Collection Planning Guidelines

May 31, 2006

Prepared by:

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

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

Contents

1	Introduction	3
2	Organizational & Policy Considerations	8
3	Creating a Web Collection Plan	18
4	Mission & Scope	21
5	Selection Activities	23
6	Web Site Acquisition	26
7	Descriptive Metadata Creation	29
8	Presentation and Access Requirements	31
9	Maintenance and Weeding	35
10	Preservation	37
11	Collection Plan Appendices	39
Appendix A.	Web Collection Plan Outline	41
Appendix B.	Glossary	45
Appendix C.	Preservation Projects: Selected Examples	49
Appendix D.	International Consortia	51
Appendix E.	Reference Model & Key Standards	53
Appendix F.	Compiled Collection Planning Resources	60

1 Introduction

Archives are repositories of content for which someone or some organization has accepted preservation responsibility. All manner of organizations have archives, including libraries, businesses, governments, and universities. The content in physical archives varies widely, from papers to artifacts, from pens to airplanes. Archivists typically view the content of their archives as consisting of one or more collections, which contain materials that are related in some fashion. In a like-manner, archives themselves may contain a number of related collections. Archivists develop and maintain descriptive inventories of their collections as well as records about the provenance of each collection. Archivists also ensure that access to their collections is in accord with both best practices within their profession and with the terms specified by those who donate collections. Further, archivists are responsible for making selection and deaccession decisions about their collections.

A *web archive*, whose contents are comprised of web-published materials, shares much in common with a traditional archive containing physical materials. Implicit in the designation "archive" are the archive agency's responsibilities for preserving the integrity of the contents over time and for permitting access in accord with legally binding arrangements, such as copyrights and agreements with content providers. Like traditional archives, the content of web archives includes a range of materials, from discrete objects such as digitally-formatted text documents to aggregates of related objects such as web sites. Curators of web archives have many similar responsibilities regarding collections of web-published materials within web archives as archivists have with their collections of physical materials. These include describing the collections and their contents and ensuring compliance with access restrictions. As with archives of physical materials, it is possible that some content in web archives might be preserved and that no access might be allowed, apart from the archive agency preserving the materials.

Collection management responsibities for all material types, including web-published materials, can be formalized in collection plans. Collection plans describe the activities necessary to create and manage a collection of materials for a specific user group or entity within an organization or a library, such as a particular agency within state government or a specific discipline within a university. Most collection plans address the preservation of materials in the collection. Web-published materials have unique preservation requirements that often cannot be met by a library. It may be that an organization has an archive that can preserve web-published materials in its libraries' collections. In this case the organization's archive policy will provide guidance to curators as they describe preservation activities in their collection plans. If an organization does not have an archive in which to preserve the web-published materials in their libraries' collections, the libraries might contract with an outside agency for archive services.

The Web-at-Risk project is developing a Web Archiving Service (WAS) that will enable the project's partner institutions to act as archive agencies that will assist the project's curators in building and managing archived collections of web-published materials. With the exception of one curator who works in a state library, the project's curators work in large academic libraries. Many of them work in government information departments while others are subject specialists in the areas of public policy, trade unions, and political movements. All of the curators have collection management responsibilities and select print materials, electronic resources, and web-published materials for their collections. However, most of the project's curators do not currently have plans in place for managing and preserving web-published materials. The guidelines in this document are intended to assist the project's curators in developing plans for the collections they will create using the project's

Web Archiving Service. The guidelines may also be useful to librarians, archivists, and curators who are not involved in the Web-at-Risk project.

It is helpful to note that some libraries refer to collection plans as collection policies. Also, some pioneer web archiving programs have created specific preservation policies for web-published materials. Local practice will specify the proper document in which to address the content areas discussed in these guidelines. Librarians will notice that some familiar concepts and practices from collection planning for print materials easily transfer to collection planning for web-published materials while some new concepts and unfamiliar practices are introduced. To effectively manage collections of web-published materials, it is good practice to either create new plans or modify existing collection plans to address these concepts and practices.

1.1 Overview of Contents

The remainder of section 1 briefly describes web archives and the Web-at-Risk project's Web Archiving Service. Section 2 of this document discusses several factors to consider and to address as appropriate in collection policies.

Section 3 identifies the key areas to include in collection plans for web-published materials. Sections 4 - 11 describe in more detail each of these areas of a collection plan. Throughout the guidelines, applicable resources and references are provided. These resources are compiled in Appendix F. Lastly, several appendices are included for background and reference.

Appendix A. Web Collection Plan Outline
Appendix B. Glossary
Appendix C. Preservation Projects: Selected Examples
Appendix D. International Consortia
Appendix E. Reference Model & Key Standards
Appendix F. Compiled Collection Planning Resources

1.2 Web Archives

A web archive contains web-published materials for which the archive agency 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 own communities. Alternatively, organizations may enter into service agreements with web archive agencies with the intention of preserving web-published materials of interest and value to their organizations. Such agreements identify the materials to be archived and delineate service terms, responsibilities, expectations, and fees for both the archive agency and the organization requesting archive services

Content
There are several approaches to identifying the scope of web-published materials for a web archive. The National Library of Australia[1] broadly defines the following four models:

1. Whole domain or comprehensive	Preserves a national or global web space *Example*: WayBack Machine - The Internet Archive [http://www.archive.org/web/web.php]
2. Selective	Preserves "defined portions of Web space or particular kinds of resources according to specified criteria" *Example*: PANDORA - Australia's Web Archive [http://pandora.nla.gov.au/]
3. Thematic	A form of selective collection which preserves content relating to a particular theme or event Example: MINERVA - Library of Congress Web Archiving Project [http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html]
4. Deposit	Preserves only materials deposited by publishers based on legal or voluntary deposit codes *Example*: Electronic Collection - Library and Archives Canada [http://www.collectionscanada.ca/electroniccollection/003008-220-e.html]

The National Library of Australia also notes that "a growing number of Web archiving programs are concluding that no one archiving model is entirely satisfactory for preserving national online heritage." As a result, many programs are using a combination of two or more of the above models.

Thus a web archive may preserve a range of web-published materials obtained in a number of different ways. For example, archived content may consist of a collection of thematically-related web sites that are captured by a web crawler and it may also include discrete web-published materials such as documents that are electronically deposited by their creators or publishers.

Access
An archive agency allows access in keeping with a web archive's user-access policies, which specify access rights to the web archive's content, including stipulating which materials are being preserved but are not accessible. There is no one mechanism for user access to a web archive's content. Common methods include searching by keyword, URL, or other criteria as well as browsing by subject categories. While cross-archive discovery of web-published materials is currently non-existent, it seems a likely future direction. For example, there are efforts underway in the digital library and government documents realms to create standard registries of materials in digital collections. These registries may provide standardized descriptions of digital collections and their contents that will enable discovery of materials across a number of digital collections. These registry models may logically extend to include

[1] National Library of Australia. (n.d.). *Web archiving*. Retrieved May 6, 2006, from http://www.nla.gov.au/padi/topics/92.html

web-published materials in web archives, enabling cross-archive discovery of web-published materials.

Collections
A collection within a web archive typically consists of a group of related web-sites but might also refer to a group of discrete web-published materials, such as a set of public policy documents related to a common subject. All collections residing in a web archive are assumed to be preserved in a manner that ensures their integrity over time and provides access to them.

Librarians and curators are familiar with building collections comprised of a range of material types. The process used to define a collection of web-published materials may depend on the archive agency and the services they offer. Some agencies might require collections to be defined prior to the acquisition of content. Other agencies may stage content after acquisition to permit curators to refine the collection definition and subsequently store the materials in the web archive. Still other agencies may provide services that enable collection definition from any materials within a web archive, as long as there are no legal or access restrictions. The Web-at-Risk project's Web Archiving Service described below represents a specific implementation of an archiving service and therefore may have characteristics that differ from other archiving service implementations.

In the future it may be possible for librarians and curators to use cross-archive registry-enabled tools to discover and evaluate web-published materials for inclusion in collections. These collections may be characterized as shared or as virtual in that not all materials in the collection will be owned by or even licensed by the library or organization building the collection. Neither will all the materials reside in the library or organization nor will the library or organization have responsibility for maintenance and preservation of the materials residing in other organizations' web archives.

1.3 The Web Archiving Service (WAS)

The Web-at-Risk project is building a Web Archiving Service (WAS), which will consist of new services and tools that integrate with and take advantage of the overall framework and resources for application development and data storage within the California Digital Library. From July 2006 to November 2007, the WAS tools will be released in stages as major functionality is implemented. The tools will enable the project's curators to specify web-published materials represented by specific URLs whose content they would like captured and preserved in the web archive provided by the WAS.

Curators will also use the WAS tools to build and manage collections of materials within the web archive. It is anticipated that these collections will be comprised of a set of related web sites stored in the archive as a result of being captured from their original web locations, for example a collection might be defined as a specific set of government and organizational web sites related to water conservation for a particular geographic area or as a set of web sites reflecting a range of perspectives related to federal immigration policy. It is conceivable that collections might be comprised of any captured web sites in the WAS archive, regardless of which curator initially requested their capture. The only caveat is that copyrights and legal arrangements pertaining to the web sites must be honored. From a system's perspective, a collection within the WAS archive consists of a set of index entries that point to captured copies of web-published materials. From a user perspective, the collection consists of the set of web sites. Curators need to understand both perspectives.

It may not be possible for curators to define collections comprised of web-published materials at a more granular level than captured web-sites. For example, a curator may request that the content of municipal government web sites be captured and stored in the WAS archive but a curator may not be able to define a collection composed of selected web-published material from within the archived web sites, such as a collection of the building code publications for the municipalities in a defined geographical region. Likewise, collections of web sites that rely on databases for their content or server-side code for their operation are beyond the planned scope of collections to be built with the WAS tools.

2 Organizational & Policy Considerations

Collection plans within a library generally articulate the role a collection has within the organization and articulate the organization's commitment to building such collections. Often collection plans identify policies, guidelines, and standards that affect collections. These might include technical standards regarding the format of web-published materials suitable for collections, web archive policies regarding the metadata required for web-published materials in the organization's web archive, or library guidelines regarding copyright clearances required for web-published materials. To successfully develop collections of web-published materials, it is important for an institution to develop policies and guidelines that support collection management activities. Such policies and guidelines will need management endorsement as well as committed support from all units and people involved in selection and ongoing maintenance of the institution's collections.

Librarians and archivists who participated in the Web-at-Risk project's needs assessment focus groups in 2005 (N=43) identified the seven factors in Figure 1 as critical to the successful implementation of web archives in their organizations. Whether a library plans to create its own web archive or utilize an external archive service, each of these factors should be explored prior to creating collection plans for web-published materials. Doing so should help libraries identify critical areas where policies or practices need to be established. The remainder of Section 2 briefly discusses each factor.

