BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 11 2008, pages 1381–1385
doi:10.1093/bioinformatics/btn127

Data and text mining

URL decay in MEDLINE—a 4-year follow-up study

Jonathan D. Wren*

Arthritis and Immunology Research Program, Oklahoma Medical Research Foundation; 825 N.E. 13th Street, Oklahoma City, OK 73104-5005, USA

Received on January 22, 2008; revised on March 11, 2008; accepted on April 6, 2008

Advance Access publication April 15, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Internet-based electronic resources, as given by Uniform Resource Locators (URLs), are being increasingly used in scientific publications but are also becoming inaccessible in a time-dependant manner, a phenomenon documented across disciplines. Initial reports brought attention to the problem, spawning methods of effectively preserving URL content while some journals adopted policies regarding URL publication and begun storing supplementary information on journal websites. Thus, a reexamination of URL growth and decay in the literature is merited to see if the problem has grown or been mitigated by any of these changes.

Results: After the 2003 study, three follow-up studies were conducted in 2004, 2005 and 2007. Unfortunately, no significant change was found in the rate of URL decay among any of the studies. However, only 5% of URLs cited more than twice have decayed versus 20% of URLs cited once or twice. The most common types of lost content were computer programs (43%), followed by scholarly content (38%) and databases (19%). Compared to URLs still available, no lost content type was significantly over- or under-represented. Searching for 30 of these websites using Google, 11 (37%) were found relocated to different URLs.

Conclusions: URL decay continues unabated, but URLs published by organizations tend to be more stable. Repeated citation of URLs suggests calculation of an electronic impact factor (eIF) would be an objective, quantitative way to measure the impact of Internet-based resources on scientific research.

Contact: Jonathan-Wren@OMRF.org

1 INTRODUCTION

The Internet age has supplemented and will probably eventually supplant the printed media for disseminating scientific information. It is not only more versatile in the types of information that can be presented (e.g. multimedia), but is cheaper to create and quicker to disseminate. However, Internet-based information, unlike the printed media, can suddenly disappear. This phenomenon is often called 'URL decay', 'link rot' or 'going 404' (so named after the '404 not found' HTTP message returned when a requested page is not available).

Prior studies have documented URL decay in several fields outside of biomedicine (Casserly and Byrd, 2003; Koehler, 1999, 2002; Lawrence and Giles, 1999; Lawrence *et al*., 2001; Rumsey, 2002; Spinellis, 2003), within biomedical subdisciplines (Carnevale and Aronsky, 2007; Dellavalle *et al*., 2003; Hester *et al*., 2004; Thireou *et al*., 2007; Thorp and Brown, 2007; Veronin, 2002) and within biomedicine as a whole (Cheung, 2001; Madani *et al*., 2006; Wren, 2004). One study of newly published papers in PubMed even found that 12.4% of URLs within the full text of articles were inaccessible at the time of publication (Madani *et al*., 2006). Subsequently, several approaches at preservation of website content published in scholarly journals have been proposed, whether as policies and procedures (Johnson *et al*., 2004; Schilling *et al*., 2004a, 2004b), or computationally such as software tools (Kahle, 1997; Eysenbach, 2006; Reich and Rosenthal, 2004; Schafer *et al*., 2001) and unique tagging/tracking measures like digital object identifiers (DOIs) (Caplan, 1998). Whether any of these proposals, changes or technologies has had a detectable effect since their inception is unknown.

A study of the reasons behind URL decay suggested that it is often outside the control of the original website creators (Wren *et al*., 2006b), suggesting that the best place for intervention would be at the time of publication. To this end, some journals are taking control of supplementary information by hosting it

*To whom correspondence should be addressed.

on their own websites. Websites with active content, however (e.g. web servers), are not suitable for such hosting for several reasons, including space, control over development and troubleshooting, platform compatibility and perhaps most importantly it is unreasonable and impractical to ask or require journals to continually upgrade and increase CPU power and memory to host an increasing array of services. Thus, some attempts at website content preservation are more likely to meet with short-term success than others and its not clear if organizational structure correlates with URL decay.

