1S-1

# HTMLの文書構造と表示制御を分離するコンバータの開発

松岡 幸典　　志田 晃一郎　　横山 孝典
武蔵工業大学

## 1　はじめに

Web上に公開する文書の標準形式となっているHTML(HyperText Markup Language)[1]は，その策定組織であるW3C (World Wide Web Consortium)によって仕様改定が続けられてきた．その経過の中で表示上の見栄えを設定する記述が策定されたが，現在W3Cはその機能には問題があるとして使用を非推奨とし，徐々に言語仕様から削除していく方向へ向かっている．

HTMLの表示制御機能の問題は，ユーザがその制御を自由に行うことが出来ない点にある．使いようによってアクセシビリティを大きく低下させる要因となる．

上記の問題の解決のためW3Cは，HTMLの記法を用いずに表示制御を行うための言語としてCSS (Cascading StyleSheet)[2]を勧告した．外部CSSを用いて表示制御を行えば，図1に示すようにユーザが任意にCSSを切り替えることによって表示の変更が容易になる．

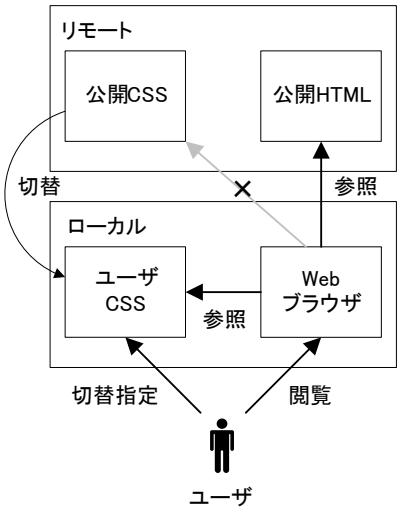


図1　ユーザによるCSSの切替

続いてW3Cは，CSSの有効な用い方やHTMLの記述についてのガイドライン(WCAG : Web Content Accessibility Guidelines 等)を策定し，それを基準とするHTMLやCSSの評価ツールを公開した．このようなHTMLの検査ツールは現在さまざまな組織によって開発，公開されている．

また，HTMLの文法やアクセシビリティの検査を行い，部分的な修正を行うツールTidy[3]がある．TidyはHTMLの文法的な誤りや推奨されない記述のうち，修正方法が自明なものについての自動修正を行うツールである．しかし，HTMLの表示制御の記述をCSSに置き換える機能は限定的であり，一部のものにしか対応していない．

そこで本研究では，表示制御に用いられるW3C非推奨タグ・属性について，Tidy未対応部分を含めCSSの代替記述に自動で置き換えるコンバータの作成を行う．

## 2　ツール概要

本研究で試作するコンバータでは，表1に示すタグ・属性をCSSの代替記述に置き換える．これらは，W3Cが非推奨とするタグ・属性のうち，同等の表示をするための代替記述が一意に特定できるものである．非推奨であるタグや属性の種類は，HTMLの仕様を定義するDTD[4]に指定されている．

本研究では，W3C非推奨のタグや属性であっても，その代替となる記述方法に主要なWebブラウザが対応していない場合は，変換を行わない仕様とした．

表1　主要な対応タグ・属性

| タグ | 属性 | Tidyの対応 |
|---|---|---|
| BODY,TABLE | bgcolor | |
| BODY | background | |
| BODY | text | |
| BODY | link | |
| BODY | alink | |
| BODY | vlink | |
| CENTER | (タグ全体) | ○ |
| FONT | (タグ全体) | ○ |
| TABLE,TH,TD, | align | |
| IMG,HR,P | align | |
| TH,TD | height | |
| TH,TD | width | |
| TH,TD | bgcolor | |
| IMG | border | |
| IMG | vspace | |
| IMG | hspace | |

HTML-CSS Converter for dividing document structure from presentation control
Yukinori Matsuoka, Koichiro Shida and Takanori Yokoyama, Musashi Institute of Technology

また，本研究では変換対象は文法に違反のないHTMLに限る．

## 3　実装

本ツールの処理手順は

1) HTML構文解析
2) 非推奨記述検知
3) 非推奨記述をHTMLからCSSへ変換
4) 非推奨記述をHTMLから削除または置換
5) 変換後HTML・CSS出力

となる．

1)の機能については，本ツールではHTMLの構造に関して行う処理は多くないので簡易的な木構造を構築するまでに留めた．また，この処理を行う過程で同時に2)の処理を行うことが可能である．

3)の処理では，HTMLとCSSの文法が異なるために単純なファイル間の文字列移動では不足である．変換の際，読み込んだタグの名前や属性を一度CSSを概念的に表したクラスとして表現し，そこからCSSの文字列を生成する方式をとった．

4)の機能について，本ツールでは変換元のHTMLに対して無用な編集を行ってしまうことを避けるために，HTML要素のデータ構造から文字列を生成しなおすのではなく，元のHTML文字列に対してテキスト処理を施すことでタグや属性の削除・置換を行う．

5)の機能については，本ツールではCSSの記述形態の中で，HTMLの外部ファイルとしてCSSを参照するようにしている．そのため，HTMLのHEAD要素内に，参照するCSSファイルを指定するためのリンク記述を追加している．Tidyでは，変換したCSSは検査したHTMLの内部に埋め込む形で出力しているが，複数のHTMLに同一の表示制御を適用させる場合などにおいては，外部ファイルに記述するほうが利便性が高い．

## 4　評価実験

サンプルデータは，実際に公開されているWebサイトのトップページ数点と，極端な例としてHTMLの表示制御のみを用いてレイアウトを行ったHTMLを用いた．これらのHTMLから，変換を行う前のもの，Tidyによる変換を行ったもの，本ツールによる変換を行ったもの，Tidyと本ツールを組み合わせて変換したものの複数のパターンを既存の検査ツールにかけて，その結果から本ツールの行う変換の有効性を確認した．

また，変換を行う前と後のHTMLをそれぞれWebブラウザで表示した際に表示崩れが起こらないことについても確認した．

## 5　今後の課題

現在は置き換えた後のCSSの個々のスタイルの名前（セレクタ）に意味の無い固定の文字列を当てている．これではCSSの利便性を完全に生かせているとは言えない．この点についての改良を進めたい．

また，HTMLの次世代フォーマットとして，XMLに準拠したXHTMLが勧告され徐々に浸透している．それに伴い，スタイルシートもCSSに限らずXSLが使用できるようになった．これらの新しい規格への対応も考慮していく．

## 参考文献

[1] "HTML 4.01 Specification", W3C, http://www.w3.org/TR/html401/

[2] "Cascading Style Sheets, level 2 CSS2 Specification", W3C, http://www.w3.org/TR/REC-CSS2/

[3] "Clean up your Web pages with HTML TIDY", Dave Raggett, http://www.w3.org/People/Raggett/tidy/

[4] "HTML 4 Document Type Definition", W3C, http://www.w3.org/TR/html401/loose.dtd