Summary Report of the Needs Assessment

June 18, 2006

Prepared by:

Kathleen R. Murray
University of North Texas
krmurray@unt.edu

Inga K. Hsieh
University of North Texas
ikh0003@unt.edu

Acknowledgements

Special thanks are extended to the people who participated in and contributed to the activities that resulted in this report. Individual participants are listed in Appendices.

Project Curators	Arizona State Library; New York University; Stanford University; University of California; University of North Texas
Information Professionals	Archivists and librarians who participated in the focus groups
End Users	Researchers and faculty from academic institutions
Content Providers	Representatives for government agencies and trade unions
Project Staff	California Digital Library; New York University; University of North Texas
Funding Source	Library of Congress: National Digital Information Infrastructure and Preservation Program

CONTENTS

Executive Summary

The Web-at-Risk project is a three-year collaborative effort of the California Digital Library, the University of North Texas, and New York University funded in 2004 by the Library of Congress under the National Digital Information Infrastructure and Preservation Program. 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.

In 2005 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. The purpose of these activities was to elicit the needs and issues librarians, curators, end users, and content providers have in relation to web archives. The key findings of these assessment activities are briefly described in this summary.

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 technical expertise, the 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 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 these content producers as either unaware of the need to preserve their web-published materials or 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. 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 state government publications is often unclear or non-existent and many publications are simply disappearing. 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 extends to local government entities, whose need for assistance in preservation of their web-published materials is quite high.

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 were echoed in focus group discussions and by survey participants. These hurdles were:

1. Technology
   - IT support
   - Preservation expertise
2. Policies
   - Lack of organizational focus for preservation of web-published materials
   - Agreement regarding which materials to archive
   - Agreement regarding what archive technology to implement
3. Management commitment
   - Senior management support for an archive effort
   - Political will to drive an archiving effort through the organization
4. Funding
   - Limited money and budget constraints
   - Staffing issues – a shortage of people and time

The project's curators estimated the magnitude of the financial challenges they will face in creating their collections of web-published materials for the Web-at-Risk project. In rank order, the top four financial challenges were: cataloging, preservation, IT support, and staff training.

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 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 was a major issue for most participants. The two basic questions librarians 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 fell into four categories:

1. Government Information
   - National
   - State
   - Regional & Local

2. Information in Support of Academic Institutions
    - Teaching & the Curriculum
    - Scholarship
    - University Operations
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: smaller 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 or the financial resources to commit to preservation.

Unit of Selection: Content v. Context

For certain research disciplines or types of research, source material context is critically important and therefore the web site would be the unit of selection. In terms of contextual importance, one historian made the analogy between a web site and a newspaper observing that placement of material on a web site has meaning much in the same way placement of an article in a newspaper has meaning. For other research fields, such as statistical research in sociology, the original web-context of the source materials is not always critical and users would be better served by interacting directly with the statistical datasets.

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 captured
3. Articulating criteria for retention of earlier versions

Authenticity

While different users assess authenticity differently, many 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 establishing "fixed" versions and dates for web-published materials is a critical area a web archive should address. Many researchers would like an archive to identify the location of original source materials.

Metadata

Survey respondents identified cataloging as the top financial and technical challenge they see in regard to building web archives. Librarians in general anticipate that in creating collections of web-published materials, the biggest challenge will be the application of metadata. Focus group participants reported their libraries currently do not have enough catalogers for their non-web-published materials. Librarians recognize that evaluating web-published materials and applying metadata requires a specialized skill set. New approaches that utilize technology and include users might apply "indicators of usefulness" to materials and provide new mechanisms, in lieu of metadata, for users to evaluate archived materials.

Organization

Librarians anticipated users would expect full-text search capability in a web archive. In fact, researchers did indicate the most important types of searches are "topic or subject" and "full-text using any keyword". Librarians also thought users would want to search by subject category and 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." Likewise, 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. Librarians agreed that preserving the content of journal articles 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 assert 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. In legal research, such a designation of authenticity for archived materials is critical. For maps and GIS data regarding environmental or natural resources and agricultural reports, both an indication of authenticity as well as version date is critical. Content providers were also 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 materials.

Meeting the Challenges

Librarians who participated in the focus group discussions were asked to identify the top three user needs web archives could address at their institutions or organizations. The most important need they identified was persistent access to the information users need for teaching and research. The participants also identified two additional needs an 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. Articulating the benefits of a web archive or institutional repository and identifying the risks of not preserving web-published materials of importance to researchers and other users should help libraries as they build business cases they can present to administrators and funding agencies.

Looking to cut expenses and realign budgets, both universities and state governments are targeting libraries for downsizing and elimination. At the same time, libraries and archives are responding to an urgent and growing need to collect and preserve web-published materials. This effort cannot be addressed without partnering both internally with other departments in their organizations and externally with other organizations and government agencies. Additionally, software tools are needed to address several aspects of web

archiving. In particular librarians and archivists need tools to help with metadata application, evaluation for selection and capture, and version comparison. Additional tools are 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. Clearly it would be of value to create a shared registry service for web archives. The benefits of such a registry service for libraries include expanding access to materials, eliminating redundancy of effort, and controlling preservation costs. A registry service in combination with collaborations, partnerships, and a web archiving service would provide a suite of solutions to the major hurdles libraries currently envision as they consider the challenges of building and preserving collections of web-published materials.

1 Introduction

The Web-at-Risk project is one of eight digital preservation projects funded in 2004 by the Library of Congress. This 3-year collaborative effort of the California Digital Library, the University of North Texas, and New York University is developing a Web Archiving Service that will enable the project's curators to build, store, and manage collections of web-published materials in repositories located at the three project partner sites. The content will be collected largely from US federal and state government agency web sites, but will also include web-published political policy documents, campaign literature, and information concerning political movements and labor unions.

The project's work is being conducted along four paths of overlapping activities: (1) Assessment, (2) Development, (3) Experimental, and (4) Partnership Building. One focus of the Assessment path is to produce guidelines to assist the project's curators and other information professionals with collection development for web archives. In support of this effort a *Needs Assessment Toolkit*[1] was published in May 2005.

The toolkit consists of implementation guidelines and data collection tools for the three types of assessment activities conducted in the second half of 2005: (a) an online survey of curators involved in the Web-at-Risk project, (b) focus groups with librarians working in a variety of settings, and (c) interviews with potential end users of web archives and providers of web content. The purpose of these activities was to elicit the needs and issues librarians, curators, end users, and content providers have in relation to web archives. This report summarizes the findings from these assessment activities. A brief description of data collection and analysis activities is provided in Appendix A. Individual reports for each assessment activity are also available. The web locations for these are listed in Appendix B.

1.1 Assessment Activities

Survey of Curators

The online needs assessment survey consisted of 58 questions divided into five sections addressing the following areas:

Section A. Respondents' Background & Collections
Section B. Selection Needs: Policy, Identification and Acquisition
Section C. Curation Needs: Description, Organization, Presentation, Maintenance and Deselection
Section D. Preservation Needs
Section E. Curator User Interface Requirements

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 (WAS).

Survey respondents were the 22 curatorial partners involved in the Web-at-Risk project at the time the survey was conducted. In all, 16 surveys were submitted. Ten curators

[1] Murray, K. R. (2005, May 31). *Needs Assessment Toolkit: Guidelines & Data Collection Tools*. Retrieved December 6, 2005, from the University of North Texas Web-at-Risk Project Web site: http://web2.unt.edu/webatrisk/na_toolkit/deliverable_na_toolkit_final_krm_31may2005.pdf

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. Survey respondents are listed in Appendix C.

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, 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 focus groups are listed in Appendix D.

Participants in the Chicago focus group were self-selected from two sources: (a) the general membership of the Law and Political Science Section (LPSS) of the Association of College and Research Libraries (ACRL) who subscribe to the LPSS discussion list and (b) members of the Chicago Metropolitan Library System who were identified by library system staff members. The participants in this focus group came from diverse disciplines in academic institutions and also included three archivists from non-profit organizations.

Participants in the Washington DC focus group were volunteers from a larger group of FDLC attendees identified by government documents librarians from the University of North Texas. Participants in this group came from a homogeneous work environment, namely, government documents departments within academic libraries.

Participants in the three project partner focus groups were volunteers identified by either project curators or project principals at their respective institutions. Participants in the groups held at NYU and UNT consisted largely of individuals with collection development or subject selection responsibilities for a variety of departments within their respective university libraries. Participants in the CDL focus group came from the campus libraries of either Stanford University or the University of California. These participants worked in a variety of departments and included three government information librarians.

Interviews with Content Providers and End Users

Interviews with potential end users of web archives and providers of web content were conducted in 2005. The purpose of the interviews was to elicit the needs and issues end users and content providers have in relation to web archives.

End User Interviews

Project team members at each of the three partner institutions interviewed researchers in the disciplines of history, political science, or law who were working at their respective institutions or archives. In all, seven interviews were conducted: four with historians, two with political scientists, and one with a professor of hospitality law and management. With

the exception of one person whose use of web-published resources was limited to archival finding aids for research, all of the participants 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. (Appendix E lists the end users who were interviewed.)

Content Provider Interviews

In all, 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 that their web sites were already being crawled and captured. (Appendix F lists the content providers who were interviewed.)

1.2 Terminology

One of the outcomes of this research was a renewed appreciation for the importance of establishing definitions of key concepts for effective discourse and inquiry. It became quite clear that in the "real" world there is a good deal of elasticity and overlap in the understanding and use of some relatively familiar terms including:

archive	digital and web
collection	digital and web
material	web-published and other
object	born-digital and digitized
repository	institutional and "trusted digital"

Appendix G is a glossary that may be helpful to readers. For purposes of reporting findings from this assessment, some terms were adopted and used throughout. These are defined below.

Digital Archive

A digital archive is a collection of digital objects that may also exist in other forms. The digital archive preserves the digital versions 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 datasets. Two types of digital objects included in digital archives are: surrogate objects, for example digitized copies of print books or audio tapes, and born-digital objects.

Repository

A repository is an umbrella term for the physical storage location and medium for one or more digital archives. A repository may contain an active copy of an archive that is accessed by users or a mirror copy of an archive that has been replicated for disaster recovery.

Institutional Repository

A repository comprised of digital collections representing the intellectual output of a single university or a group of colleges and universities. The repository captures, preserves, and provides access to these collections as a logical extension of the core mission of the university and as a vehicle for increased institutional visibility.

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. Web-published materials are both dynamic and transient. They are at risk of disappearing. Web archives preserve web-published materials. 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. Organizations, for example, national libraries, research institutions, or professional societies, may build web archives to fulfill their stated mission and to satisfy the information needs of their user community. Alternatively, organizations may enter into service arrangements with third-party archive providers or archive agencies. A Web Archive is a special case of a Digital Archive.

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. The application of the intellectual and logical processes involved in collection management by librarians and archivists results in curated web collections. All web collections residing in a web archive are assumed to be preserved.

Note: In the context of this research, web collections were 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). The web pages within a web site are often published and maintained by a single person or organization, although wider collaborations and social publishing are becoming common, for example, wikis and blogs. Hyperlinks in the form of uniform resource locators (URLs) on web pages access other web pages and specific web-published materials either within the same web site or at a different web site.

1.3 Report Content

This remainder of this report consists of four main sections: section 2 reports findings related to the current issues facing librarians as they deal with the challenges of adding web-published materials to their collections, section 3 reports findings regarding organizational, resource, and technical issues, section 4 reports findings related to collection development for web-published materials, and section 5 introduces some ideas for addressing the challenges and issues identified in this assessment.

2 Challenges in the Trenches

> *At "this university, the role of the library is very much under pressure. We're trying to prove it in because the campus plan doesn't see a need for a library in ten years."* - Librarian

2.1 Transitional Times

For the most part, librarians continue to work in the familiar world of print materials while increasingly accepting responsibilities in the ever-growing world of web-published materials. In building and maintaining collections of web-published materials, librarians confront the challenge of applying their expertise in information organization to a class of materials that behaves badly and over which they have almost no control. This challenge has emerged amid shortages of standards, staff, finances, and infrastructure and with an undercurrent of preservation urgency for web-born materials of historic and research importance that vanish from the Web at alarming rates.

Librarians are in a transition period. They continue to apply their training and skills to print and other physical materials while facing an increasing need to apply their training and expertise, in combination with a new technical skill set, to web-published materials. While interested in embracing the challenges inherent in collecting and preserving web-published materials, librarians often lack the technical expertise, the resources, or both to successfully meet the collection development and preservation challenges they encounter. Some librarians acquire new technical skills, some try new approaches to traditional practices, and some seek to collaborate with more technically trained librarians. 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.

> *"The things we're talking about are basically the things we've always done with the print collection. But I think they're just much harder with web-archived material."* - Librarian

Material Selection

Selection of web-published materials is typically viewed by librarians in academic institutions as an extension of existing collection practices. However, for web collections the bulk of the responsibility for identifying materials shifts from external publishers to internal selectors, namely librarians. This added responsibility involves identifying and selecting web-published materials for their collections as well as tracking updates and changes to those materials.

For some disciplines, materials that used to be published in print are now web-published. Formats have changed but selection has not been significantly impacted. For other disciplines, organizational web sites now offer a wider diversity of genres and formats than print materials offer. Blogs are one example of a new genre of interest to some disciplines. In addition, less-established disciplines, such as cultural studies, have a dearth of available print material. The Web has enabled an extensive amount of material in support of these disciplines to be published digitally. In disciplines such as these, selecting web-published materials in support of the curriculum is far more labor-intensive than selecting print materials.

Preservation Policies

Preservation practices for print and physical materials are well-established within most organizations, often coded into collection management practices or retention guidelines. Likewise, it is commonly known who has responsibility for preservation of print materials within most organizations. However, for web-published materials neither preservation practices nor the designation of who is responsible for implementing them are generally established. Web-published materials come and go quite easily and often no one assumes responsibility for preserving them. While print materials allowed for a delay in addressing preservation because once published the material remained viable for a period of time, web-published materials disappear quickly, often to be lost forever. The lack of preservation policies and practices for web-published materials has a direct impact on sustaining access to them over time.

Government Information

As a rule, print publications of government agencies are distinct entities. There is generally a first edition of a publication followed by mid-year or annual editions. Each printed edition can be reliably preserved as its own entity. Additionally, official legal documents published in print by commercial vendors carry an imprint attesting to their authenticity as official documents produced by some organization.

With web-born government publications, an edition cannot be relied upon to be a constant entity. Many participants have 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. In addition, there is no analog of the "authenticity imprint" for online legal documents published by commercial vendors. There is also a concern among some librarians that as more government information moves to digital-only publications, license-imposed access restrictions to government information will become more commonplace, making it more difficult in the future for people to gain access to government information.

Preservation in the Absence of Repositories

> *"I have been known to archive web publications by printing them out and having them bound in buckram and then cataloged."* - Librarian

Despite resource constraints and the sometimes daunting challenges posed by web-published materials, librarians and scholars are finding preservation solutions they can implement today and are keen to find long-term solutions. In the absence of preservation solutions for web versions of publications, such as institutional repositories or web archives, librarians are printing, binding, and cataloging web-published materials, sometimes in great quantities. Additionally, many researchers retain print versions of key web publications as a precaution against loss of the web versions.

