

津波デジタルライブラリ検索システムの実装

川本健造 今井さやか 金森吉成

群馬大学大学院工学研究科 〒 376-8515 群馬県桐生市天神町 1-5-1

概要：我々はデジタル文書管理の事例研究として、津波災害に関する新聞、本、ビデオ、CGなどをWebを通して検索、閲覧することが出来る津波デジタルライブラリを開発し、公開している。このなかで、新聞、本等の文書はテキスト化し、図や写真などと共にXML形式で構造化してXMLDBMSで管理しているが、そのスキーマは文書の構造によって様々である。そこで、ライブラリ内の全文書を検索対象とした場合に統合スキーマを利用してXML文書を検索する必要がある。本研究では統合スキーマを設計しWebからキーワード入力で簡単に検索できる津波デジタルライブラリ検索システムの実装を行った

キーワード：XML, 情報検索, XML データベース, デジタルライブラリ, スキーマ統合, 津波

1. はじめに

津波デジタルライブラリには、津波災害に関する新聞記事や災害報告書、古文書などの書籍、ビデオ映像、過去の津波記録に基づく物理的モデルによるシミュレーションCGや現地調査データなどがある。これらの書籍は、書籍データごとに定義した異なるXMLスキーマで構造化して商用XMLDBMSで管理している。XML文書に対する検索に関してはデータベース研究と情報検索研究の2つの流れで研究が行われており、従来のWebサーチエンジンと同じようにキーワード入力によってユーザの必要とするXML文書の検索結果を求める要求が大きい[8]。一方、XMLで構造化できないファイル形式のCG、ビデオ、写真なども同じ検索の枠組みで津波デジタルライブラリから検索するために、異種情報源を統合する検索システムが必要になる。このような異種情報源の統合は、これまでに多くの研究が行われてきた。例えば[6][7]が挙げられる。これらの研究ではXMLを利用して統合スキーマを表現し、統合スキーマと情報源でのローカルスキーマの関係を定義してきた。本研究でもこれらの成果と同様にXMLを利用した統合方法を採用し、検索キーワードによるXML文書の検索を行う。これにより、ユーザがデータ構造の異種性を意識することなく全てのデータを横断的に、キーワード入力で検索することができるシステムを設計し実装する。

2. 津波デジタルライブラリ

津波デジタルライブラリ[1][2]とは我々の研究グループで開発し、公開している、津波災害に関する新聞や書籍、ビデオ、CGな

どの情報をWebを通じて提供するデジタルライブラリである。図1に津波デジタルライブラリのトップページを示す。このページから5つのカテゴリ別にコンテンツを見ることができる。「津波映像」では津波の遡上シミュレーションやビデオ映像を動画として見ることができる。「津波災害対策」では災害対策のためのパワーポイントなどの資料を見ることができる。「地図検索」では地図上から地域選択することで、その地域の現地調査の写真や記録を見ることができる。「文献検索」では地方自治体による津波被害報告書や論文、その他津波に関する書籍を閲覧することができる。「新聞記事検索」では津波が発生した当時の新聞記事を見ることができる。

これらのデータのうち新聞、書籍はデータベースシステムを利用した検索閲覧システムを構築している。以降の節では特に、津波デジタルライブラリに収録されている新聞や書籍の管理の方法と検索閲覧システムについて述べる。

