XMLと情報処理の未来

檜山正幸
hiyama@HMO.iijnet.or.jp

「XML」という言葉を雑誌や新聞でも見かけるようになった。本誌においては、すでに村田真氏の解説が5月号に掲載されている。XMLは今、世界的なブームを巻き起こしている。しかし現時点では、XMLアプリケーションや実際の活用事例はまだ少ない。XMLの機能と可能性に関して、実感を持ちにくい読者も多いだろう。この記事では、XMLの全体像を描き、今後おとずれるであろうXML革命の予感を読者と共有することを目的としたい。

●XML言語：タグ付きテキストの作り方を一般的に定めたルール。
●XMLスタイルシート言語：表示や印刷の際の、XML文書の見た目や体裁を指定するための言語。XSL（XML Stylesheet Language）と略称される。
●XMLハイパーリンク機能：HTMLのアンカーを大幅に拡張したハイパーリンク機能を与える。ウェブ上のXML文書やその一部を指し示すための構文であるXPointerと、XML文書（あるいは、その一部）の間の関係を記述するXLinkからなる。
●XML文書オブジェクトモデル：プログラムやスクリプトから、XML文書を扱うときのデータ構造とAPI【※2】を定義する。XML DOM（Document Object Model）と略称される。

■XML技術のつながり
■アイデアの取り入れ
■プログラミングインターフェース

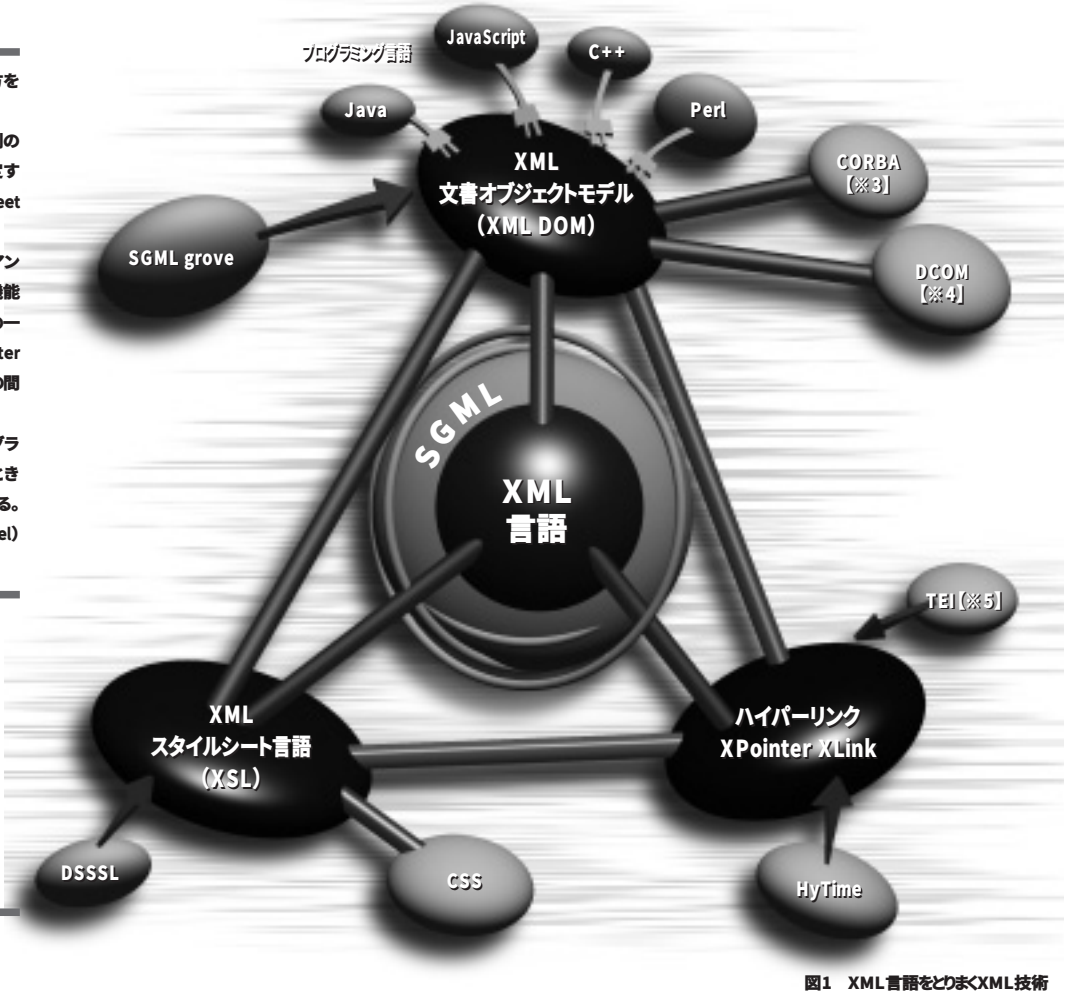


図1　XML言語をとりまくXML技術

XMLの全体像

●XML言語とXML技術

XML（eXtensible Markup Language）は、今年の2月にWorld Wide Web Consortium（W3C）【※1】勧告となったメタマークアップ言語である。この勧告は、W3CのXMLプロジェクトの大きな成果である。しかし、これはまだほんの一里塚にすぎない。XML関連技術の開発が引き続き精力的に行われている。

この記事では、XMLの名で呼ばれる2つの概念を区別するために、「XML言語」と「XML技術」という言葉を使う。「XML言語」は、W3C勧告「XML version 1.0」に定義されているメタマークアップ言語そのものを指す（XMLの「L」は「言語」を意味するので、”XML言語”は少し変な言葉だが、お許しいただこう）。一方、XML言語を中核として現在開発が進んでいる総合的な技術体系を「XML技術」と呼ぶ。

XML技術の重要な要素としては、XML言語以外に、スタイルシート言語、ハイパーリンク機能、文書オブジェクトモデルがある。

●XMLはHTMLの拡張ではない

