Turning a Marathon into a Sprint: Metadata Creation for 700+ websites of the CA.gov Web Archive of California Government Documents

Kris Kasianovitz, Stanford University; Julie Lefevre, UC Berkeley; Lucia Orlando, UC Santa Cruz; Kathryn Stine, California Digital Library

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:
  - 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.
- Tasks/work lent itself very well to virtual participation.
- Inspired by the Mozilla Global Sprint: https://mozilla.github.io/global-sprint
- Wanted participants to feel that they were able to make it to the finish line.
- We felt this would engage the community to help build a better community resource.
- Sprint Dates: December 6 - 13, 2017 (Training Session: December 5).

VOLUNTEER SPINRTERS

Volunteer Recruitment

- Email sent to key California Library Listservs (Calix, CalDoc-1) as well as the Society of American Archivists and a direct email to San Jose State University School students.
- 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 contribute to an email list.
  - 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, which was recorded.
- 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.

BROWSER WORKSPACE SETUP - CA AGENCY NAMES

<table>
<thead>
<tr>
<th>Column</th>
<th>Agency Name</th>
<th>County</th>
<th>City</th>
<th>URL Path</th>
<th>URL</th>
<th>Language</th>
<th>Title</th>
<th>Creator</th>
<th>Subject 1</th>
<th>Subject 2</th>
<th>Subject 3</th>
<th>Tags</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agency Name</td>
<td>California Department of Corrections</td>
<td>County</td>
<td>City</td>
<td>URL Path</td>
<td>URL</td>
<td>Language</td>
<td>Title</td>
<td>Creator</td>
<td>Subject 1</td>
<td>Subject 2</td>
<td>Subject 3</td>
<td>Tags</td>
<td>Notes</td>
</tr>
<tr>
<td>Title</td>
<td>California Department of Corrections</td>
<td>County</td>
<td>City</td>
<td>URL Path</td>
<td>URL</td>
<td>Language</td>
<td>Title</td>
<td>Creator</td>
<td>Subject 1</td>
<td>Subject 2</td>
<td>Subject 3</td>
<td>Tags</td>
<td>Notes</td>
</tr>
</tbody>
</table>

- Tags opened on browser: 1. CA.gov Web Archive 2. FAST (Faceted Application of Subject Terminology) for subject selection: 3. GNS (UGIS Geographic Names Information System) for place name selection: 4. Authority list of California State Agencies for name selection: 5. LCMF (Library of Congress Name Authority File) for name selection

OUTCOMES

- 150 chunks distributed (out of a total of 190)
  - 121/150 completed; 81% completion rate. Project team worked on any uncompleted chunks.
  - 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 information.

LESONS 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 not accounted for all details of the collection.
- One person managing all of the chunk distribution and questions -- we think that process.
- Communication loop - person sending out the assignments will be the point person; we had 190 chunks. So a lot for one person to manage. Would try to better utilize metadata sprint listerv.
- Considering the need for refreshing this metadata (e.g., following election cycles, as new elected officials take office/ownership over websites) – we will probably use this method again.
- Impressive how well everyone adhered to using the tools and following the instructions. SPINRTERS 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/j27B1e
- Guidelines for Enhancing Descriptive Metadata for the CA.gov collection: https://goo.gl/k2MMfI
- Contact Us: cagovarchive@lists.berkeley.edu