There are numerous niche preservation efforts underway for web-published materials. Librarians are capturing, reformatting, and preserving small web collections on CD-ROM in support of individual faculty members' research. Some researchers are also creating personal archives of web-published materials vital to their research. Others would like to do so but find the volume of information beyond their means to preserve.

Collaborative preservation efforts for web-published materials are underway among both universities and state agencies. Some state libraries are identifying essential government

publications that will continue to be published and preserved in print format. One state library is collaborating with other state agencies to leverage limited technology resources in the interests of preservation of and access to the government's web-published materials. Universities are engaged in LOCKSS[2] projects and programs to preserve scholarly journals and one university is preserving datasets used in faculty research in dark archives while the copyright dust settles.

> *"I think, just on a day-to-day basis, maybe all we can do is these little things that are of immediate use to our particular users."* - Librarian

Policies and practices for the creation and preservation of web collections are being formulated at a few universities. Identifying both the need and cost for creating and preserving these collections as well as assessing the risks of not funding them, is one strategic approach to gaining management commitment and funding. Until formal preservation policies are adopted, some university libraries are requiring subject selectors to address interim preservation strategies for web-published materials in their collection plans.

2.2 Roles & Responsibilities

The Necessity of Working Together

Librarians

While many of the activities involved in managing web-published materials are an extension of librarians' current roles and responsibilities, collecting and preserving web sites and web-published materials present unique challenges and, in most cases, increased technical and curatorial resource requirements. In a climate of uncertainty and funding constraints, libraries are often challenged not only to fulfill all of the functions they have in the past but also to adequately address additional responsibilities for web-published materials.

For the vast majority of universities, the scope of web archiving efforts is beyond the capabilities of their university libraries' IT infrastructures and staffs. For librarians to meet their curatorial responsibilities for web-published materials and collections, collaboration and support from the campus IT organization is required.

IT Staff

While librarians and archivists have expertise in preservation and curation, Information Technology (IT) personnel generally do not. At times, a clash in cultures ensues as librarians bring to bear their experience in information organization and their expectations for material preservation on an IT organization that may neither understand nor value either area.

Routine computer system backups and maintenance practices that are standard IT operations are often either unsatisfactory for or in conflict with preservation requirements. One participant related the experience of IT staff updating all the organizational web pages with current logos and dates, without regard to preserving the originally published versions. Surmounting these cultural and practice differences to implement successful web archives is a major challenge for many organizations.

[2] Lots of Copies Keep Stuff Safe: http://www.lockss.org/lockss/Home

Publishers

The preservation of web-published materials requires participation from both publishers and libraries. While preservation of materials has been a cultural responsibility assumed by libraries, many librarians question if libraries can continue their role as preservationists in regard to web-published or electronically published materials. One librarian stated: "I just don't know if libraries have the resources to do it."

Collaborations

Information products in support of university curricula and state government operations are increasingly web-published as are the information products produced by universities and state governments. The technical infrastructure and support for preservation of these web-published materials challenges both university libraries and state libraries to look for solutions beyond their own organizations and staffs. The scope of the preservation effort spans the institution or government organization and requires the clout of leadership and policy in order to be successful. Collaborative efforts among libraries, IT organizations, and other departments and stakeholders within the organization or government entity are a necessity for successful preservation programs.

Institutional Repositories

Several universities are considering creating institutional repositories. Many librarians see the institutional repository as the solution to their current need to preserve web-published materials. If the effort to build an institutional repository is undertaken by the institution, librarians thought it likely that necessary policies for material deposit and preservation would be created. Also, librarians anticipated that support would be provided by campus IT organizations and faculty would become involved. Both the policies and the support are urgent needs identified by librarians.

Institutional repositories would enable academic libraries to extend their curatorial role in two directions: (a) making their institution's scholarly web publications accessible and (b) providing assurance to users that these materials are scholarly and valid. Librarians identified a range of web-published materials important for teaching and in support of curricula and scholarship that could be preserved in an institutional repository. These materials include:

- Web-published materials of research centers. Currently, when research centers cease to exist, perhaps at the end of their funding or mission, there is often no person or organization responsible for preserving researchers' working papers or the legacy collections of the center. These materials are often lost.
- Individual faculty publications. As allowed by publishers, the institutional repository could provide an additional preservation safety net while at the same time increasing the exposure and availability of faculty research.
- Web-published course materials
- Numeric data in support of research
- Digital images referenced in monographs
- Web-published materials in subject guides

The Uncertainty of Stewardship

Responsibility for Preservation

There was general agreement among librarians that the organization or individual responsible for producing web-published materials ought to take responsibility for

preserving the original materials. Content producers should preserve the materials themselves, participate in collaborative preservation efforts, or make arrangements for someone else to preserve their materials for them. However, this is not generally the case. The current climate can be characterized as one in which publishers and web content providers do not assume responsibility for preservation of the materials they publish. This is somewhat attributable to ignorance on their part of their preservation responsibility but also reflects their long-standing position in regard to preservation. Traditionally libraries more often accepted preservation responsibility rather than publishers or authors.

Publishers

Publishers' commitment to preservation, or their lack of it, has different implications for print materials versus web-published materials. The preservation requirements posed by web-published materials, in terms of both the expertise of staff and the quantity of resources, are very different from the requirements to preserve print materials. Libraries are often hard-pressed to undertake a role as preservationists of web-published materials. Critical questions regarding preservation stewardship for these materials frame the dilemma with which libraries are struggling: Who will take responsibility for preservation and long-term access? Will publishers? Can libraries? While there was general agreement that large publishers ought to preserve their publications, there was also general acknowledgement that small publishers are unable to preserve theirs.

State Governments

With the continuing shift from print documents to web-based materials, responsibility for archiving state government publications is often unclear or non-existent and many publications are simply disappearing. State government agencies often lack policy guidance for preservation of their web-published materials. 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. Some agencies also question whether their web sites and web-published materials are really the state's responsibility to preserve beyond periods of usefulness for the constituencies they serve. Compounding the preservation confusion at some state agencies are retention guidelines that continue to require preservation of "official" publications in print but do not address web-born publications.

Universities

The fundamental research requirement for access to information over time is a primary source of university librarians' concerns for preservation of web-published materials. In regard to government information, some major research libraries are not certain they can either wait for or rely solely upon federal government preservation efforts. They express concerns regarding the sustainability of government programs in future funding cycles. This uncertainty drives these universities to assess their need for local preservation programs.

In some academic researchers' experience, there are times when the only source of required state government information is captured copies on the Internet Archive's Wayback Machine. After years of frustration with disappearing state government web publications, one researcher suggested there should to be some national mandate that state governments were required to submit standard datasets to the National Archives or some other trusted federal authority.

A research institution is primarily serving the needs of its researchers, who in turn often need access to long-term, historical government information. Should the university assume

stewardship for long-term preservation of government information when state agencies often do not retain their materials beyond a short-term time period?

Librarians' concern for preservation and access to web-published materials of federal and state agencies extends to regional and local government entities. Their need for assistance in preservation of their web-published materials is quite high. Many participants thought academic libraries were in a position to offer leadership, direction, and expertise to these entities.

Expectations of Content Providers

It is clear that the variety of web-published materials, organizational mandates and missions, and intellectual property concerns pose challenges for both content providers and web archive providers. There appears to be no one-size-fits-all in terms of approaches and agreements between these two parties. On the other hand, there are likely to be some one-size-fits-many approaches that can be identified. For example, providers of union content had much in common in their archival requirements with some state agencies.

Some content providers have long-standing relationships with archival organizations for their print materials and clearly want to extend that relationship to include their web-published materials. It is a matter of concern to them that agreements with web archive providers are explicit in terms of what is to be archived and how the materials are to be maintained. These content providers also think it is important that they retain intellectual property rights to their material. Some are adamant that their web sites not be archived without their express permission.

State government agencies are aware that their sites are already being crawled by the Internet Archive as well as by Google, Yahoo, and other commercial entities. These agencies put no additional effort into packaging their content for these organizations nor do they establish agreements with them. The agencies have some concerns about their web-published materials being captured, particularly if the information provider is reformatting and repackaging captured materials. However, in the absence of formal agreements, content providers recognize they have little control. These content providers do expect "good harvest behavior" on the part of crawlers, for example, respecting robots.txt and not impacting server performance.

A few agencies could envision a web archive as a safe back-up site for disaster recovery or as a mirror site providing alternate access to their web sites. In either case, the archive would be expected to provide equivalent operational access and functionality. These views of an archive are more in line with management strategies for operational computer systems. It seems funding challenges have prompted systems and project managers at agencies to identify opportunities for effective risk management options beyond their own organizations.

Most content providers 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. From this perspective, content providers see web archives as repositories for posterity that will enable research into historical records for analysis of change over time.

3 Organizational Issues

In a questionnaire completed immediately after the focus group discussions, participants were asked to identify the major hurdles they envisioned for their library or organization in creating a web archive. Their top five responses are depicted in Figure 1.

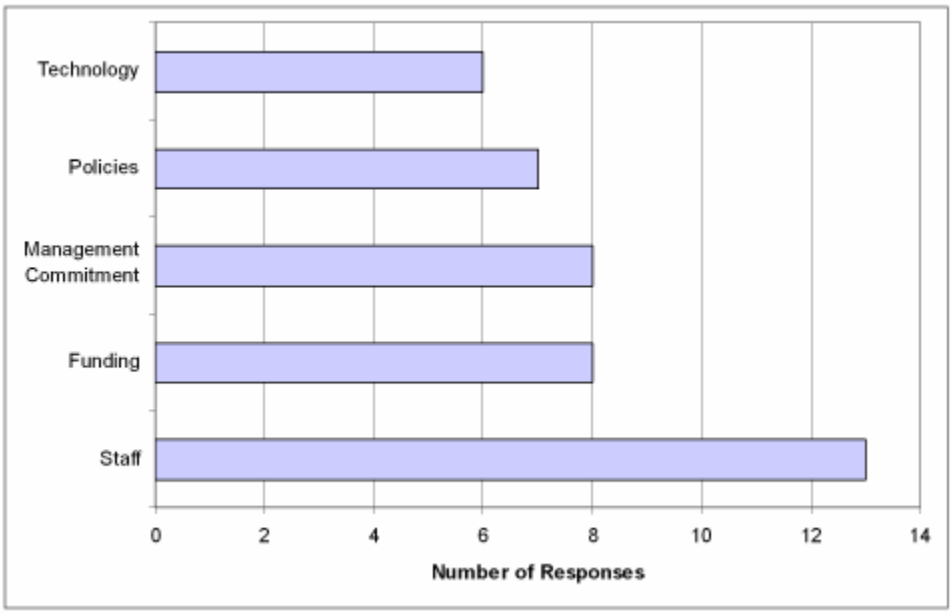

Figure 1. Hurdles to Web Archive Creation

At least six or more participants thought each the following presented major hurdles for their library or organization to surmount: (1) technology, primarily IT support and preservation expertise; (2) policies, including a lack of organizational focus for preservation of web-published materials as well as gaining agreement on which materials to archive and what archive technology to implement; (3) management commitment, getting senior management on-board with an archive effort and having them exert their political will on the organization; (4) funding, chiefly limited money and budget constraints; and (5) staffing issues, primarily a shortage of people and time.

Each of these challenging areas was echoed in the focus group discussions. Likewise, survey participants targeted many of the same areas. The remainder of this section summarizes the findings in each area.

3.1 Staff Resources

> *"There's more stuff we would like to add [to our archive]. We do not have the resources to be able to do that."* - Librarian

A common need in many organizations is for more staff resources. Libraries face dual staffing challenges: an existing staff shortage and a growing demand for staff to select and manage web-published materials. Existing staff do not have the time needed to adequately address collection and preservation of web-published materials.

> *"It's far more labor intensive than people realize to archive web stuff in my experience."* - Librarian

Not only is more staff needed but staff with the appropriate technical expertise is needed and they are harder to recruit. Either existing staff must acquire the expertise through training or new staff with the expertise must be hired.

> *"We are all selectors and we're used to doing this . . . [now] we're trying to . . . pick up a whole other type of responsibility that requires all kinds of different skills. Some of us can do it and some of us can not and really we should be looking at developing digital librarian positions in our libraries, somebody who can come in with the skill and devote the time to it."* - Librarian

Cataloging print materials is already a major resource challenge in many libraries. The ability of web crawlers to capture increasingly large numbers of web-published materials will create additional resource challenges for creating metadata. Even if grant funding was obtained and catalogers hired, there is a concern that over time the cataloging effort may become a low priority in the face of budget cuts. To address this challenge, some participants thought libraries needed metadata specialists dedicated to cataloging web-published materials. These specialists could be a shared resource and work in tandem with librarians, who provide subject expertise.

Some participants identified a need within their institutions and libraries for a central unit with expertise in the preservation of born-digital and web-published materials. This organization could be the focus for the coordinated preservation effort needed between the IT department, the library or archive department, and the overall organization or institution. Librarians could assume curatorial responsibilities for the materials but would rely on technical expertise and support from outside their organization for cutting edge technical endeavors in support of web archiving.

3.2 Financial Challenges

Survey respondents' estimated the magnitude of the financial challenges they will face in creating their collections of web-published materials for the Web-at-Risk project. Figure 2 shows the top four challenges they identified: cataloging (75%; *N*=16), preservation (63%; *N*=16), IT support (60%; *N*=15), and staff training (50%; *N*=16).

Additionally, funding is specifically required for infrastructure, the hardware and software both to create the archive and to sustain it over time through anticipated maintenance and upgrades. Some librarians identified a need for funding to support the archive's "presentation infrastructure", that is, to support the hardware and software to enable users to interact with the archived materials. There was a general concern regarding the uncertainty of future funding for web archiving. Some librarians hoped collaborative programs might provide sustainable models and a few anticipated emerging technologies would mitigate the funding burden.

