Web
Archiving
Service

A CDL Service in support of the NDIIPP Web-at-Risk Project

Requirements

July, 2005

DRAFT

Table of Contents

1 Overall Description

1.1 Purpose

For the Web-at-Risk project, the California Digital Library (CDL) and its project partners, with funding from the Library of Congress, will create the Web Archiving Service (WAS). This service enables curators to build, manage, and expose collections of government and political information harvested from the World-Wide Web. In fulfillment of this vision, the tools will be developed incrementally and in close consultation with the curators who will be using them. This project is one of eight grants funded by the National Digital Information Infrastructure Preservation Program (NDIIPP).

Web archiving is the process by which web-published materials are acquired, curated, and preserved. Acquisition is a set of activities centered on content capture, and preceded by organizational deliberation about general collection goals and specific site selection priorities. Acquisition is not considered complete until the captured material has been reviewed for quality assurance and copyright clearance.

Curation is the set of activities concerned with managing and describing the archived content, as well as making it searchable and browseable.

Preservation, the third area under web archiving, is the set of activities aimed at safeguarding the viability (intact bit streams), renderability (processable by computers), and understandability (to human beings) of the captured content. This kind of future-proofing is simultaneously the most important and least understood aspect of web archiving.

Web archiving was once a difficult and expensive undertaking. The institutional tools that we are creating, however, coupled with falling storage costs, will make it much easier for curators to build their own collections in a relatively self-service manner. They will include user interfaces that allow curators to initiate and monitor crawls, and to review, edit, and describe crawled content. Besides the project partners, the tools will be open-sourced and, when mature, distributed so that the entire digital library community can take advantage of them. With centralized, cooperating institutional repositories, there is less need for redundant collecting and for departmental repositories. Having said that, some specific, planned redundancy (replication) forms a core part of our preservation strategy.

1.2 Project Scope

[Review the timelines for the CHA, CIRT and CISA paths, and determine if they indicate clear versions and release dates for initial and subsequent releases. Specify which features, such as public searching access, will not be fully functional by the end of the project]

1.3 User Classes and Characteristics

The Web Archiving Service will be available to four different user types: end-users, curators, institution administrators and WAS administrators. The following section will define these user types, from the simplest to the most complex. These user types will correspond to separate WAS interfaces.

Consumers

Consumers are the people who will ultimately search and display materials in the archive. At the broadest definition, this group consists of the general public. These materials are being preserved in order to provide access to our digital cultural heritage, and the project is ultimately intended to serve the public. In some cases, only consumers from a particular institution or accessing from a particular workstation may be able to access materials.

Consumers will be able to search or browse the archive, and will be able to accurately display the materials in it. They cannot add or delete materials, or initiate crawls.

Curators

Curators are the users who create, populate and maintain archival collections. These users will generally be librarians at institutions served by the Web-at-Risk project partners (University of California, NYU, University of North Texas). They may also be scholars from related institutions.

Curators will be able to initiate crawls, create and update collections, and review reports of both collection-level and crawl-level activities.

Institution Administrators

Institution administrators are the users who maintain their institution's access to the Web Archiving Service. In order to limit materials by institution in any meaningful way, the system will have to provide a way for each institution to define and update the IP ranges associated with it. An institution administrator would also be responsible for adding and maintaining accounts for the curators from that institution.

Institutional Administrators may or may not also be curators. The features provided in the Institution Administrator interface would need to be accessible on a stand-alone basis outside of the curator interface.

Web Archiving Service Administrators

The WAS administrators will have access to all features available to the three previous user types, and will have additional features and reports available to them, such as adding institutional administrator accounts. This user's role is to monitor and maintain the overall functionality of the Web Archiving Service. Assuming that CDL, UNT and NYU all ultimately run their own instances of the Web Archiving Service, each instance would require at least one WAS administrator.

1.4 Operating Environment

General

The components of the web-archiving service will be compatible with the CDL Common Framework, which is a set of subsystems and interfaces that comprise CDL's technical infrastructure. This is a major leverage point, as CDL has too many services not to look for every opportunity to re-use existing, generalized technical components. Components built for this project are likewise expected to be re-usable.

