Web Citation Availability

A Follow-up Study

By Mary F. Casserly and James E. Bird

The researchers report on a study to examine the persistence of Web-based content. In 2002, a sample of 500 citations to Internet resources from articles published in library and information science journals in 1999 and 2000 were analyzed by citation characteristics and searched to determine cited content persistence, availability on the Web, and availability in the Internet Archive. Statistical analyses were conducted to identify citation characteristics associated with availability. The sample URLs were searched again between August 2005 and June 2006 to determine persistence, availability on the Web, and in the Internet Archive. As in the original study, the researchers cross-tabulated the results with URL characteristics and reviewed and analyzed journal instructions to authors on citing content on the Web. Findings included a decrease of 17.4 percent in persistence, and 8.2 percent in availability on the Web. When availability in the Internet Archives was factored in, the overall availability of Web content in the sample dropped from 89.2 percent to 80.6 percent. The statistical analysis confirmed the association between the likelihood that cited content will be found by future researchers and citation characteristics of content, domain, page type, and directory depth. The researchers also found an increase in the number of journals that provide instruction to authors on citing content on the Web.

Students and researchers look to literature citations as links between what is new and what is already known. The value of accurate and valid citations cannot be overstated, as citations act as knowledge building blocks. Citations to Web resources and documents are increasingly found in scholarly articles, and over the past several years a significant body of literature on the stability of citations to content on the Web has developed. Many of these studies document decreasing availability of cited content over time; fewer also have attempted to identify factors that contribute to the stability of these citations. Recognizing the citation stability problem and knowing the factors that contribute to Web reference stability will help authors, editors, and publishers develop policies and conventions that will ensure long-term access to cited Web content.

In 2002, the authors conducted a study of 500 citations containing URLs from articles published in library and information science journals in 1999 and 2000.[1] In this earlier study, the authors described URLs that led to cited content as "permanent." In reporting the findings of the follow-up study, they use the term "persistent" in place of permanent. Persistent is now commonly used in the literature and better describes the quality being studied. The study addressed the following questions:

- To what extent are authors currently referencing information and documents "published" on the Web?
- What percentage of cited electronic resources is available to be consulted by future scholars? How are they most often found?

Mary F. Casserly (mcasserly@uamail.albany.edu) is Assistant Director for Collections and User Services, University at Albany–SUNY. **James E. Bird** (jim.bird@umit.maine.edu) is Head, Science and Engineering Center, Raymond H. Fogler Library, University of Maine, Orono.

Submitted March 20, 2007; tentatively accepted pending revision April 30, 2007; revised and resubmitted June 30 , 2007, and accepted for publication.

- Is it possible to identify characteristics of citations to Internet resources that will help predict the availability of the content to which they refer?
- What type of guidance are authors receiving from editors and publishers?

In the earlier study, the authors found that the majority of the citations in the sample 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 percent) were persistent, and 81.4 percent were available on the Web; 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. The authors offered suggestions for updating scholarly communication citation conventions based on these findings.

The purpose of this subsequent study is to determine the changes in URL persistence included in the 1999/2000 sample, the availability of cited content, and the instructions on citing Web content provided by journals to their authors.

Literature Review

The authors reviewed the literature on Web citation persistence through 2002 as part of the original study.[2] Since then, the published research has addressed Web page persistence rates, factors related to persistence, and, to a lesser extent, instructions to authors citing Web content.

Persistence of Web Pages

Sellitto provided an extensive review of the literature on Web site persistence of cited Web resources.[3] He selected 123 papers from the 1995 to 2003 AusWeb conference series, examined the 2,168 references cited, and found Web resources in almost 50 percent of them. Of the Web resources cited, 45.8 percent were not found at the URL cited. Sellitto determined the average half-life for these Web citations to be 4.8 years. In a study of papers in dermatological journals published between 1999 and 2004, Wren and colleagues found that 81.7 percent of URLs cited in papers published in 2004 were available.[4] The percentage decreased with time to an availability rate of 65.4 percent for URLs cited in papers published in 1999.

Dellavalle and colleagues examined Internet references from the papers of three high-impact scientific journals (*NEJM*, *JAMA*, and *Science*) published over a six-week period in three different years (2001–2003).[5] They found that the number of inactive Internet references increased over the twenty-seven-month period, reaching a high of 21 percent for *JAMA,* to a low of 11 percent for *Science*. Spinellis examined the Web citations in two computer science publications from 1995 to 1999.[6] Almost 50 percent of the 4,224 references he checked were not available after four years. Dimitrova and Bugeja examined Web references from a sample of papers published in communication journals between 2000 and 2003, and found 37 percent of the URLs no longer led to the content cited.[7] In April 2004, Crichlow, Aguillo, and Prieto examined the Internet references in the articles of five major medical journals that were published in January 2004, and, by cutting and pasting the URLs into a Web browser, found that five were inaccurate (7.4 percent) and three had inaccessible sites (4.4 percent).[8]

Markwell and Brooks studied biochemistry and molecular biology Web-based educational material and found that, in a twenty-four-month period starting in August 2000, more than 20 percent of the 515 sites examined had changed content, moved with no automatic forwarding, or were broken links.[9] Ortega and colleagues compared Web content in 738 Web sites in 1997 and again in 2004. As part of this study, they examined the persistence and stability of the links within these pages; they found that 74.28 percent of the 145,092 links were "broken (linkrot) or not operative."[10] Bugeja and Dimitrova looked at Internet references cited in Association for Education in Journalism and Mass Communication online conference papers.[11] They found that after less than a year, 55 (51 percent) of the 108 citations led to the Web content cited when they clicked on the link. However, when they cut and pasted the URLs into a Web browser, the number of found citations increased to 65 (60 percent).

