神戸大学電子図書館システムにおける「電子アーカイブ」の構築

渡邊 隆弘
神戸大学附属図書館
〒 657-8501 神戸市灘区六甲台町 2-1
Tel: 078-803-7333 Fax: 078-803-7336 E-mail: watanabe@lib.kobe-u.ac.jp

概要

「神戸大学電子図書館システム」では、図書館所蔵資料等を電子化して広く公開する「電子アーカイブ」構築を中核的事業と位置づけている。本稿では、「電子アーカイブ」構築・検索システムについて、データベース設計・検索システム・コンテンツの概要を紹介する。重点対象資料の一つである阪神・淡路大震災関係資料は著しい多様性という特徴を持っており、これを念頭に開発を行なった結果、メタデータを重視したシステム設計となっている。

キーワード

神戸大学電子図書館システム, メタデータ, 阪神・淡路大震災関係資料

Construction of Digital Archives in the Kobe University Digital Library System

Takahiro WATANABE
Kobe University Library
2-1 Rokkodai-cho, Kobe, 657-8501, JAPAN
Phone: +81-78-803-7333 Fax: +81-78-803-7336 E-mail: watanabe@lib.kobe-u.ac.jp

1. はじめに

神戸大学附属図書館では、平成10年度補正予算による予算措置をうけ、平成11年2月より「神戸大学電子図書館システム」を稼働した。11年5月にはシステム披露式典を行ない、本格的なサービスを開始している[1]。システム開発は、図書館側との意見交換のもと、NTT（現NTT西日本）が行なった。

「電子図書館」の意味する範囲は大変広く、様々な方向性が見られる。電子ジャーナル等による学内研究・教育支援機能の強化、所蔵資料等の電子化による外部情報発信、ネットワーク情報源の組織化や特定分野の「サブジェクト・ゲートウエイ」への指向、などである。これらはもとより排他的に選択されるものではなく、現実の電子図書館は様々な機能をミックスしたものとなるのが普通であろう。図1に神戸大学電子図書館のシステム概念図を掲げたが、我々も様々な要素を含んだ多角的な情報サービスを指向している。

