**Web Archives Literature Review**

Sarah Pauley

ISLT 7301: Introduction to Information Technology

December 6, 2020

**Web Archives**

**Introduction**

Online content provides the artifacts future historians and archaeologists will use to write the narrative of our times. Today anyone with a device and Internet access can write a Tweet, post a video on Youtube, or build a website on Weebly. These low barriers allow ordinary people to document their lives and ideas online, providing a wealth of information and perspectives to future historians. But with ease of publication comes ease of alteration. Content creators -- including individuals and organizations -- remove Tweets, revise blog posts, delete accounts, and abandon websites. At best, this is an inconvenience; at worst, it obscures the truth. Web archives can preserve born-digital content and preserve it for consumers, researchers, fact-checkers, and future historians.

Libraries should engage with web archives both by educating patrons on accessing archives and by curating their own collections. Web archives are an excellent way for libraries to partner with their communities and preserve local or institutional history.

To find out more about web archives, I mostly relied on journal articles from the University of Missouri library. I first searched the term “web archives” and read articles related to the topic. As I read more, I began to identify topics in the literature: types of archives, the archival process, and copyright issues, for example. I referred to the bibliographies included in several articles to find additional sources. As I began to develop a schema of trends in web archives, I began to search for sources to fill in the gaps in these trends.

In this literature review, I will address four themes in the literature: the purpose of web archives, the selection criteria used when creating collections, the archival process, and legal and ethical challenges.

**Purpose of Web Archives**

Online content provides the artifacts future historians and archaeologists will use to write the narrative of our times. Milligan (2016) warns, “If we do not come to grasps with web archives, the histories that we write will be fundamentally flawed. Imagine a history of the late 1990s or early 2000s that draws primarily on print newspapers, ignoring the revolution in communications technology that fundamentally affected how people share, interact, and leave historical traces behind” (p. 80). Likewise, Finnemann (2018) writes that “the web today has become a main resource for externalised human memory weather as individual memories or as an array of shared memories… of the twenty-first century cannot be written without these archives” (p. 52). Web archives exist to preserve these “traces” of human action and memories and to ensure that “materials are still there in the same unmodified form tomorrow” (Finnemann 2018 p. 51).

According to Khan and Rahman (2019), in 2001, the average lifespan of a web page was 1,132 days; only 10 percent of web pages from that time period are accessible today (p. 71). Only six years after the World Wide Web was released to the public, Kuny (1997) warned of the possibility of a “digital dark age” in which “we are moving into an era where much of what we know today, much of what is coded and written electronically, will be lost forever” (p. 1). Nearly two decades later, Whitt (2016) noted that “as our reliance on data grows even more pervasive in every sector, massive technology and market trends—such as born-digital content, cloud computing, “big data,” and the Internet of Things—will only accelerate the scale and scope of the problem [of archiving digital content]” (p. 117).

Web content is often lost when information is updated or sites and their host institutions become defunct. In extreme cases, a hostile government may seize and destroy information. In March 2006, the Turkish government seized the opposition newspaper *Zaman* and deleted 27 years worth of digital archives (Davis 2016). Even during peaceful transitions of power, information on government websites is updated and older versions can be lost. For instance, in 2008, the National Archives and Records Administration announced that it would not archive agency websites during the transition from the Bush administration to the Obama administration; thus, the End of Term Web Archive was created to preserve at-risk administration websites (Seneca et. al. 2012). Economic factors may also come into play: Davis (2016) discusses the removal of the *Milwaukee Journal Sentinel* archives from the Google News Archive -- and its subsequent inaccessibility to the public -- due to a change in publishing contracts. Tangentially, organizations -- and therefore their websites -- may cease to exist. Pittman (2018) mentions the Labor Unions and Organization (U.S.) Collection at NYU, noting that “the sites of labor unions and organizations seem particularly vulnerable to this kind of erasure as groups disband due to successfully achieving their goals, being shut down, or running out of funds” (p. 56).

**Selection Criteria**

The Internet contains a vast amount of information -- far too much for any individual or institution to preserve single-handedly. Thus, institutions must be selective in choosing which artifacts to preserve and institutions must collaborate to preserve as much of the Internet as possible. Finnemann (2018) describes three selection strategies for web archives: canon and topic-centric selection, domain-centric selection, and time-centric (also known as event-centric) collections. Similarly, Khan and Rahman (2019) list three categories -- site-centric, topic-centric, and domain-centric -- but their definitions differ slightly from those offered by Finnemann.

