Jaffa Panken. Tracking Seeds and Crawls for Archive-It Web Archives: A Search for Best Practices. A Master's Paper for the M.S. in I.S degree. April, 2018. 45 pages. Advisor: Helen Tibbo

As web archives grow larger, institutions using Archive-It must keep track of a growing number of seeds and crawls. Managing this data often requires outside tools to create records of quality assurance efforts, scoping guidelines, and records for future colleagues and researchers to contextualize the archived websites. In an exploratory study of tracking systems for web archives, over twenty web archivists responded to a Qualtrics survey about their tracking practices as well as the reasons behind those practices. The survey revealed that only half the participants currently track seeds and crawls outside of Archive-It. Those who do track often rely upon spreadsheets, particularly for quality assurance and designing scoping guidelines. After reviewing the affordances of spreadsheets in light of participants' stated priorities for tracking, the study suggests alternative practices for tracking seeds and crawls. This study is a crucial first step towards establishing best practices for documentation of web archives.

Headings:

Web archives

Web archives -- Best practices

Archive-It

Digital preservation

TRACKING SEEDS AND CRAWLS FOR ARCHIVE-IT WEB ARCHIVES: A SEARCH FOR BEST PRACTICES

by
Jaffa Panken

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2018

Approved by

_____

Helen Tibbo

Table of Contents

Tracking Seeds and Crawls for Archive-It Web Archives 2
1. Introduction 2
2. Literature Review 6
   2.1. Digital History 7
   2.2. Creating Web Archives 8
   2.3. Search and Access 10
   2.4. Theory and Methodology 12
   2.5. Researchers' Use of Web Archives 14
3. Methods 17
   3.1. Survey 17
   3.2. Recruitment 18
   3.3. Participants 19
4. Findings 20
5. Discussion 29
   5.1. Reasons for Tracking 29
   5.2. Tools for Tracking 33
6. Conclusion 38
7. Bibliography 41

1. Introduction

Over the past twenty years, the Internet has changed the way we connect with each other and the world around us. When Pew Research Council began conducting surveys on Internet use in 2000, almost fifty percent of American adults surfed the Web. With nearly 90% of American adults using the Internet in 2016, it has become an indispensable tool for navigating modern society (Pew Research Center, 2017). As a result, the history of the early 21st Century has been largely written online. Web sites, like many born-digital materials, do not currently have the lifespan of paper documents—let alone cuneiform tablets. In fact, Internet entrepreneur and activist Brewster Kahle recently estimated that the average website lasts 92 days (PBS News Hour, 2017).

Kahle recognized the ephemeral nature of the Internet in 1996 when he founded the Internet Archive, a non-profit digital library that provides free public access to collections of digitized materials, including websites, software applications, games, music, movies, videos, moving images, and nearly three million public-domain books. The Internet Archive's collection of cached websites contains over 305 billion individual web pages (Internet Archive, 2017). Users can browse past iterations of web sites by URL or search "archived web sites" in the Internet Archive's search bar. These sites are captured by Heritrix web crawlers, bots that start with a single URL, or 'seed', and follow the links on each successive web page—often spanning several web sites. The bot records the source code

of each web page in a file format called WARC. Both Heritrix crawler software and the WARC file format—a successor to their ARC format—were developed by the Internet Archive. The Internet Archive's Wayback Machine software replays the WARC files to mimic the look-and-feel of the original web page. Users can browse archived websites, clicking on links and viewing images much like they would have the original website.

A growing number of archives and special collections subscribe to Archive-It, a service that allows the institution to specify seeds that they would like to preserve as part of their collections. The institution not only selects seeds, but determines how often the crawler captures the webpages, rules for the crawler to follow, and performs quality assurance on the results. Those collections are then maintained remotely by the Internet Archive and accessed through archive.org. Although there are other tools available for web archiving, digital archivists in the United States commonly rely on Archive-It. Unlike the National Archives in the United Kingdom, France, Denmark, and other European countries, the National Archives and Records Administration of the United States does not crawl all the websites within the country's domain. In the absence of a national initiative to archive the web, U.S. archives and special collections will continue using Archive-It for the foreseeable future.

As web archives grow larger, institutions using Archive-It must keep track of a growing number of seeds and crawls. Without some sort of tracking mechanism, it would be impossible to conduct quality assurance. Quality assurance (QA) is the process by which an archivist checks the captured pages to ensure that all crucial elements—content, formatting, style, images, other media—are displaying properly as well as taking up the correct amount of data for the material. If a video doesn't load or a certain seed is

taking up more than its share of data, it is up to the archivist to troubleshoot until the problem is resolved. To properly conduct QA on the average web crawl, the archivist must rely on the reports provided by Archive-It as well as information about prior crawls. For collections of a certain size, QA is impossible without additional information created during previous sessions. For example, an archivist solved a problem in a prior crawl by changing the scoping for that host. A colleague might have difficulty understanding the origins of those scoping guidelines if the original archivist didn't document their actions. Tracking seeds and crawls is, therefore, integral to QA work and the creation of web archives.

In addition, such documentation may answer eventual questions of provenance from researchers. Social scientists have begun using these archived websites as source material for scholarly research. Though web archives represent a tremendous resource on modern society, relatively few scholars have published research based on web archives (Lin et al, 2017; Belovari, 2017). The nature of historical study means that a number of years must have elapsed before historians start considering a given era as "historical." Historians in the U.S. and Canada began studying the 1960s during the 1980s, a temporal distance of twenty years (Lin et al, pp. 2-3). Given that the Internet Archive's earliest websites date back to the mid-1990s, historians following this pattern will soon turn to web archives for research. It is, therefore, imperative that web archives provide documentation of how the web archive was created so that scholars can perform source criticism.

To support the creation of web archives as well as future researchers, archivists must develop strategies to manage growing numbers of seeds and crawls. So as not to

alienate institutions with limited staff and financial resources, these strategies require the use of existing tools common to archival institutions. This project explores whether archival institutions with Archive-It web archives are tracking seeds and crawls, what tools they are currently using, why they are using those tools, and what features they consider important for tracking the basic components of web archives. It is a crucial step towards establishing best practices for documentation of web archives.

2. Literature Review

Web archiving began as part of a push by memory institutions toward digital preservation during the 1980s and 1990s. Kuny raised the specter of a "Digital Dark Age" in which the rapid technological development of ephemeral digital materials outpaced societal investment in preservation of those materials, leaving no trace of our society for the historical record (Kuny, 1998). With the rise of the Internet, web archives have become a crucial tool for preserving digital heritage and making it available for a variety of users. For researchers, both current and future, web archives will provide necessary sources for understanding modern society.

Histories of web archiving recount the technologies used for crawling websites, differing selection practices, and justify the continued development of web archives (Brown, 2008; Brügger, 2011; Webster, 2017). Despite the fact that web archives have existed since the mid 1990s, the value of web archives has not translated into use by historical researchers. Jane Winters suggests that "The most significant barrier to working with web archives is, quite simply, that it is difficult; it requires skills that many historians do not have, and in the short term may be unwilling to learn." (Winters, 2017, p. 174) While the historical profession does not have a promising track-record when it comes to using technology in their research, the field is progressing towards digital literacy.

2.1.Digital History

In a 2004 follow-up to a 1981 study of historians' information seeking practices, Margaret Stieg Dalton and Laurie Charnigo studied the transformative effect of electronic resources on historian's information seeking behavior. They found that historians were slow in adopting digital resources, but had made progress in using catalogs and indexes to find primary and secondary sources (Dalton and Charnigo, 2004). In 2013, Alexandra Chassanoff conducted a research study on historians' search practices and use of digitized primary sources. Through an online survey completed by 86 academic historians, Chassanoff determined that the relationship between historians and archivists is changing. Given historians use of resources like Google searches as well as finding aids, archivists must remain flexible in making both online and in-person assistance available. In addition, historians desire information about the digitization process to understand the selection and creation of digital sources. Research into historians' use of born-digital collections will be necessary as these collections become more properly historical (Chassanoff, 2013).

