Abstract:

The paper provides a review of published literature on the collection and development of Web archives, focusing specifically on the theories, techniques, tools, and approaches used to appraise Web-based materials for inclusion in collections. Facing an enormous amount of Web-based materials, archival institutions and other cultural heritage institutions need to devise methods to actively select Webpages for preservation, creating Web archives that constitute a cultural record of the Web for the benefit of users. This review outlines the challenges of collecting and appraising Web-based materials, places the theories and activities of collecting Web-based materials within the broader discourse of archival appraisal, and points out directions for future research and critical discourse for Web archives.

Keywords: appraisal | web archives | web archiving

Article:

*****Note: Full text of article below**

Colin Post*

Building a Living, Breathing Archive: A Review of Appraisal Theories and Approaches for Web Archives

DOI 10.1515/pdtc-2016-0031

Abstract: The paper provides a review of published literature on the collection and development of Web archives, focusing specifically on the theories, techniques, tools, and approaches used to appraise Web-based materials for inclusion in collections. Facing an enormous amount of Web-based materials, archival institutions and other cultural heritage institutions need to devise methods to actively select Webpages for preservation, creating Web archives that constitute a cultural record of the Web for the benefit of users. This review outlines the challenges of collecting and appraising Web-based materials, places the theories and activities of collecting Web-based materials within the broader discourse of archival appraisal, and points out directions for future research and critical discourse for Web archives.

Keywords: appraisal, web archives, web archiving

From the beginnings of the World Wide Web in the early 1990s to the present, the Web has increasingly become a locus of cultural production of artistic, scientific, political, and social importance. The Web is truly global, and access to the Internet continues to grow with the ongoing development of information communication technology infrastructure. An overwhelming amount of information is generated, exchanged, presented, and stored on the Web in the form of personal, organizational, and governmental Websites. Older media forms like newspapers and research journals have been reinvented on the Web, and entirely new forms of communication like social media continue to develop. Without any doubt, the Web constitutes a significant portion of the world's documentary heritage.

In recognition of this fact, many information institutions have developed Web archiving programs, collecting Web-based materials in addition to papers, records, and other kinds of archival collections. Information professionals have made efforts to archive the Web for almost as long as the Web itself has been around. The Internet Archive (IA), a non-profit organization dedicated to archiving the Web and host of the largest collection of Web-based materials, has been active in these efforts since 1996 (Internet Archive). However, given the vast size and constantly changing nature of the Web, the task of archiving significant web-based materials cannot be left to a single organization, and a variety of information institutions have followed suit, with numerous national libraries, university archives, and government organizations developing Web archives.

Even for the IA, the Web simply cannot be comprehensively archived—nor would this necessarily be a worthwhile undertaking. If everything from the past is saved, it becomes close to impossible to actually find significant materials. Seeking to understand the present moment, researchers will need to scour blogs, e-mails, online news stories, and organizational Webpages in Web archives, as well as more traditional archival materials like letters, photographs, and meeting minutes. To make this kind of research possible, archives need to appraise materials, determining the value of potential archival materials and strategically deploying scarce resources like storage space and staff time in order to effectively build strong collections. The problem of access may be partially assuaged as machine learning and natural language processing techniques continue to develop and improve researchers' abilities to search across large corpora of materials, and the development of these technologies should certainly influence the appraisal criteria used in building Web archival collections. However, other factors, such as storage space, staff time, and institutional resources, require that appraisal decisions be made, making it vital that archivists, librarians, and other information professionals critically consider the appraisal theories and approaches undergirding their efforts to collect the parts of the Web that will prove valuable to present and future users.

Although a variety of institutions and information professionals undertake the collection and appraisal of

*Corresponding author: Colin Post, School of Information and Library Science, University of North Carolina at Chapel Hill, 216 Lenoir Drive, Chapel Hill, NC 27599, E-Mail: ccolin@live.unc.edu

Web-based materials, I will focus this review specifically on the archival profession. Appraisal is not a new area of practice or discourse for archives, and archivists have developed a rich theoretical tradition to guide the crucial activity of selectively developing archival collections (Cook). While it is difficult for archivists to measure the success of their appraisal activity, archivists are well-practiced in sorting through an overflow of analog and digital documents to identify the sliver to be preserved and passed on. Despite this body of theoretical discourse and practical experience, the appraisal of Web-based materials presents new challenges to archivists. Web archiving remains a relatively novel endeavor for many archivists, and the profession is in the process of evaluating this existing body of appraisal theory and practice to determine how much of this applies to the appraisal of Web-based materials, and where new approaches and practices need to be employed. In this review, I investigate the range of approaches to appraising and collecting materials for Web archives, consider how these emerging approaches fit within the broader body of archival appraisal theory and practice, and suggest directions for further research and critical discourse for Web archives.

