Georgetown University Law Center
Scholarship @ GEORGETOWN LAW

2010

Breaking Down Link Rot: The Chesapeake Project Legal Information Archive's Examination of URL Stability

Sarah Rhodes
Georgetown University Law Center, sjr36@law.georgetown.edu

102 Law Libr. J. 581-597 (2010)

Breaking Down Link Rot: The Chesapeake Project Legal Information Archive's Examination of URL Stability[*]

Sarah Rhodes[**]

Ms. Rhodes explores URL stability, measured by the prevalence of link rot over a three-year period, among the original URLs for law- and policy-related materials published to the web and archived though the Chesapeake Project, a collaborative digital preservation initiative under way in the law library community. The results demonstrate a significant increase in link rot over time in materials originally published to seemingly stable organization, government, and state web sites.

Introduction

¶1 In the context of web archiving and digital preservation, one often hears that the average life span of a web page is forty-four days.[1] This statistic has been repeated among those in the digital preservation community for years, but it never seems to be accompanied by a citation. In a 2002 article by Peter Lyman, a footnote briefly explains why the source of this figure is so elusive: "These data sources were originally published on the Web, but are no longer available, illustrating the problem of Web archiving."[2] Ironically, the very source of a statistic often used to support the cause of web preservation has itself become a victim of "link rot."

¶2 Link rot refers to the loss or removal of content at a particular Uniform Resource Locator (URL) over time.[3] In other words, when an attempt is made to

---

1. *See, e.g.*, Jim Barksdale & Francine Berman, *Saving Our Digital Heritage*, WASH. POST, May 16, 2007, at A15 (giving the average life span as forty-four to seventy-five days); Gail Fineberg, *Capturing the Web: Staff Briefed on National Digital Preservation Plan*, LIBR. CONG. INFO. BULL. (Apr. 2003), *available at* http://www.loc.gov/loc/lcib/0304/digital.html; Brewster Kahle, *Preserving the Internet*, SCI. AM., Mar. 1997, at 82, 83.

2. Peter Lyman, *Archiving the World Wide Web*, *in* COUNCIL ON LIBRARY & INFO. RES., BUILDING A NATIONAL STRATEGY FOR DIGITAL PRESERVATION 38, 38 n.1 (2002), *available at* http://www.clir.org/pubs/reports/pub106/pub106.pdf.

3. *See* Wallace Koehler, *A Longitudinal Study of Web Pages Continued: A Consideration of Document Persistence*, INFO. RES., Jan. 2004, http://informationr.net/ir/9-2/paper174.html (briefly discussing link rot and other terms used to describe the disappearance of content from URLs).

open a documented link, either different or irrelevant information has replaced the expected content, or else the link is found to be broken, typically expressed by a 404 or "not found" error message. This is not an uncommon occurrence. Web-based materials often disappear as URLs change and web sites are changed, updated, or deleted.

¶3 Despite URL instability, the web remains an immediate and inexpensive publishing medium with a broad audience, and the producers of important resources, including law- and policy-related materials, have taken full advantage of the web for the dissemination of their content. As law librarians are well aware, resources ranging from government documents to sources cited in law review articles and court decisions are increasingly "born digital" and distributed only online. The prevalence of resource loss and link rot presents a challenge, especially for those who are charged with collecting, preserving, and providing patrons with access to this information.

¶4 In 2007, the Georgetown Law Library and the state law libraries of Maryland and Virginia formed the Chesapeake Project Legal Information Archive to begin preserving these important web-published law- and policy-related materials.[4] In the three years since the archive was launched, this law library collaborative has built a collection comprising more than 2300 titles and 5700 digital items, all of which were originally posted to the web.[5]

¶5 In an effort to quantify both the progress and relevance of the Chesapeake Project, an evaluation of the project's efforts has been conducted on a regular basis. Among the parameters used to evaluate the project, project participants have measured the prevalence of link rot among the original URLs for titles preserved in the archive, an analysis designed to demonstrate both the need for the project within the law library community and the instability of open access, web-published law- and policy-related materials.[6]

¶6 This article analyzes these evaluations in order to answer the following questions:

- What percentage of original URLs are impacted by link rot within two years of being harvested and archived, based on a sample of titles harvested through the Chesapeake Project in 2007–2008?
- What percentage of original URLs representing the entire digital archive collection are currently impacted by link rot, based on a sample of all titles harvested through the Chesapeake Project in 2007–2010, compared to samples from previous years?
- What are the top-level domains (such as .gov, .com, .org, or .us) of original URLs that are most impacted by link rot?

---

4. Legal Info. Archive, The Chesapeake Project, http://www.legalinfoarchive.org (last visited July 18, 2010) [hereinafter Chesapeake Project].

5. The number of titles was gathered from statistics reported by each participating library, and the number of digital items came from the project's CONTENTdm Administration module.

6. See Legal Info. Archive, The Chesapeake Project, Project Reports & Documentation, http://www.legalinfoarchive.org/custompages/documentation.php for previous evaluations and discussions of project evaluation parameters.

- What are the file format types (such as PDFs, X/HTML web pages, or Microsoft Word documents) of original URLs that are most impacted by link rot?

