Article

International Federation of
Library Associations and Institutions
2014, Vol. 40(1) 35–47
© The Author(s) 2014
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0340035214526529
ifla.sagepub.com
SAGE

Corrosion of URLs: Implications for electronic publishing

K.R. Prithviraj
Kuvempu University

B.T. Sampath Kumar
Tumkur University

Abstract

The aim of this study is to analyze the accessibility, corrosion and half-life of URLs cited in the articles of Indian LIS conference proceedings published during 2001 to 2010. A total of 5,698 URLs cited in the 1,700 articles were examined. The percentage of URLs increased from 39.10 percent in 2001 to 73.47 percent in 2009. The study found that 50.09 percent of URLs were not accessible at the time of testing and the remaining 49.91 percent of URLs were accessible. The HTTP 404 error message – "file not found" was the overwhelming message encountered and represented 53.29 percent of all HTTP messages. The study also noticed that the average half-life of URLs of missing URLs was estimated to be 4.94 years. Even though there are various retrieval tools being used to recover vanished URLs, still there is a need to improve such tools.

Keywords

URLs, corrosion, half-life, electronic publishing

Introduction

In recent years researchers have focused on the growing reliance on the Internet as a source of information and the increasing frequency with which authors cite websites and pages to document their scholarly research (Casserly and Bird 2003). The outburst of e-contents such as e-journals, e-books, etc., has facilitated access to scholarly information; thus, the nature of citations is susceptible of change. Print-to-web citations and web-to-print citations are now fairly common and thus it seems inevitable that web resources are becoming favored in scholars' communication. The cited content is considered as available if it can be found either at the URLs included in the sample citation or elsewhere on the web (Riahinia et al. 2011). During the last decade many journal articles, including refereed articles, contain citations to web sources. Despite the popularity of web citations, recent studies have documented that the problem of URL corrosion is a serious issue, not only for web masters, but also for academics who use web sources in their research (Dimitrova and Bugeja, 2007b).

This corrosion of URLs definitely will have an implication for electronic publishing. It is also noticed that the corrosion of URLs was found in earlier studies (Aronsky et al., 2007; Ducut et al., 2008; Goh and Ng, 2007; Lopresti, 2010; Sampath Kumar and Manoj Kumar, 2012; Sampath Kumar and Vinay Kumar, 2013). Today the Internet has expanded access to scholarship, but its dynamism poses many challenges to scholarly communication. Keeping in view the disappearing nature of URLs citations, the present study explores the impermanence of ULRs cited in articles of Indian conference proceedings and its impact on electronic publishing.

Previous works

There is a substantial body of literature in the field of web citations and decay. This part of the study discusses the literature related to the use of URLs as citations in scholarly literature and the corrosion of URLs.

Corresponding author:
Dr. B.T. Sampath Kumar, Associate Professor, Department of Library and Information Science, Tumkur University, Tumkur, Karnataka, India.
Email: sampathbt_2001@rediffmail.com. Website: http://www.freewebs.com/sampathkumar

In the year 2006, a study by Dimitrova and Bugeja (2007) focused on six leading communication journals and their use of online citations in articles published between 2000 and 2003. The findings showed that online citations in the .gov and .org domains were more likely to remain accessible over time. Year of publication and URL level also emerged as significant predictors of online citation permanence. More than 37 percent of the online citations had disappeared from the original source over a four-year period (2000–2003).

Goh and Ng (2007) investigated the link decay phenomenon in three leading information science journals. The study revealed that approximately 31 percent of all citations were not accessible during the time of testing and the majority of errors were due to missing content (HTTP error code 404). Citations from the .edu domain were found to have the highest failure rates (36 percent) when compared with other popular top-level domains. The results of the study indicated that link decay was a problem that cannot be ignored, and the implications for journal authors and readers are discussed.

Ducut et al. (2008) conducted a survey on the accessibility of MEDLINE URLs. Some 10,208 URL addresses were checked for errors during the initial run. A total of 2,245 (21.99 percent) were not accessible during the initial run. Of these, 544 URLs were found to have errors in formatting. A total of 163 URLs redirected to another page and the updated addresses were used in the study. In the same year, Casserly and Bird (2008) conducted a comprehensive study to determine the persistence of URL citations. They found a decrease of 17.4 percent in persistence. The researchers also found an increase in the number of journals that provide instructions to authors on citing content on the web. Wu (2009) analyzed 1,637 web references in articles in two Chinese academic journals published during 1999 to 2003. His study showed that web references accessibility has a strong negative correlation with age. Web references decay at a rate of about 9 percent to 10 percent annually. He also estimated that about 65 percent to 72 percent of web references could be accessed in newly published Chinese academic journals. Six years after publication, about 90 percent could not be accessed.

Lopresti (2010) examined citations in five leading environmental science journals for accuracy. As many as 647 (24.41 percent) of the 2,650 citations checked were found to contain errors. Of the five journals, *Conservation Biology* had the lowest percentage of citations with errors and the journal *Climatic Change* had the highest. Rhodes (2010) examined the persistence of URLs extracted from law and policy-related materials over a three year period (2008–2010). She found that URLs at .gov top-level domain are more vulnerable to link rot than those at the .org top-level domain.

Riahinia et al. (2011) analyzed 37,791 citations extracted from six LIS scholarly journals, of which 4,840 (12.8 percent) were web citations. The mean averages of web and print citations per article were 4.09 and 27.9 respectively. Of all web citations, 4,617 (95 percent) were persistent, and 5 percent returned errors and thus were not accessible. The most prevalent domain of citations was .html and the most favorable and persistent file format was .pdf. Saberi et al. (2011) analyzed the accessibility and decay of 558 web citations used in refereed articles published in the first 10 years of the *Journal of Artificial Societies and Social Simulation*, published between1998 to 2007. The study showed that 75 percent (421 URLs) were accessible and the remaining 25 percent (137 URLs) were inaccessible.

