You Can’t Get There from Here: Student Citations in an Ephemeral Electronic Environment

Karen Hovde

This study investigates 1,666 citations to Internet resources from 529 freshman English composition papers from two different years, 1999 and 2004. Web citations are examined in relation to traditional print resources, by type and domain, and with special attention to their persistence through time. After 7 months from submission of the papers for the 1999 sample, 38 percent of the links led to the document cited. After 12 months for the 2004 sample, 45 percent of the links accessed the document. A resubmission of the 1999 citations after 7 years reduced the number of successful links to 9 percent. The instability of Web citations in undergraduate student papers is discussed in the larger context of citation structure, with implications for future scholarship.

The academic community has been concerned for some time with the implications for scholarship of students’ Web use. That the quality of student research has suffered in comparison to that of past decades, that the ease of unlicensed borrowing has increased, and that the effort implied in digesting the sources pursuant to critical thinking is more grudgingly inspired—all this has been blamed, both analogously and in print, on the Internet. One subset of these investigations has been research into the idiosyncratic behavior of Web citations.

The forms and capabilities of the electronic environment have introduced a situation in which there is a poor fit between traditional inquiry and recording behavior and the larger ethical/legal structure in which it is subsumed. When the relative rigidity of citation structure wars with the ease and fluidity of Internet transactions, who is the loser? The student, who may wonder what all the fuss is about? The instructor, who can’t follow the student’s trail? The body of academic inquiry, in which truth and reproducibility are judged on the completeness and transparency of the recorded journey?

The library community, sitting as it does on the actual playing field, has been diligent in documenting student use of traditional and electronic resources. Moved by fiscal and pedagogical concerns, librarians have worked to ease the integration, navigation, and assessment of electronic resources into the student purview. Librarians, with their background in bibliometrics, were quick to appreciate the value of student citation records as a tool to investigate research behavior,

Karen Hovde is Associate Professor, Reference Librarian and Coordinator of Library Instruction in Founders’ Memorial Library at Northern Illinois University; e-mail: khovde@niu.edu.
looking at the sources, types, and implications of student citation choices. Analysis of student bibliographies by librarians can only peripherally address questions regarding the content and quality of selected resources (outside of the very general implications of authorship, authority, currency, and other factors), inasmuch as they are not ordinarily privy to the specific requirements for an assignment within a particular discipline.

**Literature Review**

Web resources have proved to be a very fertile ground for evaluation through citation analysis. That these resources have taken their place as a substantial category beside the traditional breakdown of books, journals, magazines, and newspapers is uncontested. Longitudinal studies can demonstrate the extent to which that category grows in proportion to the total, and in response to instructors’ guidelines. But in addition to documentation of the sheer volume of Web resources, there is emerging an interesting literature regarding some peculiar limitations of the character of the Web citation—indeed, of the often lamented quality of content.

Simply put, the Web is an unstable medium. *A sine qua non* of a paper’s references is stability. Wallace Koehler has provided a seminal and discriminating evaluation of the types and rates of change that Web sites undergo. Almost all Web sites (99 percent of those followed after 1 year) and Web pages (99.1 percent after 1 year) change along a continuum of significance that may be as minor as a modification to a graph or as lethal as total death and subsequent disappearance. Koehler pegs the half-life of an aging general Web collection at about two years. A site or page change may or may not be accompanied by a change in the URL, the critical element of a Web citation.

Authors have investigated the persistence of Web citations in various contexts. Mehta tracked 85 Web citations from the Education Account of LEXIS-NEXIS; only 37 percent of the citations were still active at the time of her study. Rumsey followed law review articles with Web citations for five years; in the most recent year, 2001, 61.8 percent (322 of 521) had working links. Taylor and Hudson examined the URLs and annotations of Web sites in “Internet Resources” columns for the year 1997–1998; by 1999, only about 78 percent of the links were viable. Buega and Dimitrova performed an analysis of citations spanning three years in three communications journals; of their total of 416 citations, 61 percent were still accessible. Sellitto’s study of 1043 Web citations in a set of Education and Training conference papers spanning nine years (1995–2003) found 54 percent of the links to be successful.

