Winter 2-15-2020

Longevity of URL citations Cited in LIS journal articles: A Webometric Study

Manjunatha G
Tumkur University,, manjudurga10@gmail.com

Dr. B. T. Sampath Kumar
Tumkur University,, sampathbt2001@gmail.com

Lakshmana H
Dev-In National School, lakshmanappu94@gmail.com

Longevity of URL citations Cited in LIS journal articles: A Webometric Study

Manjunatha G
Research Scholar
Department of Studies and Research in
Library and Information Science
Tumkur University, Tumakuru, Karnataka, India
Email id: manjudurga10@gmail.com

B. T. Sampath Kumar
Professor
Department of Studies and Research in
Library and Information Science
Tumkur University, Tumakuru, Karnataka, India
Email id: sampathbt2001@gmail.com

Lakshmana H
Librarian
Dev-In National School
Bangaluru, Karnataka,India
Email id: lakshmanappu94@gmail.com

Abstract

The study investigated the use of URLs as citations and their longevity, based on 966 articles published in selected LIS journals published during 2011-2015. It is found that there are 36,968 references in 966 articles with an average of 38.26 per article. Of the 36,968 references, there are 5,867 URL references. Each of the 5,867 URL citations were checked using W3C link checker to verify their accessibility. The study also found that 46.53% URLs (2,730 out of 5,867) remained active while the remaining 3,137 (53.46%) were found to be missing. The largest number of missing URLs (83.22%) cited in LIS articles are published in the year 2012. The HTTP error 404 - 'file not found' error message was the common error 40.36% In this regard the study suggested that the author(s) should check the accessibility of URL citation before it is used in the reference list of the article.

Keywords: URLs, Citations, Longevity and Webometrics

Introduction

The use of web citations has become common in conference papers, journal articles and other scholarly publications. This trend shows no signs of abating and the result is that web citations have become the norm in scholarly literature (Goh and Ng 2007). Authors being the producers as well as the consumers of information are highly influenced by the Web. It has long been recognized that references are an important part of valued academic papers and that they are significant for scientific research (Wu Z 2009).

The web is dynamic and not static like print media. Though the URLs of web sources may change or even disappear in course of time, but the URLs once cited as reference in print media cannot be changed and as a result, the later users of scientific publications in print version experience the problem of non-availability of the web links cited in them. The references that cannot be located seriously undermine the foundation of modern scientific discourse, scholarly communication and continuity of thought (Spinellis 2003).

The students and researchers look to literature citations as links between what is new and what is already known. But the foremost problem of web citation is their persistence, since citations disappear over time (Casserly and Bird 2008).

Keeping in view the disappearing nature of URLs citations, the present study made an attempt to investigate the use of URLs as citations in LIS journal articles published during the year 2011 to 2015. The study also tried to know the longevity of URL citations in LIS journals.

Review of Literature

There exists a plethora of literature regarding the use of URLs as citations in scholarly literature and also the decay of URLs. In this section, the literature has been reviewed based on the related to the use and decay of URL citation is

Goh and Ng (2007). Conducted a study on LIS journals during 1997-2003 that only 69 % of those URLs were permanent, while the remaining 31 percent had disappeared from the original web address. The 56% of error messages were “404” (page not found). The “.edu” with 36% active links was the most table domains.

Falagas et al. (2008). Study found 1,417 web citations from the 43,480 references from the New England Journal of Medicine and The Lancet. From these

two journal's articles, they could not access 16.8 % (238) web citations, Google helped them to identify references in other web sites, reducing the proportion of missing web references from 16.8 % (238) to 2.4 % (34) for the two journals. Thus 204 (85.71 %) inaccessible web citations were recovered out of 238 missing web references. In the same year, Russell and Kane (2008) found that 91 (18 %) of the 510 cited websites were inactive. But by using the Wayback Machine they could retrieve 52 (57 %) of the 91 missing websites, leaving 39 (43%) still unavailable.

A study by Bhat and Sampath Kumar (2008). on citation analysis of research articles from scholarly electronic journals in the field of LIS published during 2000-2006 shows that there is an increasing trend in web citation. They found that 81.49 % of articles published in nine selected electronic journals during 2000-2006 have web citations. Out of 25,730 references, 56.54 % were print journal references and 43.52 percent were web references.

