SpokenWeb: Metadata Scheme and Cataloguing Process

Séance de travail des membres du CRIHN

15 February, 2019

Jason Camlot, Tomasz Neugebauer, Jared Wiercinski
Outline

• Introduction to the SpokenWeb Research Program

• Metadata Task Force
  • Methodology for the Scheme Development
  • Highlights of the Results
  • Sample content
  • Contents & Title Field

• Ongoing work
  • Swallow (Spoken Web Audio Metadata Ingest System)
  • Questions
The SpokenWeb Research Program
Social Sciences and Humanities Research Council of Canada (SSHRC), Partnership Grant, (Jason Camlot, PI). “The SpokenWeb: Conceiving and Creating a Nationally Networked Archive of Literary Recordings for Research and Teaching,” 2018-2025. ($2.5 million)


Social Sciences and Humanities Research Council of Canada (SSHRC), Research Development Initiative (RDI) Grant, (Jason Camlot, Principal Investigator, Darren Wershler Co-Applicant), “The SpokenWeb 2.0: Conceptualizing and Prototyping a Comprehensive Web-Based Digital Spoken Word Archive Interface for Literary Research.” 2010-2012. ($39,692.00)

Social Sciences and Humanities Research Council of Canada (SSHRC), Image, Text Sound & Technology (ITST) Grant, (Jason Camlot, Principal Investigator), “RECITE: Exploring and Developing Digital Tools for the Analysis and Interactive Use of Literary Spoken Recordings.” 2010-2012. ($49,980.00)

• What is a Partnership Grant?

• What is The SpokenWeb? (Institutions, People, Structure)

• Aims

• Methods

• SpokenWeb Collections: The (expanding) scope of our audio corpus.

• Activities (via Task Forces)
What/Who is The SpokenWeb?

A funded research partnership and network currently made up of 15+ partners (institutional, community, organizational) across the US and Canada and 40+ researchers working on the development of digitized collections of literary sound recordings in order to make them accessible for use in research and teaching, and on research, teaching and “knowledge mobilization” activities with the digitized collections.

Present Institutional Partners

**Holding Collections**: Simon Fraser U, U British Columbia Okanagan, U Alberta, U Calgary, U Toronto, Victoria U in the U Toronto, Concordia U.

**Holding Relevant Expertise**: U Texas at Austin, UC Davis, NYU, U Pennsylvania (PennSound), U Victoria, etc.

Other Partners

Wired on Words, Archive MTL, Blue Metropolis Foundation, Compute Canada

Researchers (Co-Applicants and Collaborators)

Lit: Jason Camlot, Gregory Betts, Clint Burnham, Al Filreis, Michelle Levy, Smaro Kamboureli, Mike O’Driscoll, Karis Shearer, Darren Wershler, Jason Wiens, Cecily Devereux, Carole Gerson, Nick Mount.

Library and Archives: Rebecca Dowson, Roma Kail, Sean Luyk, Loryl MacDonald, Tomasz Neugebauer, Felicity Tayler, Jared Wiercinski, Geoff Harder, Marjorie Mitchell, Tony Power, David Seubert.

Community: Ian Ferrier, Louis Rastelli, William St-Hillaire (new: Alan Lord, Lee Gotham…)

DH: Tanya Clement, Adam Hammond, Marit MacArthur, Brian McFee, Lee Miller (as consultant), Geoffrey Rockwell, Miles Thorogood, Lee Miller.

Oral History: Steven High.

Media History: Patrick Feaster, Matthew Rubery, Jenterey Sayers, Jonathan Sterne.

Podcasting: Hannah McGregor.

Student Researchers, PDFs and “HQP” (lots of them). A Student Advisory Committee of approx. 25 students to start with.
At Concordia

2018-19

2 Librarian Co-Applicants
1 Systems developer
2 Literature scholar (PI and Co-Applicant)
1 Oral Historian Co-Applicant
3 PhD students ENGL, HIST, HUMA
2 MA students ENGL
1 Ugrad student ENGL
3 Research Affiliates
1 non-Trac RA (coder/web designer)
1 Project Manager
Governance Structure

Weekly/bi-weekly GB meetings
4+ full team meetings per year (remote participation)
1 Annual Summer Institute and Conference
(in person and remote participation)
Some Aims

1) Inventory and index existing collections in Canada, held by institutional and community partners, to establish a broad understanding of our national holdings of literary sound recordings, and, digitally preserve and aggregate these collections by developing a metadata scheme and shared repository system.

2) Study literary recordings using interdisciplinary theories and methods of “audiotextual criticism”

3) Develop and test computational methods for analyzing literary sound.