The discussion on historical scholarship in the Digital Age among historians has been marked by both excitement and caution. When Roy Rosenzweig sounded the alarm to his fellow historians with his 2003 article, "Scarcity or Abundance? Preserving the Past in a Digital Era". Despite some consideration of a Digital Dark Age, Rosenzweig concluded that historians are more likely to drown in abundant materials. Part of his reasoning stemmed from the Internet Archive's prolific collections, which "Most historians will not be interested [in] now, but in twenty-five or fifty years they will delight in searching it." (Rosenzweig, 2003, p. 751) Rosenzweig marveled at the Internet

Archives' untapped potential, but pointed out some of its pitfalls. In particular, he worried that these valuable archival resources remain in private hands—at least in the U.S. Finally, Rosenzweig pointed out the need for historians to involve themselves in the debates on digital preservation that archivists and librarians have been having for two decades. Their voices are crucial to the allocation of resources that will support historical scholarship into the future.

In an online discussion published in *The Journal of American History* in September 2008, historian Daniel J. Cohen defined digital history as "an approach to examining and representing the past that works with the new communication technologies of the computer, the Internet network, and software systems." (Cohen et al., 2008, p. 454) He went on to describe two levels of digital history: one encompassing dissemination of scholarship as well as pedagogy and the other as a methodological framework that uses technology "for people to experience, read, and follow an argument about a historical problem." Cohen's definition sparked discussion on whether digital history was a methodology—accessible to all historians—or a field—practiced by historians with specific technological knowledge. This particular debate was bound up with pedagogical and institutional concerns so that a lack of consensus among historians reflected the wide-open possibilities for digital history at the time.

2.2. Creating Web Archives

While historians have debated digital history as a whole, archivists have been focused on building collections of web archives. Case studies at various institutions have demonstrated the issues surrounding selection of websites, generating metadata, quality assurance, and making web archives accessible (Gomes et al, 2006; Slania, 2013;

Antracoli et al, 2014; Duncan, 2015; Pendse, 2016; Heil and Jin, 2017). Although all institutions struggle with implementing a web archiving program, the kind of institution and the tools at their disposal differs depending on their country. Legal deposit laws in the United Kingdom, France (Stirling et al, 2012), Denmark (Nielsen, 2016), and various others have ensured that national libraries crawl their national domain. The British Library also uses a combination of algorithms and human selection to crawl British sites outside of the .uk domain. (Milligan, "Lost in the infinite archives, 2016) Countries including Australia, Singapore, and the United States do not have legal deposit laws for websites, but do have national web archiving efforts (IIPC, 2017). Instead, most American institutions rely on Archive-It, a subscription service from the Internet Archive, to crawl their selected websites and make them available to the public (Bailey et al, 2017, p. 24).

Much has been written about the Internet Archive because it maintains the largest web archive in the world, containing 305 billion webpages and over 30 petabytes of data as of October 2017 (Internet Archive, 2017). In explaining the rationale behind choosing Archive-It to manage web archiving at Slippery Rock University and University of Scranton, Antracoli et al., mentioned its connection to the Internet Archive as well as its use of open-source software and interoperability with DuraCloud, a cloud storage space used as a digital preservation repository (p. 159). Over 450 institutions have similarly chosen Archive-It to conduct their web crawls and provide a user interface through the Wayback Machine (Internet Archive, 2017).

The National Digital Stewardship Alliance (NDSA), a consortium of American institutions that support digital preservation, found that 87% of respondents to their 2016

survey on Web Archiving in the United States subscribed to Archive-It. This most recent report by the NDSA represents the third in a series of surveys that were previously conducted in 2011 and 2013. Another survey is currently underway as of October 2017, with the report scheduled for release in 2018. These surveys are intended "to better understand the landscape of Web archiving activities in the United States by investigating the organizations involved, the history and scope of their Web archiving programs, the types of Web content being preserved, the tools and services being used, access and discovery services being provided, and overall policies related to Web archiving programs." (Bailey et al., p. 4) The survey has seen a steady increase in the number of organizations responding, from 77 in 2011, to 92 in 2013 and 104 in 2016 (p. 5). Academic institutions comprised 62% of the respondents in 2016, reflecting "the popularization of Web archiving as a core collection development and preservation activity within academic institutions." (p. 5)

2.3. Search and Access

Although collection development and web archiving practices have long dominated the literature on web archives, more recently scholars have also begun to consider issues of search and access (Jackson et al, 2016). While the 2016 NDSA Survey mostly dealt with collection development and institutional policies, it did touch upon search and access. The 2016 report noted a continued decline in organizations supporting local search and browse features and item-level access points in favor of reliance upon Archive-It's search and browse interface with access through collection-level access points and finding aids (Bailey et al., pp. 25-6). As web archives become increasingly fit into traditional archival description and discovery methods, the NDSA voiced concern

that these methods do not appreciate the "unique affordances and characteristics of Web archives" and have "the potential to stifle a notable opportunity for creativity and innovation around access and discovery." (p. 27)

As web archiving programs outsource search and browse to Archive-It and the Wayback Machine, researchers have studied these interfaces to understand their uses and limitations. Padia et al examined the means of visualizing Archive-It collections and found that the quality of a search is highly dependent on the curator's use of metadata such as groups and tags. For collections lacking such a conscientious curator, the researchers developed an alternative visualization that provided a heuristics-based categorization that groups undescribed collections by ascribing the various websites to Social Media, News Web Sites, Blogs, or Videos (Padia et al., 2012, p. 17). In the years since, members of the research team have continued to study how users interact with the Wayback Machine through web server logs. AlNoamany et al. established that "most human users come to web archives because they do not find the requested pages on the web." (2014, p. 1) This finding reflects the ephemeral nature of websites and the importance of archiving the web from the perspective of the general public.

The needs of academic researchers, however, are quite different. A historian, a librarian, a specialist in information retrieval, and a software engineer collaborated on an exploratory search interface for humanities scholars and social scientists to access web archives. Explaining their interest in search and access, Jackson et al noted that "temporal browsing," where a user specifies the URL of a website and then "move[s] forward and back in time to examine different captured versions," is of limited utility (Jackson et al., pp. 1-2). This interface depends on the user knowing the URL of the

website they want to browse. However, Jackson et al point out that this search model does not serve researchers who prefer to begin projects with "a high-level overview of what's in a collection and how it was gathered." Neither does it suit the final stages of a research project, where the scholar analyzes specific content to reach a conclusion (p. 2).

To improve upon "temporal browsing," Jackson et al. developed a search engine that allows users to start with a high-level view of the collection and gradually focus in on specific topics, then individual web pages. This prototype was not without its issues, and the team is still working out various issues before moving on to the next phase of research. One particular problem that Jackson et al. encountered concerned scholars' understanding of websites as objects of study. Interpreting the content and value of an archived website is difficult when the researcher doesn't understand how it was collected because of limited technical knowledge. They suggest that researchers must be better informed about the "technical nuances of web crawling." (p. 4)

2.4. Theory and Methodology

Another branch of the literature on web archives holds that understanding the practice of creating web archives is not enough. Instead, researchers must develop theoretical and methodological approaches to studying web archives. This includes contemplating the website as a historical object as well as case studies based on web archives. Niels Brügger provides a theoretical breakdown of the archived website as an object of study by breaking down the fundamental characteristics of the medium. Archived web material, he writes, "is an actively created subjective reconstruction." (Brügger, 2011, p. 32) By that he means that by the time a user identifies an archived website, the organization that captured it has already made a series of decisions that

resulted in its selection, look-and-feel, file types to include, the manner of its preservation, and many other qualities. These decisions result in a reconstruction of the website as it appeared on the live web—a copy that is always deficient (p. 32-33).

The archived website is deficient for many reasons, including "the dynamics of updating" by which failed captures of certain pages within the site are replaced by pages captured by another crawl (p. 34). This results in a Frankenstein's monster of a website that includes asynchronous page elements. In Archive-It, users must pay close attention to the standard header on every archived page that notes the date of capture. An archived website is deficient in other ways, including missing images, sounds, video, or interactive elements. Brügger concludes than an archived website is, therefore, a *version* and not a copy of the original website (p. 34). This has ramifications for source criticism as part of historical methodology. It is imperative that the scholar know as much about the provenance and versioning of archived websites as possible before s/he can analyze the collection.

Before studying the archived web, however, researchers must develop methods of studying the live web. Based on past projects, Schneider and Foot point out three common approaches: 1) Discursive or rhetorical analysis of web content; 2) Structural analysis of website features; and 3) Sociocultural analysis of the web as a site of interaction (Schneider and Foot, 2004, p. 116–17). Schneider and Foot report that while early research on the web were generally user studies, researchers have turned towards methods that "recognize the co-productive nature" of websites that incorporates study of both producers and users as well as their interaction (p. 119).

