University of California, California Digital Library (CDL) - Cobweb

Abstract

The Web is the preeminent medium for the political, cultural, educational, and commercial discourse of contemporary life. A full understanding of our time will not be accessible to the future, even the near future, unless proactive steps are taken now to ensure the timely acquisition and careful stewardship of our collective online heritage. Through its sheer scale, however, the demands of archiving the Web, whether in comprehensive breadth or thematic depth, exceed the capacity of any single institution. This gives greater impetus to community-based cooperation, which is dependent on the efficacy with which distributed archival responsibilities can be coordinated. Unfortunately, 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 potential duplication or gaps in coverage and siloed collections. The Cobweb project will develop an open-source collaborative collection development platform for creating comprehensive Web archives by coordinating the activities of the Web archiving community.

Cobweb is a one-year collaborative project led by the California Digital Library with Harvard Library and UCLA Library. The Cobweb platform will support three key functions: *nomination*, *claiming*, and *holdings*, all performed in the context of thematic collecting projects. The nomination function will allow any interested stakeholder to suggest URLs germane to a project theme. The claiming function will allow archival programs to indicate an intention to capture some subset of the nominated URLs. The holdings function will allow those programs to specify what material has actually been captured, along with the location and terms under which it is accessible. Note that Cobweb is a centralized aggregated catalog of web collection *metadata*; the actual web content remains hosted and accessible by the individual institutions that have captured it.

The project consists of four coordinated and complementary activities: developing the open source Cobweb system; user interface design and usability testing; community outreach and engagement; and production deployment. The performance goals are to engage the community of users to test, give feedback, and use Cobweb; ensure that Cobweb is functional and easy to use; and to lay the foundation for sustaining Cobweb after the project ends. The shared curatorial decision-making enabled by Cobweb will benefit the libraries and archives already engaged in Web archiving as well as those that have not yet started. It also lets researchers and other users more easily discover archived websites of topical interest, and participate directly in deciding what should be collected.

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

Narrative: Cobweb: A Collaborative Collection Development Platform for Web Archiving

1 Statement of Need

The Web is now the preeminent medium for the political, cultural, educational, and commercial discourse of contemporary life. A full understanding of our time will not be accessible to the future, even the near future, unless proactive steps are taken now to ensure the timely acquisition and careful stewardship of our collective online heritage. Through its sheer scale, however, the demands of archiving the Web, whether in comprehensive breadth or thematic depth, exceed the capacity of any single institution. This gives greater impetus to the desirability of community-based cooperation, which is dependent on the efficacy with which distributed archival responsibilities can be coordinated. Unfortunately, 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."[1] The Cobweb project will develop an open-source collaborative collection development platform for creating comprehensive Web archives by coordinating the independent activities of the broad Web archiving community.

Effective collaborative collection development is dependent on three conditions: (1) quickly developing a comprehensive set of URLs to use as the starting point for collecting in a given thematic area; (2) a way to apportion out the work of collecting those URLs; and (3) a way to see what content-collecting institutions actually hold and make accessible. With these conditions in place, individual Web archiving programs can make rational decisions regarding their own activities, cognizant of how their local efforts will complement and leverage the efforts of others. Towards this end, the Cobweb platform will support three key functions: *nomination*, *claiming*, and *holdings*, all performed in the context of thematic collecting projects. The nomination function will allow any interested stakeholder to suggest URLs germane to a project theme, along with relevant collection- and site-level descriptive metadata. The claiming function will allow archival programs to indicate an intention to capture some subset of the nominated URLs, along with capture and provenance metadata. The holdings function will allow those programs to specify what material has actually been captured, along with metadata related to temporal and spatial scope and the location and terms under which it is accessible. Note that Cobweb is an aggregated catalog of web collection *metadata*; the actual web content remains hosted and accessible by the individual institutions that have captured it.