Sampath Kumar and Prithviraj (2012) examined 350 conference articles published in Indian Association of Teachers of Library and Information Science (IATLIS) conference proceedings during the period 2001–2008. The study showed that overall, 45.61 percent (307) of web citations were missing from the total of 673 web citations and the percentage of missing web citations had gradually decreased from 2001 (66 percent) to 2008 (30.27 percent). Of the 307 missing web citations, the top-level domain .org had the highest percentage (30.29 percent) of missing URLs, followed by the .edu domain (21.49 percent). Mardani's (2012) study showed that out of 46,762 web citations, 40,954 (88 percent of the total) were available and 5,808 (12 percent of the total) were missing. As reported in other studies, most Internet users encounter the 404 error (page not found). The highest percentages of inactive URLs were found to be associated with the .gov top-level domain. A study by Sampath Kumar and Manoj Kumar (2012) analyzed 2,890 URL citations cited in articles in four LIS scholarly journals published between 1996 and 2009. Of the total of 2,890 URL citations, 754 (26.08 percent) encountered access errors and were designated missing web citations. Further, they found that 70.15 percent of all the missing URLs were due to HTTP error 404 (page not found). The top-level domain having the greatest number of missing URLs was the commercial domain .com (27.90 percent), followed by .org (24.03 percent).

A recent study by Sampath Kumar and Vinay Kumar (2013) investigates the availability, persistence and half-life of URL citations cited in two Indian LIS journals articles published between 2002

Table 1. Summary of URLs corrosion in previous studies.

Sl. No.	Authors	Year	Percentage of URLs corrosion
1	Rumsey	2002	51.53
2	Veronin	2002	59
3	Casserly and Bird	2003	42.6
4	Spinellis	2003	28
5	Markwell and Brooks	2003	20
6	Dellavalle, et al.,	2003	13
7	Sellitto	2004	45.8
8	McCown, et al.,	2005	29.5
9	Aronsky, et al.,	2006	11.9
10	Goh and Ng	2007	31
11	Dimitrova and Bugeja	2007	39
12	Thorp and Brown	2007	53.24
13	Ducut, et al.,	2008	21.99
14	Wagner, et al.,	2009	49.3
15	Wu	2009	55.77
16	Nagaraja et al.,	2011	17
17	Isfandyari-Moghaddam and Saberi	2011	31
18	Riahinia, et al.,	2011	5
19	Saberi, et al.,	2011	25
20	Sampath Kumar and Prithviraj	2012	45.61
21	Mardani	2012	12
22	Saberi and Abedi	2012	27
23	Sadat-Moosavi, et al.,	2012	36
24	Sampath Kumar and Manoj Kumar	2012	26.08

and 2010. Authors found that 39.84 percent of URL citations were not accessible and remaining 60.15 percent of URL citations were still accessible.

Most of the studies mentioned above clearly indicate that web citations in scholarly publications are increasing, but that the major problem of web citations is their permanence. The percentages of URLs corrosion found in the previous studies are presented in Table 1.

Research objectives and hypotheses

The present study investigates the availability and corrosion of URLs cited in conference proceedings published during 2001 to 2010. The study was carried out with the following objectives and hypotheses:

Objectives

a) To know the pattern of use of URLs by Indian LIS professionals in conference proceedings.
b) To know the rate of corrosion of URLs cited in the articles of conference proceedings.
c) To examine the top-level domains associated with active and missing URLs.
d) To identify the correlation between the path depth and corrosion of URLs.
e) To calculate the half-life period of URLs cited in conference proceedings.

Hypotheses

a) There is an increase in the number of URLs during 2001 to 2010 in selected Indian LIS scholarly communications.
b) Corrosion of URLs will increase as their age increases.
c) There is significant association between the corrosion of URLs and top-level domains.
d) The path depth and corrosion of URLs are positively correlated.

Methodology

Selection of conference proceedings and articles

The data for the present study were drawn from a selective sample of three Indian LIS conference proceedings published in printed form during 2001 to 2010. The present study has chosen the following three conferences proceedings:

CALIBER conference conducted by INFLIBNET

Convention on Automation of Libraries in Education and Research Institutions (CALIBER) is an annual convention, organised by INFLIBNET Centre in collaboration with different Universities. The convention provides a unique forum to the library and information professionals, teachers, IT professionals, consultants and users involved in automation and networking of libraries as well as information providers to come together and interact on the subjects of mutual interest (INFLIBNET, 2013)

NACLIN Conference Proceedings

The National Convention on Knowledge, Library and Information Networking (NACLIN) is the official convention of Developing Library Network (DELNET), New Delhi. NACLIN is devoted to discussing how the latest technologies can enhance library services, and help in the dissemination of information among library users including students, staff and patrons (NACLIN, 2013). DELNET is well known network in India in the field of Library and Information Science. DELNET has been established with the prime objective of promoting resource sharing among the libraries through the development of a network of libraries (DELNET, 2013).

ILA conference conducted by Indian Library Association. Indian Library Association (ILA) is the largest and renowned professional body in the field of Library and Information Science in India. It is a premier association committed to the cause of Library Movement and Development (ILA, 2013).

These conference proceedings were considered on the basis of their reputation and long publishing history.

Selection of articles and references

The study was undertaken to know the availability and corrosion of URLs cited in the articles of above conference proceedings published during 2001–2010. All articles (N = 1,700) published in these three conference proceedings were selected. All the references (N = 15,745) which are appended at the end of each of the articles under the heading ‘references’ were collected. The editorial articles, abstracts, expanded bibliographies, end notes, footnotes, e-mail links, annotations and book reviews, etc., were not considered as citations and thus not counted in the data collected for further study. In some articles citations referring to print sources and URLs were listed twice in the references sections. When this occurred they were only counted as single citations.

The 1,700 articles that formed the source of the sample citations used in this study contained a total of 15,745 citations, of which 10,047 were print citations and 5,698 were URLs.