Background

The Chesapeake Project

¶7 The origin of the Chesapeake Project is linked to the establishment of the Legal Information Preservation Alliance. In 2003, a group of Georgetown law librarians, under the leadership of Robert Oakley, then the director of the Georgetown Law Library, organized a conference called "Preserving Legal Information for the Twenty-First Century: Toward a National Agenda."[7] Oakley and his team sought to use the conference as a platform to address the vulnerability of born-digital legal materials, to explore the role of the law library community in preserving at-risk legal content, and to develop a plan of action to prevent further loss of legal information in the digital age.[8]

¶8 Conference attendees, including experts in the fields of law librarianship, legal publishing, and digital preservation, decided to form a new organization to tackle these issues: the Legal Information Preservation Alliance (LIPA).[9] LIPA was established to provide the law library community with the leadership, guidance, and organizational backing to support the preservation of legal information on a national scale.[10]

¶9 LIPA's 2006 strategic plan called for the development of a pilot project to preserve born-digital legal information.[11] To move this strategic objective forward, three LIPA-member libraries—the Georgetown Law Library, Maryland State Law Library, and Virginia State Law Library—came together as partners and established the Chesapeake Project. The project began as a two-year pilot digital preservation program to explore the feasibility of forming a collaborative, nationwide digital preservation initiative within the law library community.[12]

¶10 In 2006, the three partner libraries began defining their working relationship and selected a suite of OCLC tools and systems for the capture, access, and preservation of born-digital, web-published content.[13] On February 27, 2007, the institutions participating in the pilot began actively harvesting content from the

---

7. *See Preserving Legal Information for the Twenty-First Century: Toward a National Agenda*, 96 LAW LIBR. J. 655, 2004 LAW LIBR. J. 46.

8. *Id*. at 655, ¶¶ 1–3.

9. *Id*. at 656, ¶ 6.

10. *Id*. at 657, ¶ 8.

11. Legal Info. Preservation Alliance, Strategic Plan Outline 2 (June 20, 2006), http://www.aall net.org/committee/lipa/StratPlanFinalDraft20060620.doc.

12. Chesapeake Project, *supra* note 4.

13. Sarah Rhodes & Dana Neacsu, *Preserving and Ensuring Long-Term Access to Digitally Born Legal Information*, 18 INFO. & COMM. TECH. L. 39, 58, 60 (2009). The Chesapeake Project uses the OCLC Digital Archive for the preservation of its digital collections and an OCLC-hosted CONTENTdm interface at http://www.legalinfoarchive.org for user access to archived collections.

web and preserving this content within a shared digital archive.[14] The following year, the project's open-access CONTENTdm user interface, www.legalinfoarchive. org, was made available to the public.[15]

¶11 Due to the diversity of the three partner libraries, a strong collaborative relationship was required to ensure the Chesapeake Project's success. Not only did the project include two state law libraries and one academic law library, each with unique mandates and user groups, but the three libraries also varied in size. The Maryland State Law Library had a medium-sized staff of roughly fifteen, while the Virginia State Law Library had a small staff of five librarians and paraprofessionals. The Georgetown Law Library was significantly larger than its two partners combined, with a staff of approximately seventy, divided between two separate library buildings.

¶12 By selecting a vendor-provided digital preservation solution, the Chesapeake Project libraries were able to focus staff energies on developing the project's organizational structure, policies, and archive collections, as opposed to building and maintaining the technological infrastructure of the archive and access system. To keep up the project's momentum, the partners established a formal schedule of quarterly meetings to develop and continually reassess project policies; make shared project decisions; and share new information about the project, its tools and systems, and developments in the field of digital preservation. Each meeting was attended by a director or senior administrator, a project coordinator/digital archivist, and, when necessary, a senior cataloger or metadata specialist from each library.[16] A comprehensive collection plan, adapted from a template developed by the Web-at-Risk project,[17] was also created for the project. It described the project's mission and scope, acquisition and selection methods, metadata policies, approach to rights management, means of collection discovery and access, and digital preservation system. Project assessment and evaluation parameters, including the assessment of link rot among the original URLs for the archived titles, was also outlined in this document.[18]

¶13 In 2009, the Chesapeake Project's pilot phase came to a close, and the three partners committed to continue the project as a permanent preservation program. In addition to the project's 2009 self-evaluation, the partner libraries enlisted the Center for Research Libraries (CRL) to conduct an independent assessment of the Chesapeake Project's organization and policies, preservation strategies, and technological infrastructure based on criteria identified in the Trustworthy Repositories Audit & Certification Criteria and Checklist (TRAC).[19] The assessment, which

---

14. *See* CHESAPEAKE PROJECT, FIRST-YEAR PILOT PROJECT EVALUATION 2 (2008), http://www.legalinfo archive.org/policies/LEGAL_FirstYearProjectEvaluation.pdf.

15. Chesapeake Project, *Announcing the Chesapeake Project Web Interface,* LEGALINFOARCHIVE .ORG (Sept. 17, 2008), http://legalinfoarchive.org/custompages/news.php#20080917.

16. CHESAPEAKE PROJECT, COLLECTION PLAN 4–5, 30 (updated Jan. 2010), *available at* http://www .legalinfoarchive.org/policies/LEGAL_CollectionPlan_Updated_2010_01.pdf (providing a description of the project team and organizational structure).

17. NAT'L DIGITAL INFO. INFRASTRUCTURE & PRES. PROGRAM, COLLECTION PLAN TEMPLATE (Aug. 24, 2006), http://web3.unt.edu/webatrisk/reports/cpg_template_ikh_24aug2006.doc.

