

Pennsylvania Libraries: *Research & Practice*

Practice

Capture All the URLs

First Steps in Web Archiving

Alexis Antracoli, Steven Duckworth, Judith Silva, & Kristen Yarmey

Alexis Antracoli is Records Management Archivist at Drexel Libraries, aaa366@drexel.edu

Steven Duckworth served as Records Management Intern at Drexel University and is now an Archivist with the National Park Service, steve@stevenduckworth.com

Judith Silva is the Fine & Performing Arts Librarian and Archivist at Slippery Rock University of Pennsylvania, judith.silva@sru.edu

Kristen Yarmey is an Associate Professor and Digital Services Librarian at the University of Scranton Weinberg Memorial Library, kristen.yarmey@scranton.edu

As higher education embraces new technologies, university activities—including teaching, learning, and research—increasingly take place on university websites, on university-related social media pages, and elsewhere on the open Web. Despite perceptions that “once it’s on the Web, it’s there forever,” this dynamic digital content is highly vulnerable to degradation and loss. In order to preserve and provide enduring access to this complex body of university records, archivists and librarians must rise to the challenge of Web archiving. As digital archivists at our respective institutions, the authors introduce the concept of Web archiving and articulate its importance in higher education. We provide our institutions’ rationale for selecting subscription service Archive-It as a preservation tool, outline the progress of our institutional Web archiving initiatives, and share lessons learned, from unexpected stumbling blocks to strategies for raising funds and support from campus stakeholders.

Introduction

In Pennsylvania and elsewhere, higher education is experiencing many significant shifts as it adjusts to the new capabilities and culture of digital technology. While MOOCs and mediated classrooms dominate the news, the disruption of universities' long-established information sharing and communication practices has been mostly unacknowledged. Static analog recordkeeping is being uprooted as dynamic digital media replace the printed publications long preserved by university archivists. Some of these digital files bear a resemblance to their print ancestors, but others present much more complexity. University course catalogs, for example, may now take the form of a relational database, refreshed each year with updated digital content. Internal records, such as assessment reports and faculty senate meeting minutes, previously typed and stored in departmental filing cabinets, are now e-mailed to recipients in digital formats (often PDF). Press releases, once simply typed on university letterhead, are now dynamic Web pages, often featuring embedded media files like images and videos. Student clubs and activities, once carefully preserved in the campus newspaper and annual yearbook, have moved onto social media pages and Web apps. Alumni plan their reunions on Facebook, while current students trade photographs on Instagram and SnapChat.

Accompanying these changes in format is a similar disruption in scale. As digital information becomes easier to create and share, university departments and divisions produce an even more prolific body of records. The sheer number of born-digital documents and the frequency and regularity with which they are updated or replaced by new information simply overwhelms traditional archival practice. How can university archivists and records managers ever hope to gather, select, preserve, and manage their institutional records? The complexity of this question has proved staggeringly difficult for many institutions, forestalling necessary action.

Still, these digital preservation challenges urgently demand a response, even an imperfect one. Interestingly, a common element across many of the examples above is the World Wide Web. An institution's Web presence has historical value as a reflection of university life; it is the public face of the institution and its community. Additionally, the Web plays an important role in university recordkeeping and communication, providing access to an aggregation of digital documents while also establishing links between related materials. University-related websites, then, have the potential to serve as platforms for the systematic collection of digital university records for long-term preservation.

An increasing number of institutions are recognizing and acting upon the growing need to preserve Web content. A 2012 survey by the National Digital Stewardship Alliance (NDSA) described a "recent surge in Web archiving" that was "primarily due to universities starting Web archiving programs" (pp. 3-4). Even outside of universities, institutions of all types are beginning to consider Web archiving as a core function for archives and archivists as opposed to a side project or "digital initiative."

Each of the authors' respective institutions (Drexel University, Slippery Rock University of Pennsylvania, and the University of Scranton) has taken action to meet the challenge of capturing and preserving Web content. In this article, we introduce basic concepts and tools related to Web archiving and then describe our first steps: partnering with Archive-It's Web archiving service, obtaining buy-in and funding from our institutions, and selecting content for our collections. We also discuss more advanced steps taken at Drexel University, such as policy development and quality control, as well as future plans, from social media archiving to outreach and assessment.

What is Web Archiving?

The ubiquity and public accessibility of the World Wide Web tend to foster an impression of permanence. However, studies have demonstrated that this dynamic content is ever-changing and highly vulnerable to loss, degradation, or decreased access. Brügger (2005) estimated that roughly 80% of the material online disappears or

changes each year, making preservation a significant challenge for these dynamic and ephemeral information sources. SalahEldeen and Nelson (2012) studied the persistence of public social media posts and concluded that “after the first year of publishing, nearly 11% of shared resources will be lost and after that we will continue to lose 0.02% per day” (p. 125). Davis (2010) described an “increasingly... post-Web 2.0 world” of “highly volatile content, easily modified or removed by its authors and editors, without any guarantee that previously published versions, or any record of the change will persist.”

Link rot, or the failure of a URL to properly direct a user to the intended Web page, is a particular concern for academic research, as scholars increasingly cite Web-based sources. Indeed, the Chesapeake Digital Preservation Group (2012) found that within five years, 38% of the URLs in its sample of law and policy-related websites were broken and no longer provided access to the intended resource. A related study (Zittrain, Albert, & Lessig, 2013) had similarly devastating results: “More than 70% of the URLs within [the *Harvard Law Review* and other law journals], and 50% of the URLs found within United States Supreme Court opinions suffer reference rot—meaning, again, that they do not produce the information originally cited” (p. 167).

Web archiving, the complex process of harvesting Web content and then preserving it for the enduring future, is an attempt to address this digital preservation challenge. Web archivists aim to capture and preserve the dynamic and functional aspects of Web pages—including active links, embedded media, and animation—while also maintaining the context and relationships between files. Ideally, Web archives present documents and files with “temporal consistency”: that is, the user can navigate between Web pages that were all “live” at the same time (Ball, 2010, p.7). Far more than a simple collection of static snapshots, then, proper Web archives capture not only content but also aspects of the user environment. As such, practices for capturing and displaying archived websites must continually evolve, placing archivists in a perpetual struggle to keep pace with changing technology and practices on the live Web. As one Web archivist explained, “Right now, we’re 100 percent ready to archive the way the Web was 10 years ago” (Bleicher, 2011, p. 37).

