2003－FI－73　(6)
2003／11／13

# 多様な資料構造に対応したデジタルアーカイブシステム
## －神戸大学電子図書館アーカイブ検索システム－

依田 平[†1, †2]　渡邊 隆弘[†3]　大月 一弘[†4]　鳩野 逸生[†5]　岩杉 大輔[†6]

神戸大学附属図書館では，1999 年より電子図書館アーカイブ検索システムを構築し提供している．同システムで収容している資料は，(1)多種多様な資料を網羅的に収集しており，形態が不定形である，(2)一つの資料内に，独立した情報単位として取り扱うことのできる部分資料を重層的に含む複合型資料となっている，という特徴を持つ．

本論文では，このような不定形複合型コンテンツに対するアーカイブシステム構築方法について述べる．開発したシステムでは，部分資料に対してもすべてメタデータを作成し，これらを元に一つの資料に対するメタデータツリーを作成する．メタデータツリーを利用することにより，部分資料に対する検索も可能としている．更に，AND 検索の拡張，メタデータの記述ルールを準備することにより，適合率の向上を図っている．

# A Digital-Archive System for Various Compound Materials
## －Kobe University Library Digital Archive－

TAIRA YODA,[†1, †2]　TAKAHIRO WATANABE,[†3]　KAZUHIRO OHTSUKI,[†4]
ITSUO HATONO[†5]　and　DAISUKE IWASUGI[†6]

Since 1999 Kobe University Library has provided open Digital Archive, which consists of several collections to search, such as earthquake disaster materials, rare materials held by Kobe University, and bulletins. These materials, especially earthquake disaster materials do not have fixed format nor data structure. They also have a feature that a material includes sub-materials, which may be treated as an independent material.

This paper discusses about mechanisms for storing and retrieving semi-structured data for amorphous and compound materials. We give metadata for each candidate of an independent material and construct metadata tree, which represent a composite material. We show the mechanisms for searching materials by using metadata trees. Furthermore, we propose three extended AND query operations, which improves recall ratio and precision ratio of retrievals.

## １．はじめに

近年，計算機技術や通信技術の長足の進歩により，大量の文書・静止画・映像などを高品質で劣化しないデジタル情報の形で保存することが可能になっている．これに伴い，有形・無形の歴史・文化資産をデジタル化し，保存・蓄積・活用するデジタルアーカイブへの注目が高まっている[1].

1999 年７月以来，神戸大学附属図書館で公開しているデジタルアーカイブ（以下本アーカイブ）の中核コンテンツは，阪神・淡路大震災に関係する資料を公開する「震災文庫」である．同文庫に収容された資料の特徴のひとつは，震災に関するあらゆる資料を網羅的に収集したため，一般的な図書館資料と比べ非常に多様性に富んでおり，資料の形態・形式を定形化しにくい点である．また，もうひとつの重要な特徴として，ひとつの資料の中に個別の資料として扱うことの可能な構成要素を重層的に含む複合型資料が多い点である．

そこで，我々は，このような不定形複合型コンテンツに適したデータ格納手法ならびにその検索手法の開発を行ってきた．この際重視したことは，資料の種類をその都度意識せずとも統一的に検索ができることと，資料

†1　山形短期大学コンピュータセンター
Computer Center, Yamagata Junior College
†2　神戸大学大学院総合人間科学研究科
Graduate School of Cultural Studies and Human Science, Kobe University
†3　神戸大学附属図書館
Kobe University Library
†4　神戸大学国際文化学部
Faculty of Cross-Cultural Studies, Kobe University
†5　神戸大学学術情報基盤センター
Information Science and Technology Center, Kobe University
†6　インフォコム株式会社
Infocom Corporation

中に個別の資料として扱うことのできる構成要素を含んでいるものについては構成要素単位での情報の検索・抽出も行えるようにすることである．こうした機能を実現するために，メタデータを資料中の個々の構成要素ごとに作成するとともに，資料の全体構造に関する表現も構成要素に対するメタデータに記述するという管理手法をとった．

4 年間の運用の過程で得られた問題点やその後の技術進歩を踏まえて，神戸大学附属図書館では 2003 年 2 月より検索機能を強化した新システムの提供を開始した．強化した検索機能は主に AND 検索の部分である[3],[4]．個々のメタデータに対する検索によって構成要素単位での情報の検索・抽出ができるようになるものの，資料全体の内容が別々のメタデータに記述されることによって，元々は同一の資料に含まれていたキーワードが共起しなくなり，AND 検索において時に漏れを生じるという問題が発生する．旧システムでは特定メタデータの特定の値を他のメタデータにも継承させることによってこの問題に対応していたが，継承は十分でなかった．新システムでは構成要素間の関係を元にしたメタデータ間の横断検索を行うことによって，従来不十分だった，階層構造をまたがる AND 検索を確実に行えるようにした．さらにこれに加え，本来大きく二つの意味の違いがあると言われている[6]AND 検索におけるキーワード間の関係の違いを考慮した検索を行えるように，3 種類の AND 演算を準備し，検索の適合率を高められるようにした．