4) Create online modules, interactive aural exhibitions and new works of sound art from the digitized archive for use in research, teaching and public presentation, allowing the archive to speak and tell stories in new ways.
Methods

The SW research program is designed to explore

1) the history of literary performance and practice as discernible in sound recording media,

2) cultural effects of the circulation of analogue and digitized literary recordings, and

3) how engagement with sound recordings can have the effect of changing disciplinary practices in research and teaching, by mixing methodologies from literary, media and sound studies, oral history, computational humanities, digital libraries, and archival theory.

We’ll introduce new methodological approaches to the field of literary studies (media historical, computational, archival, etc.), revise the nature and format of literary canons, explore new forms of sociality in literature, consider the historical significance of public readings, and the changing nature of performing non-dramatic literature in public, introduce new methods for the description and public use of literary sound recordings in Canada, new methods of computational analysis of spoken audio, new methods of digital curation of spoken recordings, etc…
SpokenWeb Collections:
The (expanding) scope of our audio corpus

Time frame of audio recordings: mainly 1960-1990

SFU collections hundreds of hours of readings, lectures, interviews, literary conversations,
UBCO collection holds multiple homemade literary audio tape collections of Canadian poets including readings, compilations and literary dictation.
U Alberta collections document local reading series.
U Calgary collections hold recordings made by individual Canadian authors.
Concordia U collections hold documentary recordings of literary performance, multimedia literary/art practice, individual author tape collections
U Vic holds a rich collection of formal literary interviews
U Toronto holds collections that document practices of literary pedagogy, and French language reading series.
Potential future partners: LAC, U Ottawa, U Victoria, UNB, etc.

Wired on Words holds a “words and music” reading series that has run from 1980s to the present (hundreds of hours).
ARCMTL counterculture collections, dumpster dive materials.

Hidden Collections: We have a mandate to seek out collections that have not been deposited in archives, digitize, preserve and integrate them into our project (with the source materials to be deposited in partner archives or kept by the holder of the collection.
    A few examples:
Yellow Door Readings (36 Cassette tapes documenting a readings series from the early 1990s).
Ultimatum Festival and Readings: 39 reel to reel tapes plus several U-Matic video tapes of a bilingual punk poetry series that took place in Montreal in the mid 1980s.
We did a survey and there are MANY more such collections out there in shoeboxes, in basements.
Enough Said (50hrs of Video documenting the an early “spoken word” series in Montreal in the 1980s-90s).
As word gets out more collections come forward
Activities

Researchers pursue projects via collaborative Task Forces.

Metadata Task Force
Systems
Rights Management
Analog/Digitization
Audio Signal Analysis
Oral History
Podcasting
Pedagogy
Research Creation
Archival Events
Community Collections
RDM
etc.

(Literary or Audiotextual Research doesn’t have a task force as it is an assumed form of activity across all SW pursuits.)
Methodology for the Scheme Development
Context

• Why? Need metadata scheme to consistently and systematically capture important information as we catalogue all of the items within our partner collections.

• Capturing descriptive and (some) technical metadata to facilitate access and research.

• As partner institutions will lead their own preservation initiatives for their local collections, we will not capture preservation metadata comprehensively.
Context

• Student cataloguers use the metadata scheme to catalogue sound recordings, and record this information in our ingest system (Swallow).

• Metadata conversation started with Literary Audio Symposium, a SpokenWeb event in Dec, 2016. At the event we discovered that we all had very different ways of cataloguing our collections using different metadata scheme and that this wouldn’t work for a Partnership Grant project.
Methodology for developing our metadata scheme

• Principles:
  
  • **Build our own** vs. choosing an existing system: metadata is very tied to your chosen access system and we are not ready to choose one (Islandora, Avalon). Built our scheme with a close eye towards these systems and their schemes.
  
  • **Agile**: get the metadata scheme out quickly to students for testing and then revise based on experiential feedback
  
  • **Iterative**: iterative metadata scheme, revision process, iterative ingest system (Swallow). Any changes in metadata scheme need to be echoed in ingest system as they are paired.
Methodology for developing our metadata scheme

• Principles:
  • **Distributed**: created with input from all partners.
  • **Simplified standards**: started with standards like IASA, AACR2, MODS, Library of Congress to facilitate interoperability (LC genre) and simplified them to make them useable by student cataloguers
1.8.7. Supplied titles

**Developed from AACR2 4.1B2**

**Expanded from NSFA MAVIS Data Entry Guidelines (various sections)**

**Developed from FIAF 1.1.2.2.5, 1.5.2.6**

**See also: 1.8.1.1**

1.8.7.1. General rule Where the title cannot be established from any source of information according to 1.8.2, it is necessary to supply one.

