

NDSR Project: Web Archive Management at the New York Art Resources Consortium (NYARC)

Goal Summary

To develop policies and best practices to support the consortium's web archive collection.

The Resident will master quality assurance (QA) protocols and interview interns, project staff, external web archivists, and Archive-It support specialists to compile an inventory of known QA issues and create an open, centralized online documentation of issues and procedures to support streamlined workflows.

To identify seeds requiring capture by Hanzo Archives, which will be used as an on-demand service for sites that cannot be adequately captured by Archive-It.

The Resident will craft metadata guidelines and input workflows for preservation metadata requirements for stewardship of web archive materials.

The Resident will manage the DuraCloud implementation and produce written documentation to administer the WARC (Web ARChive) archive, noting reporting capabilities and task intervals.

Specific Objectives

To implement sustainable workflows and best practices for NYARC's web archiving program in three specific areas:

- Quality assurance
- Preservation metadata
- Archival storage

Timeframe & Deliverables

Overall — 9 months

Months 1 through 3 — Orientation and Quality Assurance (QA)

Review NYARC web archiving project reports and workflow documentation published from 2010-present. Train in existing Archive-It procedures through direct work with staff and webinars. Begin QA work. Inventory QA issues through direct experience and interviews with web archiving interns, staff, and external partners. Work with Archive-It support specialists to verify categorization of fixable/non-fixable issues. Review The Bentley Historical Web Archives: Guidelines and Procedures (<http://tinyurl.com/mzt8v8e>) as a potential model. Consult with colleagues at Columbia University and the Folger

Shakespeare Library about local QA processes that relate to long-term preservation. Test and establish improved QA workflows that take some of the guesswork out of knowing what to look for, what is fixable, etc. Create QA guidelines to document procedures, categorize Archive-It capture constraints, and identify seed characteristics for Hanzo Archive captures. Attend Archive-It Partner Meeting in Montgomery, Alabama on November 18, 2014.

Deliverables: A detailed inventory of QA issues and guidelines for improved QA workflows hosted on the NYARC wiki (<http://wiki.nyarc.org>). Presentation of QA work at NYARC stakeholder meeting.

Months 4 through 6 — Metadata

Begin research of DuraCloud features. Evaluate preservation metadata requirements (administrative, technical and structural) to support long-term preservation. Survey metadata elements contained in WARC header, the local permissions database, and DuraCloud. Conduct gap analysis and prepare recommendations for best practices. Present survey findings and recommendations to project stakeholders and metadata consultant. Based on feedback, create preservation metadata guidelines and routines.

Deliverables: Presentation of preservation metadata survey and recommendations for the program. Written documentation of preservation metadata guidelines and routines. Submit paper proposal for Art Libraries Society of North (ARLIS/NA) annual conference.

Months 7 through 9 — Archival Storage and DuraCloud

Continue DuraCloud evaluation with focus on management of archived files in cloud storage. Identify tasks necessary to verify digital backups and for ongoing repository management. Implement DuraCloud's Archive-It backup feature and produce procedural document detailing management activities to ensure long-term sustainability of the archive.

Deliverables: Implementation of DuraCloud service. Tested and documented procedures for DuraCloud managed archival storage. Present project outcomes at ARLIS/NA annual conference Fort Worth, Texas (dates to be determined).

Resources Required

2 mentors (Duncan (primary) and Pregill), 1 Resident

Access to key NYARC staff and contacts at partner institutions, specifically Archive-It, DuraCloud, and Columbia University.

Access to Archive-It, DuraCloud, Basecamp project management site, and background documentation on the NYARC web archive.

A workstation with dual monitor set-up to conduct QA work and a phone to participate in conference calls.

Context

Recognizing that publication methods for art research materials, such as auction catalogs and catalogues raisonnés, were shifting from analog to digital, NYARC began to explore web archiving as a collection building strategy. With the support of the Mellon Foundation, the consortium conducted a study in 2012 entitled *Reframing Collections for a Digital Age: A Preparatory Study for Collecting and Preserving Web-based Art Research Material*. This study investigated publication trends, web archiving technologies, infrastructure, and workflows needed to support a NYARC web archive collection. Outcomes from that study indicated that, while the tipping point from print to digital is still on the horizon, our libraries needed to take action or risk developing significant gaps in the art historical record. This exploratory study led to the current two-year Mellon grant (2013-2015) for the NYARC libraries to actively expand its web archive collection and develop workflows for administering that collection. This activity has become mission critical for NYARC.

Web archiving is increasingly being viewed as a mainstream collection development activity by both archives and libraries. The Archive-It service has over 275 partners, of which over 50 are also OCLC Research Library Partnership members. However, many challenges exist to integrate web archiving into library practice in efficient and sustainable ways. This project's tri-part structure is directly targeted at finding implementable solutions for quality assurance, metadata, and archival storage.

In the absence of automated processes, verifying the completeness of a harvested website requires manual and laborious quality assurance tasks. Identifying patterns of capture pitfalls and creating guidelines to streamline processes are local necessities, but this documentation can also assist Archive-It in tool development. The Resident's work in this area will build on lessons learned during the Folger Shakespeare Library's previous NDSR project and ongoing work at Columbia University.

Identifying the adequate level of metadata to support discovery and long-term usability of digital collections is an open question. Metadata creation is often a manual task and not easily repurposed across systems. Decisions surrounding metadata models must strike a balance between available resources and requisite elements to ensure the renderability and authenticity of digital materials. NYARC is contracting with a metadata consultant for a descriptive metadata model for web archive materials. The Resident will advise on preservation metadata requirements to support the collection.

Storing a locally managed copy of web archive files outside of the Internet Archive was a recommended digital preservation strategy that came of the 2012 NYARC study. WARC file management is new territory for many libraries, including the

NYARC libraries. The new DuraCloud service integration with Archive-It will be used as the NYARC storage solution; however management functions for maintaining that storage need to be thoroughly investigated and integrated into current workflows. The Resident's digital preservation expertise will be essential to critically examine and lead the DuraCloud implementation.

In all three areas, this project's deliverables will be shared with the broader web archiving community and the Resident will participate in outreach activities to present outcomes of their investigation. As an active member of the web archiving and art library communities, NYARC is committed to collectively finding solutions to better manage hybrid research library collections. This project expects to demonstrate solutions for managing web archives that can move this endeavor forward.

² For reports on this study see "Readying for Reframing: Reports on Web Archiving," <http://www.nyarc.org/content/readying-reframing-reports-web-archiving>.

Required Knowledge and Skills for Resident

Candidates will have an ALA-accredited master's degree in Library and Information Science or a graduate degree in Computer Science, Information Technology, or associated field. Additional qualifications include:

- Experience within an archives, library, or museum setting.
- Demonstrable organizational, analytical, and problem solving skills.
- Ability to work both independently and collaboratively with other team members in a consortial environment
- Excellent interpersonal skills, with a demonstrated ability to communicate effectively, both orally and in writing.
- Knowledge of basic web technologies, such as HTML, CSS, and CMS.
- Knowledge of metadata standards: Dublin Core, MARC, METS, PREMIS
- Inquisitive nature, sense of humor, and an eye for detail.

Preferred Knowledge or Experience

- Experience with digital preservation, web archiving and/or metadata creation in a research library or comparable environment.
- Knowledge of current trends and emerging technologies in web archiving.
- Knowledge of art history.