日本語ウェブ文書を対象とした10年間の実態調査（2002年〜2012年）

松本　章代† 　今村　真浩†† 　小西　達裕†† 　高木　朗††† 　小山　照夫††††
三宅　芳雄††††† 伊東　幸宏††

† 東北学院大学　〒 981–3193 宮城県仙台市泉区天神沢 2-1-1
†† 静岡大学　〒 432–8011 静岡県浜松市城北 3-5-1
††† 言語情報処理研究所　〒 192–0919 東京都八王子市七国 3-1-23
†††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2-1-2
††††† 放送大学　〒 261–8586 千葉県千葉市美浜区若葉 2-11
E-mail: †akiyo@izcc.tohoku-gakuin.ac.jp

あらまし　日本語のウェブ文書およそ 1 万ページ/年を対象として，(1) (X)HTML のバージョン，(2) 使用文字コード，(3) スタイル（装飾）指定方法，(4) レイアウト制御方法，(5) フレーム，(6) 動的コンテンツ，について実態調査を行ったので報告する．各調査項目について，ここ 10 年間における変化を観察し，考察する．

キーワード　ウェブ文書，HTML，CSS，スタイルシート

1. まえがき

HTML の誕生以来，その規格は変化し続け，現在に至っている．HTML4 が W3C によって勧告されたのは 1997 年のことである．HTML4 では，文書の体裁（レイアウトや装飾）をタグで記述するのではなく，スタイルシート（CSS）によって指定することが推奨され，色・大きさ・配置といった装飾・レイアウト関連の要素や属性の使用は非推奨と位置づけられた．また，ページ全体のレイアウトは，それまで `table` タグを用いて行う手法が非常にポピュラーであったが，これも文書の構造と体裁を分離されるという理念に反するため，CSS で指定する（代替する）ことが求められるようになった．しかしながら，勧告直後のウェブブラウザは，スタイルシートが十分にサポートされていない・表示速度が遅いなどの問題があったこともあり，これらの体裁の指定が，すぐに完全に CSS に置き換わることはなかった．ただし最近は，HTML5 の勧告が近づいてきたことやスマートフォンの普及を受けて HTML5 への移行が進むなど，HTML/XHTML の記述が改めて見直される傾向にあるようである．HTML タグ（非推奨要素・非推奨属性）で記述された装飾の指定を CSS に自動変換する研究 [1] も行われている．

我々はこれまで，日本語ウェブ文書を対象とした研究を行ってきており，2002 年より日本語ウェブ文書を収集し，保管している．我々は，このウェブ文書およそ 1 万ページ/年を対象として，ウェブ文書の作られ方がここ 10 年でどのように変化してきているかについて実態調査を行う．調査項目は，(1) (X)HTML のバージョン，(2) 使用文字コード，(3) スタイル（装飾）指定方法，(4) レイアウト制御方法，(5) フレーム，(6) 動的コンテンツ，とする．2 節で調査手法，3 節で調査結果について述べ，4 節でまとめる．

2. 調査手法

2.1 ウェブ文書の収集

ウェブ文書の収集は，2002 年から 2 年間隔で行っており，2012 年で 6 回目である．

2012 年は，以下の手順で収集作業を行った．なお，2012 年以前もほぼ同様の手法でそれぞれ 1 万ページの収集を行っている．

（1） Yahoo! JAPAN のカテゴリのトップページ（`http://dir.yahoo.co.jp/`）からリンクを 4 階層たどる．ただし，そのままでは yahoo.co.jp ドメインがかなりの割合を占めてしまう．それを防ぐため，3 階層たどったら，yahoo.co.jp ドメインの URL を一度すべて取り除き，残った URL のページを取得し，URL を抽出する．

（2） ここまで行うとおよそ 10 万 URL が収集される．ここから 1 万 URL をランダムに抽出しダウンロードを行う．

（3） (X)HTML 内で指定された外部 CSS についてもダウンロードする．

（4） 日本語以外の言語で書かれたファイルやバイナリファイルを取り除く．

2.2 調査項目

次の各項目について，2002 年から 2012 年まで 2 年おきに収集したデータを用いて，その実態の変遷を確認する．

2.2.1　(X)HTML のバージョン

採用されている (X)HTML のバージョンを調査する．

2.2.2　使用文字コード

使用されている文字コードを調査する．

2.2.3　スタイル（装飾）指定方法

HTML4.01 における非推奨要素・非推奨属性と CSS（外部・内部・style 属性）の使用率を確認する．

CSS は記述の方法によって大きく 3 つに分類できる.「外部 CSS」とはスタイルの記述を外部ファイルにまとめて HTML ファイルからそのファイルを参照する手法,「内部 CSS」とは HTML 中の `style` タグ中にスタイル指定する手法,「style 属性」とは HTML ファイルの各タグの style 属性の値として直接スタイルを記述する手法である. 3 つの中でもっとも望ましいのは, 独立性が最も高く再利用もし易い「外部 CSS」である. 一方,「style 属性」は構造とスタイルの分離ができていないという点で, あまり使用すべきではない, とされている.