本論文では，アーカイブ検索システムおける，主にメタデータの作成・記述方法について述べる．開発したシステムでは，独立する資料となりうる個々の構成要素に対してすべてメタデータを作成し，これらをもとにひとつの資料に対するメタデータツリーを生成する．メタデータ間の横断検索には，このツリーを利用する．また，様々な問い合わせに幅広く，かつ，効率的に対応するための，メタデータの記述ルールを定める．

第 2 章では本アーカイブのコンテンツについて述べ，その利用者要求について考察する．第 3 章では本システムの検索の特徴について述べる．第 4 章では資料構造の把握とメタデータの表現方法について述べ，第 5 章ではメタデータの検索方法について述べる．第 6 章ではメタデータの記述方法について考察する．

## 2. アーカイブに対する利用者の要求

神戸大学附属図書館では，阪神・淡路大震災関係資料，戦前期の新聞記事切抜資料，江戸期の海運海事史関係史料などの資料群をデジタル化し，神戸大学図書館デジタルアーカイブとして，1999 年 7 月から WWW 上で一般公開している（図 1 はそのトップページ）．

対象コンテンツの面からみた本アーカイブの特徴点は，前近代から現代にいたる多様な資料群を含んでいることである．わが国の図書館における資料デジタル化事業の多くは古文書等の歴史的資料に偏っており，多様な広がりに欠けていることが多い．大量の近現代資料を扱うデジタルアーカイブは，大学図書館・公共図書館を通じて，現時点では他にほとんど例がないといってよい．

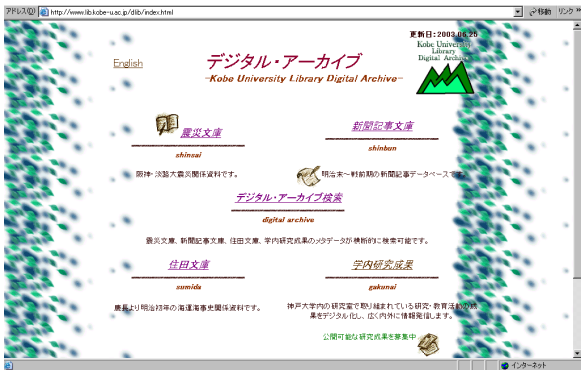


**図 1 神戸大学図書館デジタルアーカイブ**

特に，中核的コンテンツの一つである震災関係資料は，一般的な図書館資料と比べても，非常に多様性に富んだ資料群である．阪神・淡路大震災に関係する資料を網羅的に収集して公開する「震災文庫」の収集資料をベースとしており，

- 資料媒体・形態が多様である．映像・音声等のマルチメディア資料はもちろん，紙媒体資料においても，広報紙・チラシ・レジュメなどを大量に含む．
- 一部分のみが震災関連である場合，また抜刷・切抜等で収集する場合など，資料となる単位が一定しない．
- 記録を目的としない生の資料や，ミニコミ出版物などを大量に含み，内容・編集面でも多様性に富んでいる．

という特徴を持つ．

このような多様な資料からなるアーカイブに対しては，利用者の情報要求も様々である．特に，震災関係資料等においては，学術研究者から一般市民まで幅広い利用者が想定される．

まず求められることは，資料の種類をその都度意識せずとも統一的に検索ができることである．資料媒体に関わりなく，あるテーマに関する情報を横断的に検索できなくてはならない．

テーマの広狭，必要とされる網羅性の度合いなど，それぞれの要求に対応できることも重要である．この際，特に留意すべきは，様々な粒度の情報が求められるということである．利用者があるテーマの情報を網羅的に得たい場合，またごく狭いテーマの情報を得たい場合には，図書の書名など「資料」として物理的に独立した情報単位だけではなく，記事・論文・章・節といった構成要素の単位まで検索が行えなければ，十全な結果は得られない．場合によっては，広報紙上の小さなコラムや，1 枚の写真・図表などで，利用者の問題解決が果たされることもある．


