震災アーカイブにおけるメタデータの設計

渡邊　隆弘 (watanabe@lib.kobe-u.ac.jp)
神戸大学附属図書館情報サービス課

神戸大学電子図書館システムにおける阪神・淡路大震災関係資料アーカイブでは、著しい多様性を持った震災資料を適切に検索するためのメタデータを重視している。本稿では、本システムで採用しているメタデータの設計・構造を述べる。特徴の一つは、資料の構成部分を詳細に記述するための階層構造表現手法であり、もう一つは十分な表現能力と柔軟性をもったデータ項目設定である。

Design of Metadata for the Hanshin-Awaji Earthquake Digital Archives

WATANABE, Takahiro
Kobe University Library

1．はじめに

平成11年より稼動している「神戸大学電子図書館システム[注1]」では、「電子アーカイブ」と称して各種所蔵資料のデジタル化を積極的に行っているが、中でも阪神・淡路大震災関係資料をコンテンツの柱と位置づけている。これは、神戸大学附属図書館で一般公開している「震災文庫[注2]」(1995年10月開設）の所蔵資料を主たる対象とするデジタルアーカイブである。近年多くの大学図書館等で所蔵資料をもとにしたデジタルアーカイブ構築が進められているが、その対象は古典籍・古文書や紀要等の学術論文がほとんどを占めており、特定テーマに沿って収集された現代資料への取組みは他にあまり例をみない。

震災アーカイブでは様々な資料のデジタル化に取り組んでいるが、システム開発段階においては、一次情報の作成・提示手段もさることながら、検索機能を果たすに十分な表現性能を持ったメタデータの設計に最も腐心した。本発表では、メタデータの設計・構造に焦点をあてて事例報告を行うこととする。

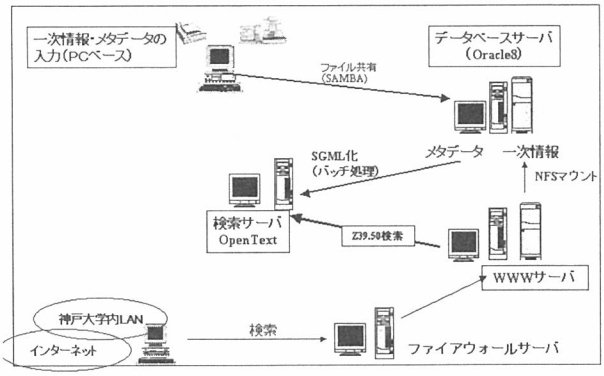
2．「電子アーカイブ」のシステム構成

本論に入る前に、「電子アーカイブ」システムの構成を略述する。本システムは、震災関係資料だけではなく、神戸大学でデジタル化して発信するその他のアーカイブをも扱う汎用的なシステムであり、一次情報とメタデータの入力・データ管理・検索・表示を行う（システム開発はNTT西日本株式会社による)。図1にシステム構成を示した。

メタデータと一次情報はPCベースで随時入力され、データベースサーバ（サーバのOSはすべてUNIX）に蓄積される。この時点のデータ管理はRDBMS（ORACLE）を用いている。

検索システムでは全文検索エンジン OpenText を使用している。このため、メタデータ（及び全文検索用テキスト）はSGML形式で格納している。SGML形式データは日次更新でRDBMS上の入力データから自動更新されている。さらにZ39.50サーバが運用されているが、今のところ外部クライアントからのZ39.50直接アクセスは許しておらず、当館で運用しているWWW-CGIインターフェースのみの検索となっている。

図１．システム構成



２．メタデータの重要性と設計の基本思想

２．１．「震災文庫」資料の特性

震災直後から関係資料の網羅的収集につとめてきた震災文庫の所蔵資料は既に 20,000 件を超え、現在も毎月 200～300 件のペースで増加を続けている。

通常の図書館資料と比較すると、震災文庫資料の特徴は「多様性」にある。図書・雑誌以外に、広報紙・チラシ・レジュメといった資料を大量に含んでおり、さらに地図・写真・ビデオ・CD といったマルチメディア資料も相当量に及ぶ。また、震災関連情報を収集するという性格上、資料の一部分のみを抽出した抜粋・抜刷・切抜なども多数を占めている。すなわち、資料の媒体・形態と資料となる単位の両面において著しい多様性をもっている。