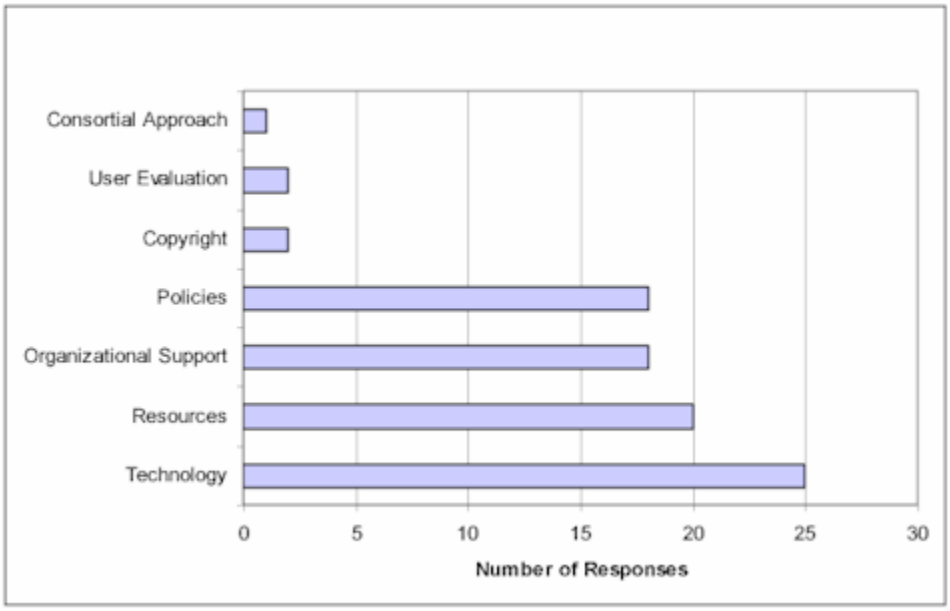

Figure 1 – Critical Success Factors for Web Archive Implementation

2.1 Consortial Approach

Libraries have a rich tradition of collaborations and consortial efforts that provide models for sharing the preservation responsibilities for web-published materials. While "preserving the Web" is an endeavor beyond the mission and resources of individual institutions and their libraries, identifying web sites that support an institution's mission and establishing procedures to preserve them is of importance to most institutions.

2.1.1 Areas to identify:

- Statement of the mission of the institution
- Statement of the institution's areas of scholarship and curriculum
- Existing consortial relationships and memberships

2.1.2 Questions to address:

1. What existing collaborative or consortial arrangements already exist that might be appropriate to involve in web archiving efforts?
    a. Within the university system
    b. Within the library community
    c. With government agencies
    d. Others
2. What new collaborations are needed with organizations external to the institution to promote preservation of web-published materials that support the mission of the institution, its research and its teaching?
3. What new collaborations are needed within the institution and is there a model for these?

2.2 User Evaluation

2.2.1 Identifying User Groups

A fundamental task in the formulation of a policy for the preservation of web materials is to define the user groups for whom information is being collected and preserved. The *Reference Model for an Open Archival Information System*[2] (OAIS)[3] refers to these groups as a *Designated Community*. Because a Designated Community may consist of disparate user groups, it is important to identify each of the user groups and evaluate their specific needs in regard to archived web-published materials.

Predictions about how a Designated Community might change over time should also be considered. For example, are other groups likely to become part of the Designated Community in the future? Finally, periodic reevaluation of both the user groups comprising a Designated Community and the changes that occur to their knowledge base over time should occur. For example, how have the terms and vernacular used by the user groups evolved?

2.2.2 Involving User Groups

Members of the Designated Community should initially be consulted to identify their information needs. Subsequently the community should be engaged in evaluating both collections of web-published materials and a web archive's effectiveness in regard to meeting the community's needs.

Within an academic institution, user groups will likely include researchers, faculty, students, members of the public, administrative staff, and alumni. Each of these user groups would be part of an institution's Designated Community but their unique information needs might predicate different requirements in regard to:

---

[2] Consultative Committee for Space Data Systems. (2002). *Reference model for an open archival information system (OAIS)* (CCSDS Publication No. 650.0-B-1). Retrieved April 27, 2006, from http://public.ccsds.org/publications/archive/650x0b1.pdf

[3] See Appendix E for a brief introduction to the OAIS.

- depth of the collection
- frequency with which materials should be captured
- level of descriptive metadata
- discovery mechanisms required in the user interface
- manner in which materials should be presented
- expectations as to preservation of the collected materials
  - certification of authenticity
  - format migration
  - retention of versions

2.2.3 Areas to identify

- Designated Community for the collection
- Needs of the Designated Community
- Frequency of refining user groups in the Designated Community
- Expectations regarding on-going evaluation

2.2.4 Questions to address

1. Who will use collections within a web archive?
2. How will they use the collections?
3. What are the salient characteristics of each user group in the Designated Community?
4. What additional information must be stored with the materials to support the needs of the users?
   a. What information will make the collected materials meaningful?
   b. What information will make the collected materials discoverable?
5. How often must the materials be captured?
6. What are the appropriate deselection policies for this collection?
7. How often might the Designated Community change?

2.3 Copyright

2.3.1 Legal Considerations

Since a web archive must honor all legal restrictions regarding copyright and intellectual property, it should be a matter of policy that web sites be evaluated in this regard. Ideally this evaluation should occur in the planning stages for a collection and prior to acquisition of the web sites. The Digital Millennium Copyright Act (DMCA) allows digital reformatting and migration over time of print source materials but does not specifically extend this allowance to born-digital source materials. The DMCA requires permissions from creators before making any copy of digital materials and encourages actions, such as encryption, password protection, and other security mechanisms, to prevent copy violations of born-digital materials from occurring in the first place. Additionally, the DCMA makes it illegal to create tools to thwart such preventative actions. These actions present challenges to the typical collection or capture method for web-published materials, specifically to web crawlers. Furthermore, the characteristics of the DMCA also present challenges to digital preservation methods such as migration and creation of redundant copies of born-digital materials.

It is critical to define as specifically as possible what rights the archiving agency has over the materials in its web archive. An archiving agency might acquire all intellectual property rights to the materials, however many content providers will not support this extent of rights transfer. At a minimum, the rights holder's responsibilities in the preservation and dissemination of the data must be defined. Preferably, the archiving agency would be

allowed to act on behalf of rights holders to execute changes to the content for clearly defined preservation activities. These types of changes might include:

- Reformatting of materials for continued access when necessary hardware and software become obsolete
- Changes to preservation metadata to record preservation activity

2.3.2 Other Considerations

Collection Perspectives
Two general approaches to capturing web-published materials can be observed in web archiving efforts: "opt-in" and "opt-out." An opt-in policy is one in which explicit permission from content owners is sought before web-published materials are captured. An opt-out policy requires that content owners explicitly request that their materials not be captured either by robots.txt exclusions or official take-down requests.

An example of an opt-in policy is that of the National Library of Australia's PANDORA project. "Because of the lack of legal deposit provisions covering online publications both at the national and State level, all PANDORA partners seek permission from publishers prior to copying publications and web sites into the Archive."[4]

The Internet Archive is an example of an opt-out policy. It includes everything that is publicly accessible unless explicitly excluded either via a robots.txt file or by an explicit request for exclusion.[5]

Embedded Information Objects
As Lyman[6] points out, it is important to realize that a web page may consist of multiple items (e.g., sounds, images, etc.) and that each of these items may also be protected by intellectual property rights.

Privacy
Lyman also identifies potential privacy issues to consider. Some web sites collect data about their customers in order to provide a customized environment. While exposure of this data to a web crawler is not likely, any collection of this data may be regarded as an invasion of privacy. Another somewhat unlikely privacy concern might be the continued collection of personal data within an archive.

Access
Another issue is that of access. Lyman reminds us: "Preservation does not threaten markets, but access might. How can the Web archive protect markets from the potential damage of competition from illegal copies preserved by the nonprofit sector?"

[4] National Library of Australia. (2005, November 11). Legal deposit. In *About Pandora*. Retrieved May 4, 2006, from http://pandora.nla.gov.au/about.html
[5] Internet Archive. (2005). *Internet Archive frequently asked questions*. Retrieved May 4, 2006, from http://www.archive.org/about/faqs.php
[6] Lyman, P. (2002, October) Archiving the World Wide Web. In *Preserving our digital heritage: Plan for the National Digital Information Infrastructure and Preservation Program,* (Appendix 2, pp. 53-66). Retrieved May 3, 2006, from http://www.digitalpreservation.gov/about/ndiipp_appendix.pdf

2.3.3 Areas to identify

- Institution's existing copyright policy and practices
- Access restrictions that might apply to collected materials
- Roles and responsibilities of content provider
- Roles and responsibilities of archiving agency

2.3.4 Questions to address

1. What are the content provider's expectations of copyright protection for their materials?
2. What are the user access requirements for web archives?
3. What collection model will be used?
    a. opt-in
    b. opt-out
    c. a hybrid of these
4. What are the roles and responsibilities required of content providers and the archiving agency in order to allow for successful acquisition, delivery, and preservation of web-published materials?

2.4 Policies

Most physical or print materials are in forms that endure for some predictable period of time allowing for preservation actions and policies to be outlined within that interval. The urgency to address the preservation of web-published materials is that their longevity is unpredictable and materials are often lost in relatively short time frames. Therefore, preservation of web-published materials must be addressed in policy in order to ensure the materials will not be lost due to a delay of action.

It is useful to examine existing policies to identify areas that impact collection planning for web-published materials. Some organizations and libraries will be able to modify or extend their existing policies to include collections of web-published materials. However, because these collections differ in many ways from traditional print collections, new policies may need to be formulated.

2.4.1 Areas to identify

It is important to review existing policies included in the list below to determine their impact on web collection development. In cases where they have an impact, this impact should be addressed in a web collection policy. If policies in the list do not currently exist, they may need to be created.

- Rights management and copyright clearance policies
- Selection policies
- Collection planning guidelines
- Document retention guidelines
- Information technology standards
    - within the institution or organization
    - within the library
- Metadata policies and standards
- Policies outlining roles and responsibilities for the organization and content providers
- Preservation policy
- Archive policies

2.4.2 Questions to consider

1. Do existing selection policies address web-published materials?
2. Are selection guidelines transferable to web-published materials?
3. Are existing rights management and copyright clearance policies transferable to web-published materials?
4. Are there depository agreements that might impact the collection?
5. Is it feasible to extend the existing standards and policies or are new policies needed?
    a. Is a level of selection specified for web-published materials? (e.g. web page, web site, organizational domain)
    b. Are acceptable types and formats of web-published materials to be included in a collection identified?
    c. Is the level and extent of required metadata identified?
    d. Is a preservation plan in place?
6. Are roles and responsibilities identified with regard to selection, acquisition, maintenance, description, presentation, and preservation of web-published materials?

2.5 Organizational Support

Creation and preservation of collections in web archives requires enormous effort and resources, spanning several departments within an organization. Successful web archival programs within an organization will require managerial commitment and sustained funding.