2.1 書籍・新聞のデジタル化

津波デジタルライブラリではなるべく実際の書籍に近い形で保存できるようにするた

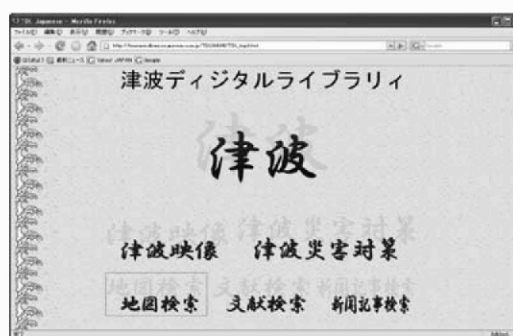
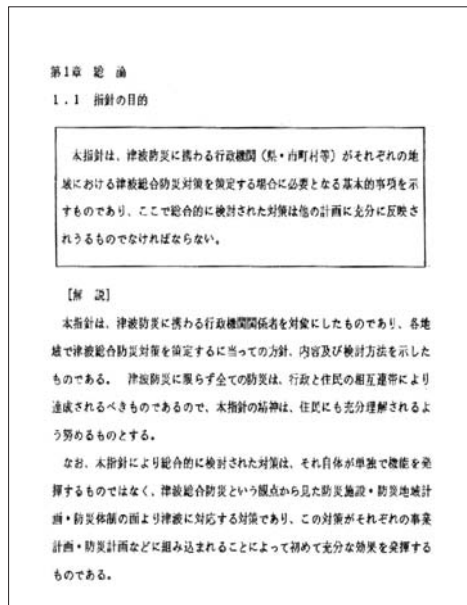


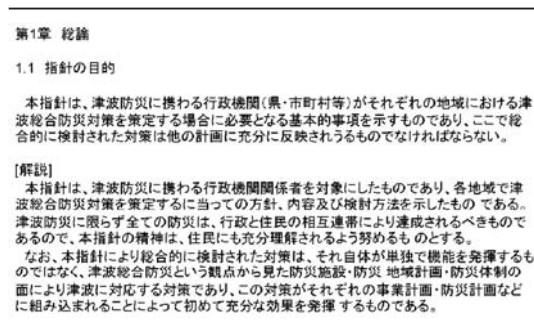
図1 津波デジタルライブラリ

めに図2に示すような画像データをテキストデータと一緒に格納する。これらは以下の手順でデジタル化を行っている。まず、紙で製本された書籍をスキャナを使って1ページごとにスキャンし画像データとしてデジタル化する(図2(a))。次に、活字印刷の場合にはOCRを利用し画像データの文字認識処理を行いテキストデータを作成する(図2(b))。OCRでテキストデータ化できないような手書きの場合にはワープロ入力でテキストデータを作成する。このように書籍の全てのページに対して1つの画像データと1つのテキストデータを作成している。

新聞の場合はマイクロフィルムで保存されているものをマイクロフィルムスキャナを使って読み取り、同様に新聞1面に対して1つの画像データを作り、ワープロ入力でテキストデータを作成する。



(a) 画像データ



(b) テキストデータ
図2 書籍のデジタル化

2.2 書籍・新聞の構造化

書籍の画像データとテキストデータは製本された1ページごとに作られるため、章や節などの文書構造は全く反映されていない。しかし、Webからデジタル書籍として検索や閲覧をするときは章ごとに読んだり、キーワードで複数の書籍の該当する部分を集めたりすることが有用である。そこで、前節のデジタル化したデータを書誌情報を含み、章構成を反映したXML文書として構造化した。構造化された書籍テキストデータの例を図3に示す。

構造化された書籍のタイトル、著者などの<metadata></metadata>で囲まれた書誌情報部分はDublin Core[4]の主要15要素に対応している。このようにして著者や発行年などの書誌情報から文書を検索することができる。本文部分は章や節などのタイトルと対応する文章をひとつのまとまりした構造を持つ入れ子構造で表現する。ここではこのまとまりは<section></section>で囲み、この中のまとまりを部分文書と呼ぶ。章の中に複数の節がある場合を例にすると、節のタイトルとその文章が1つの部分文書であり、章のタイトルと本文(節に含まれない部分)も1つの

