Received: 17 May 2017 | Revised: 11 August 2017 | Accepted: 15 August 2017

DOI: 10.1002/spe.2545

SHORT COMMUNICATION

WILEY

Toward characterizing HTML defects on the Web

Joaquim Mendes | Nuno Laranjeiro | Marco Vieira

Centre for Informatics and Systems, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal

Correspondence
Nuno Laranjeiro, Centre for Informatics and Systems, Department of Informatics Engineering, University of Coimbra, 3030-790 Coimbra, Portugal.
Email: cnl@dei.uc.pt

Funding information
EUBra-BIGSEA; European Commission under the Cooperation Programme, Horizon 2020, Grant/Award Number: 690116; Project MobiWise, Grant/Award Number: P2020 SAICTPAC/0011/2015

Summary

HTML is being massively used as an interface to provide services to users. Web developers are producing and changing sites at a high pace while trying to support the latest HTML standards. In this context, it is common to find websites that do not comply with the standards and fail to be correctly processed by browsers. Considering this dynamic environment and the increasingly large diversity of browsers with frequent updates, the appearance of problems in web pages is a common, sometimes severe, and hard-to-track problem. In this short communication, we describe the initial design of an approach that will be used to obtain information regarding the characteristics of HTML documents on the Web and extract indicators of representative errors made by their developers. Preliminary results show nearly 90% of the pages analyzed having at least one type of error and the prevalence of a small number of error types.

KEYWORDS

HTML, HTML defects, HTML validation, standards compliance, web

1 | INTRODUCTION

HTML is nowadays being used not only to hold information but also to support businesses that rely on the correctness of HTML documents to reach clients. Thus, this language is being used as an endpoint interface with clients, through which they access operations of all kinds, including entertainment or business. This became particularly true with the advent of HTML5, which adds complex features to the standard including multimedia, device access, or even semantics.[1]

The typical time-to-market constraints of web application development bring in the need for fast development of new websites and fast application of changes to existing sites so that a new or different functionality can be accommodated. The creation of new HTML standards also leads developers to change their web applications to benefit from the latest features and to be up to date with the latest trends. In any case, developers mostly concentrate on assuring correctness of existing functionality and overlook the necessary time for verification and validation activities.[2]

Nowadays, it is common to find websites that, because of small details, do not comply with the standards and fail to be correctly processed by browsers.[3] Sometimes, developers are just trying to take advantage of browser-specific features and overlook standards compliance in return of some benefit, whereas at some other times, they just make mistakes, leaving residual defects on the pages.

Browsers gradually became tolerant to small mistakes, which actually leads developers to further disregard compliance with standards, from the moment that web pages are apparently rendered correctly.[3] The problem is that the web environment is hugely dynamic and browsers are increasingly diverse, being very frequently updated. In this environment, there is no guarantee that a page holding (or prone to hold) residual defects will still work in the next version of a given browser. This is a hard-to-track problem that can have serious consequences on the service being delivered.[2]

In this paper, we present the initial design of an approach for the following: (i) analyzing HTML document characteristics (eg, size, complexity, and number of outgoing links), (ii) validating their conformance to the standards including obtaining detailed information regarding errors or bad practices in each document, and (iii) analyzing the data to extract meaningful problem indicators (eg, frequent errors and errors occurring in documents holding specific characteristics).

We used our approach to carry out an initial assessment of a small set of 1344 web pages, collected their characteristics, and identified an initial set of frequent problems. Despite the reduced size of the experiments, results tend to agree with previous research studies in this area with nearly 90% of the analyzed pages showing at least one type of error. We also observed the prevalence of a reduced number of error types with just 14 different types accounting for 76.91% of all errors found.

Although this kind of work is quite well known in the dependability community,[4] it is far less common in the web research community. In fact, the work that is closer to ours (although with different goals) has been carried out about a decade ago according to the standards that are in disuse now. Thus, there is currently no up-to-date characterization information or a reusable HTML fault model based on large-scale data that can be used by developers or researchers. This kind of information is vital for web developers to understand the reliability of their websites and to help building more reliable sites that can accommodate the presence of small mistakes (eg, introduced by some change to the site).

