HTMLに基づいた文書型定義の作成と利用

3S－1

今郷　詔
imago@ic.rdc.ricoh.co.jp
(株) リコー　情報通信研究所

1　はじめに

WWWでの文書交換形式であるHTML[1]の構文はSGML[2]のDTD(Document Type Definition, 文書型定義)として定義されており、HTMLで記述した文書はSGMLのインスタンスでもある。SGMLは文書の論理構造を記述するために使われることが多いが、HTMLは文書のレイアウト構造を記述する手段としてSGMLを利用している。

そのためHTML文書は表示のための最終形であり、そこから情報を抽出したり、別の形式に変換するには適していない。

一方我々は、共通性の高いDTDを元に簡単な指定を行なうことで新たなDTDを生成する技術であるDTD導出体系を開発した[3]。この技術をHTML DTDに適用し、導出したDTDを用いてSGML文書を作成することで、次の利点が期待できる。

- ハイパーリンクなどのHTMLの構成要素を用いてSGML文書作成ができる。
- 作成したSGML文書は自動的にHTMLに変換できる。
- 論理構造を記述しているので多様な利用ができる。

本稿ではHTML DTDから新たなDTDを導出する方法と、そのDTDのインスタンスの利用方法を述べる。

2　DTD導出方法

図1にDTD導出体系の構成を示す。HTML DTDに若干の属性定義を追加したものを拡張ベースDTDとして使用する。

拡張ベースDTDに対して表1に示すようなDTD特性指定を用意しておく。これは拡張ベースDTDにおける章・節に相当する要素型や箇条書きに相当する要素型の名前などを指定したデータである。DTD特性指定をいったん作成しておけば、新たな文書型を導出する時に作成しなければならないデータは文書構造指定だけである。

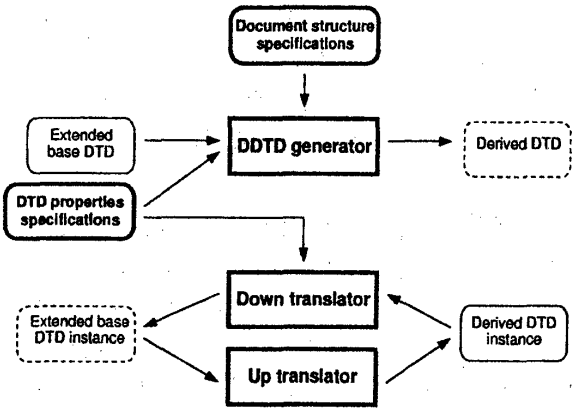

図1: DTD導出体系の構成

特性名	内容
RootGI	HTML
DivGI	DIV
DivTitleGI	H1,H2,H3,H4,H5,H6
ParagraphGI	P
ParagraphTitleGI	STRONG
ListGI	OL
BodyStartString	<TITLE DRVDGI="BASEINTR"> </TITLE>
BodyEndString	

表1: DTD特性指定の例

文書構造指定は図2に示すように特定の文書型の文書構造をインデントを利用して記述したデータである。各行が1つの要素型に対応し、インデントが1段階深い行が後続していれば、それを内容モデルの指定と見なす。出現指示子 (“*”, “?”, “+”) や接続子 (“|”, “&”) の指定も可能である。

なお文書構造指定はベースDTDからは独立であり、ベースDTDの内容をまったく意識せずに記述できる。ベースDTDを変更することがあっても文書構造指定を変える必要はない。

拡張ベースDTD・DTD特性指定・文書構造指定の3種類のデータから、導出DTDを生成でき

Creation and utilization of DTDs based on HTML.
IMAGO Satosi
RICOH Co., Ltd.

```
研究会開催通知
    研究会名
    日時
    会場
        会場名
        交通?
    議題リスト
        議題+
            発表題目
            発表者名
    備考*
```

図 2: 文書構造指定の例

る。導出 DTD は、拡張ベース DTD の内容すべてと、新たに生成した要素宣言・属性定義リスト宣言とからなる。したがって導出 DTD は拡張ベース DTD のスーパーセットに相当し、拡張ベース DTD のインスタンスは導出 DTD のインスタンスにもなる。

3 導出 DTD とそのインスタンスの利用

表 1・図 2に示した文書特性指定と文書構造指定を用いて HTML DTD から導出した DTD は次のようになる (一部省略)。

```
<!element 研究会開催通知 O O (研究会名,
    日時, 会場, 議題リスト, 備考*)>
<!element 議題リスト - - (議題+)>
<!element 議題 - - (発表題目, 発表者名)>
<!element 発表題目 - O (#PCDATA|TT|I|B|
    |CITE|A|IMG|APPLET|FONT|SCRIPT|...)*>
...
```

DTD を直接作成する場合は、個々の末端要素の内容モデルまですべて記述しなければならないが、DTD 導出体系を使うと末端要素の内容モデルはベース DTD の対応する要素型から自動的に補われるので、人間が一々記述する必要がない。

上の例では“発表題目”の内容モデルは、HTML の“P”要素型と等しくなっている。従ってハイパーリンクやアプレットの記述、イメージの指定、フォントの指定など HTML の機能を自由に使って内容を記述できる。

この DTD のインスタンスは例えば次のようになる。HTML で記述した場合とは異なり、タグによって論理構造が表現されているので、必要項目のみを取り出してデータベース化するなど、多目的に利用することができる。

```
<研究会開催通知>
  <研究会名>情報学基礎研究会
  <日時>1996 年 7 月 26 日
  <会場><会場名>奈良先端科学技術大学院大学
  </会場>
  <議題リスト>
    <議題><発表題目>SGML における DTD 導出体系
          <発表者名>今郷韶</議題>
    <議題><発表題目>題目 2
    ...
```

このインスタンスは、図 1の down translator によって HTML に変換することができる。このときに余計なパラメータやプログラミングは必要ない。また変換した HTML インスタンスは up translator によって元の SGML インスタンスに戻すこともできる。

通常の SGML インスタンスは、見やすく表示するためには DTD の内容に依存したコンバータを作成して HTML や PostScript などに変換したり、SGML ブラウザに対するスタイルシートを作成する必要がある。しかし DTD 導出体系を用いて HTML DTD から導出した DTD のインスタンスは、新たにプログラミングやスタイル指定を作成することなく、WWW ブラウザで表示できる。

4 おわりに

我々が開発した DTD 導出体系を用いて HTML DTD から新たな DTD を導出する方法を述べた。導出した DTD を用いて SGML 文書を記述することで、HTML では不可能であった論理構造の記述が可能となるとともに、HTML の構成要素を用いて SGML 文書が記述できるようになった。この方法で作成した SGML 文書は自動的に HTML に変換できるので、WWW ブラウザで表示することもできる。

今後は DTD 導出体系のベース DTD を HTML DTD とした場合に有用な付加機能・拡張機能を検討していきたい。

参考文献

[1] World Wide Web Consortium. Introducing html 3.2. http://www.w3.org/pub/WWW/MarkUp/Wilbur/, 1996.

[2] Liora Alschuler. *ABCD... SGML: A User's Guide to Structured Information.* International Thomson Computer Press, Boston, 1995.

[3] 今郷 韶, 西村美苗. SGML における DTD 導出体系. 情報処理学会情報学基礎研究会, 42, 7 1996.