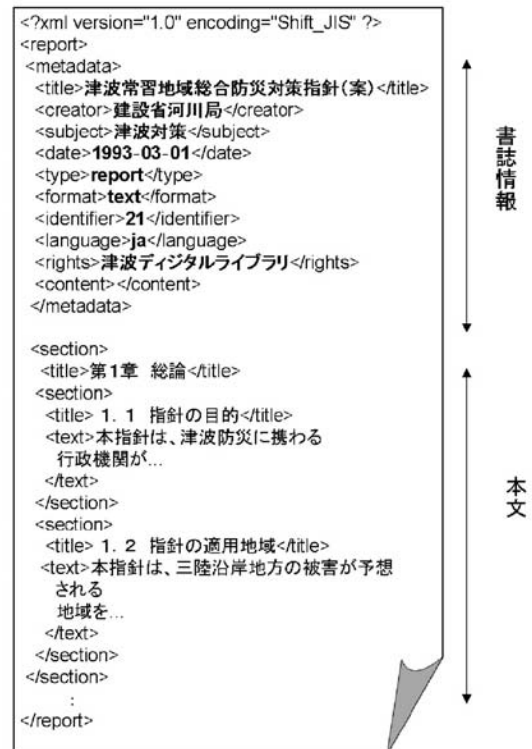


図3 XML文書化

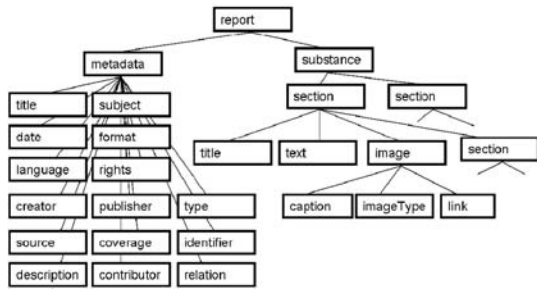


図4 XML 文書の構造

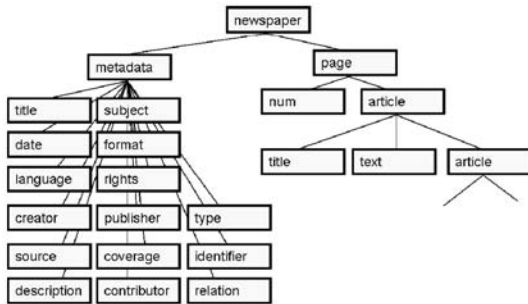


図5 新聞 XML 文書の構造

AsahiS35 . April09 . 10	<?xml><newspaper> <metadata><title> 朝日新聞 </title>...</metadata> <page> ...</page> <?xml><newspaper>
MainichiS35 . May28 . 07	<?xml><newspaper> <metadata><title> 毎日新聞 </title>...</metadata> <page> ...</page> <?xml><newspaper>
KahokuShinpous35 . June27 . 04	<?xml><newspaper> <metadata><title> 河北信報 </title>...</metadata> <page> ...</page> <?xml><newspaper>
.	.
.	.
.	.

図6 新聞 XML 文書の格納例

部分文書となる。図3の場合、書籍のタイトルが書誌情報部分にあり、第1章のタイトルが本文部分にあることがわかる。章の下の節がタイトルと文章を持った入れ子構造になっていることがわかる。書籍のXML文書のツリー構造を図4に示す。新聞の場合は1面、2面などの掲載面ごとに構造化し、書籍と同様に書誌情報部分と本文部分に分けて構造化する。本文部分は1つの記事が部分文書に相当する。新聞XML文書のツリー構造を図5に示す。書誌情報部分はDubline Coreの主要15要素で構成される。本文部分をpage要素で表し、掲載面番号(num)と記事(article)を子要素を持つ。1つの記事が1つの部分文書であり、タイトル(title)と文章(text)を子要素を持つ。

2.3 XML文書のデータベース化

XML文書を管理するXMLDBMSとしてOracle XML DB[3]を利用している。Oracle XML DBではXML文書をXMLTypeというデータ型で管理しており、1つのXMLType型のデータを1つのデータオブジェクトとして扱い、リレーショナルデータベースに格納している。

XML文書は文書構造ごとのXMLスキーマに従ってXMLデータベースに格納されており、格納の仕方は文書構造に依存する。例

えば、新聞の場合は面を単位に構造化してあり、書誌情報部分と本文部分からなる1つのXML文書としてOracle XML DBに1つのリレーションとして格納されている。新聞XML文書を格納してあるリレーションの構造は以下の様になる。

newspaper (id_xmldoc)

