Descriptive Metadata for Web Archiving

Literature Review of User Needs

Jessica Venlet, Karen Stoll Farrell, Tammy Kim, Allison Jai O'Dell and Jackie Dooley

OCLC

Descriptive Metadata for Web Archiving: Literature Review of User Needs

Jessica Venlet
University of North Carolina at Chapel Hill

Karen Stoll Farrell
Indiana University

Tammi Kim
University of Nevada, Las Vegas

Allison Jai O'Dell
University of Florida

Jackie Dooley
OCLC Research

February 2018

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 978-1-55653-003-6
DOI: 10.25333/C33P7Z
OCLC Control Number: 1021288024

ORCID iDs
Jessica Venlet https://orcid.org/0000-0002-2647-8489
Karen Stoll Farrell https://orcid.org/0000-0002-2160-583X
Tammi Kim https://orcid.org/0000-0001-9505-2601
Jackie Dooley https://orcid.org/0000-0003-4815-0086

Please direct correspondence to:
OCLC Research
oclcresearch@oclc.org

Suggested citation:
Venlet, Jessica, Karen Stoll Farrell, Tammy Kim, Allison Jai O'Dell, and Jackie Dooley. 2018. *Descriptive Metadata for Web Archiving: Literature Review of User Needs*. Dublin, OH: OCLC Research. https://doi.org/10.25333/C33P7Z.

CONTENTS

Executive Summary 4
Introduction 6
End-User Needs 7
- Who Uses the Archived Web? 7
- Provenance Metadata 8
- Limitations on Access and Re-use 9
- Discovery Systems 9
- Technical Barriers 10
- Engaging the Community 11

Metadata Practitioner Needs 11
- Scalable Practices are Needed 11
- Standards and Shared Practices 12
- Case Studies 13
- Other Metadata Types and Approaches 14

Conclusion 15
- End Users 15
- Metadata Practitioners 15

Acknowledgments 17
Appendix: Annotated Bibliography 18
Notes 42

EXECUTIVE SUMMARY

The OCLC Research Library Partnership Web Archiving Metadata Working Group was formed to recommend descriptive metadata best practices for archived web content that would meet end-user needs, enhance discovery and improve metadata consistency. To that end, we conducted a literature review to inform our development of best practices.

We selected readings that include, at minimum, a substantive section related to metadata, but most covered a wider swath of issues. This helped us learn much else about who the users of web archives are, the strategies they use and the challenges they face.

The literature falls into two clear categories: the needs of end users and the needs of metadata practitioners. This review characterizes types of end users, their research methodologies, barriers to use, discovery interfaces and the need for support services and outreach. The review of practitioner literatures addresses the need for scalable practices, the standards and shared practices currently in use, the outcomes of a variety of case studies and other approaches to metadata. The report mirrors these two categories of literature, as do our conclusions:

End Users

- The literature on end-user needs largely focuses on academic researchers in a wide variety of disciplines.
- Users express a strong need for provenance information to add context beyond standard descriptive metadata elements, reflecting a widespread desire for transparency around the decision-making process in selecting sites for capture and building collections, as well as the completeness of individual captures and changes that occur over time.
- Given the ease and ubiquity of access to the open web, restrictions on access—such as being limited to onsite viewing in a library—are both mystifying and frustrating to users. A closely related need is for clarity about intellectual property rights vis-à-vis re-use of content.
- Archived web data often is provided in a way that exceeds the limits of users' technical knowledge, constituting a widespread barrier to use. This could be alleviated by developing user-friendly tools and interfaces to enable content to be more readily accessed and repurposed.
- A need for user support services derives from the complexity of accessing and using web archives.
- Libraries and archives should actively engage in outreach to both current and potential web archives users: first, to provide an initial understanding of what web archives are, and how to find and use them; and second, to better understand users' issues in order to find ways to ameliorate their challenges.

Metadata Practitioners

- Scalable descriptive metadata practices are needed because staff resources are extremely limited at most institutions. Most web archivists have numerous other duties.
- Existing library and archival standards for data structure and content are being used for web archiving descriptive metadata. Dublin Core is most often used, which is largely attributable to the widespread use of Archive-It, the web archiving service of the Internet Archive.
- Bibliographic, archival and hybrid approaches are in use. The need to find appropriate ways to blend standard library and archival practices is widely perceived.
- In devising metadata at various levels of description (collection, site, document), practitioners should consider carefully the elements they will use at each level.
- Metadata describing archived web content is often delivered via multiple discovery systems, such as integrated within a library catalog or aggregation of archival finding aids, provided in isolation via a standalone interface for web content, and/or through Archive-It. This clearly suggests the need for smooth processes to re-use metadata generated in one system to the other systems in use.
- Experimentation with nontraditional approaches is underway.

The urgency to capture and preserve the internet increases by the day as it becomes the sole source of much human-generated knowledge and experience. Archiving of the web is gradually increasing, and the number of both end users and knowledgeable practitioners are growing in concert. Communities of practice are coming into existence, and the rich literature described in this report is a testament to the vitality of this new field.

This report is one of a complementary trio being issued simultaneously to document the work of the OCLC Research Library Partnership Web Archiving Metadata Working Group. Its siblings are *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*[*] and *Descriptive Metadata for Web Archiving: Review of Harvesting Tools.*[†]

* Dooley, Jackie, and Kate Bowers. 2018. *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group.* Dublin, OH: OCLC Research. doi:10.25333/C3005C.

† Samouelian, Mary, and Jackie Dooley. 2018. *Descriptive Metadata for Web Archiving: Review of Harvesting Tools*. Dublin, OH: OCLC Research. doi:10.25333/C37H0T.

INTRODUCTION

Historians can't do what they need to do today without access to web archives.
–Ian Milligan, University of Waterloo (2015)[*]

The OCLC Research Library Partnership Web Archiving Metadata Working Group (WAM) was formed to recommend descriptive metadata best practices for archived web content. When the group was formed early in 2016, we immediately recognized the need to develop an understanding of the descriptive metadata needs of end users. To this end, we conducted a literature review to inform our development of recommendations of descriptive metadata. We studied and abstracted 64 readings that had been published through June 2017.

The literature on end-user needs goes far beyond metadata issues, and it was necessary to limit our scope. We selected readings that include, at minimum, a substantive section related to metadata. In the process, however, we learned much more about who is using web archives, the strategies they use and the challenges they face. This report favors the issues and practices that have implications for descriptive metadata needs, with somewhat less attention paid to associated topics covered in the readings. The latter are covered more fully in the annotated bibliography that follows the report narrative.

As we began to identify sources that appeared likely to address these needs, it became clear that they fell into two clear categories: the needs of end users and those of metadata practitioners. Unsurprisingly, both types are addressed in some articles. This summary of the literature characterizes types of end users, their research methodologies surrounding web archives, barriers to use, discovery interfaces, and the need for support services and outreach activities that libraries can provide to support users of web archives. The review of practitioner literature addresses the need for scalable practices, the standards and shared practices currently in use, the outcomes of a variety of case studies, and nonstandard approaches to metadata.

Web archiving activity and practices are rapidly expanding and evolving, and so we went beyond published articles to include blog posts, slide decks, tweets and even notes taken at conferences, seeking to encompass the full range of current research and practice. Working group members each read their share of sources and prepared abstracts, and these were synthesized to prepare the narrative report.

All of our readings are included in the appended annotated bibliography, though not all are cited in the report narrative.

[*] Milligan, Ian. 2015. "Between Metadata and Content: Canadian Political Parties and Political Interest Groups using Archive-It (2005-2015)." Unpublished presentation at the symposium Web Archives 2015: Capture, Curate, Analyze, University of Michigan, Ann Arbor, Michigan, 12-13 November 2015.

This report is one of a complementary trio being issued simultaneously to document the work of the OCLC Research Library Partnership Web Archiving Metadata Working Group. Its siblings are *Descriptive Metadata for Web Archives: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group* [1] and *Descriptive Metadata for Web Archives: Review of Harvesting Tools*.[2]

End-User Needs

This summary of the literature characterizes types of end users, their research methodologies surrounding web archives, barriers to use, discovery interfaces, and the need for support services and outreach activities that libraries can provide to support web archives users.[3]

WHO USES THE ARCHIVED WEB?

A variety of authors have identified use cases by examining specific web archiving initiatives and who uses them. Our findings reveal many similarities across this literature, suggesting that some predominant information-seeking behaviors and research methodologies are employed by web archives users. Some studies have been conducted based on the heterogeneous users of national libraries and archives, but most of the in-depth research focuses on scholarly use by academic researchers.

The data from Weber and Graham's user survey indicates that graduate students are the most frequent users (38%), followed by faculty members (32%), independent researchers (11%) and post-doctoral researchers (8%).[4] The most common disciplines in which their respondents work are history, literature, English, digital humanities and computer science.

In a 2015 International Internet Preservation Consortium (IIPC) Archive Researcher Services Workshop, Jefferson Bailey laid out a topology of the end users of the Internet Archive's services for researchers[5]:

- *Legal professionals:* legal discovery, evidentiary, patent, documentary
- *Social/political scientists:* communications, politics, government, social anthropology
- *Web scientists:* web technologies, systems, protocols, benchmarking
- *Digital humanities:* historians and humanities disciplines, network graphing, text mining, topic modeling
- *Computer scientists:* information retrieval, data enrichment, technical development, technology over time
- *Data analysts:* data mining and model training, natural language processing, trend analysis, named entity recognition, machine learning

In a 2007 report on a survey of users of the web archive at Koninklijke Bibliotheek (the National Library of the Netherlands, or KB), Ras and van Bussel[6] also put forth a list of user types: historians, sociologists, linguists, journalists, owners/designers of websites who want to view previous versions, public institutions that do not archive their own websites and thus refer the public to the KB, and the general public that normally visits the library.

The KB report also includes a survey of other national libraries, each of which offered its own list of types. For example, the UK Government Web Archive listed government institutions, information

professionals from the heritage sector, researchers and the general public. The National Library of Israel, on the other hand, gave its view of potential web archives users, including instructors, lawyers, sociologists, journalists and linguists. The US Library of Congress reported that users are mainly students, instructors, website owners and US Congress staff members. Finally, an Access Working Group of the IIPC posited potential users to be "ordinary citizen, government, lawyer, patent applicant, student, researcher, genealogical researcher and internal staff members."

In summary, we see various groups of academic users, some with very distinct needs. Some use web archives primarily for content, such as 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. Beyond scholarly research, user groups such as lawyers, government agencies and the general public must be considered as potential, if not actual, users.

Users also have been characterized according to the types of research questions they ask, suggesting the different approaches and needs they bring to web archives. Overlaps exist across the various user groups, but it is also possible to consider each distinctly to explore how its needs may differ.

Costa and Silva[7] presented three categories of information needs that they believe generally cover all user groups. The authors assert that these needs can be classified into three behavioral types:

- *Navigational:* find a specific website that the user already has in mind
- *Informational:* collect information about a topic from multiple sources
- *Transactional:* perform a web-mediated activity, such as downloading a file, which is often a first step to combining large amounts of data across multiple web archives

Costa and Gomes[8] categorized use cases for the Portuguese Web Archive as being 53% to 81% navigational, 14% to 38% informational and 5% to 16% transactional. The Weber and Graham survey[9] data show that 70% of respondents use web archives for exploratory research, while only 25% undertake targeted investigations; they also found that 40% of users seek to retrieve lost documents. Smaller percentages of users perform data analysis: 34% engage in full-text analysis (data mining), 17% in network analysis, 16% in hyperlink analysis and 20% in metadata extraction.

PROVENANCE METADATA

The need for a wide variety of types of provenance metadata is a recurring theme, with multiple authors identifying this as a critical missing piece for the use of web archives. The literature relating to technical and preservation metadata is extensive, but, given our working group's overall focus on descriptive metadata, we concentrated on types of provenance information relevant in that context.

Generally speaking, provenance refers the origin of something. In the sense typically used by archivists, provenance metadata can provide information on how, why, when and by whom an object was created, as well as chain of custody. In the web archives context, this can include information such as background on site creators, the rationale used for building a collection, how a site or thematic collection has changed over time, capture dates and completeness of crawls.[10]

Murray and Hsieh[11] noted that metadata documenting both provenance and preservation activities are of value to users, while Galligan[12] concluded that researchers working with web archives are seemingly more concerned with what is not captured and preserved than are researchers working with analog materials. Jackson[13] explained that provenance information, such as crawl logs, may help users understand how and why the library captured a particular resource. In notes taken at the

2015 IIPC conference, Dooley took notes of a discussion about the need for provenance metadata to establish authenticity and trust around web archives.[14] She also recorded that users are concerned with the temporal incoherence of irregular captures, and that provenance information can alleviate concerns about why captures were made at certain times and not others.

In a somewhat different vein, Dougherty et al. (2010)[15] discussed user needs for metadata that is administrative, descriptive, and contextual or relational, as well as the ability to add metadata over time.

Overall, the literature demonstrates that provenance metadata, including information about curatorial decisions and the technical details of captures, is crucial to the use and perceived trustworthiness of web archives.

LIMITATIONS ON ACCESS AND RE-USE

Hockx-Yu[16] noted that some large-scale national web archives are able to provide access only onsite due to intellectual property issues. “The misalignment between legal requirements and user expectation is a difficult problem for web archives as the choice seems to be between the comprehensiveness of the archive and online access, not both.” In some countries, making web archives publicly available, even under license, may be regarded as republishing. She also stated that in the UK, the process of archiving can transfer certain legal risks, such as libel, from the original site publisher to archiving institutions.

Users of web archives also want clarity and transparency regarding intellectual property rights.[17] As born-digital content, archived web content can be subjected to data mining and analysis. Thus, users want to know how they can access, re-use and publish the information contained in web archives. The ubiquitous, open nature of the internet raises users’ expectations that they will be able to freely access content from anywhere, at any time. When archived web content is not accessible online, those expectations go unmet.

DISCOVERY SYSTEMS

Users want to study the content of web archives from many perspectives; discovery interfaces therefore should allow exploration based on subject content and numerous other characteristics. In 2010, Dougherty et al. stated this:

> Contextual access places artefacts in a thickly described and purposeful context. Contextual access does not place an artefact in its original context; rather it makes an artefact findable via its relationship to other objects in a research project ... Users enter an archive and view archived artefacts via the research of another. Archived artefacts, in this sense, can be seen as a collection of objects to which a research project refers. This metadata answers the question, “What is it about?” for any object in the archive, and this question can be answered differently many times over depending on the perspective and purpose of the researcher-user.[18]

Thurman and O’Hanlon[19] described the Columbia Human Rights Web Archive, which is a representative standalone interface that has numerous special features. This discovery system searches both bibliographic metadata and the full text of archived sites. Users can browse by subject, website title, URL, geographic place and languages. The system allows users to facet searches by subject, geographic focus, organization type, location where an organization is based, language, domain, date of capture and file type. Additional features include a “search other sources” feature and highlighting of keywords. Searches also can be scoped to search the

metadata on its own. The HRWA system illustrates both metadata elements tailored to the particular scope of a web archive and some of the specialized features designed to meet user needs via a discovery system tailored to the particular features of web content.

