Research Paper ■

The Prevalence and Inaccessibility of Internet References in the Biomedical Literature at the Time of Publication

DOMINIK ARONSKY, MD, PHD, SINA MADANI, MD, RANDY J. CARNEVALE, BS, STEPHANY DUDA, MS, MICHAEL T. FEYDER, BS

Abstract **Objectives:** To determine the prevalence and inaccessibility of Internet references in the bibliography of biomedical publications when first released in PubMed®.

Methods: During a one-month observational study period (Feb 21 to Mar 21, 2006) the Internet citations from a 20% random sample of all forthcoming publications released in PubMed during the previous day were identified. Attempts to access the referenced Internet citations were completed within one day and inaccessible Internet citations were recorded.

Results: The study included 4,699 publications from 844 different journals. Among the 141,845 references there were 840 (0.6%) Internet citations. One or more Internet references were cited in 403 (8.6%) articles. From the 840 Internet references, 11.9% were already inaccessible within two days after an article's release to the public.

Conclusion: The prevalence of Internet citations in journals included in PubMed is small (<1%); however, the inaccessibility rate at the time of publication is considered substantial. Authors, editors, and publishers need to take responsibility for providing accurate and accessible Internet references.

■ **J Am Med Inform Assoc.** 2007;14:232–234. DOI 10.1197/jamia.M2243.

Introduction

References in scientific publications are an important resource for interested readers to access works that are related to a study. For articles that are traditionally published in hard copy the accuracy of bibliographic references varies considerably and ranges from 3% to 60% among general and specialist biomedical journals.[1–5] As the research community continues to take advantage of the wealth of information resources, the use of Internet references in scholarly work is increasing.

The Internet is a dynamic web of computer networks that provides rapid access to information and promotes accelerated knowledge dissemination. The Internet expands the limited presentation capabilities of printed publications and facilitates access to audio, video, animated graphics, interactive web sites, databases, source code, and executable programs. Uniform Record Locators (URLs) are Internet addresses that provide the reader with a unique reference to online information. Given the transient nature of the Internet and in the absence of a permanent digital library,[6] citing URLs has the disadvantage that accessibility and content stability are not guaranteed.[7]

The gradual decay rate of URLs is a growing threat to scientific communication.[6–15] In three high-impact journals inactive URLs listed in the reference section of an article increased from 3.7% after three months to 13% after 27 months.[6] In six leading medical journals the rate of inaccurate or inaccessible URL references ranged from 0% to 22% after three months of an article's publication.[10] Inaccessibility rates for online supplementary material in six leading scientific journals increased from 4.7% after two years to 9.6% after five years of an article's initial publication.[11] In a study that examined the URLs listed in MEDLINE® abstracts, 12% were incorrectly formatted or misspelled.[12] After corrections, 18.6% of all URLs remained inaccessible, while an additional 18.8% were intermittently accessible. In specialist journals the rates of non functional URLs in oncology articles increased from 9.5% after five months to 33% after five years of the initial publication,[13] in dermatology articles from 10.9% within nine months to 34.6% after five years,[14] and in HIV/AIDS articles from 21.3% after one year to 41.7% after four years.[15] In biomedical informatics journals an average of 21.9% of bibliographic URL references were not functional and 8.9% were only intermittently accessible.[16]

Previous studies of URL reference accessibility in the biomedical domain examined references in small groups of general or specialist journals and measured frequency and inaccessibility rates after variable time intervals since an article's initial publication. This study assessed the overall rate of Internet

Affiliations of the authors: Department of Biomedical Informatics (DA, SM, RJC, SD, MTF); Department of Emergency Medicine (DA), Vanderbilt University, Nashville, TN.

The 3rd and 4th authors (RJC, SD) were supported by a Training Grant from the National Library of Medicine (T15 007450-03).

Correspondence and reprints: Dominik Aronsky, M.D., Ph.D., Department of Biomedical Informatics, Eskind Biomedical Library, Vanderbilt University Medical Center, 2209 Garland Ave., Nashville, TN 37232-8340; e-mail: <dominik.aronsky@vanderbilt.edu>.

Received for review: 8/07/2006; accepted for publication: 12/12/2006.

Table 1 ■ Characteristics of Articles and Internet References

Characteristic	Items Reviewed
Publications	3,757
Journals	844
References	141,845
References per article, mean (range)	30.2 (1–356)
URL references, unique	840
Publications with one or more URLs	403
URL references per article, mean (range)	2.1 (0–18)
Top-level domain	
.com	183
.edu	87
.gov	122
.int	38
.net	21
.org	218
other	171

citations and the frequency of inaccessible URLs in articles in all biomedical journals included in PubMed® at the time of their initial release to the public.

