Collaborative Collection Development for Web Archives
#cobwebarchive

Stephen Abrams
Kathryn Stine
Jim Vanderveen
California Digital Library

Peter Broadwell
Andrew Wallace

Lindsay Deal
Janet Taylor
Ann Whiteside
Harvard University

UCLA
CDL and Web Archiving

CDL has been involved in web archiving since 2005

- NDIIPP Web-at-Risk
- Web Archiving Service (WAS)
- End-of-Term Archive
- CA.gov Archive
  [https://archive-it.org/collections/5763](https://archive-it.org/collections/5763)

The 2015 decommissioning of WAS in favor of Archive-It decentralized web archiving activities at the University of California, so we are particularly interested in ways of promoting collaboration

So …
Imagine ...

... a fast-moving event unfolding online as much as on the ground

How can we respond as a community appropriately and responsibly?
Imagine …

Harnessing the domain knowledge and technical capabilities of the entire *community*

Enabling *local* collection development decisions based on *global* information

*Complementary, cooperative, and collaborative* collecting

Institutional participation at a level *commensurate* with local expertise and capacity

Increasing scholarly and public *awareness*
Establishment of thematic collecting *projects*

Open *nomination* of topical seed URLs by interested stakeholders

*Claiming* of seeds by archival institutions intending to harvest

*Holdings* records for seeds actually harvested

Thematic *discovery* of web archives of interest

Centralized *catalog* of collection- and seed-level metadata
Why cobWeb?

The demands of archiving the web in comprehensive breadth or thematic depth *exceeds* the technical and financial capacity of any one institution.

Curators cannot make rational collection development *decisions* without knowledge of what others have collected or intent to collect.

Relevant seed URLs can be meaningfully *contributed* by various stakeholders: curators, archivists, subject area specialists, scholars, journalists, event participants, and the public.

Apportioning collection responsibility into granular pieces encourages *participation* by smaller institutions and programs.
CA.gov archive of the California state government domain

- Collaborative effort of government information librarians at UC and Stanford and the CA State Library and CA State Archive
- More than 700 seeds (URLs) captured total
- Collecting 2007-present, major content refresh 2016-2017
- Coordinating workflows, approaches to QA and metadata
Peter Broadwell at UCLA was well into collecting “fake” news sites before it occurred to him to wonder if anyone else was doing something similar.
Peter Broadwell at UCLA was well into collecting “fake” news sites before it occurred to him to wonder if anyone else was doing something similar.

There was; Mark Graham at IA
Collaborative Web Archiving Projects

- UC Libraries, Stanford, CA State Library, CA State Archives
  - CA.gov
- Ivy Plus Libraries Web Resources Collection Program
  - CAUSEWAY, Contemporary Composers
- New York Art Resources Consortium (NYARC)
- IIPC Content Development Group
  - Olympics, Global News
- Internet Archive, LoC, Stanford, CDL, et al.
  - End of Term
- Canadian Web Archiving Coalition (CWAC)
- Preservation of Electronic Government Information (PEGI)
One-year collaborative project between CDL, Harvard University, and UCLA, funded by IMLS #LG-70-16-0093-16

Public online service hosted at CDL

Targeting initial production release in conjunction with the November 2018 IIPC General Assembly and Web Archiving Conference

https://cdlib.org/services/cobweb/
https://github.com/CobwebOrg/cobweb
<table>
<thead>
<tr>
<th>Activity</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Develop user personas and user stories, user data gathering</td>
<td>August – October 2017</td>
</tr>
<tr>
<td>Establish functional requirements, refine metadata specifications</td>
<td>August 2017 – January 2018</td>
</tr>
<tr>
<td>Develop core functionality, import sample holdings metadata</td>
<td>September – November 2017</td>
</tr>
<tr>
<td>Draft workflows and wireframes</td>
<td>October – November 2017</td>
</tr>
<tr>
<td>First prototype rollout</td>
<td>November 2017</td>
</tr>
<tr>
<td>User testing, data gathering</td>
<td>December 2017 – July 2018</td>
</tr>
<tr>
<td>Iterative development on prototype refinements, continue importing holdings metadata</td>
<td>December 2017 – July 2018</td>
</tr>
<tr>
<td>Prototype rollouts</td>
<td>May 2018, July 2018</td>
</tr>
<tr>
<td>Beta testing</td>
<td>April - September 2018</td>
</tr>
<tr>
<td>Production system rollout</td>
<td>September/October 2018</td>
</tr>
<tr>
<td>Sustainability planning</td>
<td>July – October 2018</td>
</tr>
</tbody>
</table>
Conferences/Professional Associations:
  Archive-It Partner Meeting, AoIR, Dodging the Memory Hole, ARLIS/NA, UC-DLFx, CNI, SAA (Web Archiving Section), DLF, IIPC

