Journal of Western Archives

Volume 8
Issue 2 *Web Archiving Special Issue*

Article 5

2017

Developing Web Archiving Metadata Best Practices to Meet User Needs

Jackie M. Dooley
OCLC Research, dooleyj@oclc.org

Karen Stoll Farrell
Indiana University, karsfarr@indiana.edu

Tammi Kim
University of Nevada, Las Vegas, tammiyk@gmail.com

Jessica Venlet
Massachusetts Institution of Technology, jmvenlet@gmail.com

Follow this and additional works at: http://digitalcommons.usu.edu/westernarchives

Recommended Citation
Dooley, Jackie M.; Farrell, Karen Stoll; Kim, Tammi; and Venlet, Jessica (2017) "Developing Web Archiving Metadata Best Practices to Meet User Needs," *Journal of Western Archives*: Vol. 8 : Iss. 2 , Article 5.
Available at: http://digitalcommons.usu.edu/westernarchives/vol8/iss2/5

This Article is brought to you for free and open access by the Journals at DigitalCommons@USU. It has been accepted for inclusion in Journal of Western Archives by an authorized administrator of DigitalCommons@USU. For more information, please contact dylan.burns@usu.edu.

UtahStateUniversity
MERRILL-CAZIER LIBRARY

Developing Web Archiving Metadata Best Practices to Meet User Needs

Jackie M. Dooley
Karen Stoll Farrell
Tammi Kim
Jessica Venlet

ABSTRACT

The OCLC Research Library Partnership Web Archiving Metadata Working Group was established to meet a widely recognized need for best practices for descriptive metadata for archived websites. The Working Group recognizes that development of successful best practices intended to ensure discoverability requires an understanding of user needs and behavior. We have therefore conducted an extensive literature review to build our knowledge and will issue a white paper summarizing what we have learned. We are also studying existing and emerging approaches to descriptive metadata in this realm and will publish a second report recommending best practices. We will seek broad community input prior to publication.

Two recent surveys of users and managers of archived websites have shown that lack of a common approach to creating metadata is the most widely shared challenge for this community.[1, 2] In response, OCLC Research established a Web Archiving Metadata Working Group (WAM) to develop descriptive metadata best practices.[3] At the group's first meeting in January 2016, we recognized that it would be inadvisable to develop best practices for descriptive metadata without first gaining a clear understanding of user needs and behavior in this context. We are taking this into account throughout the project.

1. Ricky Erway, "Thoughts from Partner Staff about Web Archiving," hangingtogether.org, October 29, 2015, http://hangingtogether.org/?p=5450 (accessed January 18, 2017).

2. A research team led by Matthew Weber at Rutgers University surveyed users of web archives in the winter of 2016. They expect to publish their data late in 2016.

3. "Web Archiving Metadata Working Group," OCLC Research, last modified March 10, 2016, http://oc.lc/wam (accessed January 18, 2017).

We will produce three separate reports (described further below), chief among them, a set of best practices for Web archiving metadata that is intended to be community-neutral, software-neutral, and output-neutral. The guidelines will recommend a set of data elements and provide content definitions (i.e., a data dictionary) that can be used by any type of institution in any metadata context. We recognize that needs may vary across the many types of repository that do Web harvesting. Thus, we will strive to define elements in an agnostic manner.

This paper reports on work in progress that began in January 2016. As such, it does not document final outcomes. That said, as of December 2016, we have completed much of our work other than writing the reports and selecting the final data dictionary. Therefore, we do not expect our conclusions from the two literature reviews, nor of the standards and guidelines we studied, to change in any significant way, although the reports will express them in more detail.

> **CHARGE:** The OCLC Research Library Partnership Web Archiving Metadata Working Group will evaluate existing and emerging approaches to descriptive metadata for archived websites and will recommend best practices to meet user needs and to ensure discoverability and consistency.

We have taken a multi-pronged approach, with four subgroups working in tandem:

- User needs literature review: analyzing and synthesizing selected literature that addresses user needs and behaviors as it relates to metadata needs
- Metadata readings literature review: doing the same for selected literature on approaches being taken or proposed for Web archiving metadata
- Metadata guidelines: studying and comparing three relevant descriptive standards and seven sets of institution-specific metadata guidelines
- Tools: investigating Web archiving tools that include some capability to export metadata for re-use

The working group will issue three outputs, to be published by OCLC Research in the first half of 2017:

1. A report containing our evaluations of eleven Web archiving tools will provide a view of current tools that include metadata-related features.
2. A report on user needs and behaviors will inform community-wide understanding of documented user needs and behaviors as evidence to underlie best practices for metadata for archived websites.

