Original Article

Indian Journal of Library and Information Science
Volume 14 Number 1, January-April 2020
DOI: http://dx.doi.org/10.21088/ijlis.0973.9548.14120.2

Crawling URLs for Availability and Lexical Features: An Analysis of References in Three Library and Information Science Journals

Niveditha B[1], Mallinath Kumbar[2]

How to cite this article:

Niveditha B, Mallinath Kumbar, Crawling URLs for availability and lexical features: An analysis of references in three Library and Information Science Journals. Indian J Lib Inf Sci 2020;14(1):19–31.

Authors Affiliation:

[1]UGC Junior Research Fellow, [2]Professor, Department of Library and Information Science, University of Mysore, Manasagangotri, Mysuru, Karnataka 570006, India.

Address for correspondence

Niveditha, UGC Junior Research Fellow, Department of Library and Information Science, University of Mysore, Manasagangotri, Mysuru, Karnataka 570006, India.

E-mail: niveditha.jb@gmail.com

Abstract

The present study examines the availability and recovery of Uniform Resource Locators (URLs) in scholarly Library and Information Science journals selected based on their high impact factor published between 2008 and 2017. A total of 4966 articles were downloaded and 208506 references were extracted. A PHP script was used to check 28108 URLs and extract the lexical features like path depth, character length and top-level domain. The research findings indicated that out of 28108 URL references, 15746 references contained URLs, DOIs were found in 11881 references and 481 references contained arXiv identifier/WOS article identifier. It was found that 25178 URLs were accessible and the remaining 2930 URL references were missing. The majority of errors were due to HTTP 404 error code (Not found error). The study also tried to recover the inaccessible URLs through Time Travel. Almost 60.92% of inaccessible URLs were archived in various web archives. The findings of the study will be helpful to authors, publishers and editorial staff to ensure that the URLs will be accessible in future.

Keywords: References; URL references; URLs; DOIs; PHP script; Time Travel

Introduction

The Internet has emerged as an increasingly popular information source for the academic and scientific community. The World Wide Web acts as an interface for providing easy access to information and has encouraged the researchers to make use of the benefits of the Internet. It is thus rapidly becoming an irreplaceable tool for doing scholarly work. Consequently, the use of Uniform Resource Locators (URLs) as references by authors of scholarly works has also increased. References direct the readers of a scholarly publication to resources that can help them to carry out research. References tend to be worthy only if they contain the information that the scholarly community can access. The availability of scholarly content is essential for carrying out valuable academic research. Due to the volatile nature of the Web, the URLs tend to disappear in a due course making it difficult for the researches to find reliable information. From this perspective, the present study has made an attempt to find the availability of URLs in three journals during the period 2008-2017. The URLs were examined in bulk using PHP script and the lexical features of display and destination URL were found. The paper also aims at recovering missing URLs cited in scholarly articles through Time Travel.

Review of Literature

Isfandyari-Moghaddam et al. (2010) stated that the Web has changed the citing behavior of researchers and this in turn has influenced the growth of web citations. Prithvi Raj and Sampath Kumar (2015) analyzed web citations cited in three

LIS scholarly journals published during 2001 to 2010. The web citations were used in other disciplines apart from Library and Information Science. For instance, Zhang (2007) conducted a comparison study among two journals in Communication and Media studies, and Mardani (2012) surveyed the available web citations in chemistry articles.

Wu (2008) had stated in his study that with the rise of web references, there is an increase in inaccessible web references also. The reasons for non-persistence of web citations as stated by Markwell and Brooks (2003) were broken links and restructuring the file hierarchy by some providers and Spenellis (2003) in his study stated that it was due to server problems and invalid URL host name or paths. Vinay Kumar and Sampath Kumar (2017) had also analyzed the various characteristics features of URLs like their file format, top-level domain, path depth and character length. This paper aims to extend the above said studies by examining the availability and lexical features of URLs through a PHP script.

Objectives of the Study

This study aims to investigate the availability and lexical features of URLs in Library and Information Science scholarly journals during the year 2008-2017. This study addresses the following objectives:

- To explore the proportion of URLs and DOIs used in scholarly LIS journals
- To know the percentage of inaccessible URLs.
- To examine the lexical features of URLs like top-level domain, path depth and character length.
- To differentiate the lexical features of display URLs and destination URLs
- To recover the inaccessible URLs through Time Travel.

Hypotheses

- URL references are most commonly cited in scholarly communications during 2008-2017.
- URL permanence will increase as their age decrease.
- The path depth and percentage of inaccessible URLs are positively correlated.