「HTMLの次はXMLだ」とか「XMLは進化するHTML」などのフレーズを見かける。なにごとにもキャッチフレーズは必要だから、目くじらを立てるつもりはないが、XML言語は、HTMLの「次」でもなければ、「進化する」もの、「進化した」ものでもない。HTMLはタグ付けのルールだった。それに対してXML言語は「タグ付けのルールのルール」である。つまり、マークアップ言語に対するメタ（超）ルールなのである。

具体例を出そう。あなたが会社の報告書を書くのに、HTMLを使っているとしよう。<HEAD>には<TITLE>（題名）を書けるが、報告執筆者や日付を識別するためのタグはない。これに不満を感じても、残念ながらどうしようもない。HTMLタグを決められるのはW3Cと有力なWWWブラウザーベンダーだけである。

XML言語なら、目的に応じて新しいタグを決めて使うことができる。たとえば報告書なら次のようなタグ付けができる。

```
<報告書>
<ヘッダ>
<題名>新しいサーバー導入の経過</題名>
<報告者>岡本 健二</報告者>
<宛先 敬称="様">藤崎 課長</宛先>
<報告日付 標準値="19980525">
平成10年5月25日</報告日付>
</ヘッダ>
<本文>
<節>
  <見出し>進捗状況</見出し>
  <段落>先日よりすすめています新サーバー導入計画は
... </段落>...
</節>
 :
</本文>
</報告書>
```

では、<報告書>や<宛先>などのタグが「XML version 1.0」で決められているのだろうか。もちろん、そんなことはない。無数にある応用で使われるタグを前もって決められるわけはない。タグセットを決めるのは、あなたやあなたの会社であり、固有のタグセットを作るとき、使うときに守るべきルールが「XML version 1.0」なのである。

XML言語の立場からは、HTMLも、数多くあるタグセットの1つにすぎない。もっとも、現在のHTMLはXML適合ではない（XMLのルールに従っていない）。理屈の上ではHTMLをXML適合にすることも簡単だが、圧倒的に広く使われているHTMLは別格だと考える必要がある。互換性を捨ててHTMLのXML適合を強行するのは得策ではない。つまり、XMLに従う諸々のタグセットと、XMLからはずれるが最もメジャーなタグセットであるHTMLが当面は併存していく構図となる。

●XML技術と規格

「HTMLは、国際規格SGML【※7】にのっとっている」という話を聞いた方もいるだろう（今存在するHTML文書はSGML的に正しいとは言えないのだが）。HTMLはSGMLの応用事例の1つにすぎないが、XML言語はSGML規格全体を軽量化したものである。SGMLとXMLの関係は、OSI【※8】とTCP/IP、X.500【※9】とLDAP【※10】の関係に似ている。どうも、ISO【※6】の規格は立派すぎて成功しないようだ。