3. Metadata best practices will enable practitioners to create appropriate metadata consistently and with confidence that they are following community practice. It is our hope that consistent application will help users benefit from enhanced discoverability of archived websites.

User Needs and Behaviors Literature Review

Our first step was to gather relevant literature on user needs and behavior as it relates to metadata needs. Because Web archiving is rapidly evolving and growing, we cast a wide net to identify not only published articles and reports but also blog posts, surveys, conference papers, and slide decks, thereby providing a bleeding-edge overview of the literature on metadata-related user needs. Reading, abstracting, and synthesizing 25 relevant readings has given us a firm foundation to ensure that our outcomes are in concert with current thinking.

Our synthesis revealed two broad themes: characterizing users and areas of need.

Characterizing users

A variety of authors have identified use cases by examining specific Web archives initiatives and how researchers use them. Our findings to date reveal many similarities across the Web archives literature that defines user groups. This suggests that some predominant information-seeking behaviors and research methodologies are employed by Web archives users.[4] On the other hand, the literature focuses on what might be considered "academic" users; we found no readings focused on other possible user groups such as journalists or members of the public.

Users can be characterized by the types of research questions they ask, and therefore by the different approaches and needs that they bring to Web archives. The reviewed literature reflects four types of researchers who are using Web archives: academic researchers, legal researchers, digital humanists and data analysts, and Web and computer scientists. Overlaps in user needs exist among these four, but it is also possible to consider each group distinctly and to explore how their needs may differ.

Some use Web archives primarily for a website's content; an example could be a historian examining an organization's website over time. Others need extensive data sets, such as a digital humanist doing large-scale data analysis incorporating data from multiple sites. Similarly, Miguel Costa and Mário Gomes place users of Web archives into three behavioral groups: navigational (looking for a specific website), informational (looking for information on a subject), and transactional (trying to

4. Miguel Costa and Mário J. Silva, "Characterizing Search Behavior in Web Archives," (paper presented at the Temporal Web Analytics Conference, 2011), http://ceur-ws.org/Vol-707/TWAW2011-paper5.pdf (accessed January 18, 2017).

acquire something from the Web, such as downloading a copy of a website, which is often a first step to combining data across Web archives).[5]

Areas of need

Some areas of need cut across the identified groups of users, while others are primarily a concern for only one of them. For example, the *lack of interoperability* across Web archives creates issues for users who want to aggregate and analyze content from sites that are preserved in multiple locations. This is likely to come up as a regular need for digital humanists and data analysts but does not appear to be of major concern for those who are more interested in the content of a specific website, such as legal researchers.

The formatting and organization of data within individual Web archives is an issue for academics in the humanities and social sciences who are beginning to make forays into the world of Web archives; they may need specialized training to access Web archives that do not have a user-friendly access layer. This creates a need for more user support services to help researchers grapple with the complexities of accessing and using Web archives, as well as for more effective discovery tools.[6] One specific need is for access to federated subject-based searches across multiple archives rather than only URL-based discovery systems.[7, 8]

Authors describing specific user needs speak to the need for advocacy work to connect Web archives users with the archivists and librarians who are actively working in this field. *Collaboration* is needed in multiple arenas: researchers' contributions to institutional Web archives, researchers collaborating with each other to build tailored archives, and interinstitutional collaboration to avoid duplicative work and to improve standardization of Web archives. The other advocacy piece evident in the literature is *outreach*. Authors note that libraries and archives should

5. Miguel Costa and Mário J. Gomes, "Understanding the Information Needs of Web Archive Users," (Proceedings of the 10th International Web Archiving Workshop, 2010), http://xldb.di.fc.ul.pt/xldb/publications/costa2010understandingneeds_document.pdf (accessed January 18, 2017).

6. Patrick Galligan. "WARCs! What Are They Good For? Researchers!" Bits & Bytes: News from Rockefeller Archive Center's Digital Team, April 28, 2016, http://blog.rockarch.org/?p=1502 (accessed August 15, 2016).

