DRAFT

WEBatRISK

Preserving Our Nation's Cultural Heritage

# Collection Plan Guidelines

# for

# Project Curators

May 9, 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

The Web-at-Risk project is concerned with developing a web archive service that will assist the project's curators in the creation and management of archived collections of web-based materials, or web archives. Some curators will extend existing web collections and some will create new collections of web-published materials. Collaborations are anticipated among curators who share common collection themes or subject matters of mutual interest. The project's curators primarily work in large academic libraries, with one curator from a state library. Many of the curators work in government information departments. Other areas of focus include public policy, trade unions, and political movements. The curators have collection development responsibilities and select print materials, electronic resources, and web-based resources to augment their collections. Most do not currently have policies and plans in place for managing and preserving collections of web-based materials.

Successful archive and preservation of web sites and web-based materials requires planning, which can be addressed in an organization's collection policy and specific collection plans. Collection policies and plans have inherent interdependencies and are sometimes contained within a single guideline or document. In general, policies situate web archiving into an organization's overall collection management program and articulate an organization's overall mission and strategic approach to web archiving. Collection plans typically stipulate how an organization's policy is implemented in a targeted area within an organization, such as a specific discipline within a university.

These policy and planning documents typically guide collection management within libraries and archives. Some familiar concepts and practices from collection development for non-digital materials easily transfer to collection development for web-based materials while some new concepts and unfamiliar practices are introduced. To effectively manage web collections, it is good practice to either create new policies and plans or modify existing collection policies and plans.

## 1.1 Web Archives

A web archive is a repository for web-published materials (web sites) for which the archive provider has accepted long-term responsibility for preservation as well as for access in keeping with an archive's user-access policies. Such policies specify access rights to an archive's materials and might stipulate which materials are not accessible. Institutions or organizations may enter into service agreements with web archive providers with the intention of preserving web-published materials of interest and value to the institution. Such agreements might identify the materials to be archived and might stipulate service terms, responsibilities, expectations, and fees for both the archive provider and the organization requesting services.

There are several models for collecting web content. The National Library of Australia[1] broadly defines the following:

1. Whole domain or comprehensive — preserves a national or global web space (e.g., Internet Archive)
2. Selective — preserves "defined portions of Web space or particular kinds of resources according to specified criteria"
3. Thematic — a form of selective capture which preserves content relating to a particular theme or event (e.g., Minerva)
4. Deposit — preserves only materials deposited by publishers based on legal or voluntary deposit codes

### 1.2 The Web Archiving Service (WAS)

The Web-at-Risk project is building a Web Archiving Service (WAS), which is a set of tools curators and users will employ to build, manage, and explore collections of materials captured from the World Wide Web. Three tools will be provided by the WAS:

1. Curator Tool
2. Administrator Tool
3. Search and Display Tool

The Curator Tool will provide curators with the ability to define and execute crawls, build and preserve collections, manage rights, and generate reports.

Institution Administrators and Web Archiving Service Administrators will be able to manage accounts and generate reports about accounts, captures, collections, and rights via the Administrator Tool.

The Search and Display Tool will provide curators and end users the ability to search, browse, and display collected and preserved web content.

From June 2006 to November 2007, toolsets will be released in stages as major functionality is implemented. The planned releases are identified in Table 1 and the functionality of each release is briefly described in Appendix A.

| Date | Release | Functionality |
|---|---|---|
| Jul 2006 | Release 1 | Basic Capture |
| Oct 2006 | Release 2 | Improved Search and Display |
| Dec 2006 | Release 3 | Improved Analysis and Reports |
| Feb 2007 | Release 4 | Collection Building |

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

| Date | Release | Functionality |
|---|---|---|
| May 2007 | Release 5 | Administration and Curatorial Rights Management |
| Jul 2007 | Release 6 | Event-based Capture and Enhancements |
| Oct 2007 | Release 7 | Preservation Features, Help Screens, and Reports |
| Nov 2007 | Release 8 | Integration of User Feedback and Refinement of Software and Documentation |

**Table 1 – WAS Toolset Releases**

### 1.3 Overview of Guidelines

The Web-at-Risk project curators will create specific collection plans for a set of related web sites. These plans will guide curators' interactions with the WAS toolsets as releases become available. The remainder of this document identifies and discusses the key areas to address in the collection plans. Two appendices are included for background and reference:

