



404 not found: the stability and persistence of URLs published in MEDLINE

Jonathan D. Wren

Advanced Center for Genome Technology, Department of Botany and Microbiology,
The University of Oklahoma, 620 Parrington Oval Rm. 106, Norman, OK 73019, USA

Received on June 18, 2003; accepted on June 25, 2003

Advance Access publication January 22, 2004

ABSTRACT

Motivation: The advent of the World Wide Web has enabled unprecedented supplementation of traditional journal publications, allowing access to resources, such as video, sound, software, databases, datasets too large to publish, and even supplementary information and discussion. However, unlike traditional publications, continued availability of these online resources is not guaranteed. An automated survey was conducted to quantify the growth in Uniform Resource Locators (URLs) published to date in MEDLINE abstracts, their current availability and distribution by journal.

Results: Of 1630 unique URLs identified, formatting and/or spelling errors were detected within 201 (12%) of them as published. After corrections were made, a survey revealed that ~63% of these URLs were consistently available, and another 19% were available intermittently. The rate of failure was far worse for anonymous login to FTP sites, with only 12 of 33 sites (36%) responding. This survey also shows that journals vary disproportionately in the number of web citations published, suggesting policy implementation among a few could have a profound impact overall. Out of the 306 journals with a URL published in an abstract, *Bioinformatics* published the most (12% of total).

Availability: URL database and program available by request.

Contact: Jonathan.Wren@OU.edu

INTRODUCTION

Scientists have used the Internet almost since its inception in 1969, but it was data-intensive undertakings such as the Human Genome Project (HGP) that spawned an increased reliance upon it through a community need to share, analyze and annotate data (e.g. see Nowak, 1993). The Internet was the most practical medium to provide such analytical tools and resources to the scientific community. As the cost of computer hardware dropped, connections became more widespread and basic computer literacy increased, the Internet became an increasingly popular medium to access a broad variety of scientific information and resources. The publication of links to online resources, however, is a slightly more recent development. In 1994, the first web page Uniform

Availability of URLs published in MEDLINE

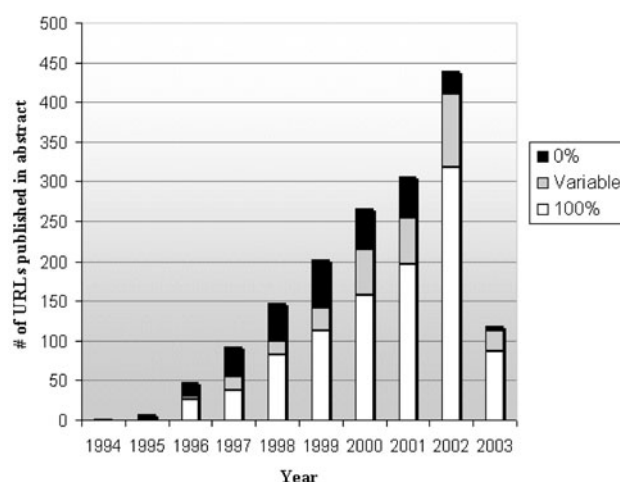


Fig. 1. The number of unique URLs published in MEDLINE, plotted as a function of time. The proportion of URLs currently accessible (as of this survey) is broken down by year and displayed as shading on each bar. Online resources are playing an increasingly important role in scientific research, as evidenced by their inclusion in the abstract. Not surprisingly, the more recent the publication, the more likely the URL is still active. Data from 2003 is only up until April.

Resource Locator (URL) was published within a MEDLINE abstract (Williams, 1994), and preceded a rapid growth in the inclusion of online resources as being of central importance to published research. Since then, the number of abstracts containing an URL has increased rapidly (Fig. 1). From 1997 to 2002 the average annual increase in the number of these URLs was 47% , compared with a consistent average annual increase of 5% in the total number of MEDLINE records over the same period.