Materials and Methods

For the present study, data was drawn from three foremost Library and Information Science scholarly journals. The journals were selected based on their high impact factor as per Clarivate Analytics' 2018 "Journal Citation Report." The journals selected for the current study are Journal of Informetrics (JOI) with impact factor of 3.484, Journal of the Association for Information Science and Technology (JASIST) with impact factor of 2.835 and Scientometrics with impact factor of 2.173. All the research articles published during the 10-year period, that is, from 2008 to 2017 were taken up for the study. Editorial notes, book reviews, short communication were excluded. The references that were adjoined at the end of each article were considered for the study. A total of 208506 references were selected from 4966 articles published in the three journals. The references that contained web links and DOIs were extracted as the study deals with the availability of the URL references. The DOIs and arXiv were first resolved to URLs using the syntax https://doi.org/. Similarly, arXiv was resolved to URLs using the syntax https://arxiv.org/. A total of 28108 URLs were extracted for checking their availability. A PHP code was developed to test bulk URLs. The code uses CURL library, a standard PHP extension that checks for the availability of URLs and also documents the error code associated with inaccessible URLs. Apart from checking the URLs, the code gives the destination URL that is got after multiple redirects as well as the lexical features of URLs like their length, top-level domain, and path depth. The study used Time Travel (http://timetravel.mementoweb.org/) to find whether the URLs were archived or not. The Time travel recovers the inaccessible URLs that are archived in Internet Archive, Library of Congress Web Archive, Archive-it, Perma-cc, etc. The URLs that were not archived were considered as missing URL references.

Result and Discussion

Year-wise distribution of URL references

A total of 4966 articles published in three LIS scholarly journals during the period 2008-2017 were examined. The articles contained a total of 2,08,506 references with 13.48 per cent (28,108) of references citing a web source. Table 1 summarizes the citation results for the 4966 research articles. The number of references and URL references cited in journal articles are positively correlated and the relations is

statistically significant (r = .953, p = .000). This was performed using Pearson's Correlation analysis. The average number of URL references per article has been increased substantially from 3.07 in the year 2008 to 9.18 in the year 2017. The percentage of URL reference by year is varied from a low of 8.43 in the year 2010 to a high of 19.23 in the year 2016. This shows that the volume of URLs references in the research journals is not consistent during the 10-year period. The statistical relation shows that even though there is a negative correlation between the age and percentage of URL references (r = -0.923) the relation is statistically significant (p = 0.000). This show that the percentage of URL references cited in articles has been continuously increased from 2008 to 2017.

Journal-wise distribution of URL references

Table 2 reflects that a total of 4966 articles were published in the three journals during the year 2008-2017. More number of articles were published in Scientometrics (2575), followed by JASIST (1744) and JOI (647). The average citation per article was high in JASIST (50.15) and low in Scientometrics (37.33). The percentage of URL references also varied among journals. Highest percentage of URL references were noticed in Scientometrics (14.83%)

Table 1: Year-wise distribution of articles, references and URL references.

Year	Total number of articles	Total number of references	Average reference per article	Total number of URL references	Average URL reference per article	Percentage of URL reference
2008	315	11140	35.37	966	3.07	8.67
2009	411	14978	36.44	1445	3.52	9.65
2010	444	16836	37.92	1419	3.20	8.43
2011	433	16596	38.33	1623	3.75	9.78
2012	474	18578	39.19	2184	4.61	11.76
2013	507	21094	41.61	2488	4.91	11.79
2014	581	24159	41.58	3067	5.28	12.70
2015	582	27092	46.55	3800	6.53	14.03
2016	614	28920	47.10	5560	9.06	19.23
2017	605	29113	48.12	5556	9.18	19.08
Total	4966	208506	41.99	28108	5.66	13.48

Table 2: Journal-wise distribution of articles, references and URL references.

Journal	Total number of articles	Total number of references	Average reference per article	Total number of URL references	Average URL reference per article	Percentage of URL reference
JOI	647	24901	38.49	3546	5.48	14.24
JASIST	1744	87468	50.15	10301	5.91	11.78
Scientometrics	2575	96137	37.33	14261	5.54	14.83
Total	4966	208506	41.99	28108	5.66	13.48

Table 3: Year-wise distribution of URLs and DOIs.

Year	URL		DOI		Others		Total URL references
	Number	Percentage	Number	Percentage	Number	Percentage	
2008	938	97.10	12	1.24	16	1.66	966
2009	1346	93.15	80	5.54	19	1.31	1445
2010	1257	88.58	117	8.25	45	3.17	1419
2011	1142	70.36	459	28.28	22	1.36	1623
2012	1474	67.49	687	31.46	23	1.05	2184
2013	1565	62.90	889	35.73	34	1.37	2488
2014	1766	57.58	1259	41.05	42	1.37	3067
2015	1734	45.63	1972	51.89	94	2.47	3800
2016	2597	46.71	2887	51.92	76	1.37	5560
2017	1927	34.68	3519	63.34	110	1.98	5556
Total	15746	56.02	11881	42.27	481	1.71	28108

and lowest the percentage of URL references were found in JASIST (11.78%).