In her study of web archives functionality, Niu found that most have search and browse features, including known-item and exploratory search with limit options, but not specialized desirable functionalities such as data mining, personalized services or the ability to reconstruct lost websites.[20] Leetaru described features of the Internet Archive's Wayback Machine such as the timeline that allows users to explore changes to sites and to the web over time, and of the Virtual Reading Room that enables visualization, mapping and cloud-based data mining.[21]

The work of Van de Sompel et al. in creating Memento[22] has been revolutionary. The "Memento solution" has two components: navigation to an archived resource via its original resource, by leveraging content negotiation; and use of a discovery API for archives that enables retrieving a list of all archived versions of a resource for a given URI.[23]

Overall, the literature in this area indicates that discovery systems should enable inclusion of the wide array of both descriptive and contextual information necessary to dig into the content of web archives, and that data analysis/mining tools are the next step in building discovery to support user needs.

TECHNICAL BARRIERS

A variety of authors indicate that the archived web data commonly served to users may be too raw or intimidatingly technical for them to manage, which clearly raises barriers to use.

Galligan described the lack of tools for accessing and analyzing the data as a major barrier for researcher use, while noting that there is forward movement in building tools.[24] For example, Bailey's description of Archive-it Research Services includes information about the packaged dataset WANE (Web Archive Named Entities).[25] This tool extracts the names of people, organizations and places from the content of websites.

According to Bailey, the Web ARChive (WARC) format is an unfamiliar format to researchers.[26]

The necessity for specialized training to access web archives that lack a user-friendly interface poses an issue for people using web archives as information and data sources. Dooley noted Bailey's statement that researchers require flexible data delivery services and formats in order to understand the context of web archives.[27] Based on usability testing of the Portuguese web archive search interface, Cruz and Gomes[28] reported that users would likely to be able to share files on social media and save them as PDFs. Bailey[29] noted that users can be daunted upon gaining access to raw files without an interface to help them sort, filter and search for datasets. Truman[30] pointed out that even when user-friendly discovery interfaces are available, the unit of retrieval is most often the URL.

Further, the lack of interoperability across web archives adds a layer of complexity for users who want to aggregate and analyze content from sites preserved in multiple locations. The web archiving community critically needs more federation of resources to enable users to discover and then manipulate content. In a research project at Old Dominion University, Thomas et al. (2010)[31] sought to extract text from archived websites. The team found a need to develop custom code to acquire, parse and process the content of pages, as well as to develop an algorithm for detection and use of character encoding.

Overall, the literature demonstrates that libraries should serve the content of web archives in formats that can readily be re-purposed by researchers without requiring extensive training or scripting.

ENGAGING THE COMMUNITY

Cruz and Gomez noted that collaboration and targeted outreach is necessary to educate users about access and re-use of data contained in web archives.[32]

Both Truman[33] and Dougherty et al.[34] emphasized the benefits of establishing communication and collaboration between researchers and web archivists. Truman established that no single community or collaboration covers all web archives users or practitioners. A study of users of the New Zealand web archive backs up these findings and points to a strong need for outreach and education.[35] Dougherty et al.[36] interviewed researchers to understand their preferred methods of working with web archives and uncovered a desire among researchers to deposit, share and/or publish their own archives, and to contribute their own metadata for the archives that they use. In a study conducted by Dougherty and Meyer in 2014,[37] collaboration between librarians and/or archivists and researchers was identified as a critical need. Finally, Hockx-Yu's 2016 study[38] of users of national-level web archives pointed to the progress that has been made by some national libraries, including the British Library, in increasing engagement with the research community.

Overall, the literature demonstrates that libraries should enable partnerships with researchers and other institutions in developing and describing web archives. Tools that support annotation and crowd-sourced metadata create a symbiotic relationship between the researcher and the archives. Dougherty and Meyer[39] refer to this explicitly as a critical need for researchers. In interviews with users, they found that, "most of all, [users] want to work with those [web] objects, enriching and annotating them on whatever level is appropriate for their analysis."

Metadata Practitioner Needs

WAM's recommendations for descriptive metadata were informed by a combination of the conclusions we drew from the literature on end-user needs, which are reported in the previous section of this report; and our analysis of current practices as revealed in institutional guidelines and sample records from a variety of libraries and archives, summarized in the companion report, *Descriptive Metadata for Web Archives: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group.* [40] The practitioner literature reviewed in this section further enhanced our sense of the descriptive metadata landscape.

Most practitioners describe fairly standard description practices already in use across the library and archives communities. They detail the full array of methods used, including content and data structure standards, levels of description, and the challenges of addressing both bibliographic (library) and archival approaches. These readings are summarized in the first two sections: Survey Findings and Case Studies.

The third section, Other Metadata Types and Approaches, summarizes readings that discuss other standards, as well as experiments undertaken to rethink standard methods for archived web content. These add further food for thought.

SCALABLE PRACTICES ARE NEEDED

A frequent theme is the need for scalable descriptive metadata practices that take into account the extremely limited human resources available. Some institutions clearly would like to create more granular metadata than they do at present, but the resources are sorely lacking.

For example, the 2015 NDSA web archiving survey revealed that more than half of respondents continue to have no more than .25 FTE staffing for this work (the same as in the 2013 data), though the percent that now have more than .5 FTE grew from 19% to 29% between 2013 and 2015.[41] Sweetser found that the amount of staff time available is the most important factor in determining which sites will be described and at what level of detail.[42] Mannheimer explored barriers to metadata creation and found that lack of staff time ranked highest, principally because most practitioners can spend only a small fraction of their time on this work.[43]

Philips and Koerbin observed in their 2004 article on Australia's PANDORA archive project (Preserving and Accessing Networked Documentary Resources of Australia) that cataloging is the second most time-consuming part of the workflow and questioned its scalability relative to existing funding.[44] Invoking one possible approach to addressing the problem, Guenther noted the importance of exploring reuse of existing metadata and the need for interoperability of schemas and tools in her recommendations to NYARC.[45]

STANDARDS AND SHARED PRACTICES

Dublin Core (DC) is by far the most widely used data standard, principally because it underpins Archive-It's metadata functionality, and so most surveys specifically address its use.[46] In an internal survey in 2013, Archive-It staff found that 90% of their users generate DC metadata at the collection level, 60% at the seed level (i.e., the harvested URL) and 15% at the document level.[47] Gibbons' 2016 survey population was broader, and 44% of her respondents use DC to describe web archives.[48] Sweetser[49] found that Description, Title, Creator, Subject, and Date were the most commonly used DC elements. In her 2013 project, Mannheimer[50] correlated practitioners' years of experience with the metadata elements that they assigned. Regardless of experience, she found that the most common DC elements were Creator, Description and Title, followed by Date and Subject. Unfortunately, no survey seems to have explored how practitioners interpret these data elements (e.g., the types of date assigned to the date element).

Researchers have explored the use of standard library catalog records and archival finding aids for describing web content. Sweetser's 2011 study of Archive-It users gathered data on the level at which users create metadata and found that only one-third add elements beyond the minimum required. She found that about 22% link to Archive-It from finding aids that describe an existing analog collection, while even fewer (17%) create finding aids specifically for web archives. About 25% create a collection-level catalog record for web archives. Mannheimer asked slightly different questions and found that 30% create catalog records for web archives and 32% create finding aids. Overall, she felt that the literature supports

contextual information as being important when describing archival materials. "Whether communicated through a finding aid, using descriptive metadata, or via social tagging, maintaining context is a vital step toward ensuring long-term preservation of archival materials, and especially the Web."[51]

The 2013 National Digital Stewardship Alliance (NDSA) survey, *Web Archiving in the United States*,[52] found a decrease in collection- and seed-level catalog records between the 2011 and 2013 surveys; the authors speculated that the addition of finding aids as a response selection could be a factor. This survey found that 20% of respondents create finding aids, 22% create collection-level catalog records and 18% create site-level descriptions. The more recent 2016 NDSA[53] survey found that respondents felt little progress has been made with regard to metadata practices since the 2013 survey, but the use of finding aids and catalog records (at the collection and seed levels) had increased somewhat.

The OCLC Research Library Partnership Metadata Managers Focus Group[54] discussed web archiving metadata during a 2016 meeting, also demonstrating the hybrid archival and bibliographic approaches employed by some institutions. For example, Yale University shared that websites described as records are governed by DACS (Describing Archives: A Content Standard), while those considered to be publications are described using RDA. The need to integrate description and discovery of web content in concert with other resources and systems was emphasized throughout.

CASE STUDIES

Guenther worked with the New York Art Resources Consortium (NYARC) in 2015 to design a Metadata Application Profile[55] for websites, and in a complementary report she outlined recommendations for further exploration[56] that contributed to the justification for aspects of our work. In part, she stressed the need for more thorough understanding of the strengths of archival and bibliographic approaches when describing web archives to aid practitioners in applying each approach appropriately. To fully understand their strengths, it would be important to test the usability of each approach, but a review of existing practices contributes some context.

In 2002, Haddad and Gatenby[57] described the Library of Australia's use of bibliographic metadata. Most seeds received a detailed catalog record, but some were described as groups when archived as topical collections. Each seed was assigned a persistent identifier. The article provides some notes on specific MARC fields used.

Prom and Swain[58] reported on the use of archival description for websites archived at the University of Illinois at Urbana Champaign. Sites were crawled in groups by record series using the harvesting tool HTTrack, which creates a list of the seed URLs that could then be used to construct a finding aid. To provide contextual information, the authors created a collection-level description that was then linked to the list of URLs. (The information captured in the collection-level record is not described in the article.)

Guenther and Myrick[59] provided another early example of how practitioners approached description of collected websites. The authors noted that the US Library of Congress used a single collection-level MARC record to describe the entire Election 2000 archive, whereas for the Election 2002 archive, the library collaborated with the SUNY Institute of Technology and Webarchivist.org to create MODS records for individual websites within the collection.

O'Dell[60] provided a detailed description of her hybrid approach to cataloging a web archive collection as "an online resource under archival control." She suggested describing web archives at the collection level. The post details O'Dell's decision-making processing in implementing MARC fields, selecting subjects (which are recorded in an abstract and subject access points), and using both RDA and DACS rules.

A set of presentations from a 2014 Society of American Archivists conference session provides a snapshot of descriptive methods in use at 11 institutions (Zhang et al. 2014).[61] Presenters discussed a variety of bibliographic and archival approaches. For example, Tuomala described the University of North Carolina's mixed methods: one curatorial area created seed-level MARC records for some collections, while another group integrated description of archived websites into existing finding aids that also describe analog materials. Bence described Emory University's method of integrating web archives into existing finding aids using the EAD element <dao> for URL links. Virakhovskaya shared the MARC fields used for describing collections of archived websites at the University of Michigan's Bentley Historical Library. Each presentation touched on the challenges of balancing archival and bibliographic approaches and integrating website description into existing workflows.

In a webinar series from the Metropolitan New York Library Council, presenters from Columbia University and the New York Art Resources Consortium described their practices. Thurman[62] briefly discussed how some institutions use finding aids to describe web archives and then focused on Columbia's fairly granular bibliographic approach to description. He noted that most of Columbia's metadata creation happens at the site level, with each described in Archive-It. Some collections and seeds are supplemented by MARC records. Likewise, Pregill[63] discussed NYARC's bibliographic metadata creation workflow and provided sample MARC records showing seed-level descriptions.

OTHER METADATA TYPES AND APPROACHES

Peterson[64] blogged about her experiment to study how web archive descriptive metadata and crawl documentation could map to archival description practices. She wanted to find a way to meet user needs for more documentation of crawls and appraisal, and she explored what might be documented for seeds, collections and crawls. Ultimately, she decided the available descriptive practices favor seeds and collections over crawls. She could not settle on a sustainable or useful method for adding crawl description to archival description.

Bernstein[65] discussed conversion of MARC records into more modern metadata practices, which could have value for converting existing records for web archives to MARCXML or other contemporary data structures.

In a 2006 paper, Wu et al.[66] examined the importance of context in describing archived websites using a bibliographic approach intended to ensure "context-sensitive annotation." The next year, the same three authors published a second article[67] in which they noted the limitations of their previous work and discussed another approach, grounded in principles of archival description, which they termed "context-aware annotation." This Web Annotation for Web Intelligence (WAWI) concept strives to establish relationships "between the metadata, the context of the web material, and the social context in which the content was produced." The WAWI approach includes use of title, alternative title, creator, subject, description, and date created. The authors also discussed the potential to use enhanced Dublin Core elements to specify CreatedDate and IssuedDate, CoverageTemporal and CoverageSpatial and others.

Although the scope of WAM's work is limited to descriptive metadata, we also reviewed several articles that looked at technical and preservation metadata. Bailey and LaCalle[68] discussed WARC file metadata in a 2015 presentation. They described the required elements, which are record identifier, content length/body size, timestamp and WARC record type. The presentation also highlights some of the challenges associated with creating preservation metadata for web archives; these include the concatenated nature of resources, the placement of resources within WARC files and the scale of web archives.

Guenther and Myrick[69] discussed the need for exploration of METS and PREMIS metadata to describe the structure of a website and to capture technical metadata generated during the crawl. Lavoie and Gartner[70] did not address web archive metadata per se, but made note of the general need for more complete understanding of the creation and management of preservation metadata in digital archiving systems. Likewise, Guenther[71] recommended exploration of the value of PREMIS for web archives. She asked important questions regarding provenance metadata and determining what metadata will be necessary to aid migration of WARC files over time.

CONCLUSION

Just as this report is in two sections, so are our conclusions.

END USERS

- The literature on end-user needs largely focuses on academic researchers in a wide variety of disciplines. A pattern of standard needs and approaches has emerged.
- Users express a strong need for provenance information to add context beyond standard descriptive metadata elements. This reflects a widespread desire for transparency around the decision-making process in selecting sites for capture and building collections, as well as the completeness of individual captures and changes that occur over time.
- Given the ease and ubiquity of access to the open web, restrictions on access—such as being limited to onsite viewing in a library—are both mystifying and frustrating to users. A closely related need is for clarity about intellectual property rights to help users understand how they can use information obtained from web archives.
- Archived web data often is provided in a way that exceeds the limits of users' technical knowledge, constituting a widespread barrier to use. This could be alleviated by developing user-friendly tools and interfaces to enable content to be more readily accessed and repurposed.
- A need for user support services derives from the complexity of accessing and using web archives.
- Libraries and archives should actively engage in outreach to both current and potential web archives users: first, to provide an initial understanding of what web archives are, and how to find and use them; and second, to better understand users' issues in order to find ways to ameliorate their challenges.

