Truman Technologies

NYARC
Reframing Collections for a Digital Age
Report from Consultant No. 2

September 11, 2012

Reviewed by referenced parties January 2013. Comments and approvals are on file.

DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.

Table of Contents

EXECUTIVE SUMMARY

In February, 2012, the New York Art Resources Consortium (NYARC) received a $50,000 grant from the Andrew W. Mellon Foundation to conduct a preparatory study for collecting and preserving web-based art research materials. NYARC consists of the Frick Art Reference Library and the libraries of the Brooklyn Museum and the Museum of Modern Art Library and Archives. Three separate outside consultants were hired to perform separate phases of the project. This report (from the second consultancy phase) is intended to inform and guide the third consultant, who will recommend the best technical solution and workflow, and prepare a funding bid to achieve this. Consultant number two was directed to recommend:

a. What information the NYARC should collect.
b. The best methods for web archiving.
c. What partners (technological, publisher, other research libraries or preservation consortia) the NYARC should be working with.
d. What legal advice is needed in order to address intellectual property, ethical and access issues.

This report is divided into five sections that address the NYARC's requests.

Section 1 describes possible scenarios, or use cases, for web archiving at NYARC. These use cases were obtained during the first consultancy phase and documented in the consultant's report (Pines, 2012):

- Auction catalogs: a perpetually accessible, central repository of online art auction catalogs for art history research.
- Citation support: stable URLs to cited web resources that are essential to the scholar.
- Catalogues raisonnés: comprehensive catalogs of an artist's work that have been archived in book form for years and are moving to an electronic format.
- Artist gallery exhibitions: ephemeral material now only available via the web that must be collected and preserved for historic research.
- Subject-based research portals: specialized groupings of topical information.
- Artist files: all the ephemeral material relating to an individual artist, previously paper-based and stored for research, now available from a web site and in need of preservation.
- Archiving for small art organizations: possible opportunity to partner with contemporary art organizations.

Use cases help us think about the workflow and technological challenges for content capture and replay, both today and into the future. For each use case scenario, we map out the necessary situations and steps that need to happen in order for the project to be successful, using color-coding to track status. We identify known barriers (shown in red and orange) along with suggestions for mitigating or getting through a barrier to move forward. This mapping of requirements, versus the ability to meet them today, allows us to prioritize use cases to focus on the first year of implementation.

Section 2 covers technology options for website capture and replay, identifies known limitations and challenges, and discusses some developments underway by the research community. With an eye to potential adoption by NYARC or its collaboration partners, these developments should be monitored for the future:

- Overarching requirements.
- Available for-fee service providers and recommendations.
- Scope and frequency of captures.
- Known and anticipated challenges.
- Promising new tools.

Section 3 suggests possible collaboration scenarios relating to access, stewardship, collection building, tool building, outreach and advocacy, best practices, and education:

- Remote copy/access copy of the web archive.
- Collection development; building community and advocacy for bringing together or linking collections within the art world.
- Toolset functional definition and co-development.
- Collaboration with technology partners, preservation consortia and publishers, with an eye toward leveraging existing tools and influencing the push of required work upstream.

Section 4 discusses the processes, workflow elements and legal implications that arise in different steps of the workflow:

- Selection and harvest.
- Cataloging and organizing.
- Quality Assurance (QA).
- Integration and discovery with other systems.
- Legal, copyright, permissions.

Section 5 is a summary of ideas, recommendations and a suggested strategy for moving forward:

- Big issues and opportunities.
- Recommendations.
- Potential Collaborations.
- ▪ Roadmap.
- ▪ Measuring success.

INTRODUCTION

Materials that art libraries have traditionally acquired in print are increasingly being produced digitally. NYARC aims to take a leadership role in developing a centralized strategy to collect, preserve and provide access to these fugitive digital materials. Collecting digital materials that now reside in portions of the web and housing them in an accessible "web archive" for future art researchers and historians is one way to help accomplish this goal.

However, web archiving for preservation and for the replay and use of information far into the future will always be a challenge so long as the publishers of Internet content are innovating. There is no way to anticipate the evolution of the web, so whatever solutions we adopt today may be short-lived or evolving themselves.

It is a bad plan that allows no room for modification. - Pubilius Syrus

Assumptions

1. There is no coordinated effort in the art community to collect these materials.

2. Unlike print, we cannot assume that we are putting a process in place that will last 10 years. We cannot become married to any process, tool or service, and we must build in flexibility.

3. The resulting program for building and sustaining a web archive of art resources will not replace any existing (e.g., catalog) systems. The program will be an addition to existing NYARC systems, and identifying solutions to integrate discovery across systems is the role of the third consultant.

4. Staffing and funding for the project will need to be managed very efficiently. Where possible, we must leverage external resources (such as work study students, MLIS interns and collaboration partners). Direct costs must cover the expense of for-fee web archiving and hosting services. There will also be labor costs as determined by the third consultant.

5. Success is possible only if we collaborate and leverage the skills, ideas and resources of all organizations.

6. We are finding that Museum and art library use cases can be different from many of the Archive-It partners, according to the staff at Archive-It web-archiving service. (Kristine Hanna pers. comm., May 3, 2012). Access, re-use, discovery, and visualization components will play increasingly more important roles.

7. The output from this project will result in best practices, documents, services and toolsets that will benefit the broader web archive community.

How do we define success of this project?

- Provides for a seamless transition from physical to digital archiving of art ephemera and collections material as needed.
- Its collections and activities are important and relevant to the scholarly community (as measured – see section 5 of the report).
- NYARC remains at the forefront of innovation for delivery of information service programs.
- Establishes NYARC as a leader in the digital realm of art resource collections development and preservation (through collaboration, partnering, advocacy and awareness).
- Contributes back to the broader web archive community (best practices, tools, etc).
- Output and activities support the stated objectives of the Consortium[1].

1. USE CASES

The use cases outlined in this section focus on areas that have traditionally been collected in print form and are moving to digital, web-site-based form. They concentrate on publications aimed at scholars or extension of the kinds of materials NYARC has always collected, like exhibition and auction catalogs. Consequently, they do not include born-digital works of art.

The table below outlines a set of universal requirements that apply, regardless of the type of use case being considered. Requirements germane to a specific use case are identified in the sections that follow for each use case.

Universal Steps Required for a Successful Program	Green, Amber or Red Light (i.e., Identify Barriers)	Resolution to Barrier
Demonstrated need for NYARC to archive these web sites.	🟩 Academic researchers/art historians will continue to look to specialized art libraries (i.e., NYARC) for content as we move from print to digital for this content.	

[1] http://nyarc.org/about

Universal Steps Required for a Successful Program	Green, Amber or Red Light (i.e., Identify Barriers)	Resolution to Barrier
Permission granted to archive each site and agreed upon embargo period for access.	The permission process warrants its own section in this report. Based on information from peer institutions, this part of the program will be time consuming and require legal counsel.	Seek legal counsel (provide examples of peer institutions' policy). Manually collect permissions data, but start to investigate co-developing a permissions workflow and tool that adds automation. Collect and store rights as metadata.
Ability to hone in on relevant content (to exclude non-art-related content from being captured and archived).	Many sites have content that is not wanted for archive collection (e.g., auction houses have extensive collections of catalogs outside of art realm like jewelry, cars, etc.)	Ensure effective scoping and QA is in place to help focus the collection.
Address any copyright complications for images (capture and access/replay).	Images that are not in public domain (concerns about access, copyright and use). Even if a site owner/organization grants permission to crawl, they themselves may not have rights to the images, or their rights may only be for a limited period.	For captured images, ensure they are public domain or permission was granted, else keep images dark or view only from inside library with appropriate legal procedures in place. Verify that the site has 'blocked' (robots.txt) the crawler from capturing protected images. Archive web pages without copyrighted image(s) and add notation to visit library for version with image (public versus private collection).
Technologically able to capture and replay the site.	Known issues are documented later in this report. Address issues that evolve over time as web publishing evolves.	Assess if it's OK to live with restriction. Use known workarounds when available (e.g., direct URLs, site maps). For replay issues, capture and archive (wait for replay technology to address).
Integration with current integrated library system (ILS) and traditional ILS-based processes.	This is essential. Silos of information are not acceptable over the long term.	This is the focus of the third consultant and will be addressed in her report.