Distribution of URL and DOI

As the permanence of URL references is of major concern to researchers, DOIs instead of URLs were used to prevent the decay of URL references. Table 3 shows the distribution of URLs and DOIs in the three scholarly journals. It was found that out of the total 28,108 URL references, 15746 were URL links, 11881 were DOIs and 481 were arXiv identifier/ WOS article identifier.

Year-wise distribution of accessible, inaccessible and recovered URLs

The URLs were tested for their availability and this is depicted in table 4. The result of the accessibility check by year indicated that of the 28108 URLs, 89.58% were accessible while the remaining 10.42% encountered accessibility error. The percentage of inaccessible URLs varied from a low of 5.11 in the year 2017 to a high of 27.85 in the year 2008. In order to know the correlation between the age and inaccessible URLs, Pearson's Correlation analysis was performed. It was found that there is positive correlation between the age and inaccessible URLs and the correlation was statistically significant ($r = 0.970$, $p = 0.000$). The table also depicts the percentage of recovered URLs by year through Time Travel. A total 60.92% of URLs were archived in various web archives. The percentage of recovered URLs varied from a low of 50.56 in the year 2008 to a high of 71.07 in the year 2013. The correlation analysis indicates that the percentage of recovered URLs and the age are negative correlated ($r = -0.350$, $p = 0.321$) and the correlation is not statistically significant.

Journal-wise distribution of accessible, inaccessible and recovered URLs

The summary of accessible and inaccessible URLs cited in journal articles is presented in the Table 5. 10.42% of URLs were inaccessible in the three journals. Further 13.32% of URLs were inaccessible in JASIST, followed by 10.80% in JOI and 8.24% in Scientometrics.

Distribution of HTTP error codes associated with inaccessible and recovered URLs

The various error codes that are encountered for inaccessible URLs are presented in the Table 6. The HTTP 404 error message "page not found" represented 7461% of all HTTP error message and it is followed by HTTP 403 "forbidden error" (13.45%), HTTP 500 "internal server error" (5.19%) and HTTP 400 (3.62%). The 2930 inaccessible URLs showing

Table 4: Year wise distribution of accessible, inaccessible and recovered URLs.

Year	Total URLs	Accessible URLs	Percentage	Inaccessible URLs	Percentage	Recovered URLs	Percentage
2008	966	697	72.15	269	27.85	136	50.56
2009	1445	1090	75.43	355	24.57	215	60.56
2010	1419	1116	78.65	303	21.35	177	58.42
2011	1623	1394	85.89	229	14.11	145	63.32
2012	2184	1847	84.57	337	15.43	207	61.42
2013	2488	2208	88.75	280	11.25	199	71.07
2014	3067	2798	91.23	269	8.77	177	65.80
2015	3800	3485	91.71	315	8.29	185	58.73
2016	5560	5271	94.80	289	5.20	172	59.52
2017	5556	5272	94.89	284	5.11	172	60.56
Total	28108	25178	89.58	2930	10.42	1785	60.92

Table 5: Journal wise distribution of accessible, inaccessible and recovered URLs.

Journal	Total URLs	Accessible URLs	Percentage	Inaccessible URLs	Percentage	Recovered URLs	Percentage
JOI	3546	3163	89.20	383	10.80	108	28.20
Scientometrics	14261	13086	91.76	1175	8.24	713	60.68
JASIST	10301	8929	86.68	1372	13.32	964	70.26
Total	28108	25178	89.58	2930	10.42	1785	60.92

Table 6: Distribution of HTTP error codes

Error Codes	Inaccessible URLs	Percentage	**Recovered URLs**	**Percentage**
400	106	3.62	71	67.0
401	2	0.07	1	50.0
403	394	13.45	238	60.4
404	2186	74.61	1293	59.1
406	5	0.17	3	60.0
408	2	0.07	0	0.0
409	1	0.03	1	100.0
410	16	0.55	12	75.0
412	2	0.07	1	50.0
416	5	0.17	5	100.0
418	3	0.10	3	100.0
429	4	0.14	4	100.0
463	2	0.07	0	0.0
479	1	0.03	1	100.0
500	152	5.19	108	71.1
502	6	0.20	5	83.3
503	41	1.40	37	90.2
521	1	0.03	1	100.0
530	1	0.03	1	100.0
Total	2930	100.00	1785	60.9

