Sara Mannheimer. Providing context to Web collections: A survey of Archive-It users. A Master's paper for the M.S. in I.S. degree. April, 2013. 54 pages. Advisor: Denise Anthony

This study describes a survey to users of the Internet Archive’s Archive-It Web-archiving tool, aiming to examine the descriptive metadata practice of archivists of the Web, how Web archives are accessed, and what variables facilitate or impede metadata implementation in Web collections.

Whereas books often contain contextual information bound between their covers, archival materials require additional explanation of context. The Web is the most transient of electronic records, and although it is currently being preserved at a higher rate than ever before, treatment of Web collections is still not up to archival standards. Through better understanding of current Web archiving metadata practices, this study hopes to help lay groundwork for future best practices.

Headings:

Web archiving

Metadata

Archives -- Administration

PROVIDING CONTEXT TO WEB COLLECTIONS:
A SURVEY OF ARCHIVE-IT USERS

by
Sara Mannheimer

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

Chapel Hill, North Carolina

April, 2013

Approved by:

Denise Anthony

Table of Contents

List of Tables

List of Figures

Introduction

In the Ben Dixon MacNeill papers (#3617, Southern Historical Collection, The Wilson Library, University of North Carolina at Chapel Hill), there is an ill-lit, badly-framed photograph of an old car on a dirt road. On its surface, this photograph seems to have little value. However, consider the following: the photographer was the first photojournalist in North Carolina, and the car is driving on the earliest constructed section of the Blue Ridge Parkway. Unlike books, which contain contextual information bound between their covers, archival materials such as this photograph frequently require additional explanation of context in order to reveal richer meaning and fuller understanding.

Some of the earliest archival thinking established the principles of provenance, original order, and *respect des fonds* as methods of providing necessary context to archival materials. (This paper assumes reader familiarity with these principles. For an overview of the concepts, see Millar, p. 94-114.) In order to portray contextual information, archivists developed the finding aid, a tool that has become the standard for archival description (Hurley, 1998).

Duff, Craig, & Cherry (2004, Spring) investigate how historians use archival resources in the research process by surveying history department faculty members at universities in Canada. The study’s findings, namely that that that “historians rate finding aids, footnotes, and archivists very highly as sources for becoming aware of and locating information in their research" (p. 7), show that contextual data provided by finding aids is necessary for both access and for full understanding of archival materials.

Table 1: DACS elements

Elements	Sub-elements or further explanation
2. Identity	2.1 Reference code 2.2 Name and location of repository 2.3 Title 2.4 Date 2.5 Extent 2.6 Creator 2.7 Administrative/biographical history
3. Content and structure	3.1 Scope and content 3.2 System of arrangement
4. Conditions of access and use	4.1 Conditions governing access 4.2 Physical access 4.3 Technical access 4.4 Conditions governing reproduction and use 4.5 Languages and scripts of the material 4.6 Finding aids (other)
5. Acquisition and appraisal	5.1 Custodial history 5.2 Immediate source of acquisition 5.3 Appraisal, destruction, and scheduling information 5.4 Accruals
6. Related materials	6.1 Existence and location of originals 6.2 Existence and location of copies 6.3 Related archival materials 6.4 Publication note
7. Notes	- Any additional information that cannot be communicated through any of the defined elements of description
8. Description control	- Sources used - Descriptive rules or conventions used - Name(s) of the person(s) who prepared or revised the record - Date(s) the record was created or revised

(From Society of American Archivists, 2004)

Processing archivists use *Describing archives: A content standard* (2004), commonly known as DACS, to standardize the elements and values used for description and provision of access in archival collections. (See Table 1). DACS is implementable by a wide range of archives, and was officially adopted as a standard by the Council of the Society of American Archivists in March 2005 (Society of American Archivists, p. ii). Assigning metadata like DACS elements to archival collections helps users identify, retrieve, and understand the meanings of archival records (McKemmish, Acland, Ward, & Reed, 2006).

Since its development in the mid-nineties (Pitti 1997), the XML language Encoded Archival Description (EAD) has been widely used to encode DACS elements and other descriptive information into finding aids that are posted to the Web, a practice that also helps standardize description and facilitate increased searchability.

As more records are being created electronically, however, archivists have had to reassess ordinary archival tasks; electronic records require new frameworks for collection development, new techniques for arranging and describing records, as well as new plans for long-term preservation. In addition, there is a phenomenon that Lyman and Varian refer to as the "democratization of data" (2000), the majority of electronic records are now being created and stored by individuals, rather than institutions. This adds another layer of complexity to the archivist's task because individuals create, edit, name, and file electronic records idiosyncratically. This added unpredictability heightens the importance of the archivist's job as context-provider.

Archiving the Web

Since the invention of the World Wide Web by Tim Berners-Lee in 1996 (Berners-Lee, 1996), increasing numbers of websites have been continually created, altered, and removed from the Web. Although this transience complicates preservation activities, it also makes the need for quality preservation all the more pressing. Without proper archiving practice for capture, description, and long-term maintenance of websites, vital cultural information is at risk. As Jeff Rothenberg wrote in 1999, “the current generation of digital records… has unique historical significance; yet our digital documents are far more fragile than paper. In fact, the record of the entire present period of history is in jeopardy” (p. 1).

The situation is somewhat less dire a decade later because of increased awareness of the problems. In addition, there are several major commercial and open-source Web archiving tools on the market to help institutions create Web collections. These include Archive-It, a service of the Internet Archive (Archive-It 2011-2012), (the service whose partner institutions were surveyed in this paper); The Web Archiving Service (WAS), from the California Digital Library (Regents of The University of California 2007-2013); Hanzo Enterprise (Teffin 2012); HTTrack Web Site Copier (Roche 2012); Teleport Webspiders from Tennyson Maxwell Information Systems (Tennyson Maxwell Information Systems 2012); and others. Each of these tools operates by using Web crawlers, also called spiders, to harvest the content of websites at scheduled times. Harvested URLs are called “seeds.” Archivists can program crawls to follow page links; the more links are harvested, the “deeper” the archive (Masanes, 2005). The archivist can also decide to pursue “internal” depth by harvesting Web pages within the main page’s domain or “external” depth through harvesting links to sites outside of the main page domain (Schneider et al, 2003). The Web archivist’s choice of different depths and extents of a Web archive can alter a future user’s perception of the site. However, after deciding what level of depth to capture, archivists must create a space for the harvested websites in the larger context of archival collections. Collections of the Web are some of the newest archival formats, for which standards and best practices are still being developed. Descriptive metadata is one way of providing context to these collections.

Two recent publications outline best practices for archives of the Web. A publication from University of Texas at Austin (2011) focuses on metadata practice, and provides best practice information using MODS, including mandatory, recommended, recommended-if-applicable, and optional metadata fields for use in archiving the Web. The publication also maps MODS elements to Dublin Core, for Web archiving services that use Dublin Core.

A publication from the Internet Archive in March 2013 provides a life-cycle model for archiving the Web. This document acknowledges that, “as with most aspects of web archiving, best practices are evolving regarding the use and creation of metadata and descriptive trends for web archives” (p. 20), and it examines Web archiving practice through two “circles,” the policy circle and the metadata/description circle. In addition, some research has been conducted to investigate optimum presentation of archived electronic records to the user. Many of these have found that the traditional finding aid may not be practical for this purpose (Duff, 1995, McKemmish et al, 2006, Wallace, 1995). However, few studies have researched how archivists might provide such information.

A study by Dellavalle, Hester, Heilig, Drake, Kuntzman, Graber, & Schilling (2003, October 31) investigates links to online references in scientific studies to measure continuing availability on the Web. The authors “examined the frequency, format, and activity of Internet references in three high-circulation U.S. journals with scientific impact” (p. 787). The study found that the percentage of inactive Web references increased from around 4% at three months to 10% at 15 months and then to 13% at 27 months after publication (p. 787). The study suggests that the Web has become a vital information source to scientists; therefore the disappearance of references from the Web is a problem that cannot be solved by banning Web references from scientific literature. Instead, the authors propose that the Library of Congress embark on Web preservation efforts, and that "Internet information cited in peer-reviewed, high-impact journals will receive priority in [these] efforts" (p. 788).

The research in this article views websites as academic evidence that must be preserved. It highlights the importance of archiving the Web in order to enable scientists and other professionals to cite Web resources in academic work without the concern that those resources may be unavailable in the future. Before one can argue that archival description and other contextual information should be provided for Web archives, it must first be established that the Web is indeed a cultural asset that must be preserved. Dellavalle et al. argue this case.

Purpose statement

This paper will investigate how archives of the Web are presented to users. By surveying users of the Archive-It service (interchangeably referred to in this paper as "users" and "partners"), it will determine what metadata is being assigned to Web collections, how Web collections are displayed and cataloged, and what factors facilitate or impeded metadata usage. This paper aims to add to the body of literature by investigating how archivists of the Web contextualize Web collections, and why Web content tends to be treated differently from other digital content. Understanding what contextual information is currently provided to Web collections will allow future archivists of the Web to determine how best to describe, catalog, and provide access to archived Web materials.

Research questions

1. What metadata do Archive-It partners assign to their collections? 2. How are archivists providing access to archives of the Web? 3. What variables facilitate or impede metadata implementation?

Literature Review

Search strategies

An initial literature search was conducted by casting a wide net across large databases, including ACM Digital Library, Articles+, Google Scholar, and Library Literature & Information Science. Initial searches consisted of general terms such as “web archiving,” “web archive metadata,” and “web archive context.” Several relevant articles were found during the initial search. These articles were then scanned for citations using the *Web of Science*, an activity that produced a sizeable amount of additional literature. It was noted that many relevant articles came from a few specific journals; for this reason, individual searches were also conducted of *American Archivist, Archivaria,* and *D-Lib.*

Descriptive Metadata for the Web

Lavoie and Gartner (2005) propose that the minimum metadata maintained by archives should include:

“1. *Provenance*, describing the custodial history of the object

2. *Authenticity*, validating that the object is what it purports to, and has not been modified

3. *Preservation activity*, describing actions taken to preserve the object
4. *Technical environment*, describing the IT environment necessary to render the object faithfully
5. *Rights management*, recording any property rights which may govern retention or publication of the object" (p. 5).

