Web Collection Plan Overview:

Considerations for Project Curators

A Companion Document to the
Web Collection Plan Template

August 24, 2006

Prepared by:

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

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

Introduction ..................................................................................................3

Section 1. Mission & Scope ...........................................................................8
Section 2. Selection ...................................................................................10
Section 3. Acquisition ...............................................................................14
Section 4. Descriptive Metadata ...............................................................16
Section 5. Presentation and Access .............................................................18
Section 6. Maintenance and Weeding .........................................................22
Section 7. Preservation .............................................................................24
Section 8. Collection Plan Appendices .......................................................26

Appendix A. Web Archiving Service: Schedule of Toolset Releases ..........28
Appendix B. Mapping of Collection Plan Sections to WAS Toolset Releases ....30
Appendix C. Glossary .................................................................................33
Appendix D. Collection Planning Resources ..............................................37
Appendix E. Collection Plan Example: UCLA Online Campaign Literature Archive ..40
Appendix F. Collection Plan Example: CyberCemetery at UNT ..................57
Introduction

The guidelines in this document are intended to assist the Web-at-Risk project’s curators in developing plans for the collections they will create using the project’s Web Archiving Service (WAS). A web collection typically consists of a group of web-sites related by a common subject, theme, or event. 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.\(^1\)

The Web Archiving Service (WAS)

The Web Archiving Service (WAS) will consist of new services and tools that integrate with and take advantage of the overall framework and existing resources for application development and data storage within the California Digital Library. Curators will use tools 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.

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 the ability to search, browse, and display collected and preserved web content.

From July 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.

\(^1\) Some libraries refer to collection plans as collection policies.
Table 1. WAS Toolset Releases

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. The planned WAS will not provide functional support for all of the collection development considerations discussed in these guidelines. A good example of a function that will not be available is the ability to specify the type and format of materials to be captured.

Collections

Curators will use WAS tools to build and manage collections of captured materials. From an application perspective, a collection within the WAS archive consists of a set of index entries that point to captured copies of web-published materials. It is conceivable that collections might be comprised of any captured materials in the WAS archive, regardless of which curator initially requested their capture. The only caveat is that copyrights and legal arrangements pertaining to the materials must be honored.

Collection examples:

- a set of government and organizational web sites related to water conservation for a particular geographic area
- a set of web sites reflecting a range of perspectives related to federal immigration policy

The WAS may not permit curators to define collections targeting specific web-published materials contained within captured web sites. For example, the WAS may not permit a curator to define a collection consisting only of web-published documents in PDF format contained within the archived web sites. Rather, curators may be required to construct collections consisting of captured sites in their entirety.
Collection Plans

As conceived in this document, web collection plans provide guidance for managing collections of web-published materials created for specific groups of users within an institution. This is not unlike the general role collection plans or collection policies often serve for traditional collections within a library. Figure 1 identifies the major phases and activities involved in web collection development. Collection plans address each of these phases. Although there is a general trajectory of activities in the web collection development process, with selection occurring early on and preservation toward the end, tasks are not necessarily completed in a linear fashion.

<table>
<thead>
<tr>
<th>PHASES</th>
<th>SELECTION</th>
<th>CURATION</th>
<th>PRESERVATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>Description</td>
<td>Preservation</td>
<td></td>
</tr>
<tr>
<td>Acquisition</td>
<td>Organization</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Presentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maintenance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deselection</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Collection Development Phases

A collection plan should provide sufficient guidance for conducting activities in each phase. The plan can be organized in sections that roughly correspond to each of the major activities involved in collection development. For each section of a collection plan discussed in this document, Appendix B identifies the WAS release(s) that is expected to include related functionality. This mapping of plan sections to WAS toolset releases should assist curators in setting reasonable expectations of the WAS as they develop and implement their collection plans.

Table 2 lists web collection development activities and indicates which activities you should consider addressing in your collection plan and which areas you will likely address using WAS tools and forms. For several areas, you may find it beneficial to include information in both your collection plan and the WAS. In areas where information is included in both places, the collection plan will likely provide more general guidance and context for the collection, while the WAS will likely contain very specific information regarding the collection. Unless it seems appropriate, it is not necessary to duplicate the specific information entered into the WAS in the collection plan. Additionally, curators may find that the WAS tools provide the capability to generate detailed reports that can later be added to the collection plan.

The collection development activities listed in Table 2 are explored in the remainder of these guidelines. Included in Appendices E and F are example plans that were created using these guidelines and the associated collection plan template. It may be helpful to reference these example plans as you create your collection plan.
<table>
<thead>
<tr>
<th>Collection Development Area</th>
<th>Collection Plan</th>
<th>WAS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Section 1. Mission &amp; Scope</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Mission Statement</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B. User Group(s)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>C. Collection Subject, Theme, or Event</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>D. Curator(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Section 2. Selection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Seed List</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>B. Capture Scope</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>C. Rights Management</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Section 3. Acquisition</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Frequency of Capture</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>B. Capture Scope</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>C. Material Types &amp; Formats</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>D. Interactive &amp; Dynamic Content</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Section 4. Descriptive Metadata</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Level of description</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>B. Metadata elements</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>C. Controlled vocabularies</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Section 5. Presentation &amp; Access</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Discovery</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>B. Access</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>C. Look-and-Feel</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>D. Dynamic Content</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>E. Multiple Types/Formats</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>F. Authenticity</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Section 6. Maintenance &amp; Weeding</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Maintenance Activities</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>B. Deselection Guidelines</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>C. Collection Evaluation</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Section 7. Preservation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Technology Obsolescence</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>B. Preservation Metadata</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Section 8. Appendices</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Submission Agreements</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>B. Web Archiving Service Agreement</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>C. Collaboration Agreements</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

**Table 2. Collection Development Considerations**
Due Dates

Curators should use the collection plan template\textsuperscript{2} to create a Web Collection Plan for the collections they plan to build as part of the Web-at-Risk project. Not all areas discussed in these guidelines will need to be included in every collection plan. Collection plans are living documents that evolve and change over time and you may not be able to include all of the areas discussed in these guidelines in your initial plan. Alternately, you may discover that some of the information you elect to include in your collection plan, while not directly relevant to the WAS, provides a broader context for web collection development within your organization.

Please email your collection plan by October 15, 2006 to:

Kathleen Murray
krmurray@unt.edu

Overview of Contents

The remainder of this document discusses each section of a collection plan and identifies aspects of web collection development that curators may need to consider as they build and manage their collections using the WAS tools. A summary of WAS toolset releases along with a mapping of collection plan sections to toolset releases are included in appendices. Some terms used in this document may be unfamiliar or unclear to readers. A glossary is provided to clarify the meaning of those terms as used in this document. Finally, a list of supplemental collection planning resources and two examples of completed web collection plans are included for reference. The following is a list of the appendices in this document:

Appendix A. Web Archiving Service: Schedule of Toolset Releases
Appendix B. Mapping of Collection Plan Sections to WAS Toolset Releases
Appendix C. Glossary
Appendix D. Collection Planning Resources
Appendix E. Collection Plan Example: UCLA Online Campaign Literature Archive
Appendix F. Collection Plan Example: CyberCemetery at UNT

\textsuperscript{2} The template is a WORD document and is available on the project wiki at: http://wiki.cdlib.org/WebAtRisk/tiki-download_file.php?fileId=171
Section 1. Mission & Scope

Overview
Web collection plans begin with articulating the mission that guides collection development, describing the user groups 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.

<table>
<thead>
<tr>
<th>Section 1. Mission &amp; Scope</th>
<th>Collection Plan</th>
<th>WAS Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Mission Statement</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>B. User Group(s)</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>C. Collection Subject, Theme, or Event</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>D. Curator(s)</td>
<td>√</td>
<td></td>
</tr>
</tbody>
</table>

Collection Planning Considerations

A. Mission Statement
What is 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 the mission of the organization or institution.

B. User Group(s)
Who are 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.

C. Collection Subject, Theme, or Event
What subject area or theme 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. How does the collection support the mission of the library, organization, or institution?
D. Curator(s)

Who is responsible for this collection? Consider including each curator’s major responsibilities within their organization or institution and contact information, as appropriate, for all collaborators involved in this collection.
Section 2. Selection

Overview

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\(^3\) 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 and are often within the same domain or sub-domain name space (e.g., unt.edu or library.unt.edu). The web pages within a web site are often published and maintained by a single person or organization, although wider collaborations and social publishing are becoming 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.

The WAS tools will allow curators to specify web-published materials, as represented by specific URLs, whose content they would like captured and preserved in the WAS archive. Some web content cannot be effectively captured. In particular, some web pages that rely on databases for their content or server-side code for their operation are outside the planned scope of the WAS capture tools.

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 Scope
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 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 scope of a crawl. From a web crawler perspective, capture scope generally refers to the number of linked URLs or link-hops away from a seed URL or local host from which a crawler should capture content.

<table>
<thead>
<tr>
<th>Section 2. Selection</th>
<th>Collection Plan</th>
<th>WAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Seed List</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>B. Capture Scope</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>C. Rights Management</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Collection Planning Considerations

A. Seed List
- Site Name
- URL(s)
- Brief Description(s)

A seed list includes one or more entry point URLs from which a web crawler begins capturing web-published materials. For each site you wish to capture, evaluate whether the site is served by a single host or multiple hosts. Some sites are generated from multiple servers, so more than one URL is required to effectively capture that site’s content. This is frequently the case when a site includes a high number of PDF, image, or multimedia files. Example:

**Site Name:**
University of California Office of the President (UCOP)

**URL(s):**
- http://www.ucop.edu/
- http://atyourservice.ucop.edu/
- http://budget.ucop.edu/

Determine if the site you wish to capture is part of a parent-site. If so, clearly distinguish the name of this site from the name of the parent-site and specify the exact subdirectory path in the URL you supply. Example:

**Site Name:**
University of California Office of the President – Division of Health Affairs

**URL:**
- http://www.ucop.edu/healthaffairs/

Identify the name of each site in the seed list and briefly describe the site. The site name may be different from the “title” that the browser displays at the top of a web page or from the information contained in HTML meta tags. (The HTML title will be captured when the page itself is captured.) The site name provides an overall descriptive identifier for the site...
and may be extracted from the site itself or created as needed by the curator who is managing the collection.

Briefly describe each site whose URL is included in the seed list.

B. Capture Scope
   - Link-hops away from the seed URL or local host

Evaluate the capture scope for each URL in the seed list by estimating the number of link-hops away from the local host from which content should be captured. As described earlier in this section, capture scope relates both to how web sites are structured and to how a web crawler captures 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. The WAS crawler will capture embedded objects, such as images and style sheets no matter what host they reside on; curators do not need to determine what hosts are serving embedded objects.