Today, the Internet has become a valuable, perhaps indispensable, resource in conducting scientific research (Lawrence and Giles, 1999), not just because of the added convenience of rapid information retrieval and sharing, but because it also provides a means of making resources available that the printed media simply cannot. Information is

often gathered in forms not amenable to the traditional printed media of journals, such as video, audio, software programs and database access. Often, data amenable to the printed media is still too large to be published in a journal. The HGP ushered in an enormous amount of sequence data that would be impractical for others to copy manually for their own use, base for base, even if printed in a journal. This same data could potentially change as more sequence was gathered and base quality checking methods improved. This trend of generating large datasets persists and now encompasses other data-intensive fields, such as microarray analysis and proteomics. The need to store, manage, distribute and analyze such information has moved the Internet from being of supplemental importance to a primary resource for many biologists and medical practitioners. In many ways, online resources are part of a new revolution in scientific research that increases the accessibility of data, broadens perspective and provides access to tools that aid scientific research. These resources, however, also present new challenges to the traditional scientific peer-review process for the publication of research, since they can change in content. Furthermore, because these online resources can disappear altogether, they also provide a unique challenge in terms of preserving our historical scientific knowledge and evaluating new resources as they arise.

Any experienced Internet user will have encountered the traditional '404 not found' message returned by a web server in response to an URL that is no longer available. Several studies to date have dealt with this general problem of 'URL decay'. Koehler, for example examined both the accessibility and content of 360 randomly chosen URLs obtained from Webcrawler over 3 years. He found that ~50% of them were still active at the end of this time and most had changed in content (Koehler, 1999, 2002). More specific to the biomedical field, another study examined 184 web sites available between 1998–1999 with information related to the herbal remedy *Opuntia* and found only 76 (41%) of them were still available in 2002 (Veronin, 2002). While both these studies found a very high rate of URL decay, the primary source of the URLs analyzed was obtained through search engines and not the peer-reviewed literature. Websites, in general, have a relatively short lifespan. Numbers vary, regarding the lifespan of the average webpage, with one author publishing ranges from 44 days (Kahle, 1997) to 75 days (http://www.archive.org/sciam_article.html). Optimally, the lifespan of peer-reviewed resources related to scholarly pursuits would be permanent, but we would at least hope their lifespan would be substantially longer than the average web page. It is therefore important to identify to what extent URL decay is a problem in the scientific literature.

Studies concerning the increase and persistence of published URLs have been conducted in other fields such as computer science and law, both of which document a rapid increase in the number of URLs referenced by papers over time and a time-dependant decay in URL availability. Spinellis

(2003) surveyed the full-text of two journals (*Communications of the ACM* and *IEEE Computer Society*) over 4 years and found that 72% of URLs contained within were still active. Lawrence and co-workers conducted a large-scale analysis of the CiteSeer database (Lawrence *et al.*, 1999), finding a time-dependent URL decay with availability rates ranging from 47% (for 1994 data) to 77% (for 1999 data) (Lawrence *et al.*, 2001). In the legal field, the URL decay rate was found to be higher within full-text documents, ranging from 30% availability (1997 data) to 62% availability of published websites in 2001 (Rumsey, 2002). However, trends have yet to be established for the biomedical literature as contained in MEDLINE, and given an increased reliance upon online resources in conducting biomedical research it is important to do so.

The aim of this report is to estimate the number, growth and current availability of all URL addresses previously published within MEDLINE abstracts, as well as provide statistical estimates of URL uptime and continuity in the biomedical literature. Abstracts were chosen for this study because they are both readily available in electronic format and intended to summarize the most important aspects of the published work. Thus, URLs appearing in abstracts are expected to be of central importance to the work as a whole.

SYSTEMS AND METHODS

The National Library of Medicine graciously provided MEDLINE records in XML format. Scripts were written in Visual Basic 6.0 (SP5) to process these records searching for the occurrence of strings within a title or abstract suggesting the presence of a URL, such as 'http://' and 'ftp://'. Gopher sites (gopher://) were not analyzed, as these text-only servers have been widely replaced with HTTP. URLs were quality-checked to eliminate unnecessary trailing characters and entered into a database (Microsoft Access 2000) along with their referring PubMed ID (PMID) and date of publication. Case was preserved when checking URL availability and all URLs within an abstract were recorded.