Bar-Ilan and Peritz conducted a broad-ranging study of Web pages concerning the subject "informetrics."[12] They looked at the number of Web sites retrieved on this subject in 1998, 1999, 2002, and 2003 and identified 866 URLs in 1998 that met their search criteria. Of those, 299 (34.5 percent) were not available in 1999, with 496 (57.3 percent) and 552 (63.7 percent) not available in 2002 and 2003, respectively. Of the 1,297 URLs identified in 1999, 643 (49.6 percent) were not available in 2002, and 769 (59.3 percent) not available in 2003. Of the 3,746 URLs identified in 2002, 682 (18.2 percent) were not available in 2003. Koehler continued his study of Web page persistence begun with 361 URLs identified in 1996 and found that by 2003, two-thirds of the original sample of URLs were gone.[13]

Kushkowski analyzed the citations in economic theses and dissertations in print and electronic format to see if different patterns emerged in the citations to Web resources.[14] He selected master's theses and doctoral dissertations from Virginia Tech, where electronic theses and dissertations were required, and Iowa State University, where electronic copies were not accepted, and found that a small percentage of citations in these works were to Web resources (Virginia

Tech had 5.4 percent, and Iowa State had 2.2 percent), with Web citations increasing between 1997 and 2002. Kushkowski also looked at the persistence of Web citations and found that approximately 55 percent of Web citations from both universities' theses and dissertations led directly to the documents cited.

Wu briefly discussed the importance of persistence in Web documents in legal research, noting, particularly, the loss of interim documents, which are of great importance to both legal research and analysis.[15] In 1995, librarians at the National Library of Australia identified fifty publications on the Web that were considered sufficiently important for preservation.[16] They found that 22 percent of these publications were still available in 2004 at their original URLs, and an additional 64 percent were still accessible on the Internet, although the content or style of some had changed. They predicted that in about five years, users will be able to find only 50 percent of the original publications.

Factors Related to Persistence

In 2004, Koehler found that three-quarters of the URLs that were still available from his 1996 sample were to navigation pages (that is, pages found at the server level and first level) as opposed to content pages (that is, pages found at the second level and below).[17] The depth of the path was linked to URL failure in Spinellis's 2003 study.[18] Wren and colleagues found that root directories were more likely to be available than those URLs with a directory depth of one.[19] They also found that the presence of a tilde or accession date did not affect URL availability. Dimotrova and Bugeja looked at four factors that could be predictors of URL stability and found more stability at the top-level sites.[20] They also found a positive relationship between year of publication and persistence, with the more recent the publication year, the higher number of accessible URLs. In addition, they discovered that the presence of a retrieval date with a Web citation was not related to URL persistence.

Wren and colleagues found that domain was an indicator of availability, with .edu sites showing the greatest availability, followed in descending order by .org, .net, .com, and .gov.[21] Sellitto's study showed that .edu sites had the highest number of URLs classified as missing, followed by .com sites.[22] Over a twenty-four-month period, Markwell and Brooks found that .gov sites were the most stable, followed by .org and .edu, respectively, with .com sites being the least stable.[23] Dellavalle and colleagues looked at the persistence of links by domain and found that after twenty-seven months, references with .com domains had the most inactive links, followed by .edu, .gov, and .org.[24] In 2005, Bugeja and Dimitrova found that .edu sites were the most stable, followed by .com and .org sites.[25] In their 2006 study, they found that .gov and .org sites were the most stable (73.0 percent and 71.0 percent active, respectively) followed by .com sites at 63.9 percent and .edu sites at 46.8 percent.[26] Bar-Ilan and Peritz found that between 2002 and 2003, 80.4 percent of .edu sites in their study were available, as opposed to 77.8 percent of .org sites and 55.5 percent of .com sites.[27] When they considered availability over a five-year period (1998 to 2003), the percentage of sites available by domain dropped to 21.6 percent, 21.9 percent, and 23.9 percent, respectively.

Instructions for Authors

Schilling and colleagues surveyed the instructions for author pages and Web sites of the one hundred highest-impact journals in science and medicine as determined by ISI's *Journal Citation Reports*.[28] They found one journal that discussed maintaining access to Internet-cited information, eleven journals that provided authors with examples for citing Internet references using digital object identifiers (DOIs), and thirty-six journals that requested dates with Internet citations. None of the journals required that their authors provide DOIs.

