Web Archiving at the New York Art Resources Consortium (NYARC): Collaboration to preserve specialist born-digital art resources

Sumitra Duncan

Sumitra Duncan
Head, Web Archiving Program
Frick Art Reference Library
The Frick Collection
10 East 71st Street
New York, NY 10021
duncan@frick.org

Web Archiving at the New York Art Resources Consortium (NYARC): Collaboration to preserve specialist born-digital art resources

Abstract

In late-2013 the New York Art Resources Consortium (NYARC), consisting of the research libraries and archives of three leading art museums in New York City (the Brooklyn Museum, The Frick Art Reference Library of the Frick Collection, and the Museum of Modern Art), implemented a program for web archiving born-digital specialist art and art historical resources. In the three years following the initiation of NYARC's collaborative web archiving program, ten collections of archived art websites have seen steady growth, with particular focus on archiving the websites of the institutions themselves, sites pertaining to New York City galleries, and those of born-digital catalogues raisonnés. NYARC works in partnership with the Internet Archive's Archive-It service to build their web archive collections and harvest web content, as well as partners with those in the galleries, libraries, archives, and museums (GLAM) community to raise awareness about the inherent need for preservation of ephemeral born-digital materials of high risk of disappearance from the live web. NYARC has pioneered the archiving of born-digital catalogue raisonné, with the collection of catalogue raisonné in Archive-It now including 40 discreet projects. NYARC's overall web archive collections now encompass approximately 3,900 sites. While web archiving is still considered an "emerging" area of focus for collection development within libraries and archives, preservation of born-digital materials remains absolutely crucial to future scholarship. Digital art history in particular depends upon the existence and accessibility of these web-native ephemeral materials, with the many use cases for web archive collections by independent scholars only just being considered as big data analysis gains momentum in the humanities. Additionally, given the rapid rate at which citations for web-based materials within scholarly publications suffer from "link rot" or drifting URLs, it is especially pertinent that web archiving becomes a more prominent practice within cultural heritage institutions, universities, and in collaboration with the GLAM community.

This paper seeks to examine the current state of web archiving of born-digital art and art history materials, with particular focus on archiving catalogues raisonnés and New York City gallery sites; NYARC's practice of providing free and public access to their complete web archive collections via the new NYARC Discovery interface; and a discussion of future collaboration for the GLAM community.

Keywords:
Web-archiving; born-digital; collection development; collaboration; GLAM

Introduction

The New York Art Resources Consortium (NYARC) consists of the libraries and archives of three leading art museums in New York City: The Brooklyn Museum, The Frick Collection, and The Museum of Modern Art. NYARC formed in 2006, with the mission to facilitate collaboration that results in enhanced resources to research communities and the shared goals of improving access to art research resources through technology, advancing the scholarly, educational, and cultural enrichment missions of the three museums, and providing leadership in the development of innovative and model information services programs.

The NYARC libraries each have highly specialized collections, and given the unique nature of their consortial holdings, the directors of the NYARC libraries were driven to continue to build collections of specialist art and art historical resources in the born-digital era. NYARC began to explore web archiving as a collection building strategy in 2010, having recognized that publication methods for art research materials, such as auction and exhibition catalogs, and catalogues raisonnés, were drifting from analog to digital formats.

Web archiving is the process of creating collections of URLs, with websites harvested via software tools that crawl the web and capture the digital content as well as the look and feel of sites at a given point in time. These captures are converted to digital files that conform to an ISO standard format called WARC (the Web ARChive file format). WARC files contain the content from a harvested website during a specific capture and web archive collections document the changes that occur in websites over time.

Current State of Web Archiving in the U.S.

Globally, many different types of institutions are engaged in web archiving, the process of collecting, preserving, and enabling access to web-native materials. This includes national

libraries and archives, some with a legal deposit mandate in place. Many university libraries and government offices engage in web archiving, as well as smaller museum libraries and consortia. These institutions engage in web archiving to collect web-native resources in their traditional collecting scope, to fulfill a records retention requirement, to document spontaneous events online, and to combat link rot and content drift. Results from the 2016 NDSA survey, "Web Archiving in the United States," show that the majority of institutions engaged in web archiving in the United States are colleges and universities.