1 Web Archiving Difficulties and Obstacles

To better understand why the appraisal of Web-based materials necessitates new approaches and methods, it will be useful to review the unique challenges that Web archivists face, and how the overall practice of Web archiving differs from archiving analog and other kinds of digital materials. In the first place, there is no easy way to define a "document," or base unit for what is being collected with Web-based materials (Pearce-Moses and Kaczmarek). Analog documents have clear boundaries, like the front and back cover of a book or the collection of pages that make up a letter. Stand-alone digital documents can also be distinguished as discrete files, each with a specific name and file size. Websites, on the other hand, contain countless Webpages, all of which are embedded in the broader context of the Web (Masanès). By its nature, the Web is an interconnected medium with Webpages hyperlinked to other Webpages located at different host domains. Removed from this context, a single website loses a great deal of its value. Of course, building Web archives requires that archivists and librarians selectively cull materials for preservation, necessarily obscuring some of that original context. Part of the challenge for archivists, then, is to decide how much material to collect in order to provide the appropriate context for archived Webpages as part of an interconnected medium, while not committing to a Sisyphean task of preserving the entire, dynamic body of the Web.

Archivists cannot rectify this difficulty by merely collecting the whole of the Web. The sheer volume of the Web frustrates any attempt to collect the entire context in which a particular website is embedded, and so librarians and archivists need to selectively assert some boundaries on their Web archival collections based on their own institutional collecting goals and the needs of their users (Hsieh, Murray, and Hartman). However, the immense size of the Web is not the only difficulty in asserting these boundaries; Web content is also dynamic, with content edited or altered over time often without trace of what was replaced. It is also ephemeral, with particular pages or entire sites liable to be removed or deleted without warning (Dougherty and Meyer; Masanès). This is especially true of social media content, which exists in a constant state of development, directed by individual users as well as the platform providers (Fansler, Gilbertson, and Petersen; Rollason-Cass and Reed). Web archivists not only need to decide what is being collected, but also how frequently to collect selected Websites.

Once archivists have decided what Websites they are collecting and how frequently they are collecting these sites, they must also contend with technical difficulties in capturing certain kinds of Web content. The overall nature of Web content has changed dramatically from the earliest days of HTML text documents to the present moment characterized by slickly designed sites filled with interactive features and streaming media (Dougherty and Meyer). To keep up with the changing nature of the Web, archivists need to update capture tools, skills, and approaches (Summer and Punzalan). Many automated Web archiving tools struggle to fully capture Websites (Duncan and Blumenthal; Masanès), often failing to incorporate streaming media and non-HTML files like Javascript and CSS files into the archived version of the site (Gray and Martin; Hsieh, Murray, and Hartman). These difficulties can result in incomplete captures of Websites; the prospect of quality assurance checks to correct these captures places onerous demands on staff attention and time. All of these obstacles make the practice of Web archiving a unique endeavor, unlike the collection of any other kind of archival material.

2 Collecting Scopes and Limitations

For the appraisal of analog and other digital materials, archivists need to consider the wealth of documentary production and decide what small piece should be preserved in their institution. For many institutions, this begins by defining a collecting scope, laying out particular focus areas to direct collecting activity. Sauer stresses the efficacy of collection development policies for archives and manuscript repositories in general to avoid the overcommitment of resources to collections of variegated, farflung and sundry materials, and similarly, Web archivists need to articulate the scope of their archives, winnowing down the immense volume of the Web to a manageable collection. Summer and Punzalan describe different collecting scopes as "crawl modalities,"[1] listing domain, topic, event, and specific website as modalities predominant in current practice.

Collecting at the level of a domain intends archiving Web-based materials either created by or pertaining to a large organization or entity, such as a university (Antracoli et al.), government body (Martin and Eubank), or even nation (Shadanpour et al.). This collecting can include all of the sites located at a particular domain extension (for example,.gov or.edu), but often also includes related materials that fall outside of that domain extension. Fansler, Gilbertson, and Petersen describe the process of developing a Web archives collection at Wake Forest University, and although their collecting scope is constrained to university-related materials, material exists on secondary, accessory sites outside of the actual university Web domain, especially including blogs and social media content. Shiozaki and Eisenschitz observe that many national libraries collect across their respective national domain (for example, .nl for the Netherlands), but supplement these crawls of the top-level national domain with topical and other focused crawls. For the Web archives at the Bibliothèque nationale de France (BnF), Lasfargues, Oury, and Wendland discuss the difficulty of defining what constitutes the "French domain," maintaining that this cannot be limited to Websites with the domain extension . fr, as many French-language and Francophone Websites also exist on .com, .net, and other domains. A domain-centric collection, then, cannot necessarily be limited to comprehensively crawling a particular top-level domain extension, but also often requires further curation from the collecting archivists. In addition to the BnF, archivists and librarians also describe this process at the National Library of Australia and the National Library of the Netherlands (Glanville), the Library and Archives Canada (Lilleniit), and the National Library and Archives of Iran (Shadanpour et al.).