In investigations of student references, Gillette and Videon checked a small number of sites (48) in English composition papers at the time of submission, of which 30 percent connected after some manipulation. Davis and Cohen’s investigation of undergraduate bibliographies from 1996 and 1999 in introductory economics classes found that 47 percent of Internet citations for the 1996 sample (69 sites), and 55 percent of the 1999 sample (175 sites), accessed the referenced document. Not found at all were 53 percent and 16 percent, respectively. Thomas tracked the persistence of Web citations in graduate student theses to find that 77.5 percent of 498 sites were available, after an average of 22 months from the date of thesis submission.

Some of the difference in percentages of active and accurate links is due to the passage of time. Generally, *in the absence of other factors*, the greater the length of time that has passed since the date of the original citation, the more change the Web has undergone, with a resulting higher percentage of unstable links. This phenomenon has earned the tag of “linkrot.” When investigating student paper bibliographies over time, there may be other factors that influence the rate of available links. If, for example,
one looks at sets of papers from the same instructor in different years, and the instructor has introduced guidelines that address the type of Web site used, and/or the form in which it is submitted (for instance, electronic paper submission that decreases errors in the URL), then the average of viable links may increase over time, although the viability of specific, individual links would still be subject to time-induced instability.

The type of Web site cited is critical to the total of viable links. Certain domains undergo change at a higher rate than others, dot-coms being the most unstable. There is variability in persistence within the dot-coms themselves. News sites, for example, change on a daily basis and differ wildly in their archiving practices. Sales sites for medical services change less frequently than commercial sites for more consumable products. The brief comparison of the various studies of viable links noted above indicates that the citations in a published literature or database (LEXIS-NEXIS, law reviews, and the Internet Resources columns) had a slightly higher rate of viability than those in the undergraduate papers. That difference is due in part to the issue of site type; undergraduates use dot-coms heavily, and their bibliographies are correspondingly vulnerable to higher rates of link loss. Finally, variance in the averages between studies of link viability may reflect the effort expended by the investigator in pursuing links that at first try did not lead to the cited page; the use of a metacrawler, attempted correction of common typographical errors, and following embedded links to archived materials would result in higher rates of success than those of one who simply stopped with the first “no page found” message.

Link instability is a given. Two current solutions to the negative consequences of link instability for the electronic academic environment are persistent uniform resource locaters (PURLs) and the digital object identifier (DOI). PURLs provide a tracking history function that relocates the URL for the user, and the DOI provides a fixed or locked URL that will not change. DOIs are attached, for example, to full-text articles, which are then made more accessible to search and retrieval software. The push for standards in the digital landscape provides peripheral benefits for those concerned with the future of electronic scholarship, but it is backed primarily by the for-profit sector of electronic publishing and database companies. It is probably the case that firm standards will first come to prevail where link instability would be most injurious, as in science publishing (where DOIs are currently prominent), and in future legal and government environments. Such standards may or may not support the activity of more garden-variety researchers who do not avail themselves of scholarly, stable sites.

**Background/Methodology**

The following study investigates the citation behavior of undergraduates in multiple sections of freshman English composition at Northern Illinois University. Samples were taken in two different years from papers that were to be discarded the year following their submission. The papers were grouped by instructor, representing approximately 23 sections, and equal numbers of papers were extracted from each of the instructors represented. Due to increased availability, the second sample (2004: 350 papers) is larger than the first (1999: 179 papers). The bibliography pages were removed from the body of the paper, thus eliminating both personal identities and grades, and citations were first examined for number and type.