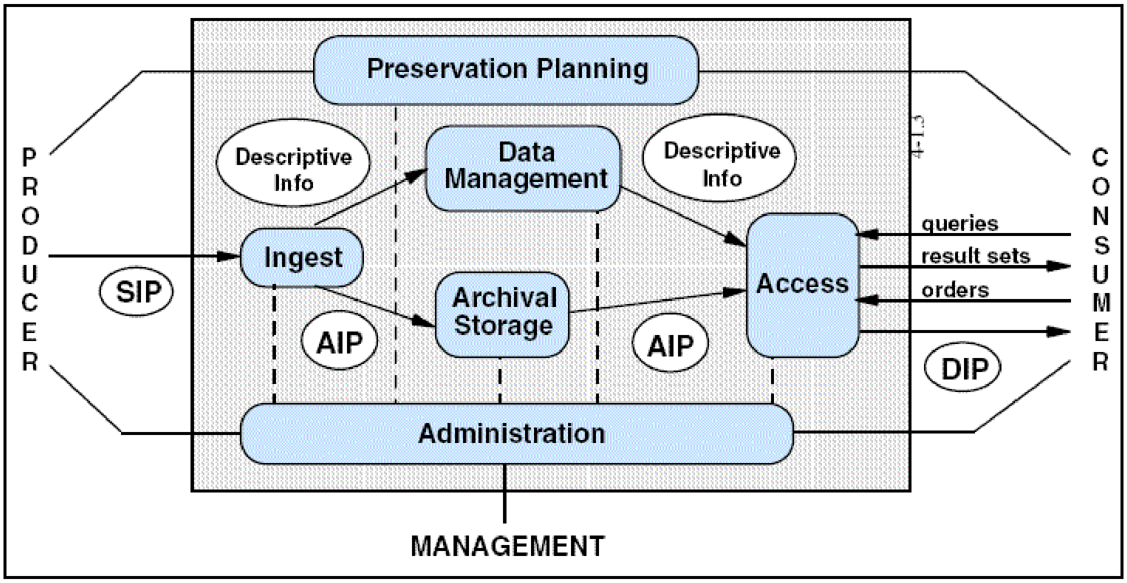
Of the many metadata schemas being applied to the Web, one of the earliest schemas to be developed is still one of the most widely-used. Dublin Core, developed in 1995 and maintained by the Dublin Core Metadata Initiative (DCMI), provides fifteen core metadata fields that aim to describe nearly any type of resource. These fields are contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, and type. According to the DCMI website, "early Dublin Core workshops popularized the idea of "core metadata" for simple and generic resource descriptions. [The schema] achieved wide dissemination as part of the Open Archives Initiative Protocol for Metadata Harvesting," and Dublin Core has since become a national and international standard (DCMI 1995-2013).

The Collaborative Digitization Program asserts that, "while... Dublin Core is relatively simple to learn and easy to use, its elements include the most essential information about a resource" (2006). The Archive-It Web archiving service provides the fifteen basic Dublin Core fields to its users, along with the option for custom fields that can provide more specific description.

The Library of Congress developed the Metadata Object Description Schema (MODS) in 2002. Although the schema can be used for a variety of purposes, it was designed specifically for use in library applications, and especially books, multimedia, and electronic library resources (MODS, 2013). MODS can be used as an extension schema to Metadata Encoding and Transmission Standard (METS), discussed below. MODS contains twenty main elements: titleInfo, note, name, subject, typeOfResource, classification, genre, relatedItem, originInfo, identifier, language, location, physicalDescription, accessCondition, abstract, part, tableOfContents, extension, targetAudience, and recordInfo. These elements, although similar to Dublin Core, are tailored more specifically to bibliographic resources and library settings. On its website, the Library of Congress describes the MODS element set as "richer than Dublin Core," but "simpler than the full MARC format" (2013).

Several different standards exist for creating and transmitting archival metadata. The Open Archive Information System (OAIS) reference model, published in 2002 by the Consultative Committee for Space Data Systems, has become an ISO standard (ISO 14721) for electronic records management in all disciplines (Lee, 2010). The model proposes a detailed conceptualization that addresses preservation planning, administration, ingest, data management, archival storage, and access for digital information. The OAIS model provides instruction for content creators as well as archival repositories. It includes terminology, concepts, architectures, and data models, and is designed to be applicable to any institution's electronic data management practices. Under the OAIS model, there are three "information packages" necessary to the digital archiving process. First, the data producer submits a data package to the repository, tailored to the archivist's specifications; this submission information package (SIP) contains data and contextual information. After the accepting repository ensures that the SIP is secure and has no harmful elements, the SIP is transferred into the archive. Second, the archival information package (AIP), contains content and metadata that are stored and managed by the repository for long-term preservation. Lastly, the model provides a structure for access, in the form of the dissemination information package (DIP). The DIP provides the information to a patron or consumer (for visualization, see Figure 2). Thomas et al. (2010), while dubbing OAIS "the most ambitious effort to date," argue that "like many such reference models, and because of its complexity, it has been adopted only in parts by archival institutions" (p. 14).

Figure 1: OAIS reference model



(From Lee, 2011)

However, because OAIS is a model for information management, rather than a specific metadata schema, full adoption of the complex process may not be necessary. According to Allinson (2006), "This conceptual nature [of the OAIS model] is seen by many as a strength and, by being light on prescriptive statements, OAIS allows those implementing the model to apply their own layers of adaptation" (p. 11). Allison provides a detailed analysis and evaluation of OAIS in the context of educational institutions, including discussion of two projects from the Joint Information Systems Committee that have adapted OAIS to fit specific needs. Allison concludes that it is beneficial for repositories to have a common framework for managing electronic records, and that the OAIS model is flexible enough to be applied to a wide range of institutions and projects.

Thomas et al. (2010) suggest that the Metadata Encoding and Transmission Standard (METS), developed by the Digital Library Federation and maintained in the Network Development and MARC Standards Office of the Library of Congress, is simpler than OAIS, and therefore has a greater likelihood of being widely adopted. METS is a structural metadata standard that provides descriptive, technical, and administrative metadata to express hierarchical relationships using XML (Digital Library Federation, 2010). However, the METS format was developed specifically for digital libraries, and therefore has some limitations when applied to Web content. (Guenther & Myrick, 2007).

There is also potential for using Linked Data approaches to metadata. In the Linked Data approach, objects and relationships are not defined by hierarchies, but rather by the Resource Description Framework (RDF), a directed graph that uses "triples," in which each digital entity is assigned either object, predicate, or subject (Gibbens 2010). Each of the objects and relationships in an RDF graph is represented by a Universal Resource Identifier (URI). There is currently an effort within the Semantic Web community to develop tools for managing large RDF graphs, and writing reasoning languages such as the Web Ontology Language (OWL) to create more meaningful graphs (McGuinness, & van Harmelen, 2004).

Wu, Heok, & Tamsir (2006) state that "the growth of the Internet continues to out-pace the speed of attempts to describe it. The emergence of the semantic web (or

Web 2.0) then becomes an appealing solution, as it mobilizes the collective effort of the public to help 'catalog the web'" (p. 20). The authors suggest that by creating a context-aware annotation system, users of websites will be able to contribute descriptive content to the Web. The authors propose that this system would "provide evidence and preserve context to the cataloged records of the materials within a web archives" (p. 20). However, annotation by users has not taken off in the way that Semantic Web proponents have hoped. The casual user may not annotate, and quality content is difficult to ensure.

Time on the Web

The Memento project (Van de Sompel et al, 2009) has identified the aspect of time as important in contextualizing the Web. The Web changes from moment to moment more than any other electronic record; over the course of years, a URL may represent completely different websites, and even if a URL contains the same site for many years, the content of that site by nature changes over time. Take as an example the *New York Times* website, in which the same URL, www.nytimes.com, displays different stories from moment to moment and from day to day. Therefore, while a website's file tree could be considered to contain some intrinsic original order (Pearce-Moses & Kaczmarek, 2005, p. 22), the same file tree might contain completely different documents if crawled at different moments in time. Simple display of a website and the crawl date are not enough to fully contextualize the site.

Web archive user studies

Although still a growing field of research, a few user studies of Web archives have been conducted in recent years. In their 2010 study, Costa and Silva ask the question, "what are the user intents and which topics are most interesting to them?" (p.

9). The authors investigated the behavior of users of the Portuguese Web Archive (PWA), an archive containing nearly 150 million web documents from 1996 to 2009, all accessible by full-text and URL search. Costa and Silva used three strategies to answer their research question: search logs, an online questionnaire to be answered by users during a search, and a laboratory study. Search logs were collected without matching IP addresses to individuals, so an exact number of participants was not calculated for this phase; 21 users responded to the online questionnaire, and the laboratory study analyzed these questionnaire results. The authors conclude that users search chiefly for known pages. Also, users generally did not restrict searches by date, but the authors speculate that this may have been due to search engine design. When users did restrict by date, they chose to view older web pages rather than newer ones. Lastly, users in the study preferred full-text over URL search.

Another study was conducted at the National Library of the Netherlands (Ras & van Bussel, 2007). This task-oriented usability study evaluated user familiarity and proficiency with search and access tools, and investigated user satisfaction with Web archive content. However, the author was unable to locate this study in English.