The tools built will be open-source so that the archiving community can benefit from this publicly funded work, and our work can improve based on community feedback (e.g., usability reports, bug fixes). Tools will also be extensible, modular, and configurable so
that our Web-at-Risk partners may be able to run and adapt the tools within their local repository environments.

The tools must also be compatible with broader, external web-archiving efforts. For archives to be able to exchange data, avoid redundancy, and conserve limited resources, it is important to use common storage formats and metadata standards.

Where possible these built components should be compatible with or augment other CDL efforts and services. An example is OAI (Open Archives Initiative) metadata harvesting, which has many similarities to web harvesting; in both cases, automated background processes remotely capture data that results in the creation, validation, and storage of well-formed information objects.

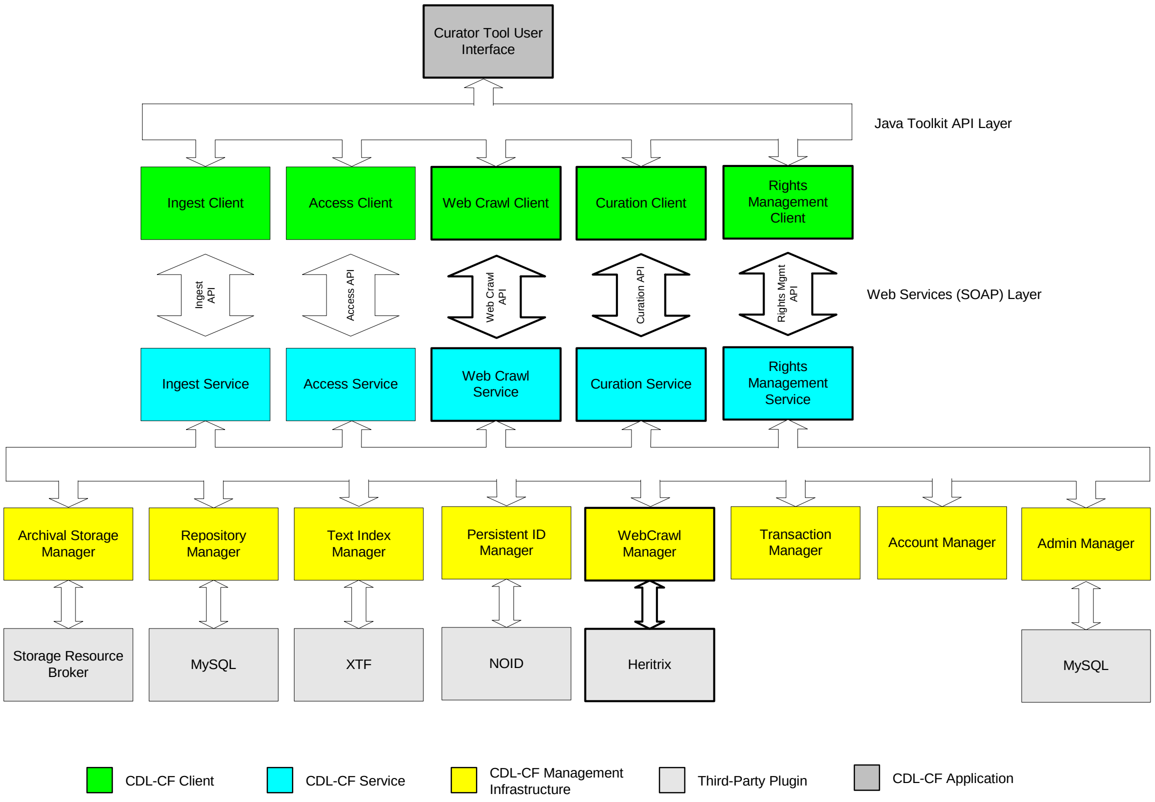
Technical Environment

CDL will build upon its existing software infrastructure, the CDL Common Framework (CDL-CF). The CDL-CF is an open, services-oriented architecture written in Java. Its fundamental principles are:

- clear separation of applications from underlying services;
- easy re-use of underlying services in multiple applications;
- simple, consistent design patterns and interfaces for services;
- consistent exposure of services through SOAP and Java Client API;
- clear separation of services from underlying data storage and other resources;

- easy integration of local and third-party solutions to specific problems through a plug-in approach;
- flexibility in service provision: shared service or end-user facing;
- scalability and flexibility at run time;
- and platform independence.

