電子書籍ケーススタディ

イースト(株) 常務取締役 下川和男 shimokawa@est.co.jp

事例その11▶「Webサービスとは何か」

今，ソフトウエアの開発方法が大きく変わろうとしている。コンピュータそのものが，パーソナル・コンピュータという計算するデバイスから，ネットワークに常時接続されたインターネット・デバイスへと衣替えの真っ最中である。11月に炭疽菌騒動の中，ラスベガスで開催されたコムデックスというコンピュータ業界最大の展示会（http://www.est.co.jp/ks/tabi/0111cmdx/）では，マイクロソフト社のビル・ゲイツ会長が，来年発売予定の「タブレットPC」(http://www.microsoft.com/windowsxp/tabletpc/)という板型のインターネット・デバイスを9種類も紹介していた。

■パソコン・ソフトの開発方法

この二十数年間で，マイクロソフト系のソフトウエア開発手法は表1のような変遷をとげている。

OSの欄にインターネットと書くのはおかしいかもしれない。正しくはWindows 2000やWindows XPなのだが，このようなクライアント・パソコンやサーバ・マシンのオペレーティング・システムが何であるかを超越して，インターネットの世界が構築されているので，あえてインターネットとした。

マイクロソフトWindowsでも，最初のCとSDKの世界から，C++とMFCという環境が構築されて，やっと本当のWindowsプログラム開発が行えるようになったが，インターネット時代に入って数年が経過し，今年ようやく.NET Framework（ドットネット・フレームワーク）という真打ちが登場した。

■Webサービスとは何か

今年登場した，インターネット時代に即したソフトウエアの考え方を「Webサービス」と呼んでいる。これは，インターネットに接続されたサーバが，固有の関数（機能）を持ち，その関数を組み合わせて各種の業務を行うという考え方である。さまざまなハードウエアやOSが混在するインターネット環境の中で縦横無尽にコミュニケーションを行う，第三世代の技術である（表2参照）。

インターネットには30年近い歴史があるが，爆発的な普及が始まったのは，1993年，世界初のブラウザ「Mosaic」の登場であった。イリノイ大学の学生，マーク・アンドリーセンが開発したこのソフトウエアは，SGMLの簡易版であるHTMLというマークアップ言語を使用し，URLを叩けば世界中の情報をビジュアルに表示できる仕組みを開発した。その後，彼はこれを事業化し，ネットスケープ社を設立した。

ブラウザはあくまでも，人間がインターネットのサーバとコミュニケーションをとるための手段であるが，最近のRossetaNet（http://www.rosettanet.gr.jp/）に代表される電子商取引では，サーバとサーバが勝手にコミュニケーションし，商品の売り買いを自動で行っている。

このようなサーバ間やサーバとクライアントなど，コンピュータ間で会話を行うための言語が，話題のXMLであり，その会話の仕組みが，IBMが中心となって策定したSOAP（Simple Object Access Protocol，http://www.atmarkit.co.jp/fxml/rensai/soap01/soap01.html）であり，このような考え方を「Webサービス」(http://www.microsoft.com/japan/net/xmlservices.asp)と呼んでいる。

これから登場する多くのWebサービスについて，どのようなWebサービスを誰が運営し，どうしたら使えるのか？という，WebサービスのYahoo!のような仕組みであるUDDI（Universal Description,Discover and Integration）もスタートしている。

■Webサービスにおけるソフトウエア開発方法

OSの表とインターネット世代の表を照らし合わせると，.NET FrameworkはWebサービスを実現するライブラリなので，この部分が合致している。つまり，インターネット時代の本物の開発環境が，今年，やっと登場したことになる。

イーストは，「パーソナル・コンピュータとともに」という

キャッチフレーズのとおり，CP/Mの時代からパソコン向けのソフトウエア開発を行ってきた。十数年前に，MS-DOSからWindowsへのソフトウエア開発環境の変化を経験したが，今回のWebサービスは，それに匹敵する大きなソフトウエアの変革だと認識している。

言語好きの日本人は，C＃（シーシャープ）ばかりに目が向いているが，言語としてのJavaを発展させたC＃は，マニュアルを読めば，その制御構造などは容易に理解できる。Webサービスは，インターネット技術やXMLの上に構築されるものなので，これらを理解し，加えて.NET Frameworkという仕組みを理解する必要がある。

十数年前，ソフトウエア開発は「設計，製造，試験」の三工程で行われていたが，最近は「調査，設計，製造，試験」の四工程となっている。インターネット技術はまさしくドッグイヤーで進歩しており，ハードウエアは相変わらず「18ヵ月で倍」という急激な成長を維持している。18ヵ月で，CPUは倍の速さになり，メモリは半額になり，ハードディスクは容量が二倍になるのである。

調査工程が全体の七割を占める作業などもあるが，.NET Frameworkなどの新しい技術を早期に会得し，今後の技術潮流に即したシステムを開発すれば，将来の機能強化に柔軟に対応でき，他システムとの連動も可能になる。

具体的には，最近さまざまな雑誌に添付されているVisualStudio.NETという開発環境を使うが，SOAPなども容易に扱えるし，Mobile Internet Toolkitという便利な仕組みも入っているので，iモードなどの携帯電話に対応したWeb配信も行える。

