URL Decay at Year 20: A Research Note

By: Fatih Oguz and Wallace Koehler

Oguz, F., & Koehler, W. (2016). URL decay at year 20: A research note. Journal of the Association for Information Science and Technology, 67(2), 477-479.

Made available courtesy of Wiley: http://dx.doi.org/10.1002/asi.23561

Abstract:

All text is ephemeral. Some texts are more ephemeral than others. The web has proved to be among the most ephemeral and changing of information vehicles. The research note revisits Koehler's original data set after about 20 years since it was first collected. By late 2013, the number of URLs responding to a query had fallen to 1.6% of the original sample. A query of the 6 remaining URLs in February 2015 showed only 2 still responding.

Keywords: URL | entropy (information)

Article:

Introduction

It is well established in the professional literatures that web documents undergo significant "decay" and change. Most papers address web document inaccessibility (e.g., Hennessey & Ge, 2013; Rumsey, 2002; Russell & Kane, 2008; Strader & Hamill, 2007; Taylor & Hudson, 2000; Tyler & McNeil, 2003; Wagner, Gebremichael, Taylor, & Soltys, 2009). Relatively few have been concerned with content changes to those documents (e.g., Bar-Ilan & Peritz, 2009; Oguz & Koehler, 2011; Payne & Thelwall, 2007, 2008a, 2008b). This research note reexamines the URL data harvested as research for a master's thesis and a number of articles published by one of the two authors of this note. The data were originally identified and collected in 1995 (Koehler, 1997). The dataset was re-examined on several occasions (Koehler, 1999a, 1999b, 1999c, 2002) and findings were reported most recently in 2004 (Koehler, 2004).

Koehler (1999a) showed that web documents underwent two forms of change. The first form, and by far the more explored, is the erosion, decay, or disappearance of links. The second is labeled "omega" or "ω" as a measure of change in content. This note reports both the 2015 status of the dataset both in terms of URL responses to queries and to content change or "ω."

Webpage Decay

The data analyzed beginning in 1995 were derived using a random URL generator (Koehler, 1997). The sample generation methodology is described in the 1999 *JASIS* article. The sample purports to represent the web at that time in a limited way. It captured a sample not inconsistent with the published distribution of then existing generic (g) and country code (cc) top-level

domains (TLD). Since 1995, the number of gTLDs has increased from five to over 400. In 1995, all 360 URLs in the selected sample responded to a query. By late 2013, the number of URLs responding to a query had fallen to 6, or 1.6 percent of the original sample. A query of the 6 remaining URLs in February 2015 showed only 2 still responding. This research note helps establish, therefore, a limit to the viability of a generalized study of links.

To determine web document viability, each original URL of the dataset was pinged. The Internet Assigned Numbers Authority (IANA) codes were collected. The IANA (2014) provides a list of codes that indicate the status of the URL. A 200-code indicates a viable link. All other code responses suggest a decayed link. The most common of these are the 404-code, file not found. All responses were rechecked including the 200- and 404-codes. In this collection, of the 106 URLs returning a 200-code, all but 6 were found to be invalid in 2013. Of these 6, only two remained in 2015. A number of URLs responded with a 301-code, moved permanently. These too on examination, were found to be defunct.

Most other studies of link decay address specific literatures or disciplines. For example, Rhodes (2010) reports link decay over three-year period for online legal and policy materials. Markwell and Brooks (2003) demonstrate link decay in the biology literature. Russell and Kane (2008) found a decline of link reliability over time for history materials. Wren (2004) found significant link decay in the medical literature. Goh and Ng (2007) showed meaningful decay in web citations in information science journals. This research note helps establish, therefore, a limit to the viability of a generalized study of links. Additional work can contribute to the definition of the limits of specific datasets.

A number of studies have reported decay rates or half-lives for web document collections. For this collection, Koehler (2004) reported a half-life of about 2 years. More specialized collection half-lives vary between about 1.4 years (Rumsey, 2002) for legal citations to 24.5 years for digital library objects (Nelson & Allen, 2002). Given reported web document half-lives it is none too surprising that the number of "extant" web documents in 2015 is two. These data and the findings of others continue to reinforce the conclusion that the longevity of web document collections or citations to web based materials is problematic at best.

