Archiving Web-Published Materials:

A Needs Assessment of Librarians, Researchers, and Content Providers

Kathleen R. Murray and Inga K. Hsieh

University of North Texas

Abstract

The Web-at-Risk project is a digital preservation project funded by the Library of Congress as part of the National Digital Information Infrastructure and Preservation Program. The project is developing a web archiving service to enable curators to build, store, and manage archived collections of web-published materials captured largely from US federal and state government agency web sites. In 2005 the project's 22 curators, as well as librarians and archivists working primarily in academic libraries (*N*=43), university researchers (*N*=7), and content providers (*N*=7) participated in a study to identify their needs in relation to web archiving. This paper summarizes the issues and challenges these groups face and discusses the need for collaborations among libraries and government entities for preserving web-published materials.

Citation

Murray, K. R. & Hsieh, I. K. (2008). Archiving Web-published materials: A needs assessment of librarians, researchers, and content providers. *Government Information Quarterly, 25* (1), 66-89.

Authors Note

Kathleen R. Murray is a Post-doctoral Research Associate at the University of North Texas Libraries in Denton, Texas and is the Assessment Analyst for the Web-at-Risk project.

Inga K. Hsieh is a Research Assistant at the University of North Texas Libraries in Denton, Texas.

This research was conducted as part of the Web-at-Risk project, a collaborative effort of the California Digital Library, the University of North Texas, and New York University. The project was funded in 2004 by the Library of Congress as part of the National Digital Information Infrastructure and Preservation Program to begin building a national preservation network for "at-risk digital materials of significant cultural and historical value to the nation" [http://www.digitalpreservation.gov/partners/project.html].

Archiving Web-Published Materials:

A Needs Assessment of Librarians, Researchers, and Content Providers

1. Introduction

Long-term access to government publications is an important component of our democracy and likewise is an important responsibility of government information librarians. As the Internet is becoming the primary information source for citizens, information professionals involved in the publication, dissemination, and preservation of government information are concerned about the loss of web-published materials. Additionally, limitations posed by librarians' current abilities and resources make it difficult for them to meet the challenges that accompany the preservation of web-publications. Web archiving may provide a solution for the acquisition and preservation of web-published government information and enable long-term information access for a wide-range of citizens, residents, and communities.

In order to identify the needs and issues librarians, curators, end users, and content providers have in relation to web archives, a needs assessment study was undertaken in 2005 as a part of the Web-at-Risk project,[1] a three-year collaborative research effort of the California Digital Library, the University of North Texas, and New York University. The project is developing a Web Archiving Service (WAS) to enable curators to build, store, and manage collections of web-published materials in web archives. The content of the collections will be captured largely from US federal and state government agency web sites, but will also include web-published political policy documents, campaign literature, and information related to political movements and labor unions. The project's 22 curators who will build collections of web-published materials using the WAS, as well as 43 librarians and archivists who primarily work in academic libraries, seven university researchers, and seven content providers

participated in needs assessment activities that included an online survey, focus groups, and interviews. While the purpose of this study was to identify the web archiving needs and issues of librarians, researchers, and content providers, one important outcome of the Web-at-Risk project is that the curators will create collections of web-published information that might otherwise be unavailable to future researchers and the general public.

Key concepts were defined at the outset of the study. They are briefly discussed in the next section but are also listed here for ease of reference.

Archive Service Provider

An archive service provider is an organization which offers repository services for other institutions or organizations that wish to create and preserve collections of web-published materials but do not have the infrastructure to capture, store and preserve the materials themselves.

Digital Archive

A digital archive is a collection of digital objects that may also exist in other forms. A digital archive preserves the digital formats for posterity and provides access to them.

Digital Object

Digital objects include interactive works such as video games, sensory presentations such as music, documents such as articles, and data such as statistical datasets. Two types of digital objects included in digital archives are: born-digital objects and digital surrogates, for example digitized copies of print books or audio tapes.

Web-published Materials

Web-published materials are accessed and presented via the World Wide Web. The materials include a range of material types from text documents to streaming video to interactive experiences. All web-published materials are digital objects.

Web Archive

A web archive contains web-published materials for which an organization has accepted long-term responsibility for both preservation and access.

Web Collection

A web collection typically consists of a group of related web sites but might also refer to a group of related web-published materials. In the context of this article, web collections are assumed to be preserved in a web archive.

Web Site

A web site consists of one or more web pages and other web-published materials that are generally related in some way and are often within the same domain or sub-domain name space (e.g., unt.edu or library.unt.edu).

1. 1. Web Archives

In this study web archives were viewed as a special case of digital archives, which are collections of digital objects that might also exist in physical formats. A digital archive preserves digital formats for posterity and provides access to them. Web archives contain web-published materials for which an organization has accepted long-term responsibility for both preservation and access. Web-published materials are any materials accessed and presented via the World Wide Web. These materials include a range of material types from text documents to streaming

video to interactive experiences. Organizations may build and manage web archives themselves or may enter into service arrangements with third-party archive providers or archive agencies.

Within a web archive, a web collection would typically consist of a group of related web-sites; however, a web collection might also be comprised of a group of related web-published materials, such as a series of discrete born-digital government documents. Librarians and archivists apply the intellectual and logical processes involved in collection management to create web collections for preservation in web archives.

Day states the case for undertaking the preservation in web archives of materials accessible via the World Wide Web as follows:

> In the short time since its invention, the World Wide Web has become a vital means of facilitating global communication and an important medium for scientific communication, publishing, e-commerce, and much else. The 'fluid' nature of the Web, however, means that pages or entire sites frequently change or disappear, often without leaving any trace. In order to help counter this change and decay, Web archiving initiatives are required to help preserve the informational, cultural and evidential value of the World Wide Web (or particular subsets of it).[2]

Since Day made these statements in 2003, several web archiving initiatives have begun. In many cases, national libraries have taken the lead in promoting projects directed at the preservation of their national cultural heritage and their government's web-publications. In the United States, the Library of Congress has undertaken several web capture initiatives pursuant to its mission to preserve "the nation's cultural artifacts" and provide "enduring access to them." [3]

As one aspect of their nation's preservation effort, the Preserving Access to Digital Information (PADI) initiative of the National Library of Australia maintains a web archiving

resource list that identifies both major web archiving projects throughout the world and web archiving collaborations among national libraries.[4] Additionally the web sites of both the International Internet Preservation Consortium (IIPC)[5] and the International Web Archiving Workshops (IWAW)[6] contain papers related to many of the technical challenges encountered in archiving the Web as well as information about the web archiving initiatives of both national libraries and researchers interested in preservation of the Web.

1. 2 Approaches to Archiving the Web

In his review of web archiving initiatives, Day[7] identified two approaches to capturing web-published materials in a web archive: a selective approach and a harvesting-based approach. In addition to these two approaches, the National Library of Australia identifies a deposit approach. They also describe a variation of the selective approach, which they identify as a thematic approach.[8] Both Day and the National Library of Australia identify a fourth approach in which two or more of the other approaches are combined.

