The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0001-253X.htm

Accessibility of online resources cited in scholarly LIS journals

A study of Emerald ISI-ranked journals

Ali Sadat-Moosavi
Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
Alireza Isfandyari-Moghaddam
Department of Library and Information Studies, Hamedan Branch, Islamic Azad University, Hamedan, Iran, and
Oranus Tajeddini
Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

Abstract

Purpose – This research aims to study the state of online resources cited in scholarly library and information science (LIS) journals which are ranked in ISI and available in the Emerald database in terms of accessibility and decay.

Design/methodology/approach – Four LIS journals published by Emerald were selected from Thomson Reuters' JCR. The journals' issues from 2005 to 2008 were downloaded directly from the publisher web site and checked in terms of decay and availability of individual cited URLs.

Findings – Original accessibility of studied online resources was 64 percent, which improved to 95 percent. The main adopted strategies that returned more results were using the Wayback Machine and Google, which revived online resources by 17 percent and 12 percent respectively.

Practical implications – To increase the rate of web citations accessibility, some recommendations, including avoiding long URLs, citing documents found in digital collections availability on the web, working through systematic checking of the web citations before publication, getting backup of cited information, using the more stable file formats and domains, and utilizing tools like WebCite®-enhanced reference and a digital object identifier (DOI®) system are suggested.

Originality/value – A study which examines the accessibility and decay of web citations used by authors of articles published in ISI-ranked LIS journals available in the Emerald database has not been already done. This paper can thus contribute to the knowledge of this field as well as quality of such literature for web content providers and publishers, authors and researchers.

Keywords Online resources, Library and information science journals, Publishers, Accessibility, URL decay, Link rot, Serials, Online access

Paper type Research paper

Introduction

The use of the internet for identifying valuable and timely information has become inevitable for most scientists and the public with access to the world wide web. Scientific and other work is created and added in digital format on the internet every day (Falagas *et al.*, 2008). In fact, the use of web links or citations is common in journal papers,

Aslib Proceedings: New Information Perspectives
Vol. 64 No. 2, 2012
pp. 178-192

0001-253X
DOI 10.1108/00012531211215196

The present research is indebted to the researchers who have inspired the authors to do this. The authors would also like to express special thanks to Professors David Nicholas and Amanda Spink for their helpful consultancy.

conference articles and other scholarly publications (Goh and Ng, 2007), which mean that due to the web the citing behavior of researchers has been influenced by the web and this has influenced the growth of web citations (Isfandyari-Moghaddam *et al.*, 2010). On the other hand, with the exponential growth of information resources in the era of Web 2.0 and e-Science (Yang *et al.*, 2010), the accessibility and persistence of online resources is a critical issue that is growing in importance. Reconstruction, terminating, merging, redirecting and expanding web sites can mean an inconsistency in web URLs. This phenomenon has been studied as web site persistence and decay (Dimitrova and Bugeja, 2007a), web site constancy and permanence (Tyler and McNeil, 2003), and web site accessibility and persistence of URLs. Meanwhile, citations are also important supports for scientific research. So we see the growth of online resources and web citations. In this paper we present findings from research investigating the accessibility of online resources cited in Emerald scholarly library and information science journals.

The web has eased the accessibility of information resources and citations. However the web sites linking to online citations have been found to disappear at increasing rates over time (Dimitrova and Bugeja, 2007a). Missing online citations linked web sites is a controversial issue for researchers and web managers. Rumsey (2002), and Tyler and McNeil (2003) state that one-third of online citations vanish from original web locations for several reasons. This phenomenon is known as "broken links", "ephemeral nature of WWW hyperlinks", or "link rot" by Markwell and Brooks (2002, 2003), "missing web-cites" by Sellitto (2004), and "going 404" (so named after the "404 not found") by Wren (2008).

This phenomenon of missing web-cites exists because the uniform recourse locator (URL) is a standard string of characters which defines the location of a file or web site. When a client requests access to an online citation, for example, by clicking related URL, HTTP (Hypertext Transfer Protocol) status code is returned by the server to the client to determine the outcome of that request (For more information and a detailed list of HTTP status-codes see: http://support.microsoft.com/kb/943891/). Traditionally, unsuccessful content retrieval is called "error codes". For example, error of "404: Not Found" means that the server has not found anything matching the Request-URI. This famous error may appear in more than a half or more webpages (Dimitrova and Bugeja, 2007b) in a sample daily web exploration.

Citation accuracy and access availability of online references, such as digitized journals, public and proprietary databases and web search engine content, are fundamental elements of reliable academic research. Yet, it is generally accepted that the uncertain nature of the web is of concern to researchers and authors alike, and the decay and annihilation of citations can hinder the accessibility of online academic materials. This is a major problem that needs to be further investigated. The study reported in this paper seeks to extend our understanding of the accessibility of web documents by examining the accessibility and decay of online resources to library and information science articles appearing in four ISI-ranked journals available in the Emerald database.

