新連載

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

事例その1▶「インターネットを使った教育システム NetLearning」

いま,「紙の本が消滅する」,「これまでたまたま紙だった」などという議論が巷を騒がせているが，紙面ではなく画面で読む「電子書籍」について，毎月一回，私がかかわったプロジェクトの事例をもとに解説していきたい。

第一回は，教科書，参考書，実用書，新書などの紙面が画面に変わり，先生までサイバースペースに存在する WBT (ウェッブ・ベース・トレーニング)のベンチャー企業，ネットラーニング社を紹介する。

■ WBT は第三世代の電子書籍

WBT を電子書籍というジャンルに含めることについては異論があるかも知れないが，私は，最先端の電子書籍だと考えている。電子書籍や電子出版をその提供形態で分類すると，以下の三世代となる。

第一世代 .. パッケージ .....広辞苑 CD-ROM, SONY 電子ブックなど
第二世代 .. ダウンロード .....PDF, Microsoft Reader など
第三世代 .. インタラクティブ.. HTML 出版, WBT など

つまり，第一世代は「メディア (媒体)」を提供し，第二世代は「ファイル」を提供し，第三世代は「サービス」を提供している。ちなみに，第〇世代が「紙」である。

この連載の中で，いまどきの電子書籍としてもてはやされている PDF や Microsoft Reader についても紹介する予定である。しかし，これは個人がハード・ディスクに情報を所有する方式で，インターネットを使った出版の最終形態とは思えない。第三世代こそが，だれでも，いつでも，インターネットが自由に使える，数年後のインターネット環境を想定した，真に「インターネット出版」と呼べるものである。

また，第三世代は，インタラクティブなだけではなく，コンテンツの一元管理という大きなメリットがある。一ヵ所のサーバにコンテンツが蓄積され，それを世界中から見る方式なので，コンテンツの更新が容易で，常に最新の情報を提供することが可能となる。

■ XML を使ったコンテンツ制作と配信

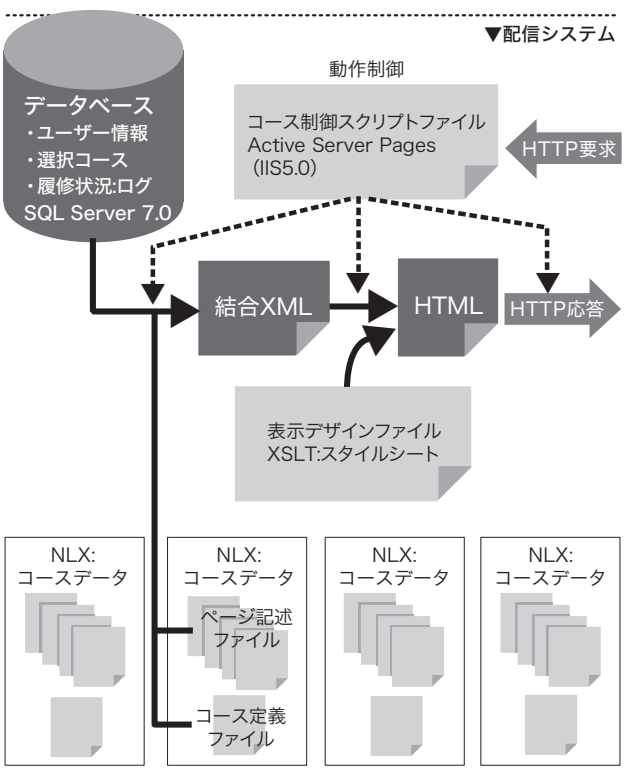
弊社が開発と運営を担当させていただいた，第三世代に属するネットラーニング社の配信システムは，図のような構成となっている。

下に並んでいる，「NLX:コースデータ」が，XML で記述された各種のコース (教科書) である。このプロジェクトは 2 年前にスタートしたが，その当時，データ構造の定義に XML を採用するのは冒険であった。しかし，その後，HTML が XHTML として XML の一部に位置づけられ，またサーバ間インターフェイスの共通言語としての地位を確立した現在，NLX は XML の成功例と自負している。

NLX (ネットラーニング XML) の仕様は非公開だが，ブラウザへの画面表示以外に，「問題の提示と採点」,「採点結果による解説文へのジャンプ」,「目次，索引からのジャンプ」などのインタラクティブな機能を持っている。

XML を採用した利点は，この他にも「コース開発が容易」,「スタイルの変更が容易」などがある。

コース開発は，新規に作成する場合と，既存の紙の教科



書から変換する場合がある。新規に作成する場合は，簡易タグを使って基本部分を作成し，簡易タグをXMLに自動変換した後，詳細な指定を行っている。XMLのエディタやツールも揃ってきたので，直接，XMLで記述する著者も増えつつある。

紙の教科書を変換する場合は，QuarkXPress, PageMaker, Wordなど使用しているDTPソフトにより，変換方法が異なる。弊社が担当する場合は，avenue.quarkによるXML変換や，HTML出力を行い，その後，各教科書の特性に合わせた変換ツールを自作して，なるべく個別の手作業を排除したコース制作を行っている。

