XML 統合スキーマを利用したディ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††　　首藤　伸夫†††

† 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
†† 群馬大学工学部情報工学科 〒 376-8515 群馬県桐生市天神町 1-5-1
††† 日本大学大学院総合科学研究科 〒 102-0073 東京都千代田区九段北 4-2-1
E-mail: †{kenzo,sayaka,kanamori}@dbms.cs.gunma-u.ac.jp, ††shuto-nobuo@arish.nihon-u.ac.jp

あらまし　我々はディ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ィ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ディ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，ディジタルライブラリ，スキーマ統合

Implementation of retrieval system for digital library using schema integration

KENZO KAWAMOTO†, SAYAKA IMAI††, YOSHINARI KANAMORI††, and NOBUO SYUTO†††

† Department of Computer Science, Graduate School of Engineering, Gunma University
1-5-1 Tenjin-cho,Kiryu,Gunma 376-8515,Japan
†† Department of Computer Science, Faculty of Enginieering, Gunma Univarsity
1-5-1 Tenjin-cho,Kiryu,Gunma 376-8515,Japan
††† Advanced Research Institute for the Sciences and Humanities,Nihon University
4-2-1 Kudankita Chiyoda-ku,Tokyo,102-0073,Japan
E-mail: †{kenzo,sayaka,kanamori}@dbms.cs.gunma-u.ac.jp, ††shuto-nobuo@arish.nihon-u.ac.jp

Abstract　We are developing “Tsunami Digital Library(TDL)” which consists of the documents of newspaper articles, reports and books, videos, and CGs on the tsunami occurred in the past, and are opened through Web. TDL is a case study of digital libraries to use XML technology. The documents which contain figures, pictures and tables are represented in XML structure, and managed by using an XML database system. The schemas of these documents depend on a kind of documents. Therefore,each document has a different schema. In order to make a traversal retrieval crossing whole documents, it is necessary to use a unified schema to integrate all schemas for XML documents. This paper describes the design of integration schema and retrieval system, and the system implementations. We can retrieves easily some requested data from TDL through keyword input on Web.

Key words　XML，Information retrieval，XML database，Digital library，Schema integration

1. はじめに

米国ではＮＳＦのディジタルライブラリ開発プロジェクト，フェーズ１，フェーズ２が終了し，様々な分野のコンテンツを対象とした実システムが稼動している [6]．ヨーロッパにおいてもＥＣＤＬの国際会議などを通して多くの実システムが稼動している様子を知ることができる [7]．今後，インターネットの高速化とユビキタス環境の整備によって，より一層ディジタルライブラリが充実，発展していくものと思われる．

ディジタルライブラリ実現には，データベースシステムの技術的側面とコンテンツ開発の両方が関係するが，現在の商用ＸＭＬデータベースシステムを利用することでかなりの技術的な課題が解決できるものと考えられる．そこで，我々は事例研究として，商用ＸＭＬデータベースシステムを利用して，津波分野のディジタルライブラリ開発を実践することにした [1] [2]．そして，具体的な実践を通して，解決されるべき技術課題を把

握するという視点で研究をしてきた．

本研究の津波ディ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 [10]．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ィ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 [8] [9] 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．

本稿の構成は以下のようになる．まず，２章で我々の開発した津波ディジタルライブラリの書籍のＸＭＬ構造化について述べて，各種書籍ＸＭＬの統合化を説明する．３章では複数の文書構造を扱うための統合スキーマの設計を述べる．４章では本研究で実装したシステムについて述べる．５章で実行例を示す．最後に 6 章でまとめと今後の課題について述べる．

2. 津波ディジタルライブラリ

津波ディジタルライブラリ [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 通じて提供するディジタルライブラリである．図 1 に津波ディ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が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，津波ディ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・新聞のディジタル化

津波ディジタルライブラリでは紙の劣化や破損による情報の消失に備えるために，なるべく実際の書籍と近い形で保存できるように図 2 に示すような画像データとテキストデータとして計算機に取り組む．これらは以下の手順でディ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ィジタル化する（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古文書の専門家が読解して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．