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 % to 10 % annually. He also estimated that about 65 % to 72 % of web references could be accessed in newly published Chinese academic journals. Six years after publication, about 90 % could not be accessed.

Yang et al. (2010) also investigated that there is an observable impact of web based sources on scholarly journals in humanities and social sciences in China and the use of web citations has grown significantly in the social sciences from 2006 to 2007.

Moghaddam and Saberi (2010) analyzed 1,761 web citations from 10,242 references, as obtained from the bibliographies of 339 articles published in "Information Research" during 1995-2008. They did not find 471 (27%) web citations out of 1,761 web citations through direct search. After searching missing web citations via Google, accessible web citations increased from 1,290 (73%) to 1,518 (86%). Conversely, inaccessible web citations decreased from 471 (27%) to 243 (14%). Thus, of the 471 missing URLs, they could recover 228 (48.41 %) and243 (51.59 %) still remained missing.

Tajeddini et al. (2011) explored the availability and decay of URLs cited in articles of six Library and Information Sciences (LIS) journals published by Emerald,

Science Direct and Sage. Original accessibility of web citations was 66 % which improved to 95 % using Way back Machine and Google.

Nagaraja et al. (2011) collected a total of 1,133 articles published from 2005-2007 in PLoS Medicine. There are 1,133 articles contained 28,177 references, with 2,503 (8.9%) identified as URLs. Non-research articles accounted for a substantially higher percentage of URL references (17.4 percent) compared to research articles (4.2%).

Riahinia et al. (2011) analyzed 37,791 citations extracted from six LIS scholarly journals, of which 4,840 (12.8 %) 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 %) were persistent and 5 % returned errors and thus were not accessible. The most prevalent domain of citations was .html and the most favourable and persistent file format was .pdf.

Kumar and Prithviraj (2012) the study 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 % (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 %) to 2008 (30.27 %). Of the 307 missing web citations, the top-level domain .org had the highest percentage (30.29 %) of missing URLs, followed by the .edu domain (21.49 %).

Kumar and Kumar (2013) conducted a study and found that 39.84 % of URL citations were not accessible and remaining 60.15 % of URL citations were still accessible. The HTTP 404 error message-"page not found" was the overwhelming message encountered and represented 54.86 % of all HTTP error messages. However, 51.06 % URLs were recovered from HTTP 404 error message. The study also noticed that the half-life of URL citations was increased from 6.33 to 13.85 years after recovering missing URLs using Wayback Machine.

Gul, Mahajan and Ali (2014) study found that majority of errors were due to the missing content (http 404-file not found) representing 52.68 % of all http error codes followed by "http 500" (24.73%) and "http 403" (19.35%). The ".com/.co" domain was also found to be the most stable and persistent domain with 95 % accessibility. The greatest number of web resources cited in the articles were found to

be of “html” and “htmls” formats and “ppt” files were found to be most stable with 100% accessibility.

Kumar and Prithviraj (2015) The findings of study shows that the average number of web citations per article ranged from a low of 1.02 in 2001, to a high of 4.58 in 2010. There was a constant and continuous increase in the number of articles with web citations over the years during 2001–2010. The average number of web citations for every article is 2.60. The most widely cited top level domains were organizational (.org) and commercial (.com) with 30.91% and 22.08% respectively. The highest percentage of web citations belonged to HTML file formats (67.30%) followed by PDF file formats (11.39%).

Vinay Kumar and Sampath Kumar (2017) URLs as citations in Library and Information Science scholarly publications. A total of 8203 research articles published in 12 LIS journals during the years 2006-2015 were studied of the 288,452 citations 42,098 were URLs. The characteristics associated with the cited URLs were also analyzed. The study revealed that an average of 5.42 URL citations per article was cited among the single- authored papers. The study also indicated the fact that URLs associated with organizational and commercial domains were highly cited and the HTML and PDF file formats were dominatingly cited.

Vinay Kumar and Sampath Kumar (2017) study found that 38.12 % (417 out of 5197) URLs were found missing and remaining 61.88 %. The recovery of vanishing URLs through Internet Archive and Google increased the active URL’s rate from 61.88 % to 87.11 % and 73.58 % respectively. The study found that Internet Archive is a most effective tool to recover vanished URLs compared to Google search engine.

