ACCESSIBILITY AND CHARACTERISTICS OF WEB CITATIONS IN JOURNAL OF COMPUTER-MEDIATED COMMUNICATION DURING 2008-2017

B. Niveditha
Mallinath Kumbar

B. Niveditha
Department of Library and Information Science, University of Mysore, Manasagangotri, Mysuru – 570006
E-mail: niveditha.jb@gmail.com
(*Corresponding Author)*

and

Mallinath Kumbar
Professor
Department of Library and Information Science, University of Mysore, Manasagangotri, Mysuru – 570006

The present paper explores the accessibility and characteristics of web citations in the Journal of Computer Mediated Communication (JCMC) during 2008-17. A total of 337 articles from JCMC were downloaded and 17946 references extracted. Out of 4548 web citations, DOIs were found in 2360 references, 2178 references contained URLs and 10 references contained arXiv, WOS article identifier, etc. It was found that 3861 web citations were accessible and the remaining 687 web citations were missing. The study also investigated the characteristic features of display and destination URLs, like the path depth, URL length, file format and top-level domain.

Keywords: References; Web citations; URLs; DOIs; Google; DOI checker

INTRODUCTION

The expanded use of the World Wide Web has led to effectively communication of academic information through computer technology and Internet. It has now become an essential mean for carrying out scientific research. Scholarly articles contain references that appear at the end of the publication. A reference is a formal mention of another work in a scientific publication[1]. The reference provides an opportunity to the academician to expand their knowledge in their areas of research. The development of the Internet and the World Wide Web during the past decade has had a profound impact on the references. Citations to books, journal articles, etc are being replaced by its electronic counterparts. The authors of research publication substitute traditional paper-based citations to books, journals, reports and notes with electronic alternatives. Moreover, with the vast quantity and easily accessible documentation available on the Web, many authors often cite Uniform Resource Locators (URLs) as part of the attribution process when it comes to acknowledging supporting material

in their publications[2-4]. As the web is extremely transient, most of its information becomes unavailable and is lost forever after a short period of time. The preservation of digital material is crucial to modern societies because web publications are extremely transient. To overcome the problem of decay, preserving the electronic resources has become inevitable to prevail over the accessibility of the resources. Web archives help in preserving the web resources permanently and enable long-term access to the URLs. Furthermore, the Digital Object Identifier is being used in research publications, allowing access to resources within an encapsulated environment and designed to work without human intervention [5]. The DOI system enables unique identification and persistence of the research publication. DOI are represented in a URL and transported by the HTTP protocol are constrained to follow standard IETF guidelines [6]. The DOIs thus resolve to a destination page that contains the research publication. This present study made an attempt to find the increasing trend of DOIs in "Journal of Computer-Mediated Communication" and examine their characteristics and accessibility.

LITERATURE REVIEW

The web has influenced the citing behavior of researchers and this in turn has influenced the growth of web citations[7]. The library and information science [LIS] authors refer to web resources as part of their increased research productivity and this has resulted in increase in the number of web citations in scholarly papers in LIS[8]. Many studies have investigated the use of URL citations in LIS scholarly journals[9-12]. In the same vein, web citations in conference proceedings of LIS were explored and it was found that web resources were becoming preferred source of information for LIS professionals[13-14].

The web citations were used in other disciplines apart from LIS. For instance, Lawrence made a comprehensive study on conference articles in computer science and related disciplines[15]. Zhang conducted a comparison study among two journals in communication and media studies [16], and Mardani surveyed the available web citations in chemistry articles [17]. This shows that web citations have become common in scholarly publications in all disciplines [18].

The rising issue of using web citations is that they tend to decay and disappear. The reasons for non-persistence of web citations were failure to maintain old links while restructuring web sites [19], broken links and restructuring the file hierarchy by some providers [20], server problems and invalid URL host name or paths [21]. However, the missing web citations could be recovered through web archives. Many studies used the Wayback machine to retrieve missing web citations [22-25]. It was noted that the accessibility rate increased after recovering the missing web citations.