さらに、資料自体の特性ではないが、震災関係情報には他の分野で利用されているような網羅的な記事索引や二次情報データベースが存在していないという事情がある。震災文庫では公開と同時に WWW サイトを開設して目録情報の提供を行ってきたが、もっぱら資料タイトルレベルの書誌情報を記述する従来の図書館目録の枠内ではきわめて不十分と言わざるを得なかった。図書館目録は各主題の様々な二次資料（書誌・索引・事典など）との連携で最終的に利用者を資料に導いているものであり、そうした連携先が十分でない場合には、利用者の情報要求を受け止めきれないのは当然である。

２．２．メタデータの重要性

電子図書館システムでは、デジタル化された一次情報が地理的時間的制約を超えて提供されることが重要なことはいうまでもなく、震災文庫でも著作権許諾処理を行いながら積極的な取組みを進めている。しかし一方、震災文庫の目録情報には前節で述べたような問題があることから、レファレンスデータベースとして使えるだけの、十分な検索機能を持ったアーカイブとすることも欠かせない要件であった。開発にあたっては、以下にあげる事情を考慮した結果、メタデータの構築とその検索・表示を最も重視することとなった。

・全資料デジタル化の困難性

著作権許諾処理が必須であることから、近い将来にすべての資料の一次情報をデジタル化できる見通しはたたず、また必ずしもニーズの高い順に処理していけるわけでもない。レファレンスデータベースとしてはデジタル化したものと紙媒体等のまま残る資料とが統合検索できなければ意味がなく、統一的な仕様に基づくメタデー

タを作成するのが適当である。

・全文検索の不確実性

一次情報がデジタル化できれば全文検索も可能であるが、十分な精度と再現率をもった検索は難しく、適切に構築されたメタデータのほうが信頼性ある検索ができると考えられる。

・マルチメディア情報の存在

震災情報では、写真・地図・映像などのマルチメディア資料も重要性が高い。このような資料の検索にはメタデータの存在が必須である。

2．3．メタデータ設計の基本思想

2.1.で述べた震災資料の特性を踏まえて、「どのような検索機能が必要であるのか」、さらに「その機能を果たすにはメタデータにどのような要件が求められるのか」を検討した結果、次の2点を重視することとなった。

・構成部分の詳細な記述と管理が可能なこと

雑誌・広報紙の論文・記事レベルや図書の章節レベル、写真資料なら一枚一枚の写真レベルの情報からも検索できなくてはならない。また、図書などの資料中に掲載される図表・統計表・地図・写真等が独立して情報要求を満たすことがあり、そうした情報も検索対象とすることが望ましい。資料に含まれる構成部分の記述が十分精細なレベルまで可能なメタデータ構造が必要である。

この結果、データベース中に様々な情報単位の記述が混在することとなるが、ピンポイント検索や資料の構造・文脈に沿った表示など多様なアクセスが実現できるよう、各単位の適切な弁別と管理が必要である。

・多様性にも対応した必要十分なデータ記述が可能なこと

メタデータが十分な表現性能を持つように、データ項目が適切に設定される必要があるのは当然である。特に、媒体や情報単位の多様性を考えると、対象となる資料・情報の種類によって最適なデータ項目を設定できる柔軟性が求められる。また、日々増加する震災資料には当初予想しえない種類の情報も現れる可能性があり、データ項目の追加・修正がシステム改造を伴わずに可能なことが望ましい。

