



IIPC 2018 Web Archiving Conference
Thursday, Nov. 15 Lightning Talk

Crowdsourcing Descriptive Metadata for Web Archives

The CA.gov Archive

CA.gov Web Archive Project and Sprint Team



University of California
CDL
California Digital Library



California
STATE LIBRARY
FOUNDED 1850
PRESERVING OUR HERITAGE, SHAPING OUR FUTURE



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



STANFORD UNIVERSITY LIBRARIES

Access & Use Issue - No metadata

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



HOME | EXPLORE | LEARN MORE | CONTACT US

The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive



Explore >> [University of California Libraries](#) >> [Archive of the California Government Domain, CA.gov](#)



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 University Libraries](#)

Narrow Your Results

Group Sort By: [Count](#) | [\(A-Z\)](#)

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.

700+ Seeds
Small Team
Project Goal

Process (~ 6 months)

Plan

Recruit

Train

Deploy

Clean

Launch



url	Coverage	Description	Creator	Subject	Subject2	Subject3	Title	Language	Note
http://sd34.senate.ca.gov/	34th District - Anaheim, Buena Park, Fountain Valley, Fullerton, Garden Grove,	The Official Website for California Senate District 34	California State Senate Majority Caucus	California. Legislature. Senate	California. Legislature	Legislators	Senator Lou Correa, Serving California Senate District 34	Chinese, Korean, Spanish, Vietnamese	Change to http://nguyen.cssrc.us/
http://www.boe.ca.gov/	California	Official website of The Board of Equalization (BOE). The BOE administers tax programs concentrated in four general areas: sales and use taxes, property taxes,	California. State Board of Equalization	California. State Board of Equalization	Taxation		California State Board of Equalization	Chinese, Vietnamese, Korean, Spanish	

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](#)

Metadata Enhancements

Marina, Monterey, Pacific Grove, San Jose (portion), Sand City, Santa Cruz, Scotts Valley, Seaside (1)
California, Assembly District 30, San Benito County, portions of Monterey, Santa Clara, and Santa Cruz Counties, Gilroy, Gonzales, Greenfield, Hollister, King City, Morgan Hill, Salinas, San Juan Batista, Soledad, Watsonville (1)
California-Alhambra, California- Arcadia, California-El Monte, California-Monterey Park, California- Rosemead, California-San Gabriel, California-San Marino, California-Temple City and portions of Montebello, and South El Monte (1)
San Gabriel Valley, Alhambra, Monterey Park, San Gabriel, South San Gabriel, Rosemead, El Monte, South El Monte, Baldwin Park, Irwindale, Industry, Avocado Heights, La Puente, Valinda, West Covina, Vincent, Azusa, Citrus, Covina, Temple City, Arcadia (1)
Santa Cruz County, Monterey County, San Luis Obispo County, Santa Cruz, Morgan Hill, Gilroy, Watsonville, Prunedale, Marina, Castroville, Monterey, Carmel, Seaside, Pacific Grove, Big Sur, Carmel Valley, Lockwood, Salinas, San Ardo, Cambria, Harmony, San Miguel, Cayucos, Morro Bay, Paso Robles, Templeton,

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/emppm/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 mollusk and fish samples at the same sites in a coordinated effort. This report contains results from the EHAP study.

Captured once 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: sw1s46h717

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

Metadata Enhancements

Some Stats...

150 chunks distributed (out of a total of 1

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 C/ 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 QA work.



SPRINT SCOPE

- Metadata Sprint subgroup managed various aspects of the work
 - Archivist 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

Field	Source	Notes	Notes	Notes	Notes	Notes	Notes	Notes	Notes
Agency Name	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names
Agency Address	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names
Agency Phone	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names
Agency Fax	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names
Agency Email	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names
Agency Website	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names
Agency Type	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names
Agency Status	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names
Agency Location	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names
Agency Contact	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names
Agency Description	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names
Agency Notes	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names	California State Agency Names

- We used familiar and accessible tools, editing the master spreadsheet in Google Sheets
- Sprinters worked in Google Sheets and in some cases came back for more!

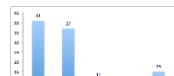
WHY A METADATA SPRINT

- Top Project Goal for FY2017: enrich descriptive metadata for the CA.gov Web Archive.
- Small project team was faced with 700 items that needed volunteer assistance.
- Wanted to complete the work within a certain timeframe: start/finish aligned with our goals.
- Format lent 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 13, 2017 (Training Session: December 5)

VOLUNTEER SPRINTERS

Volunteer Recruitment

- Email to key California Library Listservs (Calix) as well as the Society of American Archivists and a direct email to San Jose State University iSchool students
- 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 signups
- 117 Sprinters, from across the information landscape (6 came from out of state including 1 from Canada: total includes the project team).



Volunteer Training

- Had one 1 hour online training session recorded.
- Provided all volunteers a link to recording and all reference materials

Used Google Sheet metadata chunks and how to use the Metadata Creation Tools to fill in the fields

Browser Workspace Setup - CA Agency Names



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 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

LESSONS LEARNED

- Found that there was more work on the back end for us especially when knitting the all chunks files back together
- We used OpenRefine to further tidy up the data, and are evaluating whether we pursue tools to automate work
- the details of the collection.
- One person managing all of the chunk distribution and quality control is not a process.
- Communication is key: one person sending out the assignments will be the point person: we had chunks. So a lot for one person to manage. Complete 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.org/collections/5763>
- Metadata Sprint Website: <https://guides.lib.berkeley.edu/ca-sprint>
- Training Webinar Recording and Slides: <https://goo.gl/u27Eje>
- Guidelines for Enhancing Descriptive Metadata for the CA.gov Collection: <https://www.cdlib.org/ucrl/jwm/IMJ>
- Contact Us: govarchive@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