If you determine that pages from external hosts are relevant, determine how much content from those external hosts you wish to capture. You may only wish to capture the immediately linked pages, or you may wish the crawler to gather as much information as possible from the external sites. If an external site has highly relevant and critical information, you may wish to capture it separately.

Three probable scope settings have been tentatively identified for the WAS:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>This page only</td>
</tr>
<tr>
<td>2.</td>
<td>This site only</td>
</tr>
<tr>
<td>3.</td>
<td>This site + 1 link off*</td>
</tr>
</tbody>
</table>

* “1 link off” refers to links to external hosts and is not the same as link-hops discussed in the rest of this section.

Be aware that in WAS release 1, the crawler will be preset to capture content up to three link-hops away from seed URLs. Curators will not be able to change this setting. In future WAS releases it may be possible to specify the number of link-hops away from the seed URL the crawler goes. In this case, curators may need to determine how many hops away from the original seed URL the crawler must go to capture the desired content.

C. Rights Management
   - Rights evaluation
   - Rights designation
   - Rights metadata

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. It would be helpful to obtain and review any rights statements from the sites on the seed list. In the collection plan, specify any rights policies or practices that govern collection development.
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 WAS\textsuperscript{4} will likely include such categories as: “permission not needed”, “notification needed”, or “permission needed.”

Rights metadata for each seed URL might include: contact information, contact history, date permission granted. Additional rights information may be established or may be required by either a content provider or the WAS provider (the CDL).

Section 3. Acquisition

Overview

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 scope, which were identified in Selection, are refined in Acquisition.

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.

<table>
<thead>
<tr>
<th>Section 3. Acquisition</th>
<th>Collection Plan</th>
<th>WAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Frequency of Capture</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>B. Capture Scope</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>C. Material Types &amp; Formats</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>D. Interactive &amp; Dynamic Content</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Collection Planning Considerations

A. 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, 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. It may be possible to run link checking software to identify such inconsistencies in captured materials.

B. Capture Scope
   - Link-hops away from the seed URL or local host

If necessary to capture desired content, re-evaluate and refine the capture scope estimates for hosts in the seed list. Keep in mind that there is no one web site organizational structure; some web sites are organized hierarchically and some are not. Remember that the WAS crawler will automatically capture embedded objects, such as images and style sheets no matter what host they reside on; curators do not need to determine what hosts are serving embedded objects.
C. Material Types & Formats
   - Excluded types
   - Excluded formats

While the WAS may not allow curators to exclude web materials from crawls based on their type and format, consider if web material type and format are important for your collection. If one or both is important, identify in the collection plan any specific types or formats of web-published materials that are not desired for your collection. Likewise, identify any types or formats of materials that are specifically important to your collection and explain their importance. 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.

D. Interactive & Dynamic Content
   - Authentication (username/password)
   - Email links
   - Forms
   - Database-generated pages (based on user queries)
   - Dynamically or programatically generated web pages

Evaluate the web sites in the seed list for 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?

In the collection plan, estimate the importance of retaining the interactive and dynamic functionality of the original web site. Identify any web collection policies or practices related to requirements for identifying content that is no longer active, for example, creating tags to alert users to inactive email links.
Section 4. Descriptive Metadata

Overview

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 the WAS 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.

Because descriptive metadata updates and changes are costly, McCray and Gallagher\(^5\) 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?

<table>
<thead>
<tr>
<th>Section 4. Descriptive Metadata</th>
<th>Collection Plan</th>
<th>WAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Level of description</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B. Metadata elements</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>C. Controlled vocabularies</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Collection Planning Considerations

A. Level of description
   - Collection level
   - Web site level
   - Information object level

Identify the level of description required by the collection’s user group(s).

B. Metadata elements
   - Essential
   - Desirable

Identify any descriptive metadata elements of importance for information discovery by the collection’s user group(s) and rate these as either essential or desirable.

C. **Controlled vocabularies**

Identify any controlled vocabulary sources that are appropriate for the listed metadata elements.
Section 5.  Presentation and Access

Overview

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\(^6\) 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.”

---

Section 5. Presentation & Access

<table>
<thead>
<tr>
<th>A. Discovery</th>
<th>Collection Plan</th>
<th>WAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. Access</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Look-and-Feel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. Dynamic Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. Multiple Types/Formats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F. Authenticity</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Collection Planning Considerations

A. **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, are simple keyword searches sufficient or do users require advanced search functionality? 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?

B. **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.

C. **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?

D. **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?

E. 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, institutional practice or policy will likely influence the types and formats that will 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.

F. 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.7

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 an archive to be certified by some established process such as the certification process for digital repositories proposed by the Research Libraries Group8?

---

7 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.

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 (i.e., the CDL for the WAS)?
Section 6. Maintenance and Weeding

Overview

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.

<table>
<thead>
<tr>
<th>Section 6. Maintenance &amp; Weeding</th>
<th>Collection Plan</th>
<th>WAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Maintenance Activities</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>B. Deselection Guidelines</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>C. Collection Evaluation</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

Collection Planning Considerations

A. Maintenance Activities
- Seed lists
- Capture specification for seed lists
- Rights metadata
- Descriptive metadata
- Collection content

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.

B. 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?
C. 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).
Section 7. Preservation

Overview

“Technology obsolescence is generally regarded as the greatest technical threat to ensuring continued access to digital material.”\(^9\) 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.

<table>
<thead>
<tr>
<th>Section 7. Preservation</th>
<th>Collection Plan</th>
<th>WAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Technology Obsolescence</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B. Preservation Metadata</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Collection Planning Considerations

A. 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 preservation methods that are acceptable for use on this collection by an archive service provider. Identify a process for evaluating the impact of the preservation methods on the authenticity of materials and their copyright protection.

B. 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

The Open Archival Information System\textsuperscript{10} (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.

Section 8. Collection Plan Appendices

Overview

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 to include in appendices. (See the appendices included in the two example collection plans.) Alternately, the appendix might simply be a reference list of applicable agreements, policies, practices, standards, and guidelines for the collection. Some suggested appendices for Web-at-Risk collections appear in the table below. If a collection will be built as the result of collaboration amongst curators at two or more institutions, a collaboration agreement is recommended.

<table>
<thead>
<tr>
<th>Section 8. Appendices</th>
<th>Collection Plan</th>
<th>WAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Submission Agreements</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>B. Web Archiving Service Agreement</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>C. Collaboration Agreements</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

Collection Planning Considerations

A. 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.

B. Web Archiving Service Agreement

- Parties involved
- Roles & responsibilities
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.

**C. 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 Archiving Service: Schedule of Toolset Releases

<table>
<thead>
<tr>
<th>Date</th>
<th>Release</th>
<th>Functionality</th>
</tr>
</thead>
</table>
| Jul 2006 | Release 1 | Basic Capture  
- Login to WAS account  
- Access groups  
- View list of previously run crawls  
- Create new crawl with default settings (one-time crawls only for this release)  
- Select one or more seeds  
- View basic crawl reports  
- View crawl parameters  
- View capture help screens  
- Interim search and display feature to view crawl results |
| Oct 2006 | Release 2 | Improved Search and Display  
- Browse by seed URL  
- Search by keyword  
- Display search results  
- Navigate through archived websites  
- Navigate page versions (same page captured at different points in time)  
- View search help screens |
| Dec 2006 | Release 3 | Improved Analysis and Reports  
- This release will include features to help curators analyze and evaluate capture results.  
  - E.g. Allow use of mime-types report or host report to browse results  
  - E.g. Allow use of response code report to conduct a quality review |
| Feb 2007  | Release 4 | Collection Building  
- List collections  
- Add new collection and associated metadata  
- Edit a collection  
- Associate capture results with a collection  
- Disassociate capture results from a collection  
- Browse and display content by collection  
- View collection help screens |
<table>
<thead>
<tr>
<th>Date</th>
<th>Release</th>
<th>Functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 2007</td>
<td>Release 5</td>
<td>Administration and Curatorial Rights Management</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Administrator</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Create and edit user account</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Create and edit user group</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Assign user to group</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Define user’s role in group</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Curator</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Link co-curator to a collection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Select rights designation for a seed URL (&quot;permission not needed&quot;, &quot;notification needed&quot;, &quot;permission needed&quot;)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Link rights metadata to a seed URL (contact information, contact history, etc.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Display rights associated with an item</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Display a list of secondary content from a crawl that does not have associated rights information</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Display a list of hosts requiring rights action (i.e. &quot;notification&quot; or &quot;permission required&quot;)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Update a rights record (e.g. to indicate that permission to capture was granted)</td>
</tr>
<tr>
<td>Jul 2007</td>
<td>Release 6</td>
<td>Event-based Capture and Enhancements</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Create event-based capture specification</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Generate site nomination form for event capture</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Review (accept or reject) nominated seed(s) for capture</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- View help screens for event-based capture activities</td>
</tr>
<tr>
<td>Oct 2007</td>
<td>Release 7</td>
<td>Preservation Features, Help Screens, and Reports</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Mark capture specification for alternate format preservation (e.g. plain text or raster images)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Display alternate item formats</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Display checksum report</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- View preservation feature help screens</td>
</tr>
<tr>
<td>Nov 2007</td>
<td>Release 8</td>
<td>Integration of User Feedback and Refinement of Software and Documentation</td>
</tr>
</tbody>
</table>
Appendix B. Mapping of Collection Plan Sections to WAS Toolset Releases

Features and functionality 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. This appendix provides an overview of the features and functionality available in the WAS toolset releases as they apply to the sections of a collection plan.

Section 1. Mission and Scope

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.

Section 2. Selection Activities

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.

Section 3. Web Site Acquisition

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 captured 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.
Section 4. Descriptive Metadata Creation

WAS release 4 (February 2007) allows curators to build their collection within the archive and assign collection-level metadata.

Section 5. Presentation

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. 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 requiring rights action such as notification sent to the content provider or acquisition of permissions from content provider.

In release 6 (July 2007), curators will be able to display web site nominations for an event-based collection.

Section 6. Maintenance and Weeding

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

Section 7. Preservation

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.

Section 8. Appendices

Data Collaboration Agreements

The Partnership Building Path\(^\text{11}\) 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.”