Platforms/Communities:
  Internet Archive’s Archive-It, Rhizome’s webrecorder.io

Collaborative Web Archiving Projects we’ve been talking to:
  - UC Libraries, Stanford, CA State Library, CA State Archives
    - CA.gov
  - Ivy Plus Libraries Web Resources Collection Program
    - CAUSEWAY, Contemporary Composers
  - New York Art Resources Consortium (NYARC)
The Team
Project and Outreach Management       CDL
Development and Technical Oversight    UCLA
UI/UX Development and Testing         Harvard

Team Intersections
- Data model
- Functional requirements
- Metadata flows, requirements, specs
- User data, feedback
- Wireframe review
- Prototype review and walkthroughs
Our Process
Agile approach (two-week sprints), iterative and collaborative approaches:
- collecting user data
- translating user data -> workflows -> functionality/interfaces
- confirming functional reqs
- sprint prep/retros

Collaboration Tools
- bi-weekly Zoom calls
- Slack for day-to-day check ins
- Google suite for doc drafting and sharing
- GitHub for source control, planning boards, and issue tracking
cobWeb technical stack

Docker Swarm

µwsgi App Server

Python 3.6
Django 2.0

PostgreSQL DB

Solr Search Index

Docker Swarm

Archives
Current:
- Archive-It

Future:
- Webrecorder
- &c

Agile, sprint-based development
MIT software license

https://github.com/CobwebOrg/cobweb
establish/update *Project*

suggest *Nomination*

assert *Claim*

report *Holding*
Cobweb functionality - related external activities

- establish/update Project
- suggest Nomination
- assert Claim
- report Holding

(Cobweb functions)

STORAGE/ACCESS
DESCRIPTION

establish/grow Collection

reference to Collection
cobWeb functional user roles

(Cobweb functions)

establish/update Project

suggest Nomination

assert Claim

report Holding

STORAGE/ACCESS

DESCRIPTION

establish/grow Collection

PROJECT ORGANIZERS

NOMINATORS

COLLECTORS

participate in our web archive collecting project: <link to Cobweb project>!
cobWeb functional user roles

(Cobweb functions)

- establish/update Project
- suggest Nomination
- assert Claim
- report Holding

STORAGE/ACCESS

DESCRIPTION

establish/grow Collection

PROJECT ORGANIZERS

NOMINATORS

we’ve got expertise and insight to share with this project!

COLLECTORS

(all figure graphics created by Gan Khoon Lay from Noun Project)
# cobWeb functionality - addressing current workflows

<table>
<thead>
<tr>
<th>pre-Cobweb collaboration approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collecting Scope</td>
</tr>
<tr>
<td>- negotiated through ad-hoc communication channels</td>
</tr>
<tr>
<td>- established without good visibility into others’ collecting activities</td>
</tr>
<tr>
<td>- shared publicly on independent websites or nomination recruitment forms</td>
</tr>
</tbody>
</table>
### Project

**Project notes**

- **Janet Taylor, 2017/10/10**
  - Confirmed starting crawl date for description.
- **Janet Taylor, 2017/10/11**
  - This project is partially funded through a grant from the John Quincy Adams foundation. We are looking to partner with another organization to complete the project.

