XMLパブリッシングのための出版標準フォーマットへの取り組み

田原 恭二†
KYOJI TAHARA

インターネットの発達により，ネットワーク全体が大きな一つのデータベースになりつつある．ブロードバンド時代に対応するためには，データベース上の情報を，いつでも，誰でも，どこからでも取り出せ，利用できることが重要である．そのためには，データベース上の情報表現ルールが統一されていることが望ましい．XMLの利用が各分野において非常な勢いで拡大している最大の理由である．

出版分野においてもXMLの利用は広がっている．XMLを利用することにより，「コンテンツ」「文書構造」「文書体裁」を分離して持つことが可能となる．データベースに構築し，一元管理による資産管理が容易になるとともに，検索性の向上，多くのメディアへの展開，処理の効率化，時間短縮，コスト削減，正確性の向上などが可能となる．

1. はじめに

1.1 インターネットの進展

図1.に示すように，総務省の速報[1]によると2003年7月末時点で国内のインターネット接続サービス加入数は9,624万件となり，インターネット利用人口は6,942万人（2002年末）となった．中でもブロードバンドの利用は，1年間で2.4倍強の飛躍的な進展をみせており，現在はインターネット利用者の4人に1人以上がブロードバンドの利用が可能となっている．また，「カメラ付き携帯」等のヒットにより，携帯インターネットも発展しており，現在，携帯電話加入数7,780万件のうち84%（6,517万件）が，インターネットへのアクセス可能となっている．

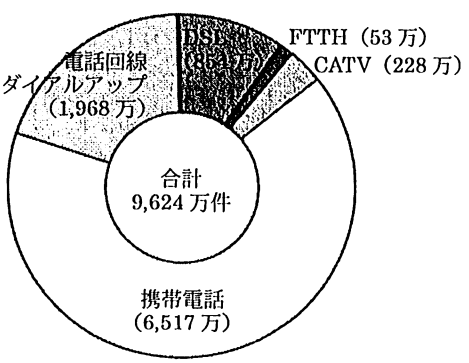


図1. 国内のインターネット接続サービス加入数
（2003年7月末現在）

このような値や，最近特に街角で携帯電話の画面を見ている人々を多く見かけるようになった，ということも含めて，インターネットがより身近なものとなりつつあることを実感させられる．

1.2 インターネットコンテンツ量の推移

インターネットコンテンツの情報量も増大している．4年前の1998年（平成10年）では，国内のインターネットコンテンツの総データ量は664ギガバイトで，総ファイル数は3,648万ファイルであったが，2002年末（平成14年）には，総データ量が10,150ギガバイト，総ファイル数が27,421万ファイルとなった．これは，この4年間で，総データ量が約15倍，総ファイル数が約8倍増加したことになり，今後もますます増大していくことが予測[2]されている．このよう

† 凸版印刷株式会社Eビジネス推進本部研究開発部
Research and Development Department, E-Business Operations, TOPPAN Printing Co.,Ltd.

なインターネットコンテンツ量の推移からも，インターネットが着実に，大きな一つのデータベースになりつつある，ということが窺える．

1.3　XMLが注目されてきた理由

図2に，インターネットの発達に伴う，新しい情報流通市場の創出，これまでの情報システムの基本的な仕組みの変化，インターネットというグローバルネットワークが大きな一つのデータベースへと進化する．という潮流を示した．

このような中では，データベース上の情報を，いつでも，誰でも，どこからでも取り出せ，利用できることが大変重要になってくる．そのためには，データベース上の情報をどのアプリケーションからでも簡単に取り扱える，共通な情報表現ルールが必要となる．このことが，現在，各分野においてXMLの利用が注目されている最大の理由である．