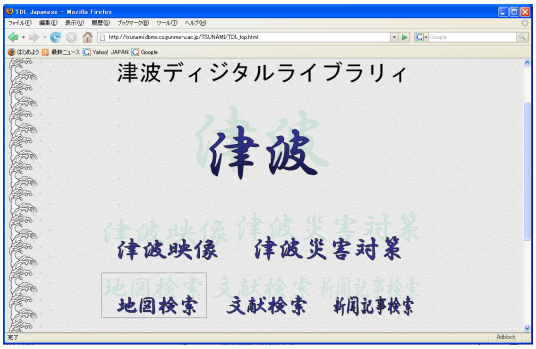


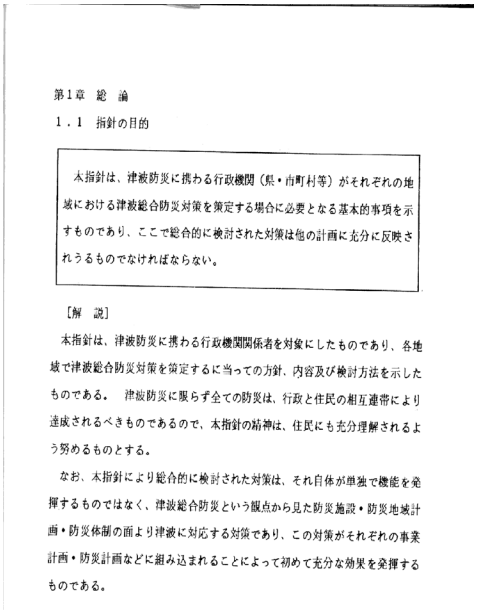
図 1　津波ディジタルライブラリ
Fig. 1　Tsunami Digital Library

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．

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 からディ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，前節のディ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．こうすることで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



第1章 総 論

1．1 指針の目的

本指針は、津波防災に携わる行政機関（県・市町村等）がそれぞれの地域における津波総合防災対策を策定する場合に必要となる基本的事項を示すものであり、ここで総合的に検討された対策は他の計画に充分に反映されうるものでなければならない。

［解 説］

本指針は、津波防災に携わる行政機関関係者を対象にしたものであり、各地域で津波総合防災対策を策定するに当っての方針、内容及び検討方法を示したものである。 津波防災に限らず全ての防災は、行政と住民の相互連帯により達成されるべきものであるので、本指針の精神は、住民にも充分理解されるよう努めるものとする。

なお、本指針により総合的に検討された対策は、それ自体が単独で機能を発揮するものではなく、津波総合防災という観点から見た防災施設・防災地域計画・防災体制の面より津波に対応する対策であり、この対策がそれぞれの事業計画・防災計画などに組み込まれることによって初めて充分な効果を発揮するものである。

(a) 画像データ

第1章 総論

1.1 指針の目的

本指針は、津波防災に携わる行政機関（県・市町村等）がそれぞれの地域における津波総合防災対策を策定する場合に必要となる基本的事項を示すものであり、ここで総合的に検討された対策は他の計画に充分に反映されうるものでなければならない。

[解説]
本指針は、津波防災に携わる行政機関関係者を対象にしたものであり、各地域で津波総合防災対策を策定するに当っての方針、内容及び検討方法を示したもの である。津波防災に限らず全ての防災は、行政と住民の相互連帯により達成されるべきものであるので、本指針の精神は、住民にも充分理解されるよう努めるも のとする。
なお、本指針により総合的に検討された対策は、それ自体が単独で機能を発揮するものではなく、津波総合防災という観点から見た防災施設・防災 地域計画・防災体制の面により津波に対応する対策であり、この対策がそれぞれの事業計画・防災計画などに組み込まれることによって初めて充分な効果を発揮 するものである。