Brewster Kahle, founder of the nonprofit Internet Archive, was one of the earliest and most vocal advocates for active preservation of the World Wide Web. Comparing the situation to the fire in the library of Alexandria and the recycling of old cinema films for their silver content, he argued that preserving websites would help avert a hole in history from forming (Kahle, 1997). As early as 1997, the one-year-old Internet Archive already held two terabytes of captured Web content, a harbinger of even more drastic changes in scale for Web archiving. In the early 2000s, as scholars increasingly expressed concern over the rapid growth of the Web and the dynamic nature of its content, Internet Archive’s collection grew by 144 terabytes in just five years (Day, 2003). By the end of 2010, the Internet Archive had swelled to 2.4 petabytes (Toyoda & Kitsuregawa, 2012), and it continues to grow at roughly 20 terabytes per month (Internet Archive, 2014).

Participation in Web archiving initiatives has grown steadily over the years (Toyoda & Kitsuregawa, 2012). Early pilots in the 1990s, such as PANDORA at the National Library of Australia and MINERVA at the Library of Congress, attempted to collect websites on a national scale. Academic libraries stepped into the ring in the 2000s, with Stanford University, the University of Michigan, the California Digital Library, and Harvard University among the early adopters. State and local governments also began exploring ways to capture their own websites in order to meet legal requirements to “preserve and administer” public records, which are increasingly created and published in digital formats (Martin & Eubank, 2007, p. 10). The establishment of several international and national organizations, such as the International Internet Preservation Consortium (IIPC) in 2003 and the NDSA in 2010, has further fostered research and collaboration on Web archiving efforts. In 2011, a global study of Web archiving by Gomes, Miranda, and Costa reported 42 existing projects and prompted the creation of a Wikipedia page (en.wikipedia.org/wiki/List_of_Web_archiving_initiatives) to track ongoing and emerging projects. In the United States, a 2012 survey by the NDSA indicated that 49 institutional respondents were already administering an active

Web archiving program, while an additional 25 were either pilot testing or planning to begin a program in the near future.

As more institutions launch Web archiving programs, librarians, archivists, and information scientists have published scholarly research relating to Web archiving issues and practices (Ayala, 2013). Many works such as Niu (2012a; 2012b) and Pennock (2013) address the various reasons to preserve Web content, the variety of ways to capture websites, the tools available to assist in this capture, the costs and infrastructure needed, and issues associated with preparing an archival initiative. The 2010 JISC PoWR report, *A Guide to Web Preservation*, discusses institutional reputation, risk management, supporting the institutional mission, and cost savings as arguments for Web archiving (Farrell, 2010). Legal concerns—including copyright, content liability, and data protection, among others—have prompted considerable discussion (e.g., Pennock, 2013), as have organizational challenges, such as selection, appraisal, collection development, staffing, and workflow (Day, 2003).

Best practices for capture have been a topic of continuing discussion. Day (2003) and Pearce-Moses and Kaczmarek (2005) highlighted various types of collecting procedures, describing a continuum of approaches. At one end of the spectrum is heavily automated, bulk harvesting (what Pearce-Moses and Kaczmarek describe as “technocentric”), such as the software-driven processes deployed by Internet Archive. Technocentric processes are by far the cheapest options but may result in more superficial, lower-quality collections. At the other extreme is the “bibliographic” approach, which resembles traditional library collection development. In this approach (employed, for example, by the Library of Congress’s MINERVA project), pages are identified, captured or printed for physical storage, and then cataloged. Such human-intensive processes result in higher-quality collections but at a significantly higher cost (Masanès, 2005). Critical of technocentric and bibliographic approaches and skeptical of user-driven submission, Pearce-Moses and Kaczmarek (2005) proposed the “Arizona Model,” which focuses on crafting layers (similar to series and subseries in a standard archival collection) that relate to directories and subdirectories of Web pages. A combination of human selection and integrated technology thus brings about a quality product with various levels of access (finding aid, controlled vocabulary indexing, and full-text searching).

Web Archiving Tools and Services

Like most other challenges in digital preservation, Web archiving inhabits a rapidly evolving landscape. A variety of tools and services, open source and proprietary, local and hosted, are currently available for consideration, and new services continue to emerge. Readers may consult the IIPC website (www.netpreserve.org) for a comprehensive list of recommended tools and software. Comparisons between Web archiving tools and services will vary by institution, depending on needs and available resources, but in all cases, close attention should be paid to the format, accessibility, and exportability of captured data in order to ensure interoperability between collections and services over time. For example, a recommended format for captured Web content is WARC (Web ARchive Container), which became an ISO standard in 2009 and is widely used across a variety of institutions.

Some institutions may prefer locally installed software, whether proprietary or open source. Proprietary tools range in functionality and robustness. Adobe Acrobat, for example, has a Web Capture feature that converts websites into PDF documents. Internal hyperlinks remain active, and external hyperlinks remain live (although the content at the destination URL may not endure), but any other dynamic functionality or navigability is lost. More specialized software options like Grab-a-Site, WebWhacker, and Teleport Pro contrive to download files from a website, including HTML and embedded content, such as images and videos. Other proprietary tools address specific, niche needs. Tweet Archivist Desktop, for example, is a Windows application that collects data from Twitter searches.

Institutions that seek greater control over Web archiving processes should explore the wide spectrum of open source software, much of which focuses on specific functions or tasks. Heritrix, for example, is a well-known and widely used open source Web crawler that was developed by the IIPC. HTTrack is a popular tool that downloads captured Web content onto a local directory. NutchWAX and Solr facilitate indexing and searching harvested content, while Wayback provides a user interface for browsing, or “replaying,” archived Web pages. Such task-specific open source tools are often combined into a single, integrated Web archiving workflow. Harvard University, for example, developed an in-house Web Archive Collection Service (WAX) that builds upon other open source tools like Heritrix, Wayback, and NutchWAX. Similar software packages include Mat Kelly’s Web Archiving Integration Layer (WAIL), Web Curator Tool (developed by the National Library of New Zealand and the British Library), and NetarchiveSuite (developed by Denmark’s Royal Library and State and University Library).

Libraries and archives unable to support open source or in-house solutions can also consider outsourcing some or all of their Web archiving needs to a growing selection of vendor services. The California Digital Library’s Web Archiving Service (WAS), established in 2009, supports California libraries primarily but offers subscriptions to libraries outside the state. OCLC’s Web Harvester is available to hosted users of OCLC’s CONTENTdm digital asset management service and Digital Archive repository. Aleph Archives offers both KEN Archiving Software, which is aimed at individual end users, as well as an institutional toolset called the Web Archiving Bucket. Some services, such as Hanzo Archives and Reed Technology’s Web Archiving Service, support forensic capture of Web content for organizations that need to comply with legal requirements or prepare for litigation.