Whole Domain Harvesting

A whole domain or comprehensive capture is employed to capture and preserve national or global web space. For a nation, this approach uses web crawlers to capture (a) the country's geographic Internet domain, such as Sweden's .se domain, as well as (b) the nation's additional web resources that are published in other Internet domains, such as the .com domain. The Internet Archive[9] uses a comprehensive capture approach in its effort to capture and archive snapshots of the global Internet. Examples of whole domain harvesting at national and global levels respectively are the National Library of Sweden's Kulturarw3 Web Archive[10] and the Internet Archive's WayBack Machine.[11]

Selective

Selective web captures seek to preserve specific portions of the Web or specific resources in accord with a set of predefined criteria or parameters for materials in collections. PANDORA, Australia's Web Archive, is an example of a selective approach.[12] Thematic web captures are a variation on the selective capture approach that preserve specific content identified by a curator as relating to a particular theme or event, such as a state's gubernatorial elections or an environmental disaster. The Library of Congress Web Archiving Project, MINERVA, is an example of a thematic approach to web archiving.[13]

Deposit

Deposit approaches preserve materials deposited in a web archive by publishers based either on government-mandated depository requirements or voluntary arrangements. For example, a commercial publisher of academic journals might deposit their published content with an archive agency such as a national library. Because content publishers play an active role in this approach to web archiving, it offers a potential solution to capturing "deep web" content, which cannot typically be captured by web crawlers. The Electronic Collection of the Library and Archives Canada is an example of a deposit approach to web archiving.[14]

Combined

A summary report of the 2005 5th International Web Archiving Workshop (IWAW), states that "the clear distinction between the two archetypal web archiving strategies of exhaustive harvesting and selective collection appears to be fading."[15] Exhaustive or whole domain harvesting initiatives are honing crawler technology to capture event-driven and thematic web sites while manual selection initiatives are seeking ways to automate their labor-intensive processes. The National Library of Australia notes that "a growing number of Web archiving

initiatives are concluding that no one archiving model is entirely satisfactory for preserving national online heritage."[16] Supporting this view in her presentation at the 2006 Digital Preservation Coalition Forum on Web Archiving, Lupovici stated that "broad extensive harvesting, focused intensive selection and harvesting, and deep web deposit are complementary techniques"[17] required for successful domain-level archiving of the Web.

Thus a web archive preserves a range of web-published materials that are captured using different approaches. For example, a web archive's content may consist of a collection of thematically-related web sites that were identified by a curator and captured by a web crawler. The archive may also include discrete web-published materials such as web-born documents that were electronically deposited by their creators or publishers.

1.3 Changing Times for Web-Published Government Information

Government information has undergone many changes in its format, distribution, and ownership enabled in large part by capabilities inherent in the Internet. In response to these changes, the Office of Innovation and New Technology of the United States Government Printing Office (GPO) is overhauling the technical infrastructure for deposit, distribution, and archive of federal government publications that are within the scope of the Federal Depository Library Program (FDLP). GPO is developing a content management system called the Future Digital System or FDsys.[18] Deployment of the system is planned for December of 2007. "Included in the FDsys will be all known Federal Government documents within the scope of GPO's Federal Depository Library Program (FDLP), whether printed or born digital."[19]

In a related effort, GPO conducted web harvesting trials of Environmental Protection Agency (EPA) web sites in 2006 to identify and capture the rising number of government publications that are published on EPA web sites but are not reported to GPO for inclusion in the

FDLP and GPO's Cataloging and Indexing Program (i.e., "fugitive documents").[20, 21] After these web sites were captured, data mining applications attempted to identify EPA publications that were within the scope of the FDLP. Comparisons of these publication results with those in the GPO catalog will hopefully result in the identification of fugitive documents that can subsequently be cataloged and included in the FDLP. The results of these and future trials will help inform the web harvesting components of the FDsys. Successful identification and capture of fugitive documents combined with standard deposit requirements for federal documents will contribute to actualizing the strategic vision for the FDsys, which is to "allow federal Content Originators to easily create and submit content that can then be preserved, authenticated, managed and delivered upon request."[22]

As with their counterparts at the federal level, state agencies are increasingly embracing web publishing and experiencing many digital preservation challenges. The Library of Congress, under the auspices of the National Digital Information Infrastructure Preservation Program (NDIIPP), is leading the effort to develop a national strategy for the preservation of digital content. One goal of the NDIIPP is to identify the preservation needs of key government information stakeholders. In 2005 NDIIPP conducted a series of workshops with state government representatives that resulted in the identification of a number of issues state libraries and state archives are facing as they grapple with capturing and preserving the digital information published by state government agencies.[23] These issues, as well as strategies for addressing them, were organized under five basic themes:

1. Identifying significant digital information
2. Learning by doing – with some help (i.e., Leading preservation initiatives that are informed by best practices)

3. Operational and technical infrastructure
4. Breaking silos: Communication, collaboration, and partnerships
5. Resources: Funding, personnel, mandates

Building on the experience of the NDIIPP workshops with the states, 30 states sent representatives to participate in an information and best practice exchange forum about digital preservation in state government held in March of 2006 and hosted by the State Library of North Carolina and the North Carolina State Archives.[24, 25] Participants thought that a strong state-wide foundation for a digital preservation program was the first order of business that state governments needed to address. The many issues and challenges discussed by participants were organized into four core components of a foundation for a state-wide digital preservation program: support and buy-in from stakeholders; "good enough" practices implemented now; collaborations and partnerships; and documentation for policies, procedures, and standards.

1.4 Assessing Curators' Needs

Government information collections in libraries are increasingly becoming virtual collections as they are of necessity extending beyond the boundaries of individual libraries' local ownership and possession. Likewise, changes in the publication and distribution of government information have predicated changes in the scope of responsibility and requisite skills involved in the roles of librarians who are responsible for government information and government documents collections within academic libraries. Shuler characterizes an emerging role of the academic librarian as one of a *curator* and suggests this role is analogous to the role of a lawyer in terms of pulling together information from distributed sources and in a variety of formats to address end users' information needs.[26] In the future, curators possessing subject expertise and

acquiring requisite technical skills may serve a vital information resource role both for citizens and the wider patron communities they serve.

The Web Archiving Service (WAS) being prototyped by the Web-at-Risk project is reaching out to a group of curators working in state and academic libraries to inform the design of a set of web archiving tools that will assist curators in their web collection management responsibilities. While using the same underlying web harvesting technology employed by large-scale, whole domain web archiving initiatives, this service plans to develop tools that will enable curators to create thematic and event-driven collections of selected web sites. Prior to full-scale development of the WAS, the Web-at-Risk needs assessment study sought to understand how librarians and archivists were meeting the challenges posed by web-published materials. A primary objective of the study was to identify the major needs and issues confronting librarians, archivists, content providers, and researchers in the interim between the formidable challenges currently posed by changes in the publication and distribution of information and the emerging solutions at the international, national, and state levels. It was anticipated that the results of this study could inform not only the development of the WAS but also provide insight into the collection management process for web collections.

2. Methodology

Three types of assessment activities were conducted in 2005. These activities included a survey of the curators participating in the Web-at-Risk project, a series of focus groups with librarians and archivists, many of whom worked in academic libraries as government information librarians, and individual interviews with academic researchers and content providers, several of whom published state government information.