**図２　検索結果一覧表示画面**

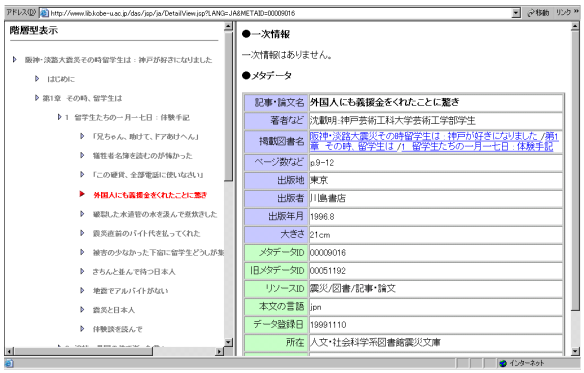


**図３　データ詳細表示画面**

## ３．アーカイブ検索システムの特徴

神戸大学電子図書館アーカイブ検索システム（以下本システム）では，「デジタルアーカイブ検索」と名付けた検索画面で，資料群を問わない統一的な検索機能を提供している．2.1 で述べた資料の多様性を十分に表現し，2.2 で述べた多様な利用者要求に応えることを目的として機能設計がなされている．大きな特徴は，

・メタデータを介した検索であること

・資料中の様々な粒度の構成要素を検索対象としていること

の２点である．

内容全文がテキストデータ化されている資料もあるが，検索の基本は図書館で入力されたメタデータに対して行われる（オプションとして，本文まで検索することも可能）．震災資料など，内容・編集面で多様性の著しい資料群を扱う本アーカイブでは，学術論文など一定の均質性を備えた資料群と異なり，全文検索で実用的な精度を得ることは困難であると判断し，メタデータを前面に据えた検索仕様としている．

資料中の構成要素を検索対象とする点については，検索結果画面の例を示して述べる．

図２は，検索結果の一覧表示画面である．本画面では，資料中の，検索要求に合致すると判断された構成要素を単位とした表示がなされる．すなわち，図書中の記事や写真など，資料中の特定部分がマッチしたと見なされれば，資料全体のタイトルではなく，当該構成要素の簡略メタデータが表示される．同一資料から，複数の構成要素が独立して表示される場合もある．資料中のより大きい部分（「第１部」など）がマッチしたと見なされればその部分が，当該資料が全体として検索要求に合致するなら資料全体が，結果一覧表示の単位となる．これにより，求める情報のピンポイント検索を可能としている．

なお，結果として様々な粒度の情報が混在して一覧され，またそもそも多様な資料からなるアーカイブであることから，各表示に「震災／パンフ／記事」（震災資料のパンフレット資料中の１記事）のように当該情報の種別を明示している．

図３は，結果一覧表示から一つを選択した際に表示される，データ詳細表示画面である．画面分割されているうち，右フレームは選択されたデータ（多くは，資料中の構成要素）の詳細メタデータと一次情報へのリンクを表示する．一方左フレームでは，当該資料の各構成要素のタイトルを，インデンションを施しながら「目次」のように並べた「階層型表示」を行っている（赤くハイライト表示されているのが，現在選択されている構成要素）．検索結果一覧表示では，ピンポイント検索のために資料中の構成要素がダイレクトに表示され，その分資料の全体像は把握しにくい．対して「階層型表示」では，各資料の資料構造が明示され，前後や上位の構成要素などをブラウジングすることが可能となっている．

## ４．資料構造の把握とメタデータ表現

### 4.1. 資料構造の把握

資料はその中に様々な要素を重層的に含んでいる．例えば，記事，章・節，写真，図などで，一つの資料はこれらの構成要素を重層的に包含して成り立っている．本アーカイブでは，これらの構成要素をそれぞれ独立した情報単位として取り扱い，この情報単位を検索・操作の対象とすることによって，資料中の一部分を検索結果として提示する．以下では，独立した情報単位として取り扱える構成要素を**部分資料**と呼ぶ．また，部分資料が資料の中でどのような構成要素として扱えるかを，資料構造上の**レベル**と表す．

### 4.2. メタデータの作成単位

本アーカイブでは，すべての部分資料に対して，それぞれを表すメタデータを作成することを原則とする．

図４，５に資料例とメタデータ記述例を示す．図４は「六甲くらし通信」という広報誌を模式化した図である．図中の点線部が部分資料に当たり，それぞれの部分資料に対してメタデータが作成される(「六甲くらし通信」全体については点線を省略しているが，この単位にももちろんメタデータが作成される)．図５は図４の中の「子供の遊び場つくり」という部分資料のメタデータの記述例である．メタデータの記述ルールは６章で後述する．

