電子書籍ファイルフォーマットの構造

記述言語と IEC TC100/TA10 の電子書籍フォーマット

小町 祐史*

キーワード　ファイルフォーマット；記述言語；スタイル指定；スキーマ言語；HTML；XML

記述言語とマーク付け

電子書籍に限らず，文書などの構造をもつデータの交換フォーマットには，ASN.1，SGML，XML などの記述言語が用いられてきた。

記述言語は文書構造記述などの特定目的のデータの記述とアクセスを指示する言語であり，指示要素の組合せによってコンピュータの多様な動作を規定するプログラム言語と比較するとき，データがテキスト形式で扱われ，制御変数をもたないことが多いなどの特徴をもつ。目的によって表 1 のように分類される。

マーク付けの一般化

文書処理の電子化は植字機（タイプセッタ）において開始され，印刷指示がタグとして文書データの中に埋め込まれた。タグは機器に依存していたため，それが埋め込まれた文書データの交換性は極めて限定されていた。

そこで文書中に印刷指示を書くのではなく，次のように文書を構成する意味的なまとまり（論理的要素）を示すタグを文書データの中に埋め込む（マーク付けする）ようになった。

<header> 環境保全の強調

<paragraph> パリで開かれていた国際環境保全会議 ...

その結果,

◇マーク付けを機器非依存にできる。

◇マーク付けに用いるタグの可読性が高く，しかも印刷の専門技術に関係しない。

◇したがって文書内容の作成者は意味内容の記述に専念できる。

ことになった。

マーク付けはさらに，ある文書クラスに共通する論理的要素とその構造を識別するようなタグ集合へと一般化され，共通マーク付け（generic markup）と呼ばれた。要素に関する属性記述をもタグに含めて，多様なアプリケーションに対応できるようにしたマーク付けも行われて，一般化マーク付け（generalized markup）となった。

このようなタグ集合の定義方法を国際的に取り決め，言語として体系付けたものが ISO（国際標準化機構）によって承認され，SGML（Standard Generalized Markup Language：標準一般化マーク付け言語）として制定された。これを用いれば，いわゆる文書に限定せず，さまざまなタイプのデータ集合（アプリケーション）に対して，一般化マーク付けを定義でき，さらに各種の補助機

表 1　目的による記述言語の分類

目的	記述言語
文書データ構造の記述	SGML，XML
文書型の記述	DTD，XML Schema
文書スタイルの記述	DSSSL，XML，CSS
文書構造変換の記述	DSSSL，XSLT
特定文書型をもつ文書データ記述	HTML，XHTML，ODF，OOXML

* KOMACHI，Yushi
大阪工業大学
情報科学部　教授
〒 573-0196 大阪府枚方市北山 1-79-1
komachi@y-adagio.com

能によってさらにマーク付けを扱うさいの利便性の向上が図られた。

なお，SGML のいくつかの追加機能は，処理系の進歩によって不要になり，その後開発された XML では簡素化が施された。

スタイル指定

論理構造（論理的要素とその構造）のマーク付けを施された文書データは，表示メディア上にフォーマット付けされて展開される必要があり，そのため論理的要素をどのようにフォーマット付けするかの指示（スタイル指定）を受ける必要がある。

なお，本稿では“フォーマット”を異なる 2 つの意味で用いる。“文書・書籍の交換フォーマット”という文脈におけるフォーマットは，交換対象データを扱う送り手と受け手との間の交換対象データに関する表記方法の取り決めであり，JIS などでは交換様式と書かれることが多い。

もう一方の表示メディア上での“文書データのフォーマット付け”という文脈におけるフォーマットは，文書データを構成する文字列等のまとまりを視覚的に見やすく表示メディア上にマッピングすることである。組版，レイアウト，スタイル付け，ページ展開などが，類似の意味をもつ。

表示機能を大きく異にする装置間での文書交換では，スタイル指定はローカルに設定しなくてはならず，交換の対象は論理的要素とその構造に限定される。しかし，充分なフォーマット付け機能と表示機能をもつ環境では，再編集の可能性を維持したまま交換による版面の一致または最適近似が要求されることが多い。その場合には，文書の論理構造に加えて，論理的要素に対するスタイル指定が交換の対象となる。

