Web Citation Availability: Analysis and Implications for Scholarship

Mary F. Casserly and James E. Bird

Five hundred citations to Internet resources from articles published in library and information science journals in 1999 and 2000 were profiled and searched on the Web. The majority contained partial bibliographic information and no date viewed. Most URLs pointed to content pages with "edu" or "org" domains and did not include a tilde. More than half (56.4%) were permanent, 81.4 percent were available on the Web, and searching the Internet Archive increased the availability rate to 89.2 percent. Content, domain, and directory depth were associated with availability. Few of the journals provided instruction on citing digital resources. Eight suggestions for improving scholarly communication citation conventions are presented.

any students regard citations as annoying details with little relevance to their work. However, individuals conducting serious research understand that long-established citation conventions help them further their own scholarship and assess the validity of other works in their field. Through citation, "researchers generously acknowledge their debts to predecessors."[1] Collectively, appropriate and accurate citations document how established scholarly works build on one another over time to transform ideas and even entire fields of inquiry.

Literature Review

Citation accuracy is critical to accessibility, and prior to the development of the World Wide Web (Web), it was well studied and documented both across and within academic disciplines. In a 1992 doctoral dissertation, Catherine Jean Sassen examined citation error case studies dating back to the mid-1800s and determined that, in the literature of the last 150 years, "citation error is a widespread problem that has impaired access to information."[2] In the early 1990s, Idrisa Pandit, Susan P. Benning and Susan C. Speer, and Nancy N. Pope conducted studies of citation errors within the library science literature and found error rates of 18 to 29 percent.[3]

More recently, researchers have focused on the growing reliance on the Internet as a source of information and on the increasing frequency with which authors cite Web sites and pages to document their scholarly research. Carol Anne Germain has suggested that Web documents published by organizations, associations, and individu-

Mary F. Casserly is Assistant Director for Collections at the University at Albany, SUNY; e-mail: mcasserly@uamail.albany.edu. James E. Bird is Head, Science and Engineering Center, in the Raymond H. Fogler Library at the University of Maine; e-mail: jim.bird@unit.maine.edu. The authors wish to thank Mary Bird, Instructor in Education, University of Maine, for her editorial assistance; Phillip Pratt, Associate Director of Institutional Studies, University of Maine, for his advice on statistics; and the two anonymous reviewers for their constructive suggestions.

als more closely resemble print fugitive materials and gray literature than the mainstream monographs and refereed journals that have for so long been the backbone of formal scholarly communication.[4] This resemblance is because these documents can easily be modified and overwritten and, in many cases, their authors are not committed to long-term storage and maintenance. They are, as Wallace Koehler described them, somewhere between "ephemera and permanent."[5] Given these limitations, it is important to examine the implications of authors' patterns of citing Web documents and to consider carefully how to integrate information residing on the Web into scholarly communication conventions.

A number of researchers have described the size and volatility of the Web, and their work provides a context for this study. In 1999, Koehler studied the permanence and constancy of a random sample of 361 Web pages and 343 Web sites and determined that they underwent measurable changes in content and availability over the year in which the study was conducted. Koehler also investigated whether site and page size, object dominance (i.e., a way of classifying Web sites by its function), domain, and various other URL markers could help predict permanence and constancy and found that inferred domain and object dominance may provide understanding of Web page behavior.[6] Koehler's 2002 analysis of the same sample of Web pages and sites over a longer, four-year period confirmed many of his earlier findings, including domain as a predictor of persistence and the Web page half-life of two years.[7]

Judit Bar-Ilan and Bluma C. Peritz conducted a study of the data on the Web in the field of informetrics. Although their sample was more stable than Koehler's, they noted that in each round of searching the character of their subject on the Web was slightly different, with documents appearing, disappearing, and changing.[8] Three percent of the digital library information objects studied by Michael L. Nelson and B. Danette Allen were no longer available, and over the period of their study the objects they downloaded changed from their baseline size 22 percent of the time.[9] Bing Tan, Schubert Foo, and Siu Cheung Hui found that 44.8 percent of the Web pages they tracked changed and 3.8 percent disappeared during the course of their study. They also found that pages with education, business, and entertainment domains were less likely to change than were those published in other domains and that text, organizational, and database pages were the most stable page types.[10] John Markwell and David W. Brooks's study of the persistence of URLs with scientific or science education content revealed that, over a fourteen-month period, 46.5 percent had either changed content or were no longer available. Their analysis of availability by domain indicated that "gov" was the most viable, followed in order by "edu," "com," and "org."[11]

In addition to examining the stability of Web sites, researchers also have begun to explore the availability, longevity, and character of scholarly references to content published on the Web. In a study conducted early in the development of the Web, Yasar Tonta found only two references to networked information in his sample of articles published in 1930 and 1994 from twenty-seven journals covering a wide range of subjects. He concluded that "networked information sources in the form of electronic journals and archives get almost no citations in print journals at all."[12] As part of their 1996 study of the impact of electronic journals on the scholarly communication process, Stephen P. Harter and Hak Joon Kim analyzed 4,317 references from a sample of 279 articles published in scholarly and peer-reviewed electronic journals and found that 1.9 percent of the references cited such electronic resources. Of the forty-seven cited references that included URLs, only two-thirds led to the text of the source, despite the fact that the data gathering took place during the same year that most of the references included in this study appeared.[13]

Philip M. Davis and Suzanne A. Cohen's study of references in undergraduate term papers indicated that the number of Web documents cited by students increased 12 percent from 1996 to 1999. This increase was

accompanied by a dramatic decline in the ability to access citations included in the older papers. The percentage of cited URLs that could not be accessed was 16 percent in papers written in 1999 but rose to 53 percent in those written in 1996.[14] In a follow-up study, Davis found that a 16 percent inaccessibility rate also applied to URLs included in papers written in 2000.[15]

Studies of URL persistence published between 1998 and 2000 by S. Mary P. Benbow, Germain, Joel D. Kitchen and Pixey Anne Mosley, Susan Davis Herring, and Mary K. Taylor and Diane Hudson identified URL availability rates ranging from a high of 89 percent to a low of 50 percent.[16] By guessing at alternate URLs or browsing the Web, Steve Lawrence and others were ultimately able to locate all but three percent of a sample of initially unavailable URLs cited in computer science journal articles and conference reports.[17] In a study of "linkrot" in law review journal articles published from 1997 to 2001, Mary Rumsey found an availability rate that declined from 61.80 to 30.27 percent, an increase in the number of Web citations per article, and a lack of parallel citations to paper sources. Rumsey's data also indicated that home pages were more likely to be persistent than document-like pages.[18]