XMLエバンジェリスト岡部恵造氏は，「XMLは規律・垂直・統制から自立・分散・協調への革命である」と自著で語られているが，このXMLを基盤としたWebサービスを使って，次のようなシステムを開発中である。

・電子辞書「取次」システム
・電子書籍「取次」システム
・書籍情報配信システム
・海外向け，書籍情報配信システム

Webサービスによって電子書籍や電子辞書がどのように進歩していくかについては，今後の連載で具体的にご紹介する予定である。

〈表1〉

OS	言語	ライブラリや実行環境
CP/M	M80（マイクロ･アセンブラ）	
	マイクロソフトBASIC	
MS-DOS	C，MASM	SDK（ソフトウエア開発キット）
Windows	C	Windows SDK
	C++（Visual Studio）	MFC（マイクロソフト基本クラスライブラリ）
インターネット	Java，Visual Basic	
	C＃，VB.NET	.NET Framework

〈表2〉

インターネット世代	サービス	マークアップ言語
第一世代（1980～）	電子メール，ftp	
第二世代（1993～）	ブラウザ	HTML
第三世代（2001～）	Webサービス	XML

電子書籍ケーススタディ

イースト(株) 常務取締役 下川和男 shimokawa@est.co.jp

事例その12▶「外字をどうするXKPとJepaX」

出版社や印刷会社の方々と一緒に，書籍や辞書のデジタル化を推進しているが，テキスト化の際，常に問題になるのが外字である。

世界中で外字が使われているのは日本だけ，という特殊事情の中で，インターネット出版やXMLでの外字の取り扱いや，公共システムでの人名外字処理についてご紹介する。

■なぜ外字なのか

外字問題のキーワードは二つある。「日本」と「インターネット」である。

パソコンやインターネットの仕組みは，そのほとんどを米国のマイクロソフト，オラクル，アドビそして，W3C，Unicode.orgなどの標準化団体が策定している。彼らの世界戦略の勝手な都合で，ユニコードという世界の文字のミニマムセットのようなコード系が策定され，JavaもXMLもWindows XPもWindows CEもユニコードをベースにしたシステムとなっている。

ユニコードには，日本，韓国，中国，台湾そしてベトナムの漢字が含まれているが，日本以外の国では，外字問題が深刻にはなっていない。

韓国はハングル全盛で，ハングルでしか自分の名前を書けない，日本でいえば「しもかわ　かずお」とひらがなでしか書けない中学生が出現するという事態に直面し，漢字への回帰がおこっているが，まだ外字問題にまでは至っていない。

台湾は，Big-5というコード系をACER，MITACなど，5社のコンピュータ・メーカで策定し，12000文字以上を揃えたので，外字の議論はそれほど発生していない。

漢字の故郷である中国は，国家が決めた標準を遵守する体制が確立しており，しかも，GBKという新しいコード系はユニコードの2万文字以上の漢字を含んでいるので，個々のユーザが外字を希望する状況ではない。

日本は，先祖代々の姓や，親がつけてくれた名前を尊重しており，戸籍のデジタル化でも，正しい表記を重視する国会決議がなされた。文学においても，様々な文字が使われており，JIS第一，第二水準の6879文字では足りない，との声が現代の作家からもあがっている。

このように，外字は日本固有の問題なので，米国の巨大コンピュータ関連企業は無関心である。

外字を，書籍に印刷するためには，外字フォントを作成すれば，どんな文字でも印刷が可能であった。しかし，書籍をテキスト化する際には，JIS文字以外には，何がしかの外字番号を入れる必要がある。

外字番号は，文字鏡研究会が策定した文字鏡番号が主流になりつつあるが，番号を決めても，実際のシステムで外字を表示しなければならない。

インターネット時代の今，外字のブラウザー画面での表示は，非常に厄介な問題をたくさん抱えている。

パソコンを中心に据えて，プリンターやモデムを周辺装置(デバイス)と呼んでいたが，最近はインターネットを中心として，パソコンを「インターネット・デバイス」と呼ぶメーカーが出現している。インターネット・デバイスには，Mac，Windows，LinuxなどのOSが入ったパソコンから，ザウルスやポケットPCなどの携帯端末，そしてiモードなどの携帯電話など，様々な機器が存在する。しかも，その標準文字セットや，文字のデザイン，文字の位置を揃えるベースラインなどが，OSやメーカー，機種ごとに異なっている。

混沌としたインターネット標準漢字環境の中で，外字を表示させる方法は，「そこだけ，画像にして送信する」のが一般的だが，表示されている書体もサイズもわからないブラウザー画面に，勝手なビットマップ・データを送りつけることになるので，一目で外字とわかる文字が表示されることになる。