The following diagram depicts portions of the CDL Common Framework architecture and illustrates how CDL and its partners will build upon this architecture:



CDL and its partners will:

- Utilize existing CDL-CF services, including Ingest and Access (Dissemination)
- Utilize existing CDL-CF resource managers, including Archival Storage,
- Repository, Text Index, Persistent Identifier, Transaction, Account and Administration
- Utilize existing CDL-developed plug-in solutions, including NOID (for persistent ARK generation) and XTF (for text indexing and searching)
- Utilize third-party solutions that have already been incorporated into the CDL-CF, including Storage Resource Broker (for archival storage), MySQL (for storage of metadata and administrative data) and JHOVE (for content file validation and technical analysis)
- Develop new services to support the project, including:
  - A Web Crawl service for initiating and monitoring web crawls and for processing web crawl results. This service will interact with a new CDLCF resource manager (Web Crawl Manager),

which in turn will interact with Internet Archive's Heritrix web crawler, utilizing Heritrix's JMX interface.
    - A Curation service for defining web crawl collections, scheduling crawls, packaging web crawl results with archival metadata, submitting crawl results to the preservation repository, etc. This service will utilize existing data management and security modules and will interact with existing services, such as Ingest.
    - A Rights Management Service for identifying and recording rights metadata for crawled content.

- Develop an interactive, web-based curator tool that will use the existing and new CDL-CF services and will communicate with these services through standard CDL-CF Java Toolkit API's
- Develop a standard encoding format for representing crawled content and associated metadata, and utilize this format for moving content between CDL-CF services, and between CDL's repositories and partners' repositories. This format will likely combine METS metadata encoding and the next generation of Internet Archive's web archive file format (WARC)
- Extend CDL's implementation of Storage Resource Broker to include remote replication of objects

As indicated above, the project will utilize existing CDL Common Framework (CDL-CF) services for archival storage, metadata management, security and authentication. New services will be developed for rights management and the project's core work of curation and web crawling. CDL-developed "plug-ins" to the CDL-CF that will be utilized include:

- eXtensible Text Framework (XTF) – for text indexing and searching
- Nice Opaque Identifier (NOID) – for minting of persistent identifiers (ARKs)

Third-party plug-in software includes:

- Storage Resource Broker (SRB, San Diego Supercomputer Center) – for management of archival storage, including remote replication
- MySQL – for storage of metadata, curation management data and administrative data
- Heritrix – for web crawling

A component of the project will be to evaluate storage hardware solutions and storage management software that can complement SRB's capabilities. CDL will evaluate several approaches, including highly integrated solutions as well as loosely clustered commodity solutions.

The CDL-CF is Java-based and is platform-independent in terms of hardware and operating systems. Currently these run on a combination of Solaris and Linux systems.

Standards

The CDL-CF, upon which the project will be based, relies on the standard specifications of Java, XML, ARK, and various file formats that appear in ingested objects. To a very large extent these are influenced and inspired by interpretation of the OAIS model, METS, Dublin Core, and PREMIS metadata.

All CDL-CF services are exposed as web services, using SOAP with attachments via HTTP. The following standards will also apply to the WAS:

WARC : WEB ARCHIVING .....

The WARC standard has been established for use by this project and for the Internet Archive Project.

The WARC will determine: WHAT

The WARC will be applied to: WHAT

WADO: WEB ARCHIVE DIGITAL OBJECT

The WADO standard is being developed specifically for the Web-at-Risk project.

The WADO will determine: WHAT

The WADO will be applied to: WHAT

OAIS: OPEN ARCHIVAL INFORMATION SYSTEM

The Consultative Committee for Space Data Systems has established OAIS to guide the development of archival systems. The CDL's existing Digital Preservation Repository, upon which the WAS Archive will be based, was developed following OAIS guidelines.

This could be right, could be wrong... seems like we should give OAIS a shout out whether it applies directly or not.

METS: METADATA ENCODING & TRANSMISSION STANDARD

METS will determine: WHAT

METS will be applied to: WHAT

1.5 Dependencies

As described in the Operating Environment section, the Web Archiving Service will be integrated into the services that make up CDL's common framework. In addition to the resources mentioned above, there will be some additional dependencies on existing CDL systems.

