Archiving websites containing streaming media: the Music Composer Project

Howard Besser, NYU
http://besser.tsoa.nyu.edu/howard/Talks/

PROBLEMS WITH HERITRIX AND ARCHIVE-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

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The page you requested has not been archived in Archive-It.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 okhouse.gov.com.

It is also possible that this site is currently being crawled and the archived pages are not yet available in the web archive. It can take up to 24 hours after a crawl has been completed for pages to be added to the web archive.

Archive-It Issues w/Streaming Media

Archive-It screenshots generated as part of our project-
• By Lorena Ramirez-López

Archive-It Issues w/Streaming Media

Archive-It Issues w/Streaming Media

Archive-It Issues w/Streaming Media

Archive-It Issues w/Streaming Media
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

Other Issues we’re trying to solve

- Discovering URLs generated by Javascript

Archiving 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 between 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

Some methods used

- Began with NPR’s list of 100 important composers under 40, and augmented the list with faculty and librarian suggestions
- Identified website infrastructures encountered and created a classification matrix

THE PROJECT

- Developed 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
**Website Infrastructure encountered**

**Project Team**
- Jefferson Bailey (Internet Archive)
- Howard Besser (MIAP)
- Lori Donovan (Internet Archive)
- April Hathcock (Lib/ScienComms)
- Nicole Greenhouse (Lib/ACM)
- Carol Kaibel (Lib/DLTS)
- Scott Statland (MIAP)
- Donald Mennerich (Lib/ACM/DLTS)
- David Milman (Lib/DLTS)
- Courtney Mumma (Internet Archive)
- Robin Press (Lib/AFC)
- Lorena Ramirez (MIAP)—special thanks!
- Michael Stoiber (Lib/ACR)
- Kent Underwood (Lib/AFC)
- Chela Scott Weber (Lib/AFC)—departed

**OUR TECHNICAL COLLABORATION: CRAWLING**

**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

Logo: Noah Levitt
Brozzler Model

- **job**: collection of seeds
- **seed**: principal unit of crawl configuration
  - one browser works on one seed at a time (politeness)
  - seed has its own configuration, also inherits from parent job
- **page**: atomic unit of crawling from brozzler perspective
- **url**: only browsers, warcprox have to deal with every url

Warcprox: WARC-writing http proxy

- man-in-the-middle for https
- asynchronous: WarcWriterThread
  - writes warc records
  - saves deduplication info
  - updates statistics

Other pieces

- python wayback
- Rethinkdb (distributed document store)

Stream capture relies on Youtube-dl

[https://github.com/youtube-dl/supportedsites.html](https://github.com/youtube-dl/supportedsites.html)
OUR COLLABORATION WITH CONTENT CREATORS, IP ISSUES

Young Composers Corpus

- Began with NPR’s 2011 list of “100 Composers Under 40”
- 91 of 100 have own self-contained sites
- Within a year of starting we had written agreements with 165 Composers (25 of them from NPR’s list)
- Planned to 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

- Worked on this with lawyers for well over a year
- Have had fairly stable language in it and many 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)

Contract Intro

- NYU and Composer wish to establish long-term preservation of the materials listed 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.

Elements in the Contract

- What is being acquired
- Terms of Transfer
- Terms of user Access
- Rights & Responsibilities (both NYU & Composer)
- Appendix describing each item (format, content, amount, other pertinent descriptors)
- Appendix with Access Restrictions-
4 possible Levels of Streaming Access

• Available for copy-protected streaming from the NYU Libraries’ website with unrestricted access by the general public.
• Available for copy-protected streaming from the NYU Libraries’ website — with access limited to registered NYU faculty and students and — to external researchers with eligibility to use NYU Libraries’ archival resources according to NYU Libraries’ general access policies, with password authentication, on or off campus.
• Available for copy-protected streaming on NYU Libraries premises, at designated workstations, with access mediated by NYU Libraries personnel.
• Not available for streaming until a designated future date.

Tentative pieces of the Contract

• The uncompressed master files of Materials licensed for inclusion will be made available to the Libraries to enable the research and development of higher quality tools and processes for archiving on the Web and successor technology. The resultant high-quality copies of Composer’s website—incorporating the best quality media files—will be preserved as historical documents in the archive, which will be accessible worldwide on the Web or successor technology as a storehouse of cultural memory and a vehicle for research and scholarship. Composer retains existing rights to his or her Materials, subject to the license granted in this Agreement.