This paper is organized as follows. Section 2 presents related work, and Section 3 overviews our approach to analyze HTML. Section 4 discusses preliminary results, and Section 5 concludes this paper.

2 | RELATED WORK

More than a decade ago, Chen et al[2] used the W3C validator[5] against 44 340 web sites composed of random sites, popular sites, and sites found using 3 search engines. Only 5% of the web pages analyzed were standards compliant. Top problems included missing DOCTYPE, the omission of the *alt* attribute, and the lack of *type* attribute. Apart from the use of home and inner pages, no filtering (eg, removing 404 error pages) or removing bias (eg, same-source pages) from the URL list appears to have been made.

Ofuonye et al[3] used *Alexa.com* as a source of 100 000 URLs to find only 3% of valid pages with the analysis per geographical origin showing significant differences. The top problems found were the lack of a DOCTYPE declaration, missing character set, and a conflict between the character set declaration and the *meta* attribute. The URL list was built by Internet Explorer users with Alexa's toolbar and could be also biased to regional variants of some sites (eg, Google). There is no description of filtering or removing bias from the list.

The Metadata Analysis and Mining Application[6] analyzed ~3.5 million URLs from DMOZ, Alexa.com top 500, and a set of W3C member companies. Domain parking sites, error pages, and non-HTTP(s) links were filtered out. Characterization included the most and least used attributes, types of encoding, the total number of hyperlinks, and images among others. Results showed that only 4.13% of the URLs were W3C valid.[6] In addition, 51% of the documents included the DOCTYPE declaration with *Strict* flavors displaying higher validation rates when compared with *Transitional* and *Frameset*. XHTML, with 13.4% validation, also fared better than HTML (with 6.6%). Of the 27 possible types of warnings, 14 were signaled. The largest amount of warnings per web page was 5, with 2 being the most common. The 17 most common errors occurred on more than 10% of the total web pages and included a wrong attribute, the absence of a required attribute, missing document type, or a close tag for a not open element. There is much space for improvement at the level of techniques such as bias removal and filtering but, especially on the richness of the results, including the analysis of the HTML documents defects.

Parnas[7] used the WDG HTML validator[8] over 2.5 million URLs gathered from the open web directory. Only 0.71% of pages were found to be valid. The most frequent errors were *no DTD declared*, followed by *nonstandard attribute specified*, and *required attribute not specified*. The most common number of errors per page was 4 with an average of 5.2.

Saarsoo[9] analyzed 1 million pages from the open web directory. Filtering techniques included removing duplicates, empty pages, and URLs not returning 200 OK status codes. The WDG HTML Validator and the W3C CSS Validator[10] revealed only 2.6% of valid web pages (with an average of 6 errors per page, with 3 being the most frequent). The top errors included nonexistent attributes, missing attributes, missing document type declaration, closing not open elements, or elements in a wrong location. Beatty et al[11] found Transitional flavors of HTML to be more frequent than their Strict versions. Pinterits et al[12] also confirmed the low percentage of successful validation. Ganzeli et al[13] analyzed *gov.br* sites and obtained slightly higher values regarding standards compliance.

Website complexity and how it impacts clients experience were studied by Butkiewicz et al.[14] The authors analyzed the complexity of 1700 sites in terms of content using number of objects, types of objects, and downloaded bytes as metrics. They also analyzed the sites complexity in terms of service (eg, distinct servers used and number of nonorigin servers used). Conclusions include that the number of objects and servers are major contributors for page load time (and load time variability). Characterization of the sites is relatively superficial, and the work has an overall focus on performance and does not consider the presence of wrong elements in the HTML code.

