URLs: Uniform Resource Locators or Unreliable Resource Locators

Carol Anne Germain

As the use of citing electronic World Wide Web sites grows, the question arises as to whether this practice has scholarly limitations due to the fact that uniform resource locators (URLs) often become inaccessible. This research studies the accessibility of sixty-four URLs cited in thirty-one academic journal articles. Results of this longitudinal study found an increasing decline in the availability of URL citations.

Ten years ago, most people had no idea that the Internet existed. Today, it is used daily by millions of people who access it for a variety of reasons. Some use it to connect with friends and family; others use it for entertainment purposes (jokes, sports and freebies); and still others use it for research. Students approaching a library reference desk often insist that the Internet be used to locate information for papers, projects, and other academic assignments. Many journal articles, including refereed articles, contain citations to Internet sources. Despite the popularity of Internet citations, we still may question the integrity of this practice. How often have we tried to link to uniform or universal resource locators (URLs) only to find a "404 NOT FOUND" or other messages denying access? These warnings let us know that the information we came to access is no longer accessible at this site. The information may have been moved to another site, equipment may be down, or the information may have been removed completely. This is frustrating because cited references need to be accessible and persistent. Citations provide the reader with an outline of the works an author has consulted to develop an article, conference paper, monograph, or other scholarly study. After a review of the importance of permanence as a feature of academic citation, this paper presents evidence of the impermanence of actual URL citations.

The Role of the Citation

What is the purpose of a citation? Why is this erudite mechanism so important? The *Oxford English Dictionary* defines the verb *to cite* as "to quote (a passage, book, or author) generally with implication of adducing as an authority."[1] Authority furnishes credibility to the written piece. Citation allows the reader to reference other works the author has cited. The reader then has the ability to verify a quotation, check the semantic connection, or confirm whether the author has included all of the materials and statistics of a study. In a sense, "citation keeps you honest."[2] It is essential that the academic community be able to rely on and utilize the studies, arguments, and findings of other scholars.

Carol Anne Germain is the Networked Resources Education Librarian at the State University of New York at Albany (SUNY); e-mail: cg219@csc.albany.edu.

Citation also provides the ability to acknowledge the works of others that support a piece of research. When using the materials of others, citation offers the opportunity to recognize the cited author. "A paper that conforms to the norms of scholarly perfection would explicitly cite every past publication to which *it* owes an intellectual debt."[3]

One of the most important functions of the citation is that it links the written work into a much larger community. When the novice physicist uses Einstein's theories to uphold an argument, a connection is established between significant works of the past and works of the present. Other physicists will evaluate this work and reach conclusions as to whether it is an addition to the field.

> Every time a scholar presents a review of the literature in her area of inquiry, or writes a bibliographic essay, or incorporates another writer's words or ideas to advance her own thesis, she maps the field of her discipline. She draws the boundaries, circumscribes the territory of her field of discourse, and determines who else is within and who is without.[4]

In other words, "she" makes herself a part of a much larger community. This community promotes intellectual growth that may, in turn, stimulate the development of new medicines and cures, novel writing techniques, or breakthroughs in technology. The dialogue that is encouraged with the usage of citation encourages a learned fellowship.

Thirty-one randomly chosen academic journal articles, containing sixty-four citations with URLs, were reviewed.

Henry Small, when describing why an author or scientist cites another text, referred to the citation as a "symbol." These "symbols of concepts or methods" function as connections to earlier works that an author-researcher has embedded as a reference in his or her writings. "This leads to the citing of works which embody ideas the author is discussing. The cited documents become, then, in a general sense, 'symbols' for these ideas."[5] Blaise Cronin summed up the need for a theory of citing very eloquently:

> Metaphorically speaking, citations are frozen footprints on the landscape of scholarly achievement; footprints that bear witness to the passage of ideas. From footprints it is possible to deduce direction; from the configuration and depth of the imprints it should be possible to construct a picture of those who have passed by, whilst the distribution and variety furnish clues as to whether the advance was orderly and purposive.[6]

The Persistence of Citations

An important feature of scholarly links is that they are available indefinitely. It is imperative that cited materials be accessible and not ephemeral. Phyllis Franklin, executive director of the Modern Language Association, stated that "the M.L.A. has concluded that scholarship depends on getting back to a source."[7] The researcher depends on cited work as a collaboration of ideas. If the locations of ideas that substantiate the author's work no longer exist, the foundation of their work is in question.