7. David Cruz and Daniel Gomes, "Adapting Search User Interfaces to Web Archives," http://sobre.arquivo.pt/sobre/publicacoes-1/Documentos-acerca-do-Arquivo.pt/adapting-search-user-interfaces-to-web-archives (accessed January 18, 2017). Megan Dougherty 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 and Technology* 65, no. 11 (2014).

8. Taking a big step forward, the Internet Archive released a beta revision of the Wayback Machine early in December 2016 that enables word searches rather than only URL searches. Vinay Goel, "Beta Wayback Machine – Now with Site Search," Internet Archive Blogs, October 24, 2016, https://blog.archive.org/2016/10/24/beta-wayback-machine-now-with-site-search/ (accessed January 18, 2017).

engage actively in outreach to both current and potential Web archives users. Basic needs could be met by informing users about the nature of Web archives and how to find and use them. Targeted outreach toward current users could include identifying specific needs for support services, introducing users to existing tools, and assisting institutions in their ongoing work toward standards, flexibility, and interoperability across Web archives.[9]

One final area of need has been found to encompass all Web archives user groups. Multiple authors identify a scarcity of what they describe as *provenance information* as a critical missing piece. As described in our readings, these include background on the site creator, how the site has changed over time, why the site was archived, capture dates, and the relevant collection development policy. This appears to be an area of concern that cuts across disciplines and approaches to Web research.[10] Although these characteristics differ in some respects from archivists' traditional understanding of provenance, the desire for contextual information is heartening.

Our forthcoming report on user needs will paint a clear picture of the identified user groups and their metadata needs. We will underscore the importance we placed on studying user needs as a prelude to preparing best practices for descriptive metadata for Web archives.

Metadata Practices Literature Review

The metadata readings subgroup analyzed 25 readings that have contributed to our understanding of the features and challenges that practitioners and scholars are implementing or highlighting for Web archives metadata. This review followed the same methodology described above for the user needs literature in terms of both the range of readings and our approach to analysis.

Two insights emerged: hybrid bibliographic and archival approaches are used across the community in describing Web archives, and the need for sustainable practices, in light of limited staff resources, poses an enormous challenge for metadata creation.

Hybrid approaches

No consistent approach to Web archiving description emerged from our readings. Many practitioners have, however, shared their practices via publications or presentations, and these provide important context. They describe a mix of

9. Cruz and Gomes, "Adapting Search User Interfaces to Web Archives;" Dougherty and Meyer, "Community, Tools, and Practices in Web Archiving."

10. Andy Jackson, "The Provenance of Web Archives," British Library UK Web Archive Blog, November 20, 2015, http://britishlibrary.typepad.co.uk/webarchive/2015/11/the-provenance-of-web-archives.html (accessed January 18, 2017).

bibliographic (MARC, Dublin Core, MODS) and archival (finding aids, DACS) approaches. A general acknowledgement emerges that the level of description (collection, website, document) to be used is a key decision point.

Two surveys offer a contrasting picture of the metadata practices of institutions that use Archive-It.[11, 12] Both surveys describe use of the most commonly used Dublin Core elements, but neither explores how institutions define or use these elements (e.g., what type of date is used in the date element?). Sara Mannheimer did not differentiate between levels of description, which she herself noted would have been useful.[13] In contrast, Michelle Sweetser gathered data on the level at which users create metadata and found that many respondents do so only at the collection level.[14] The surveys also explored whether and how practitioners provide access to metadata, with contrasting results: Sweetser found that the majority do not describe collections beyond the Archive-It interface, while Mannheimer reported that many describe collections in finding aids and library catalogs to facilitate access. While both survey samples were relatively small, the data are helpful in establishing some context for metadata creation at a variety of libraries and archives.

Practitioners have written blog posts to promulgate their ideas and experiments. For example, Allison O'Dell provided a detailed description of her hybrid approach to cataloging a Web archive collection as "an online resource, under archival control."[15] She concluded that Web archives should be described at the collection level and cites specific MARC fields, as well as RDA and DACS rules. Christie Peterson blogged about an experiment performed to understand how Web archive description and crawl documentation could map out archival description practices.[16] In a webinar series from the Metropolitan New York Library Council, presenters from Columbia

11. Sara Mannheimer, "Providing Context to Web Collections: A Survey of Archive-It Users," (Master's thesis, University of North Carolina, Chapel Hill, 2013), https://cdr.lib.unc.edu/indexablecontent/uuid:f373e421-0a31-4143-ad65-05137729d894 (accessed January 18, 2017).