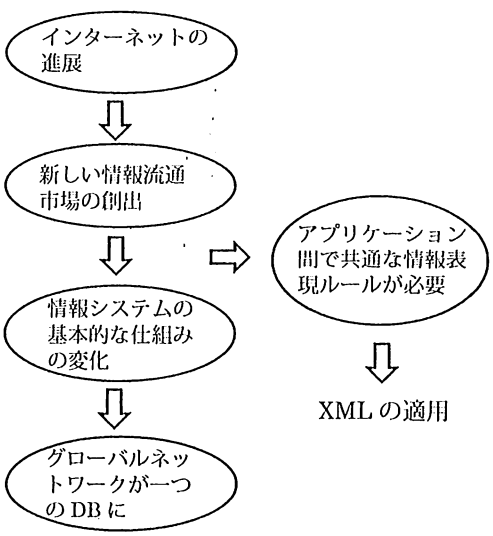
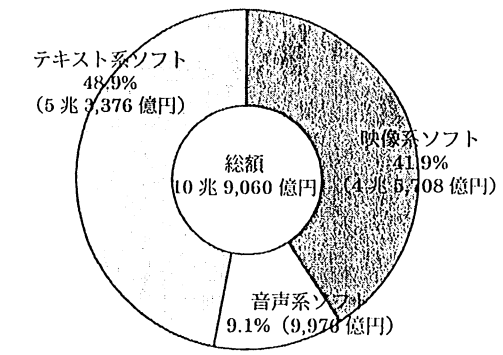


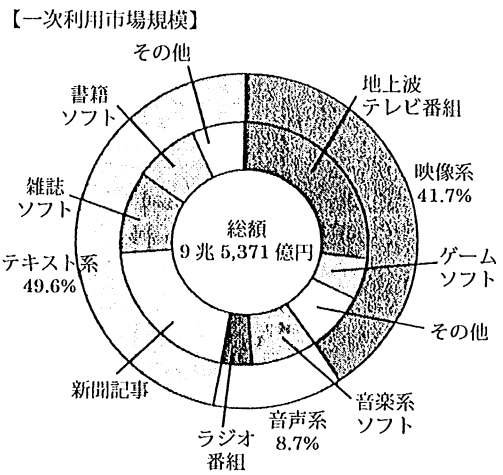
図2.　インターネットの発達による潮流

2.　メディア・ソフトの流通市場規模

図3.は，国内におけるメディア・ソフトの流通市場規模[3]を示したものである．

このうち，出版分野に大きく関わるテキスト系ソフトについては，市場の約5割を占めている





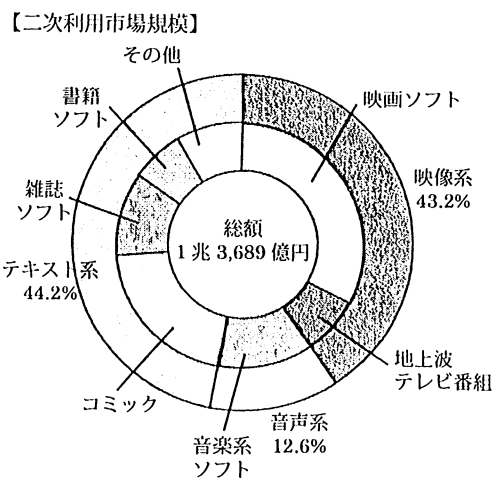


図3.　メディア・ソフト流通市場規模（2002年末）

ことがわかる．更に，コンテンツの一次利用と二次利用という視点で詳細を見てみると，テキスト系ソフトについては，新聞記事など，より即時性が求められるコンテンツは一次利用が主体であり，週刊漫画雑誌からコミック化の流れが定着している漫画コンテンツは二次利用性が高い．

また，これら以外の書籍ソフトや雑誌ソフト等の，いわゆる“文字が主体のコンテンツ”は，全体シェアは少ないものの，一次利用と二次利用の比率変化が少ないことから，再利用性が高い，及び多メディア展開が行い易いといった性質があると思われる．

3.　市販組版システムの現状

現在，書籍や雑誌の制作で多く使われている市販の組版システムでも，XML 対応のものが数多く出てきた．ネイティブな XML データベースと連動し，コンテンツ管理を容易にするとともに，自動組版による処理の効率化，時間短縮，多メディアへの展開などが実現されている．