Universal Steps Required for a Successful Program	Green, Amber or Red Light (i.e., Identify Barriers)	Resolution to Barrier
Discoverability/access.	Need enhancements to what Archive-It alone provides in the areas of linking silos, federating search and providing for easier navigation.	This is the focus of the third consultant and will be addressed in her report.

Auction Catalogs

Much of today's auction house material is exclusively online, such as catalogs and pricelists. Downloadable PDF catalogs are on the decline, replaced by very dynamic web sites. By eliminating the printed version (and, in some instances, the PDF), it allows auction houses to make frequent updates (e.g., as pricing changes or art is removed or updated) and bypass the restrictive deadlines of printed catalogs (and their associated PDFs). Many of the smaller auction houses have stopped producing printed forms of auction catalogs altogether, although the larger houses such as Christie's and Sotheby's still do.

These new forms of web-based auction catalogs and electronic pricelists remain relevant to art historians and need to be preserved in digital form. An unfortunate by-product of the ease of update is the ease of removal – and often information about unsold lots and sales results are taken down after auction. However, researchers want to track sales results and sales data over time, as well as lots that were withdrawn or were "bought in" (unsold). The need for an historical time-line for these web sites, and catalogs contained or constructed therein, points to the need for web harvesting (or crawling) of each site on a periodic basis in order to preserve it for access by scholars and historians over time.

Special Requirements: Auction Catalogs	Green, Amber or Red Light (i.e., Identify Barriers)	Resolution to Barrier
Demonstrated need for NYARC to archive these web sites.	Catalogs are a chief tool of many researcher historians.	
Identification of the 'big universe' of auction web art catalogs.	NYARC work study student Identified project, Fall 2012.	

Citation Support (Link Rot and Permanent Citations)

In art scholarship and journalism, analog and e-source are becoming equally valid as citations. This is as true for artists' web sites as for peer-reviewed journals. For this reason, stable URLs are essential to citation support for the scholar. This must encompass URLs to simple-object digital material, such as text or PDF documents that are posted on web sites, and also to a static web page or pages. Web archiving at a specific point in time for a cited URL enables persistent links to web-based research, providing stability otherwise currently unavailable.

Special Requirements: Citation Support	Green, Amber or Red Light (i.e., Identify Barriers)	Resolution to Barrier
Demonstrated need for NYARC to archive these web sites.	Broken and old links highlight the need for this.	
Identification of born-digital works cited in scholarship.	Need to identify.	Suitable for a work study student project.

Catalogues Raisonnés

Comprehensive catalogs of an artist's artwork, known as catalogues raisonnés have existed in book form for centuries, and increasingly are born-digital. Companies such as Panopticon[2] offer software and hosting services for these raisonnés. In the case of Panopticon, the sites are built using a web-accessible content management application[3] that requires a license, and access is password protected. Once behind the password/login, the underlying artist-related information is most likely not suitable for web archiving. Although it is web accessible, it rests on cataloging and database technologies. That said, Panopticon states "we can add a web site so you can publish the data you choose to an online version of your catalogue." The web site version of the raisonnés can most likely be archived (with restrictions based on art copyright).

Other online projects include Gemini G.E.L online catalogues raisonnés (in conjunction with the National Gallery of Art)[4] and Raisone.org[5], a site sponsored by Childs Gallery[6], which is still under construction. Gemini has over 250 catalogues raisonnés[7] that are accessible without a password. Images, however, are under copyright (© Gemini G.E.L. and the Artist[8]). Raisone.org has around 12 artist raisonnés online[9] – and they all require a password to access.

Special Requirements: Catalogues Raisonnés	Green, Amber or Red Light (i.e., Identify Barriers)	Resolution to Barrier
Demonstrated need for NYARC to archive these web sites.	They are a chief tool of (CUL[10]) researcher historians[11]. PUL[12] librarian indicated that access to these "before they vanish" is a chief concern[13].	
Identification of the 'big universe' of web-based catalogues raisonnés.	Not fully documented yet.	Suitable for a work study student project.

[2] http://www.panopticondesign.net
[3] http://www.panopticondesign.net/CatRaisPages/CatRais1.html
[4] http://www.nga.gov/gemini/
[5] http://www.raisonne.org/site/home
[6] http://www.childsgallery.com/
[7] http://www.nga.gov/cgi-bin/search_www.cgi?cmd=search&q=raisonne (Accessed August 23, 2012)
[8] http://www.nga.gov/fcgi-bin/gemini.pl?command=image&catnum=7.46&imgnum=2&back=essay11
[9] http://www.raisonne.org/site/artist/ (Accessed August 23, 2012)
[10] Columbia University Libraries
[11] Carole Ann Fabian, pers. comm. August 3, 2012
[12] Princeton University Libraries
[13] Sandra L Brooke, pers. comm. August 2, 2012

Special Requirements: Catalogues Raisonnés	Green, Amber or Red Light (i.e., Identify Barriers)	Resolution to Barrier
Technologically able to capture and replay the site.	Need to work around password protected areas. For sites not database driven, the content appears to be a candidate for web capture. Images likely to be blocked. Sites resting on a CMS or database likely to be problematic and not suitable.	Further investigation required.
Ability to provide versions to researchers (i.e., first version then updates over time).	Researchers must accept that not all content will replay or remain (e.g., copyright images).	Set expectations.

Artist Gallery Exhibitions

In the past, the artist or related organizations, such as galleries, would provide (or the library would gather) printed mailers, posters, and other ephemeral material relating to the artist's practice as a whole. Today's artist sends email and posts to blogs and art-related web sites, promoting the event on his or her own web site. Rarely is this material printed. Since contemporary artists often blur the line between documentation and artwork, some types of ephemera may also be treated as rare or special collections material.

Special Requirements: Artist Exhibitions	Green, Amber or Red Light (i.e., Identify Barriers)	Resolution to Barrier
Demonstrated need for NYARC to archive these web sites.	Academic librarians typically have not dealt with this more ephemeral art info (such as the first exhibit of an artist).	

Subject-Based Research Portals

In the (distant) past, a researcher could expect go to a particular shelf/area of the library to find all related content. Today, with so much content existing on the web, there could (in specific cases) be an advantage to a subject-based portal of information. However, with the virtual death of traditional subject headings, the increasing sophistication of keyword searching and auto-tagging, as well as the new cross-disciplinary and dynamism of scholarly language, it has been argued[14] that topical groupings are, in fact, not key to this project.

This would be a very specific use case, driven by the need to direct scholars to "vetted" material.

An assistant professor of art history[15] interviewed for this project indicated that a web archive collection of content selected by NYARC gives it some authenticity. Then, rather than have students entering key word

[14] Jennifer Tobias, Librarian, Readers' Services, MoMA

[15] Ellen Prokop, Adjunct Assistant Professor of Art History, NYU and Associate Photoarchivist, Frick.

searches for themselves, they can be directed to the web archive as a comprehensive, vetted resource for them to get started. Students are used to using the web for research, but this helps direct them to pre-selected sites before going onto Google.

