LIBER QUARTERLY 19 (2), December 2009 – ISSN: 1435-5205. P131–139
http://liber.library.uu.nl/
Igitur, Utrecht Publishing & Archiving Services

Missing Web References — A Case Study of Five Scholarly Journals

Mohammad Hanief Bhat

Senior Librarian, Islamia College of Science & Commerce,
Srinagar-190002 (J&K), India,
mhanief30@yahoo.co.in

Abstract

The present study attempts to ascertain the proportion of missing web references of 5–10-year-old research papers of the five leading open access (OA) journals in library and information science. The results suggest that the number of web citations has increased from 41.60% of all citations in 1998 to 53.32% in 2002. But a substantial quantity of web citations (32.09%) was found to be missing. The percentage of missing web citations goes on increasing with each passing year — ten-year-old publications having the highest number of missing citations, i.e., 39.96% and five-year-old publications having the lowest number of missing citations (25.89%). 0.92% of citations had moved to a new URL address and 74.14% of missing citations resulted in an HTTP 404 (page not found) error.

Key Words: Missing URL's; URL persistence; web citations

Introduction

Much scholarship, if not all, is based on previous work, and when new scholarly work is produced it is important that detailed and accurate information on sources consulted is provided. To facilitate referencing, scholarly works have been routinely collected and preserved in print by libraries and database producers (Veronin, 2002). With the advent of the internet large numbers of scholarly journals and other sources of information have become available online. This has resulted in an increasing usage of web references in research articles. The proportion of citations of electronic resources has

increased from less than 5% of all citations in 1995 to nearly 30% in 2001 (Rumsey, 2002).

However, compared to their paper counterparts web references have their own problems. It is now well documented that web pages and web sites come and go, and that occasionally they may resurface (Koehler, 2004). This has serious implications for researchers: whereas we are still able to read thousand-years-old written documents, the information put on the web a mere few years ago is in danger of being lost. The present study endeavours to ascertain the proportion of missing web references in research articles from 1998 to 2002 (5–10 years old) across five leading open access (OA) journals in the field of library and information science.

Objectives

The objective of the present study is to ascertain the proportion of missing web citations in 5–10-year-old research papers.

Scope

The scope of the present study is limited to web references in research articles (excluding editorials, news, reviews) from 1998 to 2002 in the following five leading OA journals in library and information science:

- Ariadne
- Library Philosophy and Practice
- Information Research
- Issues in Science and Technology Librarianship
- D-Lib Magazine

Methodology

In August 2008, the references in each research article from 1998 to 2002 of five OA journals were analysed to locate the web references. The URL

of each web reference was copied and pasted into the search box of Internet explorer to find out which references were missing. All the details were noted. To ensure that inaccessibility was not due to temporary server problems, another attempt was made to access the sites in October of 2008.

Related Literature

The web is not a particularly stable environment for the publication of long-term information and the maintenance of individual objects or items (Koehler, 2004). Taylor and Hudson (2000) found variation among domain types and subject collections of printed bibliographies of URLs and web lists. Kitchens & Mosley (2000), while discussing the ephemeral nature of web references, question the utility of printed internet guides. Germain (2000) also questions the usefulness of web resources as citations for scholarly literature due to their ephemeral nature. According to Spinellis (2003) approximately 28% of the URLs referred to in *Computer and Communications of the ACM* articles between 1995 and 1999 were no longer accessible in 2000 and the figure rose to 41% in 2002. A study by Benow (1998) found an attrition rate of 20% and 50% for websites over two- and three-year-periods. Nelson & Allen (2002) found a 3% attrition rate for digital library objects. Lawrence, Coetzee, Glover, Pennock, Flake, Nielsen, et al. (2001) also found many web references invalid after analysing 2,70,977 web references in computer science publications. The percentage of URLs that was invalid varied from 23% in 1999 to a peak of 53% in 1994. Harter & Kim (1996), after examining scholarly e-journal articles from 1993 to 1995, found that one third of the URLs were no longer accessible. Analysis of 1068 web citations by Sellitto (2005) demonstrated that 46% of all citations could not be accessed, with the HTTP 404 (page not found) message being the most common error message. A study of health-related web sites by Veronin (2002) found that 59% of the sites could not be found, 17% had moved to a new URL address and only 24% could be accessed at the original URLs. Rumsey (2002) tested citations over a span of five years in 2001 and found that 39% of URLs of 2001, 37% of 2000, 58% of 1999, 66% of 1998 and 70% of 1997 URLs were not accessible. Markwell & Brooks (2002) estimated the URL half-lives for online literature in the field of biochemistry and molecular biology at 4.6 years.