Research questions

The study has been conducted with the following research questions:

i. What proportion of URLs used as citations in the LIS journal articles?
ii. What percentage of missing URLs occurred in LIS journals articles?
iii. What is the correlation between the path depth and longevity of URL citations?

Scope of the study

This study explores the longevity of URL citations which have been cited in the scholarly journals in the field of Library and Information Science. For the present

study, the researcher has selected the following five LIS journals based on their reputation and popularity:

(i) The Library Hi-Tech

(ii) Collection Building

(iii) Information Technology and People

(iv) Journal of Documentation

(v) The Electronic Library

Methodology

The data for the present study has been drawn from selected five LIS journals published by Emerald publishing group during the year 2011-2015. References that are appeared as a list at the end of the articles under the bibliography or reference section are considered. The expanded bibliographies, endnotes and footnotes, e-mail links and annotations are not tested or counted in our dataset. After selecting all the references appended to the articles published in Emerald LIS journals, URLs are extracted for further analysis.

A total of 5,867 URLs are extracted from 36,968 references. The URLs so extracted have been tested for their availability using W3C Link Checker (https://validator.w3.org/checklink) and then URLs are categorized as active and missing URLs.

Q1: What proportion of URLs used as citations in the LIS Journal articles?

Table 1: Proportion of URLs used as citations in selected journals

Journal Name	Total no. of articles	Total no. of citations	Total no. of Print citations	Average citation per article	Total no. of URL citations	Average citations per article
Library Hi Tech	235	6390	4853	27.19	1537	6.54
Collection Building	100	1687	1376	16.87	311	3.11
Information Technology & People	112	7191	6614	64.20	577	5.15
Journal of Documentation	247	13085	11398	52.97	1687	6.82
The Electronic Library	272	8615	6860	31.73	1755	6.45

Total	**966**	**36,968**	**31,101**	**38.58**	**5,867**	**5.61**

The data presented in table-1 shows the distribution of articles, citations and URLs. It can be seen from the above table that there are 966 articles published during 2011-2015. The total number of citations in the reference is 36,968 and the average citations per article are 38.26. The percentage of articles with URL found highest in Information Technology and People (64.20%) followed by Journal of Documentation (52.97%) and The Electronic Library (31.73%). The table also reveals that Library Hi-Tech (27.19%) and Collection Building (16.87%) have very less percentage of URLs in the reference list.

Table 2: Proportion of URLs used as citations cross tabulated by year

Year	Total no. of articles	No of articles with URLs	% of articles with URLs	Total no of citations	Total no of Print citations	Total no of URLs	% of URLs	% of print citations
2011	187	157	83.95	6198	5062	1136	18.32	81.67
2012	180	152	84.44	6217	5174	1043	16.77	83.22
2013	177	139	78.53	6260	5264	996	15.91	84.08
2014	194	165	85.05	7944	6644	1300	16.36	83.63
2015	228	198	86.84	10349	8957	1392	13.45	86.54
Total	**966**	**811**	**83.95**	**36,968**	**31,101**	**5,867**	**16.16**	**83.82**

Table-2 clearly shows that the percentage of articles with URLs increased from 83.95% in the year 2011 to 86.84% in the year 2015. There are 966 articles published, out of which there are 811 (83.95%) article have URLs as citations in the articles. The percentage of URLs ranged from a low of 13.45% in the year 2015 to high of 18.32% in the year 2011.

The data clearly indicates that there an increasing trend in the use of URLs in LIS journals articles which are supported by the correlation analysis (r= .908 p=.033).

Q2. What percentage of missing URLs occurred in the selected journals?

Table 3: Percentage of missing URLs

Publication Year of the Journal	Total no. of URLs	No. of active URLs	% of active URLs	No. of missing URLs	%of missing URLs
Library Hi-Tech	1537	557	36.23	980	63.76
Collection Building	311	175	56.27	136	43.72
Information Technology & People	577	268	46.44	309	53.55
Journal of Documentation	1687	942	55.83	745	44.16
The Electronic Library	1755	788	44.90	967	55.09
Total	**5,867**	**2,730**	**47.93**	**3,137**	**52.05**

The researchers considered a web source as a missing URL if it returned with an HTTP error message. Table-4 depicts the percentage of missing URLs in articles published in the selected Emerald LIS Journals. The percentage of missing URLs cited in *Library Hi-Tech* Journal is 63.76% whereas, in *The Electronic Library* has 55.09 % of missing URLs followed by *Information Technology and People* 53.55 %, *Journal of Documentation* (44.16%).

