Building Web Archiving Collaborations to Save [More of] the Web

Anna Perricci, Web Archiving Project Librarian, Columbia University Libraries

Columbia University Libraries' web archiving activities began in 2008, following a planning phase supported by the Andrew W. Mellon Foundation in 2007-2008 and was preceded by the ongoing practice of cataloging live websites at the request of selectors and faculty. The collections built through the web resources collection program (web archiving) initially formed six thematic collections, which currently include over 1,200 seed URLs (see https://archive-it.org/organizations/304). Over 650 seed URLs (harvested quarterly) complement CUL's extensive holdings of resources for the study of human rights. The web archives on human rights can be searched and browsed through a custom-built portal, the Human Rights Web Archive (http://hrwa.cul.columbia.edu/).

Following the establishment of the web collecting program at Columbia University, multiple web archiving collaborations are taking shape within four project areas with the support of a three year grant from the Andrew W. Mellon Foundation. Through multi-faceted approaches to collaborative web archiving we are actively engaging with four categories of stakeholders: colleagues at other research libraries, software developers, website creators and users or potential users of web archives. This work has been led by project staff responsible for web archiving and project management with supervision from a steering committee for web collecting at Columbia University Libraries. So far approximately fifty people across eleven institutions have been involved as selectors for pilot collections and six teams have received funds to support the development of better web archiving tools. This short paper will briefly detail the work completed from early 2013 through mid-April 2015 on these collaborative projects then introduce some points for next steps and further discussion.

With wide recognition of the need to cooperate in order to meet the needs of users, models for collaborative collection development in research libraries are currently growing. Complex multi-institutional collaborations that influence CUL's collaborative web collecting model include HathiTrust.[1] The primary investigator for the grant funded projects described in this paper, Robert Wolven, is the Chair of Program Steering Committee for HathiTrust as well as Columbia University's Associate University Librarian for Bibliographic Services and Collection Development. Many lessons learned from HathiTrust's development into a substantial and sustainable resource have influence the planning for a coordinated web resource collecting program, particularly regarding models for support through a consortium of research libraries.

Coordinated collection development of research libraries involved in a partnership known as Ivy Plus or Borrow Direct has taken shape over the past few years with leadership of the Associate University Librarians responsible for collection development at each involved institution. The partners for Ivy Plus include the libraries of: Brown, Columbia, Cornell, Dartmouth, Duke, Johns Hopkins, Harvard, Princeton, and Yale universities, the Massachusetts Institute of Technology (MIT), and the universities of Chicago and Pennsylvania. A current initiative for collaborative collecting involves coordinated purchase of composers' scores with the intent of reducing duplication and thereby expanding the range of composers who can be included in the collections of Ivy Plus libraries. The value of building holdings cooperatively was one factor that led to the formation of the first pilot co-curated web archive, the Contemporary

[1] For more on HathiTrust: http://www.hathitrust.org/about

Composers Web Archive (CCWA). The other two experimental web collections taking shape are the Collaborative Architecture, Urbanism and Sustainability Web Archive (CAUSEWAY) and a group of websites chosen to represent significant resources on climate change and associated discussions. CAUSEWAY has been built into a collection containing over 100 harvested websites whereas the climate change collection has so far only served as a basis for further analysis. Lessons learned from the climate change nominations include processes for engaging selectors from over a dozen subject areas (including various aspects of science, area studies, geospatial data, and social sciences).

The overarching goal of CCWA is to preserve copies of present and future manifestations of the websites of notable contemporary composers in a secure digital archive to guarantee the continuing availability of these extremely important but potentially ephemeral documents for researchers and scholars. So far collecting priorities have been determined based on analysis of data about the current collecting of the scores of contemporary composers (starting with composers whose scores are collected by six or more Ivy Plus partners) and direct nominations from the Ivy Plus music librarians. CCWA is available via Archive-It (https://archive-it.org/collections/4019) and a MAchine Readable Cataloging (MARC) record being created for each seed site in CCWA. The MARC records create a path for access to the archived websites through the CUL's library catalog[2] and WorldCat.[3]

CAUSEWAY is a pilot project to archive websites devoted to the related topics of architecture, urban fabric, community development activism, public space and sustainability. Participating librarians are choosing websites that fit into the themes of CAUSEWAY: Urban Fabric (e.g. historic preservation, urban renewal, urban preservation), Public Space (e.g. parklands, community gardens), or Community Activism (e.g. historic preservation initiatives, associations). Each librarian is making nominations focused on the geographic region in which her or his institution is located. CAUSEWAY is available via Archive-It (https://archive-it.org/collections/4638) and a MARC record is being created for each seed site in CAUSEWAY to facilitate discovery.

In order to increase capacity to collect and provide access to preserved web resources, the web archiving incentive awards program has provided funds to developers working on tools designed to capture, provide access, interpret or empower users to create web archives. With guidance from an external oversight panel, six of grants of $20,000-25,000 each were distributed as sub-awards of the Mellon fund received by CUL in 2013. These software projects are nearing completion at the time of the writing of this paper and results from the work will be shared widely beginning in the summer of 2015.

Creators of web resources are being reached both through requests to website owners for permission to collect their websites, which is a standard feature in CUL's protocol, and direct outreach. Consultations on best practices for site creators have taken place within Columbia University Libraries. Guidelines for any website creator wishing to make more preservable web resources are available on this page: https://library.columbia.edu/bts/web_resources_collection/guidelines_for_preservable_websites.html.

---

[2] http://clio.columbia.edu/

[3] "WorldCat is the world's largest network of library content and services." See: http://www.worldcat.org/whatis/

The fourth project area focuses on current or potential users of web archives. This part of the projects will be covered in depth in Pamela Graham's paper, *Curating the Future, Assessing the Past: The Human Rights Web Archive*, and presentation on the panel on which this paper will be shared (*Curating the Web for Research: Emerging Practices in Libraries and Museums*). This portion of the project includes analysis of URLs cited in articles in journals focusing on scholarship about human rights. A corpus of citations was gathered from all issues of selected journals published in 2010 via web scraping and the citations were sorted using Open Refine such that citations containing URLs were isolated for further analysis. The URLs were initially assessed by support staff to determine if the cited content is still on the page listed. Technical staff devised a way to leverage an API from the Internet Archive and the Solr index for the Human Rights Web Archive to determine if a URL cited is included in either collection of archived websites (this automated link checking is still in progress). In order to develop more specific use cases for web archives scholars who use or might have a strong need for web archives are being interviewed. The results of these interviews and the citation analysis could influence collection development decisions as further seeds sites are selected for the human rights collection.

All of these collaborative efforts have led to advancements though much room for growth remains as we continue work with colleagues at other research libraries, software developers, site creators and users of web archives. The term of the grant supporting these projects will conclude at the end of 2015. Planning for the continuation of the collaborative projects, especially cooperative curation with an associated structure for sharing responsibilities and costs, has steadily remained a high priority. Currently a framework for a shared web archiving program through Ivy Plus is being discussed, including workflows for capturing web content as well as financial and governance elements needed to support this work for years into the future.

Through this work we are identifying points for further discussion as well as charting a course to address needs identified so far. With input from scholars, librarians, archivists, software developers, site creators, administrators and others in web collecting programs worldwide we will leverage shared knowledge and coordinated expertise to capture and preserve more web based content. By saving more of the web in consultation with a variety of stakeholders we strive to meet the existing and emerging needs of users of web archives in the near and distant future.