Selection of testing of URLs

The researcher extracted all URLs (N = 5,698) from the list of references. URLs that occurred in the same article twice or more were counted as a single URL. However, if the same URL was cited in other articles it was then considered as an independent URL for statistical analysis. The URLs so extracted were then tested to determine whether they were active or missing on the web. The researcher checked all of them one by one in the World Wide Web Consortium’s (W3C) Link Checker (http://validator.w3.org/checklink). The Link Checker tool was selected for its unique features to test the persistence and accessibility of URLs. Those URLs which led directly to the web source were classified as active URLs and those which received an HTTP error message were classified as missing URLs.

The exact error message was recorded and then classified according to type of error (ex: HTTP 403, HTTP 404, HTTP 500, HTTP 502, HTTP 503, etc.). Furthermore, the URL for each online citation was coded for top-level domain (ex: .com, .org, .gov, .edu, etc.), file format (ex: .html, .pdf, .doc, etc.) and URL path depth (ex: 0, 1, 2, 3, etc.).

The statistical program SPSS 19.0 for Windows was used to generate contingency tables and calculate the Pearson’s Correlation and Chi-Square values. A $p < .05$ level of significance was used for the study.

Analysis of data

Distribution of articles, citations and URLs

The 1,700 articles that formed the source of the sample citations used in this study contained a total of 15,745 citations. Of these 10,047 (63.81 percent) were print references and 5,698 (36.18 percent) were URL citations (Table 2). The data presented in Table 2 also illustrate that the percentage of articles with at least one URL has increased from 39.10 percent in 2001 to 91.67 percent in 2010. There was a constant and continuous increase in the number of articles with URLs over the years 2001–2010.In total, there were 1,011 (59.47 percent) articles with URLs among the 1,700 articles published in the selected LIS conference proceedings during 2001–2010.

The average number of URLs per article ranged from a low of 1.99 in 2001 to a high of 5.63 in 2010. The average number of URLs per article was 3.35 across all the 1,700 conference articles. Further, the years 2006 and 2010 witnessed the highest percentages of URLs with 47.24 percent and 44.85 percent respectively, which were comparatively higher than the decadal percentage of URLs (36.19 percent). The percentage of URLs had increased from the year 2001 (26.70 percent) to 2010 (44.85 percent).

A noteworthy finding of the study is that there was high use of URLs in the conference articles of the later years (2006–2010) compared to the earlier years (2001–2005) of the decade studied. This phenomenon could be attributed to the popularity of the web as an established source of information with universally accessible free digital information which might have prompted the authors to access and cite more and more web resources in their articles. Indeed, the higher numbers of URLs found in the later years of the decade studied here are in conformity with the findings of Zhang (2001), Rumsey (2002), and Spinellis (2003), who reported a relative increase of URLs in the articles published during the later years of their studies. To support this, the correlation analysis also indicates that there is a negative correlation between the year and the percentage of URLs ($r = -.777$, $p = .014$). This indicates that conference proceedings published in recent years have more URLs compared to earlier ones (Figure 1).

Table 2. Year-wise distribution of articles, citations and URLs.

Publication year of the conference proceedings	Total no. of articles	No. of articles with URLs	Percentage of articles with URLs	Total no. of citations	Total no. of print citations	No. of print citations per article as an average of all articles	Total no. of URLs	No. of URLs per article as an average of all articles	URLs as percent of all citations
2001	156	61	39.10	1165	854	5.47	311	1.99	26.70
2002	174	82	47.13	1453	1066	6.13	387	2.22	26.63
2003	182	71	39.01	1474	1025	5.63	449	2.47	30.46
2004	171	97	56.73	1519	982	5.74	537	3.14	35.35
2005	168	113	67.26	1636	1078	6.42	558	3.32	34.11
2006	195	130	66.67	1867	985	5.05	882	4.52	47.24
2007	220	138	62.73	2022	1247	5.67	775	3.52	38.33
2008	214	153	71.50	2179	1375	6.43	804	3.76	36.90
2009	196	144	73.47	2129	1269	6.47	860	4.39	40.39
2010	24	22	91.67	301	166	6.92	135	5.63	44.85
All years	1700	1011	59.47	15745	10047	5.91	5698	3.35	36.19
Average citations per article				9.26					
Percentage of citations of all the years					63.81		36.19		

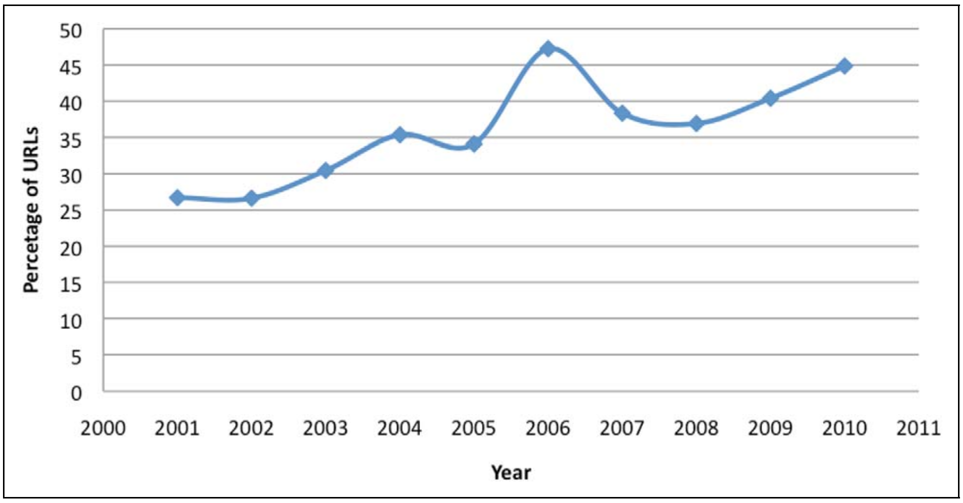

Figure 1. Year-wise distribution of URLS.

