STUDY

Uniform Resource Locator Decay in Dermatology Journals

Author Attitudes and Preservation Practices

Jonathan D. Wren, PhD; Kathryn R. Johnson, MD; David M. Crockett; Lauren F. Heilig, BA; Lisa M. Schilling, MD; Robert P. Dellavalle, MD, PhD, MSPH

Objectives: To describe dermatology journal uniform resource locator (URL) use and persistence and to better understand the level of control and awareness of authors regarding the availability of the URLs they cite.

Design: Software was written to automatically access URLs in articles published between January 1, 1999, and September 30, 2004, in the 3 dermatology journals with the highest scientific impact. Authors of publications with unavailable URLs were surveyed regarding URL content, availability, and preservation.

Main Outcome Measures: Uniform resource locator use and persistence and author opinions and practices.

Results: The percentage of articles containing at least 1 URL increased from 2.3% in 1999 to 13.5% in 2004. Of the 1113 URLs, 81.7% were available (decreasing with time since publication from 89.1% of 2004 URLs to 65.4% of 1999 URLs) ($P<.001$). Uniform resource locator unavailability was highest in *The Journal of Investigative Dermatology* (22.1%) and lowest in the *Archives of Dermatology* (14.8%) ($P=.03$). Some content was partially recoverable via the Internet Archive for 120 of the 204 unavailable URLs. Most authors (55.2%) agreed that the unavailable URL content was important to the publication, but few controlled URL availability personally (5%) or with the help of others (employees, colleagues, and friends) (6.7%).

Conclusions: Uniform resource locators are increasingly used and lost in dermatology journals. Loss will continue until better preservation policies are adopted.

Arch Dermatol. 2006;142:1147-1152

Author Affiliations:
Department of Botany and Microbiology, Advanced Center for Genome Technology, University of Oklahoma, Norman (Dr Wren); Departments of Dermatology (Drs Johnson and Dellavalle and Ms Heilig), Preventive Medicine and Biometrics (Ms Heilig and Dr Schilling), and Medicine (Dr Schilling), University of Colorado at Denver and Health Sciences Center, Aurora; Colorado School of Mines, Golden (Mr Crockett); and Dermatology Service, Department of Veterans Affairs Medical Center, Denver (Dr Dellavalle).

APPROXIMATELY 80% OF dermatologists with Internet access use the Internet for medical updating and professional purposes.[1] Locating online health information, however, can be problematic because of the inconstant nature of Internet addresses, also known as uniform resource locators (URLs).[2-7] The continual flux of information on the Internet is reflected in the changing content and disappearance of URLs, which may become unavailable because of changes in Web site organization, hardware reconfiguration, and file renaming.[8]

Previous studies[2,4,5,7,9] examined the loss of cited URLs in journals encompassing multiple academic disciplines. Unlike previous estimates of URL use and availability, this study used an automated program to examine many full-text publications. To our knowledge, this is also the first study to survey authors with unavailable URLs regarding URL content and preservation.

METHODS

URL ASSESSMENTS

All online publications from January 1, 1999, to September 30, 2004, in the 3 dermatology journals with the highest scientific impact, according to the 2003 Institute of Scientific Information Journal Citation Reports, were examined: *The Journal of Investigative Dermatology*, *Archives of Dermatology*, and the *Journal of the American Academy of Dermatology*. Advertisements were excluded. Full-text publications were downloaded to a local hard drive and saved in HTML format using an automated script (Visual Basic 6). An automated program downloaded all full-text publications and extracted all URLs that were located within text sections. Hence, URLs embedded in tables or figures were not detected for this study. The availability of each URL was determined in September 2004 using a previously described program (Visual Basic 6).[4]

Article characteristics captured included PubMed identification, journal name, and date of publication. Data recorded for each URL included text location, URL address, top-level domain (eg, ".com" or ".gov"), directory depth,

page files are available on the Web server.) Of 100 randomly chosen URLs, 39 had accession dates.

Of 204 unavailable URLs, the content of 120 (58.8%) was recoverable in some form using the IA. This increased overall recoverability of at least partial content to 92.5% of URLs in all journals for all years.

