Crowdsourcing Descriptive Metadata for Web Archives

The CA.gov Archive
CA.gov Web Archive Project and Sprint Team

Sprint Team*

Kris Kasianovitz, Stanford
(Shari Laster, UC SB)
Julie Lefevre, UCB IGS
Lucia Orlando, UC SC
Kathryn Stine, CDL

*subject specialist/curators. NEITHER techies NOR metadata specialists
Access & Use Issue - No metadata

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

- 700+ Seeds
- Small Team
- Project Goal
Process (~ 6 months)

- Plan
- Recruit
- Train
- Deploy
- Clean
- Launch
<table>
<thead>
<tr>
<th>url</th>
<th>Coverage</th>
<th>Description</th>
<th>Creator</th>
<th>Subject1</th>
<th>Subject2</th>
<th>Subject3</th>
<th>Title</th>
<th>Language</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://ad38.senate.ca.gov/">http://ad38.senate.ca.gov/</a></td>
<td>34th District - Anaheim, Buena Park, Fountain Valley, Fullerton, Glendora, Sierra Madre</td>
<td>The Official Website for California Senate District 34</td>
<td>California State Senate Majority Caucus</td>
<td>California, Legislature, Senate</td>
<td>California, Legislature</td>
<td>Legislators</td>
<td>Senator Lou Correa, Serving California Senate District 34</td>
<td>Chinese, Korean, Spanish, Vietnamese</td>
<td>Change to <a href="http://nguyen.csang.us/">http://nguyen.csang.us/</a></td>
</tr>
<tr>
<td><a href="http://www.boe.ca.gov/">http://www.boe.ca.gov/</a></td>
<td>California</td>
<td>Official website of The Board of Equalization (BOE). The BOE administers tax programs concentrated in four general areas: sales and use taxes, property taxes,</td>
<td>California, State Board of Equalization</td>
<td>California, State Board of Equalization</td>
<td>Taxation</td>
<td></td>
<td>California State Board of Equalization</td>
<td>Chinese, Vietnamese, Korean, Spanish</td>
<td></td>
</tr>
</tbody>
</table>

**Sort By:** Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

**Title:** California Commission for Jobs and Economic Growth
**URL:** http://4cajobs.com/
Captured 9 times between Oct 26, 2008 and Sep 28, 2011

**Title:** Assemblymember Scott Wilk, California Assembly District 38
**URL:** http://ad38.asmrc.org/
Description: The official website for California Assembly District 38
Captured 6 times between Oct 21, 2014 and Dec 18, 2017
Subject: Legislature California, State Assemblymembers
Creator: California State Assembly
Coverage: Los Angeles (Portions), Santa Clarita, Simi Valley
Title: Variation of Endosulfan Residues in Water and Sediment Taken from the Moss Landing Drainage of Monterey County

URL: http://www.cdpr.ca.gov/docs/empm/pubs/ehapreps/eh8702.pdf

Description: Data from previous studies conducted by the State Mussel Watch Program (SMW) implied increasing endosulfan residues in mollusks used as indicator organisms of chemical contamination in the Moss Landing Drainage area of Monterey County. As a first step in confirming a chronological trend, studies were conducted by the California Departments of Food and Agriculture (CDFA) and Fish and Game (CDFG) to determine within-site variability of endosulfan concentrations, and to estimate sample size necessary for future research. Personnel from the Environmental Hazards Assessment Program (EHAP) of CDFA collected sediment and water samples while personnel from the Pesticide Investigations Unit (PIU) of CDFG collected mussel and fish samples at the same sites in a coordinated effort. This report contains results from the EHAP study.

Captured on Jul 27, 2010

Creator: California Environmental Protection Agency Environmental Hazards Assessment Program

Language: English

Coverage: California

Note2: deactivate crawl

Max crawl seconds: 3600

Seed type: Historical

Future crawl note: pdf only

Scope: page

Note: Historical Seed

Site ID: sw1e46h717

Robots honored: TRUE

Title: Assemblymember, California State Assembly District 29 (Democrat)

URL: http://asmdc.org/members/a29

Description: The official website for California Assembly District 29

Captured 15 times between Jun 13, 2013 and Aug 22, 2018

Subject: California. Legislature. Assembly

Group: California Legislature

Creator: California State Assembly Democratic Caucus

Language: English

Coverage: California, Assembly District 29, parts of Monterey, Santa Clara, and Santa Cruz Counties, Carmel, Capitola, Del Rey Oaks, Marina, Monterey, Pacific Grove, San Jose (portion), Sand City, Santa Cruz, Scotts Valley, Seaside

Max crawl seconds: 129600

Subject2: California. Legislature

Seed type: Current

Subject3: Local Government

Scope: host+1

Robots honored: TRUE

Site ID: sw15d8qt4b
Some Stats...

- 150 chunks distributed (out of a total of 190)

- 121/150 completed; 81% completion rate
  - 15 people - 2 chunks
  - 5 people - 3 chunks
  - 1 person - 4 chunks
  - 2 people - 5 chunks

- Pre-sprint, only about 15% of all archive websites included Creator and Subject terms.
Lessoned learned

- Raised awareness and engaged with the CA library community (to increase use of collection?)

- Using low tech tools good for Sprinters; challenging for Team to clean

- Surfaced previously unknown QA issues

- Make changes to project management and
WHAT IS THE CA.GOV WEB ARCHIVE?

It is a publicly-accessible web archive of California government agency websites curated by project partners from the UC Libraries, University of California Digital Library, Stanford University, California State Library and California State Archives.