2.1 Framework

Web collection development includes three major phases: selection, curation, and preservation. By breaking down collection development into a series of activities within each phase, the functional view shown in Figure 1 emerges. Librarians will recognize the activities as those commonly employed in traditional collection planning. Although there is a general trajectory of activities in the web collection development process, with selection occurring early on and preservation toward the end, tasks are not necessarily completed in a linear fashion.

PHASES		
SELECTION ⇨	CURATION ⇨	PRESERVATION
Selection	Description	Preservation
Acquisition	Organization	
	Presentation	
	Maintenance	
	Deselection	

Figure 1. Web Collection Development Framework

2.2. Survey of Curators

The survey served two purposes: (a) to identify end user and curator needs that might impact collection development for web-published materials and (b) to identify functional requirements for the crawler and curators' tools being developed for the project's Web Archiving Service. Survey respondents were the 22 curatorial partners involved in the Web-at-Risk project at the time the survey was conducted.

The online needs assessment survey instrument consisted of 58 questions divided into five sections. The first section gathered information about the curators' backgrounds and

experiences. The following three sections identified curators' needs related to preservation and collection policies and the web collection development phases and activities identified in Figure 1. The fifth section of the survey gathered requirements for a curator interface for a Web archiving service.

In all, 16 surveys were submitted. Ten curators submitted individual surveys while 12 curators submitted a total of six surveys, each of which represented a joint effort between two curators. Four of the surveys were submitted by curators with collection responsibilities in the areas of public policy or political movements. The remaining 12 were submitted by curators with collection responsibilities in local, state, federal, or international government information. Questions in each section of the survey were analyzed individually. Where appropriate, response sets were eliminated from the analysis. For the most part, descriptive statistics were used to analyze the data.

2.3. Focus Groups with Librarians and Archivists

Five focus groups were held in the summer and fall of 2005. Two focus groups were held during national conferences for two organizations, the American Library Association conference in Chicago in June 2005 and the Federal Depository Library Conference (FDLC) in Washington, DC in October 2005. The three remaining focus groups were held at each of the three project partner institutions (New York University, the California Digital Library, and the University of North Texas). The purpose of the focus groups was to elicit the needs and issues librarians, curators, and end users have in relation to web archives.

A total of 43 people participated in the five groups. The majority (n = 39) worked in colleges or universities and 33% (n = 14) held library management positions (e.g., Department

Heads). About 25% (n = 11) of the participants indicated they had some prior experience creating web archives.

Participants in the Chicago focus group consisted of librarians working in diverse disciplines in academic institutions and three archivists from non-profit organizations. Participants in the Washington DC focus group all worked in government documents departments within academic libraries. Participants in the three project partner focus groups consisted largely of individuals with collection development or subject selection responsibilities for a variety of departments within their respective university libraries, including government information.

Each of the groups was facilitated by the Assessment Analyst for the Web-at-Risk project. Group discussions were generally one and one-half hours in length. After the participants introduced themselves, meetings generally began with a discussion of participants' needs and issues relative to the selection of web sites to archive and proceeded through the discussion topics, which corresponded to the typical phases and activities involved in collection development. At the conclusion of the group, participants completed a written questionnaire and were given a thank-you gift. Two note-takers attended each focus group and created a record of the discussion as well as a summary of key points. With the exception of the Chicago group, discussions were recorded and transcribed.

Collection development provided the umbrella framework for analyzing the focus group transcripts and notes. Based on a pilot group discussion in May of 2005 with curators involved with the Web-at-Risk project, an initial categorization of concerns and issues within each collection development phase was created. These categories were used to analyze the content of the first focus group. Additional categories were added as necessary. This process was repeated

for each of the four focus groups that followed. Working independently, two analysts categorized the contents of both the transcript and notes for each focus group. Discrepancies between the analysts were discussed and resolved.

2.4 End User and Content Provider Interviews

Project team members at each of the three partner institutions interviewed both end users and content providers. The purpose of the interviews was to elicit the needs and issues of end users and content providers in relation to web archives.

Seven end user interviews were conducted with academics: four with historians, two with political scientists, and one with a professor of hospitality law and management. With the exception of one person, all of the researchers used web-published materials in both their research and professional activities, although the extent of their usage varied widely. For some, web-published materials were more likely to be used in their teaching and for others, in their research or professional activities.

Seven content provider interviews were conducted: three with representatives of union organizations and four with representatives of state government agencies or state government-sponsored programs. The unions had existing relationships with a university archive for the preservation of their print materials and, in two cases, these relationships extended over many years. One state government agency had an existing relationship with an archive for the long-term preservation of its major web-publication. Most representatives of state government agencies were sensitive to the issues involved in archiving web-published government information and many were aware of their web sites already being crawled and captured.

Interviewers used either end user or content provider questionnaires to guide their discussions. Table 1 lists the discussion topics.

Table 1.

Interview Topics

End Users	Content Providers
Selection of materials for an archive	Web-published materials
Authenticity of archived materials	Digital archives
Interacting with materials in an archive	Access to materials
Searching an archive	Authenticity of archived materials
Preservation of archived materials	Intellectual property of archived materials
	Agreements with archive providers

Interviewers summarized the individual discussions they conducted and identified the key points that emerged. The summaries were provided to the project's Assessment Analyst, who further analyzed the content and identified the basic themes and issues for both the end users and the content providers. Three questions in the end user interview guide asked participants to select a response from a range of values (i.e., number of years or level of importance) that best matched their opinions. For each of the three questions, weighted sums were calculated to rank overall responses.

3. Findings

Subsequent to the individual analysis of data for each assessment activity, all the findings were combined and organized into four areas: (a) challenges in the current environment, (b) organizational issues, (c) collection development concerns, and (d) needs addressed by web archives.[27] Key findings in each of these areas are reported in the remainder of this section.

3.1. The Current Climate

Librarians are facing many challenges as they continue to work in the familiar world of print materials while increasingly accepting responsibilities in the ever-growing world of web-published materials. While interested in embracing the challenges inherent in web-published materials, librarians often lack the appropriate technical expertise, the required resources, or both. Most acknowledge that collection development models for print materials transfer only at great expense to web-published materials, which are expensive to select, capture, and catalog. In a climate of uncertainty and funding constraints, university libraries find the scope of the preservation effort beyond the capabilities of their libraries' IT infrastructures and staffs.

Librarians generally agree that the organization or individual responsible for producing web-published materials ought to take responsibility for preserving them. In practice, however, librarians perceive content producers either as unaware of the need to preserve their web-published materials or as unable or unwilling to accept the challenge. Libraries have traditionally accepted preservation responsibility for print publications, but they lack the resources to extend this practice to web-published materials. On the other hand, most content providers interviewed share a view of a web archive as a safe repository for specific web-published materials of historical value that are beyond the purview of providers' own retention mandates or beyond their resource ability to preserve.

With the continuing shift from print documents to web-published materials, some major research libraries are not certain they can either wait for or rely solely upon federal government preservation efforts. Additionally, librarians express concerns regarding the sustainability of government programs in future funding cycles. This uncertainty drives these libraries to assess their need for local preservation programs.