Schneider and Foot, therefore, conceptualize a fourth method: web sphere analysis. Web sphere analysis involves identifying a set of websites related to a chosen theme and considering the interaction between the producers and users of the website over time. They offer an example of a web sphere analysis that compares websites regarding the 2000 elections in the U.S. with websites about consequent elections (p. 218). Elections are a popular area for studies of websites (Schweitzer, 2005; Xenos and Foot, 2005; Larsson, 2011; Hermans and Vergeer, 2013; Miller, 2014). In addition, many articles about web archives use elections as examples of an "event-based collection." (Ankerson, 2012; Brügger, 2012; Rogers, 2017) Schneider and Foot went on to lead a research team in a study of the linking practices of candidate websites in the 2002 U.S. elections (Foot et al., 2006).

2.5. Researchers' Use of Web Archives

As researchers increasingly turn to the archived web, their theoretical considerations are blending with methodological approaches and the practical considerations of access into an allied literature regarding researchers' use of web archives. Based on existing historiography on the 1970s, Ian Milligan points out that it generally takes thirty years for the present to become history. In 2021, it will have been thirty years since the creation of the first publically accessible website. Soon thereafter historians will have to use the archived web to write history. Milligan argues that it is time for historians to "radically transform their practices." (Milligan, 2016, p. 3.)

Susannne Belovari performed a thought experiment considering the issues faced by a historian working with current web archives in the year 2050. She begins by articulating the current premise of historical methodology of analog sources, "to be able

to describe and explain phenomena, historians have to define and redefine searches, interests, and questions, moving from broad aspects to specifics and back again." (Belovari, 2017, p. 64) The future historian is perplexed by the Wayback Machine's search interface that requires URLs, how would she know a URL from decades ago? She cannot and her efforts essentially fail. Recently, the Internet Archive introduced limited keyword search, but full text searching would require more funding and support than a non-profit organization can reasonably expect. Furthermore, the abundance of content available is daunting. Belovari proposes that historians and archivists work together on establishing appraisal principles and describing the archived websites with records of provenance and useful metadata.

Historian Ian Milligan and Computer Scientist Jeremy Lin have been developing an open-source web archiving platform called Warcbase that allows for temporal browsing and provides several tools for interpreting web archives (Lin et al, 2017). This application of big data technologies to web archiving is a promising avenue for development. However, they assume that the scholar can use command-line interactions and has some familiarity with a high-level programming language. As a result, part of their argument depends upon a revolution in humanities education which they do not address in this article (p. 10). Milligan has previously suggested technological competencies and methodological approaches that must guide historical pedagogy, and is building such a program at the University of Waterloo in Canada (Milligan, 2012).

Developing tools in concert with computer scientists and training the next generation of historian to work with web archives will eventually produce results, but what is the role of archivists in these endeavors? Currently, OCLC Research has a team

of archivists developing best practices for metadata in web archiving. The Web Archiving Metadata Working Group (WAM) has put special emphasis on including provenance information on the background of a given website. After putting out a preliminary article describing their work and circulating a draft report, WAM is currently responding to comments and reworking their recommendations (Dooley et al., 2017).

It is not enough to develop metadata standards for web archiving. If web archives are heavily determined by archivists' decisions, we must understand how and why those decisions are being made. Archivists are the gatekeepers to analog materials, but they are—to a greater extent—creators of web archives. As the creators, they are ultimately responsible for how well a captured webpage represents the original and any measures aimed at eliminating distinctions might well be useful to a future researcher who must evaluate the reliability of the web archive as a source. In these early days of web archives, many collections are slowly accumulating enough seeds and crawls that a single archivist may require a system to keep track of the results and interventions. Yet, the literature and the field has not addressed the reality of this growing issue in web archiving. The NDSA Web Archiving Survey for 2018 does not ask about tools used to track seeds and crawl. This paper will correct that oversight and attempt to set guidelines on creating documentation for internal use in QA and external use by future researchers.

3. Methods

This project concerns the practice of creating web archives, considering the developing complications inherent to collections with growing numbers of seeds and crawls. To continue ensuring the quality and future utility of these valuable resources, archivists must develop systems to track the basic components of web archives. There must first be exploratory studies like this one that reviews what archivists are currently doing and what they need to be doing. It is a step towards establishing best practices for documentation of web archives.

3.1. Survey

This project studied web archiving through the experiences of the people who create web archives. A survey was chosen as the research method for this project to reach a wider spread of participants who could give more information about the "state-of-the-field" rather than the "state-of-the-art." In addition, the topic of tracking archives does not require the sort of in-depth responses that one might generate from interviews or focus groups. Instead, this project is concerned with the tools archivists use to track web archives, a practice which is still developing. It was expected that a portion of the participants would not track seeds and crawls outside of Archive-It at all. An interview with an archivist who does not track would be a short interview indeed. A survey best served the proper accounting of those archivists without tracking systems in place. Furthermore, the questions for this survey were not necessarily conducive to a

conversational interview as the techniques that archivists use to track web archives are straightforward and the systems similarly designed because Archive-It provides uniform information about documents and data.

Thirty-nine archivists working with Archive-It responded to 21 questions, of which twenty-two participants completed the survey. Considering that 104 institutions completed the NDSA survey in 2016, the goal was set at 25 participants. The NDSA had stronger institutional support, name recognition, and collected responses for months longer than this survey. As a result, almost reaching a quarter of the participants for the NDSA survey was quite ambitious. The first part of the survey asked background questions that established the scale of the web archive, the number of staff assigned to web archiving, and the number of seeds. Next, the second part of the survey asked about the tools that the institution uses to manage their seeds and crawls, as well as their reasons for tracking. Finally, third section asked participant to prioritize features that the archivist might require in a tracking system.

3.2. Recruitment

The author began recruitment efforts with a post on the Society of American Archivists' Web Archiving Discussion Group. This yielded several responses, but it became clear that other methods for advertising the study were necessary. After consulting with web historian Ian Milligan by email, the author posted on the Slack board for Archives Unleashed—a professional group that discusses web archiving—and Twitter with the hashtag "#webarchiving." At that point, other archivists began retweeting the post, including Archive-It. These retweets were instrumental in reaching a respectable number of participants. After three weeks, the survey was closed and analysis began.

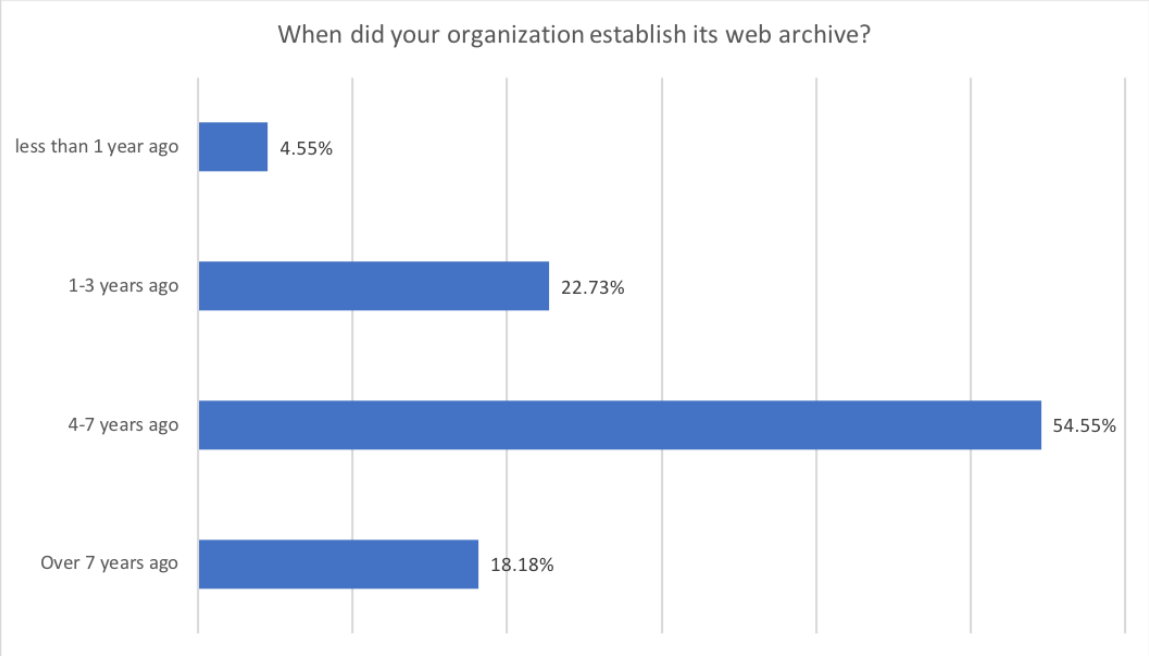
3.3. Participants

The participants for this survey were mainly archivists who were active on social media, particularly Twitter. Since the author's colleague retweeted the survey, a disproportionate number of participants work at North Carolina institutions. This does not necessarily invalidate the findings since they represent collections of different sizes and different priorities. It should be noted that archivists from a wide variety of institutions responded to the survey so that there is huge variation in the number of seeds tracked by participants, ranging from a minimum of 80 to a maximum of 5,000+ seeds. However, it is impossible to know the number of active vs. inactive seeds as well as one-time versus recurring crawling frequencies so that this particular measure has turned out to be irrelevant for determining the influence of seed numbers on tracking preferences. Instead, this variation in the number of seeds is an indication that the survey results are more representative than initially anticipated.