18. *See* CHESAPEAKE PROJECT, *supra* note 16, at 21.

19. TRUSTWORTHY REPOSITORIES AUDIT & CERTIFICATION: CRITERIA AND CHECKLIST (Feb. 2007), *available at* http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf.

included reviews of project documentation, interviews with and observation of project team members, and on-site examination of OCLC facilities, praised the project, finding that "the Chesapeake Project provides good stewardship of the web content it has identified and collected," addresses a real need in the legal research community, and uses tools and processes that are "cost-effective and focused."[20] The auditors also provided concrete recommendations for strengthening the project to ensure its future viability.

¶14 Today, the collaborative digital preservation effort that began as the Chesapeake Project is expanding. A new partner, the Harvard Law School Library, has recently joined the partnership. Additionally, the Legal Information Preservation Alliance announced in March 2010 the formation of its Legal Information Archive, which is open to all LIPA member libraries and offers subscriptions to OCLC digital preservation tools at a reduced group price. The Legal Information Archive is considered by LIPA to be an expansion of the Chesapeake Project.[21]

¶15 The collections preserved by the Chesapeake Project from its beginning through its three-year mark in 2010 were limited to born-digital law- and policy-related reports and documents, the majority of which were PDF documents, issued online via open-access web sites. All of the content preserved by the project was selected, based on collection development policies devised by each participating library, from authoritative Internet sources, such as government or organization-based web sites.[22] As such, the project archive represents a unique collection of authoritative web resources deemed by library selectors to be worthy of preservation as part of the permanent collections of the three original libraries. Examining the prevalence of link rot among the original URLs to which these archived web resources were posted provides valuable information about the stability of this content over time, while also validating digital preservation efforts aimed at safeguarding this type of web-published legal information.

The Problem of Link Rot

¶16 The typical life span of a web resource is difficult to determine with certainty. Koehler, who conducted a longitudinal study of URL permanence from 1996 through 2003, found URL stability to vary by a resource's age, discipline, domain, and field.[23] Despite our inability to pinpoint the average time it takes for a web resource to disappear from its URL, or even the overall extent of link rot within the online universe, we know that the phenomenon of link rot is indeed pervasive, and it has been well documented by studies from a variety of disciplines.

¶17 Many researchers have specifically explored the prevalence of link rot among web citations in scholarly literature. A 2003 analysis established that roughly thirteen percent of URL citations published in three leading scientific journals

---

20. Ctr. for Research Libraries, Advisory Assessment of the Chesapeake Project 2 (2009) (on file with author).

21. Legal Info. Pres. Alliance, The Legal Information Archive: A Solution for Preserving and Ensuring Long-Term Access to Digitally Born Legal Information (2010), *available at* http://listproc.ucdavis.edu/archives/law-lib/law-lib.log1003/att-0117/01-LegalInfoArchive.pdf.

22. *See* Chesapeake Project, *supra* note 16, at 8.

23. Koehler, *supra* note 3.

became inactive within twenty-seven months of the citing article's publication.[24] A study of footnotes in three influential journals in the field of journalism and communication found that only about sixty percent of web citations tested over a four-year period remained accessible.[25] In the field of medicine, a study of five biomedical journals showed the average annual link rot rate among cited URLs to be 5.4%.[26] In 2008, a study of web citation permanence among history journals found that thirty-eight percent of the cited URLs were inaccessible within seven years of an article's publication, while ten percent were inactive within months of publication.[27] And in law, Susan Lyons called attention to the proliferation of web citations in legal scholarship and the access challenges posed by link rot within law review footnotes.[28] Demonstrating that the problem of link rot transcends geographic boundaries as well as academic discipline, a New Zealand study found that thirty percent of web citations appearing within a sample of New Zealand–based scholarly journal articles published from 2002 through 2005 were no longer working by 2006.[29]

¶18 In the field of law, Coleen Barger explored link rot among URLs cited by judges in appellate court decisions. Her 2002 analysis showed that thirty-four percent of web citations from 2001 decisions had become inaccessible, and among URLs cited within 1997 decisions, nearly eighty-five percent were inactive.[30] Mary Rumsey studied the problem of link rot in law review citations appearing from 1997 through 2001. Her findings were similar to Barger's: thirty-eight percent of the URLs cited within a sample of law review articles issued in 2001 had become inactive by 2002, while seventy percent of those published in 1997 were no longer accessible.[31] Helane Davis also investigated citation link rot in a study limited to articles published by three law reviews from 2001 through 2003. By October 2004, forty percent of the URL citations analyzed by Davis had become invalid.[32]

¶19 Davis's study found link rot among citations to government web sites to be on par with that of citations to .com and .net web sites.[33] Rumsey also found federal government web citations and nongovernment web citations to be equally

---

24. Robert P. Dellavalle et al., *Going, Going, Gone: Lost Internet References*, 302 SCIENCE 787, 787 (2003).

25. Michael Bugeja & Daniela V. Dimitrova, *The Half-Life Phenomenon: Eroding Citations in Journals*, 49 SERIALS LIBR. 115, 117 (2005).

26. Randy J. Carnevale & Dominik Aronsky, *The Life and Death of URLs in Five Biomedical Informatics Journals*, 76 INT'L J. MED. INFORMATICS 269, 271 (2007).