Responsibility for preserving web-published state government publications is often unclear or non-existent and many of these publications are simply disappearing. With web-born government publications, a specific edition cannot be relied upon to be a constant entity. Many focus group participants encountered instances of web-born government publications that were altered and for which no indication of the alteration was evident either in a versioning scheme for the publication or in the creation/modification date.

State libraries are in a logical position to preserve state government publications but are often understaffed and resource-constrained, resulting in hit-and-miss efforts in regard to preserving the web-published materials of state agencies. The concern for preservation and access to web-published materials of federal and state agencies also extends to local government entities, whose need for assistance in preservation of their web-published materials is quite high.

3.2. Organizational Issues

By means of a questionnaire completed immediately after each focus group discussion, participants identified the major hurdles they envision for their library or organization in creating a web archive. The four major hurdles are technology, policies, management commitment, and funding. These hurdles emerged repeatedly in focus group discussions and were echoed in the survey responses of the Web-at-Risk project curators. In particular, the curators estimated the magnitude of the funding challenge they anticipate in creating archived collections of web-published materials. The top four funding challenges were in the areas of cataloging, preservation, IT support, and staff training.

Survey respondents also reported that it is difficult to attract and sustain management interest in digital archiving projects and therefore difficult to get staff allocated for them. It appears that a change of focus for web archiving endeavors, from a project orientation to an

organizational and consortial orientation, is necessary to attract and sustain the required resources and funding.

Both formal and informal marketing efforts to promote the concept of web archives or institutional repositories are underway within some institutions. These efforts aim to secure senior management endorsement for preservation functions within the organization or institution as well as to gain commitment to the funding required to move preservation staff and infrastructure into the core operations of the institution. Librarians recognize the need to develop their preservation case as a business case, to identify its risks, costs, and benefits. Additionally, it is understood that effective business cases for preservation efforts need to include a model for sustainability and collaboration. In this regard there is a need for consortial efforts among libraries and collaborations or partnerships between libraries and government agencies, certainly at the federal agency level where preservation efforts are underway, but especially at regional, state, and local levels.

3.3. Collection Development Concerns

While collection development activities for web-published materials conceptually parallel activities for print materials, most librarians find they are more labor-intensive. In particular the activities of selection and acquisition require more up-front work and often involve individual review of materials. These activities are especially challenging in collection development for less-established disciplines and areas of study for which web-published materials often represent the bulk of available information. Application of metadata to collected web-published materials is also challenging and often requires specialized expertise.

Selection

Identifying what to preserve is a major issue for most librarians. The two basic questions they ask in regard to identifying web-published materials for preservation are: "Should *we* save this?" and "Is *someone else* already saving it?" Overall, the important materials targeted for preservation by librarians fall into four categories:

1. Government information originating at the national, state, regional, and local levels
2. Information supporting the research, teaching, and mission of academic institutions
3. Information pertaining to key events
4. Information pertaining to or produced by organizations

Librarians identified the following materials as currently falling through the cracks of preservation programs: small journals, state and local government publications, and institutional web-published materials. The sense was that these types of publishers did not have the historical models and/or the financial resources to commit to preservation.

Unit of Selection: Content v. Context

Unit of selection refers to the granularity curators might specify for the capture of web-published materials. For certain research disciplines or types of research, the context of source material is critically important and therefore the entire web site would be the unit of selection. To illustrate the importance of capturing context, one historian made an analogy between a web site and a newspaper and observed that placement of materials on a web site has meaning in much the same way that placement of an article in a newspaper has meaning. For other disciplines the original web-context of the source materials is not always critical and users would be better served by interacting directly with collections comprised of captured documents, reports, or datasets. For example, citizens and researchers seeking to identify year-over-year differences in

a particular government agency report would be well-served by ready access to the reports themselves rather than accessing them from within an annual capture of the agency's web-site.

Acquisition

All participants were generally concerned with the frequency with which web-published materials change. Survey respondents identified three important considerations for collection building practices:

1. Assessing the change rate of the source materials
2. Establishing the interval at which collection materials will be recaptured
3. Articulating criteria for retention of earlier versions

In addition, some participants familiar with creating collections of web sites recommended curators evaluate material types and formats in web sites prior to acquisition. If done manually, this can be a daunting endeavor; automated tools are needed to support this type of analysis.

Authenticity

While each user may assess authenticity differently, many users need and most would want some authority to provide an assurance of the authenticity of web-published materials in a web archive. Survey respondents were concerned that multiple versions of source materials captured at different points in time and multiple formats of the same object might pose a threat to the authenticity of those materials. Amplifying this concern, focus group participants indicated that assigning versions and dates to captured web-published materials is a critical area a web archive should address. Additionally, many researchers would like an archive to identify the location of original source materials. It is clear that an effective archive must have policies and practices in place to address these authenticity concerns. Likewise, archives containing copies of government documents must have a means to authenticate the reliability of source documents.

Metadata

Survey respondents identified cataloging as the top financial and technical challenge they anticipate in regard to building web archives. Likewise, librarians participating in focus groups anticipated that in creating collections of web-published materials, the biggest challenge will be the application of metadata. Librarians recognize that evaluating web-published materials and applying metadata requires a specialized skill set. Focus group participants reported their libraries currently do not have enough catalogers to handle their non-web-published materials and most thought that automated metadata generation, including subject or topic classification, would be needed for web-published materials.

In lieu of hand-crafting metadata for web-published materials, some librarians suggested that "indicators of usefulness" could be included in search results. Such indicators could be automatically generated using technology to gather and analyze users' assessments of archived materials. Any user could utilize the indicators as evaluative tools for material selection.

Organization

Librarians anticipated users would expect a web archive to provide both full-text search capability as well as the ability to search the archive by subject categories. In fact, the researchers interviewed did indicate the most important types of searches to them are "topic or subject" and "full-text using any keyword." Librarians also thought it would be important to "provide some higher-level topical access, even if it is derived from the title as opposed to the actual content." Supporting this level of organization and access, researchers indicated they would like to browse a web archive via a subject directory structure.

Presentation: Look-and-Feel

For some content providers, their databases and datasets are the meat of their content and to varying extents all other content on their web sites is superfluous. These content providers do not think that replication of their web sites' "look-and-feel" is important when archived materials are presented. In a similar vein, librarians agreed that preserving the content of journal articles, versus their look-and-feel from the original publication, would suffice. However, many participants thought other types of materials would need to be presented in their original web context. This was of particular importance for historical research in many disciplines. For some librarians and researchers, web sites in an archive were basically viewed as historical records and, as such, the librarians and researchers thought that the archived web sites should be presented in such a way that they mirror the source web sites.

Presentation: Authenticity Indication

Researchers asserted that web archives should make it clear that users are interacting with archived material and not "live" material. For certain types of research purposes, a web archive must also be able to provide and present some assurance that what users are seeing is "official" information. Content providers were concerned about how an archive might represent itself; archived web sites need a statement identifying the archive as an "official" or an "unofficial" version of the source materials.

In legal research, a designation of authenticity for archived materials is critical. For maps and for GIS data pertaining to environmental or natural resources and agricultural reports, both an indication of authenticity as well as the date(s) for which the captured information was relevant is critical.