The characteristics of the accessibility of web citations like the top level domain, file format, path depth and character length were examined in many studies[26-29]. This paper aims to extend the above mentioned studies by examining the trend of Digital Object Identifier and differentiating the characteristics of display and destination URL.

OBJECTIVES OF THE STUDY

This study aims to investigate the availability and characteristics of web citations in "Journal of Computer-Mediated Communication" during 2008-17, with the following objectives: (i) To study the proportion of web citations in the Journal of Computer-Mediated Communication; (ii) To determine the percentage of URLs and DOIs in the Journal of Computer-Mediated Communication; (iii) To find the percentage of inaccessible URLs and DOIs; (iv) To identify the file format, path depth, top-level domain, character length of web citations and (v) To recover the inaccessible web citations through Time Travel.

METHODOLOGY

For the present study, data was drawn from "Journal of Computer-Mediated Communication." The journal was selected based on its high-impact factor of 4.00 as per "Journal Citation Report, 2018" from Clarivate Analytics. All the research articles published during 2008-17 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 17946 references were selected from 337 articles published in "Journal of Computer-Mediated Communication."

The references that contained web links and DOIs were extracted from the article references. A total of 4548 web citations were extracted and categorized as URL and DOI. The URLs were checked for their accessibility in the web browser and the DOIs were checked in DOI checker (https://dx.doi.org/). The W3C link checker (http://validator.w3.org/checklink) was used to report the HTTP error message for inaccessible URLs.

To find the features of the web citations, the DOIs in the study are resolved to URLs using the syntax https://doi.org/. For example a DOI name 10.1010.1234/567 would be resolved from the address https://doi.org/10.1010.1234/567. After the resolution, features such as the URL length, top-level domain, file format and path depth are found for the display and the destination URLs.

The study used Time Travel (http://timetravel.mementoweb.org/) to find whether the web citations were archived or not. The Time Travel recovers the inaccessible web citations that are archived in Internet Archive, Library of Congress Web Archive, Archive-IT, Perma-CC, etc. The study also made an attempt to find whether the inaccessible URLs were available by searching through their title present in their respective reference. The web citations that were not archived or not found through title search were considered as missing.

DATA ANALYSIS AND RESULTS

Year-wise distribution of web citations

337 research articles published in "Journal of Computer-Mediated Communication" during 2008-17 are presented in Table 1. These articles contained 17946 references with 25.34% (4548) web citations. The percentage of web citations varied from a low of 2.88% during the year 2011 to a high of 18.60% in the year 2015.

Table 1: Distribution of articles, references and web citations in JCMC during 2008-17

Year	Total articles	Total references	Percentage	Total web citations	Percentage
2008	37	2300	12.82	374	8.22
2009	49	3095	17.25	503	11.06
2010	30	1587	8.84	244	5.36
2011	20	1186	6.61	131	2.88
2012	30	1400	7.80	441	9.70
2013	28	1413	7.87	261	5.74
2014	56	2776	15.47	741	16.29
2015	39	1930	10.75	846	18.60
2016	27	1252	6.98	464	10.20
2017	21	1007	5.61	543	11.94
Total	337	17946	100.00	4548	100.00

Distribution of URL, DOI and others

As suggested in some studies [30-31], use of DOIs in place of URLs has increased in the recent years to prevent the deterioration of web citations. The DOI is defined as a character string used to identify intellectual property in the digital environment [32]. Table 2 shows the distribution of URL, DOI and others in "Journal of Computer-Mediated Communication". It was found that out of the total 4548 web citations, 2178 were URL links, 2360 were DOIs and 10 were arXiv identifier and WOS article identifier.

