Vol. 9 No. 2, January 2004

A longitudinal study of Web pages continued: a consideration of document persistence

Wallace Koehler
Valdosta State University
Valdosta, Georgia, USA

change font

Abstract

It is well established that Web documents are ephemeral in nature. The literature now suggests that some Web objects are more ephemeral than others. Some authors describe this in terms of a Web document half-life, others use terms like 'linkrot' or persistence. It may be that certain 'classes' of Web documents are more or less likely to persist than are others. This article is based upon an evaluation of the existing literature as well as a continuing study of a set of URLs first identified in late 1996. It finds that a static collection of general Web pages tends to 'stabilize' somewhat after it has 'aged'. However 'stable' various collections may be, their instability nevertheless pose problems for various classes of users. Based on the literature, it also finds that the stability of more specialized Web document collections (legal, educational, scientific citations) vary according to specialization. This finding, in turn, may have implications both for those who employ Web citations and for those involved in Web document collection development.

Introduction

The World Wide Web still is not a library. This paper offers a limited set of findings derived from the same set of URLs that have been monitored continuously since 1996. These findings have been elaborated and reported in articles published in 1999 and 2002 (Koehler 1999a, Koehler 1999b, and Koehler 2002). The first two articles reported the behaviour of 360 Web pages and 343 Web sites for a 53 week period from December 1996 to January 1998. The third paper follows the same Web pages from December 1996 to February 2001, or 214 weeks. This paper extends the analysis to May 2003.

Despite our growing comfort and familiarity with the Web both as a medium to which to publish to as well as a conduit for and as a tool for the management of information maintained in other environments, there is still a tendency to confuse one function with the other. At the same time, initiatives like the Internet Archive that seek to 'save' Web artifacts in a large, generalized digital library or search engine caches such as those generated by Google or Yahoo compete with and sometimes create potential confusion with the 'originals.'

It is now well documented that Web pages and Web sites come and go; and that somewhat more rarely, they may come back again. This propensity for Web documents to be created then disappear has sometimes been labelled in the literature as the Web page or Web site half-life. As we shall see below, different types of material appear to have different half-lives. A half-life is that period of time required for half of a defined Web literature to disappear (or for one group of isotopes or atoms to decay into another).

A second concern has less to do with half-lives and the continuing existence of any given URL, but rather with the stability of the content of persisting Web documents. The Internet Archive, for example, can allow us to compare Web page content at different points in time, just as one might compare different editions of the same title. There are a number of structural factors that appear to contribute to the prediction of content stability (see e.g. Casserly & Byrd 2003 or Koehler 1999a). In some very fundamental ways, content stability is of far greater importance to at least certain elements of the information professions than is document stability.

This paper does not dwell on content stability issues. Its focus is Web page persistence and reports findings from a continuing longitudinal study extending more than 325 weeks from December 1996 to May 2003. It is, I believe, the longest continuous study of a single set of URLs. Most URL longevity studies are relatively short and are monitored over a period of days, weeks, or sometimes months. Furthermore, if the literature suggests interest, there is a greater interest in document stability than in content

stability—there are far more papers published on the former than the latter.

If we are to understand the dynamics of the Web as a repository of knowledge and culture, we must monitor the way in which that knowledge and culture is managed. We find that the Web in its 'native form' is a far too transitory medium and that the contributions of the Internet Archive and institutions like it are absolutely essential to the preservation process. Or we may find that the Web represents a more or less stable medium in some areas but is less so in others.

Linkrot, half-lives, and the literature

It is now well established in the literature that Web pages are ephemeral. Kitchens and Mosley (2000), for example question the utility of printed Internet guides since the Web references are far too ephemeral. Taylor and Hudson (2000) explore printed biographies of URLs and Web lists. They find variation among domain types and subject collections. Benbow (1998) found an attrition rate for Web resources of 20% and 50% over two and three year periods. Germain (2000) questions URLs as citations for scholarly literature for the same reason. McMillan (2000) argues that content analysis tools can be brought to the Web, but there are problems unique to the method because of the ephemeral nature of the target. The phenomenon has been given a variety of names: 'broken links' (Markwell & Brooks 2002; Kobayashi & Takeda 2002), 'linkrot' (Denemark 1996; Taylor and Hudson 2000), 'link rot,' (Fichter 1999; Markwell & Brooks 2003), or 'decay and failure' (Spinellis 2003). Rumsey (2002) likens the use of the Web in legal research and its ephemeral nature to a 'runaway train'. Linkrot may have become the accepted spelling as it so appears in the online Webopedia although other online sources also provide the variant spelling.

