Making Web Archiving Work for Streaming Media:
Archiving the Websites of Contemporary Young Composers

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

Making Web Archiving Work for Streaming Media
• Background issues and problems
• The Project
  – Our Technical Collaboration
  – Our Collaboration with Content Creators & restrictions
  – Architectures & Workflows
  – How things may look
  – Evaluation
• Impact beyond this Project

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

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

THE PROJECT
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

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 2016)

- 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

Project Team

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

OUR TECHNICAL COLLABORATION: CRAWLING
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
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

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 had 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 over a year
- Have had fairly stable language in it and some 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—including 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

Recent Development work

- Supplying a separate audio player?
- Hiring a Digital Archivist
- Still working on precise forms of navigation between ArchiveSpace, Archive-It, and richer content within NYU’s digital repository
- 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
HOW THINGS MAY LOOK

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 Queries

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/

NYU Archive of Contemporary Composers Websites API Demo Site

About the Project

Leverage these APIs, web scraping, combinatorial search, and the extensive browser-based search of the Open World Web for musicology. The APIs are not maintained, versioned, or updated. They are intended to be used as starting points for further explorations and developments.

Featured Collections

- Julia Codette Papers
- Julia Knapp Papers (index with more)
- R. Mante Ken Papers
- Michael Friedman Papers

Besser-Berkeley Seminar 1/19/2018

From the Library Finding Aid side
http://dlib.nyu.edu/findingaids/html/fales/mss_479/

Guide to the John Cacicedo Papers MSS 479

Contents:

- Introduction
- John Cacicedo
- Notes
- Provenance
- New York University Manuscripts Collection
- Finding Aid
- Guides

Besser-Berkeley Seminar 1/19/2018

From the Library Finding Aid side

From the Library Finding Aid side (cont)

- Container List
  - Description
  - Notes
  - Provenance
  - New York University Manuscripts Collection
  - Finding Aid
  - Guides

Besser-Berkeley Seminar 1/19/2018

From the Library Finding Aid side

From the Archive-It side
http://archive-it.org/collections/7872

Welcome to the NYU Archive-It!

Avery Fisher Center: Bit Route

Search for Items

Besser-Berkeley Seminar 1/19/2018
From the Archive-It side
https://archive-it.org/collections/7872?
q=http%3A%2F%2Fwww.bitrosie.com&show=SeedVideos&fc=seedId%3A1157594

From any direction, user might need to authenticate

SOME OTHER INTERNAL TRACKING

Crawl Records

EVALUATION
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

• Jan 2018—Schedule one-on-one interviews with sets of composers and Researchers
• Feb-Mar 2017—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 2017—Construction of Evaluation Summary containing the list of improvements/changes that should be made to the Archiving project
• June-Aug 2017—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

Making Web Archiving Work for Streaming Media

• http://besser.tsoa.nyu.edu/howard/Talks/
• http://archive.org/~nlevitt/reveal.js/
• http://composers.dlib.nyu.edu/
• https://rg3.github.io/youtube-dl/supportedsites.html