といっても，「静かな」紙から，「ダイナミックな」画面を作り出すことは困難で，「カラー画像」，「音声ガイダンス」，「問題とその解答」，「シミュレーション画面」，「アニメーション画面」などは，新規に制作することになる。

■ XSLの威力

スタイルの変更は，ネットラーニング社のビジネスに大きく貢献している。ホームページ（http://www.netlearning.co.jp）を見ていただくと，Java，C++，SQLサーバ，ネットワークなどIT系のコースが並んでいるが，これはネットラーニング社のビジネスの一部である。

この「カタログ・コース」以外に，これらのコースを再販売する「OEMコース」，企業内研修などのコースを受託制作し，その企業限定で配信サービスを担当する「カスタム・コース」，配信システム自体をASP（アプリケーション・サービス・プロバイダー）として提供する「プラットフォーム販売」という四種類のビジネスモデルが事業の柱となっている。

既存コースのロゴや体裁をOEM先に合わせて変更したり，カスタム・コース用に全体の画面デザインを改訂する際に，図の右側にある「XSLT：スタイルシート」を使って，極論すればファイルを一つ修正するだけで，すべてのコースの体裁を変えられるのである。

図の右にある，XMLからHTMLへの変換は，XML未対応の4.0系ブラウザに対応するためで，この部分を変更すれば，i-ModeやPalmなどのノンPCインターネット・デバイスへの対応も可能となる。これらの処理を，ASP（アクティブ・サーバ・ページ）という言語を使ってプログラミングしている。

また，ダイナミックな企業では，社内教育資料が半年でゴミとなり，ムダな印刷を繰り返してきた。WBT方式でカスタム・コースを制作する場合，コンテンツが一元管理されているので，最新の情報を即座に全社員に提供することも可能となる。

■ WBTから電子教科書へ

図の左にあるデータベースには，受講者が「いつ，どの頁を，何分間見たか」，「どの問題を，何分間で，何問正解したか」などの情報が記憶され，これが，受講者やチューター（先生），そして企業の教育担当者に情報として提供される。

チューターには，いつでも，何でも，電子メールで質問でき，また，記述形式の問題については，チューターから細かな助言が行われる。実際のコンピュータ画面をシミュレーションしたり，音声でガイダンスが付いたり，インターネットを使ってその人に合った指導が行える。紙の本が提供していた「感動」，「情報」，「知識」などのうち，知識を提供する方法は，「電子教科書」などの議論とともにWBT化していくであろう。

「感動」を提供する小説もインタラクティブなものが主流になると思っている。XMLエディタを駆使した小説家が登場し，ハイパーリンクだけではなくBGMや絵が入り，ストーリーが分岐する小説である。話題の田中康夫さんのデビュー作は，ブランド名がたくさん登場し，その注釈の多さで話題になったが，今だったら，エルメスやボートハウスのサイトにハイパーリンクする『何となくクリスタル』を書いたに違いない。

次回は，いま話題の「120万語辞書検索e辞林」（http://www.sanseido.net）を紹介する。

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2▶「120万語辞書検索　三省堂e辞林」

電子書籍ケーススタディの第1回は，創業3年目のベンチャー企業，ネットラーニング社（http://www.netlearning.co.jp）のWBT（ウェッブ・ベース・トレーニング）システムを紹介したが，今回は，創業120年の三省堂の辞書検索システムを紹介する。

イーストは，ベンチャー企業のインターネット・サービス部分を一括受注して，サーバ・システムの企画から設計，開発，そして運用までを担当することが多いが，最近，「e辞林」のように，長い歴史を持つ会社の企業内ベンチャーの技術部門を丸ごと担当する仕事も増えつつある。

■e辞林の概要

e辞林（http://www.sanseido.net）は，三省堂が創業120年記念事業として立ち上げた，巨大な辞書検索サイトである。大辞林，新明解国語，デイリーコンサイス英和・和英，独和，仏和，地名事典など16点の辞書を引くことができ，辞書を横断的に見る「串刺し検索」や，本文中の任意の文字列を探し出す「全文検索」なども可能である。

漢字検索のデータベースも持っており，画数や読みから漢字を検索できる。その上，その漢字が使われている大辞林の見出し語まで即座に表示される。このあたりは，電子辞書の面目躍如といった機能である。

■16点の辞書をXML化

プロジェクトは2000年7月にスタートした。開発チームは辞書系とシステム系に分かれ，辞書データのXML化と，そのXMLデータを使った検索エンジンやサーバ・システムの構築を同時並行で行った。

辞書チームは，三省堂から提供されたデジタルデータを，2000年春から設計に着手したDicX（ディック・エックス，http://www. dicx.org）という辞書用XMLに変換する作業を行った。提供されたデジタルデータは，印刷会社やデータ加工会社により4，5種類に大別されたが，これを各種のデータ変換ツールを駆使して，DicX化する作業を行った。