4. Findings

Data from the survey were analyzed according to their format, with quantitative date generally displayed in bar graph visualizations and qualitative data coded by hand. More complicated analysis was not warranted due to the straightforward nature of the questions and the low number of responses. It would be irresponsible to assume that the results are representative enough to support statistical analyses of correlation. Instead, the findings are suitable for an exploratory study indicating possible directions for future studies.

Table 1: Age of represented web archives.



A wide variety of institutions responded to the survey. In addition to private and public colleges and universities, other cultural heritage institutions weighed in on their

tracking practices. Regardless of the type of organization, the vast majority of responses were from web archives that are maintained by a single staff member. This one archivist only spends a small portion of their work time on Archive-It, often assisted by one or two part-time staff members who may be students or volunteers. The age of the web archives represented in the responses indicates that most of these collections (54.5%) were established between 4 and 7 years ago, that is between 2011 and 2014. Nearly 23% were established 1-3 years ago—between 2015 and 2017—and just over 18% were established before 2011. Less than 5% were established in the past year.

Table 2: Organizations that track seeds outside of Archive-It.

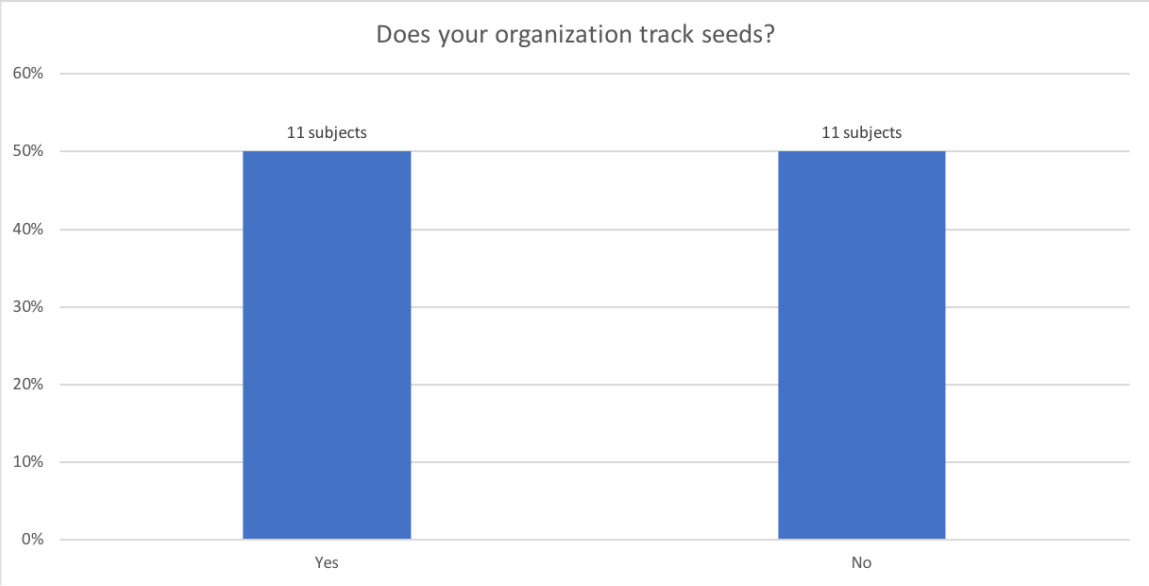
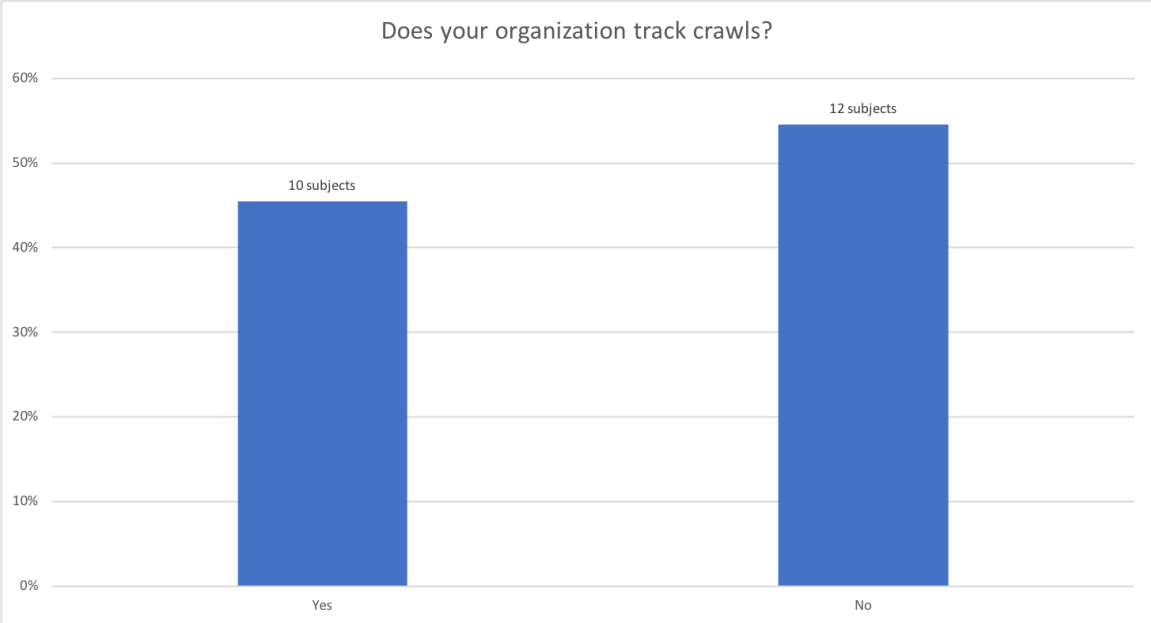


Table 3: Organizations that track crawls outside of Archive-It.



Organizations are slightly more likely to track seeds than crawls outside of Archive-It. Within Archive-It, seeds are listed by URL and ordered according to the date that they were added to the collection. While the list of seeds can be sorted by group, status, frequency, type, or access, they are most often accessed within the crawl report. Tracking seeds is part of the quality assurance process, which different institutions conduct according to the time and staff available. Some institutions rely upon Archive-It for all their tracking needs. When asked "Does your organization track seeds outside of Archive-It?" 50% of participants responded "Yes" and 50% responded "No." There was a slight difference for tracking crawls so that 10 participants captured crawls outside of Archive-It, while 12 participants did not. This means that a single institution tracks

seeds, but not crawls. The rest that do track outside of Archive-It, track both seeds and crawls.

Table 4: Tools used for tracking seeds.