Table 2: Year-wise distribution of URL and DOI in JCMC during 2008-17

Year	URL		DOI		Others		Total web citations
	Number	%	Number	%	Number	%	
2008	374	17.17	0	0.00	0	0	374
2009	502	23.05	1	0.04	0	0	503
2010	234	10.74	10	0.42	0	0	244
2011	108	4.96	23	0.97	0	0	131
2012	193	8.86	247	10.47	1	10	441
2013	125	5.74	130	5.51	6	60	261
2014	378	17.36	363	15.38	0	0	741
2015	153	7.02	692	29.32	1	10	846
2016	58	2.66	404	17.12	2	20	464
2017	53	2.43	490	20.76	0	0	543
Total	2178	100.0	2360	100.0	10	100	4548

Distribution of accessible and inaccessible web citations

Table 3 gives the distribution of accessible and inaccessible web citation by year is presented in Table 3. The URLs were checked with the web browser by clicking directly on the URL, DOIs were checked in the DOI checker "https://dx.doi.org" and arXiv identifier was checked in "https://arxiv.org/." The result of the accessibility check indicated that of the 4548 web citations, 84.89% were accessible, while the remaining 15.11% encountered accessibility error.

Table 3: Distribution of accessible and inaccessible web citations in JCMC during 2008-17

Year	Total web citations	Accessible web citations	Percentage	Inaccessible web citations	Percentage
2008	374	312	83.42	62	16.58
2009	503	262	52.09	241	47.91
2010	244	184	75.41	60	24.59
2011	131	95	72.52	36	27.48
2012	441	370	83.90	71	16.10
2013	261	222	85.06	39	14.94
2014	741	658	88.80	83	11.20
2015	846	798	94.33	48	5.67
2016	464	438	94.40	26	5.60
2017	543	522	96.13	21	3.87
Total	4548	3861	84.89	687	15.11

File format of web citations

The data illustrated in Table 4 indicates that the greatest number of cited web resources were HTML files. Out of the total 4548 web citations, 3858 were HTML files, followed by 435 PDF files, 120 ASP files, 79 PHP file and 21 CFM file.

Table 4: File format of web citations in JCMC during 2008-17

File format	Number	Percentage
HTML	3858	84.83
CFM	21	0.46
PDF	435	9.56
PHP	79	1.74
ASP	120	2.64
JSP	4	0.09
CGI	15	0.33
Others	16	0.35
Total	4548	100.0

Path depth of display and destination URL

Display URL is the URL displayed to the user. It is the web address or the DOI found at the end of the reference which is given by a central server regardless of the article's physical location. The destination URL is the URL, where the user is eventually taken after multiple redirects are involved. It is the article landing page or where the article resides. The article landing page is under the control of the publisher. In this study, an attempt has been made to distinguish the characteristics between the display and destination URL.

The distribution of display URL by path depth is shown in Table 5. Out of 4548 web citations, display URLs with path depth 2 (2769) were frequently cited, followed by 705 URLs with depth of 3. A total of 318 URLs had path depth 4, 256 URLs had path depth 5238 URLs with path depth 1 and 109 URLs with path depth 6. Unlike the display URL, almost 1094 destination URLs have a path depth of 3, followed by 1026 URLs having a path depth of 4898 URLs have a path depth of 2540 have a path depth of 6, 345 URLs have a path depth of 5.

Table 5: Path depth of display and destination URL in JCMC during 2008-17

Path Depth	Display URL	Percentage	Destination URL	Percentage
PD = 0	62	1.36	2	0.04
PD = 1	238	5.23	551	12.12
PD = 2	2769	60.88	898	19.74
PD = 3	705	15.50	1094	24.05
PD = 4	318	6.99	1026	22.56
PD = 5	256	5.63	345	7.59
PD = 6	109	2.40	540	11.87
PD = 7	54	1.19	57	1.25
PD > 7	37	0.81	35	0.77
Total	4548	100.0	4548	100.0

Character length of display and destination URL

The data on URL length are presented in Table 6. It was found that a total of 1890 display URLs had length 41-50, 1056 URLs had length of 31-40 and 580 URLs had a length of 51-60.