27. Edmund Russell & Jennifer Kane, Research Note, *The Missing Link: Assessing the Reliability of Internet Citations in History Journals*, 49 TECH. & CULTURE 420, 427 fig.2 (2008).

28. Susan Lyons, *Persistent Identification of Electronic Documents and the Future of Footnotes*, 97 LAW LIBR. J. 681, 2005 LAW LIBR. J. 42.

29. Ailsa Parker, Link Rot: How the Inaccessibility of Electronic Citations Affects the Quality of New Zealand Scholarly Literature [12] (2007), *available at* http://works.bepress.com/ailsa_parker/1.

30. Coleen M. Barger, *On the Internet, Nobody Knows You're a Judge: Appellate Courts' Use of Internet Materials*, 4 J. APP. PRAC. & PROCESS 417, 438 (2002).

31. Mary Rumsey, *Runaway Train: Problems of Permanence, Accessibility, and Stability in the Use of Web Sources in Law Review Citations*, 94 LAW LIBR. J. 27, 35 tbl.1, 2002 LAW LIBR. J. 2 tbl.1.

32. Helane E. Davis, *Keeping Validity in Cite: Web Resources Cited in Select Washington Law Reviews, 2001–03*, 98 LAW LIBR. J. 639, 646, 2006 LAW LIBR. J. 38 ¶ 24.

33. *Id.* at 661, ¶ 65.

vulnerable to link rot, despite the perception that resources published to government domains are more stable than those published to web sites hosted by commercial entities, organizations, or educational institutions.[34] Other studies have explored the viability of government URLs, with mixed results. One 2003 study found that government URLs remain stable longer than those from other top-level domains, such as .com, .edu, and .net.[35] However, more recent studies, published in 2007 and 2008, have shown resources published at .gov URLs to have a greater frequency of link rot than those from other top-level domains.[36]

Methodology

Definitions

¶20 For the purpose of the present analysis, the term "URL" describes a Uniform Resource Locator, or Internet address directing to a file site on the World Wide Web. The term "link rot" is applied to describe a URL that no longer provides direct access to files matching the content originally harvested from the URL and currently preserved in the Chesapeake Project's digital archive. The term "archived title" refers to the individual web site, document, monograph, or serial harvested from the web and ingested into the digital archive. Each archived title has a single, corresponding bibliographic record in OCLC's WorldCat catalog. A single archived title may be composed of multiple archived items, as in the case of multi-part monographs or serial web publications.

¶21 "Top-level domains" are the domain-name suffixes following the final "dot" in a web site's domain name sequence. Top-level domains include .gov, .com, .org, .edu, and .us; these suffixes can be used to indicate the type of organizational entity that hosts or publishes a web site, such as a governmental (.gov), commercial (.com), or educational (.edu) entity.

¶22 "File format type" refers to the digital manifestation of the resource located at a URL. File format types must be compatible with an operating system's platform and software applications in order to render a file's content. Examples of file format types are X/HTML web pages, PDF, and Word document files.

Samples

¶23 Three samples of archived titles were used for this analysis. Sample 1 (2007–2008) is a random sample originally generated in March 2008 from titles archived during the first year of the project, between the dates of February 27, 2007, and February 29, 2008. Sample 2 (2007–2009) is a random sample generated in March 2009 from the collection of titles in the archive harvested between February 27,

---

34. Rumsey, *supra* note 31, at 35, ¶ 25.

35. David C. Tyler & Beth McNeil, *Librarians and Link Rot: A Comparative Analysis with Some Methodological Considerations*, 3 PORTAL: LIBR. & ACAD. 615, 621–22 (2003).

36. John Markwell & David W. Brooks, *Evaluating Web-Based Information: Access and Accuracy*, 85 J. CHEM. EDUC. 458, 458 (2008); C. Rockelle Strader & Farrell D. Hamill, *Rotten but Not Forgotten: Weeding and Maintenance of URLs for Electronic Resources in The Ohio State University Online Catalog*, 53 SERIALS LIBR. 163, 174 (2007).

2007, and the project's second-year anniversary on February 28, 2009. Sample 3 (2007–2010) is a random sample generated in March 2010 from the entire archive collection of titles harvested between the project's beginning on February 27, 2007, and the project's three-year mark on February 28, 2010.

Sample 1 (2007–2008)

¶24 In March 2008, the Chesapeake Project conducted its first-year project evaluation, which included an analysis of URL link rot conducted using a sample of 579 archived titles. This sample was randomly generated from a master list of the OCLC bibliographic record numbers for all 1266 titles archived from February 27, 2007, through February 29, 2008, ensuring results at a 95% confidence level and confidence interval of +/- 3%.[37] In other words, from the entire population of 1266 titles archived as of the project's first-year mark, there is a 95% probability that a sample of 579 randomly generated titles accurately reflected the entire collection within three percentage points at the time of analysis.

¶25 The original URLs of titles in this sample were analyzed for link rot in March 2008 and reassessed at the project's second- and third-year marks in March 2009 and March 2010, respectively, in an effort to determine if additional titles in the sample had disappeared in the years following the original analysis of the sample in 2008. Results of the present 2010 study are compared to those of the initial 2008 analysis and subsequent 2009 analysis.