Methodology

The researchers searched for the content cited by the sample of 500 citations from articles published in library and information science journals in 1999 and 2000 used for the original study.[29] They began the search by looking for the content sited at the URL included in the citation. If they did not find it, they searched for it elsewhere on the Web. All citations were then searched in the Internet Archive. The methodology used and described in depth in the original study was employed in this follow-up study with one exception. In the original study, the researchers searched a second time for those URLs for which they initially received a URL unavailable or file not found message. These were not searched a second time in the follow-up study. The statistical program SPSS 14.0 for Windows was used to generate contingency tables and calculate the Pearson's Chi-Square values. A $p < 0.05$ level of significance was used for this study. The data for this follow-up study were collected between August 2005 and June 2006. The data for the original study were collected from January to July 2002.

Findings

Content Availability: URL Persistence

URL persistence decreased by 17.4 percent during the period between the original and follow-up studies. These data

are presented in table 1. In the original study, 282, or 56.4 percent, of the sample URLs, were persistent; that is, they pointed to the cited content or to Web pages that referred or redirected the researcher to the cited content. Of the 500 citations studied, the content cited by 213, or 42.6 percent, could not be found at the URLs included in the citations and, therefore, were considered to be impermanent. In the follow-up study, 195, or 39 percent, of the sample URLs were found to be persistent, and 305, or 61 percent, were impermanent.

Content Availability—Accessibility on the Web

Cited content was considered to be accessible if, after failing to find it at the URL included in the citation or at a referred page, the researchers were able to locate it elsewhere on the Web. In the original study, the researchers had to search for the content of 213 impermanent URLs, while in the follow-up study they had to search for 300. The results of these searches are presented in table 2.

The researchers found content cited in 3.8 percent of the original study impermanent URLs by truncating them, and in 4.2 percent by identifying errors that, when corrected, led to the cited information. In the follow-up study, the percentage of content found by truncation dropped slightly to 3.0 percent, with a 2.2 percent decrease in content found by correcting URL errors. In the follow-up study, the researchers had relatively less success finding content at the URL cited by browsing and searching the Web site, and relatively more success using Google to locate content elsewhere on the Web. In the original study, they located content cited in 25.4 percent of the impermanent citations by browsing or searching the site to which the URL led them, while in the follow-up study this percentage was 21.3 percent. In the follow-up study, the researchers found 30.7 percent of the impermanent URLs by using Google, an increase of 5.3 percent from the original study.

The number of citations for which the cited content could not be found rose from 83, or 16.6 percent of the 500 citations in the sample, in the original study, to 124, or 24.8 percent of the sample, in the follow-up study. This represents an 8.2 percent increase in unavailable content.

Content Availability—Internet Archive

The researchers searched the Internet Archive using the Wayback Machine (www.archive.org/index.php) to determine if the URLs included in the sample citations had been archived. The results are presented in table 3. The percentages of all citations and of those that were considered persistent that the researchers were able to find in the

Table 1. Content availability: URL persistence in original and follow-up studies

Content at cited URL or at referred page	Original study URLs		Follow-up study URLs		Change
	No.	%	No.	%	%
Found	282	56.4	195	39.0	-17.4
Not found	213	42.6	305	61.0	+18.4
Could not determine	5	1.0	0	0	-1.0
Total	**500**	**100.0**	**500**	**100.0**	

Table 2. Content availability: accessibility on the Web in the original and follow-up studies

Content on Web	Original study URLs		Follow-up study URLs		Change
	No.	%	No.	%	%
Found by truncating URL	8	3.8	9	3.0	-.8
Found by browsing and searching Web site	54	25.4	64	21.3	-4.1
Found by correcting error in URL	9	4.2	6	2.0	-2.2
Found by using Google	54	25.4	92	30.7	+5.3
Not found	83	39.0	124	41.3	+2.3
Could not determine	5	2.3	5	1.7	-.6
Total	**213**	**100.0**	**300**	**100.0**	

Internet Archive changed very little from the original to the follow-up study. The largest difference was in the accessible content category. In the follow-up study, the researchers found 64.3 percent of these URLs in the Internet Archive, almost 14 percent more than in the original study.

The researchers were able to access 39, or 47.0 percent, of the 83 citations that they could not find at the URL cited or elsewhere on the Web in the original study by using the Wayback Machine. This raised the overall availability rate of the cited content from 81.4 percent to 89.2 percent, or 7.8 percent. In the follow-up study, they found 63, or 50.8 percent, of the 124 citations in the "Content Not Found" category, raising the overall availability of cited content from 68.0 percent to 80.6 percent, or 12.6 percent.

Changing Categories

The researchers ran cross-tabulations to determine the extent to which the citations in the sample changed categories between the studies. Table 4 presents the status in the follow-up study of the URLs that were persistent, accessible, not found, or could not be determined in the original study. Of the 282 citations that were persistent in the original study, 188 were found to be persistent in the follow-up study, and all except 37 were found elsewhere on the Web by either truncating the URL (1 citation), browsing and searching the site (19 citations), or using Google (37 citations). Four of the 8 citations that were found by truncating the URL in the original study were found using this method in the follow-up study, one was found by browsing and searching the Web site, and another was found by using Google. Thirty-two of the 54 citations that were found by browsing and searching the Web site in the original study also were found using that method in the follow-up study, while 15 were found using Google, and 4 were not found. The researchers found 29 of the 54 citations for which they had to use Google in the original study by using Google in the follow-up study. They could not find the content cited by 14 of those 54 citations.