A major player in web archiving (in particular within the United States) is the Internet Archive, which offers the largest publicly available web archive in existence, via the Wayback Machine. As of mid-2017, it contains over 150 billion archived webpages and over 100 million websites in 40 or more languages. Approximately one million URLs are added to the Wayback Machine per week. In contrast, many national libraries and archives engaged in web archiving are not able to allow for public access to their web collections and instead must offer access with restrictions only in an onsite reading room.

Archive-It is the subscription web archiving service of the Internet Archive and it helps organizations in harvesting, building, and preserving collections of born-digital content. Presently they partner with over 500 organizations globally. Partner organizations use a web application to collect, catalog, and manage their collections of archived content and Archive-It offers full-text search of partner collections, with content hosted and stored at the Internet Archive data centers.

Results from the 2016 NDSA survey also identified only three percent of the institutions engaged in web archiving in the U.S. as museums (2%) or consortium (1%). A few Archive-It partners do now collect art and art history related websites, yet there remains a significant gap in collecting of websites in this subject area. Rhizome, the New York-based non-profit arts organization developing the Webrecorder tool, has recently made great strides in successfully harvesting born-digital artwoks and social media art.

Many organizations simply are not engaged in web archiving at all — particuarly the content producers. It's important that the library and archives community not assume that publishers are archiving the content that they produce/disseminate, as most content producers/publishers do not archive their digital output. For NYARC, it's particularly notable that auction houses are not archiving their own catalogs and sales results, and much of their content does not exist on their webpages for more than a few months before being removed. Galleries and artists are often also not archiving their own websites, nor are they aware that there would be a need to do so in support of future art historical scholarship. The lack of a U.S. legal deposit program for born-digital resources is also a challenge, as well as the fact that the majority of cultural institutions do not archive their websites and born-digital output.

Given that an estimated 75% of the material online will change, disappear, or drift from it's initial URL each year, preservation has become a significant challenge for ephemeral

information sources. The NYARC directors knew that their libraries were well positioned to leverage existing consortial resources in a collaborative way to be nimble in this emerging and challenging territory of web archiving. As we identified that publication methods for art research materials were shifting from analog to digital formats, NYARC began working incrementally to address the inherent challenges of the digital transition. In 2010, NYARC began to explore the efficacy of implementing a consortial web archiving program with a pilot study at the Frick Art Reference Library. The pilot study was followed by a year-long exploratory study, funded by a grant from The Andrew W. Mellon Foundation, as it was clear that establishing a sustainable program of web archiving for the NYARC libraries would require a more focused investigation about publishing trends and perceived technical and organizational challenges for building and preserving a web archive collection. In 2013 NYARC received a two-year grant from the Mellon Foundation, in support of implementing our consortial web archiving program. The implementation grant allowed the NYARC libraries to actively expand our web archive collection and develop workflows for administering it. Web archiving as a collection development activity has become highly critical for the consortium and as of mid-2017 we have archived 3.9 TB of data across ten collections with the use of the Archive-It subscription service—all publicly accessible online.

NYARC's Archive-It Collections

NYARC presently has ten collections in Archive-It and we have archived over 3,900 distinct URLs (commonly referred to as seed URLs, as they are the starting point for the web crawler to harvest a site). We are focusing on collecting subject-based websites for auction houses and their embedded catalogs, catalogues raisonnés, artists' websites, the websites of New York City galleries and art dealers, sites devoted to scholarship for the restitution of lost or looted art, and ephemeral art resources of scholarly value. We also devote a great deal of effort to archiving the institutional websites of the Frick Collection, the Brooklyn Museum, The Museum of Modern Art, and our NYARC consortium site and project sites.

We have found that our growing collections of the websites of New York City galleries and art dealers and those of born-digital catalogues raisonnés are highly unique, with no other organizations seeking to comprehensively collect these resources. At this point we are archiving the sites of 160 galleries on a monthly basis. In contacting the gallerists and discussing our interest in archiving their gallery websites, we have begun to raise awareness of the need for such work – and from these conversations it is clear that web archiving is not yet being considered very heavily in the gallery world. Gallerists have been highly supportive of our program and enthusiastic to participate.