2.5.1 Roles and Responsibilities

New organizational roles and responsibilities will emerge with web collection development and these are likely to trigger modifications to existing workflows, particularly for curators and others involved in web archiving activities. Because collections of web-published materials in archives are dependent on significant technological infrastructure, collaborations between libraries and information technology departments within organizations must occur.

Information management professionals, whether librarians, curators, or archivists, have expertise in collecting and preserving materials, but often do not have the technical expertise necessary to create and preserve an extensive collection of web-published materials. Information technology professionals do have expertise working with networks and digital storage, but rarely understand the long-term implications inherent in curation and preservation of stored content

It is clear that these organizational units will need to work together to achieve success in collecting and preserving web-published materials. Organizational commitment, especially in terms of management support and cooperation is critical to the success of this effort. In addition, organizational support in terms of a long-term commitment to funding is required for any web archiving effort to succeed. It is important to articulate how collection and preservation of web-published materials supports the organization's mission, benefits the organization, and provides a valuable service to the community it serves.

2.5.2 Areas to identify

- Benefits to the organization for undertaking collection and preservation of web-published materials
- Existing departments and personnel that can contribute expertise to a web collection and archiving effort

- Sources of funding, resources, and technology
- Roles and responsibilities for the participants in the effort
- Strategy and approach to gain management commitment and funding

2.5.3 Questions to address

1. What are the benefits to the organization for creating collections of web-published materials and preserving the materials?
2. If a library is not involved in the creation and preservation of web collections, how relevant will the library be to researchers over time?
3. What are the roles and responsibilities required to successfully create and preserve web collections?
4. What existing services might the organization or library consider abandoning in order to shift resources to the new roles and staffing positions that are required to support creation and preservation of web collections?
5. Which administrative positions in the organization must be sold on the idea of creating and preserving of web collections for it to be a success?
6. From which internal departmental or other stakeholders across the organization is it necessary to gain endorsement and cooperation?

2.6 Resources

2.6.1 The Resource Challenge

Resources primarily include money, people, and infrastructure. In many libraries, each of these is often in short supply and being stressed by ever-growing expectations from both management and end users. Generally, identifying web-published materials and making them accessible has been incorporated into library selectors' responsibilities. However, in many cases, these are labor-intensive responsibilities that have not been addressed with increases in funding or staff. Likewise, the library's IT infrastructure and internal support staff is typically unable to provide archival support for these increasingly important areas within library collections.

Material description (i.e. metadata application) is another area severely lacking in resources. Creation of collections of web-published materials implicitly involves rapid acquisition of large numbers of materials. These materials ideally would have descriptive metadata applied on an individual basis. Machine-generated baseline metadata that is captured by a crawler at the time web sites are captured is economical but often insufficient. Application of human-generated metadata after materials are captured is a very resource intensive task. Even cataloging efforts currently underway in many libraries for existing print materials are having difficulties retaining adequate resources to do the job.

2.6.2 Importance of Web-Published Materials

Prior to gaining the resources necessary to build and preserve collections of web-published materials, a library will generally need to document how collecting these materials promotes and supports the institution's mission. This process can generate internal selling points for creating these collections and identify the risks to the organization of not preserving web-published materials. It can also articulate the importance of web archives to a library's end users.

Figure 2 identifies the top three user needs that librarians who participated in the 2005 focus groups conducted by the Web-at-Risk project expected web archives could meet. The most important need they identified was persistent access to the information end users

need for teaching and research. The participants also identified an archive's ability to provide value-added information services, such as aggregation of content from disparate sources, as well as a need for preserving the institution's history and intellectual products in an institutional repository as important needs an archive could address. It may be helpful to translate the needs of user groups to selling points regarding the benefits of collecting and preserving web-published materials and the risks of not doing so.

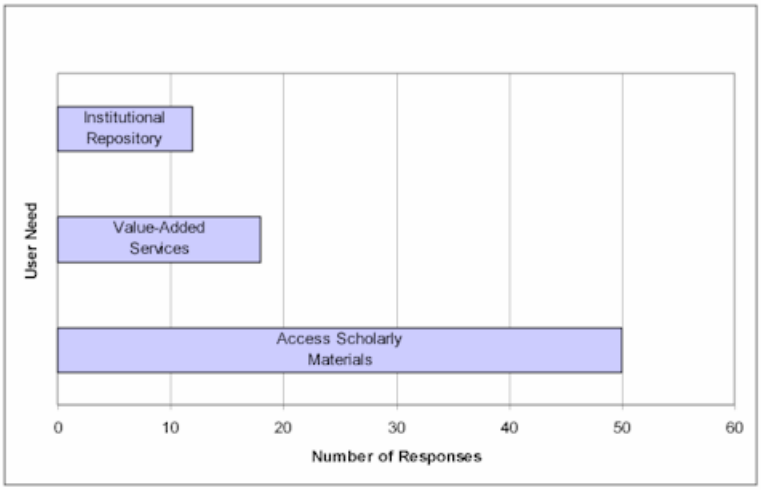

Figure 2 – Importance of Web Archives to End Users

2.6.3 Areas to identify

- Budget(s) impacting creation and preservation of web collections
- Staff required to collect, maintain, describe, present and preserve web collections
- Technical infrastructure necessary to store and provide access to web collections
- Key information needs web collections address
- Risks to the organization of not creating and preserving web collections

2.6.4 Questions to consider

1. How much staff time is spent identifying and selecting web sites?
2. How has this activity changed over the last year and over the last five years?
3. Has there been an offset in the amount of staff time required for traditional responsibilities?
4. Has the staff grown to meet the new responsibilities?
5. Is work not being done? Are user needs not being addressed?
6. Are the results of work (e.g., subject list resources) lost over time? What are the implications for staff, for faculty, for students?

2.7 Technology

Technology challenges at all stages of collection development for web-published materials are directly related to the challenges of web archiving in general. Characteristics of source materials may impact the degree to which web-published materials are successfully captured. Source materials present multiple challenges to web capture tools (i.e., web crawlers) including implementation-specific challenges such as use of Macromedia Flash, PHP, Java, and JavaScript. Some capture tools handle these challenges better than others. In addition, source materials may present challenges that no existing tools can overcome,

such as password-protected source materials and web pages generated in response to users' database queries.

Once materials are collected, it can be technically challenging to apply adequate metadata to them. Materials that are re-collected on a regular basis require that versions be separately identifiable. This presents challenges in regard to both evaluation of the differences among versions and identification of different versions in presentation.

Storage, presentation, and preservation of materials require an extensive technical infrastructure and expertise in storage media, data replication, networking, and risk management. Additionally, technical understanding of the requirements involved in content preservation is critical. These requirements include maintaining the renderability and understandability of content bitstreams as well as preserving the integrity of content over time.

2.7.1 Areas to identify

- Standards and practices
- Explicit policy regarding material formats that will and will not be collected
- Necessary infrastructure
- Necessary expertise
- Possible collaborations

2.7.2 Questions to consider

1. Will databases be included?
2. Will web site functionality be preserved (e.g., real-time database inquiries)?
3. How will external links be handled?
4. What file formats will be collected?
5. What approach to preservation will be taken?
6. What mandatory metadata is automatically generated?
7. What mandatory metadata must be added manually?

2.8 Policy Examples

Listed below are examples of policies related to the collection and preservation of web-published materials. Some of the policies are actually preservation policies for digital resources, of which web-published materials may be one example. Other policies specifically address web archiving.

Library of Congress

Collections Policy Statement: Web Site Capture & Archiving
http://www.loc.gov/acq/devpol/webarchive.html

Cornell University Library

Digital Preservation Policy Framework
http://commondepository.library.cornell.edu/cul-dp-framework.pdf

National Archives of Australia

Archiving Web Resources: A policy for keeping records of web-based activity in the Commonwealth Government
http://www.naa.gov.au/recordkeeping/er/web_records/policy_contents.html

Archiving Web Resources: Guidelines for keeping records of web-based activity in the Commonwealth Government
http://www.naa.gov.au/recordkeeping/er/web_records/guide_contents.html

The British Library

Digital Preservation Policy
http://www.bl.uk/about/collectioncare/bldppolicy1102.pdf

3 Creating a Web Collection Plan

As conceived in this document, web collection plans are the operational plans that guide activities for managing collections of web-published materials created for specific groups of users within an institution (i.e., a Designated Community in OAIS parlance). This is not unlike the general role collection plans often serve for traditional collection development within a library. A web collection plan generally articulates the role a web collection has within the organization and identifies the organization's commitment to the preservation of the collection. Figure 3 identifies the major phases and activities involved in web collection development. Collection plans address each of these phases.

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

Figure 3 – Collection Development Phases

A web collection might consist of a group of discrete but related web-published information objects but more typically it consists of a group of web-sites related by a common subject, theme, or event. The guidelines presented in this document generally assume this latter definition of a web collection to be the case. Curators may need to adapt the guidelines in this document for collections of discrete web-published information objects, for example, a collection of web-published documents in PDF format exclusive of the web site(s) in which they were published.

3.1 What to Include

Web collection plans should include the following eight sections. Considerations for each section are described in the remainder of these guidelines. Appendix A is a detailed outline for a collection plan.

Section 1.	**Mission & Scope**
	A. Mission Statement B. User Group(s) C. Collection Subject, Theme, or Event D. Curator(s)
Section 2.	**Selection Activities**
	A. Seed List B. Initial Boundary Specification C. Rights Metadata

Section 3.	**Web Site Acquisition**
	A. Frequency of Capture B. Capture Boundaries C. Material Types & Formats D. Interactive & Dynamic Content
Section 4.	**Descriptive Metadata Requirements**
	A. Level of description B. Metadata elements C. Controlled vocabularies
Section 5.	**Presentation & Access Requirements**
	A. Discovery B. Access C. Look-and-Feel D. Dynamic Content E. Multiple Types/Formats F. Authenticity
Section 6.	**Maintenance & Weeding**
	A. Maintenance Activities B. Deselection Guidelines C. Collection Evaluation
Section 7.	**Preservation**
	A. Technology Obsolescence B. Preservation Metadata
Section 8.	**Appendices**
	A. Submission Agreements B. Web Archiving Service Agreement C. Collaboration Agreements

3.2 Collection Plans for Web-published Materials: Examples

Often digital library and digital preservation policies provide guidelines that are applicable to web collection plans. The following policies include many of the web collection planning areas addressed in this document.

Canadian Heritage Information Network

Digital Preservation - Best Practice for Museums - Checklist for Creating a Preservation Policy
http://www.chin.gc.ca/English/Digital_Content/Digital_Preservation/appendixA.html
Note: Organization Items on the checklist are more in line with what this document considers under Policy.