The researchers expected that many of the citations in the sample would, over time, move from persistent to accessible to not found. However, 20 progressed in the opposite direction. These citations are identified with an asterisk in table 4. In 5 cases, URLs that were only accessible in the original study were found to be persistent in the follow-up study. This group includes 2 citations for which content was found in the original study by truncating the URL and using Google, and one whose content was found by browsing and searching the Web site. In addition, 2 of the 83 citations in the not found category in the original study were found to be persistent in the follow-up study, and 13 others were found to be accessible by truncating the URL (1 citation), browsing and searching the Web site (4 citations), or using Google (8 citations).

The researchers examined the 20 citations that moved from accessible to persistent and from not found to persistent or accessible to try to identify patterns that would explain this improvement in availability. They found that when they searched for 12 of these citations in the original study, they initially received a "URL not available" message. This required that they wait a week and then search

Table 3. Content availability: Internet archive in original and follow-up studies

	All citations		Persistent URLs		Accessible content		Content not found	
Accessible in Internet Archive	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Original study								
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	100.1	130	100.0	83	100.0
Follow-up study								
Found	340	68.0	166	85.1	110	64.3	63	50.8
Not found	156	31.2	28	14.4	61	35.7	61	49.2
Could not determine	4	.8	1	.5	0	0	0	0
Total	**500**	**100.0**	**195**	**100.0**	**171**	**100.0**	**124**	**100.0**

the URL again. The researchers found the content cited by 3 of these on the Web but not at the URL in the citation. In the follow-up study, their status changed from accessible to persistent because when the researchers looked for the content at the URL included in the citation, they were referred to a page that contained that content. The researchers also initially received a "URL not available" message in the original study for the remaining 9 of these 12 citations, but when they waited a week and searched again they were not able to find the cited content anywhere on the Web. In the follow-up study the researchers *did* find the content, but not at the URL included in the citation and not at a referred page. Therefore, the status of these cases changed from not found to accessible. Based on these findings, the improvement in availability appears to be the result of referrals and automatic redirects not in place during the original study. Improvements in Web site indexing capabilities with more sites offering tools to search their sites and enhancements to Google's Web search features also may have affected availability. Finally, in some cases, the improvement may be simply because the Web sites containing the cited content were not working at the time the original study was conducted, but were functioning during the follow-up study.

The cross-tabulation of cited content availability in the Internet Archive in the original and follow-up study is presented in table 5. Fifty of the 344 URLs found in the Internet Archive in the original study no longer led to the cited content when the follow-up study was conducted. The content for 46 of the 146 citations that were not found in the Archive in 2002 *was* found in the Archive during the follow-up study.

Characteristics Associated with Availability

The researchers ran a series of cross-tabulations to identify the characteristics of the cited URLs that could be associated with URL persistence and content availability on the Web and in the Internet Archive. Chi-Square Tests of Independence were performed to identify the statistically significant relationships. In order to run these tests, the researchers had to filter out some of the cases in the could not determine category and reclassify some of the variable values into broader categories. The results of the Chi-Square tests for the original and the follow-up studies are presented in table 6, and those that are significant at the $p < 0.05$ level are identified with an asterisk. Cross-tabulations were not run on citation content and availability in the Internet Archive variables, as 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 the archive.

Table 4. Category changes between original and follow-up study: persistence and accessibility

		Disposition in the Follow-up Study						
Category	**Original study total**	**Persistent-found at URL cited**	**Accessible-found by truncating URL**	**Accessible-found by browsing and searching Web site**	**Accessible-found by correcting error in URL**	**Accessible-found by using Google**	**Not found**	**Could not determine**
Persistent—Found at URL cited	282	188	1	19	0	37	37	0
Accessible—Found by truncating URL	8	2*	4	1	0	1	0	0
Accessible—Found by browsing and searching Web site	54	1*	2	32	0	15	4	0
Accessible—Found by correcting error in URL	9	0	0	1	5	2	1	0
Accessible—Found by using Google	54	2*	1	7	1	29	14	0
Not found	83	2*	1*	4*	0	8*	68	0
Could not determine	10	0	0	0	0	0	0	10
Total	**500**	**195**	**9**	**64**	**6**	**92**	**124**	**10**

* moved from accessible to persistent or from not found to persistent or accessible.

In the original study, the Chi-Square tests indicated that citation content, domain, and URL directory depth were associated with content availability. Specifically, the amount of information in the citation, implied domain, and directory depth were associated with content found at the URL cited or at a referred page; that is, persistence. Original and implied domain, as well as directory depth, were associated with content that was either found at the URL cited or elsewhere on the Web; that is they were "persistent or accessible." Finally, original domain and directory depth were found to be associated with content availability in the Internet Archive. Although domain, directory depth, and citation content characteristics were found to be associated with content availability in the follow-up study, they are not associated with the same types of availability as in the original study. In addition, the source journal (print or e-journal) and page type (navigation or content) were found to be associated with availability in the follow-up study, although not in the original study.