Errors in URL formatting were corrected either by a set of heuristic rules or manual editing. Inappropriate spaces in the middle of a URL were not counted as errors, but were corrected for by examining the next consecutive string after the URL for either the presence of '/' characters or a '*.*' pattern. The need for heuristic corrections was recognized as a large portion of errors seemed to fall into one of several classes, such as insertion of inappropriate spaces into the URL the use of backward '\\' instead of forward slashes '/', non-alphanumeric characters in the URL (usually from non-English speaking websites), and the inclusion of erroneous characters such as 'http:(/)' or plus signs (e.g. 'http://www+++'). URL addresses that begin with 'http@' were non-standard, but work nonetheless when typed in a browser and thus were not considered errors. Manual corrections of URLs were only

made when noticed by the author and the revised URL verified to work when typed into a browser.

A script was written to process each of these URLs, using the Microsoft Component Objects Internet Transfer Control (ITC) version 6.0. An URL was considered inaccessible if it did not respond within 60 s or returned an error message indicating that the page could not be retrieved (e.g. '404 not found', 'page was unavailable', 'file not found', etc.). Messages indicating redirection (e.g. 'you are being redirected', 'this page has moved', etc.) were considered available URLs, but flagged to indicate that a difference exists between active and published URLs. Pages that blocked automated access attempts were still considered available since no error message was returned. Each of the 1630 URLs was checked almost daily, including weekends, at random times over the course of one month from March 21, 2003 to April 21, 2003. Successful and unsuccessful attempts were recorded for each run. A bogus URL was included as a negative control and a series of 50 URLs known to be functional prior to the survey were included as positive controls. Runs were terminated and discarded if 25 consecutive access failure attempts were recorded (the presumption being that the failure is likely on this end of the communications link).

12 625 338 MEDLINE records were processed, with publication dates ranging from 1966 to April 2003 (including some abstracts appearing ahead of publication). These records contained 6 958 710 abstracts, which when analyzed were found to contain a total of 1630 unique URLs pointing to web pages (i.e. 'http://' addresses) and 33 pointing to FTP websites.

RESULTS

A preliminary survey of availability was first conducted and unavailable web pages were examined manually for any patterns. A number of errors in URL formatting were noted. Most (185) were identified and corrected by a set of heuristic string replacement rules, while 16 were the result of specific spelling errors noticed during manual examination. One spelling error was even found in a recent publication of this author (Wren and Garner, 2002), where an 'o' within the URL should have been an 'a'! A total of 201 errors were recorded and reported to the National Library of Medicine. Any errors eventually determined to be a result of data-entry or automated formatting will be corrected in PubMed (J.Rosov, personal communication). Errors ultimately attributable to authors or journals must be corrected by submission of an erratum (see <http://www.nlm.nih.gov/pubs/factsheets/errata.html>).

After corrections were made, these 1630 URLs were accessed using the ITC (see Systems and methods section) approximately once a day over the course of one month, recording whether or not each URL was accessible. A total of 30 accessibility checks were conducted. This time-course survey was used to obtain an estimate of uptime and to allow

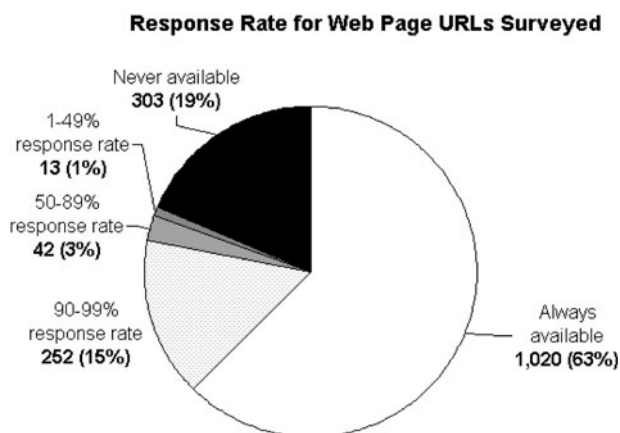


Fig. 2. Percentage of times 1630 web page URLs were available during the survey period. Most sites were available 100% of the time, and at least 78% could be considered highly accessible ($\geq 90\%$ uptime).