12. Michelle Sweetser. "Metadata Practices Among Archive-It Partner Institutions: The Lay of the Land," (slide lecture presented at the Archive-It partners meeting, October 19, 2011), http://slideplayer.com/slide/3847250/ (accessed January 18, 2017).

13. Mannheimer, "Providing Context to Web Collections."

14. Sweetser, "Metadata Practices Among Archive-It Partner Institutions."

15. Allison Jai O'Dell, "Describing web collections (I mean archived websites)," February 17, 2015, https://medium.com/@allisonjaiodell/describing-web-collections-e32b59893848#.lbg88x051 (accessed January 18, 2017).

16. Christie Peterson, "Archival description for web archives," Chaos -> Order, November 5, 2015, https://icantiemyownshoes.wordpress.com/2015/06/12/archival-description-for-web-archives/ (accessed January 18, 2017).

University and the New York Art Resources Consortium shared information about their use of MARC records and finding aids.[17, 18]

Documenting this variety of approaches illuminates practices that can be applied effectively beyond institutional use cases. This will directly pertain as we develop our best practice guidelines.

Sustainable practices and staff resources

Our second insight is the need for sustainable practices that are not out of balance with the human resources available for creating descriptive metadata. For example, in a 2004 article on Australia's PANDORA project, Margaret Philips and Paul Koerbin observed that cataloging is the second most time-consuming part of the workflow, and they questioned the scalability of the work relative to existing funding.[19] Sweetser found that the amount of staff time available is the most important factor in determining which seeds will be described, and at what level of detail.[20] Mannheimer's survey explored staffing and barriers to metadata creation, and she found that the top barrier to metadata creation was lack of staff time, in part because most practitioners have additional job duties and can spend only a small fraction of their time on this work.[21] Some institutions clearly would like to create more granular metadata than is feasible, but the scale of resources needed is daunting.[22] As WAM moves forward to develop best practices, it will be important to take sustainable practices into account.

Analysis of content standards and institutional guidelines

WAM's metadata guidelines subgroup has studied and compared three descriptive content standards and seven institution-specific metadata guidelines, and we have noted three broad patterns:

- Descriptive standards do not address the unique characteristics of websites.

17. Alex Thurman, "Web Archiving: Description and Access," Metropolitan New York Library Council webinar series, February 29, 2016.

18. Lily Pregill, "Web Archiving: Description and Access," Metropolitan New York Library Council webinar series, February 29, 2016, http://www.slideshare.net/ElizabethLilyPregill/web-archiving-description-and-access (accessed January 18, 2017).

19. Margaret E. Philips and Paul Koerbin, "PANDORA, Australia's Web Archive: How Much Metadata Is Enough?," *Journal of Internet Cataloging* 7, no. 2 (2004).

20. Sweetser, "Metadata Practices Among Archive-It Partner Institutions."

21. Mannheimer, "Providing Context to Web Collections."

22. OCLC Research Metadata Managers group. Notes contributed by members on their web archiving metadata practices and needs, January 2016. Not publicly available.

- Institutional metadata guidelines vary widely in the data elements included and how the elements are applied.
- Some record creators follow bibliographic traditions, while others take an archival approach, such as describing multiple sites in one record.
- Hybrid approaches are common.

Descriptive content standards

We studied DACS (*Describing Archives: A Content Standard*), Dublin Core, and the Program for Cooperative Cataloging guidelines for integrating resources described using RDA (*Resource Description and Access*).[23, 24, 25]

DACS is intended principally for describing groups of archival materials, taking a multi-level, hierarchical approach; for example, a description may include data elements describing the entire group of materials, a subset, or individual items. The same data elements can be applied at all levels. The fact that no data element is defined for "publisher" is one of many indicators of the archival orientation of DACS. No rules focus specifically on websites, since DACS is applicable to all types of material. DACS is designed to be output-neutral.

Dublin Core is a vocabulary of fifteen properties (i.e., elements) for use in resource description. It was devised by the Dublin Core Metadata Initiative to facilitate a standard approach to creating simplified metadata for use by any community of practice, including libraries in contexts where the MARC format is too complex to be practical or necessary. It is widely used for describing digital objects, including by Archive-It subscribers.[26] Given that Archive-It is used by a very high percentage of Web archiving implementers, Dublin Core plays a leading role in the

23. "Describing Archives: A Content Standard (DACS), Second Edition," Society of American Archivists, last modified April 22, 2016, http://www2.archivists.org/groups/technical-subcommittee-on-describing-archives-a-content-standard-dacs/dacs#.V45ZcCMrJmA (accessed January 18, 2017).