Literature review

Web citations decay

Web citations have been frequently considered, used and studied (Harter and Kim, 1996; Zhang, 1998; Koehler, 1999, 2002, 2004; Germain, 2000; Davis and Cohen, 2001; Markwell and Brooks, 2002, 2003; Casserly and Bird, 2003; Dellavalle *et al.*, 2003; Spinellis, 2003; Sellitto, 2004; Wren, 2004; McCown *et al.* 2005; Maharana *et al.*, 2006;

Wren *et al.*, 2006; Zhao and Logan, 2002; Dimitrova and Bugeja, 2007a; Falagas *et al.*, 2008; Goh and Ng, 2007; Wren, 2008; Wagner *et al.*, 2009; Wu, 2009; Isfandyari-Moghaddam *et al.*, 2010; and Isfandyari-Moghaddam and Saberi, 2011). Some of the key studies into web citations discussed as below.

Harter and Kim (1996) performed one of the oldest studies on accessibility and decay of URLs using e-journals published from 1993 to 1995. They examined 47 unique URLs and reported that 31 percent were unavailable. Germain (2000) studied the reliability of URLs in academic citations. Some 31 randomly chosen academic journal articles containing 64 citations with URLs were examined. Results showed an increasing decline in the availability of URLs. Statistically, after a three-year period, almost 50 percent of URLs could not be accessed and two-thirds of the journal articles contained corroded citations. The main error message found was "Not Found". Davis and Cohen (2001) conducted a citation analysis of undergraduate term papers in microeconomics. They found:

- a significant decrease in the frequency of scholarly resources cited between 1996 and 1999; and
- web citations checked in 2000 revealed that only 55 percent of URLs cited in 1999 led to the correct web document.

In a single approach research Wren *et al.* (2006) investigated URL decay in articles published between 1 January 1999 and 30 September 2004, in the three dermatology journals with the highest scientific impact. Of the 1,113 URLs, 81.7 percent were available (decreasing with time from 89.1 percent in 2004 URLs to 65.4 percent in 1999 URLs). Finally, they concluded that URLs are increasingly used and lost in dermatology journals. Loss will continue until better preservation policies are adopted. Dimitrova and Bugeja (2007a) researched cited URLs in journalism and communication fields. They reported persistent URLs as 61 percent and.org domain with 70 percent active links as the most persistent domain, and cited URLs' half-life mean in journalism and communication as 3.7 years. Falagas *et al.* (2008) investigated the risks of using online resources and explored the accessibility of the online medical journals – *The Lancet* and *New England Journal of Medicine*. They found that 3.9 percent of *The Lancet* and 2.5 percent of *New England Journal of Medicine* references were to web resources. The two journals' inaccessible online resources were 62.2 percent and after searching missed URLs into Google, reduced to 35 percent.

Aronsky *et al.* (2007) reported findings from investigating the prevalence and inaccessibility of Web references in the bibliography of biomedical publications when first released in PubMed. During a one-month observational study period (21 February to 21 March, 2006) web citations from a 20 percent random sample of all forthcoming publications released in PubMed during the previous day were examined. The study included 4,699 publications from 844 different journals. Among the 141,845 references there were 840 (0.6 percent) web citations. One or more web references were cited in 403 (8.6 percent) articles. From the 840 web references, 11.9 percent were already inaccessible within two days after an article's release to the public. Wagner *et al.* (2009) studied the accessibility and persistence of citations to medical healthcare management journals from 2002 to 2004. They extracted 2011 unique URLs from five dominant journals in the field. Only 50.7 percent of URLs were accessible while the rest had disappeared from the original web addresses. They reported .edu and .net as the most persistent domains with 68.4 percent and 61.5 percent, respectively.

LIS citations decay

In the LIS field, Casserly and Bird (2003) examined 500 web citations randomly chosen from LIS scholarly articles. They found that only 56.4 percent of those URLs were permanent, while the rest had disappeared from the original web address. Further, the study showed that more than half the online citations contained incomplete information and the majority did not include a retrieval date. In addition, "file not found" was the most frequent error message reported. Casserly and Bird (2003) also found that close to half of the online citations they examined were initially unavailable, but this increased in the final result to 81.4 percent available citations by using different methods, including correcting errors in the URL, browsing the parent web site or using the Google Web search engine. Koehler (2004) studied web page persistence and reported findings from a continuing longitudinal study extending more than 325 weeks from December 1996 to May 2003. His research is based on evaluating the existing literature and a continuing study of a set of URLs. Koehler (2004) found that a static collection of general webpages tends to "stabilize" somewhat after it has "aged". However, web documents are not a particularly stable media for the publication of long-term information and the maintenance of individual objects or items.

McCown *et al.* (2005) explored the availability and persistence of URLs cited in articles published in *D-Lib Magazine*. They extracted 4,387 unique URLs referenced in 453 articles published from July 1995 to August 2004. McCown *et al.* (2005) found that approximately 28 percent of those URLs failed to resolve initially, and 30 percent failed to resolve at the last check. A majority of the unresolved URLs were due to a 404 (page not found) error. Moreover, based on the data collected, they found that URLs were more likely to be unavailable if they pointed to resources in the.net,.edu.xx or country-specific domain, used nonstandard ports (i.e. not port 80), or pointed to resources with uncommon or deprecated extensions (e.g. .shtml, .ps, .txt). Goh and Ng (2007) studied accessibility and decay of URLs of three LIS journals from 1997 to 2003. They reported that:

- 31 percent of URLs were decayed;
- some 56 percent of unavailable URLs brought up 404 errors; and
- edu with 36 percent active links was the most persistent domain.