(b) テキストデータ

図 2　書籍のディジタル化

Fig. 2　Image data and text data

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）を子要素に持つ．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）を子要素に持つ．

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

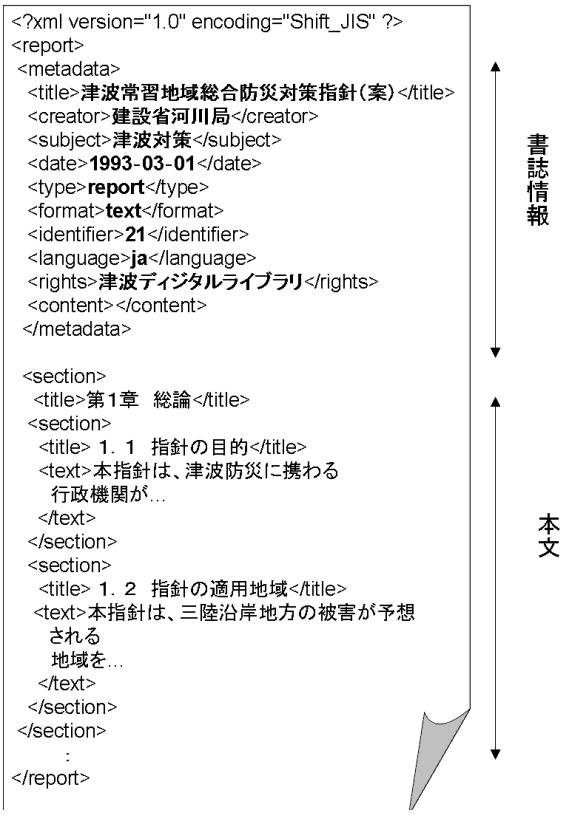


図 3　XML 文書化

Fig. 3　XML document

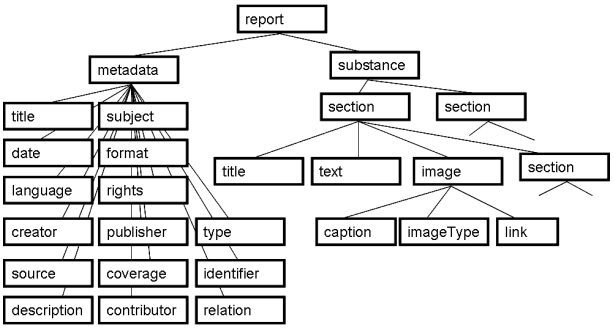


図 4　XML 文書の構造

Fig. 4　Structure of XML document for Tsunami report

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)

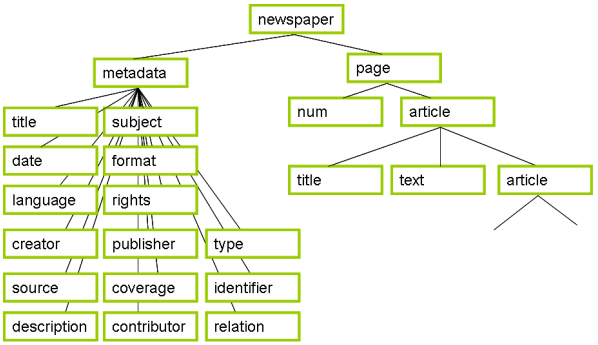


図 5　新聞 XML 文書の構造

Fig. 5　Structure of XML document for newspaper

newspaper

id	xmldoc
AsahiS35_April09_10	<?xml><newspaper> <metadata><title>朝日新聞</title>...</metadata> <page> ...</page> </newspaper>
MainichiS35_May28_07	<?xml><newspaper> <metadata><title>毎日新聞</title>...</metadata> <page>...</page> </newspaper>
KahokuShinpouS35_June27_04	<?xml><newspaper> <metadata><title>河北新報</title>...</metadata> <page></page> </newspaper>
⋮	⋮

図 6　新聞 XML 文書の格納例

Fig. 6　Newspaper XML document instances

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 桁づ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 XMLType 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．