Sanders et al[15] also characterized websites, taking into account their adaptation to different client platforms. Among other aspects, the work considers the use of the different HTML tags used to define the site structure (flow, heading and sectioning content, text-level markup, and embedded content) and to set contents and then considers the different types of tags to measure the differences among different versions. It is well known that the contents and structure of web pages are a factor that contributes to inefficiencies in the way sites are loaded by browsers.[16] This can only be aggravated if the web pages hold HTML errors that then need to be tolerated by browsers.

The aforementioned works have the large problem of being at many years of distance with none strictly focusing on HTML5, leaving researchers with no updated information of the current state of the web regarding the characteristics, validity of modern HTML documents, and typical mistakes done by developers.

3 | APPROACH FOR CHARACTERIZING HTML DEFECTS

Preparatory work that serves as input for our approach includes defining an initial set of URLs that point to the HTML documents that will be the target of analysis. The HTML files in the URL set have to be crawled and stored for analysis. The core of the approach consists of the following steps.

1. Analyze the HTML document characteristics (eg, size, complexity, and number of outgoing links).
2. Use state-of-the-art HTML validators to check the HTML code compliance against the standards including detailed information regarding errors or warnings (ie, bad practices) in each document.
3. Analyze the data to extract meaningful problem indicators (eg, frequent errors and errors occurring in documents holding specific characteristics).

The approach steps are currently supported by a prototype tool, which we depict in Figure 1, and this is based on the combined use of free existent tools and custom code. The whole solution is open-source and will be freely available for use and adaptation by other researchers once completed. The next paragraphs explain our approach, mapping its 3 steps to the tool components.

As aforementioned, the input for our tool is a set of URLs provided by the user, which are crawled by **Apache Nutch**[17] in our prototype; the HTML code is extracted and stored locally. The tool then triggers a metrics analysis (*step 1*) and a validity analysis (*step 2*) over the stored HTML files. The **metrics analysis** involves analyzing the HTML documents for characteristics such as document size, number of HTML tags present, and text to HTML ratio among others. To define this set of metrics, we collected the HTML metrics used by several authors in this domain.[2,6,9] The idea of defining a set of HTML metrics is that, by the end of the process, we have the ability of correlating HTML problems with some of the characteristics of the document (eg, document complexity with a wrong tag structure). This is actually an ongoing work, and we intend to extend and refine this set in the future. Table 1 presents the current set of metrics being used.

The second type of analysis refers to the **validation of the HTML code**. With the purpose of selecting 1 or more HTML validators for integration in our approach and tool, we analyzed 11 different validators, including the most well-known

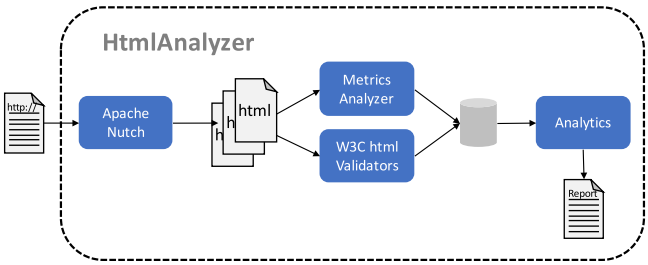

FIGURE 1 Tool overview [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Preliminary HTML metrics for individual pages

Metric	Description
fileSize	HTML file size
textSize	Size of textual content
elemSize	Size of HTML elements
textHtmlR	Ratio of text to HTML elements
version	Version of HTML in use
elemCount	Total number of HTML elements
elemFreq	List of HTML elements and the number of times each of them occurs
linkCount	Total number of hyperlinks
protLFreq	Types of protocols used in hyperlinks and the number of times each of them occurs

TABLE 2 Defect indicators for individual pages

Defect Indicator	Description
errorCount	Number of HTML errors according to the standard
errorTypes	Number of occurrences per type of HTML error
warningCount	Number of HTML warnings according to the standard
warningTypes	Number of occurrences per type of HTML warning

CSE HTML validator, Total Validator, WGD HTML, and the W3C validator, against key features such as the ability to analyze HTML5, validating pages served through HTTPS, maximum file size, and error limits. We opted for using the W3C validators[5] (which are essentially 2 tools, one for HTML5 and another for other versions) because of compliance with the key features and their overall recognition in the industry.

