# Citations to Web Pages in Scientific Articles: The Permanence of Archived References

Andrea W. Thorp, MD, David L. Schriger, MD, MPH

*From the Loma Linda University Medical Center and Children's Hospital, Loma Linda, CA (Thorp); and the University of California, Los Angeles, Los Angeles, CA (Schriger).*

**Study objective:** We validate the use of archiving Internet references by comparing the accessibility of published uniform resource locators (URLs) with corresponding archived URLs over time.

**Methods:** We scanned the "Articles in Press" section in *Annals of Emergency Medicine* from March 2009 through June 2010 for Internet references in research articles. If an Internet reference produced the authors' expected content, the Web page was archived with WebCite (http://www.webcitation.org). Because the archived Web page does not change, we compared it with the original URL to determine whether the original Web page had changed. We attempted to access each original URL and archived Web site URL at 3-month intervals from the time of online publication during an 18-month study period. Once a URL no longer existed or failed to contain the original authors' expected content, it was excluded from further study. The number of original URLs and archived URLs that remained accessible over time was totaled and compared.

**Results:** A total of 121 articles were reviewed and 144 Internet references were found within 55 articles. Of the original URLs, 15% (21/144; 95% confidence interval [CI] 9% to 21%) were inaccessible at publication. During the 18-month observation period, there was no loss of archived URLs (apart from the 4% [5/123; 95% CI 2% to 9%] that could not be archived), whereas 35% (49/139) of the original URLs were lost (46% loss; 95% CI 33% to 61% by the Kaplan-Meier method; difference between curves $P<.0001$, log rank test).

**Conclusion:** Archiving a referenced Web page at publication can help preserve the authors' expected information. [Ann Emerg Med. 2011;57:165-168.]

Please see page 166 for the Editor's Capsule Summary of this article.

Provide **feedback** on this article at the journal's Web site, www.annemergmed.com.
A **podcast** for this article is available at www.annemergmed.com.

0196-0644/$-see front matter

doi:10.1016/j.annemergmed.2010.11.029

## INTRODUCTION

It was not so long ago that all references in scientific articles were to printed materials or personal communications. Now, 1% to 19% of articles contain a reference to a virtual page on the Internet,[1-5] with an increasing number of Internet references published each year.[1,3,6] In contradistinction to print materials, which are archived in libraries, Web pages are not permanent; uniform resource locators (URLs) referenced within the scientific and medical literature become inaccessible over time.[1-12] It is estimated that 10% of URLs are inaccessible for in-press articles,[4] 27% inaccessible within 3 years from publication,[5] and 45% to 78% inaccessible within 5 years.[1,9] The rate of URL inaccessibility has been estimated at 5.4% per year.[4] Despite this decay, referencing Web pages remains an accepted practice in the scientific community.

Various ways to preserve Internet information have been suggested. One possible solution is eliminating Internet references altogether.[7] Editors will find this increasingly difficult because more sources of information are exclusively online. Expecting the reader to find the information by using a search engine, such as Google or Internet Archive's Wayback Machine, has also been suggested.[3,5,6,13] It is estimated that approximately 4% to 59% of information can be retrieved with these techniques,[3,5,6] but there is no guarantee that the information retrieved matches the information the author intended to cite.[5] Expecting authors or journals to keep a printed copy of the referenced Web page is feasible[7,6,13] when the URL is in standard print format but limits the readers' access to the information. Preserving the Web pages with digital object identifiers is gaining popularity among larger publishers, but the cost of this option dissuades smaller publishers and authors from participating.[6,7] Permanently archiving a Web page in a database maintained by an archiving service such as WebCite has been widely recommended.[1-6,10,12,13] When a URL is archived

**Editor's Capsule Summary**

*What is already known on this topic*

Authors often link to Internet sites to support or expand the study presentation. Web archiving is a tool to help maintain that link.

*What question this study addressed*

How often do article Web citations become inaccessible, and does archiving alter any decay?

*What this study adds to our knowledge*

Examining 15 months' worth of *Annals* articles revealed 15% of Web references were inaccessible at print and 35% became inaccessible by 18 months. Archiving eliminated that decay but was not possible for 4% of the Web references.

*How this is relevant to clinical practice*

Unless articles are archived at the onset, readers may find important Web links to be absent, potentially impairing the message of an article.

with WebCite, a snapshot of the Web page is preserved and a new URL is assigned. The author includes both the original and archived URLs in the reference to guarantee the reader direct access to the exact Web page the author intended to cite.

If authors or journals were required to archive Internet references, would the information be adequately preserved over time? The objective of this study is to validate the use of archiving by comparing the accessibility of published URLs with the corresponding archived URLs over time.