Appendix A. Web Archive Service: Schedule of Toolset Releases
Appendix B. Resources

# 2 Creating Your Web Collection Plan

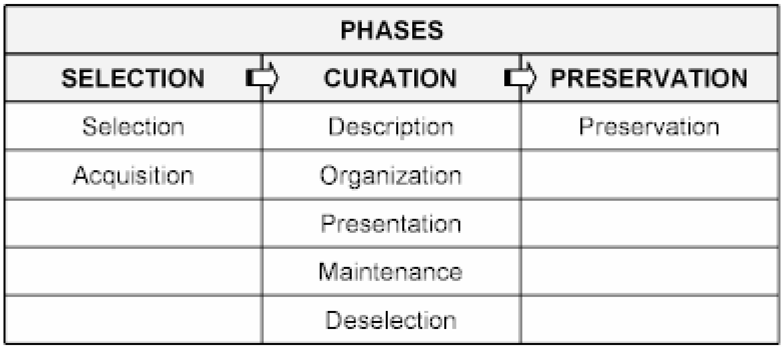
## 2.1 Terminology

Collection Policy

Policies for web collections within a library generally articulate the strategic role web collections have within an institution and identify an organization’s commitment to web collections. A web collection policy might also specify applicable guidelines and related policies, such as technical standards or retention guidelines.

Collection Plan

As conceived in this document, web collection plans are the operational plans that translate and apply the institution’s web collection policy to specific groups of users within the institution (i.e., a designated community in web archive parlance). This is not unlike the general role collection plans often serve for traditional collection development within a library. Figure 1 identifies the major phases and activities involved in web collection development.



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

**Figure 1 – Collection Development Phases**

## 2.2 Scope

It is beyond the scope of this project to expect the project’s curators to create and gain approval for comprehensive web collection policies within their respective organizations. Rather, it is anticipated that curators will create Web Collection Plans for specific collections.

Note: While curators are welcome to create comprehensive web collection plans for their collections, it is prudent for curators to keep foremost in mind that the planned Web Archiving Service will only implement a limited subset of the contents addressed in these guidelines. Furthermore, features of the WAS continue to be specified and the service should be regarded as an applied research application that will be under development over the course of the project and not as a commercially available operational system. In short, a comprehensive web collection plan might presume an ideal web archiving service while the Web-at-Risk project’s Web Archiving Service will reflect a less than ideal but nonetheless state-of-the-art service.

### 2.3 What to Include

Web collection plans should include the following sections. Not all plans will include all sub-sections. Considerations for each section are described in the remainder of these guidelines.

| | |
|---|---|
| **Section 1.** | **Mission & Scope** |
| | A. Mission Statement<br>B. User Group<br>C. Collection Subject, Theme, or Event<br>D. Curator(s) |
| **Section 2.** | **Selection Activities** |
| | A. Seed List<br>B. Initial Capture Specification<br>C. Rights Metadata |
| **Section 3.** | **Web Site Acquisition** |
| | A. Frequency of Capture<br>B. Capture Boundaries<br>C. Material Types & Formats<br>D. Interactive & Dynamic Content<br>E. Representation Metadata |
| **Section 4.** | **Descriptive Metadata Requirements** |
| | A. Level of description<br>B. Metadata elements<br>C. Controlled vocabularies |
| **Section 5.** | **Presentation & Access Requirements** |
| | A. Look-and-Feel<br>B. Dynamic Content<br>C. Multiple Types/Formats<br>D. Authenticity<br>E. Discovery<br>F. Access |
| **Section 6.** | **Maintenance & Weeding** |
| | A. Maintenance Activities<br>B. Deselection Guidelines<br>C. Collection Evaluation |
| **Section 7.** | **Preservation** |
| | A. Technology Obsolescence<br>B. Preservation Metadata |
| **Section 8.** | **Appendices** |
| | A. Submission Agreements<br>B. Web Archiving Service Agreement<br>C. Collaboration Agreements |

### 2.4 Web Archiving Service: Toolset Releases

Collection plans should be considered working documents or living guidelines. Some of the specifications or requirements identified in a plan may not be available within the WAS toolset releases. As WAS functionality is released in the 2006-2007 timeframe, collection plans may need to be adapted to the available functionality. (See Appendix A for more details.) In all cases, the functionality available within the project's Web Archive Service (WAS) will provide general guidelines and constraints within which curators' collection plans will be implemented throughout the remainder of the project.

