Building the National Radio Recordings Database: A Big Data Approach to Documenting Audio Heritage

Emily Goodmann, Clarke University
Mark A. Matienzo, Stanford University
Shawn VanCour, UCLA
William Vanden Dries, Indiana University

4th CAS Workshop
IEEE Big Data 2019
11 Dec 2019

Preprint: https://arxiv.org/abs/1911.04625

Key points

Plan of attack:

1. Project history
2. Database architecture
3. Outreach strategies

Doing critical data science work:

- Data collection strategies
- Database design strategies
- Organizational strategies

The Board should establish a subcommittee to develop strategies and tools to collect and preserve radio broadcast content. Among the subcommittee's first actions should be the convening of a symposium on the challenges to preservation of American radio broadcasts and possible solutions.

National Recording Preservation Board
OF THE LIBRARY OF CONGRESS

National Recording Preservation Plan (2012)

Radio Preservation Task Force

The Radio Preservation Task Force (RPTF), created early in 2014, grows out of the Library of Congress National Recording Preservation Plan (December 2012, see: www.loc.gov/programs/national-recording-preservation-board/), and seeks to:

1. To support collaboration between faculty researchers and archivists toward the preservation of radio history
2. To develop an online inventory of extant American radio archival collections, focusing on recorded sound holdings, including research aids
3. To identify and save endangered collections
4. To develop pedagogical guides for utilizing radio and sound archives
5. To act as a clearing house to encourage and expand academic study on the cultural history of radio through the location of grants, the creation of research caucuses, and development of metadata on extant materials

Project history

- Initial NRPB collections survey (Fall 2013 - Fall 2014)
- RPTF created Dec 2014 + survey extended thru 2015 (640 collections)
- Creation of RPTF Metadata and Network divisions (for database construction + collections outreach)
- Initial database launched in Spring 2016 (1,000 collections)
- Second version launched in Spring 2019 (2,000 collections)
- 2020: projected 3,000 collections

Data collection

- Original survey gave research teams autonomy despite original goal for shared basic information
- Original focus: larger archives with mainstream commercial content
- Eventual creation of a Google Form to gather and manage data before review

Database modeling and development

- Domain model based on survey information mapped to collection-level information in Encoded Archival Description elements
- Gaps in mapping led to additional data collection
- Database application built using Blacklight (Ruby, Apache Solr); selected because of use in library discovery and ArcLight extension for archival description

cylinder Search

Start Over | cylinder ×

Limit your search

- Content type
- Format
- Genre
- Repository/Collector
- Country (Location)
- State (Location)

« Previous | **1** - **10** of **13** | Next »

Sort by Relevance | 10 per page

1. Science Forum

Bookmark

Description: WGY; Personality: Dr. Francis Norton
Content types: Sounds
Formats: Cylinder
Extent: 1 recording
Repository/Collector: miSci (The Museum of Innovation and Science)

2. GE Global Research

Bookmark

Description: 1&2 Dedication of Ceramics Building; World Science & Technology in 1960, Dr. C. Guy Suits; Dr. C. Guy Suits, sides 1 & 4; Powder Metallurgy Awards, William Coolidge and B. Benbow; Cermet Roundtable at Knolls; Sounds of Progress
Content types: Sounds

Improvements and growing pains

- Increased pressure from NRPB required database improvements and growth
- Recruitment of second Metadata Director to respond to needs
- Data quality issues
  - Barrier to adoption of ArcLight
  - Focus on contextualization (e.g. Spotlight), improving data entry
- Development of improved search functionality and application programming interface

Ethical concerns: privacy of private media collectors

- Data collection shifting to focus on private collectors
  - Trade advantage
  - Fears of legal prosecution or theft
- Private collectors thus not motivated to contribute
- RPTF responses
  - Filtered subsets: all public; only title, owner, desc public; all private
  - Creation of Program Transcriptions Team (focused outreach group)
  - Potential finer-grained access controls per field or record

Network partners and outreach

- Growing database collection information
- Identifying additional collections and repositories
- Managing data collection/data quality control
- Managing network partners

Data collection improvements and challenges

- Improvements:
  - User-friendly collection form
  - Addition of free-text fields and prompts
  - Network team member assistance
- Challenges:
  - Network team's human resources
  - Organizational partners' human resources

Building better data and networks

- Adoption of a constituent relationship management (CRM) technology
- Better data generation:
  - More accurate, consistent
  - Yielded through just and ethical organizational practices
- Diversifying collection content accessible in the database
  - Relationship-building with underrepresented radio collections' custodians

Thank you!

Emily Goodmann, Clarke University
emily.goodmann@clarke.edu

Mark A. Matienzo, Stanford University
matienzo@stanford.edu

Shawn VanCour, UCLA
svancour@ucla.edu

William Vanden Dries, Indiana University
wvandend@indiana.edu

https://radiopreservation.org/

Preprint: arxiv.org/abs/1911.04625