Why Archive-It?

All three of our institutions chose a subscription service from Internet Archive called Archive-It (archive-it.org). The nonprofit organization Internet Archive began archiving the Web in 1996 and now preserves more than 360 billion captured URLs, all publicly accessible via the Wayback Machine (waybackmachine.org). Archive-It launched in 2005 as a subsidiary, subscription service and leverages the tools, workflow, and infrastructure developed by its parent organization. Archive-It runs on open source software, using Heritrix to crawl the Web, HTTrack to harvest content, NutchWAX to search files, and Wayback to display content to end users. All harvested data is stored in the open WARC format. As a subscription service, Archive-It hosts and maintains the software, conducts the actual crawls, and provides robust data storage in Internet Archive’s data centers (though partners can also request copies of their data for storage in a local institutional repository or an additional off-site data center). Annual fees are based on a “data budget,” that is, the amount of digital information the partner captures during crawls.

Archive-It held strong appeal for our three institutions for several reasons. Archive-It’s connection to the Internet Archive was a draw for Slippery Rock University’s library staff members; having read about Brewster Kahle’s efforts to archive the Web, they were impressed with what was already being collected and preserved in the Wayback Machine. For the University of Scranton, Archive-It’s subscription model was a good fit for institutional resources, as current staff time and expertise could not support the implementation or ongoing maintenance of a locally installed Web archiving service. At the same time, Archive-It’s use of open source software and the open WARC format were attractive, allowing the library to benefit from external support while still avoiding “vendor lock-in” via proprietary systems or data formats. Of further interest was the interoperability between Archive-It and DuraCloud, a cloud storage service that the University of Scranton employs for its digital preservation repository. Digital content captured during Scranton’s Archive-It crawls is automatically backed up in a dedicated DuraCloud space.

Archive-It is a popular favorite among Web archiving institutions, and its impressive list of users was another argument in its favor. When Slippery Rock University started looking into Web archiving in 2009, Archive-It was mentioned repeatedly, in fact almost exclusively, on professional electronic mailing lists. Indeed, the 2012 NDSA survey found “an overwhelming majority” (over 75%) of respondents using an external service for Web archiving had chosen Archive-It (p.12). At the time of this writing, Archive-It’s partner list included 312 collecting organizations, 131 of which self-identified as colleges or universities. Pennsylvania is well-represented with partnerships dating back to 2005. Pennsylvania Archive-It partners currently include:

- Bryn Mawr, Haverford, and Swarthmore Colleges – a joint partnership (2005)
- Bucknell University (2012)
- Chemical Heritage Foundation (2010)
- Curtis Institute of Music (2010)
- Drexel University (2009)
- Free Library of Philadelphia (2010)
- Gettysburg College (2013)
- La Salle University (2012)
- Penn State University (2012)
- Presbyterian Historical Society (2013)
- Slippery Rock University of Pennsylvania (2011)
- Temple University (2013)
- University of Pennsylvania Law School (2011)
- University of Pittsburgh School of Information Sciences (2014)
- University of Scranton (2012)

Recommendations by colleagues at other institutions were also considered. For Slippery Rock University, a Mid-Atlantic Regional Archives Conference (MARAC) presentation in 2011 by Rebecca Goldman of La Salle University confirmed the utility of Archive-It. Goldman suggested an Archive-It webinar as an introduction to the product and its capabilities. The University of Scranton consulted with colleagues at fellow Jesuit institution Creighton University, who endorsed Archive-It’s services and especially its customer support.

Subscribing to Archive-It also provides many advantages that would not be available by relying only on Internet Archive crawls of an institution’s Web content. These advantages include the ability to curate your own content, ensure that the full depth and breadth of your Web domain are captured, perform quality control, and crawl password-protected content. This results in a Web archive collection that is easier for your users to browse and search and is more complete and accurate than could be achieved by relying only on the Internet Archive. However, if an institution wants to include Internet Archive content in its collection, Archive-It offers a data extraction service in which content captured by Internet Archive can be added into an Archive-It user’s institutional collection. The University of Scranton took advantage of this service to supplement its Archive-It crawls of university Web pages (which began in 2012) with content harvested from Internet Archive crawls of the same URLs, which dated back to 2000. While this content had already been publicly available via the Wayback Machine, it is now full-text searchable and more easily discoverable by Scranton users.

Finally, Archive-It staff members were excellent resources during the exploration process, explaining the product and demonstrating its capabilities in introductory webinars as well as one-on-one discussions. They also set up trial accounts to allow experimentation with Archive-It’s tools, reports, and administrative interface: a significant aid in understanding the service.

Obtaining Buy-In and Securing Funding

Identifying campus stakeholders and getting buy-in are crucial first steps in the implementation of a Web archiving program. At Slippery Rock University, the first contact was the Information Technology department, partly to determine whether or not the staff was already archiving the university's website and partly to solicit its support. Other key stakeholders might be found among library colleagues, Web content creators, and, of course, administrators who hold the purse strings. At Slippery Rock University, the Office of Public Relations maintains the university's Web presence, so its staff was a natural collaborator in the identification of content worthy of capture and preservation. The concept of Web archiving may be unfamiliar to some constituents, so sharing Archive-It's introductory webinar might be helpful to create a common understanding of objectives and possibilities. At the University of Scranton, presentations and handouts helped to address stakeholders' questions about the nature and scope of the Web archiving proposal.

At each of our three institutions, positioning Web archiving as a key tool for records management was an effective way to convey the importance of digital preservation. In higher education, so much of what was once printed is now published only electronically—not only public-facing materials like college catalogs, alumni magazines, press releases, departmental journals, and newsletters but also internal documents like committee agendas, meeting minutes, and university policies. While Web archiving is not a single-arrow solution to the complex problem of digital records management, it is a handy tool and a major step forward in the process. Web archiving leverages college or university websites as aggregators of digital information about the institution, and regularly capturing these Web pages can be a systematic and efficient way to harvest digital documents and media (including photographs and videos of campus speakers and events) that would otherwise be extremely difficult for archivists to collect for preservation.

