Lisa A. Gregory. The Practice and Perception of Web Archiving in Academic Libraries and Archives. A Master's Paper for the M.S. in L.S degree. April, 2009. 58 pages. Advisor: Christopher A. Lee.

The objective of this research is to try to gain a fuller picture of Web archiving activities in libraries and archives at institutions of higher education in the United States, and the perceptions librarians and archivists have of those activities. A Web-based self-administered survey was sent to multiple listservs, and 239 respondents completed the survey.

At the time of this survey, many higher education institutions had not implemented routine Web archiving activities. Although planning and testing was being carried out, these were still in the early stages. The results of this survey reveal that archivists and librarians believe Web sites should be archived, and that cost, support for technology, and lack of trained personnel are some of the factors prohibiting them from doing so.

Headings:

College and university archives

College and university libraries

Digital Preservation

Electronic Data Archives/Conservation and Restoration

Internet/United States

Web sites

THE PRACTICE AND PERCEPTION OF WEB ARCHIVING IN ACADEMIC LIBRARIES AND ARCHIVES

by
Lisa A. Gregory

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
Library Science.

Chapel Hill, North Carolina

April 2009

Approved by

_____

Christopher A. Lee

Table of Contents

List of Tables

List of Figures

Introduction

For many institutions, intellectual property has not just gone digital, it has gone to the Web. Things that were previously issued in print are now solely electronic, making it much more likely that the Web version will be the only version. For information producers, dealing with the Web's ephemeral nature has been deemed an acceptable tradeoff for affordances that include broad dissemination and lower initial publication cost. If information professionals want to continue to ensure access to information in a manner consistent with past collections, they will have to archive Web-based materials. In information science, archival science, and library science literature, the most commonly used term for the preservation of Web sites is "archiving." This word means different things to different people. For the average technology user, archiving may be the equivalent of simply saving the data. Richard Pearce-Moses, in the entry for "archive" in the Society of American Archivists (SAA) *Glossary of Archival and Records Terminology*, acknowledges the double life the word leads, noting that the definition of the word as used in computing is simply "to store data offline" (2005). Helen Tibbo contrasts this meaning with the connotation the word typically has for archivists: "Most popular and computer-oriented usage of the term 'archiving' oversimplifies an involved process and omits any notion of responsibility for the physical and intellectual longevity, authenticity and reliability, and future usefulness of the materials being stored" (2003, p. 8). For the purposes of this study, the primary definition of "archive" from the SAA *Glossary* is used: "To transfer records from the individual or office of creation to a

repository authorized to appraise, preserve, and provide access to those records" (2005). This ensures that the idea of archiving Web sites includes the enhanced stewardship traditionally associated with archives and libraries.

Like the simple term "archive," the process of archiving Web sites can mean different things in different situations. Fundamentally, to archive a Web site means to copy a Web site to an alternate location for the purpose of using it for reference at a later date. Collecting methods can include the use of a harvester, a software program that follows links on the Web (also known as crawling), saving the data it encounters as it goes. Web sites can also be manually archived using offline browsers or by obtaining a copy of the Web site's files directly from the creator.

The way an institution collects Web sites is often related to its selection method. Adrian Brown (2006) describes the three common types of selection methods: unselective, thematic, and selective. The first goes for breadth rather than depth, harvesting entire national domains or even the entire publicly accessible Web. The Internet Archive (IA) and its efforts to collect the Web and make it accessible through the Wayback Machine is the most often cited example of long-term unselective harvesting. For instance, in 2007, the IA completed the largest crawl of the Web in history with the goal of taking a "global snapshot of the web" (IA, Around the world). The second method, thematic, chooses Web sites based on a predefined topic, creator, genre or domain. These types of archives require more human intervention and appraisal. The Library of Congress MINERVA project and the University of Heidelberg's Digital Archive of Chinese Studies are two examples of thematic archiving. Finally, selective archiving, similar to thematic archiving, follows most closely with traditional appraisal or

selection methods. Web sites are purposefully chosen for inclusion in an archives based on their applicability to that institution's mission and goals. One of the most renowned selective archives is the National Library of Australia's PANDORA project.

Regardless of collection or selection method, Web archiving carries with it a number of considerations for the archivist or librarian to navigate. Intellectual property; the interconnected and ephemeral nature of the Web; preserving context and authenticity; and selecting high quality materials are just a few. How does one distinguish a single object on the Web? What about content that is only dynamically generated when users enter a query? How should archived Web sites be presented to users so that they will understand what they see? Peter Lyman (2002) covers the cultural, technical, economic and legal territory that new Web archivists confront. The very nature of the Web, compared with more traditionally archived analog materials, means there are still many answers which information professionals can only see dimly.

When looking at the history of production of information, archiving Web sites and other digital objects seems especially relevant for universities and colleges, which not only produce abundant original research, but have also served as centrally located repositories for regional and disciplinary resources. Yet this author's informal conversations with librarians and archivists have revealed a feeling of trepidation regarding preservation of digital objects in general. This is probably no surprise if these professionals are taking their cue from digital preservation literature. Ross Harvey points out the tendency of those writing about digital preservation to describe the loss of information in dire and emotive terms, such as comparing our current situation to those on the brink of a "digital dark age" (2008, p. 1). The Internet Archive even uses this term

to substantiate their efforts (IA, Why the archive). Picking up on this trend as well, Tibbo points out that, in reality, "the questions concerning long-term preservation vastly outnumber the answers" (2003, p. 6). In order to dig deeper into possible reasons behind archivists' and librarians' reluctance to archive Web sites, the study described here asks professionals to reveal their Web archiving experiences as well as the information sources they consult regarding archiving Web sites. Specifically, the following two research questions are addressed: Are librarians and archivists at institutions of higher education currently engaged in or considering archiving Web sites? What sources do these professionals consult for information about Web archiving?

Literature Review

It is conceivable that as soon as the first Web pages started going online, archivists and librarians began considering how they would capture that information. But to what extent have considerations progressed to action? Because Web archiving initiatives are a niche within the broader field of digital preservation, a review of published surveys of digital preservation initiatives in general will help situate the results of this narrower study within a broader context. Also discussed below are published reports of universities that are archiving Web sites in some capacity. Although few, these reports demonstrate that some institutions of higher education are indeed engaged in archiving Web sites as part of their digital preservation programs. Finally, the professional information seeking practices of librarians and archivists are explored. Because it examines which sources these professionals use to find information about Web site archiving, this study can be considered a subtopic within the body of work described here.

Digital Preservation Initiatives

To date, there is no published survey that focuses specifically on Web archiving initiatives at academic libraries and archives. There are, however, a number of surveys of digital preservation initiatives in general. Cloonan and Sanett conducted one of these surveys in 2002. These authors focused on 13 archival institutions, programs and projects in the United States, Canada, Europe and Australia. Using interviews and questionnaires with open-ended questions, the researchers gathered qualitative data that they then formulated into case studies. Their goal was to try to determine what sorts of preservation strategies and techniques institutions currently use or are developing for future use. Although this survey did not focus on the types of materials being ingested, it did include an open-ended question requesting a description of the digital materials being preserved. No institutions singled out Web sites or marked-up text of any kind (Cloonan & Sanett, p. 99). Only one brief mention of Web sites was reported; respondents mentioned "web site material" as a "problem" (Cloonan & Sanett, p. 80). This study suggests that, for this particular sample, Web sites either were not being archived, or did not come to mind when considering archived formats.

Another study funded by a professional organization that focused on digital preservation was conducted by Hedstrom and Montgomery (1998). Looking at a considerably larger population than the Cloonan and Sanett study above, Hedstrom and Montgomery researched the electronic preservation practices at Research Libraries Group (RLG) member institutions. This RLG-funded study sought "(1) to gather baseline data on the nature and extent of digital preservation problems in member institutions and the status of their digital preservation programs, and (2) to identify needs and requirements of

member institutions in meeting their responsibilities for preserving digital information" (Hedstrom & Montgomery, p. 1). The survey, distributed to 160 RLG members, asked for discrete answers regarding policy, holdings, storage, training, and needs. Follow-up interviews were then performed with administrators at 15 of the responding institutions. Although the survey administered here had a much broader scope than the one proposed in this paper, it did ask whether or not "text files with markup (e.g. SGML, HTML, XML, etc.)" were among the libraries' digital holdings. Out of the 36 institutions that had digital holdings, 75% stated their collections included such files (Hedstrom & Montgomery, p. 33). The survey also sheds some light on professional attitudes toward digital preservation in general:

> The respondents . . . . were fairly evenly divided between those who found digital preservation interesting and stimulating and who had made large personal investments to keep up with the issues, and those who were concerned about the challenges, the absence of clear guidance, and the need for greater expertise. Many of the interviewees said that digital preservation was forcing them to re-examine traditional practices, change the way they administer their departments, develop more interdepartmental relationships, and learn new skills. (Hedstrom & Montgomery, p. 20)

Though the opinions described by Hedstrom and Montgomery are not regarding Web archiving specifically, they characterize the larger area of archiving digital resources and the attitudes librarians have toward professional development in this area. These attitudes seem to include both enthusiasm and concern, and members of the field feel they must try to stay current. It is possible that these feelings extend to the area of preservation of Web resources. Further, "almost 80% of the surveyed institutions stated they planned to use professional training, and almost 70% are counting on independent study" (Hedstrom & Montgomery, p. 22). Not only do librarians feel the need for

additional training, but a large proportion of these institutions seemed to expect their staff to use formal and informal sources for keeping up-to-date on this topic.

Kenney & Buckley (2005) report on surveys administered to 114 librarians and archivists participating in Digital Preservation Management Workshops. These surveys were given to participants before they attended one of two workshops held in 2003-4 and 2005. Of the respondents, 50% were from academic libraries, 27% from government institutions, with the remaining representing institutes, museums and public libraries. More than 90% of the institutions represented by the respondents stated that they had Web content in their repositories. The authors acknowledge that their population is probably already receptive to and/or involved in digital preservation initiatives; for that reason, 90% may not reflect the percentage of academic libraries overall that are engaged in Web archiving.