Findings and Discussion

A total of 8755 references appear in 630 articles across 1998–2002 in the five journals, 4001 (45.69%) of which are web references. 32.09% of web references, i.e., 1284 references are no longer accessible. The number of missing references goes on increasing with each passing year. The number of missing references for the publications of 2002 is 25.89%, which increases to 28.15% in 2001, to 37.23% in 2000, to 37.79% in 1999 and to 39.96% in 1998 (Table 1). The table also shows that with each passing year the number of web references is increasing (from 41.60% in 1998 to 53.32% in 2002). 74.14% (952 of 1284) of missing references resulted in an HTTP 404 error, while 0.93% (37 of 4001) references had moved to a new URL address with a link from the original URL.

Ariadne

Out of 1461 references appearing in 157 articles of *Ariadne* 1112 (76.11%) are web references. The number of web citations has increased from 50.0% in 1998 to 81.12% in 2002. 32.10% of web references, i.e., 357 are not accessible. The highest number of missing references (43.40%) is found in 2000 and the lowest (25.41%) in 2001. The percentage of missing references for 1998, 1999 and 2002 is 36.61%, 37.33% and 30.90% respectively (Table 2). Out of 1112 web references 11 (0.98%) had moved to a new URL address with a link from the original URL. 257 (71.98%) of missing references had an HTTP 404 problem.

Table 1. Reference statistics of five journals.

S. No.	Year	No. of articles	Total no. of references	No. of web references	Web references available	Web references missing	Missing references with HTTP 404 (page not found) error	Web references moved to a new URL address
1.	1998	126	1269	528 (41.60)	317	211 (39.96)	172 (81.51)	3
2.	1999	111	1489	553 (37.13)	344	209 (37.79)	154 (73.68)	7
3.	2000	120	1753	752 (42.89)	472	280 (37.23)	195 (69.64)	10
4.	2001	135	2050	998 (48.68)	717	281 (28.15)	210 (74.73)	9
5.	2002	138	2194	1170 (53.32)	867	303 (25.89)	221 (72.93)	8
Total		630	8755	4001 (45.69)	2717	1284 (32.09)	952 (74.14)	37

* Figures in parentheses indicate a percentage.

Table 2. Reference statistics of Ariadne.

S. No.	Year	No. of articles	Total no. of references	No. of web references	Web references available	Web references missing	Missing references with HTTP 404 (page not found) error	Web references moved to a new URL address
1.	1998	37	142	71 (50.0)	45	26 (36.61)	19 (73.07)	0
2.	1999	24	223	150 (67.26)	94	56 (37.33)	45 (80.35)	4
3.	2000	25	248	182 (73.38)	103	79 (43.40)	55 (69.62)	4
4.	2001	39	493	421 (85.39)	314	107 (25.41)	79 (73.83)	2
5.	2002	32	355	288 (81.12)	199	89 (30.90)	59 (66.29)	1
Total		157	1461	1112 (76.11)	755	357 (32.10)	257 (72.22)	11

* Figures in parentheses indicate a percentage.

Library Philosophy and Practice

Library Philosophy and Practice has published 32 articles during the period 1998–2002, which included 311 references. Out of 311 references only 70 (22.50%) are web references, 36 (51.42%) amongst which are not accessible. The highest and lowest number of missing references are in publications of year the years 2000 and 2001 respectively. Out of 34 web references 1 (2.94%) reference had moved to a new destination with a link from the original URL. 26 (72.22%) of missing references had an HTTP 404 error. The number of web citations has increased from 0% in 1998 to 31.25% in 2002 (Table 3).

Information Research

In *Information Research* the highest percentage of missing references (52.94%) is found for articles of 1998 and the lowest (29.41%) for 1999. The overall

Table 3. Reference statistics of Library Philosophy and Practice.

