Project to Production: The Web Archiving Service

DLF Fall Forum, 2010

Tracy Seneca
Today:

- Whirlwind tour
- Project to production
- Current / future work
- The DLF community
Build unique archives for local research communities

Geographically focused archives support local research
- Los Angeles
- Monterey Bay
- Orange County
- San Diego
- Santa Barbara

Topical archives support special research collections
- Guantanamo Bay / Tamiment Library
- California Water Districts / Water Resources Center Archives
Archive Ephemeral Content
Analyze Site Change

Are there new documents on this site?

Are there documents in my archive that have been removed from the live site?
Easy to Use: Simple Workflow

The Web Archiving Service enables you to define websites, capture content from those sites, and view results. In the summer of 2009, you will be able to provide public access to the sites you've captured.

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. Public Access
   - Configure your project for public access. Set the description, URL or banner images.

Recent Capture Activity
- 4 completed in past week
- 670 completed in total

Capture Publication
This project is viewable by the public.

The public can view:
- 298 sites
- 302 captures

Detailed Guides
- Getting Started
- WAS User Guide
Tailor Capture Settings

Focus on site, directory or page
Set appropriate frequency
http://was.cdlib.org

The Web has revolutionized our access to information, but Web publications are fragile, and ready access to Web resources cannot be taken for granted. The Web Archiving Service enables librarians and scholars to meet that challenge.

Search Archives
Subject experts have been capturing and preserving critical web sites to ensure that you have lasting access to web content. Public archives can be browsed or searched and provide persistent links to archived documents. Access to published archives is open to all. Click on an archive name on the right to search and view archived materials.
Learn more

Build Archives
With a WAS account you can capture, analyze and archive web sites and documents. Archives can be published or kept for private study. The WAS curator tools are easy to use, fully hosted, and allow collaborative collection building. CDL provides training and guidance for WAS curators.
Learn more

Featured Archive
Santa Cruz Mountains Wildfire Web Archive
Created by: University Library, UC Santa Cruz
Sites: 35 Oldest site: 05/23/08 Most recent site: 06/29/08
Description: The Santa Cruz Mountain Wildfire web archive collection contains images, newspaper articles, and personal blogs related to the Santa Cruz Summit Fire of May 2008

Information for
- Potential Partners
- Researchers
- WAS Curators
- Webmasters

The Archives
Move pointer over an archive name for more information
- 2003 California Recall Election
- 2007 Southern California Wildfires Web Archive
- 2009 H1N1 Influenza A (Swine Flu) Outbreak
- AFL-CIO - Change to Win: the open web archive
- African Politics Web Archive
- Alternative Mass Media / News Web Sites Web Archive
- Anarchism Web Archive
- California State Government: .ca.gov Web Archive
- California Water Districts Web Archive
- Communism, Socialism, Trotskyism Web Archive
- Guantanamo Bay Detention Camp Web Archive
California State Government: .ca.gov Web Archive

UC Government Information Librarians

Description

The .ca.gov domain is the 3rd largest domain in the U.S. government web. A great deal of that content is published by the State of California, and represents a rich source of material for researchers, students, the general public and state agencies themselves. This archive preserves access to hundreds of state agency sites that are captured on a regular basis.

State agencies utilize their websites to publish everything from press releases, agendas, minutes, events, reports and statistics. This material...

more . . .

Search

Quick Facts

Sites: 299
Oldest site: 10/24/08
Most recent site: 01/03/10
Automatic List of Sites

Los Angeles Local Government Web Archive

Home  About  Site List  Search  Help  Contact Us

Previous 1 2 3 Next

07 HomeWalk LA - Home  Show Information
1st District Gloria Molina  Show Information
2005 Greater Los Angeles Homeless Count 2  Show Information
2007 Greater Los Angeles Homeless Count  Show Information
2nd District Mark Ridley-Thomas  Hide Information

Starting URL
http://ridley-thomas.lacounty.gov/ (live link)

Captured
12/11/08 02:20 AM
12/23/08 12:51 AM
12/27/08 04:13 PM

Click the "information" icon to see further information about the site, including the dates on which each site was captured.
The site lookup box will find words anywhere in the site name.
Search Across All Sites or Target Just One

San Diego Local Governments Web Archive
UC San Diego Libraries

Search
Type words or a URL:
border

Filter by:
All sites
Search
All file types

Border Affairs - SignOnSanDiego.com
Captured: Sun Dec 06 17:31:50 UTC 2009

Mexico-U.S. Border : AFSC
Captured: Wed Apr 30 20:05:10 UTC 2008
www.afsc.org/mexico-us-border/default.htm 17.3 KB text/html
... foundations of effective cross-border organizing. Corporate Responsibility Thousands of ... site The CFO (Border Committee of Workers ...
WAS Snapshot: October 2010

Stats: Since January 2007
17 organizations
4,509 sites captured
33,619 captures run
21.7 terabytes
29 archives published
Project to Production

• Scale
  – Infrastructure: ensure continuity of service through server upgrades

• Policy
  – Setting up service agreements
Engage the users
(Needs assessment doesn’t stop)
“Production” is not just for us!

• How does WAS fit into existing workflows?
• Is there a ‘web archivist’? Student workers?

Friday November 5th 10-11:30 am

Curators from UC San Diego, NYU, University of Michigan and UC Berkeley share their approaches to web archiving.

http://was.cdlib.org

WAS News
Archiving the State of California
Learning from the archives

California State Agency Sites
- with robots: 51%
- no robots: 47%
- gone: 2%

307 sites
1363 captures
2.7 TB
Began October 2008

University of California Curation Center, California Digital Library
Archiving the Gulf oil spill
Improving support for collaboration

CDL Archives
CDL’s Web Archiving Service archives the Deepwater Horizon oil spill.

527 sites
9288 + captures
1.6 TB
Began May 5
LSU tags relevant sites in Delicious
CDL imports Delicious JSON feed into WAS
Interoperability of CDL Services
DLF Community

Tools

• WAS
  – UC libraries, NYU, Stanford
• Archive-It
  – Columbia, Rice, Indiana...
• Other
  – Library of Congress, North Texas, Los Alamos

Common issues/developments

• Memento
• Policy
  – Can / should Facebook block government pages from capture?
• Collection
• Collaboration
• Imagination
In-depth demos: Nov 8\textsuperscript{th} and 16\textsuperscript{th}
Workflows discussion: Nov 5\textsuperscript{th}

- [http://www.facebook.com/webarchiving](http://www.facebook.com/webarchiving)
- [http://was.cdlib.org](http://was.cdlib.org) > WAS News