Where a title is supplied on the basis that it is known, or has been positively or probably identified, enclose the title in square brackets, and include a note explaining the source of information (see 7.8.3). Where the information is uncertain give a question mark after it.

([String quartet, no. 12] / Shostakovich
Note: Title identified from audition and reading of score)

If no title can be found in any source, and it is not possible to positively or probably establish the title, it is necessary to devise a title. Where a title must be devised by the cataloguer there are two types of titles to consider:

- devised descriptive titles for recordings which have no given titles
- uniform, standardised or conventional form titles for specific categories of recordings

Enclose all devised titles in square brackets.

Note that for some classes of material, it may be appropriate for an archive or cataloguing agency to always supply a devised uniform, standardised or conventional title (see 1.8.7.1.2). **1.8.7.1.1. Devised descriptive titles** Devised descriptive titles should describe the recording as succinctly as possible and may include the topic, place, person and date as appropriate. Enclose the devised title in square brackets: [Bid for the 2004 Athens Olympics]

[Malta's Prime Minister, Mr Mintoff, talks on the renewal of British bases on the island, 1972-03-27]

[President Nixon's visit to China, 1972-02-18--1972-02-28]
5.7. Title

A word or phrase, usually appearing on an item (either digital file or analogue artifact or container), naming the item or the work (often as a group of individual works or recorded sounds) contained in it. In completing the Title Field you are asking yourself, what is the title of this artifact or digital recording that may contain a wide variety of performed poems, stories, or a multitude of sounds recorded on different occasions. The title may not summarize the actual contents of the recording, but it will serve as a reference term for that particular sound recording as it was captured on the audio media artifact, or on that particular occasion of digital recording.

Procedure: When possible, transcribe the title from the source material or secondary sources if source does not have sufficient information. Capitalize the first letter of the first word and of the first letter of proper nouns and additional words according to the appropriate usage in the language in which the material is catalogued.

The following is the order of preference for the source of title information:

1. the item itself (including any permanently affixed labels, or title frames, or the audio itself);  
2. accompanying textual material (e.g. cassette insert, CD slick, inlay or booklet, recording project accompanying documentation such as correspondence, donor agreements, recordist’s worksheets, script, transcript, cue sheet);  
3. a container that is an original part of the item (e.g. sound cartridge, video cassette, sleeve, container for video); or from  
4. a secondary source such as reference or research works, a publisher’s or distributor’s brochure, broadcast program schedule, abstract, index or other available finding aid, container which is not an original part of the item (e.g. a film can be used to store a reel of film, tape box for storing audio tape), or the audiovisual content of the item itself. If the information is taken from a secondary source, cite the source in a Title Source.  
5. For the unidentified information, listen to the recording.
Highlights of the Results
<table>
<thead>
<tr>
<th>Item</th>
<th>Physical Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cataloguer Last name, First Name</td>
<td>Storage Capacity of Artifact</td>
</tr>
<tr>
<td>Partner Institution</td>
<td>Extent</td>
</tr>
<tr>
<td>Contributing Unit</td>
<td>Side</td>
</tr>
<tr>
<td>Source Collection</td>
<td>Playing Speed</td>
</tr>
<tr>
<td>Source Collection Description</td>
<td>Track Configuration</td>
</tr>
<tr>
<td>Source Collection ID</td>
<td>Playback Mode</td>
</tr>
<tr>
<td>Item ID Number/Identifier</td>
<td>Equalization</td>
</tr>
<tr>
<td>Persistent URL</td>
<td>Tape Brand</td>
</tr>
<tr>
<td>Asset Image File Name</td>
<td>Title</td>
</tr>
<tr>
<td>Recording Type</td>
<td>Language</td>
</tr>
<tr>
<td>AV Type</td>
<td>Production Context</td>
</tr>
<tr>
<td>Material Designation</td>
<td>Genre</td>
</tr>
<tr>
<td></td>
<td>Rights</td>
</tr>
<tr>
<td>Locations</td>
<td>Digital Item Fields</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Address</td>
<td>Digitized File name</td>
</tr>
<tr>
<td>Venue</td>
<td>Digital File Path</td>
</tr>
<tr>
<td>Latitude</td>
<td></td>
</tr>
<tr>
<td>Longitude</td>
<td></td>
</tr>
<tr>
<td>URI</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contents</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Stamp Begin</td>
<td>Date Type</td>
</tr>
<tr>
<td>Time Stamp End</td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td>Date Source of Information</td>
</tr>
<tr>
<td>Title of Work</td>
<td></td>
</tr>
</tbody>
</table>
Time Stamping as XML

