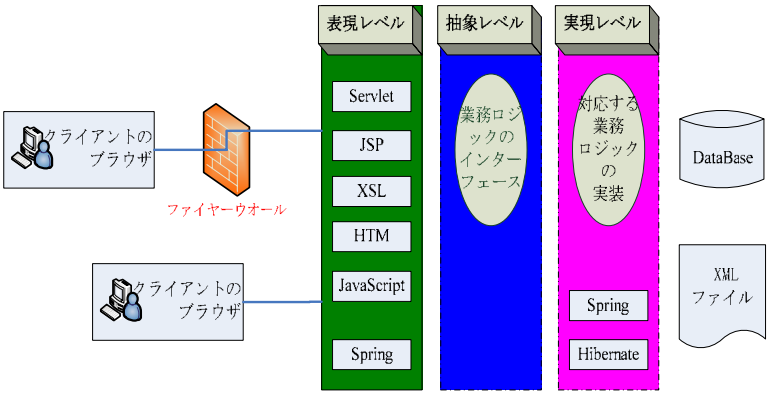

図 2　DiNeR システムの構造図

2.3 システム機能

2.3.1 データ取り込み

現在、DiNeR データは主として、新聞社から納入されるデータと新聞サイトからダウンロードしたデータから成る。データ取込みシステムによって新聞社のデータ納入、データ・

クリーニング、ウイルスチェックなどの関連業務などを行う。また、データ取込みシステムによって、システムへの人名、地名、件名表などの典拠データを取り込み、システムの最下層にある知識ベース構築のための素材を整備する。

2.3.2 データ変換と加工

システムに取り込むデータは主に、組版ファイル、版面ファイル及び PDF ファイルである。このモジュールにおいては、逆解法あるいは人的手段で、メタデータとオブジェクトデータも含むデータを統一されたファイルフォーマットに変換する。記述メタデータの作成には PRISM(Publishing Requirements for Industry Standard Metadata,《工业标准元数据的出版需求》[业界標準メタデータの発行要件])、《中文新闻信息置标语言国家标准》（GB/T 20092－2006）([中国語ニュース情報マークアップ言語国家標準]) 及び《中文新闻信息分类与代码》([中国語ニュース情報分類とコード]）などを参考にし、セマンティック記述方式を採用した。これによって、検索システムでは意味検索が可能となった。オブジェクトデータは Adobe の 2 層 PDF ファイル形式を用いて保存する。

2.3.3　データ管理

デジタル新聞コンテンツの管理は主に以下の部分から成り立っている。

（1） デジタル新聞コンテンツの分類管理：分類別に各新聞社のデジタル新聞のコンテンツを管理。

（2） デジタル新聞コンテンツの集積と統合：例えば主題など、ある一つの方式でコンテンツを再整理・組織化。

（3） デジタル新聞コンテンツの提供管理：提供コンテンツ・経路・提供形式を設定。

（4） デジタル新聞コンテンツの保存管理：デジタル新聞コンテンツの保存、バックアップ、修復を実現。全文データベース方式の採用で膨大な量のデジタル新聞情報を統合・保存・修復する機能を実現することを提案。

（5） データ管理において、コンテンツあるいは分類に基づいてそれぞれオントロジーを構築するとともに、システムに取り込んだ統制語彙表を利用して最下層の知識ベースを生成し、データ検索に提供する。

2.3.4 データの長期的保存

デジタルデータの長期的保存について言えば、単にデジタル情報のビットあるいはバイトのみを保存するだけでは全く不十分である。デジタル情報がより長い期間人々に理解され利用されるためには、デジタル情報の製作や使用に関する技術情報と環境情報を保存することがぜひとも必要である。DiNeR は、構造メタデータ標準として METS を採用し、データのカプセル化を行うほか、OCLO の PREMIS の保存メタデータ辞典を参考にして保存メタデータを作成する。

2.3.5　データ提供と検索

多くの方法でデジタル新聞のコンテンツを提示する。

１）ウェブ・ページ式デジタル新聞：紙面の内容をウェブ・ページ形式で表したもの。紙面内容の高速ブラウジング、日にち変更、全文検索、フィールドの構造化検索などが非常に簡単にできる。

２）伝統的閲覧方式の電子新聞：紙版の内容と全く同じにし、原版の形式と雰囲気そのまま表現した電子新聞で、PDF 若しくは JPEG フォーマットを用いて表示する。

３）デジタル新聞の柔軟性、閲覧の利便性を考慮してテンプレート技術を採用した。これによって、簡単にインターフェースや閲覧方式のカスタマイズ、スタイルや内容を柔軟に表現できるシステムとなった。

４）デジタル新聞のテンプレートは、ウェブデジタル新聞の表示と閲覧のテンプレートである：ウェブデジタル新聞が表現する各種コンテンツの形式及び多様な閲覧方式を、そのテンプレートを簡単にカスタマイズして実現する。

検索はこのプラットフォームの重要な構成部分で、このシステムは静的コンテンツ検索方式フォームを採用した検索サービスを提供する。ユーザーは、キーワード、時間設定、内容分類、作者の情報などを入力して組み合わせ検索ができる。検索は全文及びフィールド検索を合わせた複合検索方式で、ユーザーが検索できるコンテンツはすべてのインデックスデータベースである。検索は標題、キーワード、本文、時間、コラムなどのフィールドから、正確・迅速に必要な情報を特定する。

3．DiNeR の知的財産権問題の解決

2006 年 7 月 1 日から、中国では「情報ネットワーク伝達権保護条例」が施行された。「条例」の第 7 条は、「図書館、文書館、記念館、博物館、美術館等は、著作権者の許諾を得ずに情報ネットワークを通じ、当該館内のサービス対象者に対して収蔵している適法に出版されたデジタル作品及び法に基づき陳列又は版本の保管の必要性からデジタル化により複製された作品を提供することができる。その際、著作権者に対して報酬は支払わない。ただし、直接的又は間接的に経済的な利益を得てはならない。当事者において別途約定がある場合は、この限りでない。」＊と規定している。この条例に基づき、我々は図書館の LAN 内であれば、DiNeR を合理的に使用することができると考えている。しかし、一般利用者の利用の便のために、当館ではデータを提供する各出版者と使用条件について交渉し、一般利用者が無償で利用できる権利と、制限がある場合は特定の権限があるユーザーには利用提供できる権利を獲得しようとしている。DiNeR プロジェクトは、図書館と出版者の双方に利するものである。つまり、図書館は資源サービスの権益を得、出版者は図書館を通じて情報資源を保存し発信をすることができる。

4．結びに

新しい情報技術環境は、図書館情報サービスにとっては課題であると同時に、チャンスもたらしてくれた。このチャンスをつかみ、図書館とユーザーとの間の連携をより緊密なものにすることによって、図書館事業の発展を不断に推進しなければならない。デジタル新聞データベースの構築は、中国国家図書館のデジタル図書館建設における一つの実践であり、さらに完備しなければならない部分がまだ多くある。今後、デジタル図書館建築について、貴館と多方面での交流と協力を推し進めることによって、資源とサービスの共有が実現できるよう望むものである。

＊条例７条の翻訳は「知的財産法中日翻訳　―法令篇―」(http://legalio.com/index.html)というサイトの「情報ネットワーク伝達権保護条例(信息网络传播权保护条例)」(http://legalio.com/ip/ct-column.cgi?mode=al2&namber=8&rev=&no=1）を参考に一部付け加えています。