

標準化された文書記述言語 (XML, XSL 等) に基づく 多言語文書交換

小町 祐史

松下電送システム(株)

2001-11-07

1. はじめに

1.0 符号化文字の次の課題

アジアにおける符号化文字列に関する標準化作業は、

- ISO/IEC JTC1/SC2 および関連する参加国
- AFSIT および MLIT
- 他の関連グループ

等の活動によって概ね完了し、文字関連の標準化活動は次の段階を迎えようとしている。

1.1 文字列の表示スタイル

複数の符号化文字が集まって配列すると、それらは、そのセマンティクスに応じて、文、段落、節などを構成する。符号化文字のこの集まり(論理要素と呼ばれることが多い)は、電子文書を構成する。符号化文字のこの集まりは、そのセマンティクスの理解を容易にするために、レンダリングを施されて、特定の表示スタイルで表示される。

この表示スタイルは、固有の意味をもち、電子文書の交換に際しても保存されることが望まれる。そこでここでは、電子文書の表示における文字列の表示スタイルの重要性に着目する。

1.2 表示スタイルの課題

表示スタイルは、それぞれの国、地域、会社(出版社、新聞社)などの文化的背景に基づいて、これまでの印刷技術の中で開発されてきた。したがって、表示スタイルの調査研究に際しては、文字の調査研究に際してと同様に、文化的背景に慎重でなければならない。

表示スタイルを調べるときに直面する課題は、それぞれの国、地域、会社の表示スタイルに関するオーソライズされた参照または出版された文献が極めて少ないことである。その僅かな例として、従来の欧米の文書に関する Oxford rule, Chicago rule などがある。

もう一つの課題は、ウェブ文書に関する表示スタイルである。これについては、望ましいスタイルとして確立したものがまだ存在しない。

1.3 電子文書

表示スタイルに言及する準備として、表示スタイルが適用される対象としての電子文書の扱われ方、要件および構造を概観し、ここでの着目対象を確認する。

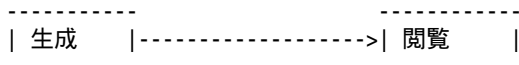
現在、数多くの電子文書のフォーマットが使われている。その例を次に示す。

- word form, RTF
- PDF, PS
- HTML

- XML, SGML

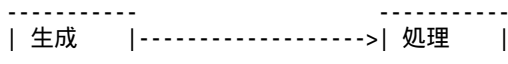
1.4 電子文書の扱われ方

(1) 処理形式1



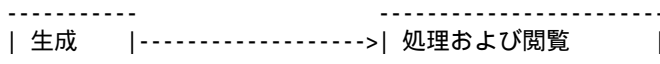
生成された文書は，転送・交換された後，受け手の画面または紙面で閲覧される。受け手が編集を行うこともある。

(2) 処理形式2



生成された文書は，転送・交換された後，受け手によって処理される。人が閲覧することはない。例えば，CADシステム。

(3) 処理形式3



生成された文書は，転送・交換された後，受け手によって処理され，閲覧される。処理には，変換，データマージ，データベースとの関連付けなどがある。

1.5 電子文書に対する要件および[文書フォーマット]

(1) 処理形式1

- セマンティクスの理解を容易にするために，表示スタイルは必要である。
- 明示的な論理構造(論理要素の構造)は必要ない。
- [編集可能形式: word form, RTF] [最終形式: PDF, PS]

(2) 処理形式2

- 明示的な論理構造が，処理のために必要である。
- セマンティクスの理解を容易にするための表示スタイルは必要ない。
- [XML, SGML]

(3) 処理形式3

- 明示的な論理構造が，処理のために必要である。
- セマンティクスの理解を容易にするための表示スタイルも必要である。表示スタイルは，対応する論理要素に関係付けられなければならない。
- EC, e-Governmentなどは，この処理形式3の電子文書を扱う。
- [論理構造: XML, SGML] [スタイル指定: CSS, XSL, DSSSL]
- [簡単な論理構造: HTML]