The categories of books, journals, newspapers, and government documents represent print sources. The print journals category does include magazine titles. As magazine titles were overwhelmingly available to these students as electronic resources, the number of print magazine titles was so low as to not merit its own label. The category of Web citations in-
cludes any citation to an electronic source, identified by the existence of a URL, or a reference to “online” or “Internet” where URLs were lacking. This category includes, then, citations to full-text journal, magazine, and newspaper articles. The parameters of this Web citation category differ from those found in studies of Web citations that are concerned with the quality of the Web resources used by students, where full-text analogues of print sources are identified, and/or Web sites may be categorized as scholarly or nonscholarly.¹⁵

The problems with access and impermanence inherent in the citation to an electronic resource with which this paper is concerned may be exacerbated by sites that fall into a low or nonscholarly category, but those problems may beset any posting to which a citation points, be it a prestigious journal or a lowly blog.

The citations to Web sites were then classified by type of domain, some of the more easily identified categories such as news sites, and full-text journal articles. Each identifiable URL was submitted manually to ascertain viability: that is, whether it still led to the page initially cited. No extra effort was made to resubmit or edit unsuccessful URLs, except in a few cases of obvious typographical errors (0 for o, lower case L for 1, left the “v” off gov). URL citations were tallied in four categories: yes, they successfully led to the cited page; no, they did not lead to the cited page (no attempt was made to tally the various error messages or other reasons for failure); site-no-page, where the URL led to the appropriate site, but the page had moved or otherwise changed; and, last, the URL either was not cited at all or was lacking enough information that it could not be submitted.

Instructor guidelines for these bibliographies were very similar in both years. Guidelines for the sections taught by graduate assistants (from which these samples came) are relatively fluid, but in conformance across sections. Papers averaged 8–10 pages in length, with an average of 8 citations. Rather than set limits on Web citations, students have generally been instructed to include minimum numbers of other, selected sources, such as two books and two journal articles. Some instructors dealt in class with the issues surrounding Web site evaluation, although not all did so. Instructor guidelines for the original papers that generated these bibliographies did not place restraints on students’ Web use that would have invalidated the use of citations to show general trends.

The academic resources that the students had available to them through the University Libraries collections include over 2 million volumes and approximately 3,200 paid serial subscriptions, split about equally between print and electronic titles. About 16,000 serial titles are represented at the present time in electronic form, including paid subscriptions and those acquired in aggregate form through database subscriptions (about 150 databases). At the time of the first sample, there were approximately 12,000 serial titles available electronically. The only other substantive change between the time of the first and second sample was the addition of the SFX software to facilitate cumulative retrieval of full-text and print holdings.

The English composition sections from which the bibliographies were extracted had all attended bibliographic instruction sessions in which they were given examples and specific practice in searching for appropriate resources through the library’s online catalog and sample general databases. Periodical Abstracts and Academic Elite were commonly used as practice databases. While the class sections all came to the library for a scheduled session, any given student might not, of course, have attended.

The first sample, consisting of 1,502 citations from 179 papers, was submitted at the end of fall term 1999 and collected in the spring of 2000. The URL tracking took place in the summer of 2000, approximately 7 months after submission. The second sample, with 2,538 citations from
350 papers, was submitted in the spring of 2004 and collected in the fall of 2004. URL tracking was completed somewhat later in this case—summer 2005—almost a year after submission.

The category of full-text articles merits some discussion. In the first, smaller sample, I was not concerned with, nor was I as aware of, the internal composition of URLs. They either worked or they didn’t. In the second sample, partly by virtue of greater numbers, it became apparent that citations to full-text articles were problematic. In some cases, the database origin of the article would show up in the character string (Firstsearch, or EBSCO). In other cases, through improper electronic citation style, the student might give the full citation information for the article but indicate that it was accessed electronically by listing only the name of the database and the date. So, that one would count as an online citation, but it couldn’t be tracked with the (nonexistent) URL. It seemed probable that there were cases where the student was using a full-text electronic source but citing it as if it were a print source. This situation may have occurred when the periodical title was locally available electronically but not in print. Access through interlibrary loan was ruled out, as our Information Delivery Services office will not process requests for which we have electronic access. And then there were those (few) cases, where the student had password access to a database or electronic collection which prohibited access through the cited URL. Actually, as will be discussed later, electronic full-text is not generally accessible through the URL, so for tracking purposes the URL was useless even where correctly rendered. The following procedure was followed in an attempt to impose some degree of uniformity on the category of electronically accessed journal. If the citation was to a journal article, but had no URL or other indication that it had been electronically accessed, it went into the Journal category. If the citation to a journal article included a URL, it went into the Web category, but a separate tally was kept of full-text online articles.