Yin Zhang examined the electronic sources cited in ten library and information science journals from 1994 to 1996 and found that 1.13 percent of the total references were e-references (i.e., references to electronic resources). Zhang's data also indicated that there was no significant difference in the proportion of e-references by year and that articles published in electronic journals had significantly higher e-reference rates than those published in print journals.[19] In a follow-up study, Zhang found that the rate of e-sources cited in print journals had increased from 0.2 to 5.2 percent between 1991 and 1998, whereas the percent of articles containing such citations rose from 1.8 to 33.9 percent.[20] In Zhang and Leigh Estabrook's 1998 study, only 30.4 percent of the e-sources cited from 1990 to 1994 were still accessible, whereas 82.2 percent of those

cited in 1996 were accessible. For papers that were "in press" as of February 1998, that figure was 81.5 percent. They also found that the access rate varied by journal format, with the e-sources cited in electronic journals being more accessible than those cited in print journals.[21]

These studies represent the growing body of research aimed at describing the extent to which scholars use Web documents and integrate them into the formal communication of their research. Researchers also have begun to explore the implications of the problems of access to cited electronic references for future scholars. However, it is not clear to what extent the publishers and editorial staffs of scholarly publications are concerned about the availability of cited electronic resources over the long term. Zhang's 2001 study surveyed the editors of eight library and information science journals and found that, although they encouraged authors to cite electronic resources, they had only begun to work on policies relevant to this practice. Indeed, Zhang's review of the journal guidelines and instructions to the authors revealed an absence of clearly stated policies and/or guidelines regarding citing electronic resources.[22] The researchers could not locate any other literature or studies that explored journal policy guidelines on citing information and documents published on the Web.

Purpose of the Study

The purpose of this study is to add to the body of knowledge about the changing landscape of scholarly communications by examining citations to Internet resources included in research articles published in the library and information sciences literature. Specifically, this study addresses the following questions:

• To what extent are authors currently referencing information and documents "published" on the Web?

• What percentage of cited electronic resources are available for consultation by future scholars? How are they most often found?

• Is it possible to identify characteristics of citations to Internet resources that

TABLE 1
Journals Included in the Study

Title	# Articles Included	Total # Citations	Average Citations/Article	# Web Citations	%Web Citations
Art Documentation	21	173	8.2	52	30.1%
Art Libraries Journal	58	470	8.1	89	18.9%
ASLIB Proceedings	78	1,443	18.5	294	20.4%
Catholic Library World	23	323	14.0	48	14.9%
College & Research Libraries	73	1,426	19.5	128	9.0%
Electronic Library	44	692	15.7	138	19.9%
Government Information Quarterly	45	1,156	25.7	192	16.6%
Information Processing & Management	81	2,536	31.3	97	3.8%
Information Services & Use	18	312	17.3	35	11.2%
Information Technology and Libraries	30	434	14.5	178	41.0%
Information Society	49	1,640	33.5	199	12.1%
Journal of Academic Librarianship	60	1,507	25.1	157	10.4%
Journal of Documentation	35	1,306	37.3	147	11.3%
Journal of Education for Library and Information Science	41	828	20.2	77	9.3%
Journal of Government Information	42	1,307	31.1	203	15.5%
Journal of Information Ethics	22	676	30.7	45	6.7%
Journal of Information, Law and Technology	22	639	29.0	234	36.6%
Journal of Information Science	88	2,272	25.8	327	14.4%
Journal of Librarianship and Information Science	33	837	25.4	67	8.0%
Journal of the American Society for Information Science	141	5,042	35.8	188	3.7%
Journal of Youth Services in Libraries	25	381	15.2	17	4.5%
Knowledge Organization	22	397	18.0	29	7.3%
Libraries & Culture	43	2,572	59.8	12	0.5%
Library Administration & Management	48	420	8.8	40	9.5%
Library & Information Science Research	36	1,445	40.1	85	5.9%
Library History	23	578	25.1	1	0.2%
Library Philosophy and Practice	12	145	12.1	25	17.2%
Library Quarterly	25	1,309	52.4	49	3.7%
Library Resources & Technical Services	23	497	21.6	76	15.3%
Libri	57	1,353	23.7	159	11.8%
Public Libraries	32	426	13.3	50	11.7%
Reference & User Services Quarterly	35	689	19.7	106	15.4%
Research Strategies	12	223	18.6	10	4.5%
Technical Services Quarterly	28	235	8.4	28	11.9%
Total	1,425	35,689	25.0	3,582	—

will help predict the availability of the content to which they refer?

• What type of guidance are authors receiving from editors and publishers?

Based on these findings, the researchers offer suggestions for updating scholarly communication citation conventions.

Methodology

The researchers chose to work with the literature of library and information science, the academic discipline in which they were trained. They anticipated that their knowledge of the subject area would be useful in searching the Web for content not found at the cited URLs. In addition, they believed that the publishing conventions used in the library and information science literature were similar to those used in other social science literatures and that, consequently, their study's findings could be extended to those disciplines.

The journals reviewed for this study were selected from the "core" list of library and information journals published in the tenth edition of *Magazines for Libraries*.[23] These core titles were examined and newsletters, bulletins, magazines, and other non-peer-reviewed titles were eliminated because it was unlikely that they would serve as a broad basis for future scholarship. In addition, several scholarly journals were excluded from the study because they were not available to either researcher and one was eliminated because its formatting practice of printing references at the bottom of each page, rather than in a list at the end of each article, posed overwhelming logistical problems for the researcher counting nonredundant citations. The remaining thirty-four journals that served as the source of articles and citations included in this study are listed in table 1.

The study was limited to citations appearing in research-level articles. Excluded, therefore, were book reviews, editorials, opinion pieces, conference reports, and other types of articles not generally subject to peer review. Similarly, articles appearing in retrospective and anniversary issues of these journals were omitted from this study. Most articles appearing in special issues of these journals also were excluded because they generally are invited, rather than refereed, papers. The researchers counted the nonredundant citations to Web and non-Web resources in the remaining 1,425 research articles published in the 1999 and 2000 volumes of the table 1 titles. These data were entered into an Excel spreadsheet from which the descriptive statistics presented in tables 1 and 2 were derived.