Change

Web document changes are well recognized but perhaps under reported. Change manifests itself in major as well as subtle ways. We have all witnessed anecdotal change. During the monitoring of these data through 2004, Koehler (2004) found that one site underwent frequent change. That change was limited to the resizing of a graphic on the page. In another instance, a hypertext link imbedded in a teaching document at one time resolved to a discussion of web document cataloging. Over time it had morphed into a pornographic site.

A number of algorithms have been developed to assess "content change." For the 1999 article (Koehler, 1999a), "byteweight" was collected and compared to prior URL assessments. Bytes per web page were reported by the software employed to collect the data for the original article. These data were collected weekly. A change in the number of bytes reported was interpreted as a change in content. Any amount of change was considered to have a value of one (1). If the byteweight stayed the same, no change value of zero (0) assigned. Frequent and periodic

byteweight changes over time (t) were noted and reported.

$$\bar{\omega} = \frac{\sum \omega}{t}$$

A number of other approaches to content change have been developed. IANA status code 304 "Not Modified" can be used to check whether a page has been updated since last time it was scanned (Fielding et al., 1999). However the support for this status code is limited especially with the increased use of content management systems, blogging software, and other database-driven applications to develop web sites. The use of checksum is another approach to detect change (Cho & Garcia-Molina, 2000). Checksum is a hash value generated based on data content of a file and is commonly used for error detection. Use of checksums is an improvement over file size as the latter would fail to detect change if a word or character is replaced with another one at identical length.

Oguz and Koehler (2011) proposed the use of the cosine similarity (Baeza-Yates & Ribeiro-Neto, 1999) to measure change. However, change in Web document content may focus on a particular type of content or topic. Payne and Thelwall (2007, 2008a, 2008b) looked at a specific type of Web document content, hyperlinks, and how they changed over time to identify general trends in academic Web sites. Bar-Ilan and Peritz (2009) found that while the number of Web pages containing a particular term increased, a large of number of Web pages were no longer exist from earlier years in their study of evolution of a topic on the Web.

For this research note, changes to web sites were assessed by inspection. Copies of the original web pages were downloaded, printed, and saved in 1995. When the extant remaining 6 webpages were examined in 2015, all were found to have undergone substantial change by visual inspection. Therefore the entire dataset had undergone change, either through decay or modification.

Conclusion

From the analyses of the original dataset of 360 URLs as well as the work of others, web documents are inherently unstable. This instability occurs in two important ways. Web sites demonstratively cease to exist over time. Web sites are almost certain to undergo change as well. Information scientists are therefore advised not to rely on the Web for even short term archiving or reference.

References

Baeza-Yates, R. A., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Bar-Ilan, J., & Peritz, B. C. (2009). The lifespan of "informetrics" on the Web: An eight year study (1998–2006). *Scientometrics*, *79*(1), 7–25. http://doi.org/10.1007/s11192-009-0401-7

Cho, J., & Garcia-Molina, H. (2000). The Evolution of the Web and Implications for an Incremental Crawler. In *Proceedings of the 26th International Conference on Very Large Data Bases* (pp. 200–209). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=645926.671679

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., & Berners-Lee, T. (1999). *Hypertext transfer protocol—HTTP/1.1* (Internet RFC). Retrieved from http://www.hjp.at/doc/rfc/rfc2616.html

Goh, D. H.-L., & Ng, P. K. (2007). Link Decay in Leading Information Science Journals. *Journal of the American Society for Information Science and Technology*, *58*(1), 15–24. http://doi.org/10.1002/asi.20513

Hennessey, J., & Ge, S. X. (2013). A cross disciplinary study of link decay and the effectiveness of mitigation techniques. *BMC Bioinformatics*, *14*(Suppl 14), S5. http://doi.org/10.1186/1471-2105-14-S14-S5

