URLs Link Rot: Implications for Electronic Publishing

D Vinay Kumar
Lecturer, Department of Library and Information Science, Kuvempu University
Jnana Sahyadri, Shankaraghatta, Karnataka, India; (E): vinay.86.kumar@gmail.com

B T Sampath Kumar
Associate Professor, Department of Library and Information Science,
Tumkur University, Tumakuru, Karnataka, India; (E): sampathbt2001@gmail.com

D R Parameshwarappa
Librarian, Apollo Research Institute, Apollo Hospitals,
Bannerghatta Road, Bangalore, Karnataka, India; (E): parmeshdr08@gmail.com

DOI: 10.18329/09757597/2015/8105

World Digital Libraries 8(1): 59–66 (2015)

Abstract

In recent years the authors of scholarly publications have relied on e-resources. But e-resources have raised the question of permanency on the web. In this context, this article investigates the availability, persistence of Uniform Resource Locator (URL) citations cited in two Library and Information Science (LIS) journal articles published by Emerald Publishers during 2008 and 2012. In total, 2477 URLs cited in 406 research articles published in two LIS journals spanning a period of five years (2008–2012) were extracted. The study found that 23.81 per cent (2,477 out of 10,400 references) of URLs were cited in these journal articles. 49.53 per cent of URL citations were not accessible and the remaining 51.47 per cent of URL citations were still accessible. The study used W3C link checker to identify HTTP errors associated with missing URLs. HTTP 500 error message—'page not found' was the overwhelming message that represented 39.18 per cent of all HTTP error messages. This study attempts to focus on URLs link rot and its implications for electronic publishing.

Keywords: Link rot, e-Publishing, Web citations, Wayback machine, HTTP errors

1. Introduction

Publishing segment has been greatly influenced by Information and Communication Technology. E-publishing is a remarkable change in the field of publishing. Scholarly world has witnessed a shift from conventional publishing to electronic publishing. This is because e-publishing has reduced the cost and time that are required for conventional publishing. Ever since the first e-book was published in 1985, electronic publishing has seen steady growth (Moorthy and Karisiddappa 1996). Today, millions of electronic resources are available and being published over the World Wide Web (WWW). E-publishing has also influenced the convenience of researcher in information search and retrieval. Despite its convenience, recent studies have documented the problem of loss of web resources or URL link rot. It is a serious issue for web masters and academic community (Dimitrova and Bugeja 2007). Hence, this article attempts to identify the permanence and URL link rot by studying the URLs cited in two LIS e-journals published by Emerald Publishers.

2. Objectives of the Study

- To know the extent of the use of web resources in LIS scholarly communication.
- To know the top-level domains and file formats associated with URL citations.
- To check the accessibility and decay of URL citations cited in LIS scholarly communication.
- To know the HTTP error messages associated with missing URL citations.

3. Methodology

The data for the present study was drawn from a selective sample of LIS scholarly journals published by Emerald Publishers. The following two journals were selected for the current study:

- Program: Electronic Library and Information Systems
- The Electronic Library

The time frame for the analysis was the publication years 2008–2012. All research articles published during the 5-year period were downloaded and saved. In totality, 10,400 references were selected from 406 articles published in the selected journals. Only those references that appeared as a list at the end of the article under the references section were considered. Editorial notes or book reviews, expanded bibliographies, end notes, foot notes, e-mail links, and annotations were not considered and were not tested or counted in our dataset.

The study aims to check the availability of URL citations and their persistence. Duplicate URLs were recognized and removed from the list. W3C link checker (http://validator.w3.org/checklink) was used to check the URLs' existence (Sellitto 2005). After the initial check, the URL citations were grouped as 'active' and 'missing' URL citations. The missing URLs were entered in the Wayback machine and the 'Take Me Back' button was clicked to retrieve the missing URLs from the Wayback machine.

4. Results

4.1 Distribution of articles, citations, and URLs

It is evident from Table 1 that the percentage of URL citations is not consistent during 2008–2012. The percentage of URL citations varied from the highest of 29.78 per cent in 2009 to a lowest of 18.06 per cent in 2012. It is also an observable fact that the percentage of URLs per article is above 5 per cent.