Web archives with topical collecting scopes focus on a particular theme or subject matter. Several art libraries and museums have developed Web archiving programs devoted to collecting online art ephemera, such as artists' Websites, auction catalogs, and exhibition publicity material (Slania). Duncan and Blumenthal describe the collaborative undertaking by the member institutions of the New York Art Resources Consortium (NYARC) to supplement longstanding collecting efforts of print ephemera by targeting the increasing amount of arts ephemera that now exists only online. Similarly, much political and election-related ephemera has moved online, and archives and special collections that have long collected this ephemera have begun to supplement these print collections with Web archives of campaign Websites, voter information sites, and political party sites (Gray and Martin; Voerman et al.).

As with topic-centric collections, event-based Web archives constrain collecting around a particular issue or theme, but also focus collecting around a specific timeline. Rollason-Cass and Reed describe this approach as a Spontaneous Events model, or a Living Archives model, as these collecting programs respond and adapt to ongoing developments, offering the example of the #blacklivesmatter Web Archives at the IA. In the same vein, the University of California Berkeley built a Web archive documenting the Ukraine/Crimea conflict from 2014–2015 (Pendse). This collection had to respond to ongoing developments in the conflict, collecting Websites that changed dramatically over the course of events, many of which were at risk of being removed from the Web altogether. In line with the more general collections of online political ephemera, some archives also implement collecting programs around particular elections; for example, the National Library of Scotland selectively archived Web content pertaining to the historic 1999 Scottish parliamentary election, the first in over 300 years (Cunnea). In addition to these thematically focused collecting scopes, Niu observes that document or media type can also serve as a collecting modality, with archives focusing on collecting online newspapers for example.

Another way to think about scope is in terms of the scale and intensity of collecting. Dougherty and Meyer identify three scales of collecting: large-scale bulk har-

[1] The act of automatic web archiving software is often referred to as "crawling," as the program crawls across the web, capturing snapshots of the Websites it has been instructed to archive.

vesting, smaller-scale topic- or event-driven in institutional collections, and idiosyncratic collections developed by individual researchers. Masanès contrasts vertical and horizontal approaches, as two different modes of collecting intensity: vertical crawls are intensive and deep, striving to collect more comprehensively from a fewer number of sites; horizontal crawls are expansive, emphasizing coverage over a larger number of sites. Although institutions rarely operate at one of these extremes, Lasfargues, Oury, and Clement describe how the French Web archive does collect more horizontally than vertically, as their primary aim is to collect a representative sample of French cultural production across the Web rather than to completely capture a few select sites.

Another important aspect of scope is the frequency of collecting. Summer and Punzalan suggest that determining the frequency of collecting is a key appraisal decision, but also hinges on logistical constraints and resource limitations. The frequency of collecting directly impacts how much material an archive has to store (Chen, Chen, and Ting; Martin and Eubank) and how much staff time needs to be devoted to appraisal activities, like reviewing archived sites for quality assurance (Antracoli et al.). As the Web is by nature dynamic, archivists do need to crawl most sites more than once as content changes, but archivists need to decide when changes are important enough to justify further collecting. Chen, Chen, and Ting describe three patterns for scheduling crawls: a one-time, immediate crawl; ongoing crawls at regular intervals (once a month, for example); or a certain frequency between two preset dates. The frequency of crawling is especially important for event-based collections. For political- and election-related Web archives, election day is a central temporal marker and will likely structure the time and frequency of crawls (Gray and Martin). For ongoing events, like the Ukraine/Crimea conflict or the #blacklivesmatter movement, the frequency of collecting may respond to pivotal moments in the course of the overall developing phenomena (Pendse; Rollason-Cass and Reed). In most cases, the frequency of crawls is determined by the archivists; however, Saad and Grançarski have proposed a computationally-driven, predictive method for determining the frequency of crawls by analyzing patterns of when Webpages change and the importance of those changes.

Determining the scope, scale, intensity, and frequency of collecting, all constitute important appraisal decisions that shape the resulting Web archives. While all of these decisions are actively made by the archivist, other inherent and extrinsic factors and limitations play an indirect role in shaping Web archival collections. Legal risks of collecting Web archival material also present challenges, since so much material is protected by intellectual property rights (Glanville; Hsieh, Murray, and Hartman; Shiozaki and Eisenschitz). For other kinds of collecting efforts, archivists often require donor agreements from previous owners of archival materials, expressly handing over control to the archival institution; however, it is often not feasible to gain the consent of copyright holders for Web-based materials due to the sheer scale of collecting and due to the fact that it may not be possible to locate the copyright holder in many cases. Although the solution to this issue will vary depending on the goals and aims of the collecting institution, Dougherty and Meyer call for thoughtful agreement on legal issues across Web archives practitioners, a discussion that remains ongoing.

