December 2016.

Tomasz Neugebauer, Digital Projects & Systems Development Librarian at Concordia University

Presentation Summary

In early December 2016 the participants of the Literary Audio Symposium (http://spokenweb.ca/events/literary-audio-symposium/) explored the literary historical study, digital development and critical and pedagogical engagement with collections of spoken recordings. Many recordings are already accessible online through repositories such as PennSound (http://writing.upenn.edu/pennsound/) and the Cylinder Archive Project (http://cylinders.library.ucsb.edu/). The potential benefits from a collaborative and coordinated development of digitized spoken-audio archives for scholars, teachers and the general public are only beginning to be realized.

Jared Wiercinski, Tim Walsh and Tomasz Neugebauer, gave a presentation (http://spectrum.concordia.ca/982042) about digital preservation and access with Avalon and Archivematica at the symposium. Concordia Library's interest in digital preservation and access to audio/video collections was primarily inspired by an ongoing discussion with Concordia's Center for Oral History and Digital Storytelling (http://storytelling.concordia.ca/.) COHDS developed a custom software solution in 2010 for managing audio/visual oral history data, called Stories Matter**.** The software has some advanced and specialized research workflows, but it also has challenges with sustainability of development, insufficient digital preservation functionality, and lack of integration with a web browser. We began our requirement gathering for supporting audio/visual research collections by looking to support most of the functionality that was developed in Stories Matter, such as:

- Ability to manage the structure of oral history metadata, which is divided into projects, interviewees, sessions, clips
- Provide for item level permission levels for access
- Attach transcripts, additional documents, interviewer observations
- Save/edit/browse Playlists

There was also specialized research workflow functionality that we would need to consider for interoperability within the library platform. For example, it should be possible to extract the data from the repository using formats that can then be ingested into research tools that can generate tag clouds and network visualizations of the data.

Another example of a research project at Concordia that is particularly relevant to the access interface features for audio was the Spoken Web project through which over 100 hours of literary audio was made available in a web archive using SoundCloud and WordPress (http://spokenweb.ca/sgw-poetry-readings/).

What is digital preservation? Digital preservation is the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability and accessibility of content over the very long term. It is necessary to deal with challenges of file corruption, media failure, and technological change. Most people associate this term with some sort of backup system for digital files. Although backup is a necessary strategy for recovering in the short term from data loss, responsible digital preservation requires management policies, strategies and activities to ensure the long term usability, authenticity, and discoverability and access to digital content. Digital preservation aims to ensure that future users will be able to:

- **discover, retrieve** (if there is no descriptive metadata that makes an object findable, if it is not retrievable, then it is not preserved)
- **interpret** (if there is no administrative metadata telling the user how the object has been changed over time, and by whom; if the context of the other objects that it is a part of is lost, then it cannot be interpreted)
- **manipulate and use** (if the object doesn't include the necessary technical components from an environment in which it was created, such as an audio codec, then it may become unusable)

A digital preservation plan needs to be maintained and implemented, specifying how long-term digital preservation strategies will be applied to digital content types. There is an excellent online tutorial (http://www.dpworkshop.org/dpm-eng/eng_index.html) and workshop series currently hosted and maintained by the Inter-university Consortium for Political and Social Research (ICPSR). The tutorial lists digital preservation strategies (http://www.dpworkshop.org/dpm-eng/terminology/strategies.html), such as the use of durable and persistent media, digital archeology, emulation, refreshing, migration, and the reliance on standards for digitization and normalization. There is no permanent solution for data storage, so the preservation plan needs to be updated periodically. CD-Rs stored in an "ideal" environment were considered best practice up until recently, when Digital Mass Storage systems became the recommended solution because they permit automatic checking for data integrity and refreshing, as well as easier migration. Emulation is usually associated with software preservation, for example, for video games, but it can also be understood in the context of historical audio technology preservation (e.g., digitizing Edison's tinfoil recordings) with the challenge of reproducing all essential characteristics of a performance available in one technological context using another context of a different design. Refreshing is to copy from one long-term storage medium to another of the same type, with no change in the bitstream, for example, from an old 4mm DAT tape to a new one of the same type. Whereas migration is to copy or convert data from one technology to another, for example, from CD-DA to WAV. Normalization is choosing one standard format to migrate all the files to, for example, the IASA (International Association of Sound and Audio Archives) guidelines (http://www.iasa-web.org/tc04/audio-preservation) recommend linear PCM (Pulse Code Modulation, interleaved for stereo) in a WAV or Broadcast WAV file. Post-migration, canonicalization is a technique designed to allow determination of whether the essential characteristics of a document have remained intact.