The cross-tabulations for the characteristics with significant Chi-Square values are presented in table 7. The cross-tabulation between source journal and persistence indicates that the URLs in the sample citations taken from print journals were found to be persistent more often than the URLs in citations taken from journals that were published only in electronic format. Specifically, 41.5 percent of the citations in the sample that came from print journals were found to be persistent, while the persistence rate for URLs from electronic-only journals was only 20.5 percent.

Page type and directory depth also were found to be associated with persistence. More than 52 percent of the content cited by navigation pages was found at the URL included in the citation, in comparison to 35 percent of the content cited by content pages. In general, persistence rates decreased as the directory depth increased. More than 62 percent of the content was found at the URL cited when the URL was at the server level, and only 13.3 percent of the content was found at the URL cited when the URL was at

Table 5. Category changes between original and follow-up study availability in Internet Archive

		Disposition in the Follow-up Study		
Category	**Original study total**	**Found in Archive**	**Not Found in Archive**	**Could not determine**
Found in Archive	344	294	50	0
Not found in Archive	146	46	100	0
Could not determine	10	0	6	4
Total	**500**	**340**	**156**	**4**

Table 6. Summary of Pearson's Chi-Square (x^2) values: citation characteristics and content availability in original and follow-up studies

	Persistent						Persistent or accessible						Archived					
	Original study			Follow-up study			Original study			Follow-up study			Original study			Follow-up study		
Characteristics	**df**	**x^2**	***p***	**df**	**x^2**	***p***	**df**	**x^2**	***p***	**df**	**x^2**	***p***	**df**	**x^2**	***p***	**df**	**x^2**	***p***
Source journal	1	.559	.455	1	6.576	.010*	1	.073	.787	1	2.207	.137	1	.754	.385	1	.070	.792
Content	2	10.050	.007*	2	4.665	.097	2	1.123	.570	2	6.544	.038*		DNA			DNA	
Date viewed	1	2.952	.086	1	1.152	.283	1	.082	.775	1	.013	.909	1	.967	.326	1	.348	.555
Original domain	5	10.780	.056	4	7.103	.131	5	11.910	.036*	4	19.478	.001*	5	11.524	.042*	4	10.364	.035*
Implied domain	5	18.784	.002*	4	8.133	.087	5	21.821	.001*	4	23.613	.000*	5	8.165	.147	4	14.245	.007*
Directory depth	5	14.165	.015*	5	27.190	.000*	5	12.738	.026*	5	7.226	.204	5	11.572	.041*	5	10.646	.059
Page type	1	2.879	.090	1	12.101	.000*	1	.000	.992	1	.359	.549	1	1.334	.248	1	5.455	.020*
Tilde (~) included	1	1.832	.176	1	.199	.656	1	.334	.563	1	.864	.353	1	1.237	.266	1	.756	.384

* significant at the $p < 0.05$ level

Table 7. Cross-tabulations: citation characteristics and content availability in the follow-up study

	Persistent		Not persistent		Total	
Characteristic	**No.**	**%**	**No.**	**%**	**No.**	**%**
Source journal (*N* = 490)						
Print journal	187	41.5	264	58.5	451	100.0
E-journal only	8	20.5	31	79.5	39	100.0
Directory depth (*N* = 490)						
0	46	62.2	28	37.8	74	100.0
1	25	40.3	37	59.7	62	100.0
2	49	36.0	87	64.0	136	100.0
3	39	33.9	76	66.1	115	100.0
4	32	43.8	41	56.2	73	100.0
5 or more	4	13.3	26	86.7	30	100.0
Page type (*N* = 490)						
Navigation page	71	52.2	65	47.8	136	100.0
Content page	124	35.0	230	65.0	354	100.0
	Available on the Web		**Not available on the Web**		**Total**	
	No.	**%**	**No.**	**%**	**No.**	**%**
Citation content (*N* = 490)						
URL only	18	62.1	11	37.9	29	100.0
URL & partial bibl. info.	181	71.8	71	28.2	252	100.0
URL & complete bibl. info.	167	79.9	42	20.1	209	100.0
Original domain (*N* = 490)						
Commercial	56	58.9	39	41.1	95	100.0
Education	77	81.9	17	18.1	94	100.0
Government	37	82.2	8	17.8	45	100.0
Organization	89	81.7	20	18.3	109	100.0
Geographic designation and other	107	72.8	40	27.2	147	100.0
Implied domain (*N* = 490)						
Commercial	64	58.7	45	41.3	109	100.0
Education	139	82.2	30	17.8	169	100.0
Government	45	75.0	15	25.0	60	100.0
Organization	97	80.8	23	19.2	120	100.0
Geographic designation and other	21	65.6	11	34.4	32	100.0
	Available in the Internet Archive		**Not available in the Internet Archive**		**Total**	
	No.	**%**	**No.**	**%**	**No.**	**%**
Original domain (*N* = 496)						
Commercial	53	55.8	42	44.2	95	100.0
Education	72	76.6	22	23.4	94	100.0
Government	31	68.9	14	31.1	45	100.0
Organization	76	69.7	33	30.3	109	100.0
Geographic designation and other	108	70.6	45	29.4	153	100.0
Implied domain (N = 496)						
Commercial	62	56.4	48	43.6	110	100.0
Education	130	76.9	39	23.1	169	100.0
Government	42	67.7	20	32.3	62	100.0
Organization	84	70.6	35	29.4	119	100.0
Geographic designation and other	22	61.1	14	38.9	36	100.0
Page type (*N* = 496)						
Navigation page	104	76.5	32	23.5	136	100.0
Content page	236	65.6	124	34.4	360	100.0