One study targeting institutions involved in archiving Web sites was identified. The Royal Library, the National Library of Denmark, sent questionnaires to 95 national libraries in March 2007. Jacobsen (2007) writes that the purpose of this survey was to get a better idea of how many national libraries are archiving Web sites, what their procedures and policies are, and what level of interoperability they believe their Web archives should achieve. Out of 37 responses, 19 stated they are currently archiving Web sites and 11 are planning to begin archiving. This means that 81% of the respondents are at least considering archiving Web sites. This survey queried national libraries. Based on the literature, national libraries seem to be tackling this issue much more assiduously

than academic institutions.[1] As a result, this percentage may not be generalizable to the academic libraries and archives targeted in this study. Yet their research design provides a useful starting point for developing further surveys.

Universities and Colleges Archiving Web Sites

Only a few articles were identified that describe archiving of Web sites in some sort of university or college environment. This makes it very difficult to discern how many institutions are undertaking Web site archiving. Even though institutions may be placing general information about their Web archiving activities on their Web sites, this information can be time consuming to locate. Lack of formal published information and aggregated data leaves practitioners to speculate on the state of the field. In 2008, the International Internet Preservation Consortium (IIPC) published a survey which asked for details about its members' Web archiving activities. Though this type of survey has many similarities with the survey used in my study, the population size was much smaller (confined to 39 IIPC member institutions) and more diverse (only five of these members classified themselves as colleges or universities, and only two of these were from the United States). Also, the IIPC members are a self-selected group of institutions with a demonstrated interest in Web archiving, so their findings are not likely to generalize to the general population of academic libraries and archives.

There are several other published accounts of Web archiving in higher education institutions. Lyle (2004) discusses preliminary sampling of the University of Michigan's domain. Selection of Web sites for harvest, even when limited to a finite domain, can be daunting and time consuming. To see if certain types of sampling could be used

[1] An analysis of eight years of proceedings of the International Web Archiving Workshop (IWAW) reveals that only 23 of the presenters were staff members from academic libraries or archives, compared with 164 from national, state or commercial institutions (http://iwaw.europarchive.org/).

effectively for selecting Web sites for harvest, Lyle applied purposive, systematic, random and mixed-mode sampling techniques to the University of Michigan domain. The author found sampling the entire domain using a stratified random sample to be the most successful option, with success being defined as providing an objective sampling method (Lyle, p. 11). Sampling on the whole was determined to be most useful for gaining an overview of a domain (Lyle, p. 12). The author does not mention whether or how this project informs other work at his institution. Given that the University of Michigan lists Web sites in their catalog, it is possible that testing sampling methods for archiving Web sites was undertaken as part of a larger archiving program.

Prom and Swain (2007) studied student organization Web sites at the University of Illinois at Urbana-Champaign to try to determine research and evidential value, and to develop capture procedures. They situate their efforts within a larger institutional context, aiming at "evaluating the potential research value of student organization websites" and "determining a 'best practice' approach for capturing website content" (Prom & Swain, p. 345). At the conclusion of their study, they propose that harvesting student organization Web sites can be a relatively cost-effective method whereby archivists can obtain valuable materials. As evidenced by their intention to reuse the search algorithm they tested and to plan for long-term preservation, the efforts described here do appear to be part of a broader Web archiving program.

Information Seeking of Librarians and Archivists

The published literature regarding the information seeking behaviors of librarians and archivists often cover this topic through the lens of professional development, focusing on motivators for seeking job-related information and perceptions of the

importance of continuing education. As will be seen below, the specific sources these professionals consult or the types of training of which they take advantage are generally a minor facet of study.

In the United Kingdom, New Zealand and Australia, the phrase "continued professional development" (CPD) refers to "a regime of training, research and contribution in the individual's own professional arena which aims to update, expand and enhance skills, knowledge and expertise" (Crockett, 2007, p. 78). When faced with new tasks in the workplace such as archiving Web sites, trying to gain new knowledge via formal or informal routes would fall under the umbrella of CPD. Crockett (2007) compares the archival profession to others that require CPD or lifelong learning for various forms of certification or licensure. Professional associations as well as different types of mentoring relationships in the workplace are two sources this author lists as significant contributors to CPD. Cossham and Fields (2007) looked at librarians' attitudes toward CPD by analyzing results from a major needs assessment survey circulated to librarians in New Zealand. [2] The needs assessment survey results reported on by Cossham and Fields suggest that information and library science staff and their administrators "expressed a preference for short, contact courses and presentations, with 'workshops' featuring as highly desirable, in contrast to conferences, online programmes, long-term study, and so on" (p. 578). These sources of information regarding new work skills may apply to all areas of the profession, including Web archiving.

[2] Although it was after the Cossham and Fields survey was distributed, it may be of interest that in 2007 the professional library association within New Zealand (LIANZA) instituted CPD requirements that librarians must fulfill in order to maintain professional status. This will have an influence on the amount of CPD in which librarians engage, and may also influence how well librarians are able to recall their CPD activities.

Investigations into the professional development of librarians and archivists in the United States yield similar trends. Varlejs (1999) found that librarians engage in both formal and informal learning activities, and that the average respondent relies most heavily on colleagues for information, but also "regularly reads four professional journals and during the preceding year spent sixteen hours attending workshops" (p. 185). An additional important point made in this article that has direct relevance for the study described here is the idea that it may be difficult for professionals to try to remember projects during which they learned something new, and that they may be more successful recalling "skill-oriented projects" than ones centered on amorphous information seeking (Varlejs, p. 177). Because of this, querying professionals about their learning related to archiving Web sites may yield more results than asking about a broader or more theoretical topic.

Two additional articles, each focusing on specific types of librarians (physical science and reference librarians), yield more sources that librarians consult when they look for information on work-related activities. Brown and Ortega (2005) studied whether or not physical science librarians consulted research literature in an effort to determine how much published research informs professional practice. Based on the results of their survey, it appears that "physical science librarians place a significantly higher value on the invisible college as an information resource for their daily activities than the journal literature" (Brown & Ortega, p. 235). The survey respondents rated personal communication and listservs as the most important sources for their daily information seeking activities. Chan and Auster (2003) also used surveys on a subset of the librarian population: reference librarians in public libraries in Toronto, Ontario. This

study looked for significant relationships between (1) updating activities and individual characteristics like age and motivation and (2) participation in updating activities and organizational factors like climate and management. The updating activities referred to in Chan and Auster's study encompass both formal (structured courses and workshops) and informal (attending conferences and self-directed research) efforts. On average, librarians reported that they engaged in approximately 300.8 hours of informal updating activities per year, as opposed to 31.5 hours spent on formal activities. Like Varlejs' suggestion that skill-oriented recall may be easier for librarians, Chan and Auster bring up some aspects of respondents' ability to recall their information seeking. Specifically, the authors point out that because librarians may believe information is freely and abundantly shared in libraries in general, they may conclude that they "do not need to consciously make efforts to learn from others" (Chan & Auster, p. 280). Another interesting side effect is that librarians' updating activities are negatively influenced by how innovative they believe their work climate to be. They appear to equate the currency of their skills with whether or not their library is up-to-date. Both of these conclusions may bear on the ability of librarians to recall their information seeking as well as their belief in the importance of keeping their skills current.

Methodology

The objective of this research was to try to gain a fuller picture of Web archiving activities in libraries and archives at institutions of higher education in the United States, and the perceptions librarians and archivists have of those activities. A Web-based, self-administered survey was chosen to try to achieve this objective. Surveys have been acknowledged as "excellent vehicles for measuring attitudes and orientations in a large

population" (Babbie, 2007, p. 244). Using the Web to deliver a self-administered survey allowed responses to come in from archivists and librarians in many different parts of the country. Allowing respondents to answer at their leisure also, hopefully, encouraged more survey completions. In contrast to interviews or focus groups, the data gained from a survey provides the desired overview of the subject instead of an in-depth look.

Sample

The population under study was librarians and archivists at institutions of higher education in the United States. To facilitate administration of the survey, an announcement was sent to the following listservs to which these professionals subscribe:

Archivist-Affiliated Lists

- Society of American Archivists' Archives and Archivists (A&A) list
- Society of American Archivists' Metadata and Digital Objects list

Librarian-Affiliated Lists

- American Library Association, Association for College and Research Libraries, University Libraries list
- American Library Association, Association for College and Research Libraries, College Libraries list
- American Library Association, Library and Information Technology Association list

Non-Affiliated Lists

- Netpreserve.org Web Curators list
- Web-Archive list

The lists above were selected purposefully to try to reach professional members of the population. Five of these lists are associated with two major American professional organizations for archivists and librarians: the Society of American Archivists (SAA) and the American Library Association (ALA). The other lists specifically discuss digital materials. It should be noted that, in addition to advertisement via these lists, one of the list members posted the call for participation on her blog (www.archivesnext.com). This may have garnered additional participation.

The survey questions were defined so that those not working at institutions of higher education in the United States would be routed to the end of the survey. Still more refinement was built in based on some of the demographic information requested, in order to enable the researcher to focus on responses from the population under study. It is acknowledged that members of the ALA and the SAA may not be representative of the entire population of librarians and archivists. Further, those who subscribe to the selected listservs may be more technologically savvy and/or more interested in topics relating to Web archiving (particularly the Web Curators and Web-Archive lists). Using the lists, however, was a convenient way to contact a large number of geographically dispersed members of the population in an efficient and timely manner, and within the budget and time constraints of the study.

Instrument

Survey administration.

The survey instrument (see Appendix A) was administered using the Qualtrics Web survey tool. This tool was chosen over other Web-based survey systems because of the lack of identifiable data automatically collected by the site. Instructions and definitions for several terms used in the survey were given throughout the survey to help respondents interpret the questions consistently without consulting outside sources (Bourque & Fielder, 2003). The instrument consisted of 15 closed-ended and 3 open-ended questions. Survey respondents had the option of skipping questions they did not wish to answer, and of discontinuing the survey at any time. The survey was available from Tuesday, February 10, 2009 through Tuesday February, 24, 2009.

Discussion of survey questions.

The first four questions asked for demographic information: institution type, institution location, institution size, and the individual's education level. Only responses from those with a bachelor's degree or higher (considered more likely to hold a librarian or archivist position) at academic institutions in the United States were analyzed for this study.