To study the availability of cited Web resources, the researchers selected a random sample of 500 from the 3,582 citations to Web resources that appeared in these 1,425 research articles. A random sample was determined to be appropriate for this study after an examination of the journals' instructions to authors revealed that the authors had been neither encouraged to cite, nor discouraged from citing, documents and information residing on the Web. The sampling error for a sample size of 500 is ± 4.0 percent.[24]

For each citation included in the sample, descriptive data on the source journal, the content of the citation, and the URL domain and directory depth (i.e., the number of levels within the URL's directory structure) were collected. The researchers then began the process of determining content availability by keying each URL in the sample into Internet Explorer 5.5 or 6.0. A URL that pointed to the Web page containing the information referenced in the article or to a referring page leading to that information was considered "permanent." When the cited URL did not lead to the referenced information, the researchers checked the URL for typographical, syntax, and other obvious errors. If an error was found, they corrected it and determined whether the corrected URL would now lead to the cited information. If they were still unable to locate the cited content, they attempted to locate it elsewhere on the Web site by entering the URL into their browser again and then removing one directory level at a time until a Web site connection was made and /or by going to the home page of the site and employing any available directories, maps, or internal search engines to locate the cited content.

TABLE 2
Citation Frequency, Range, Mean,
Median, and Mode

Frequency # Citations	Web Citations # Articles	Paper Citations # Articles	Total Citations # Articles
0	755	13	—
1–5	467	239	186
6–10	113	244	208
11–20	63	367	383
21–30	13	223	244
31–40	9	143	158
41–50	5	73	103
51–60	—	37	49
61–70	—	24	22
71–80	—	27	35
81–90	—	8	9
91–100	—	7	8
101+	—	20	20
Total	1,425	1,425	1,425
Range	0–50	0–291	1–291
	# Per Article	# Per Article	# Per Article
Mean	2.5	22.5	25.0
Median	0	16	19
Mode	0	1	10

The researchers searched the Web using Google (http://www.google.com) for cited content that could not be found using the previously described methods. Google was selected as the search engine for this process because of the large number and variety of documents to which it provides access and because the researchers believed that its relevancy ranking would be effective for the types of narrowly defined searches they would be conducting.[25] The researchers performed up to five Google searches using different combinations of titles, keywords, author names, and source information. If none of these searches returned the cited content in the first twenty-five results, that content was considered to be inaccessible. It should be noted that when the researchers encountered any type of message indicating that the URL was unavailable or that the file/page could not be found, they waited at least a week and repeated the search process.

The researchers also ascertained whether each of the URLs in the sample had been archived by the Internet Archive, a "public nonprofit 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."[26] To determine whether the referenced content was archived by the Internet Archive, the researchers entered the URL into the Wayback Machine (http://www.archive.org). From the displayed menu, they selected the appropriate date. The appropriate date was defined as the one closest to the date the author of the article viewed the cited Web content. When a "date viewed" was not included in the citation, the appropriate date was defined as the one closest to the date the article was published.

The researchers faced many challenges in trying to determine whether content at a given URL matched that viewed by the author of the article in which it was cited. For citations that included full bibliographic information and the date the author viewed the cited content, the researchers were able to determine with some certainty whether the content of the Web page matched the cited information. When the citation was less complete, the match of current content to cited content was less certain. In such cases, the researchers exercised their judgment based on the bibliographic information provided, the subject of the article, and the URL server and file names. In cases where the citation consisted only of a URL, the researchers considered the cited content to be available if the URL linked, either directly or through a referred page, to a working Web page that seemed to be consistent with the text included in the URL and/or the subject of the article. In searching both the Web and the Internet Archive, the researchers relied on the "date viewed" to match the content found to the cited content. However, these dates rarely matched those in the Wayback Machine results list and therefore the researchers could not determine with absolute certainty that the content they viewed matched that viewed by the authors.

The data on each of the citations in the sample were collected between January and July 2002. The statistical program, SPSS 10.1 for Windows, was used to generate contingency tables and calculate the Pearson's Chi-Square values. A $p < .05$ level of significance was used for this study.

Findings

The 1,425 research articles that formed the source of the sample citations used in this study contained a total of 35,689 citations. Of these, 3,582, or 10 percent, referenced information or documents residing on the Web, and 90 percent referenced paper or other nondigital resources. The distribution of these citations among the journals scanned for this study is presented in table 1. The percentage of citations that included URLs varied greatly among the journals, from a low of 0.2 percent in the articles appearing in *Library History* to a high of 41 percent in articles published in *Information Technology and Libraries*.

The citation frequencies, means, medians, and modes are presented in table 2. The average number of citations per article was 22.5 non-Web citations, 2.5 Web citations, and 25.0 total citations. For all citation categories, the medians are substantially lower than the means. In the case of the Web citations, this is an indication of the influence of the large number of articles, 755 or 53 percent, with no Web citations. Indeed, the median and the mode for this category are zero. The mean of the non-Web citations is influenced by a small number of articles that have extreme numbers of citations, including twenty that have more than a hundred.

The researchers reviewed the "Instructions for Authors" published in the journals from which the sample was drawn for the period of the study (1999–2000) and, again, as this manuscript was being prepared in order to determine whether these journals had established policies or instructions on citing digital resources. In contrast to the explicit instructions presented for tables and illustrations, the researchers found few instructions for citing digital resources published in the journals or on their Web sites. Only six of the thirty-four journals included examples of citations to electronic resources for authors to follow. Three of these also provided further instructions on citing Web resources. One additional title referred authors citing content on the Web to the American Psychological Association's Web site, APAStyle.org.[27]

Fifteen of the thirty-four journals included in this study referred authors to the fourteenth edition of *The Chicago Manual of Style,* which, having been published in 1993, does not address references to digital resources.[28] Other style manuals, including those published by the American Psychological Association (APA) and the Modern Language Association (MLA), do provide guidelines for citing electronic resources and instruct authors to include a date of publication or last revision for the Web page cited and/or the date the author last accessed or viewed the cited URL.[29] None of the "Instructions for Authors" pages in the journals studied addressed Web site permanence.

Citation Characteristics

Of the 500 sample citations, 499 included URLs that pointed to hypertext resources (i.e., those beginning with "http"). The other citation was to a listserv message. Data on variables related to the source and content of these citations and their URLs are presented in table 3.

