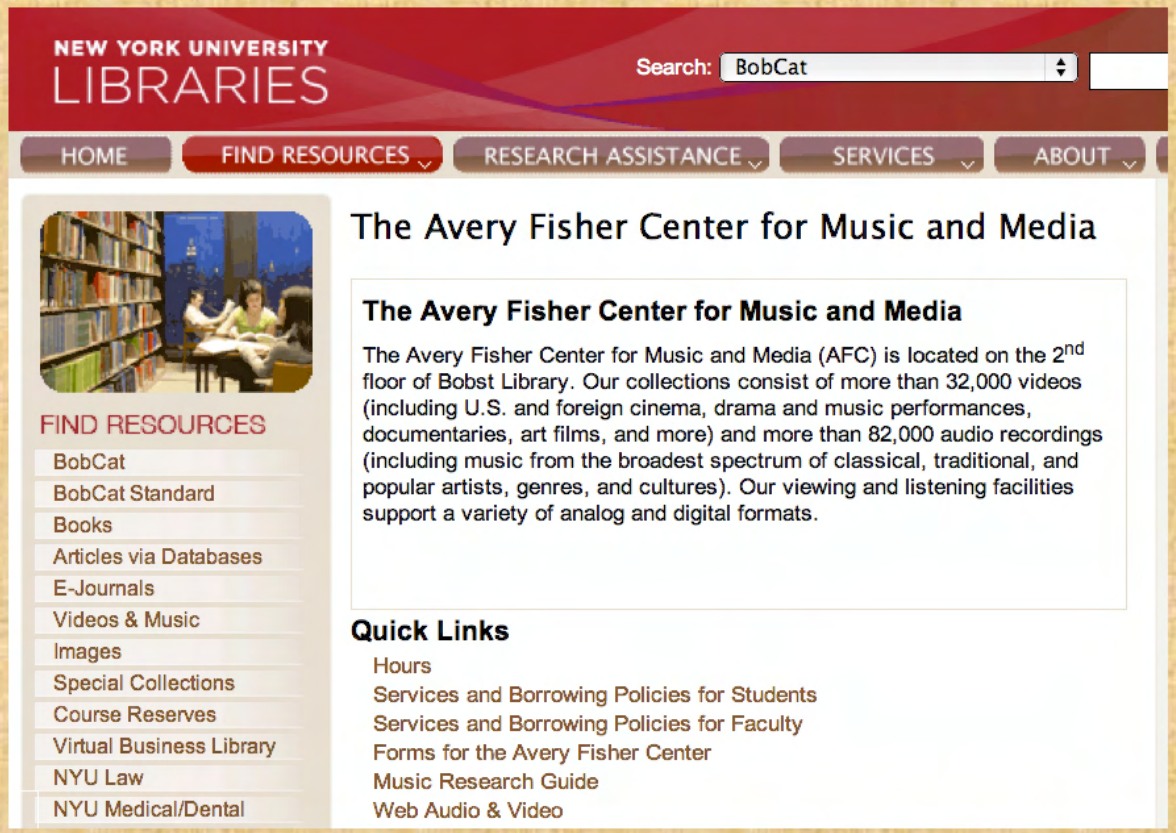
Web Archiving: Avery Fisher Center for Music & Media

Rhiannon Bettivia, Zack Lischer-Katz,
Samantha Losben & Erica Wilson

November 29, 2010
Digital Preservation

Avery Fisher Center for Music & Media



Music Librarian - Kent Underwood

Kent Meeting No. 1

- The Avery Fisher Center did not have any internal websites in need of crawling
- Suggested 3 types of music websites to pursue: artist sites, commercial sector sites, and music journalism sites
- Potential issues included decisions on how to narrow potential subjects, issues of proprietary materials and permission for capturing copyrighted objects, and sites that linked to social networking venues that are password protected

Test Crawl of AFC Website

After initial meeting with Kent Underwood:

- He informed us that he didn't actually have any internal web projects to be archived.
- Music library has one small page as part of the general library website
- The Scope was set to crawl once a day for a month
- Yielded very little information or information that is worth capturing

http://library.nyu.edu/afc/

First day of crawl

Statistics
Started October 22, 2010 1:45:07 PM
Completed October 22, 2010 1:45:37 PM
Status
Finished
Average Doc Rate 2.14 urls/sec
Average KB Rate 23.0 KB/s
Total Documents Crawled* 60
Total Data Crawled 668.0 KB
New Documents Archived 0
New Data Archived* 0.0 bytes

