# Archiving websites containing streaming media: the Music Composer Project

Howard Besser, NYU

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

# Archiving websites containing streaming media: the Music Composer Project

- The Problem with Heritrix and Archive-It
- The Project
  - Our Technical Collaboration
  - Our Collaboration with Content Creators & restrictions
  - Architectures & Workflows
  - How things look
  - Evaluation
- Impact beyond this Project

- Caveat I: This is an in-progress report; the project is unfinished
- Caveat II: I am not involved in system architecture & hand-offs, so may not be able to answer detailed questions in these areas

# 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

# Archive-It Issues w/Streaming Media

# Archive-It Issues w/Streaming Media

# Archive-It screenshots generated as part of our project-

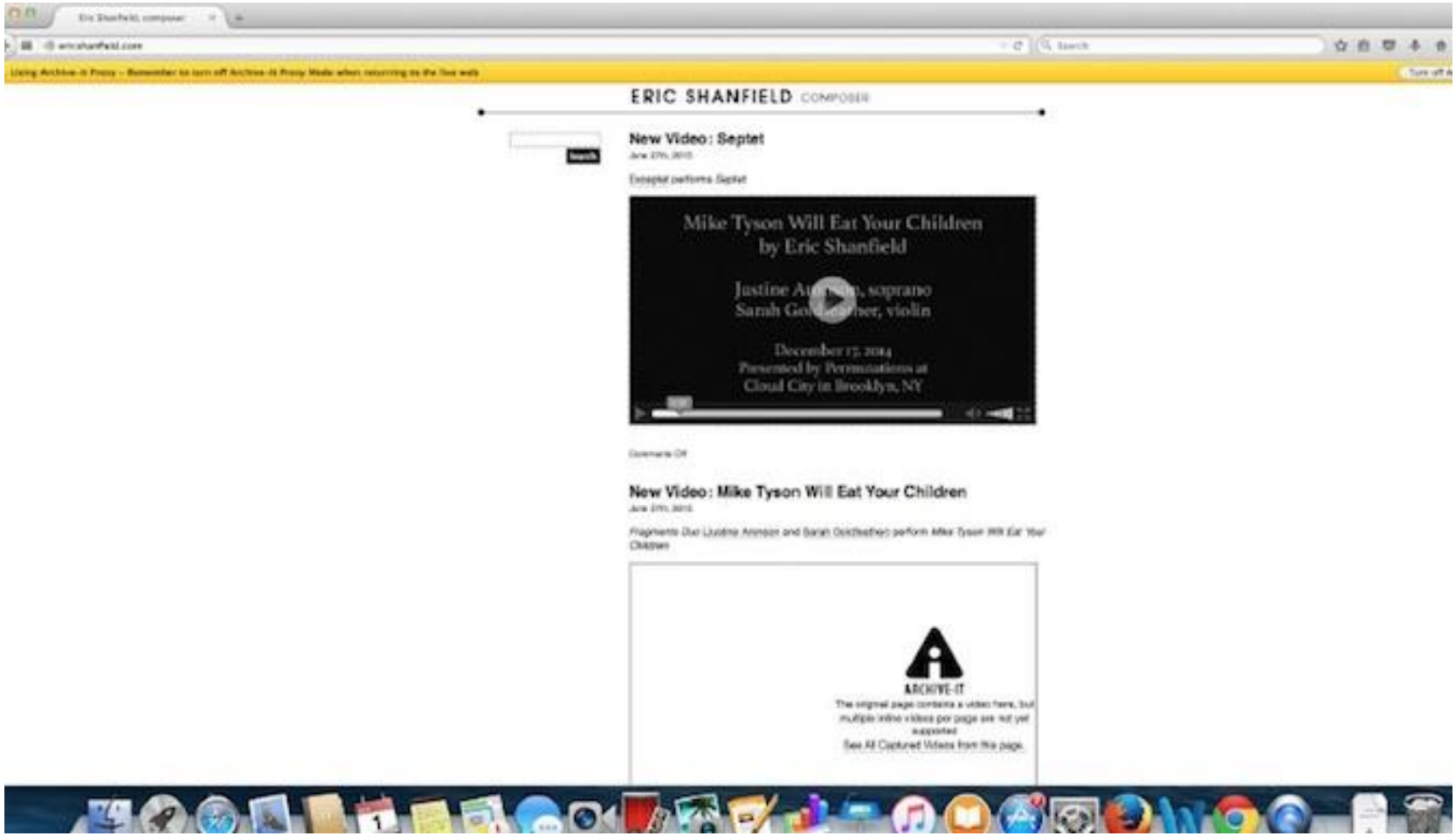
- By Lorena Ramirez-Løpez

# Archive-It Issues w/Streaming Media

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

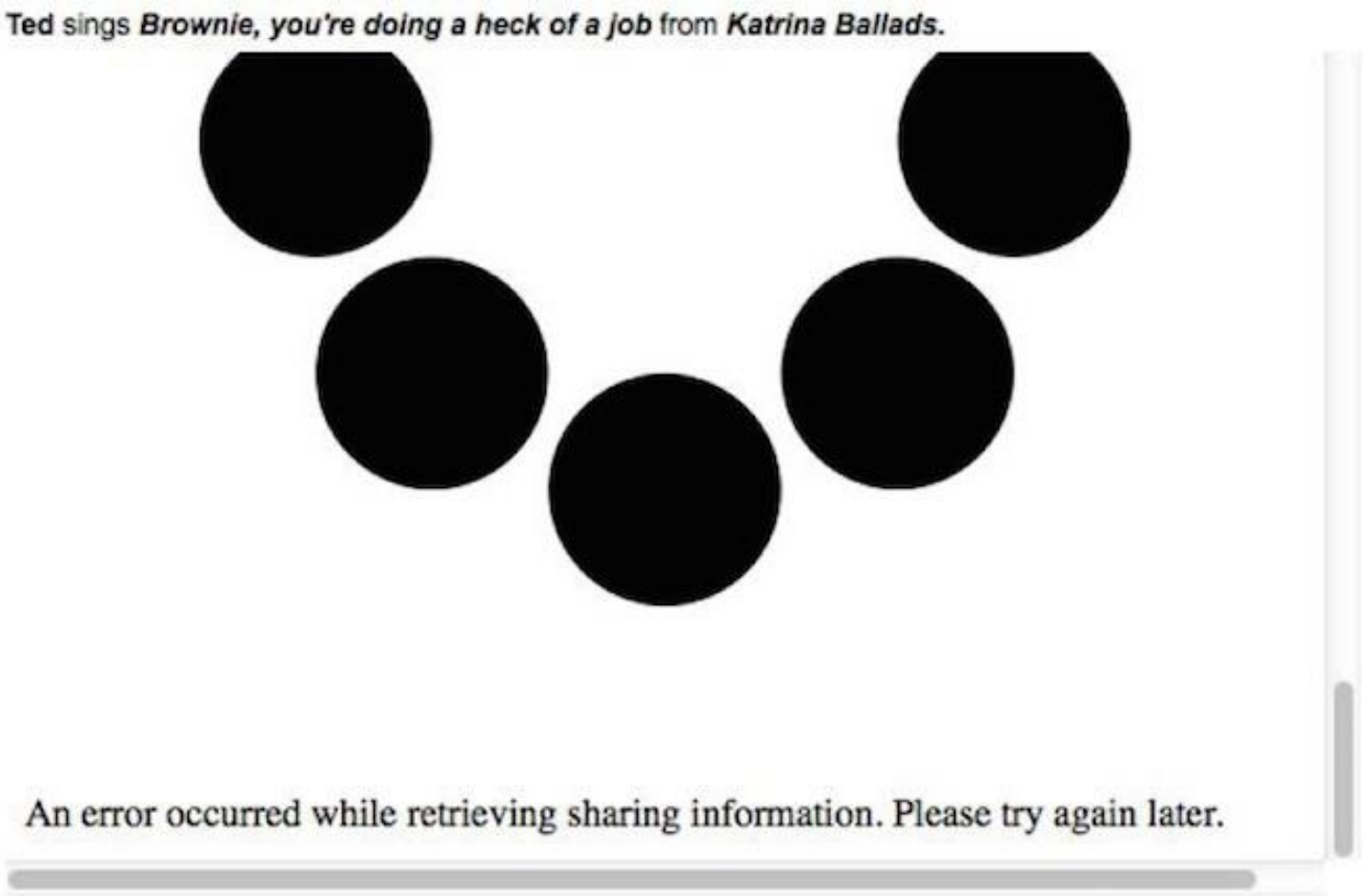
# Archive-It Issues w/Streaming Media

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



# 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 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

# Other Issues we're trying to solve

- Discovering URLs generated by Javascript

# THE PROJECT

# Archiving Composer Websites

http://www.nyu.edu/about/news-publications/news/2015/03/27/nyu-libraries-to-team-with-internet-archive-to-preserve-high-quality-musical-content-on-the-web.html

- 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 btwn 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 btwn 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 w websites generated through site building platforms such as Squarespace
- Find ways to fix those errors

# Some methods used

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

# Website Infrastructure encountered

| NYU Archive-It seed list | | | VIDEO | | | AUDIO | | |
|---|---|---|---|---|---|---|---|---|
| er's | URL | Type of website | Vimeo? | YouTube? | Embedded video | SoundCloud? | dded | Other |
| Carson Cc | http://carsoncooman.com/ | | | | | Y | Y | PDFs |
| Chris Fish | http://www.cflmusic.com/ | | | Y | | Y | | Bandcamp |
| Joby Talbc | http://www.jobytalbot.com | | | | | | Y | |
| Florent Gh | http://www.florentghys.c | Wix | Y | Y | | | Unsure | |
| Guy Baras | http://guybarash.com/ | Wordpress | | Y | Y | | | |
| Eric Shanf | http://ericshanfield.com/ | Wordpress | | Y | Y | | Y | |
| Warren Bu | http://www.warrenburt.c | Wordpress | | | | | Y | |
| Rand Stei | http://rand.info | | Y | | | | | |
| Kyle Gann | http://www.kylegann.con | Custom designed | | | | | Y | PDFs |
| John Corig | http://www.johncorigliane | Custom designed | | Y | | | Y | PDFs |
| Erling Wol | http://www.erlingwold.com/ | | | Y | Y | Y | | PDFs |
| Alex Shap | http://www.alexshapiro.org/ | | | | | | Y | PDFs |
| Paul J. Bo | http://pauljbotelho.com/ | | | | | | Y | with option to |
| Wendy Ms | http://www.wendyma | Custom designed | | | Y | | Y | downloaded |
| David Eva | http://davidevanthomas. | Custom designed | | | | | Y | |
| Tarik O'Re | http://tarikoregan.cor | Custom designed | | | | | Y | PDFs, Word docs |
| James Pri | http://jamesprimosch.co | Wordpress | Y | Y | | | Y | PDFs, Photos |
| Andrew M | http://www.andrewm | Custom designed | | | Y | | Y | |
| Timo Andr | http://www.andres.co | Custom designed | | Y | | | Y | PDFs. Word docs |
| Judith Sha | http://judithshatin.co | Custom designed | Y | Y | | | | PDFs |
| Augusta R | http://www.augustareadthomas.com/ | | | Y | | | Y | |
| Eric Lemn | http://ericlemmon.net/ | | | Y | | | Y | PDFs |
| Lauren Be | http://www.laurenbernofsky.com/ | | | Y | | | Y | PDFs |
| Oscar Bet | http://www.oscarbettison.com/ | | | Y | | Y | Y | Bandcamp |
| Zeena Par | http://www.zeenaparkins.com/ | | | | | | Y | |
| Jon Applet | http://www.appletonjon.com/ | | Y | | Y | | Y | |
| Andrea La | http://reloadsanear.com/ | | Y | Y | Y | Y | | Bandcamp |
| Mary Jane | http://www.mjleach.c | Custom designed | Y | Y | | Y | | ReverbNation, PDFs |
| Sarah Kirk | http://sarahkirklandsn | Custom designed | | Y | Y | | Y | Photos |
| David B. D | http://www.dbdoty.cc | Custom designed | | Y | Y | | Y | PDFs |
| Clara Maic | http://www.claramaid | Custom designed | | | | | Y | Photos |
| Laurie Anc | http://www.laurieanderson.com/ | | Y | Y | Y | | Y | PDFs, Photos |
| Anthony C | http://www.anthonycornicello.com/ | | | | | | | audio |
| Robert Ho | http://www.roberthonstein.com/ | | Y | | | Y | | kit |
| Keith Kirch | http://keithkirchoff.com/ | Wix | | Y | Y | Y | Y | player |

# 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 Mennerich (Lib/ACM/DLTS)
- David Millman (Lib/DLTS)
- Courtney Mumma (Internet Archive)
- Robin Preiss (Lib/AFC)
- Lorena Ramirez (MIAP) ---*special thanks!*
- Michael Stoller (Lib/C&RS)
- Kent Underwood (Lib/AFC)
- Chela Scott Weber (Lib/AFC)--departed