さらに, 非推奨要素・非推奨属性の内訳を調査する. なお, 属性によっては, 特定の要素との組み合わせのみ非推奨となるものもあるため, 配慮する必要がある.

2.2.4 レイアウト制御方法

レイアウト制御（コンテンツの配置）は, 本来 CSS で行うべきであるが, 以前は `table` タグが頻繁に用いられていた. 本来の表を表現する目的で使われている `table` タグとレイアウト制御の目的で使われている `table` タグとが混在していると, アクセシビリティや情報の再利用の面において問題がある. そこで, これらを自動分類する研究 [2] [3] やレイアウト制御目的の `table` タグを CSS に置き換える研究 [4] も行われている.

本研究では, 各ページのレイアウト制御が, CSS によって行われているか, `table` タグによって行われているかについて, ここ 10 年間の状況を確認する.

2.2.5 フレーム

かつては, フレームで分割されたウェブページをよく見かけたが,「フレームはアクセシビリティに反する」として次第に用いられなくなり, HTML5 ではついに, `frameset`, `frame`, `noframes` は廃止された. 一方, インラインフレームは, HTML4.01 Strict 及び XHTML1.0 Strict では未定義（Transitional, Frameset では定義済み）, XHTML1.1 では廃止されているものの, HTML5 では採用となった.

このような状況の中で, フレーム（`frameset` タグや `iframe` タグ）を利用しているページの割合がどのように変化しているのかを調査する.

2.2.6 動的コンテンツ

動的コンテンツがどのような技術を用いて制作されるかは, その時々の流行り廃りによって変化すると思われる.

そこで, 動的なコンテンツを含むページの割合を各技術ごとに求める. 今回は, HTML ファイルの中を確認するだけで容易に集計できる CGI・JavaScript・Flash・Java Applet・Silverlight の使用率について調査を行うこととする.

CGI については `form` タグの `action` 属性があるページの数, JavaScript については `script` タグで `text/javascript` 指定がある, または `a` タグに `href="javascript:`とあるページの数を集計する. 一方, Flash・Java Applet・Silverlight は `object` タグ（＋ `param` タグ）によって埋め込むことができるので, `object` タグの `classid` 属性や `type` 属性で確認する. ただし Flash は `embed` タグ, Java Applet は `applet` タグを利用する方法もあるため, 併せて集計する. なお, `embed` タグは HTML 4.01 では定義されておらず, `embed` タグの代わりに `object` タグを使用することが推奨されている. `applet` タグは HTML5 では廃止されており, `applet` タグの代わりに `object` タグを使用することが推奨されている.

3. 調 査 結 果

3.1 (X)HTML のバージョン

各ウェブページの DTD から (X)HTML のバージョンを確認する. 表 1・図 1 中の H は HTML, X は XHTML を表す.

表 1 (X)HTML のバージョン

	2002	2004	2006	2008	2010	2012
H2.0	0.8%	1.2%	1.2%	1.5%	0.1%	0.0%
H3.2	2.2%	1.3%	0.4%	1.0%	0.1%	0.1%
H4.0	20.1%	7.3%	3.7%	2.2%	1.6%	0.5%
H4.01	17.7%	29.2%	32.0%	36.8%	30.4%	17.3%
H5	0.0%	0.0%	0.0%	0.0%	0.5%	5.0%
X1.0	0.2%	4.5%	26.5%	38.3%	48.6%	64.8%
X1.1	0.1%	0.4%	0.8%	0.8%	1.2%	0.4%
その他	0.9%	0.6%	0.5%	0.3%	0.7%	0.3%
指定無	58.1%	55.4%	34.8%	19.2%	16.9%	11.5%

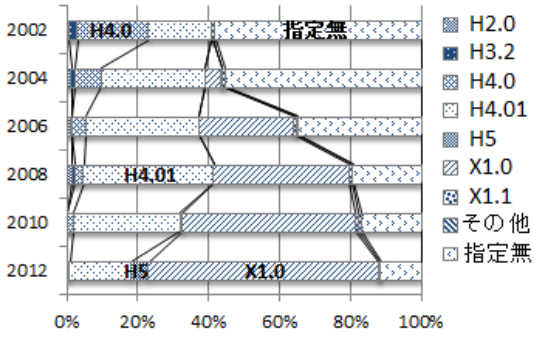

図 1 (X)HTML の変遷（グラフ）

3.2 使用文字コード

各ウェブページから `meta` タグの文字コード指定部分を抽出する. 集計結果を表 2・図 2 に示す.