The first task at this validation step is to understand the type of document being handled (eg, an HTML5 document and an XHTML 1.0 document), which should be present at the beginning of each file. The next step is to understand if the document is syntactically valid (ie, no error is found by the validator) or not. If HTML errors are detected (ie, if syntactic rules are broken such as not closing a <*div*> tag or having a wrong name in a tag), we collect detailed information regarding the errors. The W3C validators also produce warnings for less serious mistakes found in the web page (eg, using attributes that are not supported by all browsers or using attributes that can be safely omitted for particular elements).

As the validators produce errors and warnings that are specific for the code being analyzed, our tool is then responsible for identifying the generic type of issue being identified (ie, not closing a tag or a <table> tag should be reported as the same type of error). The identification of generic errors has been manually verified in our prototype tool (for the errors obtained during the experiments described in the next section).

Finally, we analyze the data collected to extract relevant indicators regarding the defects introduced by developers (*step 3*). Currently, we are considering the preliminary indicators displayed in Table 2 mostly based on empirical evaluation and on what was used in previous works.[3,6]

Again, our intention is to enrich this set of indicators and further use the data to obtain a fault model for HTML that holds representative web page defects. Such model can be helpful for verification and validation activities, especially, considering that, nowadays, HTML is being generated at runtime, which makes it more difficult for developers to assure that it will be correct.

4 | PRELIMINARY RESULTS

In this section, we discuss the results obtained during a preliminary experimental evaluation carried out using our tool and using a total of 1344 URLs as input. The URLs were randomly selected from the directory of the web (as in the work of Wilson[6]). In the future, we intend to complement this information with data gathered from known popular web sites. We first analyze the global results and then discuss the results for HTML5, which is the latest standard.

4.1 | Global data set analysis

Table 3 highlights the **overall characteristics** of the data set, as reported by our tool.

TABLE 3 Overview of the HTML data set characteristics

Characteristic	Value
Average file size	33.73 kB
Average size of text content	3.71 kB
Average size of HTML elements	30.02 kB
Average ratio of text to HTML elements	12.36%
Most common HTML version	HTML5 (30.43%)
Average number of HTML elements	342
Most frequent HTML element	<a> (17.56%)
Average number of hyperlinks	60
Most common protocol used in hyperlinks	HTTP (72.35%)

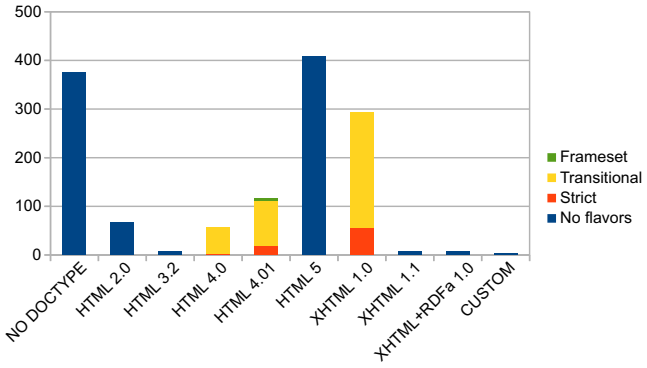

FIGURE 2 Frequency of HTML versions and flavors [Colour figure can be viewed at wileyonlinelibrary.com]

