City University of New York (CUNY)
CUNY Academic Works

Publications and Research | John Jay College of Criminal Justice

2016

The Future of Web Citation Practices

Robin Camille Davis
CUNY John Jay College

How does access to this work benefit you? Let us know!

More information about this work at: https://academicworks.cuny.edu/jj_pubs/117
Discover additional works at: https://academicworks.cuny.edu

This work is made publicly available by the City University of New York (CUNY).
Contact: AcademicWorks@cuny.edu

This is an electronic version (preprint) of an article published in *Behavioral and Social Sciences Librarian*. The article is available online at http://dx.doi.org/10.1080/01639269.2016.1241122 (subscription required).

Full citation: Davis, R. "The Future of Web Citation Practices (Internet Connection Column)." *Behavioral & Social Sciences Librarian* 35.3 (2016): 128-134. Web.

Internet Connection
The Future of Web Citation Practices

Robin Camille Davis

Abstract: Citing webpages has been a common practice in scholarly publications for nearly two decades as the Web evolved into a major information source. But over the years, more and more bibliographies have suffered from "reference rot": cited URLs are broken links or point to a page that no longer contains the content the author originally cited. In this column, I look at several studies showing how reference rot has affected different academic disciplines. I also examine citation styles' approach to citing web sources. I then turn to emerging web citation practices: Perma, a "freemium" web archiving service specifically for citation; and the Internet Archive, the largest web archive.

Introduction

There's a new twist on an old proverb: *You cannot step in the same Web twice*. The constant mutation of the Web is its boon and its bane. Websites can update every minute to show the latest news; social media is posted, reposted, and deleted; whole websites are migrated to reflect new ownership or a name change. It's not unusual to come across a broken link or a "Page Not Found" in everyday Web use. But when content from the ever-changing Web is cited in more static publications, like scholarly articles, those publications run the risk of presenting broken links to future readers. In this column, I'll examine how link rot and reference rot affect scholarly disciplines, what traditional citation styles have to say about citing the Web, and how new approaches could save citations from reference rot.

Reference Rot

In a 2015 *New Yorker* article about web preservation, the historian Jill Lepore laments, "The footnote, a landmark in the history of civilization, took centuries to invent and to spread. It has taken mere years nearly to destroy." Citing sources allows the reader to follow up on the author's claims to see if the cited evidence holds. Claims become weaker when the reader can't access the original source, and original sources often do not survive intact on the Web.

Klein et al. (2014) define the term *reference rot* as a combination of two problems: *link rot*, in which a given link no longer exists (an HTTP 404 error); and *content drift*, in which a given link's content has changed since being referenced, "even to such an extent that it ceases to be representative of the content that was originally referenced."

How much of scholarly literature is affected by reference rot? In the research on this topic, studies tend to focus on specific academic domains, and they usually measure either how many articles that cite Web sources have suffered any reference rot, or how many URLs that are cited in a corpus have suffered reference rot. These three oft-cited studies paint a bleak picture:

- In a study of scholarly articles in the domains of science, technology, and mathematics, Klein et al. (2014) found that the "vast majority of STM articles that contain references to web at large resources do suffer from reference rot. The infection rate between 2005 and 2012 oscillates between 70% and 80%."
- Reference rot plagues the law as well. Liebler and Liebert (2013) found that 29% of URLs in U.S. Supreme Court opinions between 1996–2010 suffered from link rot. Zittrain et al. (2014) conducted a study of articles published from the late 1990s to 2012 in three legal journals from Harvard, finding that more than 70% of URLs cited suffered reference rot.
- Russel and Kane (2008) conducted a study of articles published between 1999 and 2006 in two scholarly history journals. They found one year after publication, 10% of URLs cited were broken; after seven years, 38% were broken.

Five years from now, or even today, if these researchers were to revisit the cited URLs in these studies, their numbers would likely be much higher. Reference rot is always increasing, and it will always be a threat to the integrity of scholarly research in every discipline.