Percent of URLs published by year that are inaccessible between March and April 2003

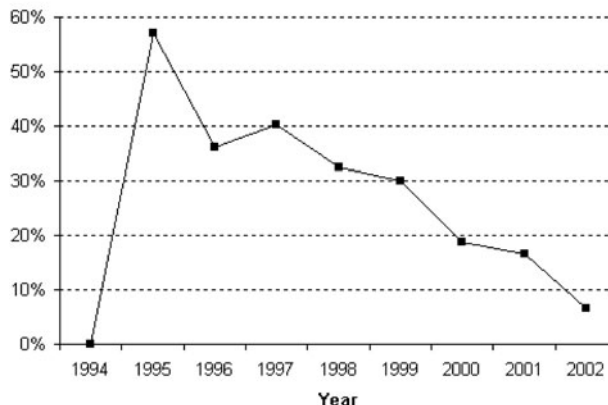


Fig. 3. The probability that a previously published URL will be available today is a function of how long ago it was published. Similar time-dependent decay functions have been noted in other studies as well (Lawrence *et al.*, 2001; Koehler, 2002; Rumsey, 2002; Spinellis, 2003). Only two URLs were published in MEDLINE abstracts during 1994, but both were consistently available during the study.

for the possibility that pages initially found to be unavailable were due to temporary, resolvable problems. A total of 1020 web pages were found to be available 100% of the time (62.6%), 307 varied in their availability (18.8%) and 303 pages were consistently unavailable (18.6%) (summarized in Fig. 2). As expected, the probability that a previously published URL was accessible during the survey was a function of how long ago it was published (Fig. 3). In 79 cases, the web page stated that the URL had changed and redirected access, indicating a departure from the original published

Response Rate for FTP URLs Surveyed

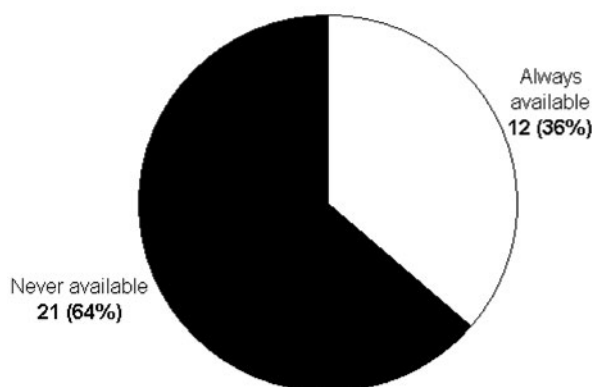


Fig. 4. Availability of the 33 FTP sites published in MEDLINE abstracts. No variability in uptime was noted during the survey.

URL but preserving accessibility. Automatic (undeclared) redirects were not recorded.

For the FTP sites, it was surprising that only 12 out of 33 sites (36%) were available for anonymous login (Fig. 4). It is not clear why the failure rate for FTP sites is so much higher than for HTTP, but one possible reason may be the relative ease by which commands can be issued from a remote client to the active server via each protocol. FTP servers can allow more mechanisms of entry to hackers than HTTP servers do, whose role is primarily to display information.

