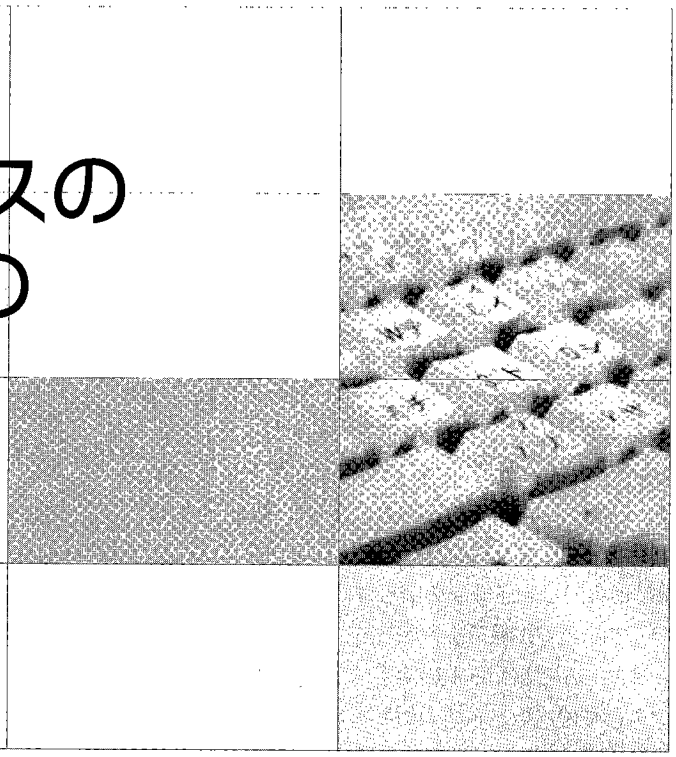
1 Webデータベースの基盤技術としてのXML



奈良先端科学技術大学院大学 情報科学研究科
吉川正俊[*1]

[*1] E-mail:yosikawa@is.aist-nara.ac.jp

XMLはWeb上のデータを地球規模のデータベースとするための基盤技術である．1998年にXML1.0が標準になって以来，多くのツール，応用システムが開発されてきている．また，周辺技術の標準化の整備や学術研究も精力的に進められている．本稿では，まず，XMLが注目される理由を技術的な観点からまとめXMLの意義について述べる．次に，XMLスキーマ言語の開発目的および名前空間の意義を説明する．さらに，XMLに基づくWebデータベースを実現するために必要な要素技術である，XMLデータの格納，問合せ言語，利用者インタフェースなどについて説明する．

XMLの位置づけ

XML（Extensible Markup Language）はデータおよび文書の書式を記述するためのメタ言語である．たとえば，大学のシラバスを表すXMLデータの例を図-1に与える．簡単にいうと，XMLデータは，文字列を（<name>のような）開始タグと（</name>のような）終了タグで囲んだ要素（element）を入れ子にしたものと，（number='523'のような）開始タグ内に記述された属性（attribute）によって，データを表現したものである．XMLデータにどのようなタグ名がどのような順序や入れ子関係で出現することを許すか，どのタグにどのような属性を許すかという文法はDTD（Document Type Definition）を用いて定義する．図-1のXMLデータのためのDTDを図-2に与える．

XMLが注目を集める理由を技術的な観点から簡単にまとめると次のようになる．

- **XMLは普遍性を持つ書式である．**

XMLは，特定のハードウェア，OS，応用プログラムな

```
<syllabus>
   <institute>
      <name>NAIST</name>
   </institute>
   <course number='523'>
      <name>Data Engineering I</name>
      <credit>2</credit>
      <outline>Introduction of XML</outline>
      <readinglist>
          <readingitem>...</readingitem>
          <readingitem>...</readingitem>
          ...
      </readinglist>
   </course>
   <course number='548'>
      ...
   </course>
   ...
</syllabus>
```