図2．検索結果一覧画面（部分）

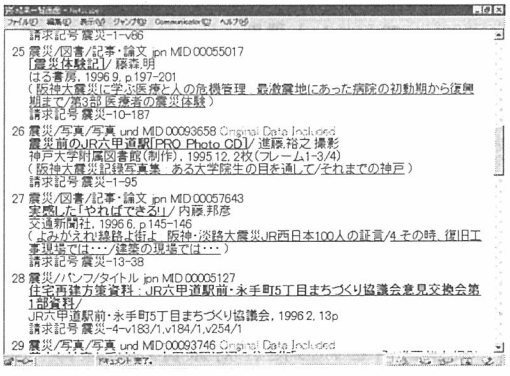
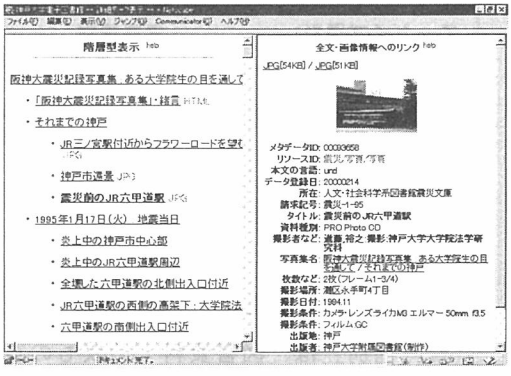


図3．メタデータ詳細表示画面（部分）



次章から、以上の点を踏まえて設計したメタデータの実際を述べる。参考のため、本システムにおける検索画面の例を図２，３にあげた。

３．レベルごとのメタデータ作成と階層構造表現

３．１．レベルごとのメタデータ作成

資料の構成部分について、精細なレベルまで十分なデータ記述を行うために、個々の構成要素ごとにそれぞれ独立したメタデータレコードを作成することとした。例えば、雑誌資料であれば、一つの「資料タイトル（誌名）レベル」、冊数分の「分冊巻号（各号）レベル」、さらに個々の「記事・著作レベル」がそれぞれ作成される。検索上の必要に応じて、「ブロックレベル（"特集"のように複数記事をまとめたレベル）」や「章節レベル」が作成され、さらには掲載された写真・図版等を「写真レベル」「図表レベル」などの形で扱うことも行われる。すなわち一つの資料から、必要と思われる限りにおいて、多数（写真資料などでは１資料で1000を超えるものもある）のメタデータレコードが作成されうる。

これらのメタデータレコードは検索システム上等価に扱われ、検索結果一覧（図２）でもメタデータ詳細表示（図３の右フレーム）でも独立して表示される（表示を自己完結させるだけのデータ項目を持っている）。これにより、必要なレベルの情報をピンポイント的に探索することが可能となっている。

３．２．「リソース種別」による管理

レベルごとにメタデータを作成すると、データベース中には様々な情報単位のレコードが混在することになる。これらの統合検索には大きな利点があるが、一方で検索結果の単位がばらばらでわかりにくいなどの問題が発生する。

このため図２の各行冒頭にある「震災／図書／記事・論文」のように、検索結果では各メタデータレコードが「どういう資料のどういうレベルのものか」を明示している。これは「リソース種別」というコード情報によって管理する。リソース種別は図５に示したように３種類のコードよりなる。ただし、資料群の種別である「アーカイブ種別」は震災資料ではすべて同一であり（本システムは他の資料群をも扱う汎用システムであるために設けられている）、「資料種別」と「エレメント種別」がメタデータの種類を表している。

図５．リソース種別一覧

アーカイブ種別	資料種別	エレメント種別
01 震災文庫	01 図書資料	01 シリーズレベル
……	02 雑誌資料	02 資料タイトルレベル
その他アーカイブ名称	03 新聞・広報紙資料	03 分冊巻号レベル
	04 パンフレット資料	04 ブロックレベル
	05 一枚もの資料	05 記事・著作レベル
	06 地図資料	06 章・節レベル
	07 写真資料	07 写真レベル
	08 映像資料	08 地図レベル
	09 音声資料	09 図表レベル
	10 コンピュータ資料	……