人名外字のXKP

このように，問題だらけの外字の世界に，イーストは古

表　JepaXで表現可能な文字一覧（JepaX仕様書：渋谷 誠 氏 制作より）

形式	Unicode環境用				Windows環境用				日本語環境汎用			
エンコーディング	UTF-8,UTF-16				UTF-8,UTF-16, ShiftJIS				制限なし			
文字表現方法	文字コード	ISO文字名	UCS番号	gi要素	文字コード	ISO文字名	UCS番号	gi要素	文字コード	ISO文字名	UCS番号	gi要素
XML予約文字	※	◎	○	○	※	◎	○	○	※	◎	○	○
ASCII	◎	×	○	○	◎	×	○	○	◎	×	○	○
非ASCII欧文	◎	○	○	○	×	◎	◎	○	×	◎	◎	○
半角カタカナ	◎	×	○	○	◎	×	○	○	×	×	◎	○
JIS1,2水準	◎	×	○	○	◎	×	○	○	◎	×	○	○
13区記号	◎	×	○	○	◎	×	○	○	×	×	◎	○
NEC拡張漢字	◎	×	○	○	◎	×	○	○	×	×	◎	○
IBM拡張漢字	◎	×	○	○	◎	×	○	○	×	×	◎	○
その他のUnicode	◎	×	○	○	×	×	◎	○	×	×	◎	○
Unicode外の文字	×	×	×	◎	×	×	×	◎	×	×	×	◎

◎：推奨　○：可能　×：不可
※：「推奨」だが、XMLの仕様上「不可」となる場合がある
ISO文字名：ISOが決めた、文字の名称
UCS番号：ユニコードの番号

くから取り組んでいた。

1995年に，官庁や地方自治体，そして金融，証券などの人名外字を処理するために，Windows NT漢字処理技術協議会（http://www.xkp.or.jp）という団体を，マイクロソフト社やコンピュータ・メーカーと共に設立した。当時は，メインフレームやオフコンの時代から，クライアント・サーバ型のコンピュータへの転換期で，Windows NTの販売を後方から支援するために組織された団体である。

Windows NTはユニコード・ベースのOSなので，2万文字以上の漢字が扱え，外字領域も，DOSやWindowsの1880文字から，6400文字に拡張されている。そのユニコードでも不足する文字を，クライアント・サーバ型の企業内システムに実装する仕組みや，TrueType形式の外字フォントデータの開発と販売を担当した。

JepaXでの外字表現

その次に，外字関連で取り組んだのが，文庫や新書などの交換用フォーマットJepaX（http://x.jepa.or.jp/jepax）である。

JepaXは，1998年に日本電子出版協会の出版データフォーマット標準化研究委員会で討議された，「出版社が，今後のインターネット時代に自社のコンテンツをどのような形式で蓄積すべきか」という命題に対して，「そりゃーXMLでしょう」ということで策定したスキーマである。

当時，電子出版界を賑わせていた電子書籍コンソーシアムが，外字問題やコミック対応で，画像化の方針を打ち出したので，その対抗として，テキストでも外字を自由に扱える仕様を目指した。外字といっても，表の通り，ベースとなるコード系によって対象範囲が異なる。

gi要素というのが，外字部分で，Glyph Imageの略である。具体的には，外字を

```
森　<gi set="mojikyo" name="58562" alt="鴎" />外
<gi set="mojikyo" name="39630" alt="&#x9127;" />小平
深<gi set="mojikyo" name="04894" alt="（土川）" />経済特区
<gi set="mojikyo" name="50021" alt="ボク（さんずいに墨）"
/>東奇<gi set="mojikyo" name="35978" alt="譚" />
```

と表記する。set=は外字番号セットの名称，name=はセット内での番号である。

この方式は官報XML化プロジェクトでも採用され，一般化しつつある。また，この方式をベースとして，ニュースの標準化であるNewsML（http://www.pressnet.or.jp/newsml/newsml.htm）や住所，氏名の標準化であるContact XML（http://www.contactxml.org/method.html），辞書データの標準化DicX（http://www.dicx.org）など，様々なXMLボキャブラリへの適用も，検討を開始した。

電子書籍ケーススタディ

イースト（株）常務取締役 下川和男 shimokawa@est.co.jp

事例その13▶「外字をどうするか？　＝今昔文字鏡＝」

先月号で，電子書籍やインターネット出版でどのように外字を処理するかを，XKPとJepaXでご説明したが，外字についての仕組みや理論をいくら振り回しても，実際にその文字が画面やプリンターに表示できなければ意味がない。今回は，10万もの文字コレクションを誇る，今昔文字鏡をご紹介する。

■文字鏡とは何か

文字鏡は，一種のコンピュータ漢字普及運動のようなプロジェクトである。推進母体は株式会社エーアイ・ネットで，調査研究的な作業は文字鏡研究会という非営利団体が担当し，販売は紀伊國屋書店が担当している。

推進者であるエーアイ・ネット社の古家社長のお話では，16年ほど前に，仏典の複雑な文字をPC-9800の画面に表示したのが起源で，その後，JISにない文字の番号付けとフォントの制作を延々と続けられている。

文字鏡は，2001年9月現在，以下の文字をサポートしている。

文字の種類	文字数
漢字	101,936字
非漢字	2,382字
梵字	1,875字
甲骨文字	3,398字
西夏文字	6,000字
合計	115,591字

文字鏡プロジェクトには，次のような製品やサービスがある。

●今昔文字鏡CD-ROM

Windows上で稼動する検索ソフトで，価格は28000円。簡単に言うと，10万文字の文字コード辞典のようなソフトで，漢和辞典風の「読み」や「部首＋画数」のほか，「部品（文字の一部分のかたち）」，「英単語」，「ピンイン」，「韓音」，「ISOコード」，「大漢和コード」から，漢字を探し出すことができる。