Currently, institutions employ a variety of approaches to face the limitations imposed by these legal risks. Shiozaki and Eisenschitz demonstrate that national policies can centralize rights clearance and thus mitigate legal risks. A number of countries have established formal systems of legal deposit. For the collection of French cultural production, Lasfargues, Oury, and Wendland describe how the framework of legal deposit has been extended to include Web-based materials. For France, this framework dates back to the 16th century, and has been successively adapted to include new media types as technology has developed over time, and continues to encompass a range of cultural products from high brow to low brow. The National Library of the Netherlands uses an opt-out policy also employed by the IA, automatically sending copyright holders a notice of an intention to archive with a deadline to opt-out (Glanville). For university archives and special collections, academic fair use has also been cited as a legal mechanism for archiving materials on the Web (Pendse).

As with any other kind of archival collecting effort, funding can indirectly affect the size, scope, and shape of collections. This can be acutely felt in the development of Web archives, which necessitate new tools, storage space, staff skills, and even infrastructure (Shiozaki and Eisenschitz). As a result, Web archiving can be tied to grants or other short term funding (Summer and Punzalan). For instance, NYARC's Web archiving efforts were significantly aided by grant funding from the Mellon Foundation and funded fellowship positions from the National Digital Stewardship Alliance (NDSA) (Duncan and Blumenthal). The reliance on short-term funding can be problematic, as many Web archiving costs are recurring, such as fees for subscription services and tools like Archive-It (Summer and Punzalan). Web archiving remains a new initiative for many archival institutions, and further research is needed to better understand the problems and solutions for sustaining these collecting programs.

3 Collecting Methods and Appraisal Activities

While archivists benefit from articulating collecting scopes that establish the broad parameters for their Web archives, much of the work of appraisal occurs on the ground in the form of a number of automated and manual tasks and activities. As Summer and Punzalan describe, Web archives are sociotechnical systems, involving the collaboration between human agents, seed lists,[2] and automated bots. Pearce-Moses and Kaczmarek suggest that Web archiving practices can range from the "technocentric," such as automated bulk harvesting of a domain, to "bibliographic," or handpicking Websites to include in a Web archives much like a librarian adds books to a library—although the authors observe the need to balance automated methods with manual input, depending on the collecting scope and goals of a given institution. Niu notes that automation can be used when criteria can be pre-established, but that topical and event-based collections may require more direct human curation. For instance, the Library and Archives Canada combines both of these approaches, using automatic bulk harvesting across top-level domains, along with the manual selection of a focused group of important sites (Lilleniit). For Gray and Martin, all stages of the process of collecting election-related materials, from identifying to capturing sites, are almost entirely manual.

For both automatic and manual capture, archivists develop seed lists to guide their Web archiving activity. These lists are often the main material trace of archivists' appraisal activity, exhibiting the archival intent for what has been deemed valuable enough to collect, and also act as the interface between archivists and automated agents like Web crawling bots (Antracoli et al.; Summer and Punzalan). Seed lists can be generated and maintained in a number of ways, such as spreadsheets that are shared and collaboratively worked on among archivists (Summer and Punzalan). Discussing Web archiving efforts at Arizona State University, Pearce-Moses and Kaczmarek describe a comprehensive taxonomy database, used to systemically organize content providers and related Websites. Seed lists are not only constructed by the archivists at the helm; several authors noted incorporating user input and recommendations for sites to be crawled (Duncan and Blumenthal; Niu; Rollason-Cass and Reed). Assessing use of collections from past crawls can also inform future crawls, if archivists recognize that certain kinds of materials are more heavily used than others (Chen, Chen, and Ting).

Martin and Eubank note that developing a seed list was the first challenge to establishing a North Carolina state government Web archives, and that developing a good seed list was a foundation for success. A number of strategies can be used to assist in this activity, and to ensure that archivists generate an effective list. For Martin and Eubank, this activity was aided by automated tools like Web Archives Workbench, manual reviews of sites, and a clearly articulated collecting scope. Creating or adopting an existing model of the target domain, theme, or event can also illustrate what Websites should be included on the seed list, including organizational hierarchies and depth charts, budgets, directories, mailing lists, and social networks (Pearce-Moses and Kaczmarek; Summer and Punzalan). To more systemically appraise sites for inclusion, the National Taiwan University uses a 10-point classification scheme to structure collecting in different areas, including arts and culture, politics, and technology; this scheme operationalizes the otherwise abstract appraisal activity of determining social, political, and cultural value (Chen, Chen, and Ting).