Traditional approaches to citation

How should a writer cite Web sources, according to current citation styles? Here, by "Web source," I mean sources found on the World Wide Web, such as blog posts, news articles, digital projects, and the like—sources that may suffer reference rot. Some documents on the Web do have DOIs (digital object identifiers), which are persistent links that do not suffer from reference rot. Citation styles therefore prefer DOIs over URLs, but DOIs are not available for most online resources. Let us consider how MLA, APA, and Chicago styles recommend the citation of Web sources that do not have DOIs, and how they fall short of preventing reference rot.

MLA

The MLA Handbook, 8th edition (published 2016), advises including the URL in every web citation. The Handbook acknowledges that URLs change often, but it points out that "[e]ven an outdated URL can be useful, however, since it provides readers with information about where the work was once found" (48). Shortlinks (e.g., Bitly links) are discouraged, as they often expire. It is optional to include the date of access, but the Handbook points out that it is "often an important indicator of the version you consulted" (53).

Sample Web source citation, taken from the MLA Handbook (48):

> Hollmichel, Stefanie. "The Reading Brain: Differences between Digital and Print." *So Many Books,* 25 Apr. 2013, somanybooksblog.com/2013/04/25/the-reading-brain-differences-between-digital-and-print/.

It should be noted that the MLA Handbook had advised including the URL in every Web citation for years until the 7th Edition (published 2009) rescinded that recommendation. The reasoning then was that URLs often expire and that typing a long URL found in a print source into a browser is burdensome (182). They noted that for resources whose URL had changed, an intrigued reader could likely find the source simply by searching the Web or a library for it, if it still existed. Unfortunately, the recommendation to drop the URL means that for MLA-style publications between 2009 and 2016, it may be difficult to track down the Web source, especially if the source has been deleted. Moreover, the 7th edition's recommendation may be why there are few studies of reference rot in humanities scholarly literature — researchers cannot gather URLs from MLA-style reference lists when authors did not include them in their citations.

APA

The APA Publication Manual suggests that if a Web source does not have a DOI, the author should always include the URL, but they should only include the date of access when it is likely that "the source material may change over time (e.g., Wikis)" (192). The only date required in the citation is that of publication or posting.

Sample Web source citation, taken from the APA Publication Manual (215):

> PZ Meyers. (2007, January 22). The unfortunate prerequisites and consequences of partitioning your mind [Web log post]. Retrieved from http://scienceblogs.com/pharyngula/2007/01/the_unfortunate_prerequisites.php

(The above URL, by the way, is a broken link. But the blog post appears to be available on the same site at a slightly different URL, found through a Google search.)

Chicago

The Chicago Manual of Style encourages authors citing website content to include the URL and "a publication date or date of revision or modification… if no such date can be determined, use an access date" (753).

Sample Web source citation, taken from the Chicago Manual of Style, notes and bibliography system (753):

> Microsoft Corporation. "WD2000: Visual Basic Macro to Assign Clipboard Text to a String Variable." Revision 1.3. Microsoft Help and Support. Last modified November 23, 2006. http://support.microsoft.com/kb/212730

Problems with traditional Web citation

Of these three most commonly used citation styles, none require the complete information about the *context* of the Web source citation: the URL and access date. That tells us exactly *what* the author accessed and *when*. This safeguards against potential source confusion, particularly if the Web source underwent content drift after it was cited. The URL and access date are necessary if a reader wants to use the largest Web archive, the Internet Archive's Wayback Machine, to view an archived page captured around the time it was cited, before any potential content drift or link rot occurred. In this scenario, even if the URL is a broken link, it is still a critical piece of metadata. The MLA style is therefore the

most informative of the three, requiring a URL and suggesting the inclusion of the date of access.

But even if a Web source citation does include the URL and access date, the page might not have been archived by the Internet Archive around the time of citation—if it has been archived at all. An enterprising reader still might not be able to see the Web source as the author saw it, which may be important for verification of a claim.

New approaches to citation

At this time, the best way to guard against reference rot is to be proactive and create a publicly accessible archival copy of the Web source at the time of citation. An archival copy means a snapshot of the webpage bundled with metadata about the date and time the snapshot was taken. To follow up on an author's claims about a Web source, it is necessary to see the original source, as close to what the author encountered as is possible. Rotten references in traditional citation styles block access to this context. But by incorporating archival copies of sources into their Web citations, scholars can proactively preserve the context of their research. A number of free services can create archival copies on demand. Two are profiled here.