図-1 XMLデータの例

```
<!ELEMENT syllabus     (institute, course+)>
<!ELEMENT institute    (name)>
<!ELEMENT course       (name, credit, outline, readinglist*)>
<!ATTLIST course       number CDATA #IMPLIED>
<!ELEMENT name         (#PCDATA)>
<!ELEMENT credit       (#PCDATA)>
<!ELEMENT outline      (#PCDATA)>
<!ELEMENT readinglist  (readingitem+)>
<!ELEMENT readingitem  (#PCDATA)>
```

図-2 DTDの例

```
...
<elementRule role="course">
        <sequence>
          <ref label="name"/>
          <ref label="credit"/>
          <ref label="outline"/>
          <ref label="readinglist" occurs="*"/>
        </sequence>
</elementRule>
<elementRule role="name" type="string"/>
<elementRule role="credit" type="integer"/>
...
<tag name="course">
        <attribute name="number" required="true"
                                        type="integer"/>
</tag>
<tag name="name"/>
<tag name="credit"/>
...
```

図-3　RELAXの例

どにはいっさい依存していない．ワープロで作成した文書ファイルが数年もするとバージョンが古くなりまったく読めなくなることがあるが，XMLデータはこのような事態とは無縁である．XMLは文字列でデータを表現するため文字コードのみに依存している．

- **XMLデータは計算機による厳密な処理が可能でありかつ人間が可読である．**
  XMLデータは，要素の名前とその内容，および属性と属性値に基づいた計算機処理が可能である．しかも，データ全体が文字列であるため人間が読むことができ，通常のエディタでもデータの作成，編集が可能である．HTMLは，人間による可読性の点ではXMLと同様であるが，終了タグの省略が可能である点など計算機による厳密な処理には適していない．

- **XMLは汎用性を持つ書式である．**
  XMLはDTDを用いて個々の応用のための言語を自由に定義することのできるメタ言語である．XMLはその汎用性ゆえに応用分野を限定しない．それに対し，HTMLはそれ自身が言語であり，最初からタグの種類とその意味が固定されている．

インターネットによって既存の組織の枠組みを超えた情報交換がきわめて容易となり，そのことはひいては組織の再編成を促すという種類の議論はすでに何度となくされている．XMLが注目を集める理由は，それが，上に述べた技術的特徴により，インターネット上で着実に進行しつつあるこのような大きな流れを加速する基盤技術となる点にある．

たとえば，全国の大学のシラバスがすべて図-1のような書式のXMLで表現され，それらがインターネット上でアクセス可能であれば，「XMLについて教えている科目名とその大学を探したい」という検索要求に即座に答えることができる．講義をインターネットを介して受講可能であれば，学生は全国の大学から受講したい科目を自由に選択することができる．そうなれば大学という組織のビジネスモデルを再構築する必要が生じることは容易に想像できる．もちろん，このようなシナリオが現実となるためには技術以前のところで多くの問題があるだろう．また，XMLによるデータの表現自体がプロジェクトの成功を約束するものではないことや，XML利用の際には確実な要求分析に基づいた書式の設計が肝要であることは，本誌でもすでに議論されているところである[7]．XMLは単なる書式に過ぎないが，インターネット上の計算機処理可能な中立性の高い共通書式は大きな社会的インパクトを与える可能性を持つという点にXMLの意義がある．

スキーマ言語

XMLは文書とデータ両方の記述のために利用されているため，DTDは，文書書式を規定する文法と，データベーススキーマ両方の役割を果たす必要がある．しかし，XMLのDTDは，基本的にSGMLのそれを踏襲しており，元来文書書式を定義するための文法記述であったため，いくつかの問題点を持つ．まず，DTDではデータ型は基本的には文字列しかない（たとえば，図-1のXMLデータの要素creditの内容は数値ではなく長さ1の文字列である）．また，モジュール化の機構がきわめて貧弱であるため大規模DTDを開発しにくいことや，独特の構文で記述されているためツールが作りにくいこと，さらに，DTDと後述する名前空間の併用は制限されるという問題点もある．

このような問題点を解消するために，（XML）スキーマ言語と呼ばれる言語がいくつか開発されている．代表的なものとしては，W3CのXMLSchemaやRELAX[4]，TREX[2]などがある．これらのスキーマ言語に共通する設計目標は次のようにまとめることができる．

