Capturing our State: The California State Government Web Archive

An Infopeople Hosted Webinar
April 24, 2017

Credits

Kris Kasianovits, 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

LSTA Grant Funded Project
(thank you & disclaimer)

This project and presentation at Best Practices Exchange 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.

Infopeople, a grant project of the Califa Group, is supported 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.
Capturing our State: The California State Government Web Archive

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

Infopeople, a grant project of the Califa Group, is supported 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.
Capturing our State: The California State Government Web Archive

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.

Infopeople, a grant project of the Califa Group, is supported 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.
Infopoeple, a grant project of the Califa Group, is supported 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.
Capturing our State: The California State Government Web Archive

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

CA.gov archive: background

https://archive-it.org/collections/5763

Workflow

- Identify Seeds
- Test Crawl
- QA
- Production Crawl
- QA
- Make Publicly Available
- Technical Support
- Documentation

Infopeople, a grant project of the Califa Group, is supported 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.
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 (1.9 Tb) 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)
7,596,972 documents — 2.7 Tb
This is a big collection by Archive-It standards [though not for the Internet Archive].
Infopeople, a grant project of the Califa Group, is supported 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.
Infopeople, a grant project of the Califa Group, is supported 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.
Capturing our State: The California State Government Web Archive

Build your own collections!

- Examples
- Using Archive-It
- Considerations

Examples of Local Government Information Web Archives using Archive-It

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/

Infopeople, a grant project of the Califa Group, is supported 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.
Capturing our State: The California State Government Web Archive

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

Infopeople, a grant project of the Califa Group, is supported 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.
Infopeople, a grant project of the Califa Group, is supported 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.