The US cultural memory community has worked together on collaborative collection projects, most notably, the End of Term archive documenting the .gov federal domain during the 2008 and 2012 presidential term transitions.[2] While a common system, the University of North Texas's nomination

[1] Gail Truman, *Web Archiving Environment Scan*, Harvard Library, 2016. http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314
[2] http://eotarchive.cdlib.org/

tool,[3] was used to manage the URL list, the distribution and coordination of subsequent partner activities relied on email, spreadsheets, and other ad hoc methods that, while ultimately successful, could gain substantially in efficiency. With the support of UNT, Cobweb addresses this issue by extending its functionality beyond that of the UNT tool to support claiming and holdings as well as nomination, and in doing so, enabling a range of desirable collaborative, coordinated, and complementary collecting activities. While the claiming and holdings features could have been provided in their own system alongside the existing UNT nomination tool, having Cobweb natively support all three functions permits more streamlined operation of the interactions necessary for effective collaboration.

The notion of crowdsourcing deeply informs the Cobweb project. Permitting public nomination, by subject area specialists as well as other interested stakeholders, is the quickest way to build up a comprehensive set of thematic URLs. Similarly, public claiming of URLs encourages widespread participation by letting individual archival programs accept responsibility for a level of activity that is commensurate with their goals and capabilities, while still contributing to global archival capacity. Centralizing the coordination of the crowd in a common platform increases the transparency and efficiency of collective action.

In addition to the benefits Cobweb provides for collection development, it also promotes better discovery of archived content. The Cobweb holdings feature will present an aggregated catalog of collection- and site-level metadata that can be used by researchers and other consumers looking for content on a thematic topic. The Internet Archive (IA), the largest single source of archived Web content, currently provides search only by known URL.[4] The Memento protocol is also concerned with the discovery of archival holdings, but again, only for known URLs. IA's Archive-It service does permit keyword searching and browsing of collection metadata, but only for those collections managed by Archive-It. Cobweb is not restricted in this sense; it will permit searching of aggregate holdings information from any willing archival source, thus providing much better high-level discovery.

Cobweb furthers IMLS's efforts towards developing a national digital platform for managing our digital heritage by helping libraries and archives make better informed decisions regarding resource allocation, and promoting effective institutional collaboration and sharing. Cobweb also addresses IMLS's strategic goals by facilitating learning through more effective discovery and use of relevant content; as well as 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.

[3] http://digital2.library.unt.edu/nomination/

[4] Internet Archive will be adding support for keyword searching for website home pages. While this will facilitate future searching for IA content, Cobweb will aggregate IA as well as other archival sources for more inclusive discovery.

2 Impact

Much of the technical infrastructure for collecting the Web already exists; Cobweb adds the missing essential piece to help libraries and archives decide *what* to collect. Cobweb provides public transparency for websites that will be or already have been collected, giving curators the knowledge they need to build archival collections that fill gaps and complement the collections of other institutions, leading to more efficient use of resources and more comprehensive collections for researchers. The shared curatorial decision-making enabled by Cobweb will benefit the libraries and archives already engaged in Web archiving as well as those that have not yet started. It also lets researchers and other users more easily discover archived websites of topical interest, and participate directly in deciding what should be collected.

Impact on Web Archiving Libraries and Archives

Cobweb will enable libraries and archives to make informed decisions about the websites to collect, allowing them to invest their staff and financial and technical resources where they can have the most impact by focusing either on unique collecting areas or supplementing, rather than unknowingly duplicating, other institutions' collections. It will also allow them to participate in multi-institutional collaborations by providing public visibility of calls to start new thematic collecting initiatives.

Impact on Libraries and Archives New to Web Archiving

Cobweb will lower the barrier to joining the Web archiving community, increasing the overall number and diversity of institutions contributing to building collaborative Web archive collections. Institutions will be able to take inexpensive, incremental steps towards contributing to collaborative efforts, starting, for example, by nominating websites related to a thematic topic or event. As institutions gain experience with the concepts and collecting of websites, they could choose to increase their level of participation in collaborative efforts by, for example, accepting responsibility to collect a portion of the websites nominated by Cobweb users for a large topical collection. The online Cobweb service will include a wiki dedicated to training and feedback that will allow newly-participating institutions to learn from others already experienced in Web archiving, broadening the community's knowledge and experience and encouraging newcomers to make useful contributions to Web archiving collaborations.