Past projects using Archive-It

There has been one previous survey of Archive-It users regarding metadata practices, available on the Archive-It website. Michelle Sweetser's 2011 study "Metadata practices among Archive-It partner institutions: The lay of the land" surveyed Archive-It partners to determine general demographics, information about the size and scope of Web archiving practice, metadata practice in Web collections, and promoting awareness of the institutions' archived Web content. Sweetser's survey was conducted recently enough that it will provide a helpful foundation of comparison for the current study.

Sweetser found that most Archive-It partners did not assign much metadata in 2011; the most-used fields were "Title" and "Description." Sweetser theorizes that three factors influenced the low level of metadata assignment reported in her survey: "1. Organizations just haven't yet gotten around to preparing metadata in Archive-It and are still in their infancy in terms of their web archiving efforts. 2. Organizations do not believe that metadata is warranted or useful to be created. 3. Organizations are focusing their metadata creation practices in areas outside the Archive-It platform" (Sweetser 2011).

Overall, the literature supports the importance of contextual information to 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.

Research design and methods

The goal of this project was to determine how metadata is being used in archives of the Web, and what factors facilitate or impede metadata implementation. Although, as previously discussed, there exist several Web-archiving services, the two major non-proprietary services currently being used in archival institutions are Archive-It, from the Internet Archive, and Web Archiving Service (WAS), from the California Digital Library. An evaluation by the Minnesota Historical Society (2009) concludes that the two services have comparable features. Therefore, the decision to survey users of Archive-It stems mainly from the convenience of access to the Archive-It community listserv.

University of North Carolina at Chapel Hill (UNC) subscribes to this service, as does Duke University and the State Archives of North Carolina. Using the Archive-It user forum listserv to disseminate the questionnaire meant that users received the survey from a trusted source, thus encouraging response rates.

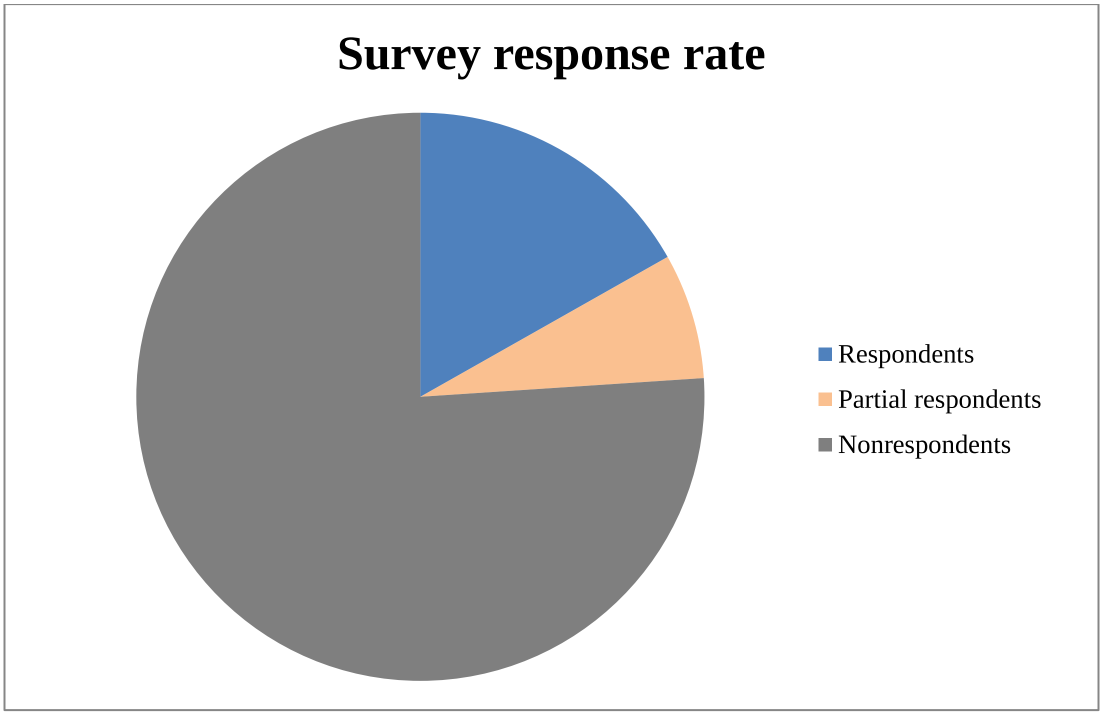
The research consisted of two stages. First, data reports provided by Archive-It showed the percentage of collections that include each specific Dublin Core field. In order to gain access to this information, the author contacted Archive-It's administrators through UNC's Electronic Records Archivist, Meg Tuomala. Lori Donovan, Partner Specialist at the Internet Archive, not only provided these reports and links to in-house studies, but also provided the author with information about Archive-It capabilities, as well as support and feedback on the survey instrument. The survey was further informed by the author's personal experience with archiving the Web at Wilson Library at UNC.

Second, a survey, created using Qualtrics Survey Software, was sent to the Archive-It listserv. The data collected from Archive-It provided a baseline for the surveys by giving concrete numbers of Archive-It partners using metadata. (For complete survey instrument, see Appendix A.)

The survey method was chosen due to its ability to provide data about a large number of institutions archiving the Web, and in order to provide a general sense of attitudes among Web archivists about how to provide context to their collections. A survey allowed the author to reach a wide audience of archivists of the Web; the fact that the survey could be administered online meant that respondents could answer at their leisure, and in the comfort of their own home or office. Because the author chose not to provide incentives beyond the satisfaction of advancement of research, she limited the survey to seventeen questions and a comments box, in order to maximize responses. Response time was estimated at five to ten minutes. In practice, response time averaged nine minutes.

As of January 2013, the Archive-It listserv had 407 subscribers (Donovan, personal correspondence 2013), from 238 different institutions (Bragg & Hanna 2013). The survey was meant to be responded to only once by each institution. If the response rate had been consistent with the field's average of 63% (Wildemuth 2009), the author would have received about 150 responses to the survey. In reality, 57 Archive-It partners responded to the survey, 40 of whom answered all questions. The total response rate of approximately 24% was a large enough number of responses to generate quality data, but a small enough number to be manageable for the time-frame and nature of this project.

Figure 2: Survey response rate



Instrument

Survey administration

The author followed a three-phase administration process adapted from Salant and Dillman (1994). The first e-mail was a short notice providing an overview of the planned research and linking to the survey instrument. (For recruitment email, see Appendix B.) One week later, a second email was sent to follow up. A week after the second email, a third and final email was sent, with a short additional note focusing on the benefits of the research. This three-phase process aimed to increase response rate without overwhelming user inboxes.

The author sent the first request for survey participation to the Archive-it listserv on February 18, 2013; this initial request generated 19 responses. A survey respondent asked the author if she could post the survey to the Society of American Archivists Web Archiving Roundtable, which generated an additional three responses. Those respondents who were directed from the Roundtable were told to specify if their institution used a Web archiving service other than Archive-It; since no respondents specified other services, the author assumes all respondents use Archive-It. The author sent a second email one week later, on February 25, 2013, which generated an additional 8 respondents, bringing the total to 30. A few responses trickled in between the second email and the final email, sent on March 4, 2013. A week later, on March 11, the survey closed, with a total of 57 responses.

The survey instrument consisted of 15 questions, two of which were elaboration questions, only administered dependent on a response to the previous question. There was also a final request for additional comments before the survey was submitted.

Answering each question on the survey was optional; only 40 respondents answered every question. Results are presented regardless of whether every respondent answered each question, with total number of respondents indicated.

Discussion of survey questions

The first four questions asked for demographic information from each respondent: Age, education level, job title, and years of experience with metadata or cataloging.

Question five asked how long the institution had been archiving the Web. Questions six, seven, and nine addressed who is in charge of metadata selection and assignment at the institution. Question eight asks about the specific metadata fields being used at the institutions, and questions 10 and 11 asked about controlled vocabularies being used to supplement these fields. Questions 12 and 13 aimed to find out the metadata and cataloging procedures at the institution for other materials, both non-Web digital materials and non-digital materials. Question 14 and 15 relate to access: how are Web collections accessed by users? And if they are accessed via the library catalog, is the metadata transformed to MARC or another cataloging standard?

The final two questions attempt to determine what factors facilitate or impede metadata implementation. The factors listed on the questionnaire were determined by the author, and informed by her personal experience with assigning metadata to Web collections. At the end of the survey, the author provided a comment box in which respondents could type any additional questions, qualifiers, or comments.

Short-comings and limitations of research

Archiving the Web is still a new enough practice that there are few standards of best practice. Each of the several Web archiving programs on the market has different features. By limiting survey participants to Archive-It users, this paper does not take into account that different Web archiving services may provide different opportunities to add contextual information.

The survey format also has unavoidable limitations. As with all surveys, response was voluntary. Those who agree to complete the questionnaire may have common attributes that may not constitute a truly random sample. By surveying a large population of potential respondents, the author hoped to improve the odds of obtaining a random sample. However, the total of 57 responses to the survey – only 40 of whom answered every question – is still a small sample size. The smaller the sample size, the less likely it is that the sample will accurately represent the population as a whole.

The last shortcoming is one of human error. As a first-time survey author and administrator, the author made an error in judgment regarding the survey's demographics questions: While the demographic information questions are directed toward the individual respondent, the main questions in the survey body are aimed at the institution as a whole. This mistake became obvious when respondents began emailing, asking whether the survey should be filled out once by each institution or by each individual receiving the survey. In retrospect, demographics information on each institution (for example, type of institution, number of employees tasked with Web archiving, or amount of training provided) would have been more useful to the final analysis.

Archive-It allows partners to create metadata on at the collection level, the seed level (the starting point URL), or the document level (specific Web pages). In the survey questions, the author did not differentiate between levels of metadata assignment, a fact that may affect the specificity of the results.

