Journal of Contemporary Archival Studies

Volume 6 Article 31

2019

Describing Web Archives: A Computer-Assisted Approach

Gregory Wiedeman
University at Albany, State University of New York, gwiedeman@albany.edu

Describing Web Archives: A Computer-Assisted Approach

Cover Page Footnote

The author would like to thank Paul Kelly for providing detailed notes and feedback on an early draft of this article. Any limitations are the responsibility of the author.

The web is a vast and important information space, and over the last twenty years the effort to preserve and document it has grown from the pioneering efforts of the Internet Archive to a community of archivists, librarians, and other practitioners who have contributed to hundreds of repositories. Still, web archives are challenging for users to discover and use outside of the limited URL-centric access of the Wayback Machine.[1] Even when users are able to find web archives, the complexity and opacity of web crawling means that they often have very little idea how or why certain captures were preserved.

Archivists have many of the tools to confront these challenges. The archival content standard, *Describing Archives: A Content Standard* (DACS) provides guidance for “all levels of description and forms of material.”[2] Archival arrangement and description focuses on providing access to meaningful aggregations of information efficiently and transparently. The web is a tremendous volume of complex and interrelated information, and web archives are a format that results from the technical process of web crawling or web recording. Web crawling at scale can be extremely noisy, meaning web archives contain both useful and useless things. Archival description empowers archivists to describe the meaningful characteristics of web archives rather than repetitively treating every web crawl as if it has the same features, context, and value.

Despite these advantages, web archives description has been dominated by bibliographic approaches. There are fundamental differences between archival and bibliographic descriptive traditions in both mission and structure.[3] Bibliographic description attempts to describe individual resources for discovery and access. In contrast, archival description attempts to provide access to records of human activity by identifying significant groupings of records, their creators, uses, and relationships. Traditionally, bibliographic records are a flat set of elements describing a single resource or group of resources, and this approach generally assumes that each

[1] Ian Milligan, “Researcher Access for Web Archives: Our Experiences,” Mid-Atlantic Regional Archives Conference, Buffalo, NY, October 2017; Maria-Dorina Costea, “Report on the Scholarly Use of Web Archives,” NetLab, 2018, accessed May 20, 2019, https://web.archive.org/web/20190416113523/netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf/.

[2] *Describing Archives: A Content Standard* (DACS), 2nd ed. (Chicago: Society of American Archivists, 2013), 5. Available at https://web.archive.org/web/20190307035606/http://files.archivists.org/pubs/DACS2E-2013_v0315.pdf. Most current version, accessed July 5, 2019, https://web.archive.org/web/20190705170945/https://github.com/saa-ts-dacs/dacs.

[3] There is a long and well-documented contrast between bibliographic and archival approaches to description in the archival literature. Richard C. Berner, “Archivists, Librarians, and the National Union Catalog of Manuscript Collections,” *The American Archivist* 27, no. 3 (July 1964): 401–9; Richard C. Berner, “Observations on Archivists, Librarians, and the National Union Catalog of Manuscript Collections,” *College and Research Libraries* 29, no. 4 (1968): 276–80. Steven Hensen described his work on Archives, Personal Papers, and Manuscripts (APPM) as “synthesiz[ing] basic archival principles into the broader framework of bibliographic description.” Steven L. Hensen, “‘NISTF II’ and EAD: The Evolution of Archival Description,” *The American Archivist* 60, no. 3 (Summer 1997): 290. During the development of EAD, Daniel Pitti contrasted SGML with MARC, which “was primarily designed to capture description and access information applying to a discrete bibliographic item.” Daniel V. Pitti, “The Berkeley Finding Aid Project,” accessed May 20, 2019, https://web.archive.org/web/20180206174035/http://archive1.village.virginia.edu/dvp4c/arlpap.htm. Terry Eastwood described archival arrangement as “essentially a process of identifying relationships, not a process of physically ordering and storing documents,” which contrasts with “the old-fashioned ‘one-thing-one-entity’ approach.” Terry Eastwood, “Putting the Parts of the Whole Together: Systematic Arrangement of Archives,” *Archivaria* 50, no. 1 (Fall 2000): 93, 116.

object in a repository requires a metadata record.[4] Dublin Core builds on this tradition in its Description Set Model, which "is a set of one or more *descriptions*, each of which describes a single *resource*."[5] While DACS allows archivists to describe materials at any level—including item-level records—it requires archivists to explicitly identify each record's relationship to its wider context of creation and use.[6] This means that, although archivists are welcome to describe any grouping of web archives using a set of bibliographic fields, to be compliant with DACS, archivists must connect these records to broader documentation about the activity(s) and agent(s) responsible for creating and using the same web archives.[7]

This poses a major challenge to archivists describing the web, as web archiving tools are typically created with a single set of fields in mind. Currently, the only path to utilizing archival description often requires an unacceptable amount of repetitive labor to update records manually. The pace of web collecting often means that archivists must update date and extent elements almost constantly. This is true even for web archives practitioners in archival repositories that rely on DACS and employ tools like ArchivesSpace. This article outlines the benefits of aggregate archival arrangement and description for web archives with DACS and offers a path to develop a tool that would be feasible for many web archives practitioners to implement that would automate the repetitive labor that is currently required.

Current Practices

The development of Archive-It in 2006 made it feasible for many libraries and archives to build their own web archives collections using the tools from the Internet Archive. Since then, other external capture services and local capture tools—such as Webrecorder and Wget—have matured and become easier to use. However, with over 400 institutional users and a growing share of users, Archive-It is still the most widely used tool by larger repositories and its tools have had a large impact on current practices.[8] The Archive-It web application allows users to describe Archive-It collections or seeds with a set of Dublin Core elements or additional user-

[4] Recent trends in bibliographic description have moved toward an entity-relationship model where the most basic unit of description has moved from bibliographic records to metadata statements. However, this thinking does not describe the current state of description for web archives. IFLA Study Group on the Functional Requirements for Bibliographic Records, "Functional Requirements for Bibliographic Records: Final Report (FRBR)," February 2009, accessed May 20,2019, https://web.archive.org/web/20190331171302/https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

[5] Andy Powell, Mikael Nilsson, Ambjörn Naeve, Pete Johnson, and Tom Baker, "DCMI: DCMI Abstract Model," June 4, 2007 [emphasis in original], accessed May 20, 2019, https://web.archive.org/web/20190504123346/www.dublincore.org/specifications/dublin-core/abstract-model/.

[6] *DACS*, 2nd ed., 10.

[7] This is made more explicit in recent updates to *DACS*, chapter 1, "Requirements for Multilevel Descriptions," accessed May 20, 2019, https://web.archive.org/web/20190520170454/https://github.com/saa-tsacs/dacs/blob/3a52ee51bf4dfd4d6cd1a7dc38bec196f4cc307f/part_I/chapter_1.md. This point has also been made informally by Maureen Callahan.

[8] "About Us," *Archive-It Blog*, accessed April 13, 2019, https://web.archive.org/web/20190328195120/https://archive-it.org/blog/learn-more/; Matthew Farrell, Edward McCain, Maria Praetzellis, Grace Thomas, and Paige Walker, "Web Archiving in the United States: A 2017 Survey," an NDSA Report, October 2018, 19–21, accessed April 13, 2019, https://osf.io/ht6ay/.

defined fields. For this reason, Dublin Core is widely used to describe web archives.[9] This is also common for archivists from college and university archives, which make up a substantial portion of web archiving institutions.[10] Archivists create Archive-It collections, which are groupings of seeds commonly based on topic, scope, or perhaps just how it was most practical to manage them. These collections may or may not correlate with archival fonds or existing description, and are typically described using the Archive-It Dublin Core elements. These web archives are often only accessible by searching this description on the Archive-It website, or through the general Wayback Machine.

Some archivists have attempted to bridge the divide between web archives and archival theory and practice. Robert Pearce-Moses and Joanne Kaczmarek outlined an approach to apply archival arrangement and description to online publications back in 2005.[11] Christie Peterson has described archival principles as "absent in discussions" on web archives. She attempted to describe web archives created using Archive-It, but was dissatisfied that "the units of arrangement, description, and access typically used in web archives simply don't map well onto traditional archival units of arrangement and description" describing how "our current interfaces and methods of describing and accessing web archives privilege *Collections* and especially *Seeds* over *Crawls.*"[12] Some archivists at other repositories have been able to use archival description to describe web archives, and provided access to them alongside other archival materials.[13] New

[9] Jackie M. Dooley, Karen Stoll Farrell, Tammi Kim, and Jessica Venlet, "Developing Web Archiving Metadata Best Practices to Meet User Needs," *Journal of Western Archives* 8, no. 2 (2017): 8.

[10] The latest NDSA Web Archiving Survey found that 60.5 percent of respondents were from colleges or universities. While not all of these practitioners come from archival repositories, a major driver for this is the public records requirements of public colleges and universities that often mandate the preservation of permanent records on the web. The same survey found that 86 percent of respondents were attempting to capture their own web content. Farrell et al., "Web Archiving," 7–8, 10. A content study of Archive-It collections found that 54.1 percent of all Archive-It collections were categorized as "self-archiving." Shawn M. Jones, Alexander Nwala, Michele C. Weigel, and Michael K. Nelson, "The Many Shapes of Archive-It: Characteristics of Archive-It Collections," *Proceedings of the 15th International Conference on Digital Preservation* (2018): 6–7, accessed April 7, 2019, https://arxiv.org/abs/1806.06878.