CERN（欧州原子核研究機構）における技術文書の交換から始まった HTML は，それが扱う要素型を極端に限定し，それらに対応するスタイル指定をもある程度規定して，SGML 宣言，文書型定義，スタイル指定の交換を不要にすることで，当時の処理系においても，ウェブ環境での軽快なナビゲーション（文書間のたどり）を可能にしてインターネットの普及に貢献した。しかしこの限定された仕様が，とくにフォーマット付けに関する要素型および属性の独自拡張を呼び，交換性が失われることが目立った。World Wide Web Consortium（W3C）は禁欲的なまでに HTML でのフォーマット付け機能を制限し，充実したフォーマット付けに関するユーザー要求はスタイル指定言語 CSS を併用することで充足した。

この戦略によって，SGML の時代から提唱されていた電子化文書，とくにウェブ文書を論理構造とスタイル指定とに分離して記述することが社会に定着していった。

XML とスキーマ

HTML の大普及の結果，当初の HTML のスコープ（適用範囲）を越えた複雑な文書までを，

```
<!DOCTYPE 参加者一覧 [
  <!ELEMENT 参加者一覧 (参加者*)>
    <!ELEMENT 参加者 EMPTY>
    <!ATTLIST 参加者
      氏名 CDATA #REQUIRED
      所属 CDATA #REQUIRED >
]>
```

図 1　スキーマ言語 XML DTD による論理構造の記述

```
  <element name="参加者一覧" xmlns=
"http://relaxng.org/ns/structure/0.9">
    <zeroOrMore>

      <element name="参加者">

        <attribute name="氏名">
          <text/>
        </attribute>

        <attribute name="所属">
          <text/>
        </attribute>

      </element>

    </zeroOrMore>
  </element>
```

図 2　スキーマ言語 RELAX NG XML syntax による論理構造の記述

HTML と同様の簡便さで交換したいという要求が現れた。この要求を満たすため，SGML のサブセットに整形式のコンセプトを導入した XML が開発された。

SGML においては，共通する論理的要素とその構造を定義するスキーマ言語として DTD だけが使われていたが，XML の普及と共に XML の構文を使ったスキーマ言語（W3C XML Schema, RELAX NG XML syntax）が利用可能になり，さらに簡素な記述を可能にする RELAX NG compact syntax が ISO/IEC 19757-2 として制定された。図 1 ～ 3 にそれぞれ XML DTD, RELAX NG XML syntax，RELAX NG compact syntax による論理構造の記述例を，その入れ子構造を図 4 に示す。データ型の規定，XML 名前空間なども次々と開発されて，XML はいわゆる文書だけでなく，一般的なデータの構造を記述する言語として，プロトコルの記述などにも広く利用されている。

電子化文書を論理構造とスタイル指定とに分けて記述することは，XML の利用においても同様であり，CSS をさらに拡張して XML の構文で表記した XSL が開発されて，印刷・出版の文化の中で発達してきた多くのフォーマット付け・組版技術の要素（文書スタイルオブジェクト）がサポートされるに至っている。

```
element 参加者一覧 {
  element 参加者 {
    empty,
    attribute 氏名 { text },
    attribute 所属 { text }
  }*
}
```

図 3　スキーマ言語 RELAX NG compact syntax による論理構造の記述

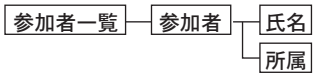

図 4　図 1 ～ 3 の各構文が示す vocaburary（要素と属性）の関係。要素と属性は入れ子構造になっている

IEC TC100 における電子書籍規格の扱い

電子書籍においては，文書としての論理構造とそのコンテンツ（文字列，画像など）だけでなく，フォーマット付けされたページイメージに対しても著作物としての扱いを受けることが多い。そこで電子書籍モデルを示す前にとくにフォーマット付けを論じる。

電子書籍におけるフォーマット付け

人の思いは通常，ことばによって表現され，文字列を使って記述されることが多い。人の思いを時間的に固定して，文字列およびその他の補助データで表現したものが文書であり，他の人（または自分自身）にその思いを伝えることを目的とする。