Questions five through twelve focused on the respondent's institution. Questions five and six helped determine the presence or absence of Web archiving activities within the institution. Questions seven and eight asked whether or not these activities are undertaken by a vendor or performed in house. Question nine was slightly more complex, in that the respondent revealed the presence or absence of planning activities at his or her institution. If the respondent answered that his or her institution is archiving Web sites, he or she was directed to the next three questions. If not, the survey took the respondent to question thirteen. Questions ten, eleven, and twelve asked for details about the type of Web sites being archived, whether or not they are being cataloged, and how they can be accessed.

The questions within the remainder of the survey dealt with the librarian's or archivist's individual experiences. Questions thirteen and fourteen helped identify whether or not the respondent him/herself archives digital materials and, if so, what types of materials. Question fifteen identified compelling reasons for archiving Web sites. As a counterpart to question fifteen, respondents were able to select reasons that may inhibit Web archiving in question sixteen.

Questions seventeen and eighteen tried to identify professional resources the archivist or librarian had used or would consider using in order to find out more about archiving Web sites. The list of resources given in these two questions was developed in part by examining professional development literature for commonly cited resources (Brown & Ortega, 2005; Cossham & Fields, 2007; Varlejs, 1999).

Incentives.

At the end of the survey, each respondent was given the opportunity to link out to a remote site and enter an email address if she or he wished to receive a copy of the final report. Respondents were also given the opportunity to enter into a random drawing for one of four $25.00 Amazon.com gift cards. Although the use of monetary incentives is sometimes eschewed by researchers (Bourque & Fielder, 2003, p. 120-1), inclusion here was intended to help draw in responses from those who may not otherwise participate due to their unfamiliarity with the subject matter. The respondents were made aware that their identifying information would be kept separate from the survey information and would not be linked in any way to their survey responses. The information was only used for sending out the final report and/or entering them into the random drawing, and was destroyed after those actions were completed.

Ethical issues.

One ethical issue associated with this survey was possible feelings of inadequacy the respondents may have experienced when questioned about a subject they know little about. If no one at their institution has discussed archiving Web sites or if they have not learned much about the topic, this survey may have led them to feel lacking, professionally. Another ethical issue involves keeping data de-identified. As mentioned

above, efforts were made to ensure respondents understood the anonymity of their responses, and that any identifying information they submitted would be done through a separate Web site and would not be used as part of the data analysis.

Survey advantages and disadvantages.

Surveys have disadvantages and advantages. Although very reliable, meaning it is likely that participants would answer in a similar way if the survey was re-administered, they are weak on validity (Babbie, 2007). It is unclear how accurate a representation of the entire population the results will give. Advertising through a listserv conveniently reaches a broad audience. However, those subscribed to a listserv may not be representative of the entire population. The sample may be biased toward those engaged in or interested in technological activities in the first place (Babbie, 2007). With no sampling frame, the researcher has no idea of the size or scope of population and, in turn, no way of calculating representativeness. Offering the survey through an uncontrolled site, as proposed here, means that there will be no way to be sure those answering are really who they say they are, or that respondents do not complete the survey more than once.

Surveys also ask for self-reported data, which can result in several problems. Unlike direct researcher observation, the accuracy of what is being reported cannot be verified in a self-administered survey. More specifically, the accuracy of a respondent's memory may be in question. Question seventeen asks for the respondent to recall sources used for seeking information about Web archiving. Remembering activities that took place in the past has been cited as a difficult procedure that may lead to guessing or erroneous responses (Chan & Auster, 2003).

Despite these disadvantages, a self-administered survey still suits the purposes of this research. It allowed for "wider geographic coverage, larger samples, and wider coverage within a sample population" (Bourque & Fielder, 2003, p. 10). Especially with Web-based deployment, such a survey was easy to implement, relatively inexpensive, and easily analyzed. The recruitment method made it likely that those taking the survey all received the announcement at relatively the same time, limiting the chance of maturation or history effects. Finally, respondents were hopefully more likely to report candidly on a self-administered instrument (Bourque & Fielder, 2003).

Results

Data gathered using the Qualtrics Web survey tool was exported to Microsoft Excel for more detailed analysis. 310 partial or complete surveys were logged. Out of that number, 239 were identified as completed by archivists or librarians working at institutions of higher education within the United States. 55 responses were from those working in archives, and 184 were from those working in libraries. Overall, the most numerous responses came from Massachusetts and Pennsylvania, each with 18 completed surveys. Responses were not received from any of the United States' territories, or from Alaska, Arkansas, Montana, Nevada, New Mexico, or Vermont. Figure 1 helps show the geographic distribution of responses. When divided by institution type, the geographic distribution of responses was a little different, as can be seen in Table 1.

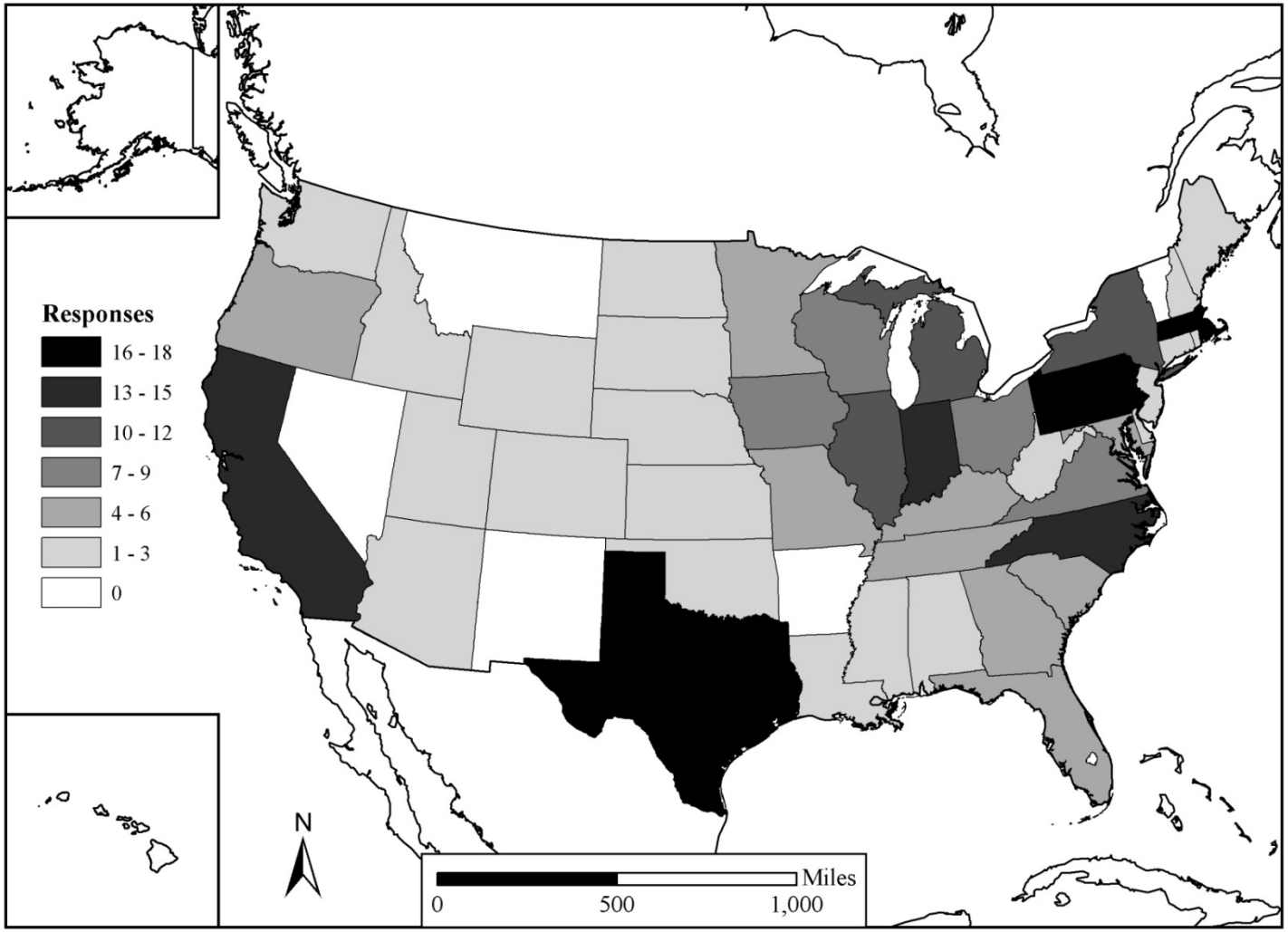

Figure 1: Survey responses by state.

Table 1: States with the Most Numerous Responses, by Institution Type

Archivists		Librarians	
State	**# of Responses**	**State**	**# of Responses**
Texas	7	California	13
Indiana	5	Massachusetts	13
Massachusetts	5	Pennsylvania	13
New York	3	Texas	11
North Carolina	3	North Carolina	10
Ohio	3	Illinois	9
Pennsylvania	3	Indiana	9
		Michigan	9

Both of these groups of responses can be very generally compared with the overall geographic distribution of professionals as represented in the Society of American Archivists' A*CENSUS (2004) results (for archivists) and the National Center for Education Statistics (NCES) supplemental data tables of the "Academic Libraries: 2006 First Look" report (2008, p. 13-4).

Table 2: Top 10 States with the Most Archivists or Librarians, in Descending Order

A*CENSUS (2004)	NCES (2006)
New York	New York
California	California
Massachusetts	Texas
Maryland	Massachusetts
Texas	Pennsylvania
Pennsylvania	Illinois
Washington, D.C.	Florida

Illinois	North Carolina
Ohio	Ohio
Missouri	Michigan

The A*CENSUS survey, done in 2004 by the Society of American Archivists, represents all archivists, not simply those at academic institutions. Still, the distribution of archivists around the country mimics the distribution of responses in my survey. The same can be said of the results from NCES, although these do specifically describe librarians at institutions of higher education. Although both the A*CENSUS and NCES surveys were done several years ago, it can still be stated generally that the respondents in my survey represent a distribution similar to that of the entire population of these professionals.

