Original scientific paper

The Lifespan of Web references: An Example in Graduate Papers at the Department of Information Sciences in Zagreb

Anina Bauer
Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
anina.bauer@gmail.com

Đilda Pečarić
Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
dpecaric@ffzg.hr

Summary

An increasing amount of students and scholars use web references as a prime source in their papers. The main concern is that those references have a short lifespan.

In retrospect to that, the aim of this article is to show how many of the web references, gathered in a corpus of 1947 web references within 362 graduate papers at the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb in the period from 2003 to 2010, are no longer active, i.e. accessible. It was also substantial to know how many of the web references have disappeared in the period between two researches.

The main hypothesis is that there has been an increase in the number of inactive web references within the aforementioned corpus. It is expected that the highest number of inactive web references is present in the graduate papers that have been written during the first half of the analysed period, i.e. more time has passed since those references have been accessed. As well as the availability of the web references, certain bibliographic data were also analysed, where it is predicted that according to the type of web references web pages are more likely to be inactive than any other types such as scientific articles, etc. In response to that, it is expected that the web sites of inactive web references are non-scientific in snature. The analysis of the presence of authors and publishers among inactive web references is also included.

Key words: web reference, lifespan, information sciences, graduate papers

Introduction

An increasing amount of students and scholars use web references as a prime source in their papers, so it has become necessary to analyse them, especially because those references have a short lifespan. The lack of constant availability of those references may be due to infrequent updating of the web sites on which the references are located, or simply because the topic in the reference is outdated and no longer useable.

One of the reasons why the web references tend to disappear has to do with the absence of a universal definition of web references[1]; disparity in the division of types of web references[2]; inability to, accurately, estimate their lifespan[3]. Even though most web references disappear within two or three years since they have been placed online, not knowing the exact date of their removal from the Internet complicates the calculation of the average lifespan.[4]

In response to the aforementioned findings, a prior research of web references in graduate papers at the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb (Bauer, Kirin, Turković, 2011) has led to a secondary research regarding the lifespan of the web references used in the initial corpus.

The aim of this article is to point out a trend of 'disappearing' web references, which are used in scientific papers. The results show how many of the web references are no longer active, whether they have been relocated or completely removed from the Internet since accessing the web references during the former research regarding the amount of used web references until the carrying out of the research this article is dealing with.

The main hypothesis is that there has been an increase in the number of inactive web references within the aforementioned corpus since the conduction of the first research, especially because more than a year has passed. By observing the

[1] Halpin, H. (2011), p. 154 and Spinellis, D. (2003), p. 72.: Is it URI (Uniform Resource Identifier) which is defined as a unique identifier of various sources on the Internet, or should one use one of his subsets such as URL (Uniform Resource Locator) for the location or URN (Uniform Resource Name) for the name of the source, or is a completely new system of identifying in order.

[2] Meyer zu Eissen, S.; Stein, B. (2004), p. 2 and Santini, M. (2007), p. 4: Most frequent is the classification according to the type of web references, as well as the origin. Since there is a number of classification schemes which allow a high level of flexibility in defining web references, one can have so called unsorted types of references, which can result in incredible sources.

[3] Research of Ashenfelder, M. (2011) and *What is the Average Lifespan of a Web Site* (2012): Even though most authors agree that the average lifespan of web references is between 44 and 75 days, one most notice that that includes isolated, personal or business web pages, not to mention a big amount of malware pages which drastically reduce it.

[4] *What is the Average Lifespan of a Web Site* (2012).

lifespan of the web references, one can see how fast the information becomes outdated, not to mention whether the information itself alters or if it simply relocates. The research included an analysis during a longer period from 2003 to 2010, where one expects that the highest number of inactive web references can be found in the graduate papers that were written during the first half of the analysed period, i.e. more time passed since those references had been accessed. It was also interesting to see whether the type of the web references, their location and the presence of authors and publishers influence their lifespan. Expectations include that web pages are more likely to be inactive than any other types such as scientific articles or records within electronic databases, that the references which are located in web sites of commercial, non-scientific nature have a shorter lifespan, and that most inactive web references have an author or publisher listed.

Among other things, this research should provide answers in which domain can the web references that disappear most frequently be found, and also according to the criteria that the active web references meet, what can be done to insure their longevity?