NYARC is presently working to archive the sites of nearly 40 born-digital catalogue raisonné projects and we are the only organization that is partnering with the producers and publishers of these born-digital catalogues in order to preserve them for future public access by researchers. We are meeting with both art historians and platform developers behind the online catalogues raisonnés to discuss the technical challenges of fully harvesting these resources, many of which employ the use of a backend database for catalogue entries.

These collections will be of rich value to researchers in the future, particularly those seeking to study the output and representation of specific artists and the evolution of the New York City gallery scene over time. In the relatively short period of time that NYARC has been archiving gallery sites we have already seen a number of galleries close permanently, move to new neighborhoods within the city, and numerous new galleries have opened. Given the extensive number of galleries in New York City, it has been a challenge to scale this collection to cover the many hundreds of sites that would be representative of all New York-based galleries. The gallery world is an area we feel is rich for collaboration with other art libraries via a shared web archiving program with a geographic distribution for stewardship.

Similarly to other existing web archiving initiatives, NYARC's life cycle for our web archiving program begins with collection development, curation, and administration, and spans from harvesting and quality assurance, to preservation, long-term storage, and description and access. Initial steps involving the selection of websites for inclusion, communications with site owners, and especially the harvesting and quality assurance of web crawls, are all quite time-intensive for staff. Quality assurance review has been particularly manual, and thus slow-moving, but we feel it is essential for building representative collections. In dealing with websites about visual art it's especially imporant to capture not only the embedded information, but also the visual appearance and the functionality/behavior of the websites.

NYARC has developed a collaborative collection development policy for websites and we now have a consortial workflow in place for our program. Websites of scholarly value are selected and nominated for inclusion in the consortium's ten curated Archive-It collections, each of which is aligned with the collecting objectives and strengths of the three NYARC libraries. Due to the rapid pace of change in the content, functionality, and features of websites, the collection development policy is reviewed and revised accordingly on a periodic basis.

