Exploring the Half-life of Internet Footnotes

Michael Bugeja and Daniela V. Dimitrova

Vanishing online references are becoming a problem for scholars. This exploratory study examines use of online citations, focusing on 2003 AEJMC conference papers accepted by the Communication Technology and Policy division. Authors analyze papers using URL reference addresses in bibliographies and document some 40% of online citations being unavailable a year later. Results show that .edu *is the most stable domain. Error messages for "dead" URL addresses also are explored. Finally authors offer much needed recommendations for researchers who use Internet citations.*

In assembling manuscripts in our area of expertise—media and new technologies—we often must cite dozens of Internet-based sources in reference pages. Anecdotal evidence shows that Internet references disappear, thus undermining the replicability and reliability of scholarly research. Some URLs lapse, some are archived under different addresses, and some simply vanish. We call this phenomenon the "half-life of Internet footnotes" because such footnotes degrade over time much like isotopes of an atom. The operative word, "half-life," is more than symbolic, however.[1] Half-life indicates a measurable interval when half of the footnotes cited in a given article lapse, vanish, or require revision to operate correctly.

Case in point: In December 2001, Microsoft Corporation displayed a values statement on "community" found at: http://www.microsoft.com/mscorp/values.htm. When accessed in 2003, visitors to that address were relocated to this one: http://www.microsoft.com/mscorp/. True, there was a link to Microsoft's "mission and values"; but the value associated with community had changed with the link.

The original citation stated:

> Microsoft and its employees recognize that we have the responsibility, and opportunity, to contribute to the communities in which we live, in ways that make a meaningful difference to people's lives.

The revised value on community read:

> Thinking and acting globally, enabling a diverse workforce that generates innovative decision-making

MIchael Bugeja (Ph.D., Oklahoma State, 1985) is a Professor and Director of the Greenlee School of Journalism and Communication, Iowa State University. Daniela Dimitrova (Ph.D., University of Florida, 2003) is an Assistant Professor at the Greenlee School of Journalism and Communication, Iowa State University. Correspondence should be addressed to the first author at bugeja@iastate.edu

> for a broad spectrum of customers and partners, innovating to lower the costs of technology, and showing leadership in supporting the communities in which we work and live. (http://www.microsoft.com/mscorp/mission)

When fact-checked in October 2003, the only URL on which the original values statement could be found was at an Australian telecommunications site (http://www.span.net.au/s04p41m18.htm).

This was disconcerting. An author quoting the 2001 values statement would seem to have "invented" the text a mere two years later when the new Microsoft "values statement"—bearing little resemblance to the original—would appear on Web browsers. Given that and other phenomena (disappearing URLs, incorrectly formatted URLs, etc.), how could scholars of the future rely on Internet citations if they had a half-life? What, indeed, was the length of time required for half of Web-based citations to decay like atoms of an isotope? What could be done to ensure that editors had copies of URLs cited in a scholarly manuscript? These were the questions that this exploratory study addressed, in the hope of discovering preliminary patterns of online citation behavior. We believe the half-life of Internet footnotes is important because no method exists, as far as we know, to recommend an archiving process for future scholars. Much of the previous work on the erosion of Internet footnotes has been done by library scientists rather than by journalism and communication educators who not only are expected to use the medium as scholars but also are responsible in large part for its diffusion into academe and beyond.

Literature Review

The Internet has been touted as opening up new opportunities for communication and information. Scholars using the Internet find sources quickly and easily, visiting library, scientific and other databases and Web sites in seconds. But fast, convenient information retrieval also has downside: citations can disappear without a trace. Several studies in library and information science in particular have examined the longevity behavior of Web sites and documented the problem of inaccessible online information (Casserly & Bird, 2003; Germain, 2000; Natriello, 1997). A 1997 study by Gary Natriello in *Teachers College Record* emphasizes two key components in citing sources in scholarly work: attribution and access. The Internet can undermine both. For instance, some Web content does not clearly designate an author. Sometimes researchers have to attribute a Web page to a "Web master." Also, there are other significant challenges with accessing online sources. The Natriello article discusses basic elements of "information integrity"—including "reference," which he defines as the ability to "locate and access electronic sources consis-

tently over time." The World Wide Web is a dynamic medium, transforming the very notion of "reference": Citations over time often cease to exist.