To assume that all cited works are easily obtainable is naive.[8] Fugitive material and grey literature are found in written works, the former being pamphlets, programs, and other literature published (not always officially) in small quantities and often produced for one-time use.[9] Materials such as these are almost impossible to retrieve and thus are generally not cited. Grey literature is literature that cannot normally be purchased through booksellers. Examples of these types of materials include conference proceedings, trade brochures, preprints, technical reports, dissertations, and government agency publications. It is often difficult to acquire these materials and frequently takes some skill to do so. The National

Technical Information Service (NTIS) provides access to technology reports, while Bell and Howell, formerly University Microfilms International (UMI), places dissertations on microform and numerous trade associations archive their professions' literature. Although it is difficult to work through resources such as NTIS and Bell and Howell, one is assured that their materials are retrievable. One of the reasons for this assurance is that various institutions and organizations, such as government agencies, union affiliations, and academic institutions, have responsibility for maintaining and preserving the materials.

With the emergence of the Internet and Internet publishing, individuals and institutions in increasing numbers are authoring and posting papers and studies on this electronic medium. One of the complications with this type of publication is that there is no guarantee that these works will be perpetually available. "Estimates put the average lifetime for a URL (the Web site location) at forty-four days."[10] A longitudinal study undertaken by Wallace Koehler reviewed the persistence of 361 randomly chosen Web sites and Web pages over one year. Results of this study found that 110 (31%) of the Web sites and Web pages failed to respond at the final test.[11] This electronic environment, though very exciting and stimulating, also is quite volatile.

The academic world should be concerned about the citation of documents that are located on the Internet. When users try to retrieve electronic sources listed in the citing publication, they often do not find the references but, instead, are faced with an "error" message. It is unfortunate, but documents found within the electronic setting have the characteristic of lacking permanency.[12] "URLs change at the whim of hardware reconfiguration, file system reorganization, or changes in organizational structure, leaving users in 404 Limbo."[13] "The Internet's holdings change every minute of the day." Students and researchers find that materials on the information superhighway can disappear "with the touch of a Webmaster's delete key."[14] In its style manual, the American Psychological Association warns those who use online information:

> The researcher has immediate access to a wealth of information but must consider the reader's access to that material: Will the information be available to the reader even if the reader follows a given retrieval path, or will the material soon be archived to tape and difficult to obtain? Is the information widely accessible or accessible only on a campus's local network? This publication recommends that if the same data is available in both print and electronic formats then the writer should use the "preferred print version."[15]

Methodology

The following study was undertaken to investigate the reliability of URLs in academic citation. Thirty-one randomly chosen academic journal articles, containing sixty-four citations with URLs, were reviewed. The academic journals used were from a variety of disciplines. Thirteen citations were from information and library science, ten from the hard sciences, seventeen from computer science, eleven from the humanities, and thirteen from the social sciences. The printed journals were published between 1995 and 1997.

To verify the persistence of the URL citations, each address was accessed to see if the site was currently active. Using a Netscape browser, the URL address was logged into the Netscape "open" window. Over a three-year period (1997–1999), this procedure was conducted once a month for three consecutive months (February, March, and April). This was to determine if each cited site still existed. Three different access days were used each year to insure against temporary interruptions. Reasons for denied access might include that the URL's host computer was down, that a Web site was not being worked on

TABLE 1
Availability of Cited URLs

	Not Accessible	Accessible	% Unavailable
1997	17	47	26.5
1998	24	40	37.5
1999	31	33	48.4

and unreachable, or that too much traffic on the Internet caused a time-out. Each of the nine testings was conducted between 8 a.m. and 9 p.m.

The content of the Web site, update information, and style format were not reviewed. In certain circumstances, some effort was made to access a site if a spelling error or misprint seemed to be within the URL. This included omitting periods where the publisher added them as style; omitting hyphens at the end of a line, within a URL; and adding a top-level domain, such as edu, to the domain name where it seemed to be absent. Only direct URL searching was done; no attempt was made to use Internet search engines to find the cited materials.

Some may say that Internet search engines provide help with locating sites, but these tools are neither authoritative nor exhaustive.

In this paper, persistence of a URL citation is understood as the ability to access a cited URL containing the Web site with the identical title of the cited work. If an index or search tool was retrieved that linked to the cited work, the URL citation also was considered persistent. Citations containing URLs that accessed a host site, but not the cited file, were not regarded as persistent. URL citations that had moved to a new URL and contained the same title/author were appraised as persistent.