More than 92 percent of the 500 citations in the sample were drawn from journals published in print format only or in print format with an electronic counterpart; 7.8 percent came from journals born digital. The citation content ranged from a URL only to a URL accompanied by complete bibliographic information. Thirty-one of the citations, or 6.2 percent, consisted only of URLs, 51 percent contained partial bibliographic information, and 42.8 percent were considered complete by the researchers. A citation was considered complete if it included, at a minimum, a title, publisher, and date of publication. Almost two-thirds (65.6%) of the citations did not include the date the author viewed the resources she or he cited.

The analysis of the URLs by their original and implied domains is similar to that

conducted by Koehler in his study of Web site and page persistence. The original domains are those that are included in the URL as it appears in the citation. Almost 28 percent of the URLs in the sample had top-level domains with geographic designations (i.e., two-letter country codes such as "au" for Australia, "cn" for China, and "it" for Italy), whereas 21.8 percent of the cited content resided on organizational, 19 percent on commercial, and 18.8 percent on educational servers. Nine percent of the content resided on government servers. Content on military and network servers and those on servers identified by Internet Protocol Number (IPN) represented less than four percent of the total.

The purpose of creating implied domains for the URLs in the sample citations was to categorize as many as possible according to the purpose of the organization hosting the content to which they refer. Although the URLs in the sample citations with generic top-level domains (g-TLD) (i.e., "com," "edu," "gov," "mil," "net," and "org") have the same original and implied domains, the researchers translated those with country code top-level domains (cc-TDLs) into g-TDLs. In this process:

ccTLDs that are identifiable as commercial (e.g., co.jp), academic (e.g., ac.uk), government (e.g., gob.mx), organizational (or.cr), or network (net.de) are folded into the gTLD classification of com, edu, and so on.[30]

Those ccTLDs that could not be reclassified were left in the "geographic designation" category. This reclassification resulted in a shift toward the "edu" domain, with 33.8 percent of the cited URLs having that implied domain.

The directory structure of the URLs in the sample citations ranged from the zero-, or server-, level domain (http://aaa.bbb.cc/) address to the seventh level (http://aaa.bbb.cc/ttt/uuu/vvv/www/xxx/yyy/zzz). Seventy-six, or 15.2 percent, of the URLs in the sample citations had no directory structure (zero level), whereas URLs with a second- or third-level structure comprised more than half of the sample. The URLs also can be categorized as either navigation or content Web pages. Navigation pages, most often found at the server level and the first-directory level, are those that help users navigate to the information the site provides, whereas content pages, usually found at the second level and above, are those that provide that information.[31] The URLs in the sample were reclassified as navigation and content pages. Nearly three-quarters of the citations in the sample (72.4%) pointed to content pages, whereas 27.6 percent pointed to navigational pages.

In conjunction with a personal name, the tilde (~) is used to indicate that individual's home directory on the server of an Internet Service Provider. "In real terms the tilde stands for a path which leads to that person's Web site on the server it is being kept. For example, http://www/best.com/~erinj says that erinj is a best.com user and that her home page is on best.com's server."[32] Thirty-seven, or 7.4 percent, of the URLs in the sample included tildes, suggesting that the content cited is maintained by an individual rather than an institution, organization, or other entity.

Content Availability: URL Permanence

For the purposes of this study, cited content was considered available if it was found either at the URL included in the sample citation (permanent) or elsewhere on the Internet (accessible). The data on availability are presented in tables 4 and 5.

The researchers considered a URL to be permanent if it led to the Web page containing the content the author cited or a Web page that referred the researcher to the page containing the cited content. As the data in table 4 indicate, 282, or 56.4 percent, of the sample URLs were found to be permanent. Of the 500 citations studied, 213, or 42.6 percent, could not be found at the URLs cited and therefore were considered to be impermanent. In five of the sample citations, the citations' text and the content to which their URLs led were in Dutch or Greek and the re-

TABLE 3
Sample Citations Characteristics

Characteristics	# Citations	% Citations
Source Journal		
Print or print with electronic counterpart	461	92.2
Electronic only	39	7.8
Total	500	100.0
Content		
URL only	31	6.2
URL and partial bibliographic information	255	51.0
URL and complete bibliographic information	214	42.8
Total	500	100.0
Date Content Viewed by Author		
Included	172	34.4
Not included	328	65.6
Total	500	100.0
URL Original Domain		
com – Commercial	95	19.0
edu – Education	94	18.8
gov – Government	45	9.0
mil – Military	3	.6
net – Network	12	2.4
org - Organization	109	21.8
IPN – Internet Protocol Number	1	.2
Geographic designation	139	27.8
Other	2	.4
Total	500	100.0
Implied Domain		
com - Commercial	111	22.2
edu - Education	169	33.8
gov - Government	62	12.4
mil – Military	3	.6
net – Network	13	2.6
org - Organization	120	24.0
IPN – Internet Protocol Number	1	.2
Geographic designation	19	3.8
Other	2	.4
Total	500	100.0
Directory Depth		
0	76	15.2
1	62	12.4
2	138	27.6
3	119	23.8

TABLE 3 (CONTINUED)		
Sample Citations Characteristics		
Characteristics	# Citations	% Citations
Directory Depth		
4	75	15.0
5	18	3.6
6	8	1.6
7	4	.8
Total	500	100.0
Page Type		
Navigation	138	27.6
Content	362	72.4
Total	500	100.0
Tilde (~) in URL		
Included	37	7.4
Not included	463	92.6
Total	500	100.0

searchers could not determine whether they matched.

Content Availability: Accessibility on the Web

The researchers considered the cited content to be accessible if, after failing to find it at the URL included in the citation, they were able to locate it elsewhere on the Web. The results of the researchers' efforts to find the content referred to by the 213 citations that did not have permanent URLs are presented in table 5.

The researchers found content cited in eight, or 3.8 percent, of the 213 sample citations that were not permanent by truncating the URL in the citation, and they identified nine errors that, when corrected, led to the cited information. They located content cited in fifty-four, or 25.4 percent, of these citations by browsing or searching the site to which the URL led them and found content cited in an additional fifty-four by using the Google search engine. The researchers failed to find the content cited by eighty-three, or 39 percent, of the 213 impermanent URLs. This is 16.6 percent of the 500 citations in the sample. The five cases categorized as "could not determine" represent cited content in a foreign language in which the researchers were not conversant. These are in addition to the five described in the previous section and presented as "could not determine" in table 4.