備考: HTMLはSGMLのインスタンス(XHTMLはXMLのインスタンス)である。スタイル指定は，HTMLの各要素に関して既に規定されている。したがって，簡単な構造をもつXML文書は，多くの場合，XSLTによってHTML文書に変換して表示できる。

1.6 処理形式3

ここでは、処理形式3の電子文書に着目し、その文書の表示スタイル指定について論じるために、次の課題を取り上げる。

- 論理構造記述の概要
- 論理要素に対するスタイル指定
- 表示スタイルの典型的な集合およびスタイル指定のライブラリ

2. 論理構造記述

2.0 論理構造

文書の主要部分は文字列から成り、文字列はそのセマンティクスに応じて組合わされて、標題(title)、節(clause)、段落(paragraph)などの要素を構成する。セマンティクスと要素構成とは、通常、文書作成者の意図に基づく。

簡単な文書は、例えば、次の要素から成る。

```
-- title --
-- author --
-- abstract --
-- heading of clause 1 --
---contents of clause 1 -----
|      -- paragraph 1 --      |
|      -- paragraph 2 --      |
-----
-- heading of clause 2 --
---contents of clause 2 -----
|      -- paragraph 1 --      |
|      -- paragraph 2 --      |
-----
```

2.1 論理構造の記述

文書の論理構造は、既存のマーク付け言語、例えば、SGML(標準一般化マーク付け言語)、XML(拡張可能なマーク付け言語)によって記述される。ここで、要素の構成は、DTD(文書型定義)によって規定され、実際の文書インスタンスは、DTDが定義する要素型をもつタグによってマーク付けされる。

HTMLは、決まったDTDをもつSGMLインスタンスである。そこで、HTMLは、簡単な文書の論理構造の記述に使用できる。

2.2 論理構造記述のための国際規格

- SGML
 - ISO 8879:1986, Standard Generalized Markup Language (SGML), 1986-10
 - Amendment 1 to ISO 8879:1986, 1988-07
 - Technical Corrigendum 1 to ISO 8879:1986, Extended naming rules, 1996-12
 - Technical Corrigendum 2 to ISO 8879:1986, Web SGML, 1999-11
- XML
 - W3C Rec., Extensible Markup Language (XML) 1.0, 1998-02
 - W3C Rec., Extensible Markup Language (XML) 1.0 (Second Edition), 2000-10
- HTML
 - RFC 1866, Hypertext Markup Language - 2.0, 1995-11
 - W3C Rec., HTML 3.2 Reference Specification, 1997-01
 - W3C Rec., HTML 4.0 Specification, 1998-04
 - W3C Rec., HTML 4.01 Specification, 1999-12
 - ISO/IEC 15445:2000, HyperText Markup Language (HTML), 2000-05
- XHTML
 - W3C Rec., XHTML 1.0: The Extensible HyperText Markup Language - A Reformulation of HTML 4 in XML 1.0, 2000-01

- W3C Rec., XHTML Basic, 2000-12
- W3C Rec., Modularization of XHTML, 2001-04
- W3C Rec., XHTML 1.1 - Module-based XHTML, 2001-05

3. 論理要素に対するスタイル指定

3.0 スタイル指定の処理

スタイル指定言語は、論理要素に対するスタイル属性の適用を記述する。そのために、次の機能を実行する。

- (1) SGML/XML文書のノード木の中で論理要素を指定する。
- (2) その論理要素に対するスタイル属性を指定する。
- (0) レンダリングに適したノード木変換が、(1)の前に実行されることがある。

3.1 標準化されたスタイル指定言語の処理機能

言語	機能(0)	機能(1)	機能(2)	備考
DSSSL ¹⁾	Y	Y	Y	SGML文書およびXML文書を対象とする。
XSL ²⁾	N	Y	Y	XML文書を対象とする。
XSLT ³⁾	Y	N	N	XML文書を対象とする。
CSS ⁴⁾	N	Y	Y	ウェブ文書対象の簡単なスタイル指定。

- 1) Document Style Semantics and Specification Language(文書スタイル意味指定言語)
- 2) Extensible Stylesheet Language(拡張可能なスタイルシート言語)
- 3) XSL Transformations(XSL変換)
- 4) Cascading Style Sheets(段階スタイルシート)