表 2 使用文字コード

	2002	2004	2006	2008	2010	2012
SJIS	67.1%	63.5%	50.9%	47.3%	40.3%	22.9%
EUC	8.3%	12.6%	14.7%	15.1%	15.6%	16.6%
JIS	2.2%	1.4%	1.0%	0.7%	0.3%	0.1%
UTF8	0.1%	3.5%	21.5%	27.7%	38.2%	56.0%
設定無	22.4%	19.0%	12.0%	9.2%	5.6%	4.4%

3.3 スタイル指定方法

まず, 次の (a)〜(f) に該当するページの割合を調査する.

(a) 非推奨要素
(b) 非推奨属性
(c) 外部 CSS
(d) 内部 CSS
(e) Style 属性
(f) (c)(d)(e) のいずれかを使用

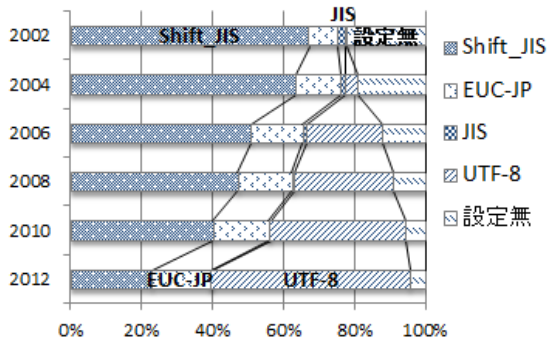

図 2　文字コードの変遷（グラフ）

表 3　スタイル指定方法

	2002	2004	2006	2008	2010	2012
(a)	85.8%	80.1%	64.9%	49.1%	44.7%	39.6%
(b)	94.7%	93.7%	93.0%	91.8%	90.9%	88.8%
(c)	21.3%	36.0%	55.2%	63.1%	68.9%	80.9%
(d)	22.8%	25.7%	26.7%	24.7%	25.3%	23.2%
(e)	27.9%	38.1%	62.2%	66.5%	72.3%	80.5%
(f)	50.4%	64.4%	78.5%	84.5%	87.8%	92.1%

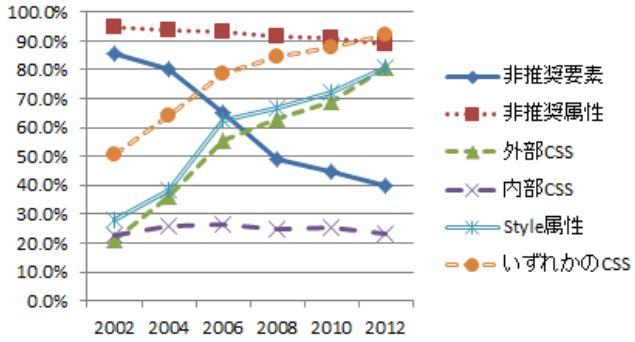

図 3　スタイル指定方法の変遷（グラフ）

調査の前には，(a)(b) の使用割合が年々低下し，CSS の使用割合が上昇していくと予想していた．しかしながら実際には，非推奨属性の使用割合と内部 CSS の使用割合はほぼ横ばい，という予想に反した結果（表 3・図 3）が得られた．

そこで，さらに詳細に検討するため，非推奨要素・非推奨属性の内訳を確認する．各ページごとに出現の有無（使用されているページの割合）を調査する．

非推奨要素の調査の結果（図 4），`font` タグと `center` タグが際立って多いことが判明した．どちらもここ 10 年の間に使用率はかなり減少しているものの，それでも今なお 3 割のページで `font` タグが，2 割のページで `center` タグが用いられていることが分かった．

一方，非推奨属性の内訳（図 5）を各属性ごとにみていくと，使用率が大きく減少しているものもあれば，ほぼ横ばいのもの，逆に増加傾向にあるものなど，さまざまであることが分かった．

3.4　`table` タグまたは CSS によるレイアウト制御

CSS によってレイアウト制御が行われているかどうかは，レイアウト制御に欠かせないプロパティである「`position`」と「`float`」のいずれかが含まれるか否かで判定を行う．集計結果を表 4 に示す．