# OUR TECHNICAL COLLABORATION: CRAWLING

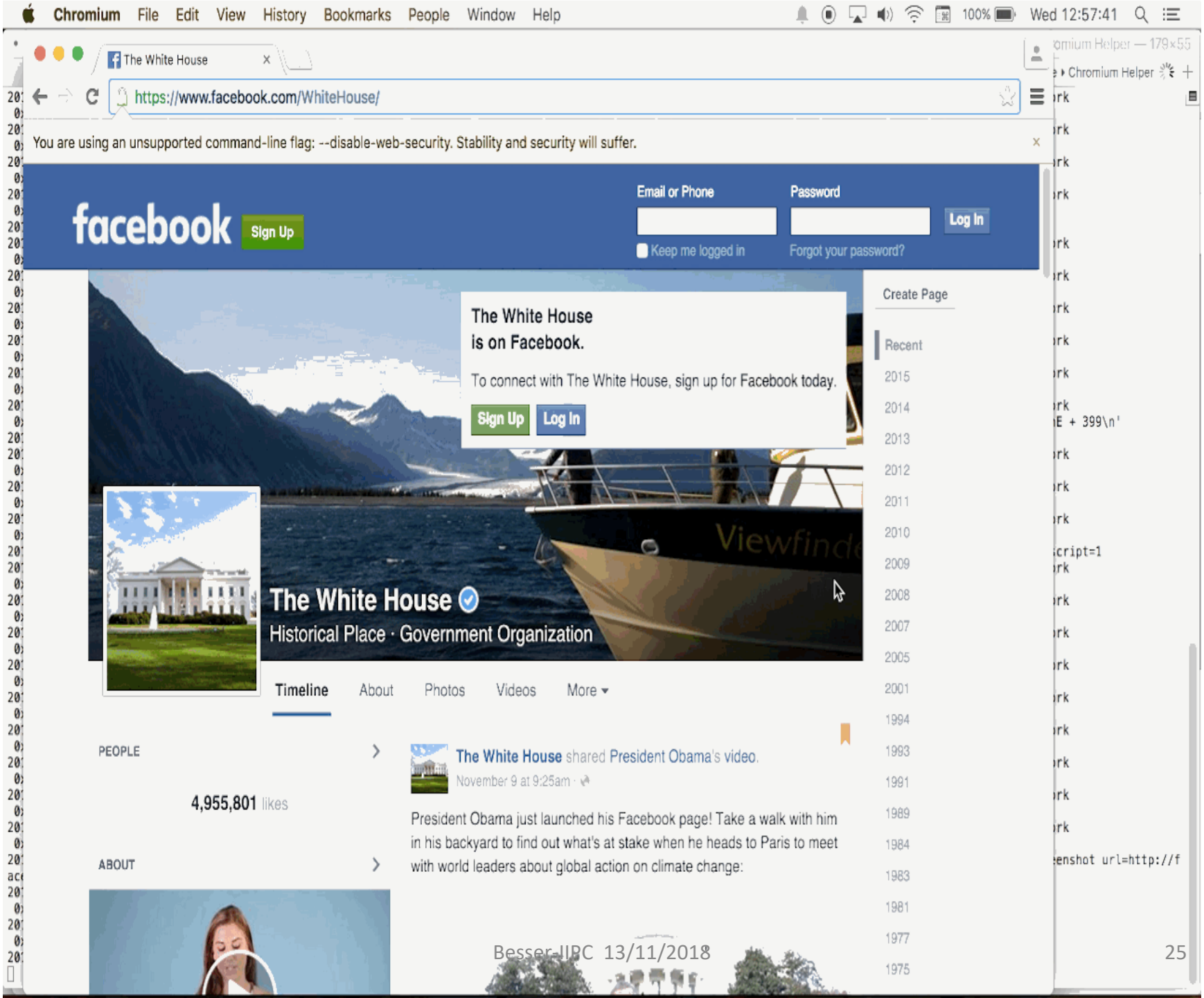
# NYU/IA Collaboration

# Traditional Crawlers

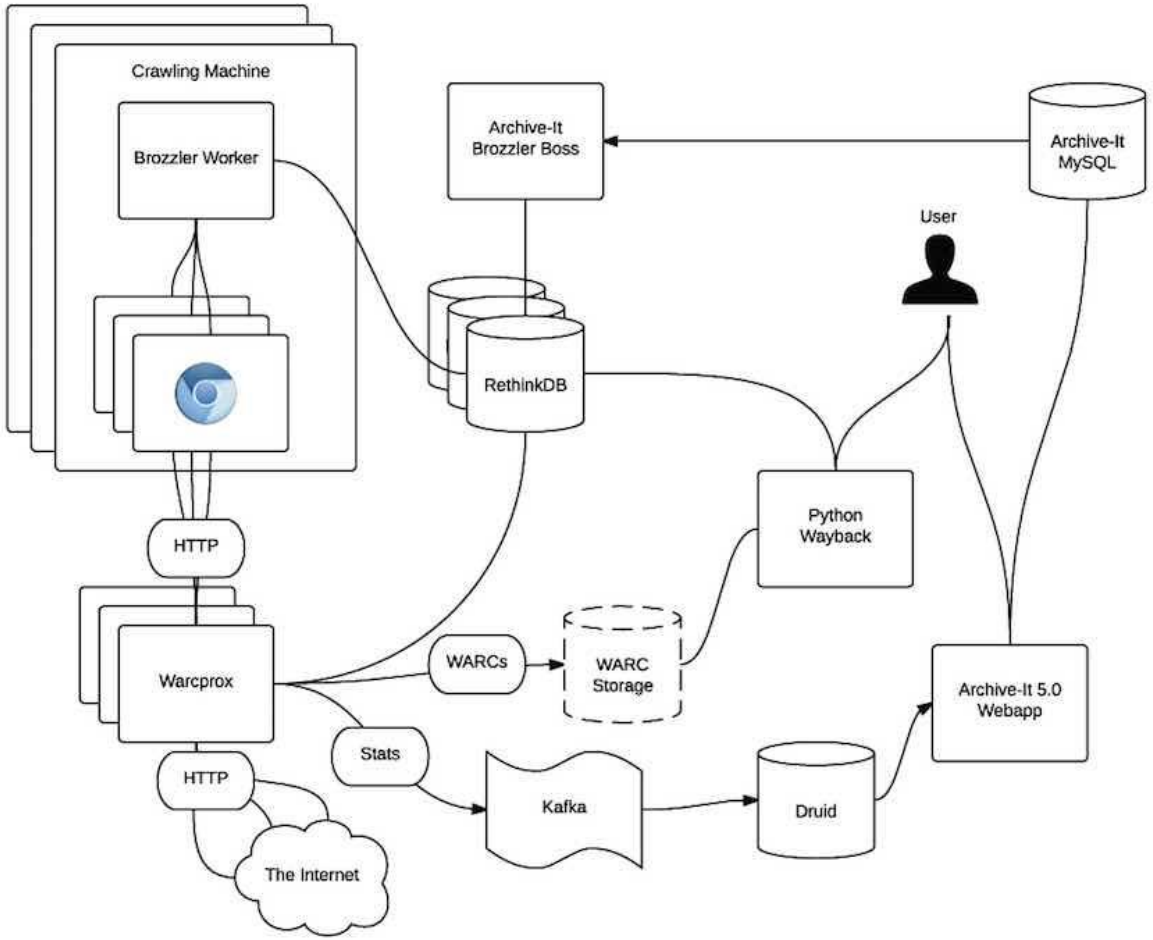
- Archive-It and other web archives use Heritrix
- Follow links, capture most web content
- Less successful with streaming video and dynamic content executed in the browser
- Umbra helps

# “browser” | “crawler” = BROZZLER

Logo: Noah Levitt


Chromium File Edit View History Bookmarks People Window Help
Wed 12:57:41
The White House
https://www.facebook.com/WhiteHouse/
You are using an unsupported command-line flag: --disable-web-security. Stability and security will suffer.
facebook
Sign Up
Email or Phone
Password
Log In
Keep me logged in
Forgot your password?
The White House is on Facebook.
To connect with The White House, sign up for Facebook today.
Sign Up
Log In
Create Page
Recent
2015
2014
2013
2012
2011
2010
2009
2008
2007
2005
2001
1994
1993
1991
1989
1984
1983
1981
1977
1975
The White House
Historical Place · Government Organization
Timeline
About
Photos
Videos
More
PEOPLE
4,955,801 likes
ABOUT
The White House shared President Obama's video.
November 9 at 9:25am
President Obama just launched his Facebook page! Take a walk with him in his backyard to find out what's at stake when he heads to Paris to meet with world leaders about global action on climate change:

# 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://rg3.github.io/youtube-dl/supportedsites.html

**youtube-dl** **Supported sites**

Here's is the list of all the supported sites, ordered alphabetically:

- **1tv**: Первый канал
- **1up.com**
- **20min**
- **220.ro**
- **23video**
- **24video**
- **3qsdn**: 3Q SDN
- **3sat**
- **4tube**
- **56.com**
- **5min**
- **6play**
- **7plus**
- **8tracks**
- **91porn**
- **9c9media**
- **9c9media:stack**
- **9gag**
- **9now.com.au**
- **abc.net.au**
- **abc.net.au:iview**
- **abcnews**
- **abcnews:video**
- **abcotvs**: ABC Owned Television Stations
- **abcotvs:clips**
- **AcademicEarth:Course**
- **acast**

# 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

tentative language

- 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

# 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 btwn 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 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