### Projects

- Open nominations project
- Milton, Massachusetts
- Mount Holyoke College
- Haguen, New York
- Dodging the memory hole project
- Propaganda, disinformation, parody, dismissal: Making sense of "Fake News".
- Zita Praha
- Zita Praha

- Recusandae quo aspernatur temporibus.
- Accusantium aperiam laborum temporibus.
- Accusantium aperiam laborum voluptates.
- Paritur vitae sequi.
- Accusantium aperiam laborum voluptates.
- Temporibus perspicilactis aspernatur laborum totam totam voluptates.
- Nostrum fugit at cum sapiente.
- Ducimus sed vel adipisci numquam.
- Modi idit voluptates dignissimos.
- FAKE News!!
- Organization Project to Organize
Propaganda, disinformation, parody, dismissal: Making sense of "Fake News".

Status: Active

Administrators:
- Janet Taylor
- Peter Brookes
- ultraman tigre
- Stephen Adams

Description:
This project captures and categorizes online sources that purport to provide news coverage of current events.

Keywords:
- disinformation
- online news
- fake news
- look in news

Nomination policy:
Open

Nominations:
- http://10percentfeedback.com/
- http://4weemaps.com/1234
- http://1ststoryteller.com/1234
- http://3newsflash.com/1234
- http://21wpr.com/
- http://380usanews.com/
- http://4theconversation.com/
- http://5news.wordpress.com/
- http://6news.wordpress.com/
- http://72.213.129.208
- http://aangfigueres.blogspot.co.uk/
- http://abcnetwork.com/
- http://abovetopsecret.com/
- http://abrilhno/
- http://aveaphile.wordpress.com/
- http://dramatic-man.com/
- http://activistpost.com/
- http://theadvocate.com/
- http://aheadofthecrowd.com/
Propaganda, disinformation, parody, dismissal: Making sense of "Fake News".

Status: Active

Description: This project captures and categorizes online sources that purport to provide news coverage of current events.

Keywords: misinformation, online news, fake news, bias in news

Nomination policy: Open

Nominations:
- http://10percentfoodup.com/ [claim]
- http://16x9mpo.com/ [claim]
- http://3ất4t.com/ [claim]
- http://4jarmyurl.com/ [claim]
- http://6ironwheat.com/ [claim]
- http://7iwyp.com/ [claim]
- http://880.sanews.com/ [claim]
- http://94revolutionarywar.wordpress.com/ [claim]
- http://10wnews.wordpress.com/ [claim]
- http://122.219.129.28/ [claim]
- http://1234fan.blogspot.co.uk/ [claim]
- http://13borwnes.com/ [claim]
- http://14norwagio.com/ [claim]
- http://15dengers.net/ [claim]
- http://16bsek.com/ [claim]
- http://17brilunny.com/ [claim]
- http://18studman.wordpress.com/ [claim]
- http://20actings-man.com/ [claim]
- http://21activatson.com/ [claim]
- http://22buildicopartersong.com/ [claim]
- http://23addictinginfo.org/ [claim]
- http://24ufflag.org/media/ [claim]
- http://26covocate.com/ [claim]
- http://27aheadoftheherd.com/ [claim]
Nomination

Project notes

Resource: http://

Project: Propaganda, disinformation, parody, dismissal: Making sense of “F 1

Status: Unclaimed

Description

Keywords

Janet Taylor, 2017/10/10
Confirmed starting crawl date for description.

Janet Taylor, 2017/10/11
This project is partially funded through a grant from the John Quincy Adams Foundation. We are looking to partner with another organization to complete the project.
Propaganda, disinformation, parody, dismissal: Making sense of "Fake News".

Status: Active

Description: This project captures and categorizes online sources that purport to provide news coverage of current events.