Website creation and maintenance largely remains in the hands of the scientific community, one that is relatively mobile as judged by a related phenomenon, the rate of corresponding author email decay (Wren *et al*., 2006a). The importance of monitoring URL growth and decay rates is directly proportional to our reliance upon URLs in scientific publishing. It has been suggested that URLs, like scientific papers, are not of equal importance, and the loss of some websites may never tangibly impact scientific research. But URL decay has thus far been documented as a general trend, and whether or not more 'important' websites are affected differentially is not known. Until recently, there were not enough URLs cited multiple times, at least in abstracts, to analyze differential decay trends with statistical confidence. Thus, we are now in a good position to revisit this issue for several reasons.

2 METHODS

URLs were automatically extracted from MEDLINE abstracts as described previously (Wren, 2004) and queried for availability. Three separate surveys were taken at pseudo-annual intervals in 2004, 2005 and 2007 (no survey was conducted in 2006). A total of 7462 URLs were recognized in the 2007 survey, 6154 of them being unique URLs. A total of 81 URLs were detected as redirects to different URLs and counted as still available.

3 RESULTS

Survey results are graphically summarized in Figure 1. Using linear least squares to estimate the slope of the decay rate and goodness of fit for each year, it is evident that the trends remain relatively unchanged from the original survey and consistent year-to-year (2007 slope $= -0.057$, $R^2 = 0.95$; 2005 slope $= -0.062$, $R^2 = 0.94$; 2004 slope $= -0.056$, $R^2 = 0.92$; 2003 slope $= -0.058$, $R^2 = 0.93$). Even the overall fraction of published URLs that are unavailable remains almost the same at ~20%. To see if any URL content preservation methods had been adopted, we examined the database for use of WebCite (which should include the URL subtext www.webcitation.org) and found no references aside of the two papers published describing the service (Eysenbach, 2006; Eysenbach and Trudel, 2005). Similarly, with the persistent URL (PURL) project, URLs will include the word 'purl' (e.g. purl.oclc.org, purl.net, purl.org, etc.) and we found no PURLs mentioned in abstracts.

The journals publishing the most URLs remain similar in rank order (Table 1), with *Bioinformatics* still leading the pack. Interestingly, the URL availability rate varies among journals, but it is not clear if this reflects random variation or a reason

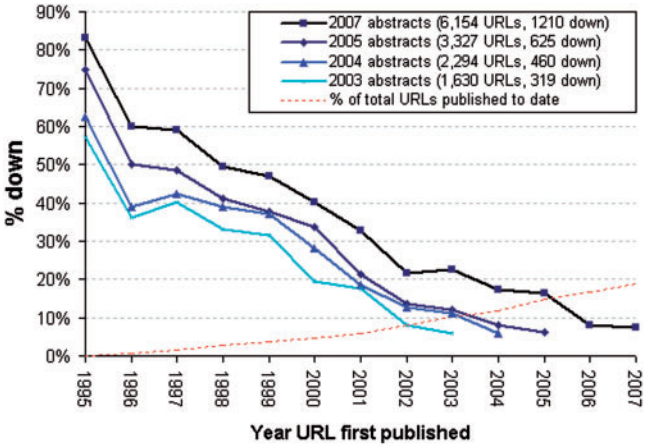

Fig. 1. Time-dependent decay of URLs published in MEDLINE abstracts. Surveys taken in 2004, 2005 and 2007 are compared to the original 2003 survey. The number of URLs published per year is displayed as a percentage of all URLs published (e.g. the 1162 unique URLs published in 2007 represent 19% of all URLs published to date).

could be attributed to it such as being a fairly new journal or having a relatively small sample size whose overall attrition rate can easily be influenced by a few URLs. *PLoS Computational Biology*, for example, has no decayed URLs, but this is likely due to the fact that it is a relatively recent journal (first abstract URL published June 2005) with few URLs. For established journals with a relatively high sample size and attrition rate, such as *Genome Research*, it is not obvious if there is a reason for the higher decay rate.