With seed list in hand, there are a number of Web archiving tools for archivists to automatically and manually capture Web sites, including Archive-It, ArchiveSocial, and Hanzo (Summer and Punzalan). Antracoli et al. list many benefits of using Archive-It, including the reputation of the IA (the service provider of Archive-It), advantages of the subscription model, the open formats used for storing archived content, interoperability between Archive-It and various preservation services, and the widespread adoption of the service across libraries and archives. Institutions, like the National Library of Australia, also contract directly with the IA to perform larger bulk harvests of Websites, complementing their own Web archiving efforts (Glanville). While tools like Archive-It are offered to institutions on a subscription model, Dougherty and Meyer note key differences in the affordances and capacities of desktop applications available to individual researchers and amateur Web archivists building their own collections. The differences between these tools can affect how and what gets captured in Web archival collections. Institutional archivists may also use multiple tools and approaches in conjunction to compensate for

[2] A seed list is the list of Websites, or URLs, to be included in a particular web archives. An archivist either manually captures these sites, or inputs this list of seeds into an automated web crawling tool.

comparative strengths and weaknesses of different capture technologies (Gray and Martin). Especially for quickly developing events and social media content, archivists may need to supplement their standard approaches with more responsive and nimbler tools. In developing the #blacklives matter Web Archives, Rollason-Cass and Reed describe using a Twitter-specific tool to glean thousands of tweets and related URLs that would have otherwise escaped their standard Web crawler. Vleck discusses using the WebAnalyzer tool to identify relevant Czech Websites that fall outside the designated .cz top-level domain, and then using Heritrix to crawl identified sites.

The activities of developing seed lists and performing crawls are iterative, ongoing, and mutually informative. Archivists perform test crawls to assess the strength of seed lists, the results of which feed back into appraisal decisions (Antracoli et al.), and can also perform patch crawls to supplement gaps left by regularly scheduled crawls (Lasfargues, Oury, and Wendland). For the Web archives at Wake Forest University, Fansler, Gilbertson, and Petersen perform test crawls and periodic crawl reviews to preempt potential scope and completeness issues, ensuring that crawls do not bring back out-of-scope content and that crawlers are capturing everything that was expected. At the State of North Carolina Library and Archives, Martin and Eubank also review crawls to assess for scope and completeness, and report actively weeding crawled materials falling outside of the collecting scope.

As the above illustrates, the diverse activities that constitute “web archiving” include many skills not typically cultivated by professional archivists. Thus, Web archiving necessarily involves interaction and communication between a variety of individuals, across departments and even organizations (Summer and Punzalan). Duncan and Blumenthal claim that a collaborative approach has been critical to the success of NYARC’s Web archiving efforts, allowing curatorial and appraisal effort to be spread across member institutions, and helping to meet a variety of Web archiving challenges, including technical difficulties and resource deficiencies. Rollason-Cass and Reed also cite the importance of collaborating across institutions to create and grow the #blacklivesmatter Web Archives. Duncan and Blumenthal suggest that similar trans-institutional collaboration could be encouraged through national organizations like the NDSA.

Detailing efforts to archive materials across several Pennsylvania universities, Antracoli et al. cite the special importance of building relationships with the webmasters administering the sites of the various departments, organizations, and schools within the collecting scope of the Web archives. Martin and Eubank also highlight the need to actively reach out to webmasters of state agencies, and educate them regarding the Web archiving efforts of the North Carolina State Library and Archives. In addition to establishing lines of communication and facilitating mutual support across departments, building this awareness can also prevent technical obstacles and address specific issues; Web admins, for example, can remove files that block access to automated bots. As Chen, Chen, and Ting observe, librarians and IT staff bring different skill sets, both of which are required to successfully collect Web archival materials. For some contexts, this kind of collaboration can be absolutely necessary. Shadanpour et al. report the need to cooperate directly with the national Iranian ICT infrastructure company to make Web archiving in this country technically feasible.

In addition to collaboration between archives and technical staff, Dougherty and Meyer call for a broader community of Web archiving practice that bridges gaps between institutions and individual researchers. These roles are increasingly blurring, as researchers develop personal Web archival collections that may have lasting value, and as institutions seek to develop Web archives that meet the needs of a variety of researchers. The authors describe divergent goals between institutions and researchers: while institutions are driven to collect Web archival material as part of a mission to preserve cultural heritage, researchers are driven to archive Web-based materials by specific research questions. Bridging these gaps between the Web archiving expectations and goals of institutions and researchers remains an area in need of further research, which might include the development of new collaborative collecting models.