Topic-centric collections, for both Finnemann (2018) and Khan and Rahman (2019), are primarily focused on the needs of future scholarly research. Finnemann alternately refers to topic-centric archives as “canon and topic-centric,” because selection is based on the quality or authority of the authors or publishers. Information is typically collected from websites associated with a particular field of knowledge (such as government sites or sites related to literature). Khan and Rahman offer several examples of topic-centric archives: the Archipol archive of Dutch political websites, the Digital Archive for Chinese Studies, and the French Elections Web archive (p. 73). Data collection may be either time-limited (if the collection is focused on a specific event, such as an election) or ongoing (if the collection is focused on the development of a topic over time). Finnemann points out that because this approach emphasizes quality and accuracy of information, it requires more human curating than other approaches.

Domain-centric or site-centric archives collect information from a particular website or list of websites. Khan and Rahman (2019) distinguish between site-centric archives, which focus on preserving multiple version of one particular website, and domain-centric archives, which cover multiple, or even all, websites within a top-level domain (.com, .edu, .org, etc) or second-level domain (.pk, .fr., etc) (p. 73) Finnemann (2018) offers a broader definition that encompasses both small collections of a few websites and massive national projects, defining domain-centric collections as “departing from a specific list of web domain-addresses” (p. 54). This strategy is often used to harvest information for national archives. The End of Term Archive (Seneca et. al 2012) is another example of a domain-centric collection, since it required archives to harvest information from government domains.

The third approach to web-archives is time- or event-centric collections. There is some overlap between topic-centric and event-centric collections, but event-centric collections tend to emphasize multiple in-the-moment perspectives of a current event, rather than scholarly research. According to Finnemann (2018), event-centric archives “collect materials related to unexpected or predictable events resulting in the creation of new pages or the appearance of materials related to the event on unexpected sites somewhere on the web” (p. 57). The Internet Archive pioneered the event-centric collection with its September 11 archive. Since 2001, event-centric archives have documented campaigns, protest movements, and health crises, among other events. Event-centric collections include the digital archives of the Occupy Wall Street movement and the Documenting Ferguson Project (Velte 2018); collections documenting Hurricane Sandy in 2012, the Boston Marathon Bombing in 2013, and the Ebola virus outbreak in 2014 (Nwala, Weigle, and Nelson 2019); and even more recently, the National Library of Medicine Global Health Events’ Covid-19 collection (Speaker and Moffatt 2020).

While it is tempting to want to preserve everything on the web, it is important for institutions to determine the scope of their web archive. Even in the 1990s, when the size of the Web was much smaller than it is today, Kuny (1997) recognized the importance of individual institutions limiting the scope of their archives, writing that “any given library will necessarily be required to select resources that they can archive and preserve according to their particular mandates and user requirements” (p. 9). Whitt (2016) notes that while collaboration is necessary in web preservation, it is a challenge because individuals and organizations often regard collaboration as antithetical to leadership and because of the free-rider problem -- organizations assume someone else is taking the lead in preserving a particular collection. However, collaboration does occur. For example, Seneca et.al (2012) describe how, beginning in 2008, multiple organizations worked together to select, manage, and store the End of Term archive.

While no one institution can save the entire Internet, many institutions -- each with their own focused collection -- can preserve a substantial amount. Outside of national archives, the trend seems to be for institutions to develop small event- or topic-centric collections. In some cases, this is undertaken at the local level, by public schools and libraries. Pittman (2018) asserts that “institutions of every size should consider having a web archiving program in place” (p. 53). Meanwhile, Freeman (2016) discusses the Archive-It K-12 Web Archiving Program, in which K-12 students learned the archival process and created their own collection of sites that were meaningful to their own lives. Through this program, students created over 200 web archive collections, preserving over 4000 URLs.

**The Archival Process**

There are several methods for capturing web resources: browser, authoring system, and web crawler (Khan and Rahman 2018). Additionally, in some cases -- especially with event-centric collections, archivists solicit information directly from content creators (Velte 2018). However, throughout the literature, the web crawler method seems to be the most prevalent. This method involves starting from a list of websites (“seeds”) and using a program (“crawler”) to capture information on those pages, and, if preferred, to follow hyperlinks included on those pages and capture information on the linked pages. Pittman (2018) addresses the pros and cons of several software-as-a-service and open-source crawlers, including Archive-It and Webrecorder, while Khan and Rahman (2019) describe three more crawlers: Heritrix, HTTrack, and Wget.

Web crawlers are preferable to capturing content via browser because crawlers can capture the behavior structure of websites in addition to the visible content. This is a crucial distinction because as Kuny (1997) notes, “documents are becoming complex, dynamic creations made of multiple objects, embedded programming and hypertext links” (p. 6). While web content includes plain text, visual content, and multimedia content -- essentially all forms of content available on media other than the Internet, web structure includes the appearance of a web page as well as the behavior -- the links and interactive elements that distinguish the Internet from previous media (Khan and Rahman 2019). Whitt (2016) similarly frames this challenge as a question of migration versus emulation, where emulation means preserving “the original form and functionality” (p. 183). While preserving form and behavior would be ideal, it is not always necessary or feasible. Both Pittman (2018) and Maches and Christensen (2020) note that Archive-It’s web crawler is better at capturing static content than interactive content. Meanwhile, both Whitt (2018) and Khan and Rahman (2019) advise web archivists to prioritize which elements are actually crucial to a collection.