しかしながら各システムは，従来の組版システムを活かした内部フォーマットを利用しており，コンテンツの受け渡しには変換プログラムが必要である．これは，データ流通という面から見るとまだまだ問題である．あわせて XML データを組版システムに取り込み，しかるべき組版編集を行った後で，組版データを元の XML データへ自動的に完全復元できないといったことも，資産の一元管理やワンソース・マルチユースといった面で，大きな問題となっている．また，XML そのものを入力，或いは編集するツール（XML エディタ）の機能が弱く，複数ページにまたがるような修正が可能なシステムが少ないのも現状である．

4.　XML パブリッシングモデル

図 4. は，現在凸版印刷が取り組んでいる多メディア対応の XML パブリッシングのモデルである．当社が考える理想のモデルは，入力から出力まで，いかなる出力媒体に対しても，XML を中心とした運用が行えることであり，その中心とな

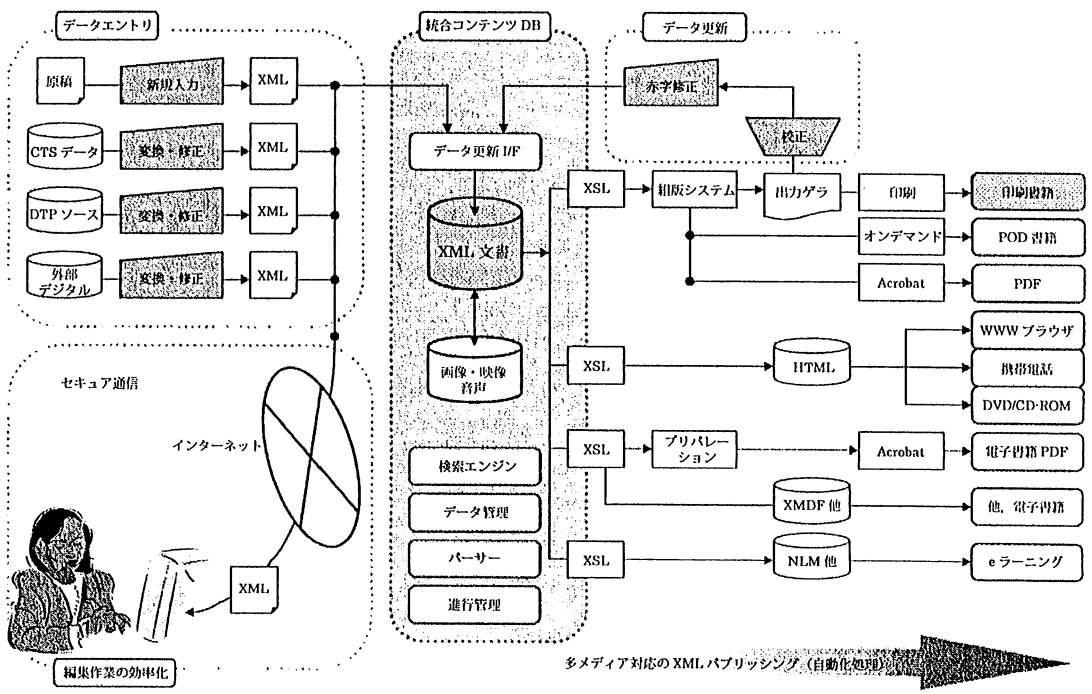


図 4.　多メディア対応の XML パブリッシングモデル

るのは，データの展開が容易に行える標準フォーマットによるXMLデータベースである．

入力側の対応は，紙原稿などからの新規入力によるXML化，既存のCTSデータやDTPデータからの変換，スキーマの異なる外部デジタルデータの取り込みなどを考えている．また，出力側は組版システム経由で，書籍印刷，オンデマンド印刷，PDF出力のほか，Web展開，CD-ROM・DVD-ROMなどのパッケージ展開，電子書籍などへはスタイルシートを介して展開しており，最近ではeラーニングへの展開も出てきている．

5. 出版標準フォーマットの現状

5.1 出版標準フォーマットの現状

現在，出版業界におけるXMLによる標準フォーマットは，残念ながら存在しておらず，先に述べた通りXML対応の組版システムが利用しているフォーマットは，いずれも独自のものである．電子出版分野においては，日本電子出版協会により1999年に策定されたJepaX（バージョン0.9）[4]があるものの，現状活発に利用されているとは言いがたい．また，JepaXが中立的な位置づけのデータ交換フォーマットであるため，一般書籍印刷用としては不足している要素があり，印刷会社でそれら不足分を追加する必要がある．