(1) 豊富なデータ型を導入する（SQLやプログラミング言語のデータ型を取り入れる）．
(2) スキーマをXML構文で記述することにより，XMLのツールをそのままスキーマ開発に利用できるようにする．
(3) 名前空間を利用し，スキーマ部品の共有と再利用の機構を導入する．

(1)，(2) の2点を例示するために，図-1のXMLデータのためのスキーマの一部をRELAXで表したものを図-3に示す．次に，(3) の名前空間について説明する．

■名前空間

XMLを活用するためには，利用目的に応じて共通のスキーマを設計しそれを普及させることが鍵となる．ただし大規模なスキーマの開発は多大な労力を要する．スキーマを部品化し組み合わせて利用することができれば，開発コストの低減化とスキーマ部品の普及を図ることができる．

スキーマ言語では，複数個のスキーマ部品を組み合わせて新たなスキーマを構築する機構が導入されている．独立に開発されたスキーマ部品の間で要素名や属性名の衝突を避けるため，スキーマ部品はURIを用いて識別する．ある1つのURIで識別されるスキーマ部品中の要素名や属性名の集まりを名前空間（namespace）と呼んでいる．スキーマの検証器は，複数の名前空間のスキーマ部品を参照するXMLデータが各スキーマ部品に対して合法か否かを検証する．たとえば，図-4に示すXMLデータの場合は，根要素を含む文書の一番外側の部分Aは名前空間1のスキーマ部品を参照するが，その部分文書であるBとEは名前空間3のスキーマ部品を参照するなど，全部で3つの名前空間のスキーマ部品を参照している．名前空間を利用したインターネット上のスキーマ部品の共有と再利用の機構により，各応用に適したカスタムメイドのスキーマを作ることができる．このような機構は従来のデータベースシステムにはなく，XMLが広く普及するためにはきわめて重要な概念である．

実際のXMLデータ中での名前空間の指定法は次のようになる．たとえば，講義のreadingitemの内容は，著作物のメタデータとして標準化が進んでいるDublin Coreの名前空間を利用して記述するならば，図-1のreadingitem要素は，図-5のようになる．図-5において，開始タグ<readingitem>中の属性xmlns:dcは，その要素自身や子要素に現れる'dc:'を接頭辞とする要素や属性は，http://purl.org/dc/elements/1.1/で識別される名前空間のものであることを宣言する．'dc'は，このXMLデータにおいてのみ有効な局所的な識別子である．

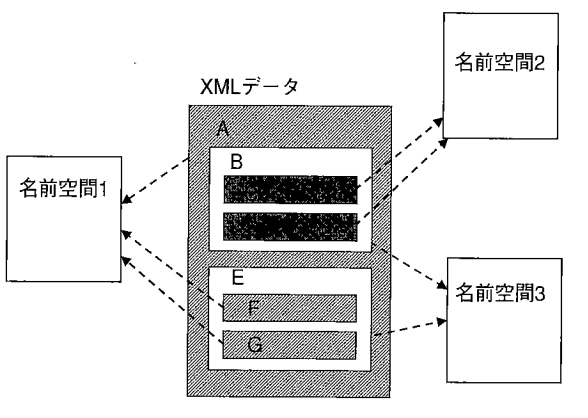


図-4　複数の名前空間のスキーマ部品を参照するXMLデータ

各図書館がこのようにDublin Coreを共通スキーマ部品として利用し書誌情報を記述していれば，科目のreadinglistと図書館の蔵書との間の照合を簡単に行えることになる．

■スキーマ言語戦争

前述のように，現在スキーマ言語としていくつかのものが提案されている．W3CのXML Schemaは仕様が巨大で複雑すぎるため実装が困難なことが懸念されている．RELAXは，基本的には，現在のDTDの意味論を少し拡張したものをXML構文で記述し，データ型を取り入れることを目的としている．スキーマ言語の普及のためには，言語の表現能力や明瞭性に加えて，ツールが充実していることも重要である．RELAXはDTDとの共存やDTDからの移行を考慮して設計されている．また，Relaxerと呼ばれるXMLからJavaクラスを自動生成するプログラムなどのツールも開発されている．

