Preserving born-digital catalogues raisonnés: Web archiving at the New York Art Resources Consortium (NYARC)

Sumitra Duncan

There has been a dramatic shift in publication of scholarly art research materials from print to digital formats. In late 2013, the New York Art Resources Consortium (NYARC) was awarded a two-year grant by The Andrew W. Mellon Foundation to address challenges of preservation and continued access for future researchers of these materials by implementing a web archiving programme. Due to the transition of many catalogues raisonnés to online from print, born-digital catalogues raisonnés are one of the five key collecting areas of focus for NYARC’s project. The intention of this article is to provide an introduction to the key components of the implementation of NYARC’s web archiving programme, to explore the far-flung implications of the digital shift, and to demonstrate the need for a community of collaboration in building curated web archive collections to benefit the future art historian.

Introduction: The digital shift

Presently there is a very real danger of a ‘digital black hole’ in the art historical record, with countless ephemeral web resources already lost to link rot, evolving technologies, and the transformation of the web over time. Materials significant to scholarship in the arts, such as auction catalogues, catalogues raisonnés and exhibition catalogues, are progressively being produced solely in digital form. Many art libraries are actively invested in addressing the myriad challenges of managing hybrid collections, yet in recent years an even greater challenge has emerged. As stated by James Cuno, President and CEO of the J. Paul Getty Trust, ‘Going forward, one of the biggest challenges scholars and curators of contemporary art and architecture face currently, and will increasingly face, is how to store, retrieve, and investigate born-digital materials.’ Equally complex is the question of who will bear the responsibility of preserving these often short-lived publications for the future; should it be the publishers or memory institutions?

Most libraries and archives are not equipped to adequately address the threat posed to the future of scholarship by acceleration in the analogue to digital shift in publication methods, and as of yet, only a small number of publishers have actively engaged in long-term preservation of ephemeral art materials. Digital stewardship of the vast and complex web, at least in the US, appears nearly insurmountable without a large-scale depository mandate in place. Yet despite the hurdles, a handful of libraries, archives and government offices are opting to grapple with this thorny territory in hopes of preserving elements of our cultural artistic heritage. Given the enormity of this task of preserving the open web, one way forward is through collaborations between entities that encompass the scholarly, governmental, and publishing spheres, with each taking responsibility for preserving a piece of the born-digital ‘pie.’ As Cuno pondered in his blog post, ‘...how can we manage these files? And why should we be working alone in trying to solve this problem?’

The actual value of curated web archive collections will not be known until an indeterminable point
in the future. The worth of web archives is demonstrated by instances in which researchers are seeking specific ephemeral art materials and discover that they have vanished from the open web, or that a large percentage of scholarly web citations lead to broken links and content that has ceased to be available online. Without the option of a print equivalent for acquisition by memory institutions, the art research community needs to work together to preserve art-rich materials for the future.

Background

At NYARC we have been working incrementally over the past several years to address the inherent challenges of the digital shift as they pertain to library and archival collections devoted to scholarship in art and art history. NYARC consists of the research libraries and archives of three 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 of facilitating ‘collaboration that results in enhanced resources to research communities.’ NYARC has historically been an active member of the art library community and is committed to collectively finding solutions to better manage hybrid research library collections.

In 2010, NYARC began to explore the efficacy of implementing a consortial web archiving programme with a pilot study at the Frick Art Reference...
Library. This pilot study sought to evaluate the usefulness of the Internet Archive's Archive-It service in capturing auction sale information, which at the time was only being distributed via online publication. Archive-It partnered with NYARC on the pilot study, offering a trial subscription to the Archive-It web archiving service (see fig. 1). Archive-It provides web-based tools, namely the Heritrix web crawler, for organizations to harvest and build collections of web-archived content for playback and access. NYARC staff and a graduate intern embarked on harvesting content from diverse auction house websites that were publishing catalogues online. As captures of the websites were initiated, the project team evaluated the harvested content for completeness and categorized the problems encountered along the way. The pilot study conducted at the Frick Art Reference Library demonstrated the many complexities of capturing certain types of web content, such as excluding unwanted content and avoiding 'crawler traps' where a web crawler would get stuck when attempting to harvest a site or crawling websites flush with dynamic content.  