Newspapers were simply too confusing to try to categorize by Web citation as full text through database access (something like NewsSource) or directly from Web news sites. Some newspapers have print formats as well as electronic access, and others were obviously small or local online venues. If the citation was to a standard newspaper title and did not contain a Web citation, it went into the newspaper category (although I suspect that a portion of these were electronically accessed because the library had no other access to those titles). If the title indicated a news component (sports updates, for example) but could not be identified as a newspaper title, and had a Web citation, it went into the general Web category. If the Web citation was to one of the commonly identified online news sites (CNN, ABC, BBC), it went into a separate tally of News Sites.

Results
Differences by category of citation between the two samples demonstrated what might be expected, given students’ known predilection for easy access. Book and journal use fell over time, while use of electronic sources increased. Web cita-

![FIGURE 1](Figure1.png)

**FIGURE 1**

Percentage by Citation Category

- 60.00%
- 50.00%
- 40.00%
- 30.00%
- 20.00%
- 10.00%
- 0.00%

- 1997
- 1999
- 2004

- Books
- Journals
- Web
- Newspapers
- Gov Doc
- Other
- Unknown
tions went up dramatically, from 28.8 percent to 48.6 percent, well over half of the total (figure 1). For comparative purposes it is worth looking at an even older set of papers from the same program, the analysis of which dates to 1998, in which books and journals each comprised a little less than half of the total, and electronic sources only 1.2 percent. Newspaper use increased (2.9 percent to 6.5 percent), a possible reflection of the addition of two full-text newspaper databases. Government documents and the unidentifieds remained almost the same. The category of “other,” comprising films, interviews, court cases, and other media events, also did not change.

The 2004 sample yielded a couple of new categories, not large enough to register as percentages, and most interesting, perhaps, in that they are so small in comparison to the size of the potential resource. There were 2 instances of use of e-books (NIU subscribes to NetLibrary) and 8 references to online encyclopedias or dictionaries.

Analysis of Web Citations
Web citations reflect the behavior of two very different agencies. The first is a decision-making process on the part of a biological entity, the college student who selects resources from an array of formats. The second type of behavior that Web citations track is electronic, as it records the success or failure of URLs to lead to a specific point in cyberspace. Both of these behaviors have consequences for scholarship, but only one of them—the biological—can be altered within academe. The behavior of students can be modified by directives and sanctions to ameliorate some of the consequences of their choices. Sanctions (instruction-and-penalty intervention) have been shown to be much more effective than directives (instruction and/or encouragement intervention). This premise is supported by the behavior of the students in this study, who responded to the gentle encouragement incurred in their bibliographic instruction sessions by incorporating a mere fraction of the scholarly full-text available to them. There are also situations such as honors or upper division classes in which intervention is not indicated because the use of electronic citations is very low. Bearing in mind, however, that full-text articles in academic journals, accessed in aggregate electronic subscriptions, run the risk of having been moved or dropped two or three years down the line, the ramifications of electronic scholarship need to be viewed independently of the quality issue.

