SGML/HTMLを考慮したドキュメント構造汎用化方式の検討

3AC−4

甲斐こずえ[*1]　藤本憲司[*1]　杉山和弘[*2]

NTTソフトウェア本部[*1]　NTTヒューマンインタフェース研究所[*2]

1.現状

電子雑誌システムでは、雑誌データはDTD(Document Type Declaration)の定義に従ってSGML(Standard Generalized Markup Language)化され、DBへ登録される。ユーザはブラウザを用いて、HTMLに変換されたデータを参照する（図1参照）。

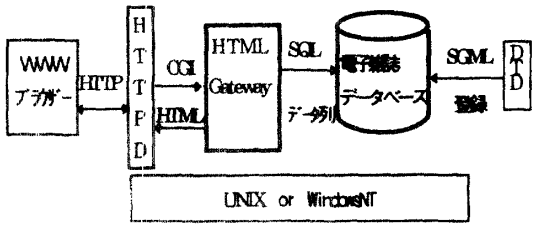

図1　電子雑誌システム構成

しかし現在、雑誌を電子化するにあたっては、雑誌毎や記事毎にDTDが定義されている。その結果、雑誌や出版社、雑誌種別毎にDTDがバラバラになり、雑誌を問わない検索が困難な状況にある。

2.目的

SGMLデータが雑誌毎にバラバラになるという問題に対処するために、汎用的なDTDを提案する。

その結果、ほとんどの雑誌について電子化された記事を雑誌の枠を越えた広い範囲で検索でき、電子雑誌化のメリットを引き出すことができると考えられる。

既存の汎用DTDも存在しているが、検討の結果、表のように、それぞれが特定分野に限った雑誌にのみ有効であることが判った。

DTD名	特徴	適用分野
NCALS general DTD *1	改版時の差分情報を重視	マニュアル向き
Elsevier Science Full Length Article DTD*2	本文情報のみを掲載	論文集向き

本稿では、ある特定分野に閉じない雑誌を対象として汎用化を試みた。また、特に利用面では検索機能を最も重要なものと位置づけた。

3.DTD試作

本検討における第一段階として、NTT社内刊行雑誌をベースとして、この範囲での汎用的DTDを試作した。第二段階として、一般的な雑誌データを試作DTDを用いてSGML化することで、その有効性を検証した。さらにこの試作DTDを基に前述の電子雑誌システムを作成し、システムの観点からその構造の再検討を行った。

3.1.雑誌構成の分析

分析対象として選択した12種類の雑誌について、その記述内容を比較／分類した。

雑誌間ではその表現形態や利用する字句の違いは見られるが、内容の8割近くを共通する要素で占めていた。違いの大半はレイアウトや配色、デザインであり、これは電子化の過程で意味を失うため、一見違う要素であるかに見えても、意味が等しければ同一と判断する。

しかし一部に、以下に示す雑誌固有の項目がある。

表現が違うが、意味が等しい情報の例：
編集と問い合わせ先と発行元、Figと図と図表

雑誌固有の情報の例：対象読者の指示（社内通達文書集）、記事毎の改版記録（規則集）

3.2.文書の構造化

雑誌における構成要素の分析の結果、図2のような骨組みが、各種雑誌に共通する構造であることが判明した。

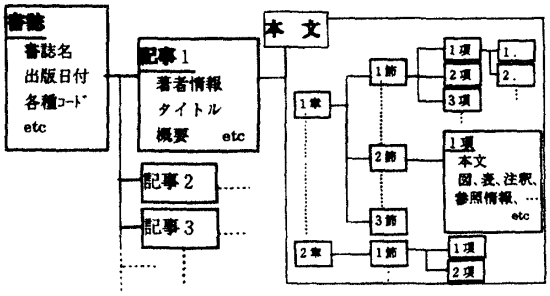

図2　雑誌構造

特定分野には限らないが、上記構造を取っていない雑誌については、今回のDTDは対象外とした。このような例としては、本文が並列な情報の併記に終始し、構造的な性質を持っておらず、記事情報を含まない形態の雑誌が該当する。

A Study of Common DTD for Magazines
Kozue Kai,Kenji Fujimoto,Kazuhiro Sugiyama
NTT Software Headquarters,
NTT Human Interface Laboratories