### 2.5 Collection Policies & Plans for Web Materials: Examples

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

# 3 Mission & Scope

Selection for a web collection begins with articulation of the mission that guides collection development, a description of the user groups served by the collection, and a statement of the information need the collection will address. Web collections will generally consist of web sites united by a common subject, theme, or event in support of the mission of the organization. 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.

## 3.1 Contents

| **Section 1. Mission & Scope** |
|---|
| A. Mission Statement<br>B. User Group(s)<br>C. Collection Subject, Theme, or Event<br>D. Curator(s) |

## 3.2 What to Address

### 3.2.1 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 or archive. For others, web collection development may be more appropriately positioned under the organization's or institution's mission.

### 3.2.2 User Group

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, beginning with identifying what to collect through metadata requirements for information discovery. Be as detailed as appropriate regarding each user group's demographic characteristics and their use of web content.

Consider assessing the web information needs of the user groups. 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.

### 3.2.3 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.

### 3.2.4 Curator

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

### 3.3 Web-at-Risk Project Considerations

The Web-at-Risk project is particularly interested in preserving web-published materials that are “at risk” of becoming lost or disappearing altogether if they are not preserved. The content of the web collections for this project are web sites related to US federal and state government agencies, political policy documents, campaign literature, and information surrounding political movements and labor unions.

### 3.4 Tools and Resources

Web-at-Risk Project: Assessment Path

Needs Assessment Toolkit: Guidelines & Data Collection Tools
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

# 4 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 service provider hosting the collection. 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 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. The initial capture of web sites for a collection will be based on a list of URLs, or *seed list*, and will be conducted using either a default or customized capture specification. The initial specification may include only limited parameters, such as the links outside the seed URL host that should be captured. Evaluation of the results of the initial capture will allow curators to refine their selection decisions.

## 4.1 Contents

**Section 2. Selection Activities**

A. Seed List
B. Initial Capture Specification
C. Rights Metadata

## 4.2 What to Address

### 4.2.1 Seed List

- URL(s)
- Brief Description(s)

Identify and describe the seed list of URLs for the web sites to be included in your collection. A seed list includes one or more entry point URLs from which a web crawler begins capturing web resources.

### 4.2.2 Initial Capture Specification

- Linked web pages within the seed URL host
- Linked web pages external to the seed URL host

Selection of web sites is generally complicated by the absence of a clearly defined *object* to be assessed, evaluated, and collected. As Lyman[2] points out: "The average Web page contains 15 links to other pages or objects and five sourced objects, such as sounds or images." Evaluate the boundaries for each seed list URL and estimate the number of layers or depth of linked pages to be captured from both within the seed URL host and from external hosts.

[2] 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

#### 4.2.3 Rights Metadata

- Rights designation
- Rights metadata
- Linked and sourced objects

For each seed URL, designate a rights category that will govern the capture of its content. The choices will include those specified for the Web Archive Service[3] or similar categories, for example: "permission not needed", "notification needed", or "permission needed." As appropriate, designate a rights category to sourced or embedded objects contained in the web sites.

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.

### 4.3 Web Archive Service Toolset Considerations

Web collections are comprised of web sites selectively identified by curators and subsequently captured by an archive service provider. Using the WAS toolset release 1 (July 2006) curators will be able to specify simple, one-time crawls of the web sites targeted for inclusion in their collections. Release 2 (October 2006) will allow additional captures of seed URL content.

WAS release 4 (February 2007) will allow curators to build their collections within the archive by associating capture results from seed URLs with a collection. Release 5 (May 2007) will give curators the ability to assign a rights designation category to seed URLs: "permission not needed", "notification needed", or "permission needed."

### 4.4 Tools and Resources

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

[3] 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

# 5 Web Site Acquisition

Typically, a web archive acquires web content by harvesting or 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 transfer the files to the archive provider.

Curators are active participants in the selection and acquisition processes. Initial capture results are 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.

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 collection's seed list.

## 5.1 Contents

| **Section 3. Web Site Acquisition** |
|---|
| A. Frequency of Capture<br>B. Capture Boundaries<br>C. Material Types & Formats<br>D. Interactive & Dynamic Content<br>E. Representation Metadata |

## 5.2 What to Address

### 5.2.1 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, which could result in inconsistencies or display problems.

