Web At Risk: Extending the Digital Curation Mission to the Web

Patricia Cruse, Director, Digital Preservation Program
Kirsten Neilsen, Digital Preservation Services Manager
California Digital Library
The Digital Preservation Program

- Established in 2002
- UC-wide program
- **Goal:** ensure long-term availability and accessibility to materials that are important to the research, teaching, and learning on the UC campuses.
- Centrally managed
- Central and external funds
- A partnership
Cornerstone of the Program: Digital Preservation Repository (DPR)

- Suite of tools & services:
  - Digital Preservation Repository
  - Documentation, guidelines, policies
- Intern’l Standards & Open Source
- Service oriented architecture: flexible, adaptable, simple
- Preservation Partnership
  - Curate
  - Preserve
Digital Preservation Repository core services

• A set of services that support the long-term retention of digital objects:
  – Submit (deposit) digital objects
  – Manage digital objects: add versions, replace, update, delete
  – Request dissemination
  – Request administrative reports (forthcoming)

• What the service is not…
Humans using browsers can interact with the DPR through the CDL-supplied HTML interface.

Applications written in Java can interact through the CDL-supplied Java client library.

Applications written in any language that supports SOAP (with Attachments) can interact through the CDL-supplied SOAP interface.
Main Menu

Submit Objects

Create new objects; or add or replace versions of objects.

Submit a single object
You may wish to validate the object first.

Submit multiple objects
You may wish to validate the objects first.

Browse

Browse by inventory
- UCD - Eastman Photographs
- UCD - Yolo Aerial Photographs
- UCD - Faculty Entomology Images

Search

Locate an object by its identifier
The locate functionality performs exact matches on identifiers.

Object ID (ARK):

ARK syntax: ark:/mmnnn/cmmmcccccc

Locate

To locate an object by its Local Object ID, choose the inventory to which that object belongs.

Locate an object to either add or replace versions, or to view or download a version.
DPR to Web Archiving Service
Web-at-Risk: NDIIPP Funds

• Build tools to allow librarians to capture, curate and preserve web-based government and political information.
  – Create topical and event-based archives
  – Capture individual sites and documents

• Assess the impact of these tools on traditional collection development practices.

• Explore web archiving service sustainability.
Project Partners

Technical Partners
- San Diego Supercomputer Center
- Computer Science Dept., Stanford University
- Sun Microsystems

Library of Congress
- Martha Anderson - Agreement Officer
  Technical Representative

California Digital Library
- Daniel Greenstein – P.I.

New York University
- James Bullen

University of North Texas
- Cathy Hartman

Curatorial Partners
- Arizona State Library
- Stanford
- UC Davis
- UCLA
- UC San Francisco
- New York University
- UC Berkeley
- UC Irvine
- UC San Diego
- UC Santa Cruz
Preserving the Web

- Why all the fuss?
- What is “Web Archiving?”
- Web Archiving Service (WAS)
  - Collecting content
  - Curating content
- Current status & future plans
The page cannot be found

The page you are looking for might have been removed, had its name changed, or is temporarily unavailable.

Please try the following:

- Make sure that the Web site address displayed in the address bar of your browser is spelled and formatted correctly.
- If you reached this page by clicking a link, contact the Web site administrator to alert them that the link is incorrectly formatted.
- Click the Back button to try another link.

HTTP Error 404 - File or directory not found.
Internet Information Services (IIS)

Technical Information (for support personnel)

- Go to Microsoft Product Support Services and perform a title search for the words HTTP and 404.
- Open IIS Help, which is accessible in IIS Manager (inetmgr), and search for topics titled Web Site Setup, Common Administrative Tasks, and About Custom Error Messages.
404 - Document Not Found

404: Page can not be found

HTTP 404: Page Not Found
• 2003 survey of the .gov domain:

  – as much as 65 percent of all government publications that are distributed to libraries through the federal depository library program are currently produced exclusively in electronic form and distributed via the web.
What is a “Web Archive?”

- Automated method to gather web content
- Collections composed of multiple sites
- Captured content preserved
- Meaningful access to content provided
  - Public or end-user access may not be available
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages</td>
<td>2 pages</td>
<td>5 pages</td>
<td>13 pages</td>
<td>15 pages</td>
<td>19 pages</td>
<td>18 pages</td>
<td>33 pages</td>
<td>51 pages</td>
<td>13 pages</td>
<td>0 pages</td>
<td></td>
</tr>
</tbody>
</table>
Domain-Based Web Archives

Nordic Web Archive  →  Nordic National Libraries
Kulturarw3  →  National Library of Sweden
National Web Archive  →  National Library of Iceland
Topical Web Archives

UCLA Library

Digital Collections

Online Campaign Literature Archive

A century of Los Angeles elections

Every American election produces thousands of campaign flyers, pamphlets, posters, and bumper stickers, generally called "campaign literature." These documents provide an important record of the campaign, its participants, issues, and tactics. Despite this value, the small size, short production period, and irregular distribution of the documents, all outside the bounds of the traditional publishing industry, put most campaign literature beyond the scope of standard library collections. These materials are seldom saved for posterity.