METADATA PRACTITIONERS

- Scalable descriptive metadata practices are needed because staff resources are extremely limited at most institutions. Most web archivists have numerous other duties.
- Existing library and archival standards for data structure and content are being used for web archiving descriptive metadata. Dublin Core is most often used, which is largely attributable to the widespread use of Archive-It, the web archiving service of the Internet Archive.
- Despite the pervasive use of standards, their application to archived web content is highly inconsistent for both the data elements employed and the contents of those elements.
- Bibliographic, archival and hybrid approaches are in use. The need to find appropriate ways to blend standard library and archival practices is widely perceived.
- In devising metadata at various levels of description (collection, site, document), practitioners should consider carefully the elements they will use at each level.
- Metadata describing archived web content is often delivered via multiple discovery systems, such as integrated within a library catalog or aggregation of archival finding aids,

provided in isolation via a standalone interface for web content, and/or through Archive-It. This clearly suggests the need for smooth processes to re-use metadata generated in one system to the other systems in use.

- Experimentation with nontraditional approaches is underway.

The urgency to capture and preserve the internet increases by the day as it becomes the sole source of much human-generated knowledge and experience. Archiving of the web is gradually increasing, and the number of both end users and knowledgeable practitioners are growing in concert. Communities of practice are coming into existence, and the rich literature described in this report is a testament to the vitality of this new field.

ACKNOWLEDGMENTS

Two subgroups of the Web Archiving Metadata Working Group read and abstracted the 64 readings on which this report is based.

These members were responsible for the end-user needs literature:

- Alexis Antracoli (Princeton University)
- Lori Dedayan (University of California Berkeley)
- Karen Stoll Farrell (Indiana University)
- Deborah Kempe (Frick Art Reference Library)
- Tammi Kim (University of Nevada, Las Vegas)
- Matthew McKinley (California Digital Library)
- Allison Jai O'Dell (University of Florida)
- Joy Panigabutra-Roberts (University of Tennessee, Knoxville)
- Olga Virakhovskaya (University of Michigan)

This group was responsible for the readings on the needs of metadata practitioners:

- Jason Kovari (Cornell University)
- Joy Panigabutra-Roberts (University of Tennessee, Knoxville)
- Dallas Pillen (University of Michigan)
- Mary Samouelian (Harvard University)
- Jessica Venlet (University of North Carolina at Chapel Hill)

Following preparation of an abstract for every article, each group synthesized the readings to develop categories that served as the basis for organizing this report. The working group members credited as authors brought it all home.

Thanks are due to the many authors and presenters, both scholars and practitioners, whose work comprises the literature on web archiving. All are pioneers.

Finally, colleagues in OCLC Research were indispensable contributors. Program Officer Dennis Massie provided wise counsel and ubiquitous support to the Working Group throughout the project. Profuse thanks also are due to those who efficiently shepherd every publication through the production process: Erin M. Schadt, Jeanette McNicol and JD Shipengrover.

APPENDIX: ANNOTATED BIBLIOGRAPHY

This annotated bibliography represents the research done by members of the OCLC Research Library Partnership Web Archiving Metadata Working Group to understand the landscape of user needs prior to preparing best practice recommendations for descriptive metadata for web archives. This work has substantially informed those recommendations. This bibliography includes citations and annotations for 64 published articles, blog and Twitter posts, conference notes and other sources.

Bailey, Jefferson. 2015. “Web Archives as Research Datasets.” Paper presented at the General Assembly of the International Internet Preservation Consortium, Stanford University, 28 April 2015. http://netpreserve.org/general-assembly/ga2015-schedule.

> Data on the web is challenging to use: steep learning curve and minimal support for non-technical users. Archive-It Research Services (AIRS) offers three types of datasets packaged for easier consumption. 1) WAT (Web Archive Transformation): Extracts provenance, text and link data from each URL into easily parsed JSON files; enables aggregate analysis, text/link mining, trend analysis over time. 2) LGA (Longitudinal Graph Analysis): Extracts links, where they link and when for ALL links in collection; enables identification of important sites, relationships over time and at a certain time. 3) WANE (Web Archive Named Entities): extracts names of people, places, organizations from text; enables identifying important entities and prevalence over time as well as correlation with external data. AiRS hopes to expand access models and enable new insights. Moving research use forward requires user comfort with technical mediation at multiple levels and increased distance between granularity and totality of object of study.

Bailey, Jefferson, and Maria LaCalle. 2015. “Don’t WARC Away: Preservation Metadata for Web Archives.” Paper presented at the ALA Annual Conference, Association of Library Collections and Technical Services, Preservation and Reformatting Section, San Francisco, California, 27 June 2015. http://connect.ala.org/files/2015-06-27_ALCTS_PARS_PMIG_web_archives.pdf.

> Bailey and LaCalle give a brief overview of the history, size and scope of the Internet Archive, as well as some context about what a web archive is and the results of several surveys of the web archiving community. The presentation then goes into detail about the WARC (Web ARCive) format, including the four required fields (record identifier, content length/body size, timestamp, WARC record type) as well as metadata that is not included in a WARC (rights and permissions, description, file format identification, etc.). The presentation then details with preservation metadata, including its concatenated nature, the arbitrary placement of resources in challenges WARC files and the volume of data. An Archive-It partner survey found that 80% of respondents do not store their WARC files locally, and that only 14% of respondents create metadata for WARC files. METS is then discussed, including the <structMap> section used to represent the logical structure of a harvested website, and PREMIS used to detail preservation events. Use of PREMIS as a method of recording preservation metadata has complications, including limited local acquisition, crawler variance and scale. Finally, practical approaches to web archives preservation

metadata are described, including favoring data redundancy over metadata granularity, using crawl reports, and simplifying events, agents and objects (presumably to make the adoption and scalability of PREMIS more feasible).

Bailey, Jefferson, Abigail Grotke, Edward McCain, Christie Moffatt, and Nicholas Taylor. 2017. *Web Archiving in the United States: A 2016 Survey*. Washington, DC: National Digital Stewardship Alliance. http://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf.

The 2016 National Digital Stewardship Alliance (NDSA) web archiving survey is the third one conducted to better understand the landscape of web archiving activities in the US. The majority of respondents represented colleges and university libraries and archives. The "Access and Discovery" section outlines how organizations are facilitating access to their web archives, and the survey revealed that organizations continue to provide many forms of access. Two key trends have emerged from the survey: the number of organizations supporting search and item-level access points has declined, and the percentage of organizations creating collection-level catalog records and finding aids continued to grow. Some conclusions include the widespread use of Archive-It, including full-text search and browsable lists in its public portal, and diminishes institutional interest in providing localized full-text or URL-based search internally.

The 2016 survey included a new question asking respondents, "Do you have active researchers utilizing your web archives?" The relative youth of many web archiving programs and limited resources revealed that many survey respondents lack knowledge of how web archives are used. Organizations that answered "yes" were asked to provide a summary of how researchers were using their web archives. Narrative responses identified historians, social and political scientists, and institutional faculty as the primary research user community.

Bailey, Jefferson, Abigail Grotke, Kristine Hanna, Cathy Hartman, Edward McCain, Christie Moffatt, and Nicholas Taylor. 2014. *Web Archiving in the United States: A 2013 Survey,* National Digital Stewardship Alliance. http://www.digitalpreservation.gov/documents/NDSA_USWebArchivingSurvey_2013.pdf.

The 2013 National Digital Stewardship Alliance (NDSA) web archiving survey gathered data from organizations in the United States actively involved with programs to archive web content. The "Access and Discovery" section outlines how organizations are facilitating access to and discovery of their web archives. Respondents were asked questions that sought to identify which web archiver viewer respondents use, discovery mechanisms and discovery interfaces. The majority of respondents use the Wayback Machine as their access platform. The survey revealed that the most common discovery mechanisms for users are full-text search, URL search and title browse lists. Less than a quarter of respondents provide catalog records at either the collection or item level, while 20% indicated that they are providing discovery through finding aids. More than two-thirds (71%) make their web archives discoverable through a dedicated interface, while others make their web archives discoverable through a common discovery environment (14%) or both a dedicated interface and a common discovery environment (15%).

Ben-David, Anat, and Hugo Huurdeman. 2014. "Web Archive Search as Research: Methodological and Theoretical Implications." *Alexandria* 25: 93-111. doi:/10.7227/ALX.0022.

This paper addresses the theoretical and methodological implications of the transition to search functionality for web archive research. It introduces "search as research" methods, practices already applied in studies of the live web, and which can be repurposed and implemented for critically studying archived web data. Such methods open up a variety of analytical practices that have been precluded by the URL as sole entry point to a web archive; examples include the re-assemblage of existing collections around a theme or an event, the study of archival

artifacts and scaling the unit of analysis from the single URL to the full archive by generating aggregate views and summaries. “Search as research” scenarios were developed by the WebART project at the University of Amsterdam and the Centrum Wiskunde & Informatica in collaboration with the National Library of the Netherlands. The paper concludes with a discussion of current and potential limitations of “search as research” methods for studying web archives and ways with which they can be overcome in the near future.

Bernstein, Steven. 2016. “MARC Reborn: Migrating MARC Fixed Field Metadata into the Variable Fields,” *Cataloging and Classification Quarterly* 54: 23-38. doi:/10.1080/01639374.2015.1075642.

Bernstein argues that instead of dismissing the MARC format entirely, it could be revised to be compatible with modern metadata practices. He suggests that MARC fixed fields be converted and migrated into variable fields and gives detailed examples of the conversion process. The conversion could be applied to existing MARC records of websites, then converted to MARCXML or other metadata structures used for web archiving.

Bragg, Molly, and Kristine Hanna. 2013. *The Web Archiving Life Cycle Model*. San Francisco, CA: Archive-It Team, Internet Archive. https://archive-it.org/static/files/archiveit_life_cycle_model.pdf.

Two pages of this 30-page paper focus on metadata, presenting a framework describing the entire web archiving lifecycle. Like the policy function, the authors state that the metadata function overlaps significantly with other activities in the lifecycle rather than being a discrete task. Data gathered internally by the Archive-It team shows that 90% of Archive-It users generate collection-level metadata, 60% seed metadata and 15% document-level metadata.

Costa, Miguel. 2014. “Information Search in Web Archives.” PhD dissertation, Universidade de Lisboa. http://xldb.lasige.di.fc.ul.pt/xldb/publications/Costa:InformationSearchIn:2014_document.pdf.

This dissertation expands on work published by Costa and Silva in 2011. The abstract for their article “Characterizing Search Behavior in Web Archives” summarizes it adequately.

Costa, Miguel, and Daniel Gomes. 2015. “Web Archive Information Retrieval.” Paper presented at the International Internet Preservation Consortium, Stanford University, 28 April 2015. http://netpreserve.org/general-assembly/ga2015-schedule.

Several slides are of primary interest for WAM. Slide 5, Classes of need: navigational (seeing how a site evolved, 50–80%), informational (a topic, 14–38%), transactional (downloading/recovering a site, 5–16%). Slides 8 and 9, two access methods: URL and full-text. Most of the presentation focuses on relevance ranking; of great importance for effective full-text searching, but perhaps not so central for WAM.

Costa, Miguel, and Mário Silva. 2010. “Understanding the Information Needs of Web Archive Users.” Paper presented at the International Web Archiving Workshop, Vienna, Austria, 23 September 2010. http://xldb.di.fc.ul.pt/xldb/publications/costa2010understandingneeds_document.pdf.

This article reports the findings of a user study done on the Portuguese Web Archives. Key takeaways: users perform mostly navigational searches (rather than informational or transactional) and generally do not restrict searches by date. The researchers defined a navigational search as one in which the user is looking for a particular web page or site, while transactional searches are defined as web-mediated activities such as downloading a file. Unsurprisingly, the authors found that users prefer full-text over URL search. The authors used search logs, an interactive questionnaire and a laboratory study to obtain their data. In addition to the findings listed above, they determined that when users search by date, they are usually

looking for the oldest documents, and that more than half of informational searches were focused on people, places or things. They also concluded that the Portuguese Web Archive did not meet users' need to see and explore the evolution of a web page or site or to find images.

Costa, Miguel, and Mário J. Silva. 2011. "Characterizing Search Behavior in Web Archives." Paper presented at the First International Temporal Web Analytics Workshop, Hyderabad, India, 28 March 2011. http://ceur-ws.org/Vol-707/TWAW2011-paper5.pdf.

This article reports on a user study of the Portuguese Web Archive search logs, for which the users are primarily Portuguese. Users cannot be usefully characterized using only the anonymous data based on IP address and language that was gathered in this study. The study is focused on users' methods of searching web archives. The primary findings are that users of the PWA "iterate less" and instead often conduct only one or two queries in a single search session. Users who conducted full-text style searches more often sought "known-items" rather than broad topics; i.e., they often included names and titles. They also tended to submit lengthy search queries. Overall, users conducted URL searches about 31% of the time and full-text 59%. Finally, the authors found that users most often searched for the oldest version of a site and tended not to look at multiple versions of the archived site.

Costa, Miguel, and Mário J. Silva. 2012. "Evaluating Web Archive Search Systems." *Web Information Systems Engineering, Lecture Notes in Computer Science* 7651: 440-454. http://doi.org/10.1007/978-3-642-35063-4_32.

At least 77 web archives have been developed to cope with the web's transience problem. The authors observe that despite the technology in use for this study has achieved reasonable maturity, retrieval effectiveness of search services still presents unsatisfactory results. The authors propose an evaluation methodology for web archive search systems based on a list of requirements compiled from previous characterizations of web archives and their users. The methodology includes design of a test collection and the selection of evaluation measures to support realistic and reproducible experiments. The test collection made it possible to measure the effectiveness of state-of-the-art retrieval technology employed in web archives, and the results confirm the poor quality of search results retrieved. The paper reports on how to combine temporal features with the usual topical features to improve search effectiveness. The test collection is available to the research community. (Adapted from the published abstract.)

Cruz, David, and Daniel Gomes. 2013. "Adapting Search User Interfaces to Web Archives." Paper presented at the 10th International Conference on Preservation of Digital Objects, September 2013. http://sobre.arquivo.pt/sobre/publicacoes-1/Documentos-acerca-do-Arquivo.pt/adapting-search-user-interfaces-to-web-archives.

This article presents findings based on several rounds of usability testing on the Portuguese web archive search interface. Findings include: Since many people are still unfamiliar with the concept of web archives, the home page should have contextual information such as examples of archived pages or collections to explore; web archives should offer one search box for both URL and full-text searching; a "datapicker" that allows for easy navigation across many years would be useful; and a frame/top bar that doesn't interface with viewing of the archived website could remind the user that she is viewing an archived version. In that frame, they suggest including the URL, date, a link to help, ways to share (such as via Twitter or Facebook) and the ability to save a web page as a pdf.

