XML を利用した文献データベースの構築

高橋　洋成

（筑波大学）

s025035@ipe.tsukuba.ac.jp

0　はじめに

近年、XML を用いたデータベース管理システムが注目を集め始めている。本稿は、将来的に GIS との連携を視野に入れた文献管理システムを構築するにあたり、書誌データを XML 化する試みである。

1　XML とは何か

まず、XML とは何かについて概略する。XML（Extensible Markup Language）は、World Wide Web Consortium（W3C）によって 1998 年に策定された汎用のマークアップ言語である。XML は、いわゆる「Web ページ」を記述する HTML によく似ているが、根本的な違いがある。HTML は、見出し（h1、h2、……）、段落（p）、箇条書き（ul、ol、dl）といった、Web ページを記述するのに必要なタグセットを定義している。一方、XML はそのようなタグセットを定義するための枠組みである。

HTML と XML の違いをもう少し具体的に見るために、XHTML で記述された Web ページの例を以下に挙げる。

```
<?xml version="1.0"?>
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="ja">
  <head>
    <title>XML を利用した文献データベースの構築</title>
  </head>
  <body>
    <h1>1. XML とは何か</h1>
    <h2>1.1 XML の誕生</h2>
```

```
    <p>XML は、World Wide Web Consortium (W3C) によって 1998 年に
策定された汎用のマークアップ言語である。</p>
    <p>XML の特徴は以下の点である。</p>
    <ul>
      <li>木構造を持つこと。</li>
      <li>テキストベースであること。</li>
      <li>拡張性と相互運用性に優れていること。</li>
    </ul>
    <h1>参考文献</h1>
    <ol>
      <li>XML 1.0, World Wide Web Consortium,
 http://www.w3.org/TR/REC-xml/</li>
    <ol>
  </body>
</html>
```

このドキュメントには、冒頭に「XML を利用した文献データベースの構築」というタイトルがある。このタイトルを、XHTML で定義された title 要素タイプとして分類し、その範囲を開始タグ <title> と終了タグ </title> で囲む。次に、本文内容は「1. XML とは何か」という見出しから始まる。この見出しを、XHTML で定義された h1 要素タイプ（Heading level 1）に分類し、その範囲を開始タグ <h1> と終了タグ </h1> で囲む。このように、ドキュメント内に存在する情報（要素）を抽出・分類し、タグを埋め込んでいくタイプのコンピュータ言語を総称してマークアップ言語と呼ぶ。

最初のレベル 1 見出し（h1 要素）に続くのは「1.1 XML の誕生」というレベル 2 の見出し（h2 要素）である。その後には段落（p 要素）が 2 個続き、順序無関係の箇条書き（ul 要素、Unordered List）と各項目（li 要素、List Item）が置かれている。このように、ドキュメントに埋め込まれたタグを追っていくことで、ドキュメントの全体構造を把握することができる。

さて、ドキュメントの末尾には「参考文献」という見出し（h1 要素）と、順序付き箇条書き（ol 要素、Ordered List）がある。通常の Web ページならば、特に何もこだわることなく、各文献を順番に並べていけば良い。だが、各文献情報をさらに分類し、タイトル・著者・所在などのタグを付けておけば、小規模ながらこれを文献データベースのように利用できることになる。

では、実際にどのような要素タイプに分類し、どのようなタグを付ければ良いだろうか。ここに至って、XHTML には適切に書誌情報を整理できる要素タイ

プが存在しないことに気付く。そこで、いったん XHTML から離れ、書誌情報を整理するための要素タイプと、実際に使用するタグセットを新たに定義する必要が生まれる。例えば「著者の名前は author 要素タイプに分類し、<author> と </author> というタグで囲むようにする」という約束事を決めなければならない。この約束事を定義するための枠組みが、XML である。実際のところ、XHTML も XML の枠組みで定義された言語である。今回作成した文献データベース用形式も、XML の枠組みで定義されている。それゆえ、XML は言語を定義するメタ言語とも呼ばれる。