[11] Richard Pearce-Moses and Joanne Kaczmarek, "An Arizona Model for Preservation and Access of Web Documents," *DttP: Documents to the People* 33, no. 1 (Spring 2005): 17–24.

[12] Christie Peterson, "Archival Description for Web Archives," *Chaos —> Order*, June 12, 2015, accessed April 14, 2019, https://web.archive.org/web/20180307061542/https://icantiemyownshoes.wordpress.com/2015/06/12/archival-description-for-web-archives/. Reposted in *On Archivy*, June 22, 2015, accessed April 14, 2019, https://web.archive.org/web/20170604110440/https://medium.com/on-archivy/archival-description-for-web-archives-1d9dce8dcef0.

[13] Mink Family Restaurants Records, Temple University Special Collections Research Center, accessed April 14, 2019, https://web.archive.org/web/20180624131829/https://library.temple.edu/scrc/mink-family-restaurants; Guide to the Mike Topp Papers, Fales Library and Special Collections, New York University, accessed April 14, 2019, https://web.archive.org/web/20190414154500/http://dlib.nyu.edu/findingaids/html/fales/mss_188/dscaspace_fd1a024cc9bb851fc2b7a610b336664b.html; "Records of the Association of American Women in Europe," Harvard University Archives, accessed April 14, 2019, https://web.archive.org/web/20190414155410/https://hollisarchives.lib.harvard.edu/repositories/8/archival_objects/2910347; "Guide to the University Archives Web Archive Collection," Duke University Archives, accessed April 19, 2019, http://library.duke.edu/rubenstein/findingaids/uawebarchive/; "University of North Carolina at Chapel Hill University Archives Collected Websites," University of North Carolina at Chapel Hill, accessed April 14, 2019, https://web.archive.org/web/20180529115702/http://finding-aids.lib.unc.edu/40417/; "Guide to the University of Chicago Web Archive Collection," Special Collections Research Center, University of Chicago Library, accessed April 14, 2019,

York University Libraries led a grant-funded effort that included an integration between ArchivesSpace and Archive-It. The project developed an ArchivesSpace plugin and worked to adapt the Archive-It web application so that basic description and a link to archival materials could be included in an Archive-It web archives record. Users browsing the Archive-It website could then see collections with related archival materials.[14]

OCLC developed and led a comprehensive effort to develop consensus guidelines for web archives metadata that would ease discoverability. The OCLC Research Library Partnership Web Archiving Metadata Working Group (WAM) gathered together many leading web archives practitioners and developed a suite of outcomes that included recommendations and a model data dictionary. The group documented a number of useful findings, such as the "strong need" for provenance information, the limited discovery of web archives across multiple systems, and need for scalable practices, as "staff resources for this work are extremely limited at most institutions."[15] The recommendations tried to develop a universal approach specific to web archives that would "provide a bridge between bibliographic and archival approaches to description."[16] While the WAM group certainly reviewed and tried to incorporate archival descriptive practices and standards, the recommended data dictionary was a fundamentally bibliographic approach. The report prioritized a single-level set of elements and seemed to envision merely using archival hierarchy to order bibliographic records instead of engaging with archival arrangement and description theory and practice. Furthermore, as the Society of American Archivists' Technical Subcommittee on Describing Archives: A Content Standard (TS-DACS) noted in their response, the data dictionary acted in some ways like a content standard that was in conflict with DACS.[17] While many of the WAM group's outcomes are useful, this makes it problematic for archivists to follow many of the recommended practices.

Web Archives as Archives: The Case for Aggregate Description

The OCLC WAM group's thorough review of current practices found "wide variation" in both the content of web archives description and the metadata fields used to describe similar

https://web.archive.org/web/20190414160018/https://www.lib.uchicago.edu/e/scrc/findingaids/view.php?eadid=ICU.SPCL.UCWEB.

[14] "Archiving the Websites of Contemporary Composers," *Archive-It Blog*, accessed April 14, 2019, https://archive-it.org/blog/projects/composers/; "New ArchivesSpace Integration," *Archive-It Blog*, accessed April 14, 2019, https://web.archive.org/web/20190407235818/https://archive-it.org/blog/post/archivesspace-integration/; "Archive It Integration," Archivesspace-DO-Plugin, May 14, 2018, https://web.archive.org/web/20190414143424/https://github.com/NYULibraries/Archivesspace-DO-Plugin/wiki/Archive-It-Integration; Avery Fisher Center: K. Marie Kim Collection, New York University Collection in Archive-It, accessed April 14, 2019, https://web.archive.org/web/20181122193415/https://archive-it.org/collections/9898.

[15] Jackie Dooley and Kate Bowers, *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group* (Dublin, OH: OCLC Research, 2018), 12–13.

[16] Ibid., 4.

[17] "Response to Best Practices for Web Archiving Metadata," Society of American Archivists' Technical Subcommittee on Describing Archives: A Content Standard, June 12, 2017, accessed April 14, 2019, https://web.archive.org/web/20190414170439/https://docs.google.com/document/d/1x5BuGYdtdjfVvbnXfbTDt7VL2ETfFwI5IfH-GUrs57U/edit.

content.[18] To the bibliographically minded, this could appear to be poor and imprecise practice. Yet, it is also possible that this was the result of well-intentioned librarians and archivists attempting to put a variety of different-shaped pegs into round holes. Archival description addresses the descriptive complexity required by web archives by providing a layer of abstraction where archivists describe the meaningful characteristics of web archives, separate from the technical process of web crawling or web recording. Describing meaning at an intellectual level allows archivists to provide access to web archives alongside other collections in different formats that have a similar context of creation, instead of keeping descriptions of web archives in a format-based silo. Focusing on the description of meaningful aggregates also empowers archivists to be efficient and thoughtful in their labor.

One core concept of archival description theory is that record formats are less meaningful characteristics than the *use* and *form* of materials, which archivists should focus on instead.[19] Web archives often contain materials with very different uses and forms. A university website might have "give" as a primary menu option, while a university library website may instead feature "services" on its menu, demonstrating that even websites created by the same organization can have vastly different priorities and purposes. Bearman and Lytle provide a particularly vivid description of form, contending that "a bank cheque, written on a watermelon, is nonetheless a cheque and even negotiable!"[20] We can see how a ledger and a spreadsheet are different formats but similar forms with similar uses. Users studying an organization's financial history are very likely to use both. Similarly, users interested in a university's branding and marketing might look at paper mailings and the university's homepage and online news, but not be interested in faculty websites in the same domain.

When librarians and archivists describe web archives, they might be describing organizational records, marketing and promotional materials, journalism of various quality, correspondence over social media, music, video, or a variety of other forms. The description for a set of online meeting minutes and a website that shows someone's personal search history one day at a time *should* be different, and users require different types of information to understand each.[21] The variability in web archives descriptive practices is understandable when one considers that the librarians and archivists describing web archives are sometimes describing a wide variety of materials with different uses, functions, and values.

Describing the wide variety and scale of web archives is difficult, but archival description is readily applicable to the challenge. DACS tells archivists to describe "meaningful aggregation[s]

[18] Dooley et al., "Developing Web Archiving Metadata," 11. The OCLC white paper stated that "institutional metadata guidelines vary widely in both the elements it included and in the choice of content within those elements." Dooley and Bowers, *Descriptive Metadata*, 14.

[19] Kathleen D. Roe, *Arranging & Describing Archives & Manuscripts*, Archival Fundamentals Series 2 (Chicago: Society of American Archivists, 2005), 17; David A. Bearman and Richard H. Lytle, "The Power of the Principle of Provenance," *Archivaria* 21 (Winter 1985–86): 14–27.

[20] Bearman and Lytle, "The Power of the Principle," 22.

[21] "Cannotsleepwithsnoringhusband," accessed April 1, 2019, https://web.archive.org/web/20180310234259/cannotsleepwithsnoringhusband.online/. As discussed by Dragan Espenschied, "The Ethics of Digital Folklore," National Forum on Ethics & Archiving the Web, New York, NY, March 23, 2018, accessed April 1, 2019, https://vimeo.com/276948412; Olia Lialina, "Do you believe in user 711391? A Search Engine Drama," *Rhizome*, March 7, 2016, accessed April 1, 2018, https://web.archive.org/web/20180816160522/rhizome.org/editorial/2016/mar/07/do-you-believe-in-user-711391/.

of records" at a higher level of abstraction than the layer at which the records are stored and managed.[22] This means archivists can physically keep records in whatever boxes or shelves they best fit, but present them in a more usable and meaningful way than how they are stored. The simplest version of this is an index, which lists materials alphabetically or chronologically, but contains references to materials stored in a different order.[23] To describe web archives effectively, archivists should often describe in aggregations that do not correlate with how they were captured. Seeds, or even groups of seeds, are not always the most meaningful material to describe. Web captures that span many different websites and web captures of large organizations are both likely to demand multiple important points of access. A common example is a university website crawled from a single seed. The creation and management of web content in many large universities is often decentralized, meaning that staff and faculty from all over the organization edit and update websites.[24] This means that while it might be effective to start a web crawl from the university home page, there are likely to be many pages throughout the site that serve different functions and have different forms, and should act as meaningful access points for users. Archivists might individually describe the page that lists the legislation and minutes from the University Senate, the academic calendar, course catalogs, or even the page of each academic department. Much like users of ledgers and spreadsheets, users interested in paper course catalogs are also likely to be interested in course catalogs captured in web archives. Describing web archives by meaningful aggregations enables archivists to describe these materials together and present them together to users. Any web page captured in a web crawl—whether that page is a seed or was just captured along the way—could be described at any point in an archival hierarchy. An archivist could describe this page as a collection, series, subseries, file, item, or even a digital representation of another instance.

