July 2021

Decay of URL References cited in DESIDOC Journal of Library & Information Technology

Sonia Bansal
soniapta@gmail.com

Decay of URL References cited in DESIDOC Journal of Library & Information Technology

ABSTRACT

The present study was conducted to examine the accessibility and corrosion of web references cited in articles of DESIDOC Journal of Library & Information Technology. A total of 1921 web references cited in 273 articles for five years from 2014-2018 were identified and checked to test their accessibility in terms of decay and persistence. Nearly 23.31% web references disappeared with passage of time. The results revealed that older the age of URL references, higher the percentage of inactive URL references. The .gov domain was found to be most constant with 95.99% accessibility rate. The average half-life for the missing web references during 2014-2018 was 6.55. The permanency of online resources is not assured because of the corrosion of web references but collective efforts are required to preclude the *corrosion of web references.*

KEYWORDS: Web references, Corrosion, URL persistence, URL decay, Missing URLs, Dead URLs

INTRODUCTION

The web is growing at gigantic speed and has become a popular source of information and research. It has become a mainstream tool for communicating research results, exchanging and sharing of information. Ever since the emergence of web, there has been substantial growth in volume of scholarly/scientific e-resources like, e-books, e-theses and dissertations, e-prints of research articles, e-journals etc. With increasing availability of e-resources on the web, the trend of citing web resources in scholarly publications has gained popularity. "The increasing number of citations to web sources in research papers is the evidence that academic and research communities are increasingly predisposed to use electronic resources for the advancement of scholarly communication[1]". The Internet has affected the citing behaviour of researchers and this has resulted in growth of URL citations.

The main problem of web citations is that they may disappear over time. There is significant relationship between age of URL citations and their accessibility. Lawrence et al. (2001)[2] through manual classification of invalid links identified several reasons for invalid URLs: restructuring of websites without maintaining old links; personal homepages tend to

disappear when researcher move and change in FTP servers to HTTP. Server shutdown; change of URL content or reconfiguration and errors in citing URLs are also the reasons for URL decay[3]. Nevertheless, many authors, though cognizant of all these limitations, cannot resist citing URLs in their articles[4]. The trend of citing URL citations in scholarly works is becoming common with the growth of web. But due to the transient and unstable nature of the Internet, the permanency and accessibility of URL citations is a cause for concern. This study makes an attempt to examine the growth, accessibility and decay of URL references over time in DESIDOC Journal of Library & Information Technology.

OBJECTIVES

To find out yearly distribution of web and print references;
To find out active and missing web references;
To ascertain error codes related with missing web references;
To find out domains related with missing URLs.

METHODS

The web references cited in the articles of DESIDOC Journal of Library & Information Technology for five years (2014-2018) were identified and checked to know the accessibility and decay of each cited URL. Editorials were not included in this study. A total of 1921 web references from 273 articles were identified. These references were transferred into an excel file and tabulated for further analysis.

LITERATURE REVIEW

Koehler[5] (1999) reported that 17.7% Web sites and 31.8% of Web pages failed to respond when retrieved after one year. Lawrence et al.[6] (2000) stated that the percentage of missing URLs referenced in academic articles within the CiteSeer (ResearchIndex) database increased from 23% in 2009 to 53% in 1994. Germain[7] (2000) investigated the accessibility of 64 URLs cited in 31 academic journal articles and found that almost half of the URLs became inaccessible and 2/3rd of the journal articles contained corroded citations. Dellavalle et al.[8] (2003) examined the frequency, format and activity of Internet references cited in three highly cited U.S. journals. The results of the study brought forward that Internet references were often inaccessible within months after publication in the highest-impact U.S. medical and scientific journals. The findings of a study by Tyler and McNeil[9] (2003) revealed that 7% of URLs (24) were inactive at an average of one year after publication and that 17% (73) were dead at two. TLD '.com' and '.edu' types were more likely to disappear and that '.gov' URLs