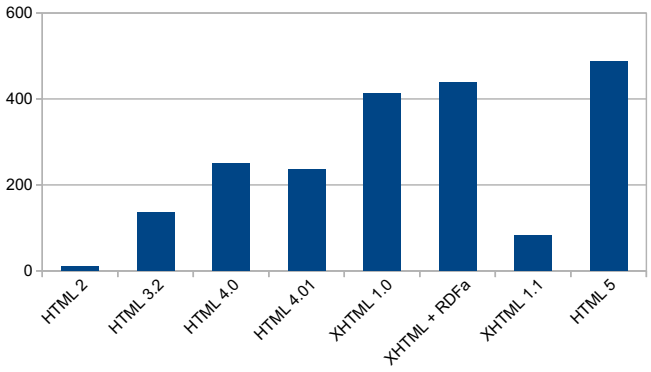

FIGURE 3 Average number of HTML elements per version [Colour figure can be viewed at wileyonlinelibrary.com]

As we can see in Table 3, the average size of text content in a page is about ten times lower than the size of that used by the HTML elements themselves, which suggests the presence of other types of elements (eg, multimedia content). HTML5 is the *most popular version* with the *average number of elements* used per page being quite high (342). The <*a*> tag defines hyperlinks and is the *most frequent element* in the data set. The next paragraphs detail these latter 3 characteristics.

Figure 2 shows the frequency of each version and the flavor of HTML found in the global data set.

As we can see in Figure 2, 28% of web pages omit the version information, leaving this task up to the browsers. HTML 5 is the most common version present in our data set (30% of the pages) followed by XHTML 1.0 (22% of the pages). We can also see (also as in the work by Beatty et al[11]) that the Transitional flavors of HTML are more popular than their Strict counterparts, which is understandable from an ease of development perspective.

It is interesting to observe that the average number of elements per page generally increases with the HTML version; this trend is shown in Figure 3. The low number observed for XHTML 1.1 is likely because of the small number (8) of XHTML web pages in the data set.

Finally, regarding the most common elements, it comes with little surprise that the <*a*> tag is the most used (17.56%), as a web page with no hyperlinks defeats the purpose of the interconnected information of the Web. It was followed by

TABLE 4 Common errors present in the global data set

Error	Frequency
Incorrectly encoded character	13.14%
Absence of a required attribute	9.88%
Omission of a close tag	7.89%
Attribute that does not belong to the standard in use	6.58%
An attribute value not contained in double quotes	6.45%

TABLE 5 Top validation warnings in the global data set

Warnings	Frequency
Using a self-closing tag on a non-HTML5/XHTML file	31.97%
Entity name with a badly encoded special character	18.25%
Attribute name with a badly encoded special character	15.24%
Attribute value with a badly encoded special character	13.56%
Incorrectly encoded character on the page content	8.67%

the <*div*> tag (17.12%), which is a generic container for data or other elements many times used to group elements and apply the same Cascading Style Sheet (CSS) styles to the whole group. The remainder of the most frequent tags is <*li*> (8.99%), which represents a list item; <*span*> (6.70%), which is a generic inline container; and <*img*> (5.30%) the tag for representing an image.

Table 4 summarizes the results for **HTML validation**, focusing on the detected errors. Similar to previous studies, we found the rate of compliance with the standards to be very low with only 11.09% of pages passing validation with no errors. The common web page seems to be plagued with errors with an average of 41 per page. A reduced number of error types (14 different error types) accounts for about 77% of all 56 257 validation errors found.

Regarding validation warnings, the 5 most frequent validation warnings represent almost 88% of the total 19 569 detected. The use of self-closing tags on document types that do not allow it was the most frequent case. The remaining top warnings were related with the use of incorrectly encoded characters in a variety of ways. Table 5 summarizes these results.

4.2 | HTML5 results

Regarding the **HTML5 documents' characteristics**, it is interesting to observe that all the metrics show higher values with the exception of the text to HTML elements ratio. Table 6 summarizes this information.

We observed that the top HTML5 elements used are the same as the ones previously reported for the overall data set. Thus, we again find <*div*> (22,38%), <*a*> (18.55%), <*li*> (11.56%), <*span*> (8.75%), and <*img*> (4.16%).