他のシステムで作成されたデータを利用する場合にも，各々のシステム毎に変換アプリケーションソフトを作成し，対応しなければならない．よって，できるだけ早く業界の出版標準フォーマットが決定され，負荷なくデータ流通がなされることが望まれている．

5.2 出版標準フォーマットの作成

当社では，このような状況を踏まえ，内部生産性の向上・コンテンツ二次展開へのスピードアップ・処理の自動化によるコスト削減・コンテンツマスターの一元化による品質保証・データ流通性の向上等の必要性から，出版標準フォーマットを作成した．この出版標準フォーマットは，出版コンテンツに対して，アクセス性はもちろんのこと，正確性，拡張性などに加え，社内の展開性，及び業界のガイドラインとして利用することも考慮するとともに，JepaXやJIS X 4052[5]等，既存の標準フォーマットと互換性が取れるように考慮しながら作成した．

6. 出版標準フォーマット解説

6.1 概要

図5.は，出版標準フォーマットの外観を示したものである．このフォーマットは，JepaXをベースにし，JIS X 4052, HTML 4.0[6], OEB1.0[7]などを参照しながら拡張を行った．

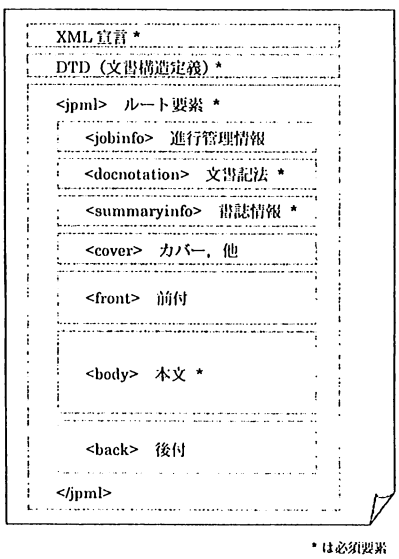


図5. 出版標準フォーマットの外観

対象としているコンテンツは，現状は，一般書籍，学術論文誌，法律書など，編・章・節・項・段落・見出し，といった文書構造や，表現体裁がパターン化できるものに限定している．

また，図4.に示したXMLパブリッシングモデルを実現するために，コンテンツの論理構造と，スタイル指示の分離に努めており，スタイル指示は利用目的に応じ，スタイルシートで指示することを前提としている．要素は，論理構造要素・ブロック要素・インライン要素とがあり，要素数は，合計102種類となっている．

6.2　拡張した仕様

このフォーマットが，他の出版関連仕様と比較して強いところは，日本語文書の論理構造単位をJepaXの持つ26種類から47種類へ細分化している点や，索引バリエーションへの対応・組方向への順応・言語混在・欧文ハイフネーションなど数多くあり，これらの拡張は，主に出力側で組版処理を容易に行えるようにするという前提によるものが大きい．次に幾つかの拡張仕様を解説する．

索引バリエーションへの対応

日本語書籍における索引は，人名索引・分野別索引・地名索引など，様々な索引が存在する．また，索引の配列も，読み五十音順・姓名五十音順・姓名五十音＋字画順・電話帳配列など，多くのバリエーションが存在する．これらに対応するためには，索引種類の別や，姓名であれば読みの姓・名の別，漢字表記の姓・名の別，漢字表記一文字に対応した読みなどを，何らかの形で把握できなければ機械的に行うことは不可能である．

今回，例1.に示すように，索引項目になる本文中の文字列に対して，配列用の「読み」「表記」属性を持たせられるようにした．

例1.　索引配列を想定した人名のマークアップ

```
<name type=" 人名 " reading=" もり，た／いち，
ろう " file-as=" 森，田／一，郎 ">
森田一郎 </name>

<name type=" 人名 " reading=" もり／た，いち "
file-as=" 森／太，市 ">
森太市 </name>

<name type=" 人名 " reading=" もり／た，ろう "
file-as=" 森／太，郎 ">
森太郎 </name>
```