were more likely to continue. Spinellis[10] (2003) concluded that the percentage of inaccessible URLs referenced in *Computer and CACM* articles between 1995 and 1999 increased from 28% in 2000 to 41% in 2002. Casserly and Bird[11] (2003) examined 500 web citations from articles published in LIS journals in 1999 and 2000 and found that 56.4% internet citations were stable and 43.6% were missing. McCown et al.[12] (2005) found that 28% of the 4387 unique URL references from the D-Lib Magazine articles between 1995 and 2004 were missing on 2004-9-9, and that number increased to 30% on 2005-2-27. Evangelou et al.[13] (2005) checked the current accessibility of online supplementary scientific information published 2 and 5 years ago in the six top scientific journals and concluded that that even in the most reputed journals, some scientific information may eventually become inaccessible when it is supplied only online; personal and institutional web pages may be particularly vulnerable. An analysis of 1126 URLs cited in articles published in five leading journals in journalism and communication between 2000 and 2003 by Dimitrova and Bugeja[14] (2007) revealed that 39% URLs became inaccessible in 2004. The average half-life of the Internet citations was estimated to be 3.17 years. Aronsky et al.[15] (2007) ascertained that 11.9% of internet references became inaccessible within two days after publication. Goh and Ng[16] (2007) investigated the URL decay phenomenon in three leading information science journals and found that approximately 31% of all citations were not accessible. Ducut, Lio and Fontelo (2008) reported that 78% URLs were available at the time of access. Wren[17] (2008) conducted three follow-up studies in 2004, 2005 and 2007 after the 2003 study and observed no significant change in the rate of URL decay among any of the studies. Wagner et al.[18] (2009) examined the problem of decay of URLs in health care management journals. Finding of the study revealed that the percentage of inactive URLs ranged from 39.2% for articles published in 2004 to 61.1% for articles published in 2002. URLs with .edu domain were found to be more stable. An analysis of decay and accessibility of URLs cited in articles of six LIS journals conducted by Tajeddini et al.[19] (2011) revealed that 34% had error messages mostly related to "File error" type. A survey conducted by Saberi and Abedi[20] (2012) to ascertain the decay and accessibility of web references in five open-access Institute for Scientific Institution (ISI) Journals. The results of the study revealed that the rate of accessible URLs increased from 73% to 89% after using complementary pathways. The ".net" domain, with an accessibility of 96 per cent was most stable. Kumar and Raj[21] (2012) examined the accessibility of URLs of 350 articles published in Indian Association of Teachers in Library and Information Science (IATLIS) conference volume (2001-2008). The results of the study revealed that 45.61% URLs were inaccessible and majority of URLs displayed HTTP Error Code 404. Habibzadeh[4] (2013) revealed that the percentage of articles citing at least one URL

had increased from 24% in 2006 to 48.5% in 2013. Accessibility to URLs perished as the references got old. Mardani and Sangari[22] (2013) investigated the accessibility of 4253 web citations in six key Iranian LIS journals published from 2006 to 2010. The results of the study showed increase in percentage of web citations from 11% in 2006 to 30% in 2010. The half-life was computed to be 4 years for Iranian LIS journals. Tajeddini et al.[23] (2018) examined the currency, disappearance and half-life of URLs of web resources cited in Iranian researchers and found that .org and .com domains were more persistent and stable.

Distribution of web and print references in DESIDOC Journal of Library & Information Technology

Table 1 Yearly distribution of web and print references in DESIDOC Journal of Library & Information Technology

Year	No. of Articles	Total References	Average References per Article	Web References	%age	Average Web References per Article	Print References	%age	Average Print References per Article
2014	60	934	15.57	297	31.80	4.95	637	68.20	10.62
2015	49	836	17.06	301	36.01	6.14	535	63.99	10.92
2016	46	782	17.00	275	35.17	5.98	507	64.83	11.02
2017	58	1112	19.17	293	26.35	5.03	819	73.65	14.12
2018	60	1359	22.65	756	55.63	12.6	603	44.37	10.05
Total	273	5023	18.40	1922	38.26	7.04	3102	61.74	11.36

The yearly distribution of web and print references is depicted in Table 1. It is apparent from above that a total of 5023 references appeared in 273 articles published in DESIDOC Journal of Library & Information Technology during 2014-2018. Of the 5023 references, 3102 (61.74%) were print references and remaining 1922 (38.26%) were web references. The no. of articles published in the year 2014 and 2018 was same, despite that web references have increased from 31.80% in 2014 to 55.63% in 2018. Lawrence et al.[2] (2001) and Kumar and Raj[21] (2012) in their research also revealed a considerable growth in the number of URL citations. Kumar and Raj[21] (2012) even indicated approximately six fold increase in web

citations in the scholarly articles. The average number of web references per article was 7.04 across all the years. This clearly reveals inclination of researchers towards web resources.