Regarding the **validation of the HTML5 documents**, we found that 93.6% of the HTML5 pages had at least one error. The average number of errors that we can expect to find in an HTML5 page in our data set is 22. The top 5 errors represent more than 60% of the total number of errors (9250). The top 10 errors would allow us to reach the same value observed for the global data set. Table 7 summarizes the top problems observed.

TABLE 6 HTML5 documents' characteristics

Characteristic	Value
Average file size	53.31 kB
Average size of text content	4.50 kB
Average size of HTML elements	48.80 kB
Average ratio of text to HTML elements	9.22%
Average number of HTML elements	487
Most frequent HTML element	(22.38%)
Average number of hyperlinks	91
Most common protocol used in hyperlinks	HTTP (76.94%)

TABLE 7 HTML5 documents' top errors

Error	Frequency
Inclusion of an attribute not allowed for an element	17.41%
Deprecated attributes related to style	16.86%
Absence of the *alt* attribute for <*img*> elements	12.26%
Element nested in a wrong element type	8.86%
Use of an *itemprop* attribute with no parent	5.23%

TABLE 8 HTML5 documents' top warnings

Warning	Frequency
Use of the deprecated *border* attribute	20.98%
Duplicate IDs	17.64%
Absence of headers	17.59%
Incorrectly described attribute that is not serializable	6.83%
Use of an unnecessary attribute that can be omitted	5.92%

We can see that 4 out of the top 5 problems involve the improper usage of attributes. It is important to emphasize that the errors shown in Table 7 are specific for this version of the standard (ie, they are not shared with the previous versions). In future work, we intend to analyze the different individual errors (provided by the W3C validators for different HTML versions) to further understand their similarities, causes, and effects. Regarding warnings, a total of 1978 were uncovered. The following five, listed in Table 8, refer to about 69% of all occurrences.

HTML5 further reduced (when compared with the previous versions) the number of style and presentation elements, delegating that type of task for CSSs. It is interesting to observe that the second most common error and the top warning both refer to the use of deprecated elements and attributes, which means that their use is still quite high among developers. This can attributed to many factors including lack of knowledge regarding the new standard or the indifference regarding HTML validation among other causes.[3]

Regarding the overall results and the HTML5 results, it is interesting to note that most of the problems are related with attributes, suggesting that the use of HTML elements is less error prone. It is our intention to further analyze the gathered data and run specialized data analytics tools so that we can get further insights over the data (which are extremely difficult to obtain, as they involve complex analysis and correlations).

5 | CONCLUSION

In this paper, we have presented the initial design of an approach and tool for characterizing and analyzing the syntactic validity of HTML documents. Preliminary results already show HTML5 as the leading version closely followed by XHTML 1.0 with the remaining versions appearing relatively less often. Web pages following newer standards also tend to show a higher number of HTML elements with the hyperlink tag being the most frequent element present in the data set.

Regarding the validation results, we found that only about 10% of the pages comply with the standards with the common web page showing, in average, tens of errors. Such low compliance was now also observed in HTML5 documents. Only 14 types of errors account for three-fourths of the errors, ie, a number that is reduced to only 10 types of errors in the case of HTML5. As validation warnings highlight, we may refer the incorrect use of self-closing tags in the global data set or incorrect use of attributes in HTML5 (this latter case is also a large source of errors in HTML5). Obviously, these results may not be generalized as the data set size is relatively small. Still, the results tend to agree with previous research studies, and the rich information overviewed in this paper provides strong indications that there is valuable research to be conducted in this topic.

As the first step of a larger experiment, besides enlarging the data set to the million scale, we intend to enrich the HTML documents characteristics and run a wider set of analyses over the data. Besides providing an up-to-date vision of the use of HTML on top websites, we intend to further understand the root causes of the errors and actually use errors for understanding how reliable, in the presence of changes, a particular website is. Such information is relevant for website developers (for training or applying website changes) but might also be helpful for browser development and optimization.

