Pave way for Semantic Digital Libraries: HTML Validity Check for Homepages of Iranian Digital and Electronic Libraries

Pegah Tajer[1] and Alireza Nikseresht [2]
[1] Department of Knowledge and Information Science, Marvdasht Branch, Islamic Azad University, Marvdasht, Iran
[2] PhD student of Department of Computer Science Engineering and IT, School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran

Abstract. Semantics and accessibility are naturally part of HTML by design. Therefore, there is a must to significantly leverage the standard markup language and write the cleanest code possible.
The purpose of this work is performing HTML validity check in terms of compliance with W3C standards for home pages of Iranian digital and electronic libraries in order to pave way for applying semantic technologies in Iranian Library websites.
This is a comparative survey. 82 library software-based digital libraries and 68 electronic libraries were identified via Google search in the domains of .ir, .ac.ir, .org and .com. Data collected via an automatic tool developed by authors in which W3C Markup Validation Service is used for validation tests. Besides, total tags of URLs are counted through it to be able to normalize HTML error frequencies for each home page. Data were analyzed using nonparametric tests via IBM SPSS Statistics 20.
Validation test results in about 50 percent failure rate. 38 out of 70 valid libraries are digital libraries and 32 are electronic ones. Mean of HTML error has dramatically decreased in 2017 in comparison with 2015; however Standard Deviation has increased sharply. Moreover, the distribution of HTML error scores is similar across digital and electronic libraries.
Regarding conformance to HTML standards, it seems essential to revise the web design of Iranian digital and electronic libraries to fulfill semantic web potentialities.

Keywords: Semantics, digital libraries, electronic libraries, Iran, W3C validation service, HTML

1. Introduction

Semantic Web technology is defined as a method of linking data between systems or entities that allows for rich, self-describing interrelations of data

Received: 14.3.2017 / Accepted: 19.12.2017 ISSN 2241-1925

available across the globe on the web (LinkedDataTools, 2005 in Abel, 2016). This means that the web is a combination of the existing hypertext-markup language (HTML) contents and contents from computer generated programming software (Abel, 2016).

Semantics have great value for digital libraries hence could provide unambiguous meaning within content, accessibility, search, and interoperability. Semantics and accessibility are naturally part of HTML by design. Therefore, there is a must to significantly leverage the standard markup language and write the cleanest code possible (Learn to Code Advanced HTML & CSS).

Nowadays there are several web standards developed by corresponding organizations that concern interoperability, accessibility and usability of web pages. A web page which is described as complying with web standards, generally has valid HTML, CSS, JavaScript, RSS or Atom news feed, RDF, metadata, XML, object embedding, script embedding, browser- and resolution-independent codes, and proper server settings. Therefore, conformance to web standards is a good way for writing clean code.

Although there is a global trend for well-designed websites that deliver standards-compliant pages, the quality of web pages in terms of conformance to web standards is really a big challenge in countries that no rules exist for developing valid and accessible webpages. Therefore, it is necessary to study the properties of websites developed in those countries to pave the way for authoring standard webpages to meet web technologies specifically semantics.
Digital libraries as primary centers for information services in Iran, regarding the second law of library science (Every reader his [or her] book) (Ranganathan, 1931) and its web version (every user has or his web resource) suggested by (Noruzi, 2004), are expected to offer valid and accessible sites. Digital libraries could not meet their mission without complying with web standard. Moreover, for moving forward with web 2.0 and 3.0 technologies and providing semantic digital libraries there is a must to develop standard digital libraries hence valid HTML codes will provide a stronger footing for reaching this goal (Breeding, 2006).

This work addressed the comparative survey to perform HTML validation test for Iranian digital and electronic libraries. The main objective in this study is to present a distribution of HTML errors with home pages of Iranian digital and electronic libraries. More specifically, this paper aims at answering the following research questions:

1- Do Iranian digital and electronic libraries' home pages pass HTML validation test?
2- What is the rate of HTML errors on homepages of Iranian digital and electronic libraries?

3- How do Iranian digital libraries' home pages differ from electronic ones in terms of the distribution of HTML error scores?