> *"A huge selling point in getting the agencies to cooperate with us is that we go out and do a little dog-and-pony show. . . . We didn't create a new place for anyone to look [for archived materials]. . . . We enhanced an existing database. In a really grim series of budget years, that's protected funding for something that is a real big ticket item and otherwise would have been an easy target, I think."* - Librarian at State Library

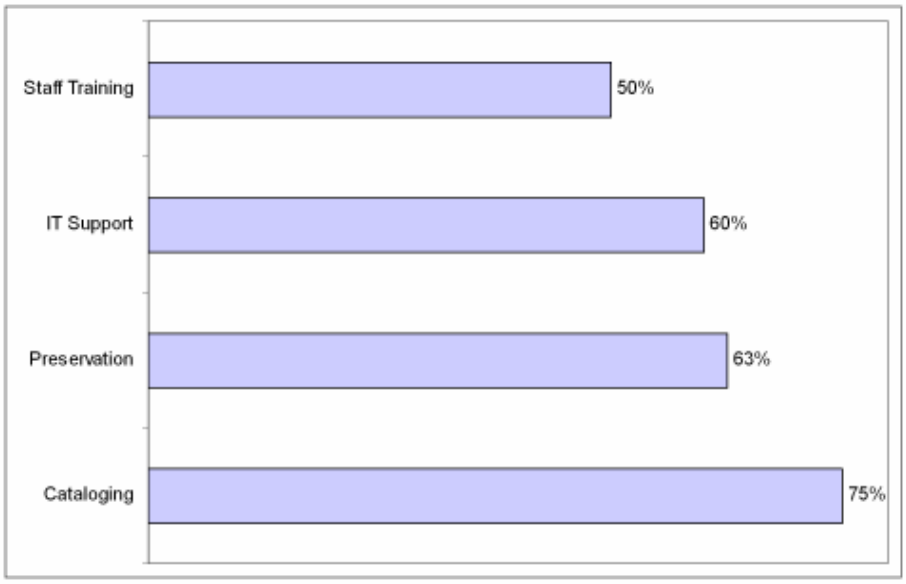

Figure 2. Estimation of Financial Challenges in Building Web Collections

State agencies, including state libraries, are strapped for funds, lack appropriations to carry out preservation programs, or operate under guidelines that constrain investments in preservation solutions. However, despite these funding challenges, there were several reports of successful archiving efforts either among state agencies or by individual agencies. On the other hand, there were also reports of lost web-published materials from agency web sites due to resource shortages.

Grant funding for preservation projects received mixed reviews. Some librarians reported valuable archival efforts that were enabled with grant funding; some reported efforts that were begun with grant funding but could not subsequently be sustained; and some were frustrated with the small return, in terms of the materials actually preserved, relative both to the amount of effort invested and the magnitude of the preservation problem.

3.3 Organizational Support

Librarians and archivists recognize that preserving the web-published or web-born materials of importance to their collections will require the endorsement of their institutions as embodied in cross-organizational archiving policies that bring together all of the key stakeholders within the organization. Among these stakeholders at universities are the libraries, the administration, the IT department, and the faculty.

An ongoing, top-down institutional commitment to a continuing relationship among IT personnel and librarians is required to successfully build and preserve web archives that are meaningful to the university community. This includes administrative procedures in support of the commitment. Support has to come from the top of the organization and then be built into the funding fiber in a sustainable fashion. Web archiving needs to be a mainstream activity of the library or organization, not a short-term project.

> *"What good is all of this [project activity] if next year there is a budget problem or somebody changes positions and the new person in that position says it's not worthwhile to keep this particular project going."* - Librarian

However, at the present time, web archiving is generally focused at the project level not at the organizational level. Survey respondents 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 will be necessary to move from a project focus to an organizational and consortial level for web archiving to have the resource scope it will require.

Both formal and informal marketing efforts to sell the concept of web archives or institutional repositories are underway within some institutions. These efforts aim to garner senior administrative management endorsement for the funding necessary 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, effective cases for preservation funding requests need to include a model for sustainability and collaboration. In this regard, librarians recognize 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 already underway, but especially at regional, state, and local levels

3.4 Technical Challenges

For librarians, technical challenges in creating web archives are entwined in staffing and funding challenges. Many frequently cite technical expertise and infrastructure as major challenges. The implied association is that funding is necessary both for staff with the requisite technical skills and for technical infrastructure. Beyond these over-arching resource needs, some technical issues related to web archiving also emerged.

Survey respondents indicated technical limitations are one of greatest hurdles they currently encounter in creating digital archives. Table 1 lists the major technical challenges they anticipate in building web collections.

Very Challenging	Somewhat Challenging
• Metadata creation • Dynamic nature of web materials • Password-protected source materials • Encrypted source materials	• Authenticity • Unclear collection boundaries in web environment

Table 1. Technical Challenges in Building Web Collections

Librarians or researchers identified the following technical challenges:

- Preservation of web materials currently involves a great number of individual decisions regarding selection, description, and deselection. Technologies and tools to streamline these activities are needed.
- Multiple versions of materials may require multiple hardware platforms, software platforms, or applications. The necessary technology to render all resource formats must be available.
- Some web servers are hostile to capture efforts and pose problems for crawlers seeking to capture their content for an archive.

- In the absence of researcher-supplied metadata for databases, over time no one may be able to meaningfully interpret data that has been saved.
- Many web sites have dynamic content that will not be functional in web archives. Alternatives to exact replication of web sites are highly desired by researchers, for whom faithful representation of web pages has a great deal of value.

3.5 Policies & Practices

There was a common sentiment among librarians that preservation of web-published materials requires collaboration among libraries and that a consortial model for web archives is needed. Such collaborations are a logical extension of current collection development practices, which include managing collections for local user groups and sharing information about local holdings among libraries to enable sharing of materials and to eliminate unnecessary duplication of effort and investments.

Focus group participants varied widely in their opinions regarding the need for new collection policies for web-published materials. Some thought their existing collection policies were format-neutral and by implication covered web-published materials. Others thought web-published materials warranted either new policies or needed to be explicitly added to existing policies. To a certain extent, librarians whose collections relied more heavily on web-published materials tended to express a need for new policies or changes in existing policies.

There was general agreement that long-term preservation of web-published materials requires new guidelines. Some state libraries and university libraries are modifying their collection plans to include preservation of web-published materials. Others are developing policies to address preservation of this new class of materials. Many librarians are hoping standard preservation practices emerge from current preservation projects and grant-funded programs.

Librarians identified five policy areas regarding web archives that need to be addressed by organizations and institutions:

1. A prioritized list of what should be archived
2. A specification of the web-published material types and formats the organization supports
3. Standards and guidelines for metadata application
4. Preservation practices
5. Terms and conditions to address in content provider agreements and other contractual arrangements

Many organizations are struggling with identifying which web-published materials they are going to archive. Most understand that it will be important to identify the material types and formats their organizations can support. Virtually everyone recognizes the need for establishing guidelines for the application of metadata.

Preservation practices should specify which material formats will be migrated and if emulation will be done to preserve access to materials. Methods for integrity assurance should be specified along with their known limitations. Any conditions under which materials will be removed from the archived should be stated as well as practices employed for material replacement, including updates or corrected versions. Lastly, policies regarding the preservation of materials outside the web archive should be specified. For example, in the

event that web sites become "broken" and can no longer be rendered, what alternative preservation methods will be employed: copying to microfilm or printing on acid-free paper?

Content providers were particularly adamant that agreements needed to ensure that intellectual property rights were respected, that access was in accord with content providers' wishes, and that the integrity of materials was safeguarded. Content providers, librarians, and researchers contributed to the following list of areas that should be addressed in submission agreements between content providers and archive providers.

- Roles and responsibilities of parties
- Services provided
  - Support and maintenance
  - Content reformatting and migration
  - Selective removal of archived materials over time
- Specification of the materials to be archived, including server-side code and back-end databases
- Specification of the data format or record construction of the materials
- Specification of the metadata to be supplied with deposited materials
- Material deposit specification (e.g. protocols and verification procedures)
- Copyright and intellectual property terms for the materials
- Access limitations

4 Collection Development Issues

While collection development activities for web-published materials conceptually parallel activities for print materials, most librarians find them more labor-intensive. In particular selection and acquisition activities require more up-front work and often involve individual review of materials. Application of metadata is challenging and often requires specialized expertise. Preserving web collections requires technical skills and infrastructure beyond the capabilities of most libraries. The activity of envisioning criteria for weeding collections of web-published materials brought home the emerging nature of web collections in that the concept of weeding materials preserved in an archive posed a bit of cognitive disjuncture.

And yet librarians had enough experience either as subject selectors who routinely identify web-published materials for their users, as participants in digitization efforts for special collections, or as participants in web harvesting and other preservation repository projects, to identify several concerns and areas that need to be addressed in the collection development process for web-published materials. A framework for this process is illustrated in Figure 3 and includes three major phases: selection, curation, and preservation. Appendix H provides a brief explanation of the activities in each phase as they apply to collection development for web-published materials.

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

Figure 3. Collection Development Framework for Web Archives

The concerns and issues in each activity area are addressed in the remainder of this section. However, librarians identified two inter-related assessments that should be made prior to building a collection: identifying the needs of a collection's users and specifying the focus and range of a collection.

Users of a Collection

Research institutions, special archives, and state libraries serve a variety of users who have unique information needs that should be understood prior to building specific web collections. Needs assessments and feedback mechanisms should be employed early in the process so that major users have a voice in the collection. These end user focused activities should ideally happen amid a culture of ongoing discussions among major stakeholders, which might include librarians, faculty, department heads, agency representatives, students, and members of the general public.

Some librarians at research institutions cautioned that there is a generation of wired students (especially freshman and sophomores) who expect everything to be accessible online. Although that may be a dominant characteristic of that set of users, this same

expectation might be found within any user group. Librarians cautioned that it will be important to address user expectations of web collections and to identify:

- Materials that are not available electronically
- Strategies to address user expectations that cannot be met
- Barriers imposed by search engines

Collection Scope

In addition to initially identifying the users of a collection, their information needs, and their expectations, it is also important to determine the focus and range for a web collection. This involves choices similar to those made for any collection. For instance, in order to appreciate a collection's value and legitimacy, its context and purpose are important. Within a particular context, even poor quality materials may be deliberately collected. It may also be helpful to identify the web-published material types that are significant representatives of the collection's focus.

The breadth of the collection, in terms of subject areas and topics, needs to be specified. Should the collection have materials representative of the range of topics in an area or materials limited to one or more topics in the area? For government information collections, librarians thought at least two selection models were necessary depending on the materials being collected:

1. Subject-centric, cross-agency models
2. Agency-centric, cross-subject models

> *"If I were making an archive I'd put all those association web sites in and the data web sites, but I might pick up a blog here and a rant there and put them in. I assume that even if nobody's going to use it today, somebody might want to use it in the future."* - Librarian

4.1 Selection

Preservation Considerations

Because it is hard to identify today the materials that will be important in the future, selection of web-published materials for preservation can be quite difficult. Librarians and researchers alike easily recount their experiences of web-published materials that vanish over time. (Appendix I is a list of materials participants identified as lost.) Issues with lost and changing web-published materials were reported by participants across organizational types including: an archivist for a non-profit organization, a selector for the Women's Studies program at a university, and a government documents librarian at a large research university who encountered problems with publications from NGOs.

What web-published materials are important to preserve? The range encompasses quite a variety, from the content of web sites listed on academic library subject lists or course resource guides to the publications of the administrative offices of US courts. Overall, the important materials targeted for preservation fell into the four categories listed in Table 2. The list primarily reflects the perspectives of the focus group participants, most of whom worked in university libraries and many of whom were government information librarians. Appendix J is a more detailed list.

Preservation Categories
1. Government Information • National • State • Regional & Local
2. Information in Support of Academic Institutions • Teaching & the Curriculum • Scholarship • University Operations
3. Information Pertaining to Key Events
4. Information Pertaining to or Produced by Organizations

Table 2. Candidates of Web-Published Materials for Preservation

The most important web-published information sources identified by researchers were journals and periodicals, databases, government information, newspapers, and the proceedings from professional meetings. For many researchers, organizational web sites are also an important information source. These web sites contain valuable newsletters, articles, brochures, and links to other information sources. Additionally, it would be of value to historians if information sources from print archives were digitized and published on the web. These source materials include manuscripts, posters, pamphlets, and photographs. Scholars in sociology generally use finding aids to determine if field trips to physical archives appear worthwhile. Web-accessible collections of sociological finding aids describing materials in sufficient detail would increase the probability that a researcher's field trip would be worth their effort.

While the value of web-published materials can often be variably measured through the eyes of the creator, the owner, the archivist, or the end user, one of the more useful findings of the needs assessment was to identify the types of information that are currently falling through the cracks of preservation programs. The primary sources were: smaller 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, for example, of the federal depository library program, or the financial resources to commit to preservation.

Currently, many librarians spend a good deal of time selecting and evaluating web-published materials. The collections in some academic disciplines are more reliant on web-published content. These disciplines tend to be relatively new, such as cultural and political studies, or disciplines that require current information, such as criminal justice or health-related programs.

The dearth of standard references for selection fosters a reliance on personal sources, including government web sites, web site reviews, archive-level descriptions of trusted archives, discussion lists, newsletters, and radio broadcasts. The two basic questions librarians ask in regard to identifying web-published materials for preservation are: "Should *we* save this?" and "Is *someone else* already saving it?" Participants struggle with these questions and seldom find easy and ready answers. One participant noted that dwelling on answering the question of what to preserve can bog down preservation efforts because no one truly knows what will be needed tomorrow and consequently what should be preserved today.

> *"We know what we're getting [with print materials and microfiche] but when we're just going out and harvesting web sites and harvesting digital information we are going to be less familiar with the individual pieces of information than we were when were able to actually handle each document. Losing that control doesn't necessarily have to be a bad thing; it's just really difficult to make that transition."* - Librarian

Federal depository libraries are grappling with preservation and archiving of documents that are published only electronically. Some librarians think redundancy of government information archives is important for their collection because "administrations change" and "funding streams change" and while the GPO "can be a trusted 3rd-party" they fear "putting all of your eggs in one basket." But what portion of a library's collection should be preserved? "We select at 80%. So what does long term preservation mean for us?"

The sheer amount of data in some areas makes if difficult to determine the institution's archiving niche. There is, for example, an enormous amount of numeric data in existence from both government and other sources. How does an institution determine what to preserve? Should the institution collect and preserve information that is deemed to be "at risk" or information that is heavily used within the institution?

> *"Census stuff is a really good example [of heavily used data]. On the other hand ICPSR [Inter-university Consortium for Political and Social Research], which [we're] a member of, [has] captured all the 2000 decennial census stuff. Do we feel like, 'OK, that's taken care of so we'll put our efforts elsewhere'?"* - Librarian

Because selection of web sites consumes an inordinate amount of time, some librarians and archivists advocated for depository or submission policies and mechanisms that would engage authors, creators, and publishers in the selection and preservation of web-published materials. Some thought authors, creators, and publishers need to select and package their content and corresponding metadata for submission to archives for preservation. Academic librarians envisioned policies requiring mandatory submission of faculty publications to institutional repositories as well as including a documented preservation process as a funding requirement for research projects.

There is general acknowledgment that, with the exception of smaller local efforts, the large-scale preservation programs required for both government information and information in support of teaching and scholarship will require collaborations. Librarians see a need for a nationally coordinated effort that would include a directory of archived web-published materials. Some localized collaborations for preservation already exist. More of these types of efforts are needed to meet the preservation challenge posed by web-published materials.

Suggested Criteria for Selection

Many participants recognize that selection criteria for the web-published materials that should be preserved need to be developed and employed. For example, when a librarian is selecting "web-published materials of long-term significance" for an institutional archive, what are the indicators or measures of long-term significance? The following criteria were identified by focus group participants.