Keywords: disinformation, online news, fake news, base in noise

Nomination policy: Open

Nominators:

- Janet Taylor
- Peter Broadway
- Stephen Aitken
- Kathryn Stine

Nominations:
- http://101percentfudup.com/
- http://16xmp.com/
- http://1st.co.uk/
- http://4knovafish.com/
- http://5twpn.com/
- http://560.anwews.com/
- http://6threvolutionswar.wordpress.com/
- http://7news.wordpress.com/
- http://82.231.123.28/
- http://9cafe.blogspot.co.uk/
- http://abornews.com/
- http://abornwagga.com/
- http://acb1dangor.net/
- http://acconiswhispers.com/
- http://adivirono.com/
- http://acufishman.wordpress.com/
- http://acting-man.com/
- http://activeapost.com/
- http://ahaidachimpanzees.com/
- http://addictinginfo.org/
- http://afflog.org/media/
- http://alobonchronicles.com/
- http://allvocate.com/
- http://aheadoftheherd.com/

Page 1 of 40
Claim

Project notes

Janet Taylor: 2017/10/18
Confirmed starting crawl data for description.

Janet Taylor: 2017/10/11
This project is partially funded through a grant from the John Quincy Adams foundation. We are looking to partner with another organization to complete the project.

Nomination*


Collection*

--------

Description


Keywords
http://newsmutiny.com/

Nomination
In Propaganda, disinformation, parody, dismissal: Making sense of "Fake News".
Keywords:
• has

Holding
In Fake News
Identifier:

Project notes

Janet Taylor; 2017/10/18
Confirmed start date for description.

Janet Taylor; 2017/10/11
This project is partially funded through a grant from the John Quincy Adams foundation. We are looking to partner with another organization to complete the project.
Resource

Project notes

Cancel  Add

Janet Taylor; 2017/10/18
Confirmed starting crawl date for description.

Janet Taylor; 2017/10/11
This project is partially funded through a grant from the John Guy and Stars Foundation. We are looking to partner with another organization to complete the project.

Holdings

http://folklore.gmu.edu/

Nominate to another project.

http://folklore.gmu.edu/

Open Folklore Web Archive

Mason's Folklore Studies Program is a leading center for the study of vernacular cultures and the historical, cultural and social context in which these cultures are rooted. The program has a distinguished 30-year history that offers students of all levels opportunities to study the broad range of human expression including narrative, dance, material culture, verbal art, cultural performances, foodways, and sense of place. Our diverse course offerings provide theoretical and historical grounding in the study of folklore with a strong emphasis in writing and ethnographic research methods.

Type:
Web Archive

Creator:
George Mason University

Subject:
Post-Secondary Education, Dance, Arts, Performance, Foodways, Storytelling

Publisher:
George Mason University

Identifier:
http://wayback.archive-it.org/2843;"http://folklore.gmu.edu/"
Cobweb is a tool for collecting communities...
cobWeb **our next steps**

- Keep learning about collaborative web archive collecting projects and workflows
  - April: Interviews about workflows (focus on roles and collaboration)
  - May: Task-based prototype testing
- Functional requirements -> wireframing -> next prototype development
- Responding to user testing findings
- Considering how Cobweb fits with other tools and workflow steps
- Focusing on metadata inputs, outputs
Cobweb is a tool for collecting communities and perhaps not just those collecting web resources...
establish/update *Project*

suggest *Nomination*

assert *Claim*

report *Holding*
Cobweb affordances for:

- Establishing collecting projects
  - collection development?
  - collection management?
- Nominating resources for...
- Claiming intentions for nominated resources
- Reporting holdings
Institute for Museum and Library Services (IMLS) #LG-70-16-0093-16

Cobweb Advisory Board:

- Jefferson Bailey, Internet Archive
- Andrea Goethals, National Library of New Zealand
- Michael Nelson, Old Dominion University
- Mark Phillips, University of North Texas
- Alex Thurman, Columbia University
- Herbert Van de Sompel, Los Alamos National Lab

Martin Klein, Los Alamos National Lab

Rosalie Lack, UCLA
Collaborative Collection Development for Web Archives
#cobwebarchive

https://cdlib.org/services/cobweb/
https://github.com/CobwebOrg/cobweb

Kathryn Stine, Cobweb Outreach Manager
Kathryn.Stine@ucop.edu

California Digital Library https://www.cdlib.org/
Harvard University Library http://library.harvard.edu/
UCLA Libraries http://www.library.ucla.edu/