Web Archiving for Music Librarians

Kent Underwood
New York University
kent.underwood@nyu.edu

MLA Cincinnati, 5 March 2016
Definitions

What is web archiving?

• Static snapshots of websites
• Preserved in a digital repository
• “monographic” or “serial”
Definitions

- Archive
- Library
- Records
- Papers
- Original order
Definitions (cont.)

- Webpage
- Website
- Web ("the")
- Internet
Definitions

Web archives, not “web archives”:

- Web caches
- Web-based collections of digitized analog originals
- Web-based collections of born-digital materials
- Hyperlink compilations
True “Web Archive”

• Contains copies of websites as archival objects unto themselves
• Treats websites as historical documents
• Snapshots preserve the contents, appearance, and behavior of each website as it existed at particular points in time
Example

New York University Archive of Contemporary Composers’ Websites
(archive-it.org)

Angélica Negrón
The Conundrum of Web Archiving

How can static snapshots preserve a dynamic system?

(its content, appearance, and behavior)?
Why archive the web?

- To preserve historically important material for posterity

- The WWW is unstable
Types of Websites of Interest to Music History

- Composers
- Performers
- Ensembles
- Performing Venues
- Festivals
- Educational institutions
- Teachers
- Agents
- Publishers
- Record Companies
- Fans
- Critics
- Historians
Component analysis of 200 composers’ websites

- Biography 95%
- List of compositions 92%
- List of recordings 84%
- Portrait photos 90%
- Calendar of events 69%
- Audio recordings 93%
- Scores (sale, perusal, download) 58%
- Social media 39%
- “Web spheres” 24%
The Web is Unstable!

Modes of death and decay:

Link rot

Updating
The Web is unstable!

Not Found

The requested URL /freetrademarksforstartups.com was not found on this server.

Apache/2.0.52 (Red Hat) Server at erikjheels.com Port 80

Error 404

Page Not Found

Mon Jul 14 2008 4:43:30 am EDT
168.192.104.2 tried to load www.revacsystems.com/images/
User Agent = Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322; InfoPath 1)
You are in a Wrong page!
Visit
Home Page
How unstable is the web?

What is the average lifespan of a webpage?

(Link rot + updating)

44-100 days
How unstable is the Web?

What percentage of the entire WWW is the same as it was a year ago?

20% same
40% changed
40% disappeared
Link Rot Vaccines

WebCite
http://www.webcitation.org/

SavePageNow
https://archive.org/web/
But isn’t the WWW archiving itself?

• What about browser caches?

• What about the Wayback Machine?
* As of March 4, 2016. (466 billion on March 1).
Passive vs. Active Curation

“Wayback Machine”
- Automatic process
- Software robot
- Broadly selective but still not exhaustive
- Passive curation

“Archive-It”
- Manual process
- Human being
- Very selective
- Active curation
Benefits of active Web archiving

• Curators (instead of a robot) can apply their skills and expertise to:
  – Selection
  – Organization
  – Description
  – Management

• Passive and active are complementary
Brief History of Web Archiving

6 August 1991

First website
Tim Berners-Lee & Robert Cailliau

(It’s been archived!)
Brief History of Web Archiving

1993

Mosaic

The first commercial web browser
Brief History of Web Archiving

1994

Web crawlers

(foundational technology for Web archiving)
Brief History of Web Archiving

1995

Amazoncom
MP3 audio
Brief History of Web Archiving

1996

Internet Archive founded

Internet Archive-Smithsonian collaboration to archive websites of the 1996 presidential campaign
Brief History of Web Archiving

2001

Wayback Machine goes live

Wikipedia launches
Brief History of Web Archiving

2004

Estimated total number of websites:

51.6 million
Brief History of Web Archiving

2006

Archive-It Partners program launches
Brief History of Web Archiving

2009

Yahoo discontinues Geocities service

Internet Archive steps in to save its history
Brief History of Web Archiving

2014

Estimated total number of websites:

1 billion

(4-fold increase since 2004)
Web Archiving Projects in Music

- L’institut national de l’audiovisuel (INA)
- Netherlands Institute for Sound and Vision
- British Library
- National Library of New Zealand
- University of Texas. Tejano and Conjunto
- Curtis Institute Social Networks
- Oklahoma Music Hall of Fame & Jazz Hall of Fame
- “Borrow Direct” Libraries’ Contemporary Composers Web Archive
- NYU Archive of Contemporary Composers’ Websites
Technological Components

• Specialized software and hardware
• Web Archiving is a type of digital archiving
• Collaboration between curators and IT within the library
• Partnering outside the institution is common
Software Tools

Web archiving crawler

• Makes a “3D” scan of a live website
• Copies both the content and relational data
• Reconstitutes the site as a simulacrum, frozen in time
• Saves and stores the copy in a digital repository
Archive-It