Ethical issues

As with all human-subject studies, ethics must be considered. The questionnaire in this study asked simple, professionally-oriented questions. Respondents were informed that all results would be published anonymously, protecting the views of participants. Answering the questionnaire was voluntary, and respondents were informed that they could skip any question or abort the survey at any time for any reason. Questions were framed in a way that the author hopes is nonbiased and neutral, eliminating the possibility of respondents feeling inadequate or uncomfortable in any way.

Findings

After survey administration was finished, the author used the Qualtrics survey tool to assist in data analysis. Qualtrics generates reports that match demographic information with questionnaire answers in order to create more meaningful data. These reports were exported to Microsoft Word and Microsoft Excel for easier analysis. In the following discussion, the metadata practices of Archive-It users and the respondents' perceptions of facilitators and barriers of metadata implementation are measured against the independent variables of work experience, education, age, and institutional experience with the Archive-It program.

Demographics

The ages of respondents were relatively evenly split until age 65, with only one respondent over 65 years old. Of those reporting, 18 respondents (39%) were under 35 years old, 14 respondents (30%) were between 36 and 50 years old, and 13 respondents (28%) were between 51 and 65 years old. There were a total of 46 responses to the question of age.

The majority of respondents (30 respondents, 70%) had a highest education level of Masters degree in Information and Library Science (two of these respondents had dual Masters degrees). Two respondents (4%) had Bachelors degrees in Information and Library Science. Of the remainder of the 46 respondents, degrees varied (see Table 2).

Table 2: Non-LIS education of respondents

Bachelor's (specify major)	Masters (specify field)	PhD (specify field)	Other (please specify
Economics	MLS and MA (Philosophy)	English & American Literature	Canadian College Diploma in Multimedia Design
Archaeology and Anthropology	History and MLIS	History with a specialty in archives	
Architecture	Environmental Studies		
Political Science and Economics	Trade Law		
History	History		

Question 3, "Job Title," had unlimited response values. For this reason, there were a wide variety of job titles entered by respondents. Some of the more common titles were University Archivist, Electronic Records Archivist, Digital Archivist, Digital Projects Librarian, and IT Manager. Most job titles indicated a wide variety of tasks performed, while very few job titles were specific to cataloging or metadata. Most respondents reported being in charge of metadata assignment at their institutions. (For full list of job titles, sorted by which titles are responsible for metadata assignment, see Appendix C). Those who reported that someone else was in charge of assigning metadata reported that the following job titles assigned metadata at their institution: Professional Cataloger (3 respondents); Processing Archivist (2 respondents); "Metadata Specialist (paraprofessional position on our digital initiatives team);" "Basic metadata by selectors (subject librarians); we are going to have a student assign more detailed metadata;" Metadata Specialist; Digital Collections Coordinator; "300 Web content authors / posters;" and "shared responsibility - cataloger, collection curator, digital collection specialists."

Survey respondents were generally experienced with metadata and cataloging, with the largest group of respondents (40%) reporting 6-10 years of cataloging experience. Only three respondents (6%) reported no cataloging experience. (For full results, see Table 3).

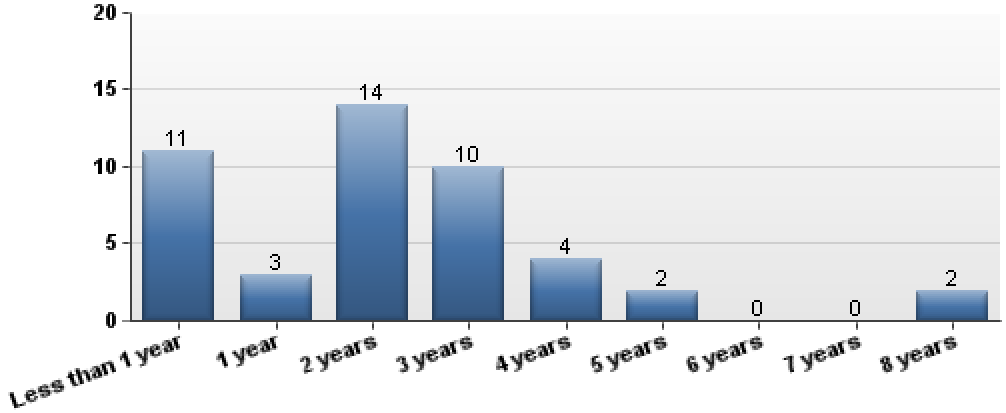
Table 3: Individual respondents' metadata or cataloging experience

Years cataloging		#	%
1-5		11	23%
5-10		19	40%
10-20		6	13%
20-30		6	13%
Over 30		2	4%
None		3	6%
Total		47	100%

Web archiving and metadata practice

In general, institutions were either extremely new to the Archive-It service, with about one quarter of institutions having used the service for less than one year, or institutions had been archiving the Web using Archive-It for 2-3 years (about half of institutions had been Archive-It partners for this time period). Two institutions had been using Archive-It since its initial deployment in 2006. The survey did not ask whether each institution had been archiving the Web with a different service before partnering with Archive-It, so the data may not reflect the institutions' total years archiving the Web, only the years using the Archive-It service.

Figure 3: Number of years institutions have used Archive-It

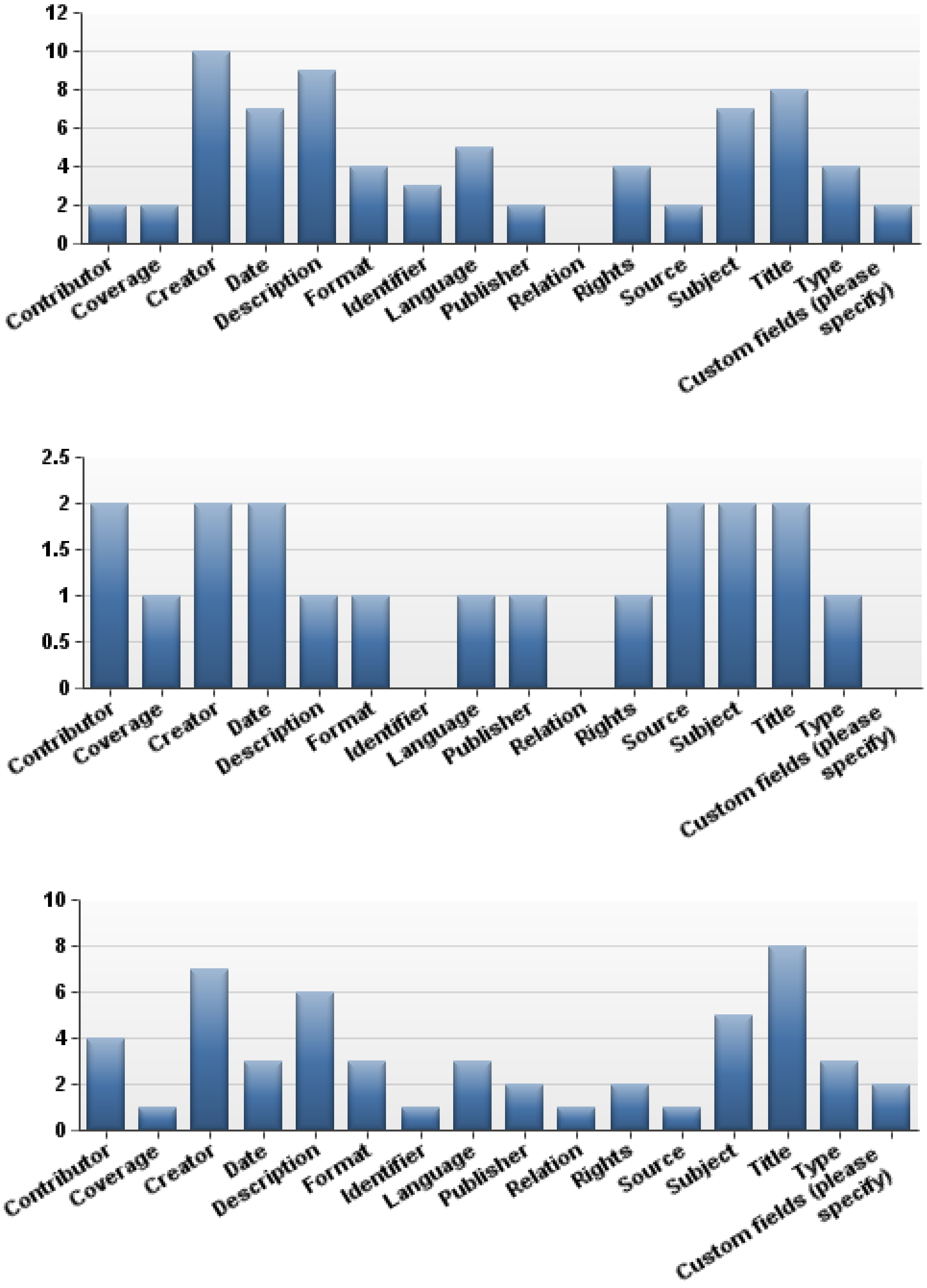


When comparing this information against what metadata fields are used by each institution, generally, the longer the institution had been archiving the Web, the more metadata fields were used. The two institutions that had been archiving the Web for eight years both used at least seven fields (Creator, Description, Format, Language, Publisher, Subject, Title). One of the institutions also used five additional fields (Coverage, Identifier, Rights, Type, and the custom field "Collector"), for a total of twelve fields.

Only one respondent had been using Archive-It for five years, and that respondent used only the Date and Description fields; the lack of rich metadata in the five-year category may be a result of the small respondent set.