Contents of URLs

A citation was considered complete if it included minimum information about the web document, i.e. name of the author, title, publication, publisher, date of publication, URL and date of access. The URLs content noted in the references lists varies from ‘Only URL’ to ‘URL accompanied by partial or complete bibliographic information’. A total of 5,698 URLs found in the selected conference proceedings were classified on the basis of the web citation content (bibliographical descriptions) and the resultant data is presented in Table 3 for further analysis.

The analysis of data presented in Table 3 shows that 33.15 percent of the 5,698 URL citations contained only URLs, whereas 56.83 percent contained URLs with partial bibliographic information and only 10.02 percent contained URLs with complete information (Figure 2).

Missing URLs by year

The researchers considered a web source as a missing URL if it returned with an HTTP error message. The data about percentage of active and missing URLs presented in Table 4 shows that 49.91 percent URLs (2,844 of 5,698) remained active while the rest 2,854 (50.09 percent) were found to be missing. The largest number of missing URLs in any one year (62.57 percent) was those cited in conference articles published in 2004, followed by 60.45 percent in 2001. The percentages of missing URLs were found to be decreasing during the later years from 2006 (48.53

Table 3. Contents of URLs.

Conference proceedings	Total no. of URLs	Only URL	percent	URL and partial bibliographic information	percent	URL and complete bibliographic information	percent
CALIBER	2489	673	27.04	1559	62.64	257	10.33
ILA	1634	545	33.35	909	55.63	180	11.02
NACLIN	1575	671	42.60	770	48.89	134	8.51
Total	5698	1889	33.15	3238	56.83	571	10.02

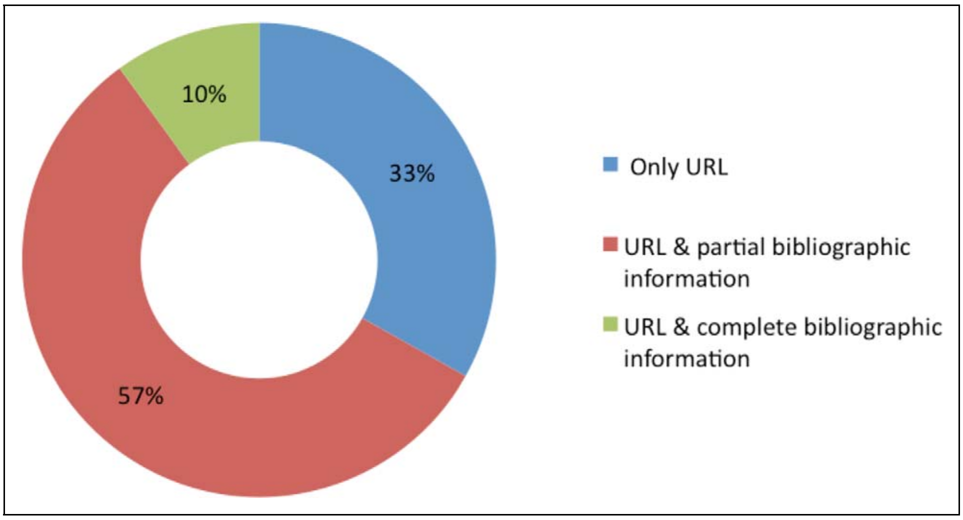

Figure 2. Contents of URLs.

Table 4. Percentage of missing URLs.

Publication year of the conference proceedings	Total no. of citations	Total no. of URLs	No. of active URLs	Percentage of active URLs	No. of missing URLs	Percentage of missing URLs
2001	1165	311	123	39.55	188	60.45
2002	1453	387	178	45.99	209	54.01
2003	1474	449	199	44.32	250	55.68
2004	1519	537	201	37.43	336	62.57
2005	1636	558	250	44.80	308	55.20
2006	1867	882	454	51.47	428	48.53
2007	2022	775	422	54.45	353	45.55
2008	2179	804	443	55.10	361	44.90
2009	2129	860	504	58.60	356	41.40
2010	301	135	70	51.85	65	48.15
All years	15745	5698	2844	49.91	2854	50.09

percent) to 2009 (41.40 percent), which is also shown in Figure 3.

Table 4 also displays the summary of active URLs and it is very clear that the year of publication is significantly related to the citations' stability, and that, specifically, the URLs from the most recently published conference articles were more likely to remain active compared to those cited in older articles. This is evident from the fact that 58.60 percent of URLs for the year 2009 were active, whereas only 39.55 percent of URLs were active for the year 2001. This shows that there was an increase in the percentage of active URLs from 2001 to 2010.

The results of this study indicate that the early-published articles (2001 to 2005) collectively had a greater number of missing web references. The percentage of missing URLs was greater in the first period – 57.58 percent (2001–2005) as compared with 45.22

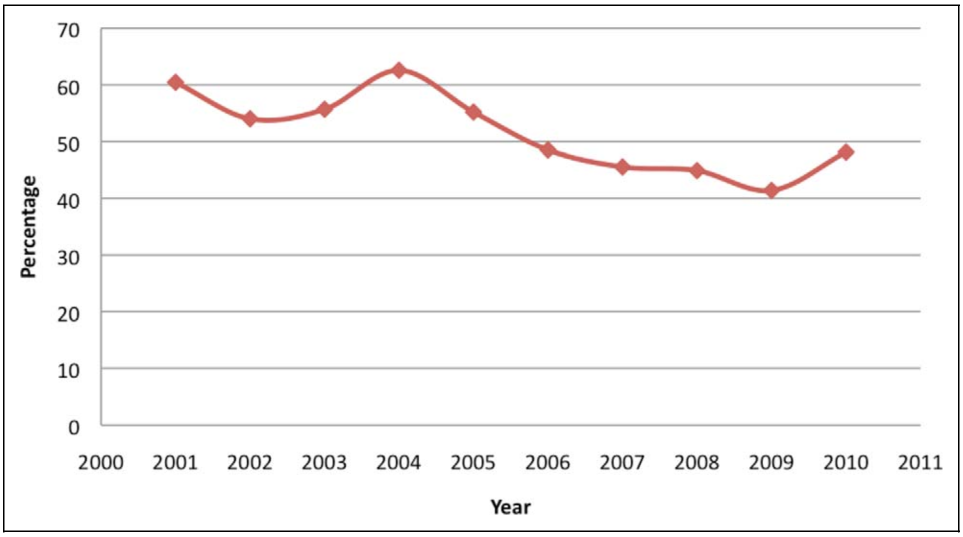

Figure 3. Percentage of missing URLs.