Table 7: Distribution of archived URLs in Time Travel

Year	Inaccessible URLs	Internet Archive	LOC	Archive it	perma.cc	archive.is	arquivo.pt	Stanford web archive	Icelandic web archive	UK web archive	Web citation memento	Bibliotheca Alexandrina	Canadian Archive Memento
2008	269	115	18	11	1	16	21	0	2	4	26	7	0
2009	355	193	20	15	2	31	28	4	2	2	22	9	1
2010	303	163	24	9	0	23	35	1	4	0	28	14	0
2011	229	128	19	12	2	28	26	4	6	0	23	12	1
2012	337	171	18	10	3	19	52	2	4	0	24	6	0
2013	280	178	17	13	5	32	66	0	5	2	25	5	0
2014	269	156	15	6	1	20	34	2	4	0	24	4	1
2015	315	145	18	13	3	26	68	5	0	2	21	9	0
2016	289	138	18	13	4	24	81	3	3	0	13	4	0
2017	284	152	22	5	6	14	24	3	3	0	19	9	0
Total	2930	1539	189	107	27	233	435	24	33	10	225	79	3

various HTTP errors were entered in the search box of Time Travel. Nearly half of the inaccessible URLs (60.9%) were archived in various web archives were recovered from Time Travel. A total of 1785 URLs could be retrieved successfully. The percentage of recovered URLs with respect to various HTTP errors is shown in Table 6. It was interesting to note that 59.1% URLs were recovered from HTTP 404 error message, 60.4% URLs were recovered from HTTP 403 error message, and 71.1% of them were recovered from HTTP 500 error message and 67% were recovered from HTTP 400 error message.

File extension associated with inaccessible and recovered URLs

The data as illustrated in Table 8 indicates that the greatest numbers of cited URLs are .html files. Out of 28108 URLs, 23070 are .html files, followed by 3900 .pho files, and 242 are .asp files

Table 8: File extension associated with inaccessible and recovered URLs.

File Extension	Total URLs	Accessible URLs	Percentage	Inaccessible URLs	Percentage	Recovered URLs	Percentage
.asp	242	189	78.10	53	21.90	26	49.06
.cfm	117	91	77.78	26	22.22	15	57.69
.cgi	53	36	67.92	17	32.08	10	58.82
.html	23070	21399	92.76	1671	7.24	1077	64.45
.jsp	65	51	78.46	14	21.54	8	57.14
.pdf	92	60	65.22	32	34.78	16	50.00
.php	3900	2880	73.85	1020	26.15	580	56.86
Others	569	472	82.95	97	17.05	53	54.64
Total	28108	25178	89.58	2930	10.42	1785	60.92

Table 9: Path depth of display and destination URL

Path Depth	Display URL	Percentage	Destination URL	Percentage
PD = 0	757	2.69	293	1.04
PD = 1	1951	6.94	2892	10.29
PD = 2	16283	57.93	7922	28.18
PD = 3	4058	14.44	5946	21.15
PD = 4	2526	8.99	5617	19.98
PD = 5	1304	4.64	3422	12.17
PD = 6	672	2.39	1182	4.21
PD = 7	325	1.16	601	2.14
PD>7	232	0.83	233	0.83
Total	28108	100.00	28108	100.00

(2.6%). File extension having the highest percent of inaccessible URLs was the .pdf (34.78%), followed by .cgi (32.08%). Low level of loss was associated with the .html (7.24%) and .jsp (21.54%). Table 8 also indicates the percentage of recovered URLs with respect to their file extension. 64.45% of .html files, 58.82% of .cgi files, 57.69% of .cfm files and 57.14% of .jsp files were recovered from Time Travel.

Path depth of display and destination URL

Table 9 summarizes the characteristics of display and destination URL. The web address which is displayed to the user regardless of the article's physical location is the Display URL. The URL after multiple redirects goes to the landing page of the article or where the article resides, which is under the control of the publisher is called the destination URL. Out of 28108 URLs, display URLs with path depth 2 (57.93%) were frequently cited, followed by URLs with path depth of 3 (14.44%) and path depth 4 (8.99%). Unlike the display URL, only 28.18% destination URLs had a path depth of 2. There was an increase in destination URLs having path depth of 3 (21.15%) and path depth 4 (19.98%).

Path depth associated with inaccessible and recovered URLs

Table 10 shows that out of 28108, URLs with path depth 2 (16283) were most frequently cited, followed by URLs with path depth 3 (4058) and path depth 4 (2526). The highest percentage of inaccessible URLs (19.68%) had path depth of 1, followed by URLs with path depth 3 (19%) and path depth 6 (18.75%). The Table also indicates the percentage of recovered URLs from the Time Travel. It indicates that URLs (76.19%) with path depth 0 were recovered the most, followed by URLs with path depth 1 (63.80%) and path depth 7 (62.26%). In order to know the relationship between the path depth of the URLs and the percentage of inaccessible URLs, Pearson's Correlation analysis was performed. It is found that the path depth and the percentage of inaccessible URLs are positively correlated ($r = 0.055$, $p = 0.888$), but the relation is not statistically significant. In case of percentage of recovered URLs and the path depth are negatively correlated ($r = -0.746$, $p = 0.020$), and the relation is statistically significant.