Dooley, Jackie. 2015. Unpublished notes summarizing two sessions on web archiving at the Annual Meeting of the Society of American Archivists, Cleveland, Ohio, 20–22 August 2015.

In a session on programs that have minimal staffing, some speakers stated that they do no descriptive metadata for lack of time, while others create extremely minimal metadata in CONTENTdm or Archive-It. In a session on engagement and outreach, the need for both broad and institution-based use cases was discussed in the context of developing a business case to build management support for a web archiving program. Many use cases exist; it would be helpful to aggregate and analyze them.

Dooley, Jackie. 2016. Twitter posts during various presentations at the Annual Meeting of the Society of American Archivists, Atlanta, Georgia, 4–6 August 2016. https://twitter.com/minniedw.

The tweets were posted during presentations about a wide variety of web archiving issues. The need for scalable practices for descriptive metadata was discussed substantially.

Dooley, Jackie. 2015. Twitter posts from the General Assembly of the International Internet Preservation Consortium, Stanford University, 27-28 April 2015. #iipcGA15; https://twitter.com/NetPreserve.

Some key points made in the tweets, which were sent throughout the conference: Users feel a need for trust (speaks to provenance metadata). They are concerned about temporal incoherence and the completeness of crawls. They need rights and access metadata. Web archives have a potential impact on history/historical methodology, so there is a need for outreach and documentation. Host institutions should provide web archiving accounts for users. Users want to analyze content of web archives, not metadata. User needs are navigational, informational and transactional. Tweets include mention of case studies of people using web archives.

Dooley, Jackie. 2015. Unpublished notes summarizing various presentations, taken at the symposium *Web Archives 2015: Capture, Curate, Analyze*, University of Michigan, Ann Arbor, 10-12 November 2015.

This conference was exceptionally rich with papers concerning types of users, their needs and behaviors. Notes were taken for these presenters: Jefferson Bailey (Internet Archive: extensive overview of the landscape), Laura Uglean Jackson (UC Irvine: web archiving program case study), Jen Bonnet (University of Maine: archiving news sources), Brenda Reyes (University of Texas grad student: internship project to classify Internet Archive help requests so they could understand what client needs they need to support), Abbie Grotke (Library of Congress: 15 years of web archiving at LC), Ed Summers (University of Maryland: social media archives of the events surrounding the death of Michael Brown in Ferguson, Missouri), Ian Milligan (University of Waterloo: Canadian political parties and political interest groups using Archive-It), Juan Cole (University of Michigan: historian/journalist covering the Iraq war on the instability of web archives), and Jefferson Bailey's workshop on data mining of web archives.

Dougherty, Meghan, and Eric T. Meyer. 2014. "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: 2195–2209. http://onlinelibrary.wiley.com/doi/10.1002/asi.23099/abstract.

This peer-reviewed article is based in part on the interviews conducted in the Dougherty et al. 2010 report and focuses on much of the same content. High-priority needs are identified as 1) support for individuals and infrastructure that explicitly links the two groups; 2) collaboration between researchers and archivists/librarians; and 3) the need to build collections linked to specific research questions that allow for in-depth focus on multiple variables and deep metadata. The authors describe the fundamental elements of a web archive: "[Researchers] want stabilized web objects that can be reliably studied and cited. They want to be able to

clearly define what the stabilized archived object represents evidence of in reference to the live web. They want to have access to archived representations of the most fine-grained features of web objects in order to suit their research needs. Most of all, they want to work with those objects, enriching and annotating them on whatever level is appropriate for their analysis. In terms of the archive itself, three things are clear: An archive must be trustworthy, long-lasting, and reliable."

Dougherty, Meghan, Eric T. Meyer, Christine McCarthy Madsen, Charles van den Heuvel, Arthur Thomas and Sally Wyatt. 2010. "Researcher Engagement with Web Archives." (London: Joint Information Systems Committee Report.) http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1714997.

This report is based on interviews with researchers and web archivists. Key recommendations include: build communities that build connections between researchers (in various disciplines) and web archivists; provide clarity to researchers regarding copyright issues; support is needed for both infrastructure and research and should connect the two; tools are needed that are shareable and user-friendly and that include flexible building blocks to suit various disciplinary models; researchers need ways to deposit/share/publish their own web archives; researchers should be able to input their own data into existing archives (akin to crowdsourcing); necessary metadata includes administrative, descriptive and contextual/relational, allowing researchers to continue to add metadata over time; and basic training and outreach to researchers is needed.

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

This blog post focuses on three researcher-centered presentations at IIPC 2016. 1) Milligan and Webber used the hackathon model to engage with researchers, providing a hands-on experience for researchers in using web archives. Researchers from multiple disciplines with varying levels of technical skill were chosen and, with the help of the presenters, explored archived web sites. 2) Webber and Graham reported the results of a survey they conducted with librarians and researchers. They received 236 responses: 157 from researchers, the rest from librarians and archivists. Three main takeaways: a) Researchers are not leveraging existing resources; b) Librarians often have knowledge that the researchers need, but the researchers are not aware of this; c) Targeted outreach can lead to meaningful engagement with researchers. Barriers to use of web archives include lack of metadata, lack of tools to access and analyze the data, difficulty in finding relevant material, and datasets that are too large for some researchers to handle. 3) Jefferson Bailey of the Internet Archive focused on program models for research resources. He noted that WARC is a format unfamiliar to researchers, which represents a technical challenge. Researchers are much more concerned about what is not being captured and preserved than are researchers who use analog materials. Research support expectations often exceed available resources. Bailey recommends that librarians focus on derivation, portability and access if they want to support web archive researchers.

Gibbons, Leisa. 2016. *Web Archiving Project 2016: Preliminary Report.* http://leisagibbons.info/wp-content/uploads/2017/03/Webarchivingbriefreport-1.pdf.

Gibbons received 54 useable survey responses from self-selected respondents throughout the world, 40% of them from academic institutions and 26% from government archives. Her instrument covered a wide array of issues, including the purpose of web archiving, formal collecting policies, permissions, access, technologies and collaboration. This abstract focuses on the metadata section, in which she asked about types of collecting, what was done with

harvested metadata and which standards were used. She found that 11% of records had no metadata, 20% had only descriptive metadata, and 34% had both technical and descriptive. 44% use Dublin Core and 24% use one or more other standards (unspecified).

Goel, Vinay. 2016. “Beta Wayback Machine-Now with Site Search!” *Internet Archive Blogs*. Posted 24 October. https://blog.archive.org/2016/10/24/beta-wayback-machine-now-with-site-search.

This blog post highlights the beta search service of the Wayback Machine, which allows users to search for home pages of over 361 million websites through descriptive keywords. To date, users had no search/browse capabilities other than entering a URL in the main search box. The new keyword search features predictive, interactive search results, multilingual results and site-based filtering.

Goethals, Andrea. 2016. “New Report on Web Archiving Available.” *IIPC netpreserve.org* (blog). https://netpreserveblog.wordpress.com/2016/03/21/new-report-on-web-archiving-available.

This blog post provides an overview of Harvard's study (see entry for Truman, “Digital access”) based on interviews with web archiving practitioners worldwide, service providers and researchers. The researchers profiled 23 institutions. Common themes and challenges fell into four broad categories: increase communication and collaboration, focus on “smart” technical development, focus on training and skills development, and build local capacity. The post also provides brief descriptions of the 22 specific opportunities to address challenges that the study identified. One of the biggest takeaways was that communication and collaboration are essential for web archiving success, and these should be increased dramatically. The report also lists web archiving tools identified during the study.

Guenther, Rebecca. 2015. “Metadata for Web Archived Resources: Recommendations for Further Exploration.” Unpublished report prepared for the New York Art Resources Consortium. http://www.nyarc.org/sites/default/files/Recommendations%20for%20further%20exploration-final.pdf.

Guenther was the architect of the *Metadata Application Profiles and Data Dictionary for Description of Websites with Archived Versions,* issued by the New York Art Resources Consortium in 2015. In this follow-up report she recommends six areas for further research relating to descriptive metadata for archived web content:

- Integrating bibliographic and archival traditions of description
- Considerations for moving to a Linked Data environment
- Extending the use of the metadata profile beyond the art community
- Understanding user discovery needs
- Repurposing existing metadata
- Considerations for long-term preservation of archived websites

Guenther, Rebecca, and Leslie Myrick. 2007. “Archiving Web Sites for Preservation and Access: MODS, METS and MINERVA.” *Journal of Archival Information* 4: 144-166.

Guenther and Myrick discuss descriptive, technical and preservation metadata for web content. Using the UC Library of Congress’s MINERVA and the California Digital Library’s PCWA projects as examples, the authors discuss data models, notably the use of METS to aggregate descriptive (MODS) and preservation (PREMIS) metadata. The paper provides a brief history of web archiving, highlighting the need for capturing web content with a focus on political and

government sites. Further, the authors give a brief overview of national library approaches (circa 2007 and earlier) to meet legal deposit requirements. They believe that METS is uniquely robust enough for use in an OAIS repository (i.e. can meet the needs of SIP, AIP and DIP) and to meet lifecycle needs of these materials. MINERVA's approach included MODS records to reflect intellectual content. The article also describes METS schema and profiles; the question of how to describe the structural components of a website arises in this context. Technical metadata derived from crawlers is raised as an important and consistent set of data. The paper highlights the difficulty of automated generation of metadata from websites themselves, given inconsistent metadata practices of website creators. PREMIS background is provided, which suggests that further work was needed to create PREMIS records for websites. The authors recommend using METS objects to capture the complexity of metadata required for these objects (e.g., descriptive metadata at the top level and technical metadata at any level of the captured site, particularly the capture level), the model for which is described on pp. 159-16.

Haddad, Peter, and Pam Gatenby. 2002. "Providing Bibliographic Access to Archived Online Resources: The National Library of Australia's Approach." Paper presented at the Bibliography and National Libraries Workshop "Bibliographic Control or Chaos," at the 68th IFLA General Conference and Council, Glasgow, Scotland, 23 August 2002. https://archive.ifla.org/IV/ifla68/papers/069-152e.pdf.

This paper is from the point of view of the National Library of Australia (NLA), which has a mandate to acquire, catalog and preserve the national documentary heritage. The author argues for a bibliographic approach to describing selected online items because this would aid users in finding all pertinent material on a topic regardless of format, but she acknowledges that neither selection nor description practices are likely to be sustainable indefinitely. A high-level catalog record was prepared for most items, but exceptions were made when many websites were collected under a single topic. In those cases, the collection was assigned a title with added entries for each organization whose site was included. The library assigned a persistent identifier to each title and recorded this in the 856 field of the MARC, and so the version of the resource in the online collection remained accessible and citable. Other metadata to describe technical characteristics of the resources and to manage the archiving and preservation was recorded in the library's Digital Archiving Management System. The author clearly feels that the metadata provided by the publishers of online content was too volatile to be harvested and used to create the national bibliography. NLA recorded the characteristics of a publication at the time it was selected and catalogued, and aimed to describe it broadly enough to avoid the need for regular amendments. The date the online resource was viewed and catalogued was placed in field 500, and this date remained unchanged. If the record was subsequently updated, it was recorded in field 515. With

regard to access requirements, the 538 field recorded that the mode of access was the internet and also stated that the technical requirements for access and display where these were not standard, e.g., when non-standard plug-ins were noted (Adobe Acrobat).

Hartman, Cathy Nelson, Kathleen Murray, and Mark Phillips. 2013. *Classification of the End-of-Term Archive: Extending Collection Development Practices to Web Archives: Final Report.* Denton, TX: University of North Texas. https://digital.library.unt.edu/ark:/67531/metadc152437.

This is the final report for the End-of-Term archive, created in collaboration by five US institutions for documenting the federal government's web presence around the time of the change in leadership of the US executive branch in 2008. An assessment of research needs for the archive consisted of a small exploratory study with 11 academics, including a questionnaire and interview sessions. All interviewees used some form of web content, whether archived and online newspapers, historical web-published content or social media. In terms of access needs for web archives: researchers want to interact with archived websites along several dimensions

and would like search capabilities that combine keyword searches with selected dimensions of the websites. Digital humanities researchers are keen to know characteristics of the content (e.g., number of files and file types) in order to structure their research methods. In terms of data extraction needs for web archives: workability is a key factor, with expressed interest in data that is easily extractable in a format that can be imported into databases for analysis. Researchers need analytic tools, particularly in the areas of OCR, text mining, natural language processing and topic modeling. Content that is quickly and consistently mineable with custom-built tools increases the viability of a project. In terms of potential research uses: interest was shown in research that investigates change over time, the effects of differential languages, and compares and contrasts content between entities and websites. In conclusion, researchers are aware of the value of web archiving, particularly social media and blogs, but they will need assistance to build the tools that will enable them to find content and analyze their results. Once specific user needs are determined, web archive providers can develop such tools.

Hockx-Yu, Helen. 2014. “Access and Scholarly Use of Web Archives.” *Alexandria*, 25:1/2. doi:10.7227/ALX.0023.

Hockx-Yu analyzes reasons for limited scholarly use of web archives; examines key characteristics of scholarly use of digital resources and translates these into a set of requirements for web archives; discusses how current web archives fail to meet the evolving scholarly requirements; and offers thoughts on moving forward. She begins by asserting that access to web archives is problematic for two primary reasons: 1) restrictions on access, often due to legal requirements, and 2) the static “document-centric” approach to how this content is collected and presented. She argues that while users are not unfamiliar with access to content with similar restrictions, due to the “ubiquitous and open nature of the web” users expect to be able to access archived websites immediately and at any time, just like the live web. She further contends that the common user interface to web archives works “well with small, curated collections, but does not scale up and provide the users with a functional way to use larger collections.” In the end, these issues may impede the use of web archives for scholarly research.

Hockx-Yu cites a user survey commissioned in 2012 to gather a scholarly perspective on the Open UK Web Archive that is provided by the British Library and its partners. The survey was focused on the content and access mechanisms that should be further developed to support research use. She notes that while all participants appreciated the potential scholarly value of the archive, they wanted a deeper understanding of the how the selection criteria and themes were established, as the perception of those using the archive for the first time could not find content relevant to their research. Hockx-Yu states that most valuable lesson learned from the survey was that the relevance of content determines whether researchers use a web archive.