[配列イメージ]
姓名五十音順（左）と，読み五十音順（右）

森　太市　　森　太市
森　太郎　　森田一郎
森田一郎　　森　太郎

組方向への順応

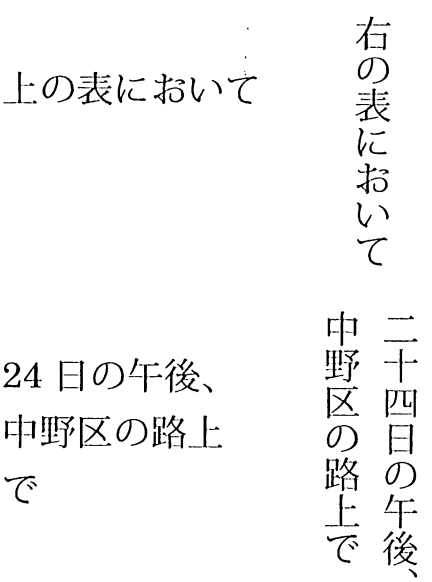
書籍の大幅な紙面変更や，コンテンツを二次利用する場合など，縦組から横組へ（またはその逆）など，組方向が変わる場合があり，本文中の「漢数字」と「算用数字」や，「上」と「右」など，表記文字そのものを変更しなければならない場合がある．このようなケースの対応負荷を軽減させるために，例2.に示すような，組方向に応じたテキストの切替えができるようにした．

例2.　組方向に応じたテキスト切替え

```
<swtext>
    <st type="ht"> 上の </st>
    <st type="vt"> 右の </st>
</swtext> 表において…

<swtext>
    <st type="ht">24 日 </st>
    <st type="vt"> 二十四日 </st>
</swtext> の午後、中野区の路上で…
```

[組版イメージ] 横書き（左），縦書き（右）



言語混在・欧文ハイフネーションの対応

日本語文章中には，欧文など，日本語以外の言語が混在する場合がある．日本語文字組版では，このような言語が混在する場合について，多くの組版ルールがあり，そのルールに基づいた細かな

対応が求められる．また，欧文組版では，単語が行末にかかる場合，分綴（ハイフネーション）と呼ばれている特別な処理を行う場合もある．

これらの対応を各出力側で行うと，その都度負荷が発生するため，今回，例 3. に示すように，言語の別（lang 要素）と欧文ハイフネーション位置（shy 要素）について文書構造の一部として捉え，これらの対応が自動的に行えるようにした．

例 3.　和欧文混在と欧文ハイフネーション

```
<div type=" 項 1">
    <head><title> つけぐすり </title></head>
    <p>
        付け薬
        <lang class="en">
            a medicine for ex<shy/>ternal
            application.
        </lang>
    </p>
</div>
```

[組版イメージ]

付け薬（つけぐすり）　a medicine for external application.
告げ口（つげぐち）　tell (on); tell tales (about).

同一内容への対応

例 4. に示すように，名簿や財務諸表等の紙面でよく見かける「同左」「同上」「〃」等の表記箇所について，表記テキストと，対応する元のテキストとを合わせて保持できるようにした．

これにより，名簿などにおいて都道府県名が同じ間は，先頭（およびページの変わり目の先頭）のみ表示させるような体裁や，財務諸表を単年で利用したい場合など，データの取り扱いが容易になっている．

例 4.　名簿における同一内容への対応

```
<same-as>
    <sab> 東京都 </sab><sat> 〃 </sat>
</same-as> 千代田区
<same-as>
    <sab> 東京都 </sab><sat> 〃 </sat>
</same-as> 江東区
<same-as>
    <sab> 東京都 </sab><sat> 〃 </sat>
</same-as> 葛飾区
<same-as>
    <sab> 東京都 </sab><sat> 〃 </sat>
</same-as> 北区
```

7.　用途サイクル

図 6. は，出版標準フォーマット文書を使った用途サイクルとして，書籍制作の例を示したものである．

大本の出版標準フォーマット文書を，スタイルシートを使って変換し，組版システムの入力データを作成する．組版システムから組版ゲラが出力され，赤字指示があればデータ修正を行い，最終的にフィルムや CTP 出力を行う．