2　なぜ XML を用いるのか

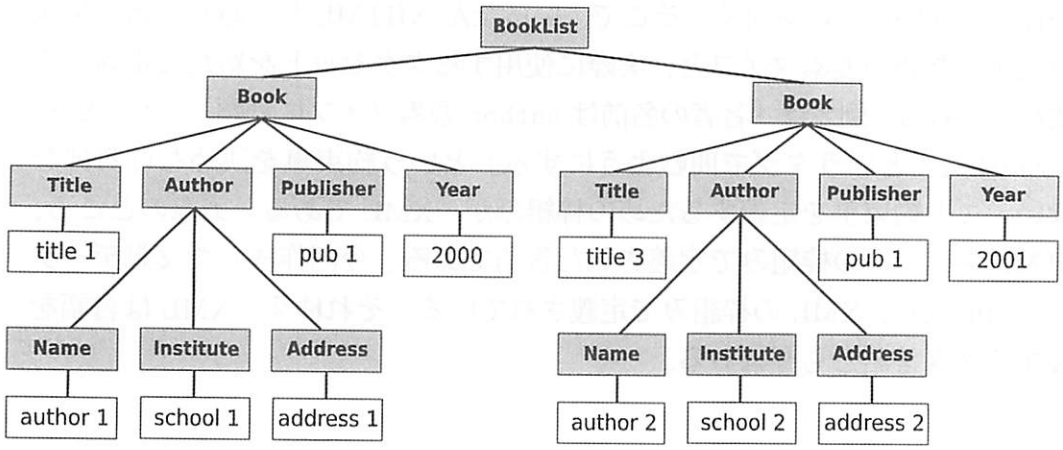
現在のデータベースの管理システムの主流は、いわゆる関係モデルを用いたものである。このモデルでは、データの集合を表として扱う。表の横軸、表の縦軸、定義域（データが持つべき値の種類・範囲）など、いくつもの項目を重ね合わせることで、条件にマッチしたデータを探し出すことができる。また、ある項目を別の表と結び付けることで、階層的なデータ検索を行うこともできる。

	Title	Author	Publisher	Year
	title 1	author 1	pub 1	2000
	title 2	author 1	pub 2	2002
	title 3	author 2	pub 1	2001

	Author	Institute	Address
	author 1	school 1	address 1
	author 2	school 2	address 2
	author 3	school 3	address 3

一方、XML は、データを主に要素の木構造として表現する。前節の XHTML の例を用いて考えてみると、このドキュメントには、最も外側の html 要素の中に head 要素と body 要素の 2 つがある。これを、html 要素が head 要素と body 要素に「枝分かれした」ものと考える。head 要素の下には、title 要素がただ 1 つだけ存在している[1]のに対し、body 要素は 7 個の子要素（h1、h2、p、p、ul、h1、ol）に枝分かれしている。このように、XML の枠組みで定義されたマークアップ言語は、ドキュメントを必ず木構造として表現するる[2]。次に挙げる図において、著者（Author 要素）の一覧を得るには、/BookList/Book/Author/Name のように枝を順次辿っていけば良い。

[1]便宜上、ここでは空白類のみのテキストノードを全て無視する。
[2]関係モデルでも木構造を表現することは可能であるが、今回は扱わない。



関係モデルがデータ集合を表として扱うのに対し、XML は木構造として表現する。ゆえに、XML は不定なデータ構造に対しても比較的強く、開発途中でデータ構造が変化しうるような状況でも十分に対処可能となる[3]。

さらに、XML が生まれたそもそもの目的として、システム間の相互運用性の向上が挙げられる。利用者が各々の目的で要素タイプを拡張していけば、その利用者にしか扱えないタグセットが増大する一方である。そこで、XML データを操作・変換するための統一規格も一緒に策定された（XSLT、XPointer、XInclude、XQuery など）。それまでは、データの利用を望む利用者が、そのデータに対応したソフトウェアを探すケースが多かった。だが、XML データに関しては、使用したいソフトウェアに合わせて XML データの方を操作・変換することが可能なのである。

以上を踏まえ、今回の文献データベースの作成にあたり、XML を用いたのは次の理由による。

1. 書誌情報の項目の増減に対処するため。何を項目として立てるかにより検索の精度も変わる。システムを運用しつつ、項目を取捨選択するには、不定なデータ構造に強い XML 形式が都合良い。
2. 将来的に GIS と連携するため。XML 形式のデータなら必要に応じて加工・修正することが容易である。
3. データ入力の容易さ。XML はテキスト形式であり、かつタグを用いて要素の範囲を示せるため、入力者に見やすいようソーステキストに空白や改行を入れても、データそのものに影響はほとんどない。また、Unicode を前提としているため、特殊な文字をそのまま入力することができる。

[3]スキーマによる強力なチェック機構を備えながら、必ずしもスキーマを必要としない柔軟性も、大きなメリットである。

4. 昨今の文献管理ソフトウェアの多くは XML 形式での入出力に対応しており、データの可搬性が高い。

3 書誌情報項目の選定

書誌情報項目の選定にあたり、以下で用いられている形式を対照した。

- BibTeX[4]
- Dublin Core Metadata Element Set（DCMES）[5]
- EndNote[6]
- RIS[7]