## MATERIALS AND METHODS

Once each week, we scanned the "Articles in Press" section in *Annals of Emergency Medicine* from March 2009 through June 2010 for research articles added since our previous check. We checked each article for Internet references and, if present, noted the number of unique references. We assessed the availability of each Internet reference by cutting and pasting the URL into an Internet browser to determine whether it produced the desired content. If no Web site was produced, we checked for typographical or obvious errors. If the URL was unable to produce a Web site or the Web site failed to produce the expected information, it was rechecked for 2 weeks and, if not available, was withdrawn from the study. We archived all Internet references that produced the intended information by using WebCite (http://www.webcitation.org). If WebCite failed to produce a usable copy of the information referenced by the original URL, that reference was no longer checked for accessibility.

To evaluate the permanence of the original and archived URLs, we attempted to access each original URL and archived WebCite URL at 3-month intervals from online publication during an 18-month study period. Because the archived Web page does not change, we used it as the criterion standard to which we could compare the original URL to determine whether the original URL Web page had changed. If on the 3-month check a URL no longer existed ("404 error," "server not found," or "page not found") or failed to contain the authors' original information, it was rechecked for 2 weeks and then declared inactive. The number of original URLs and archived URLs that remained accessible over time was totaled and compared. Kaplan-Meier statistics, 95% confidence intervals (CIs), survival curve, and log-rank statistics were calculated with the "sts" commands in Stata (version 11; StataCorp, College Station, TX). Our study contained no human subjects, and approval from the institutional review board was not required.

## RESULTS

A total of 121 articles were reviewed, with 45% (55/121; 95% CI 36% to 55%) of the articles containing a URL reference. A total of 144 Internet references were found within these 55 articles (Figure 1). None of the citations offered an alternative URL to an archival service such as webcitation.org. The number of Internet references in each article ranged from 1 to 8.

Of the 144 original URLs, 15% (21/144; 95% CI 9% to 21%) were inaccessible at publication and therefore not archived. Of the original URLs that were found, 4% (5/123; 95% CI 2% to 9%) could not be successfully archived because of incompatibility between the Web page and the archiving process. Of the 118 archived URLs, 80% (94/118; 95% CI 71% to 87%) produced an identical copy of the original Web page. The remainder produced a copy that contained all the pertinent information but was not identical. For example, the archived Web page failed to capture a graphic in a sidebar but contained all relevant text.

Of the 118 original URLs that we were able to successfully archive, 12% (14/118; 95% CI 7% to 19%) were no longer accessible at the 3-month check. An additional 9% (7/79) became inaccessible at 6 months, 6% (4/64) at 9 months, 6% (2/36) at 12 months, and an additional 1% (1/10) at 18 months (Table; Figure 2). There were no URLs that became inaccessible at the 15-month checkpoint. This represents an actual loss of 35% of the 139 citations and a projected loss of 46% (95% CI 33% to 61%) had all citations been observed for the full 18 months. This compares with a 4% (5/121) loss of archived URLs ($P<.0001$, log rank test).

## LIMITATIONS

The URLs were collected from a single journal in the emergency medicine literature; however, the loss of original URLs is similar to that reported in journals in other disciplines.[2,6] We had no formal method for catching formatting or typing errors in the printing of the original URL

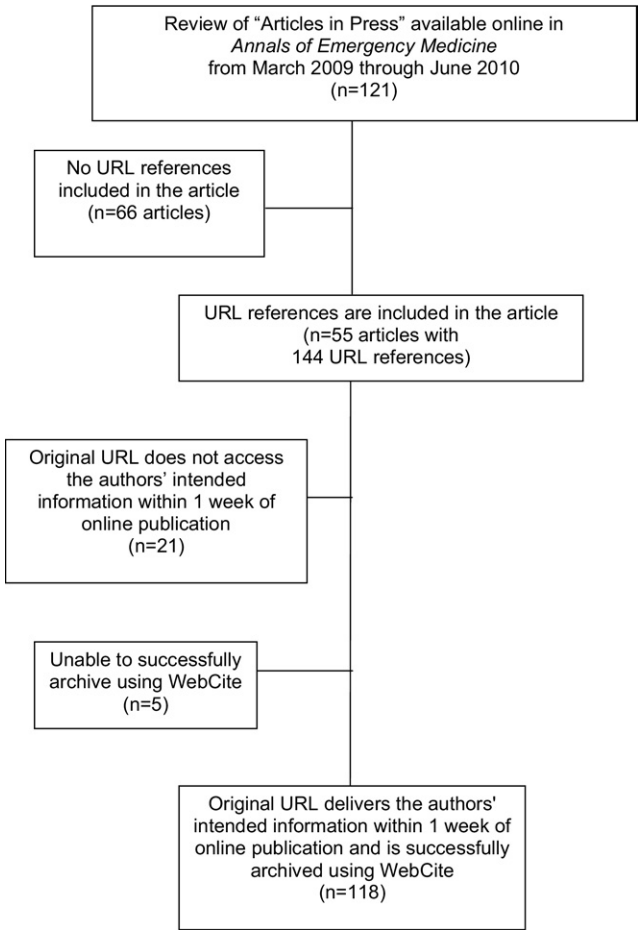


**Figure 1.** Flow diagram of URLs found in *Annals of Emergency Medicine.*

**Table.** Number of original URLs published in *Annals of Emergency Medicine* between March 2009 and June 2010 that were successfully archived and followed at 3-month intervals from online publication.