4 Frameworks for Evaluating and Directing Appraisal

The above describes the on-the-ground work of collecting and appraising Web-based materials, and archivists seem to have developed a core set of practices and approaches, even as new tools and skills are adopted to meet the changing nature of Web content. However, it is also important for archivists to develop overarching theories of appraisal to better understand and assess the overall documentary record they are creating for the long-term benefit of cultural heritage. Especially beginning in the post-war period, archivists have generated a number of appraisal theories for understanding collection development of primarily analog archival materials, but it is now necessary to evaluate the extent to which these theories apply to Web archives.

Summer and Punzalan observe that there is a significant overlap in the appraisal practices of Web archives with existing appraisal models, citing a range from documentation strategies to post-custodial approaches. Fansler, Gilbertson, and Petersen also note that their Web archiving efforts are very much continuous with their appraisal of “born-physical” materials. However, in a review of national library Web archiving efforts, Shiozaki and Eisenschitz observe significant divergence between some institutions that position Web archiving in continuity with paper-based collections and those that position Web archiving as a more or less entirely new collecting endeavor, a distinction also noted by Dougherty and Meyer. While there may be organizational, logistical, and technical reasons for positioning Web archiving as a distinct collecting effort, several authors argue for the importance of developing an overarching policy that fits Web archiving within the broader collecting goals and mission of the institution (Antracolia et al.; Hsieh, Murray, and Hartman). These policies can position Web-based material in kind with analog and other born-digital materials as cultural heritage that needs to be collected and preserved (Lasfargues, Oury, and Wendland; Lilleniit).

If archival institutions do conceive of Web-based material in kind with other archival collections, then it makes sense for archivists to apply broader appraisal theories in Web archiving efforts. Summer and Punzalan discuss the ways in which many Web archivists—implicitly and explicitly—employ a Boomsian theory, according to which archives should strive to accurately reflect the social and cultural consciousness of the time (Booms). This theory is particularly apt to the Web, which seems to always be of the moment, constructed by a broad swath of society, and demonstrative of all manner of daily life. Pearce-Moses and Kaczmarek articulate a macro-appraisal approach to Web archives, another such appraisal theory in use before the advent of Web archives. The authors describe the macro-appraisal of Web-based materials as evaluating Websites as large aggregates organized by the archival principle of hierarchically embedded series and sub-series. The archivist can then assess these larger units rather than individual URLs. Based on the model set out by Pearce-Moses and Kaczmarek, North Carolina State Library and Archives developed a macro-appraisal score to apply to and evaluate domain hosts, rating this larger aggregate on seven factors including size, originality, frequency of update, historical value, public interest, and government interest (Martin and Eubank).

Cost-benefit analysis is another framework that can be applied to evaluate Web archiving efforts, assessing whether the costs justify the purported benefits of Web archives, namely the preservation of cultural heritage (Shiozaki and Eisenschitz). As Web archives require ongoing funds to collect new material and to store and preserve already collected material, weighing the cost of these programs is perhaps a fair inclination. However, Shiozaki and Eisenschitz note that the many intangible aspects of social, cultural, and historical value make “cultural heritage” a difficult—if not impossible—entity to quantitatively measure against costs. There are other aspects of Web archives that can be quantified, and these measures can help to evaluate the relative success of a Web archiving program. Saad and Gançarski propose temporal completeness as a measure for assessing the quality of a Web archives, evaluating how thoroughly an archive has captured important changes to pages over time. Temporal completeness can also be directly tied to costs, as more thorough crawling will require more storage space and staff effort. Usage statistics of Web archives can also be used to evaluate the relative value of a collection, with collections that receive more use deemed to be more valuable. However, usage may not be the sole purpose of preserving materials, and the mere act of documenting the present moment may be deemed valuable in and of itself. Hsieh, Murray, and Hartman assert that archivists generate value just by building Web archives, adding value by selecting, organizing, describing, and bringing together disparate materials into a bounded collection. These activities alone add value to Web-based materials that they did not have before. In any event, no appraisal theory can definitively answer questions about the use and value of materials. Collections may also be created with an eye towards future use, although the notion of future use plagues all appraisal theories and approaches as a great unknown: with web-based material as with paper-based collections, archivists simply cannot tell the future. Archivists can address issues of use and value, however, by striving to build broader communities of Web archiving practice, as noted above. By entering into direct dialogue with scholars working to actively make sense of the historical and cultural value of the Web, many of whom draw on, or even build, Web archives as part of their scholarship, archivists can integrate these perspectives to broaden the discourse around the archival appraisal of Web-based materials to better understand the value of Web content to society—now and in the future.