The rest of this paper outlined as follows. After the introduction, section 2 presents the literature review of the study. Section 3 deals with the research methodology while findings exposed in the next section. Section 5 presents conclusion and future works. Finally section 6 serves as acknowledgment.

2. Literature Review

Improving effective web design through authoring websites which are compliance to web accessibility and usability standards is a paramount concern in Human-Computer Interaction research. There are several studies related to the accessibility and usability evaluation of websites regarding W3C guidelines which checked HTML and CSS codes as foremost indicators of accessibility mentioned in accessibility guidelines (Tajer, 2014), (Oud, 2012). Some of them are as follows.

The experience of web designer's usage of three tools consisted of WatchFire Bobby, W3C HTML validator and UsableNet life to evaluate and improve the usability of different Websites was performed by Ivory and Chevalier (2002). Authors showed that these tools help Web designers to identify a large number of potential problems in the Website. Furthermore, Abanumy, Al-Badi and Mayhew (2005) studied Website accessibility guidelines, website accessibility tools and the implication of human factors in the process of implementing successful e-Government websites. Saudi Arabia and Oman e-Government websites were manually checked for compliance with W3C's WCAG guidelines using a checklist made for this purpose. Then Multiweb, LYNX and W3C validator service were used for evaluation. An email survey was performed to explore the reasons behind the lack of accessibility and usability of e-Government websites. The email survey was sent to the webmasters of the government's websites. Finally researchers made recommendations for improvement of e-Government website accessibility based on findings. Pribeanu and et all (2012) reviewed Municipal web sites in Romania based on automated accessibility checking. Two web pages for each web site included the home page were selected and evaluated against WCAG 2.0 accessibility level A. The analysis of results reveals a relatively low web accessibility of municipal web sites.

Regarding usability of webpages, Atterer (2008) presented an approach for improving automated usability tool support during the development of a website, where the HTML code analyzer is applied to each web page in the website in order to detect potential problems. Author also used an automatic validator to verify usability guidelines, and as a result the researcher presented a prototype of a model-based automatic usability validator. Al-Ananbeh, and et all (2012) automatically evaluated eighty Arab university websites in their study. The main objective of study was to discover the relevancy between usability and search engine optimization based on HTML errors checking, load time, and browser compatibility problems using HTML ToolBox, PageRank Checker, and SEO PageRank. The study concluded that while usability is very important to Websites in many aspects, however, it does not necessarily improve SEO.

Some studies focused on library websites in the literature while the largest number of them focused on the accessibility of library resources such as databases and online catalogues interfaces instead (Oud, 2012). Basically they focused on University, college, public, national, and digital library web sites ((Lilly & Van Fleet, 2000), (Parker, 2001), (Schmetzke, 2001), (Kirkpatrick, 2003), (Schmetzke, Greifeneder, Comeaux, & Schmetzke, 2007), (Comeaux & Schmetzke, 2013), (Providenti & Zai III, 2007), (Mohammad Esmaeil & Kazemi, 2011), (Zarei, 2012), (Cervone, 2013)). Khan, Idrees and Mudassir (2015) assessed the accessibility of library Web sites of top ten universities of Pakistan using Web Accessibility Evaluation Tool (WAET) to examine compliance of the library Web site with Web Content Accessibility Guideline 2.0. The study also explores commonly identifies accessibility barriers in the subject Web sites. Result shows that 70 per cent of library Web sites do comply with W3C standards. However, important accessibility issues still exist in the subject Web sites.

Few works specifically focus on rate of HTML errors. Chen, Hong and Shen (2005) performed an experimental study on the validation problem of existing HTML pages in the World Wide Web. This research was resulted in about 5% of validation. Chambial and Sharma (2011) focused on an experimental study on the validation problem of top Indian websites. It is revealed that there are very few valid webpages and most of them are invalid.. Furthermore, Doulani, Hariri and Rashidi (2013) compared the quality of Iranian and British university web designs. Authors Extracted components of websites, validated web pages, and identified broken links. The 5point scale used for evaluation. The study results Iranian university websites have high rate of errors in comparison with British university websites.

