Contemporary Composers Web Archive (CCWA): Progress in Collaboratively Collecting Composers' Websites

June 24, 2015
IAML/IMS
Anna Perricci, Columbia University
Laura Stokes, Brown University
What is CCWA?
Collaborative Project

- Borrow Direct (Ivies-Plus) Music Librarians Group
  - Brown University (Laura Stokes)
  - Columbia University (Elizabeth Davis)
  - Cornell University (Bonna Boettcher, Leonora Schneller)
  - Dartmouth College (Patricia Fisken)
  - Harvard University (Sarah Adams, Sandi-Jo Malmon)
  - Johns Hopkins University (Jennifer Ottervik)
  - Massachusetts Institute of Technology (Peter Munstedt)
  - Princeton University (Darwin Scott)
  - University of Chicago (Scott Landvatter)
  - University of Pennsylvania (Dick Griscom)
  - Yale University (Suzanne Lovejoy)
  - ...and Anna Perricci, Columbia University
Ex.: Steven Stucky’s website, list of captures
Web Resources Archiving Collaboration

Many thanks to the Mellon Foundation

Building collaborations among

- Web archiving communities
- Other research libraries
- Users and potential users of web archives
- Website creators
Premise: we need radical collaboration and cross-institutional partnerships

Collaborative web archiving pilots projects supported/furthered by existing and developing cohorts of peers focused on cooperative collection development initiatives

Photo credit: http://www.concordia.ca/about/strategic-directions/james-neal.html
What is a web archiving project librarian?

Roles and goals:

• **Web** (education/outreach about web collecting, coordinating & supporting selection of web resources to archive)

• **Archiving** (helping to prioritize, obtain and save content for immediate and long-term access)

• **Project** (project manager for many moving pieces and partners)

• **Librarian** (technical services librarian and liaison roles)
Project elements

• Incentives grants to advance web archiving tools

• Collaborative collection building through Ivy Plus / Borrow Direct
  – CCWA; Collaborative Architecture, Urbanism and Sustainability Web Archive (CAUSEWAY); and climate change curation experiment

• Outreach to site creators

• Best practices for site creators

• Citation analysis
  – Humans testing 2061 URLs from citations in scholarship on Human Rights published in 9 major journals in this field in 2010
  – Around 50% don’t work or have major content drift/lead to cited content—now what?
  – Leveraging APIs to determine if cited content in the Human Right Web Archive and/or the Internet Archive; part of this process involved assistants looking for missing content on live web

• Interviews with scholars to enrich use case development
A brief introduction to web archiving

• Web archiving entails a multifaceted approach to preserve web-based materials (e.g. websites) and ensure ongoing access to collected content

• The main elements of web archiving are
  • Selection
  • Collection
  • Metadata assignment/cataloging
  • Quality assurance
  • Access
  • Long-term stewardship

• Columbia University Libraries Web Resources Collection Program
  • All of the above steps plus a policy to ask permission before collecting
Selection
From collection development policy to subject expertise

• First questions: what to collect, for whom and why?

• Collection development policy

• Defining themes and goals
  – For CCWA this was an extension of cooperative collecting of contemporary composers’ scores

• Engaged selectors
Collection / harvest using software

• How do we get the content for our collections of archived websites?
  
  – We harvest sites using software, often known as a crawler, spider or robot
  
  – For collecting and access we use Archive-It, a subscription-based service of the Internet Archive
Technical elements of curation

• Fostering and managing collaboration is a multi-faceted pursuit

• Tools for collecting, track progress and manage communications
  – Archive-It
  – Google Drive
  – Basecamp
  – Email aliases & folder rules
Cataloging & Quality Assurance

• Cataloging / Metadata assignment essential to discoverability
  – Recognition of Russell Merritt, a skilled music cataloger at Columbia, who has cataloged most of the sites in CCWA and lead the effort to create a process for cataloging the archived websites of contemporary composers

• Quality assurance testing
  – Many thanks to music librarians who did QA testing last year!
  – Sometimes errors can be corrected through patch crawls/troubleshooting

Photo credit: Anna Perricci
Access

• Who will use what we have collected and how will they access it?
  – We need more use cases
  – We need to make web archives more accessible to get use cases

• Archive-It is an access system with limitations

This is a server at the Internet Archive!
When? Now! Plus long-term stewardship...

- We need to collect websites before they disappear but we also must ensure their long-term survival and maintain access to them over time.
- So far we are saving websites in the WARC file format (the preservation standard) and temporarily relying on the Internet Archive to store the files until a repository framework can be established/chosen.
Web archives in the media

  - Lepore got a lot of things right but there was a severely incorrect assertion that most everything on the internet is already in the Internet Archive (it’s not!)
  - There was no recognition of curated web archives and their tendency to be of higher quality/fidelity to the original resource
    - Archive-It provides a set of tools to focus and improve crawls
Advocating for curated collections

• Curated collections are
  – focused rather than haphazard
  – guided by a collection development policy
  – informed by skilled selectors

• Re-crawling sites at regular intervals can show patterns and maintain a consistent flow of information

• Because we ask permission, we ignore a file that blocks crawlers (so we get things that the Internet Archive otherwise would pass over as a matter of policy)

Photo credit: Anna Perricci
Discoverability & facilitating access
CCWA in a library catalog
Collaborative work

- Selection
- Automation
- Curation / Quality control
  - This brings up content questions still to be resolved!
- Challenges faced
  - Ownership and costs
  - Leadership and consensus, esp. as the project grows
Research implications

• Loss and destruction of information in the digital world
• Curation and canon-formation
• How to retain original order when materials are removed from a linked context?

Creative Commons: “Wasteland” by Shane Gorski CC BY-ND 2.0
Effect on research

Do digital resources and tools lead to fundamentally different questions, or do they provide the basis for new answers to existing questions that could not be answered without newly emerging digital tools?

Or:
Is web archiving philosophically different from other kinds of archiving?
Concluding remarks

• This pilot project has:
  – Effectively created a robust technical framework for curated web archiving
  – Demonstrated some of the advantages (and a few pitfalls) of a collaborative approach

• Current and future media coverage and scholarly research needs demonstrate why this work is critically necessary
Thank you!

Anna Perricci
Web Archiving Project Librarian, Columbia University
anna.perricci@columbia.edu

Laura Stokes
Performing Arts Librarian, Brown University
laura_stokes@brown.edu

Photo Credit: David Niblack, Imagebase.net