Harter and Kim (1996) are perhaps among the first of a handful researchers documenting the impact of the ephemeral nature of the Web on citations and citation systems. They examined scholarly e-journal articles published primarily in 1995 - but some as early as 1993 - and found that between the writing of the articles and their analysis (1995), a third of the URLs were no longer viable. Nelson and Allen (2002) surveyed digital library (DL) objects accessible via the Web. Placement in digital libraries, they hypothesize:

> ...is indicative of someone's desire to increase the persistence and availability of an object, we expect DL objects to survive longer, change less, and be more available than general WWW content.

Nelson and Allen report a 3% attrition rate over their sample year. Rumsey (2002) addresses the legal literature and its citations. She notes that citations to electronic media have increased dramatically since 1995 from less than 5% of all citations to nearly 30% by 2001. She reports that citation viability declines rapidly. Of citations tested in mid 2001, 39% of electronic citations dated 2001 failed, 37% dated 2000, 58% dated 1999, 66% dated 1998, and 70% dated 1997 (Rumsey 2002: 35).

Markwell and Brooks (2002) concerned themselves with the online literature employed by the scientific community and more specifically in biochemistry and molecular biology (Markwell & Brooks 2003) for education purposes. They too find significant erosion in URL viability and estimate URL half-lives for these specific science education resources of some 4.6 years.

Taylor and Hudson (2000), exploring reference issues, demonstrate wide variability across academic subject area and the top-level domains of URLs.

Study	Resource type	Resource half-life
Koehler (1999 and 2002)	Random Web pages	about 2.0 years
Nelson and Allen (2002)	Digital Library Object	about 24.5 years
Harter and Kim (1996)	Scholarly Article Citations	about 1.5 years
Rumsey (2002)	Legal Citations	about 1.4 years
Markwell and Brooks (2002)	Biological Science Education Resources	about 4.6 years
Spinellis (2003)	Computer Science Citations	about 4.0 years (p. 74)

Table 1: Resource half-lives by resource type

The Nelson and Allen (2002) study offers perhaps a useful baseline against which to measure URL viabilities. They chose to measure object attrition in what they assumed a priori to be a stable environment. Their findings support their assumptions. It may be presumptuous for me to argue that my work could represent the 'right boundary' of longevity behaviour. The Harter and Kim (1996) study suggests early URL 'instability'. However, the study methodology as reported is insufficiently precise to allow precise calculations of half-lives. Our data suggest one of two possibilities. Either the online scientific literature has grown more stable since sampled in 1995 by Harter and Kim, or the biological literature used for teaching and the computer science literature as reported by Markwell and Brooks (2002) and Spinellis (2003) are exceptions to the general rule. We suspect the former, that the online scientific literature has become more 'stable'. We would also caution that an increase in the half-life of that online literature from about 1.5 years to about four years is no major gain.

Methodology

The original data set was collected between December 1996 and May 2003 to map Web page change over time. Data are collected weekly and include page size (in kilobytes) and link changes for Web pages. A number of attributes were examined to assess the growth, change, and death of those Web pages. From December 1996 until February 2001, FlashSite 1.01, a software product of Incontext was employed throughout the study for data capture. Beginning in January 2002, KyoSoft's Link Checker Pro was employed for data capture. For a variety of reasons, including systems failures, the inability of more recent versions of Windows software to support Flashsite, and other factors, data were not collected for the majority of 2001.

A sample of 361 URLs was collected in the last two weeks of December 1996 using the now defunct WebCrawler random URL generator. In late 1996, the Web contained an estimated 100 to 600 million pages (Koehler1999a). The sample was stratified to correspond to the reported distribution of Web pages by top-level domain at the time.