Table 5. Missing URLs cross tabulated by conference proceedings.

Publication year of the conference proceedings	Total no. of citations	Total no. of URLs	No. of active URLs	Percentage of active URLs	No. of missing URLs	Percentage of missing URLs
CALIBER	6144	2489	1363	54.76	1126	45.24
ILA	5949	1634	733	44.86	901	55.14
NACLIN	3652	1575	750	47.62	825	52.38
Total	15745	5698	2846	49.95	2852	50.05

Table 6. HTTP errors associated with missing URLs.

HTTP error	No. of missing URLs	Percentage of missing URLs
HTTP 300	0	0
HTTP 301	2	0.07
HTTP 302	4	0.14
HTTP 400	27	0.95
HTTP 401	0	0
HTTP 403	353	12.37
HTTP 404	1521	53.29
HTTP 406	20	0.70
HTTP 410	5	0.18
HTTP 500	910	31.89
HTTP 501	7	0.25
HTTP 502	1	0.04
HTTP 503	4	0.14
Total	2854	100

percent for the second period (2006–2010). This shows that there is a positive correlation between the percentage of missing URLs and age (r = .763, p = .017) which is consistent with the findings of other previous studies (Rumsey, 2002; Dimitrova and Bugeja, 2007, Sampath Kumar and Manoj Kumar, 2012; Sampath Kumar and Vinay Kumar, 2013).

Missing URLs cross tabulated by conference proceedings

Table 5 depicts the percentage of missing URLs in articles published in the selected conference proceedings. It was found that this percentage varied among the different conference proceedings. The percentage of missing URLs cited in ILA conference proceedings was 55.14 percent whereas in NACLIN conference proceedings it was 52.38 percent and in CALIBER conference proceedings only 45.24 percent.

HTTP Errors associated with missing URLs

The HTTP status codes of missing URLs are presented in Table 6. As reported above, as many as 2,854 (50.09 percent) cited URLs were found to be missing and the searches had resulted in different types of HTTP error message. As shown in Table 6, the HTTP error 404 – ‘file not found’ – error message accounted for more than half (53.29 percent) of all error messages. This result is comparable to those of Spinellis (2003) and Goh and Ng (2007), who found in their studies that 60 percent of all URL failures were due to HTTP 404 error. Goh and Ng (2007) stated that the reasons for obtaining this response are

Table 7. Domains associated with active and missing URLs.

Domains	Total no. of URLs	Percentage	No. of active URLs	Percentage of active URLs	No. of missing URLs	Percentage of missing URLs
.org	1799	31.57	932	51.81	867	48.19
.com	1474	25.87	881	59.77	593	40.23
.edu	717	12.58	233	32.50	484	67.50
.ac	560	9.83	247	44.11	313	55.89
.gov	279	4.90	114	40.86	165	59.14
.net	160	2.81	79	49.38	81	50.63
.ernet	64	1.12	37	57.81	27	42.19
Others*	645	11.32	321	49.77	324	50.23
Total	5698	100	2844	49.91	2854	50.09

*Others: .nic, .info, .geo domains, etc,.

varied. The problem could be due to an unreachable web server as a result of an unresolved host name or a failure to contact the target web server after a successful DNS name resolution. This error is due to changes in the URL brought about by file/directory name changes, removal of files or relocation of files. Unfortunately this HTTP code does not specify the exact nature of the problem.

The second most common error message found was HTTP 500 – 'Internet server error' – that accounted for 31.89 percent of all the missing URLs. Another significant HTTP error message was HTTP 403 – 'Forbidden' – that accounted for 12.37 percent. The other nine types of HTTP error message encountered in this study are negligible because collectively they accounted for just 2.45 percent of the missing URLs.

Domains associated with active and missing URLs

During the last decade many studies were conducted by considering domain names of URLs as a major characteristic feature to understand the relationship between the accessibility of web citations and the domain parts: portal, domain, directory and file.

In this study, eight different types of top-level domain were identified among the URLs. They were: .org, .com, .edu, .ac, .gov, .net .ernet and 'others'. Table 7 reveals that, of the 5,698 URLs, more than 50 percent were in just two domains – .org and .com – which accounted for 1,799 (31.57 percent) and 1,474 (25.87 percent) of all URLs respectively. Other domains were .edu (12.58 percent), .ac (9.83 percent), .gov (4.90 percent), .net (2.81 percent) and .ernet (1.12 percent) respectively. Table 7 also records the active and missing URLs within each domain type.

File formats associated with active and missing URLs

Similar to previous studies by McCown, et al., (2005), Maharana, et al., (2006) and Saberi and Abedi (2012), the URLs in this study have been categorized into different file formats and presented in Table 8 for further analysis.

According to the data given in Table 8 the highest percentage of cited web resources belonged to HTML files. Out of 5,698 URLs, 3,903 (68.50 percent) were of HTML files followed by 505 .PDF files (8.86 percent) URLs in second place and 232 files ending with numbers, (4.07 percent) URLs occupied the third place. The remaining file types had negligible numbers of URLs. These findings are consistent with those of the earlier studies noted above, which also reported that most of the cited web resources consisted of HTML files.

As reflected in Table 8, the highest percentage of missing URLs (URLs) were found among the URLs of file formats .TXT (100 percent), followed by .PDF (90.10 percent). The .DOC and .PPT file formats each had 77.78 percent of missing URLs. As noted by McCown et al., (2005) the .txt file extension is typically used for temporary textual data which may have been converted into a webpage later or simply disregarded as it became stale.