The crux of this article is how Hockx-Yu translates the results of the survey into tangible requirements for web archives. These include: linking geographically dispersed web archives (versus divided national domains) to increase availability; providing access information to researchers such as collection policy and scope, crawl configuration, and crawl logs for greater context; committing to an agreement on citation standards for archived websites and integration with common bibliographical tools for improved persistence and citability; providing information to researchers about what is missing from the original source (e.g., videos or other content that is challenging to crawl) to help researchers interpret incomplete sources; and giving researchers the ability to view and manipulate sources at different levels of granularity. With that said, she concedes that while it was relatively simple to arrive at these requirements, many are yet to be met.

She concludes by asserting that, as scholarly practices and methods change, providers of web archives need to respond accordingly. She challenges providers to engage with scholars in order to learn how to build collections that will meet scholarly needs and to build user interfaces that are self-explanatory, jargon-free and contain baseline information.

Hockx-Yu, Helen. 2016. *Web Archiving at National Libraries: Findings of Stakeholders' Consultation by the Internet Archive*. San Francisco, CA: Internet Archive. https://docs.google.com/document/d/1uP6DrwaUe-tbzz_SW7QMSKLTkCNpjg9nFzHTCoOwoOA/edit.

This work was based on interviews with national libraries, researchers and others. This abstract focus on section 5.3, Access and Research use: "Most national libraries' web archives are not publicly available, especially the large-scale national collections enabled by a legal mandate. These can only be accessed in the reading rooms or at premises controlled by the libraries. As a result, national libraries' web archives are mostly 'dark' and there is very little use of them. The perceived limited value of web archives sometimes leads to library management doing the minimum, investing little in web archiving. Many national libraries wish to support research use of their web archives by engaging with researchers to understand requirements and eventually embedding web archive collections into the research process. Some national libraries have shown great creativity and have made a lot of progress in engagement with the research community and developing research services, e.g., the British Library." The author sees misalignment with researchers' expectations. One of the conclusions: "... a general feeling that the community is stuck with a certain way of doing things without making any significant technological progress in the last ten years."

Jackson, Andy. 2015. "The Provenance of Web Archives." *British Library UK Web Archive Blog. Posted 20* November. http://britishlibrary.typepad.co.uk/webarchive/2015/11/the-provenance-of-web-archives.html.

This blog post provides an overview of the challenges the British Library has found in documenting provenance for web archives. The post describes their data mining project, BUDDHA, and their Shine search interface. The post lists types of provenance information, such as the crawl log, which is captured by the Heretrix3 web crawler, that can be used to understand how and why the library captured a resource. The author notes that this information requires understanding the crawler software used for capture and is therefore highly technical and challenging. The rest of the post focuses on gaps in provenance documentation and how they affect use of large collections of web archives. To meet the challenges of providing provenance information for corpus analysis, the team developed trend graphs that summarize information such as missed crawls, misclassifications and rejected URLs. The trend graphs allow researchers to view trends by year. Ultimately, the British Library wants to be able to provide researchers with trend graphs that allow them to determine whether a trend is real and to assess biases within the data set. Jackson believes that web archives should begin making large data sets available to users and allow them to work with available information, while providing as much information about crawl parameters as possible, even if it is not complete. Ultimately, he thinks that web archivists will develop practical solutions to summarizing provenance information that is useful to researchers.

Jackson, Andy. 2016. "Introducing Shine 2.0: A Historical Search Engine." *British Library UK Web Archive Blog. Posted 15* February. http://britishlibrary.typepad.co.uk/webarchive/2016/02/updating-our-historical-search-service.html.

SHINE 2.0 is a historical search engine developed by the British Library as part of the Big UK Domain Data for the Arts and Humanities (BUDDHA) project. The SHINE 2.0 search prototype searches over 3 billion records covering a date span between 1996 to April 6, 2013. As of the date this blog post was published (February 15, 2016), SHINE 2.0 can only cope with one or

two users at a time. This is a scalability issue that the project team is working with developers to resolve. The developers are encouraging users to try out the new historical archive and provide feedback on any issues. Users can propose collections to be crawled.

Jones, Shawn M., and Harihar Shankar. 2016. “Acquisition of Mementos and Their Content Is More Challenging than Expected.” *Web Science and Digital Libraries Research Group: Research and Teaching Updates from the Web Science and Digital Libraries Research Group at Old Dominion University* (blog). Posted 24 February. http://ws-dl.blogspot.com/2016/02/2016-02-24-acquisition-of-mementos-and.html.

Researchers extracted text from roughly 700,000 Memento objects from more than 20 different web archives spanning 1997–2012. Unexpected challenges made this much more difficult than text extraction from live web pages. Full findings are documented in a technical report linked from the URL, and two major difficulties are highlighted: 1) Acquiring data from WebCite archives couldn't be easily automated via command line. Necessity of cookies and content negotiation required browser simulation tool PhantomJS. Even then, it was not always reliable and took several tries. 2) Inaccurate and/or missing character encoding info in the Memento header. Researchers used an algorithm to detect and “best guess” encoding but not 100% accurate. The end result for both problems: solutions are possible but take much more time.

Takeaways relative to user needs: while Mementos are standardized as objects, lack of standardization in their storage and interpretation makes working across web archive collections and over periods of time very inefficient.

Lavoie, Brian, and Richard Gartner. 2013. *DPT Technology Watch Report*. Preservation Metadata, 2nd edition. Great Britain: Digital Preservation Coalition in association with Charles Beagrie Ltd. doi:10.7207/twr13-03.

This report updates *Preservation Metadata*, originally published in 2005. It focuses on new developments in preservation metadata made possible by the emergence of the PREMIS Data Dictionary as an international standard. While it makes no reference to web archiving specifically, the report emphasizes the importance of preservation metadata in that it facilitates “the process of achieving the general goals of most digital preservation efforts: maintaining the availability, identity, persistence, renderability, understandability and authenticity of digital objects over long periods of time.” The report provides a brief history of the evolution of preservation metadata development (including that it has moved from theory to practice), what it is and what it isn’t, what information comprises the metadata, and how well (or not) PREMIS and METS fit together. One of the most interesting conclusions is the need for future development in accumulating and consolidating best practices. The authors state that “an evidence base of detailed case studies on how preservation metadata is collected and managed within digital archiving systems would help shape consensus on a set of best practices for implementers, as well as illuminate areas of priority for technical development.” Although WAM is focusing primarily on descriptive metadata, it may be important to keep preservation metadata in mind throughout the discussions, as there is a fine line between the two.

Leetaru, Kalev. 2016. “Reimagining Libraries in the Digital Era: Lessons from Data Mining The Internet Archive.” *#bigdata* (blog). Posted 19 March 2016. http://www.forbes.com/sites/kalevleetaru/2016/03/19/reimagining-libraries-in-the-digital-era-lessons-from-data-mining-the-internet-archive/#1eeb690d6c7c.

Libraries are grappling with how to reinvent themselves in a digital world. They play a critical role in democratizing access to knowledge and ensuring a vibrant community of practitioners capable of bringing this knowledge to bear on societal needs. The need exists to bring scholars,

researchers, media and information practitioners together to reimagine libraries as centers of information innovation. Beyond preserving the web for future generations, library-based web archives offer researchers one of the few places they can work with large collections of web content. While not as extensive as the collections held by commercial search engines, library archives are accessible to scholars who lack the resources to build their own massive web-scale crawling infrastructures and uniquely allow the exploration of change over time at the scale of the web itself. As more and more academic institutions and non-profit organizations operate their own specialized crawling infrastructures focused on specific subsets of the internet, library-based archives are in a unique position to collaboratively leverage the feeds of URLs generated by these projects to massively extend their research.

As mass-scale web crawling becomes increasingly accessible, such partnerships could provide a powerful opportunity for web archives to extend their reach beyond in-house crawling infrastructures, especially to deepen their collection of highly specialized content. As the Internet Archive works toward enabling full-text search of its web holdings, countless opportunities will exist for the Internet Archive to collaborate with the research and visualization communities to think beyond simple keyword search and toward addressing user interface challenges for presenting archived web content in a meaningful way. The real challenges are how to assess relevance in a temporal index and how to visualize search results that stem from how a page has changed over time. Rather than try to build such systems in-house, libraries should reach out to the open-source and research communities for creative solutions to these challenges as libraries increasingly open their digital collections to access.

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

Mannheimer studied metadata creation by 57 Archive-It partner institutions. The author is rooted in an archival mindset and discusses the importance of context and connecting web archives to archival collections, though her survey findings provide information from bibliographic perspectives as well. The majority of respondents used Dublin Core to describe collections, which is inherent to Archive-It. The most common elements in use were Creator, Description and Title. This study doesn’t provide insight into how the fields are defined by different institutions (e.g., which date is used for Date?). The survey didn’t provide a way for respondents to differentiate between practices for collection, seed or item level, so the level of description is not identified in responses to questions about which Dublin Core fields are used. The use of controlled vocabularies was common (LCSH, LC authorities, ISO language and dates, FAST, AAT, DCMI Type, and more). Question 9 on the survey is interesting: Who determined what metadata elements are used in the web collections at your institution? Job titles vary a great deal, but those stated often were senior staff or metadata specialists. Regarding delivery of metadata, the responses to question 14 regarding access to web collections shows that most respondents provide access to web collections through the Archive-It interface and/or add metadata from finding aids, websites and library catalogs. Eleven respondents provide metadata in catalog records, and 12 do so in finding aids. Question 15 addresses metadata transformation, and eight respondents reported use of MARC for adding metadata to a catalog. The survey also explored factors that are barriers to metadata creation for archived websites; the majority of respondents reported lack of time as the top barrier. The study did not explore the reasons behind lack of time beyond noting that many respondents have job duties beyond web archiving. It would be interesting to know if the lack of time was at all related to perceptions about how much metadata is enough metadata (e.g., aggregate approach vs. item-level approach).

Masanès, Julian. 2005. “IIPC Web Archiving Metadata Set.” Presentation for paper titled *Metadata for Web Archiving* presented at the 5th International Web Archiving Workshop (IWAW), held in Vienna, Austria, 22 September 2005. http://iwaw.europarchive.org/05/masanes2.pdf.

> Although the title of this slide deck promises a metadata set, it instead outlines characteristics of crawls and relationships among file types that are useful to capture. These include technical metadata needs such as information about the document, server and crawler; the crawl itself, the selection decision; and change history. It does not include descriptive metadata elements.

Milligan, Ian. 2015. “Web Archive Legal Deposit: A Double-Edged Sword.” *Ian Milligan, Digital History, Web Archives, and Contemporary History* (blog). Posted 14 July 2015. https://ianmilligan.ca/2015/07/14/web-archive-legal-deposit-a-double-edged-sword.

> Milligan recognizes the importance of legal deposit in the context of web archiving: a recognition that born-digital sources are today’s documentary record, the need to preserve it is urgent, and institutional and legal commitment are necessary to make it happen. He raises questions about the existence of restrictions of the sort placed on rare books, such as onsite consultation only, limitations on reproduction, and a maximum of one person at a time viewing a website. His experience in accessing the UK Web Archive at the British Library (BL) was mixed for such reasons. Legal deposit requirement means the library is obligated to archive the web content, thus acquiring content not in other web archives. Use of the library’s discovery tool (Primo) allows the BL to build an interface such as SHINE, with a robust backend based on W3act tool for annotation and curation. He found restrictions on use frustrating, especially being limited to access via an onsite virtual machine, a “no photography” rule, static content, hyperlinks not visible, one-user viewing, onsite-only use and the inability to view source code. In conclusion, while he likes the idea of mandatory legal deposit for websites, he emphasizes that restrictions make it more difficult to use the web archive than the print record.

Murray, Kathleen. R., and Inga. K. Hsieh. 2008. “Archiving Web-Published Materials: A Needs Assessment of Librarians, Researchers, And Content Providers.” *Government Information Quarterly* 25(1): 66-89. https://digital.library.unt.edu/ark:/67531/metadc29322/m2/1/high_res_d/Murray-2008-Archiving_Web-Published_Materials.pdf.

> This article summarizes a needs assessment prepared as part of the Web-at-Risk project, a three-year collaborative research effort to develop the University of California’s recently retired Web Archiving Service (WAS). The purpose of the study was to identify the web archiving needs and issues of librarians, researchers and content providers. The salient points related to user needs include:
>
> - Users were interested in the unit of selection, or granularity, that curators might specify for the capture of web-published materials. It was noted that for certain research disciplines or types of research, the context of source material is critically important.
> - While each user may assess authenticity differently, many users need and want some authority to provide assurance of the authenticity of a web archive. The authors recommend that curators (or others responsible for selecting web content) assign versions and dates to captured sites and identify the location of original source materials.
> - Librarians anticipated that in creating web archives, the biggest challenge would be the application of metadata, and most thought that automated metadata generation, including subject and topic classification, would be needed.

- Researchers indicated that the most important types of searches are “topic or subject” and “full-text using any keyword.”

- Researchers asserted that web archives should make it clear that users are interacting with archival material and not “live” material. Additionally, it was recommended that archived websites include a statement identifying the archive as an “official” or an “unofficial” version of the source materials.

- Finally, researchers suggested it would be of value if an archive provided descriptions of both the provenance of materials and the material preservation activities undertaken, and descriptive tagging of interactive links.

The authors conclude that an urgent and growing user information need exists for collection and preservation of web-published materials. They hope that while the “user needs are great and the organizational resources are scarce,” the results of their survey will benefit web archiving efforts.

Neubert, Michael. 2015. “Users, Uses Cases, and Adapting Web Archiving to Achieve Better Results.” *The Signal (blog). Posted 18* May. https://blogs.loc.gov/digitalpreservation/2015/05/users-use-cases-and-adapting-web-archiving-to-achieve-better-results.

Neubert, former Supervisory Digital Projects Specialist at the US Library of Congress, writes about harvesting large websites and issues relating to completeness of a harvest. Large websites may not be captured fully, and no assurance exists that all content will be captured even with recurring captures. He posits that one complete capture of a website may be more useful to end users than several incomplete captures. The library plans to implement strategies to fully capture sites by means such as making use of RSS feeds and occasional attempts to harvest a whole site

thoroughly to fill gaps. For future users of these archived websites, this could alleviate frustrations such as encountering broken links while browsing and could provide more reliability in accessing captured content.

Niu, Jinfang. 2012. “Functionalities of Web Archives.” *D-Lib Magazine* 18. http://www.dlib.org/dlib/march12/niu/03niu2.html.

Niu’s article focuses on access to and use of web archives. She derived a checklist of functionalities from established use cases and compared this with functionality in the ten most-used web archives. Findings are that search and browse functionality, including known-item and exploratory search, with faceting options, are commonly supported, but more advanced functionalities, such as data mining, personalized services and the reconstruction of lost websites, are not.

Niu, Jinfang. 2012. “An Overview of Web Archiving.” *D-Lib Magazine* 18. doi:10.1045/march2012-niu1.

This study, based on a literature review conducted by the author, examines methods used by various universities and governmental libraries to select, acquire, describe and access web resources. This abstract is in two parts: issues related to end-user needs and those of metadata practitioners.