D-1. IP Range Tracking

The Institution Administrator Interface will allow users to record and maintain IP ranges for their institutions. This will enable curators to limit selected materials to users from their institution only. This feature will be implemented in keeping with the tools and procedures CDL already uses to limit access to online resources to individual UC campuses.

2 Functional Requirements

What follows is a brief definition of each of the major functional areas of the Web Archiving Services, followed by more detailed requirements for that area. It will be noted which users may interact with a particular feature (see 1.3 User Classes and Characteristics).

FUNCTION DEFINITIONS:

Login

Login will be required to reach the curator and administrative interfaces. When consumers attempt to reach materials that are restricted by institution, they will also be prompted to login.

Navigation

There will be separate interfaces for the curators, administrators and consumers. Each interface will provide navigation to all features available to all features available to that type of user.

Help

Users will be able to reach context-appropriate help from within the WAS system.

Crawl

A crawl represents the rules specified to execute a web crawl. This includes the seed list and crawler settings. A crawl may be saved without having been run, (so there are no associated web documents). A crawl may also be run without adding any of the resulting documents to a collection. Crawl content is accessible to curators, but not to consumers.

Collections

A collection is composed of selected crawl results and represents materials that have been selected for preservation, so preservation requirements will be defined here. Collections are defined by curators, and may be accessible to other curators as well as to consumers.

Rights Management

Curators will be able to provide rights and restrictions information about URLs either prior to crawling them, after the crawl, or after the content has been added to a collection. The WAS system will behave according to the rights and restrictions specified for an item.

Search

The search feature will be available to all users. The scope of what is searched will vary depending on the user type and affiliation.

Display

The display feature will be available to all users. The scope of what is displayed will vary depending on the user type and affiliation.

Accounts

Both institution Administrators and WAS Administrators will be able to create and update accounts for other users. Users will be able to update selected information about their own accounts.

Reports

Curators and Administrators will be able to view reports of WAS activity and holdings.

Preservation

The preservation process will begin at the moment crawled materials are added to collections. *(What do we want to say concerning long-term preservation features & activities?)*

2.1 Login

Users: curators, institution administrators, WAS administrators, consumers (as needed.)

LOG- 1. Users will be prompted to login to the WAS curator and administrative interfaces.

LOG- 2. Login username and password will be determined by the WAS administrator who created the user's account.

LOG- 3. Users will receive appropriate contact information if login fails.

LOG- 4. Consumers interacting with a specific collection may be prompted to login if they encounter materials restricted by institution.

LOG-4.a. Consumer login will be based on proxy settings / IP ranges for their institution. WAS will not provide consumer accounts.

2.2 Navigation

Users: curators, institution administrators, WAS administrators, consumers

2.3 Help

Users: curators, institution administrators, WAS administrators, consumers

2.4 Crawls

Users: curators, institution administrators, WAS administrators

CRA- 1. Prepare new URLs for crawling

CRA-1.a. Provide Entry Point URL (EPU).

CRA-1.b. Confirm that EPU is functional.

CRA-1.b.1. If any site is not reachable, prompt the user to check for accuracy and allow user to edit if necessary.

CRA-1.c. Provide rights & restriction information for EPU (see section 2.6: Rights Management for details).

CRA- 2. Create new crawls. The new crawl process will include setting the crawl parameters below. All crawl parameters will have default settings (TBD) except for seed list values.

CRA-2.a. Provide a name for the crawl.

CRA-2.b. Select Entry Point URLs for seed list.

CRA-2.c. Provide a note describing purpose of the crawl.

CRA-2.d. Specify depth of external links to crawl.

CRA-2.e. Specify maximum crawl time.

CRA-2.f. Specify frequency / duration

CRA-2.f.1. Frequency (example: daily, weekly, monthly)

CRA-2.f.2. Duration (example: weekly for 6 months)

CRA-2.f.3. Default settings will be: frequency = once, duration = N/A.

CRA-2.g. Specify number of sub-domains within domain to crawl.

CRA-2.h. There will be a link to advanced crawl parameters for experienced users to specify:

CRA-2.h.1. Maximum file size.

CRA-2.h.2. Maximum overall size of crawl.