Tentative pieces of the Contract

• non-exclusive worldwide, perpetual, irrevocable, royalty-free right to produce, use, copy, and distribute Derivative Works
• strictly limited to reformatted digital files or to excerpts and abridgements (such as thumbnails) created for the technical purposes of building, preserving, and providing access to the Web archive over the World Wide Web or its successor
• may be used only for the non-profit educational and research purposes provided under this Agreement
• Agreement does not affect or transfer any copyrights or other intellectual property rights

Architecture & Workflows

• The Finding Aids are generated from ArchiveSpace (which contains rich metadata)
• There is an overall Composers Finding Aid, as well as a separate Finding Aid for each composer (listing inventory and web archives, and link to assets)
• Web archive is stored in Archive-It; richer content in NYU Repository
• Connections built off of ArchiveSpace back-end API Demo Site

Software & Service Components

• IA’s Archive-It
• NYU digital library internal components — Aeon for workflow management — ArchiveSpace — EAD
Unfinished Development work

• Supplying a separate audio player?
• Still working on precise forms of navigation between ArchiveSpace, Archive-It, and richer content within NYU's digital repository
• What will be on the workstation for items that need to be looked at on-site?
• Issues with streams that were not captured
• Example of work done on IA's API-

Interim work on API to IA

• 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 Restrictions, 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

Query paths still under development

One option for 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 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) -

Demo from API side

http://composers.dlib.nyu.edu/

From the Library Finding Aid side

http://dlib.nyu.edu/findingaids/html/failes/mss_479/
From the Library Finding Aid side
http://dlib.nyu.edu/findingaids/html/fales/mss_460/
Besser-IIPC 13/11/2018

From the Archive-It side
https://archive-it.org/collections/7872
Besser-IIPC 13/11/2018

From the Archive-It side
https://archive-it.org/collections/7872?
q=h4p%3A%2F%2Fwww.bitrosie.com&show=SeedVideos&fc=seedId%3A1157594
Besser-IIPC 13/11/2018

From any direction, user might need to authenticate
Besser-IIPC 13/11/2018

SOME OTHER INTERNAL TRACKING
Evaluation for Improvement

• Composers and their satisfaction with the ways in which audiences will be able to view archives of their websites (improving usability)
• Researchers, and whether the content and functionality of these web archives works for them (content presentation)
• Tweaking what we do in order to better serve Creators and Researchers
• Finding out whether captures really worked

Findings still being analyzed

• Streaming captures appear more successful, but we still experience some streaming capture problems
• Need further exploration to see the precise cause of the crawler/capture issues (& rectify them if possible)
Crawler Issues (video capture failure)

Crawler Issues (Flash video issue)

Crawler Issues (video captured without audio)

Crawler Issues (broken video links)

Crawler Issues (1 audio not captured)

Crawler Issues (audio not captured)
Crawler Issues (audio failure & anchor problem)

Crawler Issues (partial capture failure)

Crawler Issues (incomplete loading)

Crawler Issues (capture issues)

Crawler Issues (unknown problems)

- Campjugle.com:
  - Any capture date if very slow, load time hard to tell if were working or not, so some subjects gave up. Site owner says this is inherent to site, so might not be a capture problem.
  - Discrepancies between when one hop out is captured or not.

- Kmarrilkim.com:
  - Sep 20, 2017 capture (latest capture): Attempts to play music from archived tumblr page from various platforms (youtube, soundcloud, etc.).

- Bitrosie.com:
  - All capture dates: links take roughly 5 minutes (assumed broken at first)

- Adelefournet.com/video/:
  - Sep 12, 2017 capture: Video error after roughly 10 seconds. Stops playing "Barranco District, Lima, Peru" (site to "Barranco District, Lima, Peru") on the Internet Archive also plays incorrect video ("Barranco District, Lima, Peru")

- Michael Robinson archived website: Error message
Evaluation Results

- The subjects were basically satisfied with the captures, but had very many suggestions for improvements with labeling, searching, display, and performance. Most also wanted additional functionality.
- Many of the subjects were confused between captured sites and the Finding Aids for them. In addition, the words “Papers of” in collection titles baffled people when they were looking for recordings, not papers.
- Both users and site owners were unclear about the scope of content that had actually been collected. One site owner expressed disappointment that reviews that they linked to were not captured. And only one subject figured out how to navigate to a suggested “live web” page that had not been not archived.

Functionality requested by users

- Most subjects wanted more metadata displayed. Examples included: displaying a description of the Composers Project and likely contents on the initial start page, display of audio/video run-time instead of file size, description, thumbnails, excerpts for material restricted to onsite use (so that they could decide whether or not they really needed to make a site visit); more fields shown in various displays (both in lists and in links to essence).
- Both site owners responded positively to the idea of providing a site map with a collapsing menu of links.
- Most subjects wanted a search box. And most wanted to be able to immediately sort a multi-column display list by any column of their choosing.
- One subject found it misleading when a restricted object linked to a new page.
- One site owner preferred that their digital objects be organized by project, rather than in an undifferentiated list of every digital object on their site.

Functionality requested for local workstations

- Ability to take screen grabs
- Access to additional browser window
- Preview frame when scrubbing (fast forwarding) through video material
- Use of their own laptop or another window
- Display of time code
- And 2 subjects specifically requested the
  – ability to slow video/audio file to transcribe
  – ability to drop pin/attach notes to specific point in video/audio file

Impact Beyond this Project

- There will be an alternative to Heritrix for capturing streaming media, and Archive-It will ideally 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 have preserved 100++ websites of young composers

Archiving websites containing streaming media: the Music Composer Project

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