Prevalence and classification of web page defects

Ejike Ofuonye, Patricia Beatty, Scott Dick and James Miller
Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada

Refereed article received
16 April 2009
Approved for publication
13 August 2009

Abstract

Purpose – The purpose of this paper is to provide an update on previous surveys that have looked at the quality of HTML documents on the worldwide web. Previous surveys have indicated that the quality of HTML documents tends to be quite poor, with most documents containing defects.

Design/methodology/approach – To determine the extent of this problem, the paper undertook a large-scale study of HTML document quality among the most popular web sites (approximately 100,000).

Findings – This paper found that the vast majority (over 95 per cent) of web sites did not adhere to the worldwide web consortium standards for HTML.

Research limitations/implications – This study represents a single investigation over a short timeframe. Hence, ideally the study needs to be replicated in the future to help generalise the findings.

Practical implications – Such poor quality may jeopardise the security or usability of a web site, making the site's users vulnerable to malware attacks. This poor level of quality has drastic implications for web usability and security.

Originality/value – This new survey undertook a more extensive examination of popular web sites than previous surveys.

Keywords Standards, Hypertext Markup Language, Worldwide web

Paper type Research paper

Introduction

Hyper Text Markup Language (HTML) is the standard language used for the majority of information exchange on the internet. It is a platform independent language used to encode documents transferred from one machine (the server) to another (the client) over the Hyper Text Transfer Protocol (HTTP). HTTP is used in the transfer of other document types such as images, JavaScript, Cascaded Style Sheets (CSS) and Platform for Privacy Preferences (P3P) (Cranor, 2002) policy documents. These document types normally constitute the bulk of HTTP traffic and are almost always organised as resources linked from a HTML document.

An HTML document is a plain text file (with a mime type text/html) containing directives in the form of tags that represent the structure and layout of a document. These tags instruct an agent (usually a web browser like Internet Explorer or Mozilla Firefox) how to graphically display the document to a user. The HTML standard denotes certain document fragments as links, headings, titles, paragraphs, images, etc. Over the years from its birth in the early 1990s, the HTML standard has continuously evolved to achieve consistent display of markup text and to embrace needed changes as the web has evolved. Currently the World Wide Web Consortium (W3C) promulgates the HTML standards as W3C Recommendation documents and also provides validating tools to the general public to check for conformance of documents

Online Information Review
Vol. 34 No. 1, 2010
pp. 160-174

1468-4527
DOI 10.1108/14684521011024182

on the internet. The current version of the HTML standard recommended by the W3C is HTML 4.01[1]. A simple example of a document coded to this standard is shown in Figure 1.

Each HTML document must begin with a DOCTYPE declaration that specifies what version of the HTML standard applies to the document (see Figure 1, line 1). In this example, the version is HTML 4.01 Strict. An up-to-date list of recommended document types is available on the W3C web site (www.w3.org/QA/2002/04/valid-dtd-list.html). The < html > </html > start and end tags are the document's root element and must enclose all other elements. The < head > </head > element is used to give information about the document that is not displayed on the web site. For example, the < title > </title > in the HTML code shown previously will not be displayed in the browser window. Both the head element and the title element are required in a valid HTML document. The next element in Figure 1 is the < body > < /body > element. This section is where the main body or content of the HTML document is placed. The body element is required in a valid HTML document and that section must not be empty. In Figure 1 the body contains just a single paragraph (< p > </p >) element. In real web pages this section can span many thousands of lines of markup text, referencing most of the available HTML 4.01 tags.

Table I provides a timeline of the evolution of the various standards of HTML leading up to the migration to the eXtensible Markup Language (XML) versions typically referred to as XHTML. Although these standards are spelt out in detailed and easily available documents, studies have repeatedly shown that the great majority of web sites do not adhere to any of these standards. For example, a recent study (Beatty *et al.*, 2008) found that up to 20 per cent of web sites did not even include the mandatory DOCTYPE declaration. The goal of our study was to provide a more up-to-date

```
1. <!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01//EN"
   http://www.w3.org/TR/html4/strict.dtd>
2. <html>
3.   <head>
4.     <title>Example HTML 4.01 Document</title>
5.   </head>
6.   <body>
7.     <p>An example of paragraph in HTML</p>
8.   </body>
9. </html>
```

Figure 1.
Sample HTML 4.01 document

Version	Year	Remark
HTML tags	Late 1991	No version number, represents the first HTML
HTML2.0	November 1995	
HTML3.2	January 1997	
HTML4.0	April 1998	Strict, Transitional, and Frameset versions
HTML4.01	December 1999	Strict, Transitional, and Frameset versions
HTML5	January 2008	Working draft
XHTML1.0	January 2000	Strict, Transitional, Basic and Frameset versions
XHTML1.1	May 2001	Based on XHTML1.0 Strict version
ISO/IEC 15445:2000	2000	Based on HTML4.01 Strict

Table I.
HTML standards evolution

investigation of the prevalence and types of defects in HTML pages. Our study population consisted of nearly 100,000 web sites, making it the largest study of its kind of which we are aware. We have aggregated our results at the level of individual nation-states and have treated the nation-state as a proxy for the technological context in which web sites are constructed. Our results indicate that overall adherence to the HTML standards is very low, but the quality of a web site does appear to be related to the nation-state of origin.