Impact on Researchers and Other Archive Users

Given the one-year timeframe of this project, the primary focus is on Cobweb adoption by collection managers. However, there are very real benefits provided by Cobweb to researchers and other consumers once institutions populate it with holdings of collection- and site-level descriptive metadata. Researchers will have a central location to search across the distributed collections of all participating institutions to locate the Web archives that match their topical interests, something that is impossible to do now. Because Cobweb will be open to nominations for collections,

researchers also will be able to communicate directly with selectors concerning websites that should be captured due to their research value. Because nominations will be transparent and contributed by any individual or institution with expertise or interest in the topic or event, the resulting Web archive collections will be more comprehensive than if a single institution tried to collect on its own, making them more valuable for research and teaching.

Measuring Success

The success of Cobweb hinges on the extent to which institutions already active in Web archiving will use it. Based on feedback gathered from Web archiving practitioners and service providers through direct conversations, an interactive presentation at the 2016 IIPC General Assembly, and an environmental scan conducted by Harvard Library, it is clear that Cobweb adoption is dependent on the ease with which Cobweb users can build site lists to crawl, discover what has already been crawled on a topical area, and import holdings metadata. The project has been designed with these institutional adoption and usability requirements at its core. The Outreach Manager will engage with the community to understand use cases, determine requirements, test prototypes, and contribute holdings metadata. Harvard Library is contributing the staff expertise and equipment of its User Research Center for usability testing.

The progress of the project and the Cobweb tool will be measured by performance indicators in the following areas: evidence of deep engagement with the community of users; evidence that Cobweb is easy to use; evidence that Cobweb is functionally suitable for the purposes for which it was designed; and evidence that the use of Cobweb can be sustained after the project. Project performance indicators and targets are shown in the table below.

Success Area	Performance Indicators	Project Targets
Deep engagement with the community who would use it	1. Presentations at representative conferences 2. Number of Web archiving curators testing Cobweb 3. Number of Archive-It partner institutions giving feedback	1. 5 conferences 2. 20 curators 3. 25 Archive-It partners
Easy to use	4. Ability to perform key tasks as measured by usability tests 5. User satisfaction	4. 90% success for last prototype tested 5. 90% positive satisfaction for last prototype tested
Functionally suitable	6. Conformance with functional	6. 100% conformance with

Success Area	Performance Indicators	Project Targets
for the purposes for which it was designed	requirements 7. Integrated with key systems and services	"must have" requirements 7. Uses prominent existing APIs (e.g., Memento, Archive-It)
Use can be sustained after the project	8. Awareness of Cobweb (citations/references in non-project presentations, etc.) 9. Deployable open-source application and documentation 10. Number of institutions indicating intention to use 11. Identification and initial talks with organizations who might sustain Cobweb post-project	8. 20 mentions 9. Deployed to Github 10. 15 institutions 11. 5 organizations

3 Project Design

The project consists of four coordinated and complementary areas of work, each led by a different project partner: (1) developing the open source Cobweb system (UCLA); (2) user interface design and usability testing (Harvard); (3) community outreach and engagement (CDL); and (4) production deployment (CDL).

Cobweb System Implementation (led by UCLA)

This part of the project will first define the functional and technical requirements related to Cobweb's three core functions: nominating site URLs relevant to a given thematic area, claiming some subset of those URLs to crawl and archive, and providing a collection- and site-level overview of existing archival holdings. Throughout the iterative development process, we will gather user feedback and conduct usability testing (coordinated by Harvard Library) to inform functionality and design.

An essential part of the development process is to leverage existing solutions. The University of North Texas (UNT) nomination tool is an important exemplar for defining the nomination process for thematic site URLs. For the holdings function, existing APIs offered by the Memento project and the Internet Archive/Archive-It will be used for the aggregation of site- and collection-level metadata of archival holdings. Archive-It APIs will be used to obtain descriptive and provenance metadata about Archive-It collections for inclusion in the Cobweb holdings function. Memento APIs

provide TimeMaps (lists of snapshots in all Memento-compliant archives) of individual URLs.[5] This data will enhance the holdings function with information on a per-URL basis. Since Cobweb's holding function is archive-agnostic, we will identify and also work with non-Archive-It institutions to automate the capture their record of holdings. Towards this end, Cobweb will employ both push and pull approaches. For the pull-based approach, a Cobweb "agent" will periodically visit the institution's interface and ask for new information. The push-based approach for pro-actively informing Cobweb about new data may rely on existing protocols such as PubSubHubbub.[6] We will also investigate the use of standardized synchronization frameworks such as ResourceSync[7] for both push and pull notifications.