Web Archiving Service Agreements

*Web Archiving Service Agreements* between the California Digital Library (CDL), as service provider, and the libraries using the CDL Web Archiving 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.
### Appendix C. Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acquisition</td>
<td>For digital materials, see Capture</td>
</tr>
<tr>
<td>Archive</td>
<td>Archives are repositories of content for which someone or some organization has accepted preservation responsibility. See also: Digital Archive and Web Archive.</td>
</tr>
<tr>
<td>Authenticity</td>
<td>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.</td>
</tr>
<tr>
<td>Automated Capture Tool</td>
<td>See Crawler</td>
</tr>
<tr>
<td>Baseline Metadata</td>
<td>Baseline metadata is machine-generated and captured by a crawler at the time of data capture.</td>
</tr>
<tr>
<td>Born-digital</td>
<td>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 analog counterpart.</td>
</tr>
<tr>
<td>Capture</td>
<td>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.</td>
</tr>
<tr>
<td>Collection</td>
<td>A group of materials related by common ownership or a common theme or subject matter. Collections are owned and/or maintained by an organization or institution.</td>
</tr>
<tr>
<td>Content Provider</td>
<td>The organization responsible for making the source material available on the Web.</td>
</tr>
<tr>
<td>Crawl</td>
<td>The activity conducted by a web crawler.</td>
</tr>
<tr>
<td>Crawler</td>
<td>Software that explores the web and collects data about its contents. A crawler can also be configured to capture web-published materials. It starts a capture process from a Seed List.</td>
</tr>
<tr>
<td>Curation Process</td>
<td>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.</td>
</tr>
<tr>
<td>Dark Archive</td>
<td>A digital archive to which no end user access is permitted.</td>
</tr>
<tr>
<td><strong>Dark Web</strong></td>
<td>See Deep Web</td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td><strong>Deep Web</strong></td>
<td>Materials available via the World Wide Web that are invisible to or inaccessible by crawlers. This may be because the materials (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.</td>
</tr>
<tr>
<td><strong>Digital Archive</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Digital Collection</strong></td>
<td>A collection consisting entirely of born-digital or digitized objects.</td>
</tr>
<tr>
<td><strong>Digital Object</strong></td>
<td>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 originally available in various formats, (e.g., print books or audio tapes) and born-digital objects.</td>
</tr>
<tr>
<td><strong>Dynamic Web Page</strong></td>
<td>A web page created automatically by software on a 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.</td>
</tr>
<tr>
<td><strong>Emulation</strong></td>
<td>A method by which newer software interacts with older digital materials and displays the result using the same commands and formatting that the software that created the materials used. Emulation provides a means of allowing digital materials to be preserved without altering their binary format.</td>
</tr>
<tr>
<td><strong>Enriched Metadata</strong></td>
<td>Enriched metadata is generally specific to an organization and contains a mixture of baseline metadata and human-generated metadata added subsequent to data capture.</td>
</tr>
<tr>
<td><strong>Entry Point URL</strong></td>
<td>See Seed URL.</td>
</tr>
<tr>
<td><strong>External Link</strong></td>
<td>A URL that links to web-published materials residing on a different host.</td>
</tr>
<tr>
<td><strong>Fixity</strong></td>
<td>The extent to which an archived object remains unchanged over time regardless of access or 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.</td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Harvest</strong></td>
<td>See Capture</td>
</tr>
<tr>
<td>Information Object</td>
<td>See Digital Object</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Ingest</td>
<td>For the Web-at-Risk project, ingest refers to the process of packaging captured materials and moving them to a repository for long-term storage.</td>
</tr>
<tr>
<td>Integrity</td>
<td>A digital object’s integrity is maintained as long as the bits contained in the object are not altered in an unauthorized manner.</td>
</tr>
<tr>
<td>Invisible Web</td>
<td>See Deep Web</td>
</tr>
<tr>
<td>Light Archive</td>
<td>A digital archive accessible to end-users.</td>
</tr>
<tr>
<td>Link-hop</td>
<td>An explicit hyper link leading from one web page to another web page. A web crawler may not consider embedded links, such as links to images and style sheets, as link-hops.</td>
</tr>
<tr>
<td>Medium</td>
<td>The delivery vehicle for the content. For example: CD-ROM, network, book, etc.</td>
</tr>
<tr>
<td>Migration</td>
<td>A method of preserving digital materials and access to those materials by copying or reformatting the materials while preserving their intellectual content.</td>
</tr>
<tr>
<td>Mime Type</td>
<td>A standard identifier included in an HTTP message that defines the type and format of the message content.</td>
</tr>
<tr>
<td>Opt-in</td>
<td>A collection policy in which the archive owner seeks explicit permission from content owners before collecting materials.</td>
</tr>
<tr>
<td>Opt-out</td>
<td>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.</td>
</tr>
<tr>
<td>Persistent Name</td>
<td>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.</td>
</tr>
<tr>
<td>Seed List</td>
<td>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.</td>
</tr>
<tr>
<td>Seed URL</td>
<td>A URL that specifies a starting address that a web crawler uses to capture content. Also called a Targeted URL or Entry Point URL.</td>
</tr>
<tr>
<td>Spider</td>
<td>See Crawler</td>
</tr>
<tr>
<td>Targeted URL</td>
<td>See Seed URL</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td><strong>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.)</strong></td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Visibility</strong></td>
<td><strong>The extent of end user access allowed to a digital archive.</strong></td>
</tr>
<tr>
<td><strong>Web Archive</strong></td>
<td><strong>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.</strong></td>
</tr>
<tr>
<td><strong>Web Crawler</strong></td>
<td><strong>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.</strong></td>
</tr>
</tbody>
</table>
| **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 D. Collection Planning Resources

Policies & Plans: Web Collections & Digital Preservation\(^{12}\)

- National Archives of Australia
  - Archiving Web Resources: A policy for keeping records of web-based activity in the Commonwealth Government
  - Archiving Web Resources: Guidelines for keeping records of web-based activity in the Commonwealth Government

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

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

- 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
  - Note: Organization Items on the checklist are more in line with what we are addressing under Policy.

- 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

\(^{12}\) The collection of digital information, especially with regard to web-published materials, remains a new and emerging practice as does the authoring of collection planning guidelines regarding these types of collections. As a result, there are few guidelines or examples that specifically address collection plans for web-published materials. The references listed are provided as general references.
Selection

Digital Preservation Coalition
Decision Tree for Selection of Digital Materials for Long-term Retention

Interactive Version of Decision Tree:

National Library of Australia
Online Australian Publications: Selection Guidelines for Archiving and Preservation by the National Library of Australia

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


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

Descriptive Metadata


RLG: Research Libraries Group
Descriptive Metadata Guidelines for RLG Cultural Materials

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/marxml/
**Authenticity**


**Digital Preservation**

National Library of Australia: PADI - Preserving Access to Digital Information


“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:


**Repositories: Institutional Repositories & Trusted Digital Repositories**


Appendix E. Collection Plan Example: UCLA Online Campaign Literature Archive

**Note:** The following sixteen pages contain the collection plan for the UCLA Online Campaign Literature Archive. It is provided as an example of a web collection plan using these guidelines and the associated template.
Collection Plan

for

UCLA Online Campaign Literature Archive

August 18, 2006

Prepared by:

Gabriella Gray & Scott Martin
UCLA
gsgray@library.ucla.edu
smartin@library.ucla.edu
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mission &amp; Scope</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Selection</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>Acquisition</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>Descriptive Metadata</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>Presentation and Access</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>Maintenance and Weeding</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>Preservation</td>
<td>11</td>
</tr>
<tr>
<td>A</td>
<td>Appendix A. Web Archiving Service Agreement</td>
<td>12</td>
</tr>
<tr>
<td>B</td>
<td>Appendix B. Seed List</td>
<td>13</td>
</tr>
<tr>
<td>C</td>
<td>Appendix C. Metadata Elements</td>
<td>14</td>
</tr>
<tr>
<td>D</td>
<td>Appendix D. Definitions of Subject Facets</td>
<td>15</td>
</tr>
</tbody>
</table>
Section 1. Mission & Scope

A. Mission Statement

The mission of the UCLA Library is to provide access to and delivery of information resources to UCLA students, faculty, and staff in support of the research and instructional mission of the university. The Library develops, organizes, and preserves collections for optimal use and provides links to remote information sources. The Library provides services, including instruction for information literacy and information management, to enable its users to fulfill their academic and intellectual needs. The Library provides resources and services to non-UCLA users to the extent possible.

B. User Groups

The primary user groups of the UCLA Online Campaign Literature Archive are UCLA students, faculty and staff. Secondary users include scholars and students across the world. The information contained in campaign web sites provide a valuable and unique primary resource for scholars researching California and more specifically Los Angeles area politics and elections. The Archive preserves and provides access to these ephemeral web sites in support of the research and instructional mission of the UCLA library.

C. Collection Subject, Theme, or Event

The Archive collects, organizes and preserves campaign web sites created for elections affecting the Los Angeles area. Web sites for local, state, and federal offices and ballot measures are collected.

D. Curator(s)

Gabriella Gray, Project Director
UCLA Online Campaign Literature Archive
Young Research Library
Los Angeles, CA  90095-1575
310-825-4567
gsgray@library.ucla.edu

Scott Martin, Computer Resource Specialist
UCLA Online Campaign Literature Archive
Young Research Library
Los Angeles, CA  90095-1575
310-206-7447
smartin@library.ucla.edu
Section 2. Selection

A. Seed List

Entire web sites are the intended materials. Each election generates a unique seed list of web sites based on the following selection criteria:

- Sites must be devoted to a specific election, not an ongoing organization or cause.
- Sites must be about candidates or measures who appear on the ballot.
- Elections for executive and legislative offices are collected, not judicial.
- Legislative candidates must be from districts which are substantially within Los Angeles County.
- State elections gathered include U.S. congressmen, California state-wide offices and measures, and California state legislators. (Note: Presidential election sites are excluded because they are captured by the Library of Congress.)
- Municipal elections gathered include those associated with Los Angeles County; the cities of Los Angeles, Santa Monica, Beverly Hills, and Culver City; and any special districts (school districts, water districts, etc.) that substantially overlap those cities.
- Election sites must be discrete, wholly contained within a single domain or directory which is not shared with other material.

The web sites included in this collection are listed in Appendix B.

B. Capture Scope

The default capture scope is “This site only”. However, an assessment of each URL on a specific election’s seed list is needed to determine the individual capture scope. When maximum links are set, it refers to the maximum number of links within the target site, usually to prevent continual downloads from calendars and other extendable database-produced pages within the site.

C. Rights Management

The web sites collected by the UCLA Online Campaign Literature Archive are protected by the U.S. Copyright Law (Title 17, U.S.C.). The materials may also be subject to publicity rights, privacy rights, or other legal interests. The Library makes these web sites available based on the principles of fair use. Transmission or reproduction of materials protected by copyright beyond that allowed by fair use requires the written permission of the copyright owners. Researchers should consult the sites themselves for information about rights, contacts, and permissions. Responsibility for making an independent legal assessment of an item and securing any necessary permissions ultimately rests with those persons wishing to use the item(s).