付録 A に掲載した対応を踏まえ、今回は次の方針を立てた。

- 今後の運用により項目の取捨選択が行われることを期待し、項目数の最も多い BibTeX 形式に準じて項目を立てておく。
- 必要に応じて他の形式にも変換しやすい設計を目指す。

4 BibTeXML の概要

BibTeX 形式の XML 化については、すでに BibTeXML プロジェクト[8]）が発足しており、関連ツールが GPL ライセンスの下で配布されている。今回、その中の DTD ファイルに適宜修正を加えつつ、書誌データのマークアップを行った。以下にマークアップ例を挙げる（`bibtex` 接頭辞は名前空間 `http://bibtexml.sf.net/` を参照するものとする）。

```
<?xml version="1.0"?>
<bibtex:file xmlns:file="http://bibtexml.sf.net/">
  <bibtex:entry id="AdamsBruce_11984">
    <bibtex:phdthesis>
      <bibtex:author>Adams, Bruce</bibtex:author>
      <bibtex:title>A Tagmemic Analysis of the Wolaitta
Language</bibtex:title>
      <bibtex:school>University of London</bibtex:school>
```

[4]組版ソフトウェア TEX とともに用いられる参照文献処理ツール。

[5]World Wide Web 上のリソース情報を統一的に記述すべく、DCMI（Dublin Core Metadata Initiative）が定義した 15 個の要素。ISO 15836 として国際標準にもなっている。

[6]Thomson ResearchSoft 社が開発している文献管理ソフトウェア。

[7]さまざまな文献管理ソフトウェアでサポートされているファイル形式。

[8]`http://bibtexml.sourceforge.net/`。

```
        <bibtex:year>1984</bibtex:year>
      </bibtex:phdthesis>
    </bibtex:entry>

    <bibtex:entry id="AlmagorUri_1972">
      <bibtex:article>
        <bibtex:author>Almagor, Uri</bibtex:author>
        <bibtex:title>Name-Oxen and Ox-Names among the Dassanetch
of Southwest Ethiopia</bibtex:title>
        <bibtex:journal>Paideuma</bibtex:journal>
        <bibtex:year>1972</bibtex:year>
        <bibtex:volume>18</bibtex:volume>
        <bibtex:pages>79-96</bibtex:pages>
      </bibtex:article>
    </bibtex:entry>
  </bibtex:file>
```

ここでは、まずルートコンテナとして file 要素が置かれ、その下に文献項目を示す entry 要素が並ぶ。データベース管理を容易にするため、entry 要素は必ず id 属性を持たねばならない。entry 要素の下には、文献の種類を示すコンテナが置かれる。その中に、書誌情報の細目が並べられる。

要素タイプの詳細は付録 B に掲載している。また、こうして作成した書誌データの活用例として、XHTML と連携し Web ブラウザ上で動作する検索システムを構築した。

5 おわりに

今回作成した XML データには、改善すべき点が残されている。

- いくつかの要素タイプにおける内容の書式。例えば、著者名、編集者名は構造化されておらず、現在は検索プログラム側でファーストネームとラストネームを切り出している。それぞれにタグを付けて構造を明示すればプログラムに依存しないデータになるが、入力が繁雑になる。
- 上記に関連し、入力システムもしくは入力支援ツールの作成が必要。
- 現在、「その他の種類」(misc 要素タイプ) に分類されているものについて、新たな種類の作成と、不要と思われる種類の廃止。

- GIS との連携を視野に入れた内容モデルの再考。

これらの問題点は、XML データを実際に使用していくに伴い、適切な形で修正されるものと思われる。

【参考文献】

Dublin Core Metadata Initiative, *Dublin Core Metadata Terms,* 2008-01-14, DCMI Recommendation.
http://dublincore.org/documentts/dcmi-terms/

Gundersen, Vidar Bronken and Zeger W. Hendrikse, *BibTeX as XML markup,* 2007-01-01.
http://bibtexml.sourceforge.net/

World Wide Web Consortium, *Extensible Markup Language (XML) 1.0,* Fourth Edition, 2006-08-16, W3C Recommendation.
http://www.w3.org/TR/xml/