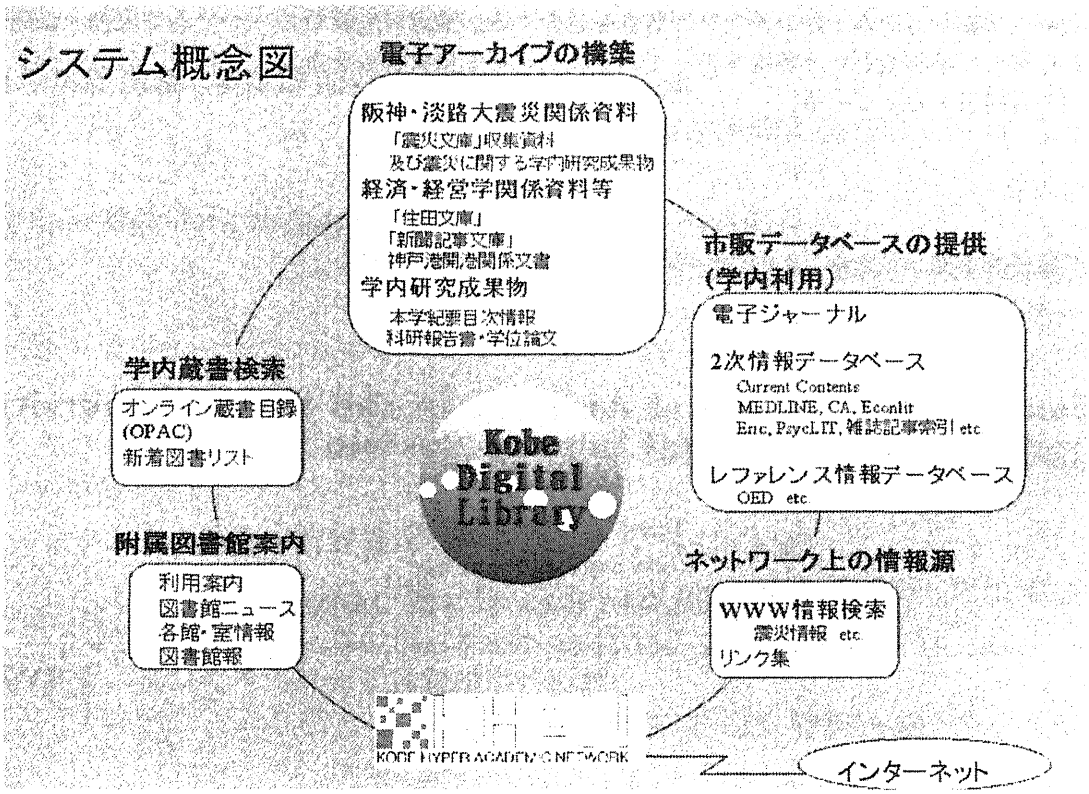

図1 システム概念図

しかしながら、各電子図書館システムにおいて、それぞれ重点の置きどころが異なっているのもまた事実である。そうした意味で、神戸大学電子図書館システムが中核的な事業と位置づけているのは、図1の中心上部にある「電子アーカイブ構築」、すなわち所蔵資料等のデジタル化による全国・世界への情報発信である。

以下本稿では、この「電子アーカイブ」の構築・検索システムについて特その概要を述べるものとする。

2. 電子アーカイブの対象資料

神戸大学電子図書館システムでは、いわゆる「電子図書館コンテンツ」のうち、図書館の所蔵する資料や神戸大学で生産された学術研究成果等を、主体的に電子化して広く公開していくものを「電子アーカイブ」

と呼称している。図書館資料といい学術研究成果といっても様々であるが、全国的にも特色のある資料群として、「阪神・淡路大震災関係資料」と「経済・経営関係資料」を2本の柱とし、その他学内の各種研究成果等も視野に入れている。

2.1. 阪神・淡路大震災関係資料

神戸大学附属図書館では、1995年1月に起きた阪神・淡路大震災に関する資料の網羅的収集活動を震災直後から続けており、「震災文庫」として広く一般公開している[2]。また、神戸大学では、被災地に位置する総合大学として、震災に関して様々な研究が行われており、その原資料なども図書館で受け入れつつある。

「震災文庫」は既に約17,000点の資料を数え（1999年11月現在。増加の勢いも衰えていない）、全国・海外からの利用・問い合わせも多く、ネットワークによる情報提供の効果は非常に大きいと考えられる。

2.2. 経済・経営関係資料

神戸大学は現在では総合大学となっているが、神戸商業大学（さらにその前は神戸高等商業学校）を前身としており、図書館蔵書においても社会科学分野に貴重資料が多い。

従来から試行的に電子化を進めてきたものに、慶長から明治初年にいたる海運海事史関係資料コレクション「住田文庫」がある。この他、神戸港開港関係の資料も多数所蔵していることから、この分野をまず重点に考えている。

また、本学経済経営研究所には「新聞切抜文庫」（明治末期からの新聞切抜資料）がある。1945年までに限っても50万コマという膨大なものであり、研究者の目で記事分類されていること、また旧植民地を含む地方紙をも対象としていることから、大変貴重な資料と評価されている。

2.3. 学内生産学術研究成果

紀要類、学位論文、科学研究費報告書などは、市場にのらない「灰色文献」であり、こうした学内研究成果の積極的な公開に寄与したいと考えている。

また、各部局・各講座に保有された研究成果・研究データなどで、電子図書館コンテンツとしてふさわしいものがあれば対象資料に加えたいと考え、学内の各教官に「電子図書館への登録資料に関するアンケート調査」を行なっている。

3. 震災資料と「電子アーカイブ」の設計

「電子アーカイブ」におけるデータベース構造の設計は、対象資料のうち、特に阪神・淡路大震災関係資料を念頭において行なった。これは震災資料が他機関の電子図書館コンテンツと比べてやや異質な特徴を持つためである。

3.1. 震災資料の多様性

「震災文庫」では関係資料の網羅的収集をめざしているが、結果として集まった資料（現在約17,000件）は多様性に富んでいる。

まず、資料媒体が多様である。図書、雑誌以外に、パンフレット・レジュメ類、チラシ・ポスター等の一枚もの資料、新聞・広報紙類、地図、写真（一般市民によるものも）、音声資料、映像資料、コンピュータ