Table 6: Character length of display and destination URL in JCMC during 2008-17

Character length	Display URL	Percentage	Destination URL	Percentage
<20	22	0.48	83	1.82
21-30	115	2.53	115	2.53
31-40	1056	23.22	214	4.71
41-50	1890	41.56	615	13.52
51-60	580	12.75	1373	30.19
61-70	275	6.05	945	20.78
71-80	234	5.15	503	11.06
81-90	139	3.06	386	8.49
91-100	93	2.04	92	2.02
>100	144	3.17	222	4.88
Total	4548	100.00	4548	100.0

Top-level domain of display and destination URL

The top-level domain associated with the display and destination URL is summarized in Table 7. It can be seen that a total of 3021 display URLs had the organizational top-level domain and 642 having the commercial top-level domain. On the other hand, a total of 2802 destination URL have commercial top-level domain and 1096 organization domain.

Table 7: Top-level domain of display and destination URL in JCMC during 2008-17

.com	642	14.12	2802	61.61
.org	3021	66.42	1096	24.10
.edu	430	9.45	220	4.84
.info	7	0.15	8	0.18
.gov	64	1.41	63	1.39
.net	56	1.23	59	1.30
Others	328	7.21	300	6.60
Total	4548	100.0	4548	100.0

HTTP errors associated with missing web citations

The missing or inaccessible URLs were checked in W3C link checker (http://validator.w3.org/checklink) to report the HTTP error codes. Table 8 shows that the HTTP 404 error message that is "Page not Found" error was the error message that mostly occurred and represented 64.05% of all the HTTP error messages. It is followed by HTTP 500 (21.54%) and HTTP 403 (11.79%) error messages.

Table 8: HTTP errors associated within accessible web citations in JCMC during 2008-17

HTTP error	Total	Percentage
HTTP 300	2	0.29
HTTP 302	1	0.15
HTTP 308	1	0.15
HTTP 400	4	0.58
HTTP 403	81	11.79
HTTP 404	440	64.05
HTTP 406	2	0.29
HTTP 416	2	0.29
HTTP 500	148	21.54
HTTP 503	6	0.87
Total	687	100.0

Distribution of archived web citations

The study intended to recover the inaccessible web citations through Time Travel tool. The inaccessible web citations were entered in the search box of "Time Travel." Table 9 depicts that out of the 687 inaccessible web citations, 400 were archived and the remaining 247 have not been archived.

Distribution of archived web citations in Time Travel

Table 10 shows the distribution of archived web citations in Time Travel. The Internet Archive recovered the highest percentage of inaccessible web citations (59.39%), followed by Library of Congress web archive (9.90) and Web Citation Memento (9.02%).

Table 9: Distribution of archived web citations in JCMC during 2008-17

Year	Inaccessible web citations	Archived	Percentage	Not archived	Percentage
2008	62	41	66.13	21	33.87
2009	241	186	77.18	55	22.82
2010	60	38	63.33	22	36.67
2011	36	23	63.89	13	36.11
2012	71	46	64.79	25	35.21
2013	39	24	61.54	15	38.46
2014	83	52	62.65	31	37.35
2015	48	23	47.92	25	52.08
2016	26	6	23.08	20	76.92
2017	21	1	4.76	20	95.24
Total	687	400	58.22	247	35.95

Table 10: Distribution of archived web citations in Time Travel in JCMC during 2008-17

Year	Internet Archive	Library of congress Web Archive	Archive it	Perma.cc	Archive.is	Arquivo.pt	Stanford Web Archive	Icelandic Web Archive	UK Web Archive	Web Citation Memento	Bibliotheca Alexandrina Web Archive	Canadian Archive Memento
2008	25	1	9	0	6	2	1	1	7	3	0	0
2009	180	42	20	0	25	22	3	7	2	32	1	1
2010	37	4	0	0	0	2	1	0	0	4	1	0
2011	21	6	0	0	6	1	0	1	0	4	2	0
2012	45	9	2	0	2	5	0	0	1	5	1	0
2013	20	0	0	0	0	2	0	0	0	5	1	0
2014	50	5	0	0	4	5	0	1	2	5	2	0
2015	23	1	5	0	1	4	0	0	0	4	0	0
2016	6	0	4	4	3	1	1	1	1	0	0	0
2017	1	0	0	0	0	0	0	0	0	0	0	0
Total	408	68	40	4	47	44	6	11	13	62	8	1