Preservation

Survey respondents suggested that a significant danger when dealing with the capture of web-published materials is data corruption. Corrupt data is of no value to users and the archiving agency needs to be capable of validating the content subsequent to its capture.

3.4 Needs Addressed by Web Archives

Librarians who participated in the focus groups were asked in a questionnaire to identify the top three user needs web archives could address at their institutions or organizations. Their responses (N = 80) are illustrated in Figure 2.

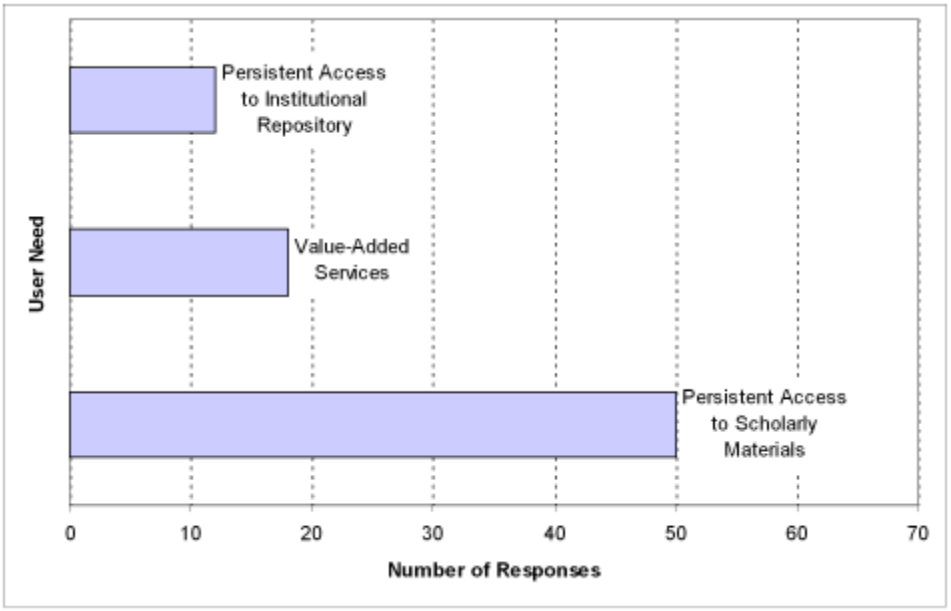

Figure 2. User Needs Addressed by a Web Archive

The most important need librarians identified was persistent access to the information users need for teaching and research. The participants also identified two additional needs a web archive could address: provision of value-added information services, such as aggregation of

content from disparate sources, and persistent access to the institution's history and intellectual products in an institutional repository.

Persistent Access to Scholarly Materials

By far, persistent access to a wide range of digital or web-born scholarly materials for research and reference is the primary user need librarians thought web archives could address. Many thought web-published materials being considered for inclusion in web archives should first be evaluated for consistency with an institution's established collection parameters. Others recommended that web archives give priority to key content that is published solely in electronic format as well as to born-digital materials from non-traditional publishers. Several librarians identified web sites and web-published material about contemporary social movements as particularly important reference materials needed by users.

Another type of materials specifically identified for persistent user access is information and materials characterized as fleeting, ephemeral, non-standard, not previously published, or not commercially available such as scholarly materials from the institution's own researchers and research centers. Likewise, government publications characterized as critical, web-born, endangered, or fugitive were identified as candidates for inclusion in a web archive. These publications include web-born government documents published independently from the issuing agencies and those not included in the FDLC program. Finally, materials characterized generally as historical records were identified as important to retain in a web archive in order to enable historical research of web-based information and to provide a safeguard for unanticipated future research needs.

Provision of Value-Added Services

Librarians thought web archives could also provide users with value-added services. One such service could be the organization of archived content based on subject concentrations and historical timelines. Another service could be access to focused collections derived from diverse sources. For government information, librarians thought it would be a significant benefit to provide citizens and other patrons with subject-centric, cross-agency access to archived materials in addition to agency-centric, cross-subject access. Participants also felt that a web archive could add value to users by providing a friendly design for content discovery and access including searchable content, authenticity indicators, and version control.

Persistent Access to Institutional Repositories

Focus group participants indicated that web archives that provided persistent access to an institutional repository would meet user's information needs as well as preserve the institution's historical record for future researchers. The following types of materials might be of value to include in the repository: university and library web pages of long-term significance, faculty published scholarship, and faculty research papers.

Additional User Needs

The researchers interviewed in this study also identified user needs that a web archive might address. In addition to adding an indication of authenticity to materials, the researchers suggested it would be of value if an archive provided: (a) descriptions of both the provenance of materials and the material preservation activities undertaken, (b) descriptive tagging of inactive links, and (c) web site maps to enable virtual reconstruction of the original web sites in support of future research.

4. Discussion

4.1 Partnerships

In several cases librarians reported that budget constraints are forcing their libraries to cut expenses and realign budgets. At the same time, libraries and archives are responding to an urgent and growing need to collect and preserve web-published materials, an effort that stresses their existing resources and an effort they acknowledge cannot be addressed without partnering both with other departments within their organizations and with external organizations.

Regarding the cost for preservation of digital materials Rusbridge comments that "the trouble is, it is a new cost, and we have not worked out how to factor it into our budgeting and business models."[28] This observation certainly applies to the costs for preservation of web-published materials, which are themselves a subset of digital materials. Collaborative partnerships could help address these costs. Findings from this needs assessment suggested models for both external and internal partnership opportunities.

External Partnerships

There is an opportunity at the state level for collaboration among state agencies, the state library, and university libraries. When university library collections include state government publications, the preservation roles and responsibilities among government agencies and the university library need to be determined. University librarians encounter confusion at some state agencies in terms of identifying who is responsible for publication and preservation of web-published materials. While state libraries are uniquely positioned to undertake preservation of materials created by state agencies, they are often severely understaffed and unable to meet preservation demands.

Similar problems regarding responsibility for preservation exist among regional and local government entities as well as civic organizations. In addition to the issue of preservation responsibility, these organizations often lack both the necessary expertise and the infrastructure to support preservation programs. These same issues confront smaller academic institutions and smaller publishing houses.

The needs of these organizations may offer larger university libraries opportunities to form external partnerships for the preservation of the web-published materials created by state and regional government agencies and smaller government entities, institutions, and publishers. A high-level diagram for external partnerships is depicted in Figure 3.