ファイルなどが含まれる。近年一般的な図書館資料の多様化も進んでいるが、それとは比較にならない割合で様々な資料が受け入れられている。たとえば、チラシ・ポスター等の一枚もの資料は4000件を超えており、全体の2割以上を占めている。

また、資料としての単位が様々である。「震災」という切り口で資料収集を行なうと、一冊の図書・雑誌といった発行時の物理単位ではなく、その一部分だけが目的に適うことがしばしばあるので、抜刷・抜粋・切抜といった形態で保持されている資料が相当数に及んでいる。

さらに、利用者による情報要求の多様性がある。どのような分野でも情報要求は多様であるが、情報探索には二次情報データベースや概説書・レビュー論文・書評・引用などいろいろな手段があり、図書館目録はそれらと総合されて利用者を資料に導くものである。ところが震災資料においては、今のところ利用者が頼るべき検索手段（とりわけ網羅性のある二次情報データベース）が乏しい。「震災文庫」がもはや他では入手できない資料も含めて相当数の蓄積を持っている以上、収集資料が十分に活用されるには、従来の図書館目録レベルを超えた、レファレンスデータベースとしての充実が必要である。具体的には、雑誌の論文タイトルからの検索などはもちろん、図書中の章節タイトル、可能ならば全文を検索対象とできること、また図書中の図表・統計表・地図・写真なども独立して検索・表示対象とできることが望まれる。

3.2. メタデータの重視

本システムでは、もちろん一次情報の入力・提供も進めていくが、一方でメタデータの整備にも力を注いでいくつもりである。

震災資料はすべてここ数年の資料であるから、一次情報の入力には著作権処理が必須である。商業出版物もかなり含まれていることから、入力作業コストの点を別にしても、すべての資料を一次情報まで電子化する見通しはたたないし、必ずしも利用ニーズの高いものから処理していけるわけでもない。また、写真・地図などテキスト以外の表現をとる情報も重要性が高く、一次情報が入力できても、全文検索という手段だけでは問題解決できない。

このような条件のもとで、震災関連情報のレファレンスデータベースとして十分なものを構築しようとすると、一次情報入力とは別に、メタデータの計画的な整備を進めていかなくてはならない。また、前節で述べた震災資料の特殊性から、メタデータには次のような要件が求められると考えた。

- 資料中の、十分精細なレベルまで表現できること
- 様々な情報単位を弁別できること
- 様々な媒体、単位に応じて柔軟性を持った記述ができること

3.3. メタデータの設計

本システムのメタデータは、1資料1レコードではなく、検索・表示対象となるレベルすべてに対して作成される。従って図書であれば、1冊の単位はもちろん、各章単位、各節単位、あるいは挿入された写真・地図・図表の1枚単位など、検索・表示に必要と判断された多数のメタデータが作成される。

そのように作成された様々なレコードを識別するために、「リソース種別」を各メタデータレコードに与える。リソース種別はレコードの基本的種別情報であり、「アーカイブ種別」「資料種別」「エレメント種別」の3種の組合せで表現される。

例） 01/02/05 ＝ 震災／雑誌／記事　（震災文庫の雑誌資料中の1記事単位）
　　 01/01/06 ＝ 震災／図書／写真　（図書資料に掲載された写真の単位）

「アーカイブ種別」は、震災文庫、住田文庫... といった資料群（アーカイブ）の別である。

「資料種別」は、図書資料、雑誌資料、新聞・広報紙類、パンフレット資料、一枚もの資料、写真資料、地図資料、映像資料... といった、対象資料の媒体をあらわす種別である。

「エレメント種別」は作成されたメタデータのレベルを表す [3]。

シリーズ単位
資料タイトル単位（通常の図書館目録の書誌単位）
分冊巻号単位
記事・著作単位
章・節単位
写真単位（一枚の写真）
地図単位
図表単位...