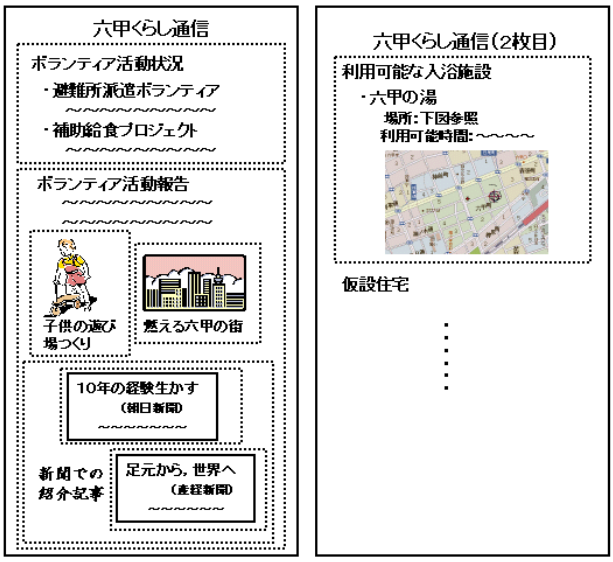


**図４　資料例（広報紙）**

### 4.3.メタデータツリーの作成

図４では部分資料がさらに細かな部分資料として認識されている．この場合，両者は全く無関係に存在するのではなく，上位の部分資料が下位の部分資料を包含しているという関係をとらえないと，適正な検索処理が行えない．また，各部分資料は資料全体（例では「六甲くらし通信」）を最上位とする階層構造の中に認識できるが，資料の全体構造を利用者に提示することも必要である．

このことから，メタデータにおいては，各部分資料の記述を行うだけでなく，部分資料間の包含関係を管理し，資料構造が把握できることも重要である．本システムでは，包含関係をもとに，各メタデータをツリー構造グラフに関連付け，資料ごとにメタデータツリーを作成する．メタデータツリーの各ノードが各部分資料のメタデータに当たる．このツリーによって，資料の全体像や包含関係に基づいた部分資料の資料構造上の位置が把握可能になる．

図６に本システムのメタデータツリーの概念図を示す．ツリーのノードが各部分資料のメタデータに当たるが，この図においてはメタデータに記述される内容を省略しており，実際には図５で示されたような内容が各ノードに記述されている．

### 4.4.メタデータツリーの表現

前節では，メタデータツリーの概念図を示したが，ここでは実際の表現方法について述べる．

システムに格納されるメタデータはXMLで記述されており，メタデータファイルは部分資料ごとではなく，１資料単位で作成される．このファイルの中に，各部分資料の情報やツリー構造を保持するための情報が記述されている．

図７にメタデータのツリー表現例を示す．図中の

```
<NODE>
  <SYSINFO>
      <METAID>00000074</METAID>
      <LEVEL>00000073</LEVEL>
                .
                .
                .
  </SYSINFO>
  <DATA>
    <JA>
      <TITLE>子供の遊び場つくり</TITLE>
      <CREATOR>六甲被災者ネットワーク</CREATOR>
      <ORGPUBLISHER>六甲被災者ネットワーク</ORGPUBLISHER>
      <PUBLISHDATE1>1995.11</PUBLISHDATE1>
                .
                .
                .
    </JA>
  </DATA>
</NODE>
```

**図５　メタデータ記述例**

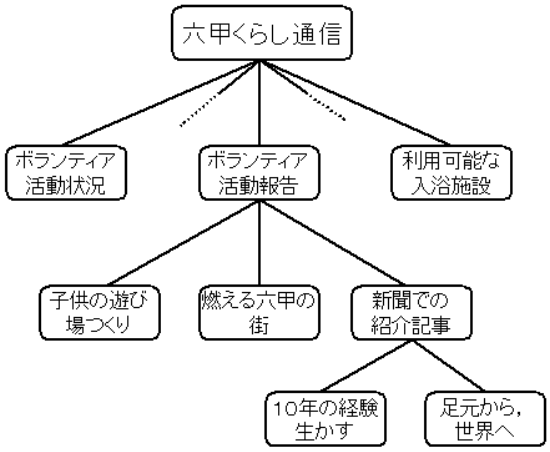


**図６　メタデータツリーの概念図**

```
<MATERIAL>
  <S>
      <NODE>……</NODE>
      <S>
          <NODE>……</NODE>
          <S>……</S>
          <S>……</S>
                .
                .
                .
      </S>
      <S>……</S>
                .
                .
                .
  </S>
</MATERIAL>
```

**図７　メタデータツリーの表現例**

**MATERIAL** は資料全体を表すタグ，**S** は資料の包含関係（構造）を保持するために各部分資料に与えられるタグである．**NODE** は各部分資料の情報を保持するためのタグで，各 **S** には必ず一つの **NODE** が存在する．**NODE** の内

容は図５で示されたような内容である．

メタデータツリーの概念図とメタデータツリーの表現においては，若干の違いがあるが，次章以下では，「メタデータツリー」は図６の概念図を，「メタデータ」は概念図における各ノードを指すものとする．

## ５．メタデータツリーの検索

本章では，４章で述べたメタデータツリーに対して，必要な情報のピンポイント検索を行う方法について説明する．なお，本システムでの検索演算については，[5]で詳しく述べている．