SURVEY OF AUTHORS WITH UNAVAILABLE URLs

A total of 102 unique corresponding authors of articles with unavailable URLs were e-mailed a survey (**Figure 2**) regarding the unavailable URLs, and 67 (65.7%) responded. Less than half (43.9%) had attempted to access the URL after publication, suggesting that most URLs become unavailable without the knowledge of the citing authors. Most (55.0%) of the cited URLs reference content outside the direct control of the authors and their coworkers. Of 60 respondents, 7 (11.7%) had direct control over URL availability.

Most authors (32 [51.6%] of 62) did not know why the URL they cited was unavailable. However, consistent with previous findings,[4] about 11% of URLs were misspelled in the final publication. Three (4.5%) indicated that the URLs became unavailable because of a lack of funding or support.

Most responding authors (63.9%) had preserved cited URL content, most commonly (29.5%) by printing it. Few (4.9%) had used an Internet-based archive for content preservation. Most (55.2%) agreed that the content of the cited URL was important to their publication, most often (60.7%) as a means of contributing to background information for the study. The most common reason for citing a URL was to provide additional information about a topic (54.1%) or to link to additional data or analyses (37.7%). Only 14.3% indicated that an alternative source of data (other than the cited URL) was available at publication.

Most often, the nature of the URL was a text-based document (46.8%), which can be backed up by several means, but 45.2% of the URL links pointed to either a database (33.9%) or a software program (11.3%), which is not as straightforward to back up.

URL POLICIES BY JOURNAL

Since January 2002, the "Instructions for Authors" of the *Archives of Dermatology* (http://archderm.ama-assn.org) has provided an example Internet reference with an accession date and has recommended that authors retain a printed copy of any referenced Internet-only information to ensure access to cited information if the URL is altered or disappears. The "Instructions for Authors" of the *Journal of the American Academy of Dermatology* (http://www.eblue.org) and *The Journal of Investigative Dermatology* (http://www.jidonline.org) do not mention an Internet referencing policy. None of the 3 journals restricted URLs to specific locations in articles.