The UCLA Online Campaign Literature Archive uses a combination of notification needed and opt-out policy. Contact information found on the original web site or other sources (such as the Secretary of State’s Certified List of Candidates) is recorded and used to notify the copyright owners that their web site has been captured and will be made accessible via the UCLA Online Campaign Literature Archive interface. Copyright owners are given the option...
to opt-out of the collection and have the captured web site or specific portions of it removed from the collection. The archive honors the right to remove specific pages or files, but will not alter content within a specific page or file.

Rights metadata includes contact information and contact history.
Section 3. Acquisition

A. Frequency of Capture

Each web site will initially be captured one time only on the day of the election or during the week leading up to the election.

B. Capture Scope

The curators will evaluate the capture results immediately after the web sites are harvested to determine if the content was successfully captured. If necessary the capture scope will be adjusted and the capture re-executed until the best possible results are attained.

C. Material Types & Formats

The Archive attempts to capture all possible material types and formats.

D. Interactive & Dynamic Content

The collection does not include password protected materials. E-mail links and comment forms are captured, but no attempt is made to insure their functionality. Users are alerted to the possible loss of functionality within the archived web sites by the technical notes associated with the entire collection and posted on the Archive's web site.

Database-generated web pages and web pages created on-the-fly are captured where possible, but converted to static html in order to minimize the effects of technological obsolescence.
Section 4. Descriptive Metadata

A. Level of description

Descriptive metadata will be applied at the web site level.

B. Metadata elements

The metadata elements are based upon Dublin Core and were selected to optimize end user discovery. Note, the UCLA Online Campaign Literature Archive also includes scanned images from the extensive print UCLA Campaign Literature Archive. Care was taken in the selection of metadata elements to ensure they would apply to both the scanned print material and web sites included in the collection. The metadata elements selected for this collection are listed in Appendix C.

C. Controlled vocabularies

The Archive employs a system of faceted subject headings using controlled vocabularies that are derived from LCSH and locally defined authority lists such as the Certified List of Candidates. Detailed definitions of the subject facets are listed in Appendix D.
Section 5. Presentation and Access

A. Discovery

The UCLA Library Catalog contains a collection level record for the Online Campaign Literature Archive but access to the collection is provided by the UCLA Digital Library Program. Keyword searching is available from the basic search screen and searches the title, subject, language, year and type fields. The advanced search screen allows users to specify that their term(s) appear within specific fields (keyword, title, subject, language, type), limit by date and sort results. Truncation, wildcard and phrase searching are available from both the basic and advanced search screens. The interface provides users with the ability to browse the collection based on subject categories. Search tips are provided on each screen.

Initial search results are returned in a brief format which includes the title of the web site, a link to the full record and a link to the archived web site. The full record displays the title, date, subjects, description, coverage, language, admin notes and a link to the archived web site (see Appendix C for a full description of these metadata elements).

B. Access

The UCLA Online Campaign Literature Archive is visible to all users with a connection to the Internet and web browser. The content of the Archive is made accessible to all users via the Fair Use doctrine.

C. Look-and-Feel

Retaining the “look-and-feel” of the captured web sites is a primary goal of the collection, second only to retaining the content. When substantial modifications are made to the captured material, users are alerted by notes in the metadata.

D. Dynamic Content

It is imperative that users be allowed to access hyperlinks to materials outside the web archive, but no attempt is made to ensure that such links are still active. Users would be alerted to the fact that they are leaving the archive only if that function was captured from the original site. E-mail links are preserved as-is.

The Archive makes every attempt to ensure that the captured site mirrors the look-and-feel of the original content, including navigation, but does not preserve the ability to interact with the site by inputting new data. Forms that interact with dynamic databases, e.g. to accept donations, collect user contact information, or solicit comments, are captured as-is, with no attempt made to maintain the functionality of the form. If the form relies on scripts or data sets within the captured website then it will be non-functional. If the form interacts with scripts or databases on external servers, then, like other external links, it will remain
functional as long as the proper scripts still exist on the host server, but no attempt is made to track or support this. In contrast, forms that are used as a navigation tool to link to other pages in the captured website, e.g. drop-down menus, must remain functional in order to retain the archived site's look and feel and to allow full access to all captured content. If necessary, scripts and other code may need to be edited so that they continue to work in a standard browser.

Users are alerted to the possible loss of functionality in e-mail addresses, submission forms and external links by the technical notes associated with the entire collection and posted on the Archive's web site.

E. Multiple Types/Formats

All types and formats of materials in the captured content should be discoverable and accessible by all users from all locations.

F. Authenticity

Users of the Archive expect the captured web sites to retain as much of the original content and look-and-feel as possible, with the understanding that some look-and-feel aspects may be altered due to incomplete capture or variable browser interpretation. The Archive relies upon the reputation of the University of California, Los Angeles as a sufficient measure of the collection’s authenticity.
Section 6. Maintenance and Weeding

A. Maintenance Activities

Seed lists and capture parameters are updated for every election. Descriptive metadata and rights management are updated only if new information is obtained or if retrospective cataloging is undertaken.

B. Deselection Guidelines

Deselection of web sites or information objects may occur at the request of copyright owners or other rights holders of the content. In this case the content would be removed from the archive.

No weeding of the collection is foreseen.

C. Collection Evaluation

Archive usage statistics (such as search and retrieval logs) and user feedback may be of use in adjusting future collection guidelines.
Section 7. Preservation

A. Technology Obsolescence

Currently, the Archive does not attempt to preserve hardware or software as a strategy for dealing with technological obsolescence. It is envisioned that as digital preservation methods and systems evolve and materialize that emulation and/or on-access migration tools will allow users to reliably and accurately access and interact with historically valuable web content from the past. To this end data is migrated on ingest into a static html format, mimicking the content and functionality from the user’s viewpoint, but removing the original database-driven data structure which uses server-side technologies such as PHP, Active Server or Cold Fusion.

B. Preservation Metadata

The WAS tool should collect preservation metadata that conforms to the categories and elements recommended by the Open Archival Information System (OAIS) reference model. This includes basic provenance, reference and fixity metadata. The Archive does not record owner/controller/creator information due to the nature of campaign material so will not require context metadata, but does recognize the value of this information for other types of web collections.

The UCLA Online Campaign Literature Archive recommends that the WAS tool create and associate detailed digital object representation metadata regarding the digital formats contained in the captured web sites on ingest. This preservation metadata should conform to the standards recommended by the Global Digital Format Registry and should include but not be limited to MIME type, hardware/software environment and relationship information.

The preservation metadata elements that will be collected are specified in the Web Archiving Service Agreement listed in Appendix A.
Appendix A.  Web Archiving Service Agreement

The CDL Web Archiving Service Agreement for the Web-at-Risk project will be available from the CDL in 2006 and should include the preservation metadata elements that will be collected by the WAS.
Appendix B. Seed List

The Seed List will be generated using the WAS when the functionality is available.
Appendix C. Metadata Elements

**DC_Title:**
- Web sites usually have an explicitly defined title provided in the `<title>` tag of the header, and this is normally used as the DC_Title. In some cases, however, the `<title>` tag is missing, empty, or contains non-descriptive information left over from a generic template (such as "Home Page"). When this happens, a title is derived from the content of the site's home page.

Definition: A name given to the resource.

**DC_Description**
Date of capture and URL of original web site
Definition: An account of the content of the resource.

**DC_Description.adminnote**
Name of the software or method used to capture the website, plus any special modifications made to the files.

**DC_Subject:** Taken from existing term lists used to assign subjects to the print collection  *See definitions of subject facets listed in Appendix D.*
Definition: A topic of the content of the resource.

**DC_Date:** Date of publication Record 4 digit year only
- Publication dates of campaign material can be problematic, so in this database DC_Date is defined as the year of the election.
Definition: A date of an event in the lifecycle of the resource.

**DC_Type:** Web Site
Definition: The nature or genre of the content of the resource.

**DC_Format:** text/html
Repeatable if multiple file types within the one object.
Definition: The physical or digital manifestation of the resource.

**DC_Identifier:** DivID unique identifier from Campaign Literature database
Definition: An unambiguous reference to the resource within a given context.

**DC_Language**
Repeatable
Definition: A language of the intellectual content of the resource.

**DC_Coverage.temporal**
Date of election
Use International Standard Date and Time Notation ISO 8601 [W3CDTF]
- This will be the year, month, day of the election e.g. 2002-03-04 vs. the date of publication being recorded in DC_date which typically is only the year.

Definition: The extent or scope of the content of the resource

**DC_Rights:** Copyright Statement
Definition: Information about rights held in and over the resource.
Appendix D. Definitions of Subject Facets

Year
- Required.
- Defines the year of the election which the material describes.

Jurisdiction
- Required.
- Format: commonly used names, no abbreviations.
- Defines the government or independent special district to which the Office or Measure in the item’s content belongs. Examples: United States, California, Los Angeles County, City of Los Angeles, Los Angeles Unified School District.
- Note: The Jurisdiction is distinct from the entity organizing the election
- Dependent special districts, which operate under the control of a county board of supervisors or a city council, are defined as part of the jurisdiction of the controlling county or city.
- If an item has content about Offices and Measures in multiple jurisdictions, then all applicable subject headings are applied.

Candidate
- Required for any item with significant content about individual candidates or which is clearly associated with one or more individual candidates’ campaigns.
- Format: names in the format used on ballots. Examples: George W. Bush, Al Gore, Charles “Chuck” Pineda, Jr. When a candidates name appears differently on different ballots, use the most complete version of the name: middle name/initial included, middle name spelled out, additional nicknames added.
- Defines the candidate(s) which the content explicitly endorses.
- If the content does not explicitly endorse a candidate, then use the candidate which the content is primarily about.
- If two candidates are endorsed by a single item, then each is listed as an individual subject. If three or more candidates are endorsed by a single item, then use the subject heading Miscellaneous Candidates.
- Authoritative versions of candidate names in state elections are determined by referring to the Certified List of Candidates, Statement of Vote, or a copy of a ballot. Authoritative versions of candidate names for municipal elections are determined by official documents where available.

Party
- Required for any item with significant content about individual candidates running for partisan office, otherwise optional.
- Format: commonly used name. Examples: Democratic Party, Green Party.
- Defines the party of the Candidate subject(s) defined above if running for partisan office. If running for non-partisan office, the candidate’s party affiliation is only given as a subject heading if it is prominently identified in the item content.
- May also define the party which sponsored the item, especially when an item is primarily about a party’s overall activities rather than that of individual candidates.
- If an item has content about multiple parties, then all applicable subject headings are applied. Special note: At certain times in California history, candidates were allowed to run under multiple party affiliations.
- Authoritative information on candidate party affiliations in state elections for partisan office is determined by referring to the Certified List of Candidates, Statement of Vote, or a copy of a ballot.
Office
- Required for any item with significant content about individual candidates.
- Format: commonly used terminology.
- Defines the office or offices for which a Candidate subject (defined above) is running.
- Offices within collegiate bodies are defined by the name of the body, not the individual office. Examples: State Assembly, City Council, Board of Education.
- President is used even though the actual votes are cast for electors.
- When an office is described by a term which may be used at multiple jurisdictional levels, it is usually prefixed by the level of government to which it belongs. Examples: U.S. Senate, State Senate, County Assessor.
- If two candidates (with separate offices) are endorsed by a single item, then each office is listed as an individual subject. If three or more candidates (with separate offices) are endorsed by a single item, then use the subject heading Miscellaneous Offices.

