User Assessment / Reality Check

Discovering Librarian Needs for Web Archiving

Tracy Seneca
California Digital Library

WEBatRISK
Preserving Our Nation’s Cultural Heritage

National Digital Information Infrastructure Preservation Program
Library of Congress

California Digital Library
New York University
University of North Texas
Problem at hand

Needs assessment

Design

User feedback

Tools available
The Web Archiving Service enables you to define web sites of interest, capture content from those sites, view results, and combine captured content from different sites into topic-specific collections.

1. Create Site
   - Provide URLs, capture settings and metadata for the sites you plan to capture.

2. Capture Site
   - Start capturing the site content. WAS will email you when complete.

3. View Captures
   - Search, browse and display archived content. Review capture reports.

4. Build Collections
   - Create topic-specific collections of web content from your captures.
Web Archiving Service: Ingredients

• CDL Digital Preservation Repository
  – Java backend
  – SRB for storage
  – MySQL

• Open source tools – (Thanks to IA)
  – Heritrix crawler
  – NutchWAX / Open Source Wayback

• Ruby on Rails curatorial interface
Search By Keyword: Oroville

116 search result(s) found for Oroville

**Title:** Oroville Facilities Relicensing - Home Page
- **Captured:** Wed May 23 22:11:20 -0700 2007
- **URL:** http://orovilrelicensing.water.ca.gov/
- **Abstract:** Text version of the Oroville Facilities Relicensing Website. © 2000 California Department of Water Resources. All Rights Reserved. Web Master: Dave Lane Site updated ...
- **Add file to collection:** California Water Management Resources

**Title:** http://fpmtaskforce.water.ca.gov/images/orodam2.jpg
- **Captured:** Wed May 23 22:16:55 -0700 2007
- **URL:** http://fpmtaskforce.water.ca.gov/images/orodam2.jpg
- **Add file to collection:** California Water Management Resources

**Title:** THE CALIFORNIA WATER DELIVERY SYSTEM: IMPACTS OF CLIMATE VARIABILITY
- **Captured:** Wed May 23 22:12:18 -0700 2007
- **URL:** http://www.climatechange.water.ca.gov/docs/sfpuw_climate.pdf
- **Add file to collection:** California Water Management Resources
Pilot User Community

• Government information focus influences vision of web archive features

• Only 44% plan to collect “at the website level.” Most plan to collect documents

• Only half think it is important to be able to render the archived material as originally seen on the web
Subject specialists traditionally apply their expertise **before** purchasing print materials.
• Web crawlers cast a wide net; you can’t know in advance what you will get
• You may not know what was at risk until it’s gone
Context: Deliberately Building a Collection of Known Sites

- Available capture settings are important
- Capture depth (comprehensiveness) is critical
- Time for
  - Descriptive metadata
  - Quality control
  - Rights management
  - Analysis/expertise
Create Site

**Capture Settings**

* Required field

**Site Name:** Angeles Housing Department

**Seed URLs:**
Ex.: http://www.example.com

**Scope:** Directory

Capture Linked Pages: No  Yes

Max. Time: Full Capture (36 hours)

**Scheduling**

**Descriptive Data**

**Rights Data**

[Cancel  Save (all tabs)]
Event: Southern California Wildfires

View Captures
Click ☑️ to view the captures associated with a site.

1-10 of 161  Prev  1 2 3 4 ... 17  Next ➤

<table>
<thead>
<tr>
<th>SITE NAME / CAPTURE DATE</th>
<th>STATUS</th>
<th>FILES</th>
<th>DURATION</th>
<th>ACTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>211 San Diego Wildfire Emergency</td>
<td>(5)</td>
<td></td>
<td></td>
<td>Compare</td>
</tr>
<tr>
<td>Blog: And Still I Persist</td>
<td>(4)</td>
<td></td>
<td></td>
<td>Compare</td>
</tr>
<tr>
<td>Blog: barboni.org</td>
<td>(5)</td>
<td></td>
<td></td>
<td>Compare</td>
</tr>
<tr>
<td>Blog: California Fire News</td>
<td>(6)</td>
<td></td>
<td></td>
<td>Compare</td>
</tr>
<tr>
<td>Blog: Cat Dirt Sez</td>
<td>(5)</td>
<td></td>
<td></td>
<td>Compare</td>
</tr>
<tr>
<td>Blog: Chronicle of Higher Education</td>
<td>(3)</td>
<td></td>
<td></td>
<td>Compare</td>
</tr>
<tr>
<td>Blog: CNN Your e-mails</td>
<td>(3)</td>
<td></td>
<td></td>
<td>Compare</td>
</tr>
<tr>
<td>Blog: Firefighter Blog</td>
<td>(5)</td>
<td></td>
<td></td>
<td>Compare</td>
</tr>
<tr>
<td>Blog: Google Earth Blog</td>
<td>(3)</td>
<td></td>
<td></td>
<td>Compare</td>
</tr>
</tbody>
</table>
Context: Event Capture

• Frequent, shallow captures
  – Top few pages on a daily basis

• Specialized settings
  – Blogs: page + linked pages (no matter what host)

• Time for selection only
Event capture: critical features

- Scheduling

- Bookmarklet for Firefox, IE: Add to WAS
  - Add sites for capture as you navigate

- Site management
Who are the web archivists?

• Event captures require you to drop everything for at least the first 3-5 days of the event

• When your target users are already juggling public service, collection development and other tasks, how do you find people who can do this?