RELAXとTREXの統合の動きもあり，複数個のスキーマ言語が併存することになるのか，あるいはいずれか1つのものに収斂するのか，戦争の帰趨は予断を許さない状況である．

XMLに基づくWebデータベース

従来，インターネット上のHTML文書の総体を漠然とWebデータベースと呼ぶことがあった．XMLの出現により，Webデータベースは，データベース管理システムが対象としてきたような，厳密に計算機処理可能なデータの集まりとなる．XMLに基づくWebデータベースとは，インターネット上でアクセス可能な実XMLデータまたは仮想XMLデータの集まりである．図-6にその概念図を与える．

■XML出版とXML変換

データ自体が物理的にはどのような形で管理されていても，それらをXMLに変換し，インターネット上でアクセス可能とする限り，（後述するXQueryなどの）XML問合せ言語によって統一的な検索が可能となる．データベース管理システムで管理されている非XMLデータをXMLに変換することはXML出版と呼ばれている．図-6では，関係データベースやオブジェクト指向データベースの非XMLデ

```
<readingitem xmlns:dc="http://purl.org/dc/elements/1.1/">
    <dc:title>Data on the Web</dc:title>
    <dc:creator>Serge Abiteboul</dc:creator>
    <dc:creator>Peter Buneman</dc:creator>
    <dc:creator>Dan Suciu</dc:creator>
    <dc:publisher>Morgan Kaufmann</dc:publisher>
</readingitem>
```

図-5　名前空間の利用例

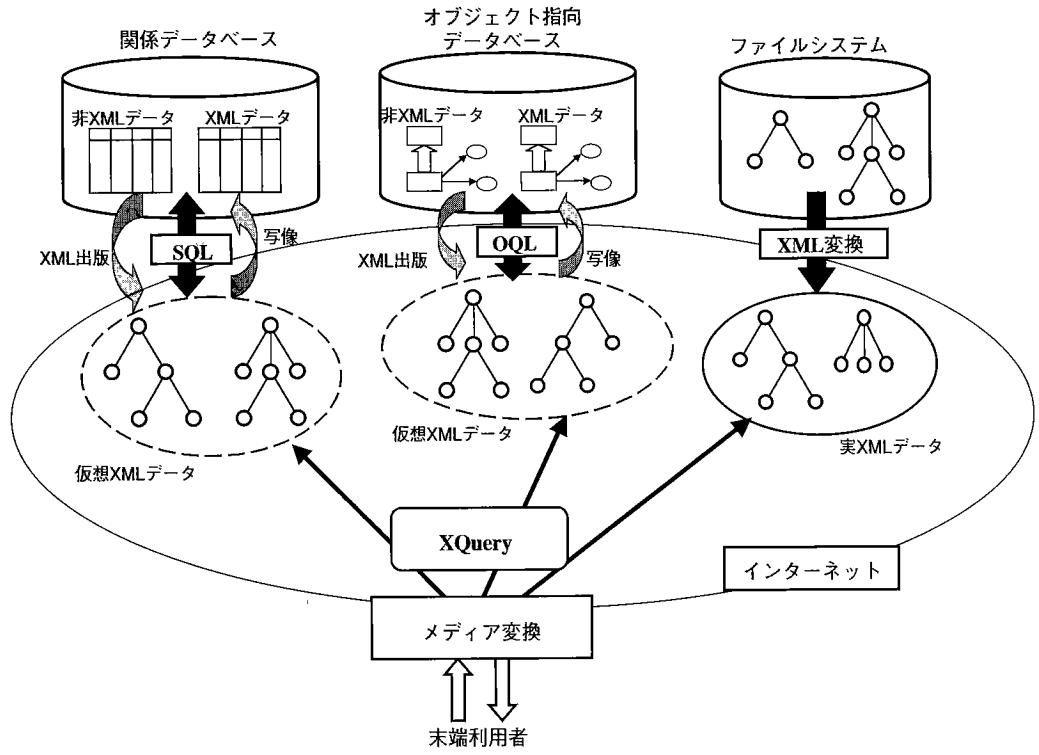


図-6　XMLに基づくWebデータベース

ータが仮想的にXMLデータとして出版されており，また，通常のファイルとして管理されているXMLデータが別の実XMLデータに変換されてインターネット上でアクセス可能となっている．