Niu uses “appraisal” synonymously with “selection,” i.e., evaluating the value of content and deciding whether and/or for how long to preserve. She cites domain, topic, event, media type and genre as selection criteria currently in use and asserts that, “Theoretically, selection based

on objective criteria can be easily automated." A WebAnalyzer can be integrated with the crawler to look for pre-defined properties of sites. Manual selection is expensive and therefore viable only for small-scale web archives.

In contrast, the Arizona model (Pearce-Moses and Kaczmarek. 2007) uses macro-appraisal theory for government websites based on aggregates of pages. Representative sampling avoids the subjectivity and bias in value-based appraisal; this approach is used by the National Library of France to determine seed list and filtering criteria prior to crawling.

The author contends that the approach to and richness of metadata must depend on the scale of the archive and organizational resources. Very large archives often rely on automatic metadata generation; timestamp, status code, size in bytes, URI and MIME type can be created/captured by crawlers or extracted from HTML metatags. The archival approach uses multilevel description, which starts at the highest level of a collection and continues to lower levels as resources permit; lower levels inherit data from higher levels. The bibliographic approach, on the other hand, uses single-level description; once again, the level of detail should be based on scale of the archives and resources available. The US Library of Congress and Harvard University both create one MARC record for a collection of many websites. Niu describes characteristics of the provenance of a site, which include URL, content producer and purpose. She combines discussion of the archival principle of original order with the concept of structure; internal structure is defined by hyperlinks within the site, while incoming/outgoing links define external structure and original order. "In a web-based archive, web pages and associated metadata are grouped and stored in container files and the original URIs and links are preserved." A non-web-based method extracts web documents from hypertext content into catalog-based access or PDF files; in this approach, authenticity and integrity are mostly lost. Level of access usually depends on the legal environment of the host country, for which she provides examples.

O'Dell, Allison Jai. 2015. "Describing Web Collections (I Mean Archived Websites)" *Medium* (stories). Posted 17 February 2015. https://medium.com/@allisonjaiodell/describing-web-collections-e32b59893848#.lbg88xo51.

O'Dell describes her thought process for developing cataloging practices for collections of archived websites. She doesn't go into detail about how the University of Miami Libraries use Dublin Core fields in Archive-It, nor specifics of how or whether archived websites are included in finding aids. Her focus is on creating a MARC record for each collection using OCLC's "books" workform. The post works through a series of questions related to MARC fields: Creator element: the library is the creator with a role of "collector." Title: O'Dell feels that leading with "University of Miami Libraries Collection of…" is misleading, and so they instead focus on the specific resources, as in "Cuban Theatre Web Collection | University of Miami. Date: crawls are repeated for many collections and so are described as for serials with formatting from DACS (YYYYMMDD). Frequency of capture is also noted. Extent: this area is particularly challenging for web archives. O'Dell provides a list of great questions but admits she didn't arrive at an opinion. At present, Miami uses "1 collection of archived websites." She briefly discusses subjects (abstract, LCSH terms, local terms). Lastly, a link to the Archive-It collection landing page is added to the record. The post provides a helpful list of MARC fields to use with specific RDA or DACS rules.

OCLC Research Library Partnership Metadata Managers Group. 2016. (Notes contributed by group members on their institutions' web archiving metadata practices and needs. Not publicly available)

This is an ongoing focus group is comprised of top-level metadata managers at large research libraries. The group selects topics of broad interest and explores them in depth. One topic investigated in 2016 was web archiving metadata practices at 31 of the libraries. Some themes:

- A fair number haven't yet started web archiving, while most that have use Archive-It.
- None are extracting descriptive metadata by automated means because it's too thin to be worthwhile, with the exception of what they've entered into Archive-It.
- One respondent is starting to explore extraction further, however, since producing site-level MODS manually is labor-intensive.
- Across the respondents, cataloging is done predominantly by archivists, while some use professional library catalogers or students.
- Several use DACS as the content standard, though one considers it inadequate for addressing web-specific issues.
- Another respondent noted that sites described as records are governed by DACS and publications by RDA; they note the need for URIs for entities to enable linked data use.
- The majority provide discovery only via Archive-It. The issue of wanting to integrate discovery of web content with other resources comes up repeatedly.
- A respondent notes they are "still discovering how researchers are searching and using web archives;" their notes include some interesting comments about user needs.
- Another respondent provided detailed description of site-level descriptive practices and noted that they would like to provide more document-level cataloging but are daunted by the scope of the undertaking.

Peterson, Christie. 2015. "Archival Description for Web Archives," *Chaos→Order*. Posted 12 June 2015. https://icantiemyownshoes.wordpress.com/2015/06/12/archival-description-for-web-archives/.

> Peterson writes about her experiment to describe web archives using Archivists' Toolkit, Archive-It, DACS and EAD. Issues discussed: The appropriate fields for descriptive elements. A lack of overlap between units of arrangement, description and access in web archives as compared to analog archives, particularly with regard to creation and provenance data. She raises the issue of needing to describe collecting practice when considering use of content (she references Columbia Web Archiving Collaboration conference talks in 2015 by Jimmy Lin and Ian Milligan), which is essential for understanding what was omitted from the dataset. She also experimented with inclusion of appraisal decisions in the context of the Archive-It metadata capture tool and provides a brief overview of Archive-It metadata structure (seed, collection, crawl). Appraisal decision data are related to crawls in Peterson's model. Using crawls as the primary unit of description allowed her to consider evolving inclusion of seeds (e.g., as new student groups emerge), essentially approaching the collection and seeds within the collection from a "complete" crawl of that collection; this is not feasible in the Archive-It structure.

Phillips, Margaret E., and Paul Koerbin. 2009. "PANDORA, Australia's Web Archive: How Much Metadata Is Enough?" *Journal of Internet Cataloging* 7(2): 19-33. doi:10.1300/J141v07n02_04.

> Phillips and Koerbin describe Australia's PANDAS (PANDORA Digital Archiving System), a web archive that selectively aggregates content from State libraries that is stored at the National Library of Australia. The authors give an overview of the project from the staffing, selection, acquisitions, copyright and legal deposit points of view; they also mention takedown practices. Metadata in PANDAS includes administrative (publisher contact details, permissions granted, crawl schedule, capture issues, etc.), as well as descriptive, harvest and IP data. A 1990s-era working group is described that specified improvements required for MARC to document digital files and preservation needs, as well as early 2000s PREMIS development.

The approach taken by NLA included MARC records to ensure discovery in the online catalog; at the time, they were investigating more efficient ways to produce metadata. The article mentions web content as similar to serials and integrating resources due to the quantity of "traditional" publications content in the archive (such as publications hosted on the web); this demands ongoing assessment of records for each crawl. Further, they sometimes created records at the site and individual title level (such as newsletters). Their internal analysis determined that cataloging was the second-most time-consuming activity in the program. Procedures and standards used include MARC and AACR2, and specific fields used are detailed (such as 583 for preservation data). The records do not include detailed system requirements, but do include a note indicating plug-in and software requirements. MARC is crosswalked to AGLS Dublin Core for the PANDAS entry page, thereby increasing scalability.

Pregill, Lily. 2016. "Web Archiving: Description and Access." Paper presented in the Metropolitan New York Library Council webinar series, 29 February 2016. http://www.slideshare.net/ElizabethLilyPregill/web-archiving-description-and-access.

Descriptive metadata and full-text indexing are essential to drive discovery and retrieval of web archives. The Wayback Machine enables multiple levels of search, combined full-text and Dublin Core metadata search on collection pages. The New York Art Resources Consortium discovery system (NYARC) includes two discovery display options: citation only or web archive collection results. The author describes NYARC cataloging workflow, which begins in WorldCat Connexion. Sample records were displayed in NYARC Discovery, the NYARC online catalog and in WorldCat.

Prom, Christopher, and Ellen Swain. 2007. "From the College Democrats to the Falling Illini: Identifying, Appraising, and Capturing Student Organization Websites." *The American Archivist* 70 (2): 344-363. http://dx.doi.org/10.17723/aarc.70.2.c8121767x9075210.

Prom and Swain explore record-keeping practices of student organizations at the University of Illinois Urbana-Champaign (UIUC), with a special focus on use of websites. The authors suggest websites are valuable records of student life that might actually make it easier for archivists to collect student organization records. The authors describe the structure and content of student organization websites at UIUC, appraisal decisions, and methods and technology used for capture (e.g., HTTrack). University websites are in one record series, with a collection-level description in the collection database. The authors note that a detailed finding aid was not needed as the technology used to capture the sites provides a "lead-in" list to all the organizations crawled. The list can be browsed or searched for access to the websites. The collection-level description is available for users online, but access to the websites is limited to the reading room.

Ras, Marce,l and Sara van Bussel. 2007. "Web Archiving User Survey." National Library of the Netherlands (Koninklijke Bibliotheek). https://www.kb.nl/sites/default/files/docs/kb_usersurvey_webarchive_en.pdf.

This report discusses findings from a 2006 survey evaluating the use of the Koninklijke Bibliotheek (National Library of the Netherlands; KB) web archive, which began that year. The report discusses design of a user survey and identification of potential end users. In particular, the authors discuss how users find web archives at KB, whether through a URL-specific search or word queries using a search engine. The central survey question asks: "What should the contents and search possibilities of the KB web archive look like, taking into account the potential users and stakeholders?" The survey methodology included presenting users with different potential use scenarios, the actions necessary to accomplish this in terms of selection, and whether these functionalities are offered anywhere else. Categories of end users surveyed include historians, sociologists, linguists, journalists, owners/designers of

websites and the general public. At the time of publication, the KB's web archive was still in its test phase and a limited number of websites have been archived. The authors speculate that web archiving will become more of a source for both scholarly research and general use as web archives grow and age. Preserving metadata is also essential so that users can identify the nature of an archived website and how it should be presented.

This report includes appendices with a list of web archives consulted and the user scenarios presented in the survey. It is also noted that a comparable qualitative survey was conducted by the Bibliothèque nationale de France (BnF), and the results of both surveys will be combined with another on Access Requirements commissioned by the British Library. The results of these three surveys were to be furnished to the Access Working Group of IIPC.

Reynolds, Emily. 2013. "Web Archiving Use Cases." Washington, DC: Library of Congress, UMSI, ASB13. http://netpreserve.org/resources/IIPC_archive-UseCases_Final.pdf.

This document provides a useful framework by placing web archiving use cases into three broad categories: 1) data mining and analysis, 2) preservation and stability, and 3) outreach and education. Data mining and analysis has four subcategories: text mining, link analysis, analysis of technology trends, and geographic analysis and mapping. Preservation and stability also has four subcategories: persistent linking, access to deleted or modified content, accountability and historic preservation.

Riley, Harriet, and Mark Crookston. 2015. *Awareness and Use of the New Zealand Web Archive: A Survey of New Zealand Academics*. New Zealand: Victoria University of Wellington and the National Library of New Zealand. http://natlib.govt.nz/files/webarchive/nzwebarchive-awarenessanduse.pdf.

This survey studied humanities and social sciences academic users of the New Zealand national web archive. The findings align with a number of others in terms of user needs: a strong need exists for outreach and education, and users prize full-text search above all other types, though the data also show that URL searches remain desirable for a steady minority of 28% of users. Each archived website has a record within the library's online catalog, which can be searched by keyword, title, subject and name. The authors note that some users may experience confusion about the nature of archived web content due to the integration of records for all types of cataloged materials; i.e., an inexperienced user may be unable to decipher whether a catalog record is referring to an archived website.

Stirling, Peter, Philippe Chevallier. and Gildas Illien. 2012. "Web Archives for Researchers: Representations, Expectations and Potential Uses." *D-Lib Maga*zine 18. doi:10.1045/march2012-stirling.

The Bibliothèque nationale de France (BnF) interviewed potential users of its web archives, particularly those studying the Internet. Types of data their subjects sought about the web include: qualitative data: online contents analyses (images or texts); quantitative data: figures and calculations, such as connection time, number of contacts, number of readers of a given site; and relational data: social networking analysis. Key findings related to user behaviors include: researchers used filtering strategies such as focusing on social networks such as Facebook, Twitter or blogs. Websites are cited as evidence or illustration of a sociological or historical phenomenon. Researchers create their own personal archives (printouts or screenshots) or amass large quantities of information and use their own textual analysis tools. Use of the Internet Archive is common. Some problems seen: ethical: it is difficult to anonymize such data; websites that have disappeared; and web data are seen as unreliable and ephemeral. The authors conclude that the study confirms that BnF's "mixed model," of large-

scale crawls complemented by focused crawls based on manual selection, seems to respond best to their demand. Tools should be put in place that allow researchers to know whether a site has been archived, navigate within the web archive and discriminate between sites.

Sweetser, Michelle. 2011. “Metadata Practices Among Archive-It Partner Institutions: The Lay of the Land.” Slide lecture presented at the Archive-It partners meeting, 19 October 2011. http://slideplayer.com/slide/3847250/.

Sweetser presents her data from a metadata survey of Archive-It partners. She includes detailed demographic information about respondents, such as institution type and staffing for web archiving. Slide 11 presents detailed information about the percentage of use of various metadata structures such as collection-level records): 35% of respondents prepare data at the Archive-It collection level, while 35% do not, and 81% do not prepare metadata for individual documents. Available time is by far the greatest factor in deciding which seeds will be described. Slide 14 shows use of each of the 12 core Dublin Core elements and whether each is used at the collection or seed level. The Description element is by far the most frequently used, followed by title, creator, subject and date. Two-thirds or more do not put metadata in their library catalog. One-third include website metadata in finding aids together with related analog material. Overall, the author suggests that practitioners generally create minimal descriptive metadata, whether in Archive-It, catalog records or finding aids. She suggests three possible reasons: 1) web archiving programs are in their infancy and institutions haven't “gotten around” to creating metadata within Archive-It; 2) organizations do not believe that metadata is useful enough to be created; and 3) organizations are focusing their metadata creation efforts outside Archive-It. An Archive-It search box is present on 41% of respondents’ websites. Roughly 50% are satisfied with users' ability to find their web content.

Taylor, Nicholas. “2013 NDSA Web Archiving Survey Report Highlights.” Paper presented at the Society of American Archivists Annual Meeting, Washington, D.C. 13 August 2014. https://www.slideshare.net/nullhandle/2013-ndsa-web-archiving-survey-report-highlights?from_action=save.

The survey covered a broad array of issues, including metadata practices. Slide 20 is a bar graph showing the granularity of metadata used by survey respondents who employ various means of discovery for archived websites, listed in descending order of their frequency of use: URL search, full-text search, browse by URI, browse by title, collection-level description, item-level description, finding aids, APIs, other. The percentages dropped in most categories from the 2011 to 2013 surveys. (See also entry for the full report under Bailey et al. 2014.)