Methods

PubMed is the U.S. National Library of Medicine's database of biomedical citations and abstracts. MEDLINE is PubMed's largest component indexing the articles of over 4,800 journals.[17] Our study examined the frequency and accessibility of Internet references in the bibliography of publications within two days of their initial release in PubMed. A publication's release in PubMed marks the first point in time when an article becomes accessible to the general public and can appear in a user's PubMed query.

During a one-month observational period (Feb 21 to Mar 21, 2006) we performed a daily download of all citations that were released in PubMed the previous day. To retrieve articles that were released in PubMed each day we used the query term *pubstatusaheadofprint AND yyyy/mm/dd[edat]*. The "pubstatusaheadofprint" term indicates that an article was made available on a Web site of a publisher or provider, and was also submitted for inclusion in PubMed.[18] The "yyyy/mm/dd[edat]" term indicates the date (year, month, day) when the citation was added to the PubMed database.

For a 20% random sample (Microsoft® Excel random generator) we obtained the full-text electronic article through our university libraries, which subscribe to more than 15,300 electronic journals, and extracted all bibliographic references. If an electronic publication was unavailable or did not include any references, it was replaced by another randomly selected reference. During article retrieval, 46 (1.2%) of the 3,757 articles were replaced because the library did not have a subscription to the journal and five (0.1%) were replaced because the third party provider did not list ahead-of-print articles. All references cited in the bibliography were copied and pasted into a spreadsheet.

A computer program that included a set of scripts identified the references containing a URL.[16] Identifying URLs in references was based on the presence of Internet protocol terms (ftp, http, //), common URL substrings (www, htm, pdf), top-level domains (com, edu, gov, mil, net, org), and other terms commonly found in URL references (available, accessed, download). The resulting list was manually reviewed to eliminate references that contained one or more of the above terms but not a true URL. We removed blank spaces within a URL string, which, for example, can occur when copying and pasting text from a PDF source file, a common format for publishing forthcoming articles. We did not correct other typographical errors in the URL string. Duplicate URL references within the same day were eliminated. URLs without a protocol term were prefixed with the protocol term "http://," as current browsers automatically do.

The URL list was submitted to GNU Wget,[19] a Web crawler, which attempted to access each Web site from two different networks (a university and a commercial Internet Service Provider network). The HyperText Transfer Protocol (HTTP) return codes[20] were recorded for all access attempts. Each day, two authors (RJC, DA) manually checked the Web sites that were inaccessible or had timed out after 30 seconds from two independent networks using two different Web browsers (Mozilla® Firefox® 1.5 and Microsoft® Internet Explorer™ 6.0). If at least one access attempt resolved in a successful web page viewing, the URL was defined functional. All initial URL access attempts and manual verifications were completed within two days of an article's appearance in PubMed.

Results

We found 840 URL references (0.6%) among 141,845 bibliographic citations collected from 4,699 publications in 844 different journals. One or more URL references were cited in 403 (8.6%) publications (max: 18 URLs). Table 1 displays the article characteristics, URL occurrences, and inaccessibility rates.

A total of 100 (11.9%) URL references were nonfunctional on the second day after PubMed appearance (Table 2). The two most frequent reasons for a nonfunctional URL accounted for 91% of inaccessible URLs and included 57 URLs that were not found and 34 URLs that timed out after 30 seconds.

Discussion

Inaccuracies in references to printed publications of leading journals ranged from 19% in 1977 to 26.5% in 1999.[1,3] Despite the capabilities of today's information technology, we found that the rate of inaccessible and inaccurate URL references at the time of publication is high (11.9%) and only moderately lower than the rate of inaccurate references found three decades ago. Inaccessible Internet references may be the result of spelling or typesetting errors, or the hosting Web site may have moved, requires user authentication, be restricted to an intranet, or have vanished for good.

References to printed publications include author names, title, journal, year, volume, and page numbers. Inaccuracies

Table 2 ■ Reason for 100 Inaccessible URL References

Reason (HTTP Error Return Code)	URL References
"Not Found" (404)	57
"Forbidden" (403)	4
"Internal Server Error" (500)	2
"Method Not Allowed" (405)	2
"Unauthorized" (401)	1
Time out after 30 seconds	34

in one element of the references in printed publications may be little more than an inconvenience to the reader, as the correct source can frequently be identified using other elements of the reference. URL references, however, consist of one character string with limited structure. Identifying the correct URL address string for non-functional URLs can be challenging, time-consuming, and frustrating. URL references following the format described in the Uniform Requirements for Manuscripts Submitted to Biomedical Journals (International Committee of Medical Journal Editors),[21] include additional elements about the information source. If readers encounter a nonfunctional URL, they may take advantage of the additional elements to locate the source through other means, such as a search engine. Currently the format and details provided in URL references, however, vary from providing the URL only to providing the full reference in the recommended format. Depending on the supplied information the amount of effort required to locate the referenced source may vary and may not lead to a successful retrieval of the information.