各メタデータレコードの入力項目は、国際標準となりつつある Dublin Core Metadata Element Set[4] の 15 項目に、版表示、出版年、数量、大きさ、請求記号など主に図書館目録で用いられてきた諸要素約 20 項目を加えたものを「基本データ項目」としている [5]。

また、写真では撮影場所や撮影日付が重要な検索・識別条件となるなど、媒体等に応じて必要になる項目があるため、「リソース種別」に応じた「拡張データ項目」も定義することができる。さらに、データベース構築過程で新たなデータ項目が必要になったりする場合に備えて、データ項目の追加も比較的柔軟に行なえるようにしている。具体的には、リソース種別ごとに各項目の有無や画面上での見出し名称・表示順などを規定した「項目定義ファイル」を外部ファイルで保持しており、入力系・検索系ともプログラム修正を伴わずに追加・変更が可能となっている。

このようにメタデータを丁寧に作成すれば、詳細なレベルの検索が可能になるが、一方で、1 資料から作られた複数のメタデータレコードから、原資料の文脈・構造に沿った表示（図書の「目次」にあたるような表示）も行なう必要がある。このため各レベルのメタデータレコードは階層的にリンクされ、最上位のものからのツリー構造を形成している。リンク情報として、メタデータレコード間の親子関係と、同位レコード内での順序情報を保持している。[6]

3.4. 一次情報の作成と管理

一次情報は原則としてメタデータと 1 対 1 でリンクされる。図書であれば章なり節なりのまとまりで一次情報ファイルを作成し、その単位のメタデータとリンクを行なうこととなる。

一次情報は最終的に外部ファイルとして扱われるので、システム上特に形式の制限はないが、テキスト文書の場合は PDF もしくは HTML、画像は JPEG を考えている。テキスト文書は OCR 処理を行なって、全文検索用テキストを生成する。文字認識の完全校正はコストがかかるが、

- 完全校正を行なってノーマル PDF もしくは HTML で表示ファイルを作成する。
- 表示ファイルはイメージ PDF とし、校正不完全の OCR テキストは検索にのみ使用する。

のどちらかを、資料に応じて選択する。[7]

電子図書館システムのデータは、単に画像やテキストが表示できるだけではなく、当該資料の論理構造を十分反映した形式である必要があるが、本システムでは全文テキストを SGML 記述するなどの構造化は当

面考えていない。検索エンジンにはOpenTextを導入しておりSGMLデータを受け入れる余地はあるが、震災資料の著しい多様性を考えると、有効なDTD設計は実際上困難である。その代わりに本システムでは、メタデータを細かく作成し、メタデータ間のリンク関係という形で、ある程度の文書構造を表現する方法を選択している。

4. システム構成と検索システム

4.1. システム構成

本システムはシステム構成上、「入力サブシステム」「データ管理サブシステム」「検索・表示サブシステム」よりなる。

「入力サブシステム」は、メタデータと一次情報双方の入力を行なうもので、DBサーバはUNIXであるが、クライアントはWin98PC上で動作する。Windows Explorerふうのインターフェースでメタデータツリーのメンテナンスができるなど、本システムのデータ特性に合わせた独自システムである。メタデータはRDB（ORACLE）で管理し、一次情報そのものは個別ファイルとして格納している。

「データ管理サブシステム」では、DBサーバ上の入力データと、検索サーバとの橋渡しを行なう。検索エンジンにOpenTextを用いており、データをSGML形式で保持する必要があるため、バッチ処理でRDB上のメタデータ（検索用全文テキストを含む）をSGML変換する。このバッチ処理は更新分を毎晩行ない、入力データは翌日検索に反映される。なお、一次情報（検索用全文テキストを除く）はDBサーバに格納されたままで、画面表示時にその都度アクセスされる。

「検索・表示サブシステム」はWWWからのCGIインターフェースを基本とする。検索サーバではZ39.50サーバシステムが稼働しており、WWWサーバからの検索要求はZ39.50の検索式として渡される。最終的にはOpenTextによりメタデータおよび検索用テキストに対する全文検索がなされる。

4.2. 検索システムの特徴

CGIによる検索システムは、ホームページ上で「電子アーカイブ検索」という名称で提供している。前述のようにデータベース中には震災文庫、住田文庫など複数のアーカイブが含まれているが、検索システムは一つで、全アーカイブの一括検索を初期設定としている（もちろん特定のアーカイブに絞ることも可能である）。検索は、検索／一覧表示／詳細表示／一次情報閲覧という一般的な流れであり、基本部分はきわめて単純なインターフェースである。また、メタデータ及び検索用全文テキストに対して、検索エンジンOpenTextにより全文検索がなされるので、検索語入力規則の点からも分かち書き等の問題がなく単純である。