図6にインスタンスを示す。新聞XML文書のリレーションでのid属性は新聞の面のXML文書を一意に識別する。これは、新聞名、発行日、掲載面番号を連結した文字列からなる。xml属性はXMLType型で格納されるXML文書である。一方、津波報告書の場合は、1報告書を単位にXMLで構造化してあるが、書誌情報部分と本文部分を分割し、別々のXML文書として別々のリレーションに格納されている。さらに本文部分は章や節などの部分文書を1タプルとしてリレーションに格納してある。津波報告書を格納してあるリレーション構造は以下の様になる。

reportmetatab (id,meta)

reportsectab (id,sid,xml)

reportmetatab表に書誌情報部分を格納し、reportsectab表に本文部分を格納する。属性idがXML文書を一意に決定するもので3桁の

数値である。属性 sid が部分文書を一意に決定するもので 16 桁の数値である。最大の桁を 1 で固定し、以下 3 桁ずつの数値で各部分文書の何番目の子部分文書であるかを示す。例えば、1,001,002,000,000,000 ならば 1.2 節であり、1,002,003,004,000,000 ならば 2.3.4 節である。meta 属性と xml 属性はそれぞれ書誌情報と部分文書の XML 文書を格納する XML Type 型である。id と sid を見ることで分割された部分文書から全文書を組み立てることができる。



図7 部分文書の検索

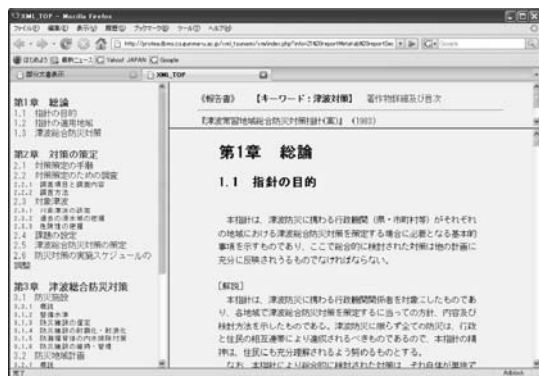


図8 全文閲覧画面



図9 新聞の検索結果

2.4 検索結果の表示

XML で構造化した津波報告書に対する検索、閲覧画面の例を図7と図8に示す。図7はキーワードによる部分文書検索の結果であり、ユーザの入力したキーワードをタイトルまたは本文に含む部分文書をXMLデータベースから検索し、その一覧を表示する。図8は文書の全文閲覧画面である。XML文書の章や節などの部分文書の関係をWebブラウザで見やすくするためにXSLTを使って整形表示している。画面の左側にある部分文書のタイトルのリストは目次に相当し章や節の構造を反映して文字の大きさを変えてある。また、リストから右下の本文にリンクされており、タイトルを選択することで、読みたい部分文書を表示することができる。各章の終わりには図表へのリンクがあり、別ウィンドウで表示することができる。右上部には書誌情報へのリンクがあり、表示することができる。また、新聞の検索結果についても同様に図9に示す。新聞の場合には津波関係記事部分のみを切り抜いた画像へのリンクがあり、それにより画像を表示することができる。

3. XLMデータベースの統合化

3.1 津波デジタルライブラリにおけるデータベースの統合及び検索

津波デジタルライブラリでは書籍、新聞などの文書の構造ごとのXMLデータベースを構築し、それぞれに対して別々に検索閲覧システムを実装している。

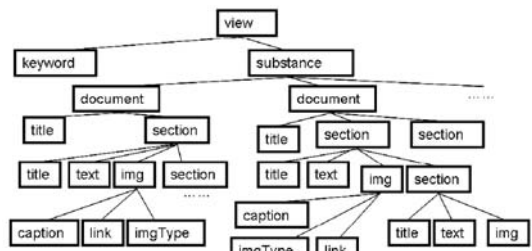


図10 統合スキーマ

また、書籍、新聞以外のCGなどのデータはデータベースで管理しておらず、HTML文書内に直接データへのリンクを記述しているだけで、XML文書と同じような検索機能をもたない。したがって、例えば、「宮古」