Aggregate description also enables archivists to employ their time efficiently and make a larger body of materials available to users. Bibliographic description traditionally focuses on a single level and often treats each resource the same by applying the same descriptive elements, regardless of its value. By this practice, if a librarian or archivist describes an individual creator or contributor for a seed or collection of seeds, they also assume that it is beneficial or even necessary to apply that same effort to every seed or collection. Instead, archival description enables archivists to match the amount of time and labor applied to any object with its value and its consistency with the overall mission of the archives. An archivist's time is finite and metadata creation is extremely laborious. As the OCLC WAM group discovered, "the top barrier to metadata creation was lack of staff time."[25] Since a majority of survey respondents stated they

[22] "Statement of Principles," in *Describing Archives: A Content Standard*, accessed July 7, 2019, https://web.archive.org/web/20190705175744/https://github.com/saa-ts-dacs/dacs/blob/master/statement_of_principles.md.

[23] Maureen Callahan provides a good example of how it is useful for archivists to keep description on a layer of abstraction separate from the containers that manage the physical materials. Maureen Callahan, "On Containers," *Chaos --> Order*, December 15, 2014, accessed April 2, 2019, https://web.archive.org/web/20160404091903/https://icantiemyownshoes.wordpress.com/2014/12/15/on-containers/.

[24] Anne Arendt and Nathan Gerber, "Dispersed Web Content Management in Higher Education," in *Educause Review*, July 30, 2009, accessed April 2, 2019, https://web.archive.org/web/20190402200808/https://er.educause.edu/articles/2009/7/dispersed-web-content-management-in-higher-education; Aaron Rester, "Web in the Higher Ed Org Chart," accessed April 2, 2019, https://web.archive.org/web/20190402194959/https://blog.aaronrester.net/2013/04/web-in-the-higher-ed-org-chart.html.

[25] Dooley et al., "Developing Web Archiving Metadata," 7.

devoted only 0.25 full time staff (FTE) to web archiving, flexibility in applying their labor may be even more important to web archivists.[26] Archival description provides archivists with the agency to apply their limited labor creatively, which might mean describing many preserved websites from a web capture in detail, or providing only the most cursory descriptive effort to an entire group of web archives. This can be more intellectually challenging than some parts of bibliographic description, where web archives professionals sometimes repeat the same information for different objects.[27] The result is a higher value for the labor of archivists, as well as the ability to describe a greater volume of materials and make more web archives discoverable to users. Leaving descriptive fields for web archives empty may feel to archivists as if they are not adequately describing the materials, but the reality is that this might be better descriptive stewardship. If archivists applied the same detailed metadata to all seeds or groups of seeds, they would create less description, and users would not be able to discover or use as many web archives.

Even the Library of Congress, with its long tradition of bibliographic leadership, has found that principles from archival description apply to web archives effectively. Grace Thomas described how the Library of Congress's Web Archiving Team discovered that "many web archives overlap between thematic collections." Different curators from throughout the library included the same preserved website in the collections they managed. Rather than duplicate captures, the team created a higher layer of abstraction in which a preserved website was stored in a single web capture, yet was described in multiple thematic collections. Descriptive titles were standardized at the web capture level, and the Web Archiving Team used a Python script to generate the thematic collection records automatically. These records do not contain subject headings and "exclude highly tailored information," but this approach has enabled the Web Archiving Team to describe 4,240 web archives and make them available to the public, many times more than they would be able to provide with traditional cataloging.[28] The Library of Congress example demonstrates how approaches from aggregate description need not be limited to repositories that employ traditional archival methods and tools, but can be useful for bibliographic systems as well.

While aggregate description for web archives provides substantial advantages for repositories, this is an idealistic vision of archival description, as not all description meets current best practices. There is also a documented history of archival access systems being a barrier for users. As the recent update to the DACS Statement of Principles concludes, "Archival description is a continuous intellectual endeavor."[29] Description of complex materials like web archives is challenging, implemented at various levels of proficiency, and can often be improved upon. Additionally, a number of user studies concluded that archival finding aids often fail to meet user

[26] Farrell et al., "Web Archiving," 14.

[27] Daniel A. Santamaria notes that "it is harder to summarize the contents of a group of boxes than it is to type up a list of items or even folders. This is where the real skill and talents of an archivist lie." Daniel A. Santamaria, *Extensible Processing for Archives and Special Collections: Reducing Processing Backlogs* (Chicago: Neal-Schuman, 2015), 19.

[28] Grace Thomas, "More Web Archives, Less Process," *The Signal*, August 3, 2018, accessed April 4, 2019, https://web.archive.org/web/20180803194118/blogs.loc.gov/thesignal/2018/08/more-web-archives-less-process.

[29] "Statement of Principles," in *Describing Archives: A Content Standard*, updated February 1, 2019, accessed July 5, 2019, https://web.archive.org/web/20190705175744/https://github.com/saa-ts-dacs/dacs/blob/master/statement_of_principles.md.

Figure 1: An idealized example of describing web archives, arranged by form and function, and connected to broader contextual information. Items in **bold** are components that represent web archives and link to a separate level of abstraction, which manages metadata about the collecting process, including technical provenance metadata, and qualitative appraisal information

- New York State Modern Political Archive
  - *appraisal/selection metadata*
  - Environmental Advocates of New York Records
    - Executive Director Records
      - Paper records, email
    - Press Releases
      - Paper records
      - …
      - **Congress Must Act to Ensure Clean Air**, Captured 2018 June 16
      - **Include all $2.5B for water infrastructure**, Captured 2019 March 26
      - …
    - Publications
      - Paper records
      - …
      - **Report 58** Captured 2016 May 1
      - **Report 59** Captured 2017 June 15
      - …
    - Web Presence
      - **Website**, 2013–2019
        - *minimal descriptive metadata*
      - Social Media accounts
  - Capitol Confidential Blog
    - *detailed descriptive metadata*
    - **Capture** 2018 April 18
    - **Capture** 2018 April 25
    - **Capture** 2018 May 2
    - **Capture** 2018 May 9
    - ...
  - Michelle Crone Papers
    - Correspondence
      - letters, email, social media posts
    - **Personal Website**, 2013–2019
      - *minimal descriptive metadata*
    - Seneca Women's Encampment for a Future of Peace and Justice
  - New York State 2018 Election Collection
    - *appraisal/selection metadata*
    - Representative 1
      - **Campaign website,** 2018 January–November
        - *minimal descriptive metadata*
      - Social media posts
    - Representative 2
    - ...

needs.[30] However, these studies examined traditional finding aids and archivists now have tools that enable them to present archival description in new ways. A number of projects have begun that utilize this new potential.[31] If archivists apply archival description to web archives, they can work to present the valuable context and provenance information from lower levels in new and intuitive ways that would otherwise not be possible with Dublin Core fields.

Web archives are large and extremely diverse in value, use, and form. Archival description enables archivists to address each preserved website from a capture according to its value. Some may deserve detailed description and many will not be described at all and be accessible only by links from other pages. Treating description as a higher-order level, apart from the technical processes of web crawling or web recording, also allows archivists to provide access to web archives according to their use and form, together with similar materials in other formats. This approach allows archivists to use their time and labor creatively and provides efficient access to the largest possible volume of materials, while avoiding repetitive tasks and challenging archivists with skilled professional-level work.

Access to Web Archives Provenance

The lack of provenance information for publicly accessible web archives is a pressing concern for users, but there are currently no commonly used methods or best practices for librarians and archivists to include information about how or why they captured these materials in web archives metadata. Researchers need this information to weigh the representativeness of their findings effectively. Maemura et al. assert that "for researchers to have confidence in the validity of their findings . . . they must also understand how the collection was created." They attempted to map the "decision space" for creating web archives to provide a foundation for empirical work on web archives provenance.[32] Littman et al. argue that when using social media data, "non-academic researchers also require documentation about the data collection process to establish the trustworthiness of the data."[33] The OCLC WAM group found that "users express a strong

[30] Anne J. Gilliland-Swetland, "Popularizing the Finding Aid: Exploiting EAD to Enhance Online Discovery and Retrieval in Archival Information Systems by Diverse User Groups," *Journal of Internet Cataloging* 4, nos. 3–4 (2001): 199–225; Elizabeth Yakel, "Listening to Users," *Archival Issues* 26, no. 2 (2002): 111–27; Christopher J. Prom, "User Interactions with Electronic Finding Aids in a Controlled Setting," *The American Archivist* 67, no. 2 (Fall/Winter 2004): 234–68; Joyce Celeste Chapman, "Observing Users: An Empirical Analysis of User Interaction with Online Finding Aids," *Journal of Archival Organization* 8 (2010): 4–30; Jody DeRidder, Amanda Presnell, and Kevin Walker, "Leveraging Encoded Archival Description for Access to Digital Content: A Cost and Usability Analysis," *The American Archivist* 75, no. 1 (Spring/Summer 2012): 143–70; Tracy M. Jackson, "I Want To See It: A Usability Study of Digital Content Integrated into Finding Aids," *Journal for the Society of North Carolina Archivists* 9, no. 2 (2012): 20–77.

[31] ArchivesSpace Public Interface Enhancement Project, accessed May 21, 2019, https://archivesspace.atlassian.net/wiki/spaces/ADC/pages/22282254/Public+Interface+Enhancement+Project; ArcLight, accessed May 21, 2019, https://web.archive.org/web/20171024160933/https://wiki.duraspace.org/display/samvera/ArcLight.