Accordingly, the half-life of online resources was five years. Goh and Ng (2007) suggest that link decay is a problem that cannot be ignored and has implications for journal authors and readers.

Isfandyari-Moghaddam and Saberi (2011) examined the URL decay in articles published in *Journal of the Medical Library Association* (*JMLA*). They extracted all issues of *JMLA* from 2002 to 2008 and calculated number of web citations (URLs). The study showed that 76 percent of *JMLA* articles included web citations and each article on average included five URLs. Of some 1,049 cited URLs approximately 31 percent produced error messages mostly related to "404 Not Found". Also, the average half-life for cited web resources is estimated to be seven years. As a practical implication, they suggested that to increase the rate of URL availability authors should retain digital backup or printed copies of cited web-only information to facilitate content recovery should a URL become unavailable.

Summary
Most studies of web citation decay are comparative and in relation to one field. Among them, Casserly and Bird (2003), McCown *et al.* (2005), and Goh and Ng (2007) are LIS-focused. The approach adopted in the study reported in this paper is relatively different from the previous studies. Apart from similar methodology utilized here, our study examines some ISI-ranked journals covered by one given database. Accordingly, this research aims to study the accessibility of online resources cited in scholarly LIS journals which are ranked in ISI and available in the Emerald database in terms of availability and decay.

Research aims and questions

Research aims
Our research aims to investigate the state and accessibility of online resources cited in scholarly LIS journals which are ranked in ISI and available in the Emerald database in terms of accessibility and decay.

Research questions

- What is the distribution of LIS articles, citations and web citations?
- What is the distribution of LIS articles which have web citations?
- What is the original accessibility and state of decay of LIS web citations?
- What is the total accessibility and decay of LIS web citations?
- What are the improvement strategies and error messages for inaccessible LIS URLs?
- How is the distribution of LIS URLs by type of file formats?

Research design

Citation analysis
Zhao and Logan (2002) stated that citation analysis is a well-known technique that has long been used to study scholarly communication. In citation analysis studies, citations in research articles, often published in journals, are analyzed as artifacts of scholarly communication representing the citing author's use of the previously published work. Our study relies on citation analysis. In fact, the methodology applied in the present study is in line with Maharana and colleagues' saying: as the web is becoming a new and powerful medium for scientific communication, citation analysis and other bibliometric techniques have been applied to the study of this new phenomenon in scholarly communication. It is important to mention that the reliability of the methodology used here to collect and analyze target data is supported by previous studies which some of them have been indicated and reviewed in the literature review section. This helps to the more generalizability of final findings.

Data collection
The study was performed during a six-month period from February 2009 to July 2009. The researchers' examined four Emerald LIS journals (ISI-ranked) appearing from the beginning of 2005 to the end of 2008. Selected journals are as follows:

(1) *Journal of Documentation* (*J DOC*).
(2) *Online Information Review* (*ONLINE INFORM REV*).

(3) *Aslib Proceedings* (*ASLIB PROC*).

(4) *Interlending and Document Supply* (*INTERLEND DOC SUPPLY*).

From the Emerald database, four journals in the LIS field (based on their JCR impact factor (IF) rankings, 2008-2009) were selected. Therefore, from the Emerald database all issues of these four journals from 2005 to 2008 were downloaded to a local disk. It should be noted that a four-year span was selected because the web is a dynamic and ever-changing medium and therefore, web citations will be gradually inaccessible after the publication of articles. Furthermore, web citations existing in newer articles are usually more accessible than those used in older articles. That is why, web citations of articles published during 2005 to 2008 were chosen for final study. It is worth saying that only publications which had reference lists were considered for analysis. Editorials, brief communications, special reports, book reviews, etc. were excluded, if they had no references. Finally, for a course of four-year period a unique set of 17,728 citations were recorded by the researchers in a spreadsheet.

Data analysis

At this stage, all online resources were extracted and their URLs hyperlinks were tested by the researchers by examining their URLs' functionality. Initially, accessibility was tested by direct click on the URLs' hyperlinks. Then two groups of URLs were recognized: accessible (without any accessibility error) and inaccessible (with accessibility errors). Two groups of accessible URLs were set: "accessible through first-check" and "accessible through second check". All URLs which were accessible through the first examination (without error) and all URLs that retrieved messages indicating redirection (e.g. "you are being redirected", "this page has been moved", etc) and returned the right citation content were considered "accessible URLs through first-check".

Other URLs which were available by adopting heuristic strategies were included in the "accessible through second check". Availability examinations were carried out at all weekends and for unavailable URLs which returned 5** errors (server errors), the availability examination was repeated four times for four weekends. If then the URL was unavailable, it was recruited for heuristic URL refinement. We tried to modify unavailable URLs. Therefore, in case we faced URL errors and we checked to see if the URL content was available through the web. Thus, as the first employed strategy, unavailable URLs were entered into the Internet Explorer 7 (IE7) and if the URL worked, was considered accessible and saved into "accessible through second-check" records.