Last Crawl

Statistics
Started November 25, 2010 9:16:04 AM
Completed November 25, 2010 9:58:24 AM
Status
Finished
Average Doc Rate 1.34 urls/sec
Average KB Rate 732.0 KB/s
Total Documents Crawled* 3,409
Total Data Crawled 1.8 GB
New Documents Archived 3,409
New Data Archived* 5.2

- Reset scope to once a Week / still nothing of great importance captured
- Feel that now it could be set to once a year

We thought the project would entail

- Cursory crawl of the Avery Fisher site
- Test crawls of 9 artists' websites recommended by Kent that were chosen as they steered clear of Facebook and MySpace
- Creation of a permission letter to send to artists

The letter, parts 1,2 and 3

- Format came from Columbia University

http://copyright.columbia.edu/copyright/permissions/requesting-permission/model-forms/#letter4

- Adapted it to be about the Internet Archive
- Edited it for length

Kent Meeting No. 2

- Showed him the letter and our test crawls
- Kent offered to make changes to the letter
- BOMBSHELL: Kent's secret ambitions and goals for the project

What the project became

- The creation of a permission letter that would encompass the scope of the new project
- Obtaining clearance on letter from Kent, then Howard, then Kent again

Final Letter

Dear _____,

The New York University library is working with graduate students in the Digital Preservation course on a pilot project to create an archive of music-related web pages. Your website is of particular interest because of the NYU library's long established commitment to collecting contemporary music.

With your permission, we would like to add your website to our growing collection which consists of personal websites maintained by independent musicians and composers. We are using software developed by the Internet Archive called *Archive-It*, which works by crawling websites and archiving copies of each page within a domain. Curated groups of archived websites become available as research collections through the Archive-It website: www.archive-it.org

The one and only goal of this initiative is to preserve historical documentation for research. Web crawling and archiving begins by taking a snapshot of a website in its current state; then by repeating the process periodically, it tracks the site's evolution as it changes over time.

We are mindful of your business concerns in an increasingly difficult market, and we want to assure you that our archiving effort is entirely scholarly in nature and that no profit is to be derived from this effort. No commercial use or any changes whatsoever will be made by us to your site. We urge you to look at http://www.archive.org/about/about.php for information about the Internet Archive, and http://www.archive-it.org for specific information about web site archiving. We would also emphasize that none of this would require any effort on your part; we and the software would be doing all the work.

The success of this project rests on our ability to crawl your website, but of course we would not want to go ahead without your permission. Once the feasibility of the pilot has been established and all the technical and administrative questions have been settled, the NYU library is prepared to commit to preserving the archive in perpetuity.

We are eager to proceed with this project, which we see as fundamental to our scholarly mission as archivists in preserving our digital heritage.

Thank you for your consideration. If you have any questions, please feel free to contact us.

With best regards,

_____on behalf of Kent and students in the Fall 2010 Digital Preservation class

Seed URLs

- http://library.nyu.edu/afc/
- http://www.elodielauten.net/
- http://www.mikelrouse.com/
- http://www.joanlabarbara.com/
- http://www.nicolascollins.com/
- http://sheanshepherd.com
- http://hannahlash.com
- http://nicomuhly.com/
- http://www.kylegann.com/
- http://www.evbvd.com/

Crawl Frequency

- Avery Fisher Center Site
  - Site changed on: 10/26, 11/4, 11/11, 11/25
  - Our subject librarian is currently only interested in one-time crawls.

Quality Assessment

- Avery Fisher Center Site Display Problems
- http://library.nyu.edu/afc/

Common File Formats

text/html	4003 URLs	41.1MB
image/jpeg	3657 URLs	232.2MB
image/gif	747 URLs	11.2MB
text/xml	329 URLs	5.6MB
application/pdf	322 URLs	796.0MB
audio/mpeg	189 URLs	1.6GB

Site Complexity and Size

- Hannahlash.com
  - 33 URLs
  - 126MB
- Nicomuhly.com
  - 3,191 URLs
  - 120.2MB
- Nicolascollins.com
  - 1,493 URLs
  - 1.7GB

Permissions

- The html code can give us clues about how the content creator feels about crawling:
  - Seanshepherd.com = “no robots!”
  - Hannahlash.com = “yes robots!”
  - <meta name="Robots" content="INDEX,FOLLOW">
    <meta name="Revisit-after" content="31 Days">

Sustainability Issues

- Server side behavior is difficult to archive:
  - Avery Fisher Site
  - NYU-Bobst CMS = OmniUpdate?
  - NYU-TSOA CMS = iOn
- Other difficult formats to archive
  - Flash
  - Javascript

Recommendations

- Letter as Template
- Partnership with Internet Archive
- Expedient need for archiving - use ARCHIVE-IT