Related work

There is little empirical evidence on adherence to HTML standards on the web. All previous studies examined smaller populations of web sites than the current study. Chen *et al.* (2005) studied HTML standards compliance using four populations of web sites:

(1) The top 10,000 web sites as ranked by Alexa.com

(2) 1,100 web sites obtained by randomly generating IP addresses and checking on port 80 for web servers.

(3) 31,540 web sites obtained by randomly generating character strings and sending them to three search engines (Google, Yahoo and Teoma). The first ten results for each search engine on each string were added to the population.

(4) The final list was obtained two months later by again randomly generating IP addresses.

The survey used W3C's HTML Validation tool to check for HTML standards compliance on the web sites in the four populations. Chen *et al.* (2005) found that only 5 per cent of the web sites studied complied with the HTML standards. A previous study by Beckett (1997) considered a population of 13,312 unique web sites found under the.uk domain. The HTML on the web sites was checked for standards compliance using the NSGMLS parser. This survey found that 6.5 per cent of web sites complied with the HTML standards. Pollach *et al.* (2006) performed a similar study on 226 web sites that focused on environmental issues. This population was obtained from the Google Directory in 2006. The survey used WebXACT to check the web sites' adherence to HTML standards and found that 4 per cent of web sites complied with the HTML standards.

We note that only the 2005 study used the official HTML Validator from the W3C. It is unclear why the 2006 study did not follow this methodology. In our view the 2005 study is the most relevant study for purposes of comparison as it was not limited by top-level domain or subject area.). Our study of a much larger survey population (nearly 100,000 web sites) comprises a more up-to-date contribution to the literature in this field.

Study methodology

The population for this study was composed of the most popular 100,000 web sites on the internet as ranked by Alexa.com (www.alexa.com), based on the geometric mean of the number of individuals visiting a web site and the number of pages they access while on the web site. According to Alexa.com, web sites that are not on this list have less than a 0.00125 per cent chance of being visited by the average internet user (Staff,

2007). We believe that this list is arguably a reasonable proxy for the "useable" web, that is, the subset of web pages that an ordinary user is likely to encounter.

The geographic location of these web sites was derived by determining their IP address using the Linux 'host' programme and then comparing this address with a database of addresses purchased from IP2Location (Staff, 2006) that maps IP addresses to a particular nation. This approach was required since the country code top-level domains are not reliable indicators of the actual host location of web sites (Black, 2005). IP2Location claim that their accuracy is above 95 per cent.

The W3C HTML Validator (http://validator.w3.org/, http://validator.w3.org/source/) was used to evaluate compliance with HTML standards. This is a freely available software tool that analyses an HTML document and reports any deviations from the HTML standard. Only the homepage of each web site was tested and as such we used the homepage as a proxy for the conformance of the entire site. Our belief is that the homepage is perhaps the most highly maintained and updated page on any web site and hence represents the page that is likely to have the highest correspondence to the standards on any site. Therefore we believe that the following results should be viewed as a conservative estimate of the deviation from the standard and that the deviation of the "average" web page is likely to be higher.

The W3C HTML Validator outputs a number of warnings, fatal errors and syntax errors. Explanations for all warnings and fatal errors are given in Table II. There are a total of 448 syntax errors, but only a few of these were ever observed. Descriptions for these syntax errors can be found at http://validator.w3.org/docs/errors.html. In general, these errors are best considered as deviations from the standard rather than problems for the client machine. Web browsers are highly tolerant of many issues and while these errors are departures from the standard, they may be invisible issues to the user.

The 100,000 web sites studied were from a total of 131 different countries. For 1,235 web sites the country of origin was unknown or the web site could not be reached at the

Error code	Description
FE0	Could not retrieve web site
FE2	Error with character set
FE3	Error transcoding document
FE4	Error with document type declaration (no or unknown FPI and a relative SI)
FE6	Error finding encoding
W01	Character set not found but content type is text so using US-ASCII
W04	No character set found, using UTF-8 instead
W06	No file mode found, using SGML by default
W07	Content type not supported
W08	No document types configuration found, mode not determined
W09	Fallback FPI, no document type found
W10	Namespace found does not match document type namespace
W11	Namespace found but document type is not XML
W12	Missing namespace
W17	No character set found
W18	Character set conflict in http header and xml
W19	Character set conflict in http header and < meta > element
W20	Character set conflict in < meta > element and xml
W21	Character set is UTF-8 and contains a byte order mark (BOM)