The literature on the topic of Internet citation availability (also known as "link rot") and online information retrieval is significant, but not entirely focused. While researchers have observed this issue, none of these studies has focused significantly on the half-life phenomenon. A sampling of such research reveals the following key findings. First, hyperlinks are unstable and disappear over time. McMillan (2001), for example, examined health-related Web sites from 1997 to 2000. She found that 27% of these sites were gone three years later. Sites created by individuals were more likely to vanish while governmental and educational sites were found to be more stable (McMillan, 2001). Germain (2000) focused on 31 randomly chosen journal articles and found that half of their URL citations were inaccessible after a three-year period. She also identified a troubling trend – the number of articles containing inaccessible citations increased from 38 percent to 68 percent during that period. Library scientists Casserly and Bird (2003) tracked down the availability of online citations in scholarly articles from 1999 to 2000. They found that only 56 percent of the citations were permanent, 81 percent were located after additional Web searches, and 89 percent were available in the Internet Archive. Online citations seem to be a problem even for medical journals. A recent study of the use of Internet references in three leading medical journals found that 13 percent of the links cited were inactive just 27 months later (Dellavalle et al., 2003). But none of these studies focused sufficiently on the cumulative impact on scholarship as the half-life phenomenon intensifies over time, with authors suggesting effective technology-based methods to address the problem.

In 1999, one of the first comprehensive studies of Web site longevity was published by Wallace Koehler in the *Journal of American Society for Information Science*. He argued in "An analysis of Web age and Web site constancy and permanence" that Web pages in general exhibit two types of longevity behavior: 1) constancy and 2) permanence. Koehler defined constancy as the notion whether or not a Web page carries the same content over time. The second characteristic, permanence of Web pages, was defined as whether Web documents carry the same URL (Web address) over time. In a later four-year longitudinal study, published in the same journal in 2002, Koehler concluded that the half-life of Web pages equaled two years. We came across a few other studies that indicated different half-life rates in scholarly and medical journals. This was precisely what we were searching for in our own discipline. As far as we could discern, however, the impermanent nature of online citations used in journalism and communications research had not been examined as a fundamental aspect of scholarship.

Hence, one of the goals of our research was to identify and examine the issue of URL citation failure in journalism and communication conference papers and to ascertain how scholars, in the short-term, may overcome them. We believe that the Association for Education in Journalism and Mass Communication—in particular AEJMC's Communication Technology and Policy division—has an ethical obligation to explore these issues, especially as the division seeks to utilize the Internet more effectively, from submission of papers to preparation of scholarly research. Moreover, we believe this will be a trend in other communication organizations, including the National Communication Association and the International Communication Association.

Methodology

A content analysis methodology was used to explore the issues of vanishing Internet references (Krippendorf, 1980). We focused on 2003 AEJMC conference papers available through an online archive at Michigan State University (http://list.msu.edu/archives/aejmc.html). The Communication Technology and Policy (CTP) division was chosen for this analysis for two reasons: first, a preliminary examination indicated that it was the division with most frequent use of Internet footnotes; second, CTP was also the first division to go to a completely online submission process this year in preparation for the 87th annual convention in Toronto, August 2004. Admittedly, this is a convenience sample, merely sufficient for raising and exploring the issue; however, a convenience sample is appropriate for a specific reason: If footnotes are lapsing in the most *current* online papers available through AEJMC, it only follows that adding more entries from previous years, to increase the sample, would result in higher percentages of the half-life phenomenon. In other words, if significant footnote decay was found in the most current available sample, that data alone would verify the severity of the problem. Further, this sample was chosen because the CTP papers focus on technology and Internet-related topics, and as a result, scholars in this division are more likely to use Internet sources.

We trained a graduate student coder who accessed all CTP division papers from AEJMC 2003 available in the online archive and recorded each Internet source cited. The URL was then coded for a number of variables identified in prior research (Casserly & Bird, 2003; Koehler, 1999). The main variable of interest was if the URL cited was still active or not. This was coded as a nominal Yes/No variable. The coder was instructed to access the link first by clicking on it and, if that did not work, by pasting the link into the browser. Each Internet source was also coded for: top-level domain of the URL; error message for "dead" footnotes; whether the content matched the footnote; whether the footnote provided a retrieval date; and presence of a "%" sign in the URL. Additionally, the code deleted "%" signs or spaces when those were present

and tested again if the link worked or not. An intercoder reliability check was conducted on five percent of the URLs, using Holsti's formula[2], resulting in .94 agreement (Holsti, 1969). Perfect agreement was established for several categories such as top-level domain (TLD). Coding differences were reconciled at a training session.

Results

We found that 108 Internet citations were used in 2003 division papers. Of those 108 citations, only 55 (51%) worked when clicking on the link. That number increased to 65 (60%) when the URL address was pasted in the browser window by the coder. Of those links that worked, only 57% matched the content given in the reference. Some links went to the main home page of the organization as opposed to the page cited in the paper—for example, going to http://www.census.gov instead of a specific demographic report. Others linked to Web pages with completely different content.