Figure 2 shows the distribution of responses based on institution size. Smaller institutions and departments were more heavily represented among respondents, with 72.8% coming from those with fewer than 50 staff members.

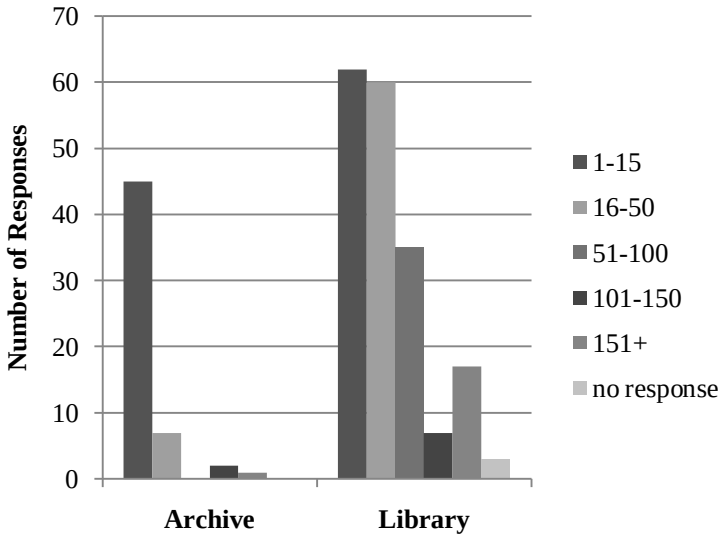

Figure 2: Survey responses by institution type and size.

Archiving Born-Digital Materials

Generally contrasted with digitized materials, born-digital materials are those that have only ever existed in a digital environment. An example would be a memo composed using Microsoft Word, or a Web site created in .html. Within archives, respondents who stated that their institutions are archiving born-digital materials outnumbered those who claimed their institutions are not. The librarians whose institutions are not archiving born-digital materials outnumbered their archivist counterparts.

Table 3: Reported Archiving of Born-Digital Materials, by Institution Type

	Archive		Library		**TOTAL**	
Yes	37	67%	83	45%	**120**	**50%**
No	16	29%	66	36%	**82**	**34%**
I am not sure.	2	4%	35	19%	**37**	**15%**
TOTAL	**55**	**100%**	**184**	**100%**	**239**	**100%**

The archivists answering this survey were much more likely to be archiving born-digital materials as part of their own jobs. Only 18% of librarians claimed to be doing so. Overall, those responding to this survey were not as likely to be involved in archiving born-digital materials.

Table 4: Respondents Who Report Archiving Born-Digital Materials in their Current Job, by Institution Type

	Archive		Library		**TOTAL**	
Yes	33	60%	33	18%	**66**	**28%**
No	22	40%	151	82%	**173**	**72%**
TOTAL	**55**	**100%**	**184**	**100%**	**239**	**100%**

Respondents also gave details about the types of born-digital materials collected at their institutions and within the scope of their own job duties. Figure 3 allows

comparisons between libraries and archives, showing the percentage of those who selected that file type. Figure 4 compares file types archived by librarians and archivists as a part of their jobs. Note especially the prevalence of those archiving .html, .xml, and .arc files. The first two file formats are for documents formatted using a specific markup language typically found on the Web. The third, .arc, is a file format created by the Internet Archive and used to aggregate multiple Web site files into a single archive.

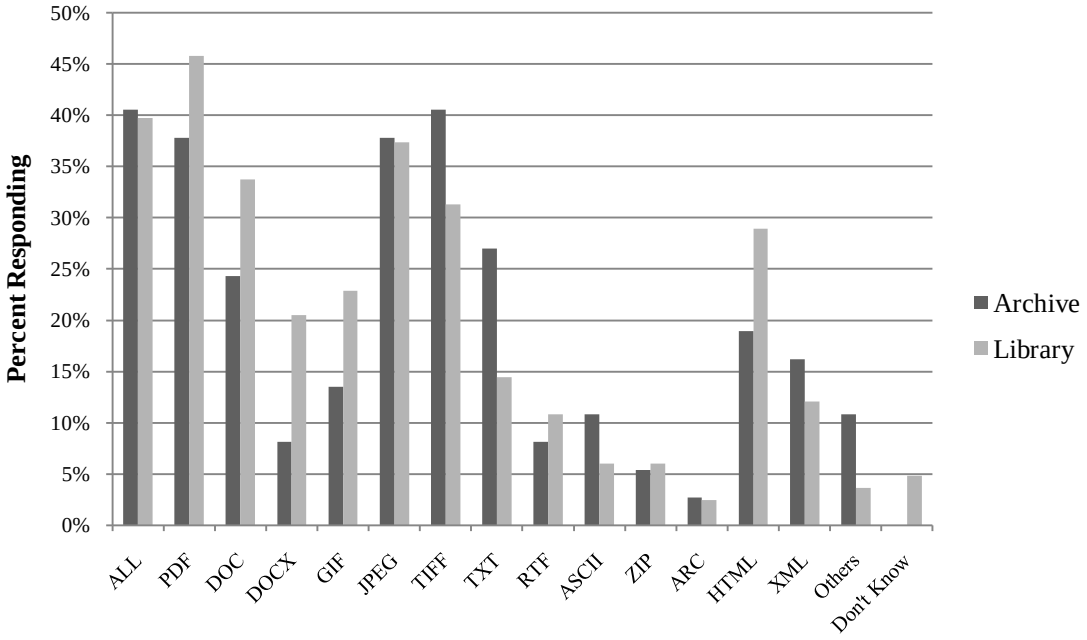

Figure 3: Percent of respondents selecting a file type their institution collects, by institution type.

Note that 31 respondents stated that their institutions archive files in .html, 16 in .xml, and 3 in .arc. The 48 who say their institutions archive all file types may be archiving Web sites or Web-related materials as well. As for personal job duties, four of the respondents report working with .arc files, 31 with .html and 17 with .xml.

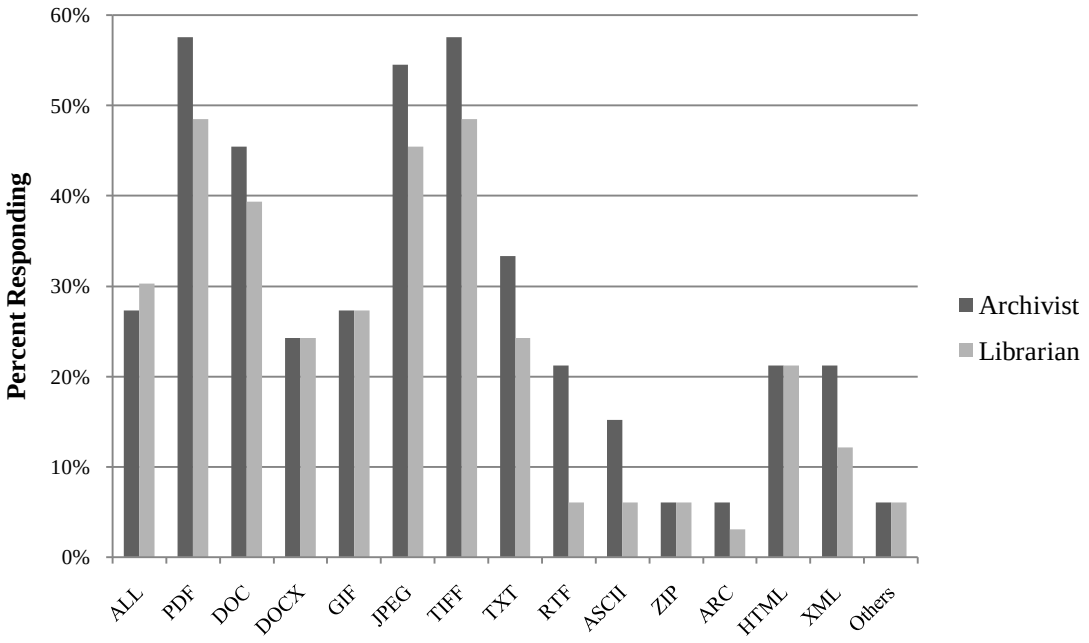

Figure 4: Percent of respondents selecting a file type they archive in their current jobs, by institution type.

Most born-digital materials appear to be archived in house. Only two of the 37 respondents whose archives are archiving born-digital materials stated they were sure that some of this was being carried out by a vendor. For libraries, only nine of the 83 respondents said the same.

Archiving Web Sites

Looking at Web site archiving from a programmatic perspective, each respondent was asked to categorize the state of Web archiving at her or his institution. For the purposes of data analysis, responses one and two have both been categorized as "non-planning," and responses three and four have been categorized as "planning."

The majority of respondents (65%) either did not know whether or not their institutions were planning to archive Web sites, or knew that their institutions had not done any planning to archive Web sites. Only 6% of all respondents knew their institutions to have implemented routine Web archiving procedures (7% of those working

at archives and 6% of those working at libraries). Only one librarian and one archivist responded that they believed that their institutions had been archiving Web sites in the past, but had ceased.

Table 5: Web Archiving Situation, by Institution Type

	Archive		Library		**TOTAL**	
I do not know if my institution has planned for archiving websites. (*Non-planning*)	10	18%	64	35%	**74**	**31%**
My institution has not planned for archiving websites. (*Non-planning*)	21	38%	61	33%	**82**	**34%**
My institution is currently planning to archive websites in the future. (*Planning*)	8	15%	17	9%	**25**	**10%**
My institution has tested some website archiving procedures. (*Planning*)	9	16%	18	10%	**27**	**11%**
My institution has implemented routine Web archiving procedures.	4	7%	11	6%	**15**	**6%**
My institution has archived websites in the past, but is no longer doing so.	1	2%	1	1%	**2**	**1%**
Other	2	4%	10	5%	**12**	**5%**
No response	0	0%	2	1%	**2**	**1%**
TOTAL	**55**	**100%**	**184**	**100%**	**239**	**100%**

The next three questions asked those whose institutions are archiving Web sites to give details about procedures, including selection criteria, cataloging, and access. Only those who responded that their institutions had tested or implemented routine Web site archiving procedures were given the opportunity to respond to this question. An open-ended question probed for details regarding selection criteria. Responses to this question are listed, according to institution type, in Appendix B and are discussed further below. Adding records to a catalog for an archived Web site is not common, with only 4%

stating that their institutions do so (see Table 6). Access to archived Web sites is variable, with no one access method heavily outweighing another (see Table 7). The "other" responses to this question give insight into the range of situations in which institutions find themselves as they continue to refine their procedures (Table 8).