Sample 2 (2007–2009)

¶26 To obtain sample 2, a master list of archived titles, comprising all titles harvested from the Internet since the start of the project, along with each title's corresponding OCLC bibliographic record number, was assembled by project participants on March 4, 2009. This list included a total of 1872 titles archived between the dates of February 27, 2007, and February 28, 2009.

¶27 From this list of 1872 titles archived during the project's first two years, a sample of 680 OCLC bibliographic record numbers was randomly selected. This sample size ensured results at a 95% confidence level and confidence interval of +/- 3%. The original URLs of titles in this sample were analyzed for link rot in March 2009.[38]

Sample 3 (2007–2010)

¶28 To obtain sample 3, a master list of archived titles, comprising all titles harvested from the Internet since the start of the project, along with each title's corresponding OCLC bibliographic record number, was assembled by project participants on March 2, 2010. This list included a total of 2372 titles archived between the dates of February 27, 2007, and February 28, 2010.

¶29 From this list of 2372 titles archived during the project's first three years, a sample of 736 OCLC bibliographic record numbers was randomly selected. This

---

37. CHESAPEAKE PROJECT, *supra* note 14, at 15.
38. CHESAPEAKE PROJECT, TWO-YEAR PILOT PROJECT EVALUATION 32–33 (2009), http://www
.legalinfoarchive.org/policies/legal_twoyearprojectevaluation_june2009.pdf.

sample size ensured results at a 95% confidence level and confidence interval of +/- 3%. The original URLs of titles in this sample were analyzed for link rot in March 2010.

Data Gathering

¶30 To determine whether or not link rot was present at an archived title's original URL, the title's OCLC number was used to retrieve the item's metadata record from the CONTENTdm system. The metadata for each archived title provided the original URL from which the archived item was harvested. These original URLs were opened in a web browser to determine whether the live content matched the archived content in the CONTENTdm system. If not, link rot was determined to be present at the URL.

¶31 A spreadsheet was created for each sample. Using these spreadsheets, researchers tracked each sample title's OCLC number, original URL, and whether or not link rot was observed at the URL. Additionally, the top-level domain and file format type for the files found at each URL were also recorded.

¶32 File format types were recorded in each item's metadata at the time of harvest to facilitate future preservation action. "Stand-alone" files, such as PDFs, were archived in their native formats, while a web harvester was launched to capture web pages, comprising multiple files, which were bundled and archived within ARC or WARC files.[39]

¶33 Researchers were given special instructions for analyzing serial or multipart monograph titles. These titles often require multiple harvests from more than one URL and are associated with multiple preservation metadata records in the digital archive. Researchers were instructed to check the original URL, top-level domain, and format type of the record appearing at the mid-point of the results list only; in other words, neither the earliest nor the most recently harvested record was analyzed.

Results

Sample 1 (2007–2008)

Prevalence of Link Rot

¶34 When sample 1 was first analyzed in March 2008, link rot was found to be present in 48 of 579 URLs. One year later, in March 2009, the sample was analyzed a second time. The second analysis demonstrated that link rot was present in 83 out of the original sample of 579 URLs. In other words, 14.3% of the archived titles had disappeared from their original URLs within two years of harvest, compared to the March 2008 analysis, which had shown link rot among the sample URLs to be 8.3%.

¶35 The present analysis of the sample showed that by March 2010, the prevalence of link rot had increased to 160 out of 579 URLs. Thus, within three years of

---

39. For more information on WARC files, see Nat'l Digital Info. Infrastructure & Pres. Program, WARC, Web ARChive File Format, http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml (last updated Sept. 2, 2009).

harvest, link rot among the sample URLs had increased to 27.6%. The ratio of URLs with link rot versus working URLs, as of March 2008, March 2009, and March 2010, is illustrated in figure 1.

2008
- URLs with Link Rot (48)
- Working URLs (531)

8%

92%

2009
- URLs with Link Rot (83)
- Working URLs (496)

14%

86%

2010
- URLs with Link Rot (160)
- Working URLs (419)

28%

72%
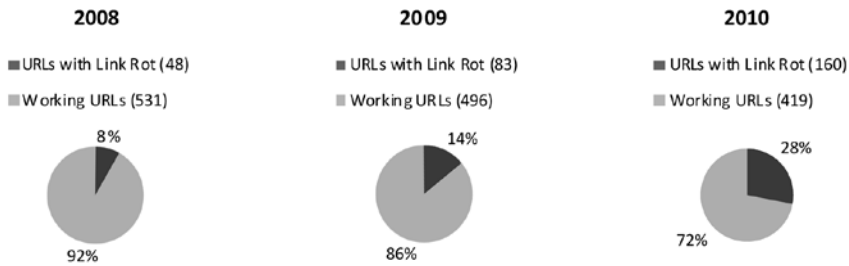
Figure 1. Ratio of URLs with Link Rot to Working URLs in Sample 1

Link Rot and Top-Level Domains

¶36 More than 90% of the top-level domains in the sample were state government (state.[state code].us), organization (.org), or government (.gov) URLs, representing approximately 41%, 32%, and 17% of the sample, respectively. Other top-level domains, which accounted for approximately 7% of the sample combined, were .edu, .com, and .net, which respectively represented 2.95%, 2.25%, and 1.9%. Less than 3% of the sample was represented by a combination of .mil, .us, .info, .uk, .au, .ca, and .int domains. The sample also included one IP address.