Perma

Begun in 2014, the Perma project was created specifically in response to reference rot in legal citations (Zittrain et al.). The Perma website, perma.cc, states simply, "Websites change. Perma Links don't. Perma.cc helps scholars, journals, courts, and others create permanent records of the web sources they cite." The service operates on a freemium model: users are offered free accounts and the option to archive up to 10 websites per month. Archived websites are accessible with a short URL. (Note that Perma is meant for citation purposes only. A user can't search Perma for any websites it may have archived.) The project is maintained at the Harvard Law School Library, with "communities" administered by other law libraries.

A sample APA-style Web source citation that links to a Perma archival copy might look like this:

> PZ Meyers. (2007, January 22). The unfortunate prerequisites and consequences of partitioning your mind [Web log post]. Retrieved from http://scienceblogs.com/pharyngula/2007/01/22/the-unfortunate-prerequisites. Archived at https://perma.cc/B9EK-7NC7.

Internet Archive

The Internet Archive (IA, archive.org) operates the largest and oldest Web archive, dating back to 1996. Billions of webpages have since been archived multiple times by IA with their Wayback Machine. A user can search for previously archived snapshots of a webpage by entering the URL into the Wayback Machine. If it has been archived, the dates of all snapshots taken are displayed as links in a header above the archived page.

Like Perma, IA offers on-demand archiving for publicly accessible webpages. The Wayback Machine's main page has a "Save Page Now" option to "[c]apture a web page as it appears now for use as a trusted citation in the future."

A sample APA-style Web source citation that links to an IA archival copy might look like this:

> PZ Meyers. (2007, January 22). The unfortunate prerequisites and consequences of partitioning your mind [Web log post]. Retrieved from http://scienceblogs.com/pharyngula/2007/01/22/the-unfortunate-prerequisites. Archived at https://web.archive.org/web/20160913232906/http://scienceblogs.com/pharyngula/2007/01/22/the-unfortunate-prerequisites/

Alternatively, an author could use the "Save Page Now" option an all Web sources cited in their publication, without including the very long link to the archival snapshot. Because the Wayback Machine is a reliable and well-known service, a curious and savvy reader would likely turn there first in case of link rot or content drift. Saving the page at the time of writing or publishing ensures that the reader accesses the page as the author saw it. A sample APA-style Web source citation that simply references an author-created IA archival copy might look like this:

> PZ Meyers. (2007, January 22). The unfortunate prerequisites and consequences of partitioning your mind [Web log post]. Retrieved from http://scienceblogs.com/pharyngula/2007/01/22/the-unfortunate-prerequisites. Archived 13 September 2016 in the Internet Archive.

Caveats

It should be noted that neither IA nor Perma can capture every webpage perfectly. Web archiving is a deceptively difficult problem. Some websites don't allow Web crawlers. Some websites might use cutting-edge technology to display content that crawlers can't fully preserve. An archived version of a page might look different from the live page. (One feature of Perma is that it also saves a screenshot as an image, in case the archived webpage doesn't save properly.) In addition, IA and Perma are not the only Web archives or archiving services. And there is no guarantee that a given Web archive will always exist and remain accessible—though users can take comfort in the fact that Perma is supported by a network of established libraries with deep pockets and the Internet Archive is 20 years old and thriving.

What scholars can do now

Study after study proves that reference rot is a serious affliction across scholarly disciplines, and it is a problem that will keep growing. It will not disappear. But it can be abated. Proactively archiving cited Web sources at the time of writing is the best way to create trusted citations of Web sources. Scholars can begin creating archival snapshots of their cited Web sources with tools like Perma or IA, and including archival information in citations. This effort is time-consuming, but it is worthwhile, as ensures that readers have a reliable way of following up on claims, preserving the integrity of the work.