5 Conclusion

The practice of Web archiving continues to mature, as a variety of institutions and individual researchers initiate, develop, and grow collections of Web-based materials. As the activities, tools, and approaches that constitute the practice of "Web archiving" solidify, archivists will also need to find ways to assess this undertaking, applying overarching appraisal theories to better understand and direct the on-the-ground work of building Web archives. As discussed above, several authors have demonstrated that archivists do see Web archives in terms of their broader collecting aims, and as part of a larger mission to preserve a significant record of cultural heritage to pass on to future generations. However, relatively little attention in the literature has been paid to articulating specifically how Web-based materials fit into this larger body of cultural heritage materials. Now that many institutions have increasingly robust Web archives in their holdings, archivists will need to broach this topic. It is not enough to say that Web archives exist in continuity with analog and other born-digital collections without articulating a more nuanced understanding of how these disparate materials contribute to the larger picture that is cultural heritage.

The Web is still a relatively young medium, barely a generation old, and so society at large continues to grapple with the importance and position of the Web within a long and diverse documentary history. Scholars are increasingly looking at both the structure and the content of the live and the archived Web to investigate questions of political, social, and cultural history (Ben-David). Archivists occupy a unique place in the discourse of history, though, deciding what materials are preserved for future generations. Given the dynamic and often volatile life of Web-based material, Web archives already provide our only insight into much of the early Web, and are thus already prominent forces shaping historical views of the Web as a form of documentary heritage. In addition to reporting on their own institutional context, detailing the specific challenges and solutions encountered in developing a particular collection, archivists also need to articulate broader theories of how to assess the social, cultural, and historical value of Web-based materials.

Reporting about the on-the-ground issues faced in Web archiving efforts, of course, remains an important research topic as well, especially as the nature of the Web and Web capture technologies continue to develop. As the Web develops, broader theoretical frameworks in which to situate these specific cases will only increase in importance, in order to understand where we are going and from where we have come. In line with these theoretical frameworks, another major research area for the appraisal of Web-based materials is the development of measures for the use, completeness, and value of Web archives. As discussed above, definite and absolute quantitative measures for these concepts are not necessarily the goal, and in many cases perhaps they are not even possible. Still, archivists can and should develop systematic ways of thinking through these concepts. Usage may not be the only goal of Web archives, but it is certainly a primary end, and so archivists need to assess how a variety of individuals are engaging with their Web archives, and whether these use patterns may influence future collection development. Completeness will likely carry different meanings depending on the goals of the archives—for some this will be comprehensive coverage of a limited set of sites and for others a representative sample across a broad range—but completeness is a concept that archivists can successfully measure if the specific parameters are clearly laid out. Yet, only a few such attempts to measure completeness have been made. Value will remain a tricky concept to pin down, but this does not lessen the importance of communicating the value of Web archives to the archival profession, institutional administrators, potential users, and beyond. Despite the inexactness of value, Web archives do require funding and staff time in order to be built and sustained, and so archivists need to find ways to express the value of these collections, even if these are qualitative, descriptive, and exploratory.

Acknowlegdment: My thanks to Barbara Wildemuth, Yuan Li, and Charlene Finley for their helpful feedback in writing this article. I also want to thank the *Preservation, Digital Technology & Culture* editors and the peer reviewers for their valuable input.

References

Antracoli, Alexis, Steven Duckworth, Judith Silva, and Kristen Yarmey. "Capture All the URLs: First Steps in Web Archiving." *Pennsylvania Libraries* 2.2 (2014): 155–70.

Ben-David, Anat. "What Does the Web Remember of Its Deleted Past? An Archival Reconstruction of the Former Yugoslav Top-Level Domain." *New Media & Society* 18.7 (August 2016): 1103–19.

Booms, Hans. "Society and the Formation of a Documentary Heritage: Issues in the Appraisal of Archival Sources." Translated by Hermina Joldersma and Richard Klumpenhouwer. *Archivaria* 24 (1987): 69–107.

Chen, Kuang-Hua, Yen-liang Chen, and Peng-fung Ting. "Developing National Taiwan University Web Archiving System." In Proceedings of the 8th International Web Archiving Workshop. Aarhus, Denmark, 2008.

Cook, Terry. "'We Are What We Keep; We Keep What We Are': Archival Appraisal Past, Present and Future." *Journal of the Society of Archivists* 32.2 (2011): 173–89.

Cunnea, Paul. "Selective Web Archiving in the UK: A Perspective of the National Library of Scotland within UK Web Archiving Consortium (UKWAC)." *SCONUL Focus* 34 (2005): 44–49.

Dougherty, Meghan, and Eric T. Meyer. "Community, Tools, and Practices in Web Archiving: The State-of-the-Art in Relation to Social Science and Humanities Research Needs." *Journal of the Association for Information Science & Technology* 65.11 (2014): 2195–2209.