Iowa State University - E-Library

Special Collections Department Information: Mission and Collection Policy
http://www.lib.iastate.edu/spcl/about/digital.html

University of Texas

Digital Library Collection Development Policy
http://www.lib.utexas.edu/admin/cird/policies/subjects/framework.html

4 Mission & Scope

Web collection plans begin with articulating the mission that guides collection development, describing the user groups, or Designated Community, served by the collection, and stating the information need(s) the collection will address. Web collections will generally consist of web sites united by a common subject, theme, or event. For example, discipline-related web sites included in curriculum subject guides support an academic library's mission to provide materials in support of faculty and student scholarship and learning.

4.1 Contents

Section 1. Mission & Scope
A. Mission Statement B. User Group(s) C. Collection Subject, Theme, or Event D. Curator(s)

4.2 What to Address

4.2.1 Mission Statement

Articulate the mission under the umbrella of which the collection is being developed. For many collections this will be the mission statement of the library. For others, web collection development may be more appropriately positioned under mission of the organization or institution.

4.2.2 User Group(s)

Define the user groups for the web collection. In many cases there will be more than one user group that will use a collection, for example faculty, students, and the general public. For web collections, a complete understanding of user groups is important so that the unique characteristics and needs of each one can influence the range of collection development activities, which include identifying what to collect and the metadata required for information discovery. Be as detailed as appropriate regarding each user group's demographic characteristics and their use of web-published materials.

Consider assessing the user information needs that could be addressed by web-published materials. Understanding how users currently use web-published materials to carry out their organizational or professional responsibilities might be helpful. Various methods can be used for this, including surveys, focus groups, and interviews. This should help identify gaps in existing collections and prioritize materials targeted for web collection development.

4.2.3 Collection Subject, Theme, or Event

State the subject area or theme that unites the web sites in the web collection. In some cases, web sites in a collection may be related to a common event, such as the Olympic Games or a national election. Describe how the collection supports the mission of the library, organization, or institution.

4.2.4 Curator(s)

Identify the curator(s) of the collection. Include a description of each curator's responsibilities within their organization or institution and their contact information.

4.3 Tools and Resources

The following toolkit was developed for the Web-at-Risk project. Two of the appendices provide questionnaires that might be useful in conducting needs assessment activities.

Web-at-Risk Project: Needs Assessment Toolkit

Appendix 13: End User Interview Questionnaire
Appendix 17: Content Provider Interview Questionnaire
http://web2.unt.edu/webatrisk/na_toolkit/deliverable_na_toolkit_final_krm_31may2005.pdf

5 Selection Activities

Policies, practices, agreements, and laws will impact web site selection decisions. These may come from the content provider, the organization creating the collection, or the archive agency or archive service provider. For example, selection may need to consider organizational or archive policies regarding acceptable subject matter, material types, and material formats. Additionally, the rights to capture and present web sites and the information objects they contain must be identified and necessary permissions must be gained. It is likely that selection will be refined over time depending on initial and subsequent web site captures.

Web Site Selection
Selection of web sites is generally complicated by the absence of a clearly defined entity to be assessed, evaluated, and collected. As Lyman[7] points out: "The average Web page contains 15 links to other pages or objects and five sourced [(i.e., embedded)] objects, such as sounds or images."

A web site consists of one or more web pages that are generally related in some way. 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 more common (e.g., wikis and blogs). A web site is located by a uniform resource locator (URL) that typically identifies the web site's *home page*. A home page is a web page designed by a web site owner as the main entry point to a web site.

Host Identification
Web pages comprising a web site access other web pages and web-published materials via URLs. These URLs may identify content that is embedded in web pages and automatically presented to a user when a web page is accessed. URLs within web pages may also be interactive hypertext references that users may activate to retrieve additional web-published materials.

The physical computers that store and serve web pages and other web-published materials identified by URLs are called *hosts*. Although a web site may be wholly contained on a single host, it is important to note that some web sites consist of materials that are stored on and served by two or more hosts. In this case, it is important during the selection process to identify each host. This information will help the curator to properly configure capture parameters for selected web sites.

Capture Depth
The initial capture of web sites for a collection will likely be based on a list of URLs, or a *seed list*, that specifies the web pages from which a capture should begin. Web crawlers extract additional candidate URLs for capture from the web pages in the seed list. These candidate URLs are evaluated by a crawler based on predefined settings such as (a) whether or not the URLs reside on the same host as a seed URL (i.e., the *local host*) or a secondary host to a seed URL (i.e., an *external host*) or (b) the desired *depth* of a crawl. From a web crawler perspective, depth refers to the number of linked URLs away from a

[7] Lyman, P. (2002, October) Archiving the World Wide Web. In *Preserving our digital heritage: Plan for the National Digital Information Infrastructure and Preservation Program,* (Appendix 2, pp. 53-66). Retrieved May 3, 2006, from http://www.digitalpreservation.gov/about/ndiipp_appendix.pdf

seed URL from which a crawler should capture content. URLs that meet the predefined settings for a crawl are added to the crawler's list of URLs to be captured.

An understanding of both the structure of the web sites to be captured and the way in which the archiving service provider's crawler works are critical to the formulation of an effective seed list. Evaluation of the results of the initial capture will allow curators to refine their selection decisions.

5.1 Contents

Section 2. Selection Activities
A. Seed List B. Initial Boundary Specification C. Rights Metadata

5.2 What to Address

5.2.1 Seed List

- URL(s)
- Brief Description(s)

Identify and describe the web sites included in the seed list of URLs. A seed list includes one or more entry point URLs from which a web crawler begins capturing web-published materials. If a web site is served by more than one host, consider including each host's URL in the seed list.

5.2.2 Initial Boundary Specification

- Depth of linked web pages within the seed URL host
- Inclusion or exclusion of linked web pages from external hosts for each seed URL host
  - Depth of linked web pages from external hosts (if included)

Evaluate the boundaries for each URL in the seed list. Evaluating boundaries consists of estimating the depth of linked pages to be captured on the local host and on external hosts. As described earlier in this section, boundary specifications are dependent both on how web sites are structured and on how a web crawler captures web content. A clear understanding of web site structure and crawler behavior, as well as some experience with web site selection, will increase the efficiency of site selection.

5.2.3 Rights Metadata

- Rights designation
- Rights metadata
- Linked and sourced objects

For each seed URL, determine the rights that will govern the capture of its content. Also, as appropriate, determine the rights of sourced or embedded objects contained in the web sites. An archive agency may provide rights categories from which an appropriate designation can be made for each seed URL and any embedded objects. For example, the

rights categories developed for the Web-at-Risk project's Web Archiving Service[8] will likely include such categories as: "permission not needed", "notification needed", or "permission needed."

Create rights metadata for each seed URL. At a minimum this might include: contact information, contact history, date permission granted. Additional rights information may be established or may be required by the content provider or the web archive service provider.

5.3 Tools and Resources

The following resources provide additional considerations that might be of interest in selecting materials for web collections.

Digital Preservation Coalition

Decision Tree for Selection of Digital Materials for Long-term Retention
http://www.dpconline.org/docs/handbook/DecTree.pdf

Interactive Version of Decision Tree:
http://www.dpconline.org/graphics/handbook/dec-tree-select.html

National Library of Australia

Online Australian Publications: Selection Guidelines for Archiving and Preservation by the National Library of Australia
http://pandora.nla.gov.au/selectionguidelines.html

University of Texas

Digital Library Collection Development Policy
Note: See *Archiving of non-University of Texas web sites*
http://www.lib.utexas.edu/admin/cird/policies/subjects/framework.html

[8] California Digital Library. (2005, September 12). *Web-at-Risk rights clearance protocol: Draft*. Retrieved May 9, 2006, from http://wiki.cdlib.org/WebAtRisk/tiki-download_file.php?fileId=128

6 Web Site Acquisition

Typically, a web archive acquires web-published materials by capturing content from web sites using a web crawler. One important exception to this might be databases, which are usually neither accessible nor friendly to a web crawler. It might be preferable for a content provider to create text-formatted data base files and make alternate arrangements to submit the files to the archive provider.

Curators are active participants in the selection and acquisition processes. Initial capture results must be evaluated and reviewed for quality. Both the seed list and capture specifications, which were identified in the Selection phase, are refined in the Acquisition phase.

Detailed capture specifications will include several parameters, which may be determined by the archive service provider. Some parameters may be required by default. Included in this section are basic parameters that might to be required for each URL in a seed list.

6.1 Contents

Section 3. Web Site Acquisition
A. Frequency of Capture B. Capture Boundaries C. Material Types & Formats D. Interactive & Dynamic Content

6.2 What to Address

6.2.1 Frequency of Capture

- Date
- Interval

Identify both when and how often each URL on the seed list should be captured. Possible capture frequencies might include: one time only, daily, every "x" number of days, monthly on a specific date, quarterly on a specific date, whenever content changes, or upon request from the content provider. It is important to note that sometimes a site will change while it is being harvested. Pages may be removed, moved, or may be changed by the content provider during the capture. This could result in hyperlink errors or semantic inconsistencies among pages of a captured web site when subsequently viewed by users.

6.2.2 Capture Boundaries

- Depth of linked web pages within the seed URL host
- Inclusion or exclusion of linked web pages from external hosts for each seed URL host
  - Depth of linked web pages from external hosts (if included)

Re-evaluate and refine the capture boundaries for hosts in the seed list. Capture boundaries refer to the depth to which a crawler will capture linked pages and embedded content and to what extent materials will be captured from external hosts. In general, specify the successive number of links or hops away from a seed URL from which linked or sourced

content should be captured. Keep in mind that there is no one web site organizational structure; some web sites are organized hierarchically and some are not. Additionally, more than one host in an organization may provide sourced objects for a web page (e.g., images or video).

6.2.3 Material Types & Formats

- Excluded types
- Excluded formats

Identify any specific types or formats of web-published materials that should not be captured during crawls of seed URLs. Material types will include such things as text, images, audio, video, and other application-specific data types. Formats refer to specific encoding schemes such as html, jpeg, gif, PDF, etc. A web-published file's type and format are identified by mime types, for example: text/html and image/gif.

6.2.4 Interactive & Dynamic Content

- Authentication (username/password)
- Email links
- Forms
- Database-generated pages (based on user queries)
- Dynamically or programmatically generated web pages

Evaluate the web sites in the seed list and identify and describe their interactive and dynamic content. Consider the following: Is a site password protected? Are email links and comment forms included? Does the web site rely on a database(s) to generate web pages? Does the web site create pages on-the-fly, possibly combining style sheets with server-side scripts or code? The archive agency may provide curators with the ability to conduct preliminary or test crawls of web sites in the seed list. Further, the agency might provide tools that can assist with an evaluation of web site interactivity based on the materials captured during test crawls. It may be possible to extrapolate these limited evaluations to characterize entire web sites.

Estimate the importance of retaining the functionality of the original web site. This information will help identify the scope of content the web collection requires. Review web collection policies to determine any requirements for identifying content that is no longer active, for example, creating tags to alert users to inactive email links.

6.3 Tools and Resources

The first reference describes what was learned with the Library of Congress' MINERVA prototype web archiving program. It briefly addresses boundary and format issues. The second reference identifies standard mime types.

