Archiving Websites Containing Streaming Media
Howard Besser, NYU
http://besser.tsoa.nyu.edu/howard/Talks/

Archiving Websites Containing Streaming Media

• Background issues and problems
• Archiving Young Composer websites
  – Our Technical Collaboration
  – Our Collaboration with Content Creators
    (preservation, access control, rights, agreements)
  – Workflows
  – How things will look
  – Evaluation
• Impact beyond this Project

BACKGROUND ISSUES AND PROBLEMS

Web Archiving poses challenges

• Any given web page may be updated frequently
• Web links constantly break (404 errors)
• Few tools/services exist for “Curated” web archiving (Archive-It, CDL’s WAS), and they require significant training/experience to learn, but we do have int’l-accepted format (WARC)

Many parameters need to be set for Web Archiving

• Frequency of crawls
• Depth of crawls (# of hops)
• Starting points of crawls (seeds)

Other issues for developing good crawls

• Quality control/assurance
• Workflows
• Fidelity to original web pages
• How end user will navigate and view it
Archive-It

- The leading application/service for curated web archiving in North America
- Run by the Internet Archive, and is much more targeted and curated than their WayBack Machine
- Is based on Crawler software developed by IA (Heritrix) in 2003-2004
- Is very poor at capturing streaming audio or video as well as inserting it properly into a composed web page.

Archive-It Issues w/Streaming Media

- 2014 Oklahoma Campaigns - State House Web Archive (Oklahoma Department of Libraries)
- Not in Archive
  The page you requested has not been archived in Archive It.
  This could be for a number of reasons.
  Most likely the page you are requesting was outside of the crawler's scope. Try another request or click to see other pages from ahousa.pecus.com.
  It is also possible this site is currently being crawled and the archived pages are not yet available in the Wayback Machine. It can take up to 24 hours after a crawl has been complete for the site to appear in the Wayback Machine. Please try again after the allotted time.
  You can also try searching for ahousa.pecus.com/ViewPublisher.php/view_412 on the FAQ, or on the general archive at the Internet Archive at www.archive.org

Archive-It Issues w/Streaming Media

- FireFox version 39.0. Screenshot of Tarik O’Regan’s site taken 2015/10/05

Archive-It Issues w/Streaming Media

- Archive-It screenshots generated as part of our project-
  - By Lorena Ramirez-Lopez
Archive-It Issues w/Streaming Media
Firefox version 39.0. Screenshot of Tarik O’Regan’s site taken 2015/10/05

Archive-It Issues w/Streaming Media
Firefox version 39.0. Screenshot of Tarik O’Regan’s site taken 2015/10/05

Some sources of streaming issues

• Problems with capturing resources residing on 3rd party services (YouTube, Vimeo, Soundcloud)
• Problems with how faithfully the A/V materials are captured and placed by Archive-It
• Problems with websites generated through site building platforms such as Squarespace

Archive-It Issues w/Streaming Media
Firefox version 39.0. Screenshot of Ted Hearne’s website taken 2015/10/05

Some sources of streaming issues

• Problems with capturing resources residing on 3rd party services (YouTube, Vimeo, Soundcloud)
• Problems with how faithfully the A/V materials are captured and placed by Archive-It
• Problems with websites generated through site building platforms such as Squarespace

ARCHIVING YOUNG COMPOSER WEBSITES

• Collect, preserve, & make available Websites of Composers
• $480,000 grant from Mellon in 2015 to NYU Library/MIAP/Internet Archive
• Dealing with the issue that contemporary composer websites go up and down (and also incorporate relationship-building btw composer and fans)
• Addressing the problems of collecting streaming media
• Also selectively collecting high-quality versions that are used to generate the streams, and allowing future researchers to see/hear the higher quality versions

Archiving Composer Websites

Archiving Composer Websites

- Develop good and ongoing relationships between Libraries and Composers
- Develop Trust
  - for developing collections, and continuing to add to them
  - for Policy reasons
- Examine what type of errors take place
  - how faithfully audiovisual materials are being captured
  - how resources that reside on third-party web services (YouTube, Vimeo, Soundcloud) are (not) displayed within Archive-It’s interface
- Issues with websites generated through site building platforms such as Squarespace
- Find ways to fix those errors

Metrics Accomplished (as of Jan 2017)

- 172 Composer sites crawled, scoped, assessed for quality, & analyzed for problems (feeding into IA development work)
- 800 QA/QC reports generated
- Initial web archiving agreement from 165 Composers (25 from NPR’s 100)
- Identified website infrastructures encountered and created a classification matrix

Website Infrastructure encountered

![Website Infrastructure](image)

Project Team

- Jefferson Bailey (Internet Archive)
- Howard Besser (MAAP)
- Lori Donovan (Internet Archive)
- April Hathcock (Lib/ScholComm)
- Nicole Greenhouse (Lib/ACM)
- Carol Kassel (Lib/DLTS)
- Scott Statland (MAAP)
- Donald Mennerich (Lib/ACM/DLTS)
- David Millman (Lib/DLTS)
- Courtney Mumma (Internet Archive)
- Robin Pruss (Lib/AFC)
- Lorena Ramirez (MAAP) — special thanks!
- Michael Stoller (Lib/G&S)
- Kent Underwood (Lib/AFC)
- Chela Scott Weber (Lib/AFC)

OUR TECHNICAL COLLABORATION: CRAWLING

NYU/IA Collaboration
NYU/IA Collaboration

Traditional Crawlers
- Archive-It and other web archives use Heritrix
- Follow links, capture most web content
- Less successful with streaming video and dynamic content executed in the browser
- Umbra helps