Special Requirements: Subject Research	Green, Amber or Red Light (i.e., Identify Barriers)	Resolution to Barrier
Demonstrated need for NYARC to archive these web sites.	Very specific use case. In order to be useful to a researcher, a subject-based portal would need to contain a critical mass of sites on given a topic.	Further investigation required.

Artist Files

An artist file consists of material relating to an individual artist – such as artwork, education, exhibitions, reviews, newspaper clippings, bibliographical information and much more.

Special Requirements: Artist Files	Green, Amber or Red Light (i.e., Identify Barriers)	Resolution to Barrier
Demonstrated need for NYARC to archive these web sites.	Academic libraries typically have-not dealt with this more ephemeral art info (such as the first exhibit of an artist).	

Archiving for Small Art Organizations

The born-digital project presents an opportunity to partner with experimental spaces, alternative publications, and other emerging group efforts to provide long-term digital archiving. It could also be an opportunity to discuss corresponding analog archiving with a given group.

Special Requirements: Archives for Art Org.	Green, Amber or Red Light (i.e.. Identify Barriers)	Resolution to Barrier
Demonstrated need for NYARC to archive these web sites.	Academic libraries typically have-not dealt with this more ephemeral art info.	

Narrowing Down

It is recommended that NYARC leadership narrow down the use cases for the first year of collecting. The decision should be based on least number of barriers and highest demonstrated need.

2. TECHNOLOGY

In this section, we provide a brief introduction to the software technology available to capture web-based content. There will always be some websites that take advantage of emerging or unusual technologies that the crawler (capture technology) cannot anticipate. A brief summary of these known restrictions is provided, but it will be necessary for those NYARC resources who administer and provide quality assurance to the captures remain educated on current restrictions and adapt their activities accordingly.

We begin this section with a high-level statement of requirements heading into the project, then look at some of the main service providers and toolsets available.

Web Archiving Components

While not an exclusive list of all components associated with web archives, this section is intended to help the reader get a basic understanding of terms used throughout the report.

Heritrix is the Internet Archive's open source software tool used to fetch (i.e., harvest or crawl), archive and analyze Internet-accessible web content. Heritrix is used to create the Internet Archive general web archive (the Wayback Machine at http://archive.org/web/web.php). It is incorporated into several open source toolsets for building web archive collections and into for-fee services. There are numerous other tools available to harvest web sites[16], but Heritrix is the most prominent and used by most services, including the Archive-It service recommended for this project.

WayBack, aside from being the name of the web archive collection at the Internet Archive, refers to the WayBack interface. This interface takes a user-supplied URL as input and returns a display showing links to snapshots (or crawls) of the site at different points in time.

The WARC (Web ARChive file) is the ISO-standard container used by web archives. Knowledge of WARC files is important, since WARC files containing NYARC collections can be ported from the service provider (i.e., Archive-It) to another institution for remote preservation or other use. For example, this might prove useful for working with partner institutions who can offer new and innovative ways of searching, analyzing and visualizing the NYARC web archive.

Search and indexing tools are used to provide full text and faceted search (e.g., using SOLR) of web archives.

Requirements

Service-level requirements (these relate to the for-fee service provider):

- Hosted service.
- Broad adoption and promise of long-term viability.
- Provision to get a copy of web archive files (remote copy) for NYARC (or its partners) to store.
- Contingencies for remote copies (disaster recovery and business continuance planning).
- Sufficient storage and quota provided.
- Contained costs and known service level agreements (no surprises).

Technology requirements (these relate more to the NYARC staff and end-user experience):

- Rendered as close to original as possible.
- Open and compatible with industry standards.
- Discoverable to all (based on permissions/access rules).

[16] http://en.wikipedia.org/wiki/Web_crawler

- Address (or roadmap to) limitations identified in the auction house web archive pilot project of October 2010 (Leahy, 2011).

Non-Requirements

- Archiving the content from inside a back-end database, or "database archiving"
  - Content delivered from a database to a web page that subsequently receives a unique URL will be a candidate.

Storage Capacity and Quota

Service providers charge based on the complexity and size of the collection. Archive-It provides each institution with a quota for annual storage capacity used for new crawls, seeds and number of documents. The quota varies on the subscription level and can be customized. A typical cost at the time of this report is $12,000 per year for one terabyte, 12 million documents and 300 seeds, with a maximum of three active collections queued for crawling at any time (inactive and active collections may be swapped in and out, and inactive collections may still be viewed and accessed).

- Once the use cases for the first year of collecting have been identified along with the desired frequency and scope of crawls, the projected capacity use, seed count and number of documents used against quota can be estimated.
- At this time, based on other known projects' collections, it is believed that this subscription level will be sufficient for at least the first year. Efficient scoping and use of the data deduplication feature will help ensure this.

Scope and Frequency of Captures

Each use case collection will be based on a series of web sites that have been deemed worthy of capture. In year one, this selection is expected to be driven by the librarians or curators at NYARC institutions. Each site must be reviewed to determine page URLs that are valuable to capture (e.g., the entire site, a portion, a page, a PDF document). The output of this exercise is a list of seed URLs that will be crawled ("seed list"). The seed lists for each use case collection can be further tuned to specify how deep a crawl will go per seed (how many hops) before stopping, and how frequently a seed is crawled. There are many tunable parameters and these are explained, along with guidance on scoping techniques, on the Archive-It wiki site[17]. There is also comprehensive online training for new subscribers to the Archive-It service that should be utilized.

Virus Scanning

Virus scanning of web archive files is the responsibility of the service provider or host. Since captured data is not executed during capture it is at low risk for virus infection. According to Archive-It (pers. comm., September 2012), exposure to Archive-It web archive files located at Internet Archive data centers is practically non-existent given extensive checks and processes in place. For Archive-It partners hosting a remote copy of the WARC files there are several implementations that address virus checking. Most use WARC tools or an equivalent WARC reader. See the following for additional information: http://netpreserve.org/events/active_solutions/4_Holden_Here%20be%20Dragons.ppt.

Available Service Providers

Several for-fee services are available to host and provide services around web archiving. The most prominent ones are highlighted here.

[17] https://webarchive.jira.com/wiki/display/ARIH/Scoping+and+Running+Crawls

Archive-It

Archive-It is the subscription service offered by the Internet Archive, a non-profit based in San Francisco, California. Over 200 partners are using the service.

California Digital Libraries Web Archiving Service

California Digital Libraries Web Archiving Service (CDL WAS) was developed and hosted by the University of California Curation Center. It is a subscription service used mostly by U.C. schools and academia, and has 19 total partners.

Hanzo Archives

Originally from London, with offices in San Francisco, Hanzo focuses on compliance, e-discovery and information governance. Its typical customers are government and corporations in the financial, insurance or pharmaceutical industries. It is a for-fee offering, based on a cloud/service model, with customers relying on the Hanzo experts to configure and carry out their web crawls.

Iterasi and Reed Technology

Iterasi teamed with Reed (part of the LexisNexis family) to offer web archive services. The service, like that of Hanzo, emphasizes e-discovery compliance and information governance and also has a strong emphasis on social media. The service organization configures and carries out web crawls for the client.

OCLC Web Harvester

A hosted solution offered by OCLC, Web Harvester is integrated with the OCLC WorldCat cataloging system (note, use of OAI-PMH feeds from other services can also accomplish this). It requires ContentDM, hosting and Web Harvester licenses. It has not been broadly adopted and does not appear to be being actively updated or enhanced. Seems more suited to PDF capture than to more complex sites.

