Summer 9-30-2019

Web Citations and Decay of URLs: A Case Study of Indian Journal of Agricultural Library and Information Services.

Seema Parmar
seemaparmar9@gmail.com

Rajive K. Pateria
rajivepateria@gmail.com

Web Citations and Decay of URLs: A Case Study of Indian Journal of Agricultural Library and Information Services.

Dr. Seema Parmar
Assistant Librarian,
Nehru Library, CCS Haryana Agricultural University, Hisar
E-mail: **seemaparm**ar9@gmail.com
and
Dr.Rajive Kumar Pateria
Professor, LIS
Nehru Library, CCS Haryana Agricultural University, Hisar
E-mail: rajivepateria@gmail.com

Abstract

The present study is conducted to identify the prevailing citation trends, durability of web citations or URLs, decay and half-life of URLs by analyzing articles' citations of Indian Journal of Agricultural Library and Information Services published during 2012 -2016. A total of 980 citations were reported in 94 articles out of which 33.16 percent were having web citations. Of web citations, 62.15 percent of URLs were accessible at the time of testing and the remaining 37.85 percent of URLs were not accessible. HTTP error message 404 "page not found'' was the irresistible error message appeared and represented 51.22 percent of all HTTP error messages. Average half-life of URLs was estimated to be 4.62 years. There was no association found between path depth and decay of URLs. Articles accessed from domain .org, .net and .co/.com were found more alive than other domains.

Keywords: Web citations, Citations, URLs, Missing URLs, URL corrosion, e-publishing, e-journals.

1. INTRODUCTION

In the present era of ICT, the society is gradually converting into web society. Every person is willing to complete all routine works with the help of Internet whether it is shopping for clothes, groceries, fruits or vegetables; bill paying; reservation or booking of tickets for travelling or movie, appointment with a doctor or any professional; selling new and old appliances; documents, and searching online catalogues of different digital libraries; and reading books online. In publishing world, most of the publishers have started e- publishing rather than print publications. They make their published content accessible on Internet either in subscription mode or free of cost under open access movement. Due to 24x7 hours availability of e-resources on web, now it is very easy for readers to explore and study their preferred documents. Scholar community in higher educational institutes make much use of e-books, e-journals, e-theses, e-lectures, e-repositories and other type of e-contents available on Internet for their academic and research purposes. "Most of the journals are available only on-line not in the

print format or dual format. E-journals are mostly in single format in these days" (Parmar, 2012). Thus, the user now prefers to read more e-research papers rather than print documents due to their multiple and advantageous features. Consequently, use of citing web resources or e-resources has been increased while writing references or bibliographies. It is mandatory to mention here that "when an author cites or refer web resources or online resources in his/her work for supporting the statements, these citations are called web citations". Every online document has its web link which is cited by a scholar in his to be published document. This web address or link of document is also called its URL. Searching of cited web document is now very easy as it is retrievable by just typing its web link or URL in web browser, but sometimes the user get frustrated when he or she do not find the cited web link due to its corrosion with passage of time. Different kind of error messages appeared on the screen in place of full text document when a user wants to access a web document by typing its URL in web browser. This way web citations have become a crucial issue for scholar society while referring e-documents in their research publications.

2. SCOPE AND NEED OF STUDY

The study has been conducted to know the ratio of print and web citations in recent years' publiction ratio of active and missing URLs in these publications. For the present study, a reputed journal in the field of Agricultural Librarianship named as 'Indian Journal of Agricultural Library and Information Services (IJALIS)' has been selected to identify the prevailing citation trends, durability of web citations over the years, association between path depth and decay of URL's, useful domain types, etc.

The present study may help any researcher to identify –i) the importance of providing complete web link or address without any typing error while giving web citations in his document; ii) the domains which have more durability; iii) importance of providing complete bibliographical information rather than only URL or partial bibliographical information.

3. OBJECTIVES

The study has been conducted with keeping in mind the following objectives:

- To identify the prevailing citation trends.

- To find out the total cited publications in IJALIS during 2012-16
- To assess year-wise distribution of print and web citations.
- To identify alive and dead URLs or web citations.
- To recognize http error codes/messages associated with inactive or missing URLs.
- To know the strong and weak domains associated with URLs.
- To identify the association between path depth and missing status of URLs.

4. METHODOLOGY

The study has been conducted to analyse the approachability and decay of URLs or web citations given in published articles of 'Indian Journal of Agricultural Library and Information Services (IJALIS)' during 2012 -2016. For collection of data, each issue of the journal was consulted to calculate the total number of articles and cited articles in them. All type of cited articles, their weblinks and other associated factors were categorised into certain codes and inserted into excel sheet. All URLs of given web citations were checked on Internet using different browsers and mentioned the access status in the specific URL's cell in excel sheet. The analysed data was then presented in tables for interpretations and findings.