ことばによって表現される思いには，必ずしも明示的ではないこともあり得るが，意味的な区分（論理構造）があり，その構造を適切に示すことによって思いの伝達が明確になる。思いを文字列で記述するとき，その論理構造をなるべくわかりやすく伝達するために，文字列を展開する表示メディア（紙など）の上で文字列を幾つものブロック（見出し，段落，注釈など）にまとめ，ブロックの境界を空白等で明らかにし，さらにブロックの中での文字の並び方，フォントなどで他のブロックと区別するというフォーマット付けが印刷・出版技術とともに発達した。

著者や編集者は彼らの思いを表示メディアの制約の範囲でなるべく適切に表現できるスタイルオブジェクトを用いて文字列を展開し，読者は紙面に展開された文字列のブロックから著者や編集者の思いをより明確に把握する。紙などのハードコピーによる文書交換においては，表示メディアに展開された文字列のブロックという著者や編集者が意図する論理構造のインスタンス（論理構造に基づく実際の値としてのデータ）があるだけであり，表示メディアの制約の変化への柔軟な対応は困難である。

文書が文字コードの列として表示メディアから独立してはじめて，その文字列に対して記述言語などを用いて論理構造の指定が可能になり，電子化された情報として論理構造が交換可能になる。

電子書籍モデル

文書の論理構造を読者に視覚的に示す技術としてフォーマット付け・組版技術があり，それは前述のとおり表示メディアに依存する。eBook（電子的な書籍）流通系の中では，多様な表示メディアの存在を許容する必要があり，表示メディアに依存しない generic format と表示メディアに依存する reader's format とを用意することが必要である。

そこで IEC 62229（マルチメディア電子出版の概念モデル）が示す TC100 の e-Publishing モデルでは，Data preparer（電子書籍を作成する組織または人。たとえば編集者）と Publisher（電子書籍を発行し，配付する組織または人）との間の交換様式として generic format を規定し，Publisher と Reader（読者）との間の交換様式として reader's format を規定することを推奨している。generic format においては，論理構造を含むだけでなく，reader's format への変換に際してのヒント情報としてのスタイル指定を含む必要がある。

reader's format においては，Reader における表示メディアに依存したスタイル指定が含まれる。そのスタイル指定をフォーマッタ（文章の整形を行うアプリケーション）によって実行した結果を reader's format とすることも可能である。

Author（著者）と Data preparer との間の交換様式として IEC/TS 62229 のモデルに含めた submission format においては，Author がとくに指定することを要求するスタイル指定を含むとともに，Data preparer との間の proofreading（文書校正処理）交換のサポートが望まれる。

```
start = ebook-g
ebook-g = ebook-g-core
        | external "bbebxylog.rnc"
        | external "xmdf.rnc"
```

図 5　IEC62448 における RELAX NG 記述の先頭部分

IEC 62448 の基本構造

IEC/TC100（国際電気標準会議のマルチメディアシステム及び機器に関する技術委員会）の e-Publishing モデルに基づく generic format として，すでに IEC 62448 が発行されている。これは我が国が TC100 固有の加速化手続きを用いて提案した規格であり，当時の e-Publishing の国際マーケットを考慮すると「統一フォーマットの国際的議論が困難である」との判断に基づき，マーケットを拡大しつつあった BBeB Xylog と XMDF のフォーマットを追認するとともに，極めて簡素な e-Book を考慮した g-core というフォーマットを規定している。g-core においては，vocaburary（要素と属性）のセマンティック（データの意味）の厳密な規定は示されておらず，スタイル指定も行っていない。

なお，最近では ISO と IEC の各種手続きが統一化される方向にあり，そのための検討が続けられているが，IEC/TC100 では ISO や ISO/IEC JTC1（ISO/IEC 合同技術委員会）とは異なる標準化手続きが認められており，加速化手続きにも TC100 固有の手続きが用意されて，新規分野の国際規格開発の効率化が図られている。

これらの規定の構文には，RELAX NG compact syntax が用いられ，RELAX NG 記述の先頭部分で図 5 のような，g-core，BBeB，XMDF の選択が行われる。今後，国際的に合意されたフォーマットもこの機構を用いて（この選択肢の追加によって）IEC 62448（電子出版の共通フォーマット）の中に導入することが可能である。■