漢和辞典といったが，その文字の解説が載っているわけではなく，その文字のJISコード，ユニコード，文字鏡番号を知ることができる。

●文字鏡WEB

今昔文字鏡CD-ROMのWEB版である。インターネットで文字鏡の文字の検索が行える。アクセス数で課金される方式で，2000回で12000円となっている。

●文字鏡フォント・サーバ

文字鏡CD-ROMや文字鏡WEBで調べた文字鏡番号を，実際にブラウザ画面に表示させるための仕組みである。インターネットへの常時接続が前提となるが，12，16，24，48，96ドットのビットマップ・フォントを高速に配信するもので，年間200万文字の配信が，100万円となっている。

文字鏡TrueTypeフォントのように，個々のパソコンのハードディスクを占有することもなく，常に最新のフォントが世界中で受信でき，かつ，外字が必要な場合のみ，リアルタイムに，このサーバが呼び出される。

さまざまな文字をブラウザ上に表示するという，至って単純で基礎的な仕組みなので，本来は国家的な機関が管理・運営すべきサーバだが，このような仕組みを国が理解するのは少し先になるので，当面は営利事業とせざるを得ない。

●筆文字鏡　楷書体

文字鏡番号に準拠した7万文字の毛筆楷書体フォントセットで，Windowsのシフト JIS，1880文字の外字領域に選択した楷書フォントを登録するツールが付いている。

●悠々漢字術2001（ISBN：4-314-10142-3）

文字鏡プロジェクトの紹介本で，付録のCD-ROMには，9万文字をシフトJISの漢字コード領域にマッピングした文字フォントが入っている。このフォントは，書体を切り替えることで，WindowsやMacintoshで画面表示や印刷が可能である。

イーストは，文字鏡WEBとフォント・サーバの開発と運

営を担当させていただいた。開発は，2000年の春から秋にかけて，半年ほどで行った。

決済システムは，紀伊國屋さんが持たれているオフラインでの仕組みを使うので，開発していない。

また，「読み」，「部首＋画数」，「部品」などの検索データベースや，文字鏡番号とユニコード，JISコードなどの変換テーブルは，エーアイ・ネット社のサーバをリアルタイムに呼び出すという，分散処理を行っている。もちろん，このインタフェースにはXMLを使用している。

決済と検索を他に依存しているシステムであるが，会員登録やアクセス数管理などの管理者画面やユーザ画面を，Javaスクリプトを使ったアクティブ・サーバ・ページというWindows 2000サーバの利用環境で動かしている。

http://www.mojikyo.com/cat/web/trial.htmで申し込めば，無料トライアルができるので，どんなソフトなのか，体験していただきたい。

文字鏡WEBの検索結果画面

■文字鏡で何ができるか

文字鏡の仕組みで嬉しいのは，文字鏡番号を調べるのは有料だが，その後の利用は無料となるところである。書籍「悠々漢字術」に添付されているTrueTypeフォントは，インターネットから誰でも無料でダウンロードすることができる。

http://www.mojikyo.org/html/download/にアクセスし，使用許諾条件文を理解した上で，圧縮形式で33ファイル，55メガバイトをダウンロードし，自分のパソコンにTrueTypeフォントとしてインストールすることにより，画面表示や印刷が可能となる。

この膨大な漢字フォントの無料配布は，欧米でも歓迎されており，スタンフォード大学仏教研究センターが，米国でのダウンロードをボランティアで担当している。

文字鏡プロジェクトのもう一つの大きなメリットは，「文字鏡にない文字には，文字鏡番号を新たに振り，そのフォントも提供される」という仕組みにある。

文字鏡研究会に出典を示して申請すれば，番号が付与される。漢字クイズ風の創作漢字は受け付けてもらえないが，「人名」，「地名」などの証票があれば，問題ない。

今までに，戦後すべての「国会議事録」や，戦後すべての「官報」などのデジタル化が完了した。イージャパン構想や電子政府を実現するために，今後もさまざまなドキュメントのデジタル化が行われるが，その際，この申請制度を使えば，「すべての文字のコード化」が可能となる。

「XMLによる画像参照交換方式」（JIS TR X 0047，http://www.y-adagio.com/public/standards/tr_lsi_xml/lsi_xml.htm）という，XMLドキュメント内での外字画像の表記方法もJIS化されており，着々と，日本固有の外字問題も解決の方向に向かっている。

電子書籍ケーススタディ

イースト(株) 常務取締役 下川和男 shimokawa@est.co.jp

事例その14▶「世界中のパソコンで日本語を ＝JiBOOKS＝」

今回は，昨年12月号でご紹介した「Webサービスとは何か」と2月号の「外字をどうするか？ ＝今昔文字鏡＝」を組み合わせたJiBOOKSをご紹介する。JiBOOKSは国立国語研究所さんからの依頼で開発を担当した，海外向け日本語情報配信サーバシステムである。

■JiBOOKSとは

JiBOOKSは，国立国語研究所の横山詔一先生が企画，推進されているシステムで，海外で日本語を勉強する人を対象として，日本語関連の情報をインターネットで提供するプロジェクトの一環として開発された。

名称にBOOKSとある通り，日本でどのような書籍が出版されているかを知るためのサイトである。書誌情報の検索には，社団法人日本書籍出版協会のご協力で，本のサーチエンジン「Books」を利用している。Booksには，いま日本で購入が可能な約60万点の書籍情報が入っており，しかも月次更新されているので，話題の新刊書なども入っている。