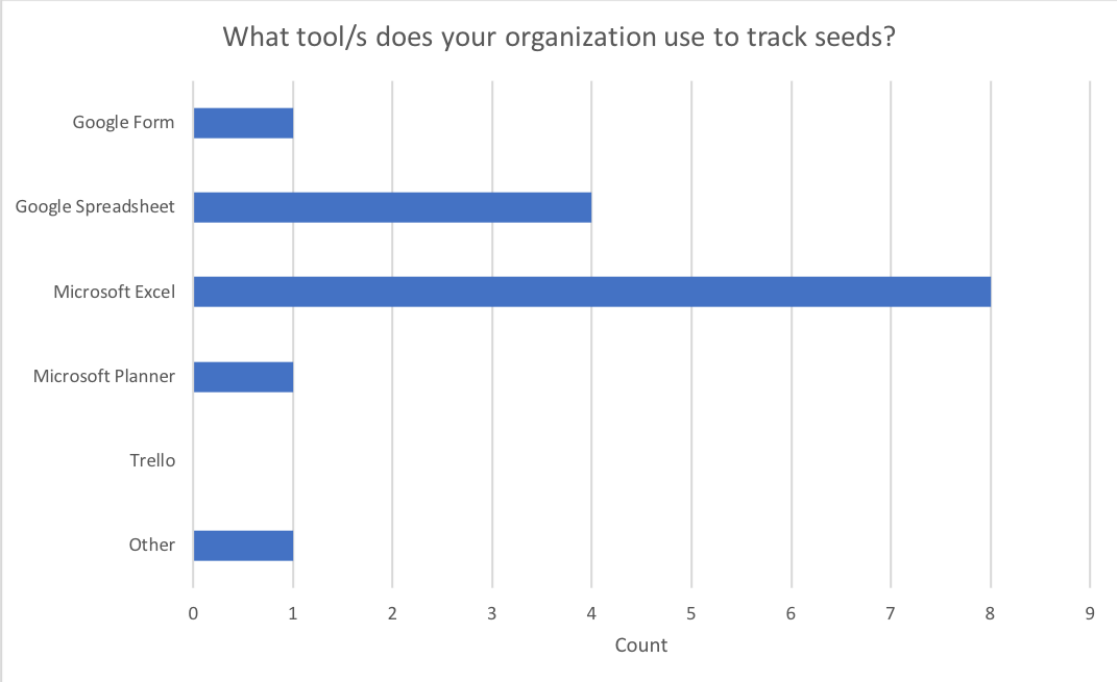
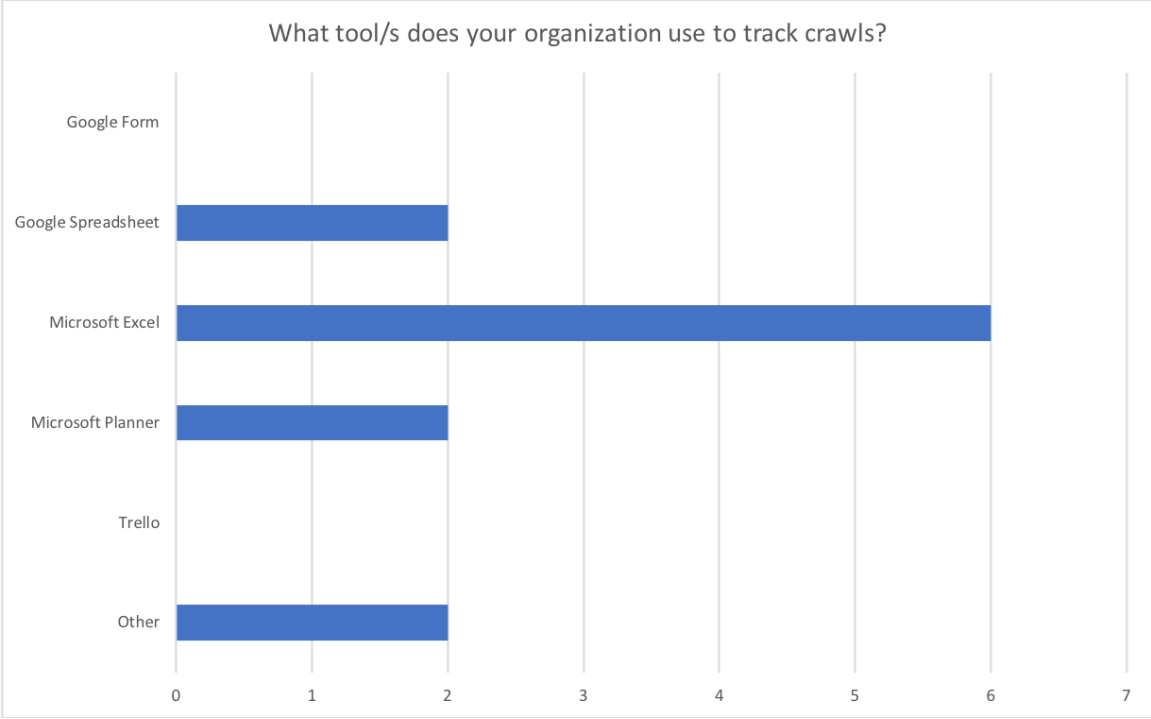


Table 5: Tools used to track crawls.



The survey then asked those that tracked seeds or crawls outside of Archive-It which tools they used. Since participants were able to choose multiple tools, more tools for tracking were indicated than participants who responded. Microsoft Excel was the most popular tool for tracking both seeds (8 participants) and crawls (6 participants), followed by Google Spreadsheet (4 participants for seeds, 2 participants for crawls). All other tools were considerably less popular for tracking seeds. However, two participants used Microsoft Planner for tracking crawls, tying with Google Spreadsheet and Other. For the two institutions that marked "Other," one used a Microsoft Access database for tracking both seeds and crawls, while the other used Microsoft Word to track crawls. No other tools were mentioned.

Table 6: Reasons for tracking seeds.

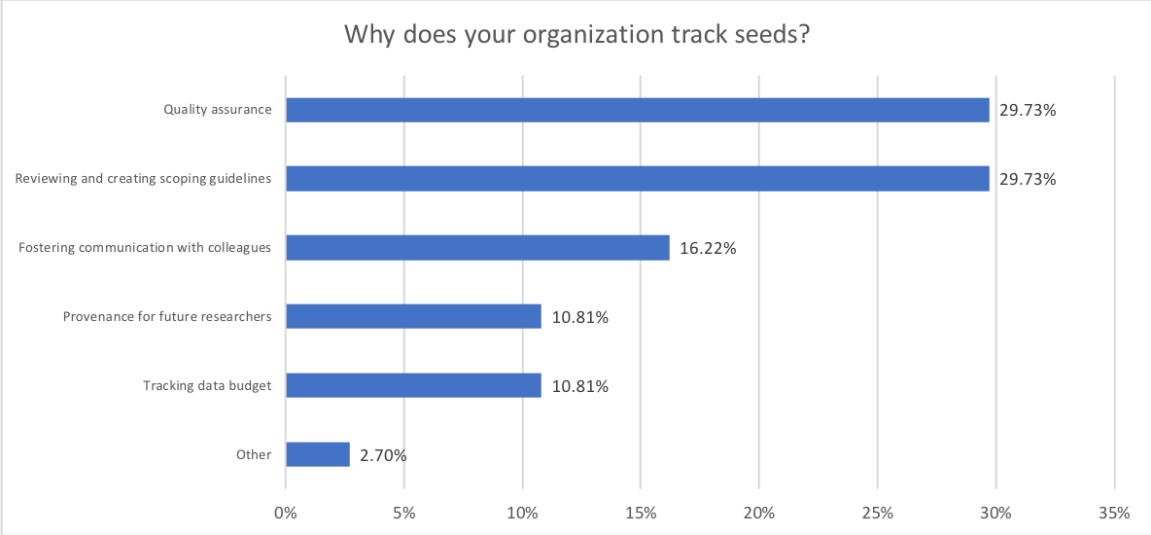
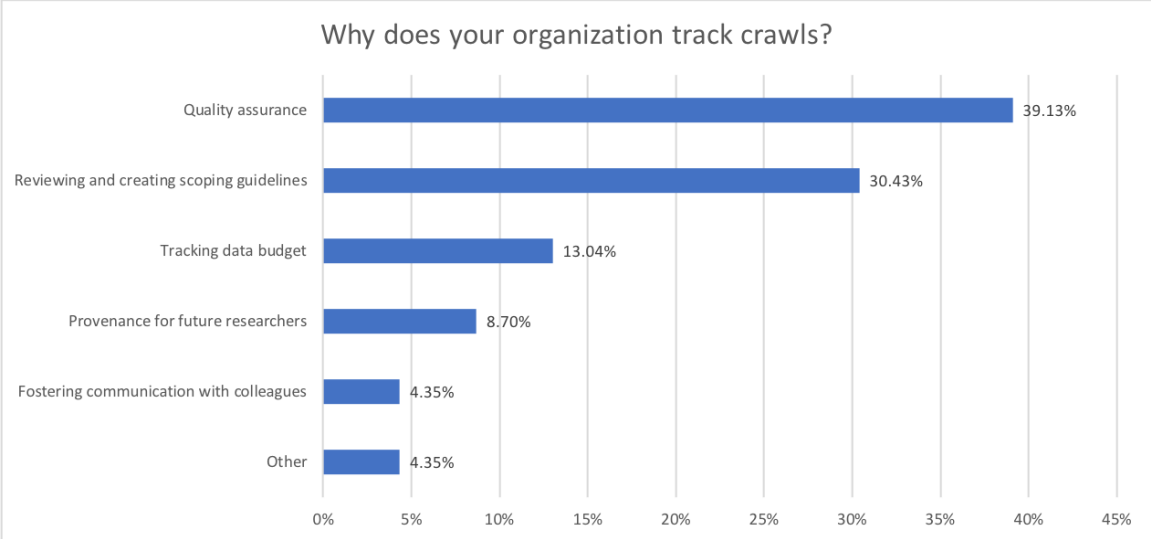


Table 7: Reasons for tracking crawls.



