INFORMATION SCIENCE

Going, Going, Gone: Lost Internet References

Robert P. Dellavalle,[1,2,3]* Eric J. Hester,[2] Lauren F. Heilig,[2] Amanda L. Drake,[2] Jeff W. Kuntzman,[4] Marla Graber,[4] Lisa M. Schilling[5]

Internet references in medical and scientific periodicals may become more common as 7 million pages of new information, including data not available elsewhere, appear daily on the World Wide Web (WWW) (*1*). The Internet, of which the Web is part, consists of a worldwide system of computer networks. The Internet promotes easy access to and revision of data and allows information formats not suitable for print media including high-resolution images, motion video, animations, simulations, and program source code. However, unlike hard copy references, Internet references may change and become inaccessible (*2*–*6*).

Nearly 20% of Internet addresses in a Web-rich high school science curriculum became inactive between August 2002 and March 2003. Addresses with ".com" and ".edu" top-level domains most frequently became inactive (*6*, *7*). In another study, 108 of 184 Internet addresses for an herbal remedy, Opuntia, became inactive within 4 years (*8*). Furthermore, no consensus on Internet reference format exists, especially with regard to providing accession (or citation) dates that report when authors viewed the information online (*9*–*11*).

Methods

Our study examined the frequency, format, and activity of Internet references in three high-circulation U.S. journals with scientific impact in the top 1% of all journals as rated by the Institute for Scientific Information (ISI) Journal Citation Reports (Science Edition of 2001) (*12*): *NEJM* (*New England Journal of Medicine*), *JAMA* (*The Journal of the American Medical Association*), and *Science* (*13*). Sampled issues of each journal were published during the 6-week period immediately antecedent to the initiation of the research on 16 January 2003. For comparison, issues from the same calendar period, 1 December to 15 January, for the preceding 2 years were also examined. A reference was defined as a numbered citation appearing at the end of an article. Advertisements and articles without references were excluded. Internet reference categories were (i) active Internet reference—Internet address citation yielding information other than an error message when accessed via an Internet browser (e.g., Internet Explorer) and (ii) inactive Internet reference—Internet address citation yielding an error message when accessed via an Internet browser.

Each inactive Internet reference was further categorized, by using www.archive.org and www.google.com, as a (i) recoverable Internet reference—inactive Internet address citation yielding recoverable information with Internet archiving systems or (ii) unrecoverable Internet reference—inactive Internet address yielding no recoverable information. The type of referenced Internet material (e.g., PDF, unpublished document, or conference proceedings) was recorded. A comparison was made between these recovered inactive Web sites and a random sample of 30 active Web sites.

Internet References

Thirty percent of articles contained one or more Internet references. Overall 2.6% of all references were Internet references. The highest density appeared in review articles (1.9 Internet references/article) and the lowest density was in letters (0.3). Internet referencing increased over a 27-month period in the medical journals (*NEJM*, 0.6 to 3.0%; *JAMA*, 2.1 to 3.2%) and decreased in *Science* from 4.2 to 1.7%. The number of articles with at least one Internet reference increased for *NEJM* (10 to 19%) and *JAMA* (21 to 42%) and declined for *Science* (47 to 21%).

Internet Reference Activity

The percentage of inactive Internet references increased from 3.8% at 3 months to 10% at 15 months and to 13% at 27 months after publication (*13*) (fig. S1). For articles 27 months old, *JAMA* had the greatest Internet reference inactivity (21%) compared with *NEJM* (13%) and *Science* (11%).

Inactive Internet references were most commonly ".com" addresses (46% lost after 27 months) followed by ".edu" (30%), other (20%), ".gov" (10%) and ".org" (5%) (see table). Book reviews had the greatest loss (17%) and opinion and news articles, the least (8%) (*13*) (table S1).

In contrast to *JAMA* and *NEJM*, some articles in *Science* contained Internet references directing readers to *Science's* Web sites for supplemental material not published in the hard copy journal. These self-references accounted for approximately 1% of Internet references in *Science*. Excluding self-references raised the loss of Internet references in Science's 27-month-old articles from 11.0 to 11.5%.