A major challenge when seeking buy-in for Web archiving (like all digital preservation projects) is demonstrating the return on investment for a tool or service that provides predominantly long-term rather than short-term benefits. At the authors' institutions, several tipping-point events established the urgency of the problem. For example, a longtime administrator retired, leaving behind a hard drive and e-mail account as the only local copy of her institutional knowledge. Assessment committees struggled to find and access internal reports, even those published digitally within just the past five years.

Archive-It's trial accounts were also very helpful in demonstrating the service and communicating its value to campus stakeholders. At the University of Scranton, an outdated academic Web server that had long been used for faculty and department Web content was scheduled for decommissioning. At the same time, the main university website underwent a significant design change. During presentations about our Archive-It trial account to campus stakeholders, demonstrating access to this archived content, no longer available on the live Web, provided a vivid example of the purpose and immediate need of Web archiving.

Furthermore, while a Web archiving subscription is a new and recurring cost for college and university budgets, it can also lead to cost savings in other areas. Effective records management can positively impact the productivity of employees, so demonstrating that Web archiving could actually be a cost-saving measure was a major step towards securing funding. At the University of Scranton, the library compared the cost of Archive-It's annual subscription fee to the much more burdensome cost of hiring or reassigning a full-time staff member to fulfill Web archiving and related records management tasks in-house.

Other sources of funding may be available, depending on institutional resources. Content creators who would benefit from the service may be willing to contribute to costs. Grant writing is another possibility, as Web archiving addresses an important, and often unfulfilled, preservation need. Lastly, one might identify a donor interested in digital preservation to help defray costs. However, given the mission-driven need to preserve Web-based information about a university and its community, sustained and recurring funding is recommended.

Content Selection

Archive-It's administrative interface gives institutions granular control over what content is crawled and how frequently crawls occur. One of the first steps in Web archiving is to select *seeds*, which are the base URLs of Web content to capture. A seed can be broad, encompassing an entire website (such as www.scranton.edu), or it can be narrow, specifying a part of a website (such as www.scranton.edu/academics/wml) or even a single Web page (such as www.scranton.edu/academics/wml/about/mission/index.shtml). Once seeds have been selected, Archive-It's scoping tools provide options for expanding or limiting the content of the crawl as well as scheduling the frequency of the crawl (e.g., one-time, daily, weekly, quarterly, and so on.) Users can also set a time or data limit to constrain crawls, and configurations can be evaluated by running test crawls, which do not count against the institution's data budget. More detailed information about seed selection and scoping may be found in Archive-It's Knowledge Center (webarchive.jira.com/wiki/display/ARIH/Knowledge+Center).

At Slippery Rock University, the library opted to test the Web archiving concept with its own content: the library homepage and the Archives' digital collections. Slippery Rock University's library is closely involved with an annual student research symposium, which provides a stepping stone beyond traditional library content along with additional stakeholders, such as the symposium's planning committee and the faculty whose students participate in the symposium. The library is also closely involved in the publication of an electronic newsletter promoting faculty research activities, which was a natural addition to its Web archiving efforts. From there, Slippery Rock added the university homepage, the University Public Relations' weekly e-newsletter, and the undergraduate and graduate catalogs. Crawls vary in frequency for different seeds. The university's home page, for example, is crawled monthly, while the campus e-newsletter *RockPride* is crawled weekly during the school semester. The undergraduate and graduate course catalogs are crawled annually. Slippery Rock University's next level of selection will include popular campus activities like athletics, student organizations, and alumni Web pages. The president's and provost's pages are natural choices, as are one-time events like anniversaries and annual events on campus.

The University of Scranton's content selection to date has similarly focused on university-owned pages. Early test crawls of the scranton.edu domain guided the selection of seeds, highlighting content that may have otherwise been overlooked. Currently, a large quarterly crawl captures the majority of the scranton.edu domain, with seeds including the main university website (www.scranton.edu) as well as related sites (admissions.scranton.edu, for example). Websites for special events such as 125th.scranton.edu (published for the University's 125th anniversary) are generally captured within this quarterly crawl, but one-time crawls are also an option for these materials if needed. A weekly crawl captures the university's *Royal News* online newsletter. Some external sites directly related to the university, such as the University of Scranton Players' website (www.thescrantonplayers.com) and alumni class pages (www.scrantonalumnicommunity.com) are also crawled on a quarterly basis.

There may, of course, be digital content of archival interest that is not appropriate for public consumption: some committee minutes or internal communication, for example. Archive-It permits crawls of intranet or otherwise password-protected content, providing that the partner supplies those credentials in the Web application. Once captured, content can be made publicly accessible or can be restricted in a private collection. The University of Scranton is currently pilot testing systematic capture and preservation of Web-based e-mails internal to the campus community, such as the university president's *Letters from Scranton Hall*, which are stored in a restricted collection.

Collection Development Policy

The initial process of selecting content raises important questions about the goals, scope, rights, and responsibilities of a Web archiving program. Decisions about content, access, and rights should be documented in a formal collection development policy, a step already taken by Drexel University (and one that the University of

Scranton and Slippery Rock University are working towards). Policy development requires both an assessment of an institution's broad collecting goals as well as new considerations related to the acquisition of digital content.

The first step in developing a collection development policy is to define the mission of the repository and then apply and interpret that mission to decisions about what Web content to crawl (and how often). Institutions take a variety of approaches to developing their scope and mission; an institution with a large collecting scope may be collecting only a small number of websites, while one with a smaller scope may want to archive websites more broadly than other types of materials. For example, the British Archives considered its mission to preserve society's cultural artifacts, rather than focusing its Web archiving efforts purely on official records (Brown & Thomas, 2005). Institutions may also take a thematic or special collections approach, archiving Web content related to a designated topic or event (see, for example, Columbia University's Human Rights Web Archive). Drexel University has approached Web archiving from both a records and a thematic perspective, ensuring that the official records of the university are preserved, while also collecting Drexel-related material from outside websites.

In addition to establishing a mission and scope, a collection development policy must define its designated community. The term "designated community" originated with the Open Archival Information Systems (OAIS) model and describes the intended audience for which the digital assets are selected and preserved, that is, the researchers expected to access the collection in the future (Consultative Committee for Space Data Systems, 2012). By defining the designated community, a repository can better predict what characteristics of the digital asset will be most important to future users and assess which of these merit long-term preservation. A clear concept of an intended research audience may thus inform decisions about preservation actions and user access issues, such as metadata and quality control. Drexel University defines its designated community as Drexel faculty, staff, students, and alumni and secondarily as general researchers interested in the history of Drexel. Other institutions may have broader or narrower designations, depending on their mission and collecting scope.