Looking ahead

In an ideal world, citations that contain URLs should trigger the creation of archival snapshots. The author, instead of taking extra time and effort to make archival copies of Web sources, can focus on their work while their sources are passively archived behind the scenes. Perhaps this could happen at one or some of these points:

- During the research process, the author's bibliographic management program requests an archival snapshot of a saved source.

- In the manuscript submission process, the publication's system requests archival snapshots and insert snapshot URLs into citations.
- In a repository ingestion process, the system could scan the submission for URLs and create a supplemental file of archival snapshots or links to publicly accessible snapshots.

There is good news: research has already been undertaken to pursue these ends. A project called Hiberlink, funded by the Mellon Foundation, developed a prototype of a Zotero plugin that archived URLs as they were added in the scholar's research process. Another Hiberlink endeavor, Hiberactive, explored the implementation of the two latter options. The Hiberlink project's grant ended in 2015, but Hiberlink's research partners continue their work in this area at Los Alamos National Laboratory Research Library and University of Edinburgh.

In the meantime, however, scholars can tackle reference rot, starting with their own references.

References

"Hiberlink - Zotero Plugin." 2016. *Hiberlink*. Accessed September 16, 2016. http://hiberlink.org/zotero.html. Archived in the Internet Archive and at https://perma.cc/9YZ8-JBEJ.

Klein, Martin, Harihar Shankar, Herbert Van de Sompel, and Richard Wincewicz. 2014. "Hiberactive: Pro-Active Archiving of Web References from Scholarly Articles." presented at the Open Repositories 2014, Helsinki, Finland, June 2014. Accessed September 12, 2016. http://www.slideshare.net/martinklein0815/hiberactive. Archived in the Internet Archive and at https://perma.cc/ED7A-AC3B.

Klein, Martin, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. "Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot." *PLoS ONE* 9 (12). doi:10.1371/journal.pone.0115253.

Lepore, Jill. 2015. "The Cobweb." *The New Yorker*, January 26. Accessed May 22, 2016. http://www.newyorker.com/magazine/2015/01/26/cobweb. Archived in the Internet Archive and at https://perma.cc/F4NH-9S8F.

Liebler, Raizel, and June Liebert. 2013. "Something Rotten In The State Of Legal Citation: The Life Span Of A United States Supreme Court Citation Containing An Internet Link (1996-2010)." *Yale Journal of Law and Technology* 15 (2). Accessed September 7, 2016. http://digitalcommons.law.yale.edu/yjolt/vol15/iss2/2. Archived in the Internet Archive and at https://perma.cc/WT3K-6CTQ.

MLA Handbook for Writers of Research Papers. 2016. 8th ed. New York: Modern Language Association of America.

"Perma [homepage]." 2016. *Perma*. Accessed September 13, 2016. https://perma.cc/. Archived in the Internet Archive.

Publication Manual of the American Psychological Association. 2009. Edited by Gary R. VandenBos. 6th ed. Washington, DC: American Psychological Association.

Rhodes, Sarah. 2010. "Breaking Down Link Rot: The Chesapeake Project Legal Information Archive's Examination of URL Stability." *Law Library Journal* 102 (4): 581–97. Accessed September 7, 2016. http://scholarship.law.georgetown.edu/digitalpreservation_publications/6/. Archived in the Internet Archive and at https://perma.cc/7G4X-PB2S.

Russell, Edmund, and Jennifer Kane. 2008. "The Missing Link: Assessing the Reliability of Internet Citations in History Journals." *Technology and Culture* 49 (2): 420–29. doi:10.1353/tech.0.0028.

The Chicago Manual of Style. 2010. 16th ed. Chicago: University of Chicago Press.

"Wayback Machine [homepage]." 2016. *Internet Archive*. Accessed September 1, 2016. https://archive.org/web/. Archived at https://perma.cc/F4UX-VF4H.

Zittrain, Jonathan, Kendra Albert, and Lawrence Lessig. 2014. "Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations." *Harvard Law Review* 127 (4): 176–99. Accessed September 1, 2016. http://harvardlawreview.org/2014/03/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations/. Archived in the Internet Archive and at https://perma.cc/Y2YD-W2L2.