- Consistent with the historical collection areas of the institution
- Supportive of collection goals

- In danger of being lost or disappearing
- Identified as "lost materials" through information requests
- Usage or demand
- Supportive of faculty scholarship and research
  - Databases used in research
  - Materials cited in research publications
- Supportive of student learning and research
- Key events that emerge amid lots of media attention and their related grassroots information sources
  - Bush/Gore vote count in Florida
- Quality
  - Packaging
  - Ease of use
  - Note: One participant cautioned it is important to remember that an individual item selected for a collection may not be of great quality in and of itself, but may represent a certain aspect of the collection

Unit of Selection

The unit of selection, for example whether a curator is collecting discrete objects such as documents or images or collecting web sites, is directly related to the goals of a collection. For the survey respondents, the unit of selection varied widely. Although 44% planned to select at a web site level, half of survey respondents planned to collect at a more granular level than this, such as the object or logical document levels. Those curators that plan to select at web site or organizational levels usually thought their end users would want to interact with archived materials in a way that mirrored the original web sites. However, for collections at more granular levels, librarians would expect their end users to interact directly with the materials in the collection and not with the web sites in which the materials were originally published.

For academic librarians, the unit of selection (e.g., image or page or web site) depends on the discipline and the purpose of a collection. For certain disciplines or types of research (e.g., anthropology and history), source material context is critically important and therefore the web site would be the unit of selection. For these disciplines, building collections comprised of "parts" of web sites (e.g., selected images or videos) would be a disservice to future scholarship and research. For other research fields (e.g., statistical research in sociology), the original web-context of the source materials is not always critical and users would be better served by the ability to interact directly with the statistical datasets.

Researchers add that the unit of selection depends on the research purpose. For example the research value of advertisements in or out of their web page contexts would be different if a researcher is comparing images of a certain character within ads over time or if the researcher is investigating the role of advertisements in web publications over time. In terms of contextual importance, one historian made the analogy between a web site and a newspaper observing that placement of material on a web site has meaning much in the same way placement of an article in a newspaper has meaning.

Assessment of Intellectual Property Rights

Since capturing web-published materials essentially involves copying them, copyrights should be evaluated prior to capture. One-half of surveyed curators were unsure if permission to copy materials would be needed for the web-published materials initially identified for their collections. It was clear from the test crawls of web sites conducted by

the Web-at-Risk project in 2005 that rights statements will need to be individually evaluated and that assumptions about rights based solely on source and type of materials will not suffice. This finding was supported by interviews with representatives from state government agencies who indicated that materials accessible from agency web sites and agency-sponsored web sites may or may not be in the public domain.

Evaluation of copyrights can decidedly impact material selection. One participant decided not to build a collection when it became too difficult and expensive to obtain copyright permissions for the photos and articles. Another participant recounted that she had a large number of newspaper articles documenting the history of her campus that she would like to digitize and archive. She was dissuaded by advice that she would have a difficult time getting the necessary releases from newspapers.

Librarians and researchers are concerned that government information remain publicly accessible. As the GPO repositions itself as a vendor or supplier of electronic information, some librarians wonder if licensing agreements will become more common for federal government publications and if such agreements will impose access conditions unacceptable to librarians. Interviews with representatives of state government agencies suggested they have a fundamental commitment to public access to their web-published materials. Most were aware that their web sites were already being captured. However, one researcher deplored the practice by some state government web sites to restrict capture of their materials by including robots.txt files on their sites.

Content providers from outside the government arena were not amenable to their materials being captured without their express involvement and permission. These organizations generally hold the property rights for the materials they publish. Capturing of content from their web sites, as well as deletions or modifications to archived copies of their web sites or web-published materials on the part of an archive agency, would require explicit permission from the organization.

4.2 Acquisition

Material Formats and Types

Some participants familiar with creating collections of web sites recommended curators evaluate material types and formats in candidate web sites. To support this analysis, automated tools are needed. Additionally, some participants cautioned that file extensions do not always accurately reflect the actual format of files, which might need to be further verified. The web sites of the content providers who participated in the study illustrate the range of materials that might be encountered. In addition to text and graphics, the materials included:

- Searchable news article databases
- Video and audio content
- Periodical publications and news content
- Web logs
- Linked content to local affiliates
- Linked content to related national affiliates or government agencies
- Databases
- Programmatically-generated web pages using database content
- Forms-based interface to collect information from visitors
- Web pages customized using personal information about the visitor

Some of these materials present formidable challenges to web crawlers or simply cannot be captured by a web crawling process. Results from the test crawls conducted in 2005 indicated that different crawlers have difficulty with different file formats, that hyperlinks embedded in java and flash coded files cannot always be followed, and that crawlers may have difficultly capturing multimedia files.

The general consensus among librarians and researchers was that all material types (e.g., a video file, an audio file, and a text file of the same speech) and possibly all formats (e.g., an image in both jpg and tiff formats) should be captured and preserved when possible even if the quality was poor because different users will require access to different material types and different formats contain different information. Researchers add that the implications of not retaining multiple types of an item can be quite significant depending on a researcher's area of study. For example, linguists researching variations in speech and psychologists studying non-verbal behavior would be thwarted if only transcripts were available. Researchers are also concerned that information may have been lost or altered, or additional information introduced, when source materials are recreated as different material types, for example, when a transcript of an audio file is created.

Even when decisions are made to limit a collection to certain file formats, complications can occur. One existing archive decided to archive only PDF formatted materials but discovered a need to establish a practice of also archiving the software to read each version of the format.

Frequency of Change

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 captured
3. Articulating criteria for retention of earlier versions

How often materials are captured and how many versions are captured involves professional judgment and there is some recognition that regardless of the frequency with which source materials are re-captured, some content will likely be lost. Assessing change can be very time-consuming and many librarians thought that change in web-published source materials would need to be assessed by the web capture system versus by a curator through human evaluation. However, while librarians definitely hope automation can help identify changes in content, they assert that automation cannot replace professional judgment regarding what content changes should trigger re-capture of materials.

Librarians and content providers agreed that change in source materials is highly variable and that not all versions of materials necessarily need to be captured. Some materials change constantly, some change infrequently, some change predictably, and some change at a variable frequency. In legal research, critical source materials change daily, (i.e., "regulations are being promulgated, court decisions are being decided, and laws are being enacted") and researchers need access to historical versions. For some government publications, capturing each version is critical to a collection. Content providers indicated that most databases changed daily but opinions regarding how frequently their databases should be captured varied from daily to yearly. Librarians thought it might be important to researchers and institutions to only preserve database versions upon which research publications are based or to only preserve final datasets at the end of research projects.

In one case the high frequency with which some web sites changed resulted in a decision to not add the sites to a collection because it would be impossible to capture them in a manner that would be representative of the material. In a related finding, the test crawl results highlighted the need for web crawlers to "be polite" in regard to consuming resources on the hosts from which they are capturing materials. Content providers echoed this concern.

For materials collected and preserved via agreements between web archive agencies and content providers, it should be possible to establish triggers for re-capture. Some content providers suggested that a routine capture schedule, such as quarterly, could typically be established as could trigger events to initiate re-captures, for example political events, conferences, or material updates. There are also unknown trigger events that might initiate captures, such as significant but unpredictable events that garner substantial attention or impact operations.

Capturing Content from External Links

Over half of survey respondents indicated it is was important to include materials from the first level of external links in their collections. Content providers thought capturing material from external links was dependent on the function of the archive. If the web sites in the archive will mirror source web sites, then external links need to be operational, which might require that the content from external links be captured. In several instances, content providers' web sites rely on content from other organizations and the archived web sites would be of less value if that external content were not captured. Users did not have general agreement regarding to what extent the content of linked materials should be included in an archive. While it would be ideal to capture as much as possible of the linked content, most thought only critical content whose absence might misrepresent the meaning or value of the source web site should be included.

The test crawl findings from 2005 indicated that the project's curators had various definitions for "level of external links" and that most of their definitions were not in concert with the behavior of the web crawler used in the test. Likewise, in many cases curators were not satisfied with the results of test crawls that captured "one level" of content from external hosts because too many extraneous web pages were captured. Those responsible for identifying web sites for a collection will need a clear understanding of both the structure of the web sites and the capture method of the crawler. In the former case curators would benefit from tools that assisted them in analyzing the structure of web sites.

Authenticity of Materials

> *"I would want the archive [to be] from an institution that I have faith and confidence in; if it's done in the university or the federal government that would satisfy me."* - Researcher

Researchers, users, and librarians generally trust that the materials libraries provide are authentic. University libraries enjoy a certain amount of attributed authority in the academic community, and their traditional reputation for assuring the authenticity of library resources is often extended to embrace web-published materials. Collections of web-published materials are often considered authoritative or authentic by virtue of being created by the library. Researchers in particular express this trust in university libraries and archives as well as in government archives such as the Library or Congress and the National Archives and in certain major publishers such as the New York Times.

Librarians and researchers are all aware that web-published materials can be altered. Librarians believe the library's contract of trust with their users ordains libraries with the responsibility of assuring users that the materials in a web archive are trustworthy, which generally translates to the materials being what they appear to be, or said another way, that the materials are genuine and have not been altered. The curators surveyed thought responsibility for guaranteeing the authenticity of web-published materials lies primarily with the content provider and secondarily with the curator or archive agency. Librarians and researchers also thought end users ultimately had to assume some responsibility for verifying the authenticity of source materials.

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 establishing "fixed" versions and dates for web-published materials is a critical area a web archive should address. Many researchers would like an archive to provide an indication of where the original source material is located.

> *"There has to be some way of having access to the original. I wouldn't be comfortable with anything else."* - Researcher

In practice, librarians observe a variance in authenticity discrimination depending on the user group, the demographic characteristics of individual users, and the practices within an academic discipline. Undergraduates will generally employ less discretion and need more cues and criteria provided to them to evaluate web-published materials. One librarian found users in the general public often accepted whatever information they found in support of their opinions. Law publications require authors to cite print sources in lieu of electronic sources and, in practice many researchers will opt to cite print sources because there are fewer questions within their discipline of their authenticity.

While different users assess authenticity differently, many need and most would want some authority to provide an assurance of the authenticity of web-published materials in a web archive. Content providers were also concerned about how an archive might represent itself. It seems in some cases there may need to be a statement clarifying the archive as an "official" or "unofficial" version of the materials. Librarians who select web-published materials sometimes adopt the archivist's concept of "certification", which is a process for assuring authenticity of materials. Many librarians saw a need for an authentication mechanism for official government and legal documents that are published online by commercial vendors.

> *"I've been thinking . . . add a symbol or icon on docs that says 'We have double-checked this against the original source' . . . like on Ebay, 'ID Verified' [indicates] authenticity verified by a human being."* - Researcher

Modification of web resources for technical and policy issues appears to be acceptable to librarians and researchers, who thought consistency in policy application would lend to the authenticity of materials in a web archive. Participants generally endorsed web archive practices that would alert visitors to changes made to the original source materials. One participant noted this "needs to be done and would mitigate the problem with authenticity." Researchers thought archives should tag web pages to indicate changes and should provide documentation explaining modifications. One researcher wondered if users would be able to assess the impact of material format changes and thought they might need the archive to provide this assessment. Another participant suggested archive agencies provide "maps" of

original web sites and thought this might provide sufficient context for future research of web sites from which some content was removed.

4.3 Description

Level of Description

One librarian thought that library catalogs are created by librarians for librarians. Users may need something less sophisticated to accomplish their locating and evaluating tasks. The counterpoint expressed by another participant is that there are a variety of users and some require the ability to locate materials using quite specific values. Some focus group participants thought that descriptions of materials in web archives were only needed at the collection level, similar to the collection descriptions in finding aids. Armed with descriptions, users could navigate the resources themselves. Several librarians thought the current level of detail and access in finding aids is of limited or no value to many users who need richer content descriptions and who expect online access to the materials in some form.

The level of acquisition may be different from the level of description needed. For example, it may be expedient to capture an entire web site but add descriptive metadata to specific web pages or objects. Many survey respondents plan to build collections of web-published materials at other than the organizational or web site level. For this reason, tools are needed to support metadata creation at levels more granular than the web site level.

In certain disciplines, applying metadata to individual objects would help discovery and increase the utility of an archive.

> *"Freshman and sophomores, in the field of popular culture, want a snapshot of a particular period and will want indexing to advertisements. It's like every ad that isn't indexed in some way is less useful to them. It's a broader brush to the way scholars may look for things."* - Librarian

Some librarians thought students might well want detailed descriptions at the object level. However, all users are used to much less granularity from web search engines, which provide very brief high-level descriptions.

Original Cataloging

Metadata creation was one of the greatest hurdles faced by owners of existing digital collections and all librarians are very concerned about creating human-generated metadata for web-published materials. Many question whether this is even feasible. Some suggested that human-generated metadata for archived materials would quickly get beyond the resource capabilities of a library as the number of web-published materials captured increases. Most librarians thought automated metadata generation would be needed, including subject or topic classification.

Some participants thought there is of necessity a dependency on the provider, creator, or owner of web-published materials for descriptive metadata for those materials. However, other participants asserted that neither individual creators nor web managers generally supply or create metadata today and these participants were not optimistic that this situation would change in the future. Neither do they expect many publishers to supply metadata with their web-published materials. In general, content providers had no issues or concerns regarding the addition of metadata to their materials by the archive agency. Some content providers want to approve the metadata that would be added.

Data bases created as part of university research programs pose a major challenge to metadata creation. Researchers and their collaborators are in the best positions, possibly the only knowledgeable positions, to create metadata for their data. Providing tools for researchers to create metadata for their data would be beneficial.

> *"…you have to know about a variety of things that you just can't necessarily glean from … definitely not just from the raw data and you may or may not be able to glean it from the … whatever documentation they give you."* - Librarian

In the government documents arena, state and local information requires original cataloging. This is true of web-published information as well. Some state libraries are cataloging state information to the extent their resources allow. Materials from federal web sites outside of the federal depository program also require original cataloging.

Overall, librarians agreed there would need to be a good deal of original cataloging of the materials in a web archive to make it useful to the range of archive users. They also thought consideration must be given to the trade-offs between the value and usefulness of metadata and the amount of time and effort required to create it. New approaches that apply technology and include users might provide "indicators of usefulness" for materials and provide new mechanisms for users to evaluate archived materials.

Breadth of Cataloging

Librarians are aware they cannot create exhaustive metadata for source materials and cannot anticipate all future resource discovery needs. Some note there is no ability to index the table of contents for print or electronic resources within the MARC record. In cataloging web-published materials at the web site level, identifying their contents would present a major resource problem.

Librarians, content providers, and researchers indicated a need to be able to identify the versions of web-published materials. Time stamps reflecting when web-published materials are harvested are needed for each archived version of a web resource. Additionally, in legal research, an indication of when the information was in force or the effective date of the material is critical. For maps and GIS data regarding environmental or natural resources and agricultural reports, version control is critical.

Some librarians thought separate catalog records will be needed for each version of captured materials. However, others anticipated this would result in a great deal of repetition within records across multiple versions. Most researchers envisioned one summary record for each web site that would include a description for the web site and list each version along with its capture date. Access to individual records for versions would be a nice feature.