出版コンテンツのXML化は，インターネット時代の出版社の急務だが，16点の辞書がXML化できたことは，大きな成果だと思う。三省堂は，ご存知のとおり，Docomo，goo，Yahoo!と提携し，そこで辞書引きが可能となっているが，三社にXML形式でのデータ提供が可能となった。

システムチームは，EXI（EAST XML Index），LaBamba（ラバンバ）という核になる検索エンジンの改良と，ユーザ・インターフェイスであるBTONIC（ビートニック），そして管理システムの開発を行った。EXI，LaBamba，BTONICについては後述するが，管理システムの開発が難航した。

e辞林は，個人ユーザと法人ユーザで管理方法が異なっている。個人は，ID，パスワードを一年間の期間限定で発行する。法人は複数ID発行方式と，固定IP方式の二種類を使っている。固定IP方式は，法人が社内LANを使っている場合，そのIPアドレス（インターネット上の番地）からの検索を無条件に受けつける方式で，特定IPからの同時ログイン数を使った課金方法が，法人売りでは一般的になりつつある。

このような多種のユーザ管理以外に，「だれが，いつ，どのような言葉で，どの辞書を検索したか」というアクセスログのデータベース管理，サイバーキャッシュ社（http://www.cybercash.co.jp）を使ったクレジットカード決裁など，多くのサブシステムを開発した。

運用も，図のように5台のサーバを使い，3台の検索サーバを並列に置いて，CPUネックにならないよう心がけた。EXIが高速検索を行うので，目標値500万アクセス/月にも耐えるシステムが構築できた。

■辞書検索の三代目

イーストでは，十数年前から各種の辞書検索システムを開発している。初代がViewIng，二代目がDTONIC，三代目がこのDicX＋EXIである。

ViewIngは，十数年前に策定されたEPWINGや電子ブック（EBXA）というCD-ROM上の辞書フォーマットに対応したパソコン・ビュアーソフトで，Windows版とMac版を開発した。出版社から販売される多くの辞書CD-ROMや，多数のWindowsパソコンにバンドルされた。

DTONICは，Windowsの思想に合った新しいデータ構造や操作方法を実現したビュアーソフトで，三省堂「インストール・シリーズ」，朝日新聞社「知恵蔵」，日経BP社「デジタル大事典」，小学館「データパル」などのCD-ROMに採用していただいた。

DicX＋EXIは，インターネット時代に即した辞書検索システムとして，最新のXML技術やインターネット技術を使って設計を行った。三代目の特長は，次のようなものである。

(1) サーバ上でもパソコン上でも，データの在り処を問わない
(2) XML（DicX）を使用しているので，データの転用や保管が行いやすい
(3) 辞書に限らず，事典，新聞や雑誌記事などに幅広く対応
(4) データの更新が出版社で行える
(5) 全文検索が可能

(1) は，検索エンジン（EXI）とユーザ・インターフェイス（BTONIC）が完全に分離しているので，EXIがXMLから生成されたインデックスを高速検索し，サーバ用BTONICがインターネットの先にあるブラウザに対してHTMLを排出する。EXIは移植性が高いので，パソコン用EXIやゲームマシン用EXIを作り，その上に各クライアント用のBTONICを作れば，同じXMLデータを個別のマシンで検索できる。

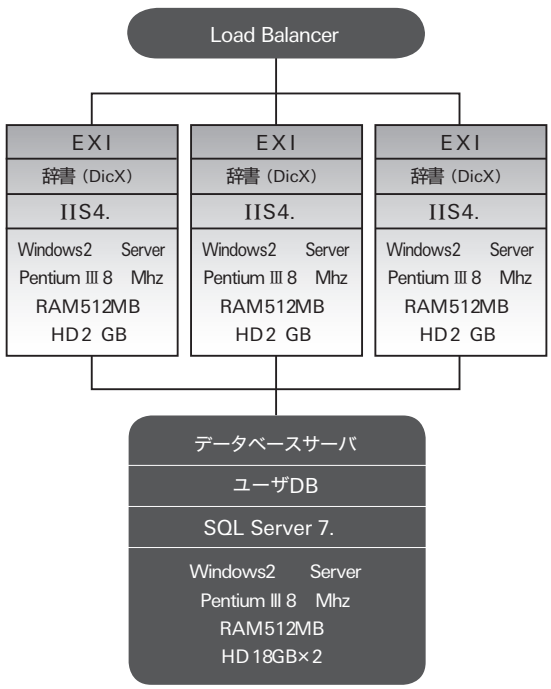
(2) は，XSLを使った強力な変換ができるし，徐々に編集ソフトも揃いつつある。

(3) は，EXIは大量ドキュメントの高速検索手法として開発したので，辞書である必要はない。試作時には，社内で2年分の官報を丸ごと検索したり，音楽事典を検索してみた。XMLデータであれば，何でも検索対象とすることができる。