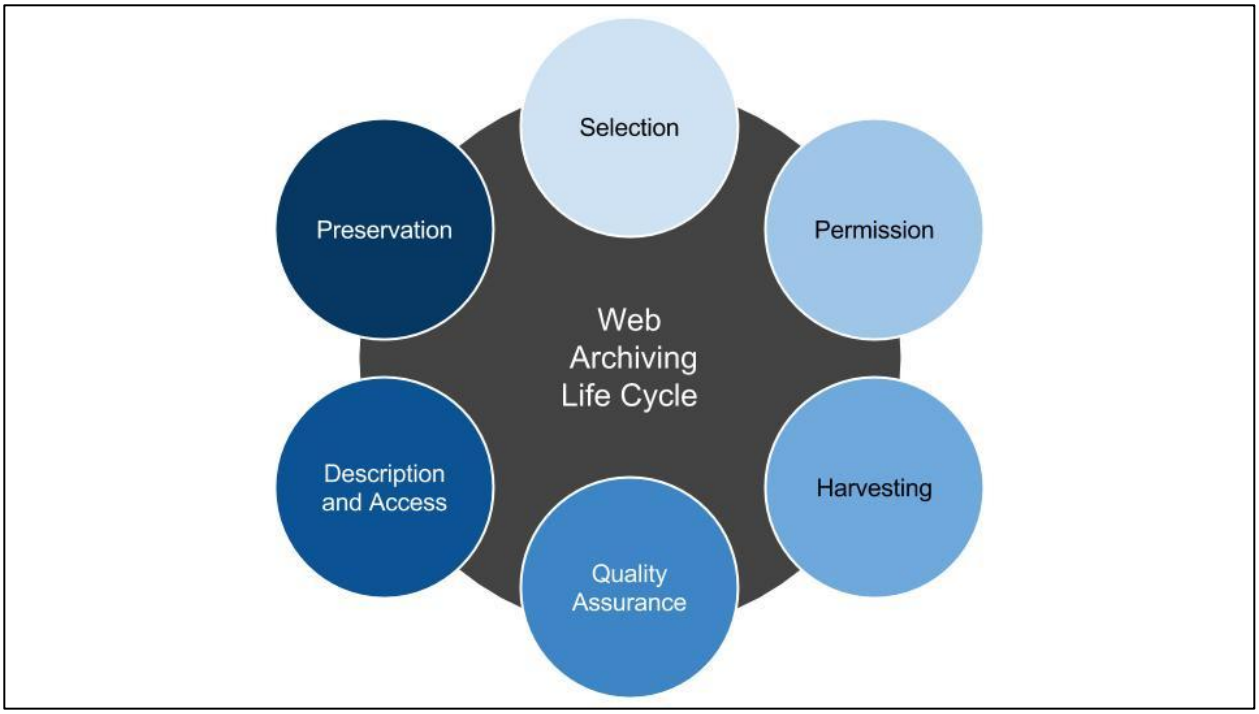

Figure 1: Web archiving life cycle at NYARC

We knew going into the implementation of our program that we faced many challenges – among them: scale, rapidly evolving and new web technologies, cost and limitations of currently available infrastructure or tools, and permissions considerations for intellectual property. Scale is perhaps the greatest challenge, but the requirements of web archiving are time intensive, more complex, and require more decision-making than we were accustomed to with collecting print materials.

Access and Discovery

In 2014, as part of our web archiving grant, we launched the NYARC Discovery research tool. The core objective for NYARC Discovery was to unite our Archive-It collections with additional scholarly resources, such as our books, periodicals, photoarchive images, journals, and e-books. The interface continues to be developed as we unite scholarly resources into this single search environment. A component of our two-year implementation web archiving grant involved review of several discovery layer interface proofs-of-concept, with the requirement of full integration of our web archive collections into the search result set. Staff from our institutions formed a Discovery Working Group to evaluate the proofs-of-concept. We selected the Primo product from Ex Libris, which features a central index, for implementation.

The development work was done by Lily Pregill (former NYARC Systems Administrator) and Ido Peled of Ex Libris, with the result of NYARC's Archive-It collections being integrated in the Primo interface through use of the already available OpenSearch API. Through the OpenSearch API, full-text results are delivered from NYARC's ten Archive-It collections (and additional Archive-It partner collections could easily be included via the API as well). The default search setting in the interface is to include the web archive results, although users may remove them should they wish. Links out to the full web archive search result set in the Archive-It interface are also provided.