District
- Required for any item with significant content about individual candidates running for membership in a collegiate body whose members represent geographic districts.
- Defines the district in which a Candidate subject (defined above) is running.
- The district of a U.S. Senator is the name of the state. Note: District names for the U.S. House of Representatives are assumed to be California unless otherwise noted. Example: Arizona 2nd Congressional District.
- If two districts are mentioned in the item, then each district is listed as an individual subject. If three or more districts are covered, then this subject is not used.
- District information for state elections is derived from the Certified List of Candidates, Statement of Vote, or a copy of a ballot. District information for municipal elections is derived from official documents where available.

Measure
- Required for any item with significant content about individual ballot measures.
- Format: names in the format used on ballots.
- If two measures are addressed by a single item, then each is listed as an individual subject. If three or more measures are addressed by a single item, then use the subject heading Miscellaneous Measures.
- Authoritative versions of ballot measure names in state elections are determined by referring to the California Voter Information Guide, Statement of Vote, or a copy of a ballot. Authoritative versions of ballot measures names for municipal elections are determined by official documents where available.

Topic
- Format: use Library of Congress Subject Headings (LCSH) with minimal subdivisions.
- Defines a general topic or issue on which the item has significant content.
Appendix F. Collection Plan Example: CyberCemetery at UNT

**Note:** The following thirty-six pages contain the collection plan for the CyberCemetery at University of North Texas. It is provided as an example of a web collection plan using these guidelines and the associated template.
Collection Plan

for

the CyberCemetery

8.21.2006

Prepared by:

Valerie Glenn
Starr Hoffman
Contents

Section 1. Mission & Scope ................................................................. 3
Section 2. Selection ........................................................................ 5
Section 3. Acquisition ................................................................. 6
Section 4. Descriptive Metadata .................................................. 7
Section 5. Presentation and Access ........................................... 8
Section 6. Maintenance and Weeding ....................................... 10
Section 7. Preservation ................................................................. 11
Appendix A. Memorandum of Understanding .......................... 12
Appendix B. Web Archiving Service Agreement ....................... 29
Appendix C. Seed List ................................................................. 30
Appendix D. UNT Metadata Schema (Descriptive Metadata) ........ 33
Appendix E. UNT Metadata Schema (Preservation Metadata) ........ 35
Section 1. Mission & Scope

A. Mission Statement

The University of North Texas Libraries Mission

The mission of the University of North Texas Libraries is to acquire, preserve, provide access to, and disseminate recorded knowledge in all its forms. Access will be provided increasingly through electronic networks and consortial arrangements. The Libraries, through traditional methods and through digital information resources, provide bibliographic, reference, and instructional support to assist the university's programs of teaching, research, scholarly and creative production, and public service.

Mission as a Federal Depository Library

Federal depository libraries must maintain information products disseminated by the Federal Depository Library Program as part of their existing collections and are responsible for providing permanent public access to this material, including digitally scanned and "born digital" documents.

B. User Groups

Because the CyberCemetery is publicly accessible online, it serves a broad user group that encompasses the students, faculty, and staff of the University of North Texas, but also includes researchers, government employees, and librarians. Because the user group is so broad and the CyberCemetery covers so many different subjects, user information needs are many and varied. Two common information needs are examples of past policies and information about specific topics.

C. Collection Subject, Theme, or Event

The CyberCemetery is a collection of government websites that have ceased operation (usually websites of defunct government agencies and commissions that have issued a final report). This collection features a variety of topics indicative of the broad nature of government information. In particular, this collection features websites that cover topics supporting the university's curriculum and particular program strengths.

This collection supports the mission of the UNT Libraries by acquiring, preserving, and providing access to recorded knowledge in the form of government websites.

This collection supports the mission of the Federal Depository Library Program by providing permanent public access to government information in the form of websites.
D. Curator(s)

Starr Hoffman
Librarian for Digital Collections
University of North Texas Libraries
Government Documents
P.O. Box 305190
Denton, TX 76203-5190

SHoffman@library.unt.edu
940.565.4150
Fax: 940.565.2599

Suzanne Ebanues
Planning and Development Specialist
Office of Planning and Development, Library Services
U.S. Government Printing Office
732 North Capitol St., NW
Mail Stop: IDLP
Washington, DC 20401

sebanues@gpo.gov
202.512.2010, x-33269
Fax: 202.512.2300
Section 2.  Selection

A.  Seed List

Suitable websites must meet the following criteria:

- created by a government agency or commission
- be at-risk (i.e., suspected of being eminently removed or changed)
  - defunct (or soon-to-be) government agency
  - commission has completed final report
  - website with material that frequently changes or has been/will be edited
    (previous information not saved)
  - website which has been untended for some time and suspected to be no
    longer active (may be decommissioned at any time)
- support the UNT curriculum and/or the needs of the community
  - example: topics relevant to UNT programs, such as geology, history, criminal
    justice, public administration, etc.
  - example: topics relevant to community members, such as tax reform,
    bankruptcy, etc.

The materials included in this collection are listed in Appendix C: Seed List (8.7.06 report). In the future, this list should be generated from the Web Archiving Service (WAS) and should be updated as new websites are captured.

B.  Capture Scope

In general, the capture scope for websites in the CyberCemetery will be the entire site (i.e., materials or content originating from the host specified in the seed URL). However, some sites may also necessitate capturing one level (link-hop) outside of the domain. Most of the websites will be hosted on .gov domains; however, there are exceptions to this rule (e.g., some government websites or specific media on some of these sites may be hosted by commercial providers). Because of these and other exceptions, each URL on the seed list must be assessed individually.

C.  Rights Management

In general, websites in this collection are copyright-exempt because they are produced by the federal government.

All materials in this collection are covered by the legal agreement between the Government Printing Office and the University of North Texas. For more information on this agreement, see Appendix A: Memorandum of Understanding.
Section 3. Acquisition

A. Frequency of Capture

Websites for this collection will be captured once. They will be captured when a government agency becomes inactive, when an expiration date has passed, and/or after a final report is issued. They will be recaptured only if information is determined to be incomplete.

Note: 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.

B. Capture Scope

[This section is not relevant to the collection.]

C. Material Types & Formats

There are no specific material types or formats to be excluded from this collection. In general, complete websites are to be captured, encompassing files of all types included on their domain. However, if a website is particularly large, the curator should discuss its capture with ITS (Information Technology Services) before capturing it.

D. Interactive & Dynamic Content

The legal agreement regarding this collection specifies that captured websites must be represented as exactly as possible within the archive. Exceptions that necessitate some minor changes to the HTML are:

- Add this text to comment forms/contact pages: "Please do not send comments. [Enter date of final commission report, if applicable.] We archived this site in [date]. If you experience any technical problems using this site, please let us know [UNT email link here]." (If it is a website for a commission, insert date that final report was submitted.)
- Email links: preserve the display of email addresses, but disable links
- Submit buttons: preserve their display, but disable
- Also, add the word "Archive" (in 8 pt. Times New Roman font and center-justified) to every page of each archived website.
Section 4. Descriptive Metadata

A. Level of description
Website level description provides ample subject access to the information.

B. Metadata elements
Our local metadata schema is listed in Appendix D: UNT Metadata Schema (Descriptive Metadata).

C. Controlled vocabularies
The CyberCemetery collection will use the Legislative Indexing Vocabulary (LIV) for subject terms. The Legislative Indexing Vocabulary is a thesaurus of over 5,000 terms created by the Congressional Research Service to describe legislative and public policy literature.
Section 5.  Presentation and Access

A.  Discovery

The web interface to the CyberCemetery collection is provided by the University of North Texas Libraries’ Government Documents Department <http://govinfo.library.unt.edu/>. The primary method of discovery is the search screen, which provides full-text searching. That search can be narrowed to search a few specific websites, such as the National Commission on Terrorist Attacks Upon the United States. The descriptive metadata includes subject terms that will be used to provide topical access (see Section 4). Users also have the option of browsing the collection by website name, branch of government, or date of expiration.

Users are likely to use both material titles and subject(s) to evaluate the results returned from a search of the collection.

B.  Access

This web collection is visible. There are no time-dependent release restrictions associated with the materials in this web collection. We try not to restrict the accessibility of captured web content; however, there may be exceptions. In some cases, a physical residence address appearing on an archived website may be removed at the resident’s request (this is done to protect the privacy of government employees and officials, as well as citizens who have publicly commented on agency proposals). In the case of such exceptions, explanatory text will be included on the page to explain the reason for the removal.

C.  Look-and-Feel

It is important to the users of this collection to retain the “look-and-feel” of the source websites. Out of necessity for identification as archived websites, clarifying text notices will be inserted on comment pages and/or content forms, and a text notice reading “archive” will be inserted at the top of each homepage.

D.  Dynamic Content

It is desirable that users be allowed to access hyperlinked materials and websites that are not located within the web archive. Users should be alerted to the fact that they are leaving the archive by an informative message, and should be provided both a link to the external website and a link back to the archived page.

Email links and form submit buttons should be displayed, but disabled. Informational text should inform the user that this website has been archived, and give contact information for the archiving institution, in case of technical difficulties with the site.

E.  Multiple Types/Formats

All types and formats of materials in the captured content of this collection will be discoverable and accessible by all users regardless of their location, assuming that the users have access to the required software. The content will always be accessible at the University of North Texas Library, which will retain the required software for discovery and access.
F. Authenticity

Authenticity will be established by the CyberCemetery interface through which users will access the websites, and by the http://govinfo.library.unt.edu/ URL stem which all these sites will share. The reputation of the University of North Texas and its legal agreement with the Government Printing Office and the National Archives and Records Administration will be a sufficient measure of material authenticity for the collection's users. This collection will not use digital stamps or watermarks as indicators of authenticity, because our legal agreement requires that the code be altered as little as possible, in order to best preserve the original “look-and-feel” of the websites. The only changes that will be made to the code are those which are necessary to indicate that the website is an archived page and those which disable actions such as forms and email links (for more on this, see Section 3, part D).
Section 6. Maintenance and Weeding

A. Maintenance Activities
Periodically, the curator will identify new websites to be added to the collection.

B. Deselection Guidelines
There are no anticipated situations under which deselection will be required; deselection is expressly denied in the legal agreement between the University of North Texas, the Government Printing Office, and the National Archives and Records Administration.