Character length of display and destination URL

Table 11 shows the URL length and it can be found that a total of 11003 display URLs had length 41-50, 6422 URLs had length of 31-40, and 3108 URLs had a length of 51-60. When the landing page of URL is reached, a total of 6232 URLs had character length of 61-70, followed by 5924 URLs with length 51-60 and 4085 URLs with character length 41-50.

Table 10: Path depth associated with inaccessible and recovered URLs.

Path Depth	Total URLs	Accessible URLs	Percentage	Inaccessible URLs	Percentage	Recovered URLs	Percentage
PD = 0	757	652	86.13	105	13.87	80	76.19
PD = 1	1951	1567	80.32	384	19.68	245	63.80
PD = 2	16283	15480	95.07	803	4.93	504	62.76
PD = 3	4058	3287	81.00	771	19.00	457	59.27
PD = 4	2526	2075	82.15	451	17.85	263	58.31
PD = 5	1304	1094	83.90	210	16.10	113	53.81
PD = 6	672	546	81.25	126	18.75	78	61.90
PD = 7	325	272	83.69	53	16.31	33	62.26
PD>7	232	205	88.36	27	11.64	12	44.44
Total	28108	25178	89.58	2930	10.42	1785	60.92

Table 11: Character length of display and destination URL.

Character length	Display URL	Percentage	Destination URL	Percentage
<20	257	0.91	325	1.16
21-30	1546	5.50	1371	4.88
31-40	6422	22.85	2929	10.42
41-50	11003	39.15	4085	14.53
51-60	3108	11.06	5924	21.08
61-70	2040	7.26	6232	22.17
71-80	1358	4.83	2247	7.99
81-90	927	3.30	3048	10.84
91-100	520	1.85	712	2.53
>100	927	3.30	1235	4.39
Total	28108	100.00	28108	100.00

Table 12: Character length of associated with inaccessible and recovered URLs

Character length	Total URLs	Accessible URLs	Percentage	Inaccessible URLs	Percentage	Recovered URLs	Percentage
<20	257	227	88.33	30	11.67	19	63.33
21-30	1546	1349	87.26	197	12.74	148	75.13
31-40	6422	6064	94.43	358	5.57	257	71.79
41-50	11003	10491	95.35	512	4.65	319	62.30
51-60	3108	2476	79.67	632	20.33	362	57.28
61-70	2040	1555	76.23	485	23.77	291	60.00
71-80	1358	1104	81.30	254	18.70	147	57.87
81-90	927	768	82.85	159	17.15	82	51.57
91-100	520	405	77.88	115	22.12	58	50.43
>100	927	739	79.72	188	20.28	102	54.26
Total	28108	25178	89.58	2930	10.42	1785	60.92

Character length associated with inaccessible and recovered URLs

Table 12 shows the percentage of accessible, inaccessible and recovered URLs. URLs with length 61-70 were found to be inaccessible more (23.77%), followed by URLs with character length 91-100 (22.12%) and 51-60 character length (20.33%). In order to know the relation between percentage of vanished URLs and the character length, Pearson's correlation analysis was performed. It was found that there is positive correlation between percentage of inaccessible URLs and the character length and this relation is statistically significant ($r = 0.670$, $p = 0.033$). This clearly indicates that more number of characters in a URLs leads to its decay. The Table also illustrates the percentage of recovered URLs. The majority of recovered URLs were having 21-30 character length (75.13%), followed by 31–40 character length (71.79%) and

less than 20 characters (63.33%). The statistical relation shows that even though there is a negative correlation between the percentage of recovered URLs and the character length (r = -0.829), the relation is statistically significant (p = .003).

Top-level domain of display and destination URL

The top-level domain associated with the display and destination URL is summarized in table 13. It can be seen that a total of 17650 display URLs had the organizational top-level domain, followed by 3557 having the commercial top-level domain. On the other hand, a total of 13380 destination URLs have commercial top-level domain followed by 7658 organizational top-level domain.