Arms, W., Adkins, R., Ammen, C., & Hayes, A. (2001, April 15). Collecting and preserving the Web: The Minerva prototype. *RLG DigiNews, 5*(2). Retrieved May 5, 2006, from http://www.rlg.org/preserv/diginews/diginews5-2.html#feature1

W3C: World Wide Web Consortium

Multimedia MIME Reference
http://www.w3schools.com/media/media_mimeref.asp

7 Descriptive Metadata Creation

Because descriptive metadata updates and changes are costly, McCray and Gallagher[9] believe it is important "to decide on the nature and number of metadata elements early in a project." Further, they state that decisions "on the basic conceptual units, or objects, the system will include" are essential in determining the level at which metadata will be assigned. Decisions regarding metadata schema and encoding method must be made, content and input rules established, and instruction regarding which extensions and qualifiers are allowed must be documented.

Because metadata is strongly related to end user information discovery, understanding the needs and salient characteristics of a collection's user group(s) is critical. Curators of web collections must determine the level(s) of description a collection's user group(s) will require; will collection-level and seed URL descriptions suffice or is a more granular level of description required?

7.1 Contents

Section 4. Descriptive Metadata Requirements
A. Level of description B. Metadata elements C. Controlled vocabularies

7.2 What to Address

- Level of description
  - Collection level
  - Web site level
  - Information object level
- Metadata elements
  - Essential
  - Desirable
- Controlled vocabularies

Descriptive metadata is information that allows end users to locate, analyze and request archived materials (e.g., author, title, subject, location). Curators may need to conform to a descriptive metadata standard established by an archive service provider or by their own organization. It may be possible to incorporate additional curator-generated metadata or other standard metadata schemas to prescribed standards.

Metadata schemas should describe the syntax and meaning of metadata element values. Controlled vocabularies specific to a collection and meaningful to a collection's intended user group(s) may exist or can be developed.

Identify the level of description required by the collection's user group(s). List any descriptive metadata elements of importance for information discovery by the collection's

[9] McCray, A. T., & Gallagher, M. E. (2001). Principles for digital library development. *Communications of the ACM, 44*(5), 48-54. Retrieved Jan 28, 2005, from ProQuest database.

user group(s) and rate these as either essential or desirable. Lastly, identify any controlled vocabulary sources that are appropriate for the listed metadata elements.

7.3 Tools and Resources

The following are metadata references for common metadata schemas.

PREMIS: Preservation Metadata: Implementation Strategies - A Working Group Jointly Sponsored by OCKC and RLG

Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group (May 2005)
http://www.oclc.org/research/projects/pmwg/premis-final.pdf

RLG: Research Libraries Group

Descriptive Metadata Guidelines for RLG Cultural Materials
http://www.rlg.org/en/pdfs/RLG_desc_metadata.pdf

DCMI – Dublin Core Metadata Initiative

http://www.dublincore.org/

MODS – Metadata Object Description Schema

http://www.loc.gov/standards/mods/

MARCXML – MARC 21 XML Schema

http://www.loc.gov/standards/marcxml/

8 Presentation and Access Requirements

Discovery
Decisions must be made regarding the discovery method user groups require. What kind of search mechanisms are needed (e.g., keyword search capability or a subject directory browse interface)? How will search results be displayed and how much information about archived content will be initially presented? When a user has located an item of potential interest, how much additional information or metadata can they access and how will the interface permit that access? For example, will users be given the capture date for each item? Will users be able to "click through" to the item once they determine that they have found something they want?

Access
In certain cases, curators may designate web collections as either *visible* or *dark*, that is, as accessible or not accessible to users. A variation on a dark archive might be a designation that a collection will become visible only at some future point in time. This might be done to protect personal privacy or to preserve a competitive market position. For example, public access to archived collections might be delayed until public access no longer has the potential to cause economic damage to the content producer.

Alternatively, an archive might restrict access to its stored information based on agreements with content producers or an archive might employ a model of the Fair-Use doctrine, requiring users of the information to formally agree to restrict use of the information to designated applications.

Presentation
In practice, most archived web collections comprised of captured web sites will likely present web sites as mirror experiences of the originally published sites. Collections comprised of selected web-published information objects, such as videos of volcanic activity, may require unique user interfaces to present this information to the collection's user groups.

Authenticity Assessment
Finally, how will users assess the authenticity and credibility of archived web sites and their contents? Thibodeau[10] cautions: "given that a digital information object is not something that is preserved as an inscription on a physical medium, but something that can only be constructed—or reconstructed—by using software to process stored inscriptions, it is necessary to have an explicit model or standard that is independent of the stored object and that provides a criterion, or at least a benchmark, for assessing the authenticity of the reconstructed object."

Identify the authenticity criteria users of the collection will require for the collection's web sites or information objects. Will user group(s) rely upon an archive's reputation or require

[10] Thibodeau, K. (2002, July). Overview of technological approaches to digital preservation and challenges in coming years. In *The State of Digital Preservation: An International Perspective: Conference proceedings*. Retrieved May 4, 2006, from http://www.clir.org/pubs/reports/pub107/pub107.pdf

an archive to be certified by some established process such as the certification process for digital repositories proposed by the Research Libraries Group[11]?

8.1 Contents

Section 5. Presentation & Access Requirements
A. Discovery B. Access C. Look-and-Feel D. Dynamic Content E. Multiple Types/Formats F. Authenticity

8.2 What to Address

8.2.1 Discovery

- Search
- Browse
- Evaluation

Identify how user groups will want to interact with the web collection for discovery and evaluation of the collection's materials. What search methods do users require, for example, advanced search screens or simple keyword searches? Will users want to browse the collection based on subject categories? What information elements or evaluation criteria do users prefer to consider in their evaluation processes?

8.2.2 Access

- Dark collection
- Time-dependent release restrictions
- Privacy concerns (redaction)

Identify the web collection as either visible (accessible) or dark (not accessible). Identify any time-dependent release restrictions associated with the web collection. List privacy practices or policies that might restrict the accessibility of captured web content.

8.2.3 Look-and-Feel

- Importance to user groups
- Removal of information objects

Curators should consider the importance of retaining the "look-and-feel" of web sites in the web collection and state the importance of this for the collection's user groups. If the web collection will consist of information objects that have been removed from their context, estimate the effect, if any, on their meaning and utility to the collection's users. In the event that some information content is removed from archived web pages for policy or legal reasons, should users be alerted to this alteration? If yes, how should users be alerted?

[11] Research Libraries Group. (2005, August). *An audit checklist for the certification of trusted digital repositories: Draft for public comment*. Retrieved April 25, 2006, from http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

8.2.4 Dynamic Content

- Type
  - Password protected
  - Email
  - Forms
  - Database-generated pages (based on user queries)
  - Dynamically or programmatically generated web pages
- Preservation State
  - Active
  - Disabled
  - Broken
- Annotation
  - Yes/No
  - Form or manner

When archived web pages retain the look-and-feel of the original sites, curators should address some functionality issues: Will the users be allowed to access hyperlinked materials and web sites that are not located within the web archive? If so, will users be alerted to the fact that they are leaving the archive? If not, will links simply be disabled or will information about links (e.g., the specified URL) be presented along with an informative message? What about preservation of email links? How will forms be addressed within the web archive? For example will the "Submit" button be disabled or will an annotated static screen shot of the original form be available?

8.2.5 Multiple Types/Formats

- Acceptable types/formats
- Restricted types/formats
- Unacceptable types/formats

When multiple types and formats of information objects contained in web sites are captured, will all the types and formats be discoverable and made accessible to users? Curators should identify the types and formats of information objects their users are allowed to access. This might vary according to a user's access location, for example, the institution's library or a user's home or office.

8.2.6 Authenticity

- Authentication process
- Indicator

Authentication of materials may result in some type of indicator that the materials in a web collection are reliable copies of source materials. This indicator might be visible to users when they view a web site in a web collection.[12]

Identify the authentication process for the materials in the collection. What type of authenticity indicator or stamp do user groups require? Is there a trusted third-party that

[12] This indicator of authenticity is different from the integrity indicator identified in the preservation section of this document. Integrity is a measure of the bits included in captured materials stored in an archive. A baseline indication of integrity is established when materials are captured, often by a checksum method. Subsequently, the integrity of materials in the archive can be verified by recalculating the checksum and comparing it to the baseline measure.

can authenticate web sites on the seed list? Can the archive service provider offer this service?

8.3 Tools and Resources

The Research Libraries Group publication proposes a certification process for digital repositories.

Research Libraries Group. (2005, August). *An audit checklist for the certification of trusted digital repositories: Draft for public comment*. Retrieved April 25, 2006, from http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

9 Maintenance and Weeding

Maintenance of the web sites in a web archive is generally a preservation activity. However, there are some curatorial responsibilities as well, particularly in regard to maintenance of seed lists, capture specifications, rights metadata, and descriptive metadata. Additionally curators may be involved in deselecting materials from the archive. In many archives, deselection or weeding will never occur, in fact, it appears to belie the essential preservation role of an archive. Yet there may be circumstances in which weeding is desirable. These circumstances might be dictated by retention guidelines, mandated by economic constraints, or result from technological obsolescence.

9.1 Contents

Section 6. Maintenance & Weeding
A. Maintenance Activities B. Deselection Guidelines C. Collection Evaluation

9.2 What to Address

9.2.1 Maintenance Activities

- Seed lists
- Capture specification for seed lists
- Rights metadata
- Descriptive metadata
- Collection membership

Identify the anticipated maintenance activities for the web collection. These may be specified by an archive service provider. Suggest the triggers for curators (or others) to conduct these activities.

9.2.2 Deselection Guidelines

- Content provider request
- Retention guidelines
- Retention practices
- Number of copies
- Currency of capture

Identify anticipated circumstances in which web sites or information objects might be removed from an archive, for example, at the request of the content provider or in accordance with a user group's judgment of a site's or an object's continuing value to the web collection. (Some information may have value for a finite period of time for the identified user group(s), perhaps one year or perhaps three years.) Consider what it means to deselect a web site or a web collection from an archive: Does it mean that the web site(s) will never be captured again? Does it mean preservation activity will be discontinued? Does it mean that the content will be removed from the archive?

9.2.3 Collection Evaluation

- Administrative data analysis
  - Usage information
  - Date of metadata creation/alteration
  - Search logs
  - Retrieval logs
- Mime type analysis
- Rights designation analysis
- User group feedback

Identify system-generated data that might assist with the evaluation of the web collection and with weeding and other maintenance decisions. Identify methods of obtaining feedback with regard to the usefulness of the web collection from its identified user group(s).

10 Preservation

"Technology obsolescence is generally regarded as the greatest technical threat to ensuring continued access to digital material."[13] Curators must be aware of the implications, with regard to authenticity and copyright, when originally captured materials are migrated due to technological obsolescence. Preservation activities also include the creation of preservation metadata.