Table 4: Year wise distribution of missing URLs

Year	Total no. of URLs	No. of active URLs	% of active URLs	No. of missing URLs	% of missing URLs
2011	1136	557	49.03	579	50.96
2012	1043	175	16.77	868	83.22
2013	996	268	26.90	728	73.09
2014	1300	942	72.46	358	27.53
2015	1392	788	56.60	604	43,39
Total	**5,867**	**2,730**	**44.35**	**3,137**	**58.70**

The data about the percentage of active and missing URLs presented in Table-4. It shows that 46.53% URLs (2,730 out of 5,867) remained active while the rest 3,137 (53.46%) are found to be missing. The largest number of missing URLs cited in LIS articles are published in 2012 (83.22%), followed by 2013 (73.09%). The percentages of missing URLs are found to be decreasing during the later years from 2015 (50.96%) to 2014 (43.39%).We performed the correlation analysis to know the

correlation between the year and missing URLs the correlation analysis indicates that there is a negative correlation between the year and the percentage of missing URLs but the correlation is not statistically significant (p=-.502, r=.389).

Q3: What is the correlation between the path depth and longevity of URL citations?

Table 5: Path depth associated with missing URLs

Path depth	No. of URLs	% of URLs	No. of missing URLs	% of missing URLs
PD=0	358	6.12	179	5.7
PD=1	646	11.04	380	12.11
PD=2	1462	25	751	23.94
PD=3	1448	24.76	726	23.14
PD=4	1007	17.22	539	17.18
PD=5	470	8.03	295	9.4
PD≥6	458	7.83	267	8.51
Total	**5,849**	**100**	**3,137**	**100**

In this study, each of the 5,849 URLs is verified for their path depth and are classified and grouped into the appropriate path depth levels from path depth of 0, 1, 2, 3, 4, 5 and any URL which has path depth level 6 and above are grouped into 6th path depth level. The analysis of data in the table-5 reveals that the path depth level 0 and 1 collectively accounted for 17.16% of the extracted URLs. On the Other hand, URL with path depth 2 accounted for 25% and another path depth level 3 accounted for 24.76%, and path depth level 4 accounted 17.22%. The path depth 5, 6 and above accounted for 8.03% and 7.83% respectively.

The path depths and corresponding missing URLs are also displayed in table-5. The URLs with path depth 2 (23.94%), 3 (23.14%) and 4(17.18%) are found to be missing as compared to the path depth 0 (5.7%) and 1(12.11%). In order to know the relationship between the path depth and the percentage of missing URLs, we performed the correlation analysis. The correlation analysis indicates that there is a positive correlation between the year and percentage of missing URLs but it is not statistically significant (p=.236, r=.652).

Table 6: HTTP Errors associated with missing URLs

HTTP Errors	Library Hi-Tech		Collection Building		Information Technology & People		Journal of Documentation		The Electronic Library		Total	
	No. of missing URLs	% of missing URLs	No. of missing URLs	% of missing URLs	No. of missing URLs	% of missing URLs	No. of missing URLs	% of missing URLs	No. of missing URLs	% of missing URLs	No. of missing URLs	% of missing URLs
HTTP-300	1	0.18	2	0.80	2	0.49	3	0.36	-	-	8	0.26
HTTP-301	1	0.18	1	0.40		-	8	0.95	2	0.19	12	0.38
HTTP-302	2	0.35	1	0.40	3	0.73	12	1.43	-	-	18	0.57
HTTP-303	-	-	15	6.00	-	-	-	-	-	-	15	0.48
HTTP-400	3	0.53	20	8.00	12	2.91	54	6.44	3	0.28	92	2.93
HTTP-401	1	0.18	12	4.80	2	0.49	21	2.50	1	0.09	37	1.18
HTTP-403	131	23.10	60	24.00	57	13.83	227	27.06	371	34.71	846	26.97
HTTP-404	283	49.91	80	32.00	220	53.40	280	33.37	403	37.70	1266	40.36
HTTP-406	2	0.35	4	1.60	-	-	2	0.24	7	0.65	15	0.48
HTTP-410	15	2.65	1	0.40	2	0.49	18	2.15	5	0.47	41	1.31
HTTP-500	116	20.46	50	20.00	112	27.18	210	25.03	273	25.54	761	24.26
HTTP-501	1	0.18	-	-	-	-	-	-	1	0.09	2	0.06
HTTP-502	5	0.88	3	1.20	-	-	4	0.48	3	0.28	15	0.48
HTTP-503	6	1.06	-	0.40	2	0.49	-	-	-	-	9	0.29
Total	**567**	**100**	**249**	**100**	**412**	**100**	**839**	**100**	**1,069**	**100**	**3,137**	**100**