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，

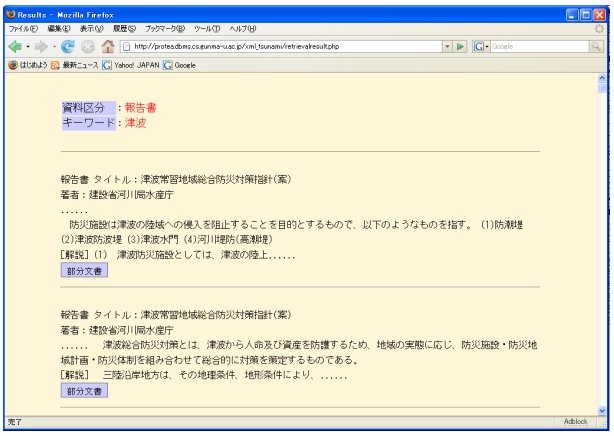


図 7　部分文書の検索

Fig. 7　Retrieval of partial documents

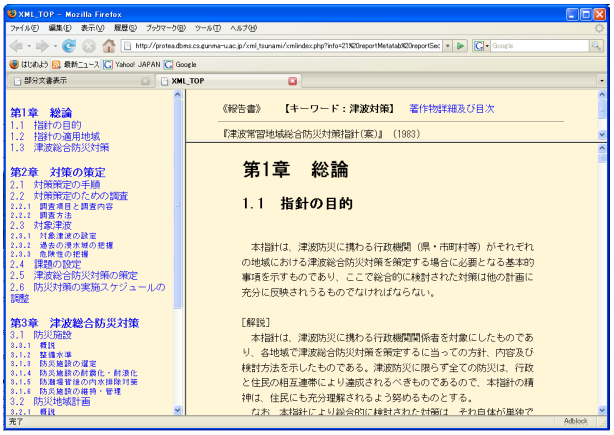


図 8　全文閲覧画面

Fig. 8　Reference screen of full document

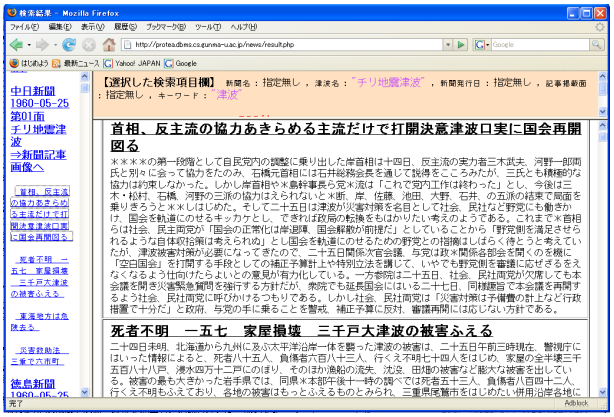


図 9　新聞の検索結果

Fig. 9　Retrieval of newspapers

新聞の検索結果についても同様に図 9 に示す．新聞の場合には記事部分のみを切り抜いた画像へのリンクがあり表示することができる．

3.　XML データベースの統合化

3.1　津波ディ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

津波ディジタルライブラリでは書籍，新聞などの文書の構造

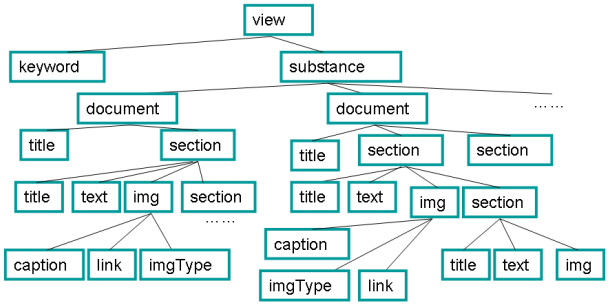


図 10　統合スキーマ

Fig. 10　Unified schema

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．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り，手間がかかる．よって，津波ディ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．本研究では津波ディ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ィ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

津波ディ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 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，