3.2 DSSSLの処理モデル

DSSSLの処理モデルを図1に示す。

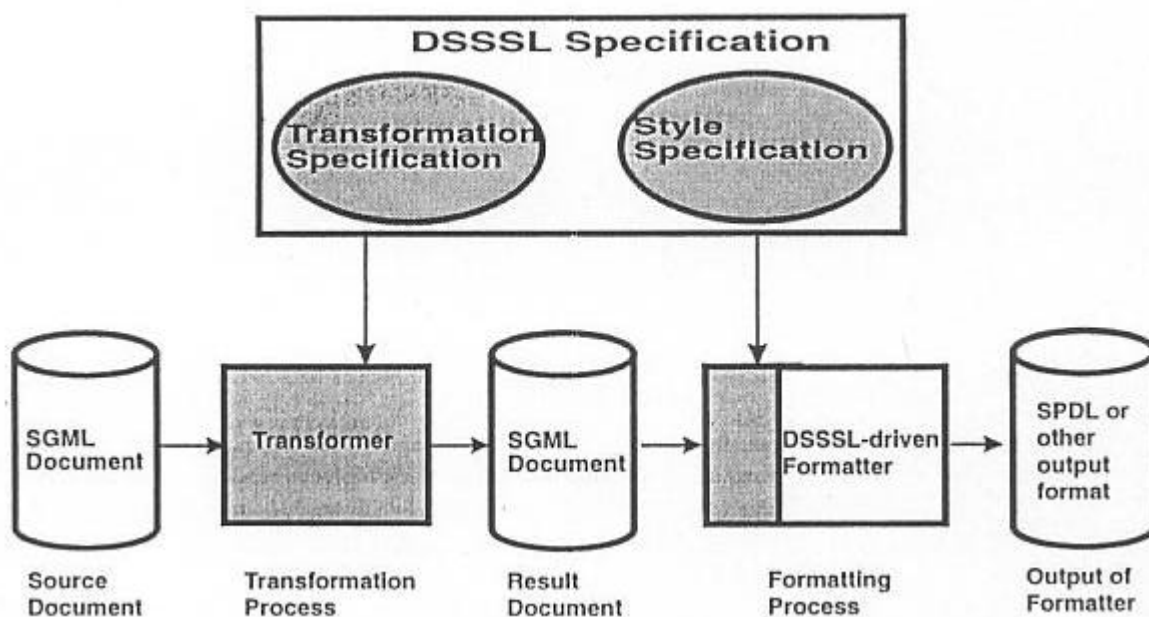


図1 DSSSLの処理モデル

3.3 CSSによるスタイル指定記述

スタイルシートは、規則の集まりであり、規則は、次の記法で記述される。
選択子 {特性: 値}

ここで、
選択枝は、要素を識別する。
{宣言}は、その要素に特性を指定する。

次のグループ化が可能である。
各選択子は、","によって分離される。
各宣言は、";"によって分離される。

次の実装が可能。
(1) リンク付きスタイルシート
(2) 埋込みスタイルシート
(3) 行内スタイルシート

3.4 CSSによるスタイル指定の例

スタイル指定記述例

```
body{
  color: black;
  font-family: helvetica, sans-serif;
  background: white;
  margin: 2em
}
h3 {
  margin-left: 1em;
  font-size: 95%
}
h4 {
  text-align: center;
  font-style: italic
}
p {
  background: yellow
}
```

3.5 スタイル指定記述のための国際規格

- DSSSL
 - ISO/IEC 10179:1996, Document Style Semantics and Specification Language (DSSSL), 1996-04
 - Technical Corrigendum 1 to ISO/IEC 10179:1996, 2001-03
 - SC34 N216, PDAM1 to ISO/IEC 10179:1996, Extensions to DSSSL, 2001-05
- XSL
 - W3C Rec., Extensible Stylesheet Language (XSL) Version 1.0, 2001-10
 - W3C Rec., XSL Transformations (XSLT) Version 1.0, 1999-11
- CSS
 - W3C Rec., Cascading Style Sheets (CSS1) level 1, 1996-12
 - W3C Rec., Cascading Style Sheets, level 2 (CSS2) Specification, 1998-05
 - W3C CR., CSS Mobile Profile 1.0, 2001-10