[32] Emily Maemura, Nicholas Worby, Ian Milligan, and Christoph Becker, "If These Crawls Could Talk: Studying and Documenting Web Archives Provenance," *Journal of the Association for Information Science & Technology* 69, no. 10 (October 2018): 1223–33, accessed May 17, 2019, https://tspace.library.utoronto.ca/handle/1807/82840.

[33] Justin Littman, Daniel Chudnov, Daniel Kerchner, Christie Peterson, Yecheng Tan, Rachel Trent, Rajat Vij, and Laura Wrubel, "API-Based Social Media Collecting as a Form of Web Archiving," *International Journal on Digital Libraries* 19, no. 1 (March 2018): 27. The importance of web archives provenance is also discussed in Pamela M.

need for provenance information to add context beyond standard descriptive metadata elements," yet defined "preservation metadata" as out-of-scope and failed to make any recommendations for improving current practices.[34]

When the web archives literature talks about provenance, the authors often think of scoping rules, time stamps, and crawler settings, which is quite different from the provenance that archivists have historically managed with description.[35] Littman et al. described three facets of social media provenance that must be documented: creation of the record, the technical capture of the record, and the selection or appraisal of the record.[36] While archival description has long documented the creation of records, archivists typically only describe appraisal at higher levels, such as the fond level or in collection development policies, and archival standards do not currently provide sufficient guidance for describing the capture process. As Maemura et al. noted, "A key challenge for archival theory is reconciling an expanded notion of provenance with a system of archival description."[37]

Describing web archives in aggregations can help bridge this divide by allowing practitioners to arrange web archives according to creation and use while enabling them to manage granular information about the collecting and selection process at a lower level of abstraction—in this case the crawl, recording, or acquisition level.[38] This collecting process metadata contains both technical metadata on the crawl process as well as qualitative information about the appraisal, selection, or intention of the acquisition. While the approach of Maemura et al. is to develop a framework to systematically document the decisions of web archivists, this approach instead focuses on the end product of these decisions—the material captured—and exposes the intentions and actions of web archivists. This holds the archivist responsible for making their actions transparent, yet places the burden on the user to infer the reasons for specific cases. Some standardization is also necessary. The complexity of web archives collecting tools means that archivists must standardize documentation of their actions outside of even granular collecting process information. This includes authorities for tools, crawl types, and scoping exclusions. While this approach is pragmatic and useful, it does reveal gaps in how archival description accounts for the actions of archivists and their tools in the collecting process.

While archivists have long been concerned with provenance, the inclusion of provenance statements in descriptive metadata originally comes from bibliographic description. The Library

Graham, "Guest Editorial: Reflections on the Ethics of Web Archiving," *Journal of Archival Organization* 14, nos. 3–4 (July-December 2017): 105; Andrew N. Jackson, Jimmy Lin, Ian Milligan, and Nick Ruest, "Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities," *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* 16 (2016): 103, accessed April 5, 2019, https://yorkspace.library.yorku.ca/xmlui/handle/10315/31236.

[34] Dooley and Bowers, *Descriptive Metadata*, 12, 7. The WAM group's midterm report also described "provenance information as a critical missing piece." Dooley et al., "Developing Web Archiving Metadata," 5.

[35] Andrew N. Jackson provides a concise overview of web archives provenance from a technical perspective in "The Provenance of Web Archives," *UK Web Archive Blog*, November 20, 2015, accessed April 10, 2019, https://web.archive.org/web/20180123212308/blogs.bl.uk/webarchive/2015/11/the-provenance-of-web-archives.html. This difference was noted by Dooley et al., "Developing Web Archiving Metadata," 5.

[36] Littman et al., "API-Based Social Media Collecting," 27.

[37] Maemura et al., "If These Crawls Could Talk," 6.

[38] The observation that the crawl level was the most appropriate to focus on was made by Christine Peterson, "Archival Description for Web Archives."

of Congress Manuscript Cataloging Division first created the "Note on provenance" for the National Union Catalog for Manuscript Collections and archivists became more comfortable with creator authorities through their adoption of the MARC for Archival and Manuscripts Control (MARC-AMC) standard.[39] While archivists have long accepted the need to describe the origin of collections explicitly, they have historically used provenance by arranging materials by their creation and use, which allows them to manage records efficiently and utilize provenance for discovery.[40] Current practices do document the creation of records, but not always in a single note, as bibliographic description might expect. The "immediate source of acquisition" note is not required by either DACS or ISAD(G).[41] This means that the provenance of archival material has been historically managed though the arrangement of material, rather than stated explicitly.

For web archives, the description of provenance via arrangement is insufficient. Moreover, multiple archivists have argued that archival description standards and prevailing practices still fail to document the nuances of records creation and the actions of archivists upon records to the standards of the recent revision to the DACS principles, which state that "archivists must document and make discoverable the actions they take on records."[42] Archivists have confronted this challenge when managing born-digital records, but have not yet worked to reconcile their descriptive standards with the granular information necessary to document, say, the numerous points of context found in a computer's file system. The biggest advancements here have come though the adoption of digital forensics tools, which preserve that context effectively using disk imaging, but do not necessarily make it discoverable—or even accessible—to users.[43] Web archives, which are created both by the authors of their content and by archivists themselves, should provide a useful challenge for how archival description documents provenance.

Using archival description for web archives enables archivists to document the provenance of the content within a WARC file via their traditional practices. Archivists can arrange and describe content within a web archives as they would for any other format. Any page, seed or otherwise, captured in a web crawl or web recording could be arranged as a collection, a series, or a file,

[39] Robert H. Land, "The National Union Catalog of Manuscript Collections," *The American Archivist* 17, no. 3 (July 1954), 201. Initially there was even some resistance to the idea. Berner, "Archivists, Librarians," 401–2.

[40] The acceptance of explicit provenance information in description generally began with Max Evans, "Authority Control: An Alternative to the Records Group Concept," *The American Archivist* 49, no. 3 (Summer 1986): 249–61. The use of provenance for access was articulated by Bearman and Lytle, "The Power of the Principle," 14–27.

[41] *ISAD(G): General International Standard Archival Description: Adopted by the Ad Hoc Commission on Descriptive Standards, Stockholm, Sweden, 21–23 January 1993; Final ICA Approved Version* (Ottawa: Secretariat of the Ad Hoc Commission on Descriptive Standards, 1994), accessed April 8, 2019, https://web.archive.org/web/20150706032159/http://www.hi.u-tokyo.ac.jp:80/personal/yokoyama/jugyo99/isad(g)e.html; *Describing Archives: A Content Standard* (Chicago: Society of American Archivists, 2004), 8, 64, accessed April 8, 2019, https://web.archive.org/web/20180916052424/files.archivists.org/pubs/DACS2E-2013_v0315.pdf.

[42] David Bearman, "Documenting Documentation," *Archivaria* 34 (Summer 1992): 33–49; Michelle Light and Tom Hyry, "Colophons and Annotations: New Directions for the Finding Aid," *The American Archivist* 65 (Fall-Winter 2002): 216–30; "Statement of Principles."

[43] Kam Woods and Christopher A. Lee, "Acquisition and Processing of Disk Images to Further Archival Goals," in *Proceedings of Archiving 2012* (Springfield, VA: Society for Imaging Science and Technology, 2012), 148, accessed April 9, 2019, https://web.archive.org/web/20170826185021/ils.unc.edu/callee/p147-woods.pdf; Christopher A. Lee, Kam Woods, Matthew Kirschenbaum, and Alexandra Chassanoff, "From Bitstreams to Heritage: Putting Digital Forensics into Practice in Collecting Institutions," Maryland Institute for Technology in the Humanities, September 30, 2013, 5, accessed April 9, 2019, https://drum.lib.umd.edu/handle/1903/14736.

together with records that have similar forms or functions. This method documents the context of creation for web archives content alongside records of any other format, through their relationship to scope and creator description at the same or higher levels. For a collection of congressional campaign records, the archivist might decide that a section of the candidate's campaign website has similar functions and use as paper mailings, e-mail newsletters, and the campaign's social media posts. A single staff member may have directed the content across each format for similar purposes. The archivist can group these into an aggregate—such as a file series—and describe the use, purpose, and scope of that material at that upper level. This would likely be more effective management of the archivist's finite resources than repeatedly describing context, like the role of the staff member, and siloing each format individually. A user could use that arrangement to discover and access materials that document the campaign's political communications.

Documenting the context of creation through arrangement is useful, but the technical complexity of capturing web content and the potential for discrepancies that it creates still make important information opaque to users. Archivists must also expose detailed *collecting process metadata* to users at the lower levels of abstraction. These are the same lower levels where archivists might manage containers, physical locations, or digital locations. This is useful because archivists might describe multiple pages within a crawl, which can all be linked to the same collecting process metadata, just as two file folders can be linked to the same box with the same location information. This is also practical because a tremendous amount of information is often required to know how and why a web archiving tool captured a specific page in a WARC file.[44] Description systems, such as ArchivesSpace, cannot currently manage granular technical information in a structured or useful way. Instead, this approach enables archivists to link description of content with more detailed collecting process description in another system.