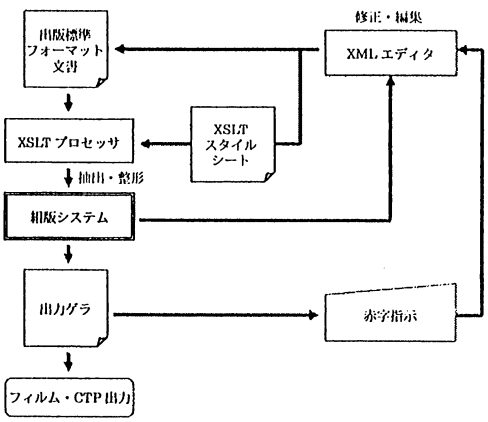


図 6.　用途サイクル概念図

このサイクルにおけるポイントは，データ修正が，常に元の XML 文書に対して行われることであり，これによってデータの一元化と，二次利用性を高めている．

また，出力ゲラ上の赤字指示に対する XML 上の修正箇所が特定しにくいため，XML データの各要素と，出力ゲラの頁・段・行を対応づけ，容易に特定できるようにし，作業効率の向上を図っている．

これらを実現のするために，今回，組版システムの拡張と，XML エディタの開発を行った．

8. XMLエディタの開発

今回，求めているXMLエディタの要件は，

- ●書籍データが入力できること
- ●初期入力（エントリ）と，データ修正を同一のツールで行えること
- ●様々なスキーマを把握できる汎用性を持っていること
- ●制作体制の構築において，大多数へツールの展開が容易に行えること
- ●進行管理システムなど，既存のシステムと親和性を高めるI/Fが取れるなど，拡張性があること

であり，これに対して市販製品も十分検討を行ったが，要件を満たすXMLエディタがなく，社内開発し，検証を行った．

8.1 XMLエディタの概観と機能

図7.は，開発したXMLエディタのメインウィンドウを示したものである．書籍制作に関わる対象入力者の平均スキルを踏まえ，XML文書を直に編集していくシンプルなテキストエディタとし，Windows®の「Notepad」並みの軽快さでテキスト編集を行うことができる．更にキーバインド設定（入力キーと入力データの関連づけ）が行えるようにしてあり，タグパターン等を簡単なキー操作で入力できるようにした．また，入力対象となるスキーマを読み込み，入力可能な文字の制限や，タグ属性値の選択肢表示，入力データの検証等，パーサー関連の機能を搭載し，入力精度が向上するようにした．

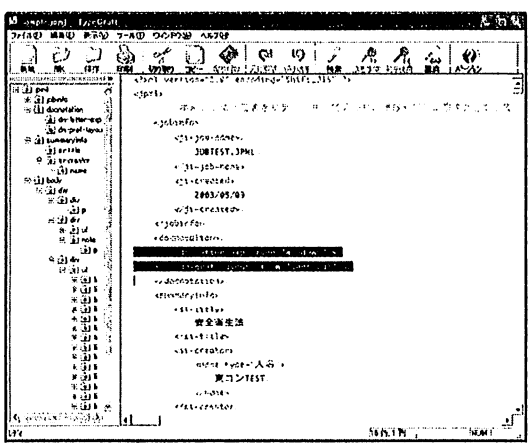


図7. XMLエディタ

出力ゲラの赤字指示に対するXMLデータの修正箇所の特定は，組版システムが生成する関連づけの情報をXMLエディタが読み込み，エディタ上で赤字箇所の頁・段・行を指定することで，編集画面が自動的スクロールし，該当箇所にジャンプするようになっている．

9. 検証・評価実験

これまで述べた出版標準フォーマット文書を使った用途サイクルとXMLエディタについて，運用面での課題や問題点を把握検証すべく，書籍制作入力者1名に参加してもらい，約2ヶ月間にわたり評価実験を行った．

9.1 評価実験環境

実験に使用した環境は，XMLデータ入力と，赤字修正作業は，Windows®のパーソナルコンピュータを使って行った．

組版システムは，社内で保有するシステムを用い，実験用のサンプルデータはパターン化しやすい構造を持ち，且つ，組版体裁が比較的高度なものを選定し，データ量は，出力ゲラで約100ページ分のデータを使用して行った．