Top level domain associated with inaccessible and recovered URLs

The analysis of total and inaccessible URLs by type of top-level domain is shown in table 14. Six main types of top-level domain have been considered in this study. They are .com, .edu, .gov, .info, .net, and.org. The top-level domain like .int, .mil and all the country top-level domains were considered in the "Others" category. The top-level domain having the greatest number of inaccessible URLs was the information top-level domain (.info) (49.46%) followed by educational (.edu) top-level domain (20.21%). A noteworthy finding is that proportionally low level of loss was associated with organizational (.org) top-level domain (4.22%). The Table also shows the top-level domains associated with the percentage of recovered URLs. The top-level domain having the greatest number of recovered URLs was the governmental top-level domain (.gov) (73.47%), followed by organizational top-level domain (67.61%). The low level of recovered URLs is associated with network (.net) top-level domain (44.86%) and educational top-level domain (56.27%).

Table 13 : Top-level domain of display and destination URL.

Top-level domain	Display URL	Percentage	Destination URL	Percentage
.com	3557	12.65	13380	47.60
.edu	1539	5.48	1627	5.79
.gov	813	2.89	820	2.92
.info	93	0.33	99	0.35
.net	719	2.56	680	2.42
.org	17650	62.79	7658	27.24
Others	3737	13.30	3844	13.68
Total	28108	100.00	28108	100.00

Table 14 : Top level domain of associated with inaccessible and recovered URLs

Top-level domain	Total URLs	Accessible URLs	Percentage	Inaccessible URLs	Percentage	Recovered URLs	Percentage
.com	3557	3001	84.37	556	15.63	322	57.91
.edu	1539	1228	79.79	311	20.21	175	56.27
.gov	813	715	87.95	98	12.05	72	73.47
.info	93	47	50.54	46	49.46	30	65.22
.net	719	612	85.12	107	14.88	48	44.86
.org	17650	16906	95.78	744	4.22	503	67.61
Others	3737	2669	71.42	1068	28.58	635	59.46
Total	28108	25178	89.58	2930	10.42	1785	60.92

Testing of Hypotheses

Table 15 illustrates the formulated hypotheses, statistical test applied to verify the hypotheses and the results. It can be seen from the table that only one hypothesis was not supported by the study results.

Conclusion

The Internet has substantially changed the way manner in which scholarly literature is accessed. It has paved the way for the academic and research community to use new form of citation practice like URL references in their scholarly work. The present

Table 15: Testing of Hypotheses

S.No	Hypotheses	Statistical test	p value	Result
H1	URL references are most commonly cited in scholarly communications during 2008-2017	Pearson's correlation	0.000	Supported
H2	URL permanence will increase as their age decrease.	Pearson's correlation	0.000	Supported
H3	The path depth and percentage of inaccessible URLs are positively correlated.	Pearson's correlation	0.888	Not supported

study confirms the use of URLs in the references cited in three journals during the year 2008-2017. But, the disappearance of URLs have impeded to access information in the Web. The URLs become unworthy if they disappear, move to a new location or change their content. It is apparent from the present study that the use of Digital Object Identifier (DOI) has reduced URL disappearance. To prevail over the problem of inaccessibility of URLs some suggestions are needed to be implemented. The URL references when used in references should be systematically checked by authors. It is obligatory for the authors to update or remove the inaccessible URL references. The editors and publishers should check the availability of the scholarly works in references before publication. Archiving of URLs should also be carried out by the authors as well as publishers. The authors, publishers and editorial team should make sure that the cited resources in the scholarly work be available without hindrance to the researchers.