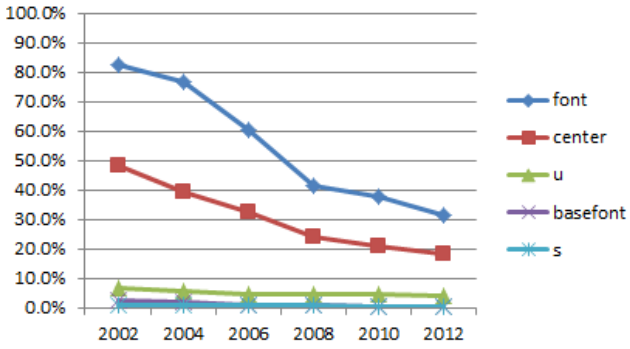

図 4　非推奨要素の使用割合（TOP5）

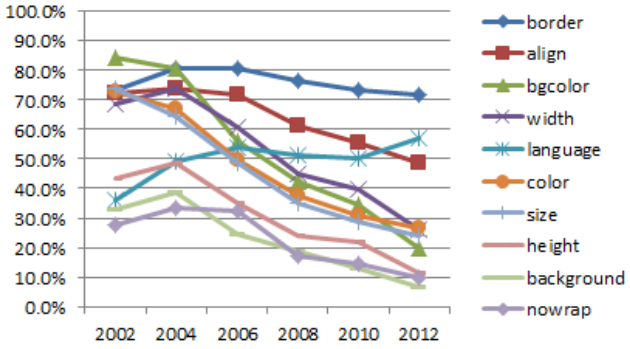

図 5　非推奨属性の使用割合（TOP10）

CSS によってレイアウト制御を行っているページの割合と単なる CSS の利用率（表 3 の (f)）とを比較すると，2002 年時点では CSS を利用していてもレイアウト制御を行っているページは 1 割強しかなかったのに対し，2012 年時点では CSS を利用しているページの 8 割以上が何らかのレイアウト制御を CSS で行っており，両者の差が年々急激に縮まっていったことが明らかとなった．

一方，`table` タグによってレイアウト制御が行われているかどうかを自動で正確に判定することは難しい．そこでまずは 2012 年に収集した約 10,000 ページに対し，以下の各項目について該当するか否かについて目視による判定作業を行った．

(A)　ページ全体の構成（コンテンツの配置）を `table` タグで行っているか

(B)　(A) には該当しないが，「表」ではない `table` タグが存在するか

(C)　(B) のうち，「ソーシャルボタン」または「カレンダー」のみに `table` タグが利用されているか

(A) に該当するページは全体の 6.2%，(B) は，ページの中の一部のコンテンツの配置を `table` タグで行っているケースなどが該当し 59.5%，(C) は「ソーシャルボタンを囲む」または「カレンダーを構成する」のみの目的で `table` タグが利用されるケースで 17.5%もあり，これを特殊ケースとみなして除くと，$(B)-(C)=42.0\%$ となる．2012 年現在でも，$(A)+(B)-(C)=48.2\%$ すなわち約半数のページにおいて，`table` タグがレイアウト制御に用いられていることがわかった．

続いて，この 2012 年のデータに基づき 2002 年〜2010 年のデータの中から (A) および (A)+(B)−(C) それぞれに該当する

ページを推定する．

判定材料（独立変数）は，以下の 4 項目を用いる．

- `table` タグの数
- `table` タグの深さの最大値
- 文書の先頭から最初の `table` タグが出現するまでのテキストの割合
- `table` タグの内側のテキストの割合

これらを使用して C4.5 で決定木を作成する．2012 年のデータに対し (A) は正解率 97.4%，(A)+(B)–(C) は正解率 90.5%の決定木がそれぞれ生成された．この決定木を用いて 2002〜2012 年のデータに適用する．推定結果を表 4 に示す．

CSS と `table` タグの調査結果を重ね合わせると，10 年間の間に割合が逆転していく様子が図 6 のとおり確認できた．

表 4　レイアウト制御手法

	2002	2004	2006	2008	2010	2012
CSS	6.2%	12.4%	29.9%	47.8%	63.7%	74.5%
(A)	43.6%	37.4%	28.8%	19.1%	13.8%	6.1%
(A)+(B)–(C)	81.7%	84.8%	78.7%	67.9%	62.6%	46.7%