While the pilot study raised more questions and concerns than were resolved, it did identify web archiving as a means of addressing the capture of ephemeral web content significant to the art historical record. It was clear that establishing a sustainable programme of web archiving for the NYARC libraries would require a more focused investigation about publishing trends and perceived technical and organizational challenges of building and preserving a web archive collection.  

In 2012, NYARC was awarded a one-year grant from The Andrew W. Mellon Foundation, called 'Reframing Collections for a Digital Age: A Preparatory Study for Collecting and Preserving Web-based Art Research Materials.' This exploratory grant allowed for research on publishing trends and input from our researchers on their current use and needs for web-based materials, engaging legal advice to develop intellectual property and fair use guidelines, investigating web archiving technologies, and evaluating existing institutional systems and workflows. The 'Reframing' study was executed in three successive phases, with separate consultants focused on the, ‘...‘tipping point,’ (i.e. the moment of critical mass) for the transition of analogue to digital formats for specialized art resources,” harvesting models, and infrastructure. The study identified five types of born-digital content as priorities for collecting at NYARC. The five areas also mirror the analogue collecting strengths at each of the NYARC institutions, and they are: born-digital catalogues raisonnés, auction houses, contemporary artists' websites, the NYARC institutional websites, and born-digital art resources falling under the categories of restitution of lost and looted art and New York City galleries or dealers.  

‘Making the Black Hole Gray: Implementing the Web Archiving of Specialist Art Resources’  

By addressing the stability of art and art history materials being made available exclusively online, NYARC identified basic risks, discovered top vendor partners, and determined the best available way in which to capture, preserve, and make accessible those resources considered of the highest priority to the NYARC institutions and the greater art research community. The findings from the ‘Reframing Collections for a Digital Age’ planning study informed the 2013-2015 NYARC implementation grant, ‘Making the Black Hole Gray: Implementing the Web Archiving of Specialist Art Resources,’ also supported by The Andrew W. Mellon Foundation. Web archiving, for our purposes, refers to the capture of born-digital materials as they appear on the live Web and storing that capture in a standard WARC file format. WARC files, or Web Archival file format, combine multiple digital resources into an aggregate archived file with related information or metadata. We aim to provide access to these materials to our researchers, both the content and functionality, and to preserve web archived materials over time. Our two-year objectives for the implementation of a web archiving programme are to capture, make accessible, and preserve two terabytes of significant art-rich websites as WARC files. In capturing this content, we will also be determining best practices and documenting a web archiving workflow for the three NYARC libraries, one which can be shared with other institutions interested in pursuing their own web archiving work.  

At this stage in the ‘Making the Black Hole Gray’ project, NYARC is working with several vendor partners, the primary partner being Archive-It, for the capture, public display, and quality assurance efforts of our web archive collections. We are also working with a commercial service, Hanzo Archives, for on-demand archiving of content that poses particular technical challenges for capture, such as dynamic content within Museum of Modern Art exhibition sites. NYARC will partner with DuraCloud during the second year of the grant to
integrate their cloud-based preservation service into our Archive-It account.

Archiving born-digital catalogues raisonnés

During the first year of our grant, NYARC partnered with several producers of born-digital catalogues raisonnés on the web archiving of their catalogues. NYARC’s inaugural partner in this endeavour was *The Isamu Noguchi Catalogue Raisonné*, with NYARC first attempting to archive the online catalogue in spring 2014 (see fig. 2). It was quickly evident that not all of the functionality of the site was easily captured via a one-time web crawl, due to the Heritrix web crawler not identifying each URL on the site as valid. Difficulty exists in archiving various types of content, such as databases, backend content management systems, and proprietary formats such as JavaScript and Flash. While efforts to fully archive the behaviour, appearance, and functionality of the Noguchi catalogue are ongoing, in the meantime we have proceeded with archiving the ‘print’ version of the catalogue as it exists as a born-digital PDF. In web archiving visual arts materials, simply capturing the text included on a site is not sufficient, thus making the challenges related to capture of the true appearance and behaviour of a site particularly important. A fall 2014 article from Pennsylvania Libraries states that,

> Ideally, the implementation of a quality assurance programme ensures that the Web content crawled by the institution has been accurately and completely archived so that the capture includes not only HTML content but also the look and feel of the site, as well as embedded documents and audiovisual materials.  