に関する書籍，新聞，CG，現地写真など全てのデータがほしい，といったような検索をしたい場合，複数の検索閲覧システムに対して「宮古」をキーワードとして含む検索を繰り返し，その都度得られるデータを整理し収集する必要がある，手間がかかる．よって，津波デジタルライブラリ内の全データを検索対象とし，XML データベースで管理されている XML 文書や HTML で記述されているデータを横断的に扱う統合システムを構築する必要がある．本研究では津波デジタルライブラリ内にある全てのデータを網羅し検索可能となる統合スキーマを設計し，キーワード入力で簡単に津波デジタルライブラリ内の全データを検索閲覧できるシステムを実装した．

3.2 統合スキーマの設計

津波デジタルライブラリに収録されている文書は新聞，論文などの文書構造ごとに定義された XML スキーマに基づいて XML 文書化されており，XML データベースに格納されている．これらの複数の XML データベースを統合的に扱うために各 XML データベース間に共通する枠組みが必要である．そこで，文書を記述する上で共通する部分を反映させた統合スキーマを定義する．本研究で定義した統合スキーマを図 10 に示し，その要素の意味を表 1 に示す．統合スキーマでは文書の内容を表す keyword 要素をルートの子要素として持つ．この keyword はユーザが検索を行う際に利用する．もう 1 つのルートの子要素の substance 要素は keyword に対応する文書に該当する document 要素を子にもつ．各 document 要素以下が 1 つの文書構造に相当する．document 要素の子の title

表 1 要素定義

要素名	意味
keyword	この統合スキーマが含むキーワード
substance	統合スキーマでまとめられた全文書
document	文書
section	部分文書
title	文書または部分文書のタイトル
text	本文テキスト
img	図表情報
caption	キャプション
link	図表ファイル名 (.jpg)
imgType	図表の種類 (figure, table, photo, map)

要素はその文書のタイトルである．section 要素は 1 つの章や節に対応しており，タイトル (title) と本文 (text) を子要素にもつ．また，自分自身 (section 要素) を子にもつ入れ子構造で実際の文書の章や節の構造を表現している．section 要素の子の img 要素は図表情報である．本文テキストとは別の要素としており，キャプション，図表の載っている画像データへのリンク (ファイル名)，図表の種類を子要素にもつ img 要素を子に持つ．なお，図表の種類は (図，グラフ，式：figure，表：table，写真：photo，地図：map) のいずれかとする．

3.3 要約機能を用いた検索結果の効果的な提示

膨大な量の検索結果の全てを精読し，詳細に内容を理解するには時間がかかる．また，時間をかけずに斜め読みをするだけでは重要な部分を読み飛ばしてしまう可能性がある．そこで，本研究では内容を把握しやすいように検索結果の文書をユーザの必要に応じて要約する機能を考案した．本研究で利用した要約アルゴリズムは [5] で述べられているのでここでは概念のみを説明する．XML 文書の部分文書の text 要素の文章の中から津波重要語の有無などで各文に重みをつける．重みの大きい文をいくつか選択して text 要素の文章の要約文とする．章構造は変えずに全ての text 要素の文章を要約したものを要約 XML 文書とする．要約 XML 文書も通常の XML 文書と同様に OracleXML DB で管理する．

3.4 XML 文書以外のデータの扱い

XML データベースで管理されている構造化文書には津波デジタルライブラリの Web ページ上で「文献検索」，「新聞記事検索」などのカテゴリに分けられたインターフェースを通してアクセスすることができる．一方，津波デジタルライブラリ内には文書以外にも写真の画像データと記録からなる現地調査データや CG，講演のパワーポイントなどのデータが存在する．これらの文書以外のデータは「津波災害対策」，「津波映像」，「地図検索」のカテゴリに分けてアクセスできるようになっており，その中のコンテンツとして HTML 内にリンクを張ることで目的のデータにアクセスできるようになっている．このような津波デジタルライブラリの Web ページからアクセスできる全ての形