Table 6: Records for Archived Web Sites Added to Catalog, by Institution Type

	Archive	Library	**TOTAL**
Yes	5	4	**9**
No	9	33	**42**
I am not sure.	1	4	**5**
TOTAL	**15**	**41**	**56**

Table 7: Method of Accessing Archived Web Sites, by Institution Type

	Archive	Library	**TOTAL**
I am not sure.	1	10	**11**
No access - the archive is completely dark.	2	2	**4**
Only staff can access these websites.	4	9	**13**
Staff and patrons can only access archived websites on-site.	1	0	**1**
Staff and patrons can access archived websites both on- and off-site.	4	9	**13**
Other (Please describe.)	3	9	**12**
TOTAL	**15**	**39**	**54**

Table 8: Open-Ended Responses Regarding Method of Accessing Archived Web Sites

Archive
to be determined
only in test phase. Not available to anyone at moment..
The websites are on a server, but copies could be created and easily accessed on the public research room computer without the researcher having access to the original files.
publicly available

Under development, currently there is limited staff access
Library
no archive yet
We are starting a university archives digitization project which is not yet available.
we do not archive websites
Still evolving
The archive is currently dark. The [ ][a] user interface should be coming in the summer.
We archive digital collections which are openly accessible but we don't archive websites.
none archived yet; we're still investigating
public access as well
Wayback Machine

[a] Removed to preserve anonymity.

Perceptions of Archiving Web Sites

In question fifteen, respondents were asked to select the most compelling reasons for archiving Web sites. They were given a list of nine options, but could also select "Other" and enter an opinion not listed. The four reasons for archiving Web sites that the respondents found to be most compelling, for those working at both archives and libraries, are (1) in order to document history, (2) for future research, (3) information online is within institution's collecting scope, and (4) to protect an institution's intellectual property. 85% of all respondents chose documenting history as a compelling reason to archive Web sites. Only 8 respondents (3%) felt that Web sites do not need to be archived.

Table 9: Compelling Reasons for Archiving Web Sites, by Institution Type

	Archive		Library		**TOTAL**	
In order to document history	47	85%	156	85%	**203**	**85%**

For future research	41	75%	122	66%	**163**	**68%**
Information online is within an institution's collecting scope	42	76%	94	51%	**136**	**57%**
To protect an institution's intellectual assets	25	45%	102	55%	**127**	**53%**
Information online may be needed for legal purposes	20	36%	73	40%	**93**	**39%**
Charged by legal mandate, such as public records law	14	25%	61	33%	**75**	**31%**
To keep up with new technological developments in archiving	17	31%	22	12%	**39**	**16%**
For novelty	4	7%	10	5%	**14**	**6%**
Other	4	7%	4	2%	**8**	**3%**
I do not feel that websites need to be archived.	0	0%	8	4%	**8**	**3%**

There was more variance between institution types when considering obstacles to archiving Web sites. Those working at archives see lack of support for technology and lack of trained personnel as the top two reasons for not archiving Web sites, while those at libraries cite cost and lack of administrative support as first and second. When considering both institution types together, the top five reasons for not archiving Web sites are (1) cost, (2) lack of administrative support, (3) lack of support for technology, (4) lack of storage space for archived sites, and (5) lack of trained personnel.

Table 10: Compelling Reasons for Not Archiving Web Sites, by Institution Type

	Archive		Library		**TOTAL**	
Cost	29	53%	115	63%	**144**	**60%**
Lack of administrative support	30	55%	106	58%	**136**	**57%**
Lack of support for technology	39	71%	96	52%	**135**	**56%**
Lack of storage space for archived sites	27	49%	102	55%	**129**	**54%**
Lack of trained personnel	36	65%	90	49%	**126**	**53%**
Information available on the Web can be collected in other ways	8	15%	38	21%	**46**	**19%**
Outside the institution's collecting scope	7	13%	19	10%	**26**	**11%**

Other	6	11%	20	11%	**26**	**11%**
Other institutions are taking care of this	0	0%	11	6%	**11**	**5%**

Seeking Information on Archiving Web Sites

Fewer than half of those surveyed (109, or 45%) responded that they had not sought information on archiving Web sites. Of those who had, regardless of institution type, the top four cited sources for information were journal articles, conference presentations, individual Web archive sites, and listservs (in that order).

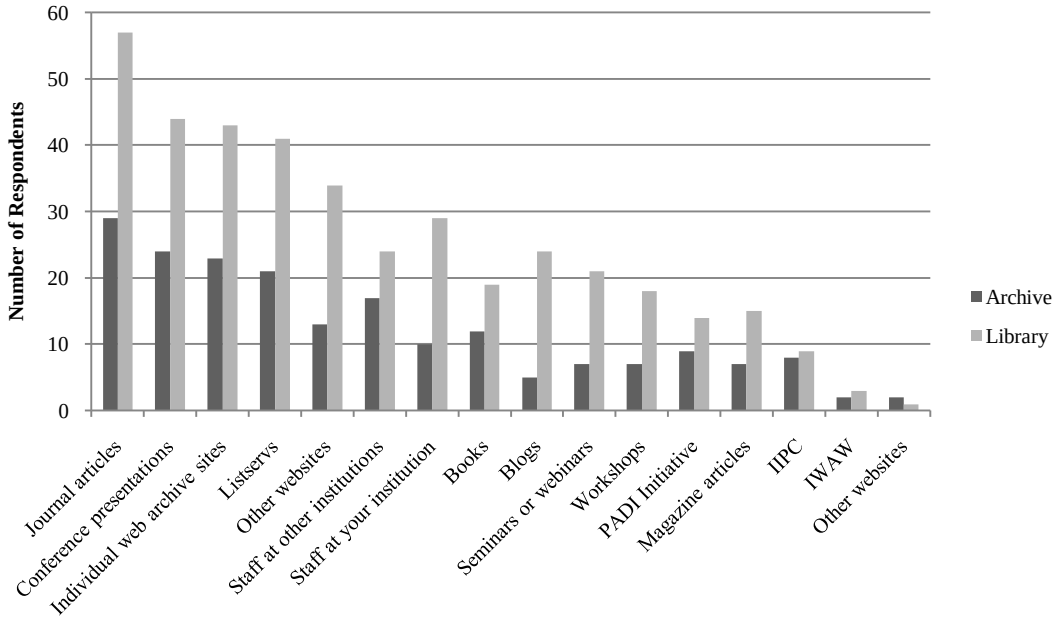

Figure 5: Resources consulted for Web site archiving information, by institution type.

When asked what resources they would consult for information regarding archiving Web sites, respondents listed similar first choices, regardless of institution. Table 11 summarizes the frequency with which different resources were chosen first, both by institution type and altogether. The top five and bottom five choices are highlighted.

Table 11: First Choice Resource, by Institution Type

Archive	
Journal articles	10
Workshops	10
Individual Web archive sites	8
Staff at your institution	7
Staff at other institutions	6
Seminars or webinars	5
PADI Initiative	4
Conference presentations	4
IIPC	3
Other websites	2
Listservs	2
IWAW	1
Other	1
Books	0
Magazine articles	0
Blogs	0

Library	
Journal articles	35
Staff at your institution	29
Staff at other institutions	23
Individual Web archive sites	22
Listservs	16
Workshops	14
Conference presentations	12
Seminars or webinars	11
PADI Initiative	9
IWAW	9
Other websites	8
Books	8
IIPC	6
Blogs	6
Magazine articles	3
Other	0

TOTAL	
Journal articles	45
Staff at your institution	36
Individual Web archive sites	30
Staff at other institutions	29
Workshops	24
Listservs	18
Seminars or webinars	16
Conference presentations	16
PADI Initiative	13
Other websites	10
IWAW	10
IIPC	9
Books	8
Blogs	6
Magazine articles	3
Other	1

The next table presents the values of the resources after being weighted and aggregated. The following scale was used: 1 = 10 pts., 2 = 8 pts., 3 = 6 pts., 4 = 4 pts., 5 = 2 pts. In this manner, a respondent rating a resource as a 1 would weigh the same as five respondents each rating the same resource with a score of 5. This gives an idea of the overall relative preference respondents had for a resource.

Table 12: Resource Choices, Weighted, by Institution Type

	Archive	Library	TOTAL
Journal articles	532	922	1454
Individual Web archive websites	590	614	1204
Staff at other institutions	312	714	1026
Seminars or webinars	382	476	858
Staff at your institution	290	554	844
Conference presentations	272	492	764
Workshops	264	456	720
Listservs	162	456	618
PADI Initiative	140	376	516
Other websites	80	382	462
Blogs	72	256	328
Books	46	276	322
IIPC	90	220	310
IWAW	80	214	294
Magazine articles	26	102	128
Other(s) (Please describe.)	10	14	24

Discussion

Status of Web Archiving Activities at Institutions of Higher Education

The results of this survey suggest that many academic archives and libraries are in the investigation or planning stages when it comes to archiving born-digital objects. Those who work in archives were more likely to state that their institutions are engaged in the archiving of born-digital materials (see Table 3). For file types, more institutions are collecting word processing documents and PDF files than digital objects formatted in a markup language (see Figure 3). This mirrors the frequent mention of PDF in the responses to question ten. There is a similar distribution for those who responded that they are archiving born-digital materials as part of their job (see Figure 4). Another interesting discovery is that 48 of the 239 (20%) respondents selected the option "We archive all file types." This may indicate that these institutions have decided not to be selective when ingesting files. Whether or not institutions are doing this because they have decided it is best practice, or because they feel it is safer to take in all types and deal with dissemination issues later, is unknown.