ISOのSGML関連の規格の1つに、スタイルを指定するDSSSL【※11】がある。XSLはこのDSSSLを取り入れるだろう。XPointerやXLinkはハイパーリンクのISO規格HyTime【※12】のアイデアを使っている。DOMには、やはりISOで開発されたSGML grove【※13】という概念が影響を与えている。つまり、大げさすぎて実用化が困難だったISOの規格群を、インターネット向けに綿密に構成し直したものがXML技術だと言える。

図2　HTMLタグセットとXML適合タグセット

現状のHTMLはSGMLから外れていることが多い

SGMLを軽量化したXMLの上に各種のタグセットが作られる

XMLがデータを統一する

●データの互換性という永遠の課題

コンピュータの役割は、広い意味のデータ処理である。では、処理対象となるデータにはどんなものがあるだろうか。最も古くから扱われてきたのは、表の形をした定型データである。住所録とか在庫目録などを思い浮かべていただきたい。一方では、企画書や論文のような文書もコンピュータで扱われる。その中間に、請求伝票、稟議書、申込書などの伝票や帳票類がある。

扱うデータの種類に応じてアプリケーションが分かれる。定型の表は主にリレーショナルデータベースの守備範囲であり、非定型の文書はワープロによって処理される。そして、それぞれのソフトウェア製品が独自のデータフォーマットを使っており、常にデータの互換性が問題となってきた。多くの試みがなされたが、いまだに満足のいく状況ではない。処理効率やベンダーの戦略的な問題などがからみ、共通のデータフォーマットは実現していない。

マイクロソフトのOLEや かつてアップル、IBMなどが開発していたOpenDocは、それぞれのアプリケーションが独自のバイナリーフォーマットを使ったままで、「コンパウンドドキュメント【※14】」という形式をベースとしてアプリケーション統合を進めようという技術だ。このアプローチは、もちろん意義を持ち続けるだろうが、今、われわれにはもう1つの道が見えている。データ記述の共通インフラとして、「タグ付きテキスト」を採用するという選択である。

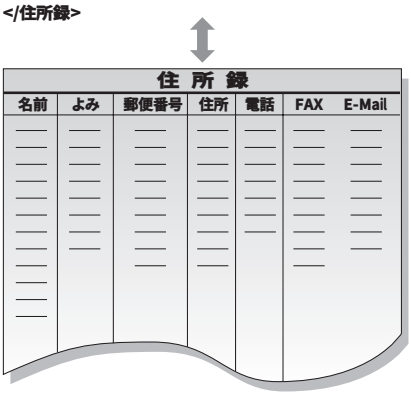
●XML言語の強力な表現力

タグ付きテキストが強力なパワーを秘めていることは、HTMLの経験を通して、すでに確認された事実である。実際、多くの人々が、OSにもアプリケーションにも依存せず、インターネット上を流通するデータフォーマットであるHTMLに飛びついた。そして、ウェブの初期段階では想像もできなかったほど多様なHTMLの応用が行われてきたわけである。

しかし、これは「やりすぎた」感がある。本来のHTMLの能力を超えて、あまりにも過大な役割を押しつけてしまった。今まで、リレーショナルデータベース、ワープロ、表計算、グループウェアなどの専用アプリケーションが扱ってきた各種のデータをHTMLに置き換えるのは、いくら何でも無理がある。

しかし、XML技術を使えば、すべてのアプリケーションの共通データ形式の実現もあながち夢ではない。XMLでは、目的に応じて、自由にタグセットとマークアップのルールを決められるから、たいていのデータフォーマットは表現できる。報告書の例で示したような「文書的なデータ」でも「表のデータ」でも、XMLのタグ付きテキストとして書ける。

```
<住所録>
<住所録項目>
  <名前>奥山 達夫</名前>
  <よみ>おくやま たつお</よみ>
  <郵便番号>169</郵便番号>
  <住所>新宿区高田馬場 x-yy-zz</住所>
  <電話>03-xxxx-yyyy</電話>
  <FAX>03-xxxx-zzzz</FAX>
  <E-Mail>t-okuyama@notexist.co.jp
  </E-Mail>
</住所録項目>
<住所録項目>
  :
</住所録項目>
  :
</住所録>
```