The Cobweb functional and technical requirements will be informed continuously with feedback from the project's advisory group. The advisory group members will provide invaluable perspectives for Cobweb since they represent content providers, archival organizations, and technology partners. Jefferson Bailey (Internet Archive) will provide guidance on the use of web archiving APIs. Herbert Van de Sompel (Los Alamos National Laboratory) and Michael L. Nelson (Old Dominion University), are editors of the ResourceSync standard specification (together with Cobweb Technical Manager Martin Klein, UCLA). Mark Phillips (University of North Texas) developed the UNT Nomination Tool.

For the implementation and testing of the Cobweb tool the project team will follow a transparent agile development process with agreed upon milestones and progress reports shared with the entire group throughout the project duration.

User Interface Design and Usability Testing (led by Harvard Library)
The goal for this part of the project is to ensure that Cobweb fully meets the needs of its target audience. Activities include testing of the functional requirements, design and testing of the Cobweb user interface, and iterative usability testing of Cobweb prototypes.

Once the functional requirements are developed, a survey will be sent out to potential Cobweb users via relevant mailing lists to verify that the functional requirements are accurate and solicit beta testers nationwide for Cobweb prototypes. The verified functional requirements will be used by the UI Designer to create wireframes, using tools such as Balsamiq, Illustrator, or Photoshop, that specify the layout of content and functional elements of the Cobweb user interface. Feedback on these wireframes will be solicited from the curator beta testers, the advisory group, and via public webinars. The final wireframes will be used by the Programmer to implement the Cobweb user interface. Under the guidance of the UX Oversight Manager, the UX Intern will design a usability test plan that will include the defined stakeholders, testing goals, research questions, testing methodology, equipment and tools, participants and recruiting methods, a test script, tasks to

[5] http://timetravel.mementoweb.org/guide/api/
[6] https://github.com/pubsubhubbub
[7] www.openarchives.org/rs/

perform, and a consent form. The tests will be conducted by the UX Intern using Skype for remote beta testers and Harvard's User Research Center for local beta testers.[8] The UX Intern will communicate all testing results to the entire project team,and the results will be used by the UCLA development team to refine Cobweb prototypes. The usability testing and Cobweb prototype development will continue to iterate throughout the project to ensure that it reflects the needs of the potential users.

Community Outreach and Engagement (Led by CDL)
Broad communication and promotion of Cobweb, and acquiring contributions from content providers are the main goals of this project activity. The primary outreach channels will be conference presentations and webinars. The five main venues planned for project promotion are the Coalition for Networked Information (CNI) Membership Meeting, International Internet Preservation Consortium (IIPC) General Assembly, Society of American Archivists (SAA) Web Archiving Roundtable, Archive-It Partners Meeting, and Association of Internet Researchers (AoIR) Conference. The project team has already introduced Cobweb to the Web archiving community at the 2016 IIPC General Assembly. The feedback was extremely positive and included suggested extensions to Cobweb functionality. All feedback and suggestions will be taken into consideration during design of the functional requirements phase.

Production Deployment (led by CDL)
The Cobweb online production service will be hosted by CDL on an Amazon AWS EC2 virtual machine. The production service will be initialized with holdings information from representative archival collections of the project partners, and other early adopters.

4 Diversity Plan

Cobweb will enable libraries, archives, and other memory institutions of all sizes and variation in resources and technical staff to contribute meaningfully to collaborative Web archiving activities. The Impact section describes how Cobweb will be useful not only to large institutions, many with well-established Web archiving programs, but also to smaller institutions that may be just starting out in this area. In addition, because Cobweb is a centrally-hosted service requiring no local infrastructure or technical staff, any institution with access to the Internet will be able to use it.

5 Project Resources: Personnel, Time, Budget

This one-year project will involve personnel from all three partner institutions. The CDL as the lead institution will take overall responsibility for the grant, the outreach activities, as well as for hosting the technical environment and the final Cobweb deployment. Harvard Library will oversee and conduct user interface/user experience tasks. The UCLA library will be responsible for all technical development.

[8] http://projects.iq.harvard.edu/harvardlibraryux/home

Principal Investigator, California Digital Library (10% time)
Stephen Abrams will be responsible for project vision, administration, and technical oversight. Abrams is Associate Director of the CDL's UC Curation Center with responsibility for strategic planning, policy, conceptual design, and innovation.