Table II.
W3C Validator warning and fatal error code

time of the *whois* query. For each country of origin we also obtained the gross domestic product (GDP) at purchasing power parity (PPP) per capita and e-readiness ranking. GDP PPP per capita, taken from the International Monetary Fund's World Economic Outlook database (www.imf.org/external/pubs/ft/weo/2005/01/data/) for 2005, was used because it is a measure that can easily be compared between countries and is an indicator of wealth within a country that takes into consideration the standard of living. The e-readiness ranking used was a 2005 ranking of 65 countries that was created by IBM and the Economist Intelligence Unit (2005). The ranking has a theoretical range of 1.00 and 10.00, however all countries were ranked between 2.00 and 8.99 in the 2005 survey. Although only 65 countries are ranked, these countries account for 98 per cent of all web sites in the Alexa list. We used these two measures as a proxy for the technological milieu in a nation-state, to investigate if there is a relationship between standards compliance and that milieu.

Limitations of the survey

As with all surveys, our survey had a number of threats to its validity. The following limitations are considered the most significant:

- the survey used a web site's homepage as a proxy for the general defect rate experienced across the web site. while the homepage tends to be actively maintained and updated, this may not be the case for pages "deep" within the structure of a web site;
- many web sites generate HTML "on the fly" by executing JavaScript, ActionScript or VBScript. This additional HTML was not analysed in our survey; and
- Alexa's list of the top 100,000 only counts visits from Internet Explorer based traffic. Hence, web sites with a large Firefox following will be under-represented in the list.

Results and analysis

Number of web sites containing valid HTML

Perhaps the most interesting result from this study was the small number of completely valid web sites in the population (recall that only the homepages were examined). The survey found that only 3,020 web sites out of 100,000 (approximately 3 per cent) contained zero errors or warnings, indicating full compliance with the HTML standards. To put this result in perspective, Table III compares our finding with the surveys previously outlined previously. Because of the differences in the web site samples, not all the results can be directly compared to each other. However, the percentage of valid web sites was similar (and very low) in every one of the studies.

Table III. Validation results from various surveys compared

Survey	Percentage valid (%)	Sample	Year	Reference
Current survey	3.0	Alexa top 100,000	2006	N/A
Alexa10k	5.0	1,100 random sites and Alexa top 10,000	2005	(Chen *et al.*, 2005)
Environmental	4.0	226 environmental sites	2006	(Pollach *et al.*, 2006)
The UK	6.5	13,312 sites under the .uk domain	1997	(Beckett, 1997)

The UK (Beckett, 1997) and Alexa10k (Chen *et al.*, 2005) surveys can be compared more closely with the results from the current survey, as we were able to identify corresponding subsets of our study population.

First we compared our results against the UK study (Beckett, 1997) by forming a population consisting of those web sites in the top 100,000 hosted from the UK (as noted, this is not equivalent to identifying sites with a.uk top-level domain). This yielded a new population of 3,536 web sites of which 8 per cent were compliant with the HTML standards. Likewise, by forming a new population from the top 10,000 web sites in the Alexa 100,000 list, we obtained a 2006 analogue of the top 10,000 list in Chen *et al.* (2005). In this new population, 3.3 per cent of web sites passed the W3C HTML Validator – a slight improvement from the full population but a large relative decrease from a year earlier. The list of the top 10,000 most popular web sites is of course not a static population, but dynamically changes over time. However, the fact that both lists were arrived at by the same methods and the same organisation, and that both populations were examined by the same validation tool, leads us to conclude that the 34 per cent drop in the number of web sites that adhered to the standard was not just a methodological artefact. The disturbing implication is that the most popular web sites seem to be taking *less* care with web standards compliance, even as standards-based web browsers are now rapidly becoming an important force in the market[2]. However, the higher rate of compliance among UK web sites indicates that this picture is nuanced, with different nations perhaps exhibiting different rates of compliance.

The first question we asked was if there was an overall trend in standards compliance versus the popularity of a site. Intuitively, one might expect that the webmasters of the most popular web sites would expend greater effort in ensuring the quality of their web pages (and particularly the homepage of the site). The graph in Figure 2 shows the number of valid web sites (per 1,000, in order of popularity) versus the ranking of the web sites. This graph shows a very slight downward trend in the number of valid web sites. However, the effect is very small and is largely obscured by the local fluctuations in the plot.

Table IV shows the number of compliant web sites per country for all countries that had more than 1,000 web sites in the study population. All other countries combined represented less than 1 per cent of all sites examined, and therefore had little individual effect on the aggregate level of standards compliance in this population.

