Capturing our State:
The California State Government Web Archive

Infopeople Webinar
April 24, 2017

Archive of the California Government Domain, CA.gov
Collected by: University of California Libraries
Archived since: Apr, 2015
Description: This archive preserves access to hundreds of California state agency sites. State agencies utilize their websites to publish everything from press releases, agendas, minutes, events, reports and statistics. This material is especially volatile as leadership changes or as time sensitive issues are no longer on agendas or in the news. The archive is maintained by government information specialists and web curators across several UC campuses, the Stanford University Libraries and the California State Library.
Subject: Government - US States, Politics & Elections, Government

This work is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc/4.0/.
Credits

Kris Kasianovitz, Government Information Librarian
Stanford University Libraries
krisk11@stanford.edu

Shari Laster, Government Information Librarian
UC Santa Barbara Library
slaster@ucsb.edu

Julie Lefevre, Digital Services Librarian
Institute of Governmental Studies Library
UC Berkeley
jlefevre@library.berkeley.edu

Lucia Orlando, Arts, Humanities, Social Sciences
and Government Information Librarian
UC Santa Cruz Library
luciao@ucsc.edu
This Infopeople presentation was supported in whole or in part by the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian.

The opinions expressed herein do not necessarily reflect the position or policy of the U.S. Institute of Museum and Library Services or the California State Library, and no official endorsement by the U.S. Institute of Museum and Library Services or the California State Library should be inferred.
Agenda

• Why CA.gov?

• Introduction to web archiving

• CA.gov web archive project overview

• Building your own archive
Outcomes

• Understand the basics of web archiving using Archive-It

• Be familiar with the scope and content of the CA.gov Web Archive

• Be able to search the CA.gov Web Archive and the WayBack Machine for state government information

• Know who to contact with questions about the project or suggest sites for inclusion in the archive

• Have ideas for web harvesting projects for your local government agencies
CA.gov archive: a collaborative effort!

- Stephen Abrams, CDL
- Julie Lefevre, UCB Institute for Governmental Studies Library
- Kris Kasianovitz, Stanford Library
- Shari Laster, UCSB Library
- Lucia Orlando, UCSC Library
- Bill Riddle, California State Library
- Mike Smith, UCSD Library
- Joseph Yue, UCLA Library
- Michael McNeil, California State Archives
Web archiving

- Goals & principles
- Internet Archive & Archive-It
- Web crawlers
Web archiving basics

Web archiving is the process of:

- collecting portions of the World Wide Web,
- preserving the collections in an archival format, and then
- serving the content for access and use.

Capture is usually accomplished with a crawler like Heritrix, which saves web pages and also follows links to additional pages.

Replay is the means of serving content for access and use: the Internet Archive’s Wayback Machine, Archive-It Collections site, and Memento are good examples.
TOOLS AND SOFTWARE

In the perspective of setting up a Web archiving chain, the following tools are recommended and used by members of the IIPC:

HTTrack WEBSITE COPIER
Free software offline browser

GNU Operating System
Sponsored by the Free Software Foundation

GNU Wget

Introduction to GNU Wget

GNU Wget is a free software package for retrieving files using HTTP, HTTPS and FTP, the most widely-used Internet protocols. It is a non-interactive commandline tool, so it may easily be called from scripts, cron jobs, terminals without X-Windows support, etc.
Crawling the web

- A web crawler is a software script that fetches, analyzes, and files information from web servers at many times the speed of a human.
- The crawler starts with a list of URLs to visit, called the seeds.
- A seed can be a top level domain, like http://www.ca.gov or a sub-directory below the top-level domain, like http://www.ca.gov/Agencies/Secretary-of-State/.
- “Scoping” rules constrain the web crawl to ensure we collect what we want, and nothing more.
Project overview

• Scope of project
• Work involved
• Using our content
CA.gov archive: background

- 2007 Collection established in California Digital Library’s Web Archiving Service (WAS)
- 2010–2012 Minimal development
- 2013 Cooperative project of the University of California/Stanford University Government Information Librarians, with representation from California State Library
- 2015 Content migrated to Archive-It (4.2Tb of data)
- 2016 LSTA grant application
  - Grant goals:
    - comprehensive capture of the State of California web domain
    - develop quality assurance workflows
    - determine a sustainable funding model
Workflow

Diagram adapted from "Automating the Web Archiving QA Process" http://toddstoffer.github.io/presentation/Archive-It-Partner-2016/
Gathering our seeds

A different kind of collection development:

- Identified and reviewed URLs from official state sources.
- Not all seeds can be captured as entered: "Site not found" errors, robots.txt can prevent capture.
- Many official state sites are missing: we add content as we find it.