It is worth mentioning that only one study focused on the rate of HTML errors on library websites. Breeding (2006) validated the home page of 123 members of the Association of Research Libraries (AEL), 136 members of Urban Libraries Council, commercial ILS vendors' sites, and some other commercial websites like Amazon, Yahoo, and Flicker by W3C tools. The survey results more than 50 percent failure rate.

All in all, it seems that there is a lack of works in a literature which address the HTML validation of digital libraries' websites. Thus, we decided to perform a survey on Iranian digital and electronic libraries. Results of this survey may flip web developers who especially work on digital and electronic libraries' projects to review HTML codes regarding to web standards and lead to notify library managers to make short-term and long-term plan to policy accessibility and usability programs in order to fulfill semantic web potentialities.

3. Methodology

Regarding the purpose of our research, this is an applied study tries to alert web developers to improve the web design of library webpages and pave way for semantic ones. It is a non-experimental descriptive study and also a comparative survey from the data collection standpoint and it compares the characteristics of two groups.

In this survey, Iranian digital libraries are the library-software based websites yet; electronic libraries are the ones which provide browsing and downloading e-books. Moreover, they are not professional library organizations indeed. For identifying the statistical population Persian equivalents of “digital library”, “electronic library” and “e-book download” were searched via Google in domains of .ir, .ac.ir, .org and .com. The first 100 out of total results for each search formula revised. To ensure that retrieved websites remained active and also for keeping the statistical population up to date, this identifying approach repeated three months later. Finally 149 URLs consist of 82 digital and 67 electronic libraries were recognized. HTML validation tests were performed for all of them.

As W3C is an internationally known institute establishes web standards and since its validation tools are not only free of charge but also easy to use, W3C Markup Validation Service was used to discover errors in this survey.

W3C Markup Validation Service also known as HTML validator is a validator that provides validating HTML, XHTML, SVG or MathML codes. Validation could be done via three options provided by W3C tools which consist of Validate by URL, File Upload and Direct Input. Validate by URL was used for this project in January 13, 2017.

The bigger the web page, the more tags and it seems the more errors. Thus, for normalizing the number of HTML errors, we had to control the effect of their size. So we developed software using C# which is able to count the total tags of each home page. Tag counts of some URLs were not allowed and so our tool reported error for them. Therefore, total tags of 102 out of 149 URLs were counted. Thus, our available samples consisted of 57 digital libraries and 45 electronic ones. We calculated HTML error score for each URL by dividing the number of HTML errors by URL total tags.

Lastly, IBM SPSS Statistics 20 was used for statistical analysis. Since Kolmogorov-Smirnov Test was statistically significant ($p<0/05$), the distribution of HTML error rates and scores were skewed (Table.1). Thus, Mann Whitney-U was performed to compare the mean score of errors between two groups. For our second question, we were dealing with the observed frequencies of HTML errors; therefore, Chi-Square Test was carried out.

Table 1. One-Sample Kolmogorov-Smirnov Test for the distribution of HTML error rates and scores

		HTM. error rates	HTML error scores
N		149	102
Normal Parameters[a,b]	Mean	.03500	.03500
	Std. Deviation	.081508	.081508
Most Extreme Differences	Absolute	.334	.334
	Positive	.312	.312
	Negative	-.334	-.334
Test Statistic		.424	3.371
Asymp. Sig. (2-tailed)		.000[c]	.000

4. Findings

4.1 Do Iranian digital and electronic libraries pass HTML validation Test?

The home page of digital and electronic libraries like all other websites is not only the initial but also the main page. Therefore, HTML validation test was performed for home page URLs of all 149 digital and electronic libraries identified on a pass-fail basis. Validation test results in about 50 percent failure rate. 38 out of 70 valid libraries are digital libraries and 32 ones are electronic libraries.

It is worth mentioning that based on previous survey done, failure rate was more than 90 percent in 2015 (Tajer, 2015). Table2 shows the distribution of valid libraries in 2015 and 2017.