(4) も，出版社にとって有用な機能である。辞書サーバ方式は，「コンテンツが世界中で一ヵ所にしかない」ので，ここを更新すれば，すべてのユーザに最新のデータが即座に提供されることになる。DicXデータを編集し，LaBambaで全文インデックスを作り，次にインデックス生成ツール操作をして，パソコンで手軽にEXI用のデータを作ることができる。DicXで記述された本体部分とインデックス部分が一緒に圧縮された一つのファイルが生成されるので，これをftp（ファイル転送プログラム）でサーバに置けば，辞書更新が完了する。新規辞書の追加も，同じ方法で可能である。

(5) には，LaBambaと名づけた全文インデックス生成ツールを使用する。その威力は，実際にe辞林のサイトで全文検索を行って確かめていただきたい。大辞林で「青森温泉」のアンド検索を行うと，瞬時に浅虫温泉，酸ヶ湯温泉，蔦温泉などが画面に表示される。

▼「辞林」の管理システム



1998年11月，MicrosoftのBill Gatesは「辞書はすでに画面で読まれている」と発言したが，インターネットにつながったパソコンから手軽に引ける「e辞林」は，年間使用料2000円という低価格も手伝って，インターネット上の標準辞書の地位を得つつある。

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3▶「Rocket BookからeBookManまで」

今回は，事例紹介といっても，電子書籍システムの開発事例ではなく，米国の読書端末の購入事例を紹介する。今までに，1998年11月にRocket Book，昨年12月にその後継機種であるRCA 1100，今年2月にフランクリン社のeBookManと三台を購入した。各々の読書端末との出会いと，使い心地をレポートする。

■Rocket Book

NuvoMedia（ヌーボメディア）社のRocket Bookを見つけたのは，1998年のComdex会場であった。その経緯は本誌1月号「eBookの最近の出来事から」でご紹介したが，数人の友人に使ってもらい次のようなコメントもらった。

「映画館や劇場で便利」
これは，バックライトの件である。紙の本だと映画館や劇場の薄暗い幕間では読めないが，読書端末の光る液晶画面なら，文字が読める！と彼は感激していた。

「書棚が覗かれる」
これは，Rocket Book用の電子書籍を実際に数冊購入した友人からのコメントである。電子書籍はBarnes and Nobleのサイト（http://www.bn.com）で購入するが，そこでは購入する権利を買うだけで，実際のダウンロードはRocket Bookのサイト（http://www.ebook-gemstar.com）で行う。このサイトには，ユーザIDとパスワードを要求される，Rocket Bookユーザ専用のホームページがあり，そこが書棚となっている。読書端末はハードディスクを搭載していないので，数冊の本しかメモリに入らない。そこで，このWeb上の書棚から，読みたい本をダウンロードすることになる。
書棚なら，他人に見せたくない本を裏にしたり，奥に押し込んだりできるが，電子書籍のe書棚は，読書端末のメーカやインターネット書店，サーバの管理者などの第三者から丸見えとなってしまう。

■RCA 1100

この読書端末は，Rocket BookのNuvoMedia社が，Gコードで有名なGemStar社（http://www.gemstar.co.jp）に買収された後，デザインしたマシンである。
GemStarはGコードで有名な会社だが，テレビ番組のビデオ録画を予約するGコードという数字のマジックで得た莫大な利益を，この15年間で三回，大きく投資している。
第一弾がTVガイド誌（http://www.tvguide.com）の購入で，この週刊誌は毎週1000万部が発行され，米国のテレビガイド誌市場を独占している。第二弾がEPG（電子テレビ番組表）で，日本でも電通，東京ニュース通信社と合弁で会社を設立している。第三弾として，2000年1月に西海岸の読書端末ベンチャー企業であるNuvoMedia社とSoftBook Press社の二社を同時に買収した。
昨年11月のフランクフルト・ブックフェアで1100とSoftBookのデザインによる1200の二機種が展示され，1100は直後に出荷された。
1100の購入は簡単で，前出のBarnes and Nobleのサイトから，洋書を買うのと同じ要領で購入できる。送料は30ドル程度なので，330ドルほどをクレジットカードで支払えば，数日で手元に届く。
電子書籍の購入は，パソコン経由でBarnes and Nobleのサイトで購入し，USBケーブルをつないで1100に送り込むこともできるし，本体に電子カタログが入っているので，そこから電子書籍を選択して，本体からダイヤルアップでインターネットに入り，直接ダウンロードすることもできる。

■eBookMan

フランクリン社のeBookMan（http://www.franklin.com/ebookman）という小さな読書端末を知ったのは，昨年10月，フランクフルトのブックフェアの直前だった。米国のeBookメールニュースにフランクリンが小さな読書端末を発表すると書いてあった。フランクフルトの小さなブースには，店番のような営業マンがポツンといるだけで，何も説明してもらえなかった。
11月に出荷とのことで，期待してラスベガスのComdexに行った。フランクリンは大きなブースを出していたが，出荷はクリスマス頃にずれ込むとのことであった。しかし，フランクリン社の技術部長が日本人だったので，詳しく話を聞くことができた。