level five or lower. However, for URLs at level four, the persistence rate (43.8 percent) was higher than that for citations containing URLs at level one (40.3 percent). This suggests that some other variable may be affecting the relationship between directory depth and persistence. This pattern also was observed in the original study, where directory level was found to be associated with availability on the Web and in the Internet Archive, but where page type was not found to be associated with availability in either place.

Citations containing complete bibliographic information along with the URL were more likely to lead to available content than those with only partial bibliographic information or those that contained only URLs. Almost 80 percent of the citations with complete bibliographic information led to content that was either at the URL cited or accessible elsewhere on the Web. When the citation included only partial bibliographic information, this percentage dropped to 71.8 percent; for citations that included only URLs, the availability rate was only 62.1 percent.

The cross-tabulations of domains with content availability indicated that content cited by URLs with original domains of .gov, .edu, and .org were more likely to be available and found in the Internet Archive than content cited by URLs with .com or other types of domains. Between 81.7 and 82.2 percent of the content cited by URLs residing on education, government, and organization servers was found either at the URL included in the citation or elsewhere on the Web. Less than 60 percent of the content cited by URLs residing on commercial servers was found to be available. For content found in the Internet Archive, the pattern was similar. The researchers found 76.6 percent of the content cited by URLs residing on education servers and only 55.8 percent of the content cited by URLs on commercial servers. URLs in the combined category include those on military and network servers and those with geographic designations. Content availability in the Internet Archive for citations with URLs in this combined category was 70.6 percent, which is higher than the percentages of available content for citations containing government and organization URLs.

The cross-tabulations indicated that content availability on the Web and in the Internet Archive also was associated with implied domain. The original domains in the sample citations' URLs were translated into implied domains by folding the country code top-level domains that included information about the organization hosting their content into the appropriate generic top-level domains. This reduced the number of sample citations in the "Geographic Designation and Other" category and increased the number in the other four domain categories. The citation URLs with implied education domains were found to have the most available content (82.2 percent), and those with commercial implied domains had the least available content (58.7 percent). Three-quarters of the content cited by URLs with government implied domains and 80.8 percent of content cited by those with organization implied domains were found to be available. This pattern of content availability also was found in the Internet Archive. The content cited by URLs in the education implied domain was found to be most accessible (76.9 percent), followed by URL citations with organization (70.6 percent) and government (67.7 percent) implied domains.

The Chi-Square test also indicated that page type was found to be associated with availability in the Internet Archive. More than three-quarters of the navigation pages (76.5 percent) were found in the Internet Archive, while 65.6 percent of the content pages were found there.

Instructions for Authors

In the original study, the researchers reviewed the instructions for authors published by the journals from which the sample was drawn for the period of the study (1999–2000), and again as that manuscript was being prepared, in order to determine if these journals had established policies or instructions on citing content on the Web. In the original study, only 6 of the 34 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). Fifteen of the journals 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. None of the instructions for authors addressed Web site persistence.

By the time the follow-up study was conducted, 2 of the 34 journals had ceased publication, and 2 had merged. Of the 31 remaining journals, 12 included examples of citations to electronic resources in their instructions to the authors. Only 5 of the journals instructed authors to use the fourteenth edition of *The Chicago Manual of Style,* whereas 6 referred authors to the fifteenth edition, which includes instructions for citing content on the Web. Three of the 31 journals included some mention of Web site instability, with 1 journal requiring authors to include digital object identifiers in citations to journal articles on the Web.

Summary and Conclusion

Persistence, as measured by the number of URLs in the sample citations that led directly or through a referred page to the cited content, degraded by 17.4 percent in the three years between the original and the follow-up studies. During the follow-up study, when searching for content not

found at the URL cited, the researchers had relatively less success searching the cited Web site and more success using Google, suggesting that the cited content was less closely associated with the Web site on which it resided at the time of the original study. Overall, cited content not found either at the URL included in the citation or elsewhere on the Web increased from 16.6 percent in the original study to 24.8 percent in the follow-up. Some cited content in this not-found category was accessible in the Internet Archive. When the content found only in the Internet Archive is factored in, the percentage of content available either on the Web or in the Internet Archive became 89.2 percent in the original study, but dropped to 80.6 percent in the follow-up study.

