7-2011

# The Life and Death of URLs: The Case of *Journal of the Medical Library Association*

Alireza Isfandyari-Moghaddam
*Islamic Azad University*, ali.isfandyari@gmail.com

Mohammad-Karim Saberi
*Islamic Azad University*

http://unllib.unl.edu/LPP/

*Library Philosophy and Practice* 2011

ISSN 1522-0222

# The Life and Death of URLs: The Case of *Journal of the Medical Library Association*

Alireza Isfandyari-Moghaddam
Department of Library and Information Studies
Islamic Azad University, Hamedan Branch
Iran

Mohammad-Karim Saberi
Department of Library and Information Studies
Islamic Azad University
Sciences and Research Branch
Tehran, Iran

## Introduction

The use of Internet for identifying valuable and timely information has become inevitable for most scientists as well as the public with access to the World Wide Web, since scientific and other work is created and added in digital format on the Internet every day (Falagas et al., 2008). Therefore, the use of Web links or citations has become common in journal papers, conference articles and other scholarly publications (Goh and Ng, 2007).

Despite the popularity of Web citations, we still may question the integrity of this practice. How often have we tried to link to URLs only to find a "404 Not Found" or other messages denying access? These warnings let us know that the information we came to access is no longer accessible at this site. The information may have been moved to another site, equipment may be down, or the information may have been removed completely (Germain, 2000). As a result, it should be accepted that due to ever-changing, instable and temporal identity of the Internet, Web citations (URLs) are exposed to the risk of decay. In a word, they disappear easily.

A common measure of URL decay is the half-life, defined as the period of time required for half of a defined Web literature to disappear (Koehler, 2004; Goh and Ng, 2007). Using this measure, several studies to date have dealt with this general problem of "URL decay". In order to make a better comparison between results of the present study and previous ones, here in the form of Table 1, we can have a look at them.

Table 1: Summary of previous studies on URL decay

| Study | Resource type | %URL decay | Half-life |
|---|---|---|---|

| | | | |
|---|---|---|---|
| (Germain (2000 | Random articles | N/A | 3.0 |
| (Rumsey (2002 | Law review articles | N/A | 1.4 |
| (Dellavalle et al (2003 | Medical Journals | 13% | N/A |
| (McCown et al (2005 | LIS Journal: D-Lib Magazine | N/A | 10.0 |
| (Goh and Ng (2007 | Information Science Journals | 31% | 5.0 |
| (Dimitrova and Bugeja (2007 | Communication Journals | 39% | 3.7 |

## Methodology

This article aims to examine URL decay of articles published in JMLA. Because JMLA is a freely available – according to DOAJ[1] (2008), it is among 3563 open access journals – international and scholarly journal, indexed in ISI, it was chosen to study. To collect related information, all issues (28 ones) of JMLA from 2002 to 2008 (a relatively seven-year time) were extracted. It should be noted that only articles which had references have been studied. Accordingly, Editorial, Case Study, Brief Communications, Special Report, Book Reviews, Electronic Resources Reviews and so on which had no references have been neglected. Consequently, 231 articles have been identified. Among them, articles which had Web citations were selected as research population. Afterwards, all URLs were extracted from these articles. Finally, in order to examine URLs decay, availability of individual cited URLs was scanned. Results of these tests were then compiled and analyzed to examine reasons for failure as well as to obtain the half-life value of Web citations in JMLA.

## Results

In the present study, the Web resources referred to by authors in JMLA articles in the references section of their papers have been studied. A total number of 1049 Web citations (URLs), as obtained from the bibliographies of 231 papers, have been analyzed and the necessary interpretations were made.

Distribution of articles, citations and Web citations. As shown in Table 2, totally, 231 articles have been published in JMLA during years 2002-2008. It is important to note that year 2005 with 42 articles and year 2008 (the last year ofpublication ofJMLA) with 12 articles have the most and the least rate of articles, respectively. It should be reminded that only 2 issues of 2008 have been released when the study was in process. According to Table 2, of 231 published articles, 175 articles (76%) have Web citations. Moreover, there were total number of 1049 citations whether printed or Web citations in 231 articles, and average "5 Web citations" was calculated per paper.