●読書端末の仕様比較

	重さ	ディスプレイ	解像度	メモリ	バッテリ寿命	周辺機器	販売価格
RCA REB 1100	500g	5.5″モノクロ	320×480	8MB＋Smart Media	20-40時間	Modem,IrDA,USB	US$ 299.00
RCA REB 1200	930g	8.2″カラー	480×640	8MB＋Compact Flash	5-10時間	Modem, LAN	US$ 699.00
eBookMan-911	208g	4″モノクロ	240×200	16MB＋MMC	単4×2,10時間	USB,イヤフォーン	US$ 229.95

Palm（http://www.palm-japan.com）が300ドル以上するのに，なぜこんなに安いのかと聞いたら，32ビットのRISCチップやOSなど，すべて自前で開発したので，他社へのロイヤリティ支払が極端に少ないとのこと。

フランクリン社は小型辞書機器のトップ企業で，セイコー電子やカシオ，ソニーなどが日本で凌ぎを削っている携帯型辞書デバイスの市場を米国で独占している企業である。しかも，辞書や百科事典以外に，電子聖書という大きな市場があるので，2200万点の販売実績を誇っている。

年末にJEPA（日本電子出版協会）でセミナー（http://www.est.co.jp/ks/dish/0012eb/nw14.htm）を行う関係で，実物を借用したが，ハードウエアもソフトもすばらしい出来栄えであった。

eBookManは，次の五つの機能を持っている。

❶小説を読む：電子書籍を読む
❷朗読を聞く：Audible.comの1万2000点が聞ける
❸音楽を聴く：MP3の再生
❹個人情報管理：予定表，ToDo，電卓など（Palmはこれだけ）
❺メッセージの録音・再生

結局，クリスマスには出荷されず，今年3月，フランクリンのサイトから直販が開始された。私は，2月に開発者向けバージョンを購入したが，これは，完全スケルトン，つまり透明なケースに入っており，基板が丸見えとなっている。

肝心の読書端末としての機能は，リーダーソフトであるMicrosoft ReaderのeBookMan版の出荷が春過ぎとなっているので，それ以降となる。Readerの出荷が始まれば，Amazon.comのeBookのページから，eBookMan対応の電子書籍が購入できる予定である。

日本語に対応するのは少し先になると思うが，今の画面解像度では，日本語での読書は無理である。Palmの160×160は論外だが，Zaurusの320×240または480×320くらいの解像度を期待したい。

eBookManは，電子書籍のダウンロード方法も見事である。写真のとおり，インターネットにつながったパソコンを経由して，デスクトップ・マネジャーというソフトが仲介役となって，インターネットから直接，USB接続されたeBookManに電子書籍がダウンロードできる。

2月に入手した開発者向けeBookManには，一切のソフトが入っておらず，フランクリンのサイトから，まずデスクトップ・マネジャーをパソコンにダウンロードし，次にOS自体を，パソコン経由でeBookManに流し込むという方法だった。この仕組みには感動した。

早い時期に日本語化されることを期待したい。

インターネットから直接，電子書籍がダウンロードできるeBookMan

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4▶「PDFかXMLか」

最近，出版社からPDFやXMLの制作依頼が，たくさん寄せられている。PDFとXMLでは，デジタルデータとしての位置づけも，用途も全く異なるものであるが，両者ともに「電子書籍」のデータフォーマットとしての利用が始まっている。

■XMLデータの作り方

XMLデータを作る場合，まずスキーマおよびDTDを決めなければならない。その書籍の構造にあわせたスキーマが必要となるが，文庫，新書などテキスト主体の書籍の場合は，日本電子出版協会（http://www.jepa.or.jp）が策定したJepaX（http://x.jepa.or.jp/jepax）がそのまま使える。辞書系の場合は，イーストで策定中のDicX（http://www.dicx.org）も参考にしていただきたい。

特殊な事典や複雑な構造を持った書籍の場合は，そのシリーズごとにスキーマを設計することになるが，最近，この設計依頼が増えている。スキーマ言語であるDTDを使った書籍の仕様書，DTD，簡易スタイルシート（XSL）をセットにして納品している。

スキーマ言語は，DTD以外にRELAX，XML Schemaなどがあるが，SGML時代からの慣れや処理系の問題でDTDが多く使われている。

DTDを作った後は，いよいよXMLへの変換作業となるが，出版社からの提供形態には「書籍」「CTSデータ」「DTPファイル」の三通りがあり，それらを自動タグ付けツール，手動タグ付け作業などを行い，XML化している。

■なぜXML化するのか？

紙の本を，なぜXML化するかといえば，ワンソース・マルチユースを行うためである。図1のように，XMLには，XSLTという強力な変換機能があり，各種のフォーマットに変換して，ホームページや電子書籍，eLearning，ドキュメント検索などが可能になる。eLearningは第一回でご紹介したNetLearning社などのインターネットを使った教育であり，ドキュメント検索は第二回でご紹介した三省堂.netのようなXML検索サービスである。