Our study did not examine whether information referenced from an Internet Web site was temporarily inaccessible or disappeared permanently. Our study simulated the case of a user performing a PubMed query at a certain point in time. In this case a URL is accessible or not, and the underlying reason for the unavailability of the referenced information becomes less relevant. In addition, our study did not correct misspelled URLs. Some misspelled URLs can include an error that can be corrected easily; others, however, are challenging or even impossible to identify, which makes a clear differentiation between inaccurate and inaccessible URLs difficult. As the general public increasingly accesses information from the scientific biomedical literature, we should not expect this audience to have the skills and knowledge to identify and correct misspelled URLs.

A generally accepted solution for a permanent digital repository may provide some relief for vanishing Web sites;[6,22] it cannot, however, address the correctness of URL references, which provide the unique key to access the referenced information resource. As the research community takes full advantage of the various electronic information modalities, it will become even more important that authors, editorial offices, and publishers pay careful attention to verify the accuracy and accessibility of URL references, including means for a permanent digital library that will guarantee accessibility and content stability for many years to come.

References ■

1. Poyer RK. Inaccurate references in significant journals of science. Bull Med Libr Assoc. 1979;67:396–8.
2. de Lacey G, Record C, Wade J. How accurate are quotations and references in medical journals? Br Med J (Clin Res Ed). 1985; 291:884–6.
3. Siebers R, Holt S. Accuracy of references in five leading medical journals. Lancet. 2000;356:1445.
4. Evans JT, Nadjari HI, Burchell SA. Quotational and reference accuracy in surgical journals. A continuing peer review problem. JAMA. 1990;263:1353–4.
5. Aronsky D, Ransom J, Robinson K. Accuracy of references in five biomedical informatics journals. J Am Med Inform Assoc. 2005;12:225–8.
6. Dellavalle RP, Hester EJ, Heilig LF, Drake AL, Kuntzman JW, Graber M, et al. Information science. Going, going, gone: lost Internet references. Science. 2003;302:787–8.
7. Koehler W. A longitudinal study of web pages continued: a report after six years, Information Research [serial on the Internet], 9 (2), paper 174 (2004). Available at: http://informationr.net/ir/9-2/paper174.html. Accessed December 10, 2006.
8. Koehler W. Web page change and persistence—a four-year longitudinal study. J Am Soc Inf Sci Technol. 2002;53(2):162–71.
9. Spinellis D. The decay and failures of web references. Commun ACM. 2003;46:71–7.
10. Crichlow R, Winbush N, Davies S. Accessibility and accuracy of web page references in 5 major medical journals. JAMA. 2004; 292:2723–4.
11. Evangelou E, Trikalinos TA, Ioannidis JP. Unavailability of online supplementary scientific information from articles published in major journals. FASEB J. 2005;19:1943–4.
12. Wren JD. 404 not found: the stability and persistence of URLs published in MEDLINE. Bioinformatics. 2004;20:668–72.
13. Hester EJ, Heilig LF, Drake AL, Johnson KR, Vu CT, Schilling LM, et al. Internet citations in oncology journals: a vanishing resource? J Natl Cancer Inst. 2004;96:969–71.
14. Wren JD, Johnson KR, Crockett DM, Heilig LF, Schilling LM, Dellavalle RP. Uniform resource locator decay in dermatology journals: author attitudes and preservation practices. Arch Dermatol. 2006;142:1147–52.
15. Olfson E, Laurence J. Accessibility and longevity of Internet citations in a clinical AIDS journal. AIDS Patient Care STDS 2005;19:5–8.
16. Carnevale RJ, Aronsky D. The life and death of URLs in five biomedical informatics journals. Int J Med Inform. 2006 Jan 30; [Epub ahead of print].
17. PubMed® MEDLINE® [database on the Internet]. US National Library of Medicine. Available at: http://www.pubmed.org. Accessed December 9, 2006.
18. PubMed® MEDLINE®. National Library of Medicine. Available at: http://www.ncbi.nlm.nih.gov/entrez/query/static/aheadofprint.html. Accessed December 10, 2006.
19. GNU Wget. Available at: http://www.gnu.org/software/wget/. Accessed December 10, 2006.
20. The Internet Engineering Task Force [homepage on the Internet]. Network Working Group: Hypertext Transfer Protocol HTTP/1.1. Available at: http://www.ietf.org/rfc/rfc2616.txt. Accessed December 10, 2006.
21. International Committee of Medical Journal Editors, Uniform requirements for manuscripts submitted to biomedical journals: sample references. Available at: http://www.nlm.nih.gov/bsd/uniform_requirements.html. Accessed December 10, 2006.
22. Eysenbach G, Trudel M. Going, going, still there: using the WebCite service to permanently archive cited web pages. J Med Internet Res. 2005;7:e60.