Otherwise, if it did not respond within 60s or returned an error message was considered as a "missed URL". For avoiding unwanted errors, the URL was directly copied and pasted into the browser. Missed URLs were rechecked for their likely errors in their strings. Thus, non-standard signs, if any, because of space, %, \\ instead of //, http:/,++, http@, non-alphanumeric characters (usually from non-English web sites) or other rare misspelling in the URL were corrected manually, and then the corrected URLs were tested again for accessibility status. If the non-standard URL worked into the IE7 browser, the URL was regarded as accessible and was saved in accessible through second-check records. Once more, if after a period of 60 seconds the yet-inaccessible URL resulted no content or returned errors (e.g. "404 (not found)",

"page was unavailable", "file not found" etc, errors), was regarded as "missed URL", otherwise, was recorded in "accessible through second-check" list. String editing was not saved to the unavailable URLs.

Path depth reduction strategy was then used for unavailable or missed URLs. Based on the assumption that the lengthy URLs could be erroneous, a unit-by-unit depth reduction was performed. Unavailable URL strings sustained depth reduction in several steps. URL path depth was specified by a "/" after the top domain. Accordingly, an URL with just a top domain string (e.g. http://emeraldinsight.com) has a path depth of 0. Comparably, a string like to http://emeraldinsight.com/journals/aslib.html has a path depth of 2.

In order to increase the validity of final results and realize more accuracy and precision, missed URLs were examined through a unique and a unit-by-unit path reduction operation. A unique operation was performed for every single of missed URLs. Therefore, after 1 unit path reduction, the URL was tested for availability. The reduction operation would be continued until either the path depth was $= 1$ or the broken URL responded. If the URL worked in any depths ≥ 1, the operation would be finished and the URL was marked as available through second-check. Otherwise, (URL with depth $= 1$ and yet unavailable), the URL was considered as unavailable. Path reduction was not saved to the missed URLs since they should be recruited for the next URL recovery strategy that was searching through an internet archive.

Thereafter, another availability check was established for the missed URLs using Wayback Machine (available at: www.archive.org/web/web/php) and then Google search. Wayback Machine is an old internet archive (IA) and likely is the most popular one. The Google is also the most popular search engine. Therefore, the missed URLs were entered in the Wayback Machine by copying the exact URL given in the online citation. If the URL was found in Wayback, the URL was recorded in the "accessible through second-check records". If the URL content could not be found even via The Wayback Machine, it was recruited for Google search strategy stage. Up to 5 keywords extracted from the citation's author(s) name(s), title, and resource were entered to Google and the first 20-retrieved results were reviewed to find the extinct content.

Overall, if the adopted strategies yielded no results, the inaccessible URL was considered as "decayed" and the related errors were recorded on specific related notes. Then, the studied journals' online resources (either accessible or inaccessible) were classified based on their top domains and file formats. Using Microsoft Excel 2007 the gathered data were analyzed and suitable tables and figures were produced.

Results

Q1: What is the distribution of articles, citations and web citations?

Table I shows that 608 articles included 17,728 citations and 2,886 (24 percent) were online resources.

Q2: What is the distribution of articles which have web citations?

Some 2886 online resources were recorded. ONLINE INFORM REV included 974 (34 percent) and INTERLEND DOC SUPPLY included 389 (13 percent) had the most and the least number of online resources, respectively (see Table II). The average of online resources per articles is 4.7 per article. ASLIB PROC included 5.6 and INTERLEND

DOC SUPPLY included 3.2, and had the most and the least number of online resources per articles, respectively.

Q3: What is the original accessibility and state of decay of web citations?
Initially, from 2,886 URLs, 64 percent (1,858) of URLs were accessible and 36 percent (1,028) were decayed. Some 71 percent of J DOC URLs' were originally accessible (best performance) while 56 percent of ASLIB PROC online resources were originally active (see Table III).

Q4: What is the total accessibility and decay of web citations?
After passing adopted refinement strategies, including considering IE7 browse, manual editing, path depth reduction, searching into Wayback Machine and the Google, the URL accessibility rate increased from 64 percent (1,858) to 95 percent (2,747) and inaccessibility decreased from 36 percent (1,028) to 5 percent (139) (see Table IV). This represented a 31 percent improvement in online resources accessibility. Figure 1 shows the improvement in results per adopted strategies.

Journal	Articles Freq.	%	Overall citations Freq.	%	Online resources Freq.	%
INTERLEND DOC SUPPLY	122	20	1,747	10	389	13
ASLIB PROC	149	25	4,154	23	830	29
ONLINE INFORM REV	176	29	5,362	30	974	34
J DOC	161	16	6,465	36	693	24
Total	608	100	17,728	100	2,886	100

Table I. Total and per journal distribution of articles, overall citations, and online resources

Journal	Articles Freq.	%	Online resources Freq.	%	Mean of online resources per articles
INTERLEND DOC SUPPLY	122	20	389	13	3.2
J DOC	161	26	693	24	4.3
ONLINE INFORM REV	176	29	974	34	5.5
ASLIB PROC	149	25	830	29	5.6
Total	608	100	2,886	100	4.7

Table II. The mean of online resources per articles

Journal	URL Accessible *n*	%	Inaccessible *n*	%	Total *n*	%
INTERLEND DOC SUPPLY	258	66	131	34	389	100
J DOC	491	71	202	29	693	100
ONLINE INFORM REV	647	66	327	34	974	100
ASLIB PROC	462	56	368	44	830	100
Total	1,858	64	1,028	36	2,886	100

Table III. Original accessibility and decay state of online resources

Q5: What are the improvement strategies and error messages for inaccessible URLs? Overall improvement of URLs functionality was 33 percent (Figure 1). URLs not recovered in a stage were recruited for other successive stages. For example, if the Wayback Machine failed to recover a dead URL, a Google search was used. The majority of (61 percent) occurred errors was 404 followed by the "403: forbidden" code (16 percent) which means the server has understood the request, but refused to fulfill it (see Figure 2). This error occurs due to filtering issue and/or firewall software priorities. 4** errors are somehow related to the client request. "401: unauthorized" (7 percent) means that the request requires user authentication or current authentication is not acceptable." 504: gateway time-out" (4 percent) means that the server, while acting as a gateway or proxy, has not received a timely response from the upstream server. "400: bad request" (2 percent) code means that the request could not be understood by the server due to malformed syntax.