When an Internet site cannot be accessed, a variety of error messages may appear. The error message "404 Not Found" appears when Netscape cannot locate the specified Web site. This is due to either relocation or removal of the site.[16] "File Not Found" is similar in nature and means that the user has reached the host computer, but the host cannot find the requested Web site file. The "Not Found " error message gives the user a variety of reasons for not being able to connect to the desired document. "Unable to Locate Server," "Socket Error," and "No Response" are error messages resulting from not being able to connect to the remote server. This may occur when the remote server is either too busy or no longer in existence. Generally, remote computers only send error messages. Unless instructed, they give no forwarding address or other indication of the materials location.

Results

It is assumed that all of the URLs found in the cited works were active Internet sites when they were cited originally. Within each test year, the results did not vary significantly over the three monthly samples; however, results of annual comparisons did produce variability. After checking for persistence of the sixty-four citations, seventeen (26.5%) could not be accessed in 1997. In 1998, twenty-four (37.5%) could not be accessed and thirty-one (48.4%) could not be reached in 1999. As table 1 shows, availability of cited URLs declined about 11 percent annually.

A review of the error messages shows that "Not Found" notices appeared nine times in 1997 and 1998 and thirteen times in the final test. Server errors were retrieved five times in 1997 and twelve times in both 1998 and 1999. Messages indicating relocation appeared three times in the first two years and six times in the final year (see table 2).

This decline in availability of cited URLs had a dramatic impact on the original articles from which these citations were drawn. Of the thirty-one original source articles, in 1997, twelve (38.7%) contained inaccessible citations; in 1998, seventeen (54%) had citations that could not be retrieved; and in the last year, twenty-one (67.7%) contained citations that could not be found (see table 3).

TABLE 2
Review of Error Messages

	1997	1998	1999
"Not Found"	9	9	13
Server Errors	5	12	12
Relocated/Unavailable	3	3	6

Conclusion

After a three-year period, almost 50 percent of the URL citations could not be accessed and two-thirds of the journal articles contained corroded citations. How can this profound loss of academic citation be explained?

Originally, some of the URL citations may have contained misspellings, incorrect domain names, or punctuation errors. Computer software requires meticulous input and is unforgiving when encountering any text or syntax error. Further decline in the accessibility of the tracked URL citations may be attributed to the vast changes in computer and institutional infrastructures. A researcher moving to another job, the purchase of a new server, or the restructuring of an academic department may change the location of a computer file and its URL. Thus, the cited URL is rendered inaccessible.

Print resources have authoritative indexes and finding aids to locate hard-to-find citations. When an author cites the incorrect volume number, the correct one can be found in a variety of sources. The Internet does not have comparative tools. Some may say that Internet search engines provide help with locating sites, but these tools are neither authoritative nor exhaustive.

An assortment of solutions for preserving Internet materials has been initiated. In the United States, Brewster Kahle and a small group of technical professionals have started a project called the Internet Archive. Over a number of years, they have taken a "snapshot" of Web pages found on the Internet.[17] Although this is a noteworthy project, there is no assurance that these records will be maintained in the future. Without adequate finding aids, it will be impossible to access information from a snapshot.

Other efforts to preserve materials found on the Internet are being developed by OCLC. This vast library consortium is working on numerous projects that involve the cataloging and archiving of resources found on the Internet. InterCAT, a project funded by the U.S. Department of Education, is one such endeavor. With the effort of libraries and institutions of higher education, the creation, implementation, testing, and evaluation of a searchable database of USMarc records that contain electronic location and access information has been initiated.[18] "This is the most traditional library-type approach to finding material on the web."[19] InterCAT uses volunteers to catalog electronic sites found on the Internet. To date, this catalog contains more than 70,000 records.[20] Another project undertaken by OCLC is implementation of persistent uniform resource locators (PURLs). A PURL is a record of URL sites that individuals or institutions have registered with OCLC. "Instead of pointing directly to an Internet location, a PURL points to an intermediate resolution service that maintains a database linking the PURL to its current

TABLE 3
Annual Comparison of Journal Articles
Containing Inaccessible URL Citations

	# Articles Containing Inaccessible Citations	% Containing Inaccessible URL Citations
1997	12	38.7
1998	17	54.0
1999	21	67.7

URL and returning that URL to the user."[21] This trial, though, may be a "short-term experiment or a long-term solution."[22]

It is ironic that a utility called "persistent" could be part of a "short-term experiment." Persistence qualifies endurance. Endurance is essential when discussing materials that are to be cited. For the scholarly community to retain its integrity, standards must be set to ensure that cited works are retrievable. In the past, this has not been such a consequential issue. Printed materials have been bought, stored, and archived in libraries for hundreds of years. It seems unlikely that a published book or journal could not be found in some library or archive in the world. Electronic data hold no such promise. With the average life span of an Internet file being less than two months, how many data and materials already have been lost?