記事は、雑誌内では順序性は無いが、唯一目次情報によってのみ、その順序性が示される。

記事と段落の関係は、書誌と記事の関係と同様であるが、同時に段落間の順序関係が重要である。

3.3.要素のまとめ方

段落における階層数や記事数は、雑誌によって最大数が異なるため、DTDにおいて最大数を制限しない

第一段階における分析の結果を基に、分析対象雑誌中、一書誌にのみ特有な要素についても、他の雑誌で類似の項目が存在する場合に備え、要素を不足無く表示するために、それぞれに対応するタグを作成した。この時、該当タグは４件あった。

また項目的には同一に括られるものの、雑誌毎に内容が異なる項目も見られた。これについては、大項目配下に予想されうる全小項目をDTDに付加した。著者項目における読み仮名、E-mailアドレス、参照URL等が該当する。すなわちそれぞれのデータを検索対象としやすい状態を保持するために、要素はなるべく独立させた。

4.検証

4.1.他雑誌からの投入

試作DTDに基づき、一般雑誌のSGML化を試みた。

新たな雑誌においても、試作DTDで全データをSGML化することができた。前項で述べた一書誌に特有な要素も４件中１件を再利用できたため、別種雑誌において、同様に使われる可能性があると考えられる。

4.2.作業者の判断が感性に任される項目

試作DTDに従い、データ投入を実施した結果、そのタグの判別が付きにくく、結果として作業者の感性に任されてしまう項目の例として、以下の点が挙げられる。

- 短文併記の章立ての本文表示と、長めの文章によるリスト表示の区別
- 図や囲み記事、表の区別
- 複数の図や表から構成された図の取り扱い

これらは雑誌に閉じていれば扱いが統一されているが、複数の雑誌に適応するがゆえに、雑誌間でその表示上の役割の境界線があいまいになっている。そのため判断時の指針を雑誌毎に作業前にまとめ、作成マニュアルとすることで対処した。

4.3.WWWへの表示の観点での検討

4.3.1.複数雑誌／記事をまとめて表示する場合

WWWブラウザ上では、検索結果は記事や出版社、著者の別なく、検索結果のみが羅列される。

さらに検索結果表示時には、雑誌内に閉じていない情報が羅列されるために、雑誌や紙面に固有な表現に対して、なんらかの処置を行わなければ、読者の混乱および電子化時のリンク情報、表示条件に混乱を生じることも判明した。よって以下の様な作業をSGML化の段階で行うことが望まれる。

- 本文中に出現する「本誌」、「本稿」を雑誌名、記事名に置換
- 参照対象を頁情報からURL等のリンク情報に置換
- 日付表示は本文での表示形式を無視し、同一形態に合せる
- リスト表示時や段落冒頭に使われている記号（・、○、●等）についてはDTD内に形式情報を保持するが、表示方式についてはビューアAPに一任することで、雑誌間の見栄えの統一を図る。

4.3.2.ディスプレイへのレイアウト

ディスプレイの発色と紙面との差を考慮し、さらに雑誌間での差異を最小化するために、レイアウト情報は表示文字の強調表示と斜体字を除いて再現しない。

また、レイアウトやデザイン情報は以下のように扱う。

- レタリングされた文字はイメージ情報として扱う。背景と一体化した文字情報についても同様。また、文字に使用されている色は無視する。強調の意図による文字色については、強調表示により対応
- 記述記号において、外字を利用している場合には、状況に応じて等価な別の文字や記号で置き換える
- 変形が著しくHTMLタグで表現仕切れない表は、状況に応じて図として扱う

5.まとめ

以上のように、一般雑誌に対する汎用的なDTDを作成できた。これを基に電子雑誌システムを構築し、現在、利用している。検証に用いた雑誌以外についてもこのシステムへ登録し、このDTDの有用性を確認し、さらに使いやすくするための改善を行っていく予定である。

参考文献

1）NCALS general DTD
Nippon CALS Research Partnership
URL：http://www.ncals.cif..or.jp

2）ElsevierScience Full Length Article DTD
Elsevier Science ,AMSTERDAM