Duncan, Sumitra, and Karl-Ranier Blumenthal. "A Collaborative Model for Web Archiving Ephemeral Art Resources at the New York Art Resources Consortium (NYARC)." *Art Libraries Journal* 41.2 (2016): 116–26.

Fansler, Craig, Kevin Gilbertson, and Rebecca Petersen. "The Missing Link: Observations on the Evolution of a Web Archive." *Journal for the Society of North Carolina Archivists* 11.1 (2014): 46–59.

Glanville, Lachlan. "Web Archiving: Ethical and Legal Issues Affecting Programmes in Australia and the Netherlands." *Australian Library Journal* 59.3 (2010): 128–34.

Gray, Gabriella, and Scott Martin. "The UCLA Campaign Literature Archive: A Case Study." In Proceedings of the 7th International Web Archiving Workshop. Vancouver, British Columbia, 2007.

Hsieh, Inga K., Kathleen R. Murray, and Cathy Nelson Hartman. "Developing Collections of Web-Published Materials." *Journal of Web Librarianship* 1.2 (2007): 5–26.

Internet Archive. "About the Internet Archive." Internet Archive. https://archive.org/about/. (accessed 12/23/2016)

Lasfargues, France, Clément Oury, and Bert Wendland. "Legal Deposit of the French Web: Harvesting Strategies for a National Domain." In Proceedings of the 8th International Web Archiving Workshop. Aarhus, Denmark, 2008.

Lilleniit, Roselyn."Archiving the Canadian Web: Experiences at Library and Archives Canada." *Serials Librarian* 53 (2007): 139–49.

Martin, Kristin E., and Kelly Eubank. "The North Carolina State Government Website Archives: A Case Study of an American Government Web Archiving Project." *New Review of Hypermedia and Multimedia* 13.1 (2007): 7–26.

Masanès, Julien. "Web Archiving Methods and Approaches: A Comparative Study." *Library Trends* 54.1 (2005): 72–90.

Niu, Jinfang. "An Overview of Web Archiving." D-Lib Magazine 18.3 (2012). At http://www.dlib.org (accessed March 1, 2017).

Pearce-Moses, Richard, and Joanne Kaczmarek. "An Arizona Model for Preservation and Access of Web Documents." *DTTP: Documents to the People* 33.1 (2005): 17–24.

Pendse, Liladhar R. "Collecting and Preserving the Ukraine Conflict (2014–2015): A Web Archive at University of California, Berkeley." *Collection Building* 35.3 (2016): 64–72.

Rollason-Cass, Sylvie, and Scott Reed. "Living Movements, Living Archives: Selecting and Archiving Web Content During Times of Social Unrest." *New Review of Information Networking* 20.1 (2015): 241–47.

Saad, Myriam Ben, and Stéphane Gançarski. "Archiving the Web Using Page Changes Patterns: A Case Study." *International Journal on Digital Libraries* 13.1 (2012): 33–49.

Sauer, Cynthia K. "Doing the Best We Can? The Use of Collection Development Policies and Cooperative Collecting Activities at Manuscript Repositories." *The American Archivist* 64.2 (2001): 308–49.

Shadanpour, Farzaneh, Saeideh Akbari Dariyan, Reza Shahrabi Farahani, Soudeh Seirafi, and Alireza Vazifehdoust. "Building an Iran Web Archive in the National Library and Archives of Iran: A Feasibility Study." *Library Philosophy & Practice* (2012): 183–95.

Shiozaki, Ryo, and Tamara Eisenschitz. "Role and Justification of Web Archiving by National Libraries A Questionnaire Survey." *Journal of Librarianship and Information Science* 41.2 (2009): 90–107.

Slania, Heather. "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 (2013): 112–26.

Summers, Ed, and Ricardo Punzalan. "Bots, Seeds and People: Web Archives as Infrastructure." In Proceedings of the 20th ACM Conference on Computer Supported Collaborative Work. Portland, Oregon: ACM, 2017.

Vleck, Ivan. "Identification and Archiving of the Czech Web Outside the National Domain." In Proceedings of the 8th International Web Archiving Workshop. Aarhus, Denmark, 2008.

Voerman, Gerrit, André Keyzer, Frank den Hollander, and Henk Druiven. "Archiving the Web: Political Party Web Sites in the Netherlands." *Information Services & Use* 23.1 (2003): 1–7.

Bionote

Colin Post

Colin Post is a doctoral student in the School of Information and Library Science at the University of North Carolina—Chapel Hill, where is also pursuing a Masters degree in Art History. His research focuses on the preservation, collection, and study of digital artworks, and in particular net-based art. He also holds a Master of Fine Arts degree in Poetry from the University of Montana. More on his poetry and research can be found at http://colincpost.info.