This method relied on a single search engine's index to generate the sample and is similar to that similar to that reported by Bharat & Broder (1998) in that the sample harvest was based on a single search engine. Other approaches to develop random samples includes a pool of URLs (Bray 1996), randomly generated IP numbers (Lawrence and Giles 1998, 1999) and a random walk of the Web using multiple search engines (Henzinger *et al.* 2000).

The 361 Web page sample was as random a representation of the Web as a whole as the WebCrawler index was representative of the Web as a whole in December 1996. The sample is not representative of the Web as a whole in 2003 nor is it intended to be. It may reflect the status of Web pages that existed in late 1996 and continue to do so.

Data were harvested employing two different commercial software packages. The use of FlashSite was discontinued when the program was no longer supported and updated by its producer. The product does not function or function well in Windows environments above Windows95. KyoSoft's Link Checker Pro has been adopted as the link check and data harvest software.

Link Checker Pro provides a text or spreadsheet cumulative and individual report of the status of hypertext links embedded in a Website. In order to use Link Checker Pro, all URLs randomly collected in December 1996 for the original research project were encoded as part of a Web page and placed on the Web. That Webpage is then scanned weekly. The software provides an Excel spreadsheet with aggregate data as well as for each individual URL.

Link Checker Pro provides the similar data to FlashSite as well as the date the site was last modified. In addition, Link Checker Pro reports the status of the missing URL: whether the request was forbidden or unauthorized (401 or 403 errors), the much more common a file not found error (the 404 error), and the no domain name server error (no DNS). The first two sets of errors indicate a restructuring or removal of content from the file structure. The latter error means either the disappearance or down status of a domain or Website.

For a number of technical reasons, there are some substantial temporal gaps. These include the transition from one operating system to another and therefore one analytic program to another. It also represents the rigours of a move and research time lost in that transition. And finally, I suffered a systems crash. The data are nevertheless sufficiently continuous and contiguous to permit analysis.

An assessment of research findings

It is suggested in the literature discussed above that the stability of Web site may be a function of URL form, domains, or of other factors. We suggest here that Web page persistence is somewhat more complex. Table 2 shows the distribution by level of the responding Web page collection when first collected in December 1996 and the remaining collection in mid-February 2001 and in May 2003.

In December 1996 about half the sample consisted of navigation pages and half of content, after four years navigation pages represented more than 60% of the remaining sample. McDonnell *et al.* (2000) have defined navigation pages as those pages that serve to guide the user through a Web site to the information the site was created to provide. They define content pages as those pages providing that information. They report that navigation pages are most often found at the server-level domain or one level removed (www.aaa.tld and www.aaa.tld/xxx).

	Dec 1996	Feb 2001	May 2003
Navigation	50.4	61.3	72.2
Content	49.6	38.7	27.8
Total N	361	124	122

Table 2: Extant sample—distribution by level in percentages
chi-square=18.95, df=2, p<=.001

While we see very little change in the size of the remaining sample between February 2001 and May 2003, there has been a continued change in the distribution between navigation and content pages. As predicted by Koehler (2002), navigation pages have shown greater 'resilience' than have content pages over time. As already indicated, in December 1996, about half the sample consisted of navigation and content pages. However, in May 2003, two-thirds of the sample had gone. Of these, three-quarters of the sample that remained constituted navigation pages. We know from experience and *ad hoc* observation that not all content pages are 'gone'; they are, rather, re-addressed on the file structure. They are on occasion, 'intermittent', that is to say, they may, for a variety of reasons, 'come and go' (Koehler 2002). From a statistical perspective, the six year distribution (1996-2003) represents a significant difference in the distribution of content and navigation pages. Although the trend continues in the 2001-2003 period, these data do not represent a statistically significant difference (chi-square=3.3, df=1, p<=.10).

It is not unreasonable to ask, from a bibliographical perspective, whether a Web document, once moved from one address to another remains the same document or whether it is metamorphosed into something different. If at the same time that the document was moved its content was changed, we would probably have little problem defining it as *different*. How shall we treat it when content is either not changed or only very slightly changed? Is each modification a new edition?