10.1 Contents

Section 7. Preservation
A. Technology Obsolescence B. Preservation Metadata

10.2 What to Address

10.2.1 Technology Obsolescence

- Policy and practice
- Preservation methods

Presentation of the original look-and-feel of web sites presents technical challenges regarding hardware and software obsolescence. Curators have a role in making such decisions as: Will obsolete hardware and software be preserved? Will the original look-and-feel be emulated with newer hardware and software? In responding to these questions, curators represent the needs and concerns of user groups in the decision processes.

Identify any policies or practices that must be considered when dealing with hardware and software obsolescence. Identify a process for determining acceptable preservation methods and evaluating their impact on the authenticity of materials and their copyright protection.

10.2.2 Preservation Metadata

- Provenance
  - Origin and history of content
  - Who has owned/controlled it
  - What changes/migrations have been done on it
- Context
  - Why content was created
  - How it relates to other content
- Reference
  - One unambiguous identifier
  - Other identifiers (e.g., URLs)
- Fixity
  - Information regarding verification/validation of data integrity of the content
  - Integrity indicator

[13] Digital Preservation Coalition. (2002). Digital preservation. In *The Handbook* (chap. 2). Retrieved May 4, 2006, from http://www.dpconline.org/graphics/handbook/

The Open Archival Information System (OAIS) reference model recommends the categories and elements identified above. They illustrate the type of metadata expected to be necessary for preservation of materials in an archive. Identify any preservation metadata elements necessary to preserve the collection. Curators might have a role in the creation of the preservation metadata. Identify who has responsibility for creating and maintaining each element.

10.3 Tools and Resources

The following are basic references that describe an archive agency's preservation responsibilities for web-published materials in an archive.

Research Libraries Group. (2002, May). *Trusted digital repositories: Attributes and responsibilities.* Retrieved May 4, 2006, from http://www.rlg.org/longterm/repositories.pdf

National Library of Australia: PADI - Preserving Access to Digital Information http://www.nla.gov.au/padi

11 Collection Plan Appendices

Appendices can include a range of materials that augment the web collection plan. What curators include is related to the web collection being built, the archive service provider, the source of the content, and a curator's institution or organization. The contents of a web collection plan suggest the types of documentation that might be helpful. Alternately, the appendix might simply be a reference list of applicable agreements, policies, practices, standards, and guidelines for the collection.

11.1 Contents

Section 8. Appendices
A. Submission Agreements B. Web Archiving Service Agreement C. Collaboration Agreements

11.2 What to Address

11.2.1 Submission Agreements

- Parties involved
- Roles & responsibilities
- Terms & conditions
  - Content included
  - Metadata provided
  - Content excluded
  - Intellectual property rights
  - Capture or submission
    - Integrity assurance
    - Error handling
  - Authenticity assurance

A content provider agreement or submission agreement specifies in some detail the legal relationship between a content provider or information producer and an archive service provider. Submission agreements need to identify what web-published content or data will be submitted and what metadata will accompany the content and data.

The agreement should also specify any procedures or protocols for web site capture by the archive service provider and, alternately, for data submission by the content provider. Additionally, procedures for verifying successful transmission and procedures for getting answers to questions about the content should be specified in the agreement.

11.2.2 Web Archiving Service Agreement

- Parties involved
- Roles & responsibilities
- Terms & conditions
  - Collection submission
  - Collection management
  - Collection use

- Capture or submission
  - Integrity assurance
  - Error handling
- Authenticity assurance

A web archiving service agreement should be contracted between the archive service provider and the institution or organization whose curator(s) is building the web collection. Such an agreement would identify the parties to the agreement and describe their respective roles and responsibilities in regard to web archiving. Additionally, the service terms and conditions should be described, including penalties for non-performance, notices of service or contract termination, verification of integrity of captured materials, and error handling procedures.

Note: If the web archive service is provided by a curator's own institution or organization, a service agreement may not be required. However, it is still important to identify organizational roles and responsibilities in the preservation effort and to ensure that supporting policies are in place within the organization.

11.2.3 Collaboration Agreements

If more than one institution is collaborating to build a web collection, one or more of the institutions may require some type of collaboration agreement. The specific terms and conditions may be dictated by the institutions as well as predicated by the type and scope of the agreement.

Appendix A. Web Collection Plan Outline

Section 1. Mission & Scope

A. Mission Statement
B. User Group(s)
C. Collection Subject, Theme, or Event
D. Curator(s)

Section 2. Selection Activities

A. Seed List
    i. URL(s)
    ii. Brief Description(s)
B. Initial Boundary Specification
    i. Depth of linked web pages within the seed URL host
    ii. Inclusion or exclusion of linked web pages from external hosts for each seed URL host
        a. Depth of linked web pages from external hosts (if included)
C. Rights Metadata
    i. Rights designation
    ii. Rights metadata
    iii. Linked and sourced objects

Section 3. Web Site Acquisition

A. Frequency of Capture
    i. Date
    ii. Interval
B. Capture Boundaries
    i. Depth of linked web pages within the seed URL host
    ii. Inclusion or exclusion of linked materials from external hosts for each seed URL host
        a. Depth of linked web pages from external hosts (if included)
C. Material Types & Formats
    i. Excluded types
    ii. Excluded formats
D. Interactive & Dynamic Content
    i. Authentication (username/password)
    ii. Email links
    iii. Forms
    iv. Database-generated pages (based on user queries)
    v. Dynamically or programmatically generated web pages

Section 4. Descriptive Metadata Requirements

A. Level of description
   i. Collection level
   ii. Web Site level
   iii. Information object level
B. Metadata elements
   i. Essential
   ii. Desirable
C. Controlled vocabularies

Section 5. Presentation & Access Requirements

A. Discovery
   i. Search
   ii. Browse
   iii. Evaluation
B. Access
   i. Dark collection
   ii. Time-dependent release restrictions
   iii. Privacy concerns (redaction)
C. Look-and-Feel
   i. Importance to user groups
   ii. Removal of information objects
D. Dynamic Content
   i. Type
      a. Password protected
      b. Email
      c. Forms
      d. Database-generated pages (based on user queries)
      e. Dynamically or programmatically generated web pages
   ii. Preservation State
      a. Active
      b. Disabled
      c. Broken
   iii. Annotation
      a. Yes/No
      b. Form or manner
E. Multiple Types/Formats
   i. Acceptable types/formats
   ii. Restricted types/formats
   iii. Unacceptable types/formats
F. Authenticity
   i. Authentication process
   ii. Indicator

Section 6. Maintenance & Weeding

A. Maintenance Activities
   i. Seed lists
   ii. Capture specification for seed lists
   iii. Rights metadata
   iv. Descriptive metadata
   v. Collection membership

B. Deselection Guidelines
   i. Content provider request
   ii. Retention guidelines
   iii. Retention practices
   iv. Number of copies
   v. Currency of capture

C. Collection Evaluation
   i. Administrative data analysis
      a. Usage information
      b. Date of metadata creation/alteration
      c. Search logs
      d. Retrieval logs
   ii. Mime type analysis
   iii. Rights designation analysis
   iv. User group feedback

Section 7. Preservation

A. Technology Obsolescence
   i. Policy and practice
   ii. Preservation methods

B. Preservation Metadata
   i. Provenance
      a. Origin and history of content
      b. Who has owned/controlled it
      c. What changes/migrations have been done on it
   ii. Context
      a. Why content was created
      b. How it relates to other content
   iii. Reference
      a. One unambiguous identifier
      b. Other identifiers (e.g., URLs)
   iv. Fixity
      a. Information regarding verification/validation of data integrity of the content
      b. Integrity indicator

Section 8. Appendices

A. Submission Agreements
    i. Parties involved
    ii. Roles & responsibilities
    iii. Terms & conditions
        a. Content included
        b. Metadata provided
        c. Content excluded
        d. Intellectual property rights
        e. Capture or submission
            - Integrity assurance
            - Error handling
        f. Authenticity assurance
B. Web Archiving Service Agreement
    i. Parties involved
    ii. Roles & responsibilities
    iii. Terms & conditions
        a. Collection submission
        b. Collection management
        c. Collection use
        d. Capture or submission
            - Integrity assurance
            - Error handling
        e. Authenticity assurance
C. Collaboration Agreements

Appendix B. Glossary

Acquisition	For digital 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, 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 or institution.
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.
Dark Web	See Deep Web
Deep Web	Resources available via the World Wide Web that are invisible to or inaccessible by crawlers. These resources may be invisible to or inaccessible by to 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.

Designated Community	A term used in OAIS. According to the OAIS recommendation, the Designated Community is "an identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities."
Digital Archive	One or more digital collections for which an institution has agreed to accept long-term responsibility for preserving the objects in the collection and for providing continual access to them in keeping with an archive's user access policies.
Digital Collection	A collection consisting entirely of born-digital or digitized materials.
Digital Object	Also called a digital information object. Digital objects can be interactive works (e.g., video games), sensory presentations (e.g., music or audio), documents, and data. Two types of digital objects included in digital archives are: surrogates of information objects in various original formats, (e.g., print books or audio tapes) and born-digital objects.
Dynamic Web Page	A web page created automatically by software at the web server. The page may be (a) personalized for the user based on identification via login or based on cookies stored on the user's computer, (b) tailored to fulfill a specific request made by the 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 (or data) 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.
Integrity	A digital object's integrity is maintained as long as the bits contained in the object are not altered in an unauthorized manner.
Invisible Web	See Deep Web
Light Archive	A digital archive accessible to end-users.
Medium	The delivery vehicle for the 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: 1) 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 2) 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.
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. Also called a Targeted URL or Entry Point URL.
Spider	See Crawler
Targeted URL	See Seed URL
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 collection of web-published materials that an institution has either made arrangements for or has accepted long-term responsibility for preservation and access in keeping with an archive's user access policies. Some of these materials may also exist in other forms but the web archive captures the web versions for posterity. 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. However, a web collection might also refer to a group of related web-published digital objects. Web collections can be preserved and/or curated. All web collections residing in a web archive are assumed to be preserved. The application of collection development processes to archived collections results in curated web collections whose content users can discover and access.
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 (i.e., files formatted for presentation via a web browser) within a web site are often published and maintained by a single person or organization, although wider collaborations and social publishing are becoming more common (e.g., wikis and blogs). Hyperlinks in the form of uniform resource locations (URLs) on web site pages access other web pages and specific web-published information objects (e.g., documents or images). Both pages and objects within a web site and external to a web site can be linked.
Web-based Resources	See Web-published Materials.
Web-published Materials	Web-published materials are accessed and presented via the World Wide Web. The materials span the cultural heritage spectrum and 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.

Appendix C. Preservation Projects: Selected Examples

National Digital Information Infrastructure and Preservation Program (NDIIPP)

Library of Congress
http://www.digitalpreservation.gov/

The Digital Preservation Program (NDIIPP) seeks to provide a national focus on important policy, standards and technical components necessary to preserve digital content. Investments in modeling and testing various options and technical solutions will take place over several years, resulting in recommendations to the U.S. Congress about the most viable and sustainable options for long-term preservation.