Following an initial test crawl, we reviewed the data, corrected some errors, and completed our first snapshot (1.9Tb) in December 2016.

Future crawls using Archive-It will de-duplicate against the content we have collected.
Managing our collection

Distributing & shifting project work is a learning process!

- Check seed reports, contact website administrators to request permission to crawl blocked content
- Quality assurance on captures, conduct patch crawls
- Troubleshoot problems, request assistance from Archive-It to solve issues
- Communicating process and progress within the group & broader communities
Our collection

This year so far:

- 413 active seeds (URLs)
- 7,598,972 documents -- 2.7Tb

This is a big collection by Archive-It standards (though not for the Internet Archive!).
Archive of the California Government Domain, CA.gov

Collected by: University of California Libraries

Archived since: Apr, 2015

Description: This archive preserves access to hundreds of California state agency sites. State agencies utilize their websites to publish everything from press releases, agendas, minutes, events, reports and statistics. This material is especially volatile as leadership changes or as time sensitive issues are no longer on agendas or in the news. The archive is maintained by government information specialists and web curators across several UC campuses, the Stanford University Libraries, the California State Library, and the California State Archives.

Subject: Government - US States, Politics & Elections, Government

Collector: California Digital Library, California State Library, California State Archives, University of California Libraries, Stanford Univeristy Libraries

Narrow Your Results

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here
Title: Acupuncture Board
URL: http://www.acupuncture.ca.gov/
Captured 16 times between Oct 24, 2008 and Dec 20, 2016

Title: Arbitration Certification Program - CA Department of Consumer Affairs
URL: http://www.dca.ca.gov/acp/
Captured 15 times between Oct 26, 2008 and Dec 23, 2016

Title: Area VI Developmental Disabilities Board
URL: http://www.areaboard6.ca.gov/
Captured 8 times between Oct 24, 2008 and Jun 13, 2013

Title: Arts Council
URL: http://www.cac.ca.gov/
Captured 26 times between Oct 24, 2008 and Dec 21, 2016
Videos: 14 Videos Captured

Title: Assembly Democratic Caucus
URL: http://asmdc.org/
Captured 29 times between Feb 1, 2010 and Dec 20, 2016
<table>
<thead>
<tr>
<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>10 pages</td>
<td>3 pages</td>
<td>4 pages</td>
<td>2 pages</td>
<td>0 pages</td>
<td>2 pages</td>
<td>3 pages</td>
<td>0 pages</td>
<td>2 pages</td>
<td></td>
</tr>
</tbody>
</table>

* denotes when page was updated
Welcome

GRANTEE SPOTLIGHT

MALASHOCK DANCE COMPANY
Artists in Schools Grantee

STEM + Arts = STEAM. In this #MyCreativeCA video, watch San Diego’s Malashock Dance Company teach students math through movement at a local

YUBA SUTTER ARTS
Veterans Initiative in the Arts Grantee

In Marysville, artists are working with veterans to create a large-scale, public art installation that honors the military experience. The third video on our #MyCreativeCA series.
Build your own collections!

- Examples
- Using Archive-It
- Considerations
Examples of Local Government Information Web Archives using Archive-It

ABAG was created by local governments to meet their planning research needs related to land use, environmental and water resource protection, disaster resilience, energy efficiency and hazardous waste mitigation, and to provide risk management, financial services and staff training to local counties, cities and towns.

The Unified Command for the Refugio oil spill response advises drivers be closure, will be in effect from July 6 to 9, in the southbound direction of U.S. Route 101, California from 8 a.m. - 6 p.m. due to heavy equipment op
Getting started

Advantages to subscribing to Archive-It:

- The Internet Archive handles the technical aspects of capture, preservation, and playback
- You control what's collected, including frequency & depth of capture
- Content can be exported to include in a digital archive
- Community of Archive-It partners working on projects

You can contribute in other ways, too:
http://blog.archive.org/2017/01/25/see-something-save-something/
Archive... with a plan!

Determine the scope of your collection based on your user communities.

Once you have set up & tested your crawl, schedule it to run on regular intervals.

Keep an eye on your space: Archive-It de-duplicates content, but some projects may require waiting until the next subscription year to complete an initial capture.

Be ready for the unexpected!
Get in Touch

Have questions about the CA.gov web archive?

Want to suggest site(s) for inclusion?

Let us know!

email: cagovarchive@lists.berkeley.edu
THANK YOU!

Discussion and Questions