Table 2: Distribution of articles, citations, and Web citations in JMLA

| Year | Articles | Articles with Web citation | Web citation | Average Web citation |
|---|---|---|---|---|
| 2002 | 37 | (67%)25 | (17%)181 | 5 |
| 2003 | 33 | (75%)25 | (9%)91 | 3 |

| | | | | |
|---|---|---|---|---|
| 2004 | 40 | (56%)23 | (13%)139 | 3 |
| 2005 | 42 | (69%)29 | (14%)149 | 4 |
| 2006 | 35 | (94%)33 | (23%)239 | 7 |
| 2007 | 32 | (90%)29 | (19%)194 | 6 |
| 2008 | 12 | (91%)11 | (5%)56 | 5 |
| All years | 231 | (76%)175 | (100%)1049 | 5 |

The Decay and availability of URLs. Table 3 demonstrates that of 1049 cited URLs, 69% were accessible, while only 31% were inaccessible. In fact, decay rate (31%) in the present research is more than decay rate estimated by Dellavalle et al (2003) namely 13%. Also, our finding in relation to URL decay is fully compatible with Goh and Ng' results (Goh and Ng, 2007). They have reported that URL decay in leading Information Science journals equates to 31%.

Table 3: URL decay and availability in JMLA

| Year | Active URLs | Inactive URLs | Total |
|---|---|---|---|
| 2002 | (49%)89 | (51%)92 | (100%)181 |
| 2003 | (53%)48 | (47%)43 | (100%)91 |
| 2004 | (65%)91 | (35%)48 | (100%)139 |
| 2005 | (66%)99 | (34%)50 | (100%)149 |
| 2006 | (78%)187 | (22%)52 | (100%)239 |
| 2007 | (80%)155 | (20%)39 | (100%)194 |
| 2008 | (89%)50 | (11%)6 | (100%)56 |
| All years | (69%)719 | (31%)330 | (100%)1049 |

Error messages at inaccessible URLs. The HTTP protocol defines 24 different errors that can occur within an HTTP exchange. In addition, some errors can occur before the client and server get a chance to communicate (Spinellis, 2003). In practice, whenever a URL was inaccessible an error message (HTTP code) appeared. In general, when URLs were checked we were faced by the following errors:

1. 401 Unauthorized

The request requires user authentication. The parameter to this message gives a specification of authorization schemes which are acceptable.

2. 403 Forbidden

The server understood the request, but is refusing to fulfill it. In this case the request is for something forbidden. Authorization will not help and the request SHOULD NOT be repeated.

3. 404 Not Found

The server has not found anything matching the URL given. This error is typically generated when Web site maintainers change file names that are part of the given URL path or entirely remove the referenced material.

4. 406 Not Acceptable

The resource identified by the request is only capable of generating response entities which have content characteristics not acceptable according to the accept headers sent in the request.

5. 410 Gone

The requested resource is no longer available at the server and no forwarding address is known.

6. 500 Internal Server Error

The server encountered an unexpected condition which prevented it from fulfilling the request. This error can occur when a server is wrongly configured, or, more commonly, if a program or database that is used to serve dynamic content fails.

7. 503 Service Unavailable

The server cannot process the request due to a system overload. This should be a temporary condition.

The distribution of HTTP codes for the inaccessible URLs is shown in Figure 1.

Figure 1: Distribution of HTTP codes at inaccessible URLs

As shown in Figure 1, error 404 (Not Found) with 44% is the most recorded error message. In addition, error messages 401 (Unauthorized) with 22%, 500 (Internet Server Error) with 14%, 410 (Gone) with 13%, 403 (Forbidden) with 5%, and 406 (Not Acceptable) and 503 (Service Unavailable) with 1% were recorded.