In September 2003, *NEJM*'s and *JAMA*'s, but not *Science*'s, instructions to authors requested accession dates for an Internet reference. Over the study period, accession dates became more common in *NEJM* (0 to 93% of Internet reference citations), less common in *JAMA* (100 to 73%), and changed minimally in *Science* (1 to 0%).

Although our study may not be representative of the entire field (*13*), Internet references occurred frequently and were often inaccessible within months after publication in the highest-impact U.S. medical and scientific journals. The problem of impermanent Internet references calls for an immediate response. It arises because Internet addresses, also known as Uniform Resource Locators (URLs) (*14*), serve as both the identifier (name) and the location (address) of the information. A URL might disappear for innumerable reasons such as servers shutting down because of business failure or URL reconfiguration so that the content is reached via a different logical pathway. Policies and resources for addressing URL maintenance may be addressed from multiple perspectives.

The simplest publisher response to Internet reference loss would be to ban or severely restrict Internet reference use. Although examples of journals prohibiting the use of URLs in endnotes do exist (e.g., *Cancer Research*), the vast majority of medical and scientific journals, including those in this study, currently have no such restrictions. Pro-

[1]Department of Veterans Affairs Medical Center, Denver, CO, 80220, USA. [2]Department of Dermatology, [3]Department of Preventive Medicine and Biometrics, [4]Denison Memorial Library, [5]Department of Medicine, University of Colorado Health Sciences Center, Denver, CO, 80262, USA.

*To whom correspondence should be addressed. E-mail: robert.dellavalle@uchsc.edu

hibiting URLs is undesirable given that rapidly developing fields, such as human genetics, rely on extensive information available only on the Internet. For reference content available in both published print and Internet form, publishers might call for citation of both. Publishers might also call upon authors to obtain appropriate copyright permissions so they can submit all Internet reference information with manuscript submission. However, this policy change to promote author and publisher electronic content preservation might not be feasible owing to copyright laws, operational limitations, and other considerations (*15*).

Editors and publishers might implement new systems for identifying online resources such as Digital Object Identifiers (DOIs) (*16*). A DOI is a permanent, unique identifier [similar to an Universal Product Code (UPC) bar code] that can be incorporated directly into a URL. Many print and online publishers have adopted the DOI system and participate via the nonprofit organization Cross Ref (www.crossref.org). However, a barrier to DOI acceptance is the cost for small publishers and individual authors (*17*, *18*).

Alternative solutions to the DOI system include the Uniform Resource Name syntax (URN) and the Persistent Uniform Resource Locator (PURL) (*19*–*21*). Like a DOI, these identifier systems also await a "critical mass" of acceptance within the online publishing community.

The Library of Congress might reconsider digital preservation options with the Open Archival Information System (OAIS) (*22*). The quantity of information involved renders it unrealistic currently to expect the Library of Congress to preserve all lost Internet information. Still, it is hoped that Internet information cited in peer-reviewed, high-impact journals will receive priority in preservation efforts.

Resources for readers encountering inactive Internet references include Internet Archive (www.archive.org), a nonprofit organization seeking to provide permanent access to Internet information; and Internet search engines such as Google (www.google.com), which cache Web pages. Internet Archive provided recoverable information for 31 of the 60 inactive Internet references; Google provided recoverable information for two. Readers, however, cannot be assured that the information captured by Internet Archive, or Google, or even an active URL is unchanged compared with the information originally captured by the authors.

Another option for obtaining lost Internet references is correspondence with authors. Successful correspondence requires that authors have (i) the original Internet information, (ii) the means to provide the information, and (iii) the desire to share the information with the requestor. We believe that the best current solution to improve access to Internet references is for publishers to require capture and submission of all Internet information at the time of manuscript consideration. Although this policy change should facilitate access to the content of lost URLs, it will not prevent URLs from becoming inactive or migrating. To slow the rate of the introductions of inactive Internet addresses into the literature, authors should also be required to provide accession dates and to verify Internet reference activity immediately before publication. Publishers, authors, and libraries must adopt better Internet reference policies and archiving strategies to limit the loss of Internet reference information in current medical and scientific literature.