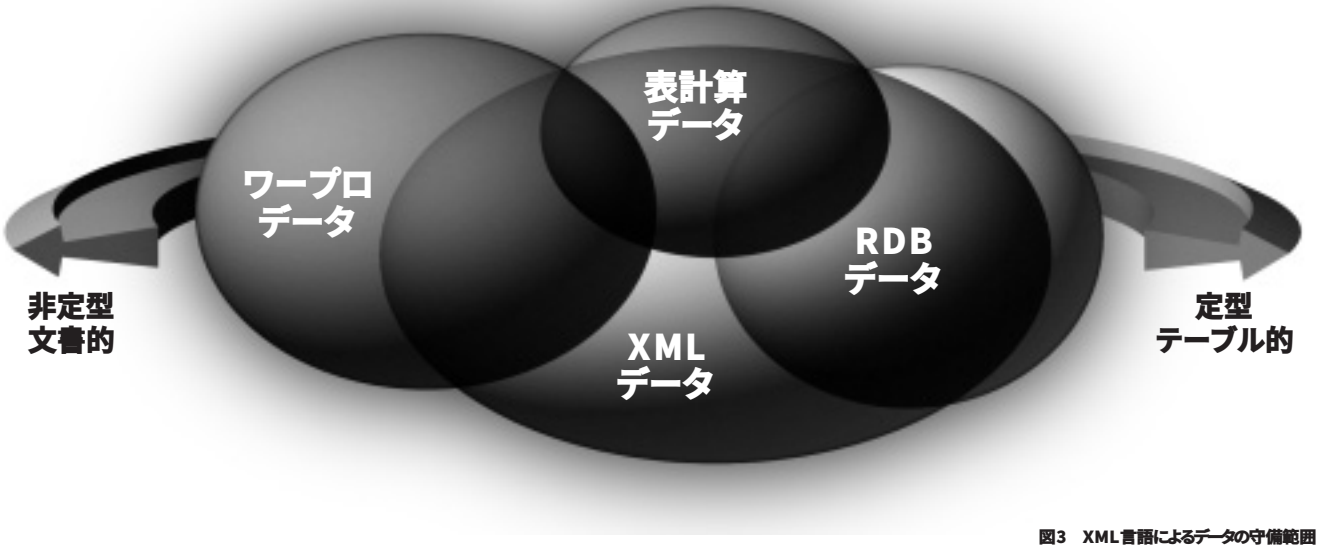


図3　XML言語によるデータの守備範囲

●データを必要なだけ利用する

XML言語で書かれた文書には、きわだった特徴がある。プログラムがタグに関する情報をまったく持たなくても、その構造ツリーを構成できるのだ（図4）。未知のタグ付き文書に対しても、特定の名前を持つタグの内容を抜き出したり、文字列を検索したり、ある属性（タグの中に書かれる「名前=値」という形の指定）の有無をチェックしたりできる。また、プログラムが知っているタグだけを処理の対象にすることもできる。OLEなどのコンポーネント技術が、コンポーネント内のデータをまったく外に見せないのとは対照的である。

これは、WWWブラウザーがHTMLを処理する際に、知らないタグを読み飛ばす動作や、プラグインにより機能を追加する方法などの延長にある。完全にデータ交換ができなくても破綻はしない。最悪のときでも、人の目で見てある程度の解釈、判断ができる。また、あるXML文書で使っているタグセットの定義やスタイルシートはURLで示される。どう処理すべきかの情報がインターネットから入手可能なのだ。

XML技術では、タグ付きテキストという単純で透明なフォーマットを採用し、応用分野、使用目的ごとのタグセットを許す。そして必要に応じて、タグセットの仕様、処理モジュールなどをインターネットからダウンロードできる。異なるアプリケーション同士がデータを交換したり共有したりする（たとえば、ワープロが住所録を扱う）場合でも、用途と能力に応じた取り扱いが可能だ。この方式により、アプリケーションとデータの関係はまったく新しい局面を迎えることだろう。