Path depth and corrosion of URLs

The path depths of the URLs and corresponding active and missing URLs are displayed in Table 9. The URL's path depth could be associated with link failure due to increasing complexity as the length of a URL increases (Goh and Ng, 2007). In this study we called this link failure 'URL corrosion'. One of the objectives of this study was to verify whether there is any relation between the path depth and URL corrosion. To determine how URL path depth influences URL corrosion rates, the researcher calculated the file path depth for each active and missing URL as per the method followed by Spinellis (2003) and McCown, et al., (2005). In order to calculate path

Table 8. File formats associated with active and missing URLs.

File types	Total no. of URLs	Percentage	No. of active URLs	Percentage of active URLs	No. of missing URLs	Percentage of missing URLs
html files	3903	68.50	2156	55.24	1747	44.76
.pdf	505	8.86	50	9.90	455	90.10
Ends only with number	232	4.07	105	45.26	127	54.74
.asp	76	1.33	24	31.58	52	68.42
.shtml	62	1.09	30	48.39	32	51.61
.php	45	0.79	11	24.44	34	75.56
.doc	36	0.63	8	22.22	28	77.78
.ppt	27	0.47	6	22.22	21	77.78
.txt	13	0.23	0	0	13	100
Others*	799	14.02	454	56.82	345	43.18
Total	5698	100	2844	49.91	2854	50.09

*Others: .cfm, .nsf, .pl, .cgi, etc,.

Table 9. Path depth and corrosion of URLs.

Path depth (PD)	Total no. of URLs	Percentage	No. of active URLs	Percentage of active URLs	No. of missing URLs	Percentage of missing URLs
PD=0	1415	24.83	1055	74.56	360	25.44
PD=1	761	13.36	332	43.63	429	56.37
PD=2	1368	24.01	625	45.69	743	54.31
PD=3	1073	18.83	433	40.35	640	59.65
PD=4	681	11.95	253	37.15	428	62.85
PD=5	226	3.97	81	35.84	145	64.16
PD≥6	174	3.05	65	37.36	109	62.64
Total	5698	100	2844	49.91	2854	50.09

depth, the researcher added one to the depth for every directory or file after a domain name. For example, http://tumkuruniversity.in/ has a path depth of 0, http://tumkuruniversity.in/dept has a depth of 1, tumkuruniversity.in/dept/lis has a depth of 2, etc. Likewise the URL path depth could increase to levels 3, 4, 5, etc.

In this study each of the 5,698 URLs was verified for their path depth and classified and grouped according to path depth level for path depths of 0, 1, 2, 3, 4, 5 and 6. Any URLs having path depth level 6 and above were grouped into path depth level 6. The resultant data is presented in Table 9.

The purpose of classifying the extracted URLs into their respective path depth categories was to assess the association between URL path depth and the level of URL corrosion. The analysis of data presented in Table 9 brings home the fact that there was a strong association between URL path depth and corrosion i.e. the greater the path depth, the higher the proportion of missing URLs.

This is substantiated with the fact that, during the test that was carried out in the above mention period, the URLs with ‘0’ level path depth (PD = 0) were the least to miss (25.44 percent), followed by PD = 1 (56.37 percent), PD = 2 (54.31 percent), PD = 3 (59.65 percent), PD ≥ 6 (62.64 percent) and PD = 4 (62.85 percent) respectively in an ascending order among the URLs of the respective groups. And the URLs with path depth 5 were the highest to miss (64.16 percent) (Figure 4).

We performed the correlation analysis to know the correlation between the path depth and the web corrosion. Not surprisingly the correlation analysis indicates that there is significant association between the path depth and web corrosion, i.e. the longer the path depth, the higher the level of web corrosion ($r = .773$, $p = .042$).

Half-life of URLs

The half-life of URLs is the time required for half of all online citations (URLs) in articles to disintegrate. In order to estimate the half-life of the URL citations examined in this study, the researcher adopted the procedure used by Koehler (1999); Tyler and McNeil (2003); Dimitrova and Bugeja (2007);

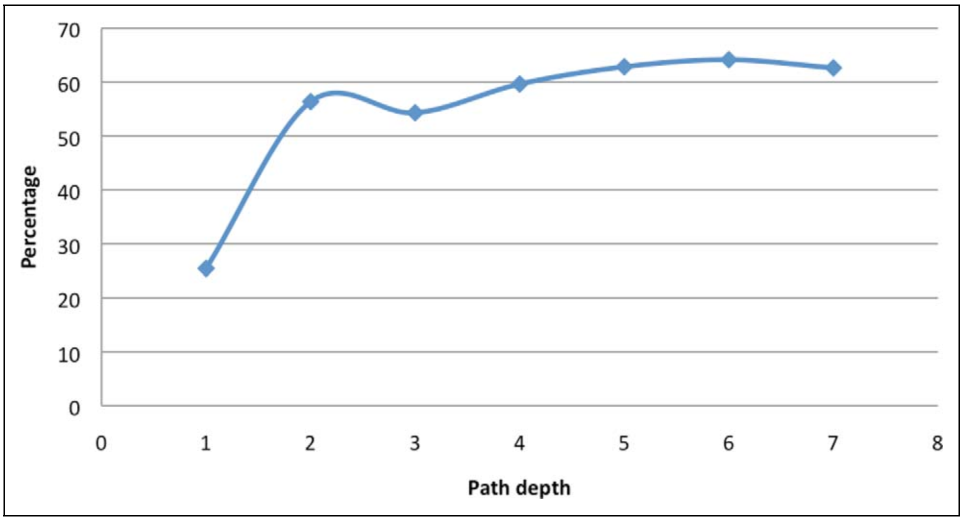

Figure 4. Path depth and corrosion of URLs.

Table 10. Half-life of URLs.