話題のMicrosoft Readerのような電子書籍も，XMLから米国

図　XMLの場合

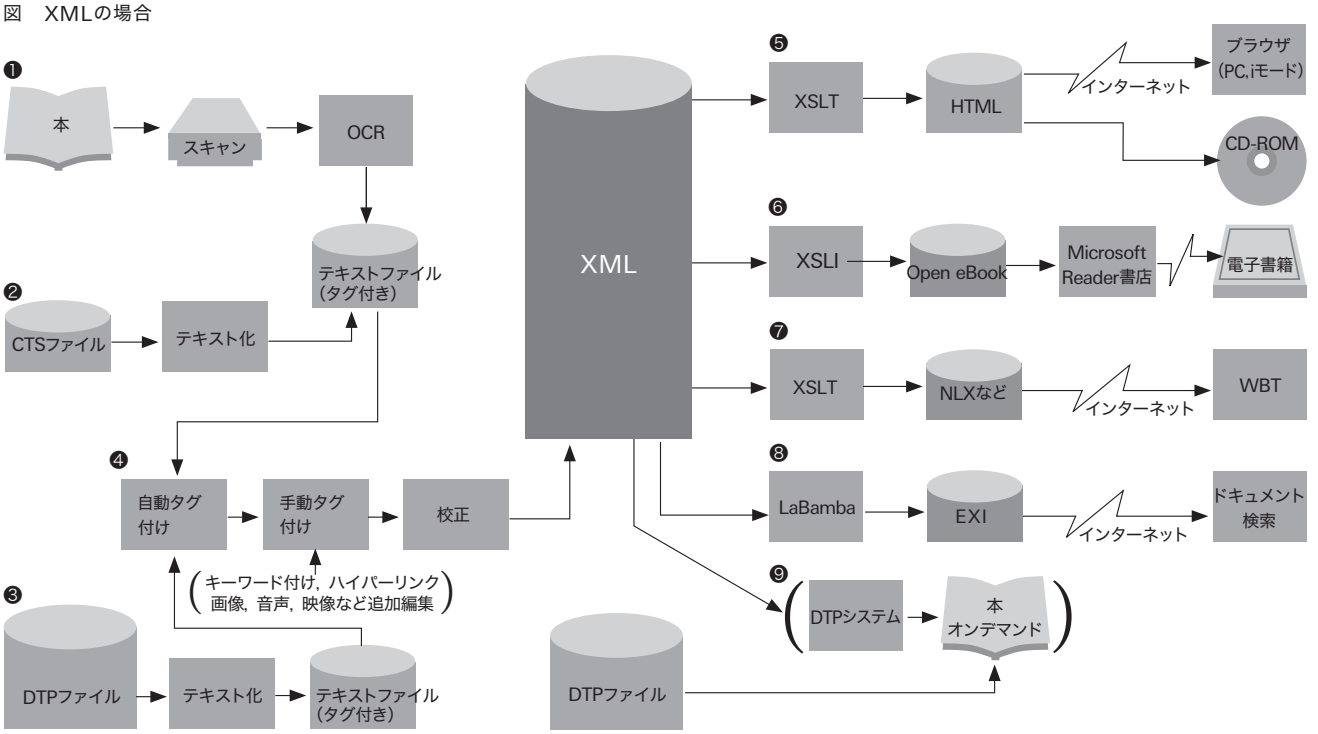
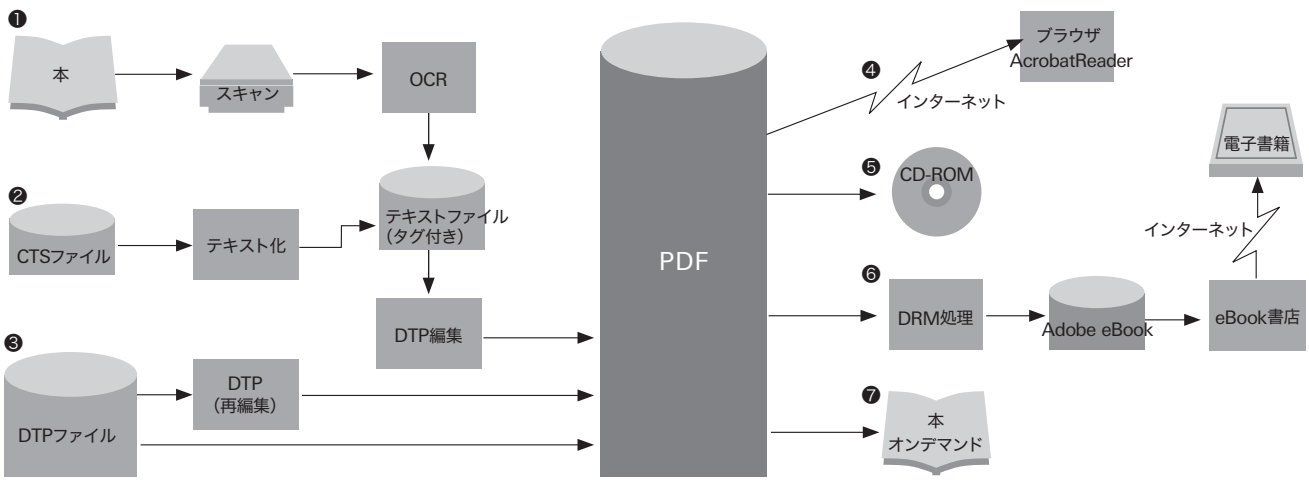