Some information may be removed by specific request, but entire websites will not be removed (see Section 5-B).

C. Collection Evaluation
[This section is not applicable to this collection.]
Section 7. Preservation

A. Technology Obsolescence

Preserving original hardware and/or software may not effectively preserve the collection, because the majority of users access the collection off-campus, using their own hardware and software. Emulation may be a better method; we are currently considering our options. For more on the terms of the preservation service, see Appendix B: Web Archiving Service Agreement.

B. Preservation Metadata

See Appendix E: UNT Metadata Schema (Preservation Metadata).
Appendix A. Memorandum of Understanding

National Archives and Records Administration
8601 Adelphi Road
College Park, Maryland 20740-6001

May 4, 2006

Judith C. Russell
Managing Director, Information Dissemination
Superintendent of Public Documents
732 North Capitol Street, NW
Washington, DC 20401

Dear Ms. Russell:

Enclosed you will find 3 addendums signed by the Archivist of the United States amending the August 12, 2003, Memorandum of Understanding between the Government Printing Office and the National Archives and Records Administration. Per our discussion, please obtain signature within your office and forward on to Cathy Hartman at the University of Texas (UNT) to obtain signature from UNT’s Vice President for Finance and Business Affairs. I will send Cathy an e-mail requesting that she send each of us one of the signed addendums for our official records.

Thank you for agreeing to forward the addendums on to UNT. We look forward to this three party partnership and the mutual benefits and efficiencies that we hope to achieve.

If I can be of further assistance, please do not hesitate to contact me at (301) 837-0993.

Sincerely,

[Signature]

Carla Patterson
Project Manager, Policy and Planning Staff

NARA’s web site is http://www.archives.gov
ADDENDUM TO THE
August 12, 2003, MEMORANDUM OF UNDERSTANDING
BETWEEN THE GOVERNMENT PRINTING OFFICE AND THE
NATIONAL ARCHIVES AND RECORDS ADMINISTRATION

1. In accordance with the terms of the August 12, 2003, Memorandum of Understanding between the U.S. Government Printing Office (GPO) and the National Archives and Records Administration (NARA), GPO and NARA hereby agree to modify that agreement to incorporate, as Appendix B, the November 3, 1999, Memorandum of Understanding between the U.S. Government Printing Office and the University of North Texas Libraries (UNT Libraries), as amended on March 17, 2005.

2. By this addendum, the UNT Libraries shall be an affiliated archives of NARA, subject to the terms of the August 12, 2003, GPO/NARA MOU.


Effective Date
This addendum shall be effective as of the date of the last signature.

Signatures

BRUCE R. JAMES
PUBLIC PRINTER OF THE UNITED STATES

ALLEN WEINSTEIN
ARCHIVIST OF THE UNITED STATES

PHIL DEIBEL
VICE PRESIDENT FOR FINANCE AND BUSINESS AFFAIRS
UNIVERSITY OF NORTH TEXAS
Memorandum of Understanding
Between the
United States Government Printing Office
and the University of North Texas Libraries

Amendment to the Memorandum of Understanding

The U.S. Government Printing Office (GPO) and the University of North Texas Libraries (UNT Libraries) agree to revise the existing terms of agreement under the aegis of the GPO/UNT Libraries Memorandum of Understanding dated October 1999, and reaffirmed in 2004 with an extension to 2009. The specific section amendment is defined below.

PURPOSE

The purpose of this Amendment is to modify the existing partnership terms of agreement, which provides for permanent public access of the electronic Government information products in the public domain from Government agencies, boards, commissions, etc., which have completed business and ceased to operate and have been identified by the FDLP for inclusion in the FDLP/EC. With this Amendment, data migration for Government information files shall become admissible; however the original informational content will remain available to secure the integrity and authenticity of the files.

AMENDMENT

Section 2.a.(6) under Terms of Agreement is replaced by:

Provide measures to secure the authenticity of the specified electronic Government information products, such as manual review of transferred files, and to secure the integrity of the files, which will assure that the original informational content is made available without modification. GPO and UNT will work jointly to develop a guideline with specifications for data migration. These specifications will serve as a guideline for data migration of UNT Libraries CyberCemetery resources.

EFFECTIVE DATE

This amendment is effective as of the date of signature by all representatives indicated below.

AGREEMENT CONTACTS

UNT

Cathy N. Hartman
Head, Digital Projects Dept.
Fellow, Texas Center for Digital Knowledge
University of North Texas Libraries
P.O. Box 305190
Denton, TX 76203-5190
cartman@library.unt.edu
Phone: 940-565-3269

Gretchen Schlag
Director, Program Planning and Coordination Service
Information Dissemination (Superintendent of Documents)
U.S. Government Printing Office
732 North Capitol St, NW
Washington, DC 20401
(p) 202-512-

ACCEPTANCE BY

B. Donald Grose, Dean of Libraries.
University of North Texas Libraries

Judith C. Russell, Managing Director
Information Dissemination (Superintendent of Documents)
U.S. Government Printing Office

Date: 3/23/05
Date: 3/17/2005
MEMORANDUM OF UNDERSTANDING (MOU)
BETWEEN THE GOVERNMENT PRINTING OFFICE AND THE NATIONAL
ARCHIVES AND RECORDS ADMINISTRATION

Purpose
The purpose of this agreement is to establish an affiliated relationship between the U.S.
Government Printing Office (GPO) with preservation and access responsibilities codified in 44
U.S.C. Chapter 41, and the National Archives and Records Administration (NARA) which operates
under the authority of 44 U.S.C. Chapters 21, 29, 31, and 33. To implement NARA Directive 1501,
Custody of Federal Records of Archival Value, dated February 28, 2003, NARA and GPO agree to
the following conditions for the continued deposit of the electronic government records of the
this agreement will eliminate potentially duplicative preservation activities at both GPO and NARA
for records within the scope of this MOU.

Scope
This agreement encompasses the content (hereinafter referred to as “the records”) on GPO Access,
GPO’s publicly accessible Internet site. As authorized in 44 U.S.C. 4101, this content includes the
online versions of the Congressional Record (beginning with volume 140, 1994), the Federal
Register (beginning with volume 59, 1994), the Code of Federal Regulations (beginning with
1996), and other appropriate publications distributed by the Superintendent of Documents, and
subsequent records as GPO adds them to GPO Access.

All records subject to this agreement shall be additionally defined in Records Schedules to be
established between NARA and the GPO. The records are designated as part of NARA Record
MOU provides for the deposit of the records outside of the physical custody of NARA.

Actions

1. The records will be transferred into the legal custody of NARA and will be legally part of
   the National Archives of the United States.

2. GPO will retain physical custody of the complete set of the records for public access and
   preservation purposes. GPO agrees to provide for the permanent preservation of this set of
   the records.

3. GPO will provide public access to the records and the information in the records including
   (if necessary):
   a. special expertise needed to interpret the records and the recordkeeping system.
   b. special equipment or specially designed facilities needed to provide ready access to
      records.
   c. records metadata, user or systems documentation or other essential research resources
      necessary to provide essential context or usability for the records.


d. access through GPO’s geographically-dispersed network of Federal depository libraries which provide specialized assistance and services to local users.

4. NARA will refer researchers of the records to GPO and the GPO web site.

5. In accordance with NARA guidance in effect on the signing date of this MOU, GPO will manage the records to ensure their long-term preservation by storing and managing them under preservation conditions that are at least equal to those that would be provided by NARA. See Appendix.

6. In accordance with NARA guidance in effect on the signing date of this MOU, GPO agrees to follow NARA reference, arrangement, description, preservation, and security guidance. See Appendix.

7. GPO will permit NARA to examine the records and review how they are being managed, preserved, and made available for research.

8. GPO will report annually on management activities for these records and quarterly on the level of public use.

9. GPO will notify NARA as early as possible of any risk that GPO might not be able to continue to preserve or provide electronic public access to the records.

10. If for any reason, GPO is unable to maintain, preserve, and provide electronic public access to the records, GPO will transfer the access and preservation sets of records to NARA.

11. NARA may, after providing notification and an opportunity to correct problems, take physical custody of the access and preservation sets of the records if they are not maintained in accordance with this MOU.

12. If both the access and preservation copies of the records are transferred to NARA, NARA shall continue permanent public access to the records comparable to the online access provided by GPO Access.

13. When GPO, at the request of a publishing agency, terminates access to a record subject to this MOU, GPO will transfer the preservation copy of that record to NARA.

14. When there is a major change in the underlying software or user interface to GPO Access GPO will transfer the preservation copy of the non-current version to NARA.
Effective Date, Modifications, Termination

1. This agreement shall be effective upon signature, and shall remain in effect until terminated.

2. This agreement may be modified or terminated by agreement of both parties.

3. Material changes in the NARA requirements in the Appendix are considered modifications of this MOU, and require the consent of both parties.

Signatures

BRUCE R. JAMES
PUBLIC PRINTER OF THE UNITED STATES
Date 8/12/02

JOHN CARLIN
ARCHIVIST OF THE UNITED STATES
Date 8-12-03
Appendix
NARA Requirements

Governing Directives:
36 CFR 1234.30 and 1234.32 (Electronic Records Management)
http://www.archives.gov/about_us/regulations/part_1234.html

NARA. NARA 1501, Custody of Federal Records of Archival Value. NARA: Washington, DC.


http://staffonly.nara.gov/directives/nara1571.html

NARA. NARA 1701, Loans of Holdings in NARA’s Physical and Legal Custody. NARA: Washington, DC:
http://staffonly.nara.gov/directives/nara1701.html

NARA. NARA 1702, Transporting Holdings in NARA’s Physical and Legal Custody. NARA: Washington, DC,
A. Preservation and General Collections Management:
1. Ensure security of electronic records from theft, vandalism, terrorism, damage, and loss through a 24 hour a day and 365 day a year systematic security and monitoring program as specified in NARA 1571.21.
2. Meet or exceed NARA standards for archival storage environmental conditions specified in Table 1.
3. Monitor and record environmental conditions in all storage spaces in accordance with NARA 1571.9d; consult with building engineers to improve conditions when environmental set points are exceeded.
4. Obtain, store, and use appropriate preservation, housing, and storage supplies and equipment that meet NARA standards and specifications.
5. Arrange for duplication/migration/refreshing of any at risk materials.
6. Obtain and manage offsite storage as specified in Table 1.