日本の書籍の検索なら，世界中のインターネットに接続されたパソコンから，BookWEBでもAmazonでもアクセスできる。しかも，インターネット・エクスプローラの場合，たとえば北京大学を見ようと思って，http://www.pku.edu.cn/と入力すると，簡体字のフォントをインストールしますか？というメッセージが表示される。高速回線なら数十秒で中文フォントがダウンロードされ，ブラウザー画面に中国語が現れる。各国語のサイト閲覧には，まことに便利な機能である。

私は頻繁に海外の展示会に行くが，プレスセンターにはインターネットに繋がったパソコンがずらりと並んでおり，自由に使うことができる。そこで，日本のニュースを知りたいと思い，http://www.asahi.comにアクセスすると，「日本語フォントをダウンロードしますか？」と英語でメッセージが表示され，ダウンロードできる。しかし，これは幸運なケースで，「ダウンロードできません」とのメッセージで，なにやら訳のわからないアルファベットが並んだ文字化けしたAsahi.comのページが表示される場合も多い。

前者と後者の違いは，ネットワーク管理者が真剣に仕事をしているか，その場限りの展示会なので，とりあえずの仕事をしているかの違いである。公共の場にあるパソコンに勝手に数メガバイトのデータをダウンロードするのは，本来は禁止すべきだし，特に外国語のフォントをインストールすると文字化けの原因となる。

また，個人のパソコンでもダイヤルアップで，数メガバイトのファイルをダウンロードするには，多くの時間とプロバイダー料金が必要となる。「欧米はブロードバンドでしょ」と言われそうだが，韓国以外のアジアの国々は，ADSLの整備は遅れているし，光ファイバーなど論外である。

JiBOOKSは，そのような海外の図書館，学校などの公共の場所や，個人の貧弱な回線や古いパソコンでも日本語の入力や表示ができるように工夫されたシステムで，入力はローマ字，表示はビットマップ・フォントを使い，海外のパソコンでの日本語書籍検索を実現している。

百聞は一見にしかず，以下の手順で実際に使ってみていただきたい。

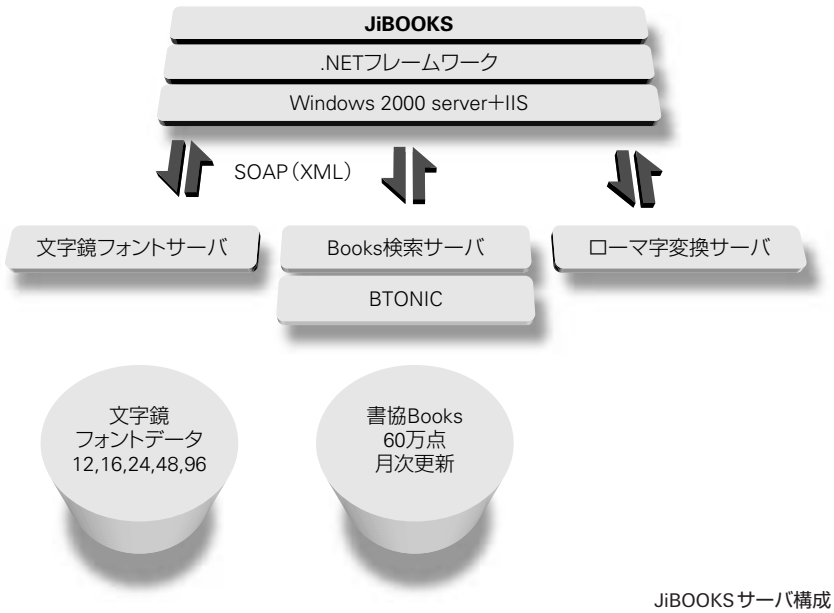
1. http://www.kokken.go.jp/jibooks へ
2. click hereで，検索画面を表示させる
3. Title:やAuthor:欄に，「ローマ字」で検索したい書名や著者名を入れる
4. Convert into Kanaボタンを押すと，ローマ字がひらがなで表示される
5. Searchボタンを押す
6. 該当する日本の書籍が5点ずつ大きな漢字で表示される

日本語が表示できるブラウザーで操作しても何の驚きもないが，アジアの片隅のインターネット・カフェで，なみなみと注がれた熱いチャイでも飲みながら，386パソコンの遅い回線からポッポッポッと出てくる大きな漢字を見れば，これは感動ものである。

■Webサービスを利用した開発

このシステムは,「ローマ字変換」,「Books検索」,「フォントサーバ」という三種類のWebサービスの上に構築されている。

ローマ字変換は, 4. でローマ字をひらがなに変換する部分だが, こんな簡単な処理をなぜWebサービスにしたかというと, 将来の拡張を考慮したためである。かな漢字変換の辞書を搭載し, 本格的な日本語入力Webサービスを, いつの日か実現させたいと思っている。



JiBOOKSサーバ構成

Books検索は, このシステムの核の部分であるが, ここは最初から別サーバとして設計され, JiBOOKS以外にも, いくつかの案件で使用することになっている。検索部分には, 昨年3月号の「三省堂 e辞林」でご紹介した, XMLドキュメントの全文検索エンジン「BTONIC」を使っている。