The reliance on standards is especially important in digitization and migration. IASA recommended standard sampling rates and bit depth should be followed when digitizing. METS (Metadata Encoding and Transmission Standard) (http://www.loc.gov/standards/mets/) is recommended for interoperability. Ensuring the enduring usability, authenticity and discoverability and access would not be possible without metadata. Establishing a repository requires choosing a set of XML namespaces for your metadata (for example: Dublin Core Terms set, MARC relators, audioMD technical metadata, etc.) METS allows externally developed metadata schemes to be used inside of its sections. There are three main types of metadata:

Descriptive, Structural and Administrative, and they are all contained in the METS file. Descriptive metadata is necessary for discoverability and access. Structural metadata is used to display or navigate an object in a digital repository, for example, time coded sections of an oral history interview; it is required for ensuring that an object is usable. Administrative metadata contains management and preservation information useful in interpreting objects and establishing authenticity, for example, the date that a recording was digitized or migrated and the name of the agent (software) used in the migration.

PREMIS, Preservation Metadata: Implementation Strategies, (http://www.loc.gov/standards/premis/) data dictionary is helpful in providing some standard/interoperable nomenclature useful in structuring administrative/preservation metadata in a way that has commonality across archives. It defines the concept of an intellectual entity (an intellectual unit for purposes of management and description) as being represented by objects with related rights and events. The object entity has three subtypes in a hierarchy: representation, file, and bitstream. A representation is a set of files, for example, a series of WAV files and a structure file describing their order. A file has a size and a format, and can contain one or many bitstreams: for example, a Broadcast WAV with a PCM bitstream.

Another key resource for digital preservation is the Trustworthy Repository Audit & Certification (https://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf). This important standard establishes criteria and a checklist for ensuring “trust” through a framework of attributes in a repository. The checklist includes three areas:

1. Organizational infrastructure. For example, organizations must be sustainable and allocate ongoing budget and responsibility for managing digital preservation.
2. Technical infrastructure and security. For example: system must have sufficient storage, duplicate data without loss, demonstrate reliability, do error checking, store metadata and reliably link it to data objects.
3. Digital object management technologies. Repository software must conform to the Open Archival Information System (OAIS) model (https://public.ccsds.org/pubs/650x0m2.pdf) .

The OAIS model is a point of view, a meta theory of repositories developed by the space science community. It defines three types of information packages (submission, archival and dissemination) and six main repository functions: ingest, archival storage, data management, administration, access, and preservation planning. OAIS doesn’t replace more specific models for research workflow interfaces, but it is useful in structuring a repository system and understanding its boundaries.

Archivematica (https://www.archivematica.org/en/) is a free open source product with a strong user community. The software bundles micro services such as FFMPEG (https://ffmpeg.org/) and JHOVE (http://jhove.sourceforge.net/) in a transparent way, with the possibility of modifying or replacing the microservices used. It follows the OAIS model closely, structuring its interface using the same repository functions and information packages defined in the standard.

A walkthrough of preservation activities with Archivematica begins with an initial transfer of the metadata and digital objects, where checksums are created in the submission information package. Micro services are used to verify, identify, characterize and validate. In the process, unique identifiers are assigned to each object, checksums are created and verified, files are identified through PRONOM (http://www.nationalarchives.gov.uk/PRONOM/), information such as bit depth and encoding is extracted, and the files are validated with JHOVE. The normalization step is carried out per the preservation plan rules for what is to be the preservation and access format, and the audio file is transcoded if needed to create the preservation and/or access copy. Descriptive metadata that is stored in the resulting METS file can also be added during ingestion. The resulting archival information package includes logs, checksums, packaging information, the metadata and the digital objects. Lastly, the dissemination information package contains the access digital files and metadata that can be passed along to an access system and users.

Archivematica is a proven open source digital preservation solution, with a sustainable development model and a large worldwide community. Currently, it models objects down to the level of the file, but with the introduction of MediaConch, future development should model down to the bitstream level, showing the different tracks/streams within an AV file.

Avalon Media System specializes in providing access to audio and video collections. In an ideal world, digital preservation and access for all formats would all be accomplished within the same system, but the reality is that different access systems exist for target formats and audiences. The Avalon Media System is designed with the purpose of providing access to large collections of digital audio and video, with a community of educational, media and open technology institutions. Given that audio/video collections are of particular interest to Concordia Library, this is one of the key strengths of Avalon, but also a challenge, given that we do also need to provide access to text and image objects. The strengths of Avalon include the fact that it is born out of the successful Variations Digital Music Library project from Indiana University Bloomington, and its strong development community with a proven track record of feature-packed releases.