Collaborative Collection Development Partnerships
http://www.digitalpreservation.gov/partners/project.html

On Sept. 30, 2004, the National Digital Information Infrastructure and Preservation Program welcomed its first formal partners by making cooperative agreements with eight institutions to begin building a digital preservation network. These eight lead institutions, which have joined with other institutions and organizations in their efforts, have agreed to identify, collect and preserve digital materials within a nationwide digital preservation infrastructure. These awards from the Library are being matched dollar-for-dollar by the winning institutions in the form of cash, in-kind or other resources. The institutions will share responsibilities for preserving at-risk digital materials of significant cultural and historical value to the nation.

Digital Archiving and Long-Term Preservation (DIGARCH)
http://www.digitalpreservation.gov/partners/research.html

On May 4, 2005, the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) and the National Science Foundation awarded 10 university teams a total of $3 million to undertake pioneering research to support the long-term management of digital information. These awards are the outcome of a partnership between the two agencies to develop the first digital-preservation research grants program.

The Archive Ingest and Handling Test (AIHT)
http://www.digitalpreservation.gov/technical/aiht.html

The Archive Ingest and Handling Test (AIHT), was designed to identify, document and disseminate working methods for preserving the nation's increasingly important digital cultural materials, as well as to identify areas that may require further research or development.

The AIHT was a joint effort of The Library of Congress; Old Dominion University, Department of Computer Science; The Johns Hopkins University, Sheridan Libraries; Stanford University Libraries & Academic Information Resources; and Harvard University Library "to explore strategies for the ingest and preservation of digital archives." [http://www.digitalpreservation.gov/about/pr_060904.html]

Final reports are available on The Library of Congress Digital Preservation web site (URL above). In addition, Clay Shirky published an article in the December 2005 issue of D-Lib Magazine (Volume 11, Number 12) discussing the "overall observations from the operation of that test." [http://www.dlib.org/dlib/december05/shirky/12shirky.html]

LOCKSS

Program Initiated by the Stanford University Libraries
http://lockss.stanford.edu/

LOCKSS (for "Lots of Copies Keep Stuff Safe") is open source software that provides librarians with an easy and inexpensive way to collect, store, preserve, and provide access to their own, local copy of authorized content they purchase. Running on standard desktop hardware and requiring almost no technical administration, LOCKSS converts a personal computer into a digital preservation appliance, creating low-cost, persistent, accessible copies of e-journal content as it is published. Since pages in these appliances are never flushed, the local community's access to that content is safeguarded. Accuracy and completeness of LOCKSS appliances is assured through a robust and secure, peer-to-peer polling and reputation system.

The MINERVA Web Archiving Project

Library of Congress
http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html

An ever-increasing amount of the world's cultural and intellectual output is presently created in digital formats and does not exist in any physical form. The MINERVA Web Preservation Project was established to initiate a broad program to collect and preserve these primary source materials. A multi disciplinary team of Library staff representing cataloging, legal, public services, and technology services is studying methods to evaluate, select, collect, catalog, provide access to, and preserve these materials for future generations of researchers.

The American Memory Program

Library of Congress
http://memory.loc.gov/ammem/

American Memory provides free and open access through the Internet to written and spoken words, sound recordings, still and moving images, prints, maps, and sheet music that document the American experience. It is a digital record of American history and creativity. These materials, from the collections of the Library of Congress and other institutions, chronicle historical events, people, places, and ideas that continue to shape America, serving the public as a resource for education and lifelong learning.

Appendix D. International Consortia

International Internet Preservation Consortium (IIPC)

http://netpreserve.org/

According to a 2004 press release by the IIPC:

> In acknowledgement of the importance of international collaboration for preserving internet content for future generations, the International Internet Preservation Consortium was formed in 2003.
>
> Led by the National Library of France, the Consortium also comprises National libraries of Australia, Canada, Denmark, Finland, Iceland, Italy, Norway, Sweden, The British Library (UK), The Library of Congress (USA) and the Internet Archive (USA).

The IIPC has identified three goals:

- To enable the collection of a rich body of Internet content from around the world to be preserved in a way that it can be archived, secured and accessed over time.
- To foster the development and use of common tools, techniques and standards that enable the creation of international archives.
- To encourage and support national libraries everywhere to address Internet archiving and preservation.

International Web Archiving Workshops (IWAW)

http://bibnum.bnf.fr/ecdl/

> International Web Archiving Workshops are held in association with the European Conferences on Digital Libraries (ECDL) since 2001. The aim of these workshops is to bring together researchers, practitioners, graduate students and IT developers with expertise and interest in building web archives. These workshops provide a forum for interaction among librarians, archivists, academic researchers and industrial researchers interested in establishing effective methods and developing improved solutions for web archiving.

Each workshop features presentations and papers from professionals around the world who are involved with the collection and preservation of digital objects including those that are web-published.

Digital Library Federation (DLF)

http://www.diglib.org/

DLF was formed as a result of the "belief that problems and issues inhibiting the formation of digital libraries are best resolved through collaborative practical activity rather than through further theoretical discussion."

Pursuant to this belief, a federation of libraries was formed. Membership fees and grants pay for DLF activities which include member services such as forums and newsletters. Fees and grants also support DLF investment in initiatives such as a Global Digital Format Registry and Guidelines for the Cataloging of Cultural Objects and other activities such as a Workshop on Standards for Electronic Resource Management.

Appendix E. Reference Model & Key Standards

E.1 Architecture

E.1.1 OAIS – Open Archival Information System

Reference Model for an Open Archival Information System (OAIS)
http://public.ccsds.org/publications/archive/650x0b1.pdf

The Open Archival Information System (OAIS) reference model is an archival framework on which many systems can be based. It defines the fundamental components that come together to create a successful archival system. However, the OAIS is not a standard and does not specify how those components should be implemented.

The Consultative Committee for Space Data Systems (CCSDS) created the OAIS as a reference model to address preservation functions for archival information with special focus on digital information. "An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community".

According to CCSDS, the Designated Community is "an identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities".

As shown in Figure 4, the OAIS defines a model for ingest of materials from the producer, management of the materials in the archive, and dissemination of the materials to the consumer.

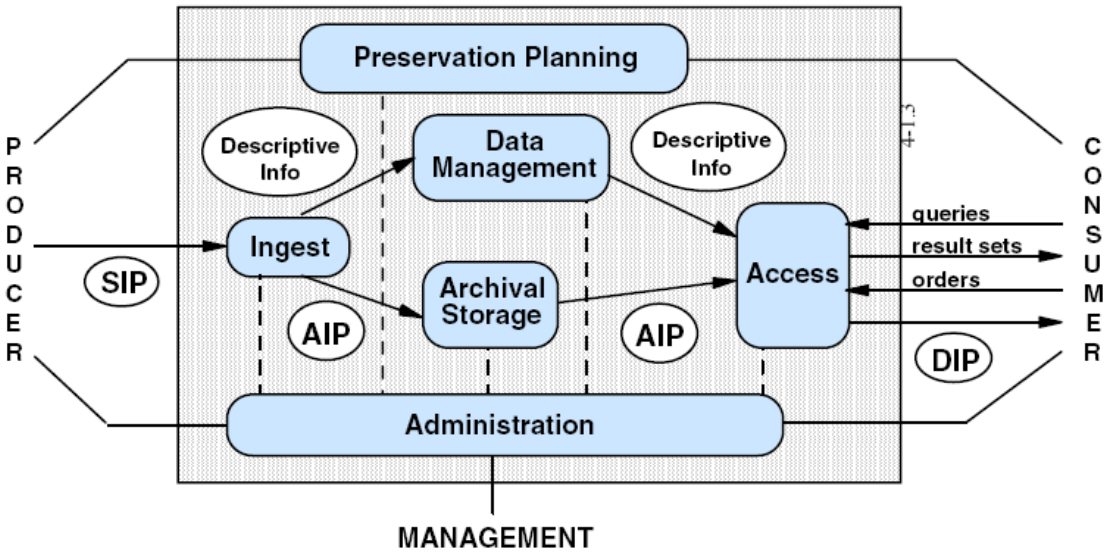

Figure 4 – OAIS Functional Entities

The basic unit of ingest, preservation, and dissemination is an information package, of which the model defines the following three:

- A SIP, or Submission Information Package, is: "An Information Package that is delivered by the Producer to the OAIS for use in the construction of one or more AIPs".

- An AIP, or Archival Information Package, is: "An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS".

- A DIP, or Dissemination Information Package, is: "The Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the OAIS".

The OAIS also describes these two important concepts:

Designated Community

The designated community is an "identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities."

Content Information

Content information is "the information which is the initial target of preservation." It consists of the bits and the representation information necessary to make those bits understandable by the designated community.

E.2 Metadata

E.2.1 OAIS Metadata Components

Reference Model for an Open Archival Information System (OAIS)
(See sections 4.2 "Information Model" and 3.2.6 "Makes the Information Available")
http://public.ccsds.org/publications/archive/650x0b1.pdf

In summary, the OAIS model recommends the following types of metadata (referred to as "information" in the OAIS recommendation) for materials in an archive:

- Preservation Description Metadata
  - Provenance – Origin and history of content. Who has owned/controlled it, and what changes/migrations have been done on it.
  - Context – Why content was created and how it relates to other content elsewhere.
  - Reference – Identifiers (especially one unambiguous identifier)
  - Fixity – Information regarding verification/validation of data integrity of the content. Authenticity indicator.

- Descriptive Metadata – Information to allow consumers to locate, analyze and request archived materials.

The OAIS also discusses the importance of adherence to the legal agreements between the OAIS and the content producer. This implies another category of metadata that should be maintained:

- Rights Metadata – Information about copyright, access and other legal restrictions.

E.2.2 METS – Metadata Encoding and Transmission Standard

http://www.loc.gov/standards/mets/

The Metadata Encoding and Transmission Standard (METS) defines a standard XML document format for encoding descriptive, administrative, and structural metadata in a repository. The METS standard is maintained by the Library of Congress' Network Development and MARC Standards Office and is being developed as an initiative of the Digital Library Federation (DLF). As a standard for transmitting various types of metadata, METS supports interoperability among digital repositories.

In regard to METS use with OAIS: "a METS document could be used in the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP) within the Open Archival Information System (OAIS) Reference Model."

E.2.3 Descriptive Metadata Element Sets and Guidelines

DCMI – Dublin Core Metadata Initiative

http://www.dublincore.org/

"The Dublin Core Metadata Initiative (DCMI) is an organization dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems." The Dublin Core metadata standard is a set of fifteen metadata elements for describing a wide range of resources.

MODS – Metadata Object Description Schema

http://www.loc.gov/standards/mods/

"The Library of Congress' Network Development and MARC Standards Office, with interested experts, has developed a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications." The Metadata Object Description Schema (MODS) is maintained by the Library of Congress' Network Development and MARC Standards Office and defines an XML schema for encoding descriptive metadata. The office also maintains crosswalks for MARC to MODS and for Dublin Core (simplified) to MODS (and vice versa). MODS v3.1 is the current schema.