CRA-2.h.3. Maximum number of connections.

CRA-2.h.4. Maximum transfer rate (bytes / second).

CRA-2.h.5. Number of multiple connections (flow control).

CRA-2.h.6. Number of retries.

CRA-2.h.7. Timeout limit (in seconds).

CRA- 3. Confirm settings and create crawl (CREATE button).

TBD: are there any settings or setting combinations that are totally unacceptable? Specify all error prevention measures.

CRA- 4. Review existing crawls. Crawl modification will include the following activities:

CRA-4.a. Review all crawls previously executed in a chronological list that includes the crawl name, date created, and date(s) executed (if any).

CRA-4.b. Select any crawl in the list to view crawl parameters.

CRA-4.b.1. If the crawl has been previously run, view any reports resulting from previous crawl executions.

CRA-4.b.2. If the crawl has been previously run, view any metadata applied to the crawl.

CRA-4.b.3. If the crawl has been added to a collection, see details concerning collection.

CRA- 5. Modify existing crawls. Once the curator has selected an existing crawl, she will be able to modify the elements below. All existing values will display to the user.

CRA-5.a. Edit seed list

CRA-5.a.1. Remove URLs from seed list.

CRA-5.a.2. Temporarily deactivate URLs in seed list.

CRA-5.a.3. Add URLs to seed list.
CRA-5.a.4. Edit the URL, Title, Notes or Rights Info for any URL in the seed list.

CRA-5.b. Change crawl name.
CRA-5.c. Edit crawl note.
CRA-5.d. Change depth of external links to crawl.
CRA-5.e. Change maximum crawl time.
CRA-5.f. Change frequency / duration settings.
CRA-5.g. Change the number of sub-domains to crawl.
CRA-5.h. Link to separate screen to change advanced crawl parameters.
CRA-5.i. Review edits and update crawl (UPDATE button). System provides a date created value for the crawl when user clicks CONFIRM button.

CRA- 6. Execute Crawl

CRA-6.a. After user clicks the EXECUTE button, but before the crawler engages, the system will Q.A. the seed list and crawl settings.
CRA-6.b. Validate that each URL is reachable.
CRA-6.c. If user confirms that URLs are correct, but remote server is simply unavailable, allow user to either postpone entire crawl, or go ahead with only partially functioning seed list.
CRA-6.d. Once the Q.A. process is complete, the crawler will engage, and the user will be routed to the crawl status screen.

TBD: How do you monitor a crawl-in-progress? What do you see?

CRA- 7. WAS will create a crawl-job that links the crawl name and parameters with this instance of the executed crawl.

CRA-7.a. System will supply a run-date for the crawl-job.
CRA-7.b. System will supply metadata as materials are gathered from crawl.

CRA- 8. The “user-agent” crawler value will be set to supply the name of the logged-in curator, the curator’s institution, and a link to a web page supplying information about the project and how content owners may opt out of having materials archived (see requirement RM- 3 for related details).

2.5 Collections

Users: curators, institution administrators, WAS administrators

The WAS collections represent what has been archived and will be preserved. Until data resulting from crawls has been added to a collection, it is not searchable, does not have image or plain text versions of pages, and cannot be viewed by anyone other than the curator who initiated the crawl.

This area will only show collections that the logged in curator has the ability to update. It is not intended to be a UI for the entire WAR contents, but is the tool to help the curator manage her own materials.

COL- 1. Review collections. From Collections main screen, the curator will see a list of basic details about the collections she has created or can contribute to. Navigation will be provided to Create, Delete, Edit, Add Material, Reports and Transfer. Collection Review List will include:

- COL-1.a. Collection name.
- COL-1.b. Number of files.
- COL-1.c. Date created.
- COL-1.d. Owner (curator name).
- COL-1.e. Link to details. (Collection Details screen will include TBD).

COL- 2. Create New Collection.