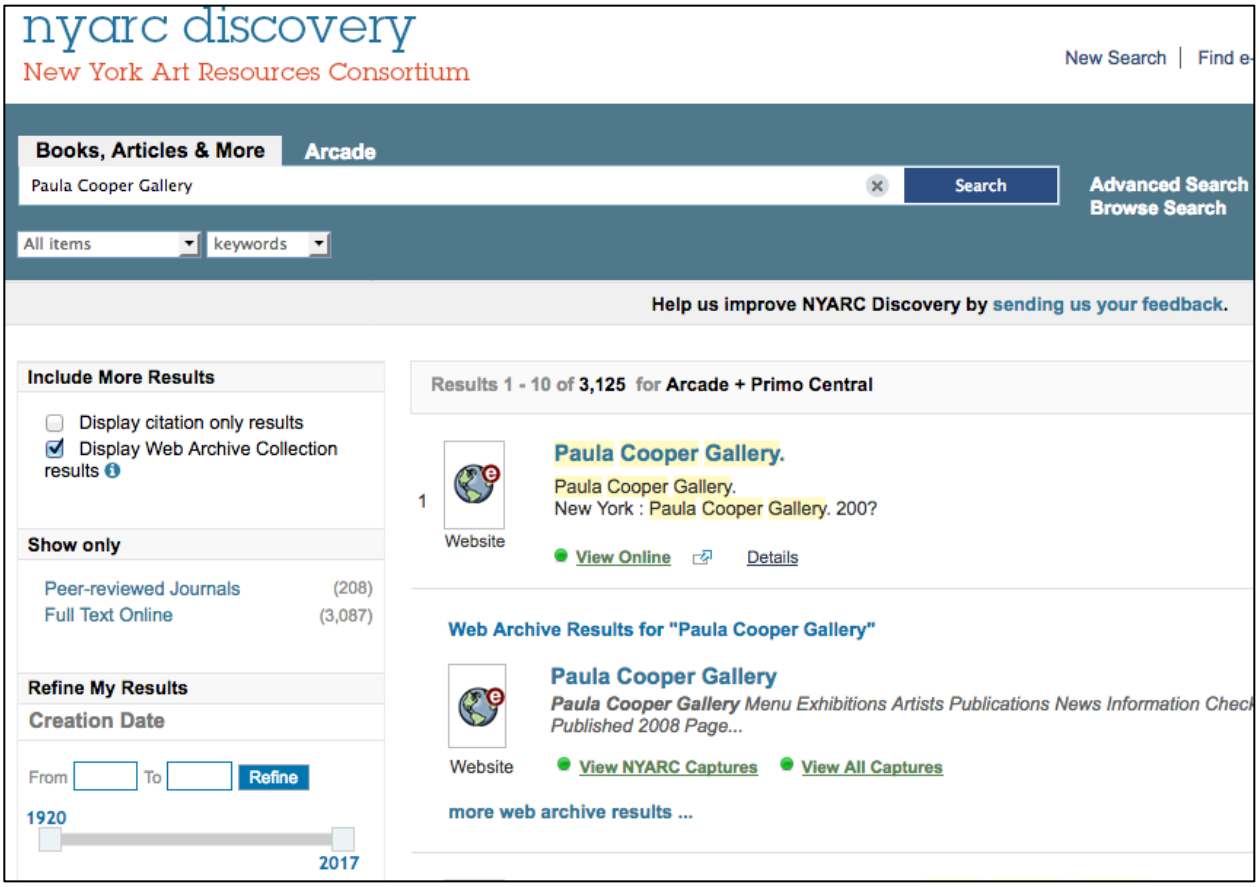

Figure 2. NYARC Discovery interface: single search environment includes web archive collection results

NYARC Discovery's hybrid approach in promoting access to our Archive-It collections is via two main access points. The first point of access is the MARC records that our catalogers are creating for both the live and archived versions of all seed URLs in our Archive-It collections (these records are also all made available in OCLC's WorldCat database). The second is via the full-text search results from Archive-It which are achieved with the OpenSearch API integration (this is a dynamic display, thus the result shows in the second display place and no faceting is available due to the API not being included in the central index of Primo). Users now discover

full-text search results from NYARC's Archive-It collections with the same ease as identifying a monographic title in NYARC's research collections, and all of the web archive collection content is freely and publicly available via the NYARC Discovery interface and through NYARC's partner page.

In March 2017, in collaboration with Martin Klein of Los Alamos National Laboratory, NYARC also integrated the MEMENTO API into NYARC Discovery. When users search via keyword, they now also have the option to "view all mementos" in addition to NYARC's captures of a site in Archive-It. When the user clicks the link to "view all captures" they are directed to the Memento Time Travel interface in a new window and this interface offers the option to view all available captures of the particular archived resource brought together from 25 different web archives.

The next phase of investigation for NYARC Discovery will pertain to researcher usage and expanded web archive content offerings, including a usability study of the NYARC Discovery interface which we have just begun in collaboration with the School of Information at Pratt Institute in New York City. Additionally, we will work to promote use of web archive collections with institutional staff, researchers, and for data analysis with special projects.

Collaboration in the GLAM Community

The challenges of digital preservation are not unique to art historical scholarship or to our institutions – they are issues faced by a much larger community with an investment in digital preservation and providing access to born-digital materials. Many would agree these issues are better addressed collaboratively by the community in a way that avoids duplication of effort. We've benefited greatly from being active in the web archiving community and sharing information with our colleagues who are undertaking similar work. NYARC is involved with organizations such as the Art Libraries Society of North America (ARLIS/NA), the National Digital Stewardship Alliance (NDSA) in the U.S., the OCLC Web Archiving Metadata Working Group, the Archive-It New York Users Group, and the Metropolitan New York Library Council (METRO) Web Archiving Group. In collaboration with members of the ARLIS/NA Web Archiving Special Interest Group (SIG) and the ARLIS/NA Digital Humanities SIG, a meeting was held at the most recent ARLIS/NA annual conference in New Orleans to discuss a project to create a digital art history registry, which would include web archiving digital art history project sites. This initiative would promote dissemination and centralize access to the variety of projects currently in development and/or publicly accessible.

NYARC is also actively pursuing a collaborative partnership to extend our web archiving program activities to a broader geographic network of art libraries in the U.S., and perhaps ideally those that are already involved with the ARLIS/NA network. We will be working to expand the web archiving of North American gallery websites, as well as institutional websites of the art libraries, in order to more comprehensively address the long-term accessibility of these unique web-based resources.