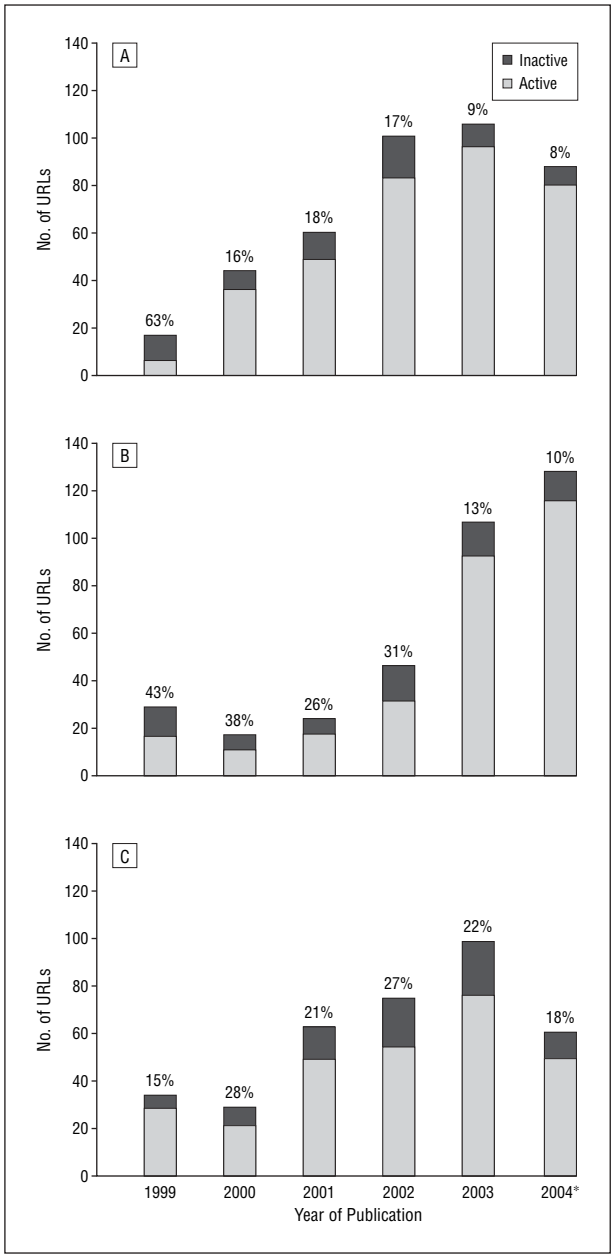

Figure 1. Uniform resource locator (URL) use in the *Archives of Dermatology* (A), the *Journal of the American Academy of Dermatology* (B), and *The Journal of Investigative Dermatology* (C) from 1999 to 2004. Percentages indicate unavailable URLs for each year. The asterisk indicates that data in 2004 were only from January through September.

COMMENT

This study confirms that URLs are increasingly cited as sources of scholarly information in dermatology journals, and that a significant portion of cited information is no longer available. Of 1113 URLs examined, 18.3% were unavailable. The probability a URL would become unavailable was significantly associated with increasing time since publication, journal, top-level domain, and greater directory depth, but not with the presence of a tilde or an accession date. These associations support the findings of Casserly and Byrd[2] in information science journals. Of unavailable URLs, 58.8% were recoverable in some form in the IA, and an assessment of content relevance of randomly selected URLs yielded no irrelevant information content. This study also corroborates findings that 12% of URLs in MEDLINE abstracts contain spelling or formatting errors that render the published URL unavailable.[4]

1. Since publication, have you or a coauthor attempted to access the currently inactive Internet reference (URL)?

Response	Response %	Response Total
Yes	43.9	29
No	**56.1**	**37**
Total respondents		**66**
(Skipped this question, 1)		

2. Why is the Internet (URL) inaccessible?

Response	Response %	Response Total
Misspelled in publication	11.3	7
Server down	0	0
URL is currently active	8.1	5
Don't know	**51.6**	**32**
Other (please specify)	29.0	18
Total respondents		**62**
(Skipped this question, 5)		

3. How have you or a coauthor preserved the content of this Internet reference at the time of citation? Please answer this question regardless of the URL's current status. (Check all that apply.)

Response	Response %	Response Total
Hard copy (eg, printed information)	29.5	18
Single digital copy (eg, stored on floppy disk, CD, hard drive)	19.7	12
Multiple digital copies (eg, backup tapes, copies kept on multiple computers)	9.8	6
Internet-based archive (eg, Internet Archive, FURL, arXiv)	4.9	3
Has not been preserved	**36.1**	**22**
Other (please specify)	18.0	11
Total respondents		**61**
(Skipped this question, 6)		

4. Please indicate your level of agreement with the following statements.

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Not Applicable	Response Average
The URL content is important to the publication.	20.7% (12)	**34.5% (20)**	27.6% (16)	10.3% (6)	1.7% (1)	5.2% (3)	**2.35**
The URL content strengthens scientific arguments in the publication.	14.3% (8)	**28.6% (16)**	**28.6% (16)**	10.7% (6)	1.8% (1)	16.1% (9)	**2.49**
The URL content contributes to study background.	16.1% (9)	**44.6% (25)**	17.9% (10)	3.6% (2)	1.8% (1)	16.1% (9)	**2.17**
The URL content contributes to study methods.	16.1% (9)	**25.0% (14)**	12.5% (7)	16.1% (9)	5.4% (3)	**25.0% (14)**	**2.60**
The URL content contributes to study results.	16.1% (9)	**26.8% (15)**	14.3% (8)	10.7% (6)	10.7% (6)	21.4% (12)	**2.66**
The URL content contributes to study conclusion.	16.1% (9)	**23.2% (13)**	19.6% (11)	14.3% (8)	8.9% (5)	17.9% (10)	**2.72**
Total respondents							**58**
(Skipped this question, 9)							

5. Why was the URL referenced? (Check all that apply.)

Response	Response %	Response Total
Provide more information on a topic discussed	**54.1**	**33**
Provide access to additional data or analysis	37.7	23
Provide access to measurement instruments (eg, survey)	3.3	2
Provide more information on a product used	4.9	3
Other (please specify)	19.7	12
Total respondents		**61**
(Skipped this question, 6)		

6. What type of content was referenced by this URL? (Check all that apply.)

Response	Response %	Response Total
Database	33.9	21
Software program	11.3	7
Text document	**46.8**	**29**
Image	4.8	3
Video	1.6	1
Audio track	0	0
Other (please specify)	11.3	7
Total respondents		**62**
(Skipped this question, 5)		

7. Was an alternative to a URL reference available at the time of manuscript submission (eg, a print version of the reference)?