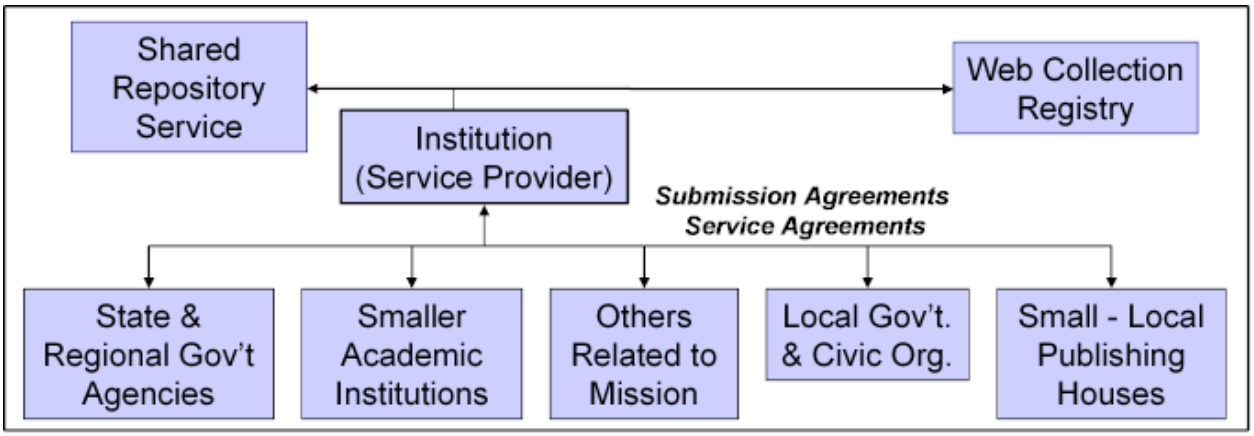

Figure 3. External Partnership Model

This partnership model involves forming a community of creators among a group of partner organizations. These creators produce a range of web-published materials, which might include web sites, discrete text publications, maps, or other materials. A large research institution could act as a service provider, offering repository services to partner organizations and including their collections of web-published materials in a web collection registry. The registry

would enable resource discovery for the materials in the shared repository. The research institution might offer its partners a range of other services, such as the provision of metadata standards and tools for partners to create metadata records for their materials. Additionally, the institution and its partners could establish service agreements that stipulate the terms and conditions of any services provided to the partners by the institution. Submission agreements would address the roles and responsibilities of all partners regarding any materials deposited in the repository.

A primary motivation for institutions to create such partnerships is to preserve the web-published materials these organizations create. Additional benefits of such partnerships to an institution include:

1. Fostering a sustainable business model to address resource requirements for long-term preservation
2. Promoting a commitment to stewardship among creators and publishers of web-published materials
3. Preserving historical records in areas of interest to the institution and its user community
4. Fulfilling the institution's mission to serve the community in which it is situated

Likewise, content providers partnering with institutions can ensure that materials of importance to their user communities remain accessible. This might be particularly true at the local government level, where web archiving resources and expertise might not be available. In many locations, both citizens and researchers might benefit from their local governments forming such external partnerships. For larger organizations and institutions, internal partnerships offer additional collaborative opportunities and benefits.

Internal Partnerships

To address the changing roles and responsibilities within an organization or institution in regard to the preservation of the organization's history, publications, scholarship, and intellectual products, internal partnerships are needed. Such partnerships need backing from top management and support from key stakeholders within an organization. For universities, backing is needed from university administrations and support is required from key stakeholders, which will likely include the IT department, faculty, library administration, and research center directors. Figure 4 depicts a model for internal partnerships.

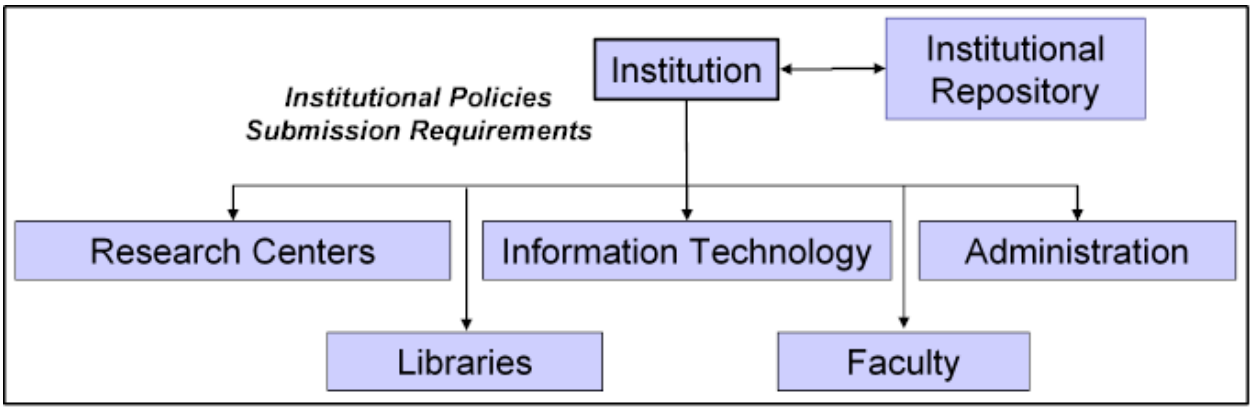

Figure 4. Internal Partnership Model

In this model, institutional policies in support of preservation could be developed to guide the effort and ensure that the organization as a whole moves forward in concert. For such a partnership, an institutional repository would be created and appropriate roles and responsibilities for all stakeholders would be identified. These would tap into and leverage the unique expertise each stakeholder group can contribute to the overall preservation effort. For a university the intellectual products would typically be provided by research centers and faculty. Other materials and records related to the history of the institution might be provided by

administrative staff. Submission requirements, including metadata requirements, would be established and tools would be developed to enable creators to deposit their web-published materials in the institutional repository. Curatorial and preservation responsibilities would be shared by librarians and IT staff.

University libraries are motivated to form internal partnerships because they recognize the web-published materials constituting the history and intellectual products of their institutions are being lost and they know the library cannot preserve these materials on its own. The expected benefits from such partnerships include:

1. Fulfilling the core mission of the institution
2. Preserving a record of the institution's history
3. Fostering a sustainable preservation model
4. Extending stewardship for intellectual products to creators
5. Leveraging preservation staff resources within the organization
6. Building a platform to promote increased visibility for the university

Successful internal partnerships involve collaboration among key stakeholders within an organization or institution as well as endorsement from department heads or senior managers. For many organizations, these criteria for success predicate changes and challenges to their organizational culture. However, the benefits of such partnerships to the organization as a whole suggest that the effort is worthwhile.

4.2 Registry Service for Web Collections

A question many of the librarians who participated in this research wanted to be able to answer was: Is some organization already archiving these materials in a manner that meets the needs of my patrons? Clearly it would be of value to create a shared directory or registry service for web collections. Small and medium institutions generally do not have the resources to engage in web preservation activities yet they could benefit from the preservation work of larger institutions. A registration service would help larger universities eliminate duplication of effort in the preservation of web-published materials and maximize the preservation efforts of their own scarce resources.

A registry of web collections might be an extension of existing consortial efforts among libraries. Bibliographic registries, such as OCLC and RLIN, are examples of existing shared cataloging services libraries use to locate materials in other libraries. The Digital Library Federation has defined the "need for and the requirements of a service that registers the existence of persistent digitally reformatted and born digital monograph and serial publications."[29] This registry service for digital masters suggests inclusion of the following information for each digital master in the registry.