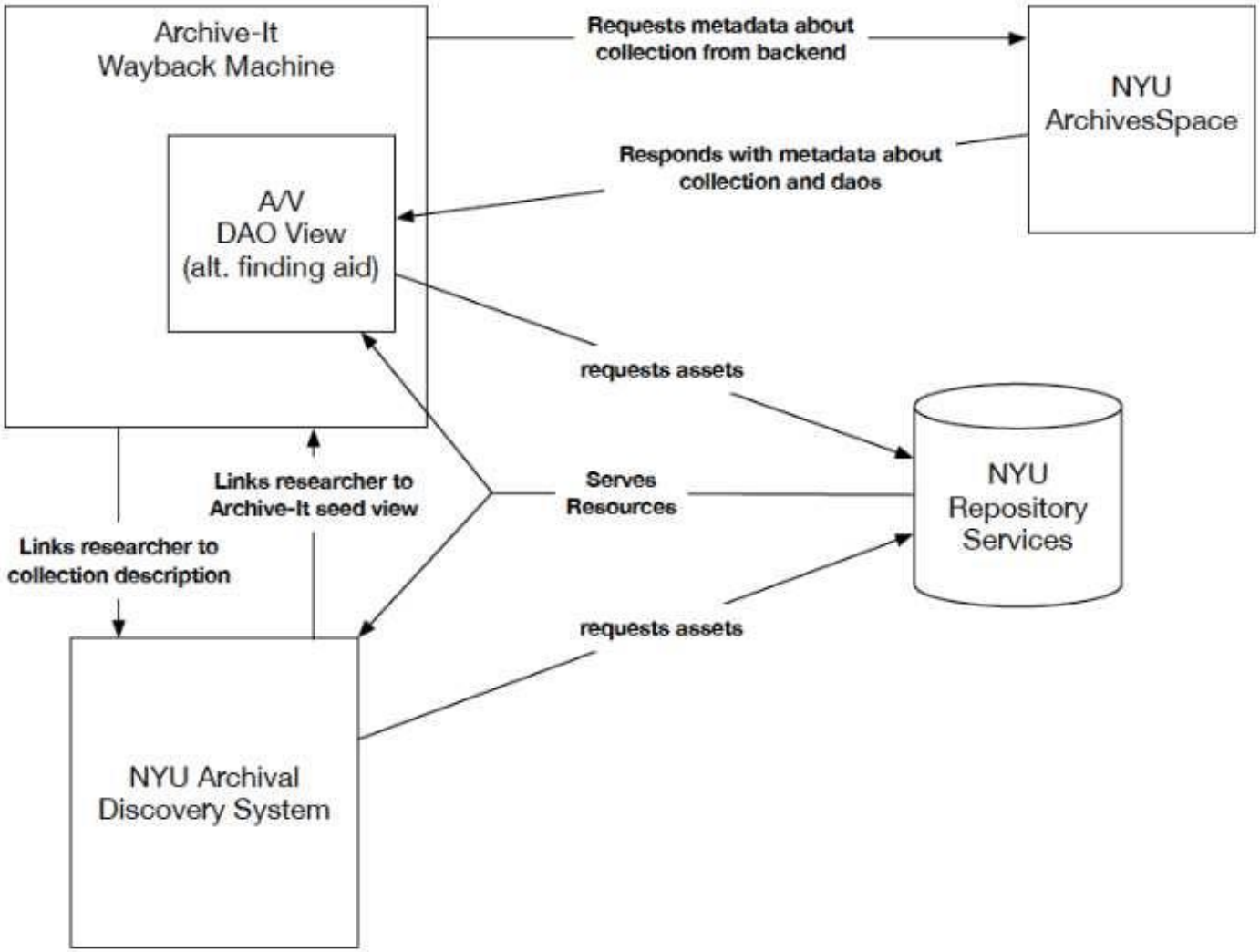
# HOW THINGS MAY LOOK

# 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)

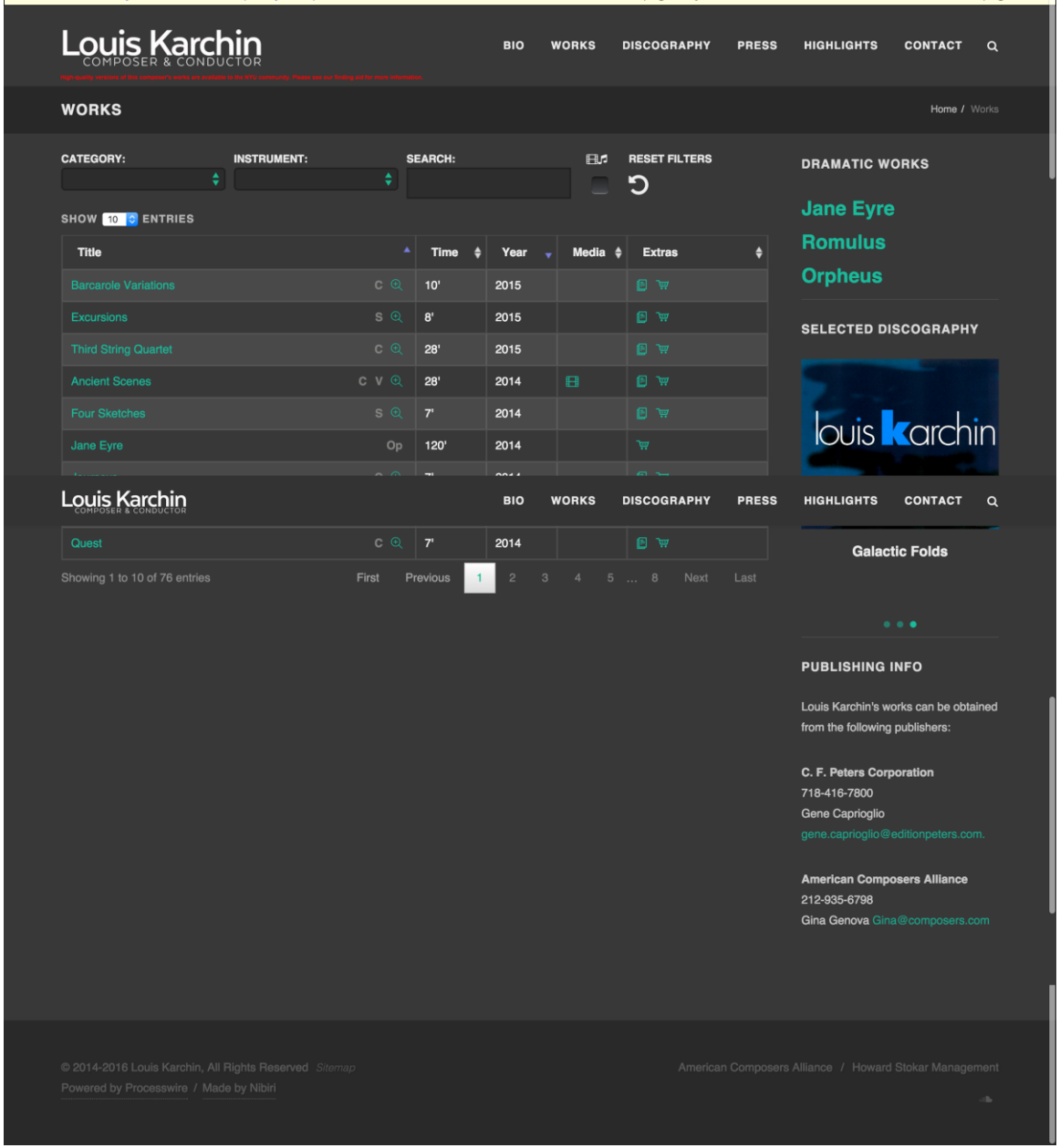
# One option for Queries



# 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)-

You are viewing an archived web page, collected at the request of New York University using Archive-It. This page was captured on 3:51:47 Feb 04, 2016, and is part of the New York University Archive of Contemporary Composers' Websites collection. The information on this web page may be out of date. See All versions of this archived page. hide



# Demo from API side

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

## NYU Archive of Contemporary Composers Websites API Demo Site

### About the Project

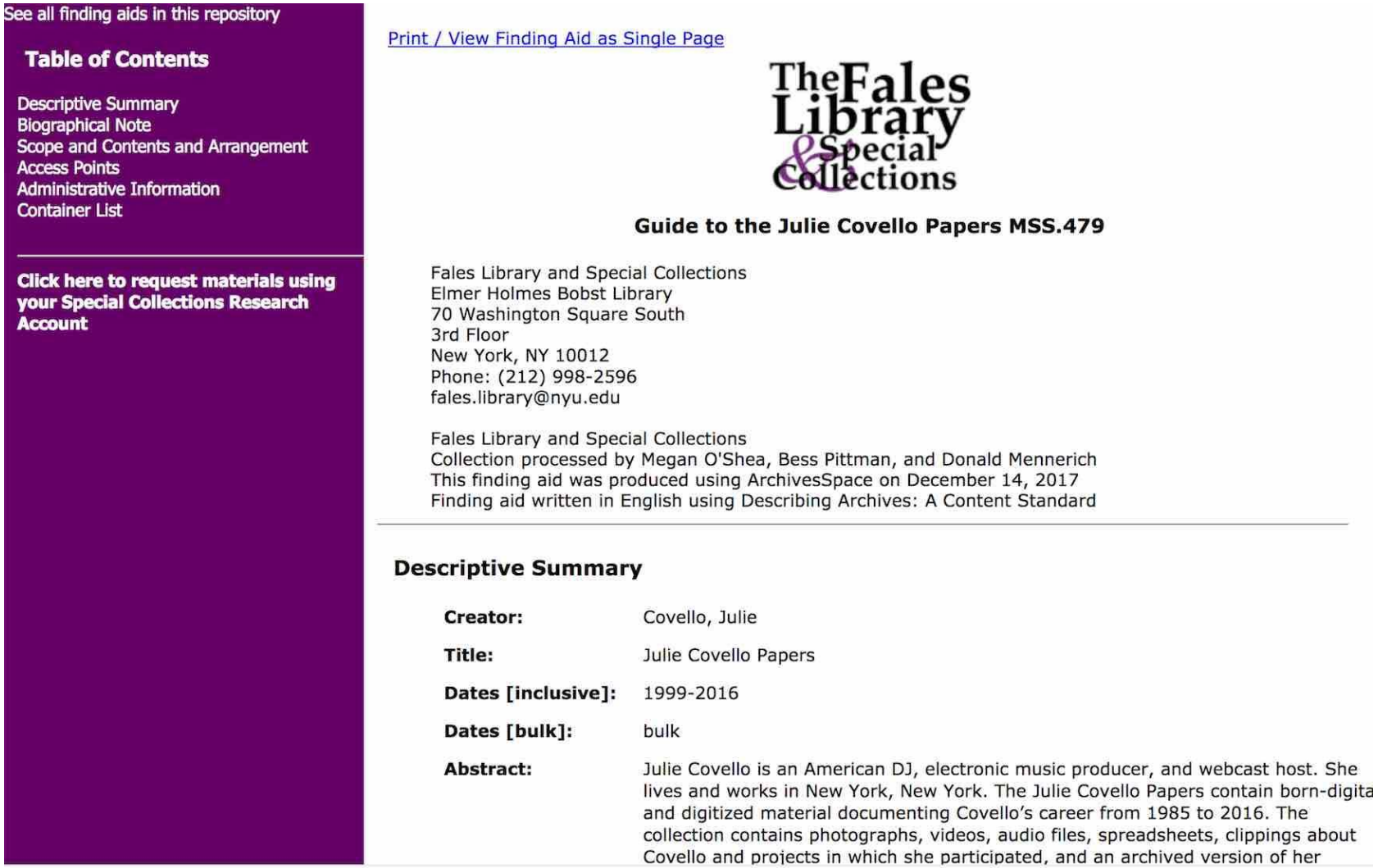
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

### Featured Collections

- Adele Fournet Papers
- Adele Fournet Papers on the Bit Rosie Web Series
- Julie Covello Papers
- Julie Kathryn Franco I Am Snow Angel Papers
- K Marie Kim Papers
- Michael Robinson Papers

# From the Library Finding Aid side

http://dlib.nyu.edu/findingaids/html/fales/mss_479/



See all finding aids in this repository

**Table of Contents**


Descriptive Summary
Biographical Note
Scope and Contents and Arrangement
Access Points
Administrative Information
Container List


**Click here to request materials using your Special Collections Research Account**

Print / View Finding Aid as Single Page

The Fales Library & Special Collections

**Guide to the Julie Covello Papers MSS.479**


Fales Library and Special Collections
Elmer Holmes Bobst Library
70 Washington Square South
3rd Floor
New York, NY 10012
Phone: (212) 998-2596
fales.library@nyu.edu


Fales Library and Special Collections
Collection processed by Megan O'Shea, Bess Pittman, and Donald Mennerich
This finding aid was produced using ArchivesSpace on December 14, 2017
Finding aid written in English using Describing Archives: A Content Standard

**Descriptive Summary**

| | |
|---|---|
| **Creator:** | Covello, Julie |
| **Title:** | Julie Covello Papers |
| **Dates [inclusive]:** | 1999-2016 |
| **Dates [bulk]:** | bulk |
| **Abstract:** | Julie Covello is an American DJ, electronic music producer, and webcast host. She lives and works in New York, New York. The Julie Covello Papers contain born-digita and digitized material documenting Covello's career from 1985 to 2016. The collection contains photographs, videos, audio files, spreadsheets, clippings about Covello and projects in which she participated, and an archived version of her |

# From the Library Finding Aid side

(cont)

| | |
|---|---|
| **Dates [bulk]:** | bulk |
| **Abstract:** | Julie Covello is an American DJ, electronic music producer, and webcast host. She lives and works in New York, New York. The Julie Covello Papers contain born-digita and digitized material documenting Covello's career from 1985 to 2016. The collection contains photographs, videos, audio files, spreadsheets, clippings about Covello and projects in which she participated, and an archived version of her website, campjulie.com. The website was first captured in May 2017. The collection documents her solo work as DJ Shakey, as the host of *The Near Death Experience Show* on pseudo.com, and as a collaborator in Asstronauts, Daughter of Darrr, FreeBassBK, The Funky Beatbots, and Opal Crocus. |
| **Quantity:** | 73.72 Gigabytes in 508 computer files and 1 archived website. |
| **Quantity:** | 1 websites in 1 archived website. |
| **Quantity:** | 2.88 Gigabytes in 430 graphic files. |
| **Quantity:** | 2.34 Gigabytes in 42 audio files. |
| **Quantity:** | 67.5 Gigabytes in 24 video files. |
| **Quantity:** | 10.9 Megabytes in 9 documents. |
| **Quantity:** | 175 kilobytes in 3 spreadsheets. |
| **Language:** | Materials in English. |
| **Call Phrase:** | MSS.479 |

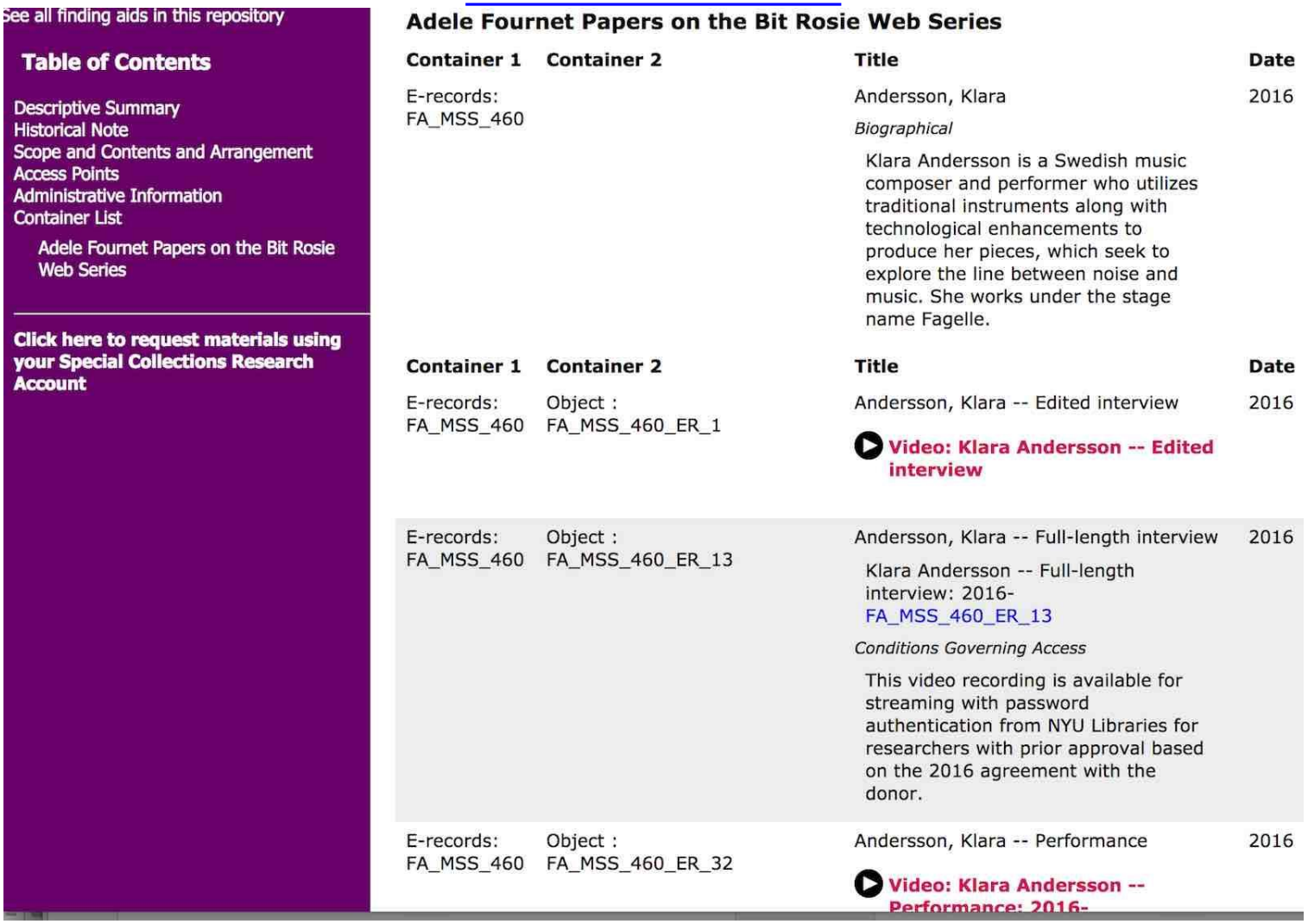
# From the Library Finding Aid side

(Container List)

| | | | |
|---|---|---|---|
| E-records: FA_MSS_479 | Object : https://wayback.archive-it.org/4049/*/http://www.campjulie.com | Archived website<br>Archived Website of Julie Covello https://wayback.archive-it.org/4049/*/http://www.campjulie.com | March 2017-ongoing |
| **Container 1** | **Container 2** | **Title** | **Date** |
| E-records: FA_MSS_479 | Object : FA_MSS_479_21 | Asstronauts -- Photographs of performances<br>Asstronauts photographs of performances: 2014-2015 FA_MSS_479_ER_21<br>*Conditions Governing Access*<br>This material is available for streaming with password authentication from NYU Libraries for researchers with prior approval based on the 2016 agreement with the donor. | 2014-2015 |
| E-records: FA_MSS_479 | Object : FA_MSS_479_ER_41 | Asstronauts -- Video recordings of performances<br>Asstronauts video recordings of performances: 2014-2015 FA_MSS_479_ER_41<br>*Conditions Governing Access*<br>This material is available for streaming with password authentication from NYU Libraries for researchers with prior approval based on the 2016 agreement with the donor. | 2014-2015 |