Table2. The distribution of valid libraries based on HTML validation test in 2015 and 2017

Valid Libraries by year	Frequency	Percentage
2015	3	2%
2017	70	47%

4.2 What is the rate of HTML errors on homepages of Iranian digital and electronic libraries?

Table3 shows descriptive statistics of HTML error rates for Iranian digital and electronic libraries' homepages. Chi-Square Test revealed that there is a statistically significant difference (p<0/05) between the frequency distributions of HTML errors across 149 digital and electronic libraries. Expected frequency of HTML errors assumed equal for Chi-Square Test (Table3).

Table3. Statistics of HTML error rates for digital and electronic libraries

N	Mean	Std. Deviation	Minimum	Maximum	Chi-Square	Df	Sig.
149	47.50	105.958	0	892	1939.443[a]	61	.000

In comparison with 2015, Mean of HTML error dramatically decreased in 2017; however Standard Deviation increased sharply in 2017 (Table4).

Table4. Comparison of Mean and Standard Deviation of HTML error rates for digital and electronic libraries in 2015 and 2017

Mean		Standard Deviation	
2015	**2017**	**2015**	**2017**
124.24	47.50	23.08	105.958

4.3 How do Iranian digital libraries differ from electronic ones in terms of the distribution of HTML error scores?

In order to be able to compare two groups of homepages, as discussed in methodology section we calculated HTML error scores based on their total tags. Therefore, calculated HTML error score is a digit between 0 and 1. Table 5 shows descriptive statistics of HTML error scores for libraries' home pages.

Table5. Descriptive statistics of HTML error scores for digital and electronic libraries

N	Mean	Std. Deviation	Minimum	Maximum
102	.03500	.081508	.000	.629

Findings reveal that Mean Rank of HTML error scores for digital libraries (Mean Rank= 54.06) is slightly higher than electronic ones (Mean Rank= 48.26). Moreover, Mann-Whitney U test states that two groups are not statistically significantly different (p-value>0.05). Thus, the distribution of HTML error scores is similar in both groups of homepages (Table6).

Table 6. Mann-Whitney U test

	HTM error scores
Mann-Whitney U	1136.500
Wilcoxon W	2171.500
Z	-1.000
Asymp. Sig. (2-tailed)	.318

5. Conclusion and future works

This survey results in about 50 percent failure rate in HTML validation of Iranian digital and electronic libraries' home pages meanwhile this rate was 90 percent based on previous survey done in 2015 (Tajer, 2015). Thus, it seems that Iranian library web developers have tried to write cleaner HTML codes. In addition, 38 out of 70 valid libraries are digital libraries and 32 are electronic ones. Mean of HTML error has dramatically decreased in 2017 in comparison with 2015; however Standard Deviation has increased sharply.

The distribution of HTML error scores is similar across digital and electronic libraries. Since Iranian digital libraries are professional organizations that consist of several scientific and technological committees are expected to pay more attention to web standards. It seems that establishing validation working groups for these institutions is vital. In other words, in comparison with non-professional ones like e-book providers' websites it is expected that such organizations do dramatically and significantly better in terms of writing HTML codes.

This work has focused on libraries' home pages. For investigating deeper, we will design a project to capture HTML error codes of not only homepages but also all URLs the whole library web sites in the near future. Furthermore, we will try to classify common HTML problems on Iranian library websites.

All in all, the practice of writing semantic and accessible code is growing; however adoption at large has not yet been achieved. Thus, regarding conformance to W3C standards specifically for HTML codes, it is essential to revise the web design of Iranian digital and electronic libraries in order to be able to make an efficient use of web technologies such as semantics in the near future.

Acknowledgement

This paper is fully funded by Islamic Azad University (Contract Code: 8337/92/د, 92/11/16 مورخ). I wish to profoundly thank Research and Technology Deputy of Islamic Azad University - Marvdasht Branch.

References

Abanumy, A., Al-Badi, A., & Mayhew, P. (2005). e-Government Website accessibility: in-depth evaluation of Saudi Arabia and Oman. *The Electronic Journal of e-Government, 3*(3), 99-106.