| Months From Online Publication | Total No. of URLs Published | Number of Accessible URLs, Months | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | At Publication | 3 | 6 | 9 | 12 | 15 | 18 |
| 18 | 16 | 12 | 12 | 11 | 10 | 10 | 10 | 9 |
| 15 | 22 | 20 | 17 | 16 | 16 | 15 | 15 | — |
| 12 | 21 | 15* | 12 | 10 | 10 | 9 | — | — |
| 9 | 41 | 33* | 28 | 28 | 24 | — | — | — |
| 6 | 16 | 13* | 10 | 8 | — | — | — | — |
| 3 | 28 | 25* | 25 | — | — | — | — | — |

*A total of 5 original URLs were unable to be archived within these groups, and the original URL was excluded from these totals.

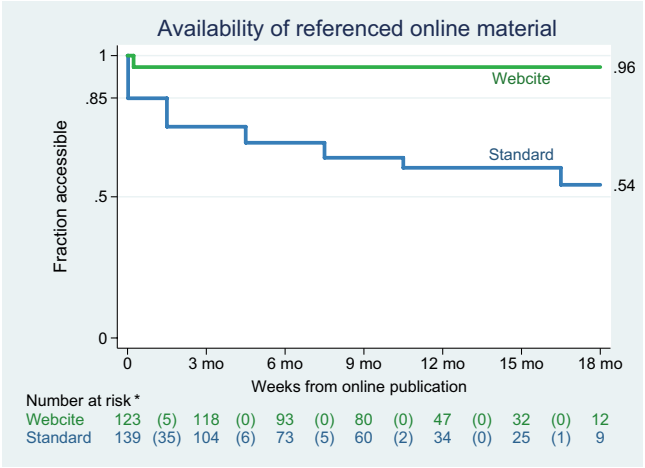


**Figure 2.** The numbers in parentheses represent the number of URLs that disappeared during that period. Of the 144 original URLs, 21 were already lost when checked on the week after their online publication. An additional 14 original URLs were lost between 1 week and 3 months. The 139 starting point represents the 144 minus 5 that could not be archived with WebCite. Of the 123 URLs that were active at 1 week, 5 could not be Web cited. The remaining 118 WebCite URLs remained operational throughout the 18-month study period.

references. Previous studies have reported that 12% to 24% of published URLs contain formatting or spelling errors.[2,10] Although possible, it is doubtful that all the inaccessible original URLs at publication were due to errors.

Even though the original URLs were archived when the article was available for online publication, it is possible that the authors' intended information had already changed before archiving. If this were true, we underestimated decay. Finally, we did not include the 5 references that could not be archived in the survival data for the original references because we could not determine whether they had changed. Four of the 5 are still active. However, including these in analysis does not change the percentage actually lost, which remains 35% (50/144).

## DISCUSSION

Published URL references became inaccessible over time, but those that were successfully archived at publication remained extant during the entire study period. It has been estimated that URLs become inaccessible at a rate of 5% per year,[4] but of the original URLs that became inaccessible in our study, a majority were inaccessible within 6 months, demonstrating a fast rate of decay near publication and a slower rate of decay thereafter.

Many authors have proposed using archival databases to preserve Internet references used within scientific publications.[1-6,10,12,13] Archiving with WebCite has many advantages for the reader, author, and publisher. The service is hosted through the University of Toronto and is financially supported through the WebCite consortium by publishers and libraries that have an interest in preserving information.[14] Authors do not pay to archive a Web page and readers do not pay to access it.[13]

The URL that is provided through the archiving process provides a "permalink" that produces a "snapshot" of the exact

Web page that the author intended to cite.[1,13,14] Similar to printed text within a journal, this snapshot Web page provides the reader the opportunity to view and formulate conclusions about the cited information. The archived Web page can be retrieved by entering the archived URL directly into the search engine or by using WebCite's search functions.[13] According to WebCite, hundreds of journals are using the service, and journals are starting to amend the instructions for authors to require archiving of Internet references.[13,14] Wren[2] found only 2 articles that included http://www.webcitation.org notations, and these articles were describing WebCite archiving.