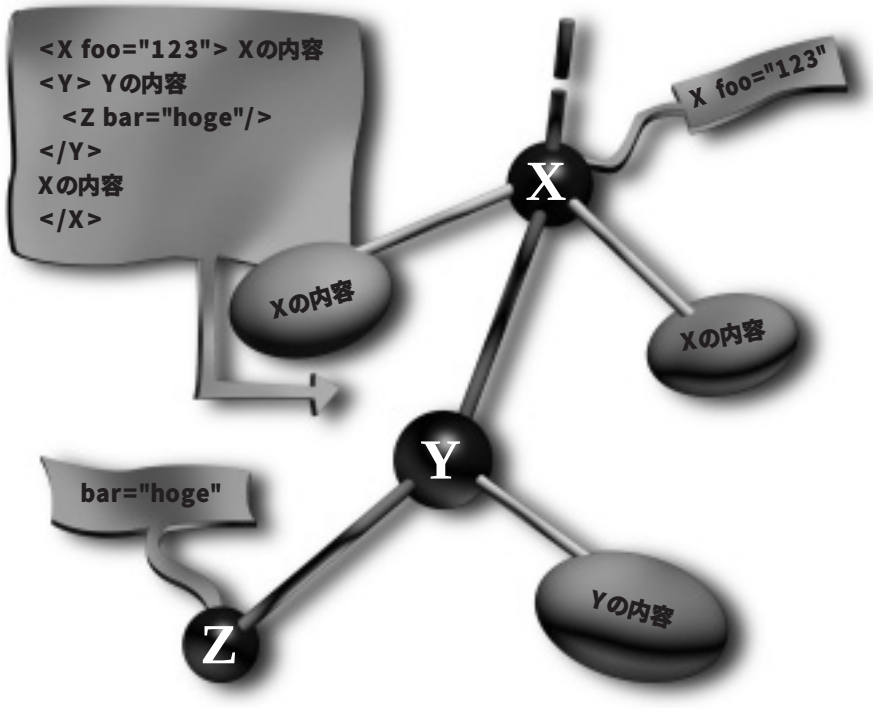


図4　知らないタグでもツリーを構成できる。

■用語解説

【※1】W3C（World Wide Web Consortium）
WWWの相互運用性を高めるため、新しい技術を開発し、データフォーマットやプロトコルを標準化する団体。

【※2】API（Application Programming Interface）
プログラムから使用するデータ構造や、そのデータ構造へのアクセス方法などの取り決め。

【※3】CORBA（Common Object Request Broker Architecture）
Object Management Groupが策定した分散オブジェクト技術のアーキテクチャー。

【※4】DCOM（Distributed Component Object Model）
マイクロソフトの分散オブジェクトアーキテクチャー。

【※5】TEI（Text Encoding Initiative）
文学作品なども含めて、文献のSGML化に取り組むプロジェクト。また、そこで開発された技術。ハイパーリンクに関して優れた提案がある。

【※6】ISO（International Organization for Standardization）
国際標準化機構。日本では、JISがISOと連携して国内規格を定めている。

【※7】SGML（Standard Generalized Markup Language）
マークアップ言語に関する国際規格。HTMLはその応用事例であり、XMLはそのサブセットである。「ISO 8879:1986」。

【※8】OSI（Open System Interconnection）
ISOによるネットワークアーキテクチャーと規格群。7階層モデルは有名。製品は少なく、ほとんど使われていない。

【※9】X.500
ISOによるディレクトリー規格。あまり製品化が進んでいない。

【※10】LDAP（Light-weight Directory Access Protocol）
X.500をベースに、インターネットでの使用のために軽量化したディレクトリーの仕様。「RFC1798」。

【※11】DSSSL（Document Style Semantics and Specification Language）
SGML文書に対して、体裁指定を行う言語の規格。「ISO/IEC 10179:1996」。

【※12】HyTime（Hypermedia/Time-based Structuring Language）
ハイパーメディアと関連する技術の規格。「ISO/IEC 10744:1992」。

【※13】SGML grove
DSSSLなどで使われたSGMLデータ構造。ツリー構造であり、各ノードはプロパティーを持つ。

【※14】コンパウンド・ドキュメント
入れ物（コンテナ）となる枠組みに、文字、表、画像、動画、音声などを表す部品を配置して構成する文書。マルチメディア文書の実現形態の1つであり、コンポーネント技術の構成要素でもある。

XMLが生み出す無限の可能性

●どんな環境でも使えるXML

XML文書は、XML文書オブジェクトモデルを介することにより、Java、JavaScript、C、C++、Perlなどの言語で操作可能となる。XML技術では、特定のプログラミング言語を仮定しない。プログラミング言語に中立なデータフォーマットなのである。そしてもちろん、XML文書は、携帯情報端末から大型汎用機まで、そのプラットフォームを選ばない。XML技術は、ハードウェア、OS、プログラミング言語、文字コードなどの一切から独立している。

