RESEARCH NOTE

The Missing Link

Assessing the Reliability of Internet Citations in History Journals

EDMUND RUSSELL and JENNIFER KANE

One of the bedrock values of professional historians is reliance on verifiable documentation. We have well-developed methods for storing, citing, and finding many media, including books, journals, and archival material. Recently, historians have relied on a new medium, the internet, to document the sources of their information. Scholars in the sciences, however, have raised alarms about the frequency with which internet sources have disappeared after their citation in journals. An influential article in *Science* found that 13 percent of internet citations in three leading journals were inactive within twenty-seven months of publication. In five leading medical journals, 4.4 percent of internet citations were inaccessible within three months of publication. In six oncology journals, 33 percent of internet citations decayed within twenty-nine months.[1]

Edmund Russell is associate professor of science, technology, and society and history at the University of Virginia. Jennifer Kane graduated from the University of Virginia in 2007 with a major in social and political thought. She is now a legislative correspondent for U.S. Senator Tom Carper. The authors thank Kathryn Soule and Fred O'Bryant, librarians at the University of Virginia, for searching databases. They benefited from discussions of this issue with the Committee on the History of Environment and Technology group at the University of Virginia and Frank Smith of Cambridge University Press. Thanks also go to the two anonymous referees and to John Staudenmaier, S.J., for their helpful suggestions.

0040-165X/08/4902-0006/420–29

1. Robert P. Dellavalle et al., "Going, Going, Gone: Lost Internet References," *Science* 302 (31 October 2003): 787–88; Renee Crichlow and Nicole Winbush, "Accessibility and Accuracy of Web Page References in 5 Major Medical Journals," *JAMA* 292 (2004): 2723–24; Eric J. Hester et al., "Internet Citations in Oncology Journals: A Vanishing Resource?" *Journal of the National Cancer Institute* 96 (2004): 969–71; Victoria Reich and David Rosenthal, "Preserving Today's Scientific Record for Tomorrow," *British Medical Journal* 328 (2004): 61–62; Evangelos Evangelou, Thomas A. Trikalinos, and John P. A. Ionnidis, "Unavailability of Online Supplementary Scientific Information from Articles Published in Major Journals," *FASEB Journal* 19 (2005): 1943–44; Carmine Sellitto, "The Impact of

Do humanists and social scientists face the same problem? This research note marks the first published attempt to answer that question in any field of the humanities or social sciences.[2] We examined the reliability of worldwide web citations in two leading history journals (*Journal of American History* and *American Historical Review*) over seven years and found that 18 percent of web links cited over that period were inactive. The problem increased over time. In articles published seven years earlier, 38 percent of web citations were dead. A digital archive enabled us to locate 57 percent of the missing web pages, leaving 43 percent unavailable even to scholars who use the archive. These findings suggest that historians (and probably other humanists) face a major problem in scholarly practice: we are citing internet sources as though they were permanent, when in fact they are ephemeral.

Readers who have used the internet already know that web links die. The contribution of this research note is to quantify the extent of the problem and place it on our professional agenda. Historians of technology are well positioned to consider this problem because of our focus on the interaction between tools and social practice.

In the first section, we briefly survey the development of source citations as integral parts of historical practice. The key point here is that historians over time increasingly valued comprehensive, verifiable references for information in their texts. The footnote evolved in tandem with this desire, especially as history professionalized in the nineteenth century. Central to the success of this system was the reliance on a particular technology, paper. The second section comments on the introduction of a new technology, the internet, as a historical source. In the third section we turn to the heart of this research note, our quantitative analysis of the reliability of the internet as a source for historians. The fourth section discusses some efforts to address the problem of "link rot" by creating archives of web sites, and it describes the way we tested the completeness of the archive. We conclude with a call for professional societies and journals to create better means for ensuring the durability of internet citations.

Impermanent Web-Located Citations: A Study of 123 Scholarly Conference Publications," *Journal of the American Society for Information Science and Technology* 56 (2005): 695–703.