Table 2 illustrates the summary of active and missing URL citations. The percentage of missing URL citations varied from a highest of 69.15 per cent in 2009 to a low of 33.61 per cent in 2012. A total of 49.53 per cent of URL citations were not accessible during the initial check of URLs in W3C link checker. It is evident from

Table 1: Year-wise distribution of articles, citations, and URLs

Year	Total articles	Total citations	% Average citations per article	URL citations	% of URLs	% of URLs per article
2008	75	1859	24.78	477	25.65	6.36
2009	98	2122	21.65	632	29.78	6.44
2010	84	1782	21.21	452	25.36	5.38
2011	78	2090	26.79	456	21.81	5.84
2012	71	2547	35.87	460	18.06	6.47
Total	**406**	**10,400**	**25.61**	**2,477**	**23.81**	**6.10**

Table 2: Summary of year-wise active and missing URL citations

Year	Total URLs	Active URL citations		Missing URL citations	
		Number	%	Number	%
2008	477	256	53.66	221	46.34
2009	632	195	30.85	437	69.15
2010	452	251	55.53	201	44.47
2011	456	263	57.67	193	42.33
2012	460	310	67.39	150	33.61
Total	**2,477**	**1,275**	**51.47**	**1,202**	**49.53**

the table that as the age of the URLs increases their permanency is questionable. The decreasing trend of missing URLs from 2009 to 2010 could be witnessed in Table 2 and Figure 1.

4.2 Recovery of missing URL citations with HTTP error messages

Around 1,202 missing URL citations were entered in the Wayback machine and the 'Take me back' button was clicked. About 40.60 per cent of missing URLs are recovered back from Wayback machine. The search in the Wayback machine yielded a total of 488 URLs that have been retrieved successfully (Table 3).

The 'HTTP 500 error message' was the overwhelming message encountered and represented 39.18 per cent of all HTTP error messages, followed by HTTP 404

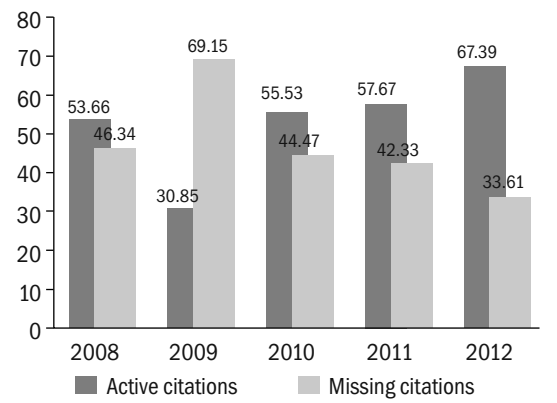

Figure 1: Summary of active and missing citations

(36.52 per cent) and HTTP 403 (12.56 per cent). It is interesting to note that 43.74 per cent of URLs were recovered from HTTP 500 error message and 43.05 per cent of missing URL

citations were recovered from HTTP 403 error message. 36.22 per cent of missing URLs with HTTP 404 error message were recovered from the Wayback machine.

The top-level domain having the greatest number of missing URLs was the .gov domain (67.33 per cent) followed by .edu (57.09 per cent). It is notable that a low level of URL link rot is associated with the .net (29.91 per cent).

Table 4 illustrates the distribution of different domains associated with missing and retrieved URLs. URLs with .com/.co (45.24 per cent) and geo domains (44.31 per cent) are the highest retrieved missing URLs followed by URLs with .org (42.18 per cent) and .net (37.14 per cent) domains. It is interesting to note that the URLs with academic (30.34 per cent) and educational (34.32 per cent) domains are the least recovered URLs.

The data as illustrated in Table 5 indicates that the greatest numbers of cited URLs in LIS scholarly communication web resources with HTML/HTM files (65.48 per cent). Out of 2,477 URLs, 1,622 cases are HTML files, followed by 689 (27.82 per cent) that are PDF files and 112 that are PHP files (4.52 per cent).

Table 5 also depicts the distribution of recovered URLs having different file types. URLs with PDF file types (41.21 per cent) and HTML file types (40.93 per cent) recorded highest recovery whereas URLs with .jsp file types and .doc recorded 15.38 per cent and 23.08 per cent respectively, which is a very low recovery rate. Missing URLs with PDF files format showed highest recovery than of other file types.