Top level domain	Dec 1996	Feb 2001	May 2003
com	32.1	37.9	26.4
edu	29.4	28.2	24.0
gov	5.0	4.0	9.1
ccTLD (uncl)	16.3	9.8	9.9
mil	3.3	4.8	1.7
net	10.5	10.5	24.0
org	3.3	4.8	5.0
Total N	361	124	122

Table 3: Extant sample—distribution by implied TLD in percentages
chi-square=17.3, df=12, p<=.20

While there has been a change in the distribution of navigation and content pages, a similar conclusion cannot be reached for Web pages classed by top-level domain (TLD). Table 2 presents data for Web pages according to 'implied' TLD. An 'implied TLD' is a country code TLD (ccTLD) 'reclassed' as a generic TLD (gTLD) for purposes of analysis based on the second level domain (2LD)—for example co.jp as commercial, ac.up as educational, or.cr as organizational, and gob.mx as governmental. Those that cannot be 'reclassed', are not. The ccTLDs are those that carry a country or regional abbreviation as the TLD. The gTLDs are those that carry the seven original and seven newer 'areas of competence or application' abbreviations - .com for commercial, .edu for educational, .museum for museums and so on.

Perhaps there been shifts in the sample for some gTLDs, the net domain, for example. On balance (and statistically), it would appear that the distribution in the sample by gTLD and implied gTLD has remained relatively unchanged. This implies that the distribution of Web documents by TLD that existed at one point in time does not change over time for that same cohort of documents, the distribution of documents that existed in 1996 and continue to exist in 2003, are similar. That does not mean, however, that the distribution of documents that existed in 1996 and that existed in May 2003 are the same. It has been well established that the size of the Web, the source of documents, and the number of documents have increased dramatically, perhaps geometrically over that period.

These data suggest that as a static URL collection ages, it may become in time relatively stable. Figure 1, borrowed from Koehler (2002) shows the rate at which the URL set eroded between December 1996 and February 2001.

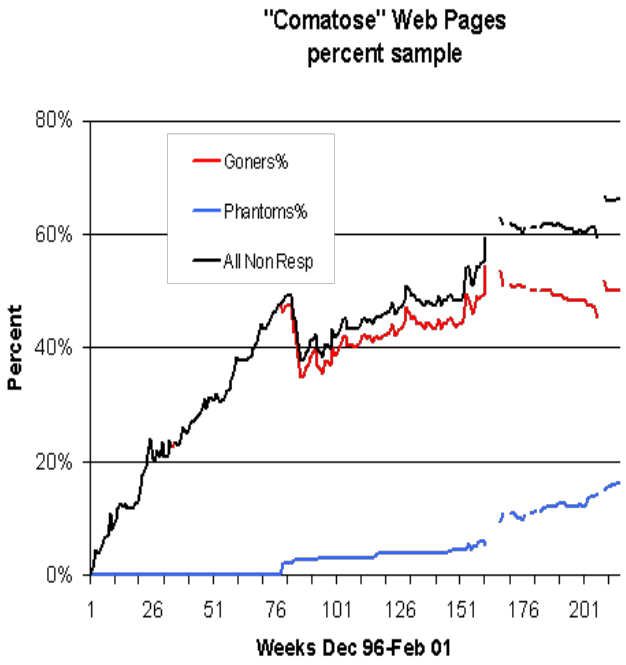

Figure 1: 'Comotose' Web pages

In December 1996, 100% of the sample was "present," but by February 2001, it had eroded to 34.4 percent of its original size. By May 2003, it had been reduced to only 33.8 percent of the original sample size, in, however, as Tables 1 and 2 suggest, a somewhat restructured configuration. This restructuring can result, as Koehler (2002) reports, from the periodic resurrection of Websites and Web pages, sometimes after protracted periods of time.