Half-life of Web citations. In order to calculate half-life of Web resources cited in JMLA articles, the procedure used in previous research (McCown et al., 2001; Rumsey, 2002; Goh and Ng, 2007) was utilized. As mentioned earlier, "half-life" is defined as the period of time required for half of all Web citations in a journal to disappear or disintegrate. This amount of time may differ for different disciplines or different years (Koehler, 1999; Koehler, 2004; Goh and Ng, 2007).

Figure 2 shows the half-life of JMLA Web citations, obtained by determining the proportion of accessible Web citations organized by age of article. Here, *Active* indicates a successful access and *Inactive* otherwise.

Figure 2: Successful Web citation access versus age of publication

On the basis of Figure 2, the more age of Web citations increases, the more number of inaccessible URLs increases. Accordingly, 89% of one-year URLs were accessible when the test was in process. Yet, this amount was decreased to 80% for two-year URLs. When article age was three years (three-year URLs), the proportion of successful URL accesses dropped to around 78% and in the four-year URLs, this value dropped to 66%. Finally, when article age was five years, the proportion of successful URL accesses dropped to around 65% and in the six year, this value dropped to 53%. It is worth saying that in the seven-year URLs, value attained was 49%. Our results thus suggest that the half-life value of JMLA

Web citations is approximately seven years.

## Discussion and Conclusion

Phenomenon "URL decay" is a relatively new topic surveyed highly in recent years. The reason behind such an increase is increased the number of Web citations in scholarly papers, because the Web has become the first choice for finding information on current research, for breaking scientific discoveries and for keeping up with colleagues at other institutions (Zhao and Logan, 2002; Maharan et al., 2006). Considering the Internet as the first choice of researchers is not just because of the added convenience of rapid information retrieval and sharing, but because it also provides a means of making resources available that the printed media simply cannot (Wren, 2004). Therefore, even though the authors may appreciate the risk of future inaccessibility of Internet references, they cannot easily avoid their use in their publications (Falagas et al., 2008). In spite of such advantages the Internet has, Web resources led us to an emerging challenge, since they are constantly being threatened by decay and disappearance. Our research indicates that 31% of Web citations have disappeared from the original Web address, while URL decay in research done by Dellavalle et al (2003), Goh and Ng (2007), and Dimitrova and Bugeja (2007), was 13%, 31% and 39%, respectively. In addition, comparing results of the present research and previous ones particularly Dellavalle et al (2003), indicates that URL decay in the field of "Library and Information Science" (31%) is more than field "Medicine" (13%). It may be due to importance and identity of medical information resources.

It was also recognized that the half-life value of JMLA Web citations is approximately seven years. This demonstrates that JMLA URLs are more stable than URLs of previous studies including Rumsey (2002), 1.4 years, Dimitrova and Bugeja (2007), 3.7 years, Germain (2000), 3 years, and Goh and Ng (2007) with 5 years.

As concluding remark, it can be said that the Internet may prove to be an inhospitable medium, especially for web-based research, because Web citations are speedily as well as constantly fading away. Nevertheless, it should be accepted that Internet research is vital to scholarship because the medium serves as a convenient electronic warehouse of data accessible at all hours and in great quantities, thereby increasing the scope and breadth of scholarship (Dimitrova, and Bugeja, 2007).

In order to increase the rate of availability of URLs, it has been already suggested that publishers, editors, and authors should work together through:

1. Requiring authors to retain digital backup or printed copies of cited Internet-only information to facilitate content recovery should a URL become unavailable;

2. Advocating the inclusion of referenced Internet content in an online archive;

3. Checking URLs systematically before publication to minimize unavailability due to spelling errors or misprints (Wren et al., 2006; Dimitrova, and Bugeja, 2007).