Abel, E. E. (2016). Overview of Semantic Web Technology: The Formulation of Semantic Web Agent System Model to Assist the Blind and Visually Impaired. *International Journal of Science and Technology*, *6*(1).

Al-Ananbeh, A. A., Ata, B. A., Al-Kabi, M., & Alsmadi, I. (2012). Website Usability Evaluation and Search Engine Optimization for Eighty Arab University Websites.

Atterer, R. (2008). *Model-based automatic usability validation: a tool concept for improving web-based UIs.* Paper presented at the Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges.

Breeding, M. (2006). Web2.0? Let's Get to Web 1.0 First. *Computers in Libraries, 26*(5), 30-34.

Cervone, H. F. (2013). Selected practices and tools for better accessibility in digital library projects. *OCLC Systems & Services: International digital library perspectives*, *29*(3), 130-133.

Comeaux, D., & Schmetzke, A. (2013). Accessibility of academic library web sites in North America: Current status and trends (2002-2012). *Library Hi Tech, 31*(1), 8-33.

Khan, A., Idrees, H., & Mudassir, K. (2015). Library Web sites for people with disability: accessibility evaluation of library websites in Pakistan. *Library Hi Tech News*, *32*(6), 1-7.

Kirkpatrick, C. H. (2003). Getting two for the price of one: accessibility and usability. *Computers in Libraries, 23*(1), 26-29.

Learn to Code Advanced HTML & CSS. "Lesson 10: Learn to Code Advanced HTML & CSS", http://learn.shayhowe.com/advanced-html-css/semantics-accessibility/

Lilly, E. B., & Van Fleet, C. (2000). Measuring the accessibility of public library home pages. *Reference & User Services Quarterly*, 156-165.

LinkedDataTools (2005). "Introducing Linked Data And The Semantic Web", http://www.linkeddatatools.com/semantic-web-basics.

Mohammad Esmaeil, S., & Kazemi, S. (2011). Accessibility of Websites of National Libraries in the Middle East. *FASLNAME-National Library, 22*(1), 56-68.

Noruzi, A. (2004). Application of Ranganathan's Laws to the Web. *Webology, 1*(2).

Oud, J. (2012). How Well Do Ontario Library Web Sites Meet New Accessibility Requirements? *Partnership: the Canadian Journal of Library and Information Practice and Research, 7*(1).

Parker, A. (2001). *Measuring the Accessibility of New Zealand Polytechnic Library Home Pages: A Report on Research in Progress*. Paper presented at the NZARE Conference, NZARE Conference. http://repository.digitalnz.org/system/uploads/record/attachment/212/library_website_accessibility__a_case_study.pdf

Providenti, M., & Zai III, R. (2007). Web accessibility at Kentucky's academic libraries. *Library Hi Tech, 25*(4), 478-493.

Ranganathan, S. R. (1931). The five laws of library science. Madras: Madras Library Association.

Schmetzke, A. (2001). Web accessibility at university libraries and library schools. *Library Hi Tech, 19*(1), 35-49.

Schmetzke, A., Greifeneder, E., Comeaux, D., & Schmetzke, A. (2007). Web accessibility trends in university libraries and library schools. *Library Hi Tech, 25*(4), 457-477.

Tajer, P. (2014). *Web Design Validation for Digital Libraries: A Must!* Paper presented at the 1th National Conference on Digital Libraries (INCDL), IRANDOC, Tehran, 185-195.

Tajer, P. (2015). *Web Design Validation for Iranian Digital and Electronic Libraries*. Paper presented at 7th Interntional Conference on Qualitative and Quantitative Methods in Libraries (QQML2015), IUT Universite Paris Descartes, Paris, France.

W3C Quality Assurance Tools. "About W3C QA Tools". http://www.w3.org/QA/Tools/ (Retrieved Jun-12-2014).

Zarei, H., Bagheri Garmaroudi, Forouzan. (2012). Website Accessibility Evaluation of Central Libraries of 20 Superior Universities Associated with Ministry of Science

,Research ,and Technology Based on World wide Web Consortium Guidelines (W3C). *Journal of Epistemology, 5*, 49-64.