When selecting reasons for tracking seeds and crawls, there were again slight differences. Quality assurance was a major reason for tracking both seeds (29.73%) and crawls (39.13%). An equal percentage of organizations track seeds for reviewing and creating scoping guidelines (29.73%) as for quality assurance. However, reviewing and creating scoping guidelines was a secondary reason for tracking crawls at 30.43%. Fostering communication with colleagues was more important for tracking seeds (16.22%) than for tracking crawls (4.35%). However, tracking the data budget was slightly more associated with tracking crawls (13.04%) than seeds (10.81%).

Provenance for future researchers figured into tracking seeds and crawls figured into approximately 10% of organizations' reasoning. As many organizations considered provenance for future researchers as important as tracking the data budget for tracking seeds. However, slightly more organizations prioritized tracking the data budget over provenance for future researchers when tracking crawls. The only "Other" reason for

tracking seeds and crawls concerned keeping data for annual reports, which figured into tracking practices for a single institution.

Archive-It has a potentially useful feature for downloading seed and crawl reports directly into CSV format. The survey asked participants whether they use this feature and found that less than one-third of organizations used them. Those participants who indicated that they did used Archive-It generated seed lists listed a variety of reasons. The most frequent response indicated that these lists were helpful for identifying seeds that were redirecting due to updated URLs. Other textual responses revealed that few archivists understood the question or were unfamiliar with this feature. In future surveys, a screenshot would clarify where to find the Archive-It generated seed lists.

One of the final questions asked for observations about the organizations' current tracking system. Only two participants were pleased with their current tracking system, while the remaining participants were varying degrees of content with their arrangements. No participant was outright displeased with their tracking system. Their comments explaining their assessments reflect three main issues: lack of resources, disorganization, and lack of interoperability. With regard to the lack of resources, participants noted that there was not enough staff time nor financial resources to make a new system. As one participant put it, "We do not need anything more complicated nor do we have the time or the resources to implement a new system when the current system wo[r]ks well." Other participants were dissatisfied with what one called "a hobbled together process." Participants specifically noted that the spreadsheets were "cumbersome and somewhat error prone." Yet another participant considered this disorganization as par for the

course, “This is pretty messy work, and so it makes sense that tracking would be messy also.”

Finally, a number of participants noted that any current system might not be comprehensible to a co-worker who had not created it. Several participants noted that they had created systems of varying complexities themselves. One of the more complicated tracking systems was “a purpose-built database that allows me to manage and produce reports for all aspects of the web archiving program, including sites, seeds, crawls, QA progress, QA personnel assignments, metadata, and accessioning.” Less complicated systems were simply a series of spreadsheets maintained by a single archivist. Both the creator of the database and managers of spreadsheets agreed that interoperability was a problem, “It does the job but wouldn’t necessarily make sense to another staff person.”

To consider an improved web archiving tracking system, the survey asked participants to prioritize possible features. Participant were given six features—simple user interface, ability to upload seed lists from Archive-It, task tracking, ability to track quality issues for seeds, open source, and generates reports for individual seeds over the course of multiple crawls. While the format of this question was not idea for a simple quantitative analysis, the data did reveal some important results. Archivists were particularly interested in tracking quality issues for seeds, with 70% of participants ranking this feature in the top two. Almost half the participants (40%) prioritized the ability to upload seed lists from Archive-It or existing spreadsheet in the top two, although more than half ranked that feature in the bottom half of the rankings. An open source system was the least important feature, with 80% of participants relegating it to 5th

place or dead last. Participants were most divided on prioritizing the generation of reports for seeds over multiple crawls, as this feature earned either 15% or 20% of the participants for each ranking.

5. Discussion

If these findings indicate anything, it is that not all institutions have reached the point where it has become necessary to track seeds and crawls outside of Archive-It. For some, this lack of need might derive from the age of their web archive as younger or more limited collections may have fewer seeds. A certain number of seeds may be manageable through Archive-It tracking features that make seeds sortable by group, status, frequency, type, and access. However, as collections conduct more crawls and add more seeds, it is reasonable to assume that more robust tracking systems will become necessary. There are two ways that this may be accomplished: 1) At some point, Archive-It may add features that allow for more sophisticated tracking than is currently available; and 2) Collections of a certain size may have to create and transfer their reports into an outside tracking system.

5.1. Reasons for Tracking

In both of these scenarios, studies of current best practices will lead the way to define the parameters of a proposed system for tracking seeds and crawls. According to this limited survey, quality assurance will be a primary driver for development as it was ranked as a major reason for tracking both seeds and crawls. In addition, quality assurance takes an inordinate amount of time and manpower because it requires the archives to inspect the captured page and, on occasion, compare it to the live webpage. Although Archive-It has an add-in called Proxy Mode on Firefox that facilitates the

inspection of a captured webpage by blocking all the live content, actually using Proxy Mode is confusing due to a lack of documentation and often misleading because it sometimes blocks content that actually functions properly. As a result, quality assurance requires the archivist to use multiple tabs and, often, multiple tools to inspect a single webpage.

According to the survey findings, reviewing and creating scoping guidelines is the other top reason why archives have tracking systems. Scoping guidelines are complicated to design because Heritrix captures content from so many different hosts in such a haphazard manner. It can be tempting to solve problems by limiting data or documents, but such a method is like using a blunt weapon where a scalpel is needed. Rather than limiting the capture entirely, it is preferable to prevent the crawler from capturing irrelevant material by excluding certain hosts based upon their URL. For example, URLs with a different domain than the seed are harvested based on their proximity to the original seed. If the crawler is already seven links out from the original seed (let's say, cnn.com) and encounters a link to a different domain (nbcnews.com), then the crawler will likely consider the nbcnew.com link to be out-of-scope. Hosts can be scoped in or blocked by referring to the composition of the URL through a regular expression or using key words that appear within the URL. While the regular expression method requires technical sophistication, it is more exact and less likely to have unintended consequences than using words or phrases. In fact, Archive-It integrates regular expressions that scope out common crawler traps, such as infinite calendars and URLs that repeat themselves ad infinitum. To facilitate useful scoping guidelines, a

system would likely require a means of tracking hosts for a single seed when the circumstance arises.

Despite similar reasons for tracking seeds versus tracking crawls, slightly more effort is required to track seeds outside of Archive-It. For one, Archive-It already has useful tools for comparing crawls to each other (Figure 1). However, it is more difficult to compare seeds within different crawls in the Archive-It interface. Comparing the same seed captured from different crawls is also integral to quality assurance, particularly when troubleshooting unexplained spikes in new data or documents. It is also useful for reviewing and creating scoping guidelines, another highly-ranked reason for tracking seeds and crawls. Any design for a tracking system must further include a feature that allows archivists to track quality issues for seeds within a given crawl, as indicated by the prioritization for possible features.

Figure 1: Archive-It interface for comparing two crawls. Courtesy of the University Archives at the University of North Carolina at Chapel Hill Libraries.