Title-wise search for inaccessible URLs

An attempt was made in this study to search the inaccessible URLs by their title present in their respective cited reference. It is found from Table 11 that out of the 687 URLs, 468 URLs were found through title search. The percentage of URLs found through title search varied a low of 60.24% cited during the year 2014 to a high of 90.48% cited in the year 2017.

Table 11: Title-wise search for inaccessible URLs in JCMC during 2008-17

Year	Inaccessible web citations	Found	Percentage	Not found	Percentage
2008	62	41	66.13	21	33.87
2009	241	150	62.24	91	37.76
2010	60	44	73.33	16	26.67
2011	36	26	72.22	10	27.78
2012	71	55	77.46	16	22.54
2013	39	29	74.36	10	25.64
2014	83	50	60.24	33	39.76
2015	48	33	68.75	15	31.25
2016	26	21	80.77	5	19.23
2017	21	19	90.48	2	9.52
Total	687	468	68.12	219	31.88

SUMMARY AND CONCLUSION

The study concludes that Digital Object Identifier was used most frequently in the references cited in the Journal of Computer-Mediated Communication during the year 2008-17. The DOI was used as they permanently identify an article or a document in the digital environment. Further, it was found from the study that the changes in path depth and character length of a display and destination URL is because the DOI server translates the DOI-based link into its actual URL on the publication server or where the article resides. If the article is moved to a different location, the DOI-based URL should redirect to the new location. It is thus the responsibility of the publisher to ensure that the current information is registered for each DOI. The study also found error in the displayed DOI when searched by title. Thus, the publishers, editors, and authors should work together through systematic checking of the web citations before publication [33].

REFERENCES

1. Ding, Y., Liu, X., Guo, C. and Cronin, B. The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, 2013, 7(3), 583-92.

2. Germaine, C. A. URLs: Uniform Resource Locators or Unreliable Resource Locators?. *College & Research Libraries*, 2000, 61(4), 359-65.

3. Rumsey, M. Runaway train: Problems of permanence, accessibility and sustainability in the use of web sources in Law Review citations. *Law Library Journal*, 2002, 94(1), 27-39

4. Spinellis D. The decay and failures of web references. *Communications of the ACM*, 2003, 46(1), 71-77.

5. Scott-Wilson, E. Identifiers and interoperability. In. A. Gilchrist and B. Mahon, eds. Information architecture: Designing information environments for purpose, London; Facet Publishing, 2004, 161-73.

6. https://www.doi.org (accessed on 3 January, 2019).

7. Isfandyari-Moghaddam, A., Saberi, M. K., and 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.

8. Zhao, D. and 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), 449-72.

9. Maharana, B., Nayak, K. and Sahu, N. K. Scholarly use of web resources in LIS research: A citation analysis. *Library Review*, 2006, 55(9), 598–607.

10. Prithvi Raj, K. R. and Sampath Kumar, B. T. Web Citation Trends in Indian LIS Journals: A Citation Analysis. *COLLNET Journal of Scientometrics and Information Management*, 2015, 9(2), 295–310.

11. Vinay Kumar, D., Sampath Kumar, B. T. and Parameshwarappa, D. R. URLs Link Rot: Implications for electronic publishing. *World Digital Libraries - An International Journal*, 2015, 8(1), 59–66.

12. Vinay Kumar, D. and Sampath Kumar, B. T. Finding the unfound: Recovery of missing URLs through Internet Archive. *Annals of Library and Information Studies*, 2017, 64(3), 165-171.