Standards and Guidelines

Librarians generally agreed that standards and guidelines could be developed that identify both a set of metadata elements and the format for values of the elements that would be applicable to the variety of materials in a web archive. Librarians reported archive experiences that included creating subject headings, a thesaurus, and authority lists as well as using modified Dublin Core with some qualifiers and enhancements. There was also a general sentiment that establishing guidelines for web-published materials presented certain

challenges. For example, what constitutes the title of a web page: the page title in the URL, the title included in the displayed content, or the title listed in the embedded metadata?

Some librarians were sensitive to the reality that some emerging disciplines use source materials differently from how catalogers are accustomed to describing them, for example, the importance to students of popular culture to identify an advertisement within a trade publication rather than the articles in the publication. These librarians suggested materials in a shared web archive might be cataloged in accord with guidelines customized for library-specific sets of users. If finding aids are used to describe collections of web-published materials, this might result in multiple collection level descriptions based on the same materials.

4.4 Organization

User Expectations

Librarians anticipated users would expect full-text search capability in a web archive. In fact, researchers indicated the most important types of searches are "topic or subject" and "full-text using any keyword". They also identified the following as desirable search criteria for materials in a web archive:

- Author
- Title
- Original URL
- Publication Date
- Date archived
- Organization
- Description

Librarians thought users would also want to search by subject category and 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." Likewise, researchers indicated they would like to browse a web archive via a subject directory structure.

A unique user need in regard to government information, was for access across agencies to comprehensive government information specific to a local area. Currently government information and materials are not indexed in a manner that supports this type of discovery.

Evaluating Search Results

When evaluating results of archive searches, Google and the Wayback Machine were both mentioned as models of search results that had been effective for the users interviewed. Researchers varied regarding the display of all versions of a web site available in an archive. One thought it would be "overwhelming" and another thought it was "extremely important".

Most researchers thought that a model using a single record to represent a web site was best. This single record should list available versions of the web site. Some desired a separate record for each version, but thought that these could be linked from a single summary record. For web-published materials described at a more granular level than the web site, a single record model was also best. This record should include the available formats for the material described. Researchers also want to be able to identify the date of capture.

4.5 Presentation

Look and Feel

Presentation of web-published materials depends on the materials themselves, on the academic discipline of researchers, as well as on one's general perspective of an archive. For instance, librarians agreed that preserving the content of journal articles would suffice. When the content of articles was coupled with an understanding of its structure, the articles could be presented within any suitable interface or context. 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 are not concerned with replicating their web sites' "look-and-feel".

However, many participants thought other types of materials would need to be presented in their original web context, that is, exactly as they appeared prior to capture. 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 archived web sites should be presented in such a way that they mirror the source web sites.

In some fields, the absence of the original web context might be unfortunate and limit research but in other fields the loss of context is irrelevant for research purposes. It appears that the research or scholarly need for capturing and rendering the original web context of source materials will vary with research requirements and will be related to the field of research. Researchers in some disciplines need to experience the original context of the source materials in order to understand them. Researchers in other disciplines are best served by working directly with extracted portions of original source materials. All researchers appreciated the historical value of faithfully preserving the original web sites but some are primarily interested in the content. For those researchers who are primarily interested in content, the same materials presented in a different context would be acceptable.

About half of survey respondents felt their end users would expect to interact with archived materials in a way that mirrored the original web sites. These curators were generally planning to collect at web site or organizational levels and presenting mirrored versions of the web sites is fundamentally sensible. However, many survey respondents plan to build collections of web-published materials at other than the organizational or web site level. Presentation of the materials in these collections would seemingly be done outside of their original web context.

Dealing with Active Content

The types of problematic content participants identified included:

- Interactive elements, such as forms and email
- Hyperlinked content, both archived and not archived
- Dynamically generated web pages, such as those resulting form searches
- Database-driven web pages
- User-specific content, which has no static presentation
- Programmatically generated web pages
- Web pages generated using a combination of code and style sheets

> *"I would either extract just the docs or data you need and toss out the navigation structure or completely duplicate the web site the way it was, including the database, the navigation . . ."* - Librarian

Content providers' existing web sites generally provide search functionality within their site. Some provide email alerts based on registered visitors' personal preferences. Some provide enhanced search and alert services for a fee. While all participants understood that some alerts and services would not be retained in the web archive, they wondered if others might be. Some users thought email links should be deactivated but thought the original link information should be retained for research purposes. Some researchers thought that any code or script-based functionality that could be replicated in an archive should be replicated. Some participants thought retaining the functionality to generate customized web pages would be desirable if possible.

Most survey respondents felt that broken links, that is, links that point outside the archive but no longer work, should remain active and that a standard browser message or customized message should be presented to users. Most researchers and users thought that hyperlink information should be preserved in an archive even if the hyperlinks are disabled or no longer valid. If external links were disabled, this information would enable users to access the sites using a web browser. In the interest of presenting the most "faithful representation of web pages" whose forms were no longer operational in an archive, some researchers suggested it would be good practice to provide screen shots of the original forms in addition to explanations as to why they no longer work.

Authenticity and Version Indication

Researchers assert that web archives should make it clear that users are interacting with archived material and not "live" materials. 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. In legal research, such a designation of authenticity for archived materials is critical. For maps and GIS data regarding environmental or natural resources and agricultural reports, both an indication of authenticity as well as version date is critical.

Access

Archives have always accepted collections that have "embargo periods" associated with them, for example, not accessible for 50 years. Similarly, some web collections might require a type of "dark archive" or one to which no end user access is permitted. One participant reported that at present their organization is creating a dark archive out of expediency and necessity. This is being done so that critical information in support of research is not lost. While not a preference, one participant indicated their library would continue to archive state government information even if they were unable to provide access to it. There was a general sentiment that collecting web materials for a dark archive would have value but many librarians thought preserving materials without allowing access to them would not be in keeping with their libraries' missions and would be difficult to sell within their organizations.

Content providers were generally committed to making the information on their web sites publicly available. However, for privacy reasons, some web sites currently require password access to a portion of the content, such as staff lists and customer data. If this content is archived, then some type of controlled access to materials in the archive might be required.

4.6 Maintenance

Technology Roles & Responsibilities

For cutting edge technologies, librarians generally acknowledge their need for technical support. It is not feasible for librarians to become technical experts. Successful archives involve a partnership between librarians or curators and information technology (IT) professionals. Each of these groups comes with its own values and skills and the roles and responsibilities of curators and IT professionals blur in some areas.

However, decisions regarding archival maintenance need to be made by the appropriate organization. While curators need to understand the impact and trade-offs for hardware and format migrations, IT staff should not make curatorial decisions. Applying standard IT data retention practices to web archives is unacceptable. For instance, removing materials based on the length of time they have been in an archive would be a disservice to research in general since change over time is precisely the concern of many scholarly researchers. IT professionals need curators to identify whether multiple versions of web sites should be retained in backups or whether newer versions should replace previous versions.

Access

Curators thought their users would require access to archived materials for the foreseeable future. Researchers concur with this and generally believe that preservation implies perpetuity of access. However, some librarians pointed out that over time some archived web sites will become "brittle" or broken. This may happen because the format of the content can no longer be rendered. Archives need the ability to identify these web sites.

In the case of archived copies of official materials, some content providers are concerned that database back-ends be secured so that the content cannot be altered without the express permission of the agency that originally published the materials. At the same time, archived versions of official materials may need to be corrected. Mechanisms for this function need to exist. For example, there may need to be a notification service in place so that content providers can alert archive agencies of changes.

4.7 Deselection

Many librarians had not considered deselection of materials from a web archive but could anticipate that this might become required or even desirable. Storage is relatively cheap and that has contributed to a lack of focus on weeding digital collections. Some librarians thought buying additional storage might be a more economical solution than weeding a web archive. Others thought as emphasis moves from *selection* of materials for an archive to *maintenance* of archived materials, more consideration might be given to weeding collections. It may be that hybrid practices that merge deaccession from the archive world and weeding from the library world will offer alternatives for managing long-term storage and preservation of materials in web archives.

Evaluation of Materials

Most survey respondents cited copyright violations and legal reasons as criteria for removing materials from a web archive. Approximately one half of the curators also planned to employ storage costs, usage, and the sensitive or offensive nature of materials as deselection criteria. Respondents also identified the following as possible deselection criteria:

- Value of material in relation to all available material
- Takedown requests from owners
- Data corruption
- Relevance to collection goals
- Availability elsewhere
- Duplication within the archive

One researcher suggested the following factors should influence retention of archived materials:

- Something that seems extraordinary (e.g., an unusual event)
- An unusual kind of record (e.g., an expensive autobiography or a rare diary)
- The source (e.g., a person of importance at one time)
- Something likely to generate interest (e.g., a great unpublished collection of cartoons)

Determining the value to researchers of a special collection is done in part through curators' interaction with the researchers who visit and use the materials in the collection. One participant wondered if this feedback mechanism would ultimately vanish as researchers move away from a reliance on physical objects in libraries and archives to a reliance on web-accessible archives and collections.

Researchers generally viewed collections of web-published materials as being retained forever. Some researchers could identify web-published materials that may not need to be preserved for as long as others. For example, some thought that proceedings of professional meetings and unpublished works might not need to be retained for more than three years.

In some cases, different formats of the same material (e.g. a video formatted for viewing with different plug-ins) is for the convenience of users and is not important for either archival or research purposes. If forced to remove some materials, researchers suggested it might be good practice to retain the original format and the most recent format of materials that had gone through several format migrations over time.

Frequency of Use

Some librarians expressed their belief that frequency of use should not be applied as a deselection factor or at least not as a sole deciding factor. Usage of an archive may not be important as an indicator of its value. Some permanent archives contain materials deemed important by the archiving institution and there is an expectation that they will never be evaluated for deselection. Participants noted that usage statistics are readily available in a server environment, however usage reports range in value depending on their structure and specificity. At times, usage statistics can be misleading and might camouflage discovery problems.

Redaction of Confidential Information

Some librarians were concerned about the personal information that might be captured from web sites. Likewise, while most government records are public, some thought there would be a concern for the confidentiality of personally identifying information, such as names, addresses, dates of birth, and social security numbers. Redaction of some personal information might be an issue web archive agencies need to consider.

4.8 Preservation

> *"When you say you're going to preserve the web sites, what is it about the web site that you're going to preserve?"* - Librarian

Librarians recognize there are many challenges inherent in preserving web-published materials. Among these are preserving the functionality embedded in web sites, validating file formats, preserving poor quality "derivative assets" as opposed to preservation quality masters, verifying the integrity of captured materials, identifying versions of the same materials, and dealing with file formats that become obsolete over time.

Integrity

Integrity of the bit stream is a preservation responsibility. 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. Bit stream integrity needs to be differentiated from authenticity. For example, a web archive could replicate an inauthentic copy of material in a file with perfect integrity but this would not change the fact that the material was inauthentic.

Versions

Depending on the degree of change to the materials, survey respondents thought users would expect multiple versions of materials to be retained in the archive. Survey respondents were evenly divided on how accepting users would be if newer versions of materials supplanted older versions. Librarians indicated that while some disciplines are typically interested in the most recent information, researchers generally value versions of material from different points in time for both longitudinal studies and historical research. Researchers thought that retention of multiple versions of a web site should be based on how much each version contributes to the body of knowledge about its subject. If the web context in which materials were originally captured could not be preserved over time, researchers and librarians agreed that the content should still be preserved and perhaps accompanied by a detailed map of the original web site to provide researchers with contextual information about the preserved content.

Migration

Migration of materials in a web archive was discussed in relation to authenticity of those materials. The migration activities that survey respondents thought were the greatest threats to the authenticity of materials were migration to a different operating system, to a different hardware platform, and to a different file system within the same operating system. Curators thought that migrating materials to a different format was only a small threat to authenticity. However, content providers and researchers were concerned that over time reformatting materials would compromise their authenticity.

Software Tools for Curators

Curators indicated that features and flexibility in the following areas are important for web archiving tools:

- Level of selection
- Frequency of reacquisition
- Specification of crawl configuration parameters
- Application of metadata

- Migration
- Validation

Generally, respondents were not concerned that embedding metadata would threaten the authenticity of materials, so the automatic generation and application of as much metadata as possible by a web archiving tool set would likely be of significant benefit at all levels of selection.

For some materials, currency is essential. Tools that can recognize when new materials appear at a source would be a valuable asset for curators. Findings of the test crawl supported this and also indicated that identifying when significant changes to a site have occurred would be a desirable function for web archiving tools.

Additional findings of the test crawl suggested features that would be of value in curator tools. These features included:

- Determining an optimal level for a crawler to retry capturing content from a link so that the crawler does not get caught in loops or the crawler process does not hang
- Understanding how a crawler handles links so that material critical to a collection is captured
- Limiting crawls to specific domains so that an abundance of extraneous content is not captured and essential content is captured
- Identifying any file formats that challenge crawler tools so that critical file formats are captured

5 Discussion

5.1 Needs Addressed by Web Archives

Librarians who participated in the focus groups conducted by the Web-at-Risk project were asked in a questionnaire to identify the top three user needs web archives could address at their institutions or organizations. Their responses are illustrated in Figure 4. The most important need they identified was persistent access to the information users need for teaching and research. The participants also identified two additional needs an 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. The outline that follows expands on the needs librarians expressed in each area.