# From the Library Finding Aid side

http://dlib.nyu.edu/findingaids/html/fales/mss_460/dscaspace_7951feea619b6c41436c556e0674d1c8.html



See all finding aids in this repository

**Table of Contents**


Descriptive Summary
Historical Note
Scope and Contents and Arrangement
Access Points
Administrative Information
Container List
Adele Fournet Papers on the Bit Rosie Web Series

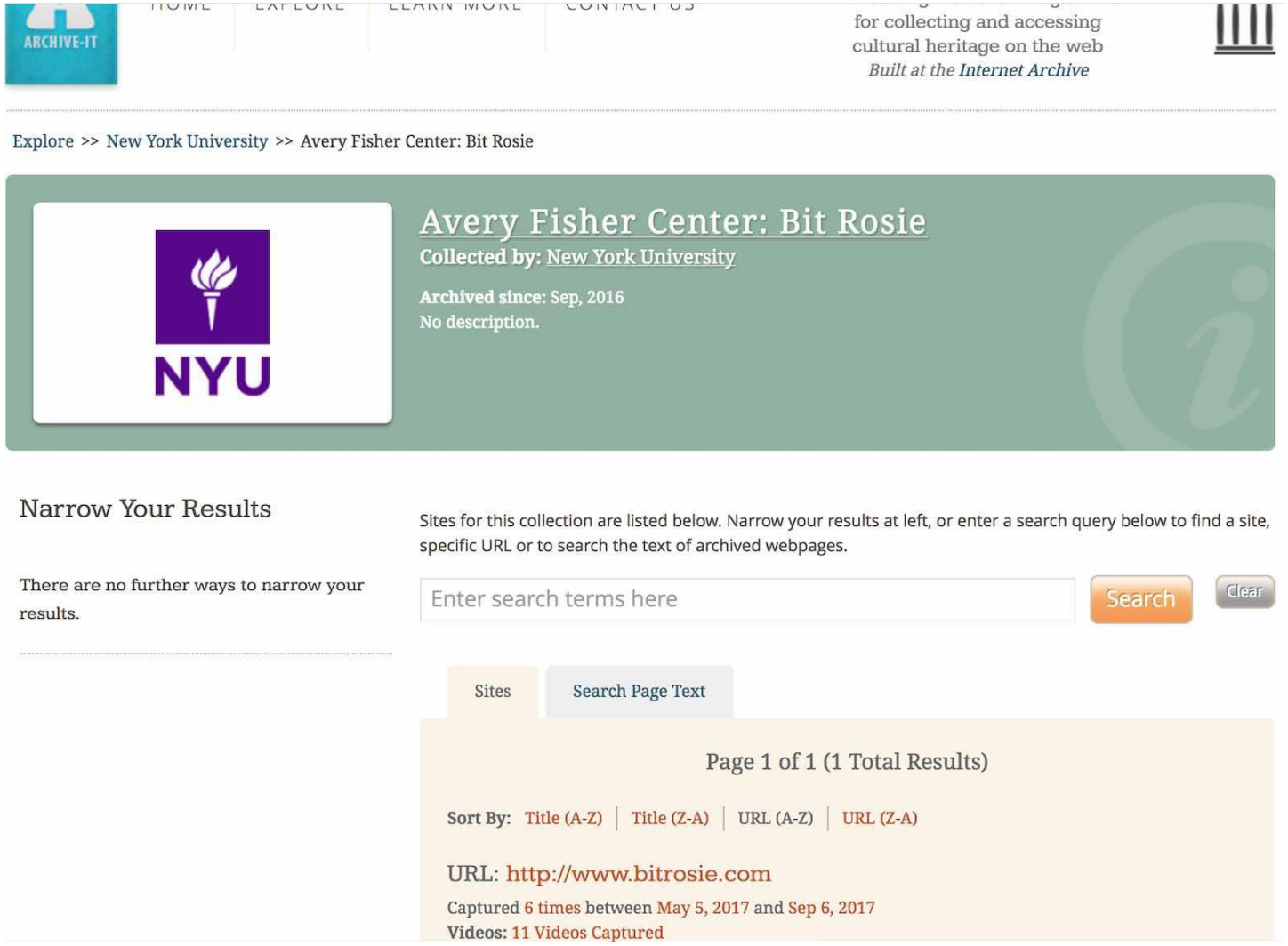

**Click here to request materials using your Special Collections Research Account**

**Adele Fournet Papers on the Bit Rosie Web Series**

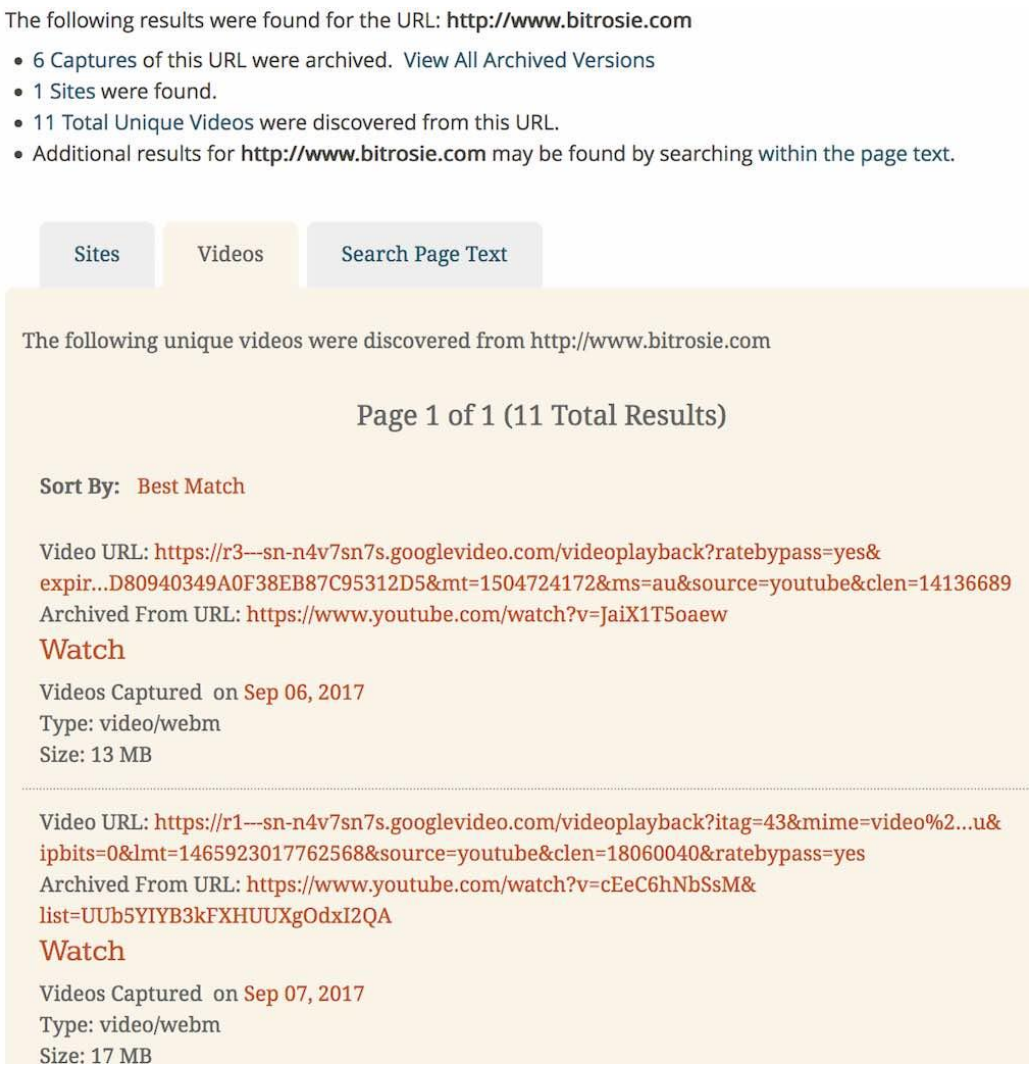
| Container 1 | Container 2 | Title | Date |
|---|---|---|---|
| E-records: FA_MSS_460 | | Andersson, Klara<br>*Biographical*<br>Klara Andersson is a Swedish music composer and performer who utilizes traditional instruments along with technological enhancements to produce her pieces, which seek to explore the line between noise and music. She works under the stage name Fagelle. | 2016 |

| Container 1 | Container 2 | Title | Date |
|---|---|---|---|
| E-records: FA_MSS_460 | Object : FA_MSS_460_ER_1 | Andersson, Klara -- Edited interview<br>**Video: Klara Andersson -- Edited interview** | 2016 |
| E-records: FA_MSS_460 | Object : FA_MSS_460_ER_13 | Andersson, Klara -- Full-length interview<br>Klara Andersson -- Full-length interview: 2016-<br>FA_MSS_460_ER_13<br>*Conditions Governing Access*<br>This video recording is available for streaming with password authentication from NYU Libraries for researchers with prior approval based on the 2016 agreement with the donor. | 2016 |
| E-records: FA_MSS_460 | Object : FA_MSS_460_ER_32 | Andersson, Klara -- Performance<br>**Video: Klara Andersson -- Performance: 2016-** | 2016 |

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



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



The following results were found for the URL: **http://www.bitrosie.com**

- 6 Captures of this URL were archived. View All Archived Versions
- 1 Sites were found.
- 11 Total Unique Videos were discovered from this URL.
- Additional results for **http://www.bitrosie.com** may be found by searching within the page text.

Sites | Videos | Search Page Text

The following unique videos were discovered from http://www.bitrosie.com

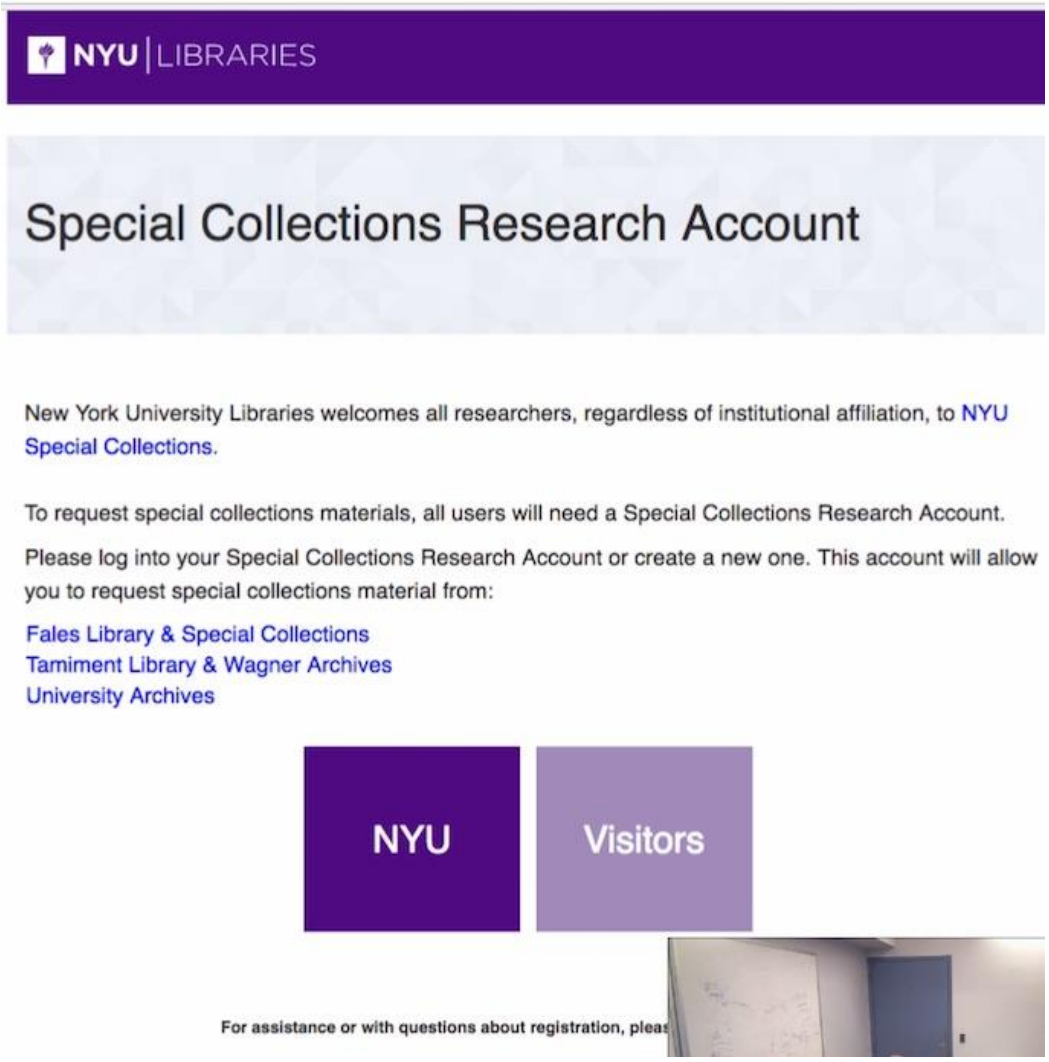
Page 1 of 1 (11 Total Results)