Some of these non-responding URLs were found to be the consequence of simple formatting errors and/or variability in server search capability. For example, some servers require a slash ('/') to be added to the end of a URL to distinguish directory structures from file names, while others attempt automatic resolution. Other formatting errors involved deviations from a standardized format, such as URLs kept by the Digital Object Identifier (DOI) Foundation (<http://www.doi.org/index.html>), which is used by publishers such as Springer-Verlag (<http://www.springer.de/>). In this survey, 54 out of 57 published URLs at the DOI were detected as working. The three URLs that returned errors were easy to recognize and fix manually only because the other 54 URLs served as a template for the correct format. One error involved elimination of part of the domain name (<http://dx.org/10.1007/s00134-002-1542-9>, which should be <http://dx.doi.org/10.1007/s00134-002-1542-9>) (Bosman *et al.*, 2003), while the other was a failure to include dashes in the appropriate places in the URL (<http://dx.doi.org/10.1007/s008940020092y> should be <http://dx.doi.org/10.1007/s00894-002-0092-y>) (Friedemann *et al.*, 2002), and the other erroneously included the prefix 'www'. However, users cannot be reasonably expected to anticipate which permutation of the published URL is correct, since most of the necessary corrections would not be apparent to the average reader.

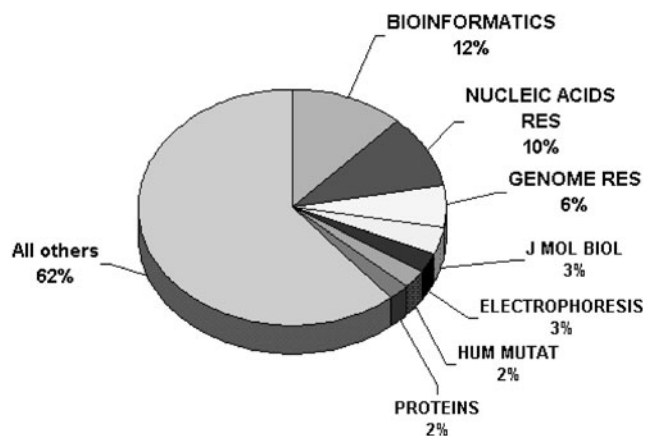


Fig. 5. Journals vary in the number of URLs published in their abstracts. A total of 306 different journals published at least one URL in an abstract, but the top three of these journals together comprise 28% of all URLs published in abstracts. Presumably, URLs published within the abstract are of central or primary importance to the published research.

Finally, the distribution of URLs published in abstracts was examined by journal name. It was found that a relatively small subset of journals account for a disproportionately large number of published URLs (Fig. 5), with the journal *Bioinformatics* accounting for 12% of the total (this number includes abstracts from when the journal was known as *Computer Applications in the Biosciences*). Not too surprisingly, the top three journals on this list were those with a focus or reliance upon bioinformatics methods and resources.

DISCUSSION

If a web page can be considered generally available when it is up at least 90% of the time, then the overall availability rates of URLs published in MEDLINE abstracts is 78% (see Fig. 2). It is difficult to compare this rate directly to the computer science and law studies mentioned previously since only abstracts were analyzed, but it seems reasonable to assume that more resources or attention would be expended to ensure the continuity of a primary resource (e.g. software tool, database) than a tangential, supplemental or secondary resource. However, 19% of published links to online resources were completely inaccessible during this survey. While this rate might be considered low compared with other fields, it still represents the loss of approximately one in five published resources. As more and more URLs are published, it is not clear how this rate will be affected as time goes by.

If this URL decay is to be prevented or at least mitigated it seems clear that some component of the current system must be changed. Authors currently bear most of the responsibility for URL continuity, and are not always in a position to ensure it. Authors often leave institutions in which they

develop online resources, and such institutions will rarely have any official interest in maintaining the author's work (and in some cases are not qualified to). While the author's interest in URL maintenance may continue, for legal reasons many institutions will even have policies against former employees being able to change web site content. Reviewers seem ill equipped to handle this burden, as they are only in a position to evaluate URLs as they stand at the time of publication, and unable to judge change or future availability. It seems that, despite the additional administrative burden, journal publishers would be best suited to assume yet another 'gate keeping' function. Fortunately, the disproportionate use of web citations among journals suggests that a relatively broad impact on overall URL decay could be obtained through policy implementations in the relatively few journals whose interests frequently involve online resource links.