As discussed earlier, “seeds” are the starting points for a web crawl, frequently used in domain-centric archives. However, with event-centric collections increasingly using social media sources, Nwala et al (2019) explored methods of gathering seeds from social media. Social media sources can be found by “scraping” search engine results pages (SERPS), but Nwala et al experimented with finding “micro collections” directly on social media platforms such as Twitter, Reddit, and Scoop.It, and using those collections as seeds to find additional resources on the event. They found that while sources extracted directly from SERPs tend to be more relevant to the search topic, micro-collections tend to yield sources that a) are not found in the top ten results pages, b) are older (dating back to the commencement of an event), and c) are more diverse (including more non-HTML URIs). While some archivists might need to expand the scope of a crawl to identify more relevant sources, Matches and Christenson (2020) point out that “a lack of careful scoping can result in capturing large amounts of unwanted data” and explain that their organization limited the scope of most crawls to the immediate domain URL (p. 11).

The archival process does not end once resources have been collected and preserved on a platform such as Archive-It. Rather, web archiving, like any kind of curation, is an ongoing process and an institution should continually fine-tune their collection and management processes, while improving the public’s access to the archives. Matches and Christensen (2020) share the UC San Diego Library’s process of creating clear web archiving policy and workflows. They indicated that although an institution may use an automated crawler, there is still a need for human curation; the library’s workflow plan includes refining the scope of a crawl to ensure it meets the achieve’s goals and conducting quality control on web crawls to check for missing content.

Along similar lines, it is also valuable for archivists and researchers to evaluate the effectiveness of larger networks of archives. AlSum et al (2014) evaluated the effectiveness of Memento Aggregator, which gives researchers the opportunity to search multiple archives simultaneously. They created an algorithm to search Memento more efficiently and eliminate redundancy in results. Meanwhile, Fernando, Merenzi, and Nejdl (2018) designed ArchiveWeb, a platform for searching and saving resources from archives. This platform was also built with collaboration in mind: it allows users to make copies of and edit collections, either individually or as groups.

While much attention has been paid to the collection process, relatively few sources explore the accessibility of web archives for end users -- whether academics or the public at large. In a heuristic analysis of the Archive-It platform, Abrams et al (2019) found that the Archive-It interface has several areas for improvement. For example, participants in their study were often confused by the search bar because it was not clear what information was actually being searched -- full text or metadata. While Abrams et al suggest immediate solutions, including help features and clearer metadata, Whitt (2018) argues that archivists should focus on the preservation process rather than on the final outcome or use of the archive. Taking the long view, he asserts that because today’s archivists do not know how tomorrow’s researchers will use the archives, the current goal should be to establish processes to preserve as much information as possible in as accessible a way as possible (p. 192).

**Legal and Ethical Concerns**

Because web archives collect, copy, and provide access to vast amounts of born-digital information, archivists face legal and ethical challenges. The most frequent challenges are copyright violations and privacy concerns.

Although the literature contains little mention of web archives actually being sued for copyright law violations, the web archive community has long been concerned about potential conflict. Kuny (1997) states that “increasingly restrictive intellectual property and licensing regimes will ensure that many materials never make it into library collections for preservation” (p. 3). Whitt (2016), takes a more optimistic approach: he suggests that rights holders and archivists recognize their symbiotic relationship. Web archivists, he writes, should repeat a new mantra: “no value in copyright without effective preservation” (p. 209).

There have been copyright lawsuits associated with web archives, but they have not substantially impacted preservation methods. This is especially salient as more social media posts are incorporated into web archives. In Field v. Google (2006), Field, a content creator sued Google for caching documents on his website. However, the court found that Google complied with the fair use doctrine because the cached version served a different purpose from Field’s original publication. While the original publication was intended to entertain, the cached -- or archived -- version enables users to access otherwise inaccessible content and detect changes to content over time. Much more recently, in 2020, the Internet Archive was sued by four major publishers after creating a “National Emergency Library” that included unlimited access to books published by these companies (Dwyer 2020). The preservation of born-digital content was not actually at issue here, though, because the books in question were also published in print, with many copies already existing.

Although the literature does not show evidence of wide-scale clashes between web archives and copyright law, creators and rights holders have asked archives to remove content from public collections. One potential way to find balance between preservation and creator’s rights is to preserve documents, but not make them publicly available. Davis (2016) explains that a “dead man’s switch” would “automatically preserve content should something dire happen to the original archive” (p. 47). Similarly, Velte (2018), who interviewed three archivists who incorporate activist social media accounts into their collections, relays how some archivists have solved this problem: they preserve the documents, but then a) do not open the data to the public or b) provide on-site-only access.