Other Available Open Source Toolsets

It is worth covering a couple of additional options that are available, but do not meet the requirement of being available as a hosted service:

Web Curator Tool

The Web Curator Tool (WCT) is an open source tool that was developed by the National Library of New Zealand and the British Library to manage the selective archiving of websites. WCT software is available for download at: http://webcurator.sourceforge.net/ and can be installed on any platform that supports Apache Tomcat requirements. WCT has several modules built around the Heritrix crawler at its core. It supports permissions authorization, selection and scheduling, basic description, harvesting, quality review, and archiving. WCT would require that it be hosted (and maintained) by NYARC, which is out of scope for this project.

Netarchive Suite

The Netarchive Suite is an open source tool originally developed by the two national deposit libraries in Denmark. The French National Library and the Austrian National Libraries joined the project in 2008. The software is available for download at https://sbforge.org/display/NAS/Releases+and+downloads. Like WCT, it has several modules and is built around the Heritrix crawler. Since it requires that it be hosted and maintained by NYARC, it is out of scope for this project.

Recommendation

Toolsets that require hosting do not meet the project requirements. Consequently, WCT and NetArchive Suite are removed from consideration at this time. Both Hanzo and Reed services are considered unsuitable for clients like NYARC who wish to collaborate on collection development and have hands-on access to the crawls.

The subscription services of the Internet Archive Archive-It, CDL WAS and OCLC Web Harvester all offer hands-on access and collaboration and are hosted solutions. OCLC Web Harvester seems more geared to PDF document than the more complex sites and QA that NYARC will require, and it is not widely adopted. CDL WAS seems more geared towards academics and has many fewer partners than the Archive-It service. Archive-It meets the stated requirements and is the recommended service offering for this project. Since the Archive-It service allows for export of WARC files, NYARC is not locked in to this decision for the long term if requirements or the viability of the Archive-It service change over time,

Addressing Limitations Identified In Pilot

In February 2011, Sean Leahy issued a report at the request of NYARC assessing how Archive-It functioned as a tool for harvesting auction houses. The pilot itself was conducted in October 2010 and used the 3.5[18] feature release and 3.6[19] bug-fix release of Archive-It (issued July 10, 2010 and September 10, 2010 respectively).

Since that study, the Archive-It development team has released two major feature releases (4.0[20] and 4.5[21]) and is currently on 4.6[22] (a minor release released May 16, 2012). A week prior to issuing this report, the tentative roadmap and date (Q1 2013) for the 4.8 release was made public.[23]

The following lists the main limitations of the pilot and Archive-It's progress in addressing them, as of this report:

1. Crawl efficacy was the biggest challenge at the time of the pilot. Specifically, avoiding irrelevant pages and refining the harvest, crawler traps, and the inability to delete extraneous archived content.
   - Release 4.0 added a URL report to help with scoping decisions.
   - Release 4.0 expanded host constraint rules to allow for better granularity when scoping. It added document limits and the ability to block specific types of URLs (helpful in avoiding traps).
   - Release 4.5 introduced additional scoping improvements (add/edit rules in bulk, ability to activate or de-activate collection-level, expand scope rules to control which crawl each rule applies to).
   - Release 4.0 introduced the ability to delete seeds. While this may not remove some extraneous content from the archive, the additional use of test crawls (which do not store the data in the archive) should help.

2. Difficulty capturing complete content of content-rich site (i.e., discovering catalogs older than two or three years), assumed to be because either the time limit or document quota limit was reached.

[18] https://webarchive.jira.com/wiki/display/ARIH/Archive-It+3.5+Release+Roadmap

[19] https://webarchive.jira.com/wiki/display/ARIH/Archive-It+3.6+Release+Notes

[20] https://webarchive.jira.com/wiki/display/ARIH/Archive-It+4.0+Release+Notes

https://webarchive.jira.com/wiki/display/ARIH/4.5+Release+Notes

[22] https://webarchive.jira.com/wiki/display/ARIH/Archive-It+4.6+Release+Notes

[23] https://webarchive.jira.com/wiki/display/ARIH/Archive-It+4.8+Release

- Release 4.0 introduced variable crawl duration for one-time crawls. The default remains three days, but one- and seven-day crawl durations are available. The extended time should help with time limit issues.
- Release 4.5 added patch crawls to the QA process to allow for re-running the crawl to capture any URLs that were not captured for seeds URLs.

3. Miscellaneous issues that should be discussed with the Archive-It team on a per site basis to see if there are ways to set up the crawl to resolve them.
   - Does not appear to be able to capture specific auction house catalogs on a web site (we need specific examples to better understand this).
   - Selectable portal views (e.g., by country) with identical content, yet unique URLs (/aus/, /uk/ etc.), create captures for each location – how to reduce redundant content.
   - A single issue encountered due to Flash on a site.
   - Robot.txt – in the pilot, they did not bypass the robot.txt file (an option that is available on Archive-It), resulting in some missing images, etc. This is thought to be due to the material being copyrighted – and is what we would want to have happened to prevent copyright violations in publicly accessible NYARC collections.

4. Password-protected sites, which are on the roadmap for 2013[24]as part of release 4.8.

5. Unaddressed Java script and XML capture issues as of this report resulting in dead-end links or buttons not displaying correctly.

In summary, many of the crawl scoping and QA issues reported have now been addressed. However, the main technical challenges, for the most part, remain. These are discussed in the following section.

Technical Challenges We Know About

The Internet-Archive publishes a list of challenges to Web Archiving, and has done so since May 2008. The most recent of list of challenges is included in Appendix A. This section expands on the challenges as document by Internet Archive (and Sean Leahy) to help the reader understand the issues. Refer to the Quality Assurance section for additional information about the kinds of technical problems likely to be revealed during the quality review of a crawl.

Dynamic Content: Forms, Database-Driven Content

Dynamic content means that some or all of the page is generated at run-time by a program executing either on the client or on the server. Note that we are discussing content rather than display, which is covered next. Information on a site that has dynamic content may be available, but only if an explicit link to that information was crawled. If information was available on the site, but the user had to specify parameters to view the information, that information will not be collected by the web crawler for archiving because there is no unique URL generated. There are three types of content dynamism to consider:

1. User or client-based dynamism. The page generated for each user is based on client-side cookie information or logins to determine how to customize the page based on the user's credentials. For example:

   - NYARC use case this might affect: auction site catalogs or catalogues raisonnés with user logins/accounts that display different information (e.g., artists' portfolios and lots) based on the user's account profile. Other possibilities are auction house location-specific portals that display content based on a location.

[24] https://webarchive.jira.com/wiki/display/ARIH/Archive-It+4.8+Release

- Challenges: A crawler needs to be given the necessary login (user name and password) or cookie information.
- Mitigation: For login, get permission from the site to allow crawler to bypass password control, or to supply the login information.
- Crawling content behind a login or password is planned for the 4,8 release 2013.

2. Input dynamism. The web server returns database content to a page based on user input and usually requires text to be keyed into a search box or a submission form. This is also called the “Deep Web” or “Hidden Web”.

   - NYARC use case this might affect: Auction catalogs with forms.
   - Challenges: Requires interaction with the site, so creates problems for crawlers.

3. Temporal dynamism. Any page containing time-sensitive content exhibits this, such as a site displaying art news headlines. The more content changes, the more likely that only a ‘sample’ of time can or should be captured. There might also be times when a site changes from the start time to the end time of an actual crawl. In this situation, the actual capture of the site is never a true representation of the site at any point in time (known as temporal incoherency).

   - NYARC use case this might affect: Auction catalog sites (frequently changing with updates to content); most art-related web sites will exhibit temporal dynamism on at least the main landing page.
   - Challenges: Today’s crawlers can and do crawl temporally dynamic pages. The key issue in crawling such pages is freshness.