4. 表示スタイルの典型的な集合

4.0 表示スタイルの現状

論理構造およびスタイル指定のための言語は、既に国際的に承認され、実際に実装されている。しかしそれは、誰もが自分の文書の表示スタイルを記述できることを意味するわけではない。

1.2に示すとおり、表示スタイルについては十分な参照文書がない。国、文書環境の文化的背景を考慮して、表示スタイルに関する参照文書を作成することが望まれる。

スタイル指定は、誰もが記述するにはあまりにも複雑である(3.4は、極めて簡単な例)。この問題に対する一つの解が、スタイル指定のライブラリである。

ここでは、日本で開発された標準情報 TR X 0010の概要を示す。その英語版(DSSSL library for complex compositions)は、ISO/IEC JTC1に提出され、ISO/IEC TR 19758としてDTR(draft technical report)投票を受けている。

4.1 DSSSLライブラリ (ISO/IEC TR 19758)

[1. 適用範囲]

このTRは、SGMLまたはXMLで記述された文書に対してスタイルを指定できるDSSSLライブラリを提供する。このライブラリは、DSSSLに関する特別な知識、組版規則に関する特別な知識をもたなくても、SGMLまたはXMLで記述された文書のDSSSL指定を記述することを容易にする。

備考：このTRの開発を開始した当時、XSLの原案はまだ発行されていなかった。CSSは、このTRの利用者要求を充足するには不十分である。

[2. 引用規定] および [3. 定義]

ISO/IEC指針に従って、幾つかの引用規定およびこのTRで用いる用語の定義を与える。

[4. フォーマット化オブジェクトおよびフォーマット化属性]

通常の出版物に用いられる主要なおよび比較的複雑なフォーマット化オブジェクトおよびフォーマット化属性(つまり、表示スタイル)が整理され定義される。

その項目を次に示す。

- 4.1 用紙サイズ(Paper size)
- 4.2 用紙の向き(Paper placement)
- 4.3 単位(Unit)
- 4.4 基本組体裁(Basic composition style)
- 4.5 基本組体裁モデル(Model of basic composition style)
- 4.6 フォント(Font)
- 4.7 文字サイズの単位(Unit of character size)
- 4.8 柱(Headline)
- 4.9 ノンブル(Page number)
- 4.10 注(Note)
- 4.11 割注(Inlinenote)
- 4.12 圏点(Emphasizing mark)
- 4.13 添え字[Superscript/Subscript (Superior/Inferior)]
- 4.14 字取り(Word-length adjustment)
- 4.15 字割り(Character space adjustment)
- 4.16 節(Clause)
- 4.17 箇条(List)
- 4.18 表(Table)
- 4.19 見出し(Heading)
- 4.20 ルビ(Ruby)
- 4.21 段落字下げ(Paragraph indentation)
- 4.22 スコア(Score)
- 4.23 罫(Rule)
- 4.24 行内行(Inline)

[5. DSSSLライブラリ]

DSSSLライブラリは、複雑な組版に求められるDSSSL指定の記述を容易にする。ここでは、DSSSLライブラリの構成が示され、ライブラリファイルの内容が後続の節で、次の順に規定される。