With permission from staff of *The Isamu Noguchi Catalogue Raisonné*, NYARC has included the PDF version in our Archive-It collection, where it serves as documentation of the catalogue as it existed in summer 2014. We will continue to archive this catalogue raisonné over time, with frequency of capture mirroring the schedule of updates to information made publicly viewable within the catalogue. As the Noguchi website states: “This progressive approach to publishing allows for research to continue while information presently
known can be made available to the public for reference purposes.

Technical services staff at the NYARC institutions have collaborative procedures in place for the creation of Arcade catalogue records for born-digital catalogues raisonnés (Arcade is the collective library catalogue for NYARC), such as the record created for *The Isamu Noguchi Catalogue Raisonné*, which are also made available to researchers via OCLC WorldCat (see fig. 3). Additional born-digital catalogues raisonnés that NYARC has actively begun archiving with the use of the Archive-It service include those for George Bellows, William T. Trego, Edwin Dickinson, and Germini G.E.L., from the National Gallery of Art. NYARC is presently working with the Jack Tworkov Estate, an enthusiastic partner interested in having *The Jack Tworkov Catalogue Raisonné* web archived and made available as part of NYARC's research collections. Collaborative discussions have also been ongoing between NYARC and Artifex Press, the innovative publisher of born-digital catalogues raisonnés of the artists Chuck Close, Jim Dine, and Tim Hawkinson. Artifex Press is comparably interested in preserving their catalogues for the future scholar, and likewise they must find the most logical and sustainable avenue for doing so – whether that be through a partnership with another institution or finding an internal solution to ensure long-term preservation.

**Building preservable websites**

As part of a recent event of the Catalogue Raisonné Scholars Association (CRSA) held at the Brooklyn Museum library, NYARC staff presented on their efforts to preserve and ensure continued access to archived versions of catalogues raisonnés. NYARC additionally shared with this audience how publishers of born-digital catalogues raisonnés and other materials can take certain steps to ensure that their websites will be preservable for future researchers. It is recommended within Columbia University Libraries' *Guidelines for preservable websites* that content such as pages, images, and videos, be created as standard links in HTML/ XHTML formats, versus proprietary formats such as JavaScript or Flash. Crawling non-HTML URLs has proven to be particularly challenging, as the crawler either cannot identify the content at all or deems it to be invalid (thus not possible to be captured). Site owners may also opt to create text-
only HTML to accompany pages that utilize Flash, to promote the ability to fully capture the website via web archiving.10 The Smithsonian Institution Archives also emphasizes the importance of providing a sitemap for web crawlers and researchers to more easily access an entire website, inclusive of those pages buried deep within it. Allowing for browsing and including links for 'viewing all' content will also promote greater overall accessibility.11

Collaboration to achieve sustainability and scalability

As Ricky Erway, Senior Program Officer, OCLC Research, succinctly stated, "...there are very few instances where users can access born-digital collections. Through effective communication and collaboration and by taking basic first steps, progress will be made toward that goal."12

The large scale and cost of web archiving undergird a strong argument for shared responsibility amongst the community to shape a preservation plan, inclusive of access points, for born-digital collections of web content. As there is far too much content being produced for any sole entity to be responsible for even one categorical piece, NYARC remains keen to capitalize on opportunities to foster future collaboration geared to preserving and cultivating access points to born-digital art research materials. This is an exciting time to be part of the innovative digital stewardship community, as many of the present players are already making significant contributions to building the collections that will be required for future digital scholarship in the arts.

According to a recent Archive-It blog posting, ‘To date in 2014, 326 Archive-It partners have created 2700 public collections on a diversity and range of topics, subjects, events and domains. These collections have become integral to these organizations' collecting strategies and have helped to raise awareness and understanding about why web archiving is so important.’13

NYARC will continue to welcome new inter-institutional projects and publisher partnerships, especially in the areas of tool development, establishing collaborative collection development models, and shareable documentation.

References

2. Ibid.

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