Privacy concerns are another challenge. Milligan (2016) ponders the ethics of using personal sites of living individuals for research when the users a) most likely do not know that their web pages were archived and b) have no practical way to opt out, since GeoCities only exists in archived form now. In some cases, privacy is a matter not only of preference, but of safety. According to Velte (2018), archives have been “used by state agencies to conduct surveillance against vulnerable groups” (2018). One way archivists in Velte’s study have mitigated this concern is by proactively collaborating with activist groups, not only during times of tension, but before activism turns into conflict.

**Conclusion**

In the face of an ever-changing media landscape, web archives provide an indispensable service: preserving a record of those changes. Web archives are valuable both to present communities and to future historians. However, web preservation is not as simple as printing off copies of born-digital content: there are challenges archivists must take into consideration. These include the interactive nature of the Internet, the impossibility of a single institution saving the whole Internet, the usability of user platforms, and copyright and privacy concerns.

The next steps for libraries and information agencies depend on their current level of engagement with web archives. Small institutions, such as public and school libraries, who do not yet engage with web archives should begin by exploring existing collections. They should also consider creating small collections to preserve local or institutional history. Larger institutions that already maintain web collections should focus on collation and access. Web preservation is best achieved when different institutions develop their own niche; thus, institutions should collaborate and communicate to determine who is preserving what content. Institutions, especially those with ample resources, should also collaborate to improve the accessibility of user interfaces. Finally, libraries of all types and sizes should partner with their communities to promote public participation in the use and creation of web archives. Like the Internet itself, web archives should have low barriers of entry.

Works Cited

Abrams, S., Antracoli, A., Appel, R., Caust-Ellenbogen, C., Denison, S., Duncan, S., & Ramsay, S. (2019). Sowing the seeds for more usable web archives: A usability study of Archive-It. *American Archivist*, 82(2), 440–469.

AlSum, A., Weigle, M., Nelson, M., & Sompel, H. (2014). Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries*, 14(3/4), 149–166.

Davis, S. J. (2016). Disappearing news archives. *Online Searcher*, 40(6), 46.

Dwyer, C. (2020, June 3). Publishers sue Internet Archive for 'mass copyright infringement.' *NPR.* https://www.npr.org/2020/06/03/868861704/publishers-sue-internet-archive-for-mass-copyright-infringement

Fernando, Z. T., Marenzi, I., & Nejdl, W. (2018). ArchiveWeb: Collaboratively extending and exploring web archive collections—How would you like to work with your collections? *International Journal on Digital Libraries,* 19(1), 39–55.

Field v. Google Inc., 2006 WL 242465 (D. Nev. 2006).

Finnemann, N. O. (2019). Web Archive. *Knowledge Organization*, 46(1), 47–70.

Freeman, J. (2016). Seen but not heard: A case study of K-12 web archiving and the importance of student participation in the archives. *Archival Issues*, 37(2), 23.

Khan, M., & Rahman, A. U. (2019). A systematic approach towards web preservation. *Information Technology & Libraries*, 38(1), 71–90.

Kuny, T. (1997). A digital dark ages? Challenges in the preservation of electronic information. *63RD IFLA Council and General Conference.*

Maches, T., & Christensen, M. (2020). Web archiving: Policy and practice. *Journal of Digital Media Management*, 8(3), 201-214.

Milligan, I. (2016). Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives. *International Journal of Humanities & Arts Computing: A Journal of Digital Humanities*, 10(1), 78–94.

Nwala, A., Weigle, M., & Nelson, M. (2019). Using micro-collections in social media to generate seeds for web archive collections. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Digital Libraries (JCDL), 2019 ACM/IEEE Joint Conference on, JCDL*, 251–260.

Pittman, B. (2018). Why archive the web? *Online Searcher*, 42(6), 63–66.

Seneca, T., Grotke, A., Hartman, C. N., & Carpenter, K. (2012). It takes a village to save the web: The end of term web archive. *DttP: Documents to the People*, 40(1), 16–24.

Speaker, S. L., & Moffatt, C. (2020). The National Library of Medicine Global Health Events web archive, coronavirus disease (COVID-19) pandemic collecting. *Journal of the Medical Library Association*, 108(4), 656–662.

Velte, A. (2018). Ethical Challenges and Current Practices in Activist Social Media Archives. *American Archivist*, 81(1), 112–134.

Whitt, R. S. (2016). *Through a glass, darkly: Technical, policy, and financial actions to avert the coming digital dark ages.* Santa Clara High Technology Law Journal, 33(2), 117–229.