The number of times each URL was cited in the literature was heavily skewed, with a small fraction being cited numerous times. The 56 URLs cited 6 or more times represent slightly <1% of the total URLs published but garnered ~10% of all citations. URL decay rates from the 2007 survey were broken down by the number of times each URL was cited and examined separately for their availability (Table 2). It is evident that URLs cited multiple times are more likely to be still available. Of course, there is quite likely a bias towards authors only citing URLs they know are available at the time of their publication and, as expected, older URLs are more likely to have been cited multiple times compared to recently published URLs (e.g. only 3% of multiply cited URLs were published in 2007, but 19% of all URLs were published in 2007).

In terms of the type of content being lost, three samples of 50 decayed URLs were classified based upon the description published in the abstract they were identified within. Content was classified into three categories: programs (including web servers), databases and scholarly content (e.g. surveys, studies, raw data, multimedia, etc.). If a URL was described as having more than one category (e.g. database and programs) then the predominant one was assessed and used. The most common type of lost content was programs (43% ± 3%), followed by scholarly content (38% ± 4%) and then databases (19% ± 6%). Three sets of 50 URLs that were still available were also queried and then the two sample sets were compared using a two-tailed *t*-test to see if any type of content was

Table 1. The rank order of journals publishing the most URLs remains relatively consistent since the 2003 survey, with the noteworthy exception of BMC *Bioinformatics* taking the #2 spot

Journal name	No. of URLs published	Times URLs cited	No. of up	No. of down	Percentage of down	Impact
Bioinformatics	441	464	364	77	17	0.052
BMC Bioinformatics	289	301	259	30	10	0.042
Nucleic Acids Res.	282	302	234	48	17	0.071
Genome Res.	96	115	68	28	29	0.198
Proteins	93	106	74	19	20	0.140
Hum. Mutat.	76	87	62	14	18	0.145
BMC Genomics	49	51	43	6	12	0.041
Protein Sci.	48	51	37	11	23	0.063
J. Mol. Biol.	47	54	32	15	32	0.149
Proteomics	45	59	36	9	20	0.311
PNAS	38	39	28	10	26	0.026
Plant Physiol.	35	41	33	2	6	0.171
Stud. Health Tech. Inf.	33	58	27	6	18	0.758
J. Am. Coll. Cardiol.	33	34	31	2	6	0.030
Electrophoresis	31	34	20	11	35	0.097
J. Comput. Biol.	29	29	22	7	24	0.000
Methods Mol. Biol.	29	40	25	4	14	0.379
J. Bioinform. Comput Biol.	27	28	20	7	26	0.037
J. Proteome Res.	26	30	24	2	8	0.154
Cochrane Syst. Rev.	26	60	18	8	31	1.308
Oncologist	26	33	23	3	12	0.269
In Silico Biol.	26	27	23	3	12	0.038
Genome Biol.	24	25	22	2	8	0.042
Biochem. Biophs. Res. Com.	23	25	20	3	13	0.087
PLoS Comput. Biol.	23	23	23	0	0	0.000
Genetics	23	24	18	5	22	0.043
Pac. Symp. Biocomput.	22	22	13	9	41	0.000
Gene	22	23	16	6	27	0.045
Comp. Meth. Prog. Bio.	21	22	16	5	24	0.048
DNA Res.	21	35	16	5	24	0.667

Impact is calculated as the # of URLs published that were cited again later

Table 2. URLs cited multiple times are more likely available than those cited only once

Times mentioned	No. of up	No. of down	Percentage down
1	4466	1166	21
2	321	36	10
3	64	3	4
4	22	1	4
5	18	1	5
6	11	1	8
7	11	2	15
≥8	31	0	0
Total	4944	1210	20