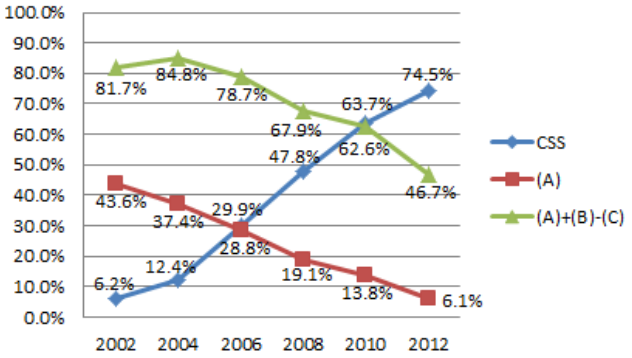

図 6　レイアウト制御手法の変遷（グラフ）

3.5　フレーム

フレームについては `frameset` タグ，インラインフレームについては `iframe` タグの使用率を調査した．結果を表 5 に示す．

フレームの使用状況については，予想どおり減少傾向が確認された．一方，インラインフレームは，急激に利用が伸びていることがわかった．

表 5　フレーム

	2002	2004	2006	2008	2010	2012
`frameset` タグ	1.6%	1.5%	0.6%	0.7%	0.8%	0.8%
`iframe` タグ	12.4%	14.0%	20.7%	23.0%	28.7%	42.4%

3.6　動的コンテンツ

動的コンテンツについては，HTML ファイルの中を確認するだけで容易に集計できる CGI・JavaScript・Flash・Java Applet・Silverlight の使用率について調査を行った．結果を表 6 に示す．

全体的な傾向として，「Web2.0」という言葉が流行し一般家庭のブロードバンド化が進んだ 2000 年代中頃に，ウェブサービスは大きく変化したことが裏付けられる結果が得られた．

まず，CGI を用いたページが 10 年の間に非常に増えたことが確認できる．その要因の一つに，サイト内検索ができる仕組みを持ったページが増えていることが挙げられる．

JavaScript は，Ajax が注目された 2006 年頃から急激に利用するページが増加している．たとえば，広告を動的に選んで掲載する仕組みに JavaScript が用いられており，急増の一因となっている．

Flash は，インタラクティブなウェブサイトを作成するための技術として登場したが，現在は動画を配信するために非常に広く用いられている．2000 年代中頃から動画を配信するページが増えたため，Flash の利用率もそれに伴い変化している．Flash の競合としては，Silverlight や HTML5 などがあるが 2012 年時点では Flash の利用率が圧倒的に高い．ただし，Flash は今後 HTML5 に置き換わっていくとみられ，その動向が注目される．

表 6　動的コンテンツ

	2002	2004	2006	2008	2010	2012
CGI	25.7%	37.0%	53.0%	59.6%	61.0%	68.3%
JavaScript	13.5%	36.1%	62.9%	77.5%	83.4%	89.7%
Flash	2.5%	4.7%	6.8%	8.6%	9.1%	6.2%
Java Applet	0.5%	0.2%	0.2%	0.2%	0.1%	0.1%
Silverlight	0.0%	0.0%	0.0%	0.4%	0.3%	0.6%

4.　ま　と　め

ここ 10 年間における日本語ウェブ文書の変化をさまざまな角度から調査した．

採用されている HTML のバージョンや使用される漢字コードがここ 10 年で大きく様変わりしていることから，情報の内容のみならず，ウェブページそのものが新しいページに次々置き換わっている状況だということがわかる．その一方で，一部の「非推奨属性」や「`table` タグによるレイアウト制御」など「非推奨」とされていても作り変えに多大な労力がかかる部分に関しては利用され続けてしまっている状況も確認された．

文　献

[1] 曽山裕，松本 章代，Martin J. Dürst：相関ルールを利用した CSS 自動生成手法の提案－ HTML からスタイル情報の分離－，第 1 回データ工学と情報マネジメントに関するフォーラム DEIM (2009).

[2] Y. Wang and J. Hu., “Detecting tables in HTML documents”, In LNCS, Vol. 2423, pp. 249–260, Springer-Verlag, (2002).

[3] 松本章代，小西達裕，高木朗，小山照夫，三宅芳雄，伊東幸宏：表構造における意味的関係に基づく WWW 検索性能の向上，電子情報通信学会論文誌 D，Vol.J91-D，No.3，pp.560-575 (2008).

[4] A. Mao, J.R. Cordy and T.R. Dean, “Automated Conversion of Table-based Websites to Structured Stylesheets Using Table Recognition and Clone Detection”, Proc. CASCON’07, pp.12–26 (2007).