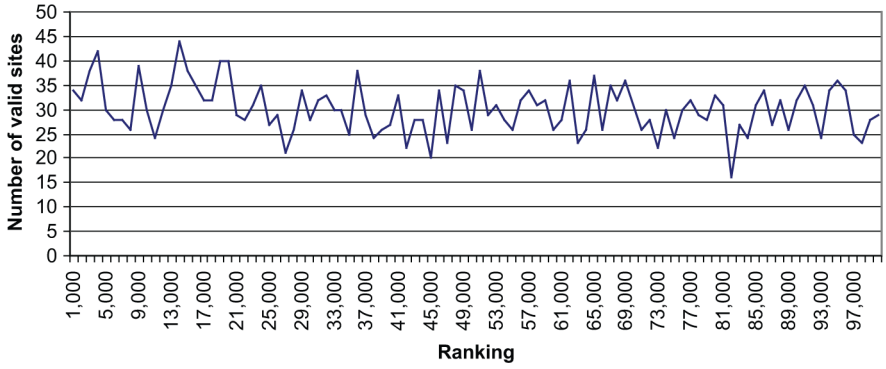

Figure 2.
Number of valid sites per 1,000 by ranking

Country	Number of web sites	Valid	Percentage valid
The USA	45,212	1,343	3.0
China	17,764	140	0.8
Japan	6,807	292	4.3
The UK	3,536	284	8.0
Canada	2,576	91	3.5
Germany	2,434	144	5.9
France	1,858	83	4.5
Republic of Korea	1,755	8	0.5
The Netherlands	1,520	60	3.9
Hong Kong	1,437	34	2.4
Taiwan	1,340	19	1.4
Spain	1,243	83	6.7

Table IV.
Valid web sites per country of origin

There were extensive differences in the percentage of standards-compliant web sites in each country, ranging from 8 per cent in the UK to 0.5 per cent in the Republic of Korea. Oddly enough, this difference appeared to be associated with continental location, with Asian countries having the lowest percentage of valid web sites and European countries having the highest percentage of valid web sites. We investigate this possibility further in Table V, which presents the percentage of standards-compliant web sites for each continent (determined by summing the results for each of the 131 countries in the Alexa top 100,000 list by the continent that country belongs to). In this more comprehensive analysis, we again see that Africa and Asia had the lowest percentage of compliant web sites, Europe and Oceania had the highest percentage compliant and the Americas were positioned in the mid-range. The web sites in the unknown category were dominated (987 sites) by web sites that were unreachable when the *whois* database was queried to find their country of origin.

To quantify the significance of these differences among continents, we employed Fisher's exact test. Table VI presents the p-values obtained from this test (recall that there is no associated test statistic) for a pair-wise comparison of the six continents and the "unknown" category. When we used the traditional significance level of $\alpha = 0.05$, we found that the great majority of the differences were significant. However, there were a few exceptions. The differences between Africa and Asia ($p < 0.193$) as well as Africa and the "unknown" category ($p < 0.704$) were not statistically significant. The difference between Oceania and Europe ($p < 0.267$) was not statistically significant and nor was the difference between North and South America ($p < 0.921$). No simple

Continent	Number of web sites	Valid	Percentage valid
Africa	227	1	0.4
Asia	31,895	538	1.7
South America	886	27	3.0
North America	48,240	1,445	3.0
Europe	16,617	939	5.7
Oceania	900	59	6.6
Unknown	1,235	11	0.9

Table V.
Valid web sites by continent

explanation for these differences was immediately apparent. Europe and North America are popularly considered to be the most "technologically advanced" (in as much as such a statement can be made about a continent) and yet North America was in the "middle of the pack" on standards compliance, while Europe was statistically tied with Oceania. Africa is commonly considered the poorest continent and yet web sites in Asia (with a number of robust and advanced economies) were statistically no more likely to be standards-compliant than their African counterparts.

The differences in the percentage of compliant web sites between countries and continents may be better explained by examining GDP PPP per capita. This measure is well defined for individual nations and the most technologically advanced nations also tend to be the wealthiest (we formalised this statement when we examined the e-readiness statistic). Table VII shows the percentages of compliant web sites versus a categorisation of the GDP PPP for each country. The unknown category consists of web sites for which the country was not known as well as web sites from countries that are not members of the International Monetary Fund. There was a noticeable trend in that the percentage of compliant web sites increased as the GDP PPP increased. The exception to this trend was when the GDP was greater than US$40,000. This category was dominated by the USA, which comprised 99 per cent of these sites. Removing the United States from this category would give a total of 385 web sites remaining of which 24 (6.2 per cent) were compliant, which matched the trend.