図2 PDFの場合



の標準電子書籍フォーマットであるOpen eBook Publication StructureにXSLで変換し，マイクロソフトが提供しているReader SDK（Software Development Kit）を使って，litファイル（Reader用の書籍データ）への変換が可能である。

XMLでデータを保管しておけば，改訂や体裁の変更が簡単に行えるのだが，XMLの編集ツールがまだ整備されていないので，辞書などの大量データや複雑な構造の書籍の場合は，修正に手間がかかる。

夏ごろ出荷予定のMicrosoft Office XPでも，データベース系のExcelやAccessではXMLファイルの読み書きが可能となったが，Wordでの XMLの読み書きは，その次のバージョンを待たなければならない。

XMLから電子出版，インターネット出版への道はたくさん開かれているが，XMLから紙への印刷は，まだデコボコ道である。XMLは，本来ドキュメントの論理構造を取り扱うものなので，細かな体裁の指定には向いていない。スタイルはXSLの担当となっているが，DTPソフトのような細かな指定は行えない。XMLデータをDTPソフトに流し込む際にも，多少の手間がかかる。

Quark XPressの場合，avenue.Quarkを使ってXMLでの取り出しは頁単位で可能であるが，XMLの読み込みは現行バージョンではサポートされていない。

■PDFならすぐにビジネス

読書ソフトAdobe Acrobat eBook Readerや電子書籍販売サーバAdobe Contents Serverの発表があったためか，PDFファイルの制作依頼も増加している。

図2のとおり，DTPで作られた書籍の場合は，ボタンを押すだけで，いとも簡単にPDFファイルを作ることができる。しかし，画面表示用の画像調整，MacintoshとWindowsのフォントの差異，外字の作成と設定など，電子書籍として製品化するには，出版社では手の負えない作業も多い。

PDFだと，話題のオンデマンド出版にもシームレスに対応できるし，Webでの公開も容易であるが，XMLのような汎用性，拡張性はない。

また，PDFから電子書籍を作る場合，画面サイズや解像度の問題で，二頁表示を行うには40字×20行程度となってしまう。新書判や文庫判ならまだしも，二段組の大判の書籍は再編集が必要となる。

しかし，読書端末のハードウエア・テクノロジーは飛躍的に進歩している。数年後にはA4判の液晶画面が登場するので，それを待つのが得策である。

ということで，すぐに電子出版ビジネスを立ち上げるならPDFをお奨めするし，将来もそのデータを有効利用するならXML化を検討していただきたい。

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5▶「DTPからXMLへ」

前回,「PDFかXMLか」というテーマで,XMLとPDFの現状を説明したが,今回はその続編で,DTPファイルをXMLに変換する具体的な方法をご紹介する。

弊社が採用している方法は,いかにもソフトウエア会社らしい方法だが,今後,需要が急増するXML化作業の参考にしていただきたい。

■はじめに官報ありき

弊社は,インターネットやWindowsに関連したソフトウエアを開発する会社だが,1999年12月に,妙な縁で官報のXML化をお手伝いし,その延長で,ドキュメントのXML化作業を今でも毎月1万頁ほど行っている。

官報のXML化は,戦後すべての官報,全88万頁について,各頁のコピーを受領し,XMLファイルを納品する作業が当時の大蔵省から発注された。イーストは,落札業者数社から,合計12万頁のXML化作業を受注した。弊社が担当したのは,デジタル化済みのテキストファイルから,官報DTDに沿ったXMLタグ付きのデータを作成する部分である。

この作業は,大きく次の工程で行った。

1. 自動タグ付け→ 2. 手動タグ付け→ 3. 表や図の設定→ 4. 最終確認

「タグは機械が処理するものなので,機械で付ける」,「膨大なドキュメントの処理は,コンピュータが行うべき」という考え方から,徹底的なシステム化を行った。

1.の自動タグ付けは,プレーンなテキストデータを読み,文字パターンの検索や前後の文章から,可能な限りタグ付けを行うもので,官報の年代やジャンルに合わせた,多数の自動変換プログラムを作成した。

2.は自動変換が不可能なタグについて,アルバイトでも使えるようなシンプルな専用エディタを開発し,これを使って人海戦術で行った。

3.は官報には決算書などの複雑な表が多数存在するので,その熟練工を養成し,図も含めて別工程とした。

4.は文字校正は弊社の責任外だったので,レイアウトやタグについての校正を,当時ベータ版が登場していたInternet Explorer 5.5の縦書き表示機能と文字鏡URLフォント(http://font.mojikyo.com)による外字表示機能を使って,できるだけ平易な作業にして生産性を高めた。