表 1　要 素 定 義

Table 1　Element difinition

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）

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 Oracle XML 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ィ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，津波ディ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．このような津波ディ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のデータを検索対象にして，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

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

Fig. 11　HTML document instances

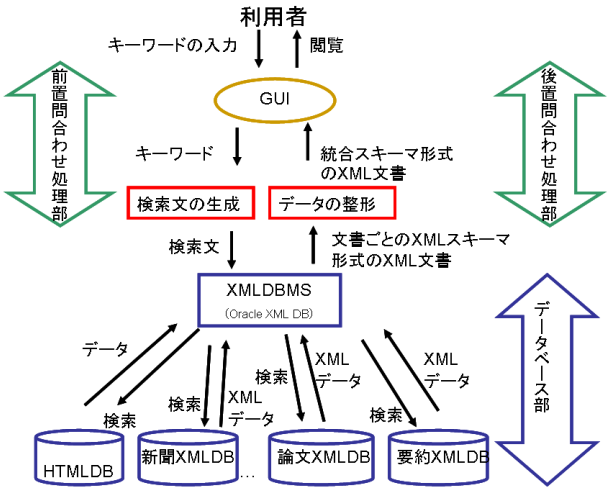


図 12　システムの概要

Fig. 12　System overview

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，津波ディ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 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ィ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，津波ディ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．津波ディ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対応表を作成した．「つなみ」という言葉には「津波」と「津浪」とう 2 つの漢字表記があり，時代によっては「海嘯」と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ので,「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 とした場合に以下の検索式として書くことができる．

/view/subatance/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ィ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．

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

```
/view/substance/document/                → /report/substance/
/view/substance/document/title           → /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
Fig. 13　Relation between unified schema and document schema

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ィジタルライブラリィ内の 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 操作用の関数に（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．Oracle XML 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 で示したリレーションの 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 で与えられる部分文書の子要素のパスである．新聞では/newspaper/page//article/title と/newspaper/page//article/text になる．したがって，図 14 の SQL 文は部分文書内の文字列にキーワードが含まれている場合，その部分文書を文字列として返してくる．以上の手順で生成された新聞の XML データベースへの検索文を図 15 に示す．WHERE 句の contains 関数が 2 つ必要な理由は部分文書のタイトルと本文の両方に対して，キーワードが含まれるかを検索する必要があるからである

```
SELECT   ＄id,extract(＄xmltype,＄section).getClobVal()
FROM     ＄table
WHERE    contains(＄xmltype, ＄KEY inpath(＄sectionchild)）
         orderby ＄id
```

図 14　Oracle XML DB への SQL
Fig. 14　SQL for Oracle XML DB