¶37 In the original 2008 analysis, link rot was present in 10.8% of URLs with state top-level domains, 10% of URLs with government top-level domains, and 3.8% of URLs with organization top-level domains. Although .edu and .com URLs represented a much smaller portion of the sample, both domains were found to have relatively high link rot levels of 11.8% and 15.4%, respectively, in 2008.

¶38 In 2009, the prevalence of link rot increased among URLs with state, government, organization, education, network (.net), and military (.mil) domains. Among URLs in the sample with state top-level domains, link rot increased by five percentage points from 2008 to 2009. While .gov URLs were shown to have relatively little increase in link rot between 2008 and 2009, the 2009 analysis demonstrated a significant increase among URLs with education top-level domains, from 11.8% to 35.3% over the one-year period, while no increase in link rot among commercial URLs was observed.

¶39 The current 2010 analysis of the sample showed that link rot was present in one-third of the URLs with a state government top-level domain. The prevalence of link rot among these state URLs more than doubled in the year following the 2009 analysis, and it nearly tripled in the two years following the original 2008 analysis. Link rot was found in more than 22% of .org URLs, nearly double the link rot observed in 2009, and almost six times the link rot found among URLs with an organization top-level domain in the 2008 analysis. Twenty-five percent of government URLs were found to have link rot in 2010, an increase from 13% in 2009 and

10% in 2008. Although they represented only a small fraction of the sample, commercial and network URLs both experienced a jump in link rot, from 15.4% in both 2008 and 2009 to 30.8% among .com domains, and from zero in 2008 and 9.1% in 2009, to 27.3% among .net domains in 2010. A list of all top-level domains found in the sample, along with link rot detected in 2008, 2009, and 2010, is available in table 1.

Table 1

Top-Level Domains and Link Rot Frequency in Sample 1

Top-Level Domain	Total in Sample	Link Rot Frequency 2008	Link Rot Frequency 2009	Link Rot Frequency 2010
.state.__.us	240	26 (10.8%)	38 (15.8%)	77 (32.1%)
.org	184	7 (3.8%)	21 (11.4%)	41 (22.3%)
.gov	100	10 (10%)	13 (13%)	25 (25%)
.edu	17	2 (11.8%)	6 (35.3%)	6 (35.3%)
.com	13	2 (15.4%)	2 (15.4%)	4 (30.8%)
.net	11	0	1 (9.1%)	3 (27.3%)
.mil	3	0	1 (33.3%)	1 (33.3%)
.us	3	0	0	0
.info	2	1 (50%)	1 (50%)	1 (50%)
.uk	2	0	0	1 (50%)
.au	1	0	0	0
.ca	1	0	0	0
.int	1	0	0	0
[IP address]	1	0	0	1 (100%)
TOTAL	**579**	**48**	**83**	**160**

Link Rot and Format Types

¶40 More than 95% of the titles in the sample were PDF files posted to the web. Of these titles, link rot was found to be present in 8.2% in the original 2008 analysis, a figure that increased to 14.1% in 2009 and to 27% in the current 2010 analysis.

¶41 A much smaller portion of the sample, 4%, was represented by X/HTML web page files. These items in 2008 were found to have a link rot rate of 8.7%; this figure jumped to 17.4% in 2009 and to 34.8% in 2010.

¶42 Other format types found in the sample included combination HTML/PDF files and Microsoft Word documents (DOC). No change in link rot rates among HTML/PDF combination and Word document files was observed in 2009, but the 2010 analysis showed link rot to be present among *all* of the HTML/PDF combination files. A list of all format types found in the sample, along with their link rot rates in 2008, 2009, and 2010, is available in table 2.

**Table 2**

Format Type and Link Rot Frequency in Sample 1

Format Type	Total in Sample	Link Rot Frequency 2008	Link Rot Frequency 2009	Link Rot Frequency 2010
PDF	552	45 (8.2%)	78 (14.1%)	149 (27%)
X/HTML	23	2 (8.7%)	4 (17.4%)	8 (34.8%)
HTML/PDF	3	1 (33.3%)	1 (33.3%)	3 (100%)
DOC	1	0	0	0
TOTAL	**579**	**48**	**83**	**160**

Sample 2 (2007–2009)

Prevalence of Link Rot

¶43 Ninety-three out of 680 URLs in the 2009 sample were found to be inactive during the March 2009 analysis. Given the total of total of 1872 titles archived, it can be inferred with a 95% confidence level and a confidence interval of +/- 3% that 13.7% of the original URLs of all titles harvested and archived during the first two years of the Chesapeake Project had become inactive as of March 2009. The ratio of active to inactive URLs in the sample is illustrated in figure 2.

Sample 2 (2007-2009)

■ URLs with Link Rot (93)

▨ Working URLs (587)

14%

86%
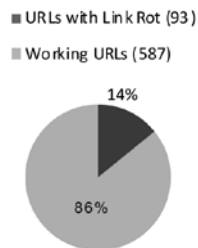
Figure 2. Ratio of URLs with Link Rot to Working URLs in Sample 2

Link Rot and Top-Level Domains

¶44 More than 88% percent of titles in the 2009 sample came from state, organization, and government top-level domains, which represented approximately 35%, 31%, and 23% of the sample, respectively. Of these three top-level domains, link rot was found to be present among 15.7% of URLs with state top-level domains, 13.7% of URLs with organization top-level domains, and 11% of URLs with government top-level domains.