「資料種別」は、図書資料・一枚もの資料・映像資料といった、対象資料の媒体・形態を表すコードである。当該資料全体をとらえての区分であり、同一資料から作られた各レ

ベルのメタデータの資料種別は同じになる。

「エレメント種別」は、当該メタデータのレベルを表すコードである。厳密にいうと、「シリーズ」「分冊巻号」「章節」のような構成上のレベルのみを表すものと、「写真」「図表」のようにレベルというより「当該構成部分データの資料種別」ともいうべきものとが混在しており、理論上は明快でない。両者とも重要な情報であり、リソース種別を４コード構成にすることも検討したが、構造が複雑になりすぎるので現在の形で実装している。

リソース種別の設定によって各メタデータレコードの種類が明らかになり、利用者の助けとなるとともに、種別を限った検索や一覧リスト作成などの処理にも役だっている。また、4.4 で述べるメタデータ項目の柔軟性という点においても、設定に関わる基本情報である。

３．３．階層構造情報の保持

各レベルのメタデータレコードがばらばらに存在していては、構成部分への検索は可能であっても、1 資料の全体像がとらえられない。ピンポイントアクセスの一方で、図３の左フレームにあるような資料の文脈・構造を再現する表示も必要である。すなわち「目次」に相当するようなツリー構造表示である（本システムでは「階層型表示」と呼ぶ）。

本システムでは、1 資料から各々独立して作られたメタデータレコード群に、相互の階層構造情報を持たせている。ここで必要な階層構造情報とは、メタデータ間の上下関係情報と、同位メタデータ内での順序情報である。具体的には図４にある〈TOPMID〉（ツリー最上位 ID）と〈LEVEL〉（ツリー内階層パス）により実装している。

３．４．階層構造と検索

資料内の各構成要素を独立したメタデータレコードと扱うと、階層をまたがった検索に問題が生じることがある。

例）　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）

このようなツリー構造の場合に、「神戸大学 AND 附属病院」ではタイトルレベルレコードも第２章のレコードも単独では条件を満たさず検索されないかもしれない。これを防ぐため、ツリーの上位階層にあたる部分のタイトルを下位レコードに埋め込むなどして上位情報からも検索を可能にしている。

ただし一方で、上位情報から検索できるがゆえに「神戸大学の被災状況」で各章・節のデータまでが別個にヒットしては、一覧表示が混乱して正確なヒット件数もつかめなくなる。このため、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んで表示するという処理を行っている。

４．メタデータ項目の設定と柔軟性

４．１．データ項目の設定

本システムの基本設計を行ったのは 1998 年であるが、当時すでにダブリン・コア[注3]（Dublin Core Metadata Element Set　以下、DC と略）がネットワーク情報資源のための

メタデータ記述規則の標準になっていくとの見通しがあり、メタデータ項目の設定もDCを検討するところから出発した。しかし、レファレンスデータベースとして必要十分なデータ項目を確保したいという要件も一方にあり、必要な項目を洗い出していった結果、非常に多くの項目となった（図４にSGML形式メタデータ記述例をあげた）。

図4. SGML形式メタデータの例（省略したところがある）

```
<KUMETATBL>
 <MANAGETBL>
  <METAID>00093658</METAID>
  <AID>01</AID>
  <DID>06</DID>
  <EID>07</EID>
  <CREATEDATE>20000214</CREATEDATE>
  <TOPMID>00093654</TOPMID>
  <LEVEL>00032719_00000020_00000030</LEVEL>
  ……
 </MANAGETBL>
 <METATBL>
  <JA>
    <TITLE>震災前の JR 六甲道駅<TRANSCRIPTION>しんさいまえの jr
ろっこうみちえき</TRANSCRIPTION></TITLE>
    <CREATOR>進藤,裕之@@@撮影@@@神戸大学大学院法学研究科</CREATOR>
    <LANGUAGE>und</LANGUAGE>
    <IDENTIFIER>V1003.jpg</IDENTIFIER>
    <IDENTIFIER>V1004.jpg</IDENTIFIER>
    <ORGPUBLISHER>神戸大学附属図書館</ORGPUBLISHER>
    <PUBLISHDATE1>1995.12</PUBLISHDATE1>
    <EXTENT>2 枚(フレーム 1-3/4)</EXTENT>
    <TREE>阪神大震災記録写真集：ある大学院生の目を通して@@@それまでの神戸</TREE>
    <EX16>灘区永手町 4 丁目</EX16>
    <EX17>1994.11</EX17>
      ……
   </JA>
   <EN>
    <TITLE>JR Rokkomichi Station prior to the earthquake</TITLE>
    <CREATOR>Shindo, Hiroyuki…</CREATOR>
    ……
   </EN>
  </METATBL>
</KUMETATBL>
```