Table VIII shows the number of valid web sites as well as the percentage of compliant web sites by categories derived from the e-readiness statistics. GDP PPP and e-readiness are linearly related, that is, national wealth is related to the technological infrastructure and technological milieu in a country. Thus a similar trend was seen here as was seen with the GDP data. The percentage of compliant web sites increased as the e-readiness score increased, with the exception of e-readiness scores of over 7.00. Web sites from the United States, which corresponded to the GDP PPP category of over

	Africa	Asia	Europe	North America	Oceania	South America	Unknown
Africa	1.000	0.193	0.000	0.017	0.000	0.029	0.704
Asia		1.000	0.000	0.000	0.000	0.005	0.030
Europe			1.000	0.000	0.267	0.000	0.000
North America				1.000	0.000	0.921	0.000
Oceania					1.000	0.001	0.000
South America						1.000	0.000
Unknown							1.000

Table VI. Fisher's exact test (two-tail) for compliant web sites by continent

GDP (PPP) per capita (USD)	Number of web sites	Valid	Percentage valid
<10,000	20,557	192	0.9
>10,000	2,547	57	2.2
>20,000	6,230	190	3.0
>30,000	23,687	1,202	5.1
>40,000	45,597	1367	3.0
Unknown	1,382	12	0.9

Table VII. Valid web sites by GDP PPP per capita (USD)

US$40000, dominated the web sites in the e-readiness category of 8.xx and had a lower percentage of valid web sites. However web sites in the e-readiness category of 7.xx deviated from the trend in the GDP PPP data. This category also had a lower percentage of valid web sites, which does not correspond to the GDP PPP data.

Number of warnings

Table IX shows the number of web sites that had each type of warning. Note that not all warnings given in Table IX were found in this survey. Of the warnings that were found, W04 and W09 were found on a large number of web sites. These warnings correspond to a missing character set and missing document type respectively when default values are used. It is possible that this is due to web developers relying on browsers to assume a default character set and document type instead of specifying one. If a character set or document type was not found by the W3C Validator a fallback option was tried in order to tentatively validate the web site. The only other warning that affected more than 1 per cent of the web sites was W19, which referred to a conflict between character set declarations in the http header and < meta > element. Missing or invalid character sets and document types repeatedly appeared as common problems in the web sites in this study.

The total number of warnings found as a function of web site ranking is shown in Figure 3. No clear trend was seen here.

Number of fatal errors

A surprisingly large number of web sites (19.1 per cent) produced a fatal error while being validated. Table X shows the fatal errors that were encountered and the number

Table VIII.
Valid web sites by e-readiness category

E-readiness	Number of web sites	Valid	Percentage valid
2.xx	18	0	0.0
3.xx	18,977	150	0.8
4.xx	1,801	38	2.1
5.xx	1,560	41	2.6
6.xx	1,670	98	5.9
7.xx	14,147	524	3.7
8.xx	59,791	2,137	3.6
Unknown	2,036	32	1.6

Table IX.
Number of web sites with each warning

Warning type	Number of web sites
W04	12,703
W06	809
W07	44
W09	42,495
W11	858
W18	61
W19	4,296
W20	14
W21	420

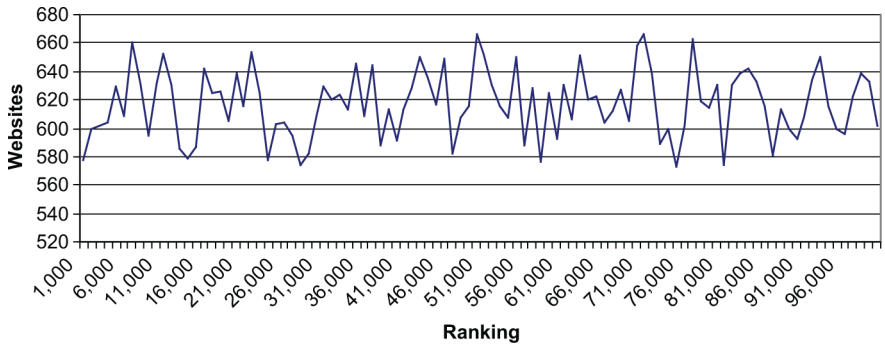

Figure 3.
Total number of warnings versus web site ranking

Fatal error type	Number of web sites
FE0	4,790
FE2	1,824
FE3	12,439
FE4	55

Table X.
Number of web sites with each fatal error

of web sites that had each error. The fatal errors were dominated by problems with the character set (FE2) and problems with transcoding the document (FE3) as well as many unreachable sites (FE0). The error FE3 occurs when a web page has characters in it that are not in the character set the HTML document specifies and so the web page can not be transcoded.

Figure 4 shows a very slight increase in the number of fatal errors as the ranking of web sites decreases. Further analysis of Figure 4 revealed that most of these fatal errors were associated with character set problems.