```
SELECT   id,extract(xmldoc ’/newspaper/page/article ’ ).getClobVal()
FROM     newspaper
WHERE    contains(xmldoc ’＄KEY inpath(/newspaper/page/article/title ’ )) or
         contains(xmldoc ’＄KEY inpath(/newspaper/page/article/text) )
         order by id
```

図 15　新聞 XML データベースへの検索文
Fig. 15　Query for newspaper XML DB

4.3　後置問合わせ処理部

4.3.1　データの整形

XML データベースから得られた検索結果は各々の XML スキーマにしたがった形式であり，検索結果の部分文書ごとに異なる．したがって，部分文書を 1 つの文書にまとめて統合スキーマに沿った形式に整形する必要がある．新聞を例にあげると，新聞は掲載面ごとに構造化されており図 15 にある検索文によって記事が 1 つの単位として得られる．この 1 つの記事を 1 つの部分文書（section 要素）とし 1 面で 1 文書（document 要素）になるように統合スキーマに沿った形式に整形する．図 13 で示したスキーマ間の関係から，部分文書を構成する要素名を新聞スキーマの形式（article,title,text）から，統合スキーマの形式 (section,titel,text) に直す．次に，id が共通するものでまとまり，元の XML 文書構造を再構築して 1 つの文書にする．このとき，キーワードを含まず検索されなかった部分文書の XML 文書構造の位置を考慮して，祖先子孫関係を再構築する．たとえば，ある部分文書の親が検索結果に含まれないときはその部分文書が親の位置に繰り上がる．

4.3.2　GUI

津波ディジタルライブラリ内の全文書から検索して得られた結果の部分文書は膨大な量になり，それをただ並べただけでは見にくい．そこで，ユーザの利便性を向上させる為に整理する必要がある．本研究では，XSLT を利用して閲覧画面を表示し，各文書を要約する機能を備えた GUI を実装した．

（1）閲覧機能

統合スキーマの形にまとめた検索結果を XML 文書としてそのまま Web ブラウザ上に表示してもタグがついたままであったり，入れ子構造がわかりづらく章や節の文書構造が理解しづらい．そこで，XSLT を使いユーザの理解しやすいような表示を行なえる HTML 形式に変換する．

（2）要約機能

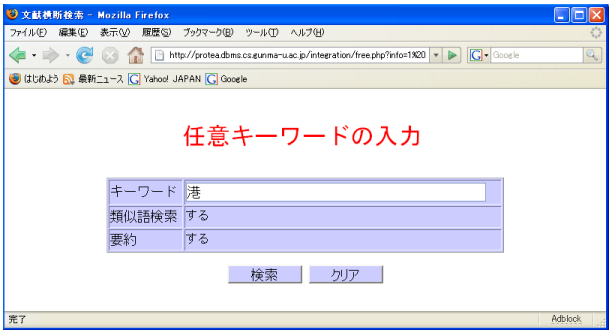


図 16　キーワードの入力画面

Fig. 16　Input screen of keyword

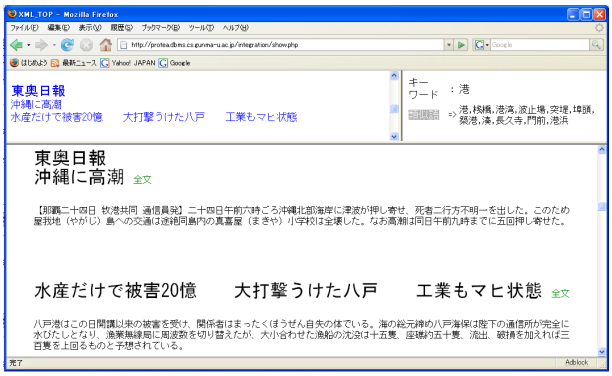


図 17　要約文検索結果閲覧画面

Fig. 17　Reference screen of summary document

キーワードを入力する際に要約するかどうかを選択し、要約する場合は通常の XML 文書を検索してくる代わりにキーワード含む要約文を要約 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「桟橋」「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ィ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ィ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，津波ディ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．

謝辞

本研究は文部科学省科学研究費補助金基盤研究 (B)（課題番号 15300029），文部科学省大都市大震災軽減化特別プロジェクト III-3 の助成を受けて行った．

また，本研究を行うにあたり，新聞記事要約文作成プログラムと書籍要約文作成プログラムについて，群馬大学工学部情報工学科金森研究室の小菅広太さんと中島秀明さんにご協力頂いた．ここに記して謝意を表する．

文　　献

[1] 津波ディジタルライブラリ
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/xmldb/

[4] Dublin Core Metadata Initiativ
http://dublincore.org/

[5] 宮崎明美，今井さやか，金森吉成，首藤伸夫：津波ディ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] Digital Libraries Initiative Phase2
http://www.dl12.nsf.gov/

[7] Reserch and Advanced Tecnology for Digital Libraries,10th European Conference,ECDL 2006,LNCS 4172,September 2006.

[8] Ioana Manolescu,Daniela Florescu,Donald Kossmann：Answering XML Queries over Heterogeneous Data Sourcec,Proceedings of the 27th VLDB Conference,Roma,Italy,2001.

[9] 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.

[10] 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(TOD 22) pp.255-273，2004.