以下では全般的な機能紹介は行なわず、特徴的な点のみを列挙する。

4.2.1. Z39.50プロトコルの使用

情報検索プロトコルZ39.50を導入している。Z39.50ではサーバークライアント間で検索セッションが維持されるので、検索履歴の再利用などが比較的容易に実現できている。しかし、本システムでは複数アーカイブを単一のDBにおさめているので、Z39.50の大きな特徴である、検索・返戻等の文法規定による複数DB間・複数サーバ間の横断検索機能は有効性を発揮していない。

現在のCGI検索だけでなく、OPACや他機関DBとの同時検索が実現されるなどしてはじめて、Z39.50の機能を十分生かしたことになると思われる。

4.2.2. メタデータツリー構造と検索

前述のように、1資料より作られた複数メタデータレコードはツリー構造を形成しているが、検索システムの「データ詳細表示」では、「目次」にあたる「階層型表示」と、選択されたメタデータの詳細項目とが左右のフレームに分かれて表示される形をとる。

また、このような構造をとるため、ツリーの上位階層にあたる部分のタイトルからも検索を可能にしている。

神戸大学の被災状況　　（資料タイトルレベル）
　第1章　概説　　（章・節レベル）
　第2章　附属病院の活動　　（章・節レベル）

といった構造の場合に、「神戸大学 AND 附属病院」で第2章のデータが検索される必要があるためである。一方で「神戸大学 AND 被災状況」の検索で各章・節レベルのデータまでが別個にヒットしては、一覧表示が見にくく正確なヒット件数もつかめないので、階層関係をなす複数データがヒットした場合は、そのうち最上位のものだけに絞り込むという処理を行なっている。

4.2.3. 類義語辞書の使用など

類義語辞書を搭載しており、検索オプション設定により、入力検索語の類義語も含めて検索することができる。

また、漢字について新旧字体の同時検索も同様に可能である。

4.2.4. 英語版検索機能

検索画面には、日本語版と英語版があり、各言語版の図書館トップページからたどることができる。類義語辞書など一部を除いて、機能・画面展開は同じである。

日・英版は単に画面表示言語の違いではなく、検索対象データを異にしている。本システムのＤＢでは一部コード情報を除く全項目についてそれぞれ日英のフィールドを設けてバイリンガルとしている（OpenText 用の SGML 形式メタデータでは、日英のフィールド群がそれぞれ<JA>...</JA>, <EN>...</EN>という上位タグで囲まれた形になっている）。日本語版検索では日本語フィールドのみを、英語版検索では英語フィールドのみを検索する。

運用上は、日本語フィールドには欧文タイトル等も記述して、日本語版検索では日本語・欧文あわせた検索を可能にしている。英語フィールドはいまのところ、欧文資料と、欧文タイトルが併記された日本語資料にのみ作成しており、英語版は検索できる情報が限られている。今後、翻訳などにより英語フィールドを増強すれば、海外への情報発信がより充実することになる。

5. コンテンツの現状と見通し

2. で電子アーカイブの対象資料について概観したが、ここでは具体的な公開状況（1999 年 11 月現在）と進行中の計画事項について述べる。

5.1. 阪神・淡路大震災関係資料

従来から目録公開を行なっていたので、所蔵資料（約17,000件）のメタデータは既に作成済であるが、多くは資料タイトルレベルまでの簡略データである。精細レベルのメタデータを一気に作成することは難しいが、まずは図書・雑誌の記事・著作レベルについて今年度の重点項目として入力を進めている。完了すれば、ひとまず論文レベルまでの検索が保証されることとなる。さらに写真資料の情報や広報紙類の記事などに取組んでいく予定である。写真資料については写真一枚ごとの情報を入力していくわけであるが、一次情報に言語の障壁がないという特質に着目して、翻訳による英語版メタデータの整備にも着手したいと考えている。