1. Which library has the material
2. Format of the material
3. Terms of use
4. Library or institution responsible for preserving the original source material
5. Library or institution taking responsibility for preserving the electronic copy

Two registry efforts are underway for digitization projects dealing with government documents: the GPO Registry of U.S. Government Publication Digitization Projects and the

American Library Association Government Documents Round Table (GODORT) Clearinghouse of Government Documents Digital Projects. The GPO registry "contains records for projects that include digitized copies of publications originating from the U.S. Government" [30] and the GODORT registry "provides information to librarians and others about digitization projects for local, state, federal, and international government documents." [31]

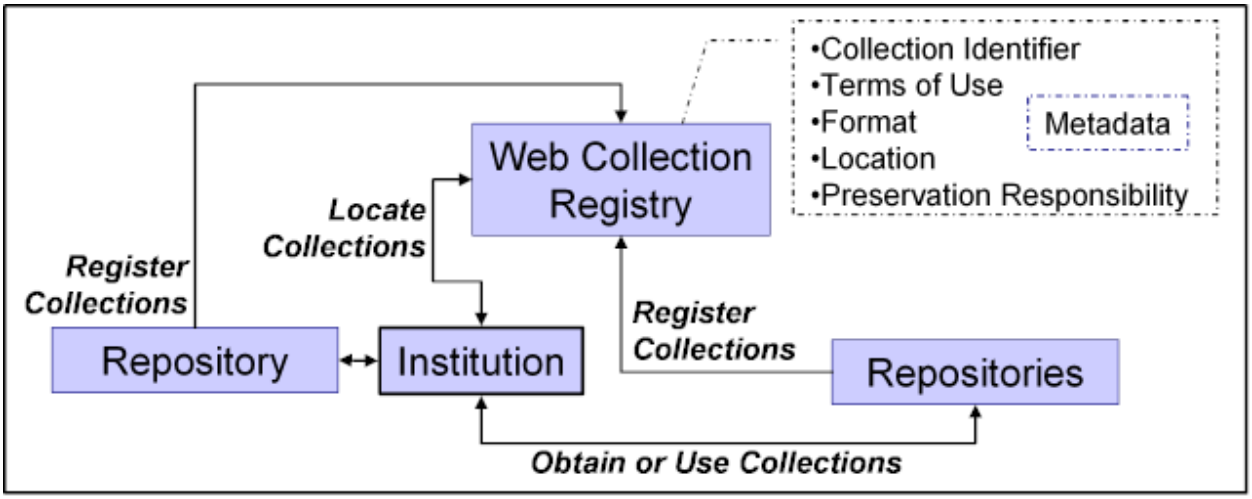

Figure 5. Registry Service Model for Web Collections

A model of a registry service for web collections is depicted in Figure 5. The registry itself would contain standardized metadata records describing web collections. These records would be submitted by the organization or institution that had accepted preservation responsibility for the materials in each collection. Typically the organization would store the collections themselves in either an institutional repository or a shared repository. Organizations with a need for collections of web-published materials could check the registry to see if some other organization had already preserved the materials. Organizations or institutions could locate

collections within the registry and either acquire copies or arrange to access the collections in other repositories.

Registry services for web collections provide an answer to the librarians' quest to know if some other organization is already preserving a particular collection of web-published materials. The benefits of such a registry service for libraries include expanding access to materials, eliminating redundancy of effort, and controlling preservation costs.

4. 3 Preservation Applications and Mandatory Deposits

Understanding the enormity of the preservation task for web-published materials, some participants in this study suggested ideas that sought to drive responsibility for preservation to the creators of web-published materials. Doing so has the advantage of many hands doing preservation work that appears as if it might not otherwise get done in its entirety. To some degree, these suggestions are already being implemented. Fundamentally, these suggestions require preservation features to be incorporated in application software and preservation requirements to be codified in organizational and funding policies.

For individual files, a method for the acquisition of all versions of user-created files was envisioned. This involves integrating a background *Save in Repository* feature as part of the typical *Save* functionality found in common application software such as word processors. If mandated by organizational policy or implemented by default in organizational software installations, this functionality, when applied to works-in-progress by their creators, would ensure all versions of files were captured in an institutional repository with little effort on the part of creators. Libraries could then offer a value-added service that essentially tracked and provided access to the various versions of these files in the institutional repository.

A second idea that emerged in the study was the development of web site packager applications. In addition to packaging a web site(s) and supporting files for submission to an institutional repository, the application would also include preservation features. These applications might combine existing functionality in web site creation applications, such as the ability to analyze web sites in order to identify working links, outline directory structures, and list files by size and type, with functionality to capture content external to the web site.

Web site managers would be responsible for packaging the web site and for determining the extent of the internal and external links that would be packaged. Features to add metadata and copyright information would be incorporated. Additionally packaged materials might include annotations resulting from deactivation of email links and hyperlinks, provenance information, and/or authenticity certifications. The final submission package would include data (the web site and its related content and code) and information about the data (metadata, rights data, and provenance). Some individuals or smaller groups within an organization might need services that analyze their web sites and individuals who consult with them about the results of the analysis prior to packaging web sites for submission to a repository.

It was also suggested that funding agencies include a funding prerequisite that recipients identify a preservation plan for any digital and web-published materials resulting from funded projects. There is a model for this in the United Kingdom. The Arts and Humanities Data Service (AHDS) receives digital material deposits from grant recipients as mandated by a number of funding organizations. Digital resource deposits are mandated as follows:

> If you have received a grant from the AHRC [Arts and Humanities Research Council] or the British Academy it will be a condition of the award that you offer relevant data and documentation for deposit with the AHDS. If you have received

> a research grant from the Carnegie Trust, the Council for British Archaeology (CBA), the Economic and Social Research Council (ESRC), the Leverhulme Trust, the Natural Environment Research Council (NERC), or the Wellcome Trust's History of Medicine Programme you are either required or recommended to offer relevant data for deposit with the AHDS. [32]

Combining a deposit mandate codified in funding policies and organizational policies with preservation features in software applications for creators and preservation packaging applications for web site managers, grassroots preservation is enabled within organizations. As some librarians suggested, if it's mandated and easy to do there's a chance of success.

5. Conclusion

At a time when both universities and state governments are targeting their libraries for downsizing and elimination, libraries are confronting an urgent and growing user information need for the collection and preservation of web-published materials. Substantial funding requirements for preservation resources and infrastructure exist as do significant organizational challenges. For government information librarians, the planned systems from GPO for federal government information cannot come too soon and the preservation foundations being articulated by state governments suggest movement in a positive direction. Despite these positive developments for web-published and web-born government information, librarians still face substantial hurdles to creating and preserving collections of web-published materials that fall outside the scope of these developments but fall solidly within the information needs and requirements of their patrons.

The information needs of citizens and library patrons in general provide a strong basis for articulating the benefits of creating web archives and institutional repositories as well as for

identifying the risks to an institution of not preserving web-published materials of importance to users. Creating business cases for preservation activities that are driven by user needs should enhance the likelihood of gaining endorsement and funding from administrators and funding agencies. Additionally, partnerships hold out the hope of establishing sustainable funding models for web archiving activities.

Beyond solutions to funding problems, librarians are also in need of software tools and services to address several aspects of web collection development and web archiving. In particular librarians need tools to help with metadata application, evaluation of web-published materials for selection and capture, and version comparison. Tools are also needed for preservation of web-published materials, including tools for file format validation and integrity assurance. Lastly, registry services for web archives would provide an answer to librarians' need to know if some other organization is already preserving a collection of web-published materials.