Additionally, a Web archiving collection development policy should address intellectual property considerations. Each of our Web archiving initiatives has primarily focused on content created and hosted by our own institutions, for which copyright is less of a challenge. However, any archives that crawls and preserves third-party content will need to address the issue of intellectual property. The Section 108 Study Group of the Library of Congress and the U. S. Copyright Office (section108.gov) advises that libraries and archives may capture publicly available Web content for preservation in order to ensure that valuable cultural resources are not lost. Password-protected content and content protected by robots.txt files, which block crawlers, are exceptions and should not be crawled without permission by the site owner(s). In keeping with the Section 108 Study Group guidelines, the Drexel University Archives allows non-Drexel content owners to opt out upon request, honors password protection and robots.txt files, and provides banners that describe content clearly as archival as an additional measure to protect intellectual property.

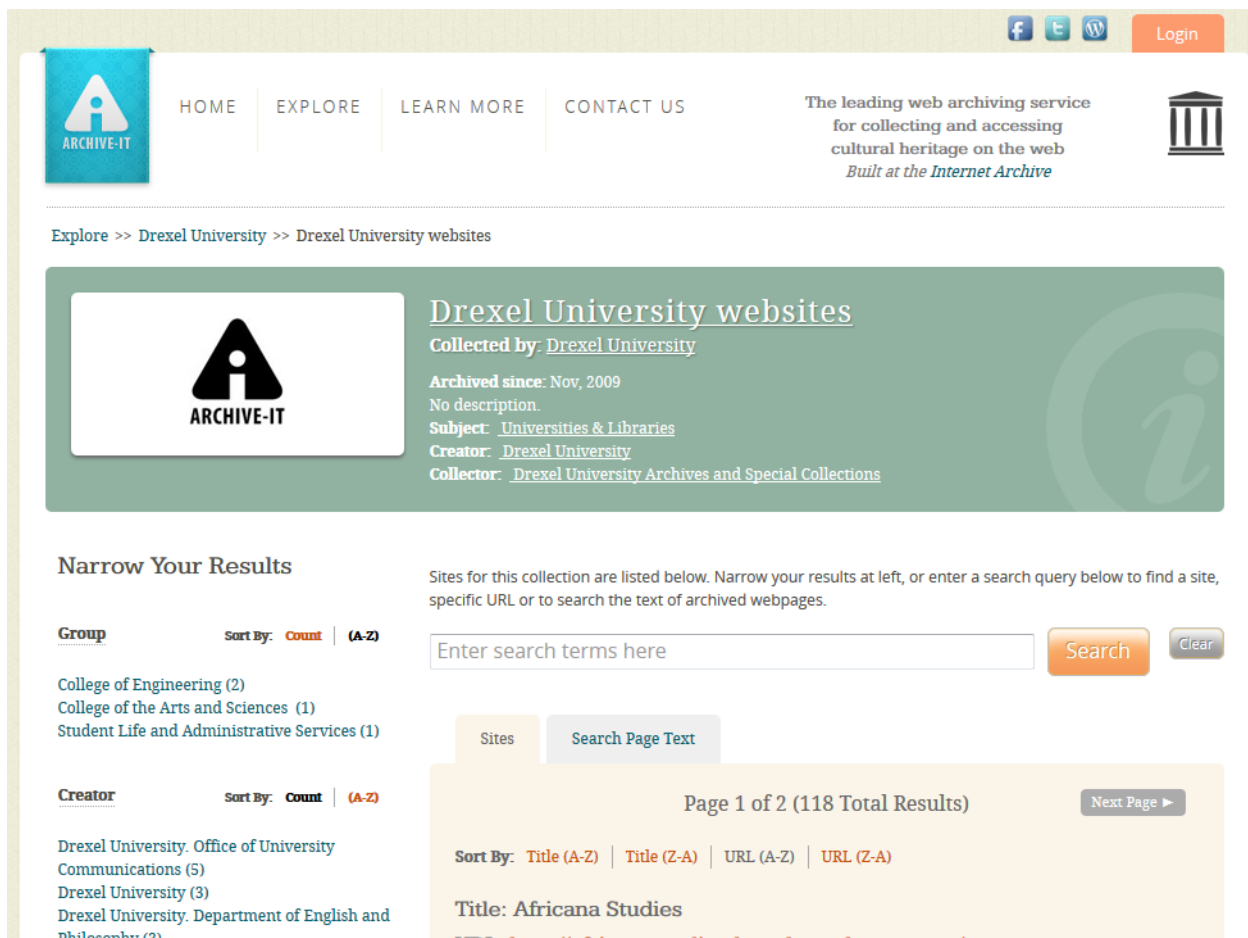
Finally, collection development policies should document access decisions. Important considerations include whether (and when) to provide public access to all archived content (for example, some content may not be appropriate for immediate release) and how to establish relationships between archived Web content and other descriptive information, such as archival finding aids. When deciding its access policy, the Drexel University Archives considered the current accessibility of content on the live Web, staffing resources for providing various access points, and the probable level of use by the designated community at this time. Choices of other institutions will likely differ according to local circumstances.

After developing a collection development policy, Drexel University Archives tackled the more practical matters of determining which websites to crawl, how frequently to crawl them, and how to manage crawl content. The archives reviewed all the seeds that were initially entered into Archive-It in order to determine which seeds should remain active for future crawls and which seeds were no longer active. This process led to the creation (and

ongoing maintenance) of an up-to-date list of seeds that are regularly crawled. Drexel University Archives plans to monitor the environment for new websites and URL changes on the drexel.edu domain on a regular basis.

Access and Metadata

Archive-It offers several methods of access for end users, who can browse and full-text search captured Web content. Each partner has a page on Archive-It's own website (archive-it.org) that provides search and browse access to its public collections (see Figure 1). Partners can also create their own portal pages, embedding a search box for their Archive-It collections. To support more in-depth research, Archive-It's internal administration also gives partners the option to add or enhance descriptive metadata, which can be applied at the collection, seed, and document levels. Each Archive-It capture also includes a banner at the top of the page that indicates who collected the page, which collection it is a part of, when it was captured, a link to metadata about the capture, and a link to all crawled versions of the page (see Figure 2).