5. REVIEW OF PREVIOUS STUDIES

Spinellis (2003) examined 4224 URL references from 2471 computer science articles published during 1995-99 and found 27 percent URLs not accessible while half of them became inaccessible in 4 years from the date they were published. Moosavi, Moghaddam and Tajeddini,Oranus (2012) studied four LIS journals viz. Journal of Documentation, Online Information Review, Aslib Proceedings, and Interblending and Document Supply . The journals' issues from 2005 to 2008 found original accessibility of studied online resources 64 percent, the majority of occurred errors was 'Error 404' (61%), 99 percent URLs of domain .net were accessible followed by .gov and .org with 97 percent accessibility each while .com/.co had the most decay with 93 percent unavailability. P. Habibzadeh (2013) studied five general medical journals from 2006 to 2013 and compared the trends in mainstream journals with small journals. He found number of citations to URLs increased in the journals with passage of time. URLs decayed as the references became older and half life ranged from 2.2 years to 5.3 years for

different journals. The ratio of citation to URLs in the studied mainstream journals, as well as the ratio of URLs accessible were significantly ($p<0.001$) higher than the small medical journals. Sife and Bernard (2013) examined theses and dissertations of Sokoine National Agricultural Library and revealed 58 percent web citations were inaccessible, error message '404 File Not Found' was highest to 92.7 percent, domain .com contained 28.2 percent missing URL and average half-life for the URLs was 2.5 years. Prithviraj and Kumar (2014) revealed in his study of Indian LIS conference proceedings published from 2001 to 2010 that 50.09 percent of URLs of web citations were not working and, average half-life of missing URLs was estimated to be around five years. Gul, Mahajan and Ali (2014) studied web magazine ARIADNE during 2010-2012 and found 'http 404-file not found' was most resisting error representing 52.68 percent of all http error codes and domain ".com/.co" was most stable and persistent domain with 95 percent accessibility, the big number of web citations were of "html" and "htmls" formats and "ppt" files were found to be most stable with 100 percent accessibility. Vinay, Divyashree and Sampath Kumar (2019) studies URL citations reported in 10 LIS journals during 2016 & 2017 and found 52.38 percent web citations active; domains .org and .com contain more number of missing URLs; majority of error were associated with HTTP 403 'File not found' (32.10%) and web citations having path depth 2 had largest number of missing URLs

6. DATA ANALYSIS

Analysis of the data collected through IJALIS has been presented by using tables and diagrams to visualize the data at a glance. The results drawn on the basis of analysis and interpretation has also been discussed at the very end of every table.

Table 6.1

Year-wise distribution of articles, citations / URLs

Year	TP	TC	TPC	%	APCPP	TWC	%	AWCPP
2012	16	131	85	64.89	5.31	46	35.11	2.88
2013	18	169	107	63.31	5.94	62	36.69	3.44
2014	22	226	165	73.01	7.50	61	26.99	2.77

2015	18	210	140	66.67	7.78	70	33.33	3.89
2016	20	244	158	64.75	7.90	86	35.25	4.30
Total	94	980	655	66.84	6.97	325	33.16	3.46

(TP= Total Paper; TC= Total Citations; TPC= Total Print Citations; APCPP= Average Print Citations Per Paper; TWC= Total Web Citations; AWCPP= Average Web Citations Per Paper)

Year-wise Distribution of Articles, Citations and URLs

Table 6.1 shows year wise distribution of articles and citations appeared during 2012-2016. Overall, a total of 980 citations were reported in 94 articles published in IJALIS during a period of five years. Out of total citations reported, 66.84 percent were print citations and 33.16 percent were web citations which indicate that that print citations were almost double than web citations. A constant increase in number of citation was also noticed over the year 2012-2016. In 2012 the total citation reported were 131 (13.37%) which were increased up to 244 (24.89%) in 2016. Average web citations per article in 2012 were 2.88 which were increased to high of 4.30 in 2016. It is noteworthy finding that last two years witnessed a high use of URLs or web citations compared to early years.