Appendix

PHP script to crawl URLs for availability and extracting lexical features

```
function parseUrl($url) {
        $r  = "^(?:(?P<scheme>\w+)://)?";        //scheme/host/subdomain/domain/extension
        //$r .= "(?:(?P<login>\w+):(?P<pass>\w+)@)?";
        $r .= "(?P<host>(?:(?P<subdomain>[\w\.]+)\.)?" . "(?P<domain>\w+\.(?P<extension>\w+)))";
        //$r .= "(?::(?P<port>\d+))?";
        $r .= "(?P<path>[\w/]*/(?P<file>\w+(?:\.\w+)?)?)?";
        $r .= "(?:\?(?P<arg>[\w=&]+))?";
        $r .= "(?:#(?P<anchor>\w+))?";
        $r = "!$r!";                                // Delimiters
        preg_match ( $r, $url, $out );
        return $out;      }
function isvalid($url){
                $response = "";
                $handle = curl_init($url);
      curl_setopt($handle, CURLOPT_RETURNTRANSFER, true);
      curl_setopt($handle, CURLOPT_SSL_VERIFYPEER, false);
      curl_setopt($handle, CURLOPT_HEADER, true);
      curl_setopt($handle, CURLOPT_NOBODY, true);
      curl_setopt($handle, CURLOPT_USERAGENT, true);
      $headers = explode(" ",curl_exec($handle));
                $http_code = curl_getinfo($handle, CURLINFO_HTTP_CODE);
      curl_close($handle);
```

```
        if(!empty($headers[0])){

                $response = $headers[0]." ".$headers[1];

                if(preg_match("/([A-Z]+.[0-9].[0-9] 2[0-9]+)/", $response)){
                        $flag[0]=1;
                        $flag[1]= $response;
                        $flag[2]= "Valid";
                        $flag[3]= $headers[1];
                        $flag[4]= "Found";
                        return $flag;
                }elseif(preg_match("/([A-Z]+.[0-9].[0-9] 3[0-9]+)/", $response)){
                        $flag[0]=2;
                        $flag[1]= $response;
                        $flag[2]= "Valid";
                        $flag[3]= $headers[1];
                        $flag[4]= "Found";
                        return $flag;
                }else{
                        $flag[0]=0;
                        $flag[1]= $response;
                        $flag[2]= "Not-Found";
                        $flag[3]= $headers[1];
                        $flag[4]= "Not Found";
                        return $flag;
                }
        }else{
                        $flag[0]=0;
                        $flag[1]= "HTTP/1.1 404";
                        $flag[2]= "Not-Found";
                        $flag[3]= "404";
                        $flag[4]= "Not Found";
                        return $flag;
                }
    }

/**
 * get_destination_url()
 * Gets the address that the URL ultimately leads to.
 * Returns $url itself if it isn't a redirect.
 * @param string $url
```

```
 * @return string
 */
        function get_destination_url($url){
                $ch = curl_init();
                curl_setopt($ch, CURLOPT_URL, $url);
                curl_setopt($ch, CURLOPT_HEADER, true);
                curl_setopt($ch, CURLOPT_FOLLOWLOCATION, true);
                curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
                curl_setopt($ch, CURLOPT_FAILONERROR, true);
                curl_setopt($ch, CURLOPT_NOBODY, 1);
        // Read the headers as provided by cURL.
                $headers = curl_exec($ch);        //print_r($headers);die;
          $url = curl_getinfo($ch, CURLINFO_EFFECTIVE_URL);
                return $url;
        }
=>Main Function
while($row = mysqli_fetch_row($result1)) {

                $query1="";
                $url=trim($row[1]);

                $res = isvalid($url);

                if($res[0]==1){
                        $url_info = "";
                        $url_info = parseUrl($url);
                        $checkdate = date("F j, Y, g:i a");
                        $query1 = "UPDATE ".$tablename." SET Src_status='".$res[1]."' ,
Src_message='".$res[2]."',Src_Error_type='".$res[3]."',
                        Src_Found_Notfound='".$res[4]."', destination_url=\"".$url."\",
Des_status='".$res[1]."', Des_Found_Notfound='".$res[4]."',
                        Des_message='".$res[2]."',Des_Error_type='".$res[3]."',
Des_HTTP_DOI='".$url_info[scheme]."',
                        Des_host = '".$url_info[host]."', Des_File_format
='".get_filetype("$url")."',Des_Sub_Domain= '".$url_info[subdomain]."',
                        Des_Domain='".$url_info[domain]."',
Des_TLD='".$url_info[extension]."',Des_PathDepth='".url_depth("$url")."',
                        Des_Characterlength='".strlen("$url")."',
Availability='Available',checkdate='".$checkdate."' where id=".$row[0]."";

                }elseif($res[0]==2){
```

```
                $url_info = "";
                $availability="Available";
                $destination_url="";
                $checkdate = date("F j, Y, g:i a");
                $destination_url = get_destination_url($url);
                $url_info = parseUrl($destination_url);

                $destination_res = isvalid($destination_url);
                if($destination_res[4]!="Found") $availability="Not Available";
                //echo $availability;
                //echo "<br>"; print_r($destination_res); echo "<br>";die;

                $query1 = "UPDATE ".$tablename." SET Src_status='".$res[1]."' ,
  Src_message='".$res[2]."' ,Src_Error_type='".$res[3]."' ,
                    Src_Found_Notfound='".$res[4]."'  ,  destination_url=\"".$destination_url."\",
Des_status='".$destination_res[1]."',
                    Des_Found_Notfound='".$destination_res[4]."',  Des_message='".$destination_
res[2]."',Des_Error_type='".$destination_res[3]."',
  Des_HTTP_DOI='".$url_info[scheme]."', Des_host = '".$url_info[host]."', Des_File_format ='".get_
filetype("$destination_url")."', Des_Sub_Domain= '".$url_info[subdomain]."',
                    Des_Domain='".$url_info[domain]."',
  Des_TLD='".$url_info[extension]."',Des_PathDepth='".url_depth("$destination_url")."',
                    Des_Characterlength='".strlen("$destination_url")."',
  Availability='".$availability."', checkdate='".$checkdate."' where id=".$row[0]."";
            }else{
                $url_info = "";
                $url_info = parseUrl($url);
                $checkdate = date("F j, Y, g:i a");
                $query1 = "UPDATE ".$tablename." SET Src_status='".$res[1]."' ,
  Src_message='".$res[2]."',Src_Error_type='".$res[3]."',
                Src_Found_Notfound='".$res[4]."', destination_url=\"".$url."\",
  Des_status='".$res[1]."', Des_Found_Notfound='".$res[4]."',
                Des_message='".$res[2]."',Des_Error_type='".$res[3]."',
  Des_HTTP_DOI='".$url_info[scheme]."',
                Des_host = '".$url_info[host]."', Des_File_format
  ='".get_filetype("$url")."',Des_Sub_Domain= '".$url_info[subdomain]."',
                Des_Domain='".$url_info[domain]."',
  Des_TLD='".$url_info[extension]."',Des_PathDepth='".url_depth("$url")."',
                Des_Characterlength='".strlen("$url")."',
  Availability='Not Available',checkdate='".$checkdate."' where id=".$row[0]."";
        }
```

```
        echo $query1."<br>";
        $result = mysqli_query($conn,$query1);
//die;
        }
```