XML言語によるデータ記述が一般化すれば、たとえば次のような応用が考えられる。初対面の人と、お互いのPDAにある名刺情報を赤外線で交換する。名刺情報は、もちろんXMLフォーマットである。そして、そのフォーマットは住所録データベースとも互換性があり、パソコンのアプリケーションでも処理できる。PDAに蓄えられた新しい名刺情報は、自宅のパソコンの住所録にも反映され、会社のサーバー上のデータベースとも同期される。年賀状のシーズン、宛名印刷がしたくなったら、ワープロのXML処理機能を使っても、Javaで書かれたアプレットをダウンロードして使ってもよい。名刺情報の形式が公開されれば、さまざまな処理プログラムがそろうに違いない。

●XMLとウェブエージェント

もう1つ考えられる応用を示しておこう。今や、求人、就職活動にインターネットが使われるのは珍しくない。企業は、会社案内や求人情報をウェブページとして公開している。就職活動をしている学生は、自分をアピールするためにページを作る。しかし、HTMLで書かれたこれらの情報が処理可能な形でウェブに載っているわけではない。ある条件を満たす人を探したり、逆に希望する企業の候補を挙げるには、人間の手と目を使わざるをえない。

ウェブエージェント（ロボット）を使ったところで、せいぜい、キーワード（たとえば、“求人”、“コンピュータ関係”、“神奈川”など）をたよりにするくらいで、正確な検索や絞り込みは無理である。しかし、もし求人、就職という特定の目的にふさわしいタグセットがXMLに従って策定されれば、精度が高い情報をインターネットから得られることになる。

●インターネットに秩序を与えるXML

ウェブは分散データベースと呼べなくもない。しかし、ひどく効率が悪く、無構造なデータベースである。データベースを「構造化された情報貯蔵庫」と定義するなら、データベースという言葉を使えないかもしれない。秩序がないことがインターネットの特質とも言えるが、目的によっては秩序が必要になる。

インターネットに秩序を与えるのがXML技術だと言える。XMLの仕様には、HTMLの気楽さ（悪く言えばズボラさ）も許容する配慮がされているが、基本的には厳格である。運用の仕方によっては、非常に精密な構造を持つ情報を作り出すことができる。インターネット上で情報処理を行うプログラムは、XMLが作り出す構造を前提に正確な処理ができる。たとえば、企業の求人ページを回って「神奈川県内で新横浜から通勤時間40分以内、コンピュータ関係ベンチャーで、特にJavaプログラマーを募集している会社」を正確に列挙できる。