Methods

The research was based on the analysis of data that was collected in the previous survey regarding the use of web references in graduate papers at the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb conducted by the end of June in the year 2011. The corpus included a set of 362 graduate papers, presented in the period between 2003 and 2010, that were available online in the Digital Repository of the Faculty Library, from which 1947 web references were extracted. Since the graduate papers were written by the students from the Department of Information Sciences, disciplines included were: Archive and Documentation Science, Library Science, Information Science and Museology.[5]

While the main object of the preceding study was to find out the amount of web references students use in their graduate papers, the following research formed around discovering which of the used web references were no longer active. The aforementioned study was carried out by the end of January in the year 2013 using the data that had been previously manually entered into a Microsoft Access database containing web references in graduate papers at the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb.

The availability of the web references was determined according to the URL addresses that were used for the citation of references in the graduate papers. The results show the state of the availability of the used web references, which

[5] see Bauer, A.; Kirin, M.; Turković, M. (2011), p. 70.

was followed by the analysis of results per access year (when the web references were used), as well as the analysis of the type of web references, web sites and the presence of authors and publishers among inactive web references. The general assumption of the availability of the web references and the per year analysis includes the comparison of data recorded in June 2011 and January 2013, while the distinction according to bibliographic data deals with the data recorded in January 2013.

The Lifespan of Web references

For the purpose of this research, 362 graduate papers from the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb that included 6525 references, from which 1947 (29.8%) were web references, were analysed.[6]

Even though the type of web references was not identified for 36.7% of them, for the purpose of this research, it was not important whether they had adequate bibliographic data or not (further results lay out the assumptions regarding their type), but whether the URL address, which was listed in the papers, was still valid.

In retrospect to that, the main goal was to find out how many of the analysed web references have become inactive once the first research (June 2011) was carried out, and how many more have become inactive until the conduction of the second research (January 2013).

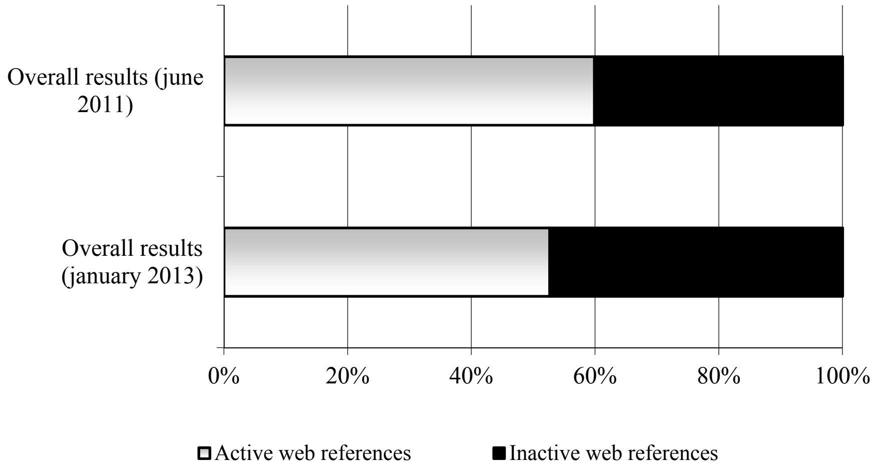

Figure 1. The overall percentage of active and inactive web references

[6] A. Bauer, M. Kirin, M. Turković (2011), p. 70.

As seen in Figure 1, the results show that by the end of June 2011, 781 (40.1%) web references were unavailable, while by the end of January 2013, the number of inactive web references went up to 922 (47.3%).

Considering that the lifespan of web references tends to shorten with time, one expects that there has been an increase in inactive web references by the time of the conduction of this research, and most of them are expected to have disappeared, i.e. been removed from the Internet.

The hypothesis has been confirmed since the number of web references that became inactive in between the two studies has gone up 7.2%, and most of the 922 inactive web references disappeared or became unavailable, while just a few were either relocated, or in some way altered.

Availability of Web references per Year of access

The following included an analysis of the availability of the web references according to the year in which they were accessed. Since not all of the web references have a date of access listed, the year in which the papers, where the references are used, were presented, i.e. published, is used as the access year.

In the first research, among other findings, it was listed how many web references were used in which year. On top of that, for this research it was also included how many of them have been inactive by the end of June 2011 and by the end of January 2013, per year of the analysed period.