Thomas, Arthur, Eric T. Meyer, Meghan Dougherty, Charles van den Heuvel, Christine Madsen, and Sally Wyatt. 2010. “Researcher Engagement with Web Archives: Challenges and Opportunities for Investment.” *Joint Information Systems Special Report*. London: JISC. https://papers.ssrn.com/sol3/papers2.cfm?abstract_id=1715000.

This report looks at challenges and opportunities of web archiving for researchers and of moving web archiving technology and practices to higher levels of comprehensiveness and usefulness. Major challenges stem from the rapid pace of technological development of the Internet due to new web content formats, such as multimedia and dynamic and executable content; social media platforms; the “deep web” of database content that defies crawlers; and mobile devices and apps that use proprietary communication protocols rather than HTTP. The report also suggests new opportunities in seven areas:

1) Put more effort into supporting users to survey, annotate, contextualize, and visualize repositories, and focus more on the thematic content of the users’ interests.

2) Develop means to blur the distinction between archival and live content, thus capitalizing on the powerful search, annotation, visualization and analysis tools available for the live web.

3) Invest in technical methods for collecting the new content types via web services, using workflow tools that allow collaboration between archivists and users to build new thematic collections.

4) Apply semantic web and linked data technologies, especially the use of the Resource Description Framework (RDF) and associated query and inference languages that support crosslinking of information from separate archives into meta-archives.

5) Incorporate these new metadata methods to allow more precise semantic search of large-scale archives.

6) Use cloud storage architectures to achieve better economies of scale and improved data management practices.

7) Work with the web science community, which has already explored similar issues related to web development, content creation and other internet architecture issues such as trust and privacy.

Thurman, Alex. 2016. “Web Archiving: Description and Access.” Paper presented in the Metropolitan New York Library Council, Web Archiving Series, Part 3, 29 February 2016.

> Thurman details the approaches that the Columbia University Libraries have taken to create descriptive metadata for web archives. Thurman first makes the case for the importance of descriptive metadata for searching and browsing web archives and then details some factors that affect description, including an institution's rationale or use case for archiving web content and the granularity with which web archives are described. He details several levels and methods of granularity, including the collection-level metadata in Archive-It, MARC collection-level records in Columbia's catalog, seed-level metadata in Archive-It, and finding aids that provide collection level description and a list of seeds. He describes how Columbia has made use of Archive-It's custom metadata fields to be more systematic and granular in its description of seeds within its New York City Religions collection. He also details the ways in which Columbia creates MARC records for site-level records, which include links to all seed URLs under which a site was captured (i.e., if a seed has changed, the MARC record provides links to captures under the earlier and current URLs) and a link to the current site on the live web (if it is still available). The presentation continues with examples of other ways in which Columbia attempts to provide access to archived web content, including additional uses of MARC fields in Columbia's catalog, access portals for specific collections, and the addition of links to Columbia's archived web content to Wikipedia articles.

Thurman, Alex, and Eric O'Hanlon. 2015. “Solr-Powered Full-Text and Metadata Search in the Human Rights Web Archive.” Paper presented at the General Assembly of the International Internet Preservation Consortium, Stanford University, 30 April 2015. http://netpreserve.org/general-assembly/ga2015-schedule.

> This slide deck provides an overview of the custom system developed at Columbia for their Human Rights Web Archive. The system searches both bibliographic records and full text of the sites captured. Users can browse by subject, website title, URL, geographic place and language. The systems allow users to facet searches by domain, date of capture, file type, geographic focus, location where the organization is based, language, subject and organization type. Users also can limit searches to the descriptive metadata.

Tillinghast, Beth. 2011. “Access to Our Archived Web Collections: Can We Be Doing More?” Paper presented at the Archive-It Partner meeting, Honolulu, Hawaii, 19 October 2011. http://slideplayer.com/slide/7705814/.

This is a brief presentation on access points and other aspects of metadata for archived websites. The author gives examples of various ways that metadata can be repurposed, including on an institutional website, Archive-It's partner page, general web search, the institution's online catalog and WorldCat. This is a good summary of how and where descriptive metadata can be used for discovery of archived websites.

Truman, Gail. 2016. “Digital Access to Scholarship at Harvard: Web Archiving Environmental Scan.” *Harvard Library Report*. Cambridge, MA: Harvard University. https://dash.harvard.edu/handle/1/25658314.

This extensive report states that “The ultimate goal of this survey is to identify opportunities for future collaborative exploration.” The author explores a wide range of issues in her environmental scan.

- Discovery: We can’t expect researchers to know the URLs of sites, so other query types are a paramount need. Lack of good metadata and quality control also are critical issues. We need more federation of resources. Memento might help but is still just a URL-based tool. Staffing: According to the 2013 NDSA survey, average staff devoted to web archiving is only .25 FTE, though the importance and complexity of web archiving requires at least dedicated 1 FTE.

- Location: where web archiving “lives” administratively (archival vs. bibliographic) in an institution influences web collecting policy. More communication is needed between collector types to make sure user needs are addressed despite location of the web archive.

- Community: no single collaboration/community covers all types of web archiving practitioners. Very little overlap exists between researchers and collectors, which leads to insufficient dialogue. Funded collaborative projects can help collectors get feedback on requirements/constraints of researchers. She suggests leveraging RESAW and European IIPC member overlap to facilitate formal collector/researcher projects.

- Collection Development: She found frustration at lack of coordination between institutions, which results in fragmentation and duplication of content. “How are we facilitating who's collecting what? Is there a shared knowledge base?” Institutions need to share at least a finding aid for restricted-access collections. To enable broader capture, the IIPC and similar organizations could pressure Google, YouTube and other internet firms on the importance of supporting web archiving by libraries and archives via both policies and the technical design of web content.

- Tools: Most are specific to narrow media types (such as tweets) or analysis type (such as link analysis) and are not designed to be used together. No off-the-shelf tool exists for initial collection development. Early capture tools are outdated and need refreshing.

- Researcher use: Other than using the Internet Archive’s Wayback Machine, most researchers prefer to create their own collections. Researchers generally are more interested in data/programmatic tools than URL/keyword exploration but need more training and expertise. More than anything, researchers want to understand the curatorial decisions made; i.e., as much documentation as possible on how and why a site was archived. To do this at scale we need to automate and develop a standard.

- Infrastructure: More and more institutions use Archive-It, which makes the most sense in terms of cost, but this may affect large dataset use. “Computing resources are best deployed close to the data,” but this is difficult with distributed hosting.

Weber, Matthew, and Pamela Graham. 2016. “Internet and Web Archives Research Use Survey.” (Unpublished survey data, 24 February 2016.) [Project Announcement: https://networks.h-net.org/node/73374/announcements/110145/survey-internet-and-web-archives-research-use-study]

> Weber and Graham conducted a survey in 2016 to study use of web archives by scholars and librarians. They received 157 responses, mostly from graduate students and faculty. As of January 2018, the data has not been published, so this abstract includes only a few data points. History and digital humanities were the most frequent disciplines of respondents. Half of respondents build their own web archives. The metadata elements cited as being of most value, in descending order of frequency, are title, subject, owner/creator/publisher, descriptive keywords, dates of capture, rationale for selection, language and geographic scope. The most widespread challenge was lack of descriptive metadata for archived content.

Webster, Peter. 2016. “What Do We Need to Know About the Archived Web?” *Webstory: Peter Webster's blog.* Posted 18 April 2016. https://peterwebster.me/2016/04/18/what-do-we-need-to-know-about-the-archived-web/.

> Webster argues for five broad categories of enhanced metadata and documentation that users need: 1) history of the collecting organization, 2) how the domain for the crawl was determined, 3) specific criteria for selecting the sites crawled, 4) when the first and last attempts were made to crawl both the domain and each specific site, and 5) whether a 404 response was ever received.

Wu, Paul H.J., Ichsan P. Tamsir, and Adrian K.H. Heok. “Applying Context-Sensitive Web Annotation in Evidence-Based Collaborative Web Archives Cataloging.” Paper presented at the International Web Archiving Workshop, Alicante, Spain, 21-22 September 2006. http://dl.acm.org/citation.cfm?id=2173354.

> This paper preceded the authors’ 2007 article “Annotating Web Archives Structure, Provenance, and Context Through Archival Cataloguing” (below) in which they largely disavow this 2006 paper. Here, they propose that the use of context-sensitive annotation will help ensure the consistency and quality of the web archives cataloging process in an institutional environment. They advocate that this is a better option than context-free annotation, as it provides evidence of the context of the semantic content. They argue that “without context, it is almost impossible to verify cataloging results.” Context-sensitive annotation establishes the relationship between the metadata and the content of the web material. The article evolves into a rather technical piece as the authors describe in detail the WAWI (Web Annotation for Web Intelligence) annotation system, which they developed to facilitate evidence-based cataloging for web archives, and which can be used as part of the web curation process. The authors suggest this as a new approach to web archives cataloging. The final section provides a brief survey of existing web archives catalog and access systems, which may be the most useful aspect of the article; they conclude that taking a bibliographic view of arranging online materials has its limitations because online publications are volatile in that they change regularly, but the changes oftentimes are not evident.

Wu, Paul H.J., Ichsan P. Tamsir, and Adrian K.H. Heok. 2007. “Annotating Web Archives Structure, Provenance, and Context Through Archival Cataloguing.” *New Review of Hypermedia and Multimedia* 13:55-75. http://www.tandfonline.com/doi/pdf/10.1080/13614560701423620.

> The authors describe their concept for WAWI (pronounced “wowie”), or Web Annotation for Web Intelligence, by which they mean cataloging and description. They refer to this as “context-aware annotation,” which “establishes the relationship between the metadata, the content of the web material, and the social context in which the content was produced,” an approach that follows standard archival principles, although the authors do not acknowledge this. A context-aware approach reveals “the working behind how a decision was taken to annotate web materials ... made visually obvious.” They describe organizing related pages into collections and then utilizing the hierarchical relationships among units of an organization and types of content/outputs for web pages are made to present the metadata in a tree model. Data elements for each unit of description are Title, Alternative Title, Creator, Subject, Description and Date Created. They also mention drilling down to use expanded Dublin Core elements such as CreatedDate, IssuedDate, CoverageTemporal, CoverageSpatial and others. Metadata is meant to be coded in XML and fully searchable. They declare their approach “at once more research-oriented and flexible ... copes with the changing needs of users.” They hinge the WAWI work on the Arizona Model (Pearce-Moses, Richard and Joanne Kaczmarek. 2007. *An Arizona Model for Preservation and Access of Web Documents*, http://www.digitalpreservation.gov/multimedia/documents/azmodel.pdf). They note that their previous paper (above) demonstrated a bibliographic approach, which they don't recommend due to its “bibliocentricity” that establishes no relationships among pages.

Zhang, Jane, Michael Paulmeno, Meg Tuomala, Benn Joseph, Polina Ilieva, Jennifer Wright, John Bence, Olga Virakhovskaya, Anna Perricci, Rick Fitzgerald, and Rosalie Lack. “From Crawling to Walking: Improving Access to Web Archives.” Eleven brief papers presented at the Society of American Archivists Annual Meeting, Washington, D.C., 16 August 2014, session 703. https://archives2014.sched.com/event/1hIEcE2/session-703-from-crawling-to-walking-improving-access-to-web-archives.

> Eleven short presentations related to improving access to web archives were presented in a single conference session. Each presenter detailed ways in which their institution has approached access to web archiving using descriptive metadata.
>
> 1) Zhang provides a brief introduction to web archiving, web archiving initiatives and web file formats. Metadata and access via theme-based collections include a collection overview, description and other elements, while provenance-based collections include metadata about the site owner, collection description and others.
>
> 2) Paulmeno discusses providing access to web archives through the library catalog. The presentation details some challenges to providing access to web archives, in particular that archival description in general is not fully compatible with library catalogs and that different metadata and content standards between libraries and archives makes integration difficult. The presentation concludes with some reasons to be hopeful, in particular for institutions that have integrated description between libraries and archives, and that that integration will benefit access to web archives.
>
> 3) Tuomala details the ways in which the University of North Carolina at Chapel Hill has attempted to meet the descriptive and access needs of multiple web archives collections between the libraries and archives. Technical services for both libraries (bibliographic) and archives (archival) were consolidated, which required modification of workflows. This was because the library provides access at the item level in the library catalog, while the archives provide collection-level access through a catalog

record and an EAD. Library web materials are cataloged individually, while archival web collections were integrated into existing finding aids, with an additional catalog record created that links back to the finding aid.

4) Joseph describes a process used at Northwestern to describe web archives (crawled using the California Digital Library's erstwhile Web Archiving System) in Archon. Item records for seeds are created in WAS, and a digital object with a link to the seed is added to the collection description in Archon.

5) Ilieva discusses appraisal and use of web archives. The presentation gives an overview of why web archives are collected, how they are appraised and how access is provided via a catalog record. She notes the need to know how data and collections are used in order to find optimal ways to improve access.

6) Wright describes the ways in which the Smithsonian Archives has attempted to integrate web archiving into existing practices for accessioning and access. Web archives are cataloged in the collection management system, and access is provided through an EAD finding aid and Archive-It.

7) Bence describes how Emory University has integrated web archives built using CDL's WAS system into its online finding aids and web presence. Harvested URLs are added as <dao> elements, and a 2013 website migration allowed for the integration of a CDL WAS search interface within Emory's website.

8) Virakhovskaya outlines the Bentley Historical Library's goal of providing access to archived websites through the library catalog without confusing researchers. The Bentley achieved this goal by using descriptive standards within their description for web archives. The author showed a mapping of web archives description in CDL WAS to the MARC fields used in the Bentley's catalog records.

9) Perricci discusses the access provided by Columbia University to its Contemporary Composers Web Archive by showing descriptive metadata used in Archive-It, the process of creating MARC records for web archives, a patron view of a web archives catalog record; and a cataloger's view of a web archives catalog record.

10) Fitzgerald discusses the rethinking of descriptive practices for web archives at the US Library of Congress as a result of a migration effort began in 2013. The library moved web archives from a standalone web application to the library-wide discovery system, which involved both metadata and content migration, and resulted in an improved interface and enhanced discoverability for the web archives.

11) Lack detailed providing access to web archives by "tearing down silos," creating finding aids for web archives, adding links to web archives to existing finding aids, providing a search page for the web archive collection, creating catalog records for web archives and sending those records to OCLC, and integrating access to various formats in their discovery systems.