Some sites may display more than one type of dynamism, such as a page that requires user login, then based on the user profile and preferences, suggests art recommendations that are time-based from the latest action catalog or lot.

Dynamic Display or Appearance

The way a web site displays can be controlled from the server or the client side. Client side DOM[25] scripting (which implies Java Scripts) and dynamic html (DHTML) are ways of programming the browser to dynamically modify the visibility or appearance of objects for navigation. Examples include pull-down menus, floating and “mouse over” images. Although the content may be able to be archived, reproducing these menus may not be possible.

- NYARC use cases this might affect: All use case will likely have sites that include Java Scripts
- Challenges: These may pose a problem for web crawlers to capture, although the links themselves will usually work for replay.
- Mitigation strategy: Identify alternate versions of the page/site to crawl that do not have the script if/when available. Contact Archive-It for assistance with specific sites to see if they have recommendations or ‘tricks.’

Databases

Archiving of actual databases (i.e., the content that resides inside the back-end database) is not something that can be accomplished without mapping the content into a standard schema and using additional tools to allow access. This is out of scope for the project, but has been accomplished by the French National Library (BnF) using tools it developed (DeepArc and Xing).

[25] Document Object Model

A note about database-driven web pages

Database-driven web pages are built from databases that pull content from the database into a web page, sharing it on pages as needed and excluding it from other pages. Drupal, a popular open source web content management system, is a good example of how this works. Drupal pulls blocks of content from the database, such as People (e.g., Artist) profile pages, Contact information blocks and "today's news" blocks (each site will use different names for their blocks of content) and assembles them together. In this scenario, the content is intrinsically static even though pages are dynamically generated.

NYARC uses Drupal to build and display its web site. Since content blocks are pulled together to create web pages (each with their own URL), these pages can easily be crawled. [As noted above, the exception will be any additional dynamic content pulled from a database of information accessed from the site via a form, search bar, or other mechanism based on client- or user-based dynamic input.]

HTML5 and "Web Sockets"

HTML5 is a newer form of HTML that is growing in popularity. All the modern browsers support rendering HTML5, it is the optimal choice for smart phone browsers, and more and more sites are adopting it.

HTML5 web socket addresses the latency and overhead problems associated with traditional polling mechanisms in client-server applications. Web socket is a bi-directional and full duplex communication standard used to build next generation applications requiring real-time interaction – such as for web games or apps sharing financial data. Server and client can exchange messages over the single channel and the application can to do a server-side push to the client browser. Web socket is supported natively on most modern browsers (or via a plug-in for Internet Explorer). Since the browser opens the page and content comes in from servers, it changes dynamically, which presents a problem to web site capture.

HTML5 is the topic of a forthcoming investigation by the Archive-It service team (Kristine Carpenter Negulescu, pers. comm., August 21, 2012).

FLASH

Web sites with FLASH can be captured and viewed, but can present a problem if the FLASH file provides navigation for the web site. When a FLASH file is the only way to navigate within the site, the site will not be accessible. The FLASH file information that sends the user into the site cannot be corrected in order to redirect to the archived site. Consequently, the user will either be directed to the live site or to a "404 Not Found" message if that site is no longer available.

Streaming or Downloadable Media

When content (video) is being pulled from a media server, it currently cannot be archived. If the media is downloaded, it can be captured. However, immediate QA is recommended to ensure the content was adequately captured. Copyright of the content may be an issue and authority to create a download web archive copy of the media should be verified.

Blocked Content

Sites may elect to block the web crawler from crawling the site using a robots.txt[26] exclusion (the name of the Archive-It crawler is archive.org_bot). This is useful to protect copyright infringement, so many sites that NYARC will crawl block content in this way. If the site is to be crawled after obtaining permission from the site, the robots.txt restriction can be bypassed. Any bypass should be done with the expressed permission of the site owner and be documented as part of the permissions process.

[26] http://www.robotstxt.org/robotstxt.html

Mobile Publishing User Experiences

Increasingly, art museums and art sites are creating mobile versions of their web sites. These user experiences are different, due in part to their being created for the small screen. These versions of web sites are out of scope for this project unless a mobile version can be displayed from a conventional browser and crawled as part of the main collection. Mobile applications (for IOS and Android smart phones and tablets) are offered for most museums and, while applications are out of scope, they will need to be on the NYARC radar for potential capture and preservation in the future.

Promising New or Existing Tools

Several tools are being developed that warrant a mention, with an eye to their potential use by NYARC and its community.

- WARCreate, a research project out of Old Dominion University, is a Google® Chrome browser plugin that captures web pages. It is about to go into beta release. Future development work is required, since today's tool captures just the page and not an entire site, but it holds promise for personal archiving. WARCreate can encrypt the WARC file, making it particularly interesting for personal sites (Facebook®, etc). If the tool lives up to its promise, it may be something that NYARC partners (such as artists) can use to submit WARC files into a personal collection managed by NYARC.
- Visualization tools (see next section).

3. COLLABORATION & PARTNERING

NYARC has a track record of successfully collaborating to achieve specific objectives, and can point to itself as a proof point. Several areas for collaboration present themselves for this project.

Aggregating and Building a Collection

There is no central place where an artist, historian or researcher can go to learn whether a site is being archived, and by whom. This gap could be filled by a range of aggregation methods – from the creation of a wiki site (e.g., Wikipedia) or a web portal, to a more comprehensive solution. NYARC is well positioned with its ties to ARLIS and to numerous art institutions to lead a collaborative effort that can identify what is being archived and create a place where people can find this out.

Another area for collaboration is around collection building, where a small group of peer institutions select the sites that need to be collected. Potential candidates for this include:

- Ivy Art and Architecture Group (IVAAG). Conversations have begun with two Ivy League institutions and should be continued.
- National Gallery of Art, specifically in the area of catalogues raisonnés.

Stewardship and Access

Content collected through Archive-It is stored with primary and backup copies and is periodically indexed into the Internet Archive's general archive. Internet Archive stores two copies online. They are working with partners to keep redundant copies in other locations, such as DuraCloud and the Bibliotheca Alexandrina in Egypt. To responsibly manage and protect its web archive assets, it is prudent for NYARC to have a copy in more than just the Internet Archive, especially to address the needs for Access.

A remote copy allows for new opportunities to access and navigate through the information. Simply providing access via WayBack, topical browsing or full-text search as offered by the Archive-IT service today, is unlikely to meet the needs of researchers. The section on toolsets covers this in greater detail.

Very early conversations (by this consultant[27]) have been conducted with DuraCloud about possibly adding visualization services to its cloud offering. DuraCloud is currently storing copies of WARC files for the Archive-It service. Should enough institutions be interested in different kinds of access and visualization of their collections, a DuraCloud service offering could be a viable option. These conversations should be continued. There might also be an opportunity for one of the Ivy Libraries to host a copy of the NYARC web archive in return for other services, such as access to its datasets for mining. These conversations need to occur.

Toolset Development

There is an opportunity to collaborate on two immediate areas of toolset development that could benefit the broader research and web archive community:

1. Visualization.
2. Permission and rights tracking.

Visualization tools: NYARC has a vision to improve the Archive-It service by adding a visual front end for access by researchers. Analysis and visualizations of historic data collected over time may be analyzed to understand relationships, linkages, and provide insights into social, cultural and historical forces relating to the art world. Sites like the UK web archive[28] show promise in this direction with a selection of visualizations as access points into data. As part of its doctoral program, the Computer Science Department of Old Dominion University (ODU) has a series of visualizations it is running against Archive-It web collections. Dr. Michele Weigle and her team are applying for National Endowment for Humanities (NEH) funding in early 2013 to continue this research, and there is an opportunity for NYARC and ODU to work together. NYARC would provide access to its web collection and help define requirements, and ODU would further the development of its toolset. The conversation with ODU has started and should be continued.