### 5.1.個々のメタデータに対する検索

本システムではまず，個々のメタデータに対して次の検索を用意しており，部分資料を単位とするピンポイント検索が行える．

**(1)単純検索**

単一キーワードによる単一メタデータに対する検索で，これは利用者の入力したキーワードが記述されているメタデータを求める検索である．

**(2)単純ＡＮＤ**

複数キーワードによる単一メタデータに対する検索で，これは利用者の入力したキーワードの組が記述されているメタデータを求める検索である．

### 5.2.ＡＮＤ検索の拡張

#### 5.2.1.拡張の必要性

ある資料に重層的に含まれる各部分資料，即ち１メタデータツリーに含まれる各メタデータは，資料構造の中で互いに関係を持ちながら存在しているので，前節で述べた単純な検索には問題がある．具体的には，AND 検索において，各キーワードがメタデータツリー上の別々のメタデータに分かれて出現する場合も考慮しないと，検索漏れを生じることがある．ただし，２つのメタデータが，ツリー上で上位・下位の階層関係をなしている場合と，同一ツリー内とはいえ遠く離れた位置にある場合とでは，両者の関係の密接さは同じでない．このため，メタデータツリー（即ち資料全体）を単位として AND 検索を行うという単純な拡張では，著しいノイズを生じる危険性が高い．

本システムでは，検索者の意図と資料構造（メタデータ間の関係）の双方を勘案していくつかの検索方法を使い分ける方式を採用している．

#### 5.2.2.キーワード間の意味の違いとキーワードの分布

AND 検索におけるキーワード間の関係には，本来大きく二つの意味の違いがあると言われている[6]．一つは２語が修飾関係になる場合（例えば「神戸 AND 被害」），もう一つは２語が並立関係になる場合（例えば「神戸 AND 大阪」）である．通常の AND 検索では，これらの関係の違いは考慮されずに単に同一の AND 検索として処理されているが，本システムでは両者を区別し，メタデータツリーに対して異なった処理を行う．

**(1)修飾関係の AND 検索**

２語が修飾関係にある AND 検索では，同一のメタデータに両キーワードが含まれる場合のほかに，一方を含むメタデータと他方を含むメタデータがメタデータツリーにおいて階層関係にある場合も適合の可能性が高いと考えられる．例えば，

・「神戸」を含む「章」のメタデータの下位に，
　「被害」を含む「節」のメタデータ
・「被害」を含む「章」のメタデータの下位に，
　「神戸」を含む「節」のメタデータ

が存在する資料はいずれも「神戸の被害」に関する資料である可能性が高い．この際，両者の順序を考慮する必要はない．

一方，同一資料内とはいえ階層関係にないメタデータに両キーワードが分散している場合には，適合の可能性は薄く，ノイズと見なしたほうが適切である．

**(2)並立関係の AND 検索**

２語が並立関係にある AND 検索では，それぞれのキーワードを含むメタデータが同一メタデータツリー上に存在すれば，両者が階層関係になくとも適合の可能性がある．例えば，「神戸」を含む「章」のメタデータと「大阪」を含む「章」のメタデータがともに存在すれば，その資料は「神戸と大阪」に関する資料といえる．ただし，非常に多数の部分資料からなる巨大なメタデータツリーを考えると，あまりに遠く離れた位置にある場合はノイズの可能性も高く，一定の「近傍性」を考慮するのが望ましい．

#### 5.2.3.拡張 AND 検索

本システムでは，前項で述べた２つの AND 検索に対応するため，修飾関係に対応する「直列 AND」と，並立関係に対応する「親戚 AND」と「兄弟 AND」という拡張 AND 検索を用意している．

**(1)直列 AND**

修飾関係に対応する AND 検索として，直列 AND を用意する．この検索では，直列的な階層関係をなすメタデータ群（最上位から最下位まで）を対象に，キーワードの共出現を検出する．

直列 AND のメタデータ及びメタデータツリーの操作は次の通りである．まず，各キーワードについて単純検索を行う．双方の結果をメタデータツリー上に突き合わせたものから，直列 AND は，それぞれのキーワードが記述されているメタデータがメタデータツリーにおいて同一パス上に存在する場合，最も下位に当たるメタデータを求める．下位に位置するものを検索結果とするのは，下位に位置するもの