一次情報としては、現在チラシ・ポスター等の「一枚もの資料」約1,200点について、画像を公開している。これについては一点一点郵送による利用許諾をとっており、許諾を頂いたものから順次公開を行なっている。さらに、一般の方の撮影による写真資料や学内教官の著作物、ボランティア団体関係の資料など、著作権処理を進めながら多様な資料のデジタル化に取組んでいきたいと考えている。

5.2. 経済・経営関係資料

「住田文庫」（海運海事史関係史料）については、約100点5000画像をデジタル化・公開している。古資料の電子化にあたっては、今のところ4x5cmのカラーフィルムによる写真撮影を行ない、ネガとProPhotoCDによるディジタル画像を保存している（インターネット上での公開にあたっては200kバイト程度の画像に抑えている）。今後は、資料の特質に応じた検索・表示方法の検討や解題情報の整備等を、関係教官の協力を得て進めることになっている。

「新聞切抜文庫」（明治末期からの新聞切抜資料）については、本年度よりデジタル化事業を開始するが、新聞記事はその性質からいって全文検索やテキストベースの二次利用の必要性が高いと考えられるため、記事全文のテキスト入力を含めたデジタル化を検討している。

5.3. 学内研究成果

現在、紀要類の記事情報（二次情報）を検索可能としている。これは学術情報センターの「目次速報データベース」事業への入力データを利用しているものであるが、相当数の紀要類で創刊号まで遡及入力が終わり、約14,000件の記事データが検索できる。また、学位論文、科学研究費報告書などの二次情報も準備中である。今後は、学内研究者の理解・協力を得て、教育・研究に係わる各種一次情報を公開していくことが課題である。

6. おわりに

以上、「電子アーカイブ」の構築について、システム設計の考えかた、システム構成と検索機能の概要、コンテンツの現状を述べた。

システムがいちおうの稼働をみたので、今後はコンテンツ作成の推進が当面の課題である。多様なコンテンツの搭載を進める過程で、今回のシステム設計の当否が明らかになり、システム面の課題も浮き彫りになってくるはずである。コンテンツ作成にも保存形式の選択や著作権処理など難しい問題が山積しているが、一歩一歩充実させていきたいと考えている。

注

[1] http://www.lib.kobe-u.ac.jp/

[2] 稲葉洋子「震災資料の保存と公開－神戸大学「震災文庫」を中心として」『大学図書館研究』55, 1999.3. pp.54-64

[3] 厳密にいうとエレメント種別には、「章・節」のような構成上のレベルのみを表すものと、「写真」のように「当該レベルの資料種別」の要素を含んだものとが混在しており、ややわかりにくくなっている。リソース種別を4段階にすることも検討したが、データ構成が複雑になりすぎるので結局現在の形となった。

[4] 杉本重雄「Dublin Core Metadata Element Set について―現在の状況と利用例」『ディジタル図書館』14, 1999.3. http://www.dl.ulis.ac.jp/DLjournal/No_14/1-sugimoto/1-sugimoto.html

[5] Dublin Core の15項目のうち、Contributor（寄与者）や Coverage（対象範囲）など、枠はあっても全く用いていない項目がいくつかある。また Publiser（出版者）、Date（日付）も定義上「現在の」形での公開者・日付となっており、一次情報を作成した場合にはその公開者（図書館）・公開日付が入るべきと考えられるので、所蔵資料の出版者・出版年月は別途フィールドを設けた。当初 Dublin Core 準拠という仕様でスタートしたが、最終的には他の項目がふくれあがり、そうは言い難い項目構成となった。

[6] その他、「関連メタデータ ID」「関連 URL」項目を設けており、他のデータやネットワーク上の情報源との水平的リンクも一応可能である（ただし、相互性チェックなどの仕組みは用意されていない）。

[7] メタデータと一次情報の1対1対応が前提なので、ページイメージの場合も、JPEG 等のページ単位画像を章・節などの意味的まとまりでイメージ PDF にまとめないとうまく表示されない。しかしイメージを何ページもまとめるとファイルサイズが大きくなりすぎて、現実には提供が難しい。ページごとに独立したメタデータを作成すれば1ファイルごとのサイズは抑えられるが、現在の検索システムでは表示が冗長になり一覧性に問題がある。今後の課題である。