Table 6.2

Summary of active and missing web citations

Publication year	TWC	AWC	%	MWC	%
2012	46	32	69.57	14	30.43
2013	62	38	61.29	24	38.71
2014	61	35	57.38	26	42.62
2015	70	40	57.14	30	42.86
2016	86	57	66.28	29	33.72
Total	325	202	62.15	123	37.85

(AWC=Active Web Citations; MWC=Missing Web Citations)

Summary of Active and Missing Web Citations

Web citation is considered as missing web citation URL if it appears on web with a specific kind of error message like page not found, file not found, site can't be reached, account suspended, etc. Table 6.2 indicates towards the number of active and missing web citations. Out

of total web citations reported in IJALIS articles, 37.85 percent web citations were missing URLs while almost double of those (62.15%) were active web citations. Highest number of active web citations (69.57%) was reported in the articles published in 2012 followed by 2016 (66.28%). While during 2014 and 2015 majority of missing web citations were found 42.62 percent and 42.86 percent respectively which clearly indicates that the earlier studies during 2012 & 2013 had greater number of active web citations. During middle two years (2014 & 2015) the number of active web citations was decreased to 57 percent while it was increased upto 62.15 percent in 2016.

Table 6.3

Distribution of web citations by domain types

Domain	Total no. of URLs	%	No. of active URLs	%	No. of missing URLs	%
.ac	36	11.08	18	50.00	18	50.00
.co/.com	90	27.69	62	68.89	28	31.11
.edu	32	09.85	13	40.63	19	59.38
.gov	15	04.62	06	40.00	09	60.00
.net	15	04.62	11	73.33	04	26.67
.org	104	32.00	76	73.08	28	26.92
other	33	10.15	16	48.48	17	51.52
	325	100.00	202	62.15	123	37.85

(Other includes. in, .ir, .de, .info, .pt, .res. .es,.eu,.fr, .hr,.ie, .nl,.pt, .ru,.ca,. ernet,. nic etc.)

Distribution of Web Citations by Domain Types

In the present study mainly six domain types viz. .ac, .co/.com, .edu, .gov, .net, .org and other domains are taken into consideration however, the domains which did not fall under these six categories are included under “other categories”. Table 6.3 lists the domain wise total, active and missing web citations. It is clear from the table that highest numbers of web citations in IJALIS articles were associated with domain type .org (32%) followed by domain type .co/.com (27.69%). More than 10 percent of web citations were associated with .ac domain (11.08%) and other domain types (10.15%) while a very less number of web citations were associated with

domain .edu (9.85%), .gov and .net (4.62% each). Highest missing web citations were associated with four domains i.e. .gov (60%), .edu (59.38%), others (51.52%) and .ac (50%). Least missing citations were associated with domain type .net (26.67%) and .org (26.92%). It is very clear that more active links were associated with domain .org (73.33%) and .net (73.08%) followed by .co/.com (68.89%).

Table 6.4

Scheme wise yearly distribution of missing URLs/web citations

Year	HTTP					HTTPS				
	TWC	Active	%	Missing	%	Total	Active	%	Missing	%
2012	24	15	62.50	9	37.50	22	17	77.27	5	22.73
2013	25	13	52.00	12	48.00	37	25	67.57	12	32.43
2014	37	16	43.24	21	56.76	24	19	79.17	5	20.83
2015	32	13	40.63	19	59.38	38	27	71.05	11	28.95
2016	39	22	56.41	17	43.59	47	35	74.47	12	25.53
Total	157	79	50.32	78	49.68	168	123	73.21	45	26.79

Year wise Distribution of Missing URLs

Table 6.4 shows scheme wise yearly distribution of active and missing URLs. Of total web citations with http scheme, almost half of URLs (49.68 %) were found missing where as a small number (26.79 %) of total web citations with https scheme were found missing. In 2015 majority of URLs in http scheme were missing URLs (59.38%) followed by 2014 (56.76%) where as in https status, 2013 accounted for highest missing web citations (32.43%) followed by 2015 (28.95%).

Table 6.5

URLs scheme distribution of web citations

Status	http	https	Total
Accessible	79	123	202
Inaccessible	78	45	123
Total	157	168	325

Inaccessible (%)	**49.68**	**26.79**	**37.84**

URLs Scheme Distribution

Table 6.5 shows the distribution of URL status or scheme in the set of 325 web citations. A total of 157 URLs relied on http status out of which 49.68 percent were not accessible and 50.32 percent were accessible. A little large number of URLs (168) relied on https status out of which 73.21 percent were accessible and 26.79 percent were inaccessible.