Discussion

In general, the results in this study confirm previous findings – the level of compliance to the HTML standard among our population of web sites is very poor at only 3 per

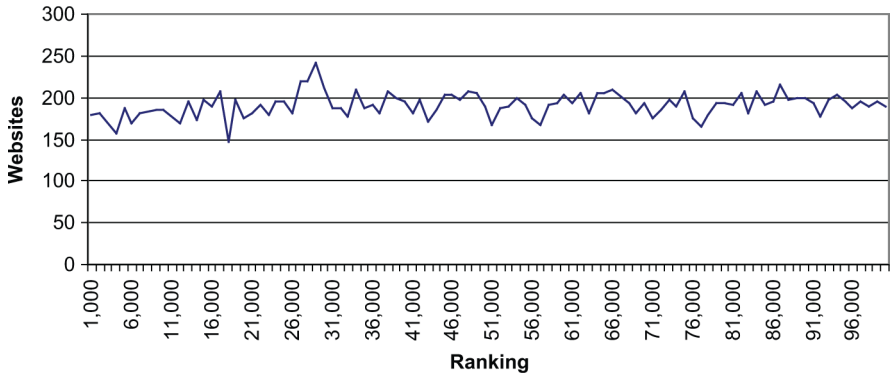

Figure 4.
Total number of web sites with a fatal error versus web site ranking

cent. There was some evidence of a slight relationship between the popularity of a web site and the level of compliance – the most popular web sites were slightly more likely to post compliant HTML documents than less popular ones. In addition, as web sites became less popular, we observed an increased number of fatal errors as well as an increased number of invalid or missing character sets and document types.

When we broke our study results down by country of origin, some stronger trends emerged. In an exploratory analysis, GDP PPP was found to be related to the level of HTML standards compliance in that country – generally speaking, the wealthier a country is in GDP PPP terms, the greater the percentage of standards-compliant web sites. However, there were two caveats. First, the USA was an exception to this pattern – the compliance rates in that nation were similar to nations with half the GDP PPP. As the USA hosted over 45,000 of the 100,000 sites in this study's population, this had a very significant effect on our overall compliance results. Second, even the nation with the highest compliance rate (the UK) only reached 8 per cent compliance. When we undertook a similar exploratory analysis for the e-readiness rankings, a similar trend emerged.

The average user, however, would not say that 97 per cent of the web sites they visit do not work. Indeed, the web functions quite well despite the problems we observed in this study. We submit that there are two reasons for this. First, many of the errors in HTML code only result in small mistakes on the web site, which the users may not be aware of. Some errors, such as a missing " > " after a tag, may simply cause text after that tag to not display in the web browser. An example of this would be forgetting the end " > " when creating a table:

⟨tableThis text would not be displayed⟨/table⟩

In this case a user would have no way of knowing that content was missing from the web page. Similarly, errors that result in a small change in the layout or formatting of a web page would probably be unnoticeable to anyone except the creator of that web page.

Second, web browsers are remarkably fault-tolerant software systems. Most errors, such as using attributes or tags that do not exist in the formal HTML specification, are simply ignored. For other errors (such as missing required attributes, i.e. the *type* attribute from the script tag) web browsers would use a default value or attempt to infer a correct value. Thus the incentive for HTML document authors to adhere to the HTML standards is considerably reduced because the non-compliant pages they currently post are "good enough" to be rendered essentially as intended.

Consequences, reasons and remedies of errors

Consequences

Although most web site authors appear to benefit from the excellent fault tolerance of browsers, there are consequences for producing malformed web pages. Broadly speaking, we can divide these consequences into usability issues and security issues.

Web site usability

Implementation differences in web browsers can lead to a host of usability issues in web sites, especially when the HTML document is not fully compliant with the standards. Generally, web sites may display quite differently in different browsers and browser versions. This is because in attempting to render malformed pages, each web browser has been designed to use different default values for missing HTML elements

and/or attributes. This leads to different layouts of the same page across multiple browsers. For example, the cellspacing, cellpadding and border attributes for the HTML < table > tag can have unique default values, leading to a very different look and feel in different web browsers. Another example of default values that are set by the web browser is CSS text style properties such as colour and font family. Also, many web browsers offer browser specific HTML – non-standard tags such as marquee and bgsound in Internet Explorer and blink, multicol and *spacer* in Netscape. If these tags are used on a web site it may not display properly or at all in other web browsers. The differences in web browser implementations may cause a web page that displays properly in Internet Explorer 7 to not display at all in Firefox. To add to all these differences, each web browser also has its own set of software defects.

In recent studies that have examined factors that impact on users' trust of a web site (Fogg *et al.*, 2003; Ofuonye *et al.*, 2008), ease of use and competence were identified as leading factors that users employed in deciding whether to trust a web site. If it were not for the fault-tolerant characteristics of the major web browsers, ease of use and the appearance of competence in web site design could be severely impacted by non-compliance with the HTML standards. As increasing emphasis is placed on standards-based design in the web browser market, one naturally wonders if this extensive fault tolerance will continue to be a feature of future web browsers, and if not, what the effect on the web authoring community might be.