式的数据を検索対象にして、文書検索閲覧システムを実装する。このために文書以外のデータも統合スキーマで扱う手法を考案した。写真の画像データやCGの場合、画像処理を行わなければ直接内容についての検索を行うことはできないので、ここではコメントやキャプション、ファイル名などから検索する手法を採用した。

HTML 文書内にあるコメントやキャプション、ファイル名などの文字列を検索することで、画像データやCGなどの検索を実現した。そして、HTML 文書の title タグの値を図 10 に示す統合スキーマの section の子要素の title 要素とし、body タグ内のタグを除いた部分を text 要素として統合スキーマに合うように構造化する。こうすることで統合スキーマを利用した検索システムに組み込むことができる。すなわち、津波デジタルライブラリの全ての形式のデータへの検索を1回のキーワードの入力で行うことができる。文書と同様に Oracle XML DB で1つのオブジェクトとしてリレーショナルデータベースで HTML 文書を管理する。HTML 文書を格納するリレーションの構造は以下の様になる。

HTML (name, category, path)
 図 11 に HTML 文書の格納例を示す。

以上のように XML 文書以外のデータを

HTML		
name	category	path
CG	video	/TSUNAMI/CG.html
education	countermeasure	/TSUNAMI/saigaitaisaku/education.html
kamaishi	map	/TSUNAMI/iwate/kamaishi.html
⋮	⋮	⋮

図 11 HTML 文書の格納例

XML 文書と同じように Oracle XML DB で管理することによって津波デジタルライブラリ内の全てのデータを検索できる統合スキーマを利用した検索閲覧システムを設計した。

4. システムの実装

4.1 システムの概要

本研究で実装したシステムの概要図を図 12 に示す。本システムは前置問合わせ処理部、後置問合わせ処理部、データベース部に分けられる。本システムではまず、前置問合わせ処理部においてユーザが入力したキーワードからデータベース部への検索文を生成する。この検索文は XML スキーマごとに生成され、個別のデータベースへ送られる。データベースでは送られた検索文により検索が行われ、その結果を返す。検索結果は個別の XML スキーマに依存した構造になっているので後置問合わせ処理部において統合スキーマに合った形式に整形する。整形された結果の XML 文書を XSLT を用いて Web 上に HTML 形式に出力する。このとき、津波デジタルライブラリ内の全文書を検索した結果は膨大な量になるので、要約機能を使って整理することで利便性を向上させた。要約は検索結果の <text></text> 部分に該当する部分を要約データベースから検索し、同様に整形を行う。

データベース部については前節で述べたように文書構造ごとの XML スキーマで構造化された XML 文書と文書以外のデータを公開する HTML 文書が XMLDBMS で管理されている。

4.2 前置問合わせ処理部

4.2.1 キーワードの入力

任意キーワードの他に津波重要語と地名をキーワードとしている。津波デジタルライブラリに収録されている文書中に出てくる頻度の高い単語をあらかじめ調査しておいて、津波重要語とする。ユーザはその重要単語の中からキーワードを選択することで利用しやすくなる。一方、ある特定の地域での被害状況を検索したい場合などは、地名をキーワードとして扱うことが便利である。表示された地図上で地域を選択することで直感的な地名選択を可能にする。また、入力したキーワードの類似語による検索も行えるように類似語