“browser” | “crawler” = BROZZLER

Brozzler System Architecture v1

OUR COLLABORATION WITH CONTENT CreATORS
Young Composers Corpus

- Began with NPR’s 2011 list of “100 Composers Under 40”
- 91 of 100 have own self-contained sites
- As of 5/2016 have written agreements with 165 Composers (25 of them from NPR’s list)
- Will recruit 10 of them for enhanced archiving (uncompressed; better than what is on website)
  - This will require an added appendix to contract/agreement (which may involve dark archiving and/or restricted access)

Building relationships with Composers

- Engage them with the idea of preserving their Website
- Are they willing to give us richer versions of content on their site?
- Are they willing to make all (or just part) of the content freely accessible? Do they want to embargo some content in a dark archive?
- Donor Agreement/Contract

Donor Agreement/Contract

- Have been working on this with lawyers for approximately one year
- Have had fairly stable language in it for 6 months, and 2 contracts already signed and returned
- Does default to allowing us complete rights for reformatting and for allowing researchers to see/hear all high quality versions at minimum on-site
  - And thusfar all Composers contacted have agreed to those principles (but not necessarily to the contractual language)

Long Term Preservation for Scholarship

NVU and Composer wish to establish long term preservation of the materials listed in Appendix One (the “Materials”) at the highest possible quality. The Parties wish to enter into this Agreement to establish guidelines and standards with regard to ongoing and future library processes related to such preservation. In consideration of the mutual promises contained in this Agreement, the Parties agree as follows.

Composer indicates restrictions on Access

- Describe access restrictions for each item listed in materials:
  1. Available for streaming from NYU Libraries’ website
  2. Available for streaming with password authentication: Does NYU Libraries’ for researchers with prior approval
  3. Only available for streaming on-site at Butter Library or off-site for NYU faculty, staff, and students
  4. Only available for streaming from designated on-site terminals in Butter Library
  5. Not available for streaming until a future date
Who are certified users?

Appendix Two, NYU will make the Materials available to those people that NYU, in its sole discretion, deems qualified according to NYU’s then current rules and regulations for access to archival materials.

ARCHITECTURE & WORKFLOWS

Architecture & Workflows

• Full copy of all websites (incl richer content) stored in NYU Repository and accessible through NYU Finding Aids
• Metadata is in ArchiveSpace
• Connections built off of ArchiveSpace back-end API

Current Development work

• Supplying a separate audio player?
• Hiring a Digital Archivist
• Precise forms of navigation btwn ArchiveSpace, Archive-It, and richer content within NYU’s digital repository
• API-

API

• What IA needs from NYU API
  – API URL
  – Credentials(username, password) -> Authentication Token []
  – Repository ID
  – Resource ID
• What IA will return as JSON array
  – Unit Title
  – Creator
  – Data Expression
  – Extent Statement
  – Tech Characteristics
  – [Something Based on Access Restriction, i.e. can it be streamed]???
• We Speak Etruscan, 1993 May 21, 23.5 MB, 1 AIFF file Stereo uncompressed 16 bit/44.1K
• The Dream of Innocence III, 1998 March 26, 150 MB, 1 AIFF file Stereo uncompressed 16 bit/44.1K

HOW THINGS MAY LOOK
User Queries

- User browses through Archive-It
- User sees that A/V content exists (and in some cases, it will include richer content, but some of that might be access-restricted)
- Archive-It hands off user to NYU (either directly to A/V content, or to Finding Aid)

One option for Queries

```
Archive & Shackleton Machine

AV
Dash Board
(AV, Finding aid)

Linking information to archive content

NYU Archival Services

Finding Aid

Responsibilities:

- Respond to inquiries about collections
- Provide access to content
```

One option for high quality content

- On archived website page listing composer’s content, user sees a message that higher quality content is available, with:
  - Access restrictions, if applicable
  - Link to relevant finding aid
  - (looking like following image)

Evaluation for Improvement

- Composers and their satisfaction with the ways in which audiences will be able to view archives of their websites
- Researchers, and whether the content and functionality of these web archives works for them
- Tweaking what we do in order to better serve Creators and Researchers
Schedule and Methodology for Evaluation

- Dec 2017—Schedule one-on-one interviews with sets of composers and researchers
- Jan-Mar 2018—One hour individual sessions with 10 Composers and also with 10 Researchers, having them look at the user interface and conduct queries
  - Composers: Are they satisfied with how audiences will be able to view the archival copies of their websites? Is it better or worse than their own live sites? Are they satisfied with the audio and video placement and quality (as well as options)? Are they content with the Donor Agreement? What changes/improvements might be made to any of these?
  - Researchers: Can they find what they need in the web archive? Is it difficult (clunky) to use? What parts don't work well or aren't intuitive? We want to identify what changes in the content, functionality, or navigation features would improve their user experience
- Apr-May 2018—Construction of Evaluation Summary containing the list of improvements/changes that should be made to the Archiving project
- June-Aug 2018—Implement the changes

Impact Beyond this Project

- Archive-It will be able to better handle streaming media, and display it in proper context
- We will have architectures and workflows for Archive-It to interact with richer local resources (as well as examples of how interaction and navigation can proceed btwn Archive-It, ArchiveSpace, Finding Aids, and an internal digital repository)
- Models for interaction btwn creators and collecting organizations will have been developed (incl donor agreements)
- We will have preserved 100+++ websites of young composers

Archiving Websites Containing Streaming Media

- http://besser.tsoa.nyu.edu/howard/Talks/