The researchers expected availability to degrade from persistent to accessible to not found over time, given the dynamism of the Web. However, some content moved in the opposite direction; that is, it was not persistent or accessible in the original study, but was found to be so in the follow-up study. These instances provide further evidence of the volatility of content published on the Web and substantiate the "but it was [or wasn't] there yesterday" experience. The results of searching the sample URLs in the Internet Archive also provide evidence of this volatility. Contrary to the permanence and stability suggested by the term "archive," this study demonstrates that content appears in and disappears from the Internet Archives. Fifty, or 14.5 percent, of the 344 URLs found in this archive in the original study *were not* found in the Internet Archive when the follow-up study was conducted, while 46, or 31.5 percent, of those not found in the Internet Archive in the original study, *were* in that archive during the follow-up study.

Two characteristics, citation content and implied domain, were associated with persistence in the original study, but not the follow-up. The only citation characteristic that remained associated with persistence through both studies was directory depth. In general, citations with URLs at the server level were the most likely to be persistent; as the directory depth increased, persistence decreased. An anomaly at the fourth level also appeared in the original study and may suggest that the relationship between directory depth and persistence is not linear.

Page type and source journal were found to be associated with persistence in the follow-up study, but not in the original. The findings that page type, which is derived from directory depth, is associated with persistence and that URLs pointing to navigation pages are more likely to be persistent than those pointing to content pages are logical. In contrast, the finding that citations from print journals are more often persistent than those from journals published only in electronic format is not easily interpreted. This finding suggests that article authors in electronic-only journals chose, or needed, to cite content that was less persistent than the content cited by authors who published in print journals. The researchers do not believe that, in terms of policy value, this is a useful finding.

The characteristics that were associated with availability—that is, with persistence or accessibility elsewhere on the Web—in the follow-up study were citation content, original domain, and implied domain. Directory depth was associated with availability in the original study, but not in the follow-up. Citation content was found to be inversely associated with persistence in the original study. In reporting the results of that study, the researchers hypothesized that the inverse relationship between the amount of information in the citation and persistence was a consequence of the study 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 or not they had found the cited content.[30]

The researchers employed this same methodology in the follow-up study, but found a direct relationship between the citation's completeness of information and cited content availability on the Web. Although this finding is more logical than the inverse relationship found in the original study, the methodological limitation remains; as a consequence, persistence rates may be overstated in this study as well as in the original. Future researchers could compensate for this limitation by consulting the source text to determine if they have found the cited content when searching incomplete citations or by limiting their samples to citations that include both URLs and complete bibliographic information.

The follow-up study underscores the association between domain, especially implied domain, and the likelihood that cited content will be found by future researchers. Original and implied domain were associated with persistence or accessibility in both studies and with availability in the Internet Archive in the follow-up study. Content cited on education, government, and organization servers were more likely to be found at the URL included in the citation or elsewhere on the Web, and in the Internet Archive, than content on commercial or other types of servers.

The follow-up study indicates that journal editors are providing more instruction to their authors regarding citing content on the Web. Instructions to authors are more likely to include examples of citations to Web sites and referrals to

style guides that prescribe citations with full bibliographic information, URLs, and dates viewed.

This study provides further documentation of the decline in citation persistence to content on the Web over time. Beyond that, by employing statistical tests of signification to citation characteristics related to persistence and availability, it provides strong support for the findings of previous research that were based on descriptive statistical analyses. This study supports the various findings by Koehler, Dimotrova and Bugeja, and Wren and colleagues of higher persistence rates of URLs at, or near, a Web site's directory level.[31] It also supports a number of studies, including those by Bugeja and Dimotrova, Bar-Ilan and Paritz, Ramsey, and Tan, Foo and Hui, that found URLs in the education domain to be the more persistent or stable than URLs in other domains.[32] However, other studies have found .gov sites to be the most stable, and more research is needed to sort out the relationship between domain and persistence.[33]

The original study and this follow-up are unlike other studies of persistence in that the researchers looked for the content cited by their sample's URLs in the Internet Archive. The studies' findings provide strong evidence that the Internet Archives is not a reliable source for cited content that is no longer available on the Web and underscore the need for both technical solutions and peer policies to address the problems associated with URL persistence.[34] The follow-up study confirms the importance of many of the recommendations made by the researchers based on the findings of the original study. Specifically, the professions and academic disciplines need to develop new citation conventions; journal publishers and editors need to better instruct authors about citing content and enforce new conventions as they are established; and authors, editorial staff, and publishers need to work together to develop and implement technologies to ensure that cited content is preserved and remains accessible to researchers and students over the long term.