Web site security

Usability issues aside, more worrisome are the security and privacy issues that erroneous HTML introduces to the already complex topic of web security. In order for a browser to be secure, it must follow a strict rule set of what inputs are valid and which are *not*. However, since browsers must currently cope with a staggering number of possible malformed inputs, they must be quite promiscuous in their input validation. Combining this requirement with the need for security results in huge complex software systems (e.g. the Mozilla project consists of over ten million lines of source code), with the concomitant increase in software defects this size entails.

Web server and web browser vulnerabilities are the most commonly discussed sources of security breaches in web applications, as exploits of these platforms typically lead to more devastating consequences. However, there have been recent exploits that have raised the profile of client application logic vulnerabilities. The Samy worm (Samy, 2008) and Yammaner virus (Hoffman and Sullivan, 2008) almost brought the social networking site MySpace and email giant Yahoo, respectively, to a standstill. Commonly, these exploits use a class of client-side exploit called Cross Site Scripting (XSS) (Anderson, 2008) attacks to perpetuate their havoc. Although XSS attacks primarily reflect security issues in the web application itself, further HTML errors in a web site add to the chances of these exploits going unnoticed by many clients. For example, the HTML fragment

⟨img“”“⟩⟨script⟩alert(“XSS”)⟨/script⟩”⟩

contains a sample XSS vector from the RSnake list of vectors that renders in almost all modern browsers. Clearly this fragment is not standards-compliant, but the eager nature of web browsers to render almost any piece of gibberish leads to this exploit being executed anyway (Hansen, 2008).

Reasons for web site errors and their remedies

In many cases the methods used to produce Cascading Style Sheets (CSS) and HTML documents may contribute to these documents not adhering to the W3C standards. There are two methods for producing CSS and HTML documents. The first method is to use a text editor and write the documents by hand. This method is time consuming and unless the developer is an expert with CSS and HTML, it may be difficult to produce documents without errors. The second method is to use a WYSIWYG (What You See Is What You Get) editor such as Microsoft FrontPage, Microsoft Expression Web or Adobe Dreamweaver. With these types of editors, the developer graphically develops the web page and the editor then automatically generates the CSS and HTML code for it. However, many of these editors produce markup text that does not adhere to the W3C standards. Dreamweaver 8.0, for instance, failed the Acid 2 test (The Web Standards Project, 2008). Errors generated by these authoring tools have included non-SGML characters being present (due to characters such as an apostrophe not being translated properly), non-standard elements being used and no HTML document type assigned, as well as the use of non-standard properties in these files. This may explain, in part, why so few HTML and CSS documents validate.

In summary and without rigorous recourse to empirical studies, it can be posited that most errors on web sites occur as a result of one or more of the following reasons:

- *Use of old web-authoring tools*. This may be due to licensing concerns or downright resistance to change. Plainly put, there are tools that were not designed for the Web 2.0 environment and using them to generate content for this class of applications results in errors.
- *Lack of awareness.* As web developers come and go, different web pages are added to a web site following different HTML standards. When there is no effort to update HTML documents that have been written in the past, the incorporation of new HTML documents (possibly written with different authoring tools) can lead to a patchwork of HTML dialects on a web site. Plainly, errors may creep into web pages if developers are not familiar with the older standards and the idiosyncrasies of older web browsers that prompted the use of non-standard HTML elements.
- *If it works, don't fix it.* As the browser, with its highly tolerant algorithm, creates the illusion of correct web pages, web developers may well move on to other tasks. Ferreting out errors in HTML is likely to be a time consuming and expensive process and one which might add little up-front business value. Removing defects is after all a loss-mitigation activity, not a revenue-generating activity.
- *Little or no validation.* Most pages today are written by individuals who are not aware of the subtleties of HTML standards. Therefore it is not surprising that very little effort goes into validating web pages using standard tools like the W3C Validator.

What needs to be done and why is self-evident. There is a need for more validation of web sites by their authors, use of modern and appropriate tools for the class of web page to be developed, constant and ongoing testing across multiple browser implementations and versions, and entrenchment of standards compliance in the development process. These are the necessary steps in improving standards compliance on web pages and ultimately in improving the usability and security of the web.

Conclusions and future research

This paper has presented the results of our study of HTML standards compliance on the homepages of a large number of very popular web sites on the internet. Our findings revealed that only 3 per cent of web sites in the study's population had fully valid markup text. However, modern web browsers incorporate advanced fault tolerant algorithms and a set of default values that allows most of these pages to render despite these errors. While this feature of web browsers is benign, there are situations where default values may result in the distortion of the page (especially across different browsers). In some cases, HTML errors also create security vulnerabilities that malicious authors can use to exploit client machines. We argue that web developers need to perform extensive validation of their pages to ensure that their HTML documents will render as intended and will not be open to security weaknesses.