B. Operational Procedures to Ensure Data Longevity for Electronic Records:
1. Perform weekly backups of the entire system (ie. data, server operating system, and configurations) that holds these records and daily incremental backups. Retain two months of weekly backups before the medium is reused.
2. Store backups in an offsite location on tape, cartridges, or other NARA approved storage medium.
3. Store back-up and master tapes, cartridges, and other magnetic media upright in NARA approved containers in stable environments that are dust-free and meet the standards specified in Table 1.
4. Monitor the environment (e.g., temperature and relative humidity) using NARA-approved monitoring equipment and report the results to NARA quarterly.
5. Read a statistical sample of all permanent data sets stored on magnetic tape annually to detect any loss of data in accordance with 36 CFR 1234.30(g)(2).
6. Copy records onto new or recertified tapes at least once every decade or more frequently when problems are noted to prevent the physical loss of data or technological obsolescence of the medium.
7. Use magnetic tape (e.g., reels, cartridges) as the only approved storage media for the GPO public access set of the records.

C. Arrangement and Description:
1. The records constitute one series that has already been described in Archival Research Catalog (ARC). That series description will need to be updated once to reflect the existence of electronic records and then updated annually to reflect changes in dates of holdings. NARA will provide GPO with Archival Research Catalog (ARC) worksheets and instructions for completing them.

D. Reference and Access:
Provide equal access to all public research customers.
Temperature and Relative Humidity Standards for Archival Records

Table 1. Records Storage Areas

<table>
<thead>
<tr>
<th>Records medium</th>
<th>Dry Bulb Maximum Temperature</th>
<th>Relative Humidity Set Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnetic/electronic media including computer tapes and disks, video tapes, audio tapes, optical disks</td>
<td>65°F*</td>
<td>35%± 5%</td>
</tr>
</tbody>
</table>

*Cooler temperature set points and within the specified range drier relative humidity set points should be used for these media whenever possible in order to improve the preservation of the records. Magnetic/electronic media should not be stored at a temperature lower than 46°F.
MEMORANDUM OF UNDERSTANDING BETWEEN
THE U.S. GOVERNMENT PRINTING OFFICE
AND
THE UNIVERSITY OF NORTH TEXAS LIBRARIES

Purpose and Scope

This memorandum of understanding (MOU or agreement) establishes the respective responsibilities of the U.S. Government Printing Office (GPO) and the University of North Texas Libraries (UNT) in a partnership arrangement, under the auspices of the Federal Depository Library Program (FDLP), designed to insure permanent storage and accessibility to certain electronic Government information products considered part of the FDLP Electronic Collection (FDLP/EC).

This MOU applies to electronic products in the public domain from Government agencies, boards, commissions, etc., which have completed business and ceased to operate and have been identified by the FDLP for inclusion in the FDLP/EC. These products are considered to be Government publications, as defined in 44 U.S.C. 1902.

Specific sites, products, and electronic publications included in this agreement are listed in Appendix "A".

Background

The partnership developed herein places electronic Government information products from agencies no longer operating, under the auspices of the FDLP, assuring permanent public access through the FDLP/EC.

Terms of Agreement


   a. The United States Government Printing Office, as administrator of the FDLP, has final responsibility for and custody over all Federal Government information products held or made accessible by FDLP libraries or partner institutions.

   b. The partners recognize that Federal Government information contained in agency sites is in the public domain and may be made available for access by any party, including other FDLP partners providing access under the auspices of the FDLP. This agreement does not grant the UNT Libraries exclusive rights to provide access to, or ownership of, any Government information products,
c. All software developed to provide access to the publications included in this MOU is considered to be an integral part of the FDLP Electronic Collection. All references to publications included in the MOU are assumed to include such software.

2. Specific Partner Requirements.

a. The UNT Libraries shall:

(1) Provide no fee, unrestricted public on-line access to the electronic Government information products specified in Appendix A of this agreement.

(2) Impose no restrictions on re-dissemination of the specified electronic Government information products.

(3) Provide no-fee technical support for FDLP access to the specified electronic Government information products. The UNT Libraries shall have the authority to levy fees when fulfilling requests for specialized processing, such as for copying files to tangible media. Any such fees shall be reasonable and set so as to recover only the costs of fulfilling the requests.

(4) Assure that the server(s) on which the specified electronic Government information products are located will support at least ten (10) simultaneous remote connections.

(5) Assure that under normal operating conditions the server(s) on which the specified electronic Government information products are located will be available for remote public access at least ninety five percent (95%) of the time, excluding scheduled down time.

(6) Provide measures to secure the authenticity of the specified electronic Government information products, such as manual review of transferred files, and to secure the integrity of the files, which will assure that the informational content is not modified or reconstituted.

(7) Assure that a mechanism is in place which provides a backup for software and data files associated with the specified electronic Government information products, such as routine on-site backup to magnetic or optical storage medium in a separate physical facility.

(8) Assure that Internet addresses for files within this agreement are persistent, regardless of the relocation or renaming of files.
(9.) Consult with Superintendent of Documents staff on any plans to migrate, reconstitute, or otherwise significantly modify the structure of the information content of or access software related to the specified electronic Government information products, other than routine refreshing of the resource. Any modification to or migration should take into consideration industry format standards and must not undermine or limit access to the product.

(10.) Compile usage statistics and other performance measures upon the request of the GPO, and make those statistics available to the GPO as FDLP administrator.

(11.) Notify GPO in writing in the event the UNT Libraries can no longer perform its responsibilities under the terms of this memorandum at least ninety (90) days before the performance of responsibilities is terminated, and continue, for a period not to exceed one hundred and eighty (180) days from the date the letter was sent, to make electronic Government information products specified in Appendix A available until GPO has arranged for an alternative server location.

(12.) Upon request from GPO, provide GPO a functioning copy of current and historical software and data files that constitute the specified electronic Government information products, to include documentation regarding the products, particularly if the UNT Libraries is no longer able to support access to the product as per the terms of this memorandum of understanding.

(13.) Acknowledge on the UNT Libraries' World Wide Web pages associated with the specified electronic Government information products that the UNT Libraries is working in partnership with the FDLP to provide permanent public access as an FDLP partner.

(14.) Work with GPO to assure that access to all Government information products in the FDLP comply with requirements of the Americans With Disabilities Act (Public Law 101-336). GPO considers basic HTML as ADA-compliant by default: i.e., it is textual and readable by any browser including large-font, audio-only, etc. When designing web pages, standard ADA needs must be considered (for example, including ALT tags when necessary for images).

(15.) Review the list of electronic Government information products in Appendix A at least semi-annually in cooperation with GPO, and add products as appropriate.
b. GPO shall:

(1.) Function as the coordinator for all remotely accessible electronic Government information products under the auspices of the FDLP providing administrative coordination and oversight of electronic Government information products distributed among FDLP partner institutions.

(2.) Recognize the UNT Libraries as the official FDLP site for the electronic Government information products specified in Appendix A.

(3.) Provide Internet pointers and other locator mechanisms on the GPO Access World Wide Web site to identify electronic Government information products on UNT servers.

(4.) Make publicly available through the FDLP a functioning copy of all current and historical software and data files associated with the specified electronic Government information products, to include documentation regarding the products, in the event that the UNT Libraries can no longer meet the terms of this memorandum of understanding.

(5.) Provide general user support through the FDLP for access to electronic Government information products under the auspices of the FDLP specified in Appendix A.

(6.) Work with partner FDLP libraries and originating agencies to assure that access to all Government information products in the FDLP complies with requirements of the Americans With Disabilities Act (Public Law 101-336).

(7.) Provide a notice on the GPO Access World Wide Web site acknowledging the partnership between the GPO and the UNT Libraries to provide permanent public access through the FDLP to electronic Government information products specified in Appendix A.

(8.) Identify and recommend for inclusion on UNT servers sites from Government agencies, commissions, and other entities which have completed business and ceased operations, and to review Appendix A at least semi-annually in cooperation with UNT Libraries, adding electronic Government information products as appropriate.
3. Financial Terms and Payment: The UNT Libraries assumes the financial obligation to provide for the maintenance, security, and public accessibility to the specified electronic Government information products as outlined above.

4. Contacts:

GPO:
George D. Barnum
Electronic Collection Manager
gbarnum@gpo.gov
202-512-1114

UNT Libraries:
Cathy Nelson Hartman
Documents Librarian: Electronic Resources Coordinator
University of North Texas Libraries
chartman@library.unt.edu
940-565-3269

5. Effective Date/Duration/Amendments

This agreement is effective as of the date of signature by all the authorized representatives indicated below and shall last for two years thereafter. The list of electronic Government information products in appendix A shall be reviewed and updated at least semi-annually by both parties. The MOU may be extended or amended by mutual agreement of the parties. Any party may withdraw from this agreement upon one hundred and eighty (180) days written notice to the other parties.
Acceptance by:

Francis J. Buckley Jr.  
Superintendent of Documents  
U.S. Government Printing Office  
Date: 10/15/99

B. Donald Grose  
Director, University of North Texas Libraries  
University of North Texas  
Date: 11/1/91

Phil Deibel  
Vice President for Finance and Business Affairs  
University of North Texas  
Date: 11/3/92
Appendix A

List of Electronic Government Information Products Included in the Memorandum of Understanding

August 1, 1999

Title: Advisory Commission on Intergovernmental Relations
URL or PURL: http://www.library.unt.edu/gpo/acir/acir.html
Added: 9/28/97

Title: National Civil Aviation Review Commission
URL or PURL: http://www.library.unt.edu/gpo/ncarc/ncarc.html
Added: 3/8/99

Title: Commission on Structural Alternatives for the Federal Courts of Appeals
URL or PURL: http://app.comm.uscourts.gov
Added: 8/1/99
Appendix B. Web Archiving Service Agreement

To be included when available.
Appendix C. Seed List

Seed List as of 8.7.2006

Access America (2001)
   http://govinfo.library.unt.edu/accessamerica/index.html

Advisory Commission on Electronic Commerce (2000)
   http://govinfo.library.unt.edu/ecommerce/index.htm

Advisory Commission on Intergovernmental Relations (1996)
   http://www.library.unt.edu/gpo/acir/acir.htm

Advisory Commission on Online Access and Security (2000)
   http://govinfo.library.unt.edu/acoas/index.htm

Advisory Committee on Public Interest Obligations of Digital Television Broadcasters (1998)
   http://govinfo.library.unt.edu/piac/pubint.htm

Amtrak Reform Council (2002)
   http://govinfo.library.unt.edu/arc/

Columbia Accident Investigation Board (2003)
   http://govinfo.library.unt.edu/caib/default.html

Census Monitoring Board–Congressional Members (2001)
   http://govinfo.library.unt.edu/cmb/cmbc/index.htm

Census Monitoring Board–Presidential Members (2000)
   http://govinfo.library.unt.edu/cmb/cmbp/index.htm