What might this technical provenance information look like? While we should be careful not to place too much emphasis on Archive-It, the relative openness of the data in the Archive-It web application provides a useful example of what provenance data might mean in practice, particularly for Heritrix-style web crawling. The Archive-It Partner Data Application Programming Interface (API) now makes data on individual collections, Dublin Core metadata, seeds, crawls, scoping rules, and more, available for other uses.[45] Focusing on crawl-level information enables us to narrow down what data is useful to document the provenance of a crawl. There are four categories of data that are useful: the type of crawl, crawl time stamps, crawl limits, and crawl results. Crawl types contain the Archive-It crawl identifier, and both Boolean and string fields describing the crawl's recurrence and if it was a test crawl or a PDFs-only crawl. Here, a standard authority for crawl tools and types would be helpful. Then, only stating that this is an "Archive-It Standard Weekly Crawl" and a version number might suffice, as users could use the authority to learn more about what that entails. The crawl time stamps

[44] Andrew N. Jackson, "The Provenance of Web Archives." Similarly, Littman et al. described how "the sheer quantity of potential provenance metadata can be overwhelming." Littman et al., "API-Based Social Media Collecting," 27.
[45] Jillian Lohndorf, "Archive-It Access Integrations," accessed April 10, 2019, https://web.archive.org/web/20190410170654/https://support.archive-it.org/hc/en-us/articles/360001231286-Archive-It-Access-Integrations. The Partner Data API is not documented by Archive-It, but information on using certain queries is available here: UAlbany Archives, "Describing Web Archives," https://github.com/UAlbanyArchives/describingWebArchives.

```
Crawl Type:
crawl: 304306
type: WEEKLY
recurrence_type: WEEKLY
pdfs_only: False
test: False

Crawl Time Stamps:
start_date: 2017-06-01T13:56:34Z
original_start_date: 2017-06-01T13:56:34Z
last_resumption: None
processing_end_date: 2017-06-03T16:04:52Z
end_date: 2017-06-03T15:51:53Z
elapsed_ms: 179669650

Crawl Limits:
time_limit: 259200
document_limit: None
byte_limit: None
crawl_stop_requested: None
Scoping Rules:
   Limit host twitter.com to 500 documents
   Limit host twimg.com to 100 documents
   Block host accounts.google.com

Crawl Results:
status: FINISHED
discovered_count: 123390
novel_count: 62424
duplicate_count: 60966
resumption_count: 0
queued_count: 0
downloaded_count: 123390
download_failures: 1575
warc_revisit_count: 60966
warc_url_count: 123363
total_data_in_kbs: 4086366
duplicate_bytes: 3278106680
warc_compressed_bytes: 413854915
```

Figure 2: Example of provenance information for a web crawl extracted from the Archive-It Partner Data API

document the start time, elapsed time, end time, and if the crawl was resumed. Crawl limits document the boundaries of the crawler, including both the data or document limits for the crawl, and scoping rules that were applied. Scoping rules and other limits also have the potential for standardization. Crawl results include the reason the crawl concluded—by either a data or document limit or reaching the end of the links—and the quantities of pages and data that were and were not captured. The crawl results are particularly useful to users trying to decipher how a page was included or excluded in a WARC. A crawl limited by data or a document count might have a number of "queued" pages that were not captured. The crawl information will also show how many pages were captured in a previous crawl and may be available, but were not included in this specific WARC. It also shows how many pages may have simply failed to download.

In addition to technical provenance information, it might be also be useful to include the archivist's intended purpose as part of the collecting process metadata. In Archive-It, archivists could include an appraisal note as a user-defined field in the Dublin Core metadata attached to the crawl's seed or collection. While archival description already manages this information quite well, pages that were not seeds may not have been actively crawled for a reason specific to that page. Here, the appraisal decision is actually made at the crawl level. Archivists also can extract the appraisal information from the crawl level and add it to the higher-level archival description, and it is feasible automate this using the Partner Data API.

Much of the collecting process metadata discussed here can be crosswalked to the "decision space" elements proposed by Maemura et al. The qualitative appraisal note could include the "motivation," "focus," and "access and discovery" elements, while the crawl time stamps and crawl times would be similar. Standardized tools and capture profiles would include crawl configuration, and scope or limit authorities would encompass inclusions, exclusions, and permissions. Crawl results details would include process elements, and higher-level description and repository information would contain context elements.[46]

While aggregate description and the Archive-It Partner Data API provide archivists with a sustainable path to make valuable detailed collecting process information available to users, substantial work remains before archivists can reasonably implement this approach. Web archiving tools must change over time, such as the data model for the Archive-It web application, which may change as new features are incorporated. Furthermore, Archive-It is far from the only web archiving tool, and archivists also need to document provenance information while using Webrecorder, Wget, or other tools that are available now or in the future. Future work should focus on developing authorized tool and capture profiles that categorize and sufficiently explain different methods of capturing web archives. Profiles for scoping rules are also needed. This would ensure interoperability by not requiring archivists to focus on Archive-It data, but instead format and match this data, its future versions, and that of other tools to a standard capture profile. Finally, this approach poses an interesting challenge to DACS and current descriptive practices. While it is practical to manage detailed collecting process metadata at a lower level, the DACS rules prescribe that much of this information applies certain elements at the intellectual level, such as the immediate source of acquisition note. Managing this information like container or location data could potentially reduce or limit its presence in access systems. Still, while in some cases archivists could also describe detailed provenance information in DACS elements, they must question whether it would be feasible or useful to researchers there.

[46] Maemura et al., "If These Crawls Could Talk," 16–17.

Technical Barriers and a Proof of Concept

Librarians and archivists can currently utilize aggregate description for web archives, but this still involves substantial repetitive work. Many archives schedule crawls in Archive-It, and they would have to constantly update some description—such as dates and extents—for new captures. The openness and flexibility of tools like Archive-It and ArchivesSpace make it possible to automate this process, but only in ways that are still infeasible for many archives. A proof of concept Python script demonstrates how this approach would work in practice and provides some insight on how to integrate these tools so that a broader set of repositories can utilize aggregate description for web archives.

Many librarians and archivists have substantial barriers to utilizing technology to advance their work. Unlike other professions that actively manage data, libraries and archives have not historically utilized mainframes or servers, and these tools are not commonly included as part of their basic infrastructure. Librarians and archivists commonly have to request server access, and many never get it. The paranoia of many system administrators also means that many librarians and archivists do not even have the privileges to install and test software at will, limiting both the types of tools they can use, and their ability to explore and learn.

It is feasible to automate the repetitive tasks of describing many web archives captures with open tools such as ArchivesSpace and the Archive-It web application. This is because both of these tools have application programming interfaces, or APIs.[47] This type of API is almost like a website for computers. While both tools have websites that humans can log into and view or manipulate data, APIs present the data from that same web page as code. Another application can then read and manipulate that data just like a human edits and submits a web form. This means that not only can humans create a resource or archival object in ArchivesSpace, or read and edit metadata in Archive-It, but other applications can too.

These APIs make it possible for an outside process to automate updates for human-created description in ArchivesSpace for web crawls using data from Archive-It. The process first requires archivists to scope and schedule crawls using Archive-It, arrange and describe the materials captured using ArchivesSpace, and configure how the external process will create the automated description. Next, an automated process can use the ArchivesSpace API to find all web archives records, and then query the Archive-It CDX API to get a list of captures with their time stamps, crawl identifiers, and checksums. For archives that are able to support a separate system to manage lower-level collecting process data, this process can also ask that system for a link to crawl-level provenance data. The external process can then update the description in ArchivesSpace by changing date and extent records, and possibly adding child archival objects or digital objects for each capture, depending on the configuration. File versions could link directly to the Wayback Machine and/or the crawl-level system. For seeds, the external process can also query the Partner Data API to get any relevant Dublin Core metadata, such as an appraisal note, and add this information to ArchivesSpace as well. A second optional process

[47] The Archive-It API is not very well known or sufficiently documented. As of writing, the Archive-It staff has explained that this is because the application is actively changing. ArchivesSpace API Reference, accessed April 5, 2019, https://archivesspace.github.io/archivesspace/api/.

would query the Partner Data API to extract crawl-level data for each crawl independent of any description. This process could also potentially download WARC files for each crawl.