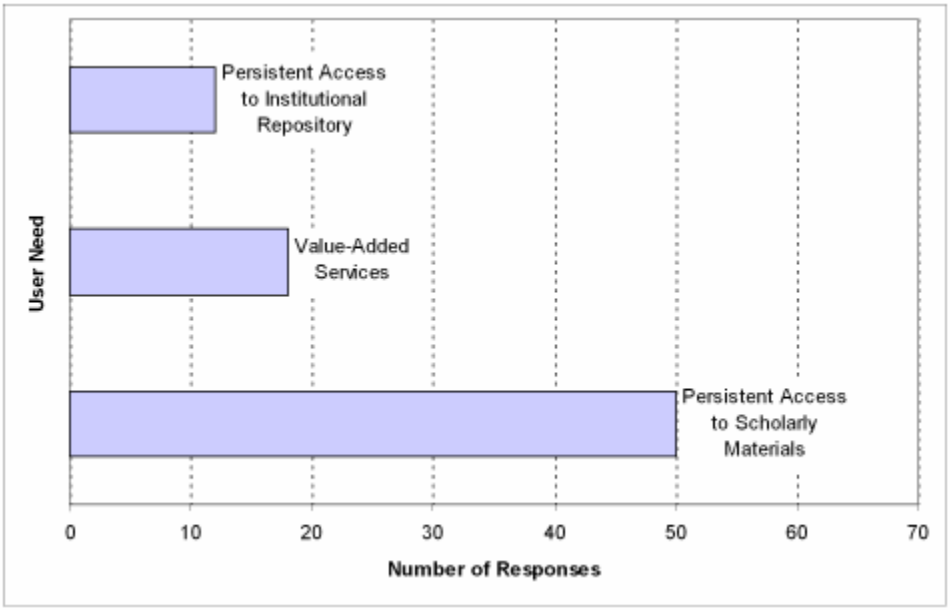

Figure 4. User Needs Addressed by a Web Archive

1. Persistent access to a wide-range of digital or web-born scholarly materials for research and reference

    a. Content consistent with collection parameters of the institution
        - Core electronic-only content
        - Born-digital materials from non-traditional publishers
        - Web sites and web-published materials about contemporary social movements
        - Unique and valuable digital collections
    b. Information and materials characterized as fleeting, ephemeral, non-standard, not previously published, or not commercially available
        - Scholarly materials from the institution's researchers and research centers
        - Web-published resources cited in faculty publications
        - Other materials created by the institution

c. Government publications characterized as critical, web-born, endangered, or fugitive
   - State and federal information from pre-web era
   - Information no longer maintained by government agencies
   - Web-born government documents published independently from the issuing agencies
   - Web-born government publications from first web publication date forward

d. Historical records to enable:
   - Historical research of web-based information
   - Unanticipated research needs

2. Provision of value-added services to satisfy user-defined needs

   a. Organization of content
      - Structured, well-organized information sources in subject concentrations
      - Historical access by subject
   b. Focused collections from diverse sources
      - Aggregation from multiple and disparate sources (e.g., newspapers from around the world)
      - Access to disparate digital collections
        1. Consistent interface to variety of numerical databases/datasets
        2. Interoperability with other repositories and indexes
   c. Friendly design for content discovery and access
      - Searchable content
      - Descriptive context for materials (e.g., author, publisher, and creation date)
      - Authentication and version control

3. Persistent access to an institutional or organizational repository to preserve its historical record, including its scholarship, for future research

   a. University and library web pages of long-term significance
   b. Faculty published scholarship (e.g., E-journals and articles)
   c. Faculty research projects

These needs were identified by librarians and archivists who participated in focus groups. Additional user needs that a web archive might address were identified by researchers. These included adding an indication of authenticity to materials, providing descriptions of the provenance of materials and the preservation activities undertaken, adding descriptive tagging of inactive links, and providing web site maps to enable a virtual reconstruction of web sites in the future. Taken together, these needs could form a springboard for librarians and other information professionals to articulate the benefits of collecting and preserving web-published materials for their user communities as well as for identifying the risks of not doing so. Campaigns within institutions for additional resources to address the challenges of collecting and preserving web-published materials might be able to translate these benefits and risks into selling points to present to administrators and funding agencies.

5.2 Partnerships

Looking to cut expenses and realign budgets, both universities and state governments are targeting libraries for downsizing and elimination. 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 Chris Rusbridge[3] comments:

> *The trouble is, it is a new cost, and we have not worked out how to factor it into our budgeting and business models. My guess is that in the long term, we will realize that print preservation is very expensive, while digital preservation is comparatively cheap.*

Partnerships could address preservation costs for web-published materials. Findings from the needs assessment led to models for both external and internal partnership opportunities.

External Partnerships

There seems to be 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 stewardship, these organizations often lack preservation 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 5.

This partnership model involves forming a community of creators among the partner organizations. These creators produce a range of web-published materials, which might include web sites, discrete text publications, maps or other materials. The institution acts as a service provider, offering repository services to partners and including their collections of web-published materials in a web collection registry, such as the registry discussed later in this document. The institution might offer a range of 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 would stipulate the terms and conditions of any services the institution would offer in service agreements. Submission agreements would address the roles and responsibilities of all partners regarding any materials deposited in the repository or captured by the service provider.

[3] Rusbridge, C. (2006, February). Excuse me . . . some digital preservation fallacies? Retrieved April 17, 2006 from the ARIADNE web site. http://www.ariadne.ac.uk/issue46/rusbridge/

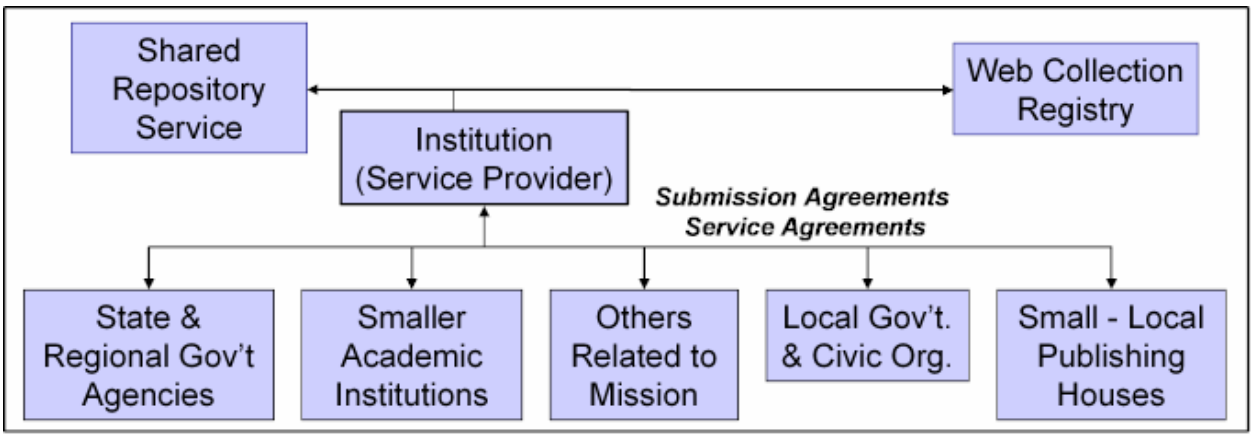

Figure 5. External Partnership Model

A primary motivation for university libraries to create such partnerships is to preserve the web-published materials these organizations create, in particular materials of importance to libraries' user communities. The benefits of such partnerships to an institution include:

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

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 6 depicts a model for internal partnerships.

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. Typically, the intellectual products of the university would be provided by research centers and faculty. Materials and records comprising 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.

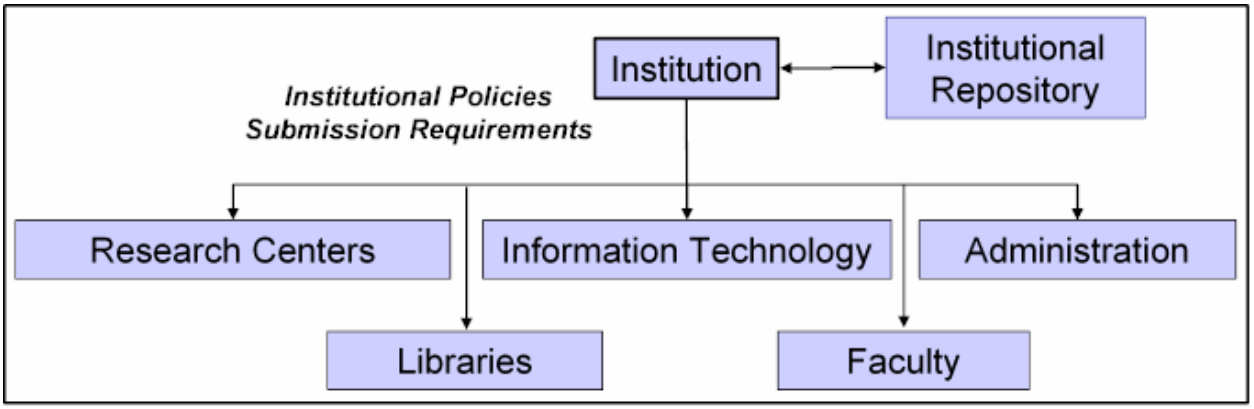

Figure 6. Internal Partnership Model

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:

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

5.3 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 user groups? Clearly it would be of value to create a shared directory or registry service for web collections. Small and medium institutions do not have the resources to engage in web collection activities yet could benefit from the work of others. Large universities would eliminate unnecessary duplication of effort in the preservation of web-published materials and maximize the preservation efforts of their 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[4] 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". This registry of digital masters would include the following information for such materials.

- Which library has the material
- Format of the material
- Terms of use

[4] Digital Library Federation: More Access at Less Cost: The Case for a Digital Registry; Updated March 23, 2006, Retrieved April 24 from http://www.diglib.org/collections/reg/regcase.htm

- Library or institution responsible for preserving the original source material
- Library or institution taking responsibility for preserving the electronic copy

Two registry efforts are underway for digitization projects dealing with government documents: the Government Printing Office (GPO) Registry of U.S. Government Publication Digitization Projects[5] and the American Library Association Government Documents Round Table (GODORT) Clearinghouse of Government Documents Digital Projects[6]. The GPO registry "contains records for projects that include digitized copies of publications originating from the U.S. Government" and the GODORT registry "provides information to librarians and others about digitization projects for local, state, federal, and international government documents".

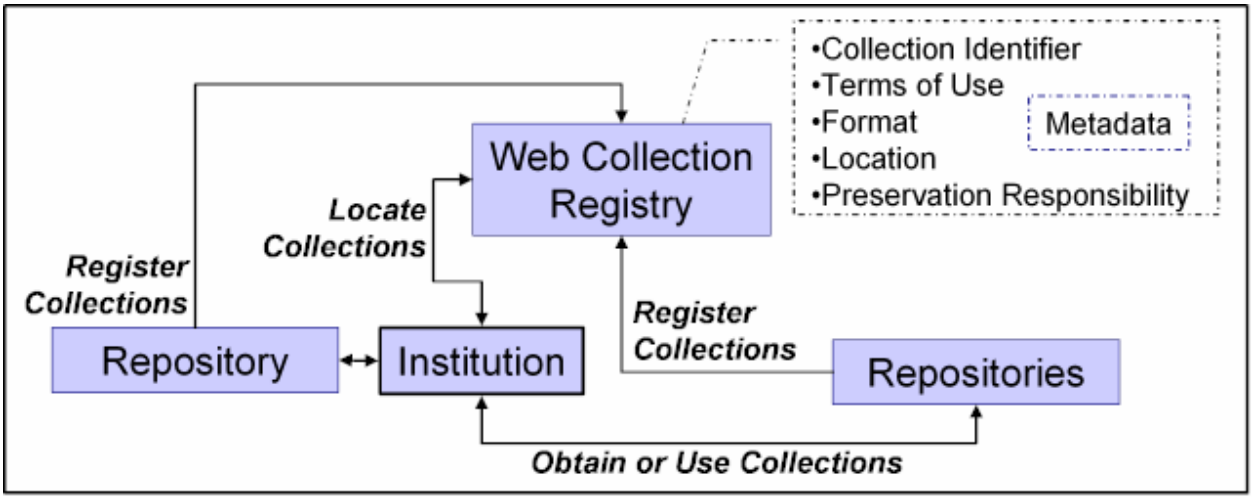

Figure 7. Registry Service Model for Web Collections

A registry service for web collections is depicted in Figure 7. The registry itself would contain standardized metadata records of web collections submitted by the organization or institution that had accepted preservation responsibility for the materials in the collection. Typically the organization would store collections 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 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.

5.4 Preservation Applications and Mandatory Deposits

Understanding the enormity of the preservation task for web-published materials, some participants suggested ideas that sought to drive responsibility for preservation to the

[5] GPO Access. (Updated January 20, 2006.) . Retrieved May 17, 2006 from http://www.gpoaccess.gov/legacy/registry/index.html
[6] GODORT. Clearinghouse of Government Documents Digital Projects. Retrieved May 17, 2006 from http://www.gl.iit.edu/services/ref/diggovclearinghouse.htm

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 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 their 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 is 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 mailto links and hyperlinks, provenance information, 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 should require applicants to submit a preservation plan as a prerequisite for funding approval. There is a model for this in the United Kingdom. The Arts and Humanities Data Service[7] (AHDS) receives material deposits from grant recipients as mandated by a number of funding organizations. 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.

Combining a deposit mandate codified in funding policies and organizational policies with preservation features in software applications for creators and preservation packaging

[7] UK: Arts and Humanities Data Service. Includes: *Why Deposit*, *How to Deposit* and *Waiver of Deposit*. Retrieved May 7, 2006 from http://ahds.ac.uk/depositing/index.htm

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.

6 Closing

Researchers and other end users of academic libraries and archives need persistent access to web-published materials that are at-risk of disappearing. Librarians and archivists want to preserve these materials in web archives or institutional repositories and are in need of a turnkey web archiving service with a simple, user-friendly, but efficient user interface. The challenge is to build the service curators need so that researchers and other users have access over time to the web-published materials they need.

Some librarians remarked that archiving the web was a daunting endeavor beyond the capability of libraries while others thought the technical hurdles were surmountable and in fact paled next to the resource demands and organizational challenges. A few participants expressed skepticism that a web archiving service that met their needs would emerge from grant-funded projects and many expressed reservations regarding ongoing support for a web archive within their institutions, particularly in regards to organizational commitment and funding for staff and infrastructure.

In the questionnaire completed by the librarians and archivists who participated in the Web-at-Risk project's focus groups (*N*=43) the seven factors in Figure 8 were identified as critical to the successful implementation of web archives in their organizations.

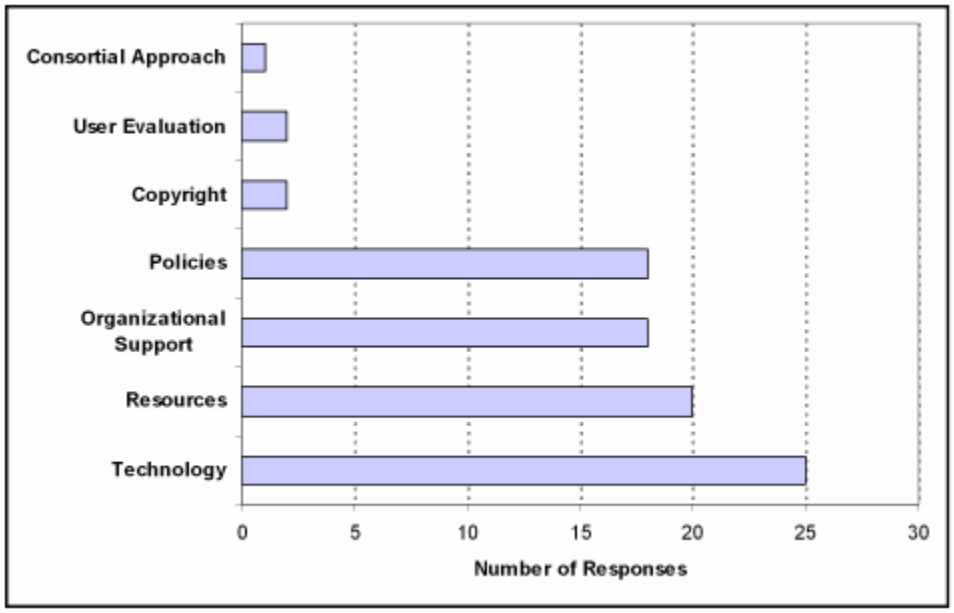

Figure 8. Implementation Success Factors for Web Archives

Reinforcing much of what was reported earlier in this report, the major success factors librarians identified were:

Technology:	Infrastructure; Support; Tools & Methods
Resources:	Money; Staff with Technical Expertise
Organizational Support:	Administration; Library; IT
Policies:	What to Archive; Selection; Library-wide in Scope

For a few organizations, understanding and dealing with copyright issues or collaborating with faculty and students to identify their needs was a factor. Only one anticipated that in the absence of a consortial project, their organization would not be able to implement a web archive.