### 5.2.2 Capture Boundaries

- Linked pages within the seed URL host
- Linked web pages external to the seed URL host
- Linked content (information objects) outside the seed URL host

Identify the capture boundaries for hosts in the seed list. Capture boundaries refer to the range or depth or level to which a crawler will capture linked pages and sourced or embedded content. 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).

### 5.2.3 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.

### 5.2.4 Interactive & Dynamic Content

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

Consider the following: Is the site password protected? Are email links and feedback or 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?

Evaluate the web sites in the seed list and identify and describe their interactive and dynamic content. Estimate the importance of retaining the functionality of the original web site. This information will help identify the scope of the content the web collection requires.

### 5.2.5 Representation Metadata

- Technical details involved in web site design
- Software name and version used to create any content
- Search engine details
- Application code
- Viewers or plug-ins
- Structure
- Meaning

Representation metadata is the information that defines the structure (e.g., mime type) and meaning (e.g., latitude/longitude pairs in a database) of web site content and back-end databases. It consists of both information about how to read the file itself (i.e. file format) and information about the data contained within the file.

For some collections it is important to identify the representation information that must be acquired along with web sites and databases to ensure that a web site can be used and interpreted by user groups. End users should be able to understand and use the information based solely on accompanying representation information and not require additional training or explanations.

### 5.3 Web Archive Service Toolset Considerations

WAS toolsets will allow curators to evaluate the content of captured web sites and to refine the parameters for subsequent web site captures. The ability to effectively evaluate captured web sites will improve as WAS toolset releases become available to curators between June 2006 and November 2007.

In 2006, search and display tools will enable curators to evaluate captured web sites for such qualities as completeness and look-and-feel. With WAS toolset release 1 (July 2006) curators will be able to specify simple, one-time crawls and conduct basic searches of the WAS archive to display crawled web sites. WAS toolset release 2 (October 2006) will offer curators improved search and display functionality, including browsing by seed URLs, searching the archive using keywords, and navigating web sites captured at different points in time.

Release 3 (December 2006) will offer curators tools for more in-depth analysis and evaluation of captured web sites. For example, curators will be able to generate reports based on mime-types and response codes. These reports can assist in the evaluation of content object types and formats as well as in quality assessment.

Specific curator tools will be provided for event-driven collections in release 6 (July 2007). Curators will be able to set event-based capture parameters, create a form that allows others to nominate web sites for inclusion in the collection, and accept or reject the nominations.

### 5.4 Tools and Resources

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

# 6 Descriptive Metadata Creation

Because descriptive metadata updates and changes are costly, McCray and Gallagher[4] 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 within the archive, understanding the needs and salient characteristics of a collection's designated community 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?

## 6.1 Contents

| **Section 4. Descriptive Metadata Requirements** |
|---|
| A. Level of description<br>B. Metadata elements<br>C. Controlled vocabularies |

## 6.2 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, keywords). Curators may need to conform to a descriptive metadata standard established by an archive service provider, who may provide curators with the flexibility to add curator-generated or other standard metadata schemes.

Metadata schemes 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

[4] 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.

### 6.3 Web Archive Service Toolset Considerations

The WAS release 1 (July 2006) allows curators to assign descriptive terms to seed URLs and WAS release 4 (February 2007) allows curators to build their collection within the archive and assign collection-level metadata.

### 6.4 Tools and Resources

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/

# 7 Presentation

In practice, some archived web collections may present their web sites as mirror experiences of originally published web sites while other collections may be comprised of selected web-published information objects, such as video of volcanic activity, uniquely organized for the collection's user groups. Additionally, curators may designate web collections as either *visible* or *dark*, that is, as accessible or not accessible to users.

A variation on dark access might be a designation that a collection will become visible at a specific point in time in the future. This might be done to protect personal privacy or to protect markets from competition. 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.

Decisions must be made regarding the content discovery method user groups require. What kind of search mechanisms are needed (e.g., keyword search capability or subject directory 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 will they be given access to and how will the interface permit that access? For example, will users be given the capture date for each item or will users be able to "click through" to the item once they determine that they have found something they want?