フォントサーバは, 先月号でご紹介した文字鏡フォントサーバに, Webサービスのインタフェースを追加した。希望する文字のユニコードやシフトJISコードと文字サイズをXMLでこのサーバに問い合わせると, 該当する文字のビットマップ・データがもらえる, という仕組みである。

VisualStudio.NETという開発環境とC#(シーシャープ)というJavaを拡張した言語を使い, 三ヵ月ほどで開発を完了した。とは言っても, Books検索とフォントサーバは一年以上前から開発を行っていたので, 全体システムの組み込みと稼動確認試験, そしてブラシュアップが主な作業である。

Webサービスの場合, サーバごとに独立した機能を開発し, それらのサーバを繋ぐ形となるので, プログラムの独立性が従来のサブルーチンやサブシステムよりも高くなり, 個々のサーバ試験も行いやすい。

JiBOOKSは今後, さまざまな分野への応用が検討されているが, Books検索部分を他の検索サイトや情報サイトとのWebサービスに変更するだけで, ひらがなの入力と漢字の表示が可能となる。しかも文字鏡の10万文字が使えるので, 中国語や外字にめっぽう強い。

今後, PDAや携帯電話が進化したインターネット・デバイスが多数登場するが, それらがインターネット・エクスプローラのような各国語フォントのダウンロード機能を持っていることは稀なので, JiBOOKSの需要は日増しに増加する。今, 私の心配は, 一般のテキストに比べて数十倍の容量となるビットマップ・フォントの配信に, イーストが所有する実質20メガの回線がいつまで耐えられるかである。

日本語を勉強する多くの人に使ってもらいたいという気持と, 回線パンクの不安が交錯して, 複雑な心境である。

電子書籍ケーススタディ

イースト（株）常務取締役 下川和男 shimokawa@est.co.jp

事例その15▶「辞書サーバをXMLでつなぐ = JapanKnowledge.com =」

先月号では，Webサービスと外字を組み合わせた海外向け日本語配信システムJiBOOKSをご紹介したが，今回は，昨年12月号の「Webサービスとは何か」と，昨年3月号の「三省堂　e辞林」と同じXML辞書検索エンジンBTONICを組み合わせた，JapanKnowledge.com向けのSOAPとXMLを使用した辞書データ配信システムをご紹介する。

■JapanKnowledge.comとは

「日本の知識ドットコム」という立派なドメイン名を持つこのサイトは，小学館さんが50%を出資した，ネットアドバンス社が運営する辞書検索サイトである。

「ナレッジピープルのための知識発見サイト」というキャッチフレーズで，ビジネスマンの机をイメージしたサイトで，知識のナビゲーションを行ってくれる。ここでは，以下の辞書が引ける。

日本大百科全書（ニッポニカ）	新語探検2002
データパル1991～2001	Internet New Words
Multimedia Internet事典	Encyclopedia of Japan
JKN Who's Who	NNA：アジア経済情報
現代用語の基礎知識	新・日本国の研究
日経BPデジタル大事典	ワールド・クロニクル
	IT書斎術百科
大辞泉	
プログレッシブ英和中辞典	56万冊の書誌データ bk1
プログレッシブ和英中辞典	ニッポニカURLセレクト

百科事典として一世を風靡したニッポニカは13万項目，画像も豊富で，「君が代」をパソコンで聞くこともできる。ニッポニカ以下の6点が事典（ことてん）で，事柄を調べるものである。大辞泉は，22万項目の百科＋国語辞典で，簡単明瞭な解説が表示される。大辞泉以下の6点が辞典である。その中には，Encyclopedia of Japanという，日本を英文で紹介する辞書も入っている。既知の事柄を英文で読めるので，英語の勉強に最適である。

NNA：アジア経済情報以下の4点が，ニュースや論説などで，現代用語の基礎知識や新語探検も含めて，「今」を意識した品揃えとなっている。

bk1（http://www.bk1.co.jp）はオンライン書店で，関連した書籍を検索し，その場で購入することも可能である。最後のURL集も重要で，関連したWebサイトのアドレスを表示してくれる。Google（http://www.google.com）などのサーチエンジンで検索しても，どれが適切なサイトなのかは，自分で一つずつ開いて確かめなければならないが，このURL集は編集者の目で確認されたものなので，適切なホームページを案内してくれる。

しかも，ワンルックという機能があり，検索語に該当する見出し項目が一覧表示されるので壮観である。ここまでは，一般の方々も操作できるので，ぜひ試していただきたい。

画面は，「インターネット」ということばを検索した場合のワンルック画面である。4903項目がヒットし，その中から日経BPデジタル大事典の「インターネット」をクリックすると，画面左の解説が別ウィンドウで表示される。

価格は毎月1500円，百科事典ニッポニカだけをとっても非常に安い金額である。法人向けには，社内のLAN環境で使い放題となるIP固定方式での販売も行われている。

■Webサービスを利用した辞書の配信

JapanKnowledge.comのサーバは，ネットアドバンス社に出資している富士通さんが開発したもので，UNIXを使い，C言語で作られている。大半の辞書データは富士通のサーバに入っており，その検索ロジックを使っているが，bk1はブックワン社の書籍データベース・サーバをhttpで呼び出して使っている。