The screenshot shows the Archive-It website interface. At the top, there is a navigation bar with links for HOME, EXPLORE, LEARN MORE, and CONTACT US. The Archive-It logo is on the left, and a tagline "The leading web archiving service for collecting and accessing cultural heritage on the web" is on the right. Below the navigation bar, the breadcrumb trail reads "Explore >> Drexel University >> Drexel University websites". The main content area features a large banner for the "Drexel University websites" collection, collected by Drexel University in Nov, 2009. Below the banner, there is a "Narrow Your Results" section with filters for Group and Creator. A search bar is present with a "Search" button and a "Clear" button. The results section shows "Page 1 of 2 (118 Total Results)" and a list of results, with the first result titled "Africana Studies".

Figure 1
Drexel University Archives partner page on Archive-It website

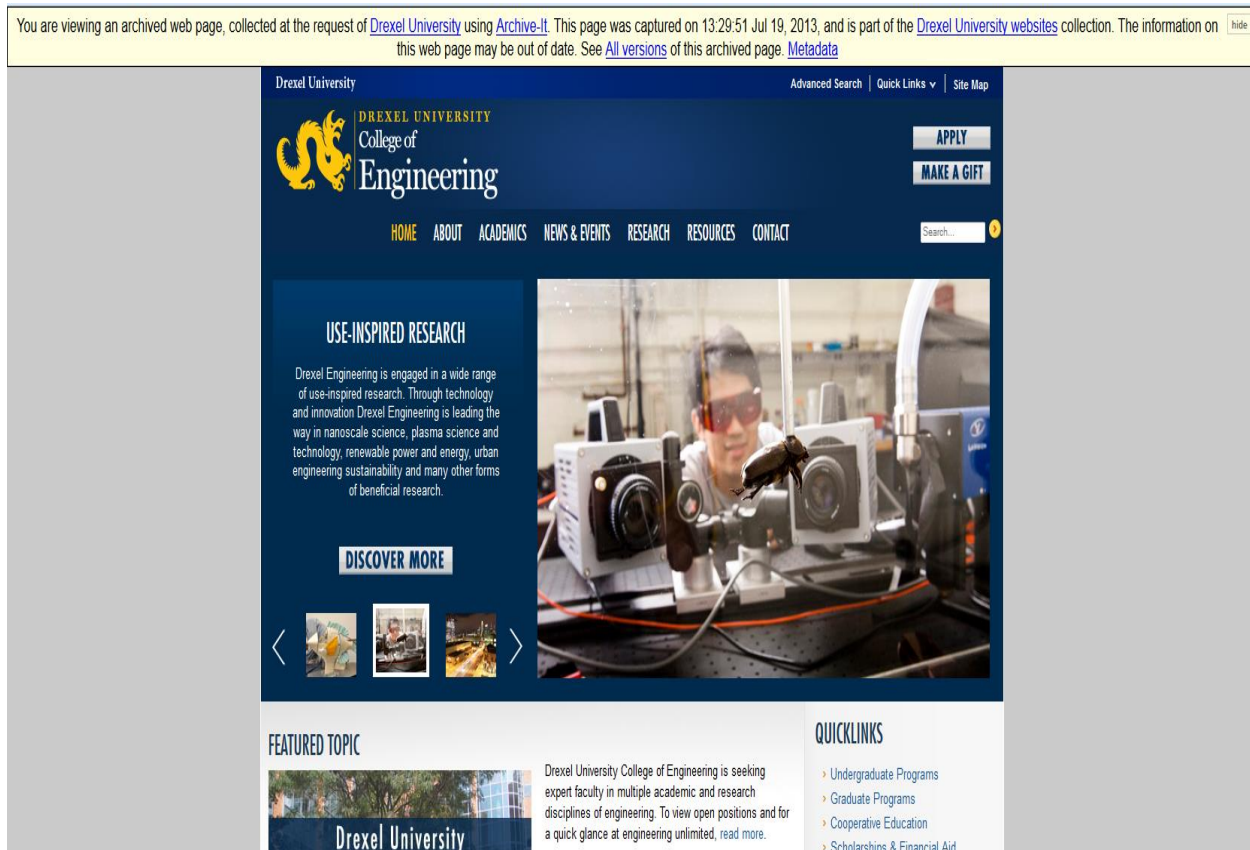


Figure 2

Example of an end user display for a Web page captured by Archive-It. The top banner notifies the user that they are viewing an archived version of the page and provides information about the conditions of capture.

Due to resource constraints, the Drexel University Archives has adopted a streamlined approach to description. Metadata is applied only at the collection and seed levels, and while Archive-It supports 16 standard Dublin Core fields and an unlimited number of custom fields, Drexel chose to use only four: title, creator, description, and collector. This basic set of standardized metadata is targeted at the needs of the user community: Most users are Drexel faculty and staff, and being able to search by the name of the office and the creator would likely be the most useful to the greatest number of researchers. Creator names are standardized to ensure accurate and consistent search results. For the description field, archives staff members pull descriptions from the websites themselves. The final field indicates that the Drexel University Archives is the collector in order to document provenance. In the end, the metadata chosen proves to be useful at creating access by school, department, and college. When combined with Archive-It's search feature, allowing keyword searching across the collections for researchers interested in themes or subjects embedded in the description or title, this simple schema demonstrates that even a minimum of metadata can be useful.

In addition to the search, browse, and metadata features provided by Archive-It, libraries can explore other ways to help users discover and navigate Web archives. Some Archive-It partners create MARC records for their seeds, which are then incorporated into the library catalog. The University of Scranton is currently pilot testing the use of CONTENTdm (the library's digital collections platform) to catalog Web archives at the seed level, with an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) feed pulling those records into the library's central catalog. Other institutions have created finding aids to describe Web archives collections, and ongoing research on visualization tools suggests that new methods of exploring and analyzing Web archives may emerge

(Padia, AlNoamany, & Weigle, 2012). Many Archive-It partners integrate access to their Web archives into their institution's 404 Not Found error page, so that users who seek an inactive URL (or follow a dead or broken link) are referred to the Web archives for access to an archived copy of the page from the past. At Scranton, at the request of the library, public relations staff added brief text onto the university website's 404 error page that links to an Archive-It Web archives portal.

Quality Assurance

Due to the complexity of the Web archiving process, crawl results are frequently imperfect, making quality review an important step for the long-term success of a Web archiving program (Pennock, 2013). Ideally, the implementation of a quality assurance program ensures that the Web content crawled by the institution has been accurately and completely archived so that the capture includes not only HTML content but also the look and feel of the site, as well as embedded documents and audiovisual materials. Like other aspects of Web archiving, approaches for quality assurance (QA) range from time-consuming manual review (whether comprehensive or sampling) to automated reports on the success of a crawl. Archive-It provides automated QA reporting while also facilitating more manual review via the Wayback QA tool. However, it is up to the institution to develop a workflow and process for regularly reviewing and acting upon quality reports.