**Sort By:** Best Match

Video URL: https://r3---sn-n4v7sn7s.googlevideo.com/videoplayback?ratebypass=yes&expir...D80940349A0F38EB87C95312D5&mt=1504724172&ms=au&source=youtube&clen=14136689
Archived From URL: https://www.youtube.com/watch?v=JaiX1T5oaew
Watch
Videos Captured on Sep 06, 2017
Type: video/webm
Size: 13 MB

Video URL: https://r1---sn-n4v7sn7s.googlevideo.com/videoplayback?itag=43&mime=video%2...u&ipbits=0&lmt=1465923017762568&source=youtube&clen=18060040&ratebypass=yes
Archived From URL: https://www.youtube.com/watch?v=cEeC6hNbSsM&list=UUb5YIYB3kFXHUUXgOdxI2QA
Watch
Videos Captured on Sep 07, 2017
Type: video/webm
Size: 17 MB

# From any direction, user might need to authenticate



# SOME OTHER INTERNAL TRACKING

| | | | | |
|---|---|---|---|---|
| faculty<br>URL: EAD<br>URL: HTML | caterpillar<br>URL: EAD<br>URL: HTML | aia_047<br>URL: EAD<br>URL: HTML | blashfield<br>URL: EAD<br>URL: HTML | arc_025_cortelyou<br>URL: EAD<br>URL: HTML |
| falnes<br>URL: EAD<br>URL: HTML | champagne<br>URL: EAD<br>URL: HTML | aia_048<br>URL: EAD<br>URL: HTML | blumenthal<br>URL: EAD<br>URL: HTML | arc_026_townsend<br>URL: EAD<br>URL: HTML |
| fas<br>URL: EAD<br>URL: HTML | charyn<br>URL: EAD<br>URL: HTML | aia_049<br>URL: EAD<br>URL: HTML | bob_adelman<br>URL: EAD<br>URL: HTML | arc_027_volck<br>URL: EAD<br>URL: HTML |
| ferris<br>URL: EAD<br>URL: HTML | chelsea<br>URL: EAD<br>URL: HTML | aia_051<br>URL: EAD<br>URL: HTML | botein<br>URL: EAD<br>URL: HTML | arc_028_eberhard_faber<br>URL: EAD<br>URL: HTML |
| ford<br>URL: EAD<br>URL: HTML | cjournal<br>URL: EAD<br>URL: HTML | aia_052<br>URL: EAD<br>URL: HTML | bourne<br>URL: EAD<br>URL: HTML | arc_029_de_martini_navy_yard<br>URL: EAD<br>URL: HTML |
| founding<br>URL: EAD<br>URL: HTML | collun<br>URL: EAD<br>URL: HTML | aia_053<br>URL: EAD<br>URL: HTML | bowers<br>URL: EAD<br>URL: HTML | arc_030_hamilton<br>URL: EAD<br>URL: HTML |
| freidsoneliot<br>URL: EAD<br>URL: HTML | cooper<br>URL: EAD<br>URL: HTML | aia_054<br>URL: EAD<br>URL: HTML | boyd<br>URL: EAD<br>URL: HTML | arc_031_flatbush<br>URL: EAD<br>URL: HTML |
| frost<br>URL: EAD<br>URL: HTML | corno<br>URL: EAD<br>URL: HTML | aia_055<br>URL: EAD<br>URL: HTML | boysclub<br>URL: EAD<br>URL: HTML | arc_032_hispanic_communities_documentati<br>URL: EAD<br>URL: HTML |
| gadsby<br>URL: EAD<br>URL: HTML | cover<br>URL: EAD<br>URL: HTML | aia_056<br>URL: EAD<br>URL: HTML | bracklow<br>URL: EAD<br>URL: HTML | arc_033_graham<br>URL: EAD<br>URL: HTML |
| gallatinaward<br>URL: EAD<br>URL: HTML | creativetime<br>URL: EAD<br>URL: HTML | aia_057<br>URL: EAD<br>URL: HTML | bradley<br>URL: EAD<br>URL: HTML | arc_034_kranjec<br>URL: EAD<br>URL: HTML |
| gallatinpapers<br>URL: EAD<br>URL: HTML | dames<br>URL: EAD<br>URL: HTML | aia_059<br>URL: EAD<br>URL: HTML | bradstreet<br>URL: EAD<br>URL: HTML | arc_035_hamilton_literary_association<br>URL: EAD<br>URL: HTML |
| gantt<br>URL: EAD<br>URL: HTML | darcangelo<br>URL: EAD<br>URL: HTML | aia_060<br>URL: EAD<br>URL: HTML | brady<br>URL: EAD<br>URL: HTML | arc_036_haviland<br>URL: EAD<br>URL: HTML |
| gitlow<br>URL: EAD<br>URL: HTML | darinka<br>URL: EAD<br>URL: HTML | aia_061<br>URL: EAD<br>URL: HTML | broadwaytabernacle<br>URL: EAD<br>URL: HTML | arc_037_hayes<br>URL: EAD<br>URL: HTML |
| glee<br>URL: EAD<br>URL: HTML | davidovich<br>URL: EAD<br>URL: HTML | aia_062<br>URL: EAD<br>URL: HTML | brooklyn_friends_school_mural<br>URL: EAD<br>URL: HTML | arc_038_heneolotter<br>URL: EAD<br>URL: HTML |
| godt<br>URL: EAD<br>URL: HTML | dean<br>URL: EAD<br>URL: HTML | aia_064<br>URL: EAD<br>URL: HTML | brown<br>URL: EAD<br>URL: HTML | arc_039_hegeman<br>URL: EAD<br>URL: HTML |
| goldfeld<br>URL: EAD<br>URL: HTML | demoura<br>URL: EAD<br>URL: HTML | aia_066<br>URL: EAD<br>URL: HTML | browning<br>URL: EAD<br>URL: HTML | arc_040_harnickell<br>URL: EAD<br>URL: HTML |
| gorn | desana | aia_068 | burling | arc_041_grace_hillard |

# Crawl Records

ARCHIVE-IT | Home | Collections | Crawls | Archives | ARS

Home / Collections / Active

## New York University

Active | Inactive

Create a Collection...

Active Collection List (92 Active Collections)

Type to Filter Active Collections

Download Collection List

| Collection Name | Last Crawl | Data (this period) | Docs (this period) | Data (all time) | Docs (all time) | Active Seeds |
|---|---|---|---|---|---|---|
| African American Issues | #469778: Oct 21, 2017 | 312.6 GB | 2,221,510 | ••• | ••• | ••• |
| All the Russias' blog | ••• | 8.1 GB | 428,211 | ••• | ••• | ••• |
| APA Yellow Peril | ••• | 0 bytes | 0 | ••• | ••• | ••• |
| Avery Fisher Center: Adele Fournet | ••• | 0 bytes | 0 | ••• | ••• | ••• |
| Avery Fisher Center: Bit Rosie | ••• | 1.9 GB | 13,382 | ••• | ••• | ••• |
| Avery Fisher Center: Frank London | ••• | 596.9 MB | 34,948 | ••• | ••• | ••• |
| Avery Fisher Center: Guy Klucevsek | ••• | 1.2 GB | 2,110 | ••• | ••• | ••• |
| Avery Fisher Center: Julie Covello | ••• | 32.5 MB | 382 | ••• | ••• | ••• |
| Avery Fisher Center: Julie Kathryn | ••• | 66.3 MB | 564 | ••• | ••• | ••• |
| Avery Fisher Center: K. Marie Kim | ••• | 104.6 MB | 9,470 | ••• | ••• | ••• |
| Avery Fisher Center: Marc Ribot | ••• | 2.4 GB | 4,414 | ••• | ••• | ••• |
| Avery Fisher Center: Michael Robin | ••• | 0 bytes | 0 | | | |
| Avery Fisher Center: Roy Nathanso | ••• | 884.8 MB | 19,343 | | | |
| Avery Fisher Center: Shelley Hirsch | ••• | 29.5 MB | 470 | | | |
| DLTS: ArchivesSpace Wiki | ••• | 0 bytes | 0 | | | |
| Edible Magazines (New York) | ••• | 59.1 GB | 2,615,476 | | | |
| El clasico | ••• | 0 bytes | 0 | | | |
| EU and Ukraine | ••• | 0 bytes | 0 | | | |