Archivists considered fostering communication with colleagues as a more compelling reason for tracking seeds (16.22%) than crawls (4.35%). There are several ways to read this finding. The nature of tracking seeds is slightly different than tracking crawls because seeds are the building blocks of crawls. Tracking crawls is about the

distribution of data through the seeds, while tracking seeds is about the quality of the capture. First, one must consider with which colleagues the archivist is communicating. Since most archivists are working with Archive-It alone or with another part-time staff member, the most likely reading of "colleague" is actually "future colleague." This understanding is further bolstered by some of the free-response answers in which archivists implied that their current systems would be difficult for a successor to understand. Even if some of the participants were concerned about working with current colleagues, the vast majority of respondents did not weigh "task tracking" highly on their prioritization of possible features for a future tracking system. For these reasons, it follows that archivists are concerned about creating documentation about their web archive collections—specifically the seeds—for their future collaborators or successors.

5.2. Tools for Tracking

If inheritors of web archiving responsibilities are to be considered, it is necessary to reconsider the dominant tool for tracking web archives. According to the survey, spreadsheets—particularly Microsoft Excel, but also Google Spreadsheets—are the most commonly used tools outside of Archive-It. However, spreadsheets are lists of numbers and short text entries, not intuitive or particularly readable on their own. In fact, the use of spreadsheets was explicitly connected with the "messiness" of tracking web archives in multiple comments. If spreadsheets are neither readable nor easily updated, then why do so many archivists rely on spreadsheets for tracking purposes?

One reason for using spreadsheets is that Archive-It seed and host lists are downloadable as CSV files for every collection and crawl. These downloadable lists make it simple to keep track of how much new data and documents were captured in each

crawl. Another downloadable spreadsheet contains similar information about the hosts for every crawl, with additional information about blocked, queued, and out-of-scope hosts. While this same information is available on Archive-It, the CSV files normalize the data so that it is measured in bytes rather than a mix of megabytes, kilobytes, and gigabytes. This would presumably allow archivists to visualize the data with software like Tableau Public.

The survey revealed, however, that these downloadable CSV files are rarely used either because archivists don't know that they exist or don't have the time to make use of them. Even if an archivist did have some elusive free time, it's not clear that these lists would be of any more use than strategic sorting and filtering of the lists within the Archive-It interface. However, the fact that this information is made available in spreadsheets rather than in data visualizations speaks to the way that archivists store copious amounts of data. Spreadsheets offer some level of security that the data about seeds and crawls can be kept locally by the organization rather than relying upon Archive-It.

The question remains, however, how useable this data is to current web archivists, their successors, and future researchers. Consider the use of spreadsheets for the purposes of quality assurance and scoping guidelines—the top two reasons why archivists use these tools. Quality assurance requires the archivists to inspect each webpage to ensure that the look-and-feel of the original website has been captured, including videos, images, fonts, styling, and formatting. Links should be navigable as long as the content is relevant to the collection, a judgement call based on knowledge of the collection and its purpose. When Heritrix has deemed a certain link "out-of-scope," it

is the decision of a bot and not a human. Quality assurance serves as a human check on the web crawler's programming so that archived webpages are intelligible (i.e. stylistically sound) and useful to researchers—a much more difficult property to gauge.

For our purposes, let's agree that a webpage is useful when it is relevant to the collection of which it is a part. For an event-based collection about the 2016 U.S. Presidential Election, a webpage about the infamous Access Hollywood tape of Donald Trump and Billy Bush would be in-scope. A retrospective on Billy Bush's Access Hollywood career would be out-of-scope. Heritrix might not be able to differentiate between the two, but an archivist worth her salt would not prioritize the capture of "Bush's Best Broadcasts."

There *are* technical aspects of quality assurance where an archivist, tipped off by irregularities in the webpage presentation, examines the levels of new data and scrutinizes the host reports for blocked URLs. That work is routinely done directly from the Archive-It interface rather than a spreadsheet, though an archivist might keep track of her findings in a spreadsheet if she is reviewing many seeds. More likely, the archivist will act immediately by running a patch crawl to pick up missing documents, decide that the issue requires a scoping adjustment, or—most likely—chalk it up to the imperfect nature of web archiving. None of these possibilities would be uniquely served by the affordances of a spreadsheet.

When evaluating a periodic crawl such as a monthly or quarterly crawl, a narrative report or logbook on problematic seeds and proposed actions is often more helpful than a spreadsheet with docs and bytes. Each crawl entry should note the crawl start date, id number, problematic seeds, any immediate actions taken, results, and

longer-term observations on the scope of the crawl. This report should be stored on a network drive so that it is accessible to colleagues and can be used for training purposes. It is important that archivists review recent entries in the report prior to performing quality assurance on a new crawl so that continuing issues can be addressed.

Scoping guidelines require more experimentation and creativity to perfect. While scoping rules exist at both the collection and seed levels, most scoping is done at the seed level. This is mostly due to the unforeseen consequences of making blanket rules for many different websites. Creating and reviewing scope is a nuanced process that often requires multiple tests to perfect. The eventual result, a series of rules that includes desirable content and excludes irrelevant or unnecessary content, can become less effective over time. When scoping guidelines outlive their usefulness, it is important for the archivist to understand why and when those rules were created in the first place. Instead of creating a running report about all the scoping decisions or generating a document for each seed, there is a convenient utility within the Archive-It interface that could be used for scoping changes, including frequency and type. Each seed has a notes section that allows archivists to explain the changes that were made and the reasons behind those decisions. Each of these notes is timestamped and signed with the archivist's login.

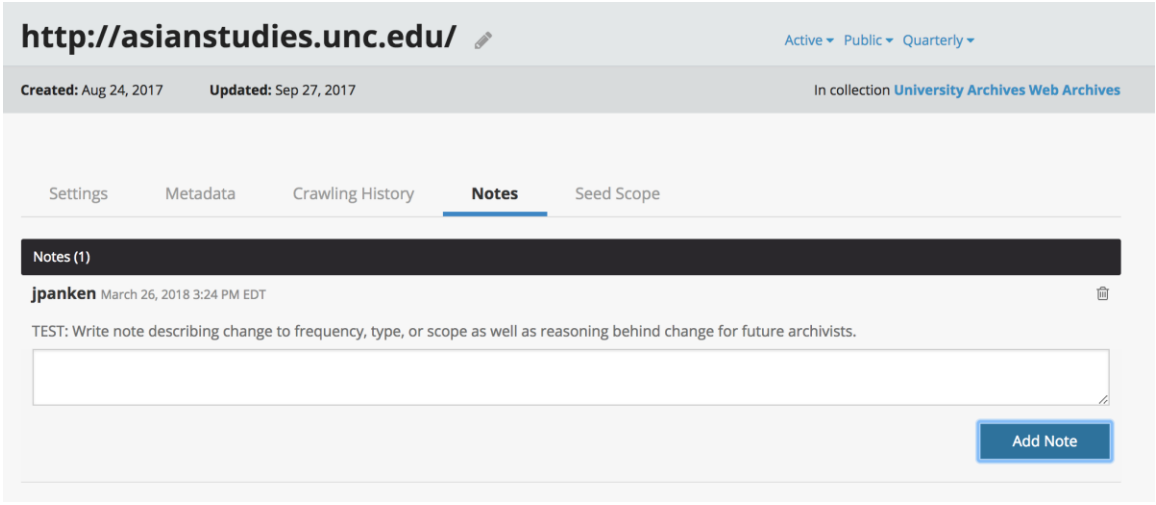


Figure 2: Sample note to future archivists within Archive-It interface for a selected seed regarding changes to frequency, type, or scope

The notes feature will allow archivists to track their own decisions as well as those of their predecessors. In time, it may become a useful resource for researchers attempting to reconstruct the provenance of the collection.

6. Conclusion

This paper considered the current tracking practices at archival institutions with Archive-It web archives and found that about half of the represented collections do not track seeds and crawls outside of Archive-It at all. The small sample size makes it difficult to say whether this finding is related to the number of seeds in the collection, whether these archives do not have the staff resources for tracking, or if the archivists do not believe that outside tracking is necessary. Establishing a baseline understanding of tracking practices, however, is a valuable exercise in understanding the current documentation about web archives as well as the possibilities for future documentation.

In light of the survey findings, this paper makes three recommendations for web archivists to consider adopting:

1) Archive-It generated seed lists and crawl reports should be compiled in spreadsheets for export into data visualization software such as Tableau Public.
2) Keep narrative on crawls to track actions and their effects.
3) Use Archive-It seed notes to track changes to frequency, type, or scoping rules for future web archivists. Make these notes available as provenance for future researchers.

These recommendations emanate from a consideration of the reasoning behind the adoption of certain tools, as much as from the affordances of the tools themselves. In certain cases, the reasoning for using the tool did not match up with the affordances of the tool, in particular the use of spreadsheets.