In essence, half of the Internet footnotes cited less than a year ago were dead. If extra effort was made to copy and paste the link into the browser, the percentage of dead hyperlinks decreased to 40%. This number is comparable to the statistics given in previous studies (e.g., Germain, 2000). When papers are scrutinized individually, the number of dead hyperlinks often approaches or even surpasses the number of working hyperlinks. For example, more than half of the 30 Internet sources offered in one paper were dead.

The most frequent error message for a dead hyperlink was a message that the page was not found—sometimes explicitly saying "404 Error" or just "Page Not Found." Some sites expanded the message to "The requested document was not found" or "The requested page could not be found." Some attempts to get to Internet citations informed the researcher that they required an account, user name and password. Subscription requirements were not that common, but happened often enough to make them noteworthy.

A dead link hosted on the PBS Web site, for example, brought up the following message: "The file you requested is unavailable. If you got here by following a link on one of our pages, please send us an e-mail at newshour@pbs.org." The Poynter Institute was another common site linked to by authors. The Institute offered this message for missing pages: "The requested page cannot be found. Thanks for using Poynter Online. We recently redesigned our site and can't find the page you are looking for. Please try locating it using the site map below, and if you continue to have problems, contact us." A site directory similar to a table of contents was offered as recompense.

The Internet references used in our sample of conference papers linked to various top-level domains (TLDs), including country-code domains. Comparisons between the online citations show that there were

significant differences across top-level domains. The *.edu* domain emerged as the most stable domain for URLs in our sample, followed by *.com* and *.org*.

More than one-fifth of the URLs (21%) were not hyperlinked correctly: for instance, they contained a spelling error, a space in the URL address, missing "/" sign, or had a wrong file extension. Also, it is interesting to note that more than 10% of the URLs did not provide a retrieval date, despite explicit instructions to do so by the main citation style guides, such as the American Psychological Association Manual.

Discussion

Our results are consistent with previous research and confirm that there is reason for concern in citing Internet sources. We find it disconcerting that much previous research was not conducted by researchers in journalism and communication because our disciplines, traditionally, have upheld rigorous standards of accuracy, from quotations of sources to citations of scholarly works. Of particular concern in the Communication Technology and Policy division is the increasing use of online footnotes coupled with a substantial number of citations without retrieval dates, perhaps an early (and as yet untested) warning indicating poor scholarship methods resulting from lack of universal standards associated with easy initial Internet access but later difficult retrieval access. Even when retrieval is possible, some citations succumb to mere naming of "Web master," again indicating the absence of standards on what should or should not be included as reference in academic research.

Also associated with standards is the possibility of invention—fabricating citations and data. Without attention paid to rudimentary aspects of scholarship, in particular, the ability to access citations, our own discipline may unwittingly be contributing to the erosion of standards, inviting invention by less scrupulous practitioners. Consider the fact that substantial number of online citations was gone just after a year. URLs can be invented and then seem to have gone dead without the reliability of fact-check, especially troubling to journalism and communication researchers seeking to set basic standards in the discipline. While no incident of invention was uncovered in the papers analyzed in our study, several citations could not be found, indicating again that absence of guidelines allows at the minimum the opportunity for invention or carelessness, especially since others seeking to verify citations have no reliable method to identify and locate dead URLs. Worse, because those lapsed URLs can be cited in subsequent research, promulgation of inherent error is possible, undermining the very essence of replication that is at the core of the scientific method. Accuracy and access, as previously noted, are primary elements of dependable academic research without which

basic scholarship in every discipline—from master's theses to post-doctoral works—will cease to serve future generations as authentically as in the past.

We believe that Internet research is vital to scholarship because the medium serves as a convenient electronic warehouse of data accessible at all hours and in great quantities, thereby increasing the scope and breadth of scholarship. By no means are we discounting the medium's capacity to disseminate information quickly, affordably, and on demand. However, without conscientious gatekeepers of scholarship utilizing Internet, setting standards and methods to ensure accuracy and long-term access, the medium's drawbacks eventually will outweigh its benefits, especially as citation is concerned.

This outcome impacts research involving human subjects in the media and natural sciences. Lack of online citations can undermine replication and methodology, essential in these disciplines. While it is true that medical and scientific databases often store information able to be retrieved from electronic archives, our preliminary study also has shown that archiving, in and of itself, is at the root of termination of Internet footnotes. Social scientists also face the same dilemma as counterparts in medical and natural sciences because accuracy and scientific method are requisite standards of research in sociology, political science and communication disciplines, which Internet eventually may undermine without attention to what the medium facilitates and what it impedes. Finally, because ours is an exploratory study, we concede that alternative standards and methods may emerge as the half-life of Internet footnotes is analyzed more thoroughly in the short- and long-term. However, we also emphasize our study has garnered sufficient data to recommend interim methods that may help maintain the integrity of scholarship involving Internet.