For the Web Archiving Service under development by the Web-at-Risk project perhaps the touch stone for librarians and archivists is summed up in the following comment from one of the focus group participants.

> *"I know how to preservation-photocopy an item and have it bound and catalogued. I want to create a digital resource just as easily!"* - Librarian

While this librarian refers to creating digital resources, there is an urgent desire among librarians for preservation tools and collection building tools for web-published resources that will enable librarians to easily integrate those functions into their current work flow.

Skepticism and uncertainty generally accompany new technology trials and service introductions. The Web-at-Risk curators, as well as many of the librarians, researchers, and content providers that participated in the project's needs assessment, are early technology adopters helping the project's development team in the creation of a new service: a web archiving service. Nascent technologies and services enter the mainstream only after the early adopters have forged the path to more proven, perhaps even standard, products and services. With broader adoption of new technologies and services, a critical mass is achieved and business models that take advantage of economies of scale emerge.

Librarians see the need for a web archiving service today and are concerned about the absence of business models that will ensure the service thrives over the longer term. Their concerns are understandable. However, there are many preservation efforts afoot. The Web-at-Risk project is developing a Web Archiving Service that includes tools to assist curators with building and managing web collections. Many institutions are experimenting with preservation efforts and are creating their business cases and models for institutional repositories. All of these efforts will surely lead somewhere better than today's frustrating web preservation situation.

Appendix A. Data Collection & Analysis

Survey

Each curator was assigned a user name and password to access the online survey. Upon accessing the survey web site, participants were advised to print a hardcopy of the survey instrument to review, as necessary, with their colleagues before completing the survey online.

Prior to logging in participants were presented a consent letter. If users agreed to the terms of the survey as described in the consent letter, they were presented with a login screen. After logging in to the survey, they were presented with a second opportunity to print a hardcopy of the survey instrument as well as the opportunity to print a hardcopy of the glossary of terms used in the survey. Proceeding from this screen took the participants to the first section of survey questions.

Upon submission of each section, responses were stored in a MySQL database. If a participant was forced to abandon the survey for technical or other reasons, he or she could reenter the survey at a later time and would be positioned at the beginning of the last unsubmitted section. Participants were not permitted to re-access any submitted survey section.

Questions in each section of the survey were first analyzed individually. Where appropriate, response sets were removed prior to further analysis. For the most part, descriptive statistics (i.e., numbers and percentages of responses) were used to analyze the data.

Due to the small number of respondents and the categorical nature of most of the data, statistical calculations were used infrequently. In a few cases, Spearman's Rho was calculated to evaluate the relationships between responses to two separate questions. A significance level of .05 was required in each case.

Focus Groups

Each of the groups was facilitated by the Assessment Analyst for the Web-at-Risk project. Group discussions were generally one and one-half hour in length. The discussions were guided by the focus group discussion guide, which was developed in accord with the Collection Development Framework for Web Archives (Appendix H). After the participants introduced themselves, meetings generally began with a discussion of needs and issues relevant to the selection of web sites to archive and proceeded through topics associated with each subsequent collection development activity. At the conclusion of the group, participants completed a written questionnaire and were given a thank-you gift.

With the exception of the Chicago group, discussions were recorded and transcribed. Additionally, two note-takers attended each focus group and created a record of the discussion as well as a summary of key points.

Interviews with Content Providers

The interviews with content providers were conducted by project team members, who used an interview questionnaire to guide the discussion. Six topics were discussed:

1. Web-published Materials
2. Digital Archives
3. Access to Materials
4. Authenticity of Archived Materials
5. Intellectual Property of Archived Materials
6. Agreements with Archive Providers

Each topic provided information related to one or more of the web collection development activities.

Interviewers summarized the discussions 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 themes and issues.

Interviews with End Users

The interviews were conducted by project team members, who used an interview questionnaire to guide the discussion. Five topics were discussed:

1. Selection of Materials for an Archive
2. Authenticity of Archived Materials
3. Interacting with Materials in an Archive
4. Searching an Archive
5. Preservation of Archived Materials

Each topic provided information related to one or more of the web collection development activities.

Interviewers summarized the discussions 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 themes and issues. Three questions asked participants to select values that best matched their opinions. For each of the three questions, weighted sums were calculated to rank responses.

Appendix B. Individual Assessment Reports

ALL REPORTS

Web-at-Risk Project wiki at CDL - Assessment Reports
http://wiki.cdlib.org/WebAtRisk/tiki-index.php?page=assessmentActivities

Web-at-Risk Project at UNT - Reports
http://web2.unt.edu/webatrisk/delivs.php

INDIVIDUAL REPORTS

Needs Assessment Survey Report
http://web2.unt.edu/webatrisk/na_toolkit/Reports/survey_data_analysis_final_05Jan2006.pdf

Focus Group - ALA - Chicago - June 2005
http://web2.unt.edu/webatrisk/na_toolkit/Reports/ala_jun2005_fg_summary_final_16feb2006_r2.pdf

Focus Group - University of North Texas - Denton - August 2005
http://web2.unt.edu/webatrisk/na_toolkit/Reports/unt_aug2005_fg_summary_final_16feb2006.pdf

Focus Group - California Digital Library - Oakland - August 2005
http://web2.unt.edu/webatrisk/na_toolkit/Reports/cdl_aug2005_fg_summary_final_21mar2006.pdf

Focus Group - New York University - New York City - September 2005
http://web2.unt.edu/webatrisk/na_toolkit/Reports/nyu_sep2005_fg_summary_final_24mar2006.pdf

Focus Group - Federal Depository Library Conference - Washington DC - October 2005
http://web2.unt.edu/webatrisk/na_toolkit/Reports/fdlc_oct2005_fg_summary_final_28mar2006.pdf

Content Provider Interviews: Summary Report
http://web2.unt.edu/webatrisk/na_toolkit/Reports/cp_interview_summary_final_10apr2006.pdf

End User Interviews: Summary Report
http://web2.unt.edu/webatrisk/na_toolkit/Reports/eu_interview_summary_final_17apr2006_2.pdf

Appendix C. Participants – Survey

Web-at-Risk Project Curators: Public Policy and Political Movements	
Gabriella Gray	Curator Online Campaign Literature Archive Young Research Library UCLA
Ronald J. Heckart (collaborating with Nick Robinson)	Director Institute of Governmental Studies Library Institute of Governmental Studies UC Berkeley
Terence K. Huwe	Director Library and Information Resources Institute of Industrial Relations UC Berkeley
Peter Filardo (collaborating with Michael Nash)	Tamiment Archivist Tamiment Library New York University
Michael Nash (collaborating with Peter Filardo)	Head Tamiment Library & Robert F. Wagner Labor Archives New York University
Nick Robinson (collaborating with Ronald J. Heckart)	Librarian Institute of Governmental Studies Library Institute of Governmental Studies UC Berkeley

Web-at-Risk Project Curators: Local, State, Federal, and International Government Information	
Sherry DeDecker (collaborating with Janet Martorana)	Head Government Information Center Davidson Library UC Santa Barbara
Charles Eckman	Head Social Sciences Resource Center Green Library Stanford University
Valerie Glenn (collaborating with Arlene Weible)	Electronic Resources Coordinator Government Documents Department University of North Texas Libraries
James R. Jacobs	Local, State, and International Government Information Librarian Social Sciences and Humanities Library UC San Diego

Web-at-Risk Project Curators: Local, State, Federal, and International Government Information	
Kris Kasianovitz	Reference and Instruction Local and State Government Information Librarian Young Research Library UCLA
Amy Kautzman (handed over to Jim Church)	Head, Research Reference and Collections Doe/Moffitt Libraries UC Berkeley
Jim Church (in lieu of Amy Kautzman)	International Documents Librarian Doe/Moffitt Libraries UC Berkeley
Linda Kennedy (collaborating with Juri Stratford)	Head Government Information and Maps Department Shields Library UC Davis
Ann Latta	State and Local Documents Bibliographer Social Sciences Resource Center Green Library Stanford University
Janet Martorana (collaborating with Sherry DeDecker)	Local & California Documents / Environmental Sciences Librarian Davidson Library UC Santa Barbara
Lucia Orlando	Government Information Librarian University Library UC Santa Cruz
Richard Pearce-Moses	Director Digital Government Information Archives and Public Records Arizona State Library
Lynne Reasoner	Government Publications Librarian UCR Libraries UC Riverside
Juri Stratford (collaborating with Linda Kennedy)	Government Information Librarian Shields Library UC Davis
Yvonne Wilson	California and Orange County Government Information Librarian Langson Library UC Irvine
Arlene Weible (collaborating with Valerie Glenn)	Head of the Government Documents Department University of North Texas Libraries

Appendix D. Participants – Focus Groups

Table D1. Focus Group Participants by Location and Sector

Focus Group	College or University	Non-Profit Organization	State Government	Totals
ALA	5	3	0	8
UNT	7	0	0	7
CDL	11	0	0	11
NYU	8	0	0	8
FDLC	8	0	1	9
Totals	39	3	1	43

Notes. UNT: University of North Texas; CDL: California Digital Library; NYU: New York University; ALA: American Library Association; FDLC: Federal Depository Library Conference.

Focus Group Participants	
Kathy Amen	Government Information Librarian St. Mary's University, Blume Library San Antonio, Texas
Beth Arthur	Archivist National Association of Realtors Chicago, IL
Gayla Byerly	Reference Librarian Liaison, English & Women's Studies University of North Texas Libraries Denton, TX
Tim Byrne	Government Publications Library University of Colorado Boulder, Colorado
Danielle Cain	Electronic Resources Acquisitions Librarian Co-liaison, Philosophy University of North Texas Libraries Denton, TX
Angela Carreno	Collections Coordinator, Collections & Research Services Social Science Bibliographer Subject Specialist: Latin America New York University, Bobst Library New York, NY
Elizabeth Cowell	Librarian State and Local Government Information Stanford University Stanford, CA
Donna Davey	Librarian Tamiment Library and Robert F. Wagner Labor Archives New York University New York, NY

Focus Group Participants	
Harrison Dekker	Data Services Librarian University of California - Berkeley Doe/Moffitt Library - Social Science Data Berkeley, CA
Jackie Druery	Head, Donald E. Stokes Library for Public & International Affairs and The Ansley J. Coale Population Research Collection Wallace Hall, Princeton University Princeton, NJ
Teri Embrey	Librarian Pritzker Military Library Chicago, IL
Tracey Erwin	Geospatial Librarian Earth Sciences Library Stanford University Stanford, CA
Alicia Estes	Head, Business & Government Documents Center Subject Specialist: Business, Hospitality, & Tourism New York University, Bobst Library New York, NY
Paula Feid	Undergraduate Librarian, Undergraduate Services New York University, Bobst Library New York, NY
Peter Filardo	Archivist Tamiment Library and Robert F. Wagner Labor Archives New York University New York, NY
Eboni A. Francis	Resident Librarian The Ohio State University Food, Agricultural and Environmental Sciences Library Columbus, OH
Deborah B. Gaspar	Instruction and Collection Development Librarian Gelman Library The George Washington University Washington, DC
Dave Green	Associate University Librarian for Collections and Information Services Ronald Williams Library - Northeastern Illinois University Chicago, IL
Cass Hartnett	U.S. Documents Librarian University of Washington Libraries Seattle, Washington
W. Gerald Heverly	Reference Librarian, Humanities & Social Sciences Services Subject Specialist: Philosophy & Classics New York University, Bobst Library New York, NY

Focus Group Participants	
Susanna Hinojosa	Librarian Doe/Moffitt Library - State and Local Documents, Latin American Documents, Spanish & Portuguese University of California - Berkeley Berkeley, CA
Chuck James	Librarian & Information Services Manager Earthquake Engineering Research Center University of California - Berkeley Berkeley, CA
Debbie Jan	Head Public Health Library University of California - Berkeley Berkeley, CA
Leora Kemp	Head, Virtual Library, Dallas Campus University of North Texas Dallas, TX
Julie Linden	Government Information Librarian Yale University-Seeley G. Mudd Library New Haven, Connecticut
Elisabeth Long	Co-Director, Digital Library Development Center University of Chicago Library Chicago, IL
Scott Matheson	Reference and Government Documents Librarian Yale University Law Library New Haven, Connecticut
Marcia Meister	Bibliographer Shields Library - Federal Government Information University of California - Davis Davis, CA
Jo Monahan	Liaison, College of Education University of North Texas Libraries Denton, TX
Sue Parks	Head, Media Library Liaison, Radio, Television & Film University of North Texas Libraries Denton, TX
Colleen Parmer	Chair, Collections and Technical Services Head, Government Documents Bowling Green State University Libraries Bowling Green, Ohio
Jason Phillips	Reference Associate, Humanities & Social Sciences Services Subject Specialist: Sociology, American Studies, Gender & Sexuality Studies New York University, Bobst Library New York, NY

Focus Group Participants	
Aimee Quinn	Asst. Professor, Richard J. Daley Library Assistant Documents Librarian University of Illinois-Chicago Chicago, IL
Missy Roser	ICON Project Coordinator Center for Research Libraries Chicago, IL
Ann Sanders	Head, Government Documents Michigan State Library Lansing, Michigan
Beth Sibley	Librarian Doe/Moffitt Library - Political Science, Sociology, & Women's Studies University of California - Berkeley Berkeley, CA
Bill Sleeman	Asst. Director for Technical Services Coordinator, Collection Development University of Maryland Law Library Baltimore, Maryland
Martha Tarlton	Head, Reference & Information Services Liaison, Merchandising, Hospitality Management University of North Texas Libraries Denton, TX
Linda Vida	Director/Head Librarian University of California - Office of the President Water Resources Center Archives Berkeley, CA
Jennifer Vinopal	Reference Librarian, Humanities & Social Sciences Services Subject Specialist: French & Italian Language & Literature Interim Manager, Digital Library Program New York University, Bobst Library New York, NY
Will Wheeler	Curator Social & Behavioral Sciences Stanford University Stanford, CA
Gay Woods	Head, Research Park Library Liaison, Material Science, Engineering Technology, & Mathematics University of North Texas Libraries Denton, TX
Alice Youmans	Head of Reference Boalt Law Library University of California - Berkeley Berkeley, CA

Appendix E. Participants – End User Interviews

Table E1. End Users by Discipline

Academic Positions & Disciplines
1. Doctoral Candidate (ABD) - Political Science 2. Assistant Professor - Political Science 3. Assistant Professor - 20th Century American History 4. Assistant Professor - History of American Business & Labor 5. Associate Professor - African America & Labor History 6. Professor - History of the Jewish American Left 7. Professor - Hospitality Law & Management

End Users	
Jim Battista, Ph.D.	Assistant Professor - Political Science University of North Texas
Joan M. Clay, Ph.D.	Professor - Hospitality Management (Business & Law) University of North Texas
William Jones, Ph.D.	Associate Professor - History (African American & Labor History) University of Wisconsin - Madison Scholar in Residence - Schomburg Center for Research in Black Culture, New York Public Library
Tony Michels, Ph.D.	Associate Professor - History (History of the Jewish American left) University of Wisconsin Scholar in Residence - Tamiment Library & the Goldstein-Goren Center for American Jewish History at NYU (2005-2006)
Todd Moye, Ph.D.	Assistant Professor - History (20th Century American History) University of North Texas
Gerhard Peters, Doctoral Candidate - ABD	Graduate Student - Political Science UC Santa Barbara
Kimberly Philips-Fein, Ph.D.	Assistant Professor - History (History of American Business & Labor) NYU Gallatin School for Individualized Study

Appendix F. Participants – Content Provider Interviews

Table F1. Content Providers by Type

Organizations Interviewed
Labor Unions 1. United Federation of Teachers 2. Transport Workers Union of America 3. American Federation of State, County, & Municipal Employees District Council 37 - New York City
State Government Agencies 4. CA Spatial Information Library & CA Environmental Resources Evaluation System* 5. California Legislative Data Center 6. Office of the Texas Secretary of State 7. Texas Building & Procurement Commission

* Interview was with the UC researcher who maintains these two agency web sites.