Table 6.6

Description of web citations during 2012-16

Year	TWC	WC with only URL	%	WC with URL and partial bibliographic information	%	WC with URL and complete bibliographic information	%
2012	46	21	45.65	23	50.00	2	4.35
2013	62	23	37.10	30	48.39	9	14.52
2014	61	19	31.15	26	42.62	16	26.23
2015	70	30	42.86	19	27.14	21	30.00
2016	86	8	9.30	49	56.98	29	33.72
Total	325	101	31.08	147	45.23	77	23.69

Description of Web Citations during 2012-16

A citation or web citation is considered complete citation if it includes maximum searchable information about the web documents. In the present study, web citations vary from Only web address/URL to web address/URL with partial bibliographic information and web address/URL with complete bibliographic information. Table 6.6 shows types of web citations appeared in the articles published in IJALIS during 2012-2016. Overall, there were 31.08 percent web citations appeared as only URL Web citations, 45.23 percent were web citations with partial bibliographic information and 23.69 percent appeared as web citations with complete bibliographic information.

Table 6.7

Description of active and missing web citations

Types	Total	Active	Inactive	%
Only URL	101	71	30	29.70
Web citations with URL and partial bibliographic information	147	92	55	37.41
Web citations with URL and complete bibliographic information	77	39	38	49.35

Description of Active and Missing Web Citations

Table 6.7 indicate towards types of URLs having missing web citations. Out of total 325 web citations, 29.70 percent web citations contained Only URLs whereas 37.41 percent contained URLs with partial bibliographic information. There were almost half of web citations (49.35 %) contained URLs with complete bibliographic information.

Table 6.8

Path depth and decay of URLs

Path Depth(PD)	Total no. of URLs	%	No of active URLs	%	No. of missing URLs	%
PD=0	107	32.92	83	41.09	24	19.51
PD=1	43	13.23	21	10.40	22	17.89
PD=2	71	21.85	42	20.79	29	23.58
PD=3	41	12.62	18	8.91	23	18.70
PD=4	38	11.69	21	10.40	17	13.82
PD=5	15	4.62	11	5.45	4	3.25
PD≥6	10	3.08	6	2.97	4	3.25
Total	325	100.00	202	100.00	123	100.00

Path Depth and Decay of URLs

Table 6.8 indicate towards the association of path depth of web citations and their active and missing status. According to Goh and Ng (2007) "The URLs path depth could be associated with link failure due to increasing complexity as the length of a URL increases". In the present study each of 325 URLs path depth was categorised into different levels-0, 1, 2, 3...6. The purpose of this method is to find out the association between weblink path depth and decay of URLs. It is apparent from the table that more missing URLs (19.51%) were having 0 level path depths than those having 6 level path depth. Similarly, missing URLs having 1 level path depth were more (17.89%) than those having 5 level path depth. The significant finding drawn from this table is contrary to the statement of Goh and Ng (2007); Prithviraj and Sampath Kumar (2014) and other many authors conducted same studies for other journals, as the present table shows no association between the path depth and decay of URLs but the finding is supported by a recent study of Vinay, Divyashree and Sampath Kumar (2019).

Table 6.9

Half-life of URL

Publication Year	**Time (t)**	**Total No. of URLs (W_o)**	**No. of Active URLs (W_t)**	**Half Life(T_h)**
2012	5	46	32	9.549968
2013	4	62	38	5.663566
2014	3	61	35	3.743195
2015	2	70	40	2.477225
2016	1	86	57	1.685276
All years		325	202	4.623846

Half-Life of URL

"The half life is the time required for exactly half of the web citations in journals in journal to decay. The procedure adopted by Koehler (1999); Tyler and McNeil (2003); Dimitrova and Bugeja (2007); Mardani and Sangari (2013); and Bansal and Parmar (2019) is used to calculate half-life of URLs. The half-life of URLs (t_h) has been calculated using following formula:

$$t_h=[t \ln(0.5)] / [\ln W(t) – \ln W(0)]$$

where t_h is the estimated number of years it takes for 50 percent of the published internet citations to stop working, W (0) is the number of working online citations at the time of publication, W(t) is the number of working online citations at some later time t ". Using this formula the half-life has been calculated in MS-Excel and the data is presented in Table 6.9 which indicates that the average half life for the missing URLs was estimated to be 4.62 for all the five years under study which means half of the URLs may be vanished in approximately 5 years (t_h=4.62).