Titles:

```xml
<item label="title of the item">
  <span label="Alex" begin="00:02:35.00" end="00:04:06.00"/>
  <span label="Continuum" begin="00:00:00.00" end="00:00:00.00"/>
  <span label="Rilke" begin="00:00:00.00" end="00:00:00.00"/>
</item>
```

Hierarchical Structure + Titles:

```xml
<item label="Multiple Events on Same Tape">
  <div label="Part 1: Tomasz Reading">
    <span label="metadata" begin="00:00:00.00" end="00:02:30.00"/>
    <span label="today" begin="00:02:30.00" end="00:04:30.00"/>
  </div>
  <div label="Part 2: Sean Reading">
    <span label="metadata" begin="00:05:00.00" end="00:05:30.00"/>
    <span label="tomorrow" begin="00:05:30.00" end="00:07:30.00"/>
  </div>
</item>
```
How about if something bad happened. Is there something that you'd cling to?

Annotation
Silence

Ira Glass
You mean in terms of a belief?

Jay Allison
Mmm hmmm?

Annotation
Silence

Ira Glass
I mean, I take comfort in the thought that when things seem really sad, it's a comfort to me that; well, everybody's gonna go through this— everybody's gone through this— and, and the problem is, it's too much of a set of truisms to actually be good enough for your series!

Jay Allison
But your show is always looking for a conflict, and something to happen, and something to change— I mean, maybe even THIS show is going to be about how something changes.
Sample Content

Unknown Introducer
08:30:48
One moment. [Silence.], problem that we have here at Sir George. We did try to get a larger hall but it was impossible. To accommodate the overflow, we have set up loud speakers in the little gallery here, Howard [Fink], and in the other one too?

Howard Fink
08:30:48
Outside.

Unknown Introducer
08:30:48
Outside, there are loud speakers. So please don’t all crowd into the room. If you are going to lean against the paintings, we shall never be able to get this room again for poetry readings. Because this, this is a gallery which belongs to the Fine Arts department, we had great difficulty getting it, these paintings are very precious, particularly to the artists themselves [laughter]. I would ask you please to stay away from the paintings [distraction from mic.] That must have been the artist. [Laughter.] We are also waiting for the arrival of someone else, so please be patient. Howard— [laughter] can you ask the security people to turn on the cooling system, the hall is going to be too hot.
Swallow (Spoken Web Audio Metadata Ingest System)
## Going Beyond the Spreadsheet

### Collection Catalogue

<table>
<thead>
<tr>
<th>Cataloguer Last Name</th>
<th>Creator Role 1</th>
<th>Contributor 1 Last, First Name</th>
<th>Contributor 1 URL</th>
<th>Contributor Role 2</th>
<th>Contributor 2</th>
<th>Contributor 2 URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barillaro, Ali</td>
<td>Author/Performer</td>
<td>Unspecified</td>
<td></td>
<td>Series organizer</td>
<td>Unspecified</td>
<td></td>
</tr>
<tr>
<td>Barillaro, Ali</td>
<td>Author/Performer</td>
<td>Unspecified</td>
<td></td>
<td>Introducer</td>
<td>Unspecified</td>
<td></td>
</tr>
<tr>
<td>Barillaro, Ali</td>
<td>Author/Performer</td>
<td>Hoffman, Stanton, ???</td>
<td></td>
<td>Series organizer/Intro</td>
<td>Unspecified</td>
<td></td>
</tr>
<tr>
<td>Barillaro, Ali</td>
<td>Author/Performer</td>
<td>Hoffman, Stanton, ???</td>
<td></td>
<td>Series organizer/Prese</td>
<td>Unspecified</td>
<td></td>
</tr>
<tr>
<td>Barillaro, Ali</td>
<td>Author/Performer</td>
<td>Kiyooka, Roy, 1926-1994</td>
<td><a href="http://viaf.org/viaf/307844">http://viaf.org/viaf/307844</a></td>
<td>Series organizer/Prese</td>
<td>Unspecified</td>
<td></td>
</tr>
<tr>
<td>Barillaro, Ali</td>
<td>Author/Performer</td>
<td>Layton, Irving, 1912-2006</td>
<td><a href="http://viaf.org/viaf/307844">http://viaf.org/viaf/307844</a></td>
<td>Series organizer/Prese</td>
<td>Unspecified</td>
<td></td>
</tr>
<tr>
<td>Barillaro, Ali</td>
<td>Author/Performer</td>
<td>Francis, Wynne, 1918-2000</td>
<td><a href="http://viaf.org/viaf/307844">http://viaf.org/viaf/307844</a></td>
<td>Series organizer/Prese</td>
<td>Unspecified</td>
<td></td>
</tr>
<tr>
<td>Barillaro, Ali</td>
<td>Author/Performer</td>
<td>Fink, Howard, 1924</td>
<td><a href="http://viaf.org/viaf/307844">http://viaf.org/viaf/307844</a></td>
<td>Series organizer/Prese</td>
<td>Unspecified</td>
<td></td>
</tr>
</tbody>
</table>
Deposit a new Item
Browse Items (all items, filter by institution, collection)
Review and edit previously deposited items
Delete Items
Import (From spread sheets)
Export (To JSON, CSV or other required formats)
Manage Collections information (Create, Read, Update)
We need a system that is