ACKNOWLEDGEMENTS

This work has been partially supported by the project EUBra-BIGSEA, funded by the European Commission under the Cooperation Programme, Horizon 2020 grant agreement 690116, and Project MobiWise: from mobile sensing to mobility advising (P2020 SAICTPAC/0011/2015).

ORCID


Nuno Laranjeiro http://orcid.org/0000-0003-0011-9901

REFERENCES

1. World Wide Web Consortium. HTML5. https://www.w3.org/TR/html5/. Published 2014 [Accessed on 5 May 2017].
2. Chen S, Hong D, Shen VY. An experimental study on validation problems with existing HTML webpages. Paper presented at: Proceedings of the 2005 International Conference on Internet Computing, ICOMP'05; 2005; Las Vegas, USA.
3. Ofuonye E, Beatty P, Dick S, Miller J. Prevalence and classification of web page defects. *Online Inf Rev*. 2010;34(1):160-174.
4. Durães J, Madeira H. Emulation of software faults: a field data study and a practical approach. *IEEE Trans Softw Eng*. 2006;32(11):849-867. https://doi.org/10.1109/TSE.2006.113
5. World Wide Web Consortium. The W3C Markup Validation Service. https://validator.w3.org/. Published 2017 [Accessed on 15 May 2017].
6. Wilson B. MAMA - Metadata Analysis and Mining Application. http://maqentaer.github.io/devopera-static-backup/http/dev.opera.com/articles/view/mama/index.html. Published 2008 [Accessed on 15 May 2017].
7. Parnas D. How to cope with incorrect HTML. MSc Thesis. Bergen, Norway: University of Bergen, Norway; 2001.
8. Quinn L. WDG HTML Validator. http://www.htmlhelp.com/tools/validator/. Published 2007 [Accessed on 15 May 2017].
9. Saarsoo R. Coding Practices of Web Pages. http://triin.net/2006/06/12/Coding_practices_of_web_pages. Published 2006 [Accessed on 15 May 2017].
10. World Wide Web Consortium. The W3C CSS Validation Service. https://jigsaw.w3.org/css-validator/. Published 2009 [Accessed on 15 May 2017].
11. Beatty P, Dick S, Miller J. Is HTML in a race to the bottom? A large-scale survey and analysis of conformance to W3C standards. *IEEE Internet Comput*. 2008;12(2):76-80.
12. Pinterits A, Treiblmaier H, Pollach I. Environmental websites: an empirical investigation of functionality and accessibility. *Int J Technol Policy Manage*. 2006;6(1):103-119.
13. Ganzeli HdS, Bressan G, Moreiras AM. ICT web: Analysis of the Brazilian governmental web. Paper presented at: Proceedings of the 18th Brazilian Symposium on Multimedia and the Web; 2012; São Paulo, Brazil. http://dl.acm.org/citation.cfm?id=2382715
14. Butkiewicz M, Madhyastha HV, Sekar V. Understanding website complexity: Measurements, metrics, and implications. Paper presented at: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference. IMC '11; 2011; New York, NY. https://doi.org/10.1145/2068816.2068846
15. Sanders S, Sanka G, Aikat J, Kaur J. *The influence of client platform on web page content: measurements, analysis, and implications*. Web Information Systems Engineering – WISE 2015. Lecture Notes in Computer Science. Cham: Springer; 2015:1-16. https://doi.org/10.1007/978-3-319-26187-4_1
16. Wang XS, Krishnamurthy A, Wetherall D. Speeding up web page loads with Shandian. Paper presented at: 13th USENIX Symposium on Networked Systems Design and Implementation; 2016; Santa Clara, CA.
17. Apache. Apache Nutch. http://nutch.apache.org/. Published 2017 [Accessed on 15 May 2017].

How to cite this article: Mendes J, Laranjeiro N, Vieira M. Toward characterizing HTML defects on the Web. *Softw Pract Exper*. 2018;48:750–757. https://doi.org/10.1002/spe.2545