The institutions that had been archiving the Web for three or four years tended to use substantially more descriptive metadata elements, with most using at least five elements. Those who have archived the Web for less than two years tended to use somewhat fewer elements, although "Creator," "Description," and "Title" were used by nearly all respondents, regardless of experience with archiving the Web.

Figure 4: Institutional use of Dublin Core elements by number of years using Archive-It

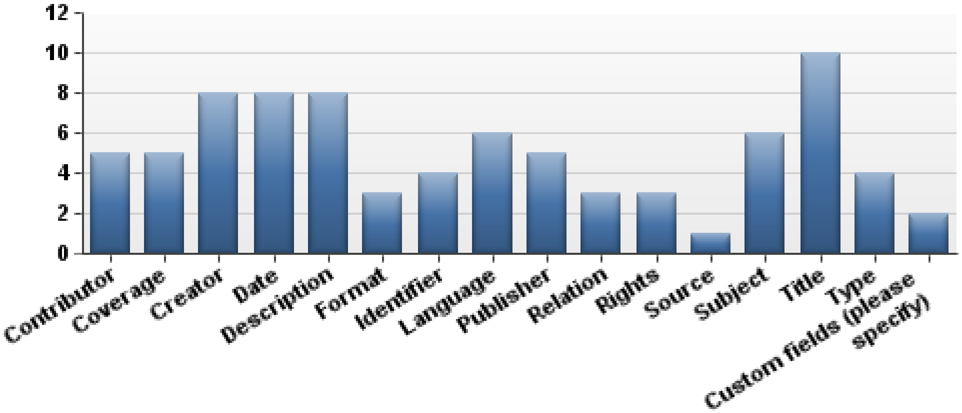


Less than one year:

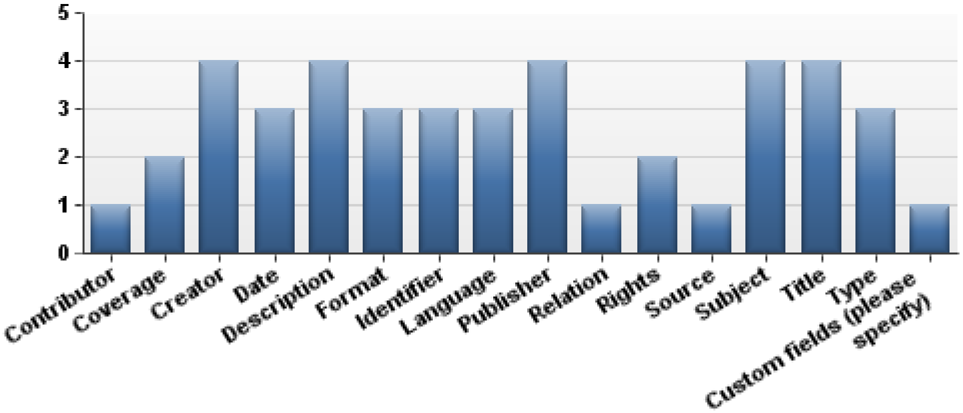
One year:

Two years:

Three years:



Four years:



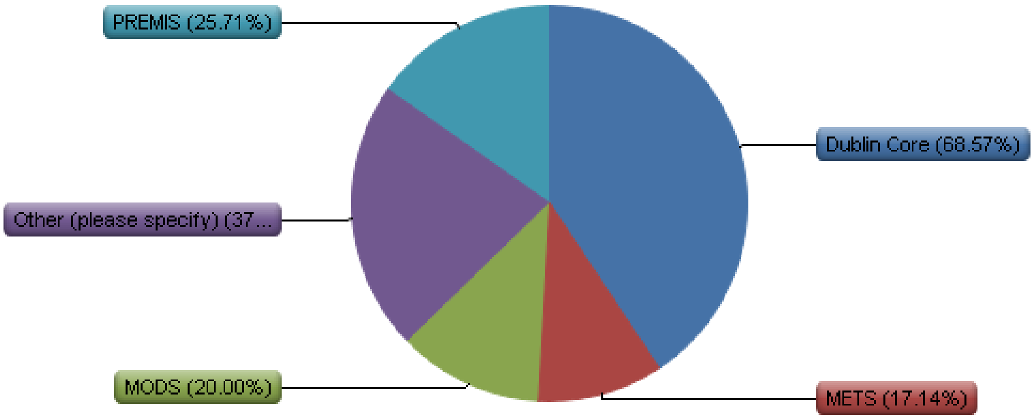
Archive-It provided the following basic statistics on metadata usage: Out of 238 partners, 216 (90%) have metadata at any level. 199 (84%) use more than one Metadata field, and 18 (8%) partners use custom fields at any level. Although these statistics are very general, they verify the information reported by respondents to this paper's questionnaire (Donovan, personal correspondence, 2013).

There was substantial variety in the reporting of Question 9, "Who determined what metadata elements are used in the Web collections at your institution?" In the majority of institutions, either senior staff or metadata specialists determined element usage. Many respondents reported collaborative efforts in the decision making process. (For a complete list of answers to Question 9, see Appendix D).

About two-thirds of respondents reported using controlled vocabularies when assigning metadata. 64% of these institutions used Library of Congress Subject Headings, and half used Library of Congress Authority Records. There were a few that used ISO language or date standards. For those who reported using other controlled vocabularies or thesauri, responses were varied, and included: FAST; "local list from the materials themselves;" DCMI Type; AAT; TGN; Su Doc Classification; MARC Code List for Languages; State Archives/OSPI; MIME Media Types; North Carolina Thesaurus; Thesaurus of Graphic Materials; and ThinkMap Visual Thesaurus.

Although a majority of respondents reported using Dublin Core to describe non-Web born-digital materials, the other responses were quite diverse. 20% reported using MODS, 17.4% reported using METS, and 25.71% reported using PREMIS. Of those respondents who reported "Other" (37%), several different schemas were specified, including: "filenaming structure only;" MARCXML; EAD; MARC; and DSpace. Five of the 35 respondents to this question either reported that their institution did not collect or assigned no metadata to born-digital materials; due to the structure of the survey, these five responses were logged as "Other." Figure 5 shows the distribution of responses. Note that respondents were able to choose multiple responses, so the total percentage sums to more than 100%.

Figure 5: Metadata used to catalog non-Web born-digital materials



Thirty-six respondents answered Question 13: "What metadata standards do you use when archiving non-digital materials?" Thirty-three respondents reported the use of either MARC, AACR2, EAD, DACS, or RDA. Of the remaining two, one respondent used DCRM (Descriptive Cataloging of Rare Materials), and two respondents reported no current metadata use for non-digital materials, with plans to someday adopt RDA or MARC. This result suggests that institutions value the generation of cataloging data. The uniformity of responses indicates that there are only a few accepted and commonly used standard metadata schemas for non-digital materials.

Providing access to Web collections

Although the majority (84%) of respondents reported providing access to their archives of the Web through the Archive-It website, the author expected this number to be near 100%, since all live, non-restricted Web collections are available on the Archive-It website. The reason for the low reported rate of access through Archive-It's website is unclear – either respondents were confused by the question, Web collections were restricted, or Web collections hadn't yet gone live.

Most respondents also provided access through another portal – either through the library’s catalog, via an online finding aid, or through a search box on the institution’s website. Of the respondents who chose “Other,” a few specified “Google;” one institution had not gone live with their Web collections; one respondent wrote, “we are developing a separate web application for our Archive-It crawl results;” another response stated “We currently have a web page with links to the archived collections, but are also looking into other ways, such as providing collection level metadata in the library catalog, and harvesting metadata to our discovery layer (we don't know if it's possible).” As with many questions in the survey, respondents could choose multiple answers, so the total percentage exceeds 100%.

Table 4: Access to Web collections provided by instituions

Presentation venue		#	%
Through the institution’s catalog		11	30%
Through an online finding aid		12	32%
Via the Archive-It website		31	84%
Via a search box on the institution’s website		11	30%
Other (please explain)		8	22%

One respondent explained, “we have a single catalog record in our online catalog for our web archiving, but also include a link to our Archive-It holdings for each agency on its agency history page. For example, there is a link on the agency history page of the Secretary of State to archived holdings of its web page. In addition, if an agency has substantial state publications on its web page, we place a link on that agency's state publications catalog record also.” This reflects the practice of about a dozen institutions who use more than one access method for their Web collections.

Although only eleven respondents reported providing access through the institution’s catalog, 24 respondents answered Question 15, “If you include Web collections in your catalog, do you transform Web collection metadata into MARC or another bibliographic cataloging standard?” The reason for the imbalance of answers is not clear. Question 14 (“How do you provide access to Web collections at your institution?”) may have been confusing to some respondents, and the complexity of Question 14 meant Qualtrics skip logic could not be used to skip Question 15 (a supplement to Question 14) if the respondent did not report providing Web collection access through the catalog. In any case, of the respondents to Question 15, one third (eight respondents) reported transformation into MARC, while two thirds (16 respondents) reported that they did not use MARC or another cataloging standard. An explanatory text box was not provided for a “no” answer, so it is unclear how respondents who reported no transformation include Web collections in the catalog.

Facilitators and barriers to metadata usage

To better understand the reasons why archivists of the Web use metadata, Question 16 and 17 asked respondents to rate some facilitators and barriers to metadata usage, from “do not affect me” to “affect me greatly. There are many factors that contribute to metadata creation, both positively and negatively; understanding how each of these factors affect archivists is important to the goal of contextualizing Web sites for future use. The facilitators reported as affecting respondents the most were “metadata education;” “recognition of the importance of metadata;” “controlled vocabularies;” and "automatic generation." The factors that least affected respondents were "institutional help" and "metadata registries." A few respondents chose "other," but none specified facilitators; the author therefore did not include these responses in the report.