が両キーワードを満たす部分資料と評価できるからである．

**(2)親戚 AND**

並立関係に対応する基本的な AND 検索として，親戚 AND を用意する．この検索では，資料全体を対象に，キーワードの共出現を検出する．

親戚 AND のメタデータ及びメタデータツリーの操作は次の通りである．各キーワードの単純検索の結果をメタデータツリー上に突き合わせたものから，親戚 AND は，入力キーワードが記述されているそれぞれのメタデータがメタデータツリーにおいて共通の上位ノードを持つ場合，最近の上位ノードに当たるメタデータを求める．それが両キーワードを満たす最小の部分資料と評価できるからである．資料全体に対する AND 検索でありながら，検索結果が部分資料となること，キーワードが含まれる部分資料の相対的な位置関係に基づいてのみ結果の選定が行われること，が特徴的な点である．

**(3)兄弟 AND**

並立関係の AND 検索の特別な場合として，兄弟 AND を用意する．この検索では，メタデータツリーにおいて共通の親ノードを持つメタデータ群を対象に，キーワードの共出現を検出する．キーワードを含むそれぞれの部分資料が資料構造上において同一レベルにあるとき，つまりキーワードが記述されたメタデータが共通の親ノードを持つ場合に，並立関係が特に強く表れると考えたからである．

兄弟 AND のメタデータ及びメタデータツリーの操作は次の通りである．各キーワードの単純検索の結果をメタデータツリー上に突き合わせたものから，兄弟 AND は，入力キーワードが記述されているそれぞれのメタデータがメタデータツリーにおいて共通の親ノードを持つ場合，その親ノードに当たるメタデータを求める．

### 5.3.検索結果の集約

以上の検索では，ある検索条件に対して，同一メタデータツリーに属する複数のメタデータが検索結果となりうるが，全てを表示すると冗長に過ぎる場合がある．そこで本システムでは，各種の検索で得られた結果集合に対して，メタデータツリーに基づく再評価を行っている．具体的には，

・上位ノードがヒットしていた場合には下位ノードは表示しない
・共通の上位ノードを持つ下位ノードが多数ヒットしていた場合には，共通上位ノードに代表させる

といった検索結果の集約を行っている．

### 5.4.各種 AND 検索の使い分け

単純 AND，直列 AND，親戚 AND，兄弟 AND のうち，直列 AND をデフォルトとしており，検索画面の入力欄に複数キーワードを空白で区切って羅列し検索すれば直列 AND が行われる．これは，2 語以上のキーワード入力では，両者が修飾関係にある場合が最も多いからである．一方，その他の検索は，特定の演算子を用いて検索式を表現することで行うことができる（複数の検索を組み合わせた検索式も可能）．なお，本システムでは詳細な検索条件が指定できる ”詳細検索 ” も用意しており，ここでは各種 AND 検索はプルダウンメニューから指定することができる．

## ６．メタデータの記述方法と特徴

本章では，本システムにおけるメタデータの記述ルールと，このルールを取ることによって生じるアーカイブの作成時と検索時の特徴を述べる．

### 6.1.メタデータの記述に対する要件

メタデータ記述ルール策定にあたっては，次の諸点を重視した．

**(1)**部分資料の情報を詳細に記述できるよう，必要十分なデータ項目を備えること．

**(2)**記述対象の種別ごとに適切なデータ項目を設定するとともに，種別によらない統一的な取り扱いも可能であること．ここでいう「種別」には，資料群の種類，資料媒体の種類，部分資料の資料構造上のレベル，などの諸側面がある．

**(3)**適切な部分資料の提供と資料構造の把握を可能とする情報を管理すること．

本稿では，上記(1)に関する個々の記述項目（「タイトル」「著者」...）の必要性や記述文法には踏み込まず，(2)(3)に対応する全体的なルールについてのみ述べる．

### 6.2.メタデータのデータ記述ルールと特徴

部分資料ごとに作成される各メタデータには，対応する部分資料に関する情報(以下内容情報)と，資料中での当該部分資料の位置を示す情報（以下構造情報）を記述する．ここでいう構造情報とは，検索処理を高速化させるために各メタデータに付与している情報である．これが表す構造は，4 章で述べた **S** タグ等を利用したメタデータツリー表現と同じである．本章では，内容情報に限って記述を行う．

内容情報の記述に関しては，原則として以下のルールを用いる．

**□ルール１：当該部分資料"固有"の内容を記述する．**

情報要求に合致した部分資料をピンポイントで提示できるよう，各メタデータには，当該部分資料に固有の内容情報を必要十分に記述する．ここでいう，固有の内容情報とは，部分資料を独立した一つの資料としてみた場合に，その部分資料の各データ項目に対して付与され

る値のことを指す.