S. No.	Year	No. of articles	Total no. of references	No. of web references	Web references available	Web references missing	Missing references with HTTP 404 (page not found) error	Web references moved to a new URL address
1.	1998	3	0	0 (00.00)	0	0 (00.00)	0 (00.00)	0
2.	1999	6	97	20 (20.61)	13	7 (35.00)	4 (57.14)	1
3.	2000	7	58	4 (6.89)	3	1 (25.00)	0 (00.00)	0
4.	2001	6	76	21 (27.63)	7	14 (66.66)	10 (71.42)	0
5.	2002	10	80	25 (31.25)	11	14 (56.00)	12 (85.71)	0
Total		32	311	70 (22.50)	34	36 (51.42)	26 (72.22)	1

* Figures in parentheses indicate a percentage.

Table 4. Reference statistics of Information Research.

S. No.	Year	No. of articles	Total no. of references	No. of web references	Web references available	Web references missing	Missing references with HTTP 404 (page not found) error	Web references moved to a new URL address
1.	1998	14	293	17 (5.80)	8	9 (52.94)	8 (88.88)	0
2.	1999	15	475	34 (7.15)	24	10 (29.41)	6 (60.00)	0
3.	2000	23	678	101 (14.89)	55	46 (45.54)	41 (89.13)	4
4.	2001	25	701	90 (12.83)	55	35 (38.88)	24 (68.57)	0
5.	2002	24	594	147 (24.74)	94	53 (36.05)	40 (75.47)	2
Total		101	2741	389 (14.19)	236	153 (39.33)	119 (77.77)	6

* Figures in parentheses indicate a percentage.

percentage of missing references is 39.33%. The total number of references for 101 research articles published during 1998–2002 is 2741, 389 (14.19%) out of which are web references. 119 (77.77%) missing references had an HTTP 404 error, whereas 6 (2.54%) web references had moved to a new URL address. The number of web citations increased from 5.80% in 1998 to 24.74% in 2002 (Table 4).

Issues in Science and Technology Librarianship

The number of web citations in *Issues in Science and Technology Librarianship* has increased from 28.09% in 1998 to 33.80% in 2002. Out of 833 references appearing in 97 research articles 249 (29.89%) are web references. 55 (22.08%) web references are not accessible. The percentage of missing web references for the years 1998, 1999, 2000, 2001 and 2002 are 20.33%, 18.42%, 18.75%, 28.57% and 22.53% respectively (Table 5). Out of 194 web references 4 (2.06%)

Table 5. Reference statistics of Issues in Science and Technology Librarianship.

S. No.	Year	No. of articles	Total no. of references	No. of web references	Web references available	Web references missing	Missing references with HTTP 404 (page not found) error	Web references moved to a new URL address
1.	1998	21	210	59 (28.09)	47	12 (20.33)	9 (75.00)	0
2.	1999	16	96	38 (39.58)	31	7 (18.42)	6 (85.71)	0
3.	2000	17	163	32 (19.63)	26	6 (18.75)	5 (83.33)	0
4.	2001	20	154	49 (31.81)	35	14 (28.57)	7 (50.00)	2
5.	2002	23	210	71 (33.80)	55	16 (22.53)	9 (56.25)	2
Total		97	833	249 (29.89)	194	55 (22.08)	36 (65.45)	4

* Figures in parentheses indicate a percentage.

had moved to a new URL address with a link from the original URL and 36 (65.45%) of missing references displayed an HTTP 404 error.

D-Lib Magazine

The highest number of references (3409) appear in 243 research articles in *D-Lib Magazine*, 2181 (63.97%) out of which are web references. A total of 683 (31.31%) of references is no longer accessible. The highest percentage of missing references is for the publications of 1998 (43.04%) which decreases to 41.47% in 1999, 34.18 in 2000, 26.61 in 2001 and 20.50 in 2002. 514 (75.25%) of missing references had an HTTP 404 error, while 15 (1.0%) web references had moved to a new destination with a link from the original URL. The number of web citations increased from 61.05% in 1998 to 66.91% in 2002 (Table 6).