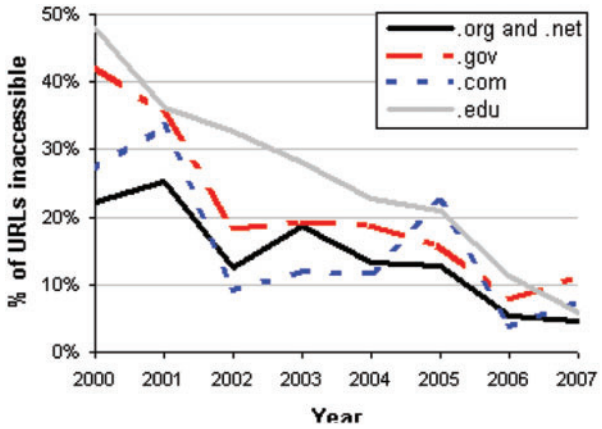

Fig. 2. Decay rates broken down by entity. As shown here, educational institutions have a relatively high rate of decay relative to .org/.net, .gov and .com addresses.

disproportionately likely to decay. None of the content types were significantly overrepresented in either dataset using a cutoff of $P<0.05$.

Furthermore, the relative stability of URLs by entity of origin (.com, .gov, .org or .net, .edu) was analyzed and the results are shown in Figure 2. Strangely, this is in contrast to the relative stability of corresponding author emails from a previous study (Wren *et al.*, 2006a) whereby emails from .org, .gov and .com addresses decayed faster than those from .edu addresses. In that study, 45, 44 and 41% of the .org, .gov and .com email addresses, respectively, were invalid at the time of the survey versus 34% of all .edu addresses. Here, we see that websites published at .edu addresses are the least stable. One possible explanation for this is that corresponding authors tend to be lab mentors, whereas creators of websites would likely be students and/or post-doctoral fellows, who would be more likely to leave.

To see if there was any variation that could be possibly attributed to funding levels or infrastructure, country top-level domains (TLDs) were examined. Table 3 shows that the country-based URL decay rates are similar, suggesting that culture and funding levels may not be significant determinants. Spain, for example, just recently received a research funding boost to address their historically low levels of funding compared to the rest of the EU (1.1% of GDP versus 1.8% for the EU average) (Editorial, 2007), yet their rate of URL decay is among the lowest.

Finally, a Google search was conducted to see if some of the decayed URLs could be found and get an idea of what fraction of decayed URLs are lost versus relocated (without automatic redirection). Thirty URLs were chosen, 10 each from the programs, databases and scholarly content categories. Based upon the MEDLINE abstract, Google searches were conducted first using resource names (if given) and then using phrases from the article that, as specifically as possible, describe the resource (including the corresponding author's last name to narrow the search if too many results were returned). Only the first search page returned was examined. A total of 11 out of 30 (37%) websites could be located using Google, 30% of the

Table 3. Overall URL decay rates for 20 countries publishing the most URLs

TLD	No. of up	No. of down	Total	Percentage down
org	1123	147	1270	12
edu	787	269	1056	25
com	464	81	545	15
gov	377	91	468	19
.uk	311	103	414	25
.de	297	78	375	21
.fr	154	50	204	25
.jp	144	33	177	19
net	125	14	139	10
.ca	97	22	119	18
.it	76	25	101	25
.au	60	25	85	29
.ch	69	13	82	16
.dk	63	15	78	19
.nl	57	11	68	16
.se	50	17	67	25
.cn	51	15	66	23
.es	51	7	58	12
.be	42	14	56	25
.in	41	10	51	20

scholarly content, 30% of the databases and 50% of the programs.

4 DISCUSSION

The sample size is relatively small, so firm conclusions cannot be drawn regarding how many websites may have been relocated, but it is nonetheless encouraging that some of the website content is not 'truly' lost even if the original URL pointer does not direct to it. Website content preservation is a burden. Researchers are used to focusing on completing their work, and preservation is often an afterthought. Journals, traditionally, have been the preservers of information through publication, but external content is not something journals have traditionally dealt with. Efforts to place the responsibility of URL preservation on either journals or authors will be difficult because journals operate independently and authors will have limitations in terms of time, funds and expertize they could dedicate to preserving dynamic content. However, it seems that if static URL content is to be preserved, the most reasonable point of intervention would have to be the part of the process where authors and journals are most focused on the paper—at the time of publication.