3.で別工程などと簡単に書いたが,一般的な事務作業で,工程の変更や細分化は難しい。これらのデジタル化作業のすべてを,vfolder(http://www.est.co.jp/vfolder)という工程とコンテンツを管理するサーバを使って行い,成果をあげることができた。

■DTPからXMLへ

官報のXML化を担当した理由は,JepaX(http://www.est.co.jp/ks/dish/jepax)により,印刷業界で多少は社名が知れていたためだが,官報XML化プロジェクトの受注により,その後も,書籍からXMLへの変換作業の依頼がきている。

日本では,DTPで作られた書籍は少ないが,技術系の出版社はDTPを多用しており,そのような出版社のDTPファイルをXMLに変換する仕事が多い。

DTPと一言でいっても,Quark,PageMaker,InDesign,Wordなど,アプリケーションソフトによって変換方法が異なる。問題集などは,Accessのデータで提供される場合もある。

これらのDTPファイルを,可能な限りコンピュータを使って変換している。

Quarkの場合は,図のとおり,avenue.quark(http://www.quark.co.jp/products/avenue)を使ってXML変換を行っているが,「頁単位の変換で,操作が面倒」,「日本語タグに対応していない」などの問題が発生した。

前者は,avenue.quarkが,書籍ではなく新聞や雑誌などの変換を想定しているので,今後のバージョンアップを待つしかない。後者は,XMLはUnicodeベースなので,日本語タグも難なく設定できるが,外国製のツールを使うときに問題となる。しかし,これも,英語タグを適当に決めて変換し,その後,テキスト・コンバータで一気に日本語タグへの変換を行っている。

WordやPageMakerの場合は,HTMLフォーマットでの一括書出し機能使ってHTMLファイルを作成している。これをJavaScriptで変換プログラムを作成し,XMLに落とし

ている。Perl（http://www.psl.ne.jp）やVisual Basic（http://www.microsoft. com/japan/developer /vbasic）ではなくJavaScript（http://www.justnet.ne.jp/ javascript）を使う理由は，DOM（ドム）が扱えるからである。

DOM（Document Object Model）は，XMLドキュメントを操作するためのアプリケーション・インタフェースで，官報XML化プロジェクト以来，これを使ってXMLドキュメントの解析や生成を行っている。

avenue.quarkの場合は，書籍の構造をそのままXMLにしてくれるので，しっかり文書構造を決めて製作された書籍であれば，すぐにXML化できる。しかし，HTMLの場合は，「構造」と呼べるほどの情報が含まれていないので，書籍のレイアウト上の特徴や，特有の文字列などを手がかりにして，一冊，一冊，JavaScriptのプログラムを作成している。

一冊ごとにプログラムを作るなんて!と思われるかもしれないが，プログラマーが100人以上いる会社なので，お手のものである。何回か試行錯誤の後，プログラムが完成すれば，1000頁の書籍でも数秒でXMLファイルが生成される。

この後，表や図の貼込みや最終確認などの作業を行うが，目次作成はXSLで簡単に行えるし，索引作成には，事例その2「三省堂 e辞林」でご紹介した，LaBambaという全文インデックス生成ツールが応用できる。

Adobe InDesignからのXML変換は，いくつかのルートがあるので，弊社にとっての最短ルートを現在調査中である。Accessからの変換は，問題集程度であれば，CSVファイルをJavaScriptでXMLに変換している。

■ XMLからPDFへ

4月の東京国際ブックフェアでAdobe社がContent Server（http://www.adobe.co.jp/products/contentserver）とAcrobat eBook Reader（http://www.adobe.com/products/ebookreader）という電子書籍の仕組みを発表した関係で，Adobe eBookの基本フォーマットであるPDF製作の依頼も増えている。

前出のQuark，PageMaker，Wordなどの DTPソフトで作られた書籍の場合は，Acrobat（http://www.adobe.co.jp/products/acrobat）を使って，数回のクリックでPDFファイルが作成できるが，事例その4（前号）「PDFかXMLか」でご紹介したとおり，JepaXやHTMLからDTPソフトへの流し込みはちょっと厄介である。QuarkやInDesignには，入出力が可能なオリジナルのタグが用意されているので，JavaScriptを使って各フォーマットへの変換を行っている。

DTPソフトに流し込んだドキュメントは，手作業でページレイアウトの調整を行い，eBook化を行っている。この際，Acrobat eBook Readerの特徴である二頁表示を行うには，一頁を20行×40文字程度に抑える必要がある。

eBookを作る場合，外字のインライン化，画像の低解像度化，表紙画像の作成，コピーや印刷などの許諾範囲の設定を行い，DRM（デジタル著作権管理）処理を入れる。このファイルをAdobe Content Serverに対応したオンラインショップに登録すれば，電子書籍の販売がスタートできる。

Quarkのデータはavenue quarkを使ってXML変換