Outreach Manager (to be hired), California Digital Library (50% time)
The Outreach Manager will be responsible for communication and engagement with the Web archiving community by keeping them informed of project progress, gathering feedback and input throughput the development cycle, and recruiting Cobweb adopters.

Project Advisor, Harvard Library (2% time)
Andrea Goethals, Manager, Digital Preservation and Repository Services, will coordinate the usability testing at Harvard, contribute to the functional requirements, hire the User Experience intern, work with the Outreach Manager to recruit individuals across the US to participate in the testing, and help publicize the project and tool.

User Interface Designer, Harvard Library (13% time)
Janet Taylor, Usability and Interface Librarian, will design the wireframes and Cobweb UI.

User Experience Intern (to be hired), Harvard Library (312 hours over 9 months)
The intern will design and conduct user research studies to verify the functional requirements, and test the usability of wireframes and Cobweb prototypes.

User Experience Oversight, Harvard Library (2% time)
Amy Deschenes, User Experience Specialist, will supervise the User Experience Intern, and provide training on data collection, analysis, and tools available at the User Research Center.

Technical Project Manager, UCLA Library (50% time)
Martin Klein, Technical Manager, will coordinate and oversee the definition of the functional and technical specifications and the development and testing of the Cobweb tool. He will further contribute to the usability testing as coordinated by the Harvard Library, support the documentation writing process, and help to broadly publicize the project and Cobweb tool.

Programmer (to be hired), UCLA Library (100% time)
The programmer will be responsible for all Cobweb technical development including gathering use cases, generating requirements and specifications, and designing, documenting, and testing the project APIs.

Systems Administrator, California Digital Library (5% time)
Jim Vanderveen will serve as systems administrator to set up and monitor the development, staging, and production environments. Vanderveen is a CDL Senior DevOps Administrator with over 25 years of experience.

Advisory Group

Alex Thurman (Columbia University), Herbert Van de Sompel (Los Alamos National Laboratory), Michael L. Nelson (Old Dominion University), Jefferson Bailey (Internet Archive), and Mark Phillips (University of North Texas), all of whom have deep experience in Web archiving initiatives with synergistic relationships with Cobweb.

Schedule

One year is a realistic timeframe for technical development given that we will be drawing on pre-existing tools and standards. The Schedule of Completion details how we will use the project time. Briefly, outreach, community building, and project promotion will happen throughout the entire project period. Technical work will begin with functional and technical specifications, then move quickly to iterative development and prototypical deployment. The user interface/user experience will also be an iterative process that begins early with designing and testing the wireframes, and then usability testing throughout the development process.

Budget

The project partners are requesting $244,894 in IMLS funds and providing $149,874 in voluntary 38% cost share. The Budget and Budget Justification provides a detailed summary of requested funds and cost sharing for project activities.

6 Communications Plan

Engaging the Web archiving community and gathering feedback throughout the development lifecycle will be critical to Cobweb's success. Engaging the Web archiving community is essential for attracting a critical mass of Cobweb adopters, while gathering feedback will allow us to ensure adoption by creating a tool that is designed at the appropriate level of simplicity and convenience. We will use three approaches to engage the community and gather feedback.

1. First, we will reach target audiences through presentations at professional conferences including CNI, IIPC, SAA, Archive-It Partners meeting, and AoIR. We have already presented on Cobweb at IIPC 2016, where we had a very productive group discussion and gained useful insights from the potential Cobweb user community.

2. Second, we will use community mailing lists and social media platforms. In particular, we will engage with the SAA Web Archiving Roundtable list and the IIPC list. We have also created our own open project list that we will use to encourage participation while keeping people informed and engaged. In addition, we will use blogs and social media to keep the community informed and to solicit feedback.

3. Finally, we will conduct a series of interactive webinars as a forum for project updates, prototype demos, advocacy for engagement, and feedback. The webinars will be recorded to accommodate viewing at a later date

All of these channels will target the following audiences: the Web archiving community (nationally and internationally); Web archiving curators, librarians, and archivists; and Internet researchers. We will also seek to engage other important stakeholders communities, including journalists, policy makers, and interested members of the general public.