Permissions tracking tools: The Library of Congress (LoC) has developed DigiBoard[29], a robust tool for centralized permissions management, including sending and tracking notification and permissions letters (emails). During a webex meeting and demonstration, September 4th 2012, the LoC indicated that the tool might be planned for eventual open source and that the two organizations should remain in touch. In year one, NYARC should investigate bringing a group of Art libraries together to collaborate on furthering this tool for their particular requirements and workflow.

Outreach, Education and Awareness

NYARC must keep abreast of a quickly changing environment by being active in the professional and research network around web archives, discovery and visualization. Discovery of new tools and insights into understanding how procedures are being developed outside of NYARC are essential to moving forward in this rapidly changing landscape. Collaboration with, and outreach to, technology partners, preservation consortia, and publishers has begun and needs to be a continued focus going into the first year of the project. These are the institutions with tools and systems that NYARC can leverage and influence, with an eye to new tools and capabilities.

Organizations to consider joining:

- National Digital Stewardship Alliance (NDSA) out of the LoC is a vibrant network of digital preservation experts from universities, consortia, professional societies, commercial businesses, government agencies, and more. The NDSA is open to any organization that has demonstrated a commitment to digital preservation and that shares the stated goals of the consortium.

[27] Gail Truman

[28] http://www.webarchive.org.uk/ukwa/

[29] http://www.loc.gov/webarchiving/technical.html

- IIPC (if money were no object – starts at 2,000 Euros)

Listservs and online communities:

Not an exhaustive list – a place to start.

- https://listes.cru.fr/sympa/info/web-archive
- https://lists.sourceforge.net/lists/listinfo/archive-access-discuss
- http://tech.groups.yahoo.com/group/archive-crawler/
- http://netpreserve.org/about/curator.php
- Linked in groups:
  - Archive-It
  - Internet Archive
  - Web Archiving
- http://blog.archive.org/
- http://britishlibrary.typepad.co.uk/webarchive/
- http://www.iterasi.com/blog
- http://ws-dl.blogspot.com/

Events to consider attending (learn, network, and educate about NYARC web archives)
These are in addition to the traditional events NYARC staff attend:

- Ivy League Art Libraries meeting - November, Princeton.
- Half-day Archive-It partner meeting – 12:00 Monday, December 3rd, Annapolis, Maryland (Loews hotel).
- Archive-It Partner meeting and the best practices exchange – December 4-6, Annapolis, Maryland (Loews hotel).
- Joint Conference on Digital Libraries 13 – July 22 – 26, Indianapolis.
- Society of American Archivists 13 – August, New Orleans.
- Personal web archiving meeting – Spring 2013, Maryland.
- Digital Preservation 13 – July 23 – 25, Washington, D.C. (part of NDIPP/NDSA).
- IIPC General Assembly – Spring 2013. The first day is usually open to the public.

4. PROCESS & WORKFLOW

How the process ties into the existing workflow at NYARC is the subject of the report by Lily Pregill, the third consultant for this project.

Selection and Harvest

The challenge for site selection is to bring the process into the NYARC workflow without adding significant process or work for librarians. At project start, the initial site selection for each use case will need to be a focused endeavor, carried out by those librarians who are most familiar with the collections. Ongoing selection over the first year and into the second should start to expand, since selection that is limited to within the consortium does not allow the project to scale.

Open nomination should be a goal heading into the second year of the project, such that the nomination and selection can be extended to librarians at peer institutions and, ultimately, to a website owner or third party. Web-based, open nomination forms are in use by other institutions to solicit ideas for sites to collect from the public.

- UNT[30] has a seed URL Nomination tool[31] it runs as a service for collaborative projects where there are a number of entities that are interested in submitting candidate urls to a centralized list

[30] The University of North Texas

for crawling. The tool allows for each “project” to have different levels of metadata assigned to them based on what is important to the project. At this time the tool has no concept of Permissions.

More innovative methods for identifying web sites using behaviors and URL recommendations from the crowd are starting to emerge, and might be incorporated over time as resources allow.

Opt-in, self-archiving by artists who have sites they wish to be harvested, or the submission of WARC files by these artists, may present a for-fee service offering that NYARC could offer, and should be investigated further. Tools being developed, such as WARCreate, offer a glimpse of where personal archiving is heading, and their progress and adoption should be followed.

Cataloging and Organizing

The level of description influences the access and how we feed web archives into existing systems for discovery. It also influences how long it takes to create a collection. Archive-It provides 15 Dublin Core metadata fields that are used for adding metadata at the collection, seed and document level. Optional custom metadata fields are also provided and not all metadata fields are required. While per seed of per document description may be desirable for discovery, this most likely will prove too cumbersome without a level of automation.

More automated ways to collect metadata from sites must be found and should be investigated heading into the first year of the project. Tagging by the community at large to add descriptive data (such as what is being done by Flickr® and YouTube®) offers an interesting goal for future years in the area of building out each collection once it is online.

The newly announced 4.8 release of Archive-It includes the ability to import metadata. No further details are available as of this report.

Quality Assurance

Although the Archive-IT documentation and guidelines for QA are very thorough, experience with the tool indicates that the QA process will be time consuming. NYARC could potentially augment the QA process with work study students or MLIS interns, adopting a 10-question checklist that others, such as University of Texas at Austin, have used (Columbia U. Web Summit Meeting. May 2012). Archive-It service offers a checklist (Archive-It 2012) for the QA process.

Rhizome works directly with artists to determine what constitutes a successful capture. (The Signal Digital Preservation blog 2012). There might be areas of the NYARC collection that could be a joint QA endeavor.

Technical problems seen during the QA process will be either display or harvest problems.

- If a problem is a display problem, it should be considered temporary. The content was harvested successfully, but there are limitations in the QA playback software that are expected to be fixed in the future. These problems might include:
    - Links to other blocks in the same page do not work.
    - Navigation menus that should expand or collapse but don’t.
    - Drop-down menus that don’t work correctly.
    - Garbled text due to incorrect character encoding.
    - Links that lead to live sites.

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

- Harvest problems tend to be more severe because content was not acquired. This can be due to crawler restrictions, crawler problems or scope problems. These problems cannot be fixed for particular harvests and might include:
  - Not enough of the site was harvested (crawler timed out?).
  - FLASH landing page or another issue prevented navigation through the site.
  - Crawler restrictions.
  - Linked pages to *404 Not Found Error* page.

Access and Discovery

Unfortunately, not everyone is familiar with web archives or has heard of the Archive-It service. Although it is possible to access the Archive-It service through a portal or link from the NYARC web site, the web archive itself remains its own silo of information, not integrated with the Arcade ILS or WorldCat. Access to this silo must be via the WayBack interface, topical browsing, or a full-text search. Having to know the URL or a date is not that useful for people who don't always know what they are looking for.

> Siloed resources run the risk of becoming underutilized; if resources are not easily found they are ignored – Lily Pregill

Integration and Discovery With Other Systems

This is the focus of the third consultant's report.

ILS/Arcade Local Integration

This is the focus of the third consultant's report.

WorldCat and OAI-PMH

According to Archive-It online documentation, (Archive-It. 2012) Archive-IT provides metadata records for the OCLC WorldCat catalog via its OAI-PMH feed. By default, WorldCat harvests Archive-It metadata records monthly. During these updates, new records are added to WorldCat, enduring ones are updated, and those no longer exposed to the OAI-PMH feed are removed. Once these records have been harvested, they are displayed and searchable in the WorldCat catalog, each with a link back to the corresponding Archive-It public page.