Until universal standards can be implemented ensuring accuracy and access, we recommend citation when appropriate from journals and books rather than their online or electronic counterparts. The next best thing to the original journal article is the pdf version available through most university libraries. A pdf in many ways is a picture of the journal much like in previous years the facsimile was a picture. A pdf version, admittedly, may be altered electronically; however, that won't be a concern in a typical university databank. The authors do not recommend using online text or html versions of journal articles for citations as these formats are not representative pictures of journals, but merely reformatting of their contents in a different platform. Conversely, we acknowledge, this approach might hamper research in Internet-based studies. Thus, in future studies, we will investigate universal methods to safeguard standards in our discipline, especially as they relate to Internet footnotes and the extension of their half-life.

In the meantime we recommend printing two hard copies of Web citations (one for the author's files and one for the editor's files, to be sent upon request). We also recommend storing digital copies of source documents. (Web documents can be saved simply by using the "File>Save As" option in a browser.) This method not only makes available upon request citations by editors and researchers; it also safeguards authors against accusations of invention or inferior scholarship when URLs lapse.

We also considered an option that might have copyright implications—posting sections of cited materials in one author-generated URL per paper or study. This method has been in use in somewhat altered form by freelance writers querying editors electronically. To get an assignment, freelance writers typically provide clippings of past work. Because many editors prefer not to open attachments, writers seeking assignments post previously published articles on a Web site. Borrowing on the idea, researchers can assemble citation URLs for each paper, article or book, reprinting pertinent sections (paragraphs, tables, charts, etc., replete with proper citations and retrieval dates). This would enhance credibility even when Internet footnotes lapse. Granted, these sections also can be doctored or invented but not without risk, for the likelihood of discovery would be substantial as other researchers utilize search engines for original sources and happen upon the posted URL. It is important to note, however, that posting printed material online may impact publishing contracts and permission fees restricting such dissemination. Nonetheless, this method would become increasingly appropriate if associations and organizations would encourage such use, perhaps by hosting citation URLs in a public archive, setting the standard and a trend. In some sense, the reputations of associations such as AEJMC and NCA depend on the reliability of scholarship. And while this has not yet caused major replication problems as the half-life phenomenon is relatively new, the phenomenon can only worsen over time with potentially disastrous, cumulative effects unless proper attention is given to it.

Certainly, other alternative methods to secure online citations should be considered. Researchers in journalism and communication should investigate and propose methods as they understand the intricacies of Internet and how the medium alters messages and intentions—in this case, ones that impact the credibility of research across disciplines. In any case proper use of Internet should be a focus of undergraduate and graduate programs, scholarly associations, academic journals, and educational publishing—all of which, in the end, stand to lose credibility simply because the new medium cannot facilitate basic research.

Endotes

1 Wallace Koehler (1999) uses the notion of half-life in a similar fashion in his study of URLs.

2 Holsti's intercoder reliability (IR) formula was used as follows: $IR=2M/(N_1+N_2)$, where M equals the number of agreements between the coders, N_1 is the total number of coding decisions made by Coder 1 and N_2 is the total number of coding decisions made by Coder 2.

References

Casserly, M. F., & Bird, J. E. (2003). Web citation availability: Analysis and implications for scholarship. *College & Research Libraries*, 64(4), 300-317.

Dellavalle, R. P., Drake, A., Graber, M., Heilig, L., Hester, E., Kuntzman, J., & Schilling, L. (2003, October). Going, going, gone: Lost Internet references. *Science,* 302(5646), 787-788.

Germain, C. A. (2000). URLs: Uniform resource locators or unreliable resource locators. *College & Research Libraries*, 61(4), 359-365.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Koehler, W. (1999). An analysis of web page and web site constancy and permanence. *Journal of American Society for Information Science*, 50(2), 162-180.

Koehler, W. (2002). Web page change and persistence—a four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, 53(2), 162-171.

Krippendorf, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage.

Markwell, J., & Brooks, D.W. (2002). Broken links: The ephemeral nature of educational WWW hyperlinks. *Journal of Science Education and Technology*, 11(2), 105-108.

McMillan, S. (2001). Survival of the fittest online: A longitudinal study of health-related web sites. *Journal of Computer-Mediated Communication*, 6(3). Retrieved September 4, 2004 from http://www.ascusc.org/jcmc/vol6/issue3/mcmillan.html

Natriello, G. (1997). Attribution and access: Citing electronic sources. Teachers College Record, 98(3), 373-380.

Vaughan, L., & Shaw, D. (2003). Bibliographic and web citations: What is the difference? *Journal of the American Society for Information Science and Technology*, 54(14), 1313–1322.