Content Availability: "Not Available" Messages

The researchers received some type of URL or file "not available" message for 158, or 31.6 percent, of the URLs in the 500 sample citations. These were searched a second time, and as the data in table 6 indicate, three, or 1.9 percent, of these 158 led to the content cited and were considered permanent. Ninety, or 57 percent, of the URLs in the sample for which the researchers received a "not available" message were eventually found on the Web by truncating the URL, browsing or searching the Web site to which the URL led, correcting an error in the URL, or using Google. These were considered accessible. Sixty-five, or 41.1 percent, of these 158 URLs were not found on the Web.

Content Availability: Internet Archive

The researchers searched the Internet Archive using the Wayback Machine to

TABLE 4
Content Availability: URL Permanence

Content at Cited URL	# URLs	% URLs
Found	282	56.4
Not found	213	42.6
Could not determine	5	1.0
Total	500	100.0

determine whether the URLs included in the sample citations had been archived. The results are presented in table 7. The researchers found 344, or 68.8 percent, of the URLs in the 500 sample citations in the Internet Archive. This includes 84.8 percent of the content found at the cited URL (permanent) and 50.8 percent of those found elsewhere on the Web (accessible). Further, by using the Wayback Machine, the researchers were able to access an additional thirty-nine cited Web pages. This is 47 percent of the eighty-three URLs in the sample citations that were neither permanent nor accessible.

Characteristics Associated with Availability

The researchers ran a series of cross-tabulations on SPSS to try to identify the characteristics of the cited URLs that could be associated with URL permanence and content availability on the Web and in the Internet Archive. Chi-Square Tests of Independence were performed to identify the statistically significant relationships. To run these tests, it was necessary to filter out some of the missing data and/or reclassify some of the variable values into broader categories. The results of the Chi-Square tests are presented in table 8. Cross-tabulations were not run on citation content and availability in the Internet Archive variables because the Wayback Machine only accepts URLs and, therefore, the presence or absence of additional bibliographic information in the

citation could not affect, or be associated with, availability in that archive.

Unlike Zhang and Estabrook, who found that citations from articles published in electronic journals were more likely to be permanent than those from articles in print journals, this study found that permanence and the source journal were independent variables.[33] In other words, the source of the citation was not an indication of whether the resource could be found at the URL included in the citation. The presence of a tilde and the page type, both studied by Koehler as possible predictors of permanence, as well as the inclusion of the date the author viewed the content, were independent of permanence. However, the Chi-Square tests indicate that citation content, URL domain, and URL directory depth were associated with content availability.

The cross-tabulations for the characteristics with Chi-Square values that were significant at the $p < .05$ level are presented in table 9. The cross-tabulation between citation content and permanence indicates that URLs in "URL only" citations were found to be permanent more often than URLs accompanied by partial or complete bibliographic information. Specifically, 82.8 percent of the URLs in the "URL only" citations were found to be permanent, whereas the permanence rates for URLs accompanied by partial and complete bibliographic information were 58.1 and 52.1 percent, respectively.

The cross-tabulations of domains with content availability suggest that content at URLs with original domains of "edu" and "org" is more likely to be permanent or accessible than is content located on other

TABLE 5
Content Availability: Accessibility on Web

Content on Web	# URLs	% URLs
Found by truncating URL	8	3.8
Found by correcting error in URL	9	4.2
Found by browsing or searching Web site	54	25.4
Found by using Google	54	25.4
Not found	83	39.0
Could not determine	5	2.3
Total	213	100.0

TABLE 6
Content Availability: URL or File
Not Available Messages

Content on Web	# URLs	% URLs
Found at cited URL	3	1.9
Found by truncating URL	8	5.1
Found by browsing or searching Web site	40	25.3
Found by correcting error in URL	3	1.9
Found by using Google	39	24.7
Not found	65	41.1
Total	158	100.0

types of servers. Almost 90 percent of the content cited by URLs on organizational servers was found at the URL cited or elsewhere on the Web. This was the case for 87.9 percent of the content cited by URLs on educational servers. Content cited by URLs with "edu," "org," and geographic designation original domains also is more likely to be found in the Internet Archive. Three-quarters of the "edu," "org," and geographic designation original domain URLs were found in that archive.

When the domains are reclassified from original to implied, content at "edu" and "org" servers is most likely to be permanent and permanent or accessible. The permanence rates for content cited by URLs with "edu" and "org" implied domains were 64.5 percent and 64.2 percent, whereas 89.3 percent of the content at URLs with "edu" implied domains were found at the URL cited or elsewhere on the Web. This was the case for 90 percent of the URLs with "org" implied domains.

The cross-tabulations between directory depth and availability on the Web and in the Internet Archive suggest that the relationships are inverse and nonlinear. Content cited by URLs with five or more levels is less likely to be permanent, permanent or accessible, and available in the Internet Archive than is content cited by URLs with zero to four levels. For content cited by URLs with five or more levels, 30 percent was found at the URL cited, 73.3 percent was found at the URL cited or elsewhere on the Web, and 43.3 percent was found in the Internet Archive. The availability rates for content cited by URLs with zero to four directory levels were uniformly higher than those for URLs with five or more levels. However, these rates do not consistently decrease as the levels increase. For example, 62.2 percent of the content cited by URLs with four levels is permanent in contrast to 56.4 percent of that cited by URLs with three levels. It should be noted, too, that page type is based on directory level. Navigation pages are those found at the server and first levels; content pages are those found at the second level and above. The fact that page type was not associated with availability on the Web or in the Internet Archive supports the idea that the relationship between directory depth and availability is nonlinear.