XML変換のためには，XSLTやXDuce[3]などの言語を用いることができる．XDuce（'transduce'と発音）は正規表現を表す型と正規表現パターンマッチの機能を持つ静的型付きプログラミング言語である．正規表現型は，DTDやスキーマ言語に通常みられる（*，?，|などを用いた）正規表現を表す型であり，正規表現パターンマッチは，関数型言語にみられるような通常のパターンマッチ機能に，繰り返しと選択を入れて拡張したものである．

■XMLデータの格納

大量に流通するXMLデータの格納管理は重要な課題である．関係データベースやオブジェクト指向データベースを利用したXMLデータの格納法がいくつか提案されている[6]．

一例として，XRel[5]では，関係データベースを利用してXMLデータを格納する．XRelでは，XMLデータの各要素や属性に関する情報を根要素からの経路と先頭文字からのバイト数の組合せで表現する．たとえば，図-1のXMLデータは，図-7の関係表で表現する．問合せはXPath 式として与えられSQLに変換される．XRelでは，関係スキーマは格納するXMLデータの構造と独立しているため，DTDに無関係にすべてのXMLデータを格納できる．また，経路自体を文字列として関係表に格納しているため，他の多くの格納法とは異なり，XPathの'//'を含む式で表現されるような曖昧な問合せを処理するために再帰問合せや多くの結合を必要としないことも利点である．

XMLデータの格納は今後とも重要な研究課題である．XML問合せ処理のベンチマークも開発されはじめているため，詳細な評価実験をもとにした各手法の比較が行われるようになることが予想される．

■問合せ代数と問合せ言語

XMLデータは，関係データベースやオブジェクト指向データベースとはデータモデルが異なるため，新たな問合せ言語を必要とする．W3Cでは，XMLのための問合せ代数と問合せ言語XQueryを開発中である．問合せ代数は，XQueryの意味論を与え，種々の書き換え規則により問合せ最適化プラン生成の指針を与える．

XML問合せ代数は静的な型システムである．すなわち，ある問合せの出力データの型は，その問合せの解析時に決定できる．したがって，型誤りの検出や，型情報を利用した問合せ最適化が可能である．また，対象とするデータは順序付き森（ordered forest）である．これは，XMLデータは木で表現できることと，文書を表現する場合に構成要素間の順序が重要であることによる．XQueryは以下のような思想に基づいて設計されている．

- 従来データベースとみなされていた情報と従来文書と

Element

docID	*pathID*	*Start*	*end*	*index*	*reindex*
1	1	0	5628	1	1
1	2	10	50	1	1
1	3	21	38	1	1
1	4	51	272	1	1
1	6	72	102	1	1
1	7	103	120	1	1
1	8	121	158	1	1
1	9	159	263	1	1
1	…	…	…	…	…

Attribute

docID	*pathID*	*Start*	*end*	*value*
1	5	52	52	523
1	5	274	274	548
1	…	…	…	…

Text

docID	*pathID*	*Start*	*end*	*value*
1	3	27	31	NAIST
1	6	78	95	Data Engineering I
1	7	111	111	2
1	8	130	148	Introduction to XML
1	9	172	249	…
1	…	…	…	…

Path

pathID	*pathexp*
1	#/syllabus
2	#/syllabus#/institute
3	#/syllabus#/institute#/name
4	#/syllabus#/course
5	#/syllabus#/course#@number
6	#/syllabus#/course#/name
7	#/syllabus#/course#/credit
8	#/syllabus#/course#/outline
9	#/syllabus#/course#/readinglist
10	#/syllabus#/course#/readinglist#/readingitem