2. As far as we could determine, no humanities or social science journal has published an article quantifying the rate of link decay within its field. Fred O'Bryant and Kathryn Soule, librarians at the University of Virginia, searched databases for us using the following terms: link rot, persistence and URL, permanence and URL, web-site links, and web sites/maintenance. The databases were Historical Abstracts, MLA Bibliography, Web of Science (including the Social Sciences and Arts/Humanities components), and Education Full Text. In Web of Science, they also did a citation search for Dellavalle et al., "Going, Going, Gone," for all disciplines.

APRIL
2008
VOL. 49

Documentation and Paper

Before the modern era, historians felt little need to document their sources in footnotes. Political historians wrote under a rhetorical tradition that focused on lessons in virtue and vice, and they prized the conveyance of moral and political lessons that would be valid in all times and places. They alluded to authorities, but often eschewed citing chapter and verse and expected readers to trust their veracity. Some fields, notably law, developed systems for citing sources as early as the Middle Ages, but these practices apparently had little impact on historiography.[3]

The footnote flowered in the nineteenth century as a way to prove historical arguments. During the eighteenth century, the footnote had been an entertaining form of literary art, and its popularity in fiction helped the footnote spread to historical writings. Its importance grew in the nineteenth century as history professionalized. Archives and libraries opened their collections to more scholars, including younger historians, who now had to demonstrate their mastery of a literature rather than allude to authorities. A professional culture developed in which historians made their arguments in the text and proved them in the footnotes.[4]

If the purpose of a footnote was to prove assertions, other historians needed to be able to examine the same material. Scholars developed conventions that enabled them to retrace one another's steps. Publication information sufficed for books and journals because readers could find them in multiple bookstores and libraries. Citations of archival material required enough data to locate both the archives and the documents within them. Oral historians created a durable record of ephemeral words by depositing transcripts and tapes of interviews in archives. When using material stored by no institution, authors cited it as belonging to their personal collections.[5]

Almost all these practices placed a premium on a particular technology: paper. Paper did not provide truly permanent storage—libraries and archives burned or were bombed, insects chewed their way through pages, acidic paper rotted of its own accord, and the trash beckoned for administrators who ran out of room for books and files. But paper had a number of virtues. In proper conditions, it survived for long periods even if neglected. People could access the information on a page without the aid of other technologies (although some, such as eyeglasses, proved useful). Printing presses made it easy to store copies in multiple places, increasing the odds of survival.

3. Anthony Grafton, *The Footnote: A Curious History* (Cambridge, Mass., 1997), 23, 30; Chuck Zerby, *The Devil's Details: A History of Footnotes* (New York, 2002).

4. Grafton, 4, 220, 225.

5. Citation methods varied, but the University of Chicago Press standardized the system historians have come to use most frequently (University of Chicago Press, *The Chicago Manual of Style*, 15th ed. [Chicago, 2003]); see also Martha Howell and Walter Prevenier, *From Reliable Sources: An Introduction to Historical Methods* (Ithaca, N.Y., 2001).

Computers and the Internet

The development of a new technology—personal computers—revolutionized some fields of history in the late twentieth century. Historical demographers, for example, could collect and analyze vast amounts of data. Personal computers made it easy to store and copy files, and they saved the tedium of calculating how much space to leave for footnotes at the bottom of a page of typescript. But computers also made it harder to retrace a historian's steps. Examining source data usually meant asking for another historian's computer files. One could not read the early data directly; it required a machine to translate binary data into numbers, letters, or other code legible to human beings. The machines that produced and read these data became obsolete in short order, making older files hard or impossible to access. Storage media, such as magnetic tapes and disks, rotted away in a matter of years. Data preservation required frequent copying and updating to suit new machines (or saving old machines for occasional use).