The HTTP status codes of missing URLs are presented in table-6. The above table shows that the HTTP error 404 - 'file not found' error message accounted for (40.36%). And the second most common error message found was HTTP error 403- 'Forbidden' that accounted for 26.97% of all error messages. Another significant error message was HTTP 500-'Internet server error' that accounted for 24.26% of all the missing URLs.

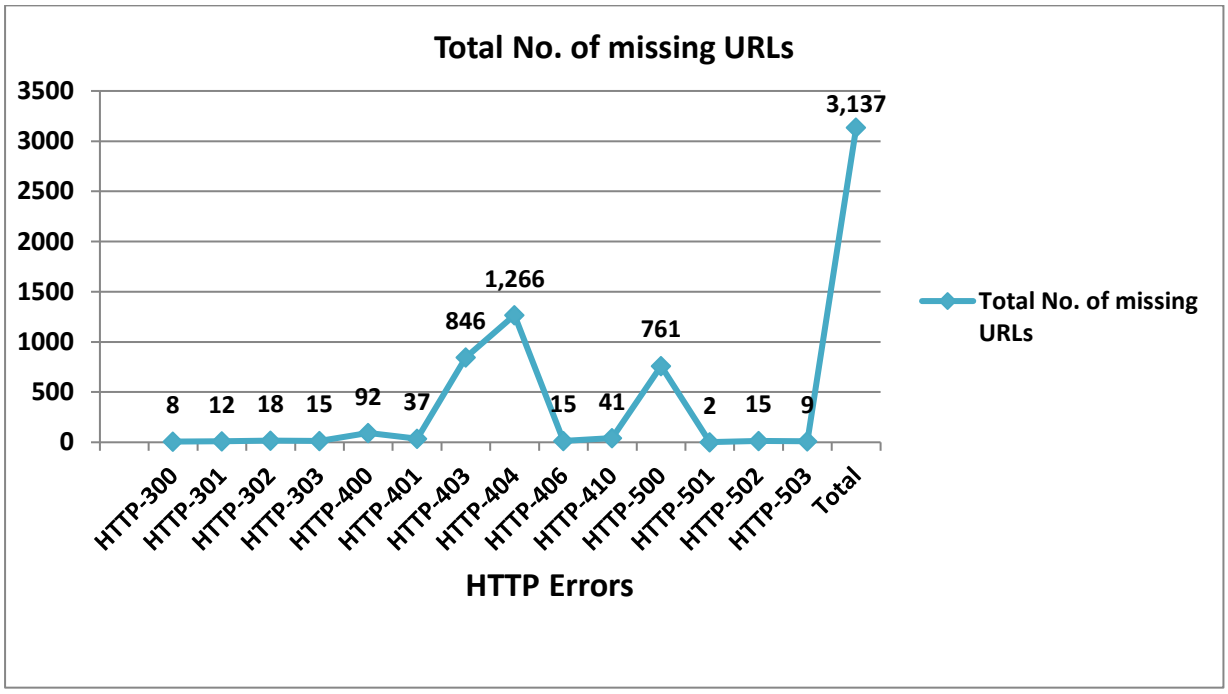

Figure-1: HTTP errors associated with missing URLs

Conclusion and Discussion

The current study on the Longevity of URLs used as citations in the emerald LIS journal articles shows that there is a constant growth in the use of URLs as citations in selected LIS journals. Though the overall percentage of the web references in the LIS journal articles slowly and steadily increased, yet they are not a match to print references. This study also found that 53.46 % of URLs are not accessible the authors of scholarly articles should determine whether there is a print version of the URL source they intended to cite. If available, parallel citations to both sources should be provided. The citations to web content should be complete and include full bibliographic information plus the date on which the site was accessed by the author, the date on which cited web page was created and last revised.