The Outreach Manager, working on the project at 50% time, will play the leading role in Cobweb communications, but we intend that all project partners, the advisory group, and key stakeholders will also provide significant input. We will gather metrics on the impact of our communications and continually adjust our strategy to ensure the broadest reach.

Finally, the technical development will be done in an open technical environment. Cobweb will be an open source product, using Github to provide access to the source code and technical documentation.

7 Sustainability

The CDL is committed to hosting Cobweb for a minimum of two years following the conclusion of the funded project. This will give us sufficient time to identify a permanent host. Given that Cobweb's mission extends beyond the institutional level, it is important for Cobweb to reside eventually with a national or international organization. During the grant year, we will engage our advisory group and other key stakeholders to explore viable options for long-term hosting. Widespread adoption of Cobweb by the Web archiving community is essential for near- and long-term success. During the grant year, the Outreach Manager, along with the project partners and the advisory group, will focus on building robust community support for Cobweb.

At the April 2016 IIPC meeting (attended by U.S. and international libraries), we solicited feedback on our Cobweb concept paper. Meeting attendees expressed a great deal of interest and enthusiasm for Cobweb, and feedback confirmed a pressing need for a tool such as Cobweb. Attendees also expressed the importance of ease of use of such a service to ensure adoption. We have a two-pronged approach to ensuring ease of use. First, our User Interface (UI)/User Experience (UX) staff will be an integral part of the production lifecycle. We will work with staff from Harvard's User Research Center to ensure conformance to good design practices and to conduct iterative usability testing throughout the design process. Second, we will continue to solicit community input early on as we finalize Cobweb's functional requirements. We will also share early versions of Cobweb for community feedback throughout the design process. We already plan on getting feedback, for example, at next year's IIPC conference and at Archive-It's partner meeting, as well as by hosting a series of webinars throughout the year.

The code and all documentation will also be open source and available from Github for others to contribute source code. See the Digital Stewardship document for more details.

University of California, California Digital Library (CDL) - Cobweb

Schedule of Completion (One-year project from November 1, 2016 through October 31, 2017)

	2016		2017									
Activity	N	D	J	F	M	A	M	J	J	A	S	O
Setup project website and other communication channels	■											
Conduct outreach, community building and project promotion (e.g. webinars)	■	■	■	■	■	■	■	■	■	■	■	■
Hire and/or identify programmer and outreach manager	■											
Setup development environment	■											
Develop functional requirements and technical specifications	■	■										
Develop Cobweb tool		■	■	■	■	■	■	■	■	■		
Test implementation of multiple prototypes			■	■	■	■	■	■	■	■	■	
Hire and train UX intern		■										
Survey users to verify functional specifications and recruit testers		■										
Refine functional requirements/specifications			■	■	■							
Design and test UI wireframes			■	■								
Refine UI based on testing					■	■						
Prototype usability testing						■	■	■	■	■	■	
Present at conferences: IIPC GA, Archive-It Partner Meeting, SAA Roundtable, AOIR, and CNI						■				■		■
Deploy to production											■	
Write technical documentation									■	■	■	■
Write final report											■	■

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded research, data, software, and other digital products. The assets you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products is not always straightforward. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and best practices that could become quickly outdated. Instead, we ask that you answer a series of questions that address specific aspects of creating and managing digital assets. Your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions
If you propose to create any type of digital product as part of your project, complete this form. We define digital products very broadly. If you are developing anything through the use of information technology (e.g., digital collections, web resources, metadata, software, or data), you should complete this form.

Please indicate which of the following digital products you will create or collect during your project
(Check all that apply):

	Every proposal creating a digital product should complete …	Part I
	If your project will create or collect …	**Then you should complete …**
☑	Digital content	Part II
☑	Software (systems, tools, apps, etc.)	Part III
☐	Dataset	Part IV

PART I.

A. Intellectual Property Rights and Permissions

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the intellectual property status of the content, software, or datasets you intend to create? Who will hold the copyright? Will you assign a Creative Commons license (http://us.creativecommons.org) to the content? If so, which license will it be? If it is software, what open source license will you use (e.g., BSD, GNU, MIT)? Explain and justify your licensing selections.