A collection-level record is not granular enough for NYARC and art researcher requirements when it comes to discovery. Each collection (auction house site) will probably include numerous catalogs, each of which, in the equivalent paper world, have their own catalog with controlled access points to discover each catalog individually. In the web archive word, the level of manual curation and description at the document or page URL level is daunting and would not scale. This points to the need for automated metadata extraction and for NYARC and its collaboration partners advocating for the use of schema.org, linked data, etc.

Web Integration

Integration with Google®, Bing® and other search engines will be based on recommendations by the third consultant.

Restricting Access

Due to copyright limitations for some of the crawled material (such as images), NYARC will need to offer restricted access to some of its collections. The Internet Archive is planning to introduce IP authentication at the collection level with the upcoming 4.8 release of Archive-It, planned for Q1 2013. No further details are available as of this report. Restricted access can also be handled by using web-site login and password restrictions.

Legal, Copyright and Permissions

In 2008 the Section 108 Study Group Report reexamined the exceptions and limitations applicable to libraries and archives under the Copyright Act, in light of digital technologies. Its recommendations for changes to the Copyright Act are found in its report[32] and should be of interest to NYARC legal counsel.

For this report the most influential recommendations of the Section 108 Study Group appear to be the following:

1. Libraries and archives have the right to capture and archive publicly available web sites without requesting permission to do so in advance.
2. Because this content was originally freely available online, the Study Group believes libraries and archives should also be permitted to make the captured content available remotely to their users, but only after a reasonable period of time has elapsed and only if it is marked as an archived copy.

Suggested framework

Subject to guidance from legal counsel, the following approach is suggested. This approach is based on best practices used by peer institutions. See Code of Best Practices in Fair Use for Academic and Research Libraries p. 26 (2012), http://www.arl.org/pp/ppcopyright/codefairuse/code/index.shtml

- Respect "no-crawl" directives in robots.txt files[33]
- Issue notifications in advance of crawling. Identify the URL, include statement of intent to crawl, ability to opt out, and take down request procedures.
- If content is at risk, capture it even without advance notification under a Fair Use argument.
- For sites where images or video are not in public domain (nor are they blocked from crawling using robots.txt exclusions) obtain permission to create offsite access (i.e. to republish).
- If there is an embargo period before content is made accessible to public, consider the gap between notification of intent to crawl (or receipt of permission) and the capture as part of the embargo period.
- In general, attempt to collect scholarly significant materials. If personal information is collection as part of the scholarly record, obtain permission to republish.
- Where permission to republish is required but not obtained, limit access to inside the libraries (with appropriate in-place usage rules based on existing policy).
- Provide appropriate contact details and instructions for copyright holders who believe their rights have been infringed by inclusion of their work in the NYARC archive.

Notifications and Permissions

Based on the type of site, the level of notification and permission will differ. Notification and permission requests will need an email letter with legally approved wording. Letters used by the LoC are forthcoming based on the September 4th meeting.

- Crawling the site
    - No notification required
    - Notification required
    - Permission required
- Creating offsite access copy (republish)

[32] http://www.section108.gov/docs/Sec108StudyGroupReport.pdf

[33] A robots.txt file can be observed by appending robots.txt to the site URL (e.g. http://nyarc.org/robots.txt)

- No notification required
- Notification required
- Permission required

For example, the NYARC.org site (and those of the consortium members) would be classified as "No notification, No notification". Sites that are likely to have copyrighted images and content will likely need to be classified as "Permission required, Permission required", unless the copyright images have been excluded from crawling by the site owner (using robots.txt exclusions). Any site that requires a login will require that this login information be provided. These sites will be classified as "Permission required, Permission required". The following should be completed based on legal counsel. Some suggestions are included.

Creating Offsite Copy (Republish)				
Crawling the site		**No Notification**	**Notification**	**Permission**
	No Notification	NYARC.org and member sites		
	Notification		Citation support (works cited in scholarship) Artist files and exhibits (where no copyright images are being captured) Professional facebook or social sites	Any (pages containing) images that are not public domain Personal facebook or social sites
	Permission			Auction catalogs and catalogues raisonnés that are password protected News sites

Process

- Prior to crawling a site, a letter (via email) should be sent to the site owner (unless no notification is required). The site owner typically is not the web master (see "about us" page for contact info).
- For notification only – no bounce-back will count as successful notification.
- "Contact us" forms are suitable for sending notification to.
- If a site granted permission in the past, there is no need to ask for it again.
- Each reponse must be tracked.

Embargo

The reason to embargo material is to not compete with the live web. As noted earlier, the section 108 study group recommends that "a reasonable period of time has elapsed" before it is made available. This has been interpreted as six months by many institutions, including the California Digital Library Web Archiving Service and the Internet Archive. Note that Archive-It provides an option to make content available following a crawl, but the content is held for six months before it is moved over to the Internet Archive's general, WayBack archive. Columbia University has indicated that it does not embargo content (pers. comm., email July 2011), while Harvard WAX service is understood to have a three-month embargo.

For any institution, including NYARC, there will be occasions where the live web material is not longer accessible - such as a result of a "take down" or site removal. And there will be sites with rapidly changing content. When information is of immediate value to a researcher or scholar this will influence the embargo period. In any case, the archive version will be marked as an archive copy.

Obtaining Permission

Copyright and use information is generally found in each website's footer information by following the "copyright" or "terms and conditions" links. For example, the Bonhams auction house terms and conditions at http://www.bonhams.com/legals/9944/ state, "We own or license the copyright in this site and in material published on it (including descriptions and photographs of articles). Those works are protected by copyright laws and treaties around the world. All our rights are reserved. You may print off one copy of any page(s) from our site for your personal reference and you may draw the attention of others within your organization to material posted but you may not reproduce or permit anyone else to reproduce such material without our prior written consent. Our status (or that of any identified contributors) as the authors of material on this site should always be acknowledged."

For instances where a web site contains images that are neither in the public domain, nor can the owner of the site grant permission for the image rights, steps can, and should, be taken to seek permission from the holder of the rights:

- The University of Texas Harry Ransom Center and the University of Reading Library in England jointly created WATCH. WATCH is a database of contact names and addresses of copyright holders or contact persons for English-language authors and artists. http://norman.hrc.utexas.edu/watch/
- A list of other collective licensing agencies (agencies that centralize copyright ownership information for their respective industries) is found at Columbia University Libraries Copyright Advisory Office http://copyright.columbia.edu/copyright/permissions/collective-licensing-agencies/

Other Considerations and Legal Resources

Some web sites include photographs of people. Aside from being concerned about the ownership of the photo and its permission to use, it is advisable to understand the rights of the people or person in the photo:

- Legal counsel should advise whether a release is required from the person(s) to protect against any claim against privacy rights.

Additional legal resources:

- http://www.section108.gov/docs/Sec108StudyGroupReport.pdf
- Fair Use evaluator: http://librarycopyright.net/resources/fairuse/
- Creative Commons reuse http://creativecommons.org/licenses/
- Lawyer Lesley Ellen Harris' web site http://www.copyrightlaws.com/us/legally-using-images/
- The United States Copyright web site http://www.copyright.gov/laws/ and form for a search (fee required) http://www.copyright.gov/forms/search_estimate.html
- Columbia University Copyright Advisory Office Director, Dr Kenneth Crews completed a study of museum policies and licenses funded by the Samuel H. Kress Foundation. http://copyright.columbia.edu/copyright/2011/06/27/copyright-museums-and-licensing-of-art-images/
- List of licensing agencies office http://copyright.columbia.edu/copyright/permissions/collective-licensing-agencies/
- Database of artist copyright holders http://norman.hrc.utexas.edu/watch/

5. SUMMARY

Opportunities

NYARC will be unable to sustain or scale this project unless it partners and collaborates. There are several opportunities for NYARC to form an ongoing coalition of art libraries and technology partnerships to explore how to share the burden of responsibility.