Finally, how will users assess the authenticity and credibility of archived web sites and their contents? Thibodeau[5] cautions that "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 criterion users of the collection will require for the collection's web sites or information objects.

## 7.1 Contents

**Section 5. Presentation & Access Requirements**

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

[5] 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

## 7.2 What to Address

### 7.2.1 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 an archived collection and state the importance of this for the collection's user groups. If a 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, how should users be alerted to this alteration?

### 7.2.2 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 originals, curators should address some functionality issues: Will the archive allow users to access hyperlinked materials and web sites that are not located within the archive? If so, will users be alerted to the fact that they are leaving the archive? If not, will the link simply be disabled or will link information be presented along with an informative message? What about preservation of email links? How will forms be addressed within the archive? For example will the "Submit" button be disabled or will an annotated static screen shot be available?

### 7.2.3 Multiple Types/Formats

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

For collections comprised of information objects or for web sites containing multiple types and formats of information objects, will all types and formats of the objects be discoverable and made accessible to users? Is the collection subject to practice or policy guidelines specifying accessible formats and types for archived information objects that user groups will be allowed to access?

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.

#### 7.2.4 Authenticity

- Authentication process
- Indicator

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

#### 7.2.5 Discovery

- Search
- Browse
- Evaluation

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

#### 7.2.6 Access

- Dark collection
- Timed release collection
- Privacy concerns (redaction)

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

### 7.3 Web Archive Service Toolset Considerations

Between June 2006 and December 2007, curators will be the "end users" of the WAS archive. The toolsets released during this period will provide curators with search and display functions to build and manage their web collections. Metadata about web sites and collections will be created by curators as well as automatically generated by the WAS. Minimal descriptive metadata about URLs on the seed will be input beginning with WAS release 1 (June 2006). Indexes and other tools will enable meaningful discovery of captured web sites.

The WAS release 3 (December 2006) will have a report generation capability, for example, a host-level report that identifies host characteristics of interest (e.g., number of files or size) and mime-type reports (e.g., text/html or image/jpeg). These reports will provide curators with a browse interface to their captured web sites.

Release 5 (May 2007) will provide curators with the ability to assign rights metadata to seed URLs and will provide enhanced display options related to rights management for certain materials in their collections. These display options include: a list of secondary content from a crawl that does not have associated rights information and a list of hosts designated as either "permission implicit" or "permission required".

In release 7 (October 2007), curators will be able to display web site nominations for an event-based collection.

### 7.4 Tools and Resources

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 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 would 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 obviated through technological obsolescence.

## 8.1 Contents

| **Section 6. Maintenance & Weeding** |
|---|
| A. Maintenance Activities<br>B. Deselection Guidelines<br>C. Collection Evaluation |

## 8.2 What to Address

### 8.2.1 Maintenance Activities

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

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

### 8.2.2 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 object's continuing value to a collection. (Some information may have a time-related value to the identified user group(s), perhaps one year or perhaps three years.) Consider what it means to deselect a web site or a collection from an archive: Does it mean that the web site(s) will not be captured ever again? Does it mean the preservation will halt? Does it mean that the item will be removed from the archive?

#### 8.2.3 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 a collection and with weeding and other maintenance decisions.

### 8.3 Web Archive Service Considerations

The WAS release 4 (February 2007) allows creators to disassociate captured web sites from collections.

# 9 Preservation

"Technology obsolescence is generally regarded as the greatest technical threat to ensuring continued access to digital material."[6] Preservation also embraces rights management and the creation of preservation metadata. Curators must be aware of the implications, with regard to authenticity and copyright, when originally captured materials are migrated due to technological obsolescence.

## 9.1 Contents

| **Section 7. Preservation** |
|---|
| A. Technology Obsolescence<br>B. Preservation Metadata |

## 9.2 What to Address

### 9.2.1 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 old 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.

### 9.2.2 Preservation Metadata

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

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

Curators might have a role in the creation of the preservation metadata. The Open Archival Information System (OAIS) reference model[7] 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. Identify who has responsibility for creating and maintaining each element.

### 9.3 Web Archiving Service Toolset Functionality

Release 7 (October 2007) provides curators with a set of tools specific to preservation of a collection. Curators will have the ability to indicate that captured materials should be preserved in alternate formats (e.g., plain text or raster images) and to display checksum reports.