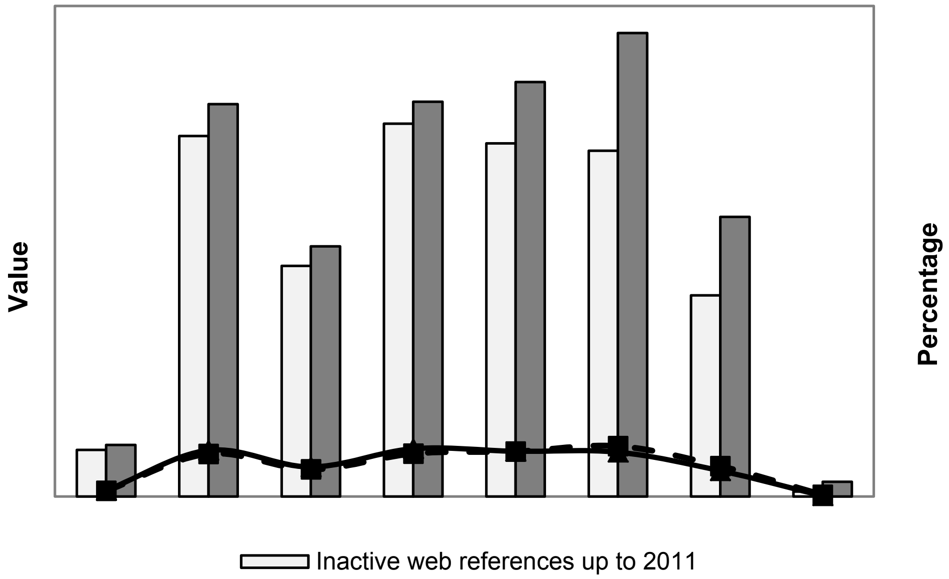

Figure 2. The change of inactive web references (from 2011 to 2013)

As depicted in Figure 2, there were more inactive than active web references in the period from 2003 to 2006, respectively, there were more active than inactive web references in the period from 2007 to 2010. Regarding the change in their availability, a higher number of previously active web references in the period from 2007 to 2010 have become inactive by the end of January 2013.

The premise was that the highest number of inactive web references is present in the first half of the analysed period. For instance, since the web references used in graduate papers in the year 2003 were accessed, until their availability was checked in the year 2013, ten years have passed, so one expects that most of them have disappeared.

The given assumption has been confirmed, i.e. there are more inactive than active web references from the first half of the analysed period, and also more web references that were active from the second half of the period became inactive in between the two studies, which indicates the rapid growth of inactive sources.

Inactive Web references according to the type of references

The data which was gathered for the second research (January 2013) was further analysed according to the type of inactive web references. The assessment of the type was based on the visual attributes of the web site where the reference was found (in some cases it was evident if there had been an article or some other electronic document on the web site). Another way to assess the type of references was possible by examining the bibliographic data provided in the graduate papers, and the type was identified with certainty for those references that had been active during the first research (June 2011).

If we categorise publication into monographs and periodicals, among 922 inactive web references there are 585 (63.4%) monographs, apropos 326 (35.4%) periodicals, while 11 (1.2%) of inactive web references were unidentifiable. Exactly 141 (15.3%) of inactive web references can be accurately identified by their type, since those are the ones that had been active by the end of the first research.

Table 1: Categorized types of monographs and periodicals (January 2013)

Monographs			**Periodicals**		
Type	**No.**	**%**	**Type**	**No.**	**%**
Web Pages	468	80.0	Articles on Web Pages	298	91.4
Documents /Databases on Web Pages	55	9.4	Scientific Articles	21	6.4
Legislation / Guidelines / Standards	41	7.0	Entries / Articles in Dictionaries	7	2.2
E-books / Manuals	15	2.6	-	-	
Lectures / Videos	6	1.0	-	-	
Total	**585**		**Total**	**326**	

The reason why there are so many monographs among inactive web references is that most of them are isolated web pages, which comprise 468, i.e. 80.0% of all monographs. On the other hand, articles on web pages comprise 298, i.e. 91.4% of all periodicals. The representation of the remaining inactive web references is shown in Table 1.

The assumption regarding the relation between the lifespan and the type of web references indicates that the isolated web pages have a shorter lifespan than any type of articles or other documents.

This was confirmed, as the majority of inactive web references are individual web pages, followed by the articles found on web pages. All other types are represented at a much lower rate.