事典の最後の二つ，「現代用語の基礎知識」と「デジタル大事典」は，辞書データ自体が，代々木のイースト本社に設置されたサーバから配信されている。JapanKnowledgeサイトのことだけを考えれば，富士通のサーバにこの二つの辞書データも同居させた方が，検索ロジ

「インターネット」という言葉で検索したときの画面

ックも統一できるし，扱いやすい。

それなのに別のサーバを使っている理由は，「コンテンツの一元管理」のためである。データを提供した場合，辞書を更新するたびに，データの再提供が必要になる。コンテンツの提供元である自由国民社としては，複数の辞書引きサイトへの提供や自社サイトでの辞書検索サービス，iモード対応など，まさしくワンソース・マルチユースを計画されており，一ヵ所のデータを更新すれば，すべての辞書引きサービスが最新データになる仕組みを希望された。

これを実現するためにイーストでは，XMLを使ったWebサービス方式で辞書検索サイトを構築した。

作業手順は以下のようなものである。

1.DicX仕様を使った辞書データのXML化

2.XMLドキュメント全文検索エンジンBTONICでの稼動

3.BTONICの上位にマイクロソフト社の.NETフレームワークを組み込み，Webサービスの実現

開発は，昨年の8月から12月まで，5ヵ月間で行った。1と2はほぼ完成していたので，3が主な作業であった。以下の三種類のメソッドと呼ばれる，インターネット上のサーバ呼び出し関数の実装である。

GetDicList	使用可能な辞書一覧の取得
SearchDicItem	辞書項目の検索（取得）
GetDicItem	辞書項目の取得

GetDicListで，「あなたに対してサービスするのは，現代用語とデジタル大事典ですよ」という応答を返す。

SearchDicItemがいちばん重要なメソッドで，検索方法（前方一致，後方一致，完全一致），検索対象（見出し語，本文，キーワード），そして検索語などをパラメータとしてもらい，検索結果の項目一覧を返す。

次に，GetDicItemで指定された項目の本文を返す，というものである。

この一連のサーバ間通信には，SOAP（Simple Object Access Protocol）というIBMやマイクロソフトが推進している仕組みを使い，実際に応答するデータはXML形式となっている。

UnixとWindowsという異なるアーキテクチャのサーバをSOAPで結合し実用で使用した，日本で最初の事例だと思う。

画面の通り，テキストだけではなく画像の配信も行えるし，セキュリティ関連の機能も入っている。以下の辞書サービス（V05）で，具体的なインタフェースを公開しているので，参考にしていただきたい。

【参考URL】

BTONIC

http://www.est.co.jp/btonic

DicX仕様サイト

http://www.dicx.org/

辞書Webサービス（V05）

http://btonic.est.co.jp/NetDic/NetDicv05.asmx

辞書Webサービス（V05）のWSDL

http://btonic.est.co.jp/NetDic/NetDicv05.asmx?WSDL

サービス動作検証用検索サイト

http://btonic.est.co.jp/NetDicTest/TestV05.aspx

電子書籍ケーススタディ

イースト（株）常務取締役 下川和男 shimokawa@est.co.jp

事例その16▶「書籍検索サーバ＝Books.or.jp＝（上）」

4月16日，東京国際ブックフェアの前々日，社団法人日本書籍出版協会（http://www.jbpa.or.jp，以下書協）の新宿区袋町にある立派な会館の4階会議室で，「本のサーチエンジンBooks（http://www.books.or.jp）」のリニューアル発表会が開催された。NHKのテレビカメラも入り，記者団30名，出版関係者100名以上が参加し，書協のデータベース委員会の佐藤委員長（新潮社社長），凸版印刷 Eビジネス本部の秋山取締役，マイクロソフト社の安藤部長そして私が新Booksの概要を説明した。

■Books開発の経緯

旧Booksは，5年前の平成9年9月9日，午前9時9分に一般公開した。

平成9年の初め，書協の前田副理事長（三修社社長）の発案で，「日本書籍総目録」のWeb版を制作することになった。凸版印刷さんが管理している印刷用のデータを入手し，SQLサーバに入れて，アクティブ・サーバ・ページという仕組みで，ソフトウェアを開発した。試作はたったの三日で行った。

旧Booksの開発には半年ほどかかった。取りあえず書籍の検索が行えるだけのシステムは簡単に開発できるが，月次でのデータ更新や毎月10万回の想定検索数に耐えるシステム作りで苦労した。

盛大な発表会が行われ，新聞やテレビでも報道されたため，アクセス数はグングン上昇し，ピーク時には60万検索に達した。Booksにそれほどのアクセスが集中したのは，「書協が運営しているので中立的である」，「今，販売されている書籍のみが入っている」，「シンプルな操作で，検索が容易」などの理由によるものだが，想定ユーザ数の6倍ものアクセスで，サーバシステムはパンク状態になり，応答に数分もかかるケースも発生した。

また，新刊の登録は月次で行うため，当時大ヒットしていた渡辺淳一の「失楽園」が見つからず，ミルトンの「失楽園」しか検索されない時期があり，多くのユーザからお叱りをうけた。

ホームページのアクセス数は，以下のように，いくつかの数え方がある。

トップページ·ビュー	トップページが表示された回数
ページ·ビュー	各ページが表示された合計数
ファイル·ビュー	ファイルがサーバから送信された回数
ユニーク·ユーザ数	そのサイトを訪れた人の数
検索数	検索が行われた回数