24. "Dublin Core Metadata Element Set, Version 1.1," Dublin Core Metadata Initiative, last updated November 5, 2013, http://dublincore.org/documents/dces/index.shtml (accessed January 18, 2017).

25. *Integrating Resources: A Cataloging Manual; Appendix A to the BIBCO Participants' Manual and Module 35 of the CONSER Cataloging Manual, 2015 Draft Revision* (Washington, D.C.: Program for Cooperative Cataloging, 2015), http://www.loc.gov/aba/pcc/conser/word/Module35.doc (accessed January 18, 2017).

26. *Add, edit, and manage your metadata* (San Francisco, CA: Internet Archive), https://support.archive-it.org/hc/en-us/articles/208332603-Add-edit-and-manage-your-metadata (accessed January 28, 2017).

creation of Web archiving descriptive metadata.[27] It is also employed by OCLC's digital collections system CONTENTdm.[28]

RDA is the successor to AACR2 (*Anglo-American Cataloging Rules*, 2nd edition) and takes a fully bibliographic approach. We reviewed the Program for Cooperative Cataloging's application of RDA for integrating resources, the scope of which includes websites. While it is tailored for resources that are regularly updated, some of the rules seem not to recognize that it may be unfeasible to edit the metadata every time a site's content changes (e.g., Rule IR2.6.2.3: "Change the title proper to reflect the current iteration of an integrating resource if there is a change on a subsequent iteration"). RDA is closely allied with the MARC 21 bibliographic format.[29]

Institution-specific metadata guidelines

Early in 2016, the working group put out a call for Web archiving descriptive metadata guidelines, and we received seven documents developed by specific institutions and organizations. The subgroup analyzed each of these relative to the standards referenced and the data elements included. Links are provided below for those that that are publicly available.

Archive-It: This widely-used tool for harvesting and describing websites is available by paid subscription from the nonprofit Internet Archive.[30] It is relevant to our scope because it includes metadata guidelines for its users, and it is used by several institutions whose local guidelines we are studying. As described above, its metadata is based on Dublin Core. Fifteen data elements are defined; in addition, users are free to define custom elements.

Columbia University: These guidelines are used by several campus repositories, including the Human Rights Portal, University Archives, Avery Art and Architecture Library, and Burke Divinity Library. Each repository has added several data elements

27. Jefferson Bailey, et al. *Web Archiving in the United States: a №º# Survey*, National Digital Stewardship Alliance, September 2014, http://www.digitalpreservation.gov/documents/NDSA_USWebArchivingSurvey_2013.pdf (accessed January 18, 2017). The National Digital Stewardship Alliance's 2011 and 2013 surveys both showed that approximately 70 percent of respondents use Archive-It. Although the 2015 survey has not yet been published, its data show that this percentage has risen to about 80% due to the withdrawal of the California Digital Library's WAS tool and migration of some of its users to Archive-It.

28. "Build, Showcase and Preserve Your Digital Collections," CONTENTdm, http://www.oclc.org/en-US/contentdm.html (accessed January 18, 2017).

29. "MARC 21 Format for Bibliographic Data: Including Guidelines for Content Designation," Library of Congress, November 2016, https://www.loc.gov/marc/bibliographic/ (accessed January 18, 2017).

30. "Archive-It," https://archive-it.org/ (accessed February 13, 2017).

tailored to the needs of its subject area. Records are based on RDA and created in MARC 21.[31]

Harvard University Archives: These guidelines are based on DACS, and descriptions are made at the collection level (i.e., multiple sites in a single record).[32] The records are fully archival in approach.

Library of Congress: Metadata for individual sites is created in MARC 65 and exported in the MODS format for public access.[33] Three types of URL are defined. A summary description is derived from the HTML header for a site's homepage.

New York Art Resources Consortium (NYARC): Three New York City museums collaborate to harvest websites documenting various segments of the art world. Site-level metadata records are created in Archive-It and then repurposed as MARC records in OCLC's WorldCat for export to the NYARC Discovery catalog.[34, 35] The data dictionary maps to MARC 21, Dublin Core, MODS, and several other standards.