4.3 Path depth associated with missing and recovered URL citations

The data presented in Figure 2 illustrates the missing and recovered URLs with different path depths. It is clear from the analysis that higher the path depth of the URLs, less is the recovery whereas, lower the path depth of URLs, higher the recovery. Path depth-1 records 30.43 per cent of recovery whereas URLs with path depth of 8 recorded very less recovery, i.e., 15.78 per cent. URL with path depth of 4 and less shows highest recovery, whereas URLs with the path depth of 5, 7 and above recorded less recovery.

Table 3: HTTP errors associated with missing and recovered URL citations

HTTP error	Missing URLs		Recovered missing URLs	
	Number	%	Number	%
302	4	0.33	2	50.00
303	2	0.17	2	100.00
400	118	9.82	50	42.37
403	151	12.56	65	43.05
404	439	36.52	159	36.22
406	5	0.42	1	20.00
410	1	0.08	0	0.00
412	1	0.08	0	0.00
415	1	0.08	0	0.00
500	471	39.18	206	43.74
501	3	0.25	1	33.33
502	4	0.33	2	50.00
503	1	0.08	0	0.00
600	1	0.08	0	0.00
Total	**1,202**	**100**	**488**	**40.60**

Table 4: Domains associated with missing and recovered URL citations

Domains	Total URLs	Missing URL citations		Recovered URL citations	
		Number	%	Number	%
.org	821	422	51.40	178	42.18
.com/.co	620	252	40.65	114	45.24
.edu	296	169	57.09	58	34.32
.gov	101	68	67.33	24	35.29
.ac	191	89	46.60	27	30.34
.net	117	35	29.91	13	37.14
Geo domains	331	167	50.45	74	44.31
Total	**2,477**	**1,202**	**48.53**	**488**	**40.60**

Table 5: File types associated with missing and recovered URLs

File types	Total URLs		Missing URLs		Recovered URLs	
	Number	%	Number	%	Number	%
Html/htm/shtm/xml	1,622	65.48	733	45.19	300	40.93
.pdf	689	27.82	398	57.76	164	41.21
.doc	22	0.89	13	59.09	3	23.08
.xls	1	0.04	0	0.00	0	0.00
.jpg	0	0.00	0	0.00	0	0.00
.cgi	5	0.20	7	60.00	3	100.00
.jsp	26	1.05	11	50.00	2	15.38
.php/txt	112	4.52	40	37.50	15	35.71
Total	**2,477**	**100.00**	**1,202**	**48.53**	**488**	**40.60**

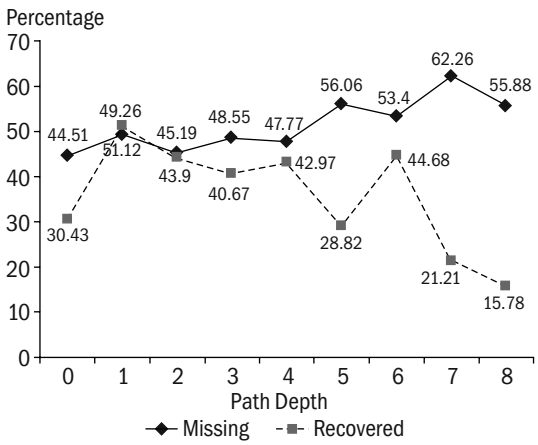

Figure 2: Path depth associated with missing and recovered URL citations

The distribution of URLs by character length is given in Table 6. It illustrates the percentage of missing URLs and recovered URLs from various URLs having number of character length. The highest percentage of missing URLs (61.30 per cent) was found in the URLs having 51–60 characters followed by >70 character length (55.48 per cent).

It is clear from the analysis of Table 6 that the highest character length of URLs lowers the recovery rate. The data indicated in Table 6 that the Wayback machine could recover only 24.41 per cent of missing URLs with the character length of >70 followed by 37.50 per cent of URLs with 61–70 character length which recorded lowest recovery rate; whereas, the recovery is highest among the URLs with character length of 31–40 (53.33 per cent) and 21–30 (49.31 per cent).