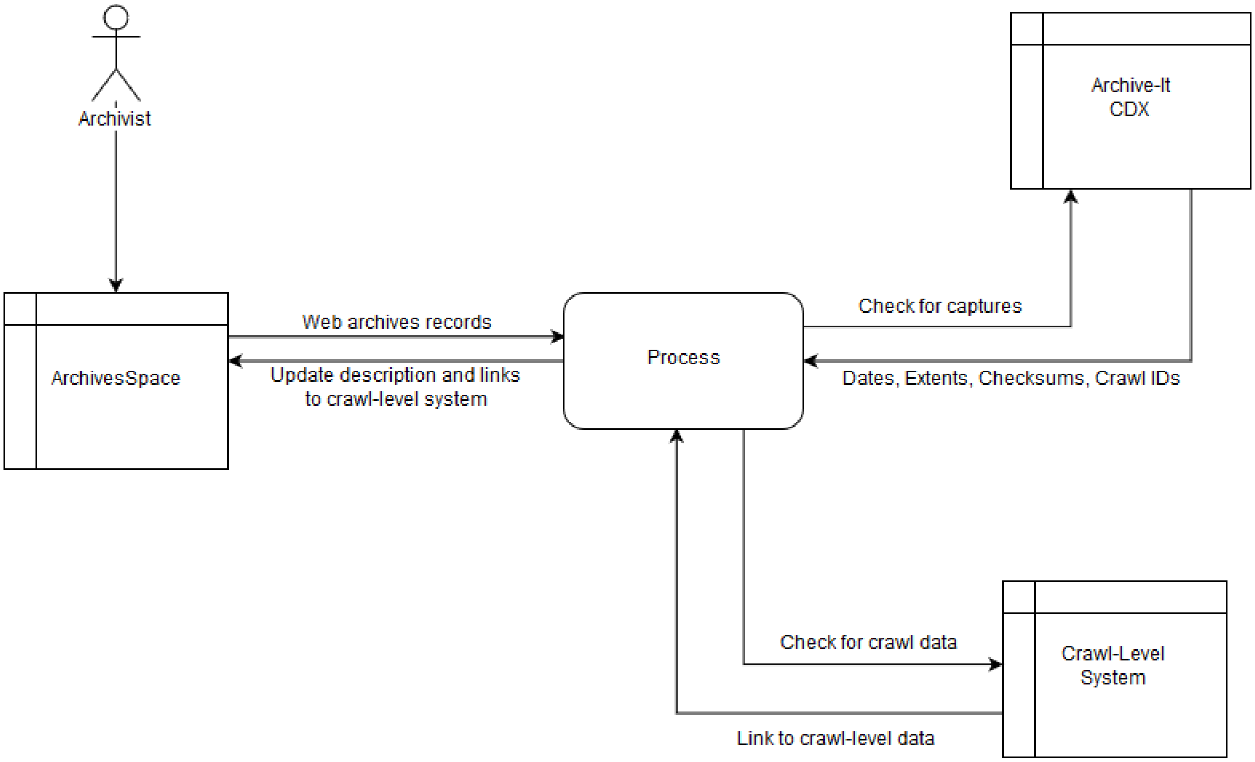

Figure 3: Diagram of external process for augmenting description of web archives

Archivist Steps

1. Crawl and scope collections using the Archive-It web application
2. Describe captured web pages in ArchivesSpace
3. Configure how captures are recorded in ArchivesSpace

Automated Process Steps

4. Get all web archives records in ArchivesSpace
5. Query Archive-It CDX API to get inclusive dates, extents, checksums, and crawl IDs for each unique capture
6. Query optional crawl-level system to check if provenance data is available
7. Add or update description in ArchivesSpace
   - Update dates and extents for web archives records and parent description
   - Possibly add archival objects or digital objects for each capture
   - Add links to crawl-level provenance data if available

Figure 4: Steps required to automate repetitive description of web archives

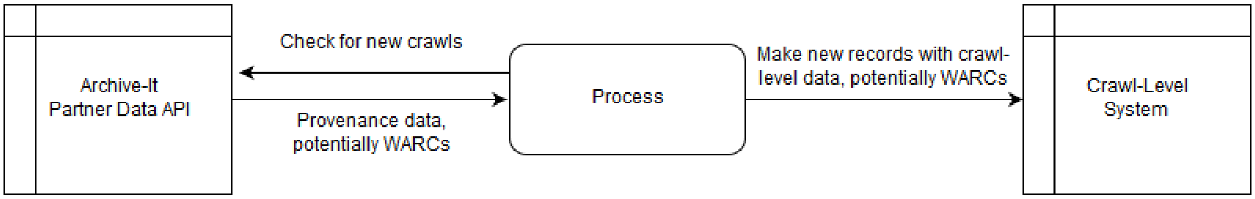

Figure 5: Diagram of second optional external process to query the Archive-It Partner Data API to store crawl-level provenance data and make it available to users

Steps for Second Optional Automated Process

1. Query the Archive-It Partner Data API to check for new crawls
2. Extract data for crawl type, time stamps, limits, and results
3. Optionally download WARC files for each crawl
4. Create new crawl records in a crawl-level system with technical provenance information

Figure 5: Steps required for a second optional external process to query the Archive-It Partner Data API and store crawl-level provenance data in a separate system

I have written a proof-of-concept Python script that automates description in ArchivesSpace using data contained in the Archive-It CDX and Partner Data APIs.[48] The Python script acts as the external process and automatically adds and updates dates, extents, and provenance information for web archives records. In this version, the script adds provenance information from the Partner Data API to ArchivesSpace in unstructured immediate source of acquisition notes. This turned out to be undesirable because of the complexity of the information, and the fact that the script repeated the data for multiple descriptions of pages that were captured during the same crawl. Running the script automatically requires a server to schedule it as a task or Cron job, or it can be run manually from the command line with any computer that has Python. Only two archival repositories have implemented this script, either for testing or as a short-term solution. Since the tool has substantial limitations and maintenance costs, and does not have a large set of active users or supporters, it should serve only as a proof of concept and archivists must seek more sustainable and user-friendly options for automating aspects of aggregate description for web archives.

[48] “Describing Web Archives,” accessed July 5, 2019, https://web.archive.org/web/20190705191907/https://github.com/UAlbanyArchives/describingWebArchives.

DACS Elements	**Source of Information**
Required	
Reference Code	ArchivesSpace manual entry
Title	ArchivesSpace manual entry
Date	Archive-It CDX
Extent	Archive-It CDX
Name of Creator(s)	ArchivesSpace or Archive-It metadata fields
Scope and Content	ArchivesSpace manual entry
Conditions Governing Access	ArchivesSpace manual entry or Archive-It rights statement
Languages and Scripts of the Material	ArchivesSpace manual entry
Optimal	
Administrative/Biographical History	ArchivesSpace manual entry
Access points (Subjects, etc.)	ArchivesSpace or Archive-It metadata fields
Added Value	
Appraisal	ArchivesSpace or Archive-It metadata fields
Technical Access	URL of web page, manual entry in ArchivesSpace

Figure 6: Metadata crosswalk showing the application of DACS to both human-created and machine-generated metadata

In this example, it is useful to examine how DACS was sufficient to describe a variety of web archive collections at multiple levels of detail. Web archives that were created by a distinct agent and had enough research value to be described as a collection were crawled and scoped in Archive-It and described as an ArchivesSpace resource. The DACS-required title, reference code, scope and content, creator, conditions governing access, and languages note were described manually in ArchivesSpace according to DACS and local guidelines as if it was any other collection of records or manuscripts. The DACS optimal and added value elements, a historical note, subject access points, and an appraisal note were also described manually and the URL for the page being described was included as a technical access note. The Python script then added the DACS-required date and extent notes with data from the Archive-It CDX API. When the script is run in the future, it will update these fields for new captures. It would also be possible to update the script to add some of the DACS elements to ArchivesSpace using the Dublin Core metadata entered in Archive-It. Appraisal notes are perhaps the most appropriate to describe in Archive-It, as they usually apply to the crawl level.

Not all captures, seeds or otherwise, deserve description or are appropriate to describe at the collection level. Many captures that were collections and/or seeds were described in Archive-It as archival objects, often at the series level. A common use case was a website for an organization that had an existing collection of records. In many cases, captures that were not seeds or collections in Archive-It were described as part of existing archival collections, such as course catalogs captured as part of a wider crawl, but described with their paper equivalents. For all of these cases, many of the DACS elements were inherited from higher levels of description. A title and a technical access note with a URL were included, and the Python script added a date and an extent. In this case, the Python script also added child archival objects for each unique capture in later crawls using the checksums found in the Archive-It CDX API. This often resulted in hundreds of child objects for regularly scheduled crawls with description for each individual capture—which may or may not be desirable and should be configurable. The script also demonstrated that it is feasible to extract metadata from the <meta> tags of individual captures using Beautiful Soup and include them as lower-level titles or scope notes.[49] The OCLC WAM was correct in stating that embedded HTML metadata was often incomplete or unhelpful, but while it is not useful to rely on this data as a primary access point, this information can be useful as part of lower-level description of a more detailed higher-level aggregate—such as linking this messy information to detailed series description.[50] In all of these cases, DACS proved to be sufficient in describing web archives in aggregate, aided by automated description extracted from the Archive-It APIs.

Toward Maintainable Integrations

Some significant but feasible alterations to ArchivesSpace and Archive-It as well as a more defined process for integrating them should make it practical for archivists and librarians to use aggregate description for web archives. This could be the case even for those without access to servers and the time and interest to learn Python or the command line. Most importantly, there must be an external process to update ArchivesSpace records that is feasible for many archives to run. Running that process from the Archive-It application or developing a basic desktop application may be workable options. Additionally, there is currently no ideal field within ArchivesSpace with an API endpoint to denote to another system that it is a web archives record. Finally, there has to be a place to store configuration data for the process, in either ArchivesSpace, Archive-It, or serialized on a local file system.

Archivists have an interest in keeping their tools maintainable, as this reduces the amount of time, labor, and subscription fees we collectively spend on keeping them functional. One way of

[49] Beautiful Soup documentation, accessed April 11, 2019, https://web.archive.org/web/20190326123119/https://www.crummy.com/software/BeautifulSoup/.
[50] Dooley and Bowers, *Descriptive Metadata*, 13–14.

ensuring this is to keep tools simple and focused on a few specific core functions. While it is understandable for archives to desire the inclusion of new functionality in their existing tools, it can often be more effective to separate tools by function and envision a network of simple, interoperable systems, instead of one or more large monolithic systems that become challenging to manage. ArchivesSpace is already a large and complex system. While the ArchivesSpace program staff has focused on making the software appear simple to set up and implement, the system actually contains multiple different web applications that rely on different frameworks, all packaged into one. While ArchivesSpace is open and customizable, archivists have an interest in keeping it simple and focused. Since extracting metadata for web archives is not part of ArchivesSpace's core purpose, relying on an external process to update ArchivesSpace records is likely to be a simpler and more maintainable path than including that functionality in ArchivesSpace.