Even fewer respondents described their institutions as having added Web site archiving to their born-digital archiving program (see Table 5). The majority of respondents (65%) stated that their institutions are in a "non-planning stage." These were either unaware of Web archiving plans or activities or knew that none had occurred. 21% are in a "planning" stage, which includes the knowledge of some planning or testing. Only 6% indicated that a routine Web archiving procedure is in place at their institutions. This is striking. Respondents volunteered to complete a survey regarding an activity that, at this point in time, appears to rarely take place on a regular basis. This is even more

interesting when considering that the call for participation went to two listservs specifically about archiving Web sites. These professionals are interested in this topic, even if (or perhaps because) it is not a routine part of their institutions' activities.

When broken down by institution type, a greater percentage of respondents stated that their archives are in a planning stage (31%, compared with 19% for libraries). The following table looks at Web archiving among those institutions currently archiving born-digital materials. Many still are in the non-planning stage compared with a planning stage or implementation.

Table 13: Status of Web Archiving Activities among Respondents Whose Institutions are Archiving Born-Digital Materials

	Archive	Library	**TOTAL**
I do not know if my institution has planned for archiving websites.	5	23	**28**
My institution has not planned for archiving websites.	13	21	**34**
My institution is currently planning to archive websites in the future.	6	8	**14**
My institution has tested some website archiving procedures.	8	14	**22**
My institution has implemented routine Web archiving procedures.	4	10	**14**
My institution has archived websites in the past, but is no longer doing so.	1	0	**1**
Other (Please describe.)	0	6	**6**
No response	0	1	**1**

From these results, it appears that even those who have ventured into archiving born-digital materials, whether systematically or on an ad hoc basis, are still not targeting Web sites. Those who are simply testing have not yet implemented a routine program.

Even without routine programs in place, respondents were able to give a picture of some of the Web archiving procedures their institutions have used to date. Of the 50 answers to question ten, an open-ended question seeking selection criteria used when

archiving Web sites, 24 respondents mentioned the use of selective collection criteria (see Appendix B). Thematic (choosing Web sites based on a predefined topic, creator, genre or domain) and unselective (harvesting for breadth rather than depth) criteria are much less prevalent, with 2 and 1 responses, respectively. University- or college- created sites are of highest interest, with 20 of the 27 collecting institutions listing that as their sole or primary interest. These answers show that institutions are looking to safeguard their own content in a Web archive before looking further. This instinct may grow out of the mandates that many college or university archives have, requiring them to house and give access to their own institutional records and publications, and the long tradition of their library counterparts to support the curriculum and faculty. Staff members at some of these institutions are applying the same collection criteria to Web sites that they use for print or other electronic materials. As is true with other materials, when information is "going to go away" or is at "risk of disappearing," the collectors have been spurred to act. A few also mentioned obtaining permission from the rights holder(s) before archiving. It is unclear whether or not obtaining permission extends to sites produced by their own institutions, or if these comments were in reference to external sites.

Few are archiving Web sites on a routine basis, which accounts for the small number of responses to questions eleven and twelve. Most who are do not add records to their catalogs or give full access to archived sites at this point in time (see Tables 6 and 7). The emerging nature of Web archiving is reflected in some of the comments entered as a response to question 12, such as "under development," "still evolving" and "we're still investigating."

Out of all of the reasons the respondents felt were most compelling for archiving Web sites, the one chosen most often was to document history (see Table 9). This awareness of historical value may stem from the fact that a number of institutions of higher education are some of the oldest institutions in the nation. When comparing archives to libraries, respondents from archives more highly ranked the fact that the site falls within the institution's collecting scope, and were more likely to select "to keep up with new technological developments in archiving." These selections may have been due to the phrasing of these choices: archivists may be more familiar with the term "collecting scope" and may have resonated with the idea of archiving technology. Similarly, those from libraries privileged protecting intellectual assets, a term possibly more familiar to librarians who actively preserve scholarly communication, supporting the work of their faculty members. The idea of having a Web site available for future research or for legal purposes was considered similarly important regardless of institution type. Out of all 239 respondents, only eight stated that they do not feel Web sites need to be archived. Even though so few feel that Web sites do not need to be archived, so many are not yet doing so on a regular basis, signaling the gulf between the ideal and current practice.

The overall top five choices for reasons prohibiting archiving Web sites were tightly clustered (see Table 10), with a considerable difference between choice five and choice six. Cost ranked first for librarians and third for archivists, although it can also be tied into some of the other selections. When considered by institution type, archivists were more concerned with lack of support for technology and lack of trained personnel. Based on results from Figure 2, this may be due to the number of smaller institutions

represented among respondents. These archives may be dependent on a parent department for technology support or, with fewer personnel, may find it more difficult to cover all desired tasks. Few libraries or archives consider redundancy of information as a reason not to archive Web sites. Only 11 felt as though other institutions could be counted on to take care of archiving Web sites (and three of those 11 had responded that they do not feel Web sites need to be archived). These professionals either do not feel inclined to shift the responsibility, or they recognize that, despite the necessity, no one is adequately fulfilling this role.

Information Seeking about Web Archiving

A good proportion of respondents to this survey were not engaged in working with born-digital materials (40% in archives and 82% in libraries). In terms of generalizability of the survey's findings, this is important: it means that the opinions of those who do not archive Web sites as a part of their jobs are represented.

109 respondents stated they had not sought information regarding archiving Web sites. Those who had sought information turned most often to journal articles, conference presentations, or individual archive Web sites (see Figure 5). Listservs also figured in prominently to this list (although this may be due to the fact that the survey was advertised via listserv). Staff members either at the respondent's institution or elsewhere were consulted frequently. Less cited were resources devoted solely to digital materials or archiving Web sites, such as the Preserving Access to Digital Information (PADI) initiative, the International Internet Preservation Consortium (IIPC), or the International Web Archiving Workshop (IWAW). Respondents, because of their unfamiliarity with the subject, may be unaware of or less familiar with these targeted resources.

Workshops and seminars/webinars ranked higher as resources the respondents would consult in the future than they did as previously consulted resources (see Table 11). As institutions progress into or through a planning stage, these more targeted and hands-on vehicles for information may be more appealing. The top three resources librarians and archivists would consult aligns closely with those they have consulted in the past: journal articles, individual Web archive sites, and staff at other institutions. Although "staff at your institution" was more often chosen as a first choice, when weighted (see Table 12), staff at other institutions ranked higher. This may mean that respondents feel the expertise lies elsewhere, as they did when citing lack of trained personnel in question sixteen.

Limitations of the Study and Possible Future Research

The nature of this survey was exploratory, and the discussion above should be taken as an approximation of the state of the field. This is due to some of the disadvantages of the research method (mentioned earlier) and to some of the possible limitations that follow. It should be noted that because the unit of analysis (institutions) is different from the unit of study (individuals), a single institution may be represented more than once within these results, and that the representation of institutional activity is dependent on the particular knowledge and perspective of the individuals who completed the survey.

Although it appears that more archives are engaging in the archiving of born-digital materials and, specifically, Web sites, this may be misleading. The title of the survey and its instructions included the phrase "archiving," which may have drawn more archivists to respond due to their increased familiarity with the term. There is also no

way of knowing whether or not all fully understood the idea of archiving Web sites as meant in the context of this survey. Based on a few of the comments, some appeared to be thinking of archiving any digital material, or simply aggregating URLs to live Web sites. Still, an effort was made to garner responses from a varied sample and, to the extent possible, use the terms commonly assigned to Web archiving while still making the concepts accessible to all.

Results of a similar study conducted in the future would be able to show any progress to adopt or the decision to abandon Web site archiving programs in libraries and archives. Further, more specific information, garnered through focus groups or interviews, might give more insight into the exact nature of these professionals' personal opinions regarding Web site archiving. If they truly think it should happen, how can the obstacles be overcome? Or is this simply another project in a long line of priorities that have to vie for resources and personnel? Finally, for those who have either tested or implemented routineWeb site archiving programs, interviews might help define the methods they used to achieve such a result, comparing those methods across institutions and situations.

Conclusion

The rise of the World Wide Web and its widespread use by the public is generally traced back to the 1993 release of the user-friendly Mosaic browser (Campbell-Kelly & Aspray, 2004). With such a recent birth, the body of literature on archiving Web sites, compared with archiving print materials, is in its infancy. A review of the literature shows that there are few resources available that help sketch the picture of Web site archiving at institutions of higher education, let alone how librarians and archivists, who

may or may not be engaged in archiving Web sites, view the activity. Many sources deal with a broader spectrum of digital objects, of which Web sites are only a small part, or omit reference to Web sites altogether.[3] Besides anecdotal evidence, no formal studies have attempted to identify the reliable sources librarians and archivists at institutions of higher education consult for Web archiving information. It is hoped that the results of this study give practitioners and other interested professionals a general, interpretable picture of the Web archiving efforts in which colleges and universities are currently engaged and how archivists and librarians are gathering intelligence regarding archiving Web sites. Although a large number have not yet begun to plan for archiving Web sites, a good many have, with some even testing and implementing Web archiving procedures. Discovering that peer institutions are beginning to consider archiving Web sites may assist information professionals in convincing administrators to increase funding at their own institutions.

Finally, filling in some of the unknowns may help diminish the perceived obstacles that those in the profession feel toward digital preservation in general, and archiving Web sites in particular. It is clear that those who took the time to take this survey feel that archiving Web sites is worthwhile and, possibly, up to their own archives or libraries to undertake. The results of this survey may help develop more effective outreach to and training of those in the profession who need reliable resources on this emerging archival responsibility. Those most concerned with disseminating information on archiving Web sites should be encouraged to publish in journals, offer seminars/webinars, put information on the Web in conjunction with current Web

[3] Hedstrom and Montgomery (1998), Kenney & Buckley (2005), and Mugridge (2006) are just a few examples.

archives, and make themselves available to their peers as these are the resources it seems many will turn to if and when they have questions. It is hoped that librarians and archivists, aware that others are pursuing this activity and offered more abundant outlets for information will be encouraged to begin the process of archiving the Web-based materials they see as important rather than relegating Web archiving to the realm of the imagination.

Bibliography

Babbie, E. (2007). *The practice of social research* (11th ed.). Belmont, CA: Thomson/Wadsworth.