図-7　XRel における関係データベースを用いた XML データ格納法

みなされていた情報のいずれも問合せ対象とする柔軟な問合せ言語とする．

- 小さく，容易に実装可能な言語で，問合せは簡潔で簡単に理解できるようにする．
- 人間が可読な問合せ構文を採用し，XMLに基づく問合せ構文は別に定義する．

たとえば，図-1のXMLデータにsyl.xmlという名前が付けられているとすると，次のXQuery問合せは，readingitemが10以上ある科目を求める．

```
<heavyclass>
     FOR $c IN
          distinct(document("syl.xml")//course)
     LET $r := $c//readingitem
     WHERE count($r) >= 10
     RETURN $c
</heavyclass>
```

当然のことながら，XML問合せ代数やXQueryは，XML SchemaやXPathとの整合性をとりながら標準化が進められる．この整合性がうまくとれることと実装が早期に現れることがXQueryが実際に普及するための鍵となるだろう．

■利用者インタフェース

XML問合せ代数やXML問合せ言語は，問合せを厳密に定義するためにWebデータベース内部で利用されるものであり，末端利用者のためのものではない．末端利用者は現在のサーチエンジンのようにより簡便な方法で問合せ要求を入力することを好むであろう．そのような問合せ要求をもとに，XML問合せや言語や情報検索技術を利用し，適合率，再現率の高い検索システムを構築する技術は重要である[1]．

また，PDA，携帯電話，自動車など，いつでもどこでもWebデータベースにアクセスできるような環境が整いつつあり，末端利用者には利用する機器に応じて使いやすい利用者インタフェースを提供する必要がある．音声ブラウザなども開発が進んでおり，入出力のメディア変換も必要となる．一方，Webコンテンツ開発者としては，利用者の機器とは独立にコンテンツを開発しておき，機器に応じて必要な変換を行いたいという要求がある．このような問題に取り組むために，クライアント機器の能力や優先選択肢を記述するための枠組みであるW3C のCC/PP（Composite Capabilities / Preference Profiles）などの開発が進んでいる．

ブームは終焉を迎えるか？

XMLは過去2，3年の間，異常とも思える注目を集めた．しかし，一過性のブームで終わるようなたぐいの技術ではないことは明らかである．本文中に述べたように，スキーマ言語，問合せ言語など今後の技術展開のための基盤となる標準作りの努力がまさに今なされているところである．このような標準や関連技術は，データベース，プログラミング言語，文書処理システムなどの過去の成果を取り入れ，強固な理論的基盤に基づきしかも実際上有用な提案がされている．また，XMLを利用した応用別標準の開発も広がっており，情報処理に分野を限定しても，地理情報システムのためのG-XMLプロトコル，マルチメディアコンテンツ記述インタフェースMPEG-7などを挙げることができる．今後，標準化作業，研究，実践の積み重ねにより，XML技術を健全に発展させる努力が必要である．

参考文献
1) ACM SIGIR2000 Workshop on XML and Information Retrieval, http://www.haifa.il.ibm.com/sigir00-xml/ (July 2000)．
2) Clark, J.: TREX - Tree Regular Expressions for XML, http://www.thaiopensource.com/trex/.
3) Hosoya, H.: XDuce: A Typed XML Processing Language, http://www.cis.upenn.edu/~hahosoya/xduce/.
4) Murata, M.: RELAX (REgular LAnguage description for XML), http://www.xml.gr.jp/relax/.
5) Yoshikawa, M., Amagasa, T., Shimura, T. and Uemura, S.: XRel: A Path-Based Approach to Storage and Retrieval of XML Documents using Relational Databases, ACM Transactions on Internet Technology, Vol.1, No.1 (June 2001).
6) 吉川正俊: XMLとデータベース (1) (2), bit, Vol.32, No.3, 4, pp.68-73, 82-87, (Mar., Apr. 2000).
7) 鈴木純一，村田　真，奥井康弘，大野邦夫，鈴木純一: XMLの悩み：どこでどう使うべきか／乱立する新技術はXMLの実用化と関係ない／XMLは単なるシンタックスと心得よ／組織文化を克服するXML専門家の育成／各氏へのコメント，情報処理，Vol.41, No.12, pp.1398-1403 (Dec. 2000).

(平成13年5月31日受付)