- Easy to deploy
- Capable of dealing with a living schema
- Lean
- Cataloguer friendly
Technology stack

Custom Cataloguing Backend
(CRUD functionality for metadata managing)

- Bootstrap
- PHP
- MySQL
- CSV
- json
Use a no-sql approach to allow the metadata schema to evolve over time and adapt to different collections.
"schema_version": "0.1",
"steps": [
  {
    "name": "Cataloguer",
    "type": "single",
    "fields": [
      {
        "name": "catalogue Id",
        "type": "shorttext"
      }
    ]
  },
  {
    "name": "Partner Institution",
    "type": "single",
    "fields": [
      {
        "name": "collection Id",
        "type": "shorttext"
      },
      {
        "name": "persistent URL",
        "type": "shorttext"
      }
    ]
  },
  {
    "name": "Material Description",
    "type": "single",
    "fields": [
      {
        "name": "image",
        "type": "uploadFile"
      },
      {
        "name": "recording type",
        "type": "dropdown vocabulary",
        "source": "Vocabulary/RecordingType.json"
      },
      {
        "name": "AV type",
        "type": "dropdown vocabulary",
        "source": "Vocabulary/AVtype.json"
      }
    ]
  }
]
Swallow

Spoken Web Audio Metadata Ingest System

tomasz.neugebauer@concordia.ca

**********

LOG IN

Forgot Password?
User Management

**MY PROFILE**

Name: Tomasz

Last Name: Neugebauer

Email: Tomasz.Neugebauer@concordia.ca

Password: 

Confirm Password: 

Institution: Concordia University Library

Save
# Managing Collection Metadata

## Collections

<table>
<thead>
<tr>
<th>Partner Institution</th>
<th>Contributing Unit</th>
<th>Source Collection</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Concordia University Library</td>
<td>Library</td>
<td>Collection 1</td>
<td></td>
</tr>
<tr>
<td>Concordia University Library</td>
<td>Library</td>
<td>Collection 2</td>
<td></td>
</tr>
<tr>
<td>University of Alberta</td>
<td>aaa</td>
<td>bbb</td>
<td></td>
</tr>
<tr>
<td>University of Toronto</td>
<td>xxxx</td>
<td>zzzz</td>
<td></td>
</tr>
<tr>
<td>Victoria University Libraries</td>
<td>drgf</td>
<td>dfgfd</td>
<td></td>
</tr>
</tbody>
</table>
Adding an Item

ITEMS > DEPOSIT ITEM

Cataloguer
Partner Institution
Material Description
Notes
Digital File Description
Item Description
Contributors
Dates
Location
Rights
Content

MATERIAL DESCRIPTION

Image
Recording Type
Digital
AV Type
Audio
Material Designation
Cassette
Physical Composition
Lacquer
Storage Capacity
300Mb
Extent
1/8 inch
Playing Speed
16 2/3 rpm
Track Configuration
4 track
Playback Mode
Stereo
Custom Cataloguing Backend

- Web Backend
- JSON Storage
- Export API

CSV  json

Mapping scripts

Elastic Search
Islandora
Avalon
Analysis Tools
<table>
<thead>
<tr>
<th>Category (if applicable)</th>
<th>SpokenWeb</th>
<th>SpokenWest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preliminary Info</td>
<td>Cataloguer Name, Partner Institution, Contributing Unit, Source Collection, Source Collection Description</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Source Collection ID, SW Catalogue Number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Item ID Number/ Identifier, Persistent URL (if available), Asset Image File Name</td>
<td></td>
</tr>
<tr>
<td>The Physical Tape</td>
<td>Recording Type, AV Type, Material Designation, Physical Composition, Storage Capacity of Artifact, Extent, Side, Playing Speed, Track Configuration, Playback Mode</td>
<td>Technical Specifications, Recording Specifications</td>
</tr>
</tbody>
</table>
Thank you

Discussion / Questions