Digital technology—an internet linking vast numbers of computers—offered several advantages to historians: it eased and democratized access to data; it enabled libraries and archives to disseminate facsimiles of rare and unique documents while protecting fragile originals; researchers could copy electronic information quickly and reliably. Internet content expanded at an exponential rate, and search engines such as Google indexed over a billion web pages.[6] With the aid of hypertext, historians could navigate among electronic sources and break down conceptual boundaries.[7]

At the same time, the internet posed challenges. Those who posted information also could revise it, so visitors to the same site might not always find the same information.[8] Websites, or pages within them, often disappeared. No one policed the veracity of most sites. History sites came from both academics and amateurs, and the latter often had little sense of professional standards.[9] The concept of "learning all there is to know" became unrealistic in the face of information glut. As Roy Rosenzweig put it, historians faced the problem of "simultaneous fragility and promiscuity of digital data."[10] Many historians, however, did not view fragility as a problem.

6. Roy Rosenzweig, "The Road to Xanadu: Public and Private Pathways on the History Web," *Journal of American History* 88 (2001): 548–79.

7. Stephen Robertson, "Doing History in Hypertext," *Journal of the Association for History and Computing* 7 (August 2004) (n.p.), http://mcel.pacificu.edu/jahc/JAHCVII2/ARTICLES/robertson/robertson.html (accessed 1 May 2006). We are aware of the irony of citing web sources in our research note and have filed hard copies of all cited documents.

8. Deborah Lines Anderson, "Benchmarks: Controlling Digital Data," *Journal of the Association for History and Computing* 6 (April 2003) (n.p.), http://mcel.pacificu.edu/jahc/JAHCVI1/benchmarks.HTML (accessed 1 May 2006).

9. Rosenzweig, "The Road to Xanadu."

10. Roy Rosenzweig, "Scarcity or Abundance? Preserving the Past in a Digital Era," *American Historical Review* 108 (2003): 735–62.

APRIL
2008
VOL. 49

Almost half (46 percent) of historians polled in 2002 believed that internet sources were permanent enough to be cited in professional publications.[11]

Humanists developed new professional standards to adapt to this new technology. Style manuals added sections on web sources and counseled the inclusion of the Universal (now Uniform) Resource Locator (URL) as the key to enabling others to find the same data. It became standard practice to include the date when an author viewed a particular website. This measure acknowledged that posted information changed, but it did not enable researchers to find earlier versions of revised documents or sites that disappeared.

When discussing the reliability of internet citations humanists tended to focus more on students than on their peers. They rarely discussed durability and instead emphasized what they saw as a tendency among students to assume the veracity of web sources and to rely on them to the exclusion of books and journals. *Wikipedia*, the online encyclopedia that any reader can edit, became the touchstone for debates about accuracy. Instructors created rules about whether and how students could cite online sources.[12]

Testing the Reliability of Internet Citations

This research note shifts the focus from the classroom to the professor's study, and from the accuracy of information to its durability. It quantifies the rate at which internet documents disappeared after being cited in history journals. In order to give internet citations the strongest chance of proving reliable, we selected two journals known for their high standards of scholarship, peer review, and editing: *Journal of American History* and *American Historical Review*. We examined only research articles, since they undergo peer review and are expected to thoroughly document sources. We narrowed our study to the most common type of internet source, the worldwide web. Both journals have posted all their issues from June 1999 to early 2006 on the web site of the History Cooperative, so we used that site's keyword search function to find all occurrences of "www" or "http" (parts of the URL of worldwide web sites) in all research articles.[13] All searches took place between 10 and 19 of April 2006.

11. Suzanne R. Graham, "Historians and Electronic Resources: Patterns and Use," *Journal of the Association for History and Computing* 5 (September 2002) (n.p.), http://mcel.pacificu.edu/JAHC/JAHCV2/ARTICLES/graham/graham.html (accessed 3 May 2006). The question was posed in the negative, with 46 percent of historians disagreeing with the statement that internet resources lack adequate permanence to be cited.

12. Philip M. Davis and Suzanne A. Cohen, "The Effect of the Web on Undergraduate Citation Behavior, 1996–1999," *Journal of the American Society for Information Science and Technology* 52 (2001): 309–14.