メタデータ項目は、３つの部分よりなっている。「管理部」（図４の〈MANAGETBL〉）はレコードID・リソース種別・階層構造情報などの管理情報で、ほとんどの項目が入力必須でかつ繰り返しは許されない。「管理部」を除いた部分が「データ部」（図４の〈METATBL〉）で、DC15項目と追加した約20項目からなる「共通項目」と、特定のリソース種別に対して使用される「拡張項目」より構成されている。「データ部」では同一項目の不定繰り返しが許される。

４．２．メタデータ項目とダブリン・コア

本システムのメタデータ項目には上述のようにDCの15項目が含まれてはいるが、「DC準拠」とは言い難い状態である。追加項目が膨れ上がっているうえに、15項目のうちで実際上使っていない項目も多くある。

「共通項目」に追加された項目は、版表示・出版年・数量・大きさ・請求記号など図書館目録で用いられてきた諸要素が多くを占めている。DCがネットワーク情報資源を主たる対象としているのに対して、大半の震災資料の背後には（一次資料がデジタル化されているかどうかにかかわらず）図書などのパッケージ型資料があることから、DCにそのまま依拠することは難しいように感じられた。また、DCはリソースの著者がメタデータを作成することも想定して基本的な項目に絞り込んでいるが、情報専門家である図書館によるメタ

データはより詳細な記述を目指すべきではないかという考えもあり、本システムでは多数の項目を独自に設定する道をとった。

一方、システム設計時点での DC がまだ発展途上にあったことから、必要なデータ項目とのマッピングが十分にとれなかったという面もある。例えば当該データがどの場所・地域を対象としているかを統一書式で収める項目「対象地名」は、DC の「Coverage（対象範囲）」に該当する内容かと思われるが、当時は Coverage の記述方法に関する情報が十分得られず、追加項目として収めることになった。

４．３．日本語・英語のメタデータセット

震災関連情報を海外に情報発信することも重要と考えており、英語版検索システムの提供も行っている（欧文資料やタイトルが日英併記の資料のほか、写真資料のキャプションを英訳して海外からのアクセスを可能にするといった事業も行っている）。

本システムのメタデータでは、日本語版と英語版のメタデータセットを二重に持つ構造としている（図４の<JA>…</JA>と<EN>…</EN>）。SGML タグで明示しているため、検索とデータ返戻の双方においてどちらか（もしくは両方）を対象とするシステムを容易に組むことができる。実際には日英いずれでも同じ値が入るフィールドもあるが、入力時には日本語データを流用して英語データを作成できるようにしている。

４．４．データ項目設定の柔軟性

写真資料や図書中の写真レベルデータには「撮影場所」「撮影日付」「撮影条件」が必要になるなど、特定の「リソース種別」にのみ必要となる項目もあり、「拡張項目」と呼んでいる。多様性が著しく、かつ今なお日々増加し続ける震災資料を考えたとき、最初に必要なデータ項目をすべて洗い出すことは困難と思われ、システム設計にあたっては拡張項目部分が運用後にも随時に追加できる柔軟性を持つように考慮した。

具体的には、あらかじめ多数の拡張項目を適当な名前でメタデータ用 DTD に登録しておき、実際上の項目定義は外部ファイルで維持して随時更新を可能としている。この項目定義ファイルはリソース種別ごとに作ることができ、存在しうるメタデータ項目を列記する。