MARCXML – MARC 21 XML Schema

http://www.loc.gov/standards/marcxml/

MARCXML is maintained by the Library of Congress' Network Development and MARC Standards Office and defines "a framework for working with MARC data in a XML environment." A MARCXML toolkit is available on the web site in addition to several style sheets for converting MARCXML to other schemas (e.g. Dublin Core and MODS) and vice versa.

RDA – Resource Description and Access

http://www.collectionscanada.ca/jsc/rda.html

Scheduled for publication in 2008, "*RDA - Resource Description and Access* will be a new standard for resource description and access, designed for the digital world." RDA is like its predecessor the *Anglo-American Cataloguing Rules* (AACR). Rather than identifying a set of metadata elements, RDA defines the guidelines (i.e. method and syntax) for defining element values as well as guidance in identifying access points. Values identified and structured using RDA guidelines can be used with various metadata schemas.

Traditional bibliographic description in libraries has been based on the AACR. The current standard is AACR2. Comments received in response to an initial draft of part I of AACR3 prompted the Joint Steering Committee for the Revision of AACR to design a new standard in lieu of revising the existing AACR2. "RDA will provide a comprehensive set of guidelines and instructions on resource description and access covering all types of content and media [including all digital and analog resources]."

E.2.4 Preservation Metadata Element Sets

PREMIS – Preservation Metadata Implementation Strategies

http://www.oclc.org/research/projects/pmwg/default.htm

"The PREMIS working group, jointly sponsored by OCLC and RLG, was composed of international experts from institutions that had developed or were currently developing digital preservation capacity."

The objectives of PREMIS were to:

- Develop a core preservation metadata set, supported by a data dictionary, with broad applicability across the digital preservation community.
- Identify and evaluate alternative strategies for encoding, storing, and managing preservation metadata in digital preservation systems.

In fulfillment of these goals, the PREMIS working group released its final report, *Data Dictionary for Preservation Metadata*, in May 2005. The report can be found at: http://www.oclc.org/research/projects/pmwg/premis-final.pdf

National Library of Australia – Preservation Metadata for Digital Collections

http://www.nla.gov.au/preserve/pmeta.html

"Because of its pressing business needs to manage both 'born digital' and 'digital surrogate' collections, the National Library of Australia … has invested in drafting its own model: a statement of the information it believes will be needed to manage the preservation of its digital collections."

This metadata element set focuses solely on metadata necessary to preserve digital collections. Other metadata elements are not addressed.

The CEDARS Guide to Preservation Metadata

http://www.leeds.ac.uk/cedars/index.html

The CEDARS Project was funded by JISC (the Joint Information Systems Committee of the UK higher education funding councils) and conducted from 1998-2002 with the broad objective of exploring digital preservation issues. These issues included acquisition, preservation, description and access.

As part of the project, the CEDARS team used the OAIS recommendation as a framework to outline a preservation metadata specification. *The CEDARS Guide to Preservation Metadata* documents this specification and can be accessed at:
http://www.leeds.ac.uk/cedars/guideto/metadata/guidetometadata.pdf

E.3 Harvesting

E.3.1 WARC – **W**eb **ARC**hive File Format

http://www.niso.org/international/SC4/N595.pdf (ISO Working Draft: ISO TC 46/SC 4 N 595: *Information and documentation – The WARC File Format*)

The WARC file format is based on the ARC File Format which is used by the Internet Archive to record materials captured from the web. The WARC format is more generalized than its predecessor in order to accommodate the variety of materials involved in web archiving.

E.3.2 OAI-PMH – The Open Archives Initiative Protocol for Metadata Harvesting

http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) "provides an application-independent interoperability framework based on *metadata harvesting*" and based on HTTP and XML. Protocol Version 2.0 is the current version.

The protocol defines the roles and responsibilities for data providers, who expose metadata associated with their data repositories, and service providers, who harvest or collect this exposed metadata, in order to provide value-added services.

E.4 Archival Standards

E.4.1 EAD – Encoded Archival Description

http://www.loc.gov/ead/

Encoded Archival Description (EAD) is a Document Type Definition (DTD) for encoding archival finding aids in XML. Development of EAD began with the recognition of the need to establish a nonproprietary encoding standard for machine-readable finding aids.

The EAD standard is maintained by the Library of Congress' Network Development and MARC Standards Office in partnership with the Society of American Archivists.

E.4.2 DACS – Describing Archives: A Content Standard

http://www.archivists.org/catalog/pubDetail.asp?objectID=1279 (purchase required)

Describing Archives: A Content Standard (DACS) is a product of the Canadian-U.S. Task Force on Archival Description. DACS is similar to RDA (described in section E.2.3 of this document) in that it is not a metadata schema, but rather a set of guidelines including syntax and recommended sources of descriptive information in order to "facilitate consistent, appropriate, and self-explanatory description of archival materials" and their creators. Also similar to RDA, DACS "can be applied to all types of material at all levels of description." Values identified and structured using DACS can be used with a variety of metadata schemas.

E.5 Repositories

E.5.1 Trusted Digital Repositories

http://www.rlg.org/legacy/longterm/repositories.pdf

A joint report from Research Libraries Group (RLG) and OCLC, *Trusted Digital Repositories: Attributes and Responsibilities* builds on the OAIS recommendation to identify the "characteristics and responsibilities of trusted digital repositories for large-scale, heterogeneous collections held by cultural organizations."

In summary, "a trusted digital repository is one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future." This document identifies specific characteristics and responsibilities that must be addressed in order to meet this goal.

E.5.2 IMS Global Learning Consortium, Inc.

http://www.imsglobal.org/digitalrepositories/index.html

IMS Global Learning Consortium is a non-profit organization formed to develop and champion the adoption of open standards, protocols and specifications for improved interoperability in learning technology.

As part of this goal, the consortium released v1.0 of its *IMS Digital Repositories Specification* on January 30, 2003. The specification includes three documents which "provide recommendations for the interoperation of the most common repository functions."

Appendix F. Compiled Collection Planning Resources

Policies & Plans: Web Collections & Digital Preservation

Library of Congress
Collections Policy Statement: Web Site Capture & Archiving
http://www.loc.gov/acq/devpol/webarchive.html

Cornell University Library
Digital Preservation Policy Framework
http://commondepository.library.cornell.edu/cul-dp-framework.pdf

National Archives of Australia
Archiving Web Resources: A policy for keeping records of web-based activity in the Commonwealth Government
http://www.naa.gov.au/recordkeeping/er/web_records/policy_contents.html

Archiving Web Resources: Guidelines for keeping records of web-based activity in the Commonwealth Government
http://www.naa.gov.au/recordkeeping/er/web_records/guide_contents.html

The British Library
Digital Preservation Policy
http://www.bl.uk/about/collectioncare/bldppolicy1102.pdf

Canadian Heritage Information Network
Digital Preservation - Best Practice for Museums - Checklist for Creating a Preservation Policy
http://www.chin.gc.ca/English/Digital_Content/Digital_Preservation/appendixA.html
Note: Organization Items on the checklist are more in line with what we are addressing under Policy.

Iowa State University - E-Library
Special Collections Department Information: Mission and Collection Policy
http://www.lib.iastate.edu/spcl/about/digital.html

University of Texas
Digital Library Collection Development Policy
http://www.lib.utexas.edu/admin/cird/policies/subjects/framework.html

Needs Assessment

Web-at-Risk Project: Needs Assessment Toolkit
Appendix 13: End User Interview Questionnaire
Appendix 17: Content Provider Interview Questionnaire
http://web2.unt.edu/webatrisk/na_toolkit/deliverable_na_toolkit_final_krm_31may2005.pdf

Selection

Digital Preservation Coalition
Decision Tree for Selection of Digital Materials for Long-term Retention
http://www.dpconline.org/docs/handbook/DecTree.pdf

Interactive Version of Decision Tree:
http://www.dpconline.org/graphics/handbook/dec-tree-select.html

National Library of Australia
Online Australian Publications: Selection Guidelines for Archiving and Preservation by the National Library of Australia
http://pandora.nla.gov.au/selectionguidelines.html

University of Texas
Digital Library Collection Development Policy
Note: See *Archiving of non-University of Texas web sites*
http://www.lib.utexas.edu/admin/cird/policies/subjects/framework.html

Acquisition

Arms, W., Adkins, R., Ammen, C. & Hayes, A. (2001, April 15). Collecting and preserving the Web: The Minerva prototype. *RLG DigiNews, 5*(2). Retrieved May 5, 2006, from http://www.rlg.org/preserv/diginews/diginews5-2.html#feature1

W3C: World Wide Web Consortium
Multimedia MIME Reference
http://www.w3schools.com/media/media_mimeref.asp

Descriptive Metadata

PREMIS: Preservation Metadata: Implementation Strategies - A Working Group Jointly Sponsored by OCKC and RLG
Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group (May 2005)
http://www.oclc.org/research/projects/pmwg/premis-final.pdf

RLG: Research Libraries Group
Descriptive Metadata Guidelines for RLG Cultural Materials
http://www.rlg.org/en/pdfs/RLG_desc_metadata.pdf

DCMI – Dublin Core Metadata Initiative
http://www.dublincore.org/

MODS – Metadata Object Description Schema
http://www.loc.gov/standards/mods/

MARCXML – MARC 21 XML Schema
http://www.loc.gov/standards/marcxml/

Authenticity

Research Libraries Group. (2005, August). *An audit checklist for the certification of trusted digital repositories: Draft for public comment*. Retrieved April 25, 2006, from http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

Digital Preservation

National Library of Australia: PADI - Preserving Access to Digital Information http://www.nla.gov.au/padi

"The PADI web site is a subject gateway to digital preservation resources" maintained by the National Library of Australia. The site provides resources and links on many topics in support of digital preservation. These topics include Web archiving tools, rights management and digital preservation policies among others.

Of particular interest to the Web at Risk project, one section of the PADI web site is dedicated to Web archiving efforts around the world: http://www.nla.gov.au/padi/topics/92.html

Repositories: Institutional Repositories & Trusted Digital Repositories

Lynch, C. A. (2003, February). *Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age*. ARL, no. 226: 1-7. Retrieved April 24, 2006, from http://www.arl.org/newsltr/226/ir.html

Research Libraries Group. (2002, May). *Trusted digital repositories: Attributes and responsibilities.* Retrieved Jan 19, 2005, from http://www.rlg.org/longterm/repositories.pdf

Research Libraries Group. (2005, August). *An audit checklist for the certification of trusted digital repositories: Draft of public comment*. Retrieved April 25, 2006, from http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

Wheatley, P. (2004, March). *Institutional repositories in the context of digital preservation.* Digital Preservation Coalition: Technology Watch Series Report 04-02. Retrieved April 24, 2006, from http://www.dpconline.org/docs/DPCTWf4word.pdf