Originally started in 2007, we received a grant from the Institute of Museum and Library Services and the California State Library in 2017 to do a comprehensive crawl and devote time and planning to OA work.

SPRINT SCOPE

• Metadata Sprint subgroup managed various aspects of the work
  • Archive it bulk metadata wrangling (export and upload)
  • Workflow design, testing, and management
  • Metadata creation documentation
  • Volunteer training, communication and support
  • To keep the work manageable, we decided to focus on 7 Metadata fields: Coverage (geographic), Description, Creator, Subject (up to 3), Title, Language, Notes
  • Chosen to improve discovery and help characterize strengths of the collection
  • Creator and Subject terms drawn from controlled, authoritative terms (e.g., FAST, California State Agency Names)
  • Description and Titles drawn from archived website content
  • Notes provided a space for sprinters to share feedback on the content and process which will inform future crawls and metadata approaches
  • We used familiar and accessible tools, editing the master spreadsheet of URL-specific metadata in Google Sheets and then dividing these into 190 “chunks” of 4 URLs, each assigned to a sprinter.
  • Sprinters worked in Google Sheets and in some cases came back for more!

WHY A METADATA SPRINT

• Top Project Goal for FY17-18: enrich descriptive metadata for the CA.gov Web Archive.
  • Small project team was faced with 700 items that needed metadata – we needed assistance.
  • Wanted to complete the work within a certain timeframe - so a short “sprint” with clear start/finish aligned with our goals.
  • Formatted itself very well to virtual participation
  • Wanted participants to feel that they were able to make it to the finish line.
  • Inspired by the Mozilla Global Sprint https://mozilla.github.io/global-sprint
  • We felt this would engage the community to help build a better community resource.
  • Sprint Dates: December 6 through 11, 2017 (Training Session: December 5)

VOLUNTEER SPRINTERS

Volunteer Recruitment

• Email to key California Library Listservs (Calix, CalDoc-L) as well as the Society of American Archivists and a direct email to San Jose State University iSchool students - didn’t have to use Social Media!

• Metadata Sprint Website, that clearly articulated goals and expectations of volunteers
  • No previous cataloging or metadata experience necessary
  • Attend or watch training session.
  • Commit to spend 1 to 2 hours entering metadata for 8 URLS
  • Be willing to communicate and ask questions!
  • Used Google Forms to manage volunteer sign-ups
  • 117 Sprinters, from across the information landscape (6 came from out of state including 1 from Canada; total includes the project team).

Volunteer Training

• Held one 1 hour online training session via Zoom
• Provided all volunteers a link to recording and all reference materials
• Based on our workflow testing, “walked” people through the process:
  • Showed Google sheet metadata chunks and how to use the Metadata Creation Tools to fill in the fields

WHY A METADATA SPRINT

• Top Project Goal for FY17-18: enrich descriptive metadata for the CA.gov Web Archive.
  • Small project team was faced with 700 items that needed metadata – we needed assistance.
  • Wanted to complete the work within a certain timeframe - so a short “sprint” with clear start/finish aligned with our goals.
  • Formatted itself very well to virtual participation
  • Wanted participants to feel that they were able to make it to the finish line.
  • Inspired by the Mozilla Global Sprint https://mozilla.github.io/global-sprint
  • We felt this would engage the community to help build a better community resource.
  • Sprint Dates: December 6 through 11, 2017 (Training Session: December 5)

OUTCOMES

• 150 chunks distributed (out of a total of 190)
  • 121/150 completed; 81% completion rate
  • 15 people did 2 chunks
  • 5 people did 3 chunks
  • 1 person did 4 chunks
  • 2 people did 5 chunks
  • Prior to the sprint, only about 15% of all archive websites in the CA.gov collection included Creator and Subject terms. Following the week-long sprint, at least 85% of all archived websites in the CA.gov collection include Creator and Subject terms!
  • Enriched metadata will enable better discovery and use of the collection.
  • Number of volunteers far exceeded our expectations. Possible explanations:
    • Skill building - working with web archives and metadata offered sprinters a opportunity to learn something new or apply skills in the web archives context.
    • Interested in and concerned about access/disappearance of government information.
    • Interested in and want to support the collection and discovery of CA state government.

LESSES LEARNED

• Using easily available and low tech tools was good especially to “train up” sprinters; however we found that there was more work on the back end for us especially when knitting the all 190 chunks files back together.
• We used OpenRefine to further tidy up the data, and are evaluating whether we pursue other tools to automate work.
• Sprinters helped raise QA issues that we hadn’t previously encountered; we had many eyes on the details of the collection.
• One person managing all of the chunk distribution and questions – would rethink that process.
• Communication loop - person sending out the assignments will be the point person; we had 190 people for 190 chunks. So a lot for one person to manage. Completion rate 80%; project team finished any remaining chunks.
• Consider need for refreshing this metadata (e.g., following election cycles, as new elected officials take office/ownership over websites).
• Impressive how well everyone adhered to using the tools and following the instructions. SPRINTERS did an AMAZING JOB!

MATERIALS

• CA.gov Web Archive https://archive-it.org/collections/5763
• Metadata Sprint Website http://guides.lib.berkeley.edu/ca-gov-sprint
• Training webinar Recording and Slides https://goo.gl/z/2ZBie
• Guidelines for Enhancing Descriptive Metadata for the CA.gov collection https://goo.gl/2VMMAg
• Contact us - cagovarchive@lists.berkeley.edu
Thank You!

Kris Kasianovitz, Stanford University
Julie Lefevre, University of California Berkeley
Lucia Orlando, University of California Santa Barbara
Kathryn Stine, California Digital Library

Contact Us:
cagovarchive@lists.berkeley.edu