Behavior in the electronic medium can only be apprehended by a large and unwieldy consortium of corporations, foundations, governments, and other standardizing bodies. In the present study, there were numerous cases in which the original URL was automatically redirected to its new address. This makes good business sense within the commercial navigation structure of the Web, and as will be seen further on, undergraduates make overwhelmingly preferential use of commercial sites. Beyond the patchwork application of solutions, such as the use of DOIs in the science field, scholarly sites are not subject to a correspondingly cogent push to address linkrot. Although citation styles do not require the DOI designation, many students cut and paste the URL into their bibliography. There was not a single case of a DOI URL in the 1,666 citations in the student samples. Most of the literature concerned with linkage viability and its implications for general scholarship have identified and analyzed links within a published literature. Investigations of the Web behavior of students are frequently concerned with suggesting ways in which to modify the behavior of the user to mitigate the impact of poor site selection on the quality of the finished product. The majority of the literature on the Web behavior of undergraduate students is not overly concerned with the fact that their citations increasingly fail to lead to the sources cited. The citations of graduate students are another
matter, and, as Thomas’ study concludes, faculty advisors should be aware that the impermanence of this citation category may weaken research. But, just as today’s graduate students become tomorrow’s authors, producing the articles that stoke the boilers of Lexis-Nexis and Current Contents, so do many current undergraduates become graduate students.

If it is uncritically accepted that in great masses of undergraduate papers and a smaller mass of graduate papers, anywhere from one-third to one-half of the Web citations will not work a year later—that they effectively do not exist—then the currently somewhat smaller proportion of nonviable links in published papers will inevitably rise. The subsequent effectiveness of the citation record must suffer.

The Web citation behavior of the students is summarized in figure 2.

Dot-coms are the overwhelming choice, over half in both years. Dot-orgs went up, and other categories retained very similar proportions. Dot-coms can be a somewhat devious domain, in that many of the “preferred” sources (in other words, academic journal articles) are available in full-text format through databasing services that use a dot-com domain. EBSCO and Firstsearch, for example, are both dot-com entities. One truly depressing fact that emerged from this analysis with regard to journals was the apparent oblivion surrounding online articles. Only 39 of the 1,234 Web citations in 2004 could be identified as the digitized text of academic journals. This after a bibliographic orientation session that drew specific attention to them, and with a large selection of databases upon which to draw. To make the situation even worse, it is not possible to access the article by entering the URL string (generally formidable) for the full-text. An instructor wishing to access the article would have to reenter the citation information to call up the article directly from the appropriate database (or, if available, a citation finder such as SFX), or locate a print version of the text. The student benefits from the electronic advantage in a way that the instructor cannot.

One category of Web citation that saw increased use from 1999 to 2004 was that of news sites. For purposes of tallying, news sites were simply subsumed in the category of Web citations, but there were 29 news sites utilized in 1999 and 110 in 2004. News sites may be specific to a topic—sports, for instance, or conservation—or are an online analog to both major print and online news sources. News sites, by virtue of their function, can and frequently do change on an hourly basis and are a major culprit in the site-no-page category. They are next to useless for purposes of tracking a source. In the present sample, some archived their articles (old sports news articles could actually be located on some sites), others archived some types of documents, and others did not archive. Their appeal may lie largely in their authority—more specifically, in the brevity of their authority.

There were changes in the ways students used Web sites that were not apparent in the citation. In 1999, there were more unintelligible Web citations, in which the students simple put “online article” or “Internet.” This is the activity behind the difference in the “no URL” category in figure 2. The bad poster child
in 1999 was the bibliography reproduced here in its entirety.

I only had one source and that was the Internet.
P.4-Internet
P.6-Internet
P.48-Internet
P.49-Internet
P.50-Internet

By 2004, students used correct styles for their Web citations with the appropriate information appended. However, there was a new form in which students cited an entire Web site, and the link led to a multi-tiered site with numerous interior links, with no information regarding what actually came from where. This phenomenon is compounded with the passage of time, where the link continues to point to the site, but the interior arrangement of the site and its posted information have changed. This is a different case from the site-no-page links. In the latter case, the citation refers to a discrete chunk of information, such as an “article” — “Atkins Diet” — and the link led to the ABC news site that had run an article on Atkins, which was no longer available. The whole-site citation was to the “Atkins homepage” at such and such a link. A trivial enough distinction, except that it means that another portion of the “found” links are in reality “site-no-page” links.

Taking a “time snapshot” in 2005, of the two samples from 1999 and 2004, we have a situation in which 91 percent and 55.8 percent, respectively, of the links do not do the job they were intended to do.