A 付録：書誌情報項目の対照表

	BibTeX	Dublin Core	EndNote	RIS
コンテナ	(entry types)	container	RECORD	(none)
メディア	entry types	dc:format (medium)	REFERENCE_TYPE(S)	TY
ジャンル	type	dc:type	TYPE_OF_WORK	TY?
著者	author	dc:creator	AUTHOR(S)	AU (A1)
研究機関	institution	(none)	(none)	(none)
団体	organization	(none)	(none)	(none)
大学	school	(none)	(none)	(none)
提携	affiliation	(none)	(none)	AD
書名	booktitle	dc:title	TITLE, SECONDARY_TITLE	TI T2, T3
出版社	publisher	dc:publisher	PUBLISHER	PB
出版地	address	(none)	PLACE_PUBLISHED	CY
発生地	location	(none)	(none)	(none)
出版日	year, month	dc:date	YEAR, DATE	PY, Y2
編集者	editor	(none)	SUBSIDIARY_AUTHOR(S)	A2 (ED)
版数	edition	(none)	EDITION	(none)
第 1 刷日付	(none)	(none)	(none)	PY
原著	(none)	(none)	ORIGINAL_PUB	(none)
雑誌名	journal	(none)	(none)	JO, JF, J1, J2

巻数	volume	dc:identifier	VOLUME	VL
番号	number	dc:identifier	NUMBER	IS
価格	price	(none)	(none)	(none)
サイズ	size	(none)	(none)	(none)
ISBN	ISBN	(none)	ISBN	SN
ISSN	ISSN	(none)	(none)	SN
LCCN	LCCN	(none)	(none)	(none)
主題	(none)	dc:subject	(none)	(none)
要約	abstract	dc:description	ABSTRACT	N2 (AB)
目次	contents	(none)	(none)	(none)
シリーズ	series	(none)	(none)	(none)
仕事名	title	(none)	TYPE_OF_WORK	(none)
ノート	note	(none)	NOTES	N1
注釈	annote	(none)	(none)	(none)
キーワード	keywords	(none)	KEYWORD(S)	KW
範囲	(none)	dc:coverage	(none)	(none)
出版形態	howpublished	(none)	(none)	(none)
章	chapter	(none)	(none)	(none)
ページ範囲	pages	(none)	PAGES	(none)
貢献者	(none)	dc:contributor	(none)	(none)
出典	(none)	dc:source	(none)	(none)
URI	URL	(none)	URL	UR
権利	copyright	dc:rights	(none)	(none)
言語	language	dc:language	(none)	(none)
関連	crossref	dc:relation	(none)	(none)
キー	key	(none)	REFNUM	ID

B 付録：BibTeXML 要素タイプ詳細

要素タイプ名	分類	内容
file	基本構造	entry

属性	属性値	デフォルト値	備考
(none)			

説明

データのルートを示す要素タイプであり、参考文献データは全てこの要素に含まれる。名前空間 xmlns:bibtex="http://bibtexml.sf.net/" を宣言しておくのが望ましい。

要素タイプ名	分類	内容
entry	基本構造	次の中の 1 つ。article, book, booklet, manual, techreport, mastersthesis, phdthesis, inbook, incollection, proceeding, inproceedings, conference, unpublished, misc

属性	属性値	デフォルト値	備考
id	ID	(必須)	文献データの識別子

説明

1 件の文献データを包含する要素タイプ。識別子となる id 属性値を指定しなければならない。この id 属性はデータの検索、クロスリファレンスなどに利用される。

要素タイプ名	分類	内容
article	文献の種類	author, title, journal, year, (volume, number, pages, month, note)

属性	属性値	デフォルト値	備考
(none)			

説明

雑誌、定期刊行物に掲載された論文を表す。

要素タイプ名	分類	内容
book	文献の種類	author か editor, title, publisher, year, (volume, number, series, address, edition, month, note)

属性	属性値	デフォルト値	備考
(none)			

説明

出版社が明らかな書籍を表す。

要素タイプ名	分類	内容
booklet	文献の種類	title, (author, howpublished, address, month, year, note)

属性	属性値	デフォルト値	備考
(none)			

説明
書籍の形態を持つが、出版社などが不明なものを表す。

要素タイプ名	分類	内容
inbook	文献の種類	author または editor, title, chapter または pages, (publisher, year)

属性	属性値	デフォルト値	備考
(none)			

説明
書籍中の章・節など、部分的な参照であることを示す。

要素タイプ名	分類	内容
incollection	文献の種類	author, title, booktitle, publisher, year, (editor, volume, number, series, type, chapter, pages, address, edition, month, note)

属性	属性値	デフォルト値	備考
(none)			

説明
論文書内の論文など、書籍の一部であるが独立したタイトルを持つものを表す。

要素タイプ名	分類	内容
inproceedings	文献の種類	author, title, booktitle, year, (editor, volume, number, series, pages, address, month, organization, publisher, note)

属性	属性値	デフォルト値	備考
(none)			