13. For *American Historical Review*, the issues spanned volume 104, issue 3 (June 1999) through volume 11, issue 1 (February 2006). For *Journal of American History*, cov-

We first documented the number of times that articles cited the internet. Next we graphed the data and performed a regression analysis to look for trends over time. Finally, we tested the reliability of links by trying to access all of them with a web browser. We avoided introducing typographical or pasting errors by clicking directly on URLs in the online (History Cooperative) version of each article. The result of each click led us to classify a link as active or inactive. In the active category we included links that took us directly or indirectly (via a redirect) to an active site. Because our focus was on the durability of internet sites, we did not verify the accuracy of the cited information within the sites themselves, nor did we try to gauge whether the information had been revised since the author had cited it.[14]

In the inactive category, we included links that directly or indirectly (via a redirect) produced error messages. We did not try to determine whether an entire web site or just a page was unavailable, because both resulted in the inability to locate cited information. We counted sites that denied access because we lacked a password as inactive; such protection made the information as unavailable to readers as did taking it down. Our measure of inactivity is conservative because we used computers on a university network with access to large numbers of online databases that require subscriptions.

Results

A significant number of articles—a total of 132—cited the web (table 1). Between them, these articles included a total of 510 web citations. The degree of reliance differed: more than twice as many articles in the *Journal of American History* cited the web compared to the *American Historical Review* (96 versus 36), and the maximum number of citations in a single article was also twice as large (44 versus 20). The average number of citations per article, however, varied less: 4.1 versus 3.3, respectively. The overall average number of citations per article, 3.9, suggests that reliance on the web was broad but shallow.

The trend in reliance was less clear (fig. 1). After surging between 2000 and 2003, the rate dropped and leveled off during 2004–2005. Between them, the journals published nine issues per year. If the rate in the first 2006 issue of each journal continued, we would expect more than twice as many internet citations in 2006 as in 2005 (180 versus 77).

Our tests of link activity produced three findings:

---

erage extended from volume 86, issue 1 (June 1999) through volume 92, issue 4 (March 2006). All issues are available at the History Cooperative, http://www.historycooperative.org (accessed 12–18 April 2006).

14. A number of cited sources charge users for subscriptions or for individual articles. Examples include all or parts of Project Muse, History Cooperative, JSTOR, www.nytimes.com, www.theweeklystandard.com, and www.washingtonpost.com. We counted these links as active because, although not free, the information was available.

APRIL
2008
VOL. 49

TABLE 1
COMPARISON OF WEB CITATIONS IN *AMERICAN HISTORICAL REVIEW* (*AHR*) AND *JOURNAL OF AMERICAN HISTORY* (*JAH*)

Journal	*Articles citing web*	*Total web citations*	*Inactive citations*	*Percent inactive*	*Average citations per article*	*Minimum citations in one article*	*Maximum citations in one article*
AHR	36	118	19	16%	3.3	1	20
JAH	96	392	72	18%	4.1	1	44
TOTAL	**132**	**510**	**91**	**18%**	**3.9**	—	—

Notes: More than twice as many articles in *JAH* cited the internet compared to those in *AHR*. The maximum number of citations in a single article followed the same pattern. In other respects, the journals differed little.