9.2 新規データ入力作業

新規データの入力作業について，次に示す三つの方法で実施し，作業負荷・作業効率・入力精度等の観点から比較を行った．

- ●原稿を見ながら，入力者がタグを判断し，タグとテキストを同時に入力する方法（方法1）
- ●事前に原稿にタグを記入しておき，入力者は原稿通りにタグとテキストを同時に入力する方法（方法2）
- ●先にテキストを入力してから，後でタグを入力する方法．（方法3）

表 1. に，比較結果を示した．方式 3 のテキストとタグを別々に入力する方式が最も効率の良い結果となった．

表 1.　新規データ入力における入力方式の比較

方式	負荷	効率	精度	総合評価
方式 1	×	×	△	×
方式 2	△	△	×	△
方式 3	○	○	○	○

○ 良い　△ふつう　×悪い

9.3　赤字修正作業

赤字ゲラを修正原稿とし，段落内のテキスト挿入・削除・移動をはじめ，文書構造の階層の移動，別紙原稿の挿入，ルビをはじめとしたインライン要素の変更，表の挿入など，修正作業を行った．新規入力と比較すると，赤字修正作業のほうは，作業負荷が少なく，XML データを直接修正する作業も十分対応可能であった．これは，既にデータ中に参照可能なタグが入力されていることにより，新規入力よりもタグの仕様を意識する必要が少ないためであると思われる．なお，ルビや数式などの，細かなタグ付けが必要な要素の修正では，入力者の思考する傾向が強く，多くの時間を必要とした．

9.4　XML エディタの評価

現在，書籍制作では，データ入力・修正ツールとして，ワードプロセッサやテキストエディタを使用しており，今回の XML エディタの操作に対する使いづらい面はなかったが，画面上でタグとテキストを見ながらの編集になるため，入力者が，現在編集中の階層位置などの，編集状況が分かりづらい面があった．よって，もっと入力者の思考を促さない，作業に集中できるユーザインターフェースへ，改良すべきであると思われた．また，入力エラーに関して，パース機能を搭載し，エラーの防止を図ったわけだが，入力者に対し，具体的なエラー箇所の位置・原因・回避方法等が明確に通知されないメッセージが多く，入力者がその場で対応を考えなければならなく，一部の箇所で，対処方法が判断できず，作業効率を低下させる要因となった．

9.5　出版標準フォーマットの評価

現在の仕様では，文書構造の定義に論理構造単位の div 要素を多様することになってしまった．div 要素が type 属性を使って階層意味を表すため，今回のようなタグとテキストを同時に表示する XML エディタでは，視覚的に階層が判りづらくなってしまった．また，プログラムによるデータ検証も負荷が高くなるということが判明したため，この点に関しては，フォーマットそのものの再考が必要であると思われる．

10.　おわりに

今回の検証・評価実験で，出版標準フォーマットを中心とした用途サイクルをシミュレーションし，有効性や合理性を幾つか確認できた．実運用のためには，フォーマットの最適化や，XML エディタの機能拡張など，課題も把握できた，これらは引き続き検討していきたい．あわせて，進行管理システムへの組み込みや，コンテンツデータベースとの連携等，検証を行っていく予定である．なお，この出版標準フォーマットては，今後検証がまとまった段階で，多方面から評価していただく意味でも，オープン化したいと考えている．

参考文献

[1] 総務省．インターネット接続サービスの利用者数等の推移【平成 15 年 7 月末現在】(速報)．8, 2003.
[2] 総務省．情報通信統計データベース．情報通信白書．平成 15 年度版．5, 2003.
[3] 総務省．情報通信統計データベース．情報通信白書．平成 14 年度版．5, 2002.
[4] 日本電子出版協会．JepaX．3, 1999. http://www.jepax.org/
[5] 日本工業規格．日本語文書の組版指定交換形式．JIS X 4052, 10, 2000
[6] WORLD WIDE WEB consortium, HTML 4.0 Specification, 4, 1998
[7] Open eBook Forum, Open eBook Publication Structure Specification Version 1.2, 8, 2002.