Status of Active and Missing Web References

Table 2 Active and missing web references

Year	Total Web References	Active References	Percentage of Active References	Missing References	Percentage of Missing References
2014	297	183	61.62	114	38.38
2015	301	219	72.76	82	27.24
2016	275	209	76.00	66	24.00
2017	293	222	75.77	71	24.23
2018	756	641	84.79	115	15.21
Total	1922	1474	76.69	448	23.31

The status of active and missing web references is shown in above table. Of the 1922 URL references, 1474 (76.69%) were active and 448 (23.31%) were inactive/missing. The percentage of inactive URLs decreased from 38.38% in 2014 to 15.21% in 2018. It clearly reveals that there is significant relationship between age of web references and their accessibility. The URLs from articles published in recent years were more likely to remain active compared to the earlier ones. Koehler[5] (1999), Spinellis[10] (2003), Goh and Ng[16] (2007), Ducut, Lio and Fontelo[3] (2008) in their research also observed that URLs from recent years were more accessible than earlier ones.

HTTP errors associated with missing web references

Table 3 HTTP errors associated with missing web references

Year	HTTP 301	HTTP 400	HTTP 403	HTTP 404	HTTP 410	HTTP 500	HTTP 501	HTTP 503	Total
2014	0	0	3	74	1	36	0	0	114
2015	0	1	7	58	0	16	0	0	82
2016	0	1	2	44	0	18	1	0	66
2017	0	0	1	51	0	17	0	2	71
2018	1	2	5	63	0	44	0	0	115

Total	1	4	18	290	1	131	1	2	448
%age	0.22	0.89	4.02	64.74	0.22	29.24	0.22	0.45	100.00

Table 3 presents HTTP error codes related with decayed web references. Of the total 448 missing URLS, HTTP 404 error was encountered for majority (64.74%) of the missing URLs which is consistent with the findings of Spinellis[10] (2003), Goh and Ng[16] (2007) and Kumar and Raj[21] (2012). Goh and Ng[16] (2007) reported that the reasons for getting this error code are varied. The cause of this error code is 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 could be due to changes in the URL brought about by file/directory name changes, relocation of files or removal of files. About 29.24% missing URLs were due to HTTP 500 error code, followed by HTTP 403 (4.02%). The other five types of link accessibility errors encountered are negligible as they jointly accounted for only 2% of the missing URLs.

Web References by Domain Type

Table 4 Distribution of web references by domain type

Domains	**Total No. of URLs in 2014**	**Total No. of URLs in 2015**	**Total No. of URLs in 2016**	**Total No. of URLs in 2017**	**Total No. of URLs in 2018**	**Total No. of URLs**	**%age**	**No. of Active URLs**	**%age of Active URLs**	**No. of Missing URLs**	**%age of Missing URLs**
.edu/.ac	82	60	49	66	79	336	17.48	240	71.43	96	28.57
.com/.co	60	85	80	49	82	356	18.53	267	75.00	89	25.00
.net	9	7	11	16	13	56	2.91	40	71.43	16	28.57
.org	100	80	87	111	22	400	20.81	246	61.5	154	38.5
.gov	11	16	13	12	521	573	29.81	550	95.99	23	4.01
Others	35	53	35	39	39	201	10.46	131	65.17	70	34.83
Total	297	301	275	293	756	1922	100	1474	76.78	448	23.22

Five domains i.e. .edu/.ac, .com/.co, .net, .org and .gov were taken into consideration for this study. The domains not belonging to any of these categories were included in others. Of the 1922 URLs, about 50.62% were in .org and .gov domain. About 18.53% URLs were in .com/

.co domain, followed by 17.48% in .edu/.ac, 10.46% in others and 2.91% in .net domain. The highest number of active URLs 550 (95.99%) was in .gov domain, corroborating the findings of Casserly and Bird[11] (2003) and Wagner et al.[18] (2009). It was also found that .com/.co domain has 1/4th missing links.