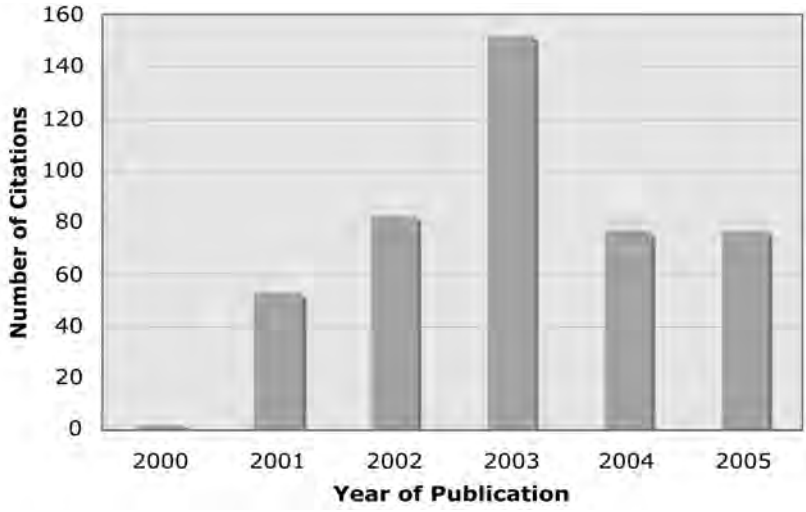

FIG. 1 Total internet citations. This figure pools the data for *American Historical Review* and *Journal of American History*. It only reflects data from 2000–2005 because all issues for those years were searchable on the History Cooperative website (versus incomplete runs for 1999 and 2006). Thus, although the growth of internet citations flattened during 2004–05, it might have increased later. Each journal had posted one 2006 issue by April 2006; those two issues included 40 web citations. If this rate continued in the seven subsequent issues for that year, 2006 would have seen 180 web citations.

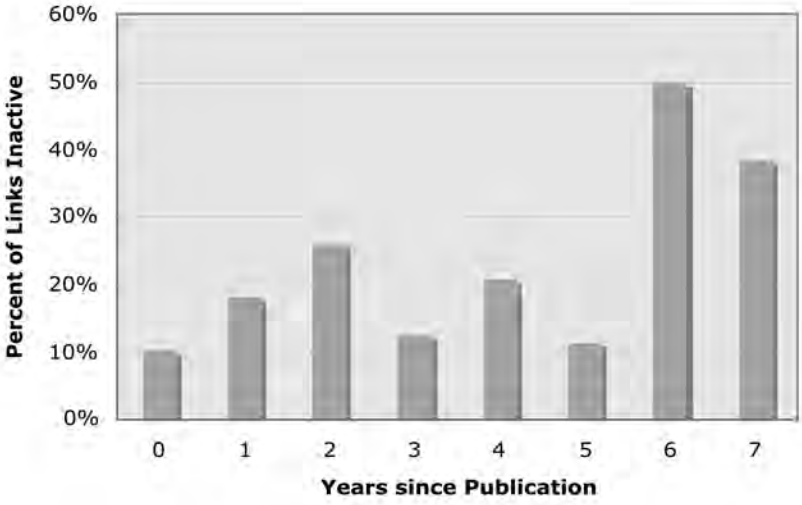

FIG. 2 Rate of link inactivity. The rate of inactivity rose from 10 percent in the year of publication to 38 percent after seven years. (We urge caution about the rate at six years because it is based on only two data points.) A linear regression of the rate of inactivity as a function of time fell just short of statistical significance ($p = 0.074$).

1. Overall, 18 percent of cited websites were inactive (table 1). The journals differed little in this regard (18 versus 16 percent). This is our most important and robust finding.
2. Inactivity began almost immediately. Within two months of publication of the first 2006 issue of each journal, 10 percent of web citations were inactive. This result is based on a small sample size (one issue for each journal), so we consider these findings to be in need of verification.
3. The rate of inactivity rose with time, though the trend was not statistically significant (fig. 2). Beginning at 10 percent in the year of publication, the rate of inactivity rose to 38 percent after seven years. A linear regression of inactivity as a function of time fell just short of statistical significance ($p = 0.074$; $r^2 = 0.44$).[15]

15. Here is a simplified primer for readers unfamiliar with statistics. We wanted to know if link decay increased with time, so we used Microsoft Excel to calculate the straight line (a linear regression) that best followed the overall trend. If the data had followed an exact stepwise pattern (e.g., by increasing 10 percent each year), the straight line would have exactly intersected each year's proportion of inactive sites. In our data, though, some data points fell above and some below that straight line. This was not surprising, since the line was a simplified, rather than an exact, model. To determine *how*

Efficacy of Web Archiving

Our findings demonstrate that historians have been citing ephemeral information from the internet. Ephemeral evidence is not a new problem for historians. One of the virtues of archives and museums is that they collect ephemera. In a parallel way, some groups have tried to address the ephemeral nature of web pages by archiving large parts of the worldwide web. These groups use the strength of information technology—its ability to store large amounts of information—to take snapshots of the web at regular intervals and store those snapshots. One of the most popular is the Wayback Machine operated by the Internet Archive.[16]

To assess the efficacy of the Wayback Machine, we searched it on 22 June 2007 for all the links we had found missing. As before, we tested only whether the URL supplied by the author of an article led us to a live link, not whether the information there was accurate. We found that the Wayback Machine had archived 57 percent of the missing web pages, leaving 43 percent still unavailable. Current archival methods ameliorate, but do not solve, the problem of link rot.