この記述ルールは著者名の扱いにおいて特に特徴的な点を持つ．例えば，ある著者が記述した資料の中に，参考資料として別の著者の文章が転載されている場合を考える．転載部分を独立した資料として見ると，この部分の著者は，資料全体の著者ではなく，あくまでも参考資料の著者である．従って，転載部分のメタデータには，資料全体の著者名ではなく，参考資料の著者名を記入する．このことから，以下のような特徴的な検索が行える．

**(1)単一著者名による検索**

その著者の著作を探せるだけでなく，ある資料に転載されているその著者の記述も探せ，部分資料単位で結果を提示できる．

**(2)著者名での AND 検索**

「著者 A AND 著者 B」という AND 検索によって著者 A と著者 B の共著を探せるだけでなく，一方の著者の著書における他方の著者の記述した部分も探せ，その部分単位で結果を提示できる．

**(3)著者名と著者名以外のメタデータ項目に対応するキーワードによる AND 検索**

「著者 A が記述した B に関する情報」という検索要求における「著者 A　AND　B」という AND 検索によって，著者 A の著書が探せるだけでなく，資料全体の中から著者 A が B に関して記述した部分も探せ，その部分単位で結果を提示できる．

このように，部分資料のメタデータにもその部分固有の著者名が付与されることから，資料全体の中から利用者が要求する著者の記述した部分が得られるため，利用者の要求に該当する部分がピンポイントで提供できる．様々な情報を寄せ集めたような資料を多く扱うアーカイブにおいては，部分部分によって著者名が異なるといったことが多く考えられるため，利用者に対して部分資料を適切に提供するためにこのデータ記述ルールは特に有効であると考えられる．

**□ルール２：各種別によらず，同種の内容には単一のデータ項目を用いる．**

例えば，標題を表す情報はすべてデータ項目「TITLE」に記述する．資料媒体ごとに「書名」「雑誌名」等を，あるいは構造上の位置に応じて「シリーズ名」，「資料タイトル」，「章タイトル」等を別個に設定することはしない．

同種の内容が単一のデータ項目に記述されていることは，検索処理上大きな利点である．利用者の多くは種別を限定しない横断的な検索を志向するからである．例えば，「A に関する B という資料」を欲するとき利用者の多くは，「章タイトルに A，節タイトルに B が入っている資料」を探すといった検索ではなく，「章タイトルでも節タイトルでもどこでもかまわないので，A に関する B という関係で A と B が入っている資料」を探すといった検索を行う．

一方，資料種別を限定した検索を行いたい利用者には，資料種別を表すデータ項目として「リソース種別」を別途設けることによって対処する．「リソース種別」は，「アーカイブ種別」（「震災資料」など資料群の種類），「資料種別」（「図書」「地図」「新聞」などの媒体種別），「エレメント種別」（「資料タイトル」「章・節」など資料構造上のレベル）という３種類のコードから成っている．例えば，検索対象を「章・節タイトル」に限定したいといった検索要求に対しては，エレメント種別が「章・節レベル」の部分資料のみを検索することで対処する．

「リソース種別」は検索結果表示やメタデータ作成者への指示画面において，データの種類を明示したり，種別にふさわしい表示項目名を用いたり，と各局面で生かされている．

また，エレメント種別の付与においては，固定的な階層構造の要求や，あるエレメント種別の下位には特定のエレメント種別のみを付与するといった制約を設けることは行っていない．例えば，第１層：シリーズ名，第２層：タイトル，第３層：章などの制約は設けないし，「資料タイトルレベル」に対応するメタデータの下位にあたるメタデータのエレメント種別には「章」しか付与できないといった制約は一切ない．このため，メタデータ作成者が複雑なデータスキーマなどを考えなくてもよいという利点もある．

**□ルール３：あるメタデータツリーにおけるあるメタデータ項目において，値が同じになるものは，下位のメタデータには記述しない．**

一部のメタデータ項目においては，特定の資料構造レベルに帰属するものがある．例えば，出版者・出版年月等の出版事項などで，これらは「資料タイトルレベル」のように，物理的に独立したレベルに本来帰属するものである．「記事・論文レベル」のような下位の構成要素にはとっても関係するには違いないが，当該部分資料固有の情報とはいえない．本システムのメタデータ記述では，こうした情報は本来帰属するレベルにのみ記述を行い，下位レベルには記述しない．

この場合，「1998 年に発表された，トリアージに関する雑誌論文」のような検索要求では階層構造をまたがった検索処理が要求されるが，５．で述べた「拡張 AND 検索」により問題がない．また，図３のデータ詳細表示画面で，「記事・論文レベル」等のメタデータを表示する際には，上位に帰属する出版事項等も表示されたほうがわかりやすいが，この点もシステムがカバーしている．

上位に帰属する情報を下位メタデータに順次コピーして格納していく方法も考えられるが（検索・表示においては同じ機能が得られる），データ修正時の労力などを考えると，本来帰属すべき位置に限定する本方式に優