University of Michigan, Bentley Historical Library: The Bentley's guidelines include aspects of archival and bibliographic approaches. Most records describe an individual website, though DACS is the designated source content standard.

University of Texas, Human Rights Documentation Initiative: HRDI creates its records using the MODS standard. A provenance note states why the site was harvested, and the University of Texas is recorded as creator.[36]

In comparing these guidelines, we tracked a variety of characteristics, including whether and how the content of each element is defined, the number of guidelines in which an element appears, and whether particular elements are designated as being mandatory or core. In some cases, the same type of content is placed in different data elements across the array of guidelines—and vice versa. Only three data elements appear in every set of guidelines: creator or contributor, title, and descriptive summary.

31. Columbia's approach can be seen in the Human Rights Web Archive, an open portal that gathers metadata from web archives worldwide. https://hrwa.cul.columbia.edu/.

32. Other Harvard units describe websites bibliographically; WAM is not studying those guidelines.

33. *MODS: Metadata Object Description Schema*, Library of Congress, last modified February 5 6450, accessed January 18, 2017, http://www.loc.gov/standards/mods/.

34. "Metadata Application Profile and Data Dictionary for Description of Websites with Archived Versions, Version 1 (June 2015)," New York Art Resources Consortium Web Archiving Initiative, accessed January 18, 2017, http://www.nyarc.org/sites/default/files/web-archiving-profile.pdf.

35. "NYARC Discovery," New York Art Resources Consortium, http://nyarc-primo.hosted.exlibrisgroup.com/primo_library/libweb/action/search.do?vid=01NYARC (accessed January 18, 2017).

36. "The Human Rights Documentation Initiative," University of Texas at Austin, https://www.lib.utexas.edu/hrdi/hr_archive (accessed January 18, 2017).

Issues revealed by the analysis

Our analysis reveals wide variation in the types of content included—or omitted—in each set of guidelines, and to which data element each type is assigned. To ensure consistency, each type of content should be assigned to a particular data element. In developing best practices, we must resolve two key questions: Which types of content are most important to include in a metadata record that describes an archived website or a group of sites? Which data element should be designated for each of these content types?

We have identified a variety of issues specific to website description:

- *Website creator/owner*: Is this the publisher? Creator? Subject? All three?
- *Host institution*: Is the institution that harvests and hosts the site the repository? Creator? Publisher? Selector? Collector?
- *Publisher*: Does a website have a publisher?
- *Title*: Should it be transcribed verbatim from the head of the site? Or edited to clarify the nature/scope of the site? Should acronyms be spelled out? Should the title begin with "Website of the ..."
- *Dates*: Which dates are both important and feasible to record? Beginning/end of the site's existence? Date(s) of capture by a repository? Content? Copyright?
- *Extent*: How should this be expressed? 5 archived website? 5 online resource? 6.25 GB? Approximately 300 websites?
- *Provenance*: In the Web context, does provenance refer to the site owner? The repository that harvests and hosts the site? Ways in which the site evolved? Frequency and dates of capture?
- *Appraisal*: Does appraisal mean the reason why the site warrants being archived? A collection of a set of sites named by the repository? The parts of the site that were and were not harvested?
- *Format*: Is it important that the description clearly states that the resource is a website? If so, how best to do this? In the title? Extent? Description?
- *URL*: Which URLs should be included? Seed? Access? Landing page?
- *MARC 21 record type*: When coded in the MARC 21 format, should a website be considered a continuing resource? Integrating resource? Electronic resource? Textual publication? Mixed material? Manuscript? Any of these?

Given these variations in practice, it will be challenging to sort out the issues and arrive at recommended best practices for data elements to serve user needs.

The subgroup informally sampled and evaluated existing descriptions of archived websites found in WorldCat (MARC records), ArchiveGrid (MARC records and finding aids), and Archive-It collections. By doing so, we can take into account the practices of some institutions that may not have formal guidelines.

Review of Tools with Metadata Features

The tools subgroup studied open-source tools that include metadata features for harvesting and managing Web archives. We began with the list of Web archiving tools that comprise the IIPC's *Tools and Software* registry.[37] We then narrowed this list of 28 tools to eleven that are actively maintained and seem to have potential for contributing to metadata creation. Since starting our work, we have learned that an IIPC group is planning to do similar evaluations, and we hope to collaborate to avoid duplication of effort.[38]

Evaluation criteria

- What is the basic purpose of the tool, and what functions does it perform?
- What objects/files can the tool take in and generate, such as WARCs or PDFs?
- What metadata profiles does it record in?
- Which descriptive elements are automatically generated?
- Which descriptive elements can be created/edited by the Web archivist/user?
- Which elements in each category are automatically generated or can be manually generated?
- What relation does it have to other tools?