• Both a utility and partnership program
• Developed and supported by the Internet Archive
• Adopted by institutions both large and small
• Subscription fees based on data storage
• Integrated suite of tools & services
• De facto national Web archive for the USA
Other Web archiving tools/services

- **Web Curator Tool** (Britain, New Zealand)
- **Archivethe.net** (Internet Memory Foundation)
- **Netarchivesuite** (Denmark, Austria)
- **California Digital Library** (defunct 2015)
Web Archiving Organizations

• Internet Memory Foundation

• International Internet Preservation Consortium (IIPC)
Curating a Web Archive: Planning

• Basic theme
  – Immediate goals
  – Scalability
  – Sustainability

• Models to emulate

• Collection Development Policy Statement
  – Standard best practice
  – ALA & IIPC Guidelines
Collection Development Policy Statement

- Name
- Summary Description
- History
- Audience & potential users (near- & short-term)
- Scope & boundaries (inclusions, exclusions)
- Relationships to other collections
- Hosting
- Quality assurance procedures
- Conditions for access and use
Curating a Web Archive: Website Component Analysis

• What are the constituent elements?
  – Text, images, audio, video, etc.
• Private/self administered vs. organizational
• Non-commercial vs. commercial
• Freely accessible vs. password
• Internal vs. imported content
• “Web spheres”
Web Spheres

• Contextual relationships to other sites and to the Web at large
• Authorial (hyperlink compilations)
• Curatorial (informed by curator’s educated knowledge and understanding)
• New and more expansive approach to provenance
Curating a Web Archive: Outreach

• Contact site authors
  – Explain project
  – Solicit their cooperation

• Not mandatory but advisable
  – Ethical
  – Courteous
  – Fosters collegial relations generally between donors and archives
NYU Outreach Letter

• Explanation and rationale for project
• Asks permission (opt in)
• Questionnaire
Contemporary composers questionnaire

• Administer the site yourself (82% yes)
• Taking preservation measures already (33% yes)
• Frequency of additions?
• Frequency of deletions?
• Agree to override robots.txt script? (92% yes)
Technical Implementation (in Archive-It)

• Seeds

• Scoping
  – Informed by the website component analysis

• Crawl frequency
  – Informed by the website component analysis
  – Capture all important “editions”
  – Sweet spot between too few and too many crawls
Technical Implementation: Quality Assurance

• Goal: Capture and preserve the contents, appearance, and behavior of the site as it exists on the live Web
• Easier said than done
• NYARC guidelines
Quality Assurance Process

• Test crawls
• Crawl reports
• Human eye making side by side comparisons
• Document the process
• Assume and plan for QA to be ongoing
Internet Archive’s “Five Challenges”

1. Javascript navigation functions
2. Recorded media
3. Password protection
4. Interactive databases and forms
5. Robots.txt scripts
Web Archiving and Copyright

• “Original works of authorship, fixed in a tangible medium of expression” (US Code)

• Is a website a protectable “work”?

• Copyright Office says yes!
Is Web archiving “fair use”? 

• Metaphysical certitude?
  – NO

• Nuanced calculations of risk tolerance and management?
  – YES

• Much depends on the type & content of site
Intellectual Property Best Practices

• Work with site authors/owners

• Work with intellectual property advisor to stay within your institution’s comfort zone

• Obtain permission as much as possible

• Take-down (opt-out) policy as second line of defense
Access

• Potential levels of access
  – Open (identical to live Web)
  – Open-plus (includes walled-off material)
  – Open selective (targeted exclusions)
  – Restricted (material complete but user access limited)*
  – Dark storage (no access until expiration date)*

  *adopt existing archival best practices
Discovery

• Finding aid ("landing page")

• Metadata
  – Collection level vs. Item level
  – Descriptive
  – Administrative
  – Technical
Metadata Schemes

• Dublin Core
  – Descriptive, Administrative, Technical
  – Collection-level & Item-level
  – Example: NYU Archive of Contemporary Composers’ Websites

• MARC
  – Example: Borrow Direct Contemporary Composers Web Archive
What next?

What causes professional Web archivists anxiety?*

- Social media 79%
- Databases 74%
- Video 73%
- Interactive Media 56%
- Audio 45%
- Blogs 36%
- Art 17%

*2012 Internet Archive survey
What next?

• More powerful back-end tools (solutions to the “5 challenges”)

• More powerful front-end tools (data & text mining)

• “Macro” coalescing with “micro” archiving (Niels Brügger)

• Music Encoding Initiative
What next?

MORE WEB ARCHIVING!

Composers
Performers
Ensembles
Performing Venues
Festivals
Educational Institutions
Teachers

Agents
Publishers
Record Companies
Fans
Critics
Historians
REFERENCES

Archive-it.org


REFERENCES

International Internet Preservation Consortium (IIPC).
http://www.netpreserve.org/