Analysis of inactive Web references with respect to the type of Web sites

There was also a distinction regarding the types of web sites on which the inactive web references were found. The majority of them are too diverse in order to be classified into one group. They make 432 (46.8%) of all inactive web references, and most of them are commercial, non-scientific in nature. On the other hand, there are 263 (28.5%) web sites for the educational, cultural and scientific institutions while 223 (24.2%) web sites are considered to be undefined.

Interesting findings include the presence of four domains that were on sale, which is why the previous reference cannot be found on the given location. It was also noted that some of the frequently visited web sites have inactive web references, such as the web site for the National and University Library in Zagreb which included 32 inactive web references, the web site for *Narodne novine* (The Official Gazette) with 24 of them, and online encyclopaedia Wikipedia with 7 inactive web references. However, the majority of those references were relocated, but without a redirection from the former URL address which doesn't make them easily accessible.

It was expected that the majority of inactive web sites would be non-scientific, seldom visited by the users. The given hypothesis was partially confirmed since most of inactive web sites are commercial in their nature. However a vast number of web sites for educational, cultural and scientific institutions was not foreseen, as those web sites should be frequently updated and maintain the access to their references whether they are archived or not, since they are scientific in their nature.

The Presence of Authors and Publishers among inactive Web references

When it comes to the presence of authors and / or publishers among inactive web references, one expects that most of them do not have the aforementioned bibliographic data listed, which would be one of the reasons why they have a shorter lifespan.

As was anticipated, majority of the web references do not have an author, exactly 657 (71.3%) of them, and even more do not have a publisher listed, that is 848 (92.0%) of inactive web references.

Possible Solutions

There is a whole range of possible solutions to how to slow down the process of the disappearance of the scientific literature found online. Firstly, one can improve the citing system by implementing stricter, more formal obligation, and also, during the process of placing the references online, one should be required to list all the necessary bibliographical data.[7] As well as that, it is suggested to list more meta-data so that the search engines can index them more frequently and comprehensively.[8]

Thus far, the least complicated would be to archive outdated sources, that is simply relocate the reference, at the same time making it just as accessible through a redirection (the user is led to another location for the reference).[9] Also, there is a need to evaluate the usability of formats in which the references are accessible, so as to prevent them becoming obsolete,[10] just as one should take into consideration the design of the web site where the reference was found, such as its functionality and attendance rate.[11]

The results of this research correspond with the given findings, as well as outline one other possibility such as (whenever possible) citing more URL addresses for the same reference, or at least listing more references for the same information, so that in case one of them is no longer available, the data itself does not lose its validity.

Other possible solutions is adding extra identificator such as permalink for dynamic web sites or the digital object identifier (DOI) for the publications published by publishing houses. Both of this solutions alow more stable linking of online document. Because, in the case of DOI the publisher only need to change metadata for the DOI when URL is changed; and in the case of permalink even if context is replaced by something new, wanted contexts is still available with the same permalink.

[7] Maharana, B.; Nayak, K.; Sahu, N.K. (2006), p. 600.

[8] Spinellis, D. (2003), p. 74.

[9] Davis, R. M. (2010.), p. 2.

[10] Stanescu, A. (2005), p. 62.

[11] Shelstad, M. 3(2005), p. 209.

Conclusion

The main hypothesis, regarding the increase in inactive web references, has been confirmed, including the rate in which the web references have become unavailable in the given period. As well as that, the assumptions having to do with the type of web references, where web pages prevail, and the lack of authors and publishers among inactive web references, have also been confirmed. The only premise that has not been confirmed, in full, has to do with the type of web sites where the inactive web references were located, due to the fact that a high percentage of web sites were for the educational, cultural and scientific institutions.

Based on the results of the research and along with the previously stated findings, the reasons for shortening the lifespan of web references and the suggestions on how to increase it, have been summed up. The largest emphasis has been placed on the problem of unavailability of relevant scientific sources by invoking the need for stricter citing systems and laying out more formal obligations while placing the sources online in order to elongate their lifespan. The other main concern had to do with the inability to accurately estimate the average lifespan because of inability to get a hold of the exact date of the disappearance of a web reference. As a way of solving that, more frequent indexing is suggested. In correspondence to that, a more quality archiving system is proposed as well as evaluating the state of the formats for the references and the functionality of the environment in which they are found. One more solution had to do with listing more URL addresses or other identifiers for the same reference. Other possible solutions is adding extra identificator such as permalink for dynamic web sites or the digital object identifier (DOI) for the publications published by publishing houses, because both of this solutions alow more stable linking of online document.