- 6. 詳細パラメタ生成プログラム
- 7. 関数群
- 8. ページモデル群
- 9. フローオブジェクト構成規則

DSSSLライブラリの構成および処理フローを図2に示す。このライブラリの動作には、SchemeプロセサおよびDSSSLプロセサの利用が必要である。

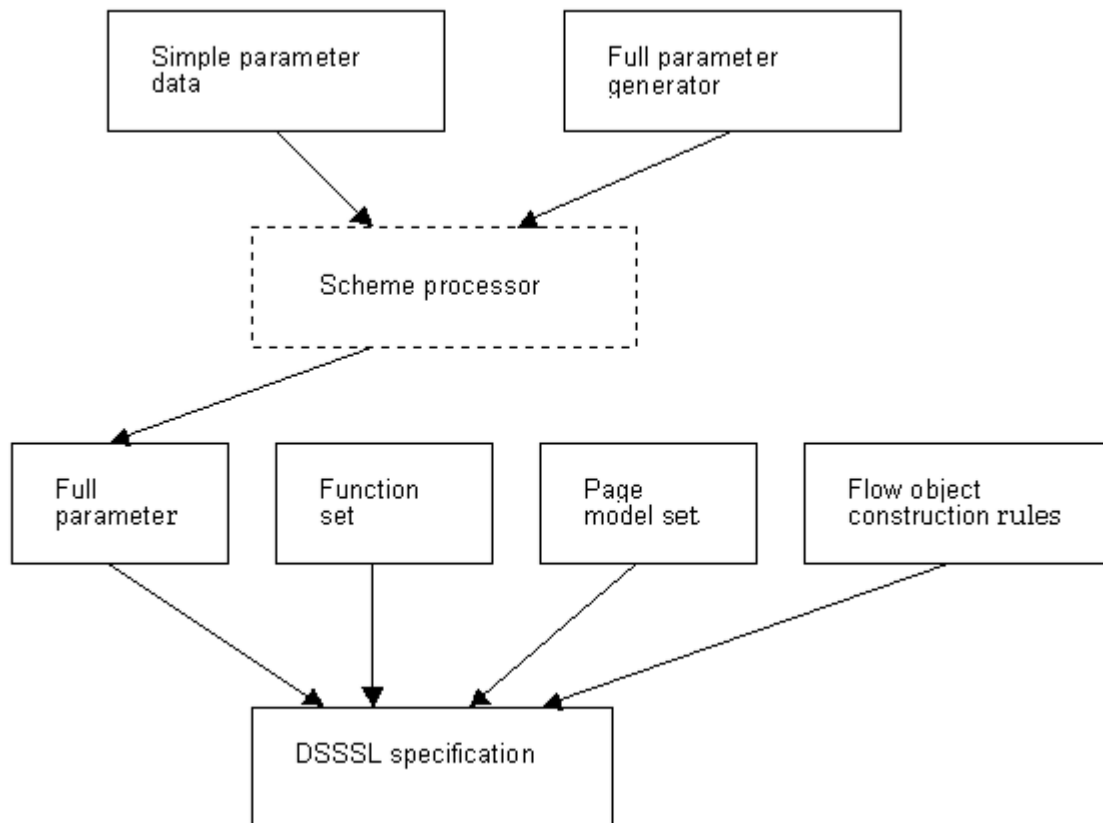


図2 DSSSLライブラリの構成および処理フロー

5. 表示スタイル指定に関する拡張要求

5.0 適用範囲の拡張

このTRの適用範囲は、必ずしもすべての表示スタイルを網羅しているわけではない。例えば、それは次の文書の表示スタイルをサポートしていない。

- 古い文書
- 日本語および英語以外の言語によって書かれた文書
- 多くの言語を含む文書(多言語文書)
- etc.

5.1 XSLライブラリおよびXSLTライブラリ

XSLのPR(Proposed Recommendation)およびCR(Candidate Recommendation)が公表された後、XSLの多くの実装または実装プランがアナウンスされている。これらのウェブ技術動向に応じて、今後はXSLライブラリおよびXSLTライブラリの開発が望まれる。

6. 作業課題案

結びとして、スタイルを保存した多言語電子文書の交換に必要な表示スタイルに関して、幾つかの新作業課題を次のとおり提案する。

- 各国で実際に使われている表示スタイルについて調査し，参照文書を開発する。
- 多言語文書に必要な表示スタイルを調査研究する。
- DSSSLおよびXSL/XSLTに基づくスタイル指定ライブラリを開発する。

これらの課題の成果は，例えば "Style Specification Library for Multilingual Compositions" という標題の国際規格案又はTR案としてまとめ，ISO/IECに提案することが望まれる。