ARTICLE AND REFERENCE CHARACTERISTICS

	NEJM	*JAMA*	*Science*
Items reviewed			
Journal issues	18	18	16
Articles	290	300	472
Total refs.	6,704	7,045	11,799
Internet refs.	97	189	386
Articles with 1 or more Internet refs.	39	93	190
Internet refs. per article, mean*	2.49 (1.35, 3.63)	2.03 (1.74, 2.32)	2.03 (1.63, 2.43)
Top-level domain			
.gov	41	103	111
.org	37	46	162
.com	6	17	14
.edu	4	8	47
.net	1	3	2
Other	8	12	50
Inactive Internet refs./all Internet refs.			
Age of article in months			
2 to 3	1/71	4/79	3/65
14 to 15	2/11	13/68	7/148
26 to 27	2/15	9/42	19/173

*Values in parentheses are 95% confidence intervals.

References and Notes

1. "Preserving our digital heritage: Plan for the National Digital Information Infrastructure and Preservation Program" (Library of Congress, Washington, DC, 2002); available at www.digitalpreservation.gov/repor/ndiipp_plan.pdf.
2. D. Fichter, *Online* **23**, 77 (1999).
3. B. Machrone, *PC Week* **16**, 81 (1999).
4. D. Pearson, *Health Info. Libr. J.* **18**, 139 (2001).
5. J. E. Grogg, *Searcher* **10**, 57 (2002).
6. J. Markwell and D. W. Brooks, *Biochem. Mol. Biol. Educ.* **31**, 69 (2003).
7. J. Markwell, and D. W. Brooks, *J. Sci. Educ. Technol.* **11**, 105 (2002).
8. M. A. Veronin, *J. Med. Internet Res.* **4**, E10 (2002); available at www.jmir.org/2002/2/e10/index.htm.
9. Institute for Scientific Information (ISI), *Journal Citation Reports: JCR Science Edition* (Institute for Scientific Information, Philadelphia, 2001).
10. Staff of the University of Chicago Press, *The Chicago Manual of Style* (Univ. of Chicago Press, Chicago, ed. 15, 2003), p. 646.
11. M. Dee, Ed., "Quote, unquote: The Harvard style of referencing published material" (Leeds Metropolitan Univ., Leeds, UK, version 2, 1998); available at www.lmu.ac.uk/lss/ls/docs/harvfron.htm].
12. Purdue University Online Writing Lab, "Using Modern Language Association (MLA) format" (Purdue Univ., Purdue, IN, 2003); available at http://owl.english.purdue.edu/handouts/research/r_mla.html, also in the *MLA Style Manual and Guide to Scholarly Publishing* (Modern Language Association, New York, ed. 2, 1998).
13. See Supporting Online Material at *Science* Online.
14. *Miriam-Webster Dictionary*; available at http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=url.
15. S. Olsen, *CNET News.com*, 9 July 2003; available at http://news.com.com/2100-1038_3-1024234.html.
16. The Digital Object Identifier System. International DOI Foundation, 2003); available at http://www.doi.org.
17. P. Caplan, *Public-Access Comput. Syst. Rev.* **9**(1), 1998); available at http://info.lib.uh.edu/pr/v9/n1/capl9n1.html
18. P. Jacsó, *Inf. Today* **19**, 30 (2002).
19. Persistent URL Home Page (OCLC PURL Service, OCLC Online Computer Library Center, Dublin, OH, 2003); available at http://purl.oclc.org/
20. K. E. Shafer, S. L. Weibel, E. Jul, *J. Libr. Admin.* **34**, 123 (2001).
21. World Wide Web Consortium (W3C), Naming and addressing: URIs, URLs, ...; available at http://www.w3.org/Addressing/
22. See (*1*), p. 37.

23. We thank D. Page, M. Degitis, M. Hrenya-Wood, C. Grey, M. Murray, and M. Wilson for helping conduct this study; and D. Norris and W. Weston for stimulating discussion. R.P.D. was supported by the National Cancer Institute grant K-07 CA92550, and E.J.H. was supported by National Institutes of Health grant T32 AR07411. Supported in part by NCI Cancer Education Grant R25 CA49981.

Supporting Online Material

www.sciencemag.org/cgi/content/full/302/5646/787/DC1
Materials and Methods
SOM Text
Fig. S1
Table S1