Table 6.10

Error messages associated with missing URLs

Year	IP address couldn't be found	Site can't be reached	Website/ URL not available	Page does not exist	HTTP error 403	HTTP error 404	HTTP error 405	Account suspended	Total missing URLs	% age of missing URLs
2012	2	0	2	0	1	9	0	0	14	11.38
2013	5	2	5	0	0	10	1	1	24	19.51
2014	0	6	7	0	0	11	1	1	26	21.14
2015	2	7	1	4	0	16	0	0	30	24.39
2016	1	6	2	2	0	17	1	0	29	23.58
Total	10	21	17	6	1	63	3	2	123	100
%	8.13	17.07	13.82	4.88	0.81	51.22	2.44	1.63	100	

Error Messages associated with Missing URLs

In the present study, there were eight type of errors reported for missing URLs. As per table 6.10, HTTP Error 404 (Page not found) was reported in majority of missing URLs (51.22%) distantly followed by HTTP Error "Site can't be searched"(17.07%), "Website not available"(13.82%) and "IP Address couldn't be found"(8.13%). HTTP Error "Account

Suspended" and HTTP Error 405 "Method not allowed" were found in least number of missing URLs i.e 1.63 percent and 2.44 percent respectively.

7. CONCLUSION

The study has shown the trends of providing citations in research journal Indian Journal of Agricultural Library and Information Services (IJALIS) during 2012 -2016. Print citations are still dominating web citations in the journal as only 33.16 percent citations are reported as web citations in the journal. Out of web citations reported in articles, 62.15 percent web citations were active while rests were inactive. The most common error message was HTTP error 404 File Not Found and highest numbers of web citations were associated with domain type .org (32%) while highest missing web citations were associated with domains .gov (60%) nearly followed by .edu (59.38.). More web citations relied on https status out of which 26.79 percent URLs were inaccessible. The study has not found any association between the path depth and decay of URLs and the average half life for the missing URLs is estimated to be 4.62 for all the five years which means half of the URLs may be vanished in approximately 5 years.

8. REFERENCES

- Dimitrova, D.V., & Bugeja, M. (2007). The half-life of Internet references cited in communication journals. *New Media and Society*, 9(5): 811–826. https://doi.org/10.1177/1461444807081226
- Gul, S., Mahajan, I., & Ali, A. (2014). The growth and decay of URLs citation: A case of an online library & information science journal. *Malaysian Journal of Library and Information Science*, *19*(3), 27–39.
- Habibzadeh, P. (2013). Decay of References to Web sites in Articles Published in General Medical Journals: Mainstream vs Small Journals. *Applied Clinical Informatics*, *04*(04), 455–464. https://doi.org/10.4338/aci-2013-07-ra-0055
- 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.

- Mardani, A., & Sangari, M. (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.
- Moghaddam, A.S., Saberi, M.K., & Esmaeel, S.M. 2010. Availability and half-life of web references cited in information research journals: A citation study. *International Journal of Information Science and Management*, 8(2): 57-75.
- Moosavi, Ali Sadat, Moghaddam, Alireza Isfandyari, Tajeddini,Oranus (2012). Accessibility of online resources cited in scholarly LIS journals A study of Emerald ISI-ranked journals. *Aslib Proceedings: New Information Perspectives*, 64 (2):178-192 .DOI 10.1108/00012531211215196 availble at https://www.researchgate.net/publication/227344542_Accessibility_of_online_resources_cited_in_scholarly_LIS_journals_A_study_of_Emerald_ISI-ranked_journals/figures?lo=1 (accessed 9 june 2019)
- Parmar, Seema (2012). Use of E-journals and CD-ROM Databases by Fraternity of CCSHAU, Hisar, India. *International Journal of Information Dissemination and Technology, 2*(1): 62-66.
- Prithviraj, K. R., & Sampath Kumar, B. T. (2014). Corrosion of URLs: Implications for electronic publishing. *IFLA Journal*, *40*(1), 35–47. https://doi.org/10.1177/0340035214526529
- Sife, A. S. , & Bernard, R. (2013). Persistence and decay of web citations used in theses and dissertations available at the Sokoine National Agricultural Library, Tanzania. *International Journal of Education & Development Using Information & Communication Technology*, *9*(2), 85–94. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=95745566&site=ehost-live&scope=site
- Spinellis, D. (2003). The decay and failures of web references. *Communications of the ACM*, 46(1), 71–77.
- Tyler, D., & McNeil, B. (2003). Librarians and link rot: a comparative analysis with some methodological considerations. *Portal: Libraries and the Academy*, 3(4): 615–632.
- Vinay, R. S., Divyashree G. N. & Sampath Kumar, B. T. (2019). Decayed URLs in LIS Journal articles : An Exploration Decayed URLs in LIS Journal articles : An Exploration.

In 9th KSCLA National Conference on Library in the Life of the User (pp. 540–544). Department of Studies and Research in Library and Information Science, Tumkur University. Retrieved from https://www.researchgate.net/profile/B_T_Sampath_Kumar2

.