In addition to considering above recommendations and also, using domains and files which are more stable and persistent, we indicate that the best solution to prevent decay or disappearance of Web citations and diminish URL decay is to make use of WebCite®-enhanced reference. WebCite®, a member of the International Internet Preservation Consortium, is an on-demand archiving system for Web references (cited Web pages and websites, or other kinds of Internet-accessible digital objects), which can be used by authors, editors, and publishers of scholarly papers and books, to ensure that cited Web material will remain available to readers in the future. A WebCite®-enhanced reference is a reference which contains – in addition to the original live URL (which can and probably will disappear in the future, or its content may change) – a link to an archived copy of the material, exactly as the citing author saw it when he accessed the cited material (WebCite, 2008). Since its official launch in October 2005, more than 100 journals are already using WebCite on a routine basis (WebCite Consortium Members, 2008). Based on WebCite Consortium Members (2008), JMLA is observed in the list of all active members of the WebCite® Consortium.

Since URLs are increasingly used and lost in scientific articles, on the one hand, and because WebCite is free of charge as well as useful, on the other hand, we recommend that all scholarly journals particularly JMLA as an international well-known journal in Library and Information Science area call for using WebCite®-enhanced reference and oblige authors to utilize WebCitation.org for all of citations referred in their articles. As a result, JMLA will serve as a pioneer and model for Library and Information Science journals in making use of WebCite®-enhanced reference.

## References

Dellavalle, R.P., Drake, A., Graber, M., Heilig, L., Hester, E., Kuntzman, J., and Schilling, L. (2003). Going, going, gone: Lost Internet references. *Science* 302(5646): 787–88.

Dimitrova, D.V., & Bugeja, M. (2007). The half-life of Internet references cited in communication journals. *New Media & Society* 9(9): 811-826.

Falagas, M.E., Karveli, E.A., and Tritsaroli, V.I. (2008). The risk of using the Internet as reference resource: a comparative study. *International Journal of Medical Informatics* 77(4): 280-286.

Germain, C.A. (2000). URLs: Uniform resource locators or unreliable resource locators. *College and Research Libraries* 61(4): 359–65.

Goh, D.H., & Ng, P.K. (2007). Link decay in leading information science journals. *Journal of the American Society for Information Science and Technology* 58(1): 15-24.

Koehler, W. (1999). An analysis of web page and web site constancy and permanence. *Journal of the American Society of information Science and Technology 59(*2): 162-180.

Koehler, W. (2004). A longitudinal study of web pages continued: A report after six years. *Information Research* 9(2), available at: http://informationr.net/ir/9-2/paper174.html (accessed 20 November 2010).

Maharan, B., Nayak, K., & Sahu, N.K. (2006). Scholarly use of web resources in LIS research: A citation analysis. *Library Review* 55(9): 598-607.

McCown, F., Chan, S., Nelson, L.M., & Bollen, J. (2001). The availability and persistence of web references in D-Lib Magazine. available at: http://www.iwaw.net/05/papers/iwaw05-mccown1.pdf (accessed 20 November 2010).

Rumsey, M. (2002). Runaway train: Problems of permanence, accessibility, and stability in the use of web sources in law review citations. *Law Library Journal* 94: 27–35.

Spinellis, D. (2003). The decay and failures of web references. *Communications of the ACM* 46(1): 71-77.

WebCite (2008). [homepage on the Internet], available at: http://www.webcitation.org/index (accessed 20 November 2010).

WebCite Consortium Members (2008). available at:

http://www.webcitation.org/members (accessed 20 November 2010).

Wren, J.D. (2004). 404 not found: The stability and persistence of URLs published in MEDLINE. *Bioinformatics* 29(5): 668–672.

Wren, J.D., Johnson, K.R., Crockett, D.M., Heilig, L.F., Schilling, L.M., & Dellavalle, R.P. (2006). Uniform resource locator decay in dermatology journals. *Arch Dermatol* 142: 1147-1152.

Zhao, D.Z., & Logan, E. (2002). Citation analysis using scientific publication on the web as data source: A case study in the XML research area. *Scientometrics* 5(3):449-472.

[1] Directory of Open Access Journals