Publication year of the conference proceedings	Time (t)	Total no. of URLs (W0)	No. of active URLs (Wt)	* Half-life (T_h)
2001	10	311	123	7.47241
2002	9	387	178	8.032442
2003	8	449	199	6.814618
2004	7	537	201	4.937482
2005	6	558	250	5.17984
2006	5	882	454	5.218736
2007	4	775	422	4.561246
2008	3	804	443	3.488823
2009	2	860	504	2.594327
2010	1	135	70	1.055373
All years		5698	2844	** 4.93553

$*Half\ life\ [t_{(h)}] = \frac{t \ln(0.5)}{\ln w_{(t)} - \ln W_{(O)}}$.

** Average half-life.

Mardani and Sangari (2013). The following formula was used to calculate the half-life of URLs for each conference year:

$$Half\ \ life\ [t_{(h)}] = \frac{t\ \ln\ (0.5)}{\ln\ w_{(t)} - \ln\ \ W_{(0)}}$$

where $W_{(0)}$ is the number of working URLs at the time of publication, $W_{(t)}$ is the number of working URL citations at some later time t.

Using this formula the half-life has been calculated using MS-Excel and the data is presented in Table 10. As can be observed in the table, the average half-life for the missing URLs was estimated to be 4.93 years. This means that it will take about 4.93 (approximately 5 years) years for half of the URL citations to vanish.

Testing of hypotheses

The formulated hypotheses were tested with the data using the statistical program SPSS (19.0 version) for Windows. The Pearson's Correlation and Chi-Square test were carried out to check the significance level. A p < .05 level of significance was used to test the hypotheses. The hypothesis, corresponding table number, test applied, p value and the result are presented in Table 11.

Discussion and conclusion

The researcher examined the corrosion of URLs cited in three Indian LIS conference proceedings. It was found the average number of URLs per article ranged from a low of 1.99 in 2001 to a high of 5.63 in 2010.

Table 11. Testing of hypotheses.

Sl. No	Hypotheses	Test	P value and result
1	There is an increase in the number of URLs during 2001 to 2010 in selected Indian LIS scholarly communications (Table 2).	Pearson's correlation	p = .014 Hypothesis Accepted
2	Corrosion of URLs will increase as their age increases (Table 2).	Pearson's correlation	p = .017 Hypothesis Accepted
3	There is significant association between the corrosion of URLs and top-level domains (Table 7)	Chi-Square	p = .000 Hypothesis Accepted
4	The path depth and corrosion URLs are positively correlated (Table 9).	Pearson's correlation	p = .042 Hypothesis Accepted

The average number of URLs per article was 3.35 across all the 1,700 conference articles. Further, the years 2006 and 2010 witnessed the highest percentage of URLs, with 47.24 percent and 44.85 percent respectively, which were comparatively higher than the decadal percentage of URLs (36.19 percent). The percentage of URLs had increased from the year 2001 (26.70 percent) to 2010 (44.85 percent). This clearly indicates that the use of URLs as citations increased during the last 10 years. Many of the citations to URLs examined in this study contained partial bibliographical details (56.83 percent) and 33.15 percent contained only URLs. Editorial staff should inform the authors of their journals to provide full bibliographical details along with URLs. Full bibliographical details should include the name of the author(s), title, URL address and date of access. This will help the future researcher to retrieve the information if the URL is not accessible. Casserly and Bird (2003) suggested that the creator/author(s) of web sources should include contact information. Further they suggested that editorial staff should require authors to adhere to citation policies, style manuals and formats established by their journals. Editors should also review the citation guidelines frequently and modify them as needed.

The study has also provided evidence of the impermanent nature (corrosion) of URLs. In our dataset we found that 50.09 percent of URLs cited in articles of conference proceedings were missing. From Sellitto's (2005) perspective the loss of a large number of URLs cited in scholarly articles has important implications for the authors and the academic community. The disappearance of previously cited URLs challenges the reader's traditional assumption of reference availability and access. Furthermore, missing references tend to stymie the ability of a reader to further investigate interesting or significant aspects of an article. Indeed the loss of cited sources tends to weaken an article's theoretical foundation – one of the fundamental objectives associated with scholarly literature.

From the above discussion it can be concluded that the corrosion of URLs in scholarly publications will have an impact on electronic publications. Corrosion of URLs cited in scholarly communication has important implications for authors, publishers, future researchers and also academic community. Hence the list of URLs cited in an article should be always accessible to the readers.

To avoid possible corrosion of URLs, the authors, editors and publishers should check the accessibility of a URL before it is used in the article. In case it is not accessible, they should try to recover the missing URL using search engines or Internet archives. Authors should adhere to the guidelines of the publishers while citing URLs in the article. Publishers need to review their reference guidelines frequently and modify them, as needed. Both authors and publishers need to archive the web content cited in the articles they publish. This will help the future researchers to contact the authors or publishers of the web document when it not accessible to them.

References

Aronsky D, Madani S, Carnevale RJ, Duda S and Feyder FT (2007) The prevalence and inaccessibility of Internet references in the biomedical literature at the time of publication. *Journal of the American Medical Informatics Association* 14(2): 232–234.

Casserly MF and Bird JE (2008) Web citation availability: a follow-up study. *Library Resources and Technical Services*. 52(1): 42–53.

Casserly MF and Bird JE (2003). Web citation availability: analysis and implications for scholarship. *College* and *Research Libraries* 64(7): 300–317.

Dellavalle RP, Hester EJ, Heilig LF, Drake AL, Kuntzman JW, Graber M and Schilling LM (2003) Going, going, gone: lost Internet references. *Science* 302(5646): 787–788.

DELNET (2013) Retrieved from (http://delnet.nic.in/).

Dimitrova DV and Bugeja M (2007) The half-life of Internet references cited in communication journals. *New Media and Society* 9(9): 811–826.

Ducut E, Liu F and Fontelo P (2008) An update on uniform resource locators (URL) decay in MEDLINE abstract and measures for its mitigation. *BMC Medical Informatics and Decision Making* 8(23). Retrieved from. http://www.biomedcentral.com/1472-6947/8/23.