It's imperative that we make research on the web persistant, as it is estimated that the average lifespan of a unique URL is less than 75 days. The ever-evolving nature of the web often means that URLs in citations no longer link to the original information. In contrast to printed materials that exist for hundreds of years, web-based content seems unreliable and thus researchers are at times reluctant to cite URLs in their own scholarly output. In order for born-digital scholarship to be considered at a level equivalent to print publications, there must be a method for insuring accuracy and permanence. Confidence in permanent access will be vital to integrating born-digital scholarship into the scholarly canon. Fortunately, there are several easy ways to create archived versions of websites and permanent links to URLs (even if you are not part of a larger program of web collecting). One such way is by saving webpages in the Wayback Machine and then referencing the new archived URL that is produced.

The Frick Art Reference Library formed a Digital Art History Lab (DAHL) in late 2014 with the objective of providing art historians with the digital tools and data necessary to explore new methodologies. The DAHL group also works to stimulate collaborations between art historians and specialists from a variety of fields, from computer science to historical Geographic Information Systems (GIS). As part of the DAHL programming, which includes regular lectures and workshops at the library and museum, we taught a workshop in 2016 on web archiving and the preservation of art scholarship. This workshop will be repeated in fall 2017 at the Frick and is open to the public. We have additionally hosted workshops on mapping tools, data visualization, online catalogue raisonné softwares, and open-source web publishing platforms.

Beyond conducting workshops about saving art historical scholarship, we also seek to work in an interdisciplinary capacity with researchers who might benefit from anaylzing the data within the WARC files created via NYARC's web archiving program. We will be able to generate data sets from our WARC files that show key metadata elements representing all crawled resources, longitudinal graph analysis files, and web archive data sets which utilize named-entity recognitions tools to generate a list of all of the people, places, and organizations mentioned in each URI in a collection along with a timestamp. The potential for data visualization is one that excites us as far as the use of these collections and we will be interested to hear from

researchers as to how they might envision collaborating to make use of NYARC's collections for data analysis.

Our initial experience in developing a web archiving program, which compliments our historic analog collection development actitivites at the NYARC libraries, leads us to believe that there is rich potential for greater collaboration within the cultural heritage sector and GLAM community. We envision our future collaboration to encompass both shared effort in the collection and preservation of born-digital scholarly resources and exploration of potential researcher use and analysis of web archive data sets.

References

New York Art Resources Consortium (NYARC): www.nyarc.org

ISO 28500:2009 Information and documentation — WARC file format: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717

Bailey, Jefferson; Grotke, Abigail; McCain, Edward; Moffatt, Christie; Taylor, Nicholas (2017): Web Archiving in the United States: A 2016 Survey. NDSA, February 2017. Available at: http://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf

Internet Archive Wayback Machine: https://archive.org/

Archive-It: https://archive-it.org

Webrecorder: https://webrecorder.io/

Jones, Shawn M.; Van de Sompel, Herbert, Shankar, Harihar; Klein, Martin; Tobin, Richard; Grover, Claire (2016). Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content. PLOS ONE, 11(12), December, 2016. Available at: https://doi.org/10.1371/journal.pone.0167475

NYARC: Three major New York City art libraries announce plans for a shared program to preserve digital art resources. Available at: http://www.nyarc.org/sites/default/files/NYARCborndigitalgrantOctober2013.pdf

Archive-It – New York Art Resources Consortium (NYARC): www.nyarc.org/webarchive

NYARC Documentation: NYARC Collection Scope and Seed URL Nomination Process: https://sites.google.com/site/nyarc3/web-archiving/3-nyarc-collection-scope-and-seed-url-nomination-process

NYARC Discovery: https://discovery.nyarc.org

GitHub – Integration of Archive-It results via OpenSearch API into Ex Libris Primo discovery platform: https://github.com/technelily/archiveit-in-primo

NYARC Documentation: Metadata for Web Archives: https://sites.google.com/site/nyarc3/web-archiving/8-metadata-for-web-archives

MEMENTO Time Travel: http://timetravel.mementoweb.org/

Save Page Now, Wayback Machine: www.web.archive.org

Frick Art Reference Library Digital Art History Lab (DAHL): www.frick.org/research/dahl