Drexel's quality control program is quite new, and staff members are still experimenting with various processes to ensure the viability of this important part of the Web archiving program. Currently, archives staff uses an Excel spreadsheet to track errors by seed within each of its quarterly crawls (which contain the most important Web content). After each crawl, a staff member documents which seeds were crawled and then reviews the Archive-It report for basic problems, such as the following yes/no questions: *crawl too large*, *data queued*, and *robots.txt*.

The staff member then looks for seed errors (for example, a notification that a seed had redirected the crawl to a new website) and embedded file problems (such as missing content or failed displays). Once this basic quality control has been completed, the records management archivist makes needed corrections or changes, conducting patch crawls or rerunning crawls to ensure that archived material authentically retains its original look and feel. It takes a single staff member approximately 10 to 20 minutes to perform quality control on each seed. Drexel crawls 43 sites per quarter, two sites per week, six sites on a semi-annual basis, and nine websites annually. In total, to complete quality control on every crawled seed would require between 2,970 and 5,940 hours per year of staff time, the equivalent of more than a single full-time staff member doing only quality control. If staffing is consistent, once a staff member is familiar with crawled websites, quality control time will likely be reduced. To further reduce this time, quality control of weekly crawls can be done on only a monthly basis to spot check for any major changes, while assuming that most informational content is captured on a regular basis. Nonetheless, quality control requires a significant amount of staff time. All updates, corrections, and changes to Archive-It configurations are carefully documented. Archive-It recently launched a new quality control feature that allows administrators to automatically run patch crawls directly from a Wayback review of a crawl, and Drexel will likely take advantage of this feature to streamline its processes, with staff running and tracking patch crawls so that the records management archivist can focus on updates and configuration changes.

Some issues revealed during a quality review cannot be directly addressed through Archive-It's functionality. For example, a site may employ robots.txt to block crawlers, or files intended for capture were missed because they were not embedded. At Drexel University, in such cases the records management archivist contacts the website administrator for the department or website in question and asks for permission to crawl the site or discusses how the content could be made more conducive to Web archiving. While this type of one-to-one quality control is

time-consuming, it increases the likelihood that ongoing Web captures will be as complete as possible and thereby most useful for future researchers.

Challenges

A significant challenge each of our institutions faces in starting and building our Web archiving programs is limited staffing. Web archiving (particularly description and quality control) requires a significant amount of staff time, although some accommodations can be made for limited institutional resources. At the University of Scranton, the library's limited staff time is managed by prioritizing Web archiving tasks. Efforts are concentrated on the front-end (identifying and selecting seeds and then scoping and testing crawls) with only superficial review of crawl results, and no resources have yet been allocated to metadata enhancement or systematic quality assurance, despite the significance of these steps. At Drexel University, full-time archives staff members manage the addition of new seeds, metadata, and correcting quality control issues, while the records management archivist addresses and corrects problems. Several paraprofessional staff members outside the department also help with quality control procedures. To best allocate staff time, Drexel decided that not every error identified during quality control would be corrected. Instead, the records management archivist considers whether an error affects the main content or look and feel of the website. Small errors, such as a missing image or two, are sometimes not corrected or cannot be corrected, and a decision has been made to tolerate these errors. Staff time is, thus, concentrated on ensuring basic usability of all the archived sites for future users, rather than getting bogged down in the small details of one site.

The configuration of crawls (and the correction of any subsequent technical problems) is itself a major challenge for Web archiving initiatives. In some cases, the design and structure of the Web content to be captured present difficulties to standard Web crawling. At Drexel University, electronic publications, especially e-mail newsletters, have been troublesome. In order to capture this content, archives staff members crawl publications pages separately, treating them as RSS feeds in order to ensure that each link off of a publication page is captured. They also worked with Archive-It support staff to carefully expand the scope of their crawls to capture links to e-mail newsletter content. At the University of Scranton, staff members have encountered issues when trying to crawl university content hosted by third-party vendors, whose sites sometimes featured robots.txt blocks or even, as in one case, blocked the IP range of Archive-It's crawler. While hardly insurmountable, resolving these kinds of technical issues requires a reasonably deep and technical understanding of the Web itself, as well as solid knowledge of Archive-It's functionality. While Archive-It's quality assurance reports and patch crawling tools can help with capturing images and JavaScript or re-crawling missing URLs, careful scoping, often employing regular expressions, may be required to ensure complete and accurate capture. Test crawls are very helpful in appropriately configuring crawls but also require detailed examination and technical interpretation. At this time, Drexel University's records management archivist manages technical difficulties in collaboration with Archive-It staff, relying on their expertise to manage crawl scope and correct difficult technical errors.

A final challenge for Web archivists is uncertainty about how future researchers will use Web archives collections. Will users primarily be internal, looking for archived versions of their own institutional or departmental websites? Will outside researchers find the collection useful? What kinds of questions will researchers try to answer with these materials? These kinds of insights would have significant impact on decisions about selection of content, description, and quality review, as well as the prioritization of these steps. The limited metadata that we now create may, in the future, result in the need for heavier reference assistance for users, and the choices we are now making when selecting seeds and scoping crawls may unknowingly exclude materials of high interest to researchers.

Next Steps

At each of our institutions, we plan to continue and expand our Web archiving initiatives. Refinements to current practices are in order. At the University of Scranton and at Slippery Rock University, plans for improvement include drafting a collection development policy, cataloging seeds to better integrate the Web archives into the library's other digital collections, and finding effective and efficient ways to enhance descriptive metadata. At Drexel University, staff members will continue to focus on the challenges inherent in maintaining their quality control program and keeping their seeds up-to-date.

We also envision expansion into new collecting areas. Social media is one area of shared interest, despite the technical and ethical complexities raised (Farrell, 2010). Scranton is currently testing crawls of the university's official YouTube, Facebook, Twitter, and Flickr accounts, since the digital media and communication stored in these accounts provide a rich insight into the life of the university and serve as visual documentation of campus events. Drexel is likewise exploring the possibility of doing regular captures of university social media accounts, either using Archive-It or another service, such as Social Archives, in order to expand the scope of the material the archives can collect about the history of the university.

