Research Ideas and Outcomes 2: e8760
doi: 10.3897/rio.2.e8760

OPEN ACCESS

Small Grant Proposal

Cobweb: A Collaborative Collection Development Platform for Web Archiving

Stephen Abrams‡, Andrea Goethals§, Martin Klein|, Rosalie Lack‡

‡ California Digital Library, University of California Curation Center, Oakland, United States of America
§ Harvard University, Cambridge, United States of America
| University of California, Los Angeles, United States of America

Corresponding author: Rosalie Lack (rosalie.lack@ucop.edu)

Reviewable v1

Received: 06 Apr 2016 | Published: 12 Apr 2016

Citation: Abrams S, Goethals A, Klein M, Lack R (2016) Cobweb: A Collaborative Collection Development Platform for Web Archiving. Research Ideas and Outcomes 2: e8760. doi: 10.3897/rio.2.e8760

Abstract

The California Digital Library, Harvard Library, and UCLA Library seek $243,765 to develop Cobweb - a lightweight open-source collaborative collection development platform supporting the creation of comprehensive web archives by coordinating the independent activities of the web archiving community.

Keywords

web archives, collection development, digital library

Our vision for the future

> *Radical collaboration to support the national digital platform emerged as a priority for libraries, archives, museums, and allied institutions*– *IMLS Focus Summary Report: The National Digital Platform* (2015)
>
> *We must, indeed, all hang together, or most assuredly we shall all hang separately* – Benjamin Franklin

Imagine a fast-moving news event, such as the Arab Spring, unfolding online via news reports, videos, blogs, and social media. Recognizing the importance of recording this event, a curator immediately creates a new Cobweb project and issues an open call for

nominations of relevant web sites. Scholars, subject area specialists, interested members of the public, and event participants themselves quickly respond, contributing to a site list that is more comprehensive than could be created by any one curator or institution. Archiving institutions review the site list and publicly claim responsibility for capturing portions of it that are consistent with local collection development policies and technical capacities. After capture, the institutions' holdings information is updated in Cobweb to disclose the various collections containing newly available content. By distributing the responsibility, more content is captured more quickly with less overall effort than would otherwise be possible.

Current challenges and related initiatives

The demands of archiving the web in comprehensive breadth or thematic depth exceed the technical and financial capacity of any single institution. This gives greater impetus to the desirability of community-based cooperation, which is dependent on automated support for facilitating coordination of distributed responsibilities. However, as identified in a recent environmental scan by the Harvard Library, there currently are no effective means for curators or researchers to know what is or is not being captured and archived by others, resulting in "duplication or gaps in coverage and siloed collections." Even the Internet Archive (IA) currently supports search by known URL only. This means that IA "will allow you to find a needle in a haystack, but only if you already know approximately where the needle is." The Memento protocol is another initiative that aids discovery, but again, only if a desired URL is known in advance.

The International Internet Preservation Consortium (IIPC), of which all three project partners are members, has tried collaborative collecting relying on a nomination tool from the University of North Texas (UNT) and other ad hoc methods such as spreadsheets and email. While a valuable resource, the UNT tool supports nomination only and does not support other critical collecting activities; in particular, it has no mechanisms for indicating either an institution's collecting intentions or its actual holdings. Archive-It (AIT), IA's subscription service, has been used often for cross-institutional projects, however, IA does not have the legal, managerial, or technical infrastructure to support large-scale, cross-institutional collecting, especially when the collaborating institutions do not already have formal AIT agreements.

Cobweb - a new collaborative collection development platform

While there are a number of tools that address some aspects of the collaborative collection development problem, they do not form a single integrated system as is envisioned with the Cobweb platform. As a centralized catalog of aggregated collection- and seed-level descriptive metadata, Cobweb will enable a range of desirable collaborative, coordinated, and complementary collecting activities by supporting three key functions: nominating, claiming, and holdings. The nomination function will let curators and stakeholders suggest

web sites pertinent to specific thematic areas and provide seed-level descriptive metadata; the claiming function will allow archival programs to indicate an intention to capture some subset of nominated sites; and the holdings function will allow programs to document captured sites along with their collection-level description, structural and temporal scope, preservation policies, and terms of use. Cobweb will leverage existing tools and sources of archival information, exploiting, for example, the APIs being developed for AIT to retrieve holdings information for over 3,500 collections from 350 institutions.

The platform will further IMLS's efforts towards developing a national digital platform for managing our digital heritage, helping libraries and archives make better informed decisions regarding the allocation of their resources, and promoting effective institutional collaboration and sharing. It also addresses IMLS's strategic goals by facilitating learning through more effective discovery, and ultimate use, of relevant content; permitting libraries and archives to be more responsive to the needs of their constituencies by letting them scale their efforts to their capabilities; and increasing the overall efficiency of collaborative solutions to common problems.

Partners and stakeholders

The Cobweb project is a partnership of the CDL (PI), Harvard Library, and UCLA Library, which have extensive expertise in web archiving, digital library infrastructure and services, collection development policy, and software development. An external advisory board will review and provide input throughout the project. The partners also will work in consultation with an informal but engaged stakeholder group for input and feedback to an iterative development process. Stakeholders include the IIPC, IA/AIT, Library of Congress, George Washington University Libraries, MIT, the New York Art Resources Consortium, Old Dominion University, Stanford University Libraries, UNT, and others interested in adopting and contributing to the platform.

Project plan, performance goals and outcomes

This one-year Cobweb project will produce an open source collaborative collection development system along with relevant policies, guidelines, and best practices to support usage and encourage adoption. The partners will employ an agile development process featuring requirements- and test-driven development with frequent iterative sprints. The platform will be hosted by the CDL and initialized with collection metadata from the partners and its stakeholder group. A mid-year release will be shared with the global web archiving community at the April 2017 IIPC General Assembly to further gather feedback and discuss ongoing sustainability. Significant outreach efforts, including public webinars and workshops, will be focused on the creation of an engaged user community and garnering support for post-grant sustainability.

Budget

This project has a total cost of $422,428 ($243,765 grant funds; $178,423 voluntary 73% cost-share). The total budget is allocated for salaries ($200,718) and fringe benefits ($81,531) for the PI, outreach manager and system administrator at CDL, and through subawards, a technical manager and developer at UCLA, and UI/UX designers and testers at Harvard. Other costs include travel ($10,000), servers and databases ($6,240), and indirect costs ($123,938).

Funding program

U.S. IMLS National Leadership Grants for Libraries

https://www.imls.gov/grants/available/national-leadership-grants-libraries