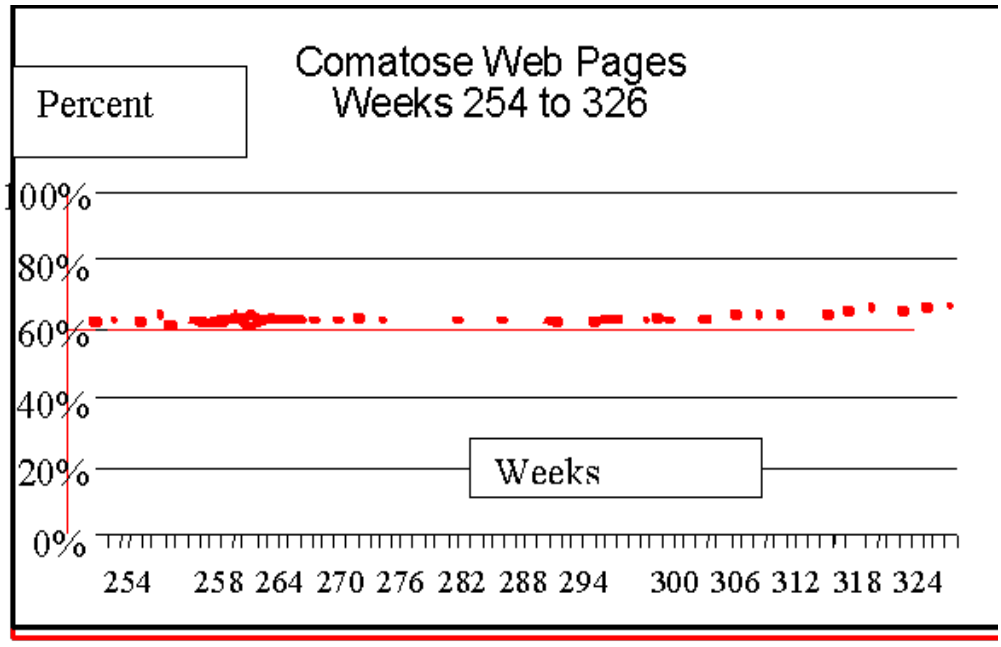

Figure 2: 'Comotose' Web pages, weeks 254 to 326

As is shown in Figure 3, 'missing', or 404 error continues to be the primary cause for Web page attrition in the sample. The legend in Figure 3 lists four 'causes' for URL failure: missing, restricted access, bad, and unknown. Missing signified the 'file not found or 404 error'; restricted access the '401 unauthorized and 403 forbidden errors'; bad, the no domain name server fault; and other, all others. In recent months there appears to be a slight trend for a slight increase in DNS errors which may not be too surprising in a sample of Web sites that is over six years old.

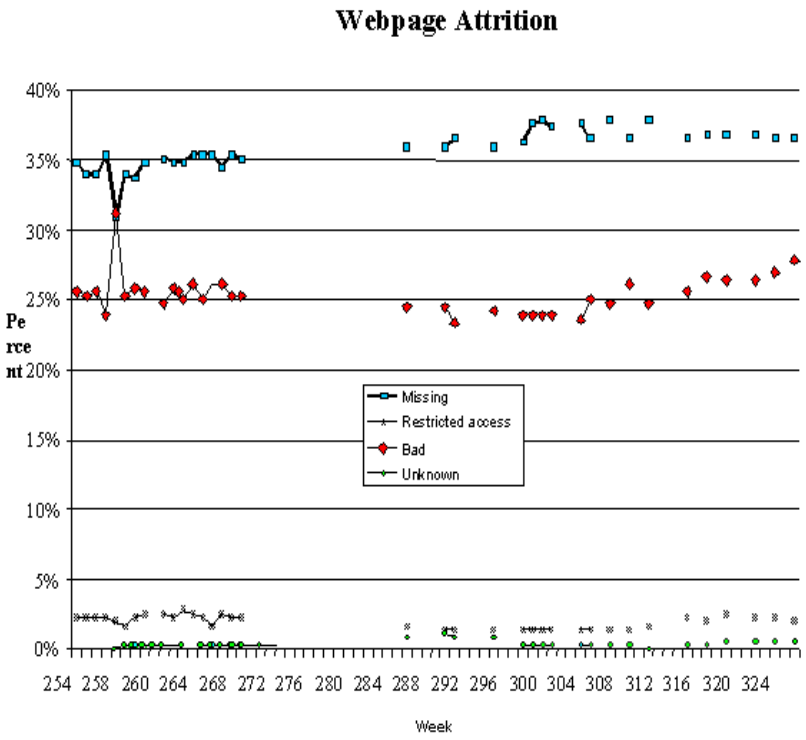

Figure 3: Web page attrition

Conclusions

Based on a relatively small number of studies, we can conclude that the Web documents are not particularly stable media for the publication of long-term information and the maintenance of individual objects or items. That said, we must distinguish between material published to the Web and material for which the Web serves as a conduit for access. For example, Nelson and Allen (2002) indicate long half-lives for online databases, where others suggest far shorter half-lives for Web published resources.

There are two interesting trends that emerge from this analysis. First, once a collection has sufficiently aged, it may stabilize in the sense at least that its URLs may become more durable in time. We have shown, for example, that Koehler's collection of randomly collected URLs remained in a fairly 'steady-state' for two years after it lost approximately two-thirds of its population over a four year period. From a collection development perspective, this period of stability has been but of short duration. Additional monitoring is needed to establish resource lifetimes.

Second, it is equally interesting to find that that the half-lives of Web resources in different disciplines, domains, and fields differ. First, not only are legal, scholarly, and educational electronic citations reported to have limited lifecycles not dissimilar to Web resources in general, but there is also variability among the disciplines.

The discussion thus far has focused on URL sets monitored over specific periods of time to determine to probability of linkrot and other variations of decline in citation efficacy. There have been a number of initiatives to try to address these matters either through strategic design or by application of some form link checking technology. The Scholarly Societies Project at the University of Waterloo is among the first to have adopted a collection strategy based on the format of the URL. They have observed that canonical URLs - those that took the form ***www.orgname.org*** and ***www.orgname.org.cc*** are more likely to persist than other non-canonical forms. Thus, as a strategic collection decision, one might preclude Web documents that do not meet some 'persistence test' such as the canonical URL in an effort to build more stable collections.

Second, one might elect to link to the index page, or what we have labelled navigation pages here. It is not uncommon for any given domain to undergo wholesale restructuring. In recent months, for example, the American Library Association, owner and manager of www.ala.org completely changed the file tree with "catastrophic consequences" for many of us maintaining links to **ala.org**. Similarly, when management of the International Federation of Library Associations and Institutions (IFLA) Web site migrated from Canada to France, the IFLA Committee on Free Access to Information and Freedom of Expression (FAIFE) Web site was subsumed under the IFLA umbrella and converted from **faife.dk** to **ifla.org/faife/**. This change had particular consequences for the stability of one monitored collection I maintain.

Links can be checked and corrected. I have adopted KyoSoft's Link Checker Pro to check the links periodically at my little web library and information professional association Web site and to prepare a report of its status. On a much larger scale, OCLC Connexion - subsuming the functions of the Cooperative Online Resource catalogue or CORC - provides a cataloguing system for electronic resources, including Web pages. It includes a periodic check for resource stability and catalogue maintenance.