Conclusion

The present study reveals that with each passing year the number of web citations increased, and so was the number of missing citations. For five-year-old publications the percentage of missing citations is 25.89% and this increases to 39.96% for ten-year-old publications. The most common error is an HTTP 404 (page not found) message.

There are many causes for failed web references, one being the wholesale restructuring of any give domain (Koehler, 2004). Although in certain cases

Table 6. Reference statistics of D-Lib Magazine.

S. No.	Year	No. of articles	Total no. of references	No. of web references	Web references available	Web references missing	Missing references with HTTP 404 (page not found) error	Web references moved to a new URL address
1.	1998	51	624	381 (61.05)	217	164 (43.04)	136 (82.92)	4
2.	1999	50	598	311 (52.00)	182	129 (41.47)	93 (72.09)	2
3.	2000	48	606	433 (71.45)	285	148 (34.18)	94 (63.51)	2
4.	2001	45	626	417 (66.61)	306	111 (26.61)	90 (81.08)	5
5.	2002	49	955	639 (66.91)	508	131 (20.50)	101 (77.09)	2
Total		243	3409	2181 (63.97)	1498	683 (31.31)	514 (75.25)	15

* Figures in parentheses indicate a percentage.

the invalid web references could be located by means of alternative searches (Lawrence, Coetzee, Glover, Pennock, Flake, Nielsen, et al., 2001), this is not advisable from scholarly point of view as it is a time-consuming process. It has been observed that canonical URLs which took the form www.orgname.org and www.orgname.org.cc are more likely to persist than non-canonical forms (Koehler, 2004).

There is an immediate need to address the problem by devising and adopting uniform standards for long-term preservation of web resources and their persistence.

Future Research

No attempt has been made to locate citations other than by means of the designated URLs. As such there it is possible that some documents are available from a different URL. This could be explored in future research.

References

Benbow, S.M.P. (1998). 'File Not Found: the problem of changing URLs for the World Wide Web'. *Internet Research: Network Applications and Policy* 8(3).

Germain, C.A. (2000). 'URLs: uniform resource locators or unreliable reliable resource locators?', *College and Research Libraries* 61(4).

Harter, S. & H. Kim (1996). 'Electronic journals and scholarly communication: a citation and reference study'. *Information Research* 2 (1). Retrieved April 23, 2009 from http://informationr.net/ir/2-1/paper9a.html

Kitchens, J.D. & P.A. Mosley (2000). 'Error 404: or, WWhat is the shelf-life of printed Internet guides?', *Library Collections, Acquisitions & Technical Services* 24(4).

Koehler, Wallace (2004). 'A longitudinal study of Web pages continued: a consideration of document persistence', *Information Research* 9(2). Retrieved November 1, 2008 from http://informationr.net/ir/9-2/paper174.html

Lawrence, S., F. Coetzee, E. Glover, D. Pennock, G. Flake, F. Nielsen, et al. (2001). 'Persistence of Web References in Scientific Research', *IEEE Computer* 34(2), 26–31. Retrieved November 1, 2008 from http://wotan.liu.edu/docis/lib/goti/rclis/dbl/

ieecom/(2001)34%253A2%253C26%253APOWRIS%253E/www.neci.nec.com%252F~lawrence%252Fpapers%252Fpersistence-computer01%252Fpersistence-computer01.pdf

Markwell, J. & D.W. Brooks (2002). 'Broken links: the ephemeral nature of educational WWW hyperlinks', *Journal of Science Education and Technology* 11(2).

Nelson, M. & B. Allen (2002). 'Object persistence and availability in digital libraries', *D-Lib Magazine* 8(1). Retrieved April 24, 2009 from http://www.dlib.org/dlib/january02/nelson/01nelson.html

Rumsey, M. (2002). 'Runaway train: Problems of permanence, accessibility, and stability in the use of Web sources in law review citations', *Law Library Journal* 94(1).

Sellitto, Carmine (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* 56(7), 695–703.

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

Taylor, M.K. & D. Hudson (2000). '"Linkrot" and the usefulness of Web site bibliographies', *Reference & User Services Quarterly* 39(3).

Veronin, Michael A. (2002). 'Where Are They Now? A Case Study of Health-related Web Site Attrition', *Journal of Medical Internet Research* V4 (2). Retrieved November 1, 2008 from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1761933