To aid preservation of static content, the most practical solution seems to be developing an automated method that publishers could incorporate into their online manuscript submission systems (e.g. Manuscript Central). If it were automated, then no additional time or effort would be necessary from editors, authors or reviewers. The basic changes would require simply scanning for URLs in a publication, automatically checking them for availability, creating a snapshot of URL content at the time of publication, and permitting authors to update URLs on the journal website should they change. Methods of preservation such as PURLs (Schafer *et al.*, 2001) and WebCite (Eysenbach, 2006) have been developed but are apparently not in widespread use. This is unfortunate, but likely due to both a lack of awareness for authors/publishers, and probably in part a lack of concern.

In the short run, the effects of URL decay on scientific research range from mild inconvenience to preventing study replication and/or loss of important data. In the long run, however, it is possible that URL decay could be looked at instead as an evolving system that was not really possible via the printed medium, whereby less important URLs are 'selected against' (i.e. are not duplicated or preserved), but the important ones are resurrected/recreated either by the authors or by others duplicating their effort out of necessity. A previous study on publications, for example, reported that across 21 different disciplines, an average of 26% of all papers were never cited within 17 years (Science Watch, 1999). Quite likely, the importance or utility of URLs follows a similar trend and, as observed in this study, some of the most cited and stable URLs are from organizations rather than individuals or single research groups (e.g. the most cited URL is www.clinicaltrials.gov with 77 citations). URLs with top-level domains of .org, .gov or .com represent only 39% of all URLs published, but comprise 84% of the most stable URLs (those cited eight or more times).

Generally, publications that overlap substantially or completely in content (Errami *et al.*, 2008) are considered undesirable, but if useful resources suddenly become unavailable, perhaps this is a circumstance whereby duplicating effort is not only permissible, but desirable. If one group of authors publishes a useful software tool that becomes unavailable within a few years, other authors might be discouraged from duplicating their effort if they knew such work was unpublishable. In this circumstance we must ask what is best for the scientific community. It seems reasonable to suggest that a stipulation of duplicate publication would be to ensure that the content was also preserved in some manner. It also seems reasonable to be pre-emptive and consider publishing software and databases that are preserved in some manner, provided that the authors can demonstrate that such website content is indeed valuable to the community. And this brings us to the final point to be discussed here—how would one objectively estimate the importance of any given website or software, such that an effective argument could be made for resurrecting it in a new publication?

Multiple URL citations in concert with the increasing use of Internet resources suggests an alternate form of impact that can be measured, perhaps in a similar way that traditional journal impact factors are. A journal's 'Electronic impact factor' (eIF) could be calculated as a function of the number of times URLs first published in that journal are cited after publication. This is somewhat similar to the idea of a PageRank algorithm, but is slightly different because it focuses specifically on URLs cited within published scientific research. Some eIF estimates are shown in Table 1, but are likely not very accurate since only abstracts were analyzed in this analysis. Abstracts are more likely to document the creation rather than the use of an electronic resource (usage information is likely to appear more

often in the material and methods or reference section of papers). Whereas traditional impact factors are intended to measure the influence of an idea, discovery or method put forth by a published paper, the eIF would better reflect the specific influence of an electronic resource on research activity or progress. But one caveat is that the first paper citing the URL may not be the originator of the idea and thus it is not clear how much credit the journal can or should take for introducing the resource (e.g. the first journal to mention NCBI or clinicaltrials.gov). A similar problem exists for published papers: Often, review articles are cited in reference to an idea, method or discovery rather than the original source cited within the review. Another problem is that URLs can change and thus citations to the new resource would not be credited to the original source. Nonetheless, this is hardly a trivial matter—as electronic resources are increasingly used in managing and analyzing data, it will be very useful for those charged with maintaining these resources to have objective, quantitative documentation of their relative importance to the scientific research community as a whole when it comes to seeking future funding. Especially since other metrics such as website hit counters and number of registered users are unreliable because they can easily be artificially inflated.

ACKNOWLEDGEMENTS

Funding: This work was funded in part by NSF-EPSCoR grant # EPS-0447262 and NIH/NLM grant 1 R01 LM009758-01.