References

1. Dimitrova DV, & Bugeja M. The half–life of internet references cited in communication journals. New Media & Society 2007;9(5):811–826.
2. Goh DH, & Ng PK. Link decay in leading information science journals. Journal of the American Society for Information Science and Technology 2007;58(1):15–24.
3. Isfandyari–Moghaddam A, Saberi MK, & Mohammad Esmaeel S. Availability and Half–life of Web References Cited in Information Research Journal: A Citation Study. International Journal of Information Science and Management 2010;8(2):57–75.
4. Maharana B, Nayak K, & Sahu NK. Scholarly use of web resources in LIS research: a citation analysis. Library Review 2006;55(9):598–607.
5. Mardani A. An investigation of the web citations in Iran's chemistry articles in SCI. Library Review 2012;61(1):18–29.
6. Prithvi Raj KR, & Sampath Kumar BT. Web Citation Trends in Indian LIS Journals: A Citation Analysis. COLLNET Journal of Scientometrics and Information Management 2015;9(2):295–310.
7. Saberi MK, & Abedi H. Accessibility and decay of web citations in five open access ISI journals. Internet Research 2012;22(2):234–247.
8. Sadat–Moosavi A, Isfandyari–Moghaddam A, & Tajeddini O. Accessibility of online resources cited in scholarly LIS journals: A study of Emerald ISI–ranked journals. Aslib Proceedings 2012;64(2):178–192.
9. Sampath Kumar BT, & Manoj Kumar KS. Persistence and half–life of URL citations cited in LIS open access journals. Aslib Proceedings 2012;64(4):405–422.
10. Sampath Kumar BT, & Prithvi Raj KR. Availability and persistence of web citations in Indian LIS literature. The Electronic Library 2012;30(1):19–32.
11. Sampath Kumar BT. and Vinay Kumar D. "HTTP 404–page (not) found: recovery of decayed URL citations". Journal of Informetrics 2013;7(1):145–157.
12. Sampath Kumar BT, Vinay Kumar D, & Prithvi Raj KR. Wayback machine: reincarnation to vanished online citations. Program 2015;49(2):205–223.
13. Vinay Kumar D. & Sampath Kumar BT. Finding the unfound: Recovery of missing URLs through Internet Archive. Annals of Library and Information Studies. 2017;64(3):165–171.
14. Vinay Kumar D, Sampath Kumar BT, & Parameshwarappa DR. URLs Link Rot: Implications for Electronic Publishing. World Digital Libraries – An International Journal. 2015;8(1):59–66.
15. Vinay Kumar D. & Sushmitha D. Recovery of Missing URLs cited in Annals of Library and Information Studies: a study of Time Travel. Annals of Library and Information Studies. 2017;66(1):24–32.
16. Wu Z. An empirical study of the accessibility of web references in two Chinese academic journals. Scientometrics. 2008;78(3):481–503.
17. Yang S, Qiu J, & Xiong Z. An empirical study on the utilization of web academic resources in humanities and social sciences based on web citations. Scientometrics 2010;84(1):1–19.
18. Zhang, Y. The Effect of Open Access on Citation Impact: A Comparison Study Based on Web Citation Analysis. Libri. 2007; 56(3):145–156.
19. Zhao, D. & Logan, E. Citation analysis using scientific publications on the Web as data source: A case study in the XML research area. Scientometrics. 2002;54(3):449–472.