Response	Response %	Response Total
Yes	14.3	9
No	**49.2**	**31**
Don't know	36.5	23
Total respondents		**63**
(Skipped this question, 1)		

8. What role could you play to help make the URL you cited accessible again?

Response	Response %	Response Total
I can do it myself (eg, I am the Webmaster)	5.0	3
I can ask someone I know (eg, employee, colleague, or friend)	6.7	4
I would have to contact someone I do not know (eg, I would have to search for contact information for the person who managed the URL)	41.7	25
Other (please specify)	**46.7**	**28**
Total respondents		**60**
(Skipped this question, 7)		

Figure 2. Dermatology article author responses regarding unavailable uniform resource locators (URLs). CD indicates compact disc. Percentages are based on the denominator of total respondents for each question. Boldface indicates the most frequent response.

The Internet serves as an invaluable network that provides global access to information. However, the average lifespan of a Web site is far from sufficient to ensure reliable long-term availability.[10,11] Because of the inconstant nature of URLs, neither publishers nor authors are able to guarantee the long-term accuracy or availability of digital information referenced in dermatology journals. Effective solutions will likely require a collaborative effort on the part of researchers, authors, and journal editors.

Digital archiving resources offer one approach to preserving digital information. The IA, a public nonprofit

organization, was constructed with the purpose of archiving Internet content and can often locate content of otherwise unrecoverable URLs, with snapshots taken on multiple dates. Unfortunately, archived versions of dynamic Web pages may not fully retain functionality, and other URLs, including those that are password protected or that block Web crawlers, are not available for archiving. Moreover, IA archiving typically takes place every couple of months, so changes made during this time will not be preserved. Thus, while 58.8% of unavailable URLs were classified as "recoverable" on the IA, the information recovered could not be verified as identical to that viewed and cited by the author.

An additional problem is the possibility of copyright infringement associated with preserving Internet content that is not the intellectual property of the citing author. In terms of scientific publications, for example, a recent study[12] demonstrated that many authors make journal article reprints available online, which may in turn be archived by the IA regardless of whether the journals want this content freely available. It is difficult, if not impossible, in many cases for the IA to ascertain what content has been legally posted and what content may be illegal. Web authors may ask to have their electronic content removed from the IA (more information is available at: http://www.archive.org/about/faqs.php), which may further limit the ability of the IA to preserve URLs.

Other efforts to remedy the problem of URL loss exist (**Table 2**). Software programs, such as Peridot (IBM Corporation, White Plains, NY)[13] and Xenu's Link Sleuth (http://home.snafu.de/tilman/xenulink.html), automate the updating of linked Web sites. Another program (FURL; LookSmart, Ltd, San Francisco, Calif) (http://www.furl.net) also serves as a digital information archive, but preserves only URL content submitted by individuals for personal archiving. Alternatively, WebCite specifically targets preservation of URLs in academic journals.

Readers commonly use additional recovery methods, such as typing the higher-level stem (beginning) of an unavailable URL or the entire URL into a search engine such as Google. About 30% of the unavailable URLs in our study yielded *prima facie* relevant information using these methods. In the end, however, the reader does not know with certainty that this retrieved information is, in fact, the originally cited information.