References

1. Mary F. Casserly and James E. Bird, "Web Citation Availability: Analysis and Implications for Scholarship," *College & Research Libraries* 64, no. 4 (July 2003): 300–17.
2. Ibid., 300–302.
3. Carmine Sellitto, "The Impact of Impermanent Web-located Citations: A Study of 123 Scholarly Conference Publications," *Journal of the American Society for Information Science and Technology* 56, no. 7 (May 2005): 695–703; Carmine Sellitto, "A Study of Missing Web-Cites in Scholarly Articles: Towards an Evaluation Framework," *Journal of Information Science* 30 no. 6 (2004): 484–95.
4. Jonathan D. Wren et al., "Uniform Resource Locator Decay in Dermatology Journals: Author Attitudes and Preservation Practices," *Archives of Dermatology* 142, no. 9 (Sept. 2006): 1147–52.
5. Robert P. Dellavalle et al., "Going, Going, Gone: Lost Internet References," *Science* 302, no. 5646 (Oct. 31, 2003): 787–88.
6. Diomidis Spinellis, "The Decay and Failures of Web References," *Communications of the ACM* 46, no. 1 (Jan. 2003): 71–77.
7. Daniela V. Dimitrova and Michael Bugeja, "Consider the Source: Predictors of Online Citation Permanence in Communication Journals," *portal: Libraries and the Academy* 6, no. 3 (2006): 269–83.
8. Reneé Crichlow, Stefanie Davis, and Nicole Winbush, "Accessibility and Accuracy of Web Page References in 5 Major Medical Journals," *Journal of the American Medical Association* 292, no. 22 (Dec. 8, 2004): 2723–24.
9. John Markwell and David W. Brooks, "'Link Rot' Limits the Usefulness of Web-Based Educational Materials in Biochemistry and Molecular Biology," *Biochemistry and Molecular Biology Education* 31, no.1 (Jan. 2003): 69–72.
10. José Luis Ortega, Isidro Aguillo, and José Antonio Prieto, "Longitudinal Study of Content and Elements in the Scientific Web Environment," *Journal of Information Science* 32, no. 4 (2006): 344–51.
11. Michael Bugeja and Daniela V. Dimitrova, "Exploring the Half-Life of Internet Footnotes," *Iowa Journal of Communication* 37, no. 1 (Spring 2005): 77–86.
12. Judit Bar-Ilan and Bluma C. Peritz, "Evolution, Continuity, and Disappearance of Documents on a Specific Topic on the Web: A Longitudinal Study of 'Informetrics,'" *Journal of the American Society for Information Science and Technology* 55, no. 11 (Sept. 2004): 980–90.
13. Wallace Koehler, "A Longitudinal Study of Web Pages Continued: A Consideration of Document Persistence," *Information Research* 9 no. 2 (Jan. 2004), http://informationr.net/ir/9-2/paper174.html (accessed June 6, 2007)
14. Jeffrey D. Kushkowski, "Web Citation by Graduate Students: A Comparison of Print and Electronic Theses," *portal: Libraries and the Academy* 5, no. 2 (2005): 259–76.
15. Michelle M. Wu, "Why Print and Electronic Resources are Essential to the Academic Law Library." *Law Library Journal* 97, no. 2 (Spring 2005): 233–56, www.aallnet.org/products/pub_llj_v97no2_2005-14.pdf (accessed June 6, 2007).
16. Wendy Smith, "Still Lost in Cyberspace? Preservation Challenges of Australian Internet Resources," *Australian Library Journal* 54, no. 3 (2005), http://alia.org.au/publishing/alj/54.3/full.text/smith.html (accessed June 6, 2007).
17. Koehler, "A Longitudinal Study of Web Pages Continued."
18. Spinellis, "The Decay and Failures of Web References."
19. Wren et al, "Uniform Resource Locator Decay in Dermatology Journals."
20. Dimitrova and Bugeja, "Consider the Source."
21. Wren et al, "Uniform Resource Locator Decay in Dermatology Journals."
22. Sellitto, "The Impact of Impermanent Web-located Citations."
23. Markwell and Brooks, "'Link Rot' Limits the Usefulness of Web-Based Education Materials in Biochemistry and Molecular Biology."
24. Dellavalle et al., "Going, Going, Gone."

25. Bugeja and Dimitrova, "Exploring the Half-Life of Internet Footnotes."
26. Dimitrova and Bugeja, "Consider the Source."
27. Bar-Ilan and Peritz, "Evolution, Continuity, and Disappearance of Documents of a Specific Topic on the Web.'"
28. Lisa M. Schilling et al, "Digital Information Archiving Policies in High-Impact Medical and Scientific Periodicals," *Journal of the American Medical Association* 292, no. 22 (Dec. 8, 2004): 2724–26.
29. Casserly and Bird, "Web Citation Availability."
30. Ibid., 315.
31. Koehler, "A Longitudinal Study of Web Pages Continued"; Dimitrova and Bugeja, "Consider the Source"; Wren et al, "Uniform Resource Locator Decay in Dermatology Journals."
32. Bugeja and Dimitrova, "Exploring the Half-Life of Internet Footnotes"; Bar-Ilan and Peritz, "Evolution, Continuity, and Disappearance of Documents of a Specific Topic on the Web"; Mary Rumsey, "Runaway Train: Problems of Permanence, Accessibility, and Stability in the Use of Web Sources in Law Review Citations," *Law Library Journal* 94, no. 1 (Winter 2002): 27–39; 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.
33. Dimitrova and Bugeja, "Consider the Source"; 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): 105–08.
34. Steve Lawrence et al., "Persistence of Web References in Scientific Research," *Computer* 34, no. 2 (Feb. 2001): 26–31.