References

- Bharat, K. & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. In: *Proceedings of the 7th International World Wide Web Conference*. (pp. 379-388). Amsterdam: Elsevier Science.
- Benbow, S.M.P. (1998). File Not Found: the problem of changing URLs for the World Wide Web. *Internet Research: Network Applications and Policy* **8**(3), 247-250.
- Bray, T. (1996). Measuring the Web. *World Wide Web Journal*, **1**(3). Retrieved 24 December, 2003 from http://www5conf.inria.fr/fich_html/papers/P9/Overview.html [Also found at http://www.w3j.com/3/s3.bray.html, but without images.]
- Casserly, M. & Byrd, J. (2003). Web citation availability: analysis and implications for scholarship. *College and Research Libraries*, **64**(4), 300-317.
- Denemark, H. (1996) The death of law reviews has been predicted: What might be lost when the last law review shuts down? *Seton Hall Law Review*, **27**(1), 1-32.
- Douglis, F, Feldmann, A., & Krishnamurthy, B. (1997) Rate of change and other metrics: a live study of the World Wide Web. In *USENIX Symposium on Internet Technologies and Systems, December 8-11, Monterey, CA.* (pp. 147-158). Berkeley, CA: USENIX. Retrieved 24 December, 2003 from http://www.usenix.org/publications/library/proceedings/usits97/full_papers/douglis_rate/douglis_rate_html/douglis_rate.html
- Fichter, F. (1999) Do I look like a maid? Strategies for preventing link rot. *Online*, **23**(5), 77-79.
- Germain, C.A. (2000). URLs: uniform resource locators or unreliable reliable resource locators? *College and Research Libraries* **61**(4), 359-365.
- Harter, S and Kim, H. (1996) Electronic journals and scholarly communication: a citation and reference study. *Information Research* **2** (1) paper 9. Retrieved May 1, 2003 from http://informationr.net/ir/2-1/paper9a.html
- Henzinger, M., Heydon, A., Mitzenmacher, M. & Najork, M. (2000). On near-uniform URL sampling. In, Proceedings of the Ninth International World Wide Web Conference (WWW9), May 15-19, 2000, Amsterdam. New York, NY: Elsevier Science B.V. Retrieved July 7, 2001 from http://www9.org/w9cdrom/88/88.html#BB
- Kitchens, J.D. & Mosley, P.A. (2000). Error 404: or, WWhat is the shelf-life of printed Internet guides? *Library Collections, Acquisitions & Technical Services*, **24**(4), 467-478.
- Kobayashi, M & Takeda, K. (2002) Information retrieval on the Web. **ACM Computing Surveys**, **32**(2), 144-173.
- Koehler W. (1999a) An Analysis of Web page and Web site constancy and permanence. *Journal of the American Society for Information Science* **50**,(2), 162-180.
- Koehler, W. (1999b) Digital libraries and World Wide Web sites and page persistence. *Information Research*, **4**(4) Retrieved 24 December, 2003 from http://informationr.net/ir/4-4/paper60.html
- Koehler, W. (2002). Web page change and persistence - a four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, **53**(2), 162-171.
- Lawrence, S. & Giles, C.L. (1998). Searching the World Wide Web. *Science*, **280**(536), 98.
- Lawrence, S. & Giles, C.L. (1999). Accessibility of information on the Web. *Nature* **400**(8 July), 107-109.
- McDonnell, J., Koehler, W. & Carroll, B. (2000). Cataloging challenges in an Area Studies Virtual Library Catalog (ASVLC): results of a case study. *Journal of Internet Cataloging* **2**(2), 15-42.
- Markwell, J. & Brooks, D.W. (2002) Broken links: the ephemeral nature of educational WWW hyperlinks. *Journal of Science Education and Technology*, **11**(2), 105-108.
- Markwell, J. & Brooks, D.W. (2003) 'Link rot' limits the usefulness of Web-based educational materials in biochemistry and molecular biology. *Biochemistry and Molecular Biology Education* **31**(1), 69-72.
- Nelson, M & Allen, B. (2002). Object persistence and availability in digital libraries. *D-Lib Magazine* **8**(1). Retrieved May 1, 2003 from http://www.dlib.org/dlib/january02/nelson/01nelson.html
- Rumsey, M. (2002). Runaway train: Problems of permanence, accessibility, and stability in the use of Web sources in law review citations. *Law Library Journal*, **94**(1), 27-39
- Spinellis, D. (2003). The decay and failures of Web references. *Communications of the ACM* , **46**(1), 71-77.
- Taylor, M.K. & Hudson, D. (2000). "Linkrot" and the usefulness of Web site bibliographies. *Reference & User Services Quarterly*, **39**(3), 273-276.

Acknowledgements: The author wishes to acknowledge and thank the two annonymous referees. Their very useful comments and criticisms have been instrumental in improving this paper.

Scholar Search

Find other papers on this subject.

Google Search

How to cite this paper:

Koehler, W. (2004) A longitudinal study of Web pages continued: a report after six years. *Information Research*, **9**(2) paper 174 [Available at http://InformationR.net/ir/9-2/paper174.html]

Check for citations, using Google Scholar

© Wallace Koehler, 2003. Last updated: 24 December, 2003

Contents 12582 **Web Counter** **Home**