# EVALUATION

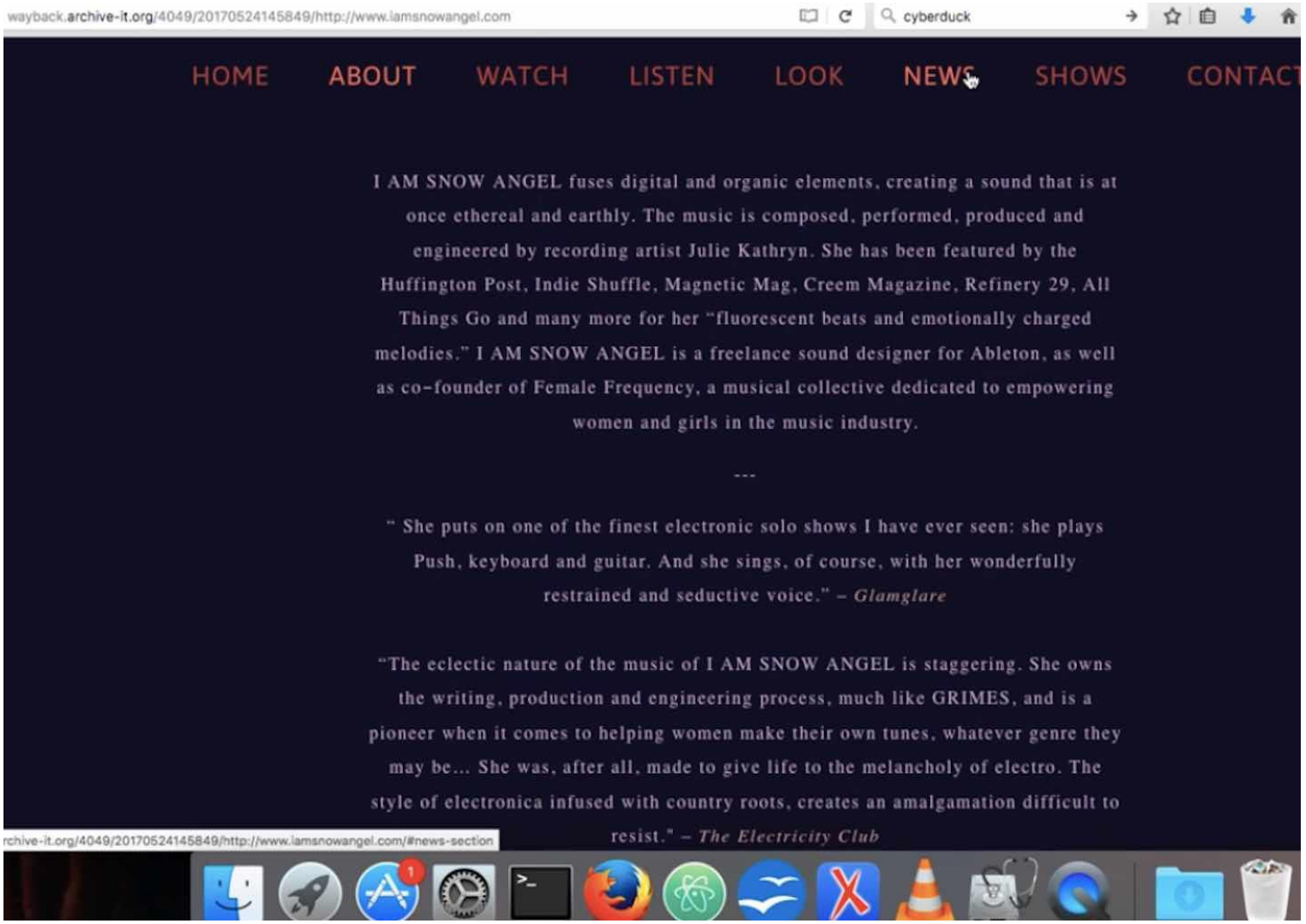
# Evaluation for Improvement

- 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

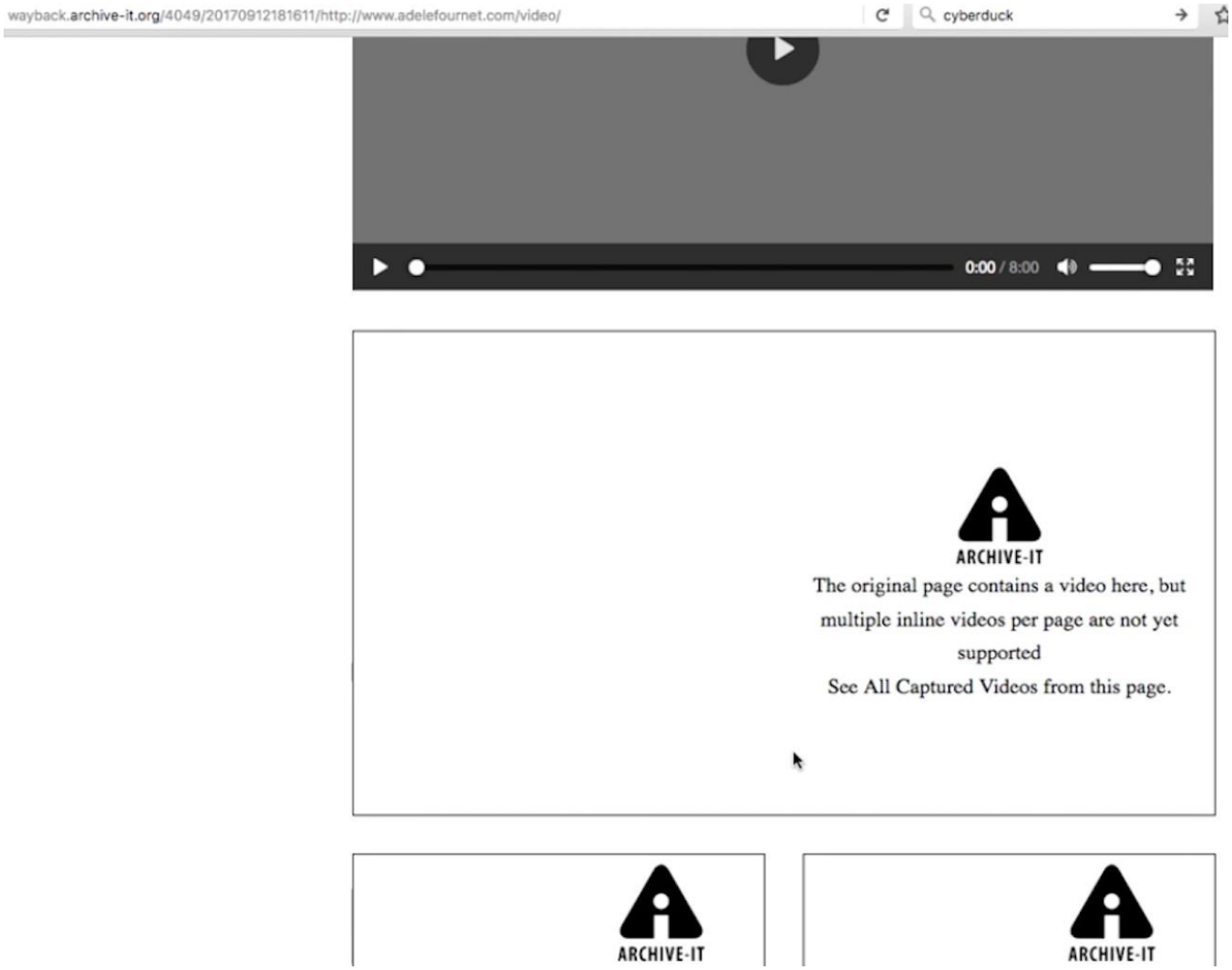
# Findings still being analyzed

- 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)

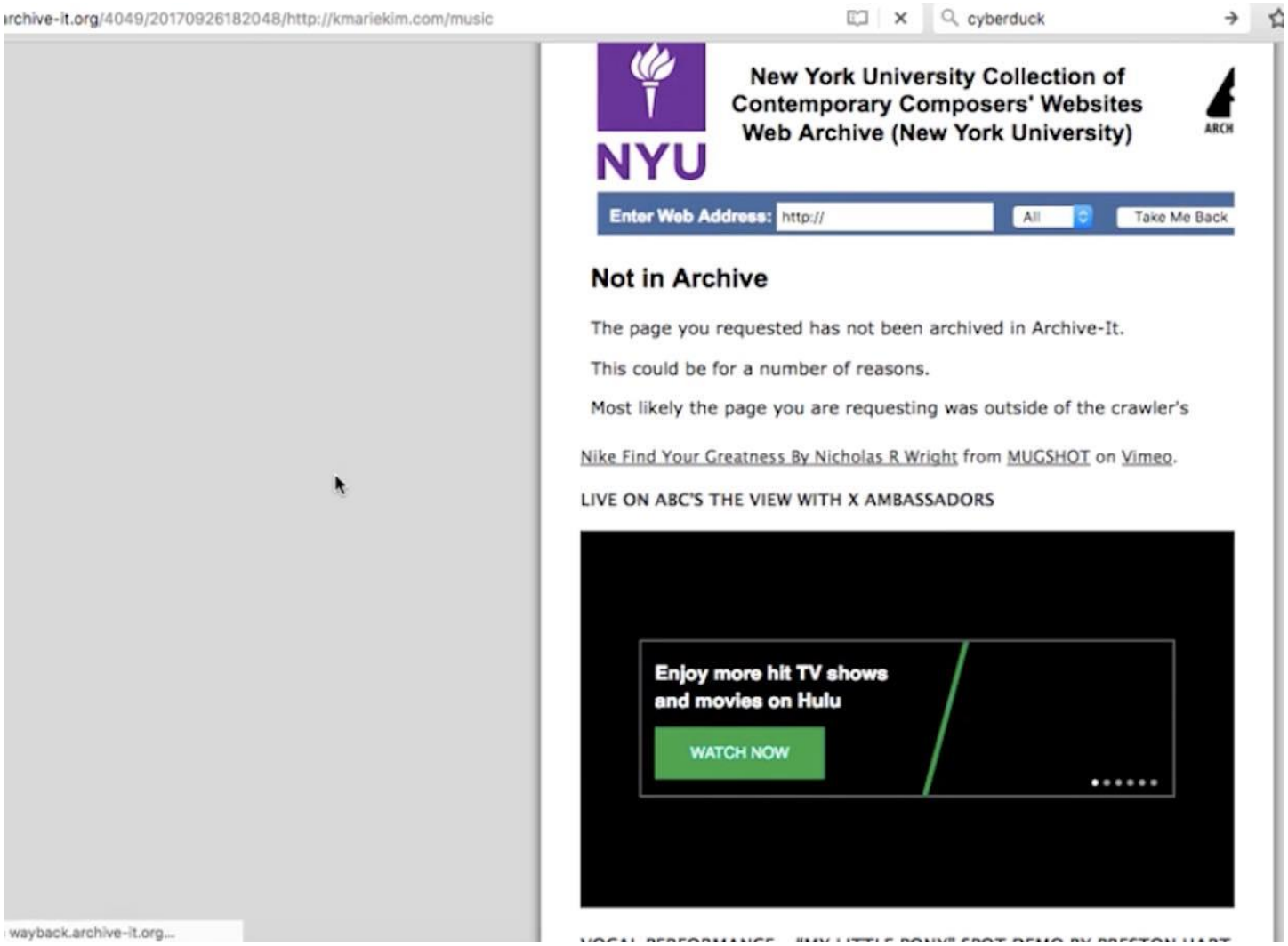
# Crawler Issues (broken header links)



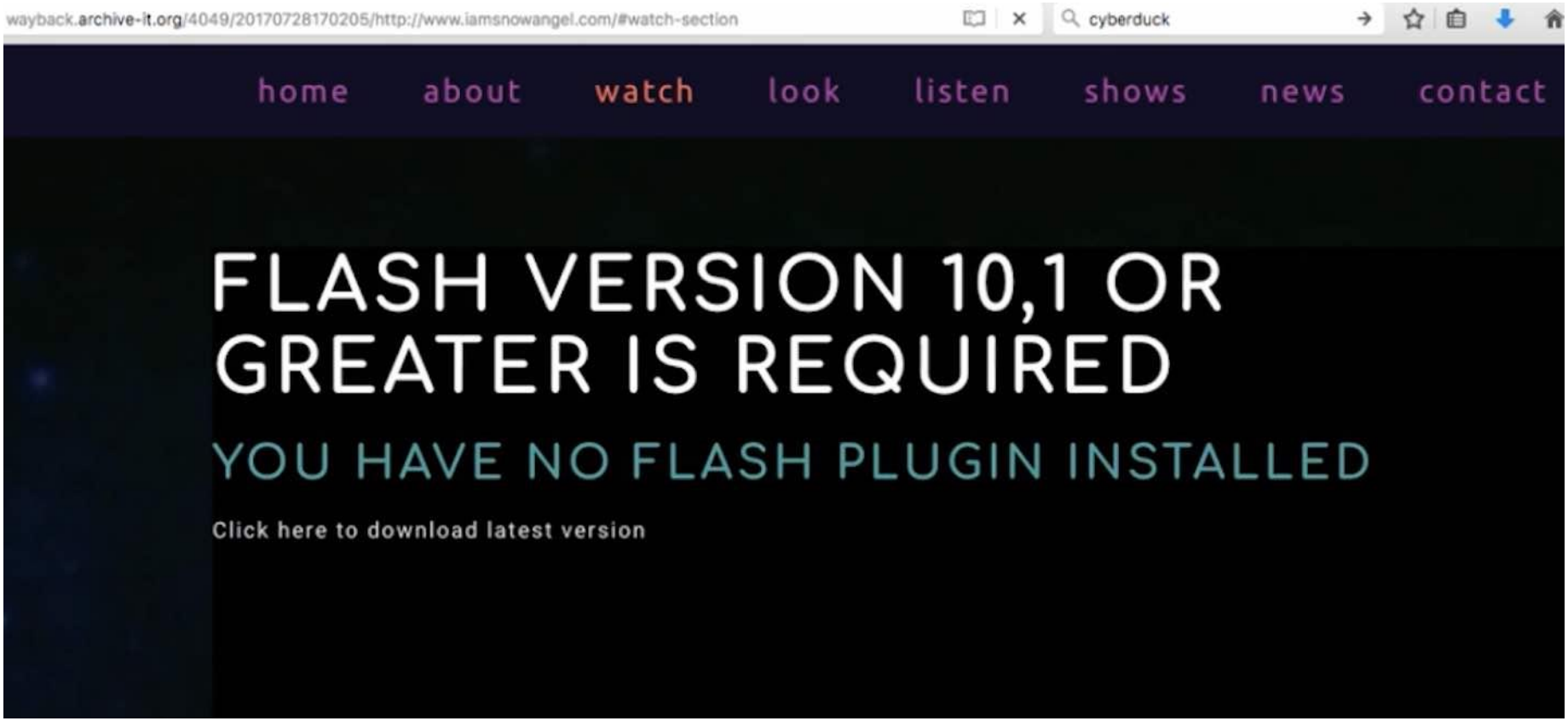
# Crawler Issues (failed video capture)



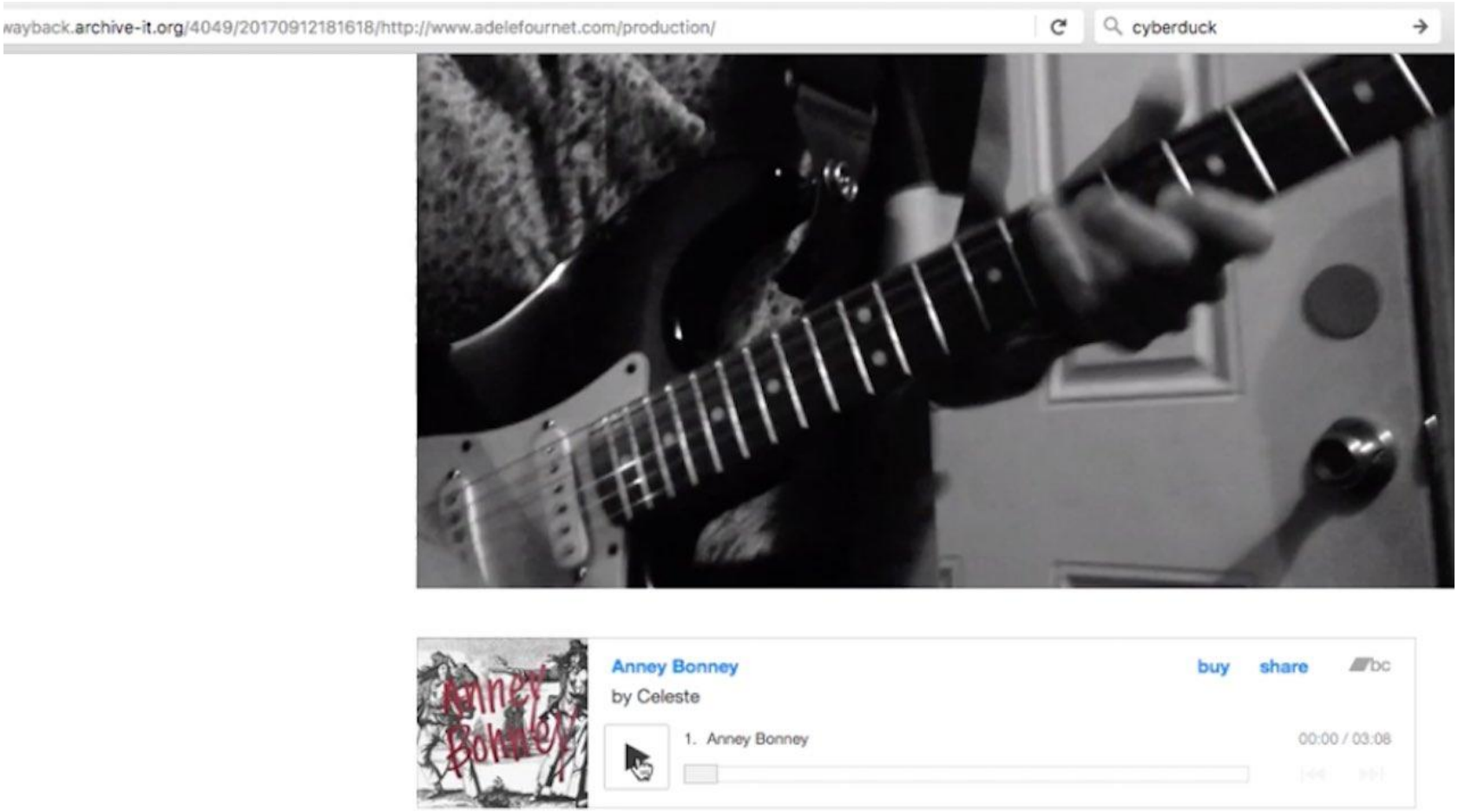
# Crawler Issues (video capture failure)