Uniform resource locator content might also be better preserved by using more permanent alternatives to URLs for locating information on the Internet. Uniform resource locators serve as the name (identifying content) and address (identifying location) for Internet resources, rendering cited content unavailable if either one changes. Alternatively, permanent URLs are associated with specific URLs, but are unchanging, effectively redirecting the Web client to the correct URL via an intermediary resolution service.[8] This process is not fully location independent, and its success depends on the reliability of permanent URL maintainers to update the associated URL if it changes.[8] Other alternatives are uniform resource names, permanent location-independent identifiers of cited resources that rely on a resolving service; and digital object identifiers, which identify a digital object by name only, using a persistent novel identifier embedded within a URL.[14]

Table 2. Tools for URL Preservation and Recovery

Tool (Web Address)	Category	Description
Internet Archive (http://www.archive.org/)	Digital archive	Regularly crawls the Internet to archive all available URLs; all archived URLs available to the general public
FURL (http://www.furl.net/)*	Digital archive	Contains only URLs submitted by individuals; consists of personal archives not available to the general public
WebCite (http://www.webcitation.org)	Digital archive	Permits authors and editors to archive selected Web pages
Peridot software† (http://en.wikipedia.org/wiki/Peridot_%28software%29)	Web site maintenance tool	Automated program that updates links on internal Web sites
Xenu's Link Sleuth (http://home.snafu.de/tilman/xenulink.html)	Web site maintenance tool	Checks Web pages for unavailable URLs
URL (http://www.w3.org/Addressing/)	Identifier	The most commonly used identifier; specifies the name and the location of Internet content
PURL (http://purl.oclc.org/)	Identifier	Uses an intermediate resolution service to redirect browsers to the correct URL; not fully location independent; requires updating
URN (http://www.w3.org/Addressing/)	Identifier	Location independent; becomes unavailable only if Internet content is deleted
DOI (http://www.doi.org/)	Identifier	Location independent; embedded within URL; relies on a reference-linking service

Abbreviations: DOI, digital object identifier; PURL, permanent URL; URL, uniform resource locator; URN, uniform resource name.
*Developed by LookSmart, Ltd, San Francisco, Calif.
†Developed by IBM Corporation, White Plains, NY.

In light of the limitations of URL preservation options, the importance of improving journal policies regarding URLs cannot be overstated. In a recent study[15] of the top 100 medical and scientific journals, as rated by the Institute for Scientific Information for scientific impact, only one, the *Archives of General Psychiatry*, had a URL preservation policy stated in the "Instructions for Authors." Of the 3 dermatology journals, only the *Archives of Dermatology* gives specific mention to Internet referencing in the "Instructions for Authors," using the same policy used by the *Archives of General Psychiatry*. The *Archives of Dermatology* also demonstrated a significantly lower rate of unavailable URLs in this study. Pub-

lishers, editors, and authors should work together to discover and implement feasible solutions to URL content loss[15-18] by (1) requiring authors to retain digital backup or printed copies of cited Internet-only information to facilitate content recovery should a URL become unavailable and (2) advocating the inclusion of referenced Internet content in an online archive (Table 2). In addition, URLs need systematic double checking before publication to minimize unavailability due to spelling errors or misprints.

The adoption of standard electronic referencing policies, the use of Internet-based archives, and collaboration between authors and publishers will hopefully lead to more permanent URL availability in dermatology journals. Ultimately, widespread acceptance and support for these easily implemented policies could serve as a model for all medical literature.