When the server encounters an unexpected condition like to configuration or system error, it sends "500: internal server error" (9 percent). "503: service unavailable" (1 percent) error occurs when temporary the server cannot process the request due to a system overload. Other errors like to "410: gone" (the requested resource is no longer available at the server and no forwarding address is known) and "301: Moved Permanently" or other 2** related errors had not occurred during the study.

Figure 3 illustrates the accessibility of URLs based on their top domains.

Table IV.
Total accessibility and decay of URLs

	URL					
	Accessible		Decayed		Total	
Journal	*n*	%	*n*	%	*n*	%
INTERLEND DOC SUPPLY	387	97	11	3	389	100
J DOC	648	94	45	6	693	100
ONLINE INFORM REV	940	97	34	3	974	100
ASLIB PROC	781	94	49	6	830	100
Total	2,747	95	139	5	2,886	100

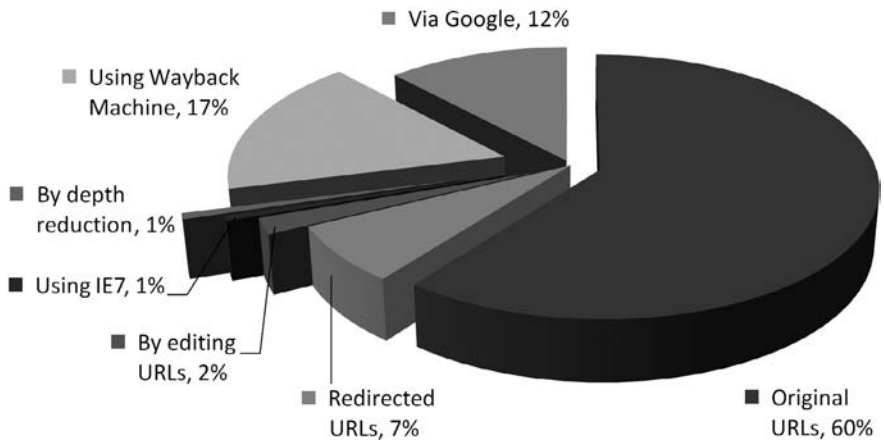


Figure 1.
Percentages of total accessible URLs

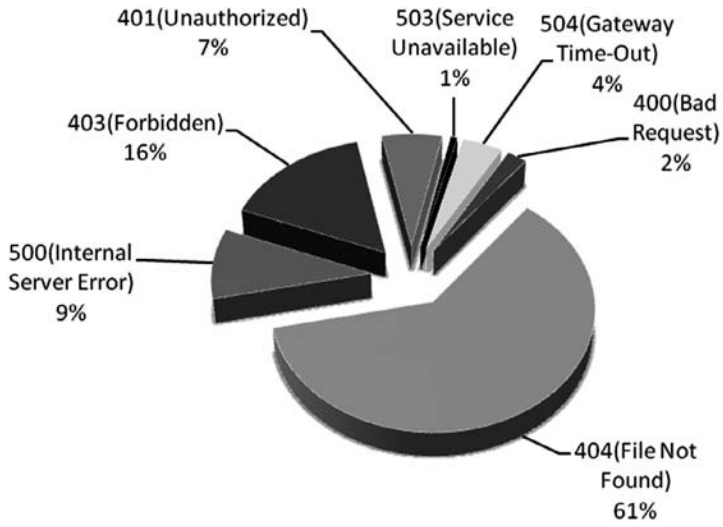
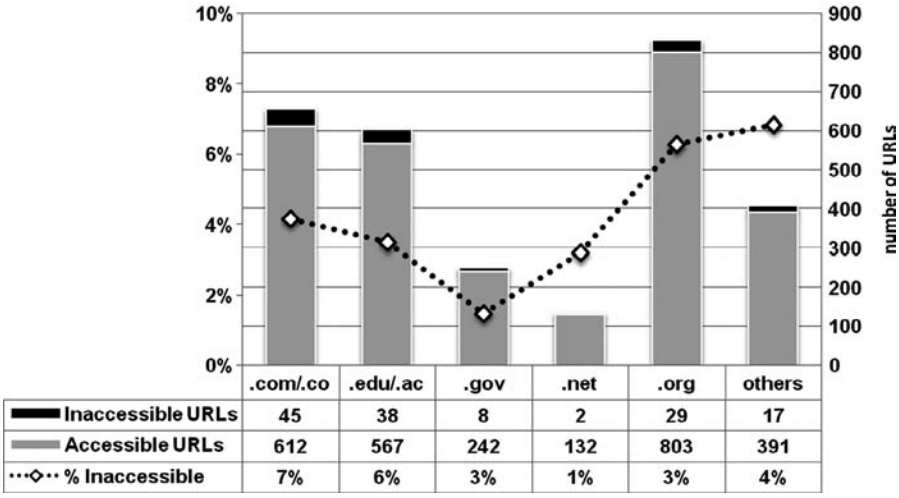