- COL-2.a. Add collection name.
- COL-2.b. System will supply logged-in curator as the collection owner.
- COL-2.c. Add co-curators.
- COL-2.d. Add collection description. This description will appear on the collection main screen. (Specify length. Big.)
- COL-2.e. Add navigation menu labels.
- COL-2.f. System will supply date-created value.
- COL-2.g. Upload an image for the collection welcome screen.
- COL-2.h. The system will supply information about the institution sponsoring the collection, based on data drawn from the collection owner and co-curator user accounts. This information will display on the collection welcome page.
- COL-2.i. Is there any kind of agreement or rights statement that needs to be linked to the collection?
- COL-2.j. CREATE button. User will be prompted to review values and either CONFIRM, EDIT, or CANCEL.

COL- 3. Edit a Collection.

- COL-3.a. Edit collection name.
- COL-3.b. Add or Delete co-curators.
- COL-3.c. Edit collection description.
- COL-3.d. Add, edit or delete navigation menu labels.
- COL-3.e. Only navigation labels that are not linked to any content can be deleted.
- COL-3.f. System will supply date-updated value.

COL- 4. Add Material to a Collection.

- COL-4.a. Select materials to add to a collection. Curators will be able to choose from:
  - COL-4.a.1. Crawls they own (crawls they conducted themselves.)
  - COL-4.a.2. Any materials they find in any other collection. (Note that rights established for the context of the other collection may not apply to the new collection. Rights information may require redetermination.)
- COL-4.b. Add the contents of an entire crawl to a collection.
  - COL-4.b.1. Navigate into the crawl details to selectively add individual seed-list URLs from a crawl to the collection. All resulting linked pages will be included.
  - COL-4.b.2. Select individual pages from the crawl to add to the collection.
  - COL-4.b.3. View any metadata gathered about crawl contents.

COL-4.b.4. Display the content before deciding to add it to the collection.

2.6 Rights Management

Users: curators, institution administrators, WAS administrators, consumers

Consider using terminology found in Nancy Hoebelheinrich's METS Rights Declaration Schema. Elements and attributes specified there may be sufficient to provide all of the functionality required for WAS rights management.

RM- 1. Rights management functionality will be accessible at appropriate points in the curatorial process.

RM-1.a. Curators will be prompted for rights information when entering new EPUs to be crawled.

RM-1.b. Curators will be prompted for rights management information when they attempt to add domains to collections that do not have any associated rights data.

RM-1.c. Curators will be prompted for rights management information when they attempt to crawl materials that are protected by robots exclusion files.

RM-1.d. Curators will be able to navigate to rights management features at any point in the archiving process; rights management features will remain available after an item has been added to a collection for preservation.

RM- 2. The curator can set a rights management scheme for a fully qualified domain name (example: http://www.usps.gov). Based on the Web-at-Risk Rights Management Protocol, there are three RM schemes for domain names: A: Open Rights, B: Permission Implicit and C: Permission Required.

RM-2.a. When a domain name is set at RM Scheme A: Open Rights, the curator will be able to add it to a collection and will not be prompted to contact the content owner.

RM-2.b. When a domain name is set at RM Scheme B: Permission Implicit, the curator will be able to add it to a collection, but will be prompted to inform the content owner at a later time.

RM-2.c. When a domain name is set at RM Scheme C: Permission Required, materials from that domain cannot be added to a collection until there is also associated rights metadata linked with it.

RM- 3. The curator will be able to enter text describing the purpose of the crawl. This text will be rendered as a web page. The URL to this page will be part of the "user-agent" field at the time the crawl is executed. (see CRA- 8 for crawl user-agent requirements.) This text will serve as the sole method of content owner notification in RM Scheme A: Open Rights.

RM- 4. Rights management information associated with a domain will be available for display.

RM-4.a. For materials with RM Scheme A: Open Rights, consumers will be notified that the material is in the public domain.

RM-4.b. For all other materials, consumers will be presented with a rights warning (language TBD).

RM-4.c. All curators will be able to see complete rights information associated with an item.

RM- 5. The curator can set the following access restrictions to either domains or individual files:

RM-5.a. Restricted to curator access only. The material will be visible to any other curator with access to the collection.

RM-5.b. Restricted to users from the archiving institution.

RM- 6. Consumers will be able to see limited rights information in the metatdata for each item.

RM-6.a. For all other materials, consumers will be presented with a rights warning (language TBD).

2.7 Search

Users: curators, institution administrators, WAS administrators, consumers

SEA- 1. Enter search text

SEA-1.a. Enter keywords

SEA-1.b. Enter URL

SEA- 2. Choose search scope

SEA-2.a. Across collections or within any combination of collections.