The editorial staff develops guidelines, for authors and references about the type of URLs sources permissible, based on consideration permanency / stability of the cited content and its scholarly importance. Editorial staff should work with authors to preserve and make available cited URLs sources. One possible strategy would be to support the development and maintenance of the Internet archive by encouraging the authors to upload the cited URLs sources.

References:

Bhat S V R and Kumar B S. 2008. Web citation behaviour in scholarly electronic journals in the field of library and information science. *Webology*, Vol. 5 no. 2: 15.

Casserly M F and Bird J E,. 2008. Web citation availability: analysis and implications for scholarship, *College & Research Libraries*, Vol. 64, no 4 :300-317.

Casserly M F and Bird J E,. 2008. Web citation availability: A follow-up study. *Library Resources & Technical Services*, Vol 52, no 1:42-53.

Falagas, M.E., Karveli, E.A,. and Tritsaroli, 2008. V I. The risk of using the Internet as reference resource: a comparative Study. *International journal of medical informatics*, Vol. 77, no 4:280-286.

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*, Vol 58, no 1: 15-24.

Prithviraj K. R. and B. T. Sampath Kumar,. 2015. Web Citation Trends in Indian LIS Journals: A Citation Analysis. *COLLNET Journal of Scientometrics and Information Management*, Vol 9, no 2: 295-310.

Mardani A., 2012. An investigation of the web citations in Iran's chemistry articles in SCI. *Library review*, Vol 61, no 1: 18-29.

Mariam Jalalifard, Yaghoub Norouzi and Alireza Isfandyari-Moghaddam, 2013. Analyzing web citations availability and half-life in medical journals: A case study in an Iranian university. *Aslib Proceedings*, Vol 65, no3: 242-261.

Moghaddam A I, Saberi M. K and Esmaeel S M, 2010. Availability and half-life of Web references cited in Information Research Journal: a citation study. *International Journal of Information*, Vol 8, no 2: 57-75.

Nagaraja A et al, 2011. Disappearing act: Persistence and attrition of uniform resource locators (URLs) in an open access medical journal. *Program*, Vol 45, no 1:98-106.

Riahinia N, Zandian F and Azimi A, 2011. Web citation persistence overtime: a retrospective study. *The Electronic Library*, Vol 29, no 5:609-620.

Russell E and Kane J,. 2008. The missing link: Assessing the reliability of internet citations in history. *Journals of Technology and Culture*, Vol 49, no 2: 420-429.

Sampath Kumar B T and Prithvi Raj K R,. 2012. Availability and persistence of web citations in Indian LIS literature. *The electronic library*, Vol 30, no 1,:19-32.

Sampath Kumar B T and Vinay Kumar D,. 2013. HTTP 404-page (not) found: Recovery of decayed URL citations. *Journal of Informetrics*, Vol 7, no 1: 145-157.

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, Vol* 56, no 7: 695-703.

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

Sumeer Gul, Iram Mahajan and Asifa Ali , 2014. The growth and decay of URLs citation: A case of an online Library & Information Science journal. *Malaysian Journal of Library & Information Science,* Vol 19, no 3:227-39.

Tajeddini, O Azimi A , Sadatmoosavi A and Sharif-Moghaddam, H,. 2011. Death of web citations: a serious alarm for authors. *Malaysian Journal of Library & Information Science*, Vol 16, no 3:17-29.

Vinay Kumar D and Sampath Kumar B T. 2017. Prevalence of URLs in Library and Information Science (LIS) Literature: A Citation Analysis. *COLLNET Journal of Scientometrics and Information Management.* Vol 11, no 2: 287-297.

Vinay Kumar D and Sampath Kumar B T,. 2017. Recovery of vanished URLs: Comparing the efficiency of Internet Archive and Google. *Malaysian Journal of Library & Information Science*, Vol 22, no 2:31-43.

Wu Z, 2009. An empirical study of the accessibility of web references in two Chinese academic journals. *Scientometrics*, Vol 78, no 3: 481-503.