### 9.4 Tools and Resources

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

[7] 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

# 10 Collection Plan Appendices

Appendices can include a range of materials that augment the collection plan. What curators include is related to the collection being built, the archive service provider, the source of the content, and a curator’s institution or organization. The contents 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.

## 10.1 Contents

| **Section 8. Appendices** |
|---|
| A. Submission Agreements<br>B. Web Archiving Service Agreement<br>C. Collaboration Agreements |

## 10.2 What to Address

### 10.2.1 Submission Agreements

- Parties involved
- Roles & responsibilities
- Terms & conditions
  - Content included
  - Metadata provided
  - Representation information 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. At a minimum, the defined representation information for any data must be delivered with the data or must be extractable by the archive service provider.

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.

### 10.2.2 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.

### 10.2.3 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.

## 10.3 Web-at-Risk Project Considerations

### 10.3.1 Data Collaboration Agreements

The Partnership Building Path[8] of the Web-at-Risk project refers to submission agreements as *Model Data Collaboration Agreements*. "Curators will use this model agreement to begin exploring relationships with content providers to capture, manage, and preserve their content using existing WAS tools."

### 10.3.2 Web Archiving Service Agreements

*Web Archiving Service Agreements* between the California Digital Library (CDL), as service provider, and the libraries using the CDL Web Archive Service for the Web-at-Risk project will be available in draft form for curators' review in June 2006. These agreements are expected to be finalized in September 2006.

[8] California Digital Library. (2006, March.) *Partnership building: Work plan* (Rev. ed.). Retrieved May 6, 2006, from http://wiki.cdlib.org/WebAtRisk/tiki-download_file.php?fileId=116

## Appendix A. Web Archive Service: Schedule of Toolset Releases

| Date | Release | Functionality |
|---|---|---|
| Jul 2006 | Release 1 | Basic Capture<br>• Login to WAS account<br>• Access groups<br>• View list of previously run crawls<br>• Create new crawl with default settings (one-time crawls only for this release)<br>• Select one or more seeds<br>• View basic crawl reports<br>• View crawl parameters<br>• View capture help screens<br>• Interim search and display feature to view crawl results |
| Oct 2006 | Release 2 | Improved Search and Display<br>• Browse by seed URL<br>• Search by keyword<br>• Display search results<br>• Navigate through archived web sites<br>• Navigate page versions (same page captured at different points in time)<br>• View search help screens |
| Dec 2006 | Release 3 | Improved Analysis and Reports<br>• This release will include features to help curators analyze and evaluate capture results.<br>  • E.g. Allow use of mime-types report or host report to browse results<br>  • E.g. Allow use of response code report to conduct a quality review |
| Feb 2007 | Release 4 | Collection Building<br>• List collections<br>• Add new collection and associated metadata<br>• Edit a collection<br>• Associate capture results with a collection<br>• Disassociate capture results from a collection<br>• Browse and display content by collection<br>• View collection help screens |

| Date | Release | Functionality |
|---|---|---|
| May 2007 | Release 5 | Administration and Curatorial Rights Management<br>Administrator<br>• Create and edit user account<br>• Create and edit user group<br>• Assign user to group<br>• Define user's role in group<br>Curator<br>• Link co-curator to a collection<br>• Select rights designation for a seed URL ("permission not needed", "notification needed", "permission needed")<br>• Link rights metadata to a seed URL (contact information, contact history, etc.)<br>• Display rights associated with an item<br>• Display a list of secondary content from a crawl that does not have associated rights information<br>• Display a list of hosts requiring rights action (i.e. "notification" or "permission required")<br>• Update a rights record (e.g. to indicate that permission to capture was granted) |
| Jul 2007 | Release 6 | Event-based Capture and Enhancements<br>• Create event-based capture specification<br>• Generate site nomination form for event capture<br>• Review (accept or reject) nominated seed(s) for capture<br>• View help screens for event-based capture activities |
| Oct 2007 | Release 7 | Preservation Features, Help Screens, and Reports<br>• Mark capture specification for alternate format preservation (e.g. plain text or raster images)<br>• Display alternate item formats<br>• Display checksum report<br>• View preservation feature help screens |
| Nov 2007 | Release 8 | Integration of User Feedback and Refinement of Software and Documentation |

## Appendix B. 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

*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