Figure 6: Facilitators to metadata usage

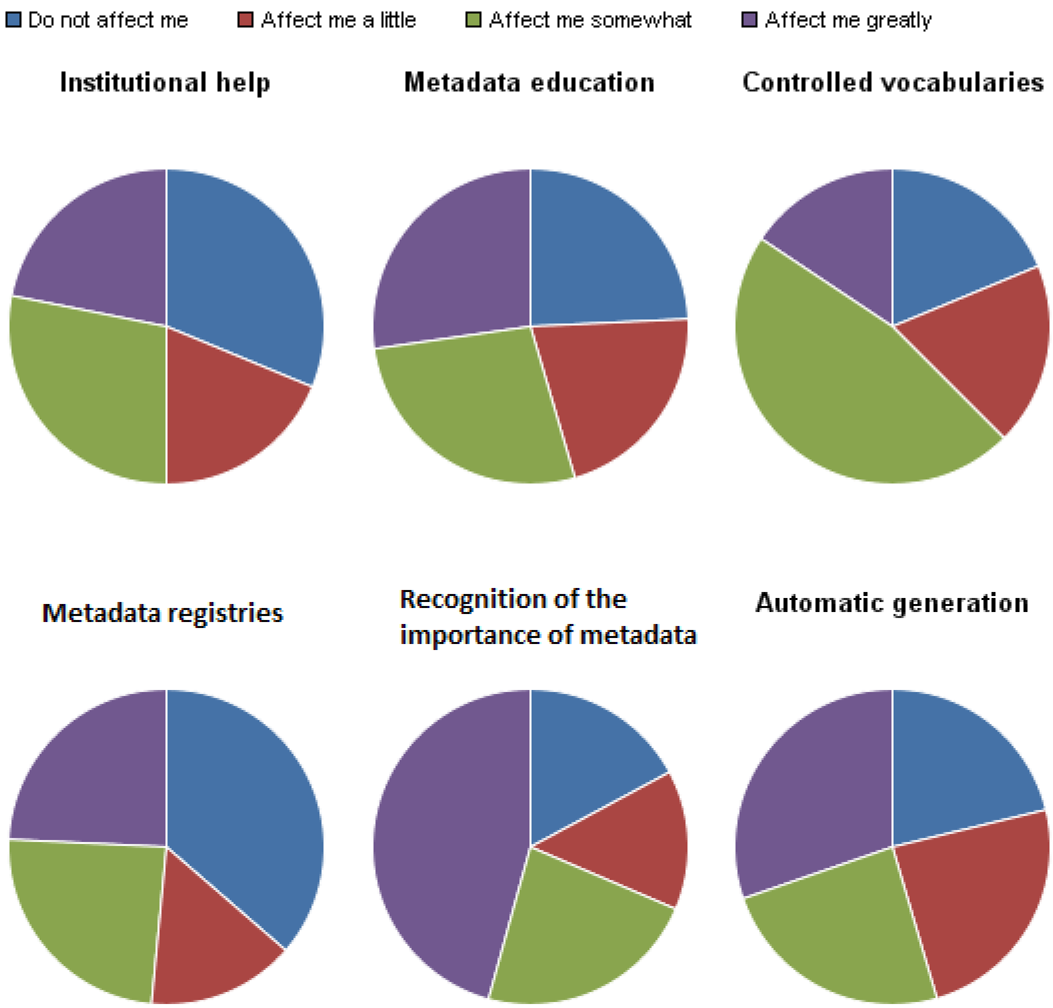
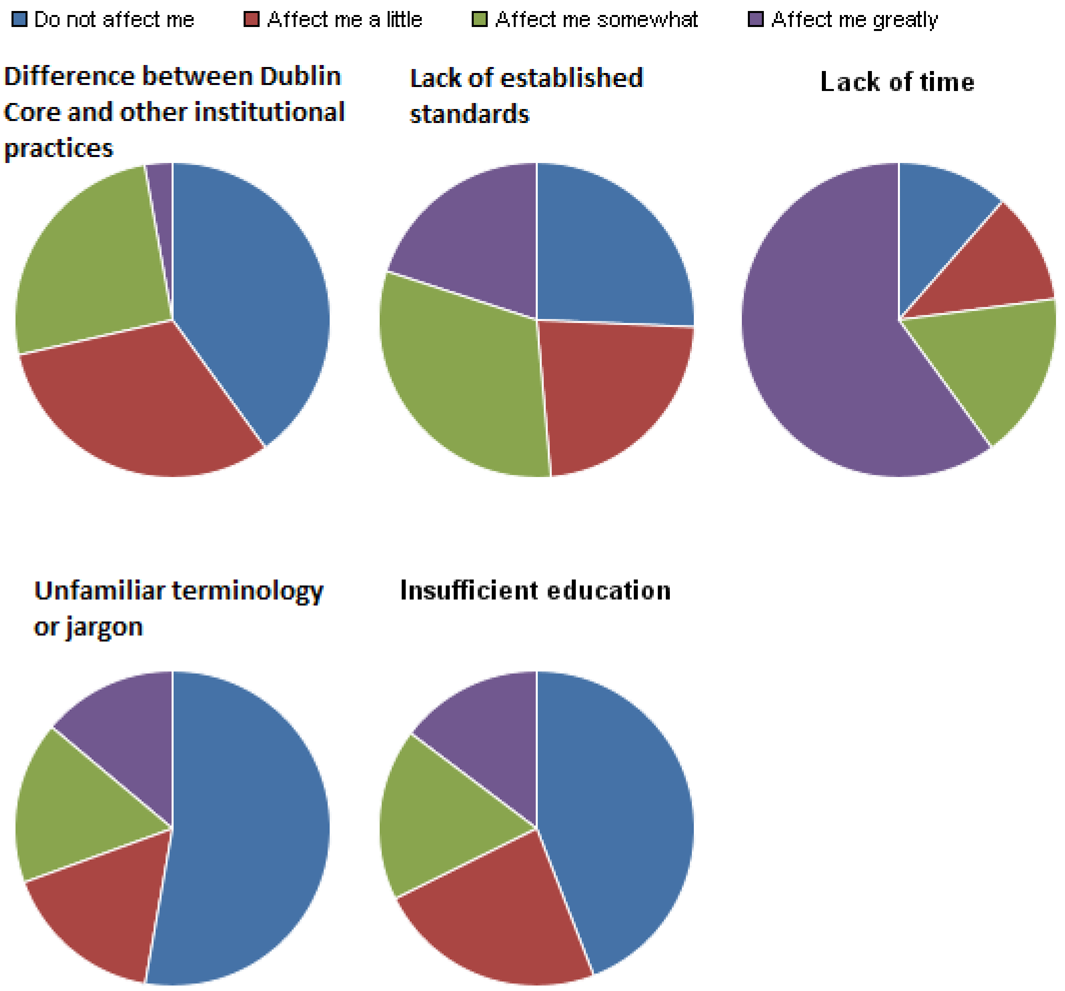


Figure 7: Barriers to metadata usage



Perhaps unsurprisingly, the principal barrier to metadata usage was reported to be "lack of time." This factor far exceeded all other factors in preventing metadata creation. A distant second and third were "lack of established standards" and "insufficient education." One respondent chose "other," specifying "No batch upload for metadata and seed URLs" as well as the difficulty of making batch edits. The respondent further expanded, "I'm thinking about not including any metadata, just because it's too difficult for me to keep it up to date. [When creating] a full D[ublin] C[ore] record for a seed URL in Archive-It, it takes about 15 minutes to capture the seed URL, do the basic descriptive work, set host constraints, etc. The inability to batch process these tasks is really prohibitive." This respondent's desire for batch uploading and editing is closely related to the barrier "lack of time." Looking back to the job titles of archivists creating metadata for the archived Web (Appendix C), most job titles constitute far more duties than simply assigning metadata to archives of the Web. For this reason, one can assume that archivists of the Web are attempting to fit metadata assignment into an already busy schedule.

It is worth mentioning that Archive-It is aware of the "lack of time" factor, and Partner Specialist Lori Donovan informed the author that Archive-It Version 4.8, slated for release in May 2013, will include features that "further automate the metadata addition process, including batch uploading seed level metadata and bulk editing document metadata" (personal correspondence 2013).

The last question, Question 18, provided respondents with a chance to make any additional comments. A few respondents expressed encountering difficulty when Web archiving activity bridged several different institutional departments; one respondent wrote: "we do assign a title to every public-facing seed and we use "Groups" to gather seeds supporting the same collection (which are easy to use in the Archive-It portal), but our/my efforts end there since I am not directly responsible for processing collections."

A few respondents indicated that their Web archiving program was still in beta, including the following comments: "I am still in the planning/testing stages, so some of my answers are predicting what I think I'll be doing by next year rather than what I'm doing this month," and, "Right now we're just starting with one field. We'll put more in the future if it seems worth the time and effort."

Additional comments focused on the as yet undefined best practices in Web archiving. For example, “while we strive to include useful metadata, establishing a standard vocabulary has been difficult, given the amount and variety in types of information included in our various web collections,” and “Often have more acceptance of file naming standards then of meta data tagging - but in the Web world generic file names (where data can be replaced) are more heavily adopted then with other electronic content.”

Discussion

Web archiving is a very new practice, and the Archive-It service has only been available for eight years. Considering the youth of Web archiving programs, the rate of metadata usage reported by respondents is remarkably high. The majority of Archive-It partners use metadata of some kind, and most use five or more Dublin Core fields, especially those who have been using Archive-It for three or more years. Nearly all respondents use at least “Creator,” “Description,” and “Title.”