NOTES

1. Dooley, Jackie, and Kate Bowers. 2018. *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group.* Dublin, OH: OCLC Research. doi:10.25333/C3005C.
2. Samouelian, Mary, and Jackie Dooley. 2018. *Descriptive Metadata for Web Archiving: Review of Harvesting Tools.* Dublin, OH: OCLC Research. doi:10.25333/C37H0T.
3. Costa, Miguel, and Mário J. Silva. 2011. "Characterizing Search Behavior in Web Archives" Paper presented at the First International Temporal Web Analytics Workshop, Hyderabad, India, 28 March 2011. http://ceur-ws.org/Vol-707/TWAW2011-paper5.pdf.
4. Weber, Matthew, and Pamela Graham. 2016. "Internet and Web Archives Research Use Survey." (Unpublished survey data, 24 February 2016.) [Project Announcement: https://networks.h-net.org/node/73374/announcements/110145/survey-internet-and-web-archives-research-use-study]
5. Bailey, Jefferson. 2015. "Web Archives as Research Datasets." Paper presented at the General Assembly of the International Internet Preservation Consortium, Stanford University, 28 April 2015. http://netpreserve.org/general-assembly/ga2015-schedule.
6. Ras, Marcel, and Sara van Bussel. 2007. "Web Archiving User Survey." National Library of the Netherlands (Koninklijke Bibliotheek), 9-10. https://www.kb.nl/sites/default/files/docs/kb_usersurvey_webarchive_en.pdf.
7. Costa, Miguel, and Mário J. Silva. 2012. "Evaluating Web Archive Search Systems." Web Information Systems Engineering, Lecture Notes in Computer Science 7651: 440-454. http://doi.org/10.1007/978-3-642-35063-4_32.
8. Costa, Miguel, and Daniel Gomes. 2015. "Web Archive Information Retrieval." Paper presented at the International Internet Preservation Consortium, Stanford University, 28 April 2015. http://netpreserve.org/general-assembly/ga2015-schedule.
9. See note 4.
10. Jackson, Andy. 2015. "The Provenance of Web Archives." *British Library UK Web Archive Blog. Posted 20* November. http://britishlibrary.typepad.co.uk/webarchive/2015/11/the-provenance-of-web-archives.html.
11. Murray, Kathleen. R., and Inga. K. Hsieh. 2008. "Archiving Web-Published Materials: A Needs Assessment of Librarians, Researchers, And Content Providers." *Government Information Quarterly* 25(1): 66-89. https://digital.library.unt.edu/ark:/67531/metadc29322/m2/1/high_res_d/Murray-2008-Archiving_Web-Published_Materials.pdf.
12. Galligan, Patrick. "WARCS! What Are They Good For? Researchers!" *Bits & Bytes: News from Rockefeller Archive Center's Digital Team (*blog), Posted 28 April 2016. http://blog.rockarch.org/?p=1502.
13. Jackson, Andy. 2015. "The Provenance of Web Archives." *British Library UK Web Archive Blog. Posted 20* November. http://britishlibrary.typepad.co.uk/webarchive/2015/11/the-provenance-of-web-archives.html.
14. Dooley, Jackie. 2015. Twitter posts from the General Assembly of the International Internet Preservation Consortium, Stanford University, 27-28 April 2015. #iipcGA15; https://twitter.com/NetPreserve.

15. Dougherty, Meghan, Eric T. Meyer, Christine McCarthy Madsen, Charles van den Heuvel, Arthur Thomas, and Sally Wyatt. 2010. “Researcher Engagement with Web Archives.” London: Joint Information Systems Committee Report. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1714997.

16. Hockx-Yu, Helen. 2014. “Access and Scholarly Use of Web Archives.” *Alexandria* 25:1/2. doi:10.7227/ALX.0023.

17. See note 15.

18. Ibid.

19. Thurman, Alex, and Eric O'Hanlon. “Solr-Powered Full-Text and Metadata Search in the Human Rights Web Archive.” Paper presented at the General Assembly of the International Internet Preservation Consortium, Stanford University, 30 April 2015. http://netpreserve.org/general-assembly/ga2015-schedule.

20. Niu, Jinfang. 2012. “Functionalities of Web Archives.” *D-Lib Magazine* 18. http://www.dlib.org/dlib/march12/niu/03niu2.html.

21. Leetaru, Kalev. 2016. “Reimagining Libraries in the Digital Era: Lessons from Data Mining The Internet Archive.” *#bigdata* (blog). Posted 19 March. http://www.forbes.com/sites/kalevleetaru/2016/03/19/reimagining-libraries-in-the-digital-era-lessons-from-data-mining-the-internet-archive/#1eeb690d6c7c.

22. http://mementoweb.org/about/.

23. Van de Sompel, Herbert, Robert Sanderson, Michael L. Nelson, Lyudmila Balakireva, Scott Ainsworth, and Harihar Shankar. 2010. “Time Maps: Metadata for Memento.” Presentation to the GSLIS Metadata Group, University of Illinois, Urbana-Champaign, 14 July 2010. https://www.slideshare.net/azaroth42/timemaps-metadata-for-memento?from_action=save.

24. Galligan, Patrick. “WARCS! What Are They Good For? Researchers!” *Bits & Bytes: News from Rockefeller Archive Center’s Digital Team (*blog), Posted 28 April 2016. http://blog.rockarch.org/?p=1502.

25. Bailey, Jefferson. 2015. “Web Archives as Research Datasets.” Paper presented at the General Assembly of the International Internet Preservation Consortium, Stanford University, 28 April 2015. http://netpreserve.org/general-assembly/ga2015-schedule.

26. The Web ARChive (WARC) file format is a standard format for web archives that combines multiple resources and metadata into an aggregate file. See: Library of Congress. 2009. “WARC, Web ARChive File Format.” Sustainability of Digital Formats: Planning for Library of Congress collections. http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml.

27. Dooley, Jackie. 2015. Unpublished notes summarizing various presentations, taken at the symposium *Web Archives 2015: Capture, Curate, Analyze*, University of Michigan, Ann Arbor, 10-12 November 2015.

28. Cruz, David, and Daniel Gomes. 2013. “Adapting Search User Interfaces to Web Archives.” Paper presented at the 10th International Conference on Preservation of Digital Objects, September 2013. http://sobre.arquivo.pt/sobre/publicacoes-1/Documentos-acerca-do-Arquivo.pt/adapting-search-user-interfaces-to-web-archives.

29. See note 25.

30. Truman, Gail. 2016. “Digital Access to Scholarship at Harvard: Web Archiving Environmental Scan.” *Harvard Library Report*. Cambridge, MA: Harvard University. https://dash.harvard.edu/handle/1/25658314.

31. Thomas, Arthur, Eric T. Meyer, Meghan Dougherty, Charles van den Heuvel, Christine Madsen and Sally Wyatt. 2010. “Researcher Engagement with Web Archives: Challenges and Opportunities for Investment.” *Joint Information Systems Special Report*. London: JISC. https://papers.ssrn.com/sol3/papers2.cfm?abstract_id=1715000.

32. See note 28.

33. See note 30.

34. See note 15.

35. Reynolds, Emily. 2013. *Web Archiving Use Cases*. Washington, D.C.: Library of Congress, UMSI, ASB13. http://netpreserve.org/resources/IIPC_archive-UseCases_Final.pdf.

36. See note 15.

37. Dougherty, Meghan, and Eric T. Meyer. 2014. “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: 2195–2209. http://onlinelibrary.wiley.com/doi/10.1002/asi.23099/abstract.

38. Hockx-Yu, Helen. 2016. *Web Archiving at National Libraries: Findings of Stakeholders’ Consultation by the Internet Archive*. San Francisco, CA: Internet Archive. https://docs.google.com/document/d/1uP6DrwaUe-tbzz_SW7QMSKLTkCNpjg9nFzHTCoOwoOA/edit.

39. See note 37, 2207.

40. Dooley, Jackie, and Kate Bowers. 2018. *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group.* Dublin, OH: OCLC Research. doi:10.25333/C3005C.

41. Bailey, Jefferson, Abigail Grotke, Edward McCain, Christie Moffatt, and Nicholas Taylor. 2017. *Web Archiving in the United States: A 2016 Survey*. Washington, DC: National Digital Stewardship Alliance. http://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf

42. Sweetser, Michelle. 2011. “Metadata Practices Among Archive-It Partner Institutions: The Lay of the Land.” Slide lecture presented at the Archive-It partners meeting, 19 October 2011. http://slideplayer.com/slide/3847250/.

43. Mannheimer, Sara. 2013. “Providing Context to Web Collections: A Survey of Archive-It Users.” Master's thesis, University of North Carolina, Chapel Hill. https://cdr.lib.unc.edu/indexablecontent/uuid:f373e421-0a31-4143-ad65-05137729d894.

44. Phillips, Margaret E., and Paul Koerbin. 2009. “PANDORA, Australia’s Web Archive: How Much Metadata Is Enough?” *Journal of Internet Cataloging* 7(2): 19-33. doi:10.1300/J141v07n02_04.

45. Guenther, Rebecca. 2015. “Metadata for Web Archived Resources: Recommendations for Further Exploration.” Unpublished report prepared for the New York Art Resources Consortium. http://www.nyarc.org/sites/default/files/Recommendations%20for%20further%20exploration-final.pdf.

46. Dublin Core Metadata Element Set, Version 1.1. Dublin Core Metadata Initiative, 2012-06-14. http://www.dublincore.org/documents/dces.

47. Bragg, Molly, and Kristine Hanna. 2013. *The Web Archiving Life Cycle Model*. San Francisco, CA: Archive-It Team, Internet Archive. https://archive-it.org/static/files/archiveit_life_cycle_model.pdf.

48. Gibbons, Leisa. 2016. *Web Archiving Project 2016: Preliminary Report.* http://leisagibbons.info/wp-content/uploads/2017/03/Webarchivingbriefreport-1.pdf.

49. See note 42.

50. See note 43.

51. Ibid., 16.

52. Bailey, Jefferson, Abigail Grotke, Kristine Hanna, Cathy Hartman, Edward McCain, Christie Moffatt, and Nicholas Taylor. 2014. *Web Archiving in the United States: A 2013 Survey,* National Digital Stewardship Alliance. http://www.digitalpreservation.gov/documents/NDSA_USWebArchivingSurvey_2013.pdf.

53. See note 41.

54. OCLC Research Library Partnership Metadata Managers Focus Group. 2016. (Notes contributed by group members on their institutions' web archiving metadata practices and needs. Not publicly available.) http://www.oclc.org/research/themes/data-science/metadata-managers.html.

55. New York Art Resources Consortium. 2015. *Metadata Application Profile and Data Dictionary for Description of Websites with Archived Versions, Version 1, June*. http://www.nyarc.org/sites/default/files/web-archiving-profile.pdf.

56. See note 45.

57. Haddad, Peter, and Pam Gatenby. 2002. "Providing Bibliographic Access to Archived Online Resources: The National Library of Australia's Approach." Paper presented at the Bibliography and National Libraries Workshop "Bibliographic Control or Chaos," at the 68th IFLA General Conference and Council, Glasgow, Scotland, 23 August 2002. https://archive.ifla.org/IV/ifla68/papers/069-152e.pdf.

58. Prom, Christopher, and Ellen Swain. 2007. "From the College Democrats to the Falling Illini: Identifying, Appraising, and Capturing Student Organization Websites." *The American Archivist* 70 (2): 344-363. http://dx.doi.org/10.17723/aarc.70.2.c8121767x9075210.

59. Guenther, Rebecca, and Leslie Myrick. 2007. "Archiving Web Sites for Preservation and Access: MODS, METS and MINERVA." *Journal of Archival Information* 4: 144-166.

60. O'Dell, Allison Jai. 2015. "Describing Web Collections (I Mean Archived Websites)" *Medium* (stories). Posted 17 February. https://medium.com/@allisonjaiodell/describing-web-collections-e32b59893848#.lbg88xo51

61. Zhang, Jane, Michael Paulmeno, Meg Tuomala, Benn Joseph, Polina Ilieva, Jennifer Wright, John Bence, Olga Virakhovskaya, Anna Perricci, Rick Fitzgerald, and Rosalie Lack. "From Crawling to Walking: Improving Access to Web Archives." Eleven brief papers presented at the Society of American Archivists Annual Meeting, Washington, D.C., 16 August 2014, session 703. https://archives2014.sched.com/event/1hIEcE2/session-703-from-crawling-to-walking-improving-access-to-web-archives.

62. Thurman, Alex. 2016. "Web Archiving: Description and Access." Paper presented in the Metropolitan New York Library Council webinar series, 29 February 2016.

63. Pregill, Lily. 2016. "Web Archiving: Description and Access." Paper presented in the Metropolitan New York Library Council webinar series, 29 February. http://www.slideshare.net/ElizabethLilyPregill/web-archiving-description-and-access.

64. Peterson, Christie. 2015. “Archival Description for Web Archives,” *Chaos→Order*. Posted 12 June. https://icantiemyownshoes.wordpress.com/2015/06/12/archival-description-for-web-archives/.

65. Bernstein, Steven. 2016. “MARC Reborn: Migrating MARC Fixed Field Metadata into the Variable Fields,” *Cataloging and Classification Quarterly* 54: 23-38. doi:/10.1080/01639374.2015.1075642.

66. Wu, Paul H.J., Ichsan P. Tamsir, and Adrian K.H. Heok. 2006. “Applying Context-Sensitive Web Annotation in Evidence-Based Collaborative Web Archives Cataloging.” Paper presented at the International Web Archiving Workshop, Alicante, Spain, 21-22 September 2006, 75-98. http://dl.acm.org/citation.cfm?id=2173354.

67. Wu, Paul H.J., Ichsan P. Tamsir, and Adrian K.H. Heok. 2007. “Annotating Web Archives Structure, Provenance, and Context Through Archival Cataloguing.” *New Review of Hypermedia and Multimedia* 13:55-75. doi:10.1080/13614560701423620.

68. Bailey, Jefferson and Maria LaCalle. 2015. “Don’t WARC Away: Preservation Metadata for Web Archives.” Paper presented at the ALA Annual Conference, Association of Library Collections and Technical Services, Preservation and Reformatting Section, San Francisco, California, 27 June 2015. http://connect.ala.org/files/2015-06-27_ALCTS_PARS_PMIG_web_archives.pdf.

69. See note 59.

70. Lavoie, Brian, and Richard Gartner. 2013. *DPT Technology Watch Report*. Preservation Metadata, 2nd edition. Great Britain: Digital Preservation Coalition in association with Charles Beagrie Ltd. doi:10.7207/twr13-03.

71. See note 45.

For more information about our work related to digitizing library collections, please visit: **oc.lc/digitizing**

OCLC

6565 Kilgour Place
Dublin, Ohio 43017-3395

T: 1-800-848-5878
T: +1-614-764-6000
F: +1-614-764-6096
www.oclc.org/research

ISBN: 978-1-55653-009-8
DOI:10.25333/C33P7Z
RM-PR-215938-WWBE 1709