Bourque, L. B., & Fielder, E. P. (2003). *How to conduct self-administered and mail surveys* (2nd ed.). Thousand Oaks: Sage Publications.

Brown, A. (2006). *Archiving websites: A practical guide for information management professionals*. London: Facet.

Brown, C. M., & Ortega, L. (2005). Information-seeking behavior of physical science librarians: Does research inform practice? *College & Research Libraries*, *66*(3), 231-47.

Campbell-Kelly, M., & Aspray, W. (2004). *Computer: A history of the information machine* (2nd ed.). Boulder: Westview Press.

Chan, D. C., & Auster, E. (2003). Factors contributing to the professional development of reference librarians. *Library & Information Science Research*, *25*(3), 265-86.

Cloonan, M., & Sanett, S. (2002). Preservation strategies for electronic records: Where we are now—obliquity and squint? *American Archivist*, *65*(1), 70-106.

Cossham, A., & Fields, A. (2007). Balancing individuals' expectations and organisational requirements for continuing professional development. *Library Review*, *56*(7), 573-84.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly*, *65*(2), 230-53.

Crockett, M. (2007). Continuing professional development and the hallmarks of professionalism: An overview of the current environment for the record-keeping profession. *Journal of the Society of Archivists*, *28*(1), 77-102.

Doney, E. (1998). Developing opinions: The attitude of ILS staff to continuing professional development. *Library Management*, *19*(8), 486-91.

Greenstein, D., & Thorin, S. E. (2002). *The digital library: A biography*. Retrieved 19 March 2008, from http://www.clir.org/PUBS/reports/pub109/pub109.pdf.

Grotke, A. (2008). International Internet Preservation Consortium 2008 member profile survey results. Retrieved 8 January 2009, from http://netpreserve.org/publications/reports.php?id=005.

Harvey, R. (2008). So where's the black hole in our collective memory? A provocative position paper. Digital Preservation Europe. Retrieved 23 September 2008, from http://www.digitalpreservationeurope.eu/publications/position/Ross_Harvey_black_hole_PPP.pdf.

Hedstrom, M., & Montgomery, S. (1998). *Digital preservation needs and requirements in RLG member institutions*. Research Libraries Group. Retrieved 23 September 2008, from http://www.oclc.org/programs/ourwork/past/digpresneeds/digpres.pdf.

Internet Archive. (2007). Around the world in 2 billion pages. Retrieved 14 November 2008, from http://wa.archive.org/aroundtheworld/.

Internet Archive. (nd). Why the archive is building an 'Internet Library.' Retrieved 14 November 2008, from http://www.archive.org/about/about.php.

Jacobsen, G. (2007). Webarchiving internationally: Interoperability in the future? In *World Library and information congress: 73rd IFLA general conference and*

council. Durban, South Africa. Retrieved 21 October 2008, from http://www.ifla.org/IV/ifla73/papers/073-Jacobsen-en.pdf.

Kenney, A. R., & Buckley, E. (2005). Developing digital preservation programs: The Cornell survey of institutional readiness, 2003-2005. *RLG DigiNews*, *9*(4). Retrieved 23 September 2008, from http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file1088.html.

Lyle, J. (2004). Sampling the umich.edu domain. In *Proceedings of the 4th International Web Archiving Workshop*. Bath, United Kingdom. Retrieved 23 September 2008, from http://iwaw.europarchive.org/04/Lyle.pdf.

Lyman, P. (2002). Archiving the World Wide Web. In *Building a national strategy for digital preservation: Issues in digital media archiving* (pp. 38-51). Washington, D.C.: Council on Library and Information Resources. Retrieved 22 August 2008, from http://www.clir.org/pubs/reports/pub106/pub106.pdf.

Martin, K. E., & Eubank, K. (2007). The North Carolina state government website archives: a case study of an American government web archiving project. *The New Review of Hypermedia and Multimedia*, *13*(1), 7-26. Accessible at: http://www.informaworld.com.libproxy.lib.unc.edu/openurl?genre=article&issn=1361-4568&volume=13&issue=1&spage=7

Mugridge, R. L. (2006). SPEC Kit 204: Managing digitization activities. Association of Research Libraries. Retrieved September 20, 2008, from http://www.arl.org/bm~doc/spec294web.pdf.

National Center for Education Statistics. (2008). Supplemental tables from academic libraries: 2006 first look report. Retrieved 27 March 2009, from http://nces.ed.gov/pubs2008/2008337_s.pdf.

Pearce-Moses, R. (2005). *A Glossary of Archival and Records Terminology*. Retrieved September 23, 2008, from http://www.archivists.org/glossary/.

Prom, C. J., & Swain, E. D. (2007). From the College Democrats to the Falling Illini: Identifying, appraising, and capturing student organization websites. *American Archivist, 70*(2), 344-63.

Society of American Archivists (2006). A*CENSUS data tabulated by state. Retrieved 27 March 2009, from http://www.archivists.org/a-census/.

Tibbo, H. (2003). On the nature and importance of archiving in the digital age. *Advances in Computers, 57*. 1-67.

Varlejs, J. (1999). On their own: Librarians self-directed, work-related learning. *Library Quarterly, 69*(2), 173-201.

Appendix A: Survey Instrument

This survey is designed to collect data regarding (1) web archiving experiences and (2) sources of information about archiving websites. **All archivists and librarians at institutions of higher education in the United States are invited to take this survey.** It should take around 10-15 minutes of your time.

If you complete the survey, you will be given the opportunity to enter a drawing for one of four $25.00 Amazon.com gift certificates. Those who complete the survey may also submit their email address in order to receive a copy of the final report from this research project.

Below are some additional details about the survey. Please review this information and then **proceed to the survey by clicking the button at the bottom of the page.** Thank you for your participation!

[QUESTIONS 1-3 (Page 2)]
The first portion of the survey contains questions asking for general characteristics about you and your institution.

Which of the following BEST describes your current employer? (Note: If you are employed by a college or university run by a government entity, please select "College- or University-Level Archive" or "College- or University-Level Library," as appropriate.)

☐ College- or University-Level Archive
☐ College- or University-Level Library
☐ Elementary, Middle, or High School Library [if selected, skip to end of survey]
☐ Corporate Archive [if selected, skip to end of survey]
☐ Corporate Library [if selected, skip to end of survey]
☐ Government Archive [if selected, skip to end of survey]
☐ Government Library [if selected, skip to end of survey]
☐ Public Archive [if selected, skip to end of survey]
☐ Public Library [if selected, skip to end of survey]
☐ Self-Employed
☐ Retired
☐ Unemployed
☐ Other (Please describe.) _____

In what state, district, or territory is your institution located?

☐ Alabama
☐ Alaska
☐ American Samoa
☐ Arizona
☐ Arkansas
☐ California
☐ Colorado
☐ Connecticut
☐ Delaware
☐ District of Columbia
☐ Florida
☐ Georgia

- ☐ Guam
- ☐ Hawaii
- ☐ Idaho
- ☐ Illinois
- ☐ Indiana
- ☐ Iowa
- ☐ Kansas
- ☐ Kentucky
- ☐ Louisiana
- ☐ Maine
- ☐ Maryland
- ☐ Massachusetts
- ☐ Michigan
- ☐ Minnesota
- ☐ Mississippi
- ☐ Missouri
- ☐ Montana
- ☐ Nebraska
- ☐ Nevada
- ☐ New Hampshire
- ☐ New Jersey
- ☐ New Mexico
- ☐ New York
- ☐ North Carolina
- ☐ North Dakota
- ☐ Ohio
- ☐ Oklahoma
- ☐ Oregon
- ☐ Pennsylvania
- ☐ Puerto Rico
- ☐ Rhode Island
- ☐ South Carolina
- ☐ South Dakota
- ☐ Tennessee
- ☐ Texas
- ☐ Utah
- ☐ Vermont
- ☐ Virginia
- ☐ Virgin Islands
- ☐ Washington
- ☐ West Virginia
- ☐ Wisconsin
- ☐ Wyoming
- ☐ Other [if selected, skip to end of survey]

Please list the number of staff members at your library or archive. You may estimate, if you wish.

_____

[QUESTION 4 (Page 3)]

What is the highest level of education you have completed?

- ☐ Less than High School [if selected, skip to end of survey]
- ☐ High School / GED [if selected, skip to end of survey]
- ☐ Some College [if selected, skip to end of survey]
- ☐ 2-year College Degree (for example: A.A., A.S.) [if selected, skip to end of survey]
- ☐ 4-year College Degree (for example: B.A., B.S.) [if selected, skip to end of survey]
- ☐ Master's Degree (for example: M.A., M.S., M.L.S., M.L.I.S.)
- ☐ Doctoral Degree (for example: Ed.D., Ph.D.)
- ☐ Professional Degree (for example: J.D., M.D.)
- ☐ Other (Please describe.)
- ☐ I would rather not say.

[QUESTION 5 (Page 4)]

The next portion of the survey contains questions asking for information regarding digital materials (including websites) and your institution.

For the purposes of this survey, the term "archive" means collecting and preserving an item.

Does your institution currently archive born-digital materials (for example Microsoft Word documents, emails, digital photographs)?

☐ Yes
☐ No [if selected, skip to Question 9]
☐ I am not sure. [if selected, skip to Question 9]

[QUESTION 6 (Page 5)]
Please list the types of file formats of born-digital materials that your institution archives. (Please select all that apply.)

☐ We archive all file types.
☐ PDF
☐ DOC
☐ DOCX
☐ GIF
☐ JPEG
☐ TIFF
☐ TXT
☐ RTF
☐ ASCII
☐ ZIP
☐ ARC
☐ HTML
☐ XML
☐ Other(s) (You may list multiple formats here.)
_____

[QUESTION 7 (Page 6)]
Does your institution contract with a vendor or outside agency to archive born digital materials?

☐ Yes
☐ No [if selected, skip to Question 9]
☐ I am not sure. [if selected, skip to Question 9]

[QUESTION 8 (Page 7)]
Please list the vendor(s) or agenc(ies) here. If you are not sure, please type "I am not sure." If you would rather not say, please type "I would rather not say."
_____

[QUESTION 9 (Page 8)]
Please select the statement below that best describes your situation:

☐ I do not know if my institution has planned for archiving websites. [if selected, skip to Question 13]
☐ My institution has not planned for archiving websites. [if selected, skip to Question 13]
☐ My institution is currently planning to archive websites in the future. [if selected, skip to Question 13]
☐ My institution has tested some website archiving procedures.