To conclude, by getting insight into possible solutions on how to increase the lifespan of web references, and what are some of the reasons why they decay, one has a better understanding of what to do to insure their availability.

References

Ashenfelder, M. The Average Lifespan of a Webpage. Library of Congress. 2011. URL: http://blogs.loc.gov/digitalpreservation/2011/11/the-average-lifespan-of-a-webpage/ (11.07.2012).

Bauer, A.; Kirin, M.; Turković, M. Web References in Graduate Papers at the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb. // *3rd International Conference "The Future of Information Sciences: INFuture2011 – Information Sciences and e-Society"*/ Billenness, C.; Hemera, A.; Mateljan, V.; Banek Zorica, M.; Stančić, H.; Seljan, S. (ed.). Zagreb : Department of Information Science, Faculty of Humanities and Social Sciences, University of, 2011, p.p. 69-75.

Coates, T. On Permalinks and Paradigms… 2003. URL: http://plasticbag.org/archives/2003/06/on_permalinks_and_paradigms (24.09.2013).

Davis, R. M. Moving Targets: Web Preservation and Reference Management. // *Ariadne*. The Web version (2010); 62. URL: http://pubs.ulcc.ac.uk/155/2/ariadne-print-issue62-davis.pdf (11.07.2012).

Franklin, J. Open access to scientific and technical information: the state of the art. // *Open Access to Scientific and Technical Information: State of the Art and Future Trends* / Gruttemeier, H.; Mahon, B. (ed.). Amsterdam : IOS Press, 2003. p. 74. URL: http://books.google.hr/books?id=2X3gW1lUvN4C&pg=PA74&hl=en#v=onepage&q&f=false (24.09.2013).

Halpin, H. Sense and Reference on the Web. // *Minds & Machines*. 2(2011), 21; p.p. 153-178. URL: http://web.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=26&hid=110&sid=20746051-f82f-44a7-81e1-383db21aa97d%40sessionmgr112 (03.05.2012).

Maharana, B.; Nayak, K.; Sahu, N.K. Scholarly use of web resources in LIS research: a citation analysis. // *Library Review*. 9(2006), 55; p.p. 598 – 607. URL: http://dx.doi.org/10.1108/00242530610706789 (09.05.2012).

Meyer zu Eissen, S.; Stein, B. Genre Classification of Web Pages: User Study and Feasibility Analysis. 2004. URL: http://www.uni-weimar.de/medien/webis/publications/papers/stein_2004c.pdf (11.07.2012).

Online Extra: How the "Digital Object Identifier" Works. *BusinessWeek*. 2001. URL: http://www.businessweek.com/printer/articles/150904-online-extra-how-the-digital-object-identifier-works?type=old_article (24.09.2013).

Santini, M. Characterizing Genres of Web Pages. // *Proceedings of the 40th Hawaii International Conference on System Sciences* (2007) / R. H. Sprague, Jr. (ed.). Waikoloa: IEEE Computer Society. URL: http://www.google.hr/url?sa=t&rct=j&q=web%20page%20genres&source=web&cd=5&ved=0CF8QFjAE&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.106.1160%26rep%3Drep1%26type%3Dpdf&ei=rKz9T5CnNfDN4QTGkvjdBg&usg=AFQjCNFz0ANLN7DD_9IpTwL5xN7Wj3Ndrg&cad=rja (11.07.2012.).

Shelstad, M. Content matters: analysis of a website redesign. // *OCLC Systems & Services*. 3(2005), 21; p.p. 209-225. URL: http://dx.doi.org/10.1108/10650750510612407 (09.05.2012).

Spinellis, D. The Decay and Failures of Web References. // *Communications of the ACM*. 1(2003), 46; p.p. 71-77. URL: http://web.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=11&hid=110&sid=3d002ef8-7047-449f-83e9-262cf1cc8873%40sessionmgr104 (03.05.2012.).

Stanescu, A. Assessing the durability of formats in a digital preservation environment: The INFORM methodology. // *OCLC Systems & Services*. 1(2005), 21; p.p. 61 – 81. URL: http://dx.doi.org/10.1108/10650750510578163 (09.05.2012).

What is the Average Lifespan of a Web Site. DialMe.com, Webmaster Tips and Resources. URL: http://www.dialme.com/blog/what-is-the-average-lifespan-of-a-web-site/ (11.07.2012).