位性がある．

以上のデータ記述ルールより，メタデータの作成は，該当部分の上位や下位の内容情報を意識せずにそれぞれの部分で自由に行える．つまり，データ作成時において複雑なデータスキーマなどを考えなくても良いことから，どのような構造の資料にも簡便に対応できる．さらに，データの統一性や一貫性を保つためのデータスキーマの制約がないため，個別に作成したメタデータツリーを統合するのも容易である．

## 7．システム概要

本システムは，様々なアーカイブを扱う汎用システムである．図8にそのシステム概要図を示す（開発はインフォコム株式会社による）．

本システムでは，データ入力時はツリー構造のデータ入力及び管理が中心となるため RDBMS を，検索時は高速なキーワード検索が可能な全文検索エンジンを，という形でバックエンドのデータベースを使い分けている．2つのデータベース間の連携は，バッチ処理によりRDBMSからXML形式でデータを出力して全文検索エンジン用のインデックスを更新する形で行っている．全文検索用インデックスは毎日更新される．

メタデータと一次情報は入力用PCから随時入力され，データベースサーバに蓄積される．データベースシステムには MyQuick を使用し，入力クライアントアプリケーションは本システムの特性に合わせたカスタマイズを施している．

利用者への検索サービスは全文検索エンジン OpenText 7 と電子図書館パッケージ InfoLib により構築されている．InfoLib では，本システムの特徴である拡張 AND 検索を実現するために，OpenText 7 から得られた検索結果ノードリストを上位アプリケーション内でマージする処理を実装した．

XML 形式のメタデータ及び一次情報はすべてデータベースサーバ上に保持されているが，利用者への提供のために WWW サーバと共有している．ただしメタデータによっては外部に非公開となっているものもあるため，プログラム（Java サーブレット）経由でのみ閲覧できるようになっている．

メタデータを XML 形式で扱うため，利用者に提示する際は XSLT スタイルシートにより HTML 形式に変換している．スタイルシート内の制御により，文脈ごとの項目表示名称やハイパーリンクなど設定できるようになっている．

また，検索サーバでは Z39.50 ターゲット機能を持たせ，蓄積されたメタデータを Z39.50 プロトコルにより検索することも可能となっている．

## 8．おわりに

本論文では，不定形複合型コンテンツに対するアーカイブシステム構築方法について，神戸大学電子図書館アーカイブ検索システムをもとに述べた．我々は，同システムのメタデータツリーを用いる検索と，メタデータの記述ルールによって，利用者が簡単な入力で欲する部分資料を取り出せるようになったと考える．

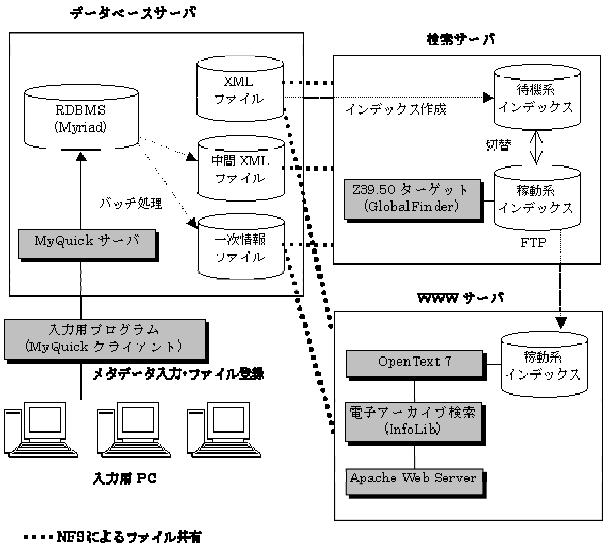


**図8　システム概要図**

## 参考文献


[1] デジタルアーカイブ白書
http://www.jdaa.gr.jp/hakusho/index.html

[2] 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.
http://www.dl.ulis.ac.jp/DLjournal/No_16/1-watanabe/1-watanabe.html

[3] 依田平, 大月一弘, 森下淳也, 清光英成, “デジタルアーカイブに対する効率的な検索の提案　―神戸大学電子図書館システムを例として―”, 情報処理学会シンポジウムシリーズ 18 号　人文科学とコンピュータシンポジウム論文集, pp.259-266, 2001.

[4] 依田平, 小椋正道, 大月一弘, 森下淳也, 清光英成, “電子図書館用デジタルアーカイブの検索方法の検討”, 情報処理学会研究報告 70 号, pp.469-476, 2001.

[5] 依田平, 大月一弘, 清光英成, 森下淳也, “ツリー型不定形文書からの部分文書の検索手法の検討”, 第 14 回データ工学ワークショップ DEWS2003, 2003.

[6] 山本昭, “ブール検索における and の使用法と意味論　―共出現の諸ケースと検索者側での対応”, 情報の科学と技術, 50 巻, 10 号, p.501, 2000.