The findings of this study identified needs and issues faced by librarians, content providers, and researchers, specifically in the areas of government information but also in the larger arena of collection development for web-published materials. These findings suggested a suite of solutions to the major funding, infrastructure, and organization hurdles libraries currently face as they consider the challenges of building and preserving collections of web-published materials. While the user needs are great and the organizational resources are scarce, web archiving efforts will hopefully benefit from the ideas generated by the participants in this study.

Notes and References

1. The Web-at-Risk project [http://www.digitalpreservation.gov/partners/project_cdl.pdf] is one of eight collection development partnership projects funded in 2004 by the Library of Congress under the National Digital Information Infrastructure and Preservation Program (NDIIPP) [http://www.digitalpreservation.gov/].
2. Day, M. (2003, February 25). *Collecting and preserving the World Wide Web: A feasibility study undertaken for the JISC and Wellcome Trust* (Version 1). Retrieved September 22, 2006 from http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf
3. Library of Congress web capture web page. (n.d.) Retrieved September 22, 2006 from http://www.loc.gov/webcapture/index.html
4. PADI (Preserving access to digital information) is a subject gateway to international digital preservation resources that is maintained by the National Library of Australia. On September 22, 2006 the PADI web site was located at http://www.nla.gov.au/padi/index.html
5. International Internet Preservation Consortium (IIPC) was formed in 2003 and is led by the National Library of France. Consortium members include the national libraries of Australia, Canada, Denmark, Finland, Iceland, Italy, Norway, and Sweden; The British Library (UK); The Library of Congress (USA); as well as the Internet Archive (USA). On September 22, 2006 the IIPC web site was located at http://netpreserve.org/
6. International Web Archiving Workshops (IWAW) have been held since 2001 in association with the European Conferences on Digital Libraries (ECDL). The workshops bring together and provide an interactive forum for librarians, archivists, academic

researchers, and industrial researchers interested in developing methods and solutions for web archiving. On September 22, 2006 the IWAW web site was located at http://www.iwaw.net/

7. Day, M. op. cit.

8. National Library of Australia. (n.d.). *Web archiving*. Retrieved September 22, 2006, from http://www.nla.gov.au/padi/topics/92.html

9. The Internet Archive was founded in 1996 to build an Internet library that provides permanent access for researchers, historians, and scholars to historical collections that exist in digital formats, including texts, audio, moving images, software, and archived web pages. On September 22, 2006 their web site was located at http://www.archive.org

10. The National Library of Sweden. (2005, March 1). *Kulturarw^3 web archive*. Retrieved September 22, 2006 from http://www.kb.se/kw3/ENG/

11. Internet Archive. (n.d.) *Wayback machine*. Retrieved September 22, 2006 from http://www.archive.org/web/web.php

12. National Library of Australia. (n.d.) *PANDORA: Australia's web archive*. Retrieved September 22, 2006 from http://pandora.nla.gov.au/

13. Library of Congress. (2005, November 22). *Minerva web archiving project*. Retrieved September 22, 2006 from http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html

14. National Archives of Canada. (2001, December). *Preservation policy*. Retrieved September 22, 2006 from http://www.collectionscanada.ca/preservation/1304/docs/preservationpolicy_e.pdf

15. Aschenbrenner, A., Brandt, O., & Strodl, S. (2005). Report on the 5th International Web Archiving Workshop (IWAW). *D-Lib Magazine, 11*(11). Retrieved September 22, 2006 from http://www.dlib.org/dlib/november05/aschenbrenner/11aschenbrenner.html

16. National Library of Australia. op. cit.

17. Lupovici, C. (2006, June 12). *IIPC activity: Standards & tools for domain scale archiving*. Paper presented at the Digital Preservation Coalition Forum on Web Archiving. Retrieved September 22, 2006 from http://www.dpconline.org/docs/events/060612Lupovici.pdf

18. U.S. Government Printing Office. (2006, August 31). *Technology: Future digital system (FDsys)*. Retrieved September 22, 2006 from http://www.gpo.gov/projects/fdsys.htm

19. U.S. Government Printing Office. (2006, April 18). *Requirements document for the Future Digital digital system* (Version 2.1). Retrieved September 22, 2006 from http://www.gpo.gov/projects/pdfs/FDsys_RD_v2.1.pdf

20. Federal Depository Library Program. (2006, March 2). GPO launches web harvesting pilot projects. Message posted to GPO-FDLP-L electronic mailing list archived September 22, 2006 at http://listserv.access.gpo.gov/cgi-bin/wa.exe?A2=ind0603&L=gpo-fdlp-l&T=0&P=67

21. Sternstein, A. (2005, May 9). Federal documents elude preservationists. *Federal Computer Week, 19*(14), 36. Retrieved September 22, 2006 from http://www.fcw.com/article88797-05-09-05

22. U.S. Government Printing Office. (2006, April 18). op. cit.

23. Library of Congress - National Digital Information Infrastructure and Preservation Program. (2005, October) *Preservation of state government digital information: Issues*

and opportunities. Retrieved September 22, 2006 from http://www.digitalpreservation.gov/partners/states_wkshps.pdf

24. Allen, C. (2006, June 15). Foundations for a Successful Digital Preservation Program: Discussions from Digital Preservation in State Government: Best Practices Exchange 2006. *RLG DigiNews, (10)*3. Retrieved September 22, 2006 from http://www.rlg.org/en/page.php?Page_ID=20952

25. Summary documents of the topics discussed in the exchange sessions are available at the State Library of North Carolina's Digital Preservation in State Government: Best Practices Exchange 2006. On September 22, 2006 the web site was located at http://statelibrary.dcr.state.nc.us/digidocs/bestpractices

26. Shuler, J. A. (2005). Informing the nation: The future of librarianship and government information service [Editorial]. *Government Information Quarterly, 22,* 146-150.

27. The unpublished summary report, *Web-at-Risk: Summary Report of the Needs Assessment,* (2006, June 18), is available from http://web2.unt.edu/webatrisk/na_toolkit/Reports/na_summary_report_final_18jun2006.pdf

28. Rusbridge, C. (2006, February). Excuse me . . . some digital preservation fallacies? Retrieved September 22, 2006 from the ARIADNE web site. http://www.ariadne.ac.uk/issue46/rusbridge/

29. Digital Library Federation. (Updated March 23, 2006). *Registry of digital masters record creation guidelines.* Retrieved September 22, 2006 from http://www.diglib.org/collections/reg/reg.htm

30. GPO Access. (Updated September 12, 2006.). *Registry of U.S. government publication digitization projects*. Retrieved September 22, 2006 from http://www.gpoaccess.gov/legacy/registry/index.html

31. GODORT. *Clearinghouse of government documents digital projects.* Retrieved September 22, 2006 from http://www.gl.iit.edu/services/ref/diggovclearinghouse.htm

32. UK: Arts and Humanities Data Service. (Updated 2004, April 30). *Information on depositing digital resources with the AHDS*. Retrieved September 22, 2006 from http://ahds.ac.uk/depositing/index.htm