APRIL
2008
VOL. 49

---

simplified, we had Excel calculate how much the data points fell above or below the line. The result of that calculation is called r^2. If the data had increased 10 percent each year, r^2 would have been 1.00 because all data points would have fallen exactly on the line. We would probably feel confident predicting that another 10 percent of links would decay next year. Our r^2 was 0.44, which told us that our simplified model "explained" 44 percent of the pattern in the data. We seem to have identified an important, but not the only, variable in predicting the rate at which links decayed.

Still, knowing what causes 44 percent of the rate at which links decay could be useful information. So, given our data, how confidently can we say that links decay with time? Statistics provide a way to answer that question. Statistics never prove that something is true, since it is always possible that chance (in other words, all the variables other than the one we are studying) caused a pattern. Instead, statistics tells us how often we would expect a pattern if chance alone were at work. If the pattern is extremely unlikely, our confidence increases that our hypothesis is true and we call the result statistically significant. By convention, scientists use 5 percent as the cutoff for significance; that is, if chance alone would have created our pattern five or fewer times out of 100, we have enough confidence that our hypothesis is correct to consider the result significant. With Excel, we calculated that chance alone would have created the pattern in our data (showing that links decayed over time) 7 percent of the time. Because that is more than 5 percent, we call it statistically insignificant—but it is not far from 5 percent, so we consider the result strongly suggestive.

16. The archive's website describes itself this way: "The Internet Archive is a 501(c)(3) non-profit that was founded to build an Internet library, with the purpose of offering permanent access for researchers, historians, and scholars to historical collections that exist in digital format. Founded in 1996 and located in the Presidio of San Francisco, the Archive has been receiving data donations from Alexa Internet and others. In late 1999, the organization started to grow to include more well-rounded collections. Now the Internet Archive includes texts, audio, moving images, and software as well as archived web pages in our collections." The Internet Archive, "About the Internet Archive," http://www.archive.org/about/about.php (accessed 22 June 2007).

Several factors could have caused the omission of that 43 percent of our missing web content from the archive. The automated web crawlers used by the machine might not have found the page when it was still posted on the web. Some sites could not be archived because they were blocked by passwords or other means. One of these means was called robots.txt, which prevented automated crawlers from accessing sites. The Wayback Machine identified several of the sites in our sample as unavailable because of robots.txt. The Wayback Machine also honored requests from site owners to remove their sites from the archive.[17]

Conclusion

The worldwide web has offered an increasingly common though ephemeral source of information. In research articles in two of the most highly respected history journals, 18 percent of web citations decayed within seven years of publication; 10 percent were inactive shortly after publication. Our findings are roughly consistent with those for science journals; we suspect that this problem extends to other humanities and social science publications. A means created to preserve internet sites—the Wayback Machine—made 57 percent of the missing articles in our sample available to scholars who knew about the archive. The other 43 percent of the missing links remained beyond the reach even of those searching the archive.

Reliance on unarchived ephemera is distressing given our commitment to a documented past. We urge professional societies, journals, and presses to create and adopt professional standards for the use of internet documents, including means for preserving materials in a way that ensures their accessibility into the indefinite future. Doing so would be a boon to current and future historians.

17. The Internet Archive, "About the Internet Archive," http://www.archive.org/about/faqs.php#5 (accessed 22 June 2007); Internet Archive, "Frequently Asked Questions," http://web.archive.org/collections/web/faqs.html#exclusions (accessed 22 June 2007).