¶45 Although URLs with education top-level domains represented a much smaller portion of the sample, they were found to have relatively high link rot levels of 26%. Commercial URLs, like education URLs, represented less than 3.5% of the

sample, but had a much lower instance of link rot, only 4.5%. A list of all top-level domains found in the sample, along with the prevalence of link rot among each, is available in table 3.

Table 3

Top-Level Domains and Link Rot Frequency in Sample 2

Top-Level Domain	Total in Sample (2007–2009)	Link Rot Frequency 2009
.state.__.us	235	37 (15.7%)
.org	212	29 (13.7%)
.gov	155	17 (11%)
.edu	23	6 (26%)
.com	22	1 (4.5%)
.net	12	0
.us	5	0
.mil	4	0
.info	3	2 (66.7%)
.uk	3	1 (33.3%)
.int	2	0
.au	1	0
.ca	1	0
.eu	1	0
[IP address]	1	0
TOTAL	**680**	**93**

Link Rot and Format Types

¶46 More than 94% of the titles in the sample were comprised of PDF files, and link rot was found to be present in 13.6% of these PDFs during the March 2009 analysis. A significantly smaller portion of the sample, 3.5%, was represented by X/HTML files. Interestingly, these titles were found to have a similar link rot rate of 12.5%. Other format types found in the sample included combination HTML/PDF files, ASCII text (TXT) files, and proprietary Microsoft Word documents (DOC) files and PowerPoint (PPT) presentations. A list of all format types found in the sample, along with the prevalence of link rot found in March 2009, is available in table 4.

Sample 3 (2007–2010)

Prevalence of Link Rot

¶47 Out of 736 titles randomly selected for the 2010 sample, link rot was found to be present in 165 URLs. A total of 2372 titles were archived from February 27, 2007, through February 28, 2010; therefore, it can be inferred with a 95% confidence level and a confidence interval of +/- 3% that 22.4% of the original URLs of all titles harvested and archived during the first three years of the Chesapeake
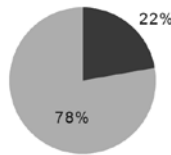
Table 4

Format Type and Link Rot Frequency in Sample 2

Format Type	Total in Sample (2007–2009)	Link Rot Frequency 2009
PDF	641	87 (13.6%)
X/HTML	24	3 (12.5%)
DOC	8	0
HTML/PDF	5	3 (60%)
TXT	1	0
PPT	1	0
TOTAL	**680**	**93**

Project had succumbed to link rot by March 2010. The ratio of working URLs to those with link rot is illustrated in figure 3.

Sample 3 (2007-2010)

■ URLs with Link Rot (165)
▨ Working URLs (571)

Figure 3. Ratio of URLs with Link Rot to Working URLs in Sample 3

Link Rot and Top-Level Domains

¶48 In the 2010 sample, 86.8% of the top-level domains were state government, organization, and government URLs, which represented 34.8%, 30.4%, and 21.6% of the sample, respectively. Of these three top-level domains, link rot was present in 30.5% of URLs with state domains, 20.1% of URLs with organization domains, and 15.7% of URLs with government domains.

¶49 URLs with .com. and .net top-level domains were found to have link rot levels of 17.9% and 13.6%, respectively, while .edu URLs were found to have a lower link rot rate of 7.1%. A list of all top-level domains found in the 2010 sample, along with their link rot rates, is available in table 5.

Table 5

Top-Level Domains and Link Rot Frequency in Sample 3

Top-Level Domain	Total in Sample (2007–2010)	Link Rot Frequency 2010
.state.__.us	256	78 (30.5%)
.org	224	45 (20.1%)
.gov	159	25 (15.7%)
.edu	28	2 (7.1%)
.com	28	5 (17.9%)
.net	22	3 (13.6%)
.us	2	0
.mil	5	2 (40%)
.info	2	0
.uk	3	2 (66.7%)
.int	2	0
.au	1	0
.eu	2	1 (50%)
[IP address]	2	2 (100%)
TOTAL	**736**	**165**

Link Rot and Format Types

¶50 More than 95% of the titles in the sample were comprised of PDF files, and link rot was found to be present in 21.2% of these PDFs during the March 2010 analysis. A significantly smaller portion of the sample, only 3%, was represented by X/HTML files. These titles were found to have a dramatically increased link rot rate of 54.5%. Other format types found in the sample included combination HTML/ PDF titles, 80% of which had been impacted by link rot, and Microsoft Word documents (DOC), which had no incidence of link rot. A list of all format types found in the sample, along with the prevalence of link rot found in March 2010, is available in table 6.

Table 6

Format Type and Link Rot Frequency in Sample 3

Format Type	Total in Sample (2007–2010)	Link Rot Frequency 2010
PDF	702	149 (21.2%)
X/HTML	22	12 (54.5%)
DOC	7	0
HTML/PDF	5	4 (80%)
TOTAL	**736**	**165**

Discussion

¶51 The present study explored the stability of URLs for legal, government, and policy-related web resources selected for preservation and harvested from the web for inclusion in the Chesapeake Project, which was initiated in late February 2007. The results demonstrate that among the original URLs from which content was harvested for the Chesapeake Project, link rot has increased steadily over time.