Content Providers	
Bill Behnk, Coordinator of the Legislative Information System Linda Heatherly, Librarian in the Office of the Legislative Counsel Three programmers	California Legislative Data Center
Tyrone Butler, Archivist & Records Manager Tom Dickson, Assistant Archivist	United Federation of Teachers
Eva Dechene, Records Management Officer Vice-Chair, Records Management Interagency Coordinating Council	Texas Building and Procurement Commission
Quinn Hart, CERES Technical Researcher, CaSIL Developer	CaSIL (California Spatial Information Library) CERES (California Environmental Resources Evaluation System)
David Paskin, Director of Research	American Federation of State, County, & Municipal Employees District Council 37 - New York City
Dan Procter, Director - Texas Register Chair, Records Management Interagency Coordinating Council	Office of the Texas Secretary of State
Dr. Robert Wechsler, Director - Education and Research	Transport Workers Union of America

Appendix G. Glossary

Acquisition	For web-published materials, see Capture
Archive	Archives are repositories of content for which someone or some organization has accepted preservation responsibility. See also Digital Archive and Web Archive
Authenticity	The genuineness of a digital object. Verification of authenticity requires ascertaining that the object is what it claims to be or is what the metadata associated with the object asserts it to be. Authenticity of a digital object is determined in several ways including provenance and digital signatures.
Automated Capture Tool	See Crawler
Baseline Metadata	Baseline metadata is machine-generated and captured by a crawler at the time of data capture.
Born-digital	Created originally in digital format (i.e., a machine-readable format). Examples include scientific databases, electronic documents, web pages, sensory data, digital photographs, and digital audio and video recordings. A born-digital resource may or may not have a counterpart analog format but, if it does, the digital version existed prior to the counterpart.
Capture	The process of copying web-published materials from their source locations for collection or archive purposes or the web-published materials copied as the result of that activity. For the Web-at-Risk project, a capture is specified by a list of one or more seed URLs in conjunction with parameters controlling the capture activity itself.
Collection	A group of resources related by common ownership or a common theme or subject matter. Collections are owned and/or maintained by an organization, an institution, or an individual.
Crawl	The activity conducted by a web crawler.
Curation Process	Collection development for web-published materials includes the selection, curation, and preservation processes. In this context, the curation process involves description, organization, presentation, maintenance, and deselection of the materials in the collection.
Dark Archive	A digital archive to which no end user access is permitted.
Deep Web	Resources available via the World Wide Web that are invisible to or inaccessible by crawlers because they (a) are contained in a database or other data store, (b) require information collected from the end user before they are created, or (c) are password protected.
Digital Archive	A collection of digital objects that may also exist in other forms. The digital archive preserves the digital versions for posterity and provides access to them.

Digital Collection	A collection consisting entirely of born-digital or digitized materials.
Digital Material	See Digital Object
Digital Object	Digital objects include interactive works such as video games, sensory presentations such as music, documents such as articles, and data such as datasets. Two types of digital objects included in digital archives are: surrogate objects, for example digitized copies of print books or audio tapes, and born-digital objects.
Digital Resource	See Digital Object
Dynamic Web Page	A web page created automatically by software at the web server. The page may be (a) personalized for a user based on identification via login or based on cookies stored on a user's computer, (b) tailored to fulfill a specific request made by a user, or (c) code-generated (e.g., using php, jsp, asp, or xml). Information used for personalization or tailoring of pages may be retrieved in real-time from a database or other data store.
Emulation	A method by which newer software interacts with older resources and displays the result using the same commands and formatting that the software that created the resource used. Emulation provides a means of allowing a digital resource to be preserved without altering its binary format.
Enriched Metadata	Enriched metadata is generally specific to an organization and contains a mixture of baseline metadata and human-generated metadata added subsequent to data capture.
Entry Point URL	See Seed URL
External Link	A URL that links to web-published materials residing on a different host.
Fixity	The extent to which an archived object remains unchanged over time regardless of access and movement due to copying. One common fixity mechanism used to establish and protect the integrity of a digital object is the result of a cyclical redundancy check (CRC). Redundancy checks are sometimes referred to as checksums.
Format	Refers to specific encoding schemes for the contents of a digital object and is frequently designated in the extension of a file, for example, html, jpeg, gif, PDF, etc.
Harvest	See Capture
Information Object	See Digital Object
Ingest	For the Web-at-Risk project, ingest refers to the process of packaging captured materials and moving them to the repository for long-term storage.

Institutional Repository	Comprised of digital collections representing the intellectual output of a single university or a group of colleges and universities. An institutional repository captures, preserves, and provides access to these collections as a logical extension of the core mission of the university and as a vehicle for increased institutional visibility.
Integrity	A digital object's integrity is maintained as long as the bits contained in the object are not altered in an unauthorized manner. See also Fixity
Light Archive	A digital archive accessible to end users.
Medium	The delivery vehicle for content. For example: CD-ROM, network, book, etc.
Migration	A method of preserving digital materials and access to those materials by copying or reformatting the materials while preserving their intellectual content.
Opt-in	A collection policy in which the archive owner seeks explicit permission from content owners before collecting materials.
Opt-out	A collection policy in which the archive owner automatically collects materials, assumes preservation responsibility for the materials, and makes them available for use unless one of the following occurs: (a) The owner of the content requests that their content be removed from the archive and that their content not be included in future collection efforts or (b) the owner of the content blocks the content from crawlers using robots.txt or Meta tags.
Persistent Name	A unique name assigned to a web-based resource that will remain unchanged regardless of movement of the resource from one location to another or changes to the resource's URL. Persistent names are often resolved by a third party that maintains a map of the persistent name to the current URL of the resource.
Repository	A repository is an umbrella term for the physical storage location and medium for one or more digital archives. A repository may contain an active copy of an archive that is accessed by users or a mirror copy of an archive that has been replicated for disaster recovery.
Seed List	One or more seed URLs from which a web crawler begins capturing web-published materials. Curators, or others responsible for building collections of web-published materials, specify seed lists for specific crawls.
Seed URL	A URL appearing in a seed list as one of the starting addresses a web crawler uses to capture content.
Spider	See Crawler
Targeted URL	See Seed URL

Trusted Digital Repository	A repository, built by either a single institution or multiple institutions, that accepts responsibility for the long-term maintenance of digital resources for depositors and provides reliable, long-term access to those resources for its users. Some institutions may contract with a third-party for storage and maintenance while retaining management of the logical and intellectual aspects of a repository.
Type	Material types include such things as text, image, audio, video, and application-specific data types. A material type may be encoded in one of several formats (e.g., an image may be encoded as gif, jpeg, tiff, etc.).
Visibility	The extent of end user access allowed to a digital archive.
Web Archive	A web archive contains web-published materials for which an organization has accepted long-term responsibility for both preservation and access. Organizations, for example, national libraries, research institutions, or professional societies, may build web archives to fulfill their stated mission and to satisfy the information needs of their user community. Alternatively, organizations may enter into service arrangements with third-party archive providers or archive agencies. A web archive is a special case of a digital archive.
Web Crawler	Software that explores the web and collects data about its contents. A web crawler can also be configured to capture web-published materials. It starts a capture process from a seed list of URLs.
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. The application of the intellectual and logical processes involved in collection management by librarians and archivists results in curated web collections. All web collections residing in a web archive are assumed to be preserved.
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). The web pages within a web site are often published and maintained by a single person or organization, although wider collaborations and social publishing are becoming common, for example, wikis and blogs. Hyperlinks in the form of uniform resource locators (URLs) on web pages access other web pages and specific web-published materials either within the same web site or at a different web site.
Web-based Resources	See Web-published Materials

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. Web-published materials are both dynamic and transient. They are at risk of disappearing. Web archives preserve web-published materials. All web-published materials are digital objects.

Appendix H. Collection Development Framework for Web Archives

POLICY SETTING	Policy factors influencing web archiving include political mandates, organizational mission, financial parameters, and technical capabilities.	
	SELECTION	
	Selection	Choice of web-published materials for archiving is impacted by the focus of the collection, unit of selection, web boundaries, copyright obligations, and authenticity of materials.
	Acquisition	Web-published materials are acquired or captured using crawling tools, which either globally or selectively capture web-published materials.
	CURATION	
	Description	Baseline metadata is machine-generated and gathered by a crawler at the time of data capture. Enriched metadata is generally specific to an organization and contains a mixture of human-generated metadata added subsequent to data capture as well as machine-generated metadata.
	Organization	Digital archives of web-published materials typically either retain the organizational structure of the materials as they existed on the web at the time of capture or modify the organizational structure to suit the archive's mission or constraints.
	Presentation	Presentation of web archive materials is related to how the content was captured and to post-harvest descriptive and organizational analysis. For example, archived materials might mirror the web at the time of their capture or might be categorized in accord with selection criteria, such as image files presented by subject.
	Maintenance	Several maintenance functions are critical to ensuring the successful use of materials in web archives: software and hardware training for archive support staff; hardware and software maintenance, performance optimization, backups, and upgrades; and duplicate detection.
	Deselection	Removal of materials from a web archive can be for several reasons: duplication, errors, legal or social considerations (e.g., offensive materials). Risks of removal and retention are weighed against policy and storage costs.
	PRESERVATION	
	Preservation	Preservation challenges are numerous. They include persistent naming, format migration and/or emulation, inventory management, volatility, replication, re-validation, curator-operator error, and storage.

Appendix I. Lost Materials

- Materials included in subject lists at academic libraries:
  - A health information bibliography
  - Materials produced at academic institutions
- University publications:
  - Materials created by university research centers
  - Working papers of researchers
- Data sets published on the web
- Association conference papers:
  - ASA - the American Sociological Association
- NGO publications
- Materials on federal agency web sites:
  - USGS
  - NOAA
  - FEMA
  - US Department of Interior
- Materials on state agency web sites:
  - State water resources reports/publications
  - State public health agency publications
  - Annual county-level statistical report
  - When agencies reorganize
  - At the change of administrations
- Materials of regional offices of federal agencies
- Materials of regional offices of federal agencies publishing materials in collaboration with:
  - Counterpart state agencies
  - University research centers

Appendix J. What to Preserve

I. Government Information

A. Information Sources
- Agency web sites
  - Web sites of original publications from government agencies and non-government organizations
  - Web sites of reformatted government publications available from private publishers
  - Documents & other publications
  - All versions of publications
  - Both print and digital formats

B. National
- Agency Examples:
- USGS
- NOAA
- FEMA
- Department of Interior
- Department of Labor
- Material Examples:
- Supreme Court briefs
- Administrative offices of US courts
- Congressional bills (1873 - 1937)
- Census data

C. State
- Water resources control boards
- Public health agency publications
- State budget
- Legislative committee memberships
- Statistical Abstracts
- State government "essential titles" list
- Documents from the state-wide shared cataloging project

D. Regional & Local
- Web-based publications of regional offices of Federal agencies
- Example: Sacramento Region of the Army Corps of Engineers
- Web-based publications resulting from collaborative efforts by regional offices of Federal agencies, their counterpart state agencies, and university research centers
- Web-published information on specific topics of local interest that may cut across state agencies, local agencies and outside advocacy groups
- State gambling issues
- State electronic voting
- Regional environmental issues:
  - Regional water quality control boards
  - Environmental Impact Statements (EIS) for a region
- Soil survey reports and maps
- Geospatial data
- Materials produced by local communities

II. Information in Support of an Academic Institution

A. Materials in Support of Teaching and the Curriculum

- Discipline-specific web publications from:
    - Faculty course materials
    - Subject lists created by librarians
    - Trusted web sites:
        - Other universities
        - Government web sites
        - Research institutes, associations, and organizations:
        - Information about regions or towns
        - Foreign, regional and local information
        - Biographies
- Some disciplines rely more on web-published source materials. This may be related to either the 'newness' of a discipline, the currency of its information, or the nature of its reference and resource materials.
    - Cultural and political studies
    - Web sites of current social organizations and activities, often include historical materials as well
    - Women's Studies
    - Criminal Justice: Databases and government web sites
    - Health Information: Databases and government publications
    - Film, Radio, & Television: Databases and web sites
- Some disciplines primarily augment their source materials with web-published materials and web sites.
    - Classics: Collections of ancient scientific images
    - Philosophy: Collections of pre-published materials
        - Philosophy and Classics:
        - Textual materials for different types of primary sources (e.g., books or glossaries)
        - Scholarship written about textual materials
    - Sociology: Statistical databases and publications (e.g., education and vital statistics)
    - Archeology: Excavation sites for particular digs
- Materials pertaining to topics and issues in support of both teaching and scholarship, for example:
    - Federal tax reform initiative
    - Immigration
    - Social welfare movement in American history & social work
    - Trade unions

B. Materials in Support of Scholarship

- Electronic resources
    - E-journals (both licensed and unlicensed)
    - E-books
    - Databases
    - Data sets
- University research center publications after they cease to exist
    - Web sites
    - Working papers from faculty and students
    - Newsletters

        - Data: Statistical and other
    - Scholarly publications of university faculty and students
        - The range of an institution's scholarly products that reflect new and emerging publishing methods and vetting mechanisms
        - Articles by faculty in e-journals
        - Student papers and publications
        - E-journals published by the university faculty
    - Web-published source materials citied in scholarly research, including:
        - Publications of public policy groups
        - Non-government organization (NGO) publications
        - Political party information and publications
        - Datasets
        - Electronic resources (e.g., journal articles)
    - Content from small publishing houses or society publishing houses
    - Association conference papers that are neither preserved by the association nor published and preserved in other publications.
    - Web materials identified as 'lost' by academic researchers and library patrons

C. Materials in Support of University Operations

    - University web sites
        - Main web site
        - Library web sites
        - Other web sties (organizations, departments, etc.)
    - Materials of long-term value
    - Locally published unique collections
        - Legacy collections

III. Information Pertaining to Key Events

- Changes in government administrations: Federal and State
- Events that emerge amid lots of media attention
- Grass-root information sources related to key events
- Examples:
    - Bush/Gore vote count in Florida
    - Howard Dean's web-based campaign
    - Labor union strikes and contract negotiations

IV. Information Pertaining to Organizations

- Organizational publications & resources:
    - Organizational web sites
    - Organizational membership lists