Still, since there is no ideal way to identify web archives records in ArchivesSpace, it may be useful to make minor alterations to the ArchivesSpace data model to allow this. The two required points of data are a way to designate web archives records with an API end point and a URL. Since URLs might be considered description, it might be appropriate to include them as a physical characteristics and technical requirements note, with a specific label. A field denoting that the resource or archival object is a web archives record should not be considered description. The current proof-of-concept Python script uses a local subject called "Web Collection." Earlier efforts experimented with using the external documents field, yet this does not have an API end point, so the external process would not be able to request all web archives records. Instead, it would have to request all records and sort through them, an unnecessary increase in the required processing power. There are a few different choices to denote web archives records in ArchivesSpace, with varying levels of difficulty. One would be to simply add an API end point for external resources. Another option would be to add a new field to archival objects and resources, or add another instance type specifically for web archives links. Updating the model for digital objects may appear to be a good option, but the required field is really only a link that denotes a web archive, and it would be inconsistent to have one digital object that denotes a web archive, and others that describe specific captures. Since all of these methods have drawbacks, it is also possible that using a local subject is most suitable, but it would be more ideal to have a field that is more consistent with its intended use.

The most important requirement to create a sustainable integration for describing web archives with ArchivesSpace and Archive-It is an external process to query the APIs, process the data, and create and update description. Building on the existing Python script, it is possible to create a desktop application with a graphical user interface (GUI) that archives could use without requiring a server. Perhaps an Electron application would be a manageable route.[51] Another option would be to integrate this functionality into Archive-It. While many of the underlying Archive-It systems are quite complex, it could be feasible to add this functionality to the

[51] Electron, accessed April 12, 2019, https://electronjs.org; Jason Ronallo, "Building Desktop Applications Using Web Technologies with Electron," *Code4Lib*, Philadelphia, PA, March 9, 2016, https://web.archive.org/web/20170505102209/2016.code4lib.org/Building-Desktop-Applications-using-Web-Technologies-with-Electron.

Archive-It front end, which is a Django web application.[52] ArchivesSnake, a Python client library for the ArchivesSpace API, could help facilitate the integration.[53] This would make it practical for Archive-It partners to describe their web archives in ArchivesSpace, and if the process were open and documented well, it would be possible for other web archiving vendors to replicate.

The final requirement to allow a wider set of web archives practitioners to describe web archives in aggregate is a persistent location to store configuration data for the external process. Once archivists describe web archives in ArchivesSpace, the configuration would specify how the external process creates automated description. It could create new child archival objects for each capture, create new digital object instances, or only update date and extent notes. For an Electron desktop application, this could simply be stored on the local file system. For a process run though the Archive-It application, it may be ideal to store these settings in ArchivesSpace. This would allow archivists to define how records are created or updated as they describe web archives. However, storing the configuration in ArchivesSpace would require more substantial changes than only adding API end points.

Conclusion

Currently web archives are not easily discoverable by users. Web collections are difficult to find and most users access web archives though the Wayback Machine and other Internet archive systems, which only expose certain information and are limited to only certain types of use. Even when librarians and archivists add detailed metadata to web archives—such as the Dublin Core elements in the Archive-It application—these materials are siloed off, away from the rest of the repositories' holdings. The acquisitions of web archives are also not transparent to users, who require detailed information on the collecting process to adequately assess the meaning of their findings.

Describing web archives in aggregations in accordance with DACS helps web archives practitioners address many of these challenges. Archival arrangement and description enables archivists to describe the most meaningful captures from web archives at a higher level of abstraction—whether or not they were seeds. This provides access to web archives alongside records of other formats that have common creators, forms, or functions. Aggregate description also empowers archivists to describe records at whatever level of detail is most effective given their finite time and labor instead of describing only what was useful to crawl from a technical perspective. Managing description at the higher intellectual level also provides a path for archivists to provide access to web archives provenance data. If a repository is able to support a separate system, detailed collecting process metadata—such as when the crawl started, scoping rules, and the crawl result—can be extracted from the Archive-It Partner Data API. If web archives practitioners can define a standard set of profiles for crawls and scoping rules, this information could be regularized for other web archives collecting tools. Archivists will then be

[52] Karl-Rainer Blumenthal, "Archive-It 5.0 Release Notes," 2017, accessed April 12, 2019, https://web.archive.org/web/20170628161340/https://support.archive-it.org/hc/en-us/articles/208110946-Archive-It-5-0-Release-Notes.

[53] ArchivesSnake, accessed April 12, 2019, https://github.com/archivesspace-labs/ArchivesSnake.

able to describe web archives independently from the technical processes of web crawling or web recording, but technical provenance information will still be available to users when they need it.

It is currently very challenging for most repositories to describe web archives in aggregate, as updating records for every new crawl is often unfeasible. The ArchivesSpace and Archive-It APIs have made it possible to automate repetitive description. However, the existing proof of concept requires skills and technologies that are not typically available to many archives. With some alterations to the current tools, it is feasible to create an integration between ArchivesSpace and Archive-It that allows web archives practitioners to describe web archives in aggregates in ArchivesSpace, while an automated integration creates and updates description for new captures. If librarians and archives are able to describe web archives in aggregates alongside related collections, more web archives will be accessible to users in more familiar places.

References

"About Us." *Archive-It Blog*. Accessed April 13, 2019. https://web.archive.org/web/20190328195120/https://archive-it.org/blog/learn-more/.

"Archive It Integration." Archivesspace-DO-Plugin. May 14, 2018. https://web.archive.org/web/20190414143424/https://github.com/NYULibraries/Archivesspace-DO-Plugin/wiki/Archive-It-Integration.

ArchivesSnake. Accessed April 12, 2019. https://github.com/archivesspace-labs/ArchivesSnake.

ArchivesSpace API Documentation. Accessed April 5, 2019. https://archivesspace.github.io/archivesspace/api/.

"ArchivesSpace Public Interface Enhancement Project." Accessed May 21, 2019. https://archivesspace.atlassian.net/wiki/spaces/ADC/pages/22282254/Public+Interface+Enhancement+Project.

"Archiving the Websites of Contemporary Composers." *Archive-It Blog.* Accessed April 14, 2019. https://archive-it.org/blog/projects/composers/.

ArcLight. Accessed May 21, 2019. https://web.archive.org/web/20171024160933/https://wiki.duraspace.org/display/samvera/ArcLight.

Arendt, Anne, and Nathan Gerber. "Dispersed Web Content Management in Higher Education." *Educause Review*. July 30, 2009. https://web.archive.org/web/20190402200808/https://er.educause.edu/articles/2009/7/dispersed-web-content-management-in-higher-education.

Bearman, David. "Documenting Documentation." *Archivaria* 34 (Summer 1992): 33–49.

Bearman, David A., and Richard H. Lytle. "The Power of the Principle of Provenance." *Archivaria* 21 (Winter 1985–86): 14–27.

Beautiful Soup Documentation. Accessed April 11, 2019. https://web.archive.org/web/20190326123119/https://www.crummy.com/software/BeautifulSoup.

Berner, Richard C. "Archivists, Librarians, and the National Union Catalog of Manuscript Collections." *The American Archivist* 27, no. 3 (July 1964): 401–9.

——— "Observations on Archivists, Librarians, and the National Union Catalog of Manuscript Collections." *College and Research Libraries* 29, no. 4 (1968): 276–80.

Blumenthal, Karl-Rainer. "Archive-It 5.0 Release Notes." 2017. https://web.archive.org/web/20170628161340/https://support.archive-it.org/hc/en-us/articles/208110946-Archive-It-5-0-Release-Notes.

Callahan, Maureen. "On Containers." *Chaos --> Order*. December 15, 2014. https://web.archive.org/web/20160404091903/https://icantiemyownshoes.wordpress.com/2014/12/15/on-containers/.

"Cannotsleepwithsnoringhusband." Accessed April 1, 2019. https://web.archive.org/web/20180310234259/cannotsleepwithsnoringhusband.online/.

Chapman, Joyce Celeste. "Observing Users: An Empirical Analysis of User Interaction with Online Finding Aids." *Journal of Archival Organization* 8 (2010): 4–30.

Costea, Maria-Dorina. "Report on the Scholarly Use of Web Archives." NetLab, 2018. https://web.archive.org/web/20190416113523/netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf/.

DeRidder, Jody, Amanda Presnell, and Kevin Walker. "Leveraging Encoded Archival Description for Access to Digital Content: A Cost and Usability Analysis." *The American Archivist* 75, no. 1 (Spring/Summer 2012): 143–70.

Describing Archives: A Content Standard (DACS). 2nd ed. Chicago: Society of American Archivists, 2013. https://web.archive.org/web/20190307035606/http://files.archivists.org/pubs/DACS2E-2013_v0315.pdf. Most current version, https://web.archive.org/web/20190705170945/https://github.com/saa-ts-dacs/dacs.

"Describing Web Archives." UAlbany Archives. Accessed July 5, 2019. https://github.com/UAlbanyArchives/describingWebArchives.

Dooley, Jackie, and Kate Bowers. *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. Dublin, OH: OCLC Research, 2018.

Dooley, Jackie M., Karen Stoll Farrell, Tammi Kim, and Jessica Venlet. "Developing Web Archiving Metadata Best Practices to Meet User Needs. *Journal of Western Archives* 8, no. 2 (2017): 1–14.

Eastwood, Terry. "Putting the Parts of the Whole Together: Systematic Arrangement of Archives." *Archivaria* 50, no. 1 (Fall 2000): 93–116.