SEA-2.b. Search against both metadata and full text.

SEA-2.c. Search metadata separate from full text.

SEA- 3. Browse collections

SEA-3.a. Browse a collection by navigation label.

SEA-3.b. Browse by site metadata: title, subject, creator, date.

SEA-3.c. Across collections or within a collection.

2.8 Display

DIS- 1. Display search results. Display features will include:

DIS-1.a. Brief records (details TBD via needs analysis). Initial display will clearly indicate that the user is viewing materials within an archive, not the live sites.

DIS-1.b. Full item record (details tbd).

DIS-1.c. Full look and feel of website.

DIS- 2. The user will be able to navigate through archived web sites.

DIS-2.a. Within a website there will be a user-friendly way to determine

DIS-2.a.1. When a link will lead to material that is also archived.

DIS-2.a.2. When a link will lead to material that is outside of the archive.

DIS-2.b. Users will be able to display any existing metadata about the page they're viewing.

DIS-2.c. Across website snapshots:

DIS-2.c.1. Provide navigation through time when many versions of the site are available in the archive.

DIS-2.c.2. Identify when archived versions in timeline are significantly different.

DIS- 3. Materials will display appropriately according to the rights associated with them.

DIS-3.a. If access to materials is limited by institution, and the user is accessing via an IP address within that institution's range, the item will display without incident.

DIS-3.b. If access to a search or browse result is limited by institution and the user is accessing via IP that is out of that institution's range, the user will:

DIS-3.b.1. See the metadata about the item.

DIS-3.b.2. See an explanation of why the material is not viewable.

DIS-3.b.3. See a username and login box allowing him/her to authenticate to that institution's proxy server?

DIS-3.b.4. If access to materials is limited to curators only, the users will retrieve any information about them at all.

DIS-3.b.5. If access to materials is limited to a particular location/workstation, users will

1. See the metadata about the item.
2. See instructions as to where to go to access the material.

2.9 Accounts

Users: curators, institution administrators, WAS administrators

ACC- 1. WAS Administrators will be able to create and edit user accounts. This will include:

ACC-1.a. User name

ACC-1.b. User Login

ACC-1.c. User Password

ACC-1.d. Institution name

ACC-1.e. user email

ACC-1.f. Institution IP range?

ACC- 2. Institution Administrators will be able to create and edit new user accounts for their own institutions only.

ACC- 3. Curators will be able to edit the following information concerning their own accounts only:

ACC-3.a. User password

ACC-3.b. User email

2.10 Reports

Users: curators, institution administrators, WAS administrators

2.11 Preservation

Users: curators, institution administrators, WAS administrators

The preservation process will begin at the point when materials are added to collections.

PRE- 1 Create preservation copies. As materials are added to collections WAS will generate copies in alternate formats.

PRE-1.a. Text versions of HTML and PDF formats will be rendered, linked to the item, and saved.

PRE-1.b. JPG versions of HTML formats will be rendered, linked to the item and saved.

PRE- 2 All preservation activities will be recorded and linked to the item. Information concerning preservation activities will be viewable to curators.

3 Glossary

Web Archiving Service

Web-at-Risk Project

NOTE: while we do have a preservation glossary, I wouldn't suggest using it here… it's kind of overkill. I would only include words that appear in this document. If a word from this document is already defined in our glossary, use that same definition here.

4 Issues List

The following issues are still to be determined:

TBD-1. The crawler user interface requirements need to be specified in more detail, based on further research on Heritrix and on the results of the user needs assessment. This assessment will be completed in September 2005.

TBD-2. How long are we promising the curators that we will preserve the materials they add to this system in the course of this next three years? What preservation features do we anticipate we'll need to account for that span of time?

TBD-3. How exactly do crawls and collections interact? If material has been crawled but never added to a collection, does it have any metadata yet? (TS thinks no.) Is it find-able by the search mechanism? (TS thinks no.) Where is it stored? (TS has no idea.) Is it linked to any rights management info yet? (TS thinks no.)

TBD-4. What level of granularity do we need for rights management? Are we applying rights status by domain? By URL? Both are impossible, but domain appears to be less impossible.