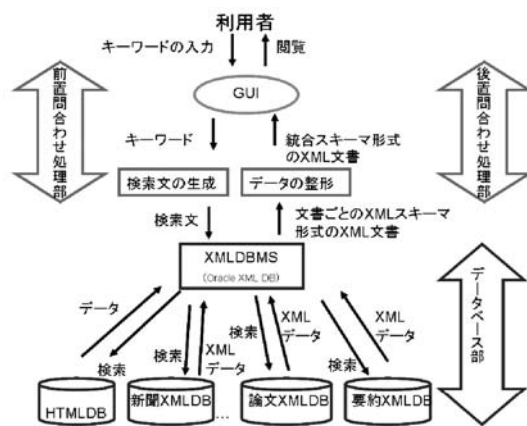


図 12 システムの概要

対応表を作成した。「つなみ」という言葉には「津波」と「津浪」という2つの漢字表記があり、時代によっては「海嘯」という言葉を使う場合もある。したがって、「津波」というキーワードが入力された場合に「津浪」や「海嘯」を類似語としてキーワードに加える。地名に対しては合併などにより変化した地名変遷表を作成し、地名での検索を新旧の地名で行えるようにした。例えば、岩手県宮古市の場合、田老村や磯鶏村などと合併しており、現在の宮古市を対象にした場合に「宮古」というキーワードだけでは旧田老村や磯鶏村に該当する地域の情報を得られない可能性があるため、「田老」、「磯鶏」を地名変遷語としてキーワードに加える。ユーザの必要に応じてこれらの類似語によるOR検索を行えるように設計した。

4.2.2 検索文の生成

前置問合わせ処理部ではユーザが入力したキーワードから各XMLデータベースへの検索文を生成する。まず、ユーザが入力したキーワードから統合スキーマへの検索文をXPathで作成する。このXPathは入力されたキーワードを変数\$KEYとした場合以下の検索式として書くことができる。

```
/view/substance/document/section[/  
key = $KEY]
```

これはキーワード\$KEYで統合スキーマから部分文書を検索するものである。すなわち、子要素であるタイトルまたは本文(図表含む)にキーワード\$KEYを含む部分文書を津波デジタルライブラリにある文書全体から検索してくることに相当する。

/view/substance/document/	→	/report/substance/
/view/substance/document/title	→	/report/substance/
		/repor/metadata/title
/view/substance/document/section	→	/report/substance//section
/view/substance/document/section/title	→	/report/substance//section/title
/view/substance/document/section/text	→	/report/substance//section/text
/view/substance/document/	→	/newspaper/page/
/view/substance/document/title	→	/newspaper/metadata/title
/view/substance/document/section	→	/newspaper/page//article
/view/substance/document/section/title	→	/newspaper/page//article/title
/view/substance/document/section/text	→	/newspaper/page//article/text

図13 統合スキーマと文書スキーマの対応関係

統合スキーマへのXPath検索式から各XMLデータベースへのSQL文に以下の手順に従って変換することで、ユーザが入力したキーワードから、各XMLデータベースへの検索文を生成することができる。実際のXML文書を管理するXMLデータベース部はOracle XML DBのリレーショナルデータベースシステムで管理されている。そこで、各XMLデータベースへの検索文はOracle XML DBに固有のSQL文で表現する必要がある。

- (1) 統合スキーマへのXPathから各XMLデータベースへのXPathへの変換
- (2) 各XMLデータベースへのXPathをOracle XML DBへのSQLに変換

(1)の変換はスキーマ間の対応関係をあらかじめ定義しておくことで行う。統合スキーマに対応する津波デジタルライブラリ内のXMLスキーマの数は決まっており統合スキーマの各要素が各XMLスキーマのどの要素に対応するかをあらかじめ定義しておくことができる。この関係より統合スキーマの部分文書を構成する要素が各XML文書スキーマのどの要素に対応するかを調べて該当するXMLデータベースに対するXPath検索式を作成する。統合スキーマと各XML文書スキーマの対応関係を図13に示す。津波被害報告書のXMLスキーマは統合スキーマの/view/substance/document//sectionが/report/substance//section/に対応する。

(2)の変換によるSQL文を図14に示す。Oracle XML DBで使用するXML操作の関数に(1)で変換した各XMLデータベースへのXPathを引数として与えることで行う。OracleXML DBへのSQLはXMLスキーマ構造によらず構文が同一である。図中で\$マークが付いている文字列は変数であり、(1)で得られるパスやXMLデータが格納されているテーブル名などが入る。新聞XMLデータベースの場合には次のようになる。\$idは部分文書を識別するための識別子である。新聞の場合は新聞名、発行日、掲載面を連結したものであり、図6で示したりレーションのid属性の値が当てはまる。\$xmltypeはXMLデータベース部でXML文書を管理しているリレーションのXMLType型の属性の名前である。新聞ではxmldoc属性に相当する。\$sectionは部分文書を表す要素のパスである。新聞の場合は図13より/newspaper/page//articleが相当する。\$sectionchildは\$sectionで与えられる