Coalition Provisional Authority (2004)
   http://govinfo.library.unt.edu/cpa-iraq/index.html

Commission on Affordable Housing and Health Facility Needs for Seniors in the 21st Century (2002)
   http://govinfo.library.unt.edu/seniorscommission/

Commission on Structural Alternatives for the Federal Court of Appeals (1999)
   http://www.library.unt.edu/gpo/csaftca/app_comm_uscourts_gov.html

   http://govinfo.library.unt.edu/osbc/

   http://govinfo.library.unt.edu/wmd/

Commission to Strengthen Social Security (2001)
   http://govinfo.library.unt.edu/csss/index.htm
http://govinfo.library.unt.edu/copacommission/

http://www.library.unt.edu/gpo/icpac/icpac.htm

Millennial Housing Commission (2002)
http://govinfo.library.unt.edu/mhc/

http://govinfo.library.unt.edu/nbrc/index.html

National Bioethics Advisory Commission (2001)
http://govinfo.library.unt.edu/nbac/nbac.html

National Civil Aviation Review Commission (1997)
http://www.library.unt.edu/gpo/NCARC/index.htm

National Commission on Terrorist Attacks Upon the United States (2004)
http://govinfo.library.unt.edu/911/

http://govinfo.library.unt.edu/ncecic/

http://govinfo.library.unt.edu/drought/default.htm

National Education Goals Panel (NEGP) (2002)
http://govinfo.library.unt.edu/negp/index.htm

National Gambling Impact Study Commission (NGISC) (2001)
http://govinfo.library.unt.edu/ngisc/index.html

National Partnership for Reinventing Government (2001)
http://govinfo.library.unt.edu/npr/default.htm

Office of Consumer Affairs (1999)
http://www.library.unt.edu/gpo/oca/default.htm

Office of Independent Counsel Donald C. Smaltz in re Secretary of Agriculture Alphonso Michael Espy (2001)
http://govinfo.library.unt.edu/oic/

http://www.library.unt.edu/gpo/ota/index.html

Office of Technology Assessment (OTA) (1995)
http://govinfo.library.unt.edu/ota/default.htm

President’s Advisory Commission on Consumer Protection and Quality in the Health Care Industry (1998)
http://govinfo.library.unt.edu/hcquality/

President's Advisory Panel on Federal Tax Reform (2005)  
http://govinfo.library.unt.edu/taxreformpanel/

President's Commission on Improving Economic Opportunity in Communities Dependent on Tobacco Production while Protecting Public Health (2001)  
http://govinfo.library.unt.edu/tobacco/index.htm

President's Commission on Moon, Mars and Beyond (2004)  
http://govinfo.library.unt.edu/moontomars/index.htm

President's Commission on the Celebration of Women in American History (2001)  
http://govinfo.library.unt.edu/whc/whc.htm

http://govinfo.library.unt.edu/mentalhealthcommission/index.htm

Presidential Advisory Commission on Holocaust Assets in the United States (2001)  
http://govinfo.library.unt.edu/pcha/

Smithsonian Institution Science Commission (2003)  
http://govinfo.library.unt.edu/sciencecommission/

Special Oversight Board for Department of Defense Investigations of Gulf War Chemical and Biological Incidents (2000)  
http://govinfo.library.unt.edu/oversight/

http://govinfo.library.unt.edu/oceancommission/index.html

United States Information Agency (USIA) [University of Illinois at Chicago] (1999)  
http://dosfan.lib.uic.edu/usia/index.html

http://govinfo.library.unt.edu/nssg/index.html

http://govinfo.library.unt.edu/tdrc/index.html

http://govinfo.library.unt.edu/whccamp/
### Appendix D. UNT Metadata Schema (Descriptive Metadata)

<table>
<thead>
<tr>
<th>DESCRIPTIVE METADATA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Title</td>
</tr>
<tr>
<td>Parallel Title</td>
<td>Alternate Title</td>
</tr>
<tr>
<td>Added Title</td>
<td>Series Title</td>
</tr>
<tr>
<td>Serial Title</td>
<td>Uniform Title</td>
</tr>
<tr>
<td>Other</td>
<td>Other Title</td>
</tr>
<tr>
<td>Creator</td>
<td>Creator Name</td>
</tr>
<tr>
<td>Creator Name</td>
<td>Creator Role</td>
</tr>
<tr>
<td>Creator Information</td>
<td>Creator Information</td>
</tr>
<tr>
<td>Contributor</td>
<td>Contributor Name</td>
</tr>
<tr>
<td>Contributor Name</td>
<td>Contributor Role</td>
</tr>
<tr>
<td>Contributor Role</td>
<td>Contributor Information</td>
</tr>
<tr>
<td>Publisher</td>
<td>Publisher Name</td>
</tr>
<tr>
<td>Publisher Name</td>
<td>Place of Publication</td>
</tr>
<tr>
<td>Publisher Information</td>
<td>Publisher Information</td>
</tr>
<tr>
<td>Date</td>
<td>Original Resource Creation Date</td>
</tr>
<tr>
<td>Date</td>
<td>Digital Creation Date</td>
</tr>
<tr>
<td>Language</td>
<td>Description</td>
</tr>
<tr>
<td>Description</td>
<td>Content Description</td>
</tr>
<tr>
<td>Description</td>
<td>Physical Description</td>
</tr>
<tr>
<td>Subject and Keywords</td>
<td>Subject and Keywords</td>
</tr>
<tr>
<td>Primary Source</td>
<td>Primary Source</td>
</tr>
<tr>
<td>Coverage</td>
<td>Coverage</td>
</tr>
<tr>
<td>Place Name</td>
<td>Place Name</td>
</tr>
<tr>
<td>Time Period</td>
<td>Time Period</td>
</tr>
<tr>
<td>Date</td>
<td>Date</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Date Range</td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td></td>
</tr>
<tr>
<td>Relation</td>
<td></td>
</tr>
<tr>
<td>Collection</td>
<td></td>
</tr>
<tr>
<td>Institution</td>
<td></td>
</tr>
<tr>
<td>Rights Management</td>
<td></td>
</tr>
<tr>
<td>Resource Type</td>
<td></td>
</tr>
<tr>
<td>Format</td>
<td></td>
</tr>
<tr>
<td>Identifier</td>
<td></td>
</tr>
<tr>
<td>URL</td>
<td></td>
</tr>
<tr>
<td>URN</td>
<td></td>
</tr>
<tr>
<td>DOI</td>
<td></td>
</tr>
<tr>
<td>ISBN</td>
<td></td>
</tr>
<tr>
<td>ISSN</td>
<td></td>
</tr>
<tr>
<td>OCLC No.</td>
<td></td>
</tr>
<tr>
<td>Report No.</td>
<td></td>
</tr>
<tr>
<td>Government Document No.</td>
<td></td>
</tr>
<tr>
<td>Accession or Local Control No.</td>
<td></td>
</tr>
<tr>
<td>UNT Catalog No.</td>
<td></td>
</tr>
<tr>
<td>RISM No.</td>
<td></td>
</tr>
<tr>
<td>Other Identifier</td>
<td></td>
</tr>
<tr>
<td>Note</td>
<td></td>
</tr>
<tr>
<td>Metadata Information</td>
<td></td>
</tr>
<tr>
<td>Metadata Creator</td>
<td></td>
</tr>
<tr>
<td>Date of Creation</td>
<td></td>
</tr>
<tr>
<td>Metadata Modifier</td>
<td></td>
</tr>
<tr>
<td>Date of Modification</td>
<td></td>
</tr>
</tbody>
</table>
Appendix E. UNT Metadata Schema (Preservation Metadata)

<table>
<thead>
<tr>
<th>PRESERVATION METADATA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>File Information</strong></td>
</tr>
<tr>
<td><strong>File Size</strong></td>
</tr>
<tr>
<td><strong>File Name</strong></td>
</tr>
<tr>
<td><strong>Format Name</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Format Version</strong></td>
</tr>
<tr>
<td><strong>File description</strong></td>
</tr>
<tr>
<td><strong>Resolution</strong></td>
</tr>
<tr>
<td><strong>Dimension</strong></td>
</tr>
<tr>
<td><strong>Duration</strong></td>
</tr>
<tr>
<td><strong>Rate</strong></td>
</tr>
<tr>
<td><strong>Tonal-Resolution</strong></td>
</tr>
<tr>
<td><strong>Color</strong></td>
</tr>
<tr>
<td><strong>Compression</strong></td>
</tr>
<tr>
<td><strong>Other File information</strong></td>
</tr>
<tr>
<td><strong>Fixity Information</strong></td>
</tr>
<tr>
<td><strong>Authentication Type</strong></td>
</tr>
<tr>
<td><strong>Authentication Result</strong></td>
</tr>
<tr>
<td><strong>Date</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>First Date</strong></td>
</tr>
<tr>
<td><strong>Last date</strong></td>
</tr>
<tr>
<td><strong>System Information</strong></td>
</tr>
<tr>
<td><strong>Software</strong></td>
</tr>
<tr>
<td><strong>Creation Application Software</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Creation Application Name</strong></td>
</tr>
<tr>
<td><strong>Creation Application Version</strong></td>
</tr>
<tr>
<td><strong>Access Application Software</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Access Application Name</strong></td>
</tr>
<tr>
<td><strong>Access Application Version</strong></td>
</tr>
<tr>
<td><strong>Other Software Information</strong></td>
</tr>
<tr>
<td><strong>Hardware</strong></td>
</tr>
<tr>
<td><strong>Creation Hardware</strong></td>
</tr>
<tr>
<td><strong>Access Hardware</strong></td>
</tr>
<tr>
<td>Other Hardware Information</td>
</tr>
<tr>
<td>----------------------------</td>
</tr>
<tr>
<td>Documentation</td>
</tr>
<tr>
<td>Structural Composition</td>
</tr>
<tr>
<td>Storage Medium</td>
</tr>
<tr>
<td>Access Inhibitors</td>
</tr>
<tr>
<td>Inhibitor Key</td>
</tr>
<tr>
<td>Functionality</td>
</tr>
<tr>
<td>Exception</td>
</tr>
<tr>
<td>Alteration History</td>
</tr>
<tr>
<td>Action Taken</td>
</tr>
<tr>
<td>Date of Alteration</td>
</tr>
<tr>
<td>Modifier</td>
</tr>
<tr>
<td>Other Alteration Information</td>
</tr>
<tr>
<td>Metadata Information</td>
</tr>
<tr>
<td>Metadata Editor/Modifier</td>
</tr>
<tr>
<td>Metadata Creation/Modification Date</td>
</tr>
<tr>
<td>Metadata Modification Action</td>
</tr>
<tr>
<td>Other Metadata Information</td>
</tr>
<tr>
<td>Comments</td>
</tr>
</tbody>
</table>