The metadata content aggregated by the Cobweb platform will be contributed by various institutions who will retain all existing intellectual property rights to their material. However, the contribution to Cobweb will be done under the terms of the Creative Commons Attribution 4.0 (CC-BY) license. The Cobweb system will maintain proper attribution for all contributed material.

The Cobweb software system developed by UCLA as part of the project will be copyrighted by the Regents of the University of California, but released under the terms of a permissive open source license, either GNU GPL or MIT. The final choice of the license is dependent on the inherited open source licenses of other technology solutions that may be incorporated into the Cobweb codebase.

A.2 What ownership rights will your organization assert over the new digital content, software, or datasets and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users about relevant terms or conditions.

Aggregated metadata content managed by the Cobweb platform will be made freely available to the public. This material will be contributed by its owners under the terms of the Creative Commons Attribution 4.0 (CC-BY) license, and the Cobweb system will maintain and expose proper attribution for all content.

The Cobweb software system will be freely available under the terms of a permissive open source license, either GNU GPL or MIT.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

Cobweb is a neutral content aggregator that only collects content that is freely and openly available on the Web and via open APIs. We can not control the nature of the content in advance but will respond promptly and disable access if and when appropriate requests are received.

Part II: Projects Creating or Collecting Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and/or collect, the quantities of each type, and format you will use.

Cobweb does not create any new content but rather aggregates existing metadata, typically obtained in XML or JSON format. We aspire to collect more than one hundred archival collections and Memento TimeMaps for several hundred URIs. TimeMaps are also provided in JSON format.

A.2 List the equipment, software, and supplies that you will use to create the content or the name of the service provider who will perform the work.

The newly developed Cobweb tool will aggregate the metadata using modern web protocols such as ResourceSync. In addition, existing APIs provided by the Internet Archive and the Memento project will be utilized.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to create, along with the relevant information on the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

All aggregated metadata is textual information. Anticipated file formats are: JSON, XML, MODS, and plain text.

B. Digital Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

Utilizing the web resource synchronization framework ResourceSync, we can take advantage of the provided audit mechanism to ensure the authenticity of transferred data. When using public APIs there is no reliable function to ensure the completeness of the obtained data and hence Cobweb deploys a best-effort approach and trusts the providing sources. Each metadata aggregation process, however, will be monitored by automated validation processes that will trigger a message to the system administrator and developer if a process failed in completion.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: You may charge the Federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the Federal award. (See 2 CFR 200.461).

Cobweb source code will be stored in the git version control system both on CDL servers hosted on the Amazon AWS EC2 cloud and publicly on github.com. CDL servers are backed up nightly with a 35 day retention period. Monthly backup snapshots are kept for one year.

Cobweb metadata is managed in mirrored relational databases replicated across multiple availability zones in Amazon's AWS RDS cloud. Databases are backed up nightly (warm) and weekly (cold).

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, or preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, or PREMIS) and metadata content (e.g., thesauri).

Since we will be aggregating metadata from various sources that may rely on divergent local standards, we will crosswalk all metadata to Dublin Core.

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during and after the award period of performance.

Cobweb metadata is managed in mirrored relational databases replicated across multiple availability zones in Amazon's AWS RDS cloud. Databases are backed up nightly (warm) and weekly (cold).

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of digital content created during your project (e.g., an API (Application Programming Interface), contributions to the Digital Public Library of America (DPLA) or other digital platform, or other support to allow batch queries and retrieval of metadata).

By utilizing ResourceSync we will provide metadata for synchronization in a standardized web-centric manner. This means, for example, a content provider can obtain Cobweb's data and in turn provide it to other digital platforms such as DPLA.

D. Access and Use

D.1 Describe how you will make the digital content available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

Cobweb will be openly available and publicly accessible via standard web browsers. In addition, Cobweb provides data discovery, synchronization, and re-use capabilities as described in C.3

D.2 Provide the name and URL(s) (Uniform Resource Locator) for any examples of previous digital collections or content your organization has created.

The CDLprovides public access to the following collections:
-Calisphere https://calisphere.org/
-eScholarship http://escholarship.org/
-Merritt http://merritt.cdlib.org/
-Online Archive of California http://www.oac.cdlib.org/

Part III. Projects Creating Software (systems, tools, apps, etc.)

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) this software will serve.

We will be creating the Cobweb tool that offers three main function:

Nomination of site URLs
Claiming of site URLs for capture
Holdings of archival collections

The audience will be web collection curators and researchers and other users of archival collections.

A.2 List other existing software that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

To the best of our knowledge, there are no other systems that offer the comprehensive function of Cobweb. The University of North Texas's Nomination tool is the exemplar for the Cobweb nomination function, but Cobweb extends its capabilities to include support for site claiming and holdings aggregation.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software (systems, tools, apps, etc.) and explain why you chose them.

Cobweb will be a web-based application and hence the planned programming language and platforms will be standard web technologies. Programming and scripting languages used will be Python, HTML, CSS, JavaScript, and JQuery and the tool will run on a Linux (virtual) machine. Further, the ResourceSync standard will be utilized to implement essential aggregation tasks.

B.2 Describe how the intended software will extend or interoperate with other existing software.

Cobweb will aggregate archival holding information by utilizing APIs offered by the Internet Archive and the Memento project. In addition, it will synchronize with other institutions to obtain holding information that is not available through these standard APIs.

B.3 Describe any underlying additional software or system dependencies necessary to run the new software you will create.

Cobweb is dependent on content contributed by institutions engaged in web archiving efforts and hence can not depend on just one single environment or piece of software. It will be built using standard web technologies and hence is dependent on the internet and connected content providers.

B.4 Describe the processes you will use for development documentation and for maintaining and updating technical documentation for users of the software.

The project team will maintain a private wiki and ticket system for internal communication and code development. This serves for the documentation of the source code and recording of implementation progress and milestones while developing the Cobweb tool. We will use a Github repository to host and share the source code, for versioning, and to facilitate the release of technical documentation. We will further maintain a Google Group to help build a community and foster discussions related to the project and tool.

B.5 Provide the name and URL(s) for examples of any previous software tools or systems your organization has created.

Project partners have created dozens of software tools and systems:
https://github.com/CDLUC3
https://github.com/UCLALibrary

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under an open-source license to maximize access and promote reuse. What ownership rights will your organization assert over the software created, and what conditions will you impose on the access and use of this product? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain any prohibitive terms or conditions of use or access, explain why these terms or conditions are justifiable, and explain how you will notify potential users of the software or system.

The Cobweb software system developed by UCLA as part of the project will be copyrighted by the Regents of the University of California, but released under the terms of a permissive open source license, either GNU GPL or MIT. The final choice of the license is dependent on the inherited open source licenses of other technology solutions that may be incorporated into the Cobweb codebase.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

The software will be hosted in a public Github repository. We will provide a link to the software from our project website and advertise the URI broadly in various presentations.

C.3 Identify where you will be publicly depositing source code for the software developed:

Name of publicly accessible source code repository: Cobweb
URL: https://github.com/CDLUC3/cobweb

Part IV. Projects Creating a Dataset

1. Summarize the intended purpose of this data, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

2. Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

3. Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

4. If you will collect additional documentation such as consent agreements along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

5. What will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

6. What documentation (e.g., data documentation, codebooks, etc.) will you capture or create along with the dataset(s)? Where will the documentation be stored, and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

7. What is the plan for archiving, managing, and disseminating data after the completion of the award-funded project?

8. Identify where you will be publicly depositing dataset(s):

   Name of repository:
   URL:

9. When and how frequently will you review this data management plan? How will the implementation be monitored?

Original Preliminary Proposal

Cobweb: A Collaborative Collection Development Platform for Web Archiving

> *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

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.

OUR VISION FOR THE FUTURE

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."[1] 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."[2] The Memento protocol is another initiative that aids discovery, but again, only if a desired URL is known in advance.[3]

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.

---

1 Gail Truman, *Web Archiving Environment Scan*, Harvard Library, 2016. http://library.harvard.edu/sites/default/files/Harvard_WA_Environmental_Scan_Jan_2016_optimized.pdf

2 Meredith Broussard, "The irony of writing online about digital preservation," *The Atlantic*, November 20, 2015. https://archive.is/ykBt0

3 Herbert Van de Sompel, Michael L. Nelson and Robert Sanderson. *HTTP Framework for Time-Based Access to Resource States - Memento*, December 2013. http://mementoweb.org/guide/rfc/

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 and fringe benefits 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 servers and databases, and indirect costs.