Figure 2.
Percentages of HTTP codes for the URLs' accessibility errors



	.com/.co	.edu/.ac	.gov	.net	.org	others
Inaccessible URLs	45	38	8	2	29	17
Accessible URLs	612	567	242	132	803	391
% Inaccessible	7%	6%	3%	1%	3%	4%

Figure 3.
Accessibility and decay of URLs based on their top domains

Accordingly, some 99 percent of.net URLs were accessible followed by.gov and.org with 97 percent accessibility, and.com/co had the most decay with 93 percent unavailability.

Q6: What is the distribution of URLs by type of file formats?
Considering online resources file formats, seven categories were recognized: slash (/) files (URLs end with "/" such as http://foo.edu/), HTM/HTML/SHML, PDF, PPT, DOC, RTF, TXT, and Other, for any other file formats (see Figure 4). Some 1,179 out of 2,886 online resources were HTM/HTML/SHTML files. PPT and RTF with 100 percent followed by PDF by 97 percent were the most accessible formats. DOC was the most unstable format with 20 percent decay.

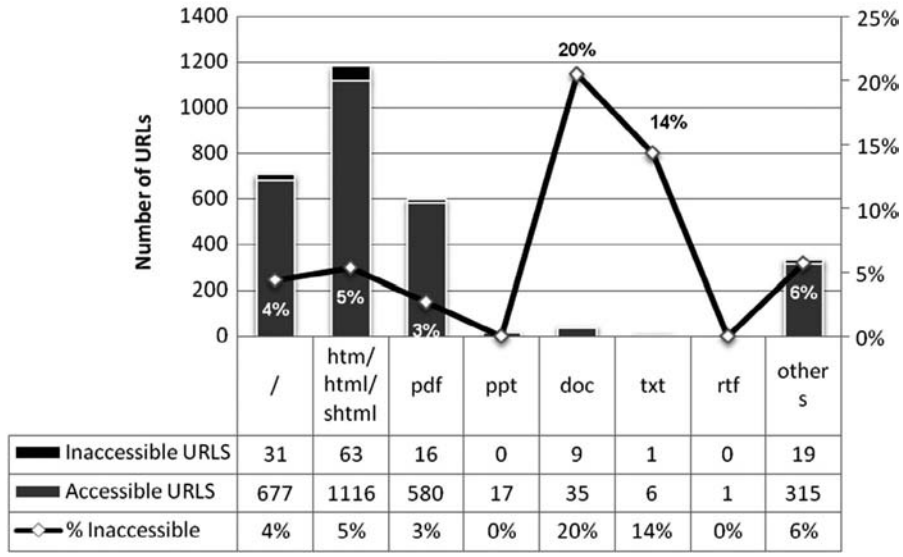


Figure 4.
Accessibility and decay of URLs' based on file formats

Discussion

The study reported in this paper investigated the state of online resources cited in scholarly LIS journals which are ranked in ISI and available in the Emerald database in terms of accessibility and decay. This study contributes to the fields of bibliometrics, citations analysis, and Webometrics particularly as our study is one of the newest studies in relation to LIS. In other words, a study which examines the accessibility and decay of web citations used by authors of articles published in ISI-ranked LIS journals available in the Emerald database has not been previously conducted. Our study can also contribute to the quality of such literature for web content providers and publishers, authors and researchers.

Investigation through the URL top domains showed that the.org has received more citations than other domains. This finding was in agreement with Dimitrova and Bugeja (2007a) and McCown *et al.* (2005). RTF and PPT with 100 percent accessibility and the PDF with 97 percent were the most stable file formats. This result did not agree with McCown *et al.* (2005) who reported HTML files as the most stable format. The 404 error was the most prevalent error. This means that in this study, the majority of errors (61 percent) were client based errors that occurred for several reasons such as filtering, connection failure status, explorer malfunctioning due to the used proxies, and intrusion of alphanumeric characters.

According to Sullivan (1999; quoted in Wu, 2009, p. 482), link rot [decay] is on the rise and over 20 percent of web page resources are affected by link rot. In the present study, the online resources decay rate was 36 percent, while in the previous studies, the online resources decay was: 50 percent (Germain, 2000), 13 percent (Dellavalle *et al.*, 2003), 45.4 percent (Casserly and Bird, 2003), 39 percent (Dimitrova and Bugeja, 2007a), 49.3 percent (Wagner *et al.*, 2009), and 55.8 percent (Wu, 2009). The decay rate was decreased to 5 percent by adopting various strategies such as Google and Wayback Machine search, URL path reduction or truncation, and manual editing.

The main strategies which revived more dead URLs were using the Wayback Machine (17 percent) and Google (12 percent). Nonetheless, we found some limitations in their services. For example, the Wayback Machine works just for HTML based URLs. Therefore, we could not search the Wayback Machine for the FTP (file transport protocol) based URLs. Also, this reputable web archive had some limitations in archiving dynamic pages and pages containing Java Scripts. The Wayback Machine did not archive the pages which had not external links to other web sites.