Electron. Accessed April 12, 2019. https://electronjs.org.

Espenschied, Dragan. "The Ethics of Digital Folklore." National Forum on Ethics & Archiving the Web. New York, NY. March 23, 2018. https://vimeo.com/276948412.

Evans, Max. "Authority Control: An Alternative to the Records Group Concept." *The American Archivist* 49, no. 3 (Summer 1986): 249–61.

Farrell, Matthew, Edward McCain, Maria Praetzellis, Grace Thomas, and Paige Walker. "Web Archiving in the United States: A 2017 Survey." Washington, DC: National Digital Stewardship Alliance, 2017. https://osf.io/ht6ay/.

Gilliland-Swetland, Anne J. "Popularizing the Finding Aid: Exploiting EAD to Enhance Online Discovery and Retrieval in Archival Information Systems by Diverse User Groups." *Journal of Internet Cataloging* 4, nos. 3–4 (2001): 199–225.

Graham, Pamela M. "Guest Editorial: Reflections on the Ethics of Web Archiving." *Journal of Archival Organization* 14, nos. 3–4 (July-December 2017): 103–10.

"Guide to the Mike Topp Papers." Fales Library and Special Collections. New York University. Accessed April 14, 2019. https://web.archive.org/web/20190414154500/http://dlib.nyu.edu/findingaids/html/fales/mss_188/dscaspace_fd1a024cc9bb851fc2b7a610b336664b.html.

"Guide to the University Archives Web Archive Collection." Duke University Archives. Accessed April 19, 2019, http://library.duke.edu/rubenstein/findingaids/uawebarchive/.

"Guide to the University of Chicago Web Archive Collection." Special Collections Research Center. University of Chicago Library. Accessed April 14, 2019. https://web.archive.org/web/20190414160018/www.lib.uchicago.edu/e/scrc/findingaids/view.php?eadid=ICU.SPCL.UCWEB.

Hensen, Steven L. "'NISTF II' and EAD: The Evolution of Archival Description." *The American Archivist* 60, no. 3 (Summer 1997): 284–96.

IFLA Study Group on the Functional Requirements for Bibliographic Records. "Functional Requirements for Bibliographic Records: Final Report (FRBR)." February 2009. https://web.archive.org/web/20190331171302/https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

ISAD(G): General International Standard Archival Description: Adopted by the Ad Hoc Commission on Descriptive Standards, Stockholm, Sweden, 21–23 January 1993; Final ICA Approved Version. Ottawa: Secretariat of the Ad Hoc Commission on Descriptive Standards, 1994. https://web.archive.org/web/20150706032159/http://www.hi.u-tokyo.ac.jp:80/personal/yokoyama/jugyo99/isad(g)e.html.

Jackson, Andrew N. "The Provenance of Web Archives." *UK Web Archive Blog*. November 20, 2015. https://web.archive.org/web/20180123212308/blogs.bl.uk/webarchive/2015/11/the-provenance-of-web-archives.html.

Jackson, Andrew N., Jimmy Lin, Ian Milligan, and Nick Ruest. "Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities." *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* 16 (2016): 103–6. https://yorkspace.library.yorku.ca/xmlui/handle/10315/31236.

Jackson, Tracy M. "I Want To See It: A Usability Study of Digital Content Integrated into Finding Aids." *Journal for the Society of North Carolina* Archivists 9, no. 2 (2012): 20–77.

Jones, Shawn M., Alexander Nwala, Michele C. Weigel, and Michael K. Nelson. "The Many Shapes of Archive-It: Characteristics of Archive-It Collections." *Proceedings of the 15th International Conference on Digital Preservation* (2018): 1–10. https://arxiv.org/abs/1806.06878.

Land, Robert H. "The National Union Catalog of Manuscript Collections." *The American Archivist* 17, no. 3 (July 1954): 195–207.

Lee, Christopher A., Kam Woods, Matthew Kirschenbaum, and Alexandra Chassanoff. "From Bitstreams to Heritage: Putting Digital Forensics into Practice in Collecting Institutions." Maryland Institute for Technology in the Humanities. September 30, 2013. https://drum.lib.umd.edu/handle/1903/14736.

Lialina, Olia. "Do you believe in user 711391? A Search Engine Drama." *Rhizome*. March 7, 2016. https://web.archive.org/web/20180816160522/rhizome.org/editorial/2016/mar/07/do-you-believe-in-user-711391/.

Light, Michelle, and Tom Hyry. “Colophons and Annotations: New Directions for the Finding Aid.” *The American Archivist* 65 (Fall-Winter 2002): 216–30.

Littman, Justin, Daniel Chudnov, Daniel Kerchner, Christie Peterson, Yecheng Tan, Rachel Trent, Rajat Vij, and Laura Wrubel. “API-Based Social Media Collecting as a Form of Web Archiving.” *International Journal on Digital Libraries* 19, no, 1 (March 2018): 27.

Lohndorf, Jillian. “Archive-It Access Integrations.” Accessed April 10, 2019. https://web.archive.org/web/20190410170654/https://support.archive-it.org/hc/en-us/articles/360001231286-Archive-It-Access-Integrations.

Maemura, Emily, Nicholas Worby, Ian Milligan, and Christoph Becker. “If These Crawls Could Talk: Studying and Documenting Web Archives Provenance.” *Journal of the Association for Information Science & Technology* 69, no. 10 (October 2018): 1223–33. https://tspace.library.utoronto.ca/handle/1807/82840.

Milligan, Ian. “Researcher Access for Web Archives: Our Experiences.” Mid-Atlantic Regional Archives Conference, Buffalo, NY, October 2017.

Mink Family Restaurants Records. Temple University Special Collections Research Center. Accessed April 14, 2019. https://web.archive.org/web/20180624131829/https://library.temple.edu/scrc/mink-family-restaurants.

“New ArchivesSpace Integration.” *Archive-It Blog.* Accessed April 14, 2019. https://web.archive.org/web/20190407235818/https://archive-it.org/blog/post/archivesspace-integration/.

Pearce-Moses, Richard, and Joanne Kaczmarek. “An Arizona Model for Preservation and Access of Web Documents.” *DttP: Documents to the People* 33, no. 1 (Spring 2005): 17–24.

Peterson, Christie. “Archival Description for Web Archives.” *Chaos —> Order*. June 12, 2015. https://web.archive.org/web/20180307061542/https://icantiemyownshoes.wordpress.com/2015/06/12/archival-description-for-web-archives/. Reposted in *On Archivy*. June 22, 2015. https://web.archive.org/web/20170604110440/https://medium.com/on-archivy/archival-description-for-web-archives-1d9dce8dcef0.

Pitti, Daniel V. “The Berkeley Finding Aid Project.” https://web.archive.org/web/20180206174035/http://archive1.village.virginia.edu/dvp4c/arlpap.htm.

Powell, Andy, Mikael Nilsson, Ambjörn Naeve, Pete Johnson, and Tom Baker. “DCMI: DCMI Abstract Model.” June 4, 2007. https://web.archive.org/web/20190504123346/www.dublincore.org/specifications/dublin-core/abstract-model/.

Prom, Christopher J. “User Interactions with Electronic Finding Aids in a Controlled Setting.” *The American Archivist* 67, no. 2 (Fall/Winter 2004): 234–68.

“Records of the Association of American Women in Europe.” Harvard University Archives. Accessed April 14, 2019. https://web.archive.org/web/20190414155410/https://hollisarchives.lib.harvard.edu/repositories/8/archival_objects/2910347.

“Response to Best Practices for Web Archiving Metadata.” Society of American Archivists’ Technical Subcommittee on Describing Archives: A Content Standard. June 12, 2017. https://web.archive.org/web/20190414170439/https://docs.google.com/document/d/1x5BuGYdtdjfVvbnXfbTDt7VL2ETfFwI5IfH-GUrs57U/edit.

Rester, Aaron. "Web in the Higher Ed Org Chart." Accessed April 2, 2019. https://web.archive.org/web/20190402194959/https://blog.aaronrester.net/2013/04/web-in-the-higher-ed-org-chart.html.

Roe, Kathleen D. *Arranging & Describing Archives & Manuscripts*. Archival Fundamentals Series 2. Chicago: Society of American Archivists, 2005.

Ronallo, Jason. "Building Desktop Applications Using Web Technologies with Electron." *Code4Lib*. Philadelphia, PA. March 9, 2016. https://web.archive.org/web/20170505102209/2016.code4lib.org/Building-Desktop-Applications-using-Web-Technologies-with-Electron.

Santamaria, Daniel A. *Extensible Processing for Archives and Special Collections: Reducing Processing Backlogs*. Chicago: Neal-Schuman, 2015.

Thomas, Grace. "More Web Archives, Less Process." *The Signal*. August 3, 2018. https://web.archive.org/web/20180803194118/blogs.loc.gov/thesignal/2018/08/more-web-archives-less-process.

"University of North Carolina at Chapel Hill University Archives Collected Websites." University of North Carolina at Chapel Hill. Accessed April 14, 2019. https://web.archive.org/web/20180529115702/http://finding-aids.lib.unc.edu/40417/.

Woods, Kam, and Christopher A. Lee. "Acquisition and Processing of Disk Images to Further Archival Goals." *Proceedings of Archiving 2012.* Springfield, VA: Society for Imaging Science and Technology, 2012. https://web.archive.org/web/20170826185021/ils.unc.edu/callee/p147-woods.pdf.

Yakel, Elizabeth. "Listening to Users." *Archival Issues* 26, no. 2 (2002): 111–27.