Content Availability: Researcher Skill

A Chi-Square test was run on content availability and researcher to determine whether researcher skill or ability to find content on

TABLE 7
Content Availability: Internet Archive

Accessible in Internet Archive	All Citations		Permanent URLs		Accessible Content		Content Not Found	
	#	%	#	%	#	%	#	%
Found	344	68.8	239	84.8	66	50.8	39	47.0
Not found	146	29.2	42	14.9	60	46.2	44	53.0
Could not determine	10	2.0	1	.4	4	3.0	0	0.0
Total	500	100.0	282	100.0	130	100.0	83	100.0

TABLE 8
Summary of Pearson's Chi-Square (X²) Values
Citation Characteristics and Content Availability

Characteristics	Permanent			Permanent or Accessible			Archived		
	df	X^2	p	df	X^2	p	df	X^2	p
Source journal	1	.559	.455	1	.073	.787	1	.754	.385
Content	2	10.050	.007*	2	1.123	.570		DNA	
Date viewed	1	2.952	.086	1	.082	.775	1	.967	.326
Original domain	5	10.780	.056	5	11.910	.036*	5	11.524	.042*
Implied domain	5	18.784	.002*	5	21.821	.001*	5	8.165	.147
Directory depth	5	14.165	.015*	5	12.738	.026*	5	11.572	.041*
Page type	1	2.879	.090	1	.000	.992	1	1.334	.248
Tilde (~) included	1	1.832	.176	1	.334	.563	1	1.237	.266

*Significant at the $p<.05$ level.

the Web may have influenced the permanence and accessibility results. The results indicate that there was no association between researcher and permanence ($X^2 = 1.845$, df $= 1$, $p = .174$) or between researcher and accessibility elsewhere on the Web ($X^2 = 1.771$, df $= 1$, $p = .183$).

Discussion and Conclusion

The researchers examined the articles published in the thirty-four core, refereed, library and information science journals during 1999 and 2000 to determine the frequency with which authors cited digital resources. They drew a sample of 500 digital resources cited in these articles in order to identify citation characteristics and explore URL permanence and availability. Statistical analyses were then conducted to identify characteristics associated with availability on the Web and in the Internet Archive.

Of the 3,582 citations examined across 1,425 articles, 10 percent were to Web documents, although in some journals this percentage was substantially higher. The analysis of the sample drawn from this 10 percent indicates that the overwhelming number of citations to Web documents in the library and information science literature published during the period of this study pointed to hypertext resources. The majority contained only partial bibliographic information and did not include the date the

author viewed the site. Most resided on servers at either educational institutions ("edu") or organizations ("org"), did not include a tilde, and could be considered content, as opposed to navigational, pages.

Whereas 56.4 percent of the sample URLs were found to be permanent, 42.6 percent of the cited content was not found at the URLs included in the citations. These findings suggest that concerns about Web content permanence and its implication for scholarly communications are well founded. The findings of this study also confirm those of Davis and Cohen and of Lawrence and others in which a substantial amount of cited Web content that could not be found at the cited URL was found elsewhere on the Web.[34] In this study, the search strategies that were most effective for locating content not found at the cited URLs were using Google to search the Web and browsing/searching the Web site to which the URL led the researchers. Correcting errors in the URLs and truncating them were less-effective search strategies. Using all of these search strategies, the researchers eventually found the content cited by an additional 125 of the URLs in the sample citations, increasing the overall availability rate from 56.4 to 81.4 percent.

This study was the first to look at the effect the Internet Archive might have on content availability. Forty-seven percent of the URLs that could not be found at

TABLE 9
Cross-Tabulations
Citation Characteristics and Content Availability

Characteristic	Content Availability	Found #	Found %	Not Found #	Not Found %	Total #	Total %
	URL Permanent						
Citation Content							
URL only		24	82.8	5	17.2	29	100.0
URL & partial bibl. info.		147	58.1	106	41.9	253	100.0
URL & complete bibl. info.		222	52.1	102	47.9	324	100.0
N = 495							
Implied Domain							
Commercial		57	51.8	53	48.2	110	100.0
Education		109	64.5	60	35.5	169	100.0
Government		26	42.6	35	57.4	61	100.0
Organization		77	64.2	43	35.8	120	100.0
Geographic designation		5	31.2	11	68.8	16	100.0
Other		8	42.1	11	57.9	19	100.0
N = 495							
Directory Depth							
0		52	68.4	24	31.6	76	100.0
1		35	56.5	27	43.5	62	100.0
2		74	54.4	62	45.6	136	100.0
3		66	56.4	51	43.6	117	100.0
4		46	62.2	28	37.8	74	100.0
5 or more		9	30.0	21	70.0	30	100.0
N = 495							
	URL Permanent or Content Accessible						
Original Domain							
Commercial		71	74.7	24	25.3	95	100.0
Education		82	87.2	12	12.8	94	100.0
Government		35	77.8	10	22.2	45	100.0
Organization		98	89.9	11	10.1	109	100.0
Geographic designation		108	83.7	21	16.3	129	100.0
Other		13	72.2	5	27.8	18	100.0
N = 490							
Implied Domain							
Commercial		80	73.4	29	26.6	109	100.0
Education		151	89.3	18	10.7	169	100.0
Government		45	75.0	15	25.0	60	100.0
Organization		108	90.0	12	10.0	120	100.0
Geographic designation		9	69.2	4	30.8	13	100.0
Other		14	73.7	5	26.3	19	100.0
N = 490							

TABLE 9 (CONTINUED)
Cross-Tabulations
Citation Characteristics and Content Availability

Characteristic	Content Availability	Found #	Found %	Not Found #	Not Found %	Total #	Total %
	URL Permanent or Content Accessible						
Directory Depth							
0		65	87.8	9	12.2	74	100.0
1		48	77.4	14	22.6	62	100.0
2		108	79.4	28	20.6	136	100.0
3		95	82.6	20	17.4	115	100.0
4		69	94.5	4	5.5	73	100.0
5 or more		22	73.3	8	26.7	30	100.0
N = 490							
	Archived						
Original Domain							
Commercial		58	61.1	37	38.9	95	100.0
Education		71	75.5	23	24.5	94	100.0
Government		26	57.8	19	42.2	45	100.0
Organization		82	75.2	27	24.8	109	100.0
Geographic designation		96	74.4	33	25.6	129	100.0
Other		11	61.1	7	38.9	18	100.0
N = 490							
Directory Depth							
0		54	73.0	20	27.0	74	100.0
1		46	75.4	15	24.6	61	100.0
2		97	71.3	39	28.7	136	100.0
3		81	70.4	34	29.6	115	100.0
4		53	71.6	21	28.4	74	100.0
5 or more		13	43.3	17	56.7	30	100.0
N = 490							

either the URL included in the citation or elsewhere on the Web were found in the Internet Archive. Overall, almost 69 percent of the URLs in the sample were found in the Internet Archive. In this study, searching the Internet Archive increased the overall availability rate of the cited content from 81.4 to 89.2 percent.