Google was unable to crawl the pages including active contents and pages using robots.txt codes. Using heuristic rules were rather tedious. The user is not always patient enough for manual editing of URL strings. Therefore, the URL decay should be considered as a critical issue for scholarly online publishing. We could not compare the Wayback Machine with the Google as they have different primary functions and carry different types of content. Dimitrova and Bugeja (2007b) showed that the Wayback Machine performance in reviving unavailable URLs was largely better than using Google. They showed that 64 percent of citations retrieved through Google were also found in the Wayback Machine and only 36 percent of citations were uniquely available through Google. In contrast, 67 percent of the citations found in the Wayback Machine did not overlap and 33 percent overlapped with Google.

Ultimately, we were faced with 5 percent absolutely unavailable online citations even with employing useful services such as Google, the Wayback Machine or the heuristic strategies. Nonetheless, the original accessibility was 64 percent. This should be considered as an issue of using online resources.

Limitations

The present study has some limitations. First, only four LIS journals were selected from which web citations were collected and analyzed. Although these are ISI-ranked and accessible journals they may not be representative of the entire population of their counterparts. In addition this study is limited to a four-year span which does not provide a relatively definite picture of the status of web citations used in the articles of mentioned journals.

Implications

The availability of online resources is critical, especially when the LIS field's journals are limited or the retrieval alternatives are confined to a few journals. URL decay can occur for several reasons. However, the accessibility of online resources should not to be ignored, since the scientific status of scholarly publications would sustain serious damages. To bridge or diminish this gap and thus increase the rate of web citations accessibility, Goh and Ng (2007) recommended some strategies including avoiding long URLs as it increases the likelihood of failure, citing documents found in digital collections on the web (e.g. digital libraries and commercial databases) rather than web sites in general. Emphasizing Goh and Ng's (2007) suggestions, we also reaffirm previous studies, by Wren *et al.* (2006); Casserly and Bird (2003); Germain (2000); and Dimitrova and Bugeja (2007a) who suggested that publishers, editors, and authors should work together through systematic checking of the web citations before publication, getting backup of cited information, and using the more stable file formats and domains. One of the best solutions to prevent decay or disappearance of web citations and increase URLs permanence is use of tools like WebCite®-enhanced

reference (www.webcitation.org) and Digital Object Identifier (DOI®) System (www.doi.org).

Conclusion and further research

URL decay happens for several reasons, among them the reorganization of web sites and likely changes in adopted domains. Based on researchers' reliance on online resources in a specific field, the decay rate differs. The decay increases with the increase in researchers' reliance on online resources. This issue is critical to those fields, including LIS, which are specific and have a limited number of online resources. Checking the availability of online resources prior to publishing submitted manuscripts could improve the availability status, since the authors are likely more informed of their used online resources and easily can modify the URL strings' errors or replace decayed URLs with live alternatives.

Spinellis's (2003, p. 77) suggest that:

> [. . .] the web has revolutionized on a global scale the way we distribute, disseminate, and access information and, as a consequence, is creating a disruptive paradigm shift in the way human scientific knowledge builds upon and references existing work. In the past, libraries could provide reliable archival services for books and other printed publications; the emergence of the web is marginalizing their role. In the short term none of the approaches toward solving the general problem of dangling URL references is likely to be a panacea. It is therefore important to appreciate the importance of web citations and invest in research, technical infrastructures, and social processes that will lead toward a more stable scientific publication paradigm.

It is hoped that editors and their colleagues will provide more clearly defined expectations in journal guidelines. Librarians can also play a role in developing such guidelines. To overcome the problem of URL decay a multi-lateral collaboration is necessary between article producers (i.e. researchers and authors) and article distributors (i.e. journals, their editorial team, and publishers).

As for future research, we need to replicate our study long-term and at a large-scale including more LIS journals so more generalizable results can be attained. Moreover, doing a comparative study investigating the accessibility and decay of web citations used in the articles of ISI-ranked and non-ISI-ranked would also be important.

References

Aronsky, D., Madani, S., Carnevale, R.J., Duda, S. and Feyder, M.T. (2007), "The prevalence and inaccessibility of internet references in the biomedical literature at the time of publication", *Journal of the American Medical Informatics Association*, Vol. 14 No. 2, pp. 232-4.

Casserly, M. and Bird, J.E. (2003), "Web resources availability: analysis and implications for scholarship", *College & Research Libraries*, Vol. 64 No. 4, pp. 300-17.

Davis, P.M. and Cohen, S.A. (2001), "The effect of the web on undergraduate citation behavior 1996-1999", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 4, pp. 309-14.

Dellavalle, R.P., Drake, A., Graber, M., Heilig, L., Hester, E., Kuntzman, J. and Schilling, L. (2003), "Going, going, gone: lost internet references", *Science*, Vol. 302 No. 5646, pp. 787-8.

Dimitrova, D.V. and Bugeja, M. (2007a), "The half-life of internet references cited in communication journals", *New Media & Society*, Vol. 9 No. 9, pp. 811-26.

Dimitrova, D.V. and Bugeja, M. (2007b), "Raising the dead: recovery of decayed online citations", *American Communication Journal*, Vol. 9 No. 2, available at: www.acjournal.org/holdings/vol9/summer/articles/citations.html (accessed 20 July 2011).