Not one of the sixty-four citations reviewed in this study was a PURL. All of the articles were published in academic journals and written by members of the scholarly community. At the final testing, twenty-one of the thirty-one articles contained citations that could not be accessed. Whether this information is available in parallel print sources is unknown. Nonetheless, it is frightening to think that the substructure of the intellectual community is relying on a medium that is so volatile.

The Internet is a very provocative environment. It provides the ability to connect, communicate, and share with members of many disciplines. However, this useful tool needs, at this point, to be viewed as a medium for exchange rather than as a library. Until there is some secure means of accessing data continuously from this resource, using the Internet as a virtual depository of cited materials is indefensible. Academic citations need to be reliable and accessible, and URL citations are not. Students and scholars should proceed with caution and utilize sources that endure.

Notes

1. *The Oxford English Dictionary*, vol. 3. (New York: Clarendon Pr., 1989), 248.
2. Mary-Claire Van Leunen, *A Handbook for Scholars*, rev. ed. (New York: Oxford University Pr., 1992), 9.
3. Manfred Kochen, "How Well Do We Acknowledge Intellectual Debts?" *Journal of Documentation* 43 (Mar. 1987): 54–64.
4. Shirley Rose, "Citation Rituals in Academic Cultures" (paper presented at the annual meeting of the Conference on College Composition and Communication, Seattle, Mar. 16–18, 1989), ERIC ED 309 434, microfiche.
5. Henry Small, "Cited Documents as Concept Symbols," in *Social Studies of Science*, vol. 8 (Beverly Hills, Calif.: SAGE, 1978), 327–40.
6. Blaise Cronin, *The Citation Process: The Role and Significance of Citations in Scientific Communication* (London: Taylor, 1984), 25.
7. Lisa Guernsey, "Cyberspace Citations," *Chronicle of Higher Education* 42 (Jan. 12, 1996):A18–21.
8. Charles Auger, *Information Sources in Grey Literature* (New Providence, N.J. : Bowker-Saur, 1994), 3.
9. Leonard Montague Harrod, *Harrod's Librarians' Glossary of Terms Used in Librarianship, Documentation and the Book Crafts and Reference Book*, comp. Ray Prytherch (Brookfield, Vt.: Gower Publishing, 1990), 263.
10. Brewster Kahle, "Preserving the Internet," *Scientific American* 276 (Mar. 1997): 82–83.
11. Wallace Koehler, "An Analysis of Web Page and Web Site Constancy and Permanence," *Journal of the American Society for Information Science* 50 (Feb. 1999): 162–80.
12. Corrinne Jorgensen and Peter Jorgensen, "Citations in Hypermedia: Maintaining Critical Links," *College & Research Libraries* 52 (Nov. 1991): 528–36.
13. K. E. Shafer, S. L. Weible, and E. Jul, "The PURL Project," *Annual Review of OCLC Research* (1996): 25–26.
14. Michael A. Arnzen, "Cyber Citations: Documenting Internet Sources Presents Some Thorny Problems," *Internet World* 7 (Sept. 1996): 2–4.

15. *Publication Manual of the American Psychological Association* (Washington, D.C.: American Psychological Association, 1994), 218.

16. *Netscape 2 Simplified* (Foster City, Calif.: IDG Books Worldwide, 1996), 35.

17. Kahle, "Preserving the Internet," 82.

18. Jeanette Woodward, "Cataloging and Classifying Information Resources on the Internet," *Annual Review of Information Science and Technology* 31 (1996): 189–220.

19. Pat L Ensor, "Libraryland Organizes the Web: An Unnatural Process?" *Technicalities* 15 (Nov. 1995): 9–11.

20. Norm Medeiros, "Making Room for MARC in a Dublin Core World," *Online* 23 (Nov./Dec. 1999): 57–60.

21. Jennifer L. Marill, "A Survey of Standards for Identifying Serial Items on the Internet," *Acquisitions Librarian* 21 (1999): 83–91.

22. Karen Schneider, "Cataloging Internet Resources: Concerns and Caveats," *American Libraries* 28 (Mar. 1997): 57.