☐ My institution has implemented routine web archiving procedures.
☐ My institution has archived websites in the past, but is no longer doing so. [if selected, skip to Question 13]
☐ Other (Please describe.) _____

[QUESTION 10 (Page 9)]
Please describe the selection criteria your institution uses for choosing websites to archive.
_____

[QUESTION 11 (Page 10)]
Does your institution add records to its catalog for archived websites?
☐ Yes
☐ No
☐ I am not sure.

[QUESTION 12 (Page 11)]
What type of access to archived websites does your institution offer?
☐ I am not sure.
☐ No access – the archive is completely dark.
☐ Only staff can access these websites.
☐ Staff and patrons can only access archived websites on-site.
☐ Staff and patrons can access archived websites both on- and off-site.
☐ Other (Please describe.)

[QUESTION 13 (Page 12)]
The remainder of the survey contains questions about digital materials and your own opinions and experience.

Do your current job duties involve archiving born-digital materials (for example Microsoft Word documents, emails, digital photographs)?
☐ Yes
☐ No [if selected, skip to Question 15]

[QUESTION 14 (Page 13)]
Please list the types of file formats of born-digital materials that you archive as part of your job. (Please select all that apply.)
☐ I archive all file types.
☐ PDF
☐ DOC
☐ DOCX
☐ GIF
☐ JPEG
☐ TIFF
☐ TXT
☐ RTF
☐ ASCII
☐ ZIP
☐ ARC
☐ HTML
☐ XML
☐ Other(s) (You may list multiple formats here.) _____

[QUESTION 15 (Page 14)]
Please answer the next two questions even if your institution is NOT currently archiving websites.

What do you feel are the most compelling reasons for archiving websites? (Select all that apply.)
☐ I do not feel that websites need to be archived.
☐ Charged by legal mandate, such as public records law
☐ Information online is within an institution's collecting scope
☐ Information online may be needed for legal purposes
☐ In order to document history
☐ For novelty
☐ For future research
☐ To protect an institution's intellectual assets
☐ To keep up with new technological developments in archiving
☐ Other (Please describe.) _____

[QUESTION 16 (Page 15)]
What do you feel are the most compelling reasons for NOT archiving websites? (Select all that apply.)
☐ Cost
☐ Lack of administrative support
☐ Lack of trained personnel
☐ Lack of storage space for archived sites
☐ Lack of support for technology
☐ Information available on the web can be collected in other ways
☐ Other institutions are taking care of this
☐ Outside the institution's collecting scope
☐ Other (Please describe.) _____

[QUESTION 17 (Page 16)]
To the best of your memory, have you used any of the following resources for obtaining information about archiving websites? (Select all that apply.)
☐ I have not sought information regarding archiving websites.
☐ Books
☐ Journal articles
☐ Magazine articles
☐ Individual web archive websites (for instance the MINERVA project at the Library of Congress)
☐ Preserving Access to Digital Information (PADI) Initiative
☐ International Web Archiving Workshop (IWAW)
☐ International Internet Preservation Consortium (IIPC)
☐ Other websites
☐ Conference presentations
☐ Seminars or webinars

☐ Workshops
☐ Listservs
☐ Blogs
☐ Staff at other institutions
☐ Staff at your institution
☐ Other(s) (Please describe.) _____

[QUESTION 18 (Page 17)]
If you wanted to learn about archiving websites (for example: procedures, tools, examples of archives, best practices), which 5 of the following resources would you be MOST LIKELY to consult? Please rank order them from 1 to 5, with 1 being the resource you would be MOST likely to consult and 5 being the resource you would be LEAST likely to consult.

☐ Books
☐ Journal articles
☐ Magazine articles
☐ Individual web archive websites (for instance the MINERVA project at the Library of Congress)
☐ Preserving Access to Digital Information (PADI) Initiative
☐ International Web Archiving Workshop (IWAW)
☐ International Internet Preservation Consortium (IIPC)
☐ Other websites
☐ Conference presentations
☐ Seminars or webinars
☐ Workshops
☐ Listservs
☐ Blogs
☐ Staff at other institutions
☐ Staff at your institution
☐ Other(s) (Please describe.) _____

[INCENTIVE INFORMATION (Page 18)]
You have now completed all of the survey questions. Your responses have been recorded.

To enter for a chance to win one of four $25.00 Amazon.com gift certificates and/or to receive a copy of the final report, please follow the link below and enter your email address. **This address will not be associated with your survey response in any way. It will be kept confidential, and will ONLY be used to send you the final report or to notify you if you are selected to receive a gift certificate.**

At the completion of this research project, all email addresses will be deleted.

[link to email address submission page here]

If you do not wish to submit your email address, you may close your browser window now.

Appendix B: Responses to Question Ten, "Please describe the selection criteria your institution uses for choosing websites to archive."

Archive

We do not have any established criteria. For the few we have tested, it was either because the site was going to go away, or because the format is sort of an "online journal" in which once a quarter it completely changes.

It's my understanding that the university saves only top-level pages from our own site.

We are still testing so we have not established selection procedures. We have identified three sites to archive as test cases after University faculty and staff contacted us about web archiving.

Quality Assessment for making the decision to archive is / done is the following steps. / Whether or not: / 1)The website has a content of our interest. / 2)The automatic harvesting system harvested the material well enough that it's worth archiving. (We archive with the spirit of 'the more the better' for the users of the archives.) / 3)The webowner has given us a consent to archive his/her material.

We have started capturing our university and university affiliated (i.e. athletics) site(s), along with a small pilot of non-university sites reflecting our current manuscript collecting areas.

Special events, such as homecoming or faculty handbook revisions. Underdocumented university groups, such as student life and student government.

By subject/topic of website content.

Complement current collecting areas. We have only really addressed this in one collecting area. Within that one area we have much finer selection criteria (ie. type of content available, frequency of change, profile of organization within community etc).

The content on the website must fit one of our collecting areas. / We must have rights to archive the website. / The website must currently be convertable to a PDF format.

At present, our institution only archives versions of its own and subsidiary sites; we do not archive any external sites.

So far, we have collected academic department web sites and a few web sites that were at risk of disappearing. We are still developing criteria for broader harvesting.

Those identified as part of a donor's records and those with significant contribution to the university's history (with permission of the site's content rights holder).

Selection is done mostly by University Archivist. Currently, we are only archiving University-related websites...to my knowledge

Library

committee designated to test and make recommendations to select the vendor/provider of the archiving service. University dean and budget committee determine final selection

We archive a focused set of Latin American Web sites, with the archived material organized into several "collections". The scope of our primary collection is Latin American government ministry Web sites.

Unknown.

Website is up to date. The site should contain the date last updated and contact information. Sites about controversial issues should contain information about differing points of view or other sites containing that information is provided. / All information is authoritative. / Sites are updated on a regular basis for quality, validity, design, and to check to make sure the site follows the college curriculum. /

Our college archives is interested in archiving websites produced or created by the college.

We only archive our own library web pages at this time. We copy the site to CD-Rom. The College's web

site is archived by default through a system of tape backup, wherein at certain intervals (not sure what that is) older tapes are retained and not written over.

We archive University sites and are looking into archiving sites from the organizations' whose papers we maintain

historical value to the university, not replicated in another format

We archive college produced websites.

undetermined

I don't know the criteria.

Unknown

Periodic site wide archives (university, student and library websites). Daily and weekly crawls of pages that change on a daily/weekly basis

Only previous versions of the library web site itself have been archived. This has typically done during major web site redesigns. No institutional web site archives have been created.

Currently we are trying to capture key university publications, such as course catalogs, that are no longer available in hard copy but have gone to online only. We are also capturing student newspapers.

We don't really archive websites, although there are some exceptions (we keep old course help pages, etc.)

We do not archive websites

None. I thought you meant previous generations of our own library website. Which I think I will not delete off my harddrive when I retire. Save it to a disk or something...

subject librarians may on occasion select online resources and request the electronic resources librarian to add them to the catalog. However, this is not something she does eagerly or frequently because of the unstable nature of URLs. There are criteria we use to try to determine the stability and longevity of a website. The most promising ones go to the "other resources" by subject on the library website. Very rarely do they get in the catalog (some free online journals).

We do not archive websites

Our archives and special collections staff has done some website archiving of our university's websites as well as Foundation websites from across the world (because there collection area is Foundations and philanthropic studies). I do not know their selection criteria beyond this.

Primarily digital exhibits from our collections

Select items related to the primary historical record of the University (for the University Archives).

Still evolving

It is selected based on the university's records retention schedule. We are a public institution so our collecting policy is made very clear through the schedule.

We push old web pages to an archive directory on our web server - it's not accessible by anyone outside of the university.

We're working on developing criteria right now. [The remainder of this response has been deleted to preserve the anonymity of the respondent.]

The library is looking at archiving its own websites, websites created in house.

Ones that relate to our university, to our local community, and ones that are created in a larger community but of are interest to our population.

At the moment, it is part of the University Archives electronic records initiative; the same criteria apply as for other University Archives material.

All our websites internal to the libraries

We aren't yet

California and Los Angeles campaign websites

We are working this out right now but our first inclination is NOT to archive any vendor hosted content or any pages that are constantly changing in minor ways (includes budgets, org charts, policies, directories, about info, faqs and forms). / / We want to archive things that are of historical relevance or will be important for future scholarship. Our digital collections are already being archived by the Texas Digital Library. Our faculty publications and intellectual capital are being archived in our institutional repository. We are considering archiving our library calendar but suspect there's nothing there worth saving. Future design scholars may be interested in the visual and information layout of our pages but the Internet Archive is already capturing all of our design templates, if not all of our pages. / / We do have a growing collection of videos and tutorials that we are not sure what to do with, whether or not to create an archiving policy that specifically addresses them. / / / / /

those policies haven't yet been firmly established.

Collection specialists make decisions based on priorities discussed by administrators and committees.

I don't know what these are.