説明
プロシーディングに掲載された論文であることを表す。

要素タイプ名	分類	内容
conference	文献の種類	title, (author, organization, address, edition, month, year, note)

属性	属性値	デフォルト値	備考
(none)			

説明
inproceedings と同じ（Scribe との互換性のため）。

要素タイプ名	分類	内容
manual	文献の種類	title, author, organization, address, edition, month, year, note

属性	属性値	デフォルト値	備考
(none)			

説明
マニュアルなど技術文書を表す。

要素タイプ名	分類	内容
mastersthesis	文献の種類	author, title, school, year, (type, address, month, note)

属性	属性値	デフォルト値	備考
(none)			

説明
修士論文を表す。

要素タイプ名	分類	内容
phdthesis	文献の種類	author, title, school, year, (type, address, month, note)

属性	属性値	デフォルト値	備考
(none)			

説明
博士論文を表す。

要素タイプ名	分類	内容
misc	文献の種類	(author, title, howpublished, month, year, note)

属性	属性値	デフォルト値	備考
(none)			

説明
定義済み種類の範囲内では分類しにくい文献。note 要素、annote 要素を利用して補足情報を入力し、後に拡張することができる。

要素タイプ名	分類	内容
proceeding	文献の種類	title, year, (editor, volume, number, series, address, month, organization, publisher, note)

属性	属性値	デフォルト値	備考
(none)			

説明
プロシーディングであることを表す。

要素タイプ名	分類	内容
techreport	文献の種類	author, title, institution, year, (type, number, address, month, note)

属性	属性値	デフォルト値	備考
(none)			

説明
学術・研究機関から発行された報告書を表す。

要素タイプ名	分類	内容
unpublished	文献の種類	author, title, note, (month, year)

属性	属性値	デフォルト値	備考
(none)			

説明
タイトルと著者が明らかだが、公に出版されていない文書を表す。

要素タイプ名	分類	内容
address	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
出版社、発行機関などの住所を示す。

要素タイプ名	分類	内容
annote	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明

文献に関する任意の注釈。この情報が検索に用いられることは少ない。

要素タイプ名	分類	内容
author	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明

著者の名前。著者が複数の場合は and を用いて Last-name, first name and first-name last-name and ... となることが望ましいが、この形式に関しては将来的に拡張する予定である。

要素タイプ名	分類	内容
booktitle	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明

文献が書籍の一部である場合、その書籍名を示す。文献の種類が book の場合、この要素ではなく title を用いる。

要素タイプ名	分類	内容
chapter	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明

章の番号を示す。

要素タイプ名	分類	内容
crossref	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
クロスリファレンスの際に用いられる参照キー。

要素タイプ名	分類	内容
edition	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
書籍の版を示す。e.g. Second

要素タイプ名	分類	内容
editor	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
編集者の名前を示す。内容形式に関しては author 要素タイプを参照のこと。

要素タイプ名	分類	内容
howpublished	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
出版形態が特殊な場合、その形態を示す。

要素タイプ名	分類	内容
institution	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
報告書を提供している機関名を示す。

要素タイプ名	分類	内容
journal	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明

雑誌などの定期刊行物の名前を示す。

要素タイプ名	分類	内容
key	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明

文献データの並び替え、クロスリファレンス、ラベル作成などに利用される参照キーを示す。

要素タイプ名	分類	内容
month	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明

出版された（未刊行物の場合は執筆された）月を示す。

要素タイプ名	分類	内容
note	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明

文献に関する補助的な情報を示す。

要素タイプ名	分類	内容
number	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
雑誌などの定期刊行物や報告書の号数、あるいはシリーズ物の巻数を示す。

要素タイプ名	分類	内容
organization	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
カンファレンスの主催機関、あるいはマニュアルの発行元を示す。

要素タイプ名	分類	内容
pages	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
ページの範囲を示す。e.g. 7，41，73-97

要素タイプ名	分類	内容
publisher	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
出版社の名前を示す。

要素タイプ名	分類	内容
school	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
修士論文、博士論文が受理された大学名を示す。

要素タイプ名	分類	内容
series	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
書籍のシリーズ名を示す。

要素タイプ名	分類	内容
title	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
文献のタイトルを示す。

要素タイプ名	分類	内容
type	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
報告書の種類を示す。e.g. Research Note

要素タイプ名	分類	内容
volume	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
定期刊行物や複数の巻からなる書籍の巻数を示す。

要素タイプ名	分類	内容
year	書誌情報	#PCDATA

属性	属性値	デフォルト値	備考
(none)			

説明
出版された（未刊行物の場合は執筆された）年を示す。