13. Chikate, R. V. and Patil, S. K. Measuring impact of web sources in ILA conference proceedings: A citation analysis. *Library Herald*, 2009, 47(2), 142-154

14. Doraswamy, M. and Janakiramaiah, M. Measuring impact of web resources in conference proceedings: A citation analysis. In. 8th International CALIBER 2011: Towards Building a Knowledge Society: Library as Catalyst for Knowledge Discovery and Management,02-04 March 2011, Goa University, Goa. 2011. pp. 541-549. http://ir.inflibnet.ac.in:8080/ jspui/handle/1944/1645(accessed on 15 March, 2019).

15. Lawrence, S. Online or invisible. *Nature*, 2001, 411(6837), 521.

16. Zhang, Y. The effect of open access on citation impact: A comparison study based

on web citation analysis. *Libri*, 2007, 56(3), 145–156.

17. Mardani, A. An investigation of the web citations in Iran's chemistry articles in SCI. *Library Review*, 2012, 61(1), 18–29.
18. Sampath Kumar, B. T. and Prithvi Raj, K. R. Availability and persistence of web citations in Indian LIS literature. *The Electronic Library*, 2012, 30(1), 19–32.
19. Lawrence, S., Pennock, D. M., Flake, G. W., Krovetz, R., Coetzee, F. M., Glover, E. and Giles, L. C. Persistence of web references in scientific research. *Computing Practices*, 2001, 34(2), 26-31.
20. Markwell, J. and Brooks, D. W. "Link rot" limits the usefulness of web-based educational materials in biochemistry and molecular biology. *Biochemistry and Molecular Biology Education*, 2003, 31(1), 69–72.
21. Spinellis, D. The Decay and Failures of Web References. *Communications of the ACM*, 2003, 46(1), 71–77
22. Dimitrova, D. V., & Bugeja, M. The half-life of internet references cited in communication journals. *New Media & Society*, 2007, 9(5), 811–826.
23. Tajeddini, O., Azimi, A., Sadat-Moosavi, A. and Sharif-Moghaddam, H. Death of web citations: A serious alarm for authors. *Malaysian Journal of Library and Information Science*, 2011, 16(3), 17-29
24. Sampath Kumar, B.T. and Vinay Kumar, D. HTTP 404-page (not) found: recovery of decayed URL citations. *Journal of Informetrics*, 2013, 7(1), 145-157.
25. Sampath Kumar, B. T., Vinay Kumar, D. and Prithvi Raj, K. R. Wayback machine: Reincarnation to vanished online citations. *Program*, 2015, *49*(2), 205–223.
26. Sampath Kumar, B. T. and Manoj Kumar, K. S. Persistence and half life of URL citations cited in LIS open access journals. *Aslib Proceedings*, 2012, 64(4), 405–422.
27. Sadat Moosavi, A., Isfandyari Moghaddam, A. and 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.
28. Jalalifard, M., Norouzi, Y. and Isfandyari Moghaddam, A. Analyzing web citations availability and half life in medical journals: A case study in an Iranian university. *Aslib Proceedings*, 2013, 65(3), 242–261.
29. Saberi, M. K. and Abedi, H. Accessibility and decay of web citations in five open access ISI journals. *Internet Research*, 2012, 22(2), 234–247.
30. Yang, S., Qiu, J. and 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.
31. Sife, A. S. and Bernard, R. Persistence and decay of web citations used in theses and

dissertations available at the Sokoine National Agricultural Library, Tanzania. *International Journal of Education and Development using Information and Communication Technology,* 2013, 9(2), 85-94.

32. Information Standards Quarterly (ISQ). "National Information Standards Organization (NISO)." *Information Standards Quarterly (ISQ)*, 2004, 16(3), 16.
33. Dimitrova, D.V. and Bugeja, M. Raising the dead: recovery of decayed online citations. *American Communication Journal*, 2007, 9(2), 1-14.

❖❖❖