¶52 In analyzing a single sample of these original URLs at annual intervals, the prevalence of link rot was 8.3% in 2008, within zero to twelve months of the content being harvested. One year later, twelve to twenty-four months after the content was harvested, link rot in the same sample was found to have jumped to 14.3%. In the most recent analysis, in 2010, link rot was found to be 27.6%. In other words, link rot increased from about one in every twelve archived titles in 2008, to one in every seven titles in 2009, and finally to about one in every 3.5 titles in 2010.

¶53 An analysis of separate samples gathered at annual intervals from 2008 through 2010 to track link rot among all titles in the archive at the time of assessment also showed an increase in link rot over time. In 2008, link rot was present in 8.3% of the resource URLs from a statistically significant sample of 579 titles. In 2009, link rot was found in 13.7% of the URLs from a sample of 680, and in 2010, the content at 22.4% of the URLs from a sample of 736 was found to have been lost to link rot.

¶54 These findings are consistent with those of previous studies demonstrating an overall increase in link rot over time. However, in comparing this study to previous studies, there appears to be little consistency in the rate and extent to which link rot can be anticipated overall, perhaps due to the distinctiveness of each collection of URLs being analyzed. As noted above, Koehler speculated that this variance may be due to the age, domain, or discipline of the web resources being studied.[40] Yet, the findings of the present study diverge even from others in the discipline of law and legal informatics. Specifically, Rumsey, Barger, and Davis found the prevalence of link rot among URLs cited within court decisions and law review articles to exceed thirty percent within one year of citation.[41] Within twenty-four to thirty-six months following harvest, the prevalence of link rot among the original URLs of web resources archived by the Chesapeake Project has yet to exceed thirty percent. Of course, Rumsey and Barger were exploring this issue in 2002, eight years prior to the current study, and Davis's study was conducted two years later in 2004; it is likely that URL stability has improved in recent years. Moreover, they were studying web citations, as opposed to a collection of authoritative web-based content selected for preservation by law libraries; certainly there are distinct differences in the resources comprising these various collections of URLs.

¶55 State, organization, and government domains represented the three most common top-level domains in the archive. Among these, state top-level domains were shown consistently to have the highest level of link rot in every sample analyzed. Sample 1, comprising only titles archived from 2007 through 2008, showed

---

40. Koehler, *supra* note 3.
41. Barger, *supra* note 30, at 438; Davis, *supra* note 32, at 646; Rumsey, *supra* note 31, at 35.

link rot to be higher in .gov domains than .org domains in 2008, 2009, and 2010. However, sample 2, comprising content harvested from 2007 to 2009, showed the prevalence of link rot among .org domains to exceed that of .gov domains in the March 2009 analysis. Likewise, the March 2010 analysis of sample 3, consisting of content harvested from 2007 to 2010, also showed link rot among .org domains to exceed that of .gov domains. Despite this variation, it is clear that content at state government top-level domains appears to be the most at risk for link rot; content at seemingly stable organization and government top-level domains is also vulnerable to link rot, and this vulnerability increases with the age of the resource.

¶56 Due to the overrepresentation of PDF files in the archive as compared to other file formats, it is difficult to determine if a relationship exists between file format type and link rot, i.e., whether individual PDF or Word documents posted to the web disappear at a higher rate than actual web pages do. Based on the present analysis, it seems that, over time, X/HTML web pages are more vulnerable than PDFs to link rot. The analysis of sample 1 in 2008, 2009, and 2010 showed an increase in link rot among PDFs to 27% in 2010 from 8.2% in 2008. The same sample showed an increase in link rot among X/HTML web pages to nearly 35% in 2010 from 8.7% in 2008. Yet sample 2, analyzed in 2009, showed link rot to be more prevalent among PDF files than web pages, though by only 1.1%. Sample 3, analyzed in 2010, showed a dramatic increase in link rot among X/HTML web pages compared to PDF files, 54.5% web page link rot compared to 21.2% PDF link rot. Clearly, this variation warrants further examination, such as an analysis of format-specific samples that account for and accurately represent the populations of various file formats present in the archive.

¶57 The results of this study are not meant to be broadly applicable or to provide a representation of link rot throughout the universe of web resources; rather, this study paints a portrait of the vulnerability of the original sources for the collections archived by the Chesapeake Project, while also providing insight into the vulnerability of law- and policy-related web resources selected by experienced law librarians from seemingly stable open-access web sites hosted by reputable organizations and state and federal governments. Thanks to the efforts of the Chesapeake Project, none of the content analyzed in this study has been truly lost; all of the content has been preserved in a digital archive, with copies accessible to users who are able to discover these titles via the participating libraries' OPACs, OCLC's WorldCat, the project's CONTENTdm interface, or a simple web search-engine search.

¶58 The Chesapeake Project set out in 2007 "to stabilize, preserve, and ensure permanent access to critical born-digital . . . legal materials on the World Wide Web."[42] The present study demonstrates that the project has been largely successful in this effort. The project was also intended to "help inspire . . . a comprehensive, collaborative, and nationwide preservation program for legal resources."[43] With the launch of LIPA's Legal Information Archive in 2010, and the invitation to law libraries throughout the country to join this collaborative effort to preserve our digital legal heritage, this vision today is within reach.

---

42. CHESAPEAKE PROJECT, *supra* note 16, at 2.
43. *Id.* at 3.