Tool evaluation summaries

Four of the eleven tools evaluated are summarized below to illustrate the work being done:

As described earlier, Archive-It (AI) is a subscription-based Web archiving service from the Internet Archive.[39] AI's functions include capturing websites, administrative

37. "Tools and Software," International Internet Preservation Consortium, http://www.netpreserve.org/web-archiving/tools-and-software (accessed January 18, 2017).

38. Tom Cramer, "What Can IIPC Do to Advance Tools Development?," netpreserve.org, April 25, 2016, https://netpreserveblog.wordpress.com/2016/04/ (accessed January 18, 2017); Jackie Dooley and Tom Cramer, email correspondences, May 2016.

39. Archive-it, https://www.archive-it.org/ (accessed August 15, 2016).

tools for collecting and managing collections, and a preservation function by storing WARC files in the Internet Archive's digital repository. AI also generates Web Archive Transformation (WAT) files, which can be used to create data analysis reports; Longitudinal Graph Analysis (LGA) files; and WANE files, which include the named entities from each text resource. AI's API makes collection and seed-level metadata available to harvesters, including OCLC's WorldCat. AI employs several tools, including Heritrix (Web crawler), NutchWAX (text search), Solr (metadata-based search), and Umbra (which works with Heritrix to detect previously unavailable URLs).

HTTrack is an open-source capture tool that uses an offline browser utility to download a website to a directory, generates a folder hierarchy, and saves content that mirrors the original website structure. HTTrack produces basic log files but does not generate WARCs or ARCs. It does not allow for input of any descriptive metadata, which would need to be created using external tools, nor does it enable extraction of technical or preservation metadata.[40]

Web Archive Discovery is an open-source capture tool that takes ARC and WARC files and manipulates them with Solr to provide full-text search of archived websites. Because it is a back-end tool without a Web interface, users may employ their own interface on top of Solr. Web Archive Discovery parses WARC files into JSON format for Solr querying and auto-captures descriptive, administrative, and preservation/technical metadata that can be extracted.[41]

Webrecorder is an open-source tool that archives Web content through interactive browsing, capturing the exact sequence of navigation through a series of webpages or digital objects, thus preserving the user's experience. The recording is archived as a WARC file, which can be replayed using Webrecorder. Descriptive metadata elements include username (labeled "creator"), title, capture date/time, archive file format, title of collection (if defined by user), and URL. Descriptive metadata is automatically generated by Webrecorder during archiving and is embedded in a downloadable WARC file. Future development plans include the ability to add user-created metadata and support for page-level annotations.[42]

Our evaluation has revealed that tools generally record and enable export of minimal metadata. Our tools report will publish the full evaluations so that readers can obtain a thorough overview of the current tools landscape.

40. HTTrack, https://www.httrack.com/ (accessed August 15, 2016).

41. "webarchive-discovery," UKWA (UK Web Archive), https://github.com/ukwa/webarchive-discovery (accessed January 18, 2017).

42. Webrecorder, https://webrecorder.io/ (accessed January 18, 2017).

Conclusion

The Web Archiving Metadata Working Group's four subgroups have abstracted and synthesized literature relevant to Web archiving metadata, including published articles, blog posts, surveys, conference papers, tweets, and slide decks. This work, as summarized in this paper, has revealed much about the needs and behaviors of the types of users described in those readings. Further, our analysis of three metadata standards and seven sets of local guidelines has confirmed a lack of shared practices across the community, as well as the common absence of some types of metadata that users value. Having accomplished this preparatory research, we are ready to begin drafting recommended best practices for Web archiving. Representatives from each of the four initial subgroups will be directly involved in this phase.

When we have a draft in hand, we will circulate it widely to gather feedback from the Web archiving community by various means, including listservs, social media, blog posts, and an OCLC Research Library Partnership webinar.

WAM's work to create best metadata practice guidelines that articulate the needs of archived websites users will move the community forward toward ensuring the discoverability of these increasingly indispensable resources.