In order to fill this gap, since the 1920s the government documents section of the UCLA Library has built and maintained a Campaign Literature Collection, containing printed ephemeral election materials distributed by campaigns for local, state, and federal offices and ballot measures affecting the Los Angeles area. In response to changes in election campaigning brought about by the internet, in 1990 the Library also began to capture and preserve copies of campaign websites.

The UCLA Online Campaign Literature Archive presents a subset of the
Event-Based Web Archives

**Hurricanes Katrina & Rita Web Archive**

Internet Archive and many individual contributors created a comprehensive list of websites documenting the historic devastation and massive relief effort due to Hurricanes Katrina. The sites were crawled between the dates of September 4 - October 17th. This collection, containing more than 25 million searchable documents, will be preserved by Internet Archive with access to historians, researchers, scholars and the general public.

Querying works generally as it does in Google with some caveats. See host queries work in the Katrina search for discussion.

**News Sites**

- The Times-Picayune

**Relief Sites**

- USA Freedom Corps

**Complete List of Crawled URLs**

---

**Election 2000 Web Archive**

The Election 2000: an Internet Library is a selective collection of nearly 900 sites archived daily between August 1, 2000 and January 21, 2001, covering the United States national election period.

The Library of Congress commissioned the Internet Archive to create the Election 2000: an Internet Library. The Internet Archive provided project coordination and quality assurance. Compaq Computer Corporation captured and stored all the digital materials collected. The Internet Archive currently hosts the collection and provides access to the archive using its Wayback Machine technology. The Election 2000: an Internet Library was made available to the public in June of 2000.
See What's New: WAS Release 2 for details about new features. See Capturing Sites: Overview for a brief guide to the capture process.

Add Site
Create entries for sites you plan to capture.
Provide seed URLs, capture settings and descriptions.

Manage Sites
Capture, edit or delete existing sites.
Review site capture history & site details.

View Results
Search and browse archived content.
Review capture job reports.
Web Archiving Lingo

- Crawler
- Host
- Site
- Seed
- Capture
- Robots.txt
General Info

Description: 
Capture Job Status: Finished
Start Time: Apr 05, 2007 - 10:02:04 AM
Finish Time: Apr 05, 2007 - 12:05:19 PM
Duration: 2h 54m
Size: 41.1 MB
Documents: 387

Add this capture job to collection:
California Water Sites

Robot Exclusions

Robots.txt files
The following robots.txt files were discovered during your capture. The robots.txt files are archived in order to document the host server policies on the date this capture job was run. Click the link to view the archived version of the robots.txt file.

- http://www.lacity.org/robots.txt
- http://download.macromedia.com/robots.txt

See Robots Help for more information.

Mimetypes

Below are the most common mimetypes for this capture job. See the Mimetypes Report for complete statistics.

<table>
<thead>
<tr>
<th>Mimetype</th>
<th>URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>application/pdf</td>
<td>282</td>
</tr>
<tr>
<td>text/html</td>
<td>66</td>
</tr>
</tbody>
</table>
Sample Collection Plan

- Section 1. Mission & Scope
- Section 2. Selection
- Section 3. Acquisition
- Section 4. Descriptive Metadata
- Section 5. Rights and Access
- Section 6. Maintenance and Weeding
- Section 7. Preservation

- Appendix A. Letter of Agreement
- Appendix B. Seed List
- Appendix C. Metadata
Flexibility in the face of uncertainty
What metadata will you need?

Title:
Subject:
Author:
Date:
Rights Management Approaches

• **Library of Congress**
  – Extensive rights management efforts
  – Permission secured for any site not clearly in the public domain
    • If no response, the site is not captured

• **Internet Archive**
  – Opt-out policy
  – Obey robots.txt

• **WAS**
  – Flexibility
Preservation

• Content preserved in the DPR
  – Bit preservation (fixity, integrity)
  – Replication
  – Desiccation

• Massive storage requirements
  – Multiple projects investigating mass storage environments
WAS: Now & into the Future

• Current Status
  – in development
  – 12/07 roll out to current curators

• Beyond 2007
  – Extending service to additional curators
  – Developing end user access
  – Exploring release of open access tools
Acknowledgements

• Tracy Seneca, Web Archiving Coordinator
  – CDL WAS development team

• Kathleen Murray
  – UNT Partners

• NDIIPP