An additional project for the future is increased outreach to campus stakeholders. At Drexel University and at Slippery Rock University, outreach efforts will include a targeted campaign to website administrators. This kind of outreach will facilitate quality control efforts, ensuring that archives staff is notified of major website changes and encouraging crawler-friendly practices that lead to fewer errors. Drexel University Archives also plans to hold an official launch in order to encourage faculty, staff, and students to make use of the resource. At the University of Scranton, future outreach will include the formation of a university-wide Web archiving task force, designed to engage campus stakeholders in the selection of seeds as well as the development of policies.

Finally, we plan to explore and implement tools and strategies for assessing the success and progress of our Web archiving initiatives. At Drexel University, archives staff plans to develop a set of criteria for the evaluation of its Web archives. Measures of success could include the following: percentage of the *drexel.edu* domain archived; awareness of the resource; and usefulness to university staff, especially website administrators. This last project is likely the farthest in the future but remains an important consideration to ensure that the Web archiving program is useful to end users and a sustainable initiative for our institutions.

References

- Ayala, B. R. (2013). Web archiving bibliography 2013. UNT Digital Library. Retrieved from <http://digital.library.unt.edu/ark:/67531/metadc172362/>
- Ball, A. (2010). Web archiving (version 1.1). Digital Curation Centre. Retrieved from <http://hdl.handle.net/1842/3327>
- Bleicher, A. (2011). A memory of webs past. *IEEE Spectrum*, 48(3), 30-37. doi:10.1109/MSPEC.2011.5719723
- Brown, A., & Thomas, D. (2005) Archiving websites. *Comma*, 2005(1), 1-9. doi:10.3828/coma.2005.1.17
- Brügger, N. (2005). Archiving websites: General considerations and strategies. Centre for Internet Research. Retrieved from http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf
- Chesapeake Digital Preservation Group. (2012). "Link rot" and legal resources on the Web: A 2012 analysis by the Chesapeake Digital Preservation Group. Retrieved from <http://cdm16064.contentdm.oclc.org/cdm/linkrot2012>
- Consultative Committee for Space Data Systems. (2012). Reference model for an Open Archival Information System (OAIS). Retrieved from <http://public.ccsds.org/publications/archive/650x0m2.pdf>
- Davis, R. M. (2010, January). Moving targets: Web preservation and reference management. *Ariadne*, 62. Retrieved from <http://www.ariadne.ac.uk/issue62/davis>
- Day, M. (2003). Preserving the fabric of our lives: A survey of Web preservation initiatives. In T. Koch & I. T. Sølvsberg (Eds.), *Research and advanced technology for digital libraries: Proceedings from ECDL 2003, 7th European Conference, Trondheim, Norway, August 17-22, 2003* (pp. 461-472). Berlin, Germany: Springer-Verlag.
- Farrell, S. (Ed.). (2010). *A guide to Web preservation: Practical advice for Web and records managers based on best practices from the JISC-funded PoWR project*. JISC PoWR. Retrieved from <http://jiscpowr.jiscinvolve.org/wp/files/2010/06/Guide-2010-final.pdf>
- Goldman, R. (2011, October). Internships that go the distance: A how-to-do-it (and how-not-to-do-it) guide to online internships. Paper presented at the meeting of the Mid-Atlantic Regional Archives Conference, Bethlehem, PA. Retrieved from <http://digitalcommons.lasalle.edu/libraryconf/1/>
- Gomes, D., Miranda, J., & Costa, M. (2011). A survey on web archiving initiatives. In *Research and advanced technology for digital libraries: Proceedings from TPDL 2011, International Conference on Theory and Practice of Digital Libraries, Berlin, Germany, September 26-28, 2011* (pp. 408-420). Berlin, Germany: Springer-Verlag.
- International Internet Preservation Consortium. (2012) Tools and software. Retrieved from <http://netpreserve.org/web-archiving/tools-and-software>
- Internet Archive. (2014). Frequently asked questions. Retrieved from <https://archive.org/about/faqs.php#9>
- Kahle, B. (1997). Preserving the Internet. *Scientific American*, 276(3), 82-83.
- Martin, K. E., & Eubank, K. (2007). The North Carolina State Government Web Site Archives: A case study of an American government Web archiving project. *New Review of Hypermedia and Multimedia*, 13(1), 7-26. doi:10.1080/13614560701423638
- Masanès, J. (2005). Web archiving methods and approaches: A comparative study. *Library Trends*, 54(1), 72-90. doi:10.1353/lib.2006.0005
- National Digital Stewardship Alliance. (2012). Web archiving survey report. Retrieved from http://www.digitalpreservation.gov/ndsa/working_groups/documents/ndsa_web_archiving_survey_report_2012.pdf
- Niu, J. (2012a, March/April). An overview of Web archiving. *D-Lib Magazine*, 18(3/4), doi:10.1045/march2012-niu1
- Niu, J. (2012b, March/April). Functionalities of Web archives. *D-Lib Magazine*, 18(3/4), doi:10.1045/march2012-niu2

- Padia, K., AlNoamany, Y., & Weigle, M. C. (2012). Visualizing digital collections at Archive-It. *JCDL '12: Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries* (pp. 15-18). doi:[10.1145/2232817.2232821](https://doi.org/10.1145/2232817.2232821)
- Pearce-Moses, R., & Kaczmarek, J. (2005). An Arizona model for preservation and access of Web documents. *DttP: Documents to the People*, 33(1), 17-24.
- Pennock, M. (2013, March). Web-archiving. *DPC Technology Watch Report*, 13(1). doi:[10.7207/twr13-01](https://doi.org/10.7207/twr13-01)
- SalahEldeen, H. M., & Nelson, M. L. (2012). Losing my revolution: How many resources shared on social media have been lost? In P. Zaphiris (Ed.), *Theory and practice of digital libraries: Second international conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012* (pp. 125-137). doi:[10.1007/978-3-642-33290-6_14](https://doi.org/10.1007/978-3-642-33290-6_14)
- Toyoda, M., & Kitsuregawa, M. (2012). The history of Web archiving. *Proceedings of the IEEE*, 100, 1441-1443. doi:[10.1109/JPROC.2012.2189920](https://doi.org/10.1109/JPROC.2012.2189920)
- Zittrain, J., Albert, K., and Lessig, L. (2014, March 17). Perma: Scoping and addressing the problem of link and reference rot in legal citations. *Harvard Law Review Forum*, 127. Retrieved from <http://harvardlawreview.org/2014/03/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations>