# Retrieving vanished Web references in health science journals in East Africa

Web references in health science journals

385

Received 26 April 2017 Revised 5 June 2017 Accepted 6 July 2017

## Alfred Said Sife

Sokoine National Agricultural Library, Sokoine University of Agriculture, Morogoro, Tanzania, United Republic of Tanzania, and

## Edda Tandi Lwoga

Directorate of Library Services, Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania, United Republic of Tanzania

#### **Abstract**

**Purpose** – This study aims to examine the availability and persistence of universal resource locators (URLs) cited in scholarly articles published in selected health journals based in East Africa.

**Design/methodology/approach** – Four health sciences online journals in East Africa were selected for this study. In this study, all Web citations in the selected journal articles covering the 2001-2015 period were extracted. This study explored the number of URLs used as citations, determined the rate of URLs' loss, identified error messages associated with inaccessible URLs, identified the top domain levels of decayed URLs, calculated the half-life of the Web citations and determined the proportion of recovered URL citations through the Internet Wayback Machine.

**Findings** – In total, 822 articles were published between 2001 and 2015. There were in total 17,609 citations of which, only 574 (3.3 per cent) were Web citations. The findings show that 253 (44.1 per cent) Web citations were inaccessible and the "404 File Not Found" error message was the most (88.9 per cent) encountered. Toplevel domains with country endings had the most (23.7 per cent) missing URLs. The average half-life for the URLs cited in journal articles was 10.5 years. Only 36 (6.3 per cent) Web references were recovered through the Wayback Machine.

**Originality/value** – This is a comprehensive study of East African health sciences online journals that provides findings that raises questions as to whether URLs should continue to be included as part of bibliographic details in the lists of references. It also calls for concerted efforts from various actors in overcoming the problem of URL decay.

**Keywords** East Africa, Journals, Wayback machine, Citation, Vanished URLs, Web references **Paper type** Research paper

#### Introduction

The volume of scholarly electronic publications, particularly electronic journals (e-journals), has increased tremendously in the past few years following spectacular transformations brought about by developments in information and communication technologies (ICTs). ICT developments have greatly revolutionized almost all aspects of scholarly publishing ranging from data collection and analysis, manuscript preparation and submission, peer review process, dissemination, accessibility and archiving. Scholarly publications are now deposited, stored and published online where users can discover and retrieve them through electronic search facilities such as Web search engines or by following their universal resource locators (URLs).



Information and Learning Science Vol. 118 No. 7/8, 2017 pp. 385-392 © Emerald Publishing Limited 2398-5348 DOI 10.1108/ILS-04-2017-0030 A URL is a Web address that specifies a location to a document and enables Web browsers to reach it on any server. Generally, URLs consist of four parts – protocol, domain, directory and file. The protocol or hypertext transfer protocol (HTTP) enables browsers to connect with Web servers. A domain is an internet address that provides connection to a Web server and it can indicate the type of organization such as *.com* for commercial organizations, *.edu* or *.ac.* for education institutions, *.gov* for government sites and *.org* for non-profit organizations (Tajeddini *et al.*, 2011). Directory is the name for the folder in which the content is located, and the file stores the content.

Scholarly writing requires authors to correctly cite others' work that they have consulted. In-text citations and matching lists of references show readers where ideas or facts were taken, and enable them to find the original source in case they want more details. Citations and references also show the context of new research in relation to earlier research, enable authors to substantiate their arguments and credit other authors' ideas. Consequently, the use of URLs has become common in scholarly communication, whereby authors are now required to include these URLs as part of bibliographic details.

Even though the Web provides wider and faster accessibility to e-resources, their continued accessibility is sometimes challenging as they often vanish from their locations. A number of previous studies (Rumsey, 2002; Markwell and Brooks, 2003; Tyler and McNeil, 2003; Sellitto, 2004) had reported that between 10 and 40 per cent of the citations tend to disappear from their Web locations for reasons such as expiry of domain names, removal of Web pages, redesigning of websites, moving of files to other locations, equipment malfunctions and internet connectivity problems. Recent studies have reported even higher proportions of vanished URLs. A study of articles published from the late 1990s to 2012 in three legal journals from Harvard found that more than 70 per cent of URLs cited suffered reference rot (Zittrain *et al.*, 2014). Klein *et al.* (2014) established that between 70 and 80 per cent of Web references of scholarly articles published between 2005 and 2012 in the domains of science, technology and mathematics suffered from reference rot.

To locate a resource on the Web, a URL must be resolved to a valid Internet Protocol (IP) address. Failure to resolve a valid IP address leads to an HTTP error message. Errors are often encountered when the requested URL cannot be understood by the server (400 Bad Request); user authentication required (401 Unauthorized); server refuses to fulfil a request (403 Forbidden); server not found anything matching the requested URL (404 Not Found) and when the server cannot send an acceptable response (406 Not Acceptable). Errors also occur when a server encounters unexpected conditions which prevent it from fulfilling a request (500 Internal Server Error); server is unable to handle a request (503 Service Unavailable); and when a server does not receive a timely response from an upstream server specified by the URLs (504 Gateway Time-out) (Powell, 2003; Spinellis, 2003). Failure to locate online references not only raises questions on the necessity of citing URLs but also undermines the foundations of academic writing in which new research is built from published literature.

A considerable body of literature (Casserly and Bird, 2003; Dimitrova and Bugeja, 2007; Goh and Ng, 2007; Falagas *et al.*, 2008; Wagner *et al.*, 2009) reports on the types of errors encountered. URL decay rates reported in previous studies range from 30 to 65 per cent and "File Not Found" is the most frequent reported error message. Some studies (Dimitrova and Bugeja, 2007) have reported that *.edu* domain ending had the most active links, while others (Goh and Ng, 2007) found that *.org* had the most active links. With regard to the half-lives of online resources, studies have established different values ranging from 1.6 years (Rumsey, 2002) to 11.5 years (Kumar and Kumar, 2012).

iournals

references in

health science

Web

Increasingly, efforts have been made by publishers, academic institutions and private organizations to find a solution for vanished URLs. At least 68 Web archiving initiatives have been undertaken worldwide to acquire, store, manage and preserve parts of the Web. Available access tools for Web archiving include Time Travel Portal, Time Travel APIs, Memento Time Travel, NutchWAX, WERA and the Wayback Machine. The Internet Archive's Wayback Machine is one of the first Web archiving tools established in 1996, and it had 376 billion archived Web pages by 2014 (this is equivalent to 13.8 petabyte) (Costaet al., 2016; IIPC, 2017).

Casserly and Bird (2003) reported that the use of the Wayback Machine enabled recovery of 47 per cent of missing URLs. Through the same techniques, Dimitrova and Bugeja (2007) retrieved 53.4 per cent of vanished URLs, while Russell and Kane (2008) retrieved 57 per cent of missing URLs. Wagner *et al.* (2009) reported that the Wayback Machine was able to retrieve 59.8 per cent of missing URLs as compared to the Google search engine which could retrieve 48.8 per cent. A study conducted by Tajeddini *et al.* (2011) found that the accessibility of online citations improved from 66 to 95 per cent by using the Wayback Machine and Google in six LIS scholarly journals. Similarly, Sadat-Moosavi *et al.* (2012) reported an increase in the accessibility rate URLs from 64 to 95 per cent by using the Wayback Machine and Google in four LIS scholarly journals.

Keeping in view the volatile nature of Web resources, the present study examined the extent of URLs vanishing in selected health sciences journals based in East Africa. It also attempted to recover the missing URLs using the Wayback Machine. One reason for this selection is that health sciences often have higher publication rates and therefore many of such journals publish their issues regularly. Even though several studies have been carried out regarding the disappearance and recovery of online citations, there is scarcity of such studies for publications based in the East African region. Specifically, this study explored the proportion of Web references used in the selected journals, determined the proportion of missing Web citations, identified error messages associated with inaccessible URLs, identified top domain levels of decayed URLs, calculated the half-life of the Web citations and determined the proportion of recovered Web citations.

#### Methods

This study was conducted in October 2015 to examine URL decay and recovery in four East Africa based health journals, namely, the *East and Central African Journal of Surgery, Journal of Applied Biosciences, Rwanda Medical Journal* and the *Tanzania Journal of Health Research*. These four journals were purposeful selected based on their easy retrieval through the internet and that they had adequate number of articles. All articles in the published issues of the selected journals covering the 2001-2015 period were downloaded and their reference lists were examined. The present study has some limitations in that we chose only four journals which might not be representative of all scholarly publications in East Africa.

All Web citations were extracted and their URLs were tested for their accessibility. These citations were considered active if an online publication was found at its original location or if automatically redirected to a new location. Because some sites might have been only temporarily unavailable, inactive links were rechecked several times in a week and if they were still inaccessible, they were recorded as vanished and the associated error message and top domain ending were recorded. The vanished URLs were recovered through the Internet Wayback Machine (https://archive.org/web/). The Wayback Machine was selected for this study because it is one of the earliest Web archiving tools with a large collection of historical website content. The half-life of Web citations in journal articles was established through a procedure previously used by Moghaddam *et al.* (2010) i.e.  $t(h) = [t\ln(0.50)]/[\ln W(t) - W(0)]$ , where t(h) is half-life of Web citations for each year, W(0) is

the number of Web citations at the time of publication and W(t) is the number of active Web citations at some later time t. Collected data were compiled using Microsoft Excel.

#### Results and discussions

The study findings in Tables I and II indicate that all four journals had produced a total of 822 articles between 2001 and 2015. These articles had a total of 17,609 references, giving an average of 21 references per article. There were only 574 (3.3 per cent) online references for all four journals, giving an average of 143.5 online references per journal. This suggests that either there is very low usage of online resources by authors publishing in these journals or there is non-adherence to citation guidelines that require inclusion of URL in the reference details. The findings indicate an increasing trend of online references between 2001 and 2015 with some fluctuations in the years in-between. However, of the 574 online references, 253 (44.1 per cent) were inactive. As pointed out earlier, some previous studies such as Davis and Cohen (2001), Falagas *et al.* (2008), Wagner *et al.* (2009) had reported even higher decay rates. Surprisingly, some online references from new articles were also inactive. For instance, nearly half (47.6 per cent) of online references from articles published in 2014 were inactive. This suggests that even newly cited Web resources can disappear rather quickly.

The vanished URLs were recovered through the Wayback Machine in which their web addresses were entered in the search box of the Wayback Machine and searched. If the submitted URL existed in this tool then it was considered as recovered. The findings indicate that only 36 (14.2 per cent) of inactive URLs were recovered through Wayback Machine. This is contrary to many previous studies (Dimitrova and Bugeja, 2007; Russell and Kane, 2008; Wagneret al., 2009; Tajeddini et al., 2011) which found that the Wayback Machine was able to recover more than 50 per cent of vanished URLs.

The East and Central African Journal of Surgery had the highest proportion (59 per cent) of active online references and Journal of Applied Biosciences had the highest proportion (63.6 per cent) of inactive online references. The proportion of recovered online references in all four journals was less than 18 per cent (Table II).

Error messages associated with decayed URLs were recorded and the 404 Not Found error was the most (81.4 per cent) encountered (Table III). This means that online references in journal

Year	Total no. of articles	Total no. of references	Online references	Active online references	Inactive online references	Recovered online references
2001	28	264	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
2002	10	109	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
2003	12	166	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
2004	33	369	4(1.1)	4 (100)	0 (0.0)	0 (0.0)
2005	15	210	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
2006	33	486	17 (3.5)	2 (11.8)	15 (88.2)	1 (6.7)
2007	44	704	12(1.7)	3 (25.0)	9 (75.0)	0 (0.0)
2008	93	1,897	59 (3.1)	26 (44.1)	33 (55.9)	5 (15.2)
2009	69	1,244	40 (3.2)	19 (47.5)	21(52.5)	3 (14.3)
2010	75	1,759	53 (3.0)	27 (50.9)	26 (49.1)	2 (7.7)
2011	93	2,414	98 (4.1)	69 (70.4)	29 (29.6)	12 (41.4)
2012	107	2,464	77 (3.1)	43 (55.8)	34 (44.2)	6 (17.6)
2013	95	2,344	65 (2.8)	37 (56.9)	28 (43.1)	3 (10.7)
2014	85	2,528	84 (3.3)	44 (52.4)	40 (47.6)	2 (5.0)
2015	30	651	65 (10.0)	47 (72.3)	18 (27.7)	2(11.1)
Total	822	17,609	574 (3.3)	321 (55.9)	253 (44.1)	36 (14.2)

**Table I.**Summary of references by years

Web

articles had disappeared from their original locations and nothing was matching the requested URLs. The 404 File Not Found error is largely caused by changes in the URL such as the removal or relocation of files as well as changes in file or directory names. These findings confirm the volatile nature of the Web as a publishing medium where Web resources can be easily altered or removed from their original locations. Errors in typing URLs can also contribute to URL decay. Many previous studies (Casserly and Bird, 2003; Goh and Ng, 2007; Wagner et al., 2009) have also reported that "File Not Found" was the most frequently encountered error message. The findings also indicate that very few (5.8 per cent) vanished URLs with error message 404 Not Found were recovered through the Wayback Machine.

The findings in Table IV indicate that of the 253 decayed URLs, top-level domains with country endings had the most (23.7 per cent) missing URLs. Similarly, one-third (33.3 per cent) of the recovered online references were those with country endings. Overall, the most recovered Web citations were those with top-level domain .int (23.5 per cent) followed by those with .net (22.2 per cent) and those with country endings (20 per cent). This is contrary to earlier studies (Dimitrova and Bugeja, 2007; Goh and Ng, 2007; Moghaddam *et al.*, 2010) which reported .org and .edu as some of the most persistent domains.

The half-life of Web citations in the four journals was computed using the formula:  $t(h) = [t\ln(0.50)]/[\ln W(t) - \ln W(0)]$ , where t(h) is the half-life of online references for each year, W(0) is the number of online references at the time of publication, and W(t) is the number of active online references at some later time t (Moghaddam et al., 2010). The results of this estimation

Journals	References	Online references	Active online references	Inactive online references	Recovered online references	
East and Central African						
Journal of Surgery	8,825	244 (2.8)	144 (59.0)	100 (41.0)	12 (4.9)	Table II.
Journal of Applied						Summary of
Biosciences	410	11 (2.7)	4 (36.4)	7 (63.6)	0 (0.0)	vanished and
Rwanda Medical Journal	1,358	87 (6.4)	47 (54.0)	40 (46.0)	6 (6.9)	recovered online
Tanzania Journal of						
Health Research	7,016	232 (3.3)	126 (54.3)	106 (45.7)	18 (7.8)	references by
Total	17,609	574 (3.3)	321 (55.9)	253 (44.1)	36 (19.6)	journals

Types of errors	Types of errors for vanished references	(%)	Types of errors for recovered references	(%)	Proportion of recovered references (%)	
401					_	
Unauthorised	1	0.4	1	2.8	100	
403 Forbidden	5	2.0	2	5.6	40.0	
404 Not Found	206	81.4	12	33.3	5.8	
406 Not						
Acceptable	11	4.3	4	11.1	36.4	Table III.
410 Gone	8	3.2	4	11.1	50.0	Types of errors
500 Internal						associated with
Server Error	13	5.1	6	16.7	46.2	vanished and
503 Service						
Unavailable	9	3.6	7	19.4	77.8	recovered online
Total	253	100	36	100		references

presented in Table V indicate that the average half-life for the cited URLs in journals was 10.5 years. This means that it takes 10 years and 6 months for half of the online references in these four journals to disappear. This average half-life is higher than those reported by Koehler (2002) (two years), Rumsey (2002) (1.6) years, Markwell and Brooks (2003) (4.6 years), as well as Goh and Ng (2007) (five years). Nevertheless, a much higher figure (14.94 years) was reported by Moghaddam *et al.* (2010). These findings again emphasize the ephemeral nature of Web resources which makes important information to go missing.

The findings in the present study show that many online references in the four East Africa health journals had disappeared from their original locations and could not be found in other online locations. This contradicts the desire of scholarly writing requirements that URLs should be included as part of bibliographic details so as to ensure continued access to the cited resources. This disappearance implies that readers cannot retrieve the cited sources in case they want to see more details. This in turn contradicts the principle that the advancement of knowledge is incremental with authors *standing on the shoulders of others in furthering their own work* (Sellitto, 2004). Missing the cited resources hampers readers from seeing the way authors used works of others to support their arguments, substantiate

Domain endings	Domain endings of inactive references	(%)	Domain endings of recovered references	(%)	Proportion of recovered references (%)
. com/.co	31	12.3	3	8.3	9.7
.net	9	3.6	2	5.6	22.2
.org	40	15.8	3	8.3	7.5
.edu	22	8.7	2	5.6	9.1
.gov/. go	23	9.1	3	8.3	13.0
Country endings	60	23.7	12	33.3	20.0
.html	14	5.5	1	2.8	7.1
.int	17	6.7	4	11.1	23.5
.info	11	4.3	1	2.8	9.1
Others e.g.	26	10.3	5	13.9	19.2
Total	253	100.0	36	100.0	

Table IV.
Summary of
vanished and
recovered online
references by domain

**Table V.**Half-life of Web citations

Year	Online references	Active online references	Half-life
2001	0	0	0.0
2002	0	0	0.0
2003	0	0	0.0
2004	4	4	0.0
2005	0	0	0.0
2006	17	2	4.5
2007	12	3	7.0
2008	59	26	11.8
2009	40	19	13.0
2010	53	27	14.4
2011	98	69	27.7
2012	77	43	16.7
2013	65	37	17.2
2014	84	44	15.0
2015	65	47	29.9
Average			10.5

Web

references in

health science iournals

claims, build upon existing works, provide contexts of research studies and compare approaches and methodologies. The findings in study have also shown that tools such as the Wayback Machine that have been designed to archive Web pages are proving to be a failure.

#### Conclusion

This study has found that the number of online publications cited in four journals published in East Africa was very low. This is an indicator of low usage of e-resources by scholars in the region. There could also be non-adherence to citation styles in these journals. Still, many online resources cited in these journals have disappeared from their Web locations. The findings have also shown that, on average, it takes 10 years and 6 months for half of the Web citations in these journal articles to disappear. Unfortunately, the Wayback Machine has not been able to retrieve back the majority of the vanished URLs. This implies that the long-term availability and accessibility of online information resources cited in these journals cannot be guaranteed. This, in turn, raises questions as to whether URLs should continue to be included as part of bibliographic details in the lists of references.

Efforts are therefore required from various actors including authors, editors, publishers, librarians, Web managers and ICT professionals to overcome the problem of vanished URLs. Because the number of online references in the four East African based journals was extremely low, efforts should be undertaken by libraries and research institutions to promote the use of e-resources by researchers. Journal editors and publishers should ensure that authors adhere to correct citation styles including providing detailed bibliographical description in their list of references. In addition, the use of citation and referencing software by authors would avoid citation errors resulting from typing. Journal publishers should make use of Web archiving tools to archive the URLs cited in the articles they publish. This study also supports a number of recommendations of previous studies that include the need for authors to retain digital backup or printed copies of cited Web resources, advocating for the inclusion of Web content in Internet archives, checking URLs systematically before publishing, using digital object identifiers (DOIs) and uniform resource names (URNs) in place of URLs and establishing institutional repositories to upload copies of scholarly material such as preprints. Future research would ensure representation of the entire scholarly community by taking into consideration a diverse range of journals in the East Africa region. Researchers could also compare the retrieval efficiency of various Web archiving tools in accessing vanished URLs.