Goh DHL and Ng PK (2007) Link decay in leading information science journals. *Journal of the American Society for Information Science and Technology* 58(1): 15–24.

INFLIBNET (2013) Retrieved from: http://www.inflibnet.ac.in/caliber2013/

ILA (2013) Retrieved from (http://www.ilaindia.net/)

Isfandyari-Moghaddam A and Saberi MK (2011) The life and death of URLs: the case of Journal of the Medical Library Association. *Library Philosophy and Practice*. Retrieved from http://digitalcommons.unl.edu/libphilprac/592/

Koehler W (1999) An analysis of web page and web site, constancy and permanence. *Journal of the American Society for Information Science* 50(2): 162–180.

Lopresti R (2010) Citation accuracy in environmental science journals. *Scientometrics* 85(3): 647–655.

Maharana B, Nayak K and Sahu NK (2006) Scholarly use of web resources in LIS research: A citation analysis. *Library Review* 55(9): 598–607.

Mardani and Sangari (2013) An analysis of the availability and persistence of web citations in Iranian LIS journals. *International Journal of Information Science and Management* 3(1): 29–42.

Mardani AH (2012) An investigation of the web citations in Iran's chemistry articles in SCI. *Library Review* 61(1): 18–29.

Markwell J and Brooks DW (2003) Link rot limits the usefulness of the web based educational material in biochemistry and molecular biology. *Biochemistry and Molecular Biology Education* 31(1): 69–72.

McCown F, Chan S, Nelson ML and Bollen J (2005) The availability and persistence of web references. *D-Lib Magazine*. Retrieved from http://iwaw.europarchive.org/05/papers/iwaw05-mccown1.pdf

NACLIN (2013) Retrieved from: http://www.naclin.org/Naclin_2013.pdf

Nagaraja A, Joseph SA, Polen HH and Clauson KA (2011) Disappearing act: Persistence and attrition of uniform resource locators (URLs) in an open access medical journal. *Program: Electronic Library and Information Systems* 45(1): 98–106.

Rhodes S (2010) Breaking down link rot: The Chesapeake project legal information archive's examination of URL stability. Retrieved from http://www.llrx.com/features/linkrot.htm

Riahinia N, Zandian F and Azimi A (2011) Web citation persistence over time: a retrospective study. *The Electronic Library* 29(5): 609– 620.

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(2): 27–39.

Saberi MK, Isfandyari-Moghaddam A and Mohamadesmaeil S (2011) Web Citations Analysis of the JASSS: the First Ten Years. *Journal of Artificial Societies and Social Simulation* 14(4): 22. Retrieved from http://jasss.soc.surrey.ac.uk/14/4/22.html

Saberi MK and Abedi H (2012) Accessibility and decay of web citations in five open access ISI journals. *Internet Research* 22(2): 234–247.

Sadat-Moosavi A, Isfandyari-Moghaddam A and Tajeddini O (2012) Accessibility of online resources cited in scholarly LIS journals: A study of Emerald ISI-ranked journals. *Aslib Proceedings* 64(2): 178–192.

Sampath Kumar and Vinay Kumar (2013) HTTP 404-page (not) found: Recovery of decayed URL citations. *Journal of Informetrics* 7,145–157.

Sampath Kumar BT and Manoj Kumar KS (2012) Persistence and half-life of URL citations cited in LIS open access journals. *Aslib Proceedings* 64(4): 405–422.

Sampath Kumar BT and Prithviraj KR (2012) Availability and persistence of web citations in Indian LIS literature. *The Electronic Library* 30(1): 19–32.

Sellitto C (2004) A study of missing web-cites in scholarly articles: Towards an evaluation framework. *Journal of Information Science* 30(6): 484–495.

Sellitto C (2005) The impact of impermanent web-located citations: a study of 123 scholarly conference publications. *Journal of the American Society for Information Science and Technology* 56(7): 695–703.

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

Thorp AW and Brown L (2007) Accessibility of Internet references in *Annals of Emergency Medicine*: is it time to require archiving? *Annals of Emergency Medicine* 50(2).

Tyler D and McNeil B (2003) Librarians and link rot: a comparative analysis with some methodological considerations. *Portal: Libraries and the Academy* 3(4): 615–632.

Veronin MA (2002) Where are they now? A case study of health-related Web site attrition. *Journal of Medical Internet Resources* 4(2).

Wagner C, Gebremichael MD and Taylor M (2009) Disappearing act: decay of uniform resource locators in health care management journals. *Journal of Medical Library Association* 97(2): 122–130.

u Z (2009) An empirical study of the accessibility of web references in two Chinese academic journals. *Scientometrics* 78(3): 481–503.

Zhang Y (2001) Scholarly use of Internet-based electronic resources. *Journal of American Society for Information Science and Technology* 52(8): 628–654.

Author biographies

K.R. Prithviraj is Research and Training Associate, Department of Library and Information Science, Directorate of Distance Education, Kuvempu University, Karnataka, India. He obtained MSc in Library and Information

from Kuvempu university in the year 2007 and PhD in Library and Information Science in the year 2013. He has published and presented papers in national and international journals and conferences respectively. His research interests include the applications of information technology in library services, webometrics, library management and Internet use. Email: prithviraj.kr@gmail.com

B.T. Sampath Kumar is currently an Associate Professor in the Department of Studies and Research in Library and Information Science, Tumkur University, Tumkur, India. He achieved First Rank in MSc Library and Information Science and received the Gold Medal. He obtained his PhD from Kuvempu University in 2003 and also obtained Post-graduate Diploma in Computer Science (PGDCS) from the University of Hyderabad, Andra Pradesh. He has received the "LAPV Verghese Award for the best research paper published in the *ILA Bulletin* in 2005. B.T. Sampath Kumar has published more than 85 research papers in national reputed journals and conferencesproceedings. His research interests include the applications of information technology in library services, World Wide Web resources and their evaluation, search engines, web designing and Internet use. Contact: Email: sampathbt_2001@rediffmail.com