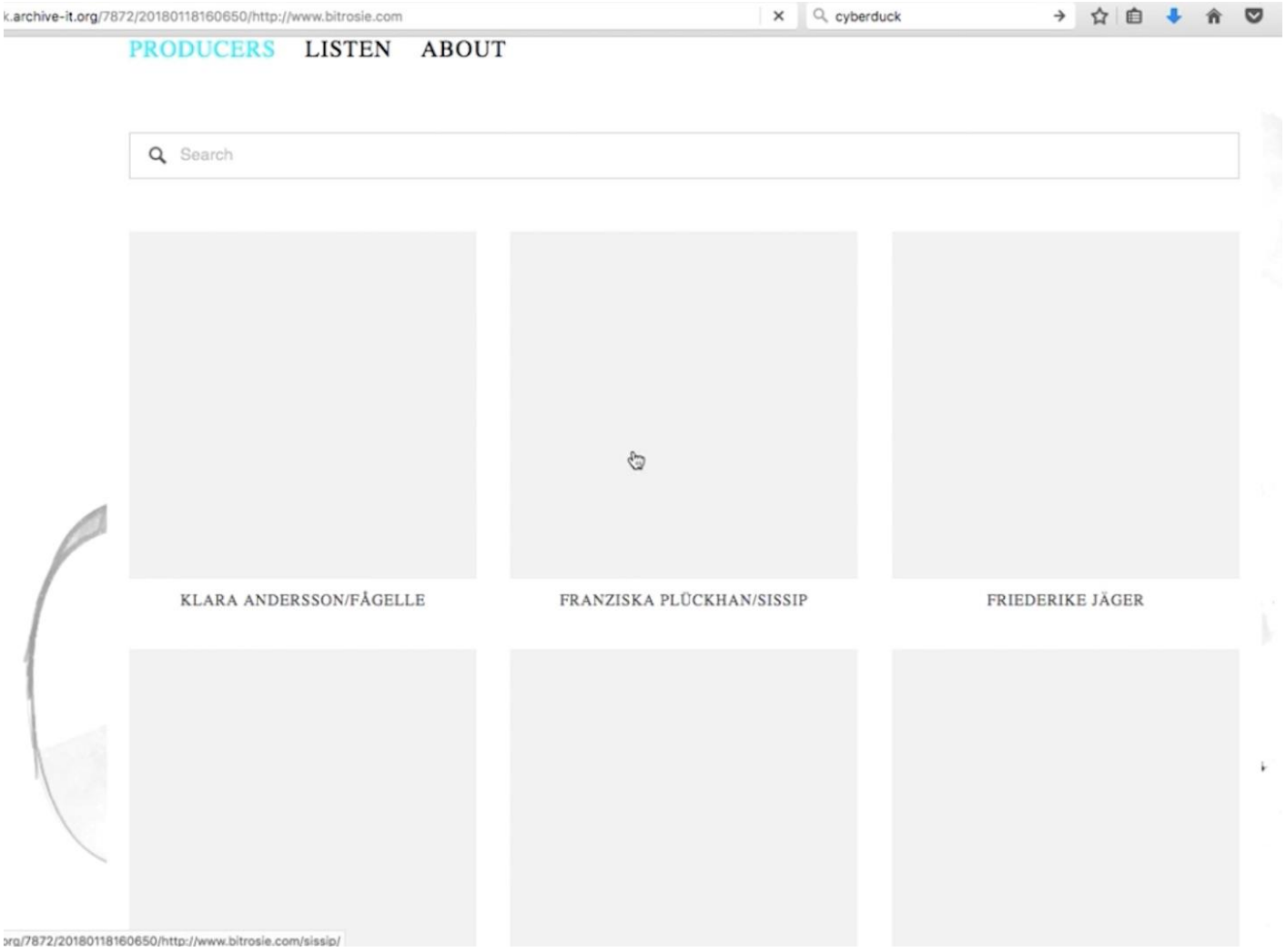
# Crawler Issues (Flash video issue)



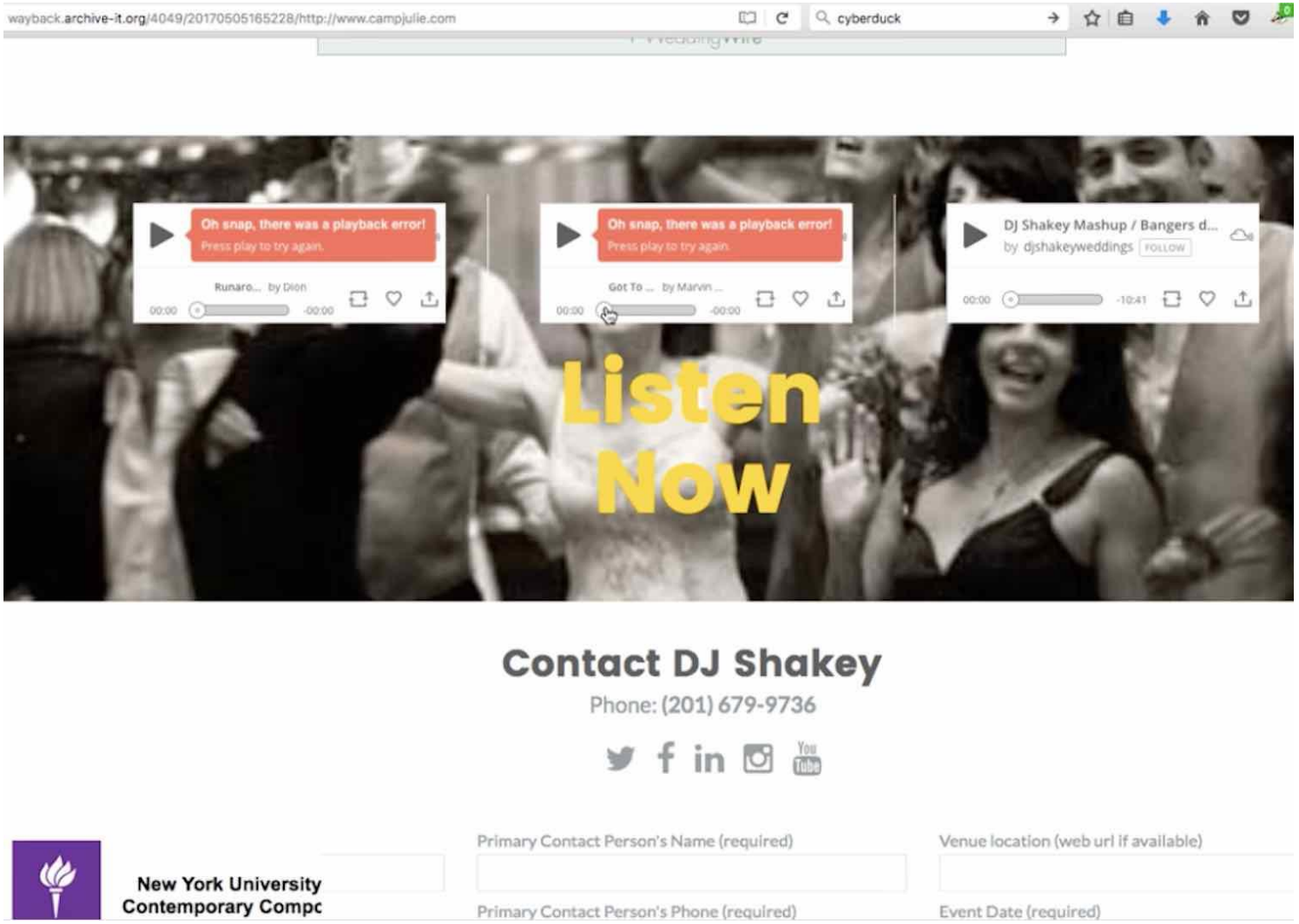
# Crawler Issues (video captured without audio)



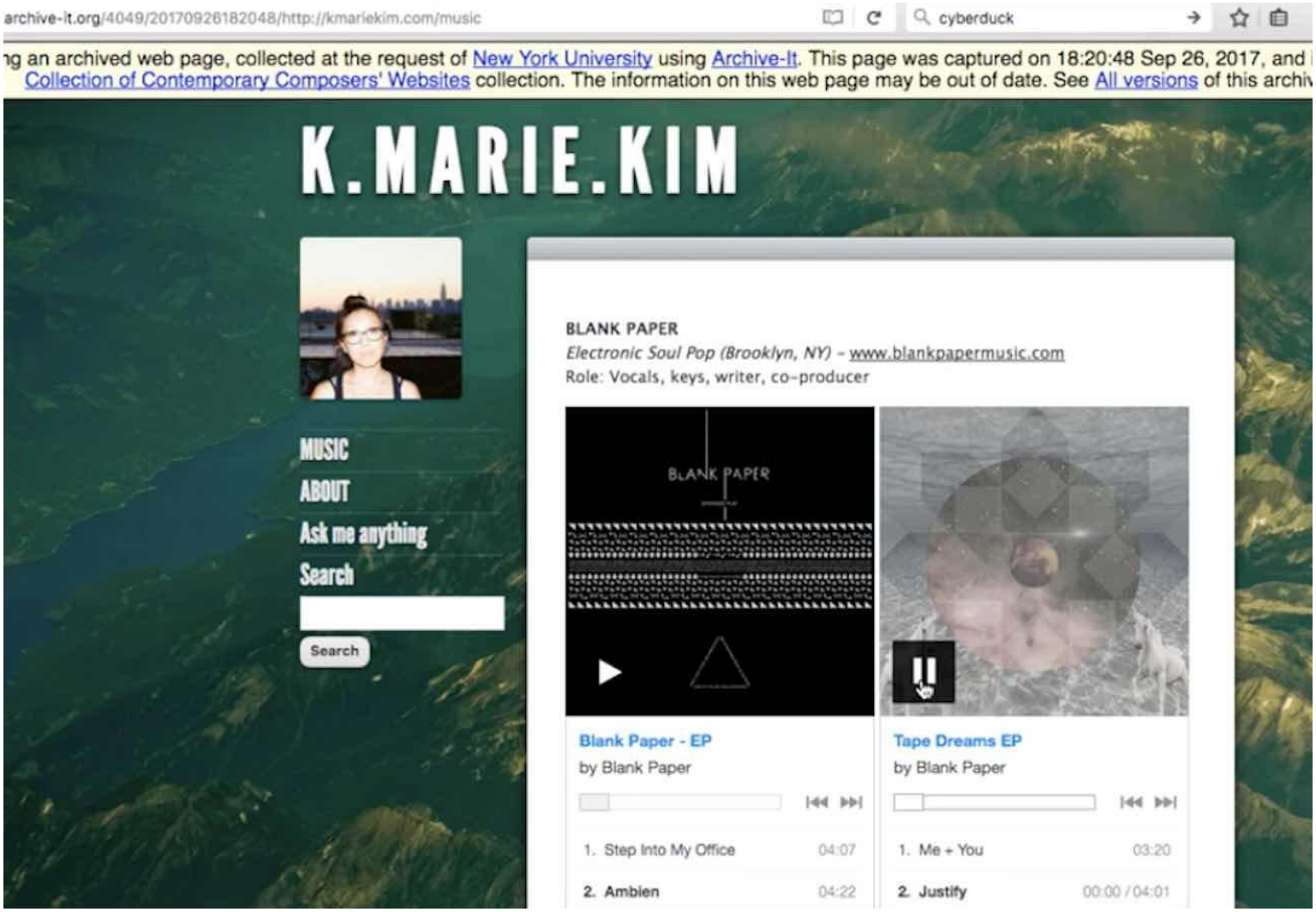
# Crawler Issues (broken video links)



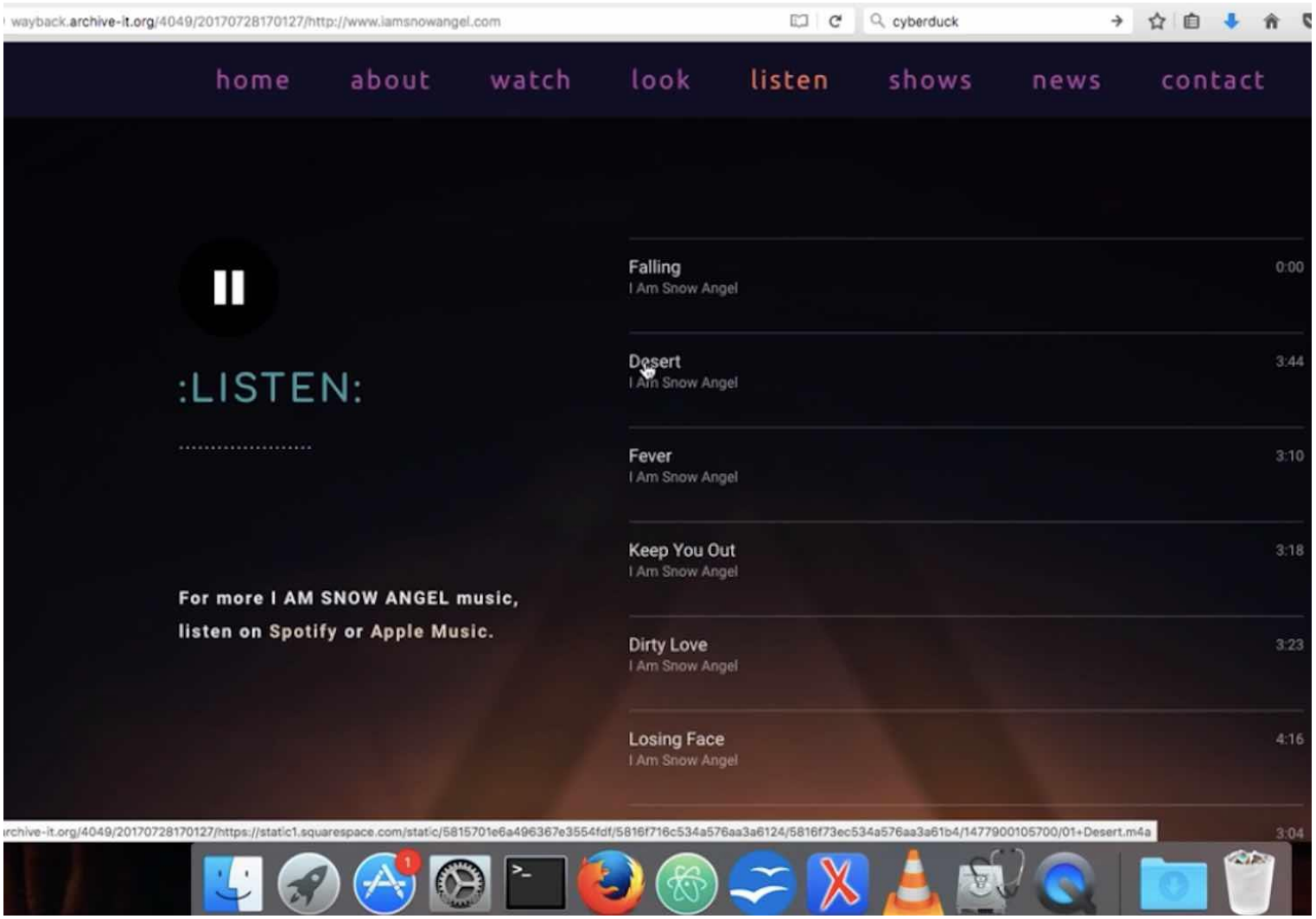
# Crawler Issues (1 audio not captured)