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*, Vol. 77 No. 4, pp. 280-6.

Germain, C.A. (2000), "URLs: uniform resource locators or unreliable resource locators", *College & Research Libraries*, Vol. 61 No. 4, pp. 359-65.

Goh, D.H. and Ng, P.K. (2007), "Link decay in leading information science journals", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 1, pp. 15-24.

Harter, S.P. and Kim, H.J. (1996), "Electronic journals and scholarly communication: a citation and reference study", *Information Research*, Vol. 2 No. 1, available at: http://InformationR.net/ir/2-1/paper9a.html (accessed 20 July 2011).

Isfandyari-Moghaddam, A. and Saberi, M.K. (2011), "The life and death of URLs: the case of *Journal of the Medical Library Association*", *Library Philosophy and Practice*, (July), (annual volume), available at: www.webpages.uidaho.edu/~mbolin/moghaddam-saberi.htm (accessed 20 July 2011).

Isfandyari-Moghaddam, A., Saberi, M.K. and Mohammad Esmaeel, S. (2010), "Availability and half-life of web references cited in *Information Research Journal*: a citation study", *International Journal of Information Science and Management*, Vol. 8 No. 2, pp. 57-75.

Koehler, W. (1999), "An analysis of web page and web site constancy and permanence", *Journal of the American Society of Information Science and Technology*, Vol. 50 No. 2, pp. 162-80.

Koehler, W. (2002), "Web page change and persistence: a four-year longitudinal study", *Journal of the American Society of Information Science and Technology*, Vol. 53 No. 2, pp. 162-71.

Koehler, W. (2004), "A longitudinal study of web pages continued: a report after six years", *Information Research*, Vol. 9 No. 2, available at: http://InformationR.net/ir/9-2/paper174.html (accessed 20 July 2011).

McCown, F., Chan, S., Nelson, L.M. and Bollen, J. (2005), "The availability and persistence of web references in *D-Lib Magazine*", available at: www.iwaw.net/05/papers/iwaw05-mccown1.pdf (accessed 20 July 2011).

Maharana, B., Nayak, K. and Sahu, N.K. (2006), "Scholarly use of web resources in LIS research: a citation analysis", *Library Review*, Vol. 55 No. 9, pp. 598-607.

Markwell, J. and Brooks, D.W. (2002), "Broken links: the ephemeral nature of educational WWW hyperlinks", *Journal of Science Education and Technology*, Vol. 11 No. 2, pp. 105-8.

Markwell, J. and Brooks, D.W. (2003), "Link rot limits the usefulness of web-based educational materials in biochemistry and molecular biology", *Biochemistry and Molecular Biology Education*, Vol. 31 No. 1, pp. 69-72.

Rumsey, M. (2002), "Runaway train: problems of permanence, accessibility, and stability in the use of web sources in law review citations", *Law Library Journal*, Vol. 94 No. 1, pp. 27-35.

Sellitto, C. (2004), "A study of missing web-cites in scholarly articles: towards an evaluation framework", *Journal of Information Science*, Vol. 30 No. 6, pp. 484-95.

Spinellis, D. (2003), "The decay and failures of web references", *Communications of the ACM*, Vol. 46 No. 1, pp. 71-7.

Tyler, D.C. and McNeil, B. (2003), "Librarians and link rot: a comparative analysis with some methodological considerations", *Portal: Libraries and the Academy*, Vol. 3 No. 4, pp. 615-32.

Wagner, C., Gebremichael, M.D., Taylor, M.K. and Soltys, M.J. (2009), "Disappearing act: decay of uniform resource locators in health care management journals", *Journal of the Medical Library Association*, Vol. 97 No. 2, pp. 122-30.

Wren, J.D. (2004), "404 not found: the stability and persistence of URLs published in MEDLINE", *Bioinformatics*, Vol. 20 No. 5, pp. 668-72.

Wren, J.D. (2008), "URL decay in MEDLINE: a 4-year follow-up study", *Bioinformatics*, Vol. 24 No. 11, pp. 1381-5.

Wren, J.D., Johnson, K.R., Crockett, D.M., Heilig, L.F., Schilling, L.M. and Dellavalle, R.P. (2006), "Uniform resource locator decay in dermatology journals: author attitudes and preservation practices", *Archives of Dermatology*, Vol. 142 No. 9, pp. 1147-52.

Wu, Z. (2009), "An empirical study of the accessibility of web references in two Chinese academic journals", *Scientometrics*, Vol. 78 No. 3, pp. 481-503.

Yang, S., Qiu, J. and Xiong, Z. (2010), "An empirical study on the utilization of web academic resources in humanities and social sciences based on web citations", *Scientometrics*, Vol. 84 No. 1, pp. 1-19.

Zhang, Y. (1998), "The impact of internet based electronic resources on formal scholarly communication in the area of library and information science: a citation analysis", *Journal of Information Science*, Vol. 24 No. 4, pp. 241-54.

Zhao, D.Z. and Logan, E. (2002), "Citation analysis using scientific publication on the web as data source: a case study in the XML research area", *Scientometrics*, Vol. 5 No. 3, pp. 449-72.

Corresponding author
Alireza Isfandyari-Moghaddam can be contacted at: ali.isfandyari@gmail.com