5. Discussion and Conclusion

E-publishing has gained momentum in the scholarly world. The low cost to produce information electronically and the speed of publication have made it an inevitable source in the scholarly environment. E-publishing has given opportunity to post any kind of information on the Internet. But, it has also raised the question of unavailability of web resources in the longer duration. Several studies have shown that URLs of web pages would disappear as their age increases (Wren *et al.*; Aronsky *et al.* 2007; Dimitrova and Bugeja 2007; Ducut *et al.* 2008; Goh and Ng 2007; Sampath and Vinay 2013). Since majority of web sources vanished from the original web location due to various reasons, the users of the web find it very difficult to see the needed web sources. There are many reasons why online sources disappear.

Table 6: Character length associated with missing and recovered URLs

Character length of URLs	Total URLs	Missing URL citations		Recovered URL citations	
		Number	%	Number	%
0–20	73	32	43.84	13	40.62
21–30	174	73	41.95	36	49.31
31–40	346	150	43.35	80	53.33
41–50	618	266	43.04	122	45.86
51–60	416	255	61.30	111	43.52
61–70	332	168	50.60	63	37.50
>70	465	258	55.48	63	24.41
Total	**2,477**	**1,202**	**48.53**	**488**	**40.60**

Some websites simply cease to exist because their website domain names expire. Other web pages are removed from the web by the website creator. Some websites are redesigned with new file structures (Dimitrova and Bugeja 2007). In the present study we found that 48.53 per cent of URL citations were not accessible and the remaining 51.47 per cent of URL citations were still accessible. The HTTP 500 error message—was the overwhelming message encountered and represented 36.52 per cent of all HTTP messages. The study also checked the rate of URL link rot associated with the URLs with different domains. URLs with .gov domain were found to be highest missing which represents 67.33 per cent of the total followed by .org domain (57.09 per cent). It is a considerable observation of the study that the electronic publication from governmental and organizational domains are most liable to disappear.

Many studies (Dimitrova and Bugeja 2007; Wren *et al.* 2006) have already recommended various strategies for the publishers, editors, and authors in order to increase the rate of availability of URLs. Wren *et al.* opined that loss of URLs would be continued until better preservation policies are adopted. Authors are in a vulnerable position to maintain the web pages who post content on the Internet. They are often ill equipped to maintain it (Sampath and Vinay 2013). But authors may give parallel citations to e-resources that are referred in their scholarly work. In the long term, all research organizations must come to the realization that, due to the dynamic nature of the Internet, certain minimum standards must be maintained in posting scholarly information there (Altman and King 2007). It is recommended to publishers of electronic web pages that they have to maintain lower path depth in the URL address, because several studies identified that lower the path depth lesser the missing citations (Goh and Ng 2007; Sampath and Vinay 2013). Apart from this, the Digital Object Identifier (DOI) system could be used with the URLs of electronic resources. It is considered as PURL—Persistence Uniform Resources Locator. It would be solution to disappearing act of web pages (DOI.org 2012).

References

Altman M and King G. 2007. **A proposed standard for the scholarly citation of quantitative data.** *D-LibMagazine.* Available at <http://www.dlib.org/dlib/march07/altman/03altman.html> (last accessed on June 25, 2013).

Aronsky D, Madani S, Carnevale R J, Duda S, and Feyder F 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* **14**(2): 232–234.

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

DOI.org. 2012. **DOI System and Persistent URLs (PURLs).** Available at <http://www.doi.org/factsheets/DOI_PURL.html> (last accessed on July 1, 2013).

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. Available at <http://www.biomedcentral.com/1472-6947/8/23> (last accessed on July 5, 2013).

Goh D H L and 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.

Moorthy L R and Karisiddappa C R. 1996. **Electronic publishing: Impact and implications on library and information centres.** pp. 15–34. In *Digital Libraries: Dynamic storehouse of digitized information*, edited by N M Malwad, T B Rajashekar Rao, I K Ravichandra and N V Satyanarayana, New Delhi: New Age International.

Sampath Kumar D and Vinay Kumar B T. 2013. **HTTP 404-page (not) found: Recovery of decayed URL citations**. *Journal of Informetrics* **7**: 145–157. Available at <http://dx.doi.org/10.1016/j.joi.2012.09.007> (last accessed on June 29, 2013).

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* **142**: 1147–1152.

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.