Path depth and Decay of URLs

Table 5. Path depth and decay of URLs

Path Depth (PD)	Total no. of URLs	Percentage	No. of Active URLs	Percentage	No. of Missing URLs	Percentage
PD=0	168	8.74	152	90.48	16	9.52
PD=1	155	8.06	103	66.45	52	33.55
PD=2	928	48.28	762	82.11	166	17.89
PD=3	266	13.84	183	68.8	83	31.2
PD=4	186	9.68	126	67.74	60	32.26
PD=5	111	5.78	76	68.47	35	31.53
PD≥6	108	5.62	72	66.67	36	33.33
Total	1922	100	1474		448	

Table 5 shows path depths of URLs and corresponding accessible and inaccessible URLs. The URL depth could be related with link failure due to the increasing complexity as the length of a URL increases (Goh and Ng, 2007). To determine how URL path length influences URL decay rates, the path depth for each active and missing URL was calculated as per methods followed by Spinellis[10] (2003) and McCown et al.[12] (2005). The path depth was calculated by adding one to the depth for every directory or file after the domain name. For example, http://www.educause.edu/ has a path depth of 0, http://www.educause.edu/ecar has a path depth of 1, etc. It is clear from the table that 928 (48.28%) URLs have path depth of 2. The highest percentage (33.55%) of inaccessible URLs were associated with PD=1, followed by PD≥6 (33.33%) and PD=4 (32.26%). A substantial increase is observed in inaccessible URLs as the path depth moves from 1 to 2. McCown et al.[12] (2005) stated that a missing URL with path depth 0 would more likely be caused by the disappearance of business or an organization (either it has changed its name or gone out of business). A missing URL with path depth greater than 0 implies reorganization of internal structure or system change. To know the relationship between path depth and missing URLs correlation was calculated and it showed negative correlation between these two ($r=-0.124$, $df=6$).

Half-life of Web References

The half-life is the time required for half of all the web references in articles to decay. The half-life of URLs has been calculated using the procedure adopted by Koehler[24] (1999); Tyler and McNeil[9] (2003); Dimitrova and Bugeja[14] (2007); Mardani and Sangari[22] (2013).The half-life of URLs (t_h) has been calculated using following formula:

$$t_h=[t \ln(0.5)] / [\ln W(t) – \ln W(o)]$$

where t_h is the estimated number of years it takes for half of the web references to vanish, W(o) is the number of working web references at the time of publication, W(t) is the number of working web references at some later time t. Half-life has been calculated using this formula and the data is presented in Table 5. The average half-life for the missing web references was highest in the year 2015. The average half-life for the missing web references during 2014-2018 was 6.55 which mean that it will take about 7 years for half of the web references to decay.

Table 6. Half-life of Web References

Year	Time (t)	Total no. of URLs W(o)	No. of Active URLs W(t)	Half-life
2014	5	297	183	7.156974
2015	4	301	219	8.717776
2016	3	275	209	7.577122
2017	2	292	223	5.142384
2018	1	756	641	4.200589
All years		1921	1475	6.558969

Conclusion

This study revealed increase in web references from 31.80% in 2014 to 55.63% in 2018. About 23.22% of web references cited in the articles vanished. The .org domain registered highest number of decayed URLs and .gov had highest number of active URLs. HTTP 404 was the most common error message associated with missing URLs. It is not possible for authors and publishers to assure the longevity of online information resources cited in journal articles because of the unstable nature of URLs. Effective solutions will likely require a collective effort on the part of authors, researchers and journal editors[22]. The permanency and

longevity of online resources is not assured because of the corrosion of web references but improved citation practices[6], Digital Object Identifiers (DOI), WebCite[25], Lots of Copies Keeps Stuff Safe (LOCKSS)[26] can preclude the erosion of web references.

NOTES

1. Bullu Maharana, Kalpana Nayak, and N.K.Sahu, "Scholarly use of web resources in LIS research: a citation analysis," *Library Review* 55, no. 9 (2006): 598-60.

2. S. Lawrence, F. Coetzee, E. Glover, D. M. Pennock, G. Flake, and F. Nielsen, "Persistence of Web References in Scientific Research," *IEEE Computer* 34, no. 2 (2001): 26- 31.

3. Erick Ducut, Fang Liu, and Paul Fontelo, "An update on Uniform Resource Locator (URL) decay in MEDLINE abstracts and measures for its mitigation," *BMC Medical Informatics and Decision Making* 8, no. 1 (2008).

4. P. Habibzadeh, "Decay of References to Web sites in Articles Published in General Medical Journals: Mainstream vs Small Journals," Applied Clinical Informatics 4, no. 4 (2013): 455–464.

5. W. Koehler, "An analysis of web page and web site constancy and permanence," *Journal of the American Society for Information Science* 50, no. 2 (1999): 162-180.