It should be noted that receipt of a file or URL "not available" message as a result of an initial search was almost always as indication that the URL was impermanent. "Not available" messages were received during the initial search for more than 30 percent of the cited URLs. In subsequent searches,

after an intervening period of at least a week, the researchers were able to find the content at the URL included in the citation for only three of the 158 URLs.

Three of the characteristics studied—citation content, URL domain, and URL directory depth—were found to be associated with availability. URLs with "edu" and "org" original and implied domains were more often found at the URL cited or elsewhere on the Web than those with other domains. URLs with "edu" and "org" domains and those with geographic designations were more often found in the Internet Archive. Although the Chi-

Square tests suggest dependence between directory depth and permanence, directory depth and availability on the Web, and directory depth and availability in the Internet Archive, the researchers are unsure of the nature of these relationships.

The cross-tabulation between citation content and permanence suggests an inverse relationship between the amount of information included with the URL in the citation and permanence. The researchers suspect that this finding is the result of the research methodology. When searching the Web for the content cited by "URL only" or "URL and partial bibliographic information" citations, the researchers, having little or no bibliographic information to provide evidence to the contrary, may have tended to accept the Web page that was retrieved as containing the content the author cited. In contrast, when they were working with "URL and complete bibliographic information" citations, the researchers were able to determine with certainty whether they had found the cited content.

Although the researchers do not believe citation content to be a valid predictor of permanence, the finding of dependence between these variables spotlights an important limitation of this study and of most previous investigations of Web citation permanence and availability. That is, the researchers relied on the information in the citation and did not refer back to the text to determine whether the content found was actually the content the author was citing. The dynamism of Web pages documented by Koehler, Bar-Ilan and Peritz, Nelson and Allen, and Tan, Foo, and Hui underscores the significance of this limitation and suggests that the permanence and availability rates reported here may be overstated.[35] Therefore, the researchers suggest that in future citation studies of URL permanence investigators consult the source text to verify that the content the author cited is included at the Web site found.

The researchers found that most citations to Web resources that appear in articles published in library and information science journals did not contain complete bibliographic information, nor did they include the date the author last viewed the cited content. The findings of this study also indicate that few of the core journals in the library and information science disciplines provide authors with instructions on citing Internet resources and generally confirm the results of Zhang's editorial policy survey, which revealed a "lack of clearly stated conventions on citing e-sources."[36] Further, the researchers agree with the observation by Lawrence and others that "the general problem of persistence and disappearance requires a combination of technical solutions and peer policies" and recommend that authors, editorial staff, and publishers work together to develop such "peer policies" to improve scholarly communication citation conventions.[37]

The following suggestions are based on the researchers' experiences in collecting data for this study and the study's findings:

• The instructions for article authors, reviewers, and referees should include information on how to evaluate an Internet site in terms of both the quality of its content and its availability over the long run. The current study's findings suggest that URLs with "edu" and "org" domains and implied domains may be more permanent than those with other domains. URLs with fewer directory depths also may be more available. However, further studies will be needed to clarify this study's findings of dependence between directory depth and availability on the Web and in the Internet Archive. Editorial staff should be aware of these relationships and instruct authors that, when there is a choice, they should cite content at the URL that is most likely to be permanent.

• Many of the citations to Web sites the researchers examined in this study were included by authors as a means of further identifying businesses, organizations, and individuals they mention in their articles. The researchers suspect that many of the "URL only" citations fall into this category. Editorial staff and authors should work together to determine when and where this type of Web content should be referenced.

• Just as some journals do not allow authors to use "pers comm" or "in prep"

papers in citation lists, editorial staff should develop guidelines to convey to authors and referees the types of Web content that are suitable for a reference list and where to place them within the article. These guidelines should be based on considerations of future availability of the content cited as well as scholarly importance.

• "Instructions for Authors" pages should include complete information on citing content that resides on the Internet or refer authors to a style manual or Web site that includes this information.

• Complete citations to Web content should include full bibliographic information plus the date the site was accessed by the author and the dates the cited Web page was created and last revised. Moreover, it may be advisable for authors to include contact information for the Web page creator or other Web site accessibility information.

• Authors should determine whether there is a paper counterpart to the Web content they are citing. If so, complete citations to both sources should be provided.

• Editorial staff should work with authors to preserve and make available cited Web content. One possible strategy would be to support the development and maintenance of the Internet Archive and require that Web content cited by authors be easily retrievable from that archive. In the case of electronic journals, another possibility would be for the journals to archive the Web content cited in the articles they publish or to partner with academic libraries for that purpose.

• Editorial staff should require authors to adhere to the citation policies, styles, and formats established by their journals. Further, they should review their citation guidelines frequently and modify them, as needed, to ensure maximum access to the Web content referenced by their authors.

The Internet has expanded access to scholarship, and its dynamism poses many challenges to scholarly communication. This study has addressed some questions about the use of citations to Web content in the library and information science literature and the availability of this content. The results suggest that authors, editorial staff, and publishers need to work together to improve existing citation conventions, promote their use, and ensure that cited resources are accessible to future researchers.

Notes

1. Joseph Gibaldi, *MLA Handbook for Writers of Research Papers*, 5th ed. (NewYork: The Modern Language Association of America, 1999), 114.

2. Catherine Jean Sassen, "Citation Accuracy in the Journal Literature of Four Disciplines: Chemistry, Psychology, Library Science, and English and American Literature" (Ph.D. diss, Univ. of North Texas, 1992), 33.

3. Idrisa Pandit, "Citation Errors in Library Literature: A Study of Five Library Science Journals," *Library and Information Science Research* 15, no. 2 (1993): 185–98; Susan P. Benning and Susan C. Speer, "Incorrect Citations: A Comparison of Library Literature with Medical Literature," *Bulletin of the Medical Library Association* 81 (1993): 56–58; Nancy N. Pope, "Accuracy of References in Ten Library Science Journals," *RQ* 32, no. 2 (1992): 240–43.

4. Carol Anne Germain, "URLs: Uniform Resource Locators or Unreliable Resource Locators," *College and Research Libraries* 61, no. 4 (July 2000): 360.

5. Wallace Koehler, "An Analysis of Web Page and Web Site Constancy and Permanence," *Journal of the American Society for Information Science* 50, no. 2 (Feb. 1999): 162.