More important, they do not fulfill the function that their very presence mandates. Of those citations, 91 percent and 55 percent are pure form, with no substance.

Conclusions
Citation structure provides a detailed map of the specific path followed by the writer. Over half of the students’ bibliographic entries were, in mapping terms, only as effective as a declaration of direction. They went west, or south — how far, and to what fixed point, we do not know. From instructor comments on the bibliography pages, it was clear that points were deducted for incorrect citation style.

Link Viability
Figure 3 indicates the percentages of links that led to the cited document. Between 1999 and 2004, the percentage of successful links increased by almost 10 percent. The not-founds and the site-no-page links were reduced by about 5 percent. A resubmission of the successful 1999 links during the same time window that the 2004 links were submitted resulted in a reduction of the found links in the 1999 sample to 8.5 percent (figure 4). This is in conformance with the deterioration rates proposed by Koehler.
Citation formatting was something that incidentally improved considerably between the two samples. The corpus thus formally clothed, however, was sadly insubstantial.

So back to the question originally posed. Does it matter? Undergraduate papers are ultimately discarded, by departments or by their authors. In many cases they serve a pedagogical function quite independent of the content of the sources used. The principles of writing, disquisition, and argument can be taught and practiced with complete indifference to the validity and authority of sources. Even within a published literature—science, social science, or humanities—the accepted truths and premises of a discipline change over time. And that is precisely why those references cited at the end of a paper fulfill a function for western scholarship that is independent of the individual identities of the entries. The volume of the citations acts as a provision that upholds the authority of the cognitive process that produced the creation they accompany. That was the original intent of the inclusion of citations in the history of written scholarship, the proof that this was a new form, neither plagiary nor borrowing, but building on the work of others who receive proper credit.

Borrowing is an essential element of learning. Classical writers were sent to their predecessors for more than inspiration. They borrowed subjects and materials as well, and it wasn’t until the 16th century that there was a real outcry against the practice of flattering imitation. It is still the case that students learn what is acceptable by reading and writing empirical reports framed by a discipline’s stylistic conventions. In an era when form is often cunningly arrayed to obscure substance, it is even more imperative that the published trail of evidence not be obscure. The modest bibliography works to that end and is the only truly universal venue to accomplish transparency.

Today’s undergraduates are tomorrow’s graduates, and they will populate the labs and academies of the future. If it is acceptable that over half of our students’ trails do not lead in the promised direction, then the percentage that behaves similarly in the body of published scholarship will passively increase. It will change how we accept that we know something, and it will change also the scales we use to measure validity and authority. If, on the other hand, we can make a case that linkrot does matter at the undergraduate level, if instructors and students can be made aware that the deficiencies exist, then improvements in student decision-making and in encouraging Web standardizing solutions may be stimulated. Internet copyright issues are undergoing similar scrutiny. The impetus for change came with the realization that there was a growing misfit between new forms and old rules. Web citation behavior is at a comparable juncture. There are tangible problems with new forms and we cannot afford simply to throw out the rules.

Notes


11. Mary K. Taylor and Diane Hudson, “‘linkrot’ and the Usefulness of Web Site Bibliographies.”

12. Phillip M. Davis, “Effect of the Web on Undergraduate Citation Behavior: Guiding Student Scholarship in a Networked Age.”


17. Phillip M. Davis, “Effect of the Web on Undergraduate Citation Behavior: Guiding Student Scholarship in a Networked Age”; Andrew M. Robinson and Karen Schlegl, “Student Bibliographies Improve When Professors Provide Enforceable Guidelines for Citations.”


ANNOUNCING: CRO2

Choice Reviews Online – Version 2

Comprehensive Content
115,000 Searchable Reviews
Enhanced Search Capabilities

Improved Interface
More Personalization
Web Exclusives

CRO2 will replace both the Password and Site License Editions of ChoiceReviews.online, combining the best features and functions of both versions!

FOR MORE INFORMATION VISIT WWW.CRO2INFO.ORG