Although the levels of reported metadata usage in this study are good, there is still room for better context provision. No librarian would consider putting a book on the shelf without creating a full catalog record, yet this practice is not standard when archiving the Web. In fact, with the sheer number of electronic records being created, it may be impossible to provide the same quality of metadata to electronic records as we do to paper ones. As indicated in Question 17, by far the greatest barrier to metadata usage is “lack of time.” Archive-It can crawl a large number of websites in relatively little time, and archivists must manage their work-days to perform many different tasks. According to this survey, high-ranking archivists are often in charge of assigning metadata to Web collections. Perhaps a better system would be to write rules for metadata assignment, then allocate the actual task to lower-level staff or professional catalogers.

The fact that many institutions use different metadata schemas for Web collections than for other born-digital records is also problematic. Without streamlined metadata schemas being used for all record formats within an institution, finding archival materials and relating them to each other is difficult.

Another area that could use improvement is that of providing online access. Only about one-third of respondents reported using a form of access other than the Archive-It website, with responses equally distributed between the library catalog, an online finding aid, or a search-box on the institution’s website. On the Archive-It website, harvested Web collections exist as a single entity, separate from related archival materials. Archival finding aids were designed to show context and relationships between materials. Without finding aids or comparable access methods, website collections are deprived of the contextual information usually provided by archival groupings.

While working with the Web archiving program at the UNC, the author encountered this problem of access first-hand. Although initially, UNC archivists had hoped to create a finding aid for all seed URLs that weren’t directly associated with a collection, even with a relatively small seed collection (UNC collects about 60 URLs in University Archives, and about 15 URLs for the Southern Historical Collection), this plan proved to be too time-consuming to implement. At a recent meeting regarding Web archive access, Technical Services archivists, curators, and the Electronic Records Archivist made an initial determination that, although Southern Historical Collection websites would be added to the finding aids, all the harvested University Archives websites will be stored in a single finding aid. A note linking to the Web archive finding aid will be included in University Archives finding aids that may have related Web content. This may be a necessary step, due to the large numbers of harvested websites and the fact that not all seed URLs match existing finding aids. However, this separation of Web content from other content could be detrimental to context. Archival materials gain contextual richness through their relationships with and proximity to related materials. For this reason, archival materials tend to lose meaning when removed from a larger collection. While Dublin Core metadata facilitates easier searching and provides important contextual information, archivists are still not doing enough to connect harvested websites with related archival collections.

Expected benefits

It is the author's hope that, by examining contextual information in Web archives, this study will draw attention to the necessity of quality descriptive metadata when preserving the Web. The study also hopes add to the body of literature investigating best practices for archiving the Web, both for archival professionals, Web archive users, and future generations. These metadata practices will affect preservation, understanding, and access, and it is important to begin to develop a standard of practice. By looking at facilitators and barriers to metadata implementation, this study hopes to bolster those facilitators and break down those barriers. This study hopes to help metadata creation for Web content become more widespread, and ultimately to help archivists provide the contextual information that is a cornerstone of archival theory and practice.

Recommendations for future research

Since the Web itself is a fairly new phenomenon, the existing research on Web archiving is limited. There is a need for more study surrounding the preservation of Web content. While this project aims to shed light on the role of contextual information in Web archives for preservation, understanding, and access, it does not investigate the long-term facility of rich descriptive metadata or the long-term consequences of poor descriptive metadata. Future research could examine how the quality of metadata affects preservation and access over the long term. These projects should focus both on Web archive users, Web archive creators, and the content itself in order to fully explore the issue.

Furthermore, this paper only briefly addresses the problem of access, and how best to present harvested websites to users. Questions abound in the consideration of this topic. Is it practical to make a finding aid for each website? Should a MARC record be created in the catalog for each website? If finding aids and/or MARC records are used, is it even necessary to assign Dublin Core metadata (other than "Title" and "Description") using the Archive-It service? What other access options exist? A study of how institutions provide access, as well as a user study about how users interpret this access, would provide vital information to institutions archiving the Web.

Summary

The digital age has sparked a renaissance of information technologies that enrich our communication, our knowledge, and our understanding of the world around us. Creating archives of the Web and other digital content is vital for preservation of these valuable cultural resources. However, the sheer amount of records being created makes the archivist’s job more difficult than ever before. In order to facilitate preservation, understanding, and access – both now and in the future – quality descriptive metadata must be assigned to digital content, and especially the Web. By determining the descriptive functionalities of Archive-It, what additional contextual information should be captured, and what variables facilitate or impede metadata implementation, this study hopes to understand how to improve the system of metadata implementation for future Web archives.

References

Archive-It. (2011-2012). Archive-It: A service of the Internet Archive. Retrieved from www.archive-it.org.

Berners-Lee, T. (1996, October). WWW: past, present, and future. *Computer* 29(10), 69-77.

Collaborative Digitization Group Metadata Working Group (2006). Dublin Core Metadata Best Practices Version 2.1.1. http://www.mndigital.org/digitizing/standards/metadata.pdf

Costa, M., & Silva, M. J. (2010). Understanding the information needs of Web archive users. In Masanès, Rauber, & Spaniol (Eds.), *10th International Web Archiving Workshop*.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative and mixed methods approaches (3rd edition)*. Thousand Oaks, CA: Sage Publications, Inc.

Dellavalle, R., Hester, E., Heilig, L., Drake, A., Kuntzman, J, Graber, M, & Schilling, L. (2003, October 31). Going, going, gone: Lost internet references. *Science Magazine*, *302*(5646), 787-788. Retrieved from http://www.sciencemag.org/content/302/5646/787.summary.

Digital Library Federation. (2010). Metadata encoding and transmission standard. Retrieved from http://www.loc.gov/standards/mets/METSPrimerRevised.pdf.

Donovan, L (Partner Specialist at the Internet Archive). Email message to author, February 6, 2013.

Dublin Core Metadata Initiative (1995-2013). Retrieved from http://dublincore.org/

Duff, W. (1995). Will Metadata Replace Archival Description? A Commentary. *Archivaria, 1*(39), 33-38.

Duff, W., Craig, B., & Cherry, J. (2004, Spring). Historians' Use of Archival Sources: Promises and Pitfalls of the Digital Age. *The Public Historian, 26*, 7-22. Retrieved from http://www.jstor.org/stable/10.1525/tph.2004.26.2.7.

Gibbins, N. and Shadbolt, N. (2010). Resource Description Framework (RDF). *Encyclopedia of Library and Information Sciences, 3rd. ed, 1*, 4539-4547.

Guenther, R., & Myrick, L. (2007). Archiving Web sites for preservation and access: MODS, METS and MINERVA. *Journal of Archival Organization, 4*(1-2), 141-166.

Hurley, C. (1998). The Making and the Keeping of Records: What are Finding Aids For?. *Archives and Manuscripts, 26*, 58-77.

Lavoie, B., & Gartner, R. (2005). *Preservation Metadata (Report 05-01)*. York, UK: Digital Preservation Coalition.

Lee, C. A. (2010). Open Archival Information System (OAIS) Reference Model. *Encyclopedia of Library and Information Sciences*,, 4020-4030.

Library of Congress. (2013). Metadata Object Description Schema (MODS) Official Website. Retrieved from http://www.loc.gov/standards/mods/

Lyman, P., Varian, H.R. (2000, December). Reprint: How much information? *Journal of Electronic Publishing.* Retrieved from http://dx.doi.org/10.3998/3336451.0006.204.

Masanès, J. (2005). Web archiving methods and approaches: a comparative study.

Library Trends, *54*, 72-90.

McGuinness, D.L., & van Harmelen, F. (Eds.) (2004, February 10). *OWL Web Ontology Language Overview*. World Wide Web Consortium. Retrieved from http://www.w3.org/TR/owl-features/

McKemmish, S., Acland, G., Ward, N., & Reed, B. (2006). Describing Records in Context in the Continuum: The Australian Recordkeeping Metadata Schema. *Archivaria*, *1*(48), 3-37.

Millar, L. (2010). Provenance, original order and respect des fonds. In *Archives: Principles and practices* (97-114). New York: Neal-Schuman.

Minnesota Historical Society. (2009). *Preserving State Government Digital Information: Web Archive Evaluations*. Retrieved from http://www.mnhs.org/preserve/records/legislativerecords/WebArchiving.htm

Monks-Leeson, E. (2011, Spring-Summer). Archives on the Internet: Representing Contexts and Provenance from Repository to Website. *The American Archivist*, *74*, 38-57. Retrieved from http://archivists.metapress.com/content/h386n333653kr83u/.

Pearce-Moses, R., & Kaczmarek, J. (2005). An Arizona Model for preservation and access of Web documents. DTTP: *Documents to the People*, *33*, 17-24.

Ras, M., & Van Bussel, S. (2007). *Web archiving user survey*. Technical report, National Library of the Netherlands (Koninklijke Bibliotheek).

Regents of The University of California. (2007-2013). The Web Archiving Service. Retrieved from http://webarchives.cdlib.org/was.

Roche, X. & other contributors. (2012). HTTrack Website Copier. Retrieved from http://www.httrack.com/.

Rothenberg, J. (1999). Ensuring the Longevity of Digital Information. Washington, DC: Council on Library and Information Resources.

Salant, P. & Dillman, D.A. (1994). *How to conduct your own survey.* New York: John Wiley.

Schneider, S.M., Foot, K., Kimpton, M., & Jones, G. (2003, August 21). Building thematic Web collections: Challenges and experiences from the September 11 Web archive and the Election 2002 Web archive. In Masanès, Rauber, and Cobena (Eds.), *Proceedings: 3rd Workshop on Web Archives*, 77-93.

Society of American Archivists. (2004). *Describing archives: A content standard*. Chicago: Society of American Archivists.

Sweetser, M. (2011). Metadata practices among Archive-It partners: The lay of the land. *Archive-It Meeting Presentations 2011.*

Taffin, N. (September 2012). Hanzo Archives. Retrieved from http://www.hanzoarchives.com/.