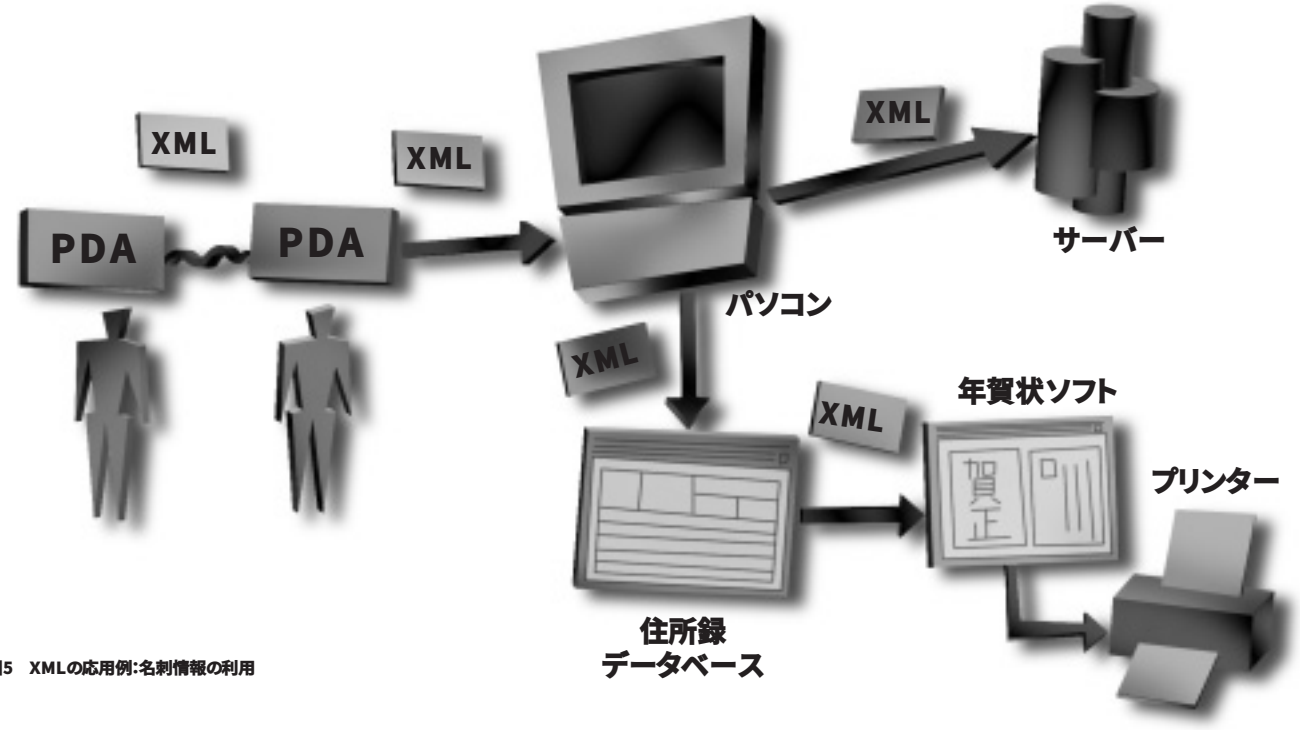


図5　XMLの応用例:名刺情報の利用

●データとプログラムのインターフェイスとしてのXML

ネットワーク上のプログラミング技術においては、異なるコンピュータ上に置かれたプログラムどうしが呼び出しを行うための方法、呼び出しインターフェイスの記述、定義されたインターフェイスの管理（インターフェイスリポジトリ）などの手法が開発されてきた。しかし、プログラムと処理対象であるデータのインターフェイスは意外にも注目されなかった。XMLは、プログラムとデータのあいだに明確なインターフェイスを与える。

XML文書は、その目的や適用分野などにより分類される。それぞれの種別をXML文書型（document type: DOCTYPEと略称される）という。報告書、名刺情報、会社案内、履歴書などが文書型の例である。この記事で「タグセット」と呼んだものが文書型である。プログラムが処理の際に、文書型に対する情報が必要なら、プログラムはそれをインターネットから読み込むことができる。たとえば、「<名前>というタグのデータを、姓と名に区切ることができるか?」、「<日付>は省略可能か?」などは、文書型の情報を見ればわかる。XML文書は、自分がどういう構造を持ち、どう処理されるべきかの情報までも持っているのである。

●XMLが拓く世界

XML文書がインターネット上で流通し処理される対象として、優れた特徴を備えていることがおわかりいただけただろう。XML言語によるデータ記述は、インターネット上の文書に限らず、PDAの内部や組み込み機器間のプロトコルにさえ使われるかもしれない（たとえば、計測器が解析装置に報告するときなど）。すべてのアプリケーション（いや、多くのアプリケーションでも十分だ）が統一されたデータフォーマットによりネットワークを介して対話するとき、われわれは新しい世界を手に入れるはずだ。

●参考文献

- 『XML入門』村田 真編著（日本経済新聞社）
- 『標準XML完全解説』SGML/XMLサロン著（技術評論社）
- 『XMLを知る』リチャード・ライト著（プレンティスホール出版）
- http://www.w3.org/TR/の各種仕様、特にhttp://www.w3.org/TR/REC-xmlは「XML version 1.0」の正式な仕様
- http://www.sil.org/sgml/xml.html
- http://www.xml.com/
- http://www.fxis.co.jp/DMS/sgml/index.html
- http://www.y-adagio.com/public/standards/xml/toc.htm