Booksは検索サイトなので検索数をカウントしているが，トップページ・ビューやページ・ビューが一般的である。Booksはトップページで検索を行うので，検索数とトップページ・ビューはほぼ同じ値となる。トップページを眺めるだけで，検索を行わない人は検索数にカウントされないが，それは稀である。

Booksは，検索（トップ）画面⇒検索結果一覧画面⇒詳細画面と遷移するので，一回の検索で3ページが表示される。60万検索は180万ページ・ビューとなる。

ファイル・ビューは，送信されたファイルの数で，これがもっとも多いカウント数である。講談社のWeb現代（http://kodansha.cplaza.ne.jp/）は週刊誌の電車広告風の画面なので，トップページを表示するだけで，50ファイル・ビューくらいになる。

ユニーク・ユーザ数は，アクセス・カウンタなどで使用する方法で，IPアドレスをチェックして，同じ人が何回そのページを見ても一回しかカウントしない，もっとも少ないカウント方法である。

旧Booksは平成9年9月の公開以降，数回の改良を行った。最初はBooksLinkである。これは，「Booksで探していた本を見つけたが，もう少し詳しい内容を知りたい」，「どこで買えるのか」というご要望をたくさんいただいて発案したものである。書籍にはISBNというユニークな番号が付いているので，これをキーにして，出版社のサイトにBooksからパイパー・リンクを行う。該当する書籍のページをダイレクトに表示して欲しいとの要求仕様を提示し，200以上の出版社サイトで対応していただいた。

Booksの詳細画面で，書名にアンダーラインが引かれている書籍をクリックすると，BooksLinkで，出版社のサイトにリンクされ，購入ボタンや目次，概要などが表示される。

次はアクセスログ管理で，毎月60万件の検索語や出版社名，著者名など，ユーザが入力した情報を分類，整理してグラフ表示する仕組みを追加した。

2年後には，サーバ・ハードウェアの増強を行った。当時最強のPentium Proの4CPU構成とし，640メガバイトのメモリを搭載して，検索速度を数倍向上させた。

■新Booksの登場

2001年，新Booksの開発プランがまとまり，一年がかりで開発を行った。コンピュータの速度向上は，ハードウェアをいくら高価な最高速マシンにしても数倍しか向上しないが，ソフトウェアのロジックを改良すれば，数百倍，数千倍も向上することがある。今回の最大のテーマは，検索速度の向上で，平均2分半（150秒）かかっていたものを，0.5秒つまり300倍の高速化を目指した。

SQL系のデータベースで部分一致検索を行うと，応答が極端に遅くなる，という問題点は判っていたので，SQL DBは使わず，三省堂.NETなどで使用している，イースト・オリジナルのXMLドキュメント全文検索エンジンBTONIC（http://www.btonic.com）を使用した。辞書のようなドキュメント系のXMLデータではなく，書名，出版社名，著者名などの項目に分かれたデータベース系のデータにBTONICを適用する最初の事例であったが，多少の機能追加で，高速検索を実現することができた。

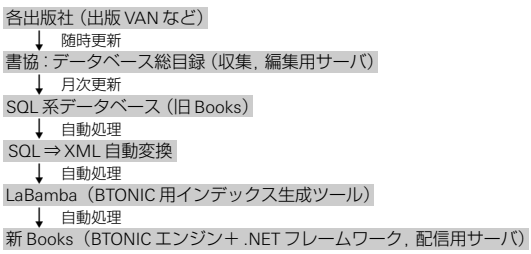
SQLでの部分一致は，実際にデータベースのインデックス部分をサーチするので，それをメモリ上に置き高速化を行った。しかし，BTONICは全文検索用のインデックスがあらかじめ生成されているので，数回のデータアクセスで，検索が完了する。CPU負荷だけを考えれば，数千倍の高速化が行える。

新Booksは，4，50万円程度の薄型サーバ二台で運用しているが，二台構成はハードウェア障害対策が目的なので，検索能力だけであれば一台で充分処理できる。理論値だが，普及型のサーバで，一時間に1万検索，月間600万検索くらいは可能である。

Amazon.comは一時間で2200万ページ・ビューとのことで，誰もがAmazonを目指してインターネット・ビジネスに参入するが，最初からサーバに数億円投資できるわけではないので，BTONICを使った安上がりの高速検索は，引合いが増えている。

BTONICは全文検索やXMLタグ（論理構造）のインデックスを事前に生成して高速検索を実現しているので，データがドンドン更新されるシステムでの検索には不向きである。

Booksでは，毎月以下のようなデータ更新作業を行っている。



書協の収集用サーバには，出版VANなどから書誌データが随時登録されるが，それを一ヶ月分まとめてSQL系の旧Booksに登録している。旧Booksでは，これで更新完了であったが，ここからBTONICまで，「SQL DBからXMLへの変換」，「XMLデータのインデックス生成」そして「新Booksへの登録」までを自動的に行っている。処理時間は5時間ほどなので，毎晩処理することにより，月次ではなく，日次更新のアプリケーションであっても，BTONIC方式での対応が可能である。

満を持して開発した新Booksは，携帯電話対応，PDA対応，オンライン書店アフィリエート機能，Webサービスなどにより，支持を増やしつつある。そのあたりについては，次号でご紹介する。