#### References

- Casserly, M.F. and Bird, J.E. (2003), "Web citation availability: analysis and implications for scholarship", *College & Research Libraries*, Vol. 64 No. 4, pp. 300-318, doi: 10.5860/crl.64.4.300.
- Davis, P.M. and Cohen, S.A. (2001), "The effect of the web on undergraduate citation behavior 1996 1999", Journal of the Association for Information Science and Technology, Vol. 52 No. 4, pp. 309-314.
- Dimitrova, D.V. and Bugeja, M. (2007), "Raising the dead: recovery of decayed online citations", American Communication Journal, Vol. 9 No. 2.
- Falagas, M., Karveli, E. and Tritsaroli, V. (2008), "The risk of using the internet as reference resource: a comparative study", *International Journal of Medical Informatics*, Vol. 77 No. 4, pp. 280-286.
- Goh, D. and Ng, P. (2007), "Link decay in leading information science journals", Journal of the American Society for Information Science and Technology, Vol. 58 No. 1, pp. 15-24.
- Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K. and Tobin, R. (2014), "Scholarly context not found: one in five articles suffers from reference rot", in Bar-Ilan, J. (Ed.), *PLoS ONE*, Vol. 9 No. 12, p. e115253, doi: 10.1371/journal.pone.0115253.

- Koehler, W. (2002), "Web page change and persistence: a four-year longitudinal study", Journal of the American Society for Information Science and Technology, Vol. 53 No. 2, pp. 162-171, doi: 10.1002/asi.10018.
- Kumar, B.T.S. and Kumar, K.S.M. (2012), "Persistence and half-life of URL citations cited in LIS open access journals", *Aslib Proceedings*, Vol. 64 No. 4, pp. 405-422, doi: 10.1108/00012531211244752.
- Markwell, J. and Brooks, D. (2003), "Link rot" limits the usefulness of web-based educational materials in biochemistry and molecular biology", *Biochemistry and Molecular Biology*, Vol. 31 No. 1, pp. 69-72.
- Moghaddam, A.I., Branch, H., 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 Science and Management Availability*, Vol. 8 No. 2, pp. 57-75.
- Powell, T.A. (2003), HTML & XHTM: The Complete Reference, McGraw-Hill, New York, NY.
- Rumsey, M. (2002), "Runaway train: Problems of permanence, accessibility, and stability in the use of web sources in law review citations", *Law Library Journal*, Vol. 94 No. 1, pp. 27-39.
- Russell, E. and Kane, J. (2008), "The missing link: assessing the reliability of internet citations in history journals", *Technology and Culture*, Vol. 49 No. 2, pp. 420-429, doi: 10.1353/tech.0.0028.
- Sadat-Moosavi, A., Isfandyari-Moghaddam, A. and Tajeddini, O. (2012), "Accessibility of online resources cited in scholarly LIS journals: a study of emerald ISI-ranked journals", Aslib Proceedings, Vol. 64 No. 2, pp. 178-192, doi: 10.1108/00012531211215196.
- Sellitto, C. (2004), "A study of missing web-cites in scholarly articles: towards an evaluation framework", *Journal of Information Science*, Vol. 30 No. 6, pp. 484-495.
- Spinellis, D. (2003), "The decay and failures of web references", Communications of the ACM, Vol. 46 No. 1, pp. 71-77, doi: 10.1145/602421.602422.
- 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, pp. 17-29.
- Tyler, D. and McNeil, B. (2003), "Librarians and link rot: a comparative analysis with some methodological considerations", *Portal: Libraries and the Academy*, Vol. 3 No. 4, pp. 615-632.
- Wagner, C., Gebremichael, M.D., Taylor, M.K. and Soltys, M.J. (2009), "Disappearing act: decay of uniform resource locators in health care management journals", *Journal of the Medical Library Association: JMLA*, Vol. 97 No. 2, pp. 122-130, doi: 10.3163/1536-5050.97.2.009.
- Zittrain, J., Albert, K. and Lessig, L. (2014), "Perma: scoping and addressing the problem of link and reference rot in legal citations", *Harvard Law Review*, Vol. 127 No. 4, pp. 176-199, available at: http://harvardlawreview.org/2014/03/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations

#### **Further reading**

Isfandyari Moghaddam, A., Saberi, M.K. and Mohammad Esmaeel, S. (2010), "Availability and half-life of web references cited in information research journal: a citation study", *International Journal of Information Science and Management*, Vol. 8 No. 2, pp. 57-74, doi: 10.1108/02640471211204042.

### Corresponding author

Edda Tandi Lwoga can be contacted at: tlwoga@gmail.com