Our survey was a snapshot of the worldwide web. The worldwide web is constantly developing and so we intend to repeat this study in the future to observe the evolution of HTML standards compliance. The standards-based web browsing movement gives us cause to believe that, over time, web designers and developers will spend additional effort on their web sites to ensure compliance with HTML standards. We also believe that there will be ongoing improvements in the HTML code generated by authoring tools. We look forward to examining – and hopefully confirming – these predictions.

Notes

1. This statement clearly ignores the existence of XHTML.
2. The Acid2 test of the Web Standards Project (www.webstandards.org/action/acid2/) is designed to help browser vendors ensure proper support for web standards in their products. Firefox 3 passes the Acid2 test, Internet Explorer 7 does not, but Internet Explorer 8 beta does.

References

Anderson, R. (2008), *Security Engineering*, 2nd ed., Wiley Publishing, Indianapolis, IN.

Beatty, P., Dick, S. and Miller, J. (2008), "Is HTML in a race to the bottom? A large-scale survey and analysis of conformance to W3C standards", *IEEE Internet Computing*, Vol. 12 No. 2, pp. 76-80.

Beckett, D.J. (1997), "30% accessible – a survey of the UK Wide Web", *Computer Networks and ISDN Systems*, Vol. 29 Nos 8-13, pp. 1367-75.

Black, H. (2005), "Letter released about Abika.com, an on-line data broker in the US", available at: www.privcom.gc.ca/legislation/let/let_051118_e.asp (accessed 19 November 2006).

Chen, S., Hong, D. and Shen, V.Y. (2005), "An experimental study on validation problems with existing HTML web pages", *Proceedings of the International Conference on Internet Computing (ICOMP'05)*, pp. 373-9.

Cranor, L.F. (2002), *Web Privacy with P3P*, O'Reilly & Associates, Sebastopol, CA.

Economist Intelligence Unit (2005), *The 2005 E-Readiness Rankings*, white paper, The Economist Intelligence Unit, London, available at: http://a330.g.akamai.net/7/330/25828/20050415154011/graphics.eiu.com/files/ad_pdfs/2005Ereadiness_Ranking_WP.pdf

Fogg, B.J., Soohoo, C., Danielson, D.R., Marable, L., Stanford, J. and Tauber, E.R. (2003), "How do users evaluate the credibility of web sites? A study with over 2,500 participants", *Proceedings of the 2003 Conference on Designing for User Experiences, San Francisco, CA*, pp. 1-5.

Hansen, R. (2008), "XSS cheat sheet, appendix of OWASP 2.0 Guide", available at: http://ha.ckers.org/xss.html (accessed 1 September).

Hoffman, B. and Sullivan, B. (2008), *Ajax Security*, Addison Wesley, Upper Saddle River, NJ.

Ofuonye, E., Beatty, P., Reay, I., Dick, S. and Miller, J. (2008), "How do we build trust into e-commerce web site?", *IEEE Software*, Vol. 25 No. 5, pp. 7-9.

Pollach, I., Pinterits, A. and Treiblmaier, H. (2006), "Environmental web sites: an empirical investigation of functionality and accessibility", *International Journal of Technology, Policy and Management*, Vol. 6 No. 1, pp. 103-19.

Samy (2008), "Technical explanation of the MySpace worm", available at: http://namb.la/popular/tech.html (accessed 1 September).

Staff, A.I.I. (2007), "Alexa web search – Top 500", available at: www.alexa.com/site/ds/top_500 (accessed 21 February).

Staff, I.L. (2006), "Geolocation IP address to country city region latitude longitude ZIP code ISP domain name database for developers", available at: www.ip2location.com/ (accessed 21 February 2007).

The Web Standards Project (2008), "Acid2 Browser test", available at: www.webstandards.org/action/acid2/ (accessed 24 September).

About the authors

Ejike Ofuonye graduated in 2003 from the Department of Electrical and Electronic Engineering at the University of Nigeria. He is currently a PhD student in the Department of Electrical and Computer Engineering at the University of Alberta, Canada having completed an MSc in 2008. His research interests are in the areas of web usability, web security and secure software development methods.

Patricia Beatty is a Network Engineer at MedExpert in San Francisco. She has an MSc in Software Engineering from the University of Alberta, Canada.

Scott Dick is an Associate Professor at the Department of Computer Engineering at the University of Alberta, Canada. He specialises in computational intelligence, machine learning and data mining, as well as the interdisciplinary application of these technologies to other fields of study.

James Miller is a Professor at the Department of Electrical and Computer Engineering at the University of Alberta. He has published over one hundred refereed journal and conference papers on web, software and systems engineering (see www.steam.ualberta.ca for details on recent directions), and currently serves on the programme chair for the IEEE International Symposium on Empirical Software Engineering and Measurement 2009. He also sits on the editorial board of the *Journal of Empirical Software Engineering*. James Miller is the corresponding author and can be contacted at: jm@ece.ualberta.ca