6. Steve Lawrence, Frans Coetzee, Eric Glover, Gary Flake, David Pennock, Bob Krovetz, Finn Nielsen, Andries Kruger, and Lee Giles. "Persistence of information on the web: Analyzing citations contained in research articles," *In Proceedings of the Ninth international conference on Information and Knowledge Management*, (2000): 235-242.

7. C.A. Germain, "URLs: uniform resource locators or unreliable resource locators," *College and Research Libraries* 61, no. 4 (2000):359–65.

8. Robert P. Dellavalle, Eric J. Hester, Lauren F. Heilig, Amanda L. Drake, Jeff W. Kuntzman, Marla Graber, and Lisa M. Schilling, "Information Science: Going, Going, Gone: Lost Internet References," *Science* 302, no. 5646, (2003): 787–788.

9. David C Tyler, and Beth McNeil, “Librarians and link rot: a comparative analysis with some methodological considerations,” *Portal: Libraries and the Academy* 3, no. 4 (2003): 615–632.

10. D. Spinellis, “The decay and failures of web references,” *Communications of the ACM* 46, no. 1 (2003):71-77.
11. M. Casserly, and J.E. Bird, “Web citation availability: analysis and implications for citations,” *American Communication Journal* 9, no. 2, (2003), https://doi.org/10.5860/crl.64.4.300

12. Frank McCown, Sheffan Chan, Nelson, L. Michael, and Johan Bollen, The Availability and Persistence of Web References in D-Lib Magazine. *arXiv preprint cs/0511077* (2005).

13. E. Evangelou, T.A. Trikalinos, J.P.A. Ioannidis, “Unavailability of online supplementary scientific information from articles published in major journals,” *Faseb Journal* 19, (2005): 1943-1944.

14. Daniela V. Dimitrova, and Michael Bugeja, “The half-life of Internet references cited in communication journals,” *New Media and Society* 9, no. 9 (2007): 811–826.

15. D. Aronsky, S. Madani, R.J. Carnevale, S. Duda, and M.T. Feyder, “The Prevalence and Inaccessibility of Internet References in the Biomedical Literature at the Time of Publication,” *Journal of American Medical Information Association* 14, no. 2, (2007): 232-234.

16. Dion Hoe-Lian Goh, and Peng Kin Ng, “Link decay in leading information science journals,” Journal of American Society of Information Science and Technology 58, no. 1 (2007): 15-24.

17. Jonathan D. Wren, “URL decay in MEDLINE--a 4-year follow-up study,” *Bioinformatics* 24, no.11, (2008): 1381-1385.

18. Cassie Wagner, Meseret D. Gebremichael, Mary K. Taylor, and Michael J. Soltys, “Disappearing act: decay of uniform resource locators in health care management journals,” *Journal of Medical Library Association* 97, no. 2 (2009): 122-130.

19. O. Tajeddini, A. Azimi, A. Sadatmoosavi, and H.S. Moghaddam, “Death of web citations: a serious alarm for authors”, *Malaysian Journal of Library and Information Science* 16, no. 3 (2011): 17-29.

20. M.K. Saberi, and H. Abedi, “Accessibility and decay of web citations in five open access ISI journals,” *Internet Research* 22 no. 2 (2012): 234-247.

21. B.T.S Kumar, and K.R.P. Raj, “Availability and persistence of web citations in Indian Lis literature,” *The Electronic Library* 30, no. 1 (2012): 22.

22. Amir Mardani, and Sangari, Mahmood, “An analysis of the availability and persistence of web references in Iranian LIS journals,” *International Journal of Information Science and Management* 3, no. 1, (2013): 29–42.

23. O. Tajedini, A. Sadatmoosavi, A. Ghazizade, and A. Tajedini, “Investigation of the currency, disappearance and half-life of URLs of web resources cited in Iranian researchers: a comparative study,” *International Journal of Information Science and Management* 16, no. 1 (2018): 27-47.

24. Wallace Koehler, “An analysis of web page and web site, constancy and permanence,” *Journal of American Society of Information Science* 50, no.2 (1999): 162–180.

25. G. Eysenbach, “Going, going, still there: Using the WebCite service to permanently archive cited web pages,” *Journal of Medical Internet Research* 7, no. 5 (2005): 60- 68.

26. V. Reich, and D. Rosenthal, “Preserving today's scientific record for tomorrow,” *BMJ* 328, no. 7431 (2004): 61-62.