6. Ibid., 162–80.

7. Wallace Koehler, "Web Page Change and Persistence: A Four-year Longitudinal Study," *Journal of the American Society for Information Science and Technology* 53, no. 2 (2002): 162–71.

8. Judit Bar-Ilan and Bluma C. Peritz, "The Life Span of a Specific Topic on the Web: The Case of 'Informetrics': A Quantitative Analysis." *Scientometrics* 46 (Nov. 1999): 371–82.

9. Michael L. Nelson and B. Danette Allen, "Object Persistence and Availability in Digital Libraries," *D-Lib Magazine* 8, no. 1 (Jan: 2002). Available online from: http://www.dlib.org/dlib/january02/nelson/01nelson.html (2 December 2002).

10. Bing Tan, Schubert Foo, and Siu Cheung Hui, "Web Information Monitoring: An Analysis

of Web Page Updates," *Online Information Review* 25, no. 1 (2001): 6–19.

11. John Markwell and David W. Brooks, "Broken Links: The Ephemeral Nature of Educational WWW Hyperlinks," *Journal of Science Education and Technology* 11, no. 2 (June 2002): 105–8.

12. Yasar Tonta, "Scholarly Communication and the Use of Networked Information Sources," *IFLA Journal* 22, no. 3 (1996): 240–45.

13. Stephen P. Harter and Hak Joon Kim, "Accessing Electronic Journals and Other E-publications: An Empirical Study," *College and Research Libraries* 57, no. 5 (Sept. 1996): 440–56; ———, "Electronic Journals and Scholarly Communication: A Citation and Reference Study," in *The Digital Revolution: Assessing the Impact on Business, Education, and Social Structures, Proceedings of the ASIS Mid-year Meeting Held May 18–22, 1996, San Diego, Calif.* (Medford, N.J.: Information Today Inc., 1996), 299–315.

14. 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, no. 4 (2001): 309–14.

15. Philip M. Davis, "The Effect of the Web on Undergraduate Citation Behavior: A 2000 Update," *College and Research Libraries* 63 (Jan. 2002): 53–60.

16. S. Mary P. Benbow, "File Not Found: The Problems of Changing URLs for the World Wide Web," *Internet Research: Electronic Networking Applications and Policy* 8, no. 3 (1998): 247–50; Germain, "URLs," 359–65; Joel D. Kitchens and Pixey Anne Mosley, "Error 404: Or, What Is the Shelf-Life of Printed Internet Guides?" *Library Collections, Acquisitions, and Technical Services* 24 (2000): 467–78; Susan Davis Herring, "Use of Electronic Resources in Scholarly Electronic Journals: A Citation Analysis," *College and Research Libraries* 63, no. 4 (July 2002): 334–40; Mary K. Taylor and Diane Hudson, "'Linkrot' and the Usefulness of Web Site Bibliographies," *Reference & User Services Quarterly* 39 (spring 2000): 273–77.

17. Steve Lawrence, et al., "Persistence of Web References in Scientific Research," *Computer* 34 (Feb. 2001): 26–31.

18. Mary Rumsey, "Runaway Train: Problems of Permanence, Accessibility, and Stability in the Use of Web Sources in Law Review Citations," *Law Library Journal* 94 (winter 2002): 27–39.

19. Yin Zhang, "The Impact of Internet-based Electronic Resources on Formal Scholarly Communication in the Area of Library and Information Science: A Citation Analysis," *Journal of Information Science* 24 (Aug. 1998): 241–54.

20. ———, "Scholarly Use of Internet-based Electronic Resources," *Journal of the American Society for Information Science and Technology* 52, no. 8 (2001): 628–54.

21. Yin Zhang and Leigh Estabrook, "Accessibility to Internet-based Electronic Resources and Its Implications for Electronic Scholarship," in *ASIS '98 Information Access in the Global Information Economy. Proceedings of the 61st Annual Meeting of the American Society for Information Science Held October 25–29, 1998, Pittsburgh, PA* (Medford, N.J.: Information Today, Inc., 1998), 463–73.

22. Zhang, "Scholarly Use of Internet-based Electronic Resources."

23. Bill Katz and Linda Sternberg Katz, *Magazines for Libraries,* 10th ed. (New York: Bowker, 2000), 914–29.

24. Edward L. Vockell and J. William Asher, *Educational Research,* 2nd ed. (Englewood Cliffs, N.J.: Merrill, 1995), 182, table 8.2.

25. Google, "Our Search: Google Technology: PageRank Explained" (2002). Available online from: http://www.google.com/technology/index.html (2 December 2002); Laura Cohen, "10 Tips for Teaching How to Search the Web," *American Libraries* 32, no. 10 (Nov. 2001): 44–46.

26. Internet Archive, "About the Internet Archive." Available online from: http://www.archive.org/about/about.php (2 December 2002).

27. American Psychological Association, "Electronic Resources." Available online from: http://apastyle.org/elecref.html (13 December 2002).

28. *The Chicago Manual of Style,* 14th ed. (Chicago: Univ. of Chicago Pr., 1993).

29. *Publication Manual of the American Psychological Association,* 5th ed. (Washington, D.C.: American Psychological Association, 2001); Gibaldi, *MLA Handbook for Writers of Research Papers*.

30. Koehler, "Web Page Change and Persistence," 167.

31. Ibid.

32. "Tilde or ~," in Vincent James and Erin Jansen, *Netlingo: The Internet Dictionary* (Ojai, Calif.: Netlingo Inc., 2002). Available online from: http://www.netlingo.com/ (2 December 2002).

33. Zhang and Estabrook, "Accessibility to Internet-based Electronic Resources and Its Implications for Electronic Scholarship."

34. Davis and Cohen, "The Effect of the Web on Undergraduate Citation Behavior 1996–1999"; Lawrence, et al., "Persistence of Web References in Scientific Research."

35. Koehler, "An Analysis of Web Page and Web Site Constancy and Permanence," 162–80; Bar-Ilan and Peritz, "The Life Span of a Specific Topic on the Web"; Nelson and Allen, "Object Persistence and Availability in Digital Libraries"; Tan, Foo, and Hiu, "Web Information Monitoring."

36. Zhang, "Scholarly Use of Internet-based Electronic Resources," 639.

37. Lawrence, et al., "Persistence of Web References in Scientific Research," 30.