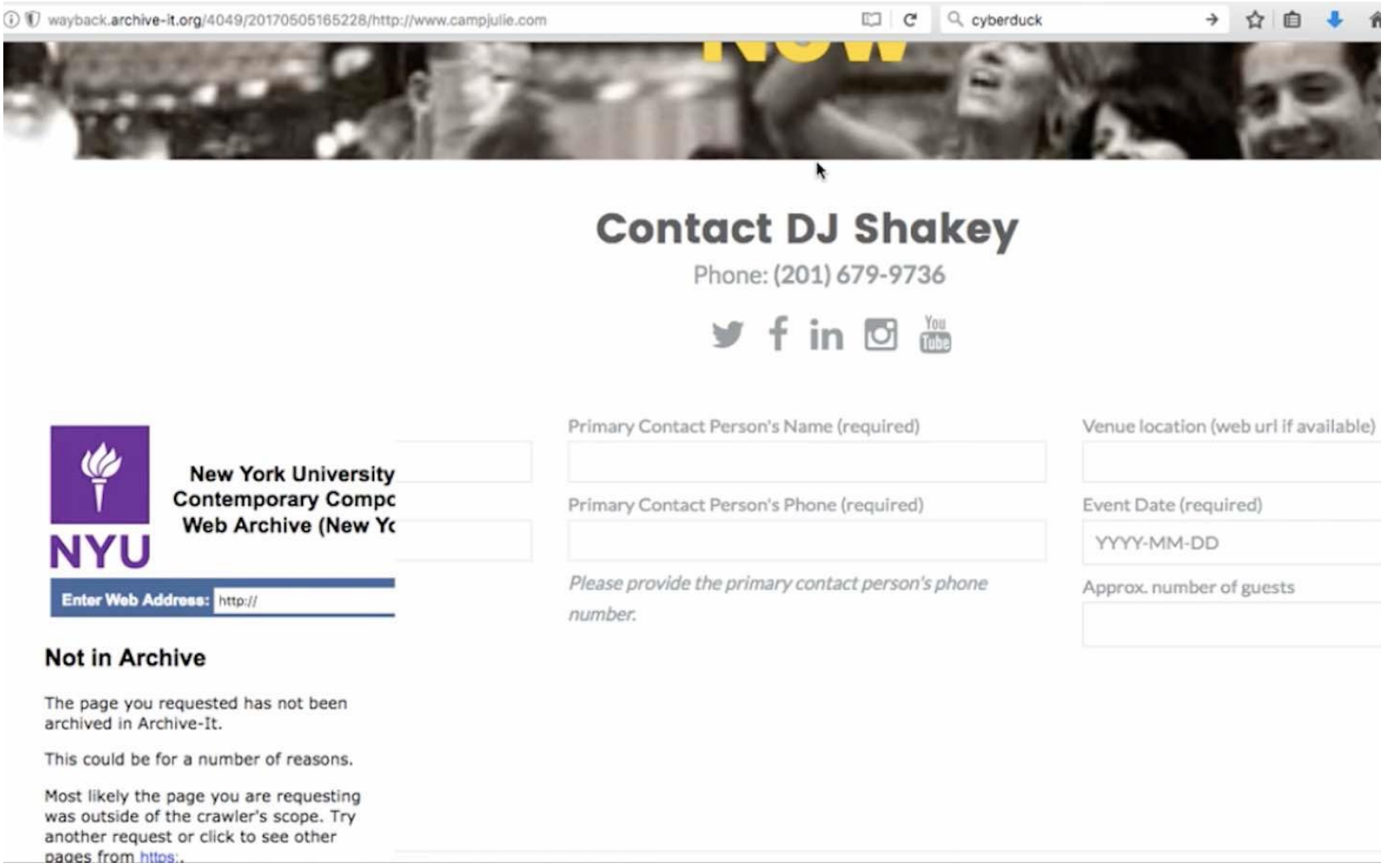
# Crawler Issues (audio not captured)



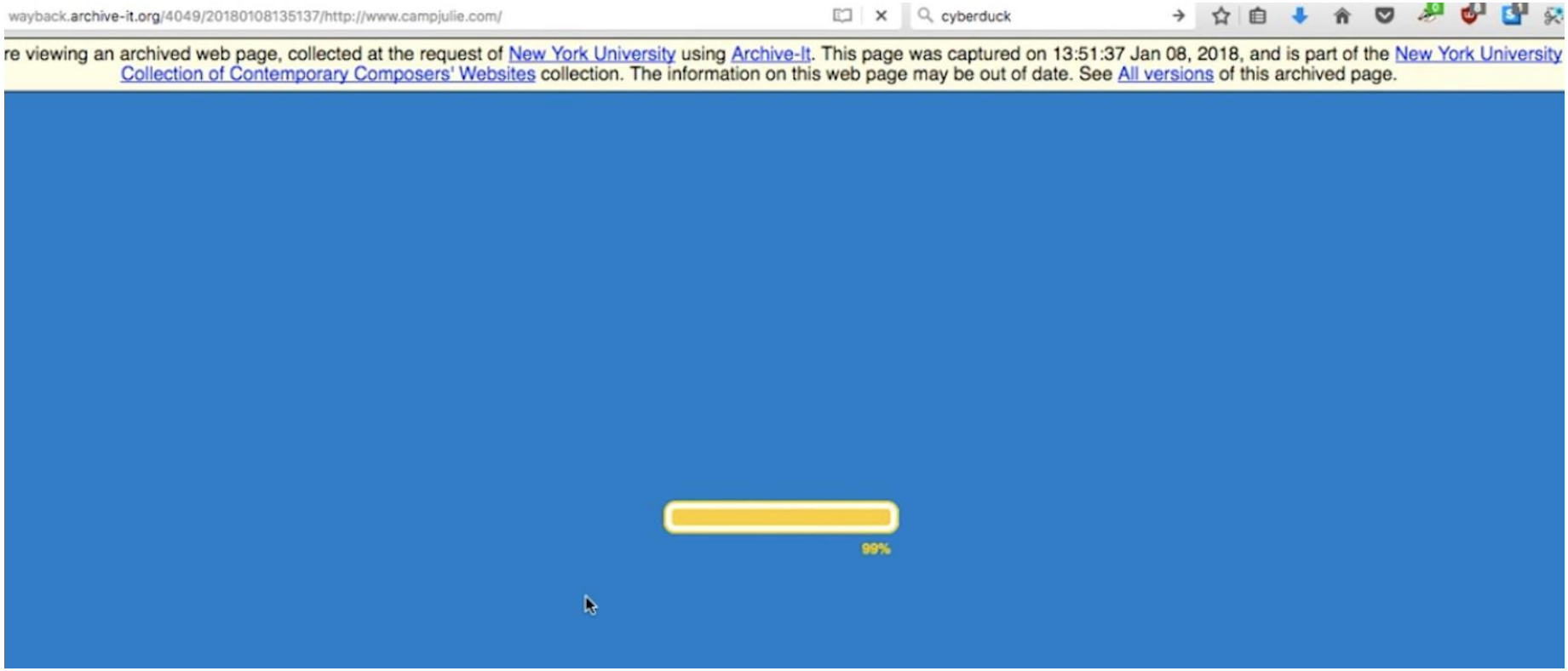
# Crawler Issues (audio failure & anchor problem)



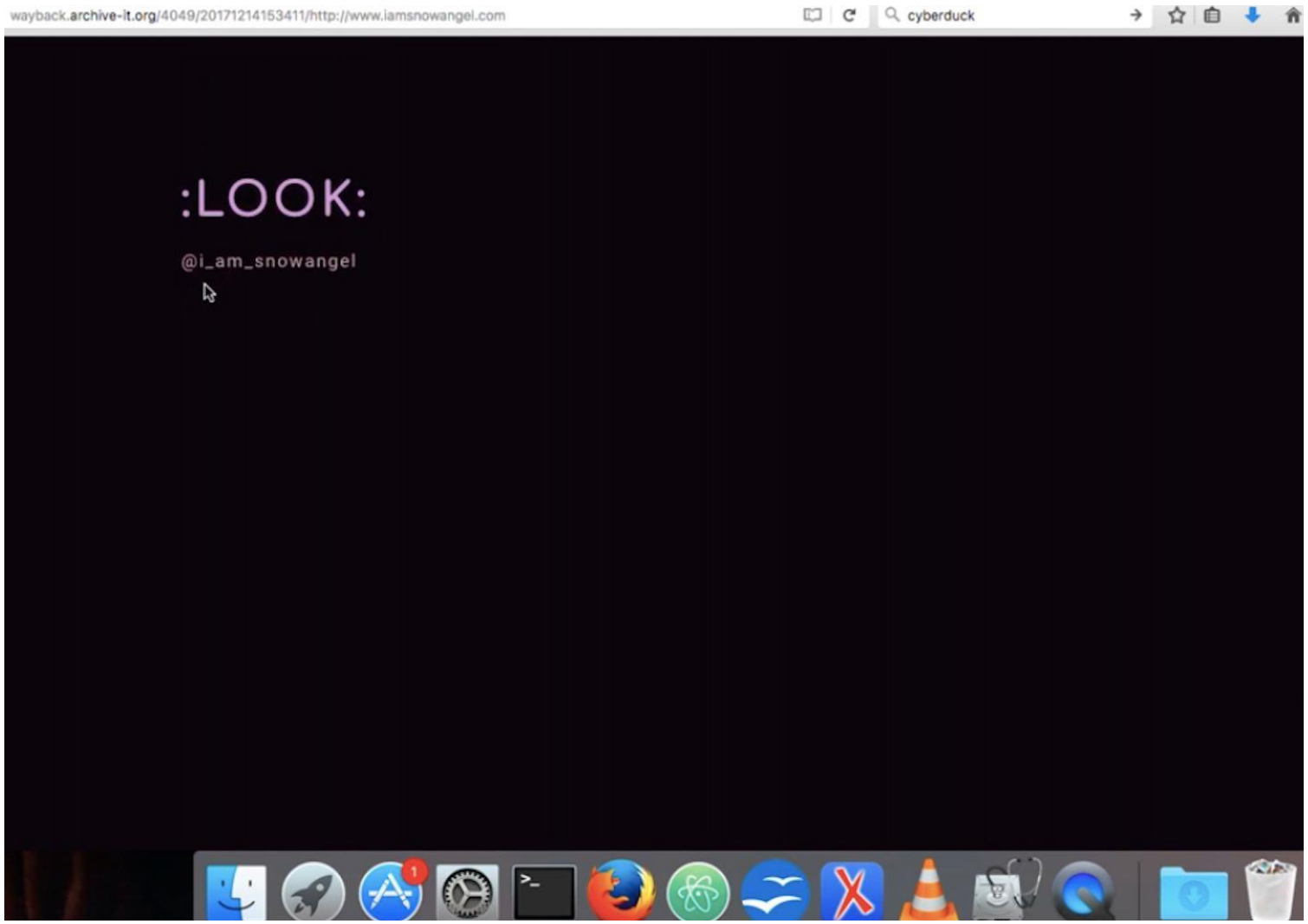
# Crawler Issues (partial capture failure)



# Crawler Issues (incomplete loading)



# Crawler Issues (Capture issues)



# Crawler Issues (unknown problems)

Sorry! Something went wrong

Is your network connection unstable or browser outdated?

I need help

# Crawler Issues

- Campjulie.com:
  - Any capture date: If very slow load time, hard to tell if was 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.
- Kmariekim.com:
  - Sep 26, 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 "Berets of Mary Jean Place", and starts playing another video with opening title "Barranco District, Lima, Peru". The rest of the videos on the page do not play. Link to "Berets of Mary Jean Place" 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 screengrabs
- Access to additional browser window
- Preview frame when scrubbing (fast forwarding) through video material
- Use of their own laptop or another window
- Display of timecode
- 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

# 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://www.nyu.edu/about/news-publications/news/2015/03/27/nyu-libraries-to-team-with-internet-archive-to-preserve-high-quality-musical-content-on-the-web.html
- http://archive.org/~nlevitt/reveal.js/
- http://composers.dlib.nyu.edu/
- https://rg3.github.io/youtube-dl/supportedsites.html