

User Assessment / Reality Check

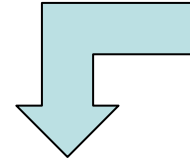
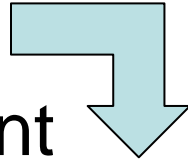
Discovering Librarian Needs for Web Archiving



Tracy Seneca
California Digital Library

National Digital Information Infrastructure Preservation Program
Library of Congress

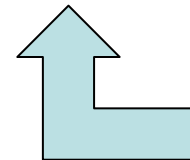
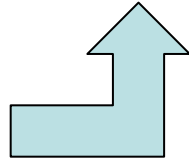
Needs
assessment



Problem at
hand

Design

User
feedback



Tools
available





Sites	Captures	Collections	Administration	
-----------------------	--------------------------	-----------------------------	--------------------------------	--

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

The screenshot shows a form with three tabs: 'Capture Settings', 'Descriptive Data', and 'Rights Data'. The 'Capture Settings' tab is active. It contains a text field for '* Site Name:' and a text area for '* Seed URLs:' with a small example 'Ex. http://www.example.com' below it.

Provide URLs, capture settings and metadata for the sites you plan to capture.

2. Capture Site

SITE INFO	STATUS	ACTIONS
Bureau of Reclamation Mid-Pacific Region The Mid-Pacific Region is one of five Reclamation.	Running Started 1 minute ago 9 files captured Stop this job	Capture
California Department of Water Resources DWR operates and maintains the State	Finished Last completed 22 days ago	Capture
California State Water Resources Cont...	Finished Last completed 22 days ago	Capture

Start capturing the site content.
WAS will email you when complete.

3. View Captures

The screenshot shows a search results page with two entries. Each entry includes a thumbnail image, a title, a file size, a date, and a 'View Details Report' link.

Search, browse and display archived content.
Review capture reports.

4. Build Collections

The screenshot shows a 'General Info' section with a description: 'Web publications from California and federal government agencies and non-profit organizations concerning water resources in California.' Below is a 'Contents' section with a tree view showing folders for 'Bureau of Reclamation Mid-Pacific Region', 'California Department of Water Resources', and 'California State Water Resources Control Board', each with a date and time stamp.

Create topic-specific collections of web content from your captures.

Recent Capture Activity

- 1 ingesting now
- 14 completed in total

Detailed Guides

- [What's New: WAS-4](#)
- [Getting Started](#)
- [WAS-4 User Guide \(PDF\)](#)

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

Search

116 search result(s) found for Oroville ◀ Prev 41-50 of 116 Next ▶

Title: Oroville Facilities Relicensing - Home Page

0 Bytes text/html

Captured: Wed May 23 22:11:20 -0700 2007

URL: <http://orovillereicensing.water.ca.gov/>

Abstract: Oroville Facilities Relicensing - Home Page Text version of the Oroville Facilities Relicensing Website. © 2000 California Department of Water Resources. All Rights Reserved. Web Master: Dave Lane Site updated ...

[Show Detailed Record](#)

Add file to collection:

California Water Management Resources



Add



Title: <http://fpmtaskforce.water.ca.gov/images/orodam2.jpg>

0 Bytes image/jpeg

Captured: Wed May 23 22:16:55 -0700 2007

URL: <http://fpmtaskforce.water.ca.gov/images/orodam2.jpg>

[Show Detailed Record](#)

Add file to collection:

California Water Management Resources



Add

Title: THE CALIFORNIA WATER DELIVERY SYSTEM: IMPACTS OF CLIMATE VARIABILITY

0 Bytes application/pdf

Captured: Wed May 23 22:12:18 -0700 2007

URL: http://www.climatechange.water.ca.gov/docs/sfpuc_climate.pdf

Abstract: ... 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 Water Year 1, 000 c f s Unimpaired Runoff at Oroville Dam Changes in Peak Flows Feather River Red Line = Construction of Oroville Dam San Joaquin River Runoff Annual Maximum 1-Day Flow 0 25 50 75 100 1900 1905 1910 1915 1920 ...

[Show Detailed Record](#)

Add file to collection:

California Water Management Resources



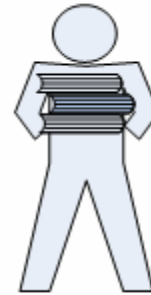
Add

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



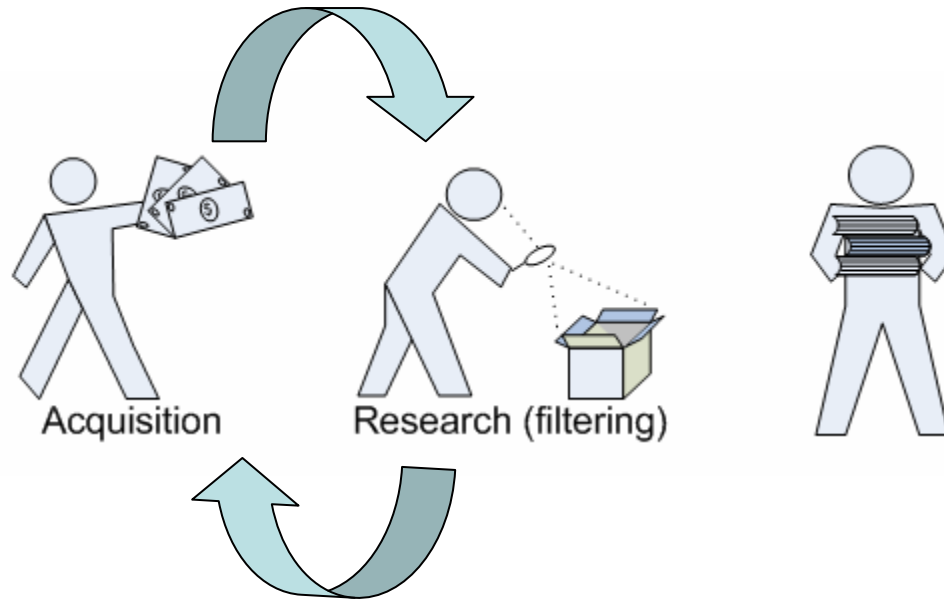
Collection-Building Paradigms



- Subject specialists traditionally apply their expertise **before** purchasing print materials



Collection-Building Paradigms



- 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

Scheduling


Descriptive Data

Rights Data

* Required field

* Site Name:

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

 Scope:

Capture Linked Pages: No Yes

Max. Time:


Cancel


Save (all tabs)



Event: Southern California Wildfires





















View Captures

Click  to view the captures associated with a site.

Capture Status 

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

Limit: ▼

SITE NAME / CAPTURE DATE	STATUS	FILES	DURATION	ACTIONS
  211 San Diego Wildfire Emergency (5)				Compare
  Blog: And Still I Persist (4)				Compare
  Blog: barboni.org (5)				Compare
  Blog: California Fire News (6)				Compare
  Blog: Cat Dirt Sez (5)				Compare
  Blog: Chronicle of Higher Education (3)				Compare
  Blog: CNN Your e-mails (3)				Compare
  Blog: Emmet Pierce Blog: Evacuating Penasquitos (4)				Compare
  Blog: Firefighter Blog (5)				Compare
  Blog: Google Earth Blog (3)				Compare



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?