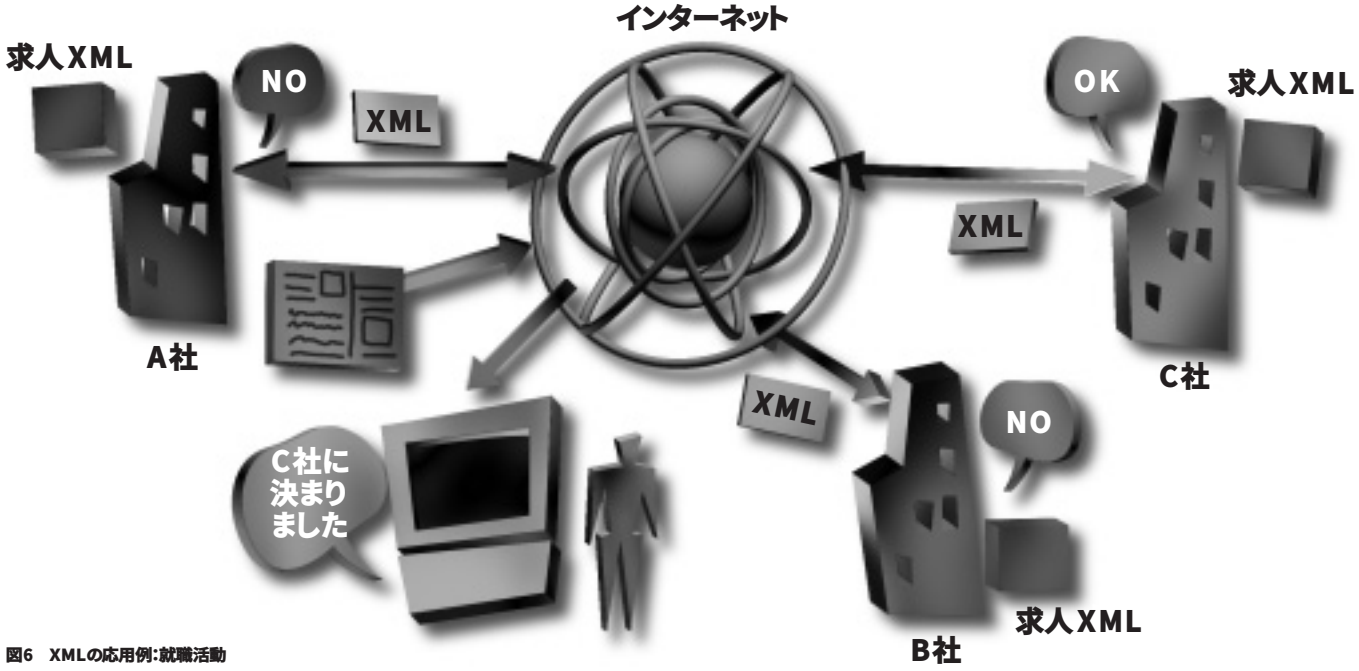


図6　XMLの応用例:就職活動

［インターネットマガジン　バックナンバーアーカイブ］ ご利用上の注意

このPDFファイルは、株式会社インプレスR&D（株式会社インプレスから分割）が1994年～2006年まで発行した月刊誌『インターネットマガジン』の誌面をPDF化し、「インターネットマガジン　バックナンバーアーカイブ」として以下のウェブサイト「All-in-One INTERNET magazine 2.0」で公開しているものです。

http://i.impressRD.jp/bn

このファイルをご利用いただくにあたり、下記の注意事項を必ずお読みください。

- 記載されている内容（技術解説、URL、団体・企業名、商品名、価格、プレゼント募集、アンケートなど）は発行当時のものです。
- 収録されている内容は著作権法上の保護を受けています。著作権はそれぞれの記事の著作者（執筆者、写真の撮影者、イラストの作成者、編集部など）が保持しています。
- 著作者から許諾が得られなかった著作物は収録されていない場合があります。
- このファイルやその内容を改変したり、商用を目的として再利用することはできません。あくまで個人や企業の非商用利用での閲覧、複製、送信に限られます。
- 収録されている内容を何らかの媒体に引用としてご利用する際は、出典として媒体名および月号、該当ページ番号、発行元（株式会社インプレスR&D）、コピーライトなどの情報をご明記ください。
- オリジナルの雑誌の発行時点では、株式会社インプレス R&D（当時は株式会社インプレス）と著作権者は内容が正確なものであるように最大限に努めましたが、すべての情報が完全に正確であることは保証できません。このファイルの内容に起因する直接的および間接的な損害に対して、一切の責任を負いません。お客様個人の責任においてご利用ください。

このファイルに関するお問い合わせ先
株式会社インプレスR&D
All-in-One INTERNET magazine 編集部
im-info@impress.co.jp