約文を要約 XML データベースから検索してくる。検索結果を通常の文書と同様に整形し、表示することで、大まかな内容を把握できる。

5. 検索実行例

この節では本研究で実装した文書検索閲覧システムの実行例を示す。図 16 はキーワード入力画面である。任意にキーワードを入力するだけで、検索を行うことができ従来の検索エンジンと同じように直感的に利用することが可能である。必要に応じて津波重要単語からキーワードを選択できたり、類似語をキーワードに加えることもできる。要約を選択すれば要約文を検索する。図 16 では任意に入力した「港」というキーワードについて要約文検索し、類似語検索も行う。この場合類似語は「栈橋」「港湾」などが対応表から取り出され検索キーワードに加えられる。図 16 で入力されたキーワードに対する閲覧画面は図 17 のようになる。画面左上に検索された全文書のタイトルと部分文書のリストが表示されている。画面右上にはキーワードが表示されており、類似語もここに表示される。画面下は部分文書の要約文が表示されており、画面左上のタイトルを選択することで読みたい部分文書を表示することができる。部分文書のタイトルの横にある全文ボタンをクリックすればその全文閲覧を行える。

6. まとめと今後の課題

本研究では我々が開発し、公開している津波デジタルライブラリの全文書に対する検索閲覧機能の実装を行った。

全文書を検索対象にするために文書構造の共通点を統合スキーマの形で表現して、文書構造ごとの XML スキーマを統合的に扱えるようにした。また、XML スキーマごとに存在する XMLDB への検索文をキーワードを入力するだけで生成できるようにした。そして、ライブラリ内の全文書からの大量の検索結果を整理するための GUI を実装した。

今後の課題としては様々な利用形態に対応した津波デジタルライブラリのコンテンツの提供環境の構築が挙げられる。現在、津波デジタルライブラリの有効利用の一例として携帯端末を用いて検索できる環境の開発をしており[5]、その他にも、ユーザごとに検索方法や結果の表示を行えるように GUI を

カスタマイズし、パーソナライゼーションを行うことでより利用しやすい環境の構築を目指す。

文献

- [1] 津波デジタルライブラリ
<http://tsunami.dbms.cs.gunma-u.ac.jp>
- [2] Sayaka Imai, Yoshinari Kanamori, and Nobuo Shuto: Tsunami Digital Library, J. Gonzalo et al. (Eds.) ECDL 2006, LNCS 4172, pp.555-558, 2006.
- [3] Oracle XML DB
<http://otn.oracle.co.jp/tech/xml/xmlldb/>
- [4] Dublin Core Metadata Initiative
<http://dublincore.org/>
- [5] 宮崎明美, 今井さやか, 金森吉成, 首藤伸夫: 津波デジタルライブラリから携帯端末へのコンテンツ配信システム開発, DEWS2007, 2007.
- [6] Ioana Manolescu, Daniela Florescu, Donald Kossman: Answering XML Queries over Heterogeneous Data Sources, Proceedings of the 27th VLDB Conference, Roma, Italy, 2001.
- [7] Michael Carey, Jerry Kiernan, Jayavel Shanmugasundaram, Eugene Shekita, Subbu Subramanian: XPERANTO: A Middleware for Publishing Object-Relational Data as XML Documents, Proceedings of the 26th VLDB Conference, Cairo, Egypt, 2000.
- [8] 絹谷弘子, 波多野賢治, 吉川正俊, 植村俊亮: キーワードを利用した XML 文書検索, 情報処理学会論文誌, Vol.45 No.SIG7 (TOD22) pp.255-273, 2004.