Internet Assigned Numbers Authority (IANA) Hypertext Transfer Protocol (HTTP) Status Code Registry. (2014). Retrieved February 2, 2015, from http://www.iana.org/assignments/http-status-codes/http-status-codes.xhtml

Koehler, W. C. (1997). *Web Site And Web Page Persistence And Change: A Longitudinal Study* (MS Thesis). University of Tennessee, Knoxville. Retrieved from http://www.sis.utk.edu/thesis/wallace-conrad-koehler-jr

Koehler, W. C. (1999a). An Analysis of Web Page and Web Site Constancy and Permanence. *Journal of the American Society for Information Science*, *50*(2), 162–180. http://doi.org/10.1002/(SICI)1097-4571(1999)50:2<162::AID-ASI7>3.0.CO;2-B

Koehler, W. C. (1999b). Classifying Web sites and Web pages the use of metrics and URL characteristics as markers. *Journal of Librarianship and Information Science*, *31*(1), 21–31. http://doi.org/10.1177/096100069903100103

Koehler, W. C. (1999c). Digital libraries and World Wide Web sites and page persistence. *Information Research*, *4*(4). Retrieved from http://www.informationr.net/ir/4-4/paper60.html

Koehler, W. C. (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. http://doi.org/10.1002/asi.10018

Koehler, W. C. (2004). A Longitudinal Study of Web Pages Continued: A Consideration of Document Persistence. *Information Research*, *9*(2). Retrieved from http://www.informationr.net/ir/9-2/paper174.html

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. http://doi.org/10.1002/bmb.2003.494031010165

Nelson, M. L., & Allen, B. D. (2002). Object Persistence and Availability in Digital Libraries. *D-Lib Magazine*, *8*(1). http://doi.org/10.1045/january2002-nelson

Oguz, F., & Koehler, W. C. (2011). Document Constancy and Persistence: A Study of Web Pages in Library and Information Science Domain. In *Proceedings of the 74rd ASIS&T Annual Meeting* (Vol. 48, pp. 1–9). New Orleans, LA. Retrieved from http://doi.org/10.1002/meet.2011.14504801059

Payne, N., & Thelwall, M. (2007). A longitudinal study of academic webs: Growth and stabilisation. *Scientometrics*, *71*(3), 523–539. http://doi.org/10.1007/s11192-007-1695-y

Payne, N., & Thelwall, M. (2008a). Do academic link types change over time? *Journal of Documentation*, *64*(5), 707–720.

Payne, N., & Thelwall, M. (2008b). Longitudinal trends in academic web links. *Journal of Information Science*, *34*(1), 3–14. http://doi.org/10.1177/0165551507079417

Rhodes, S. (2010). Breaking Down Link Rot: The Chesapeake Project Legal Information Archive's Examination of URL Stability. *Law Library Journal*, *102*(4), 581–597.

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.

Russell, E., & Kane, J. (2008). The Missing Link : Assessing the Reliability of Internet Citations in History Journals. *Technology and Culture*, *49*(2), 420–429.

Strader, C. R., & Hamill, F. D. (2007). Rotten But Not Forgotten. *The Serials Librarian*, *53*(1-2), 163–177. http://doi.org/10.1300/J123v53n01_13

Taylor, M. K., & Hudson, D. (2000). "Linkrot" and the Usefulness of Web Site Bibliographies. *Reference & User Services Quarterly*, *39*(3), 273–277.

Tyler, D. C., & McNeil, B. (2003). Librarians and Link Rot: A Comparative Analysis with Some Methodological Considerations. *Portal: Libraries and the Academy*, *3*(4), 615–632. http://doi.org/10.1353/pla.2003.0098

Wagner, C., Gebremichael, M. D., Taylor, M. K., & Soltys, M. J. (2009). Disappearing act: decay of uniform resource locators in health care management journals. *Journal of the Medical Library Association : JMLA*, *97*(2), 122–130. http://doi.org/10.3163/1536-5050.97.2.009

Wren, J. D. (2004). 404 Not Found: The Stability and Persistence of URLs Published in MEDLINE. *Bioinformatics*, *20*(5), 668–672. http://doi.org/10.1093/bioinformatics/btg465