While spreadsheets are valuable for organizing large amount of information, they are not necessarily intuitive for use in quality assurance. To address that weakness, data visualization features in Microsoft Excel or data visualization such as Tableau Public can help archivists understand the distribution of data within the seeds. As for the use of spreadsheets for creating and reviewing scoping guidelines, it's unclear how spreadsheets assist that process. Instead, archivists could consider writing narrative reports about crawls, indicating the start date, id number, problematic seeds, any immediate actions taken, results, and longer-term observations on the scope of the crawl. This information would help future colleagues and researchers understand how and why a collection was scoped.

Since narrative reports would be laborious for tracking individual seeds over time, the final recommendation concerns communication about changes to frequency, type, or scoping rules. Archive-It has an underutilized feature that creates notes about each seed within the interface. The notes feature would allow archivists to record the process of creating and reviewing scoping procedures without relying upon spreadsheets with long text entries. Archivists could then make these notes available to the researcher interested in the provenance of a certain seed. The main purpose, however, would be communication between current archivists and their successors.

These three recommendations begin to address the practical issues surrounding the creation of web archives as well as emerging concerns about provenance for future researchers. By reducing reliance upon spreadsheets, archivists may be able to interact with the data on seeds and crawls through intuitive data visualizations. Narrative reports and the Archive-It notes section will allow future archivists and researchers to reconstruct

the decisions made by past archivists as well as their reasoning. Such documentation will allow for proper contextualization of web archival collections, rather than creating useless documentation for its own sake.

7. Bibliography

AlNoamany, Y., AlSum, A., Weigle, M. C., & Nelson, M. L. (2014). Who and what links to the Internet Archive. *International Journal on Digital Libraries*, *14*(3–4).

Ankerson, M. S. (2012). Writing web histories with an eye on the analog past. *New Media & Society*, *14*(3), 384–400.

Antracoli, A., Duckworth, S., Silva, J., & Yarmey, K. (2014). Capture all the URLs: First steps in web archiving. *Pennsylvania Libraries: Research & Practice*, *2*(2), 155–170.

Bailey, J., Grotke, A., McCain, E., Moffatt, C., & Taylor, N. (2017). *Web Archiving in the United States: A 2016 Survey*.

Belovari, S. (2017). Historians and web archives. *Archivaria*, *83*(1), 59–79.

Brown, A. (2006). *Archiving websites: A practical guide for information management professionals*. Facet: London.

Brügger, N. (2011). Web Archiving - between past, present, and future. In *The Handbook of Internet Studies* (pp. 24–42). Oxford, UK: Wiley-Blackwell.

Brügger, N. (2012). When the present web is later the past: Web historiography, digital history, and Internet studies. *Historical Social Research*, *37*(4), 102–117.

Chassanoff, A. (2013). Historians and the use of primary source materials in the digital age. *The American Archivist*, *76*(2), 458–480.

Cohen, D. J., Frisch, M., Gallagher, P., Mintz, S., Sword, K., Taylor, A. M., Turkel, W. J. (2008). Interchange: The promise of digital history. *Journal of American History*, *95*(2), 442–451.

Dalton, M. S., & Charnigo, L. (2004). Historians and their information sources. *College & Research Libraries*, *65*(5), 400–425.

Dooley, J. M., Farrell, K. S., Kim, T., & Venlet, J. (2017). Developing Web Archiving Metadata Best Practices to Meet User Needs. *Journal of Western Archives*, *8*(2). Retrieved from http://digitalcommons.usu.edu/westernarchives

Duncan, S. (2015). Preserving born-digital catalogues raisonnés: Web archiving at the New York Art Resources Consortium (NYARC). *Art Libraries Journal*, *40*(2), 50–55.

Foot, K., Schneider, S. M., Dougherty, M., Xenos, M., & Larsen, E. (2006). Analyzing linking practices: Candidate sites in the 2002 US electoral web sphere. *Journal of Computer-Mediated Communication*, *8*(4).

Gomes, D., Freitas, S., & Silva, M. (2006). Design and Selection Criteria for a National Web Archive. In *10th European Conference on Digital Libraries*.

Hale, S. A., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R., & Margetts, H. (2014). Mapping the UK webspace. In *Proceedings of the 2014 ACM conference on Web science - WebSci '14*.

Heil, J. M., & Jin, S. (2017). Preserving seeds of knowledge: A web archiving case study. *Information Management*, *51*(3), 20–24.

Bibliography Continued

Hermans, L., & Vergeer, M. (2013). Personalization in e-campaigning: A cross-national comparison of personalization strategies used on candidate websites of 17 countries in EP elections 2009. *New Media & Society*, *15*(1), 72–92.

International Internet Preservation Consortium. (2017). Legal deposit. Retrieved October 15, 2017, from http://netpreserve.org/web-archiving/legal-deposit/

Internet Archive. (2017). About the Internet Archive. Retrieved October 16, 2017, from https://archive.org/about/

Jackson, A., Lin, J., Milligan, I., & Ruest, N. (2016). Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. Newark, New Jersey.

Kuny, T. (1998). The digital Dark Ages? Challenges in the preservation of electronic information. *International Preservation News*, *17*(May), 8–13.

Larsson, A. O. (2011). ”Extended infomercials” or ”Politics 2.0”? A study of Swedish political party Web sites before, during and after the 2010 election. *First Monday*, *16*(4).

Lin, J., Milligan, I., Wiebe, J., & Zhou, A. (2017). Warcbase: Scalable analytics infrastructure for exploring web archives. *Journal on Computing and Cultural Heritage*, *10*(4), 1–30.

Miller, B. (2014). Accentuating the Queer: An examination of how LGBT websites framed the 2012 presidential election. *Electronic News*, *8*(4), 260–273.

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

Milligan, I. (2012). Mining the “Internet Graveyard”: Rethinking the historians’ toolkit. *Journal of the Canadian Historical Association*, *23*(2), 21. http://doi.org/10.7202/1015788ar

Nielsen, J. (2016). *Using web archives in research*. Aarhus, Denmark: NetLab.

Padia, K., AlNoamany, Y., & Weigle, M. C. (2012). Visualizing Digital Collections At Archive-It. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries - JCDL ’12* (pp. 15–18). New York, New York, USA: ACM Press.

PBS NewsHour. (2017). Internet history is fragile. This archive is making sure it doesn’t disappear. Retrieved September 15, 2017, from https://www.pbs.org/newshour/show/internet-history-fragile-archive-making-sure-doesnt-disappear

Pendse, L. R. (2016). Collecting and preserving the Ukraine conflict (2014-2015): A web archive at University of California, Berkeley. *Collection Building*, *35*(3), 64–72.

Pew Research Center. (2017). Internet/broadband fact sheet. Retrieved September 15, 2017, from http://www.pewinternet.org/fact-sheet/internet-broadband/

Rogers, R. (2017). Doing Web history with the Internet Archive: screencast documentaries. *Internet Histories*, *1*(1–2), 160–172.

Rosenzweig, R. (2003). Scarcity or abundance? Preserving the past in a digital era. *The American Historical Review*, *108*(3), 735–762.

Bibliography Continued

Slania, H. (2013). Online art ephemera: Web archiving at the National Museum of Women in the Arts. *Art Documentation: Journal of the Art Libraries Society of North America*, *32*(1), 112–126.

Schneider, S. M., & Foot, K. A. (2004). The web as an object of study. *New Media & Society*, *6*(1), 114–122.

Schweitzer, E. J. (2005). Election campaigning online: German party websites in the 2002 national elections. *European Journal of Communication*, *20*(3), 327–351.

Stirling, P., Chevallier, P., & Illien, G. (2012). Web archives for researchers: Representations, expectations and potential uses. *D-Lib Magazine*, *18*(3–4).

Webster, P. (2017). Users, technologies, organisations: Towards a cultural history of world web archiving. In *Web 25: histories from the first 25 years of the World Wide Web*, Niels Brügger (Ed.). Peter Lang.

Winters, J. (2017). Breaking in to the mainstream: demonstrating the value of internet (and web) histories. *Internet Histories*, *1*(1–2), 173–179.

Xenos, M. A., & Foot, K. A. (2005). Politics as usual, or politics unusual? Position taking and dialogue on campaign websites in the 2002 U.S. elections. *Journal of Communication*, *55*(1), 169–185.