Accepted for Publication: January 9, 2006.
Correspondence: Robert P. Dellavalle, MD, PhD, MSPH, Dermatology Service, Department of Veterans Affairs Medical Center, 1055 Clermont St, Mail Code 165, Denver, CO 80220 (robert.dellavalle@uchsc.edu).
Author Contributions: *Study concept and design:* Wren, Schilling, and Dellavalle. *Acquisition of data:* Wren, Johnson, Crockett, Heilig, and Dellavalle. *Analysis and interpretation of data:* Wren, Johnson, Heilig, and Schilling. *Drafting of the manuscript:* Johnson, Crockett, Heilig, Schilling, and Dellavalle. *Critical revision of the manuscript for important intellectual content*: Wren, Heilig, Schilling, and Dellavalle. *Statistical analysis:* Heilig. *Obtained funding:* Dellavalle. *Administrative, technical, and material support:* Wren, Heilig, Schilling, and Dellavalle. *Study supervision:* Heilig and Dellavalle.
Financial Disclosure: None reported.
Funding/Support: This study was supported by grant EPS-0447262 from the National Science Foundation Experimental Program to Stimulate Competitive Research (Dr Wren); grant T32 AR07411 from the National Institutes of Health (Dr Johnson); in part by research grant R25 CA49981 from the National Cancer Institute Education (Mr Crockett); grant 5 D14HP00153, a Faculty Development in Primary Care Health Services Research Award (Dr Schilling); and grant K-07 CA92550 from the National Cancer Institute (Dr Dellavalle).
Previous Presentation: This study was presented at the Fifth International Congress on Peer Review and Biomedical Publication; September 16, 2005; Chicago, Ill.
Acknowledgment: We thank John Kittelson, PhD, Department of Preventive Medicine and Biometrics, University of Colorado at Denver and Health Sciences Center, for statistical advice; and Eric Hester, MD, Jennifer Myers, MD, Renee D'Ambrosia, MD, Kristy Lundahl, MBA, and Shayla Francis, MD, for their work on this project.

REFERENCES

1. Gjersvik PJ, Nylenna M, Aasland OG. Use of the Internet among dermatologists in the United Kingdom, Sweden and Norway. *Dermatol Online J*. 2002;8:1.
2. Casserly M, Byrd J. Web citation availability: analysis and implications for scholarship. *Coll Res Libr*. 2003;64:300-317.
3. Currò V, Buonuomo PS, De Rose P, Onesimo R, Vituzzi A, D'Atri A. The evolution of Web-based medical information on sore throat: a longitudinal study. *J Med Internet Res*. 2003;5:e10. http://www.jmir.org/2003/2/e10. Accessed May 9, 2006.
4. Wren JD. 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics*. 2004;20:668-672.
5. Lawrence S, Coetzee F, Glover E, et al. Persistence of Web references in scientific research. *IEEE Comput*. 2001;34:26-31.
6. Koehler W. Web page change and persistence: a four-year longitudinal study. *J Am Soc Inf Sci*. 2002;53:162-171.
7. Dellavalle RP, Hester EJ, Heilig LF, et al. Information science: going, going, gone: lost Internet references. *Science*. 2003;302:787-788.
8. Schafer K, Weibel S, Jul E. The PURL project. *J Libr Adm*. 2001;34:123. http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003338. Accessed November 14, 2005.
9. Hester EJ, Heilig LF, Drake AL, et al. Internet citations in oncology journals: a vanishing resource? *J Natl Cancer Inst*. 2004;96:969-971.
10. Kahle B. Preserving the Internet. *Sci Am*. 1997;276:82-83.
11. Spinellis D. The decay and failures of Web references. *Commun ACM*. 2003;46:71-77.
12. Wren JD. Open access and openly accessible: a study of scientific publications shared via the Internet. *BMJ*. 2005;330:1128. Accessed May 9, 2006. doi:10.1136/bmj.38422.611736.E0.
13. Twist J. Web tool may banish broken links. http://news.bbc.co.uk/1/hi/technology/3666660.stm. Accessed November 14, 2005.
14. Caplan P. DOI or don't we? http://info.lib.uh.edu/pr/v9/n1/capl9n1.html. Accessed November 14, 2005.
15. Schilling LM, Kelly DP, Drake AL, Heilig LF, Hester EJ, Dellavalle RP. Digital information archiving policies in high-impact medical and scientific periodicals. *JAMA*. 2004;292:2724-2726.
16. Johnson KR, Hester EJ, Schilling LM, Dellavalle RP. Addressing Internet reference loss. *Lancet*. 2004;363:660-661.
17. Kelly DP, Hester EJ, Johnson KR, et al. Avoiding URL reference degradation in scientific publications. *PLoS Biol*. 2004;2:e99. doi:10.1371/journal.pbio.0020099.
18. Schilling LM, Wren JD, Dellavalle RP. Bioinformatics leads charge by publishing more Internet addresses in abstracts than any other journal [letter]. *Bioinformatics*. 2004;20:2903. doi:10.1093/bioinformatics/bth385.