The success of archival services for Internet material depends on 2 factors. First, the service needs to be able to archive all types of Internet material. In this study, we were unable to archive 4% of the published URLs because of incompatibility with the archiving process. Although WebCite can accommodate many types of information, dynamic Web pages cannot be archived.[6] Web page owners reserve the right to encode archiving restrictions into the Web page and, if the Web page is archived, can insist that the Web page be removed from the archived database.[5,6,13] In staying compliant with copyright law, WebCite must honor these restrictions.[13,14]

Second, the archival service must remain extant. To the extent that such services are dependent on donations for their survival, they remain vulnerable to funding deficiencies. Just as the federal government has taken over stewardship of clinical trial registries (eg, clinicaltrials.gov), so also it may be appropriate for the National Library of Medicine to do the same for this function.

Internet references become inaccessible over time, but archiving the Web page at publication can preserve the information the author intended to cite. It is suggested that journals require authors to archive URL references at article submission but, because of copyright restrictions and limitations of the archiving process, keep a printed copy of the Web page for those readers who later request this information.

*The authors acknowledge Mickey Murano, MA, for checking the URLs and maintaining the database and Tracy Napper, BA, for help with trial tests of the WebCite.*

*Supervising editor:* Donald M. Yealy, MD

Dr. Yealy was the supervising editor on this article. Dr. Schriger did not participate in the editorial review or decision to publish this article.

*Author contributions:* DLS conceived the idea and performed the data collection. AWT and DLS designed the study, analyzed the data, and drafted the article. AWT takes responsibility for the paper as a whole.

*Funding and support:* By *Annals* policy, all authors are required to disclose any and all commercial, financial, and other relationships in any way related to the subject of this article that might create any potential conflict of interest. See the Manuscript Submission Agreement in this issue for examples of specific conflicts covered by this statement. Dr. Schriger is supported in part by a grant from the Korein Foundation. The sponsors fund the author's time for a number of projects. They have no involvement in the design, conduct, analysis, or publication of the research.

*Publication dates:* Received for publication October 25, 2010. Revision received November 11, 2010. Accepted for publication November 22, 2010.

Presented as an abstract at the Sixth International Congress on Peer Review and Biomedical Publication, September 2009, Vancouver, British Columbia, Canada.

Reprints not available from the authors.

*Address for correspondence:* Andrea W. Thorp, MD, Department of Emergency Medicine, Loma Linda University Medical Center and Children's Hospital, 11234 Anderson St, MC-A108, Loma Linda, CA 92354; 909-558-7698, fax 909-558-0121; E-mail athorp@llu.edu.

## REFERENCES

1. Thorp AW, Brown L. Accessibility of Internet references in *Annals of Emergency Medicine*: is it time to require archiving? *Ann Emerg Med*. 2007;50:188-192.
2. Wren JD. URL delay in MEDLINE: a 4-year follow-up study. *Bioinformatics*. 2008;24:1381-1385.
3. Wren JD, Johnson KR, Crockett DM, et al. Uniform resource locator decay in dermatology journals: author attitudes and preservation practices. *Arch Dermatol*. 2006;142:1147-1152.
4. Carnevale RJ, Aronsky D. The life and death of URLs in five biomedical informatics journals. *Int J Med Inform*. 2007;76:269-273.
5. Falagas ME, Karveli EA, Tritsaroli VI. The risk of using the Internet as reference resource: a comparative study. *Int J Med Inform*. 2008;77:280-286.
6. Wagner C, Gebremichael MD, Taylor MK, et al. Disappearing act: decay of uniform resource locators in health care management journals. *J Med Libr Assoc*. 2009;97:122-123.
7. Dellavalle RP, Hester EJ, Heilig LF, et al. Information science. Going, going, gone: lost Internet references. *Science*. 2003;302:787-788.
8. Wren JD. 404 Not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics*. 2004;20:668-672.
9. Carnevale RJ, Aronsky D. The life and death of URLs in five biomedical informatics journals. *AMIA Annu Symp Proc*. 2005;912.
10. Ducut E, Liu F, Fontelo P. An update on uniform resource locator (URL) decay in MEDLINE abstracts and measures for its mitigation. *BMC Med Inform Decis*. 2008;8:23. Available at: http://www.biomedcentral.com/1472-6947/8/23. Accessed December 10, 2010.
11. Aronsky D, Madani S, Carnevale RJ, et al. The prevalence and inaccessibility of Internet references in the biomedical literature at the time of publication. *J Am Med Inform Assoc*. 2007;14:232-234.
12. Veronin MA. Where are they now? a case study of health-related Web site attrition. *J Med Internet Res*. 2002;4:E10.
13. Eysenbach G. Going, going, still there: using the WebCite service to permanently archive cited Web pages. *J Med Internet Res*. 2005;7:e60.
14. WebCite. WebCite Information web page. Available at: http://www.webcitation.org/faq. Accessed October 25, 2010.