さらにこの定義ファイルでは、共通項目も含めて、項目の出現順序（メタデータ詳細表示での順序）や表示ラベル名称、検索/表示対象とするかどうかなどが記述でき、検索・表示系のさまざまな条件を操作できるようになっている。

５．課題とまとめ

以上、神戸大学電子図書館システムにおける震災資料のメタデータについて述べてきた。

有効な検索を行うためには、資料の持つ論理構造を十分反映した形式で電子図書館コンテンツ（一次情報）を作成することが望ましく、SGML、XML 等のマークアップ言語が注目される所以である。しかし、震災資料のように著しい多様性を持つアーカイブにおいては、実効ある DTD を設計することは極めて困難である。本システムで採用した階層構造表現の手法は、その代替として、ある程度の文書構造を保持して検索システムに生かしていく試みと考えている。

本システムのもとで実際にデータ登録をはじめてから約１年半であるが、現在約 20,000

件の資料に対して約45,000件のメタデータレコードが作られている。日々の増加資料に対するデータ作成とともに、精細レベルの情報を集中的に入力する事業も行っている。

ある程度の実践を経て、メタデータ設計上の問題点もいくつか明らかになってきた。ここでは2点を述べる。

一つは、SGML形式でのデータ保持を十分生かしきれていないことである。現状のメタデータは「データ部」内のレベルで考えると、一部の例外（ヨミ情報）を除いて平板に項目が並列しているだけでさらなる構造を持たない。例えば、現在「著作者<CREATOR>」項目には著者名・役割表示・所属の3要素を入力しているが、これらを1フィールドに（適当な区切り記号で）収めている。このため「神戸大学法学部」で検索すると、法学部が編集した資料以外に法学部所属教官の著した資料までがすべてヒットし、それ以上絞り込めない。SGMLではタグの入れ子構造が可能なので、3つの下位タグを定義することで解決できる問題である。現状では、RDBMS（ORACLE）を用いた入力システムからSGMLに変換しているので仕方ないところであるが、いずれは解消をはかりたい。

もう一つは、メタデータの標準化に対する問題である。本システムのメタデータはリソース種別のような独自のコード情報があるなど、かなり特殊な構造となっているため、汎用的情報検索プロトコルであるZ39.50のサーバを導入していながら、Z39.50での直接公開ができていない。また、4.2で述べたようにダブリン・コアとのマッピングにも問題がある。電子図書館のメタデータが直ちにDCを採用すべきだとは考えていないが、標準化された形式での出力を保証することは重要である。DC側では、コアエレメントの中で「Qualifier」を用いてさらに意味を特化させる方式が最近標準化の道を進みつつあるようで[注4]、そうした方向との整合も次期の課題である。

冒頭に述べたように、現在の電子図書館コンテンツには古典籍などが多く、震災資料は特異な例に属する。しかし、今後電子図書館化が公共図書館にも広がっていけば、地方行政資料などが取り組まれ、多様な形態からなる現代資料のデジタル化・データベース化が進展していくであろう。そうした動向も注意しながら、震災アーカイブのメタデータをより高品質なものにしていきたいと考えている。

[注]

[1]「Kobe Digital Library」http://www.lib.kobe-u.ac.jp/
渡邊隆弘「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
『ディジタル図書館』16, 1999.11
http://www.dl.ulis.ac.jp/DLjournal/No_16/1-watanabe/1-watanabe.html
[2]「震災文庫ホームページ」http://www.lib.kobe-u.ac.jp/eqb/
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
渡邊隆弘「『震災文庫』のこれまでとこれから　電子図書館を中心に」
『Academic Resource Guide』055, 2000.2
http://www.ne.jp/asahi/coffee/house/ARG/055.html
[3] "Dublin Core Metadata Initiative" http://purl.oclc.org/dc/
[4]杉本重雄「Dublin Coreについて－最近の動向、特にqualifierについて」
『ディジタル図書館』18, 2000.9