Situation	Opportunity for NYARC
Pockets of art web collections without a central place where people can find out what is being archived (e.g., if a small art gallery is going away, where to find out if its web presence is already being archived).	Lead effort to identify what is being archived and where people can find this out.
Permissions and rights tracking needs to be more automated.	Lead collaborative effort to develop permissions toolset with the art community (based on LoC) that has broader appeal to other web archivists.
Access and visualization needs for art patrons and researchers differ from those of traditional web archive patrons (and those currently offered by Archive-It).	• Improve on the Achive-It service by collaborating to provide a visual front-end into collections. • Lead collaborative effort to gather and share art-based requirements. • Influence visualization tool development.

Recommendations

- Select Archive-It subscription as the host and service provider.
- Investigate and pursue collaborations per below.
- Continue focus on education, networking, and outreach.
- Take on an advocacy role.
- Innovate with new tools and processes by leveraging collaborations.

Potential Collaborations

Continue conversations and exploration for opportunities with:	Discussion or activities around:
Columbia University, NY, NY Library school Avery Library CS	• Work study students (Auction House project, QA). • Art as the next piece of their Mellon grant (after HR). • Possibly hosting a remote copy of NYARC archive.
Princeton University, Princeton, NJ CS	Possibly hosting a remote copy of NYARC archive or doing DH research against data sets.
Old Dominion University, Norfolk VA CS	• Visualization tools development (NEH grant – focus on Art/NYARC as a partner). • WARCreate as a potential for personal archiving.
Ivy League Art Libraries group IDAAG (ID art and architecture group)	• Collaborative effort to identify (and document) all art resources that are being archived. • Year 2+ potential to use library resources for nomination/selection of Art sites to crawl.

Continue conversations and exploration for opportunities with:	**Discussion or activities around:**
ARLIS	Collaborative effort to identify (and document) all art resources that are being archived.
DuraSpace/DuraCloud	Visualization access service/against hosted remote copy (potential for).
ALL of above	Influencing the roadmap and priorities for year 2+.

Roadmap

Activity	Year 1	Year 2+
Community and advocacy	Start/lead collaborative effort to identify and document all art resources that are being archived. Start/lead discussion groups around: - How to link the silos. - Nomination and permissions (automated workflow/tools). - Visualization and access (tools).	Should have rolled out the results of Year 1.
	START COLLECTING:	EXPAND:
Scope of collections	Sub-set of use cases (as identified by NYARC leadership team).	Consider adding more use cases.
Site selection	• Manual in-house selection using traditional model. • Investigate larger community for form-based submissions (Ivy Art Libraries).	• Peer nomination and selection. • Investigate tools and process for crowd-based nomination.
Permissions	• Manual process (based on legal council). • Gather requirements and partners for more automated workflow and potential tool development.	Automated (using automation and workflow (tool) from Year 1.
Access/discovery	• Link via access portal into Archive-It (i.e., a Silo). • Basic search and discovery – WayBack, browse (tag), search.	• Should be ready to start using visualization tools and other access mechanisms from Year 1. • Broader discovery (based on third consultant).
Visualization	• None • Investigate art scholars and researchers requirements • Investigate and start partnerships	Use of visualization tools.
Integration with rest of the NYARC collection and bibliographic records	• This is its own project in need of funding. • A discovery layer to unite everything into a search interface.	Dependent on funding as a separate project.

Metrics: Measuring Success

One obvious metric for measuring success of a collection is the extent to which is used by others. In an email dated September 10th, 2012, Kristine Hanna explained that access statistics for Archive-It collections are available to partners on request (typically quarterly or annually). More automated way for access to these stats is under development (through PiWiks[34] open source web analytics software).

REFERENCES CITED

Archive-It, 2012. Archive-It Guide. Accessed August 23, 2012
https://webarchive.jira.com/wiki/display/ARIH/OCLC+WorldCat+Catalog

Archive-It, 2012. Archive-It Guide. Accessed August 26, 2012
https://webarchive.jira.com/wiki/display/ARIH/QA+Checklist

Archive-It, 2012 Archive-It Guide. Accessed August 25, 2012
https://webarchive.jira.com/wiki/display/ARIH/5+Challenges+of+Web+Archiving

Leahy, Sean. 2011. Archive-It and Online Auction Catalogs. A report on the functionality of Archive-It as a tool for harvesting data from auction house websites. http://nyarc.org/sites/default/files/Archive-It%20FINAL.pdf

Pines, Doralyn. 2012. Reframing Collections for a Digital Age Report.

The Signal Digital Preservation blog . Accessed August 23, 2012
http://blogs.loc.gov/digitalpreservation/2012/08/preserving-digital-culture-art-theater-video-games-and-more/

[34] http://piwik.org/

APPENDIX A

Five Challenges of Web Archiving[35]

This is reproduced in full from Archive-IT on-line documentation (Archive-It 2012).

Added by Renata Ewing, last edited by Kristine Hanna on Jun 20, 2012 (view change).

Certain types of content can be challenging to archive effectively. These difficulties affect all web crawlers, not just Heritrix. When selecting seed URL's you wish to archive and reviewing your archived content, please keep these limitations in mind:

1. Javascript: While for the most part, sites with Javascript on them can be archived without any problems, Javascript can sometimes be difficult to capture and display. Javascript is commonly used to create navigation menus (if you mouse over a word and a drop-down menu suddenly appears, Javascript is most likely in use). Often we can archive the content, however reproducing the menus and other javascript files can sometimes be difficult for the Wayback Machine. Sometimes there are slight changes that can help with the capture/replay of a site, so please feel free to contact us at archive-itsupport at archive.org to see if anything can be done for a specific site.

2. Streaming & Downloadable Media: Streaming media cannot currently be archived. Downloadable media is usually captured but can also be difficult to archive reliably in large volumes. If you plan to archive sites which include a large volume of downloadable media, we suggest immediately checking the sites after they've been crawled to make sure the media was captured to your satisfaction. Viewing your archived site in proxy mode is the most effective way to make sure the media was archived. If you notice problems with downloadable media being archived, please contact a partner specialist (email archive-itsupport at archive.org). We will notify you when methods for capturing streaming media become available.

3. Password Protected Sites: Currently Archive-It opts to crawl the public web and does not crawl information protected behind a login/password. We are looking into this capability for our 2013 development roadmap.

4. Form and Database Driven Content: If you need to interact with a site to get to the content, Archive-It can have difficulty crawling the site. There are two workarounds to archiving database driven content. If there are links into the raw content, the crawler can follow those. Also if there is an XML site map to your seed site, you can include this in your seed list. Archive-It will be able to crawl all links included on an XML site map.

5. Robots.txt Exclusions: Sometimes a webmaster will use a robot.txt exclusion to prevent certain content from being crawled. Our crawler respects all robots.txt exclusions. To see if an entire site you wish to crawl is being blocked, please check your seed status report after your crawl is complete. To check if part of your website is blocked, please check your hosts report.. If you wish to crawl a site blocked by robots we encourage you to contact the webmaster of the blocked website to allow the Archive-it crawler in. Please contact us for the user agent string. There is also a feature within the Archive-It web application that allows users to over ride robots.txt blocks.

[35] https://webarchive.jira.com/wiki/display/ARIH/5+Challenges+of+Web+Archiving