Perhaps, the most promising solution to the problem of URL decay in all fields is to provide an archive of web pages published on the Internet. While this seems like an overly ambitious project given the size and rapid growth of the World Wide Web, the Internet Archive (IA) (<http://www.archive.org/>) is committed to taking on this task, archiving an estimated 12 terabytes of data per month (<http://www.archive.org/about/faqs.php#9>). A small subset of the unavailable published URLs in this survey were found on this site, some conveniently archived at different time points. However, two downsides currently exist to the use of this engine: first, information is obtained from the Alexa search engine, and previous studies have shown that the most web pages indexed by any search engine is approximately one-third of the total web pages available (Lawrence and Giles, 1999), suggesting a significant number of sites might simply not be indexed. Second, the type of information archived is limited (e.g. database queries will no longer be available). Nonetheless, the IA represents perhaps the only solution for researchers looking for an URL that is no longer available.

CONCLUSION

Citation and publication of online resources is becoming increasingly common in biology and medicine. Many of these resources are being lost in a time-dependent manner, as URLs pointing to them are no longer able to retrieve the intended resource. The rate of failure for FTP sites was found to be far worse (64%) than for HTTP sites (19%). While online resource citation is becoming increasingly common

in all journals, the distribution of published URLs is heavily skewed towards journals whose focus involves bioinformatics methods and resources. Policy implementation by a few of these leading journals could have a profound effect upon overall trends.

ACKNOWLEDGEMENTS

I would like to thank Tyrrell Conway for his helpful review of this manuscript. This work was funded by NSF-EPSCoR grant no. EPS-0132534.

REFERENCES

- Bosman,R.J., Rood,E., Oudemans-van Straaten,H.M., Van der Spoel,J.I., Wester,J.P. and Zandstra,D.F. (2003) Intensive care information system reduces documentation time of the nurses after cardi thoracic surgery. *Intensive Care Med.*, **29**, 83–90.
- Friedemann,R., Naumann,S. and Brickmann,J. (2002) Molecular dynamics studies on the aggregation of Y-shaped fluoroalkanes. *J. Mol. Model (Online)*, **8**, 266–271.
- Kahle,B. (1997) Archiving the Internet. *Sci. Am.*, **273**, 82–83.
- Koehler,W. (1999) An analysis of Web page and Web site constancy and permanence. *J. Am. Soc. Inf. Sci.*, **50**, 162–180.
- Koehler,W. (2002) Web page change and persistence—a four-year longitudinal study. *J. Am. Soc. Inf. Sci.*, **53**, 162–171.
- Lawrence,S., Coetzee,F., Glover,E., Pennock,D., Flake,G., Nielsen,F., Krovetz,R., Kruger,A. and Giles,C.L. (2001) Persistence of web references in scientific research. *IEEE Comput.*, **34**, 26–31.
- Lawrence,S. and Giles,C.L. (1999) Accessibility of information on the web. *Nature*, **400**, 107–109.
- Lawrence,S., Giles,C.L. and Bollacker,K. (1999) Digital libraries and autonomous citation indexing. *IEEE Comput.*, **32**, 67–71.
- Nowak,R. (1993) Draft genome map debuts on Internet. *Science*, **262**, 1967.
- Rumsey,M. (2002) Runaway train: problems of permanence, accessibility, and stability in the use of web sources in law review citations. *Law Libr. J.*, **94**, 27–39.
- Spinellis,D. (2003) The decay and failures of web references. *Commun. ACM*, **46**, 71–77.
- Veronin,M.A. (2002) Where are they now? A case study of health-related Web site attrition. *J. Med. Internet Res.*, **4**, E10.
- Williams,R.W. (1994) The Portable Dictionary of the Mouse Genome: a personal database for gene mapping and molecular biology. *Mamm. Genome*, **5**, 372–375.
- Wren,J.D. and Garner,H.R. (2002) Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods Inf. Med.*, **41**, 426–434.