Conflict of Interest: none declared.

REFERENCES

Caplan,P. (1998) DOI or Don't We? *PACS Review*, **9**, http://info.lib.uh.edu/pr/v9/n1/capl9n1.html (accessed Jan 21, 2008).

Carnevale,R.J. and Aronsky,D. (2007) The life and death of URLs in five biomedical informatics journals. *Int. J. Med. Inform*, **76**, 269–273.

Casserly,M. and Byrd,J. (2003) Web citation availability: analysis and implications for scholarship. *College and Research Libraries*, **64**, 300–317.

Cheung,J. (2001) Vanishing websites are the weakest link. *Nature*, **414**, 15.

Dellavalle,R.P. *et al.* (2003) Information science. Going, going, gone: lost Internet references. *Science*, **302**, 787–788.

Editorial (2007) Independence day? *Nature*, **446**, 347–348.

Errami,M. *et al.* (2008) Deja vu–a study of duplicate citations in Medline. *Bioinformatics*, **24**, 243–249.

Eysenbach,G. (2006) Going, going, still there: using the WebCite service to permanently archive cited Web pages. *AMIA Annu. Symp. Proc.*, **919**.

Eysenbach,G. and Trudel,M. (2005) Going, going, still there: using the WebCite service to permanently archive cited web pages. *J. Med. Internet Res.*, **7**, e60.

Hester,E.J. *et al.* (2004) Internet citations in oncology journals: a vanishing resource? *J. Natl Cancer Inst.*, **96**, 969–971.

Johnson,K.R. *et al.* (2004) Addressing internet reference loss. *Lancet*, **363**, 660–661.

Kahle,B. (1997) Preserving the Internet. *Sci. Am.*, **276**, 82–83.

Koehler,W. (1999) An analysis of web page and web site constancy and permanence. *J. Am. Soc. Inform.Sci.*, **50**, 162–180.

Koehler,W. (2002) Web page change and persistence – a four-year longitudinal study. *J. Am. Soc. Inform. Sci.*, **53**, 162–171.

Lawrence,S. *et al.* (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.

Madani,S. *et al.* (2006) Prevalence and inaccessibility of URLs in the biomedical literature. *AMIA Annu. Symp. Proc.*, **1019**.

Reich,V. and Rosenthal,D. (2004) Preserving today's scientific record for tomorrow. *Br. Med. J.*, **328**, 61–62.

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.

Schafer,K. *et al.* (2001) The PURL project. *J. Libr. Adm.*, **34**, 123.

Schilling,L.M. *et al.* (2004a) Digital information archiving policies in high-impact medical and scientific periodicals. *JAMA*, **292**, 2724–2726.

Schilling,L.M. *et al.* (2004b) Bioinformatics leads charge by publishing more Internet addresses in abstracts than any other journal. *Bioinformatics*, **20**, 2903.

ScienceWatch (1999) Citations reveal concentrated influence: some fields have it, but what does it mean? *Sci. Watch*, **10**.

Spinellis,D. (2003) The decay and failures of web references. *Commun. ACM*, **46**, 71–77.

Thireou,T. *et al.* (2007) A survey of the availability of primary bioinformatics web resources. *Genomics Proteomics Bioinformatics*, **5**, 70–76.

Thorp,A.W. and Brown,L. (2007) Accessibility of Internet references in annals of emergency medicine: is it time to require archiving? *Ann. Emerg. Med.*, **50**, 188–192, 192 e1–33.

Veronin,M.A. (2002) Where are they now? A case study of health-related Web site attrition. *J. Med. Internet Res.*, **4**, E10.

Wren,J.D. (2004) 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics*, **20**, 668–672.

Wren,J.D. *et al.* (2006a) E-mail decay rates among corresponding authors in MEDLINE. The ability to communicate with and request materials from authors is being eroded by the expiration of e-mail addresses. *EMBO Rep.*, **7**, 122–127.

Wren,J.D. *et al.* (2006b) Uniform resource locator decay in dermatology journals: author attitudes and preservation practices. *Arch. Dermatol.*, **142**, 1147–1152.