Tennyson Maxwell Information Systems, Inc. (2012). The teleport webspiders. Retrieved from http://www.tenmax.com/teleport/home.htm.

Thomas, A., Meyer, E., Dougherty, M., Van den Heuvel, C., Madsen, C., & Wyatt, S. (2010). Researcher engagement with web archives: State of the art. *Joint Information Systems Committee Report*.

Van de Sompel, H., Nelson, M. L., Sanderson, R., Balakireva, L. L., Ainsworth, S., & Shankar, H. (2009). Memento: Time travel for the web. *arXiv preprint arXiv:0911.1112*.

Wallace, D. (1995). Managing the Present: Metadata as Archival Description. *Archivaria, 1*(39), 11-21.

Wildemuth, B. M. (2009). *Application of social research methods to questions in information and library science*. Westport, CT: Libraries Unlimited Press.

Wu, P., Heok, A., & Tamsir, I. (2006). Annotating the Web Archives–An Exploration of Web Archives Cataloging and Semantic Web. In *Digital Libraries: Achievements, Challenges and Opportunities 4312*, 12-21.

Appendix A

Survey Instrument:

Q1. What is your age?

❑ Under 35
❑ 36-50
❑ 50-65
❑ Over 65

Q2. Please check the box that best describes your highest level of education:

❍ High School Diploma
❍ Bachelor in Library of Information Science
❍ Bachelors (specify major) ____________________
❍ Masters in Library or Information Science
❍ Masters (specify field) ____________________
❍ PhD in Library or Information Science
❍ PhD (specify field) ____________________
❍ Other (please specify) ____________________

Q3. Job Title: ____________________

Q4. Years of experience with metadata or cataloging:

❍ 1-5
❍ 5-10
❍ 10-20
❍ 20-30
❍ Over 30
❍ None

Q5. How many years has your institution been archiving the Web using Archive-It?

❍ Less than 1 year
❍ 1 year
❍ 2 years
❍ 3 years
❍ 4 years
❍ 5 years
❍ 6 years
❍ 7 years
❍ 8 years
❍ 9 years
❍ 10 or more years

Q6. Are you in charge of assigning metadata to Web collections at your institution?

○ Yes
○ No

If Yes Is Selected, Then Skip To Please check all descriptive metadata...

Q7. Who is in charge of assigning metadata to Web collections at your institution (please check all that apply)?

□ Professional cataloger
□ Processing archivist
□ Electronic Records Manager
□ Other (please specify) ____________________

Q8. Please check all descriptive metadata elements used by your institution for websites archived with Archive-It (to the best of your knowledge):

□ Contributor
□ Coverage
□ Creator
□ Date
□ Description
□ Format
□ Identifier
□ Language
□ Publisher
□ Relation
□ Rights
□ Source
□ Subject
□ Title
□ Type
□ Custom fields (please specify) ____________________

Q9. Who determined what metadata elements are used in the Web collections at your institution?

Q10. Do you use any controlled vocabularies for assigning metadata?

○ Yes
○ No

If No Is Selected, Then Skip To What metadata standards do you use wh...

Q11. What controlled vocabularies do you use (please check all that apply)?

❑ ISO Language standard (please specify) ____________________
❑ ISO date standard (please specify) ____________________
❑ ISO country codes
❑ Other ISO standards (please specify) ____________________
❑ Library of Congress Subject Headings
❑ Library of Congress Authorities
❑ SEARS Subject Headings
❑ Thesaurus for Graphic Materials 1: Subject Terms
❑ Thesaurus for Graphic Materials 2: Genre and Physical Characteristics
❑ Thinkmap Visual Thesaurus
❑ UNESCO Thesaurus
❑ WordNet (Princeton University)
❑ DCMI controlled vocabularies (please specify) ____________________
❑ Other (please specify) ____________________

Q12. What metadata standards do you use when archiving other (non-Web) born-digital materials?

❑ MODS
❑ METS
❑ PREMIS
❑ Dublin Core
❑ Other (please specify) ____________________

Q13. What metadata standards do you use when archiving non-digital materials?

❑ MARC
❑ AACR2
❑ DACS
❑ Other (please specify) ____________________

Q14. How do users at your institution find the Web collections you maintain? (all that apply)

❑ Through the institution's catalog
❑ Through an online finding aid
❑ Via the Archive-It website
❑ Via a search box on the institution's website
❑ Other (please explain) ____________________

Q15. If you include Web collections in your catalog, do you transform Web collection metadata into MARC or another bibliographic cataloging standard?

❍ Yes (please specify cataloging standard) ____________________
❍ No

Q16. Please indicate the level to which each of the following facilitators of metadata implementation affect you:

	Do not affect me	Affect me a little	Affect me somewhat	Affect me greatly
Differences between Dublin Core and other institutional cataloging/metadata practices	□	□	□	□
Lack of established standards	□	□	□	□
Lack of time	□	□	□	□
Unfamiliar terminology or jargon	□	□	□	□
Insufficient education	□	□	□	□
Other (please specify)	□	□	□	□

Q17. Please indicate the level to which each of the following barriers to metadata implementation affect you:

	Do not affect me	Affect me a little	Affect me somewhat	Affect me greatly
Institutional help	□	□	□	□
Metadata education	□	□	□	□
Controlled vocabularies	□	□	□	□
Metadata registries (e.g. the Dublin Core registry)	□	□	□	□
Recognition of importance of metadata	□	□	□	□
Automatic generation	□	□	□	□
Other (please specify)	□	□	□	□

Q18. Thank you for taking the time to complete this survey. Please leave any additional comments below. Be sure to click the ">>" button when you are finished (below right) to submit the survey.

Appendix B

Recruitment email to the Archive-It liststerv:

Subject: Request for participation: Descriptive metadata survey

Dear __________

My name is Sara Mannheimer; I am a graduate student at the University of North Carolina at Chapel Hill School of Information and Library Science writing a master's thesis about descriptive metadata in Web collections, under the direction of Professor Denise Anthony. I am asking for your participation in a survey as part of my research.

The survey seeks to examine how Archive-It partners use descriptive metadata when archiving the Web. Answering the survey will take 5-10 minutes.

This email has been distributed to all members of the Archive-It listserv, and I have received permission from listserv administrators on the Archive-It team to contact you for this survey. Your answers will be anonymous. The study has been reviewed and approved by the University of North Carolina Institutional Review Board (http://research.unc.edu/offices/human-research-ethics/index.htm).

Please note that your participation is voluntary. You may skip any question for any reason. There are no anticipated risks to answering the survey, and there may be no direct benefit. Additionally, there is no cost or incentive associated with participation.

Please see attached research abstract for more details (abstract consistent with page 1).

Clicking on the survey link below indicates that you understand the research study, have had the opportunity for any questions to be answered, and agree to participate.

Click here to link to the survey.

Thank you in advance for your time and participation. When the paper is finished in May, it will be available online at http://dc.lib.unc.edu/cdm/landingpage/collection/s_papers.

Please do not hesitate to contact me (mannheim@live.unc.edu) or Professor Anthony (anthonyd@live.unc.edu) with any questions or concerns.

Sincerely,
Sara Mannheimer

Appendix C

Job titles in charge of metadata assignment:
Archives and Digital Collections Librarian
Archivist (2)
Archivist & Records Manager
Archivist Coordinator
Assistant Archivist/University Archivist
Collection's Archivist
College Archivist
Curator of Collections
Curator of Digital Collections
Digital Archivist
Digital Collections Archivist
Digital Projects Librarian
Director of the Library
Electronic Records Archivist (2)
Government Publications Librarian
Head of Archives and Special Collections
IT Manager (2)
Librarian
Library Specialist
Manuscripts Curator
Metadata & Cataloguing Librarian
Project Associate, Newspaper Digitization
Records and Archives Manager
University Archivist
University Records Archivist
Web Manager
Web Resources Collection Coordinator
Wisconsin Historical Society (Several people)

Job titles not in charge of metadata assignment:
Archivist/Librarian
Associate Director of Libraries
Catalog/Metadata Librarian
Department Head
Electronic Records Archivist
Government Information Reference Librarian
Head of Collection Information Services
Interim Director
Preservation Librarian
Product Manager
Project Manager
XML Database Administrator

Appendix D

Responses to Question 9: Who determined what metadata elements are used in the Web collections at your institution?

Me (Web Resources Collection Coordinator), Metadata Coordinator
Senior archivists
Me (University Records Archivist), in consultation with the Archives Department Head
Metadata Specialist (paraprofessional member of digital initiatives team)
The Electronic Records Archivist
Curator, Special Collections/University Archives
Me (University Archivist)
Metadata & Cataloguing Librarian in consultation with colleagues
I did (Project Manager)
I (Archivist & Records Manager) am building in subject and creator fields so that links to Archive-It content can be easily referenced from the main finding aids to the larger collection
Web Manager
Me (Librarian)
Metadata librarian
Primarily myself (Archivist Coordinator)
A group of librarians, including the selectors (subject librarians), digital projet librarian, and myself (metadata librarian), is currently working on establishing our metadata guideline for Archive-It.
Electronic records archivist, processing coordinators and processing archivists, cataloger
Head of Collection Information Services and Metadata Specialist
Me (Head of Archives and Special Collections)
We are in the process of determining what elements to be used as we just got the service
Digital Collections Coordinator
Information architect
Collection manager
I did (Records and Archives Manager)
Nobody specifically
Myself (Archivist Librarian) and cataloger
Archivist
I did (Library Specialist)
Archivist
Digital Collection Specialists in consultation with metadata cataloger
Digital Archivist and Head of Archives and Special Collections
Me (Director of the Library)
2 professional staff people (both archivists)
Me (Electronic Records Archivist)
It was a team effort
Group decision
Web archive team
The librarian in charge of our technology section.
Head of Collections Management?