The key features of Avalon that make it a top choice for an access system:

- Support for a hierarchical structure for objects, with units, collections, items, sections and hyperlinked labelled time-stamped start and end points.
- Sophisticated access and permission controls at collection and item level
- Robust metadata and faceted discovery
- Playlists including items and sections, public or private.
- Captions and subtitles
- Integration with HydraDam2 (http://www.avalonmediasystem.org/blog-post/hydradam2) and Spotlight exhibition tool (https://library.stanford.edu/projects/spotlight)

Our rationale for selecting Avalon as a top choice is centered on the fact that we need a system for humanities qualitative research data in audio/video format. Avalon is an open source system developed within a larger context of the Hydra project (https://projecthydra.org/) with an active development and user community and sustainable modular architecture with Fedora (http://fedorarepository.org/).

Our concerns for Avalon, in addition to its limited capabilities for non-audio/video documents, include the fact that it is comprised of a complex set of software components that can be difficult to install and maintain. Community sustainability is also a concern, given that Avalon has not yet announced any funding beyond the two-year grant ending in January 2017 from the Andrew W. Mellon Foundation that helped to secure its development. There are also missing features in Avalon, that we hope will be added, such as: support for administrative, technical and provenance metadata, transcription integration, the Oral History Metadata Synchronizer (http://www.oralhistoryonline.org/) that is mentioned in the development roadmap but which is unscheduled, specialized research workflow interface is limited compared to a platform like Databrary (https://nyu.databrary.org/), and there is a lack of support for user annotations.

Avalon is developed by a team that has experience with audio through Variations software. Avalon's excellent support for video is a key requirement for oral history, but the fact that Avalon acknowledges audio as a distinct format with its own requirements is important for spoken word literary content and other audio types. Audio research data is important for scholarly research in many disciplines; Mark R. Roosa's (2015) "Sound and Audio Archives" (In M. V. Cloonan (Ed.), *Preserving our heritage : perspectives from antiquity to the digitial age* (pp. 278-287). London: Facet Publishing.) lists audio types such as: Linguistic, Folklore, Oral history, Ethnographic, Dialectic, Music, Ethnomusicology, Bioacoustics (e.g. orthinology, etc.), Spoken word (p. 279). Although commercial tools like Soundcloud, Spotify and Mixcloud offer robust functionality, there is a large gap from these tools to how audio is supported in institutional repositories.

In the long term, ideally, a workflow for ingesting content from Archivematica's Dissemination Information Packages (i.e., the access files that are optimized for the Web) into Avalon would need to be developed. In the short term, it is possible to keep the Archivematica and Avalon workflows separate and rely on Avalon directly for all transcoding during ingesting.

Ingesting into Avalon implies creating a Unit, such as "Library Special Collections" or "Center for Oral History", adding specific Collections, such as "Irving Layton Collection" with some descriptive metadata and then creating Items with audio/video uploads that are described with a detailed set of descriptive metadata such as a list of contributors, title, time periods, location, terms of use, notes, etc. The access control at the collection and item level allows for assigning access by user, internal or external group, IP address or range, and to limit access to items by date period.

The fact that version 7 of Avalon will integrate the Spotlight exhibition tool developed at Stanford University is particularly promising. Spotlight extends the repository ecosystem by providing a means for reusing digital content in other scholarly websites, allowing for the possibility of pulling content out of Avalon and into new research contexts.

It is worthwhile to mention some of the other access systems for audio and video that offer similar or complementary functionality: Islandora, Mukurtu and Databrary.

Like Avalon, Islandora (https://islandora.ca/) is an open source front-end for a Fedora repository. It is a general use repository system that supports many formats and different audiences or knowledge domains. Islandora was originally developed by the University of Prince Edward Island's Robertson Library, but is now implemented and contributed to by an ever-growing international community. Islandora community releases solution packs which empower users to work with data types (such as image, video, and pdf) and knowledge domains (such as Chemistry and the Digital Humanities). Solution packs also often provide integration with additional viewers, editors, and data processing applications.

Islandora is an open source project with a large and sustainable user community. Like Avalon, it is built on Fedora Commons and Solr, but the front-end of Islandora is developed with Drupal (https://www.drupal.org/). Islandora has a larger user community and a greater number of implementations than Avalon. However, Islandora is a general repository solution that we would have to customize, using some set of solution packs to deliver the audio/video support that we need. One example of a particularly useful recent development in Islandora is the Oral History Solution Pack (https://github.com/digitalutsc/islandora_solution_pack_oralhistories.), which includes the type of transcript integration that we would like to see in Avalon (an example item: https://digitalscholarship.utsc.utoronto.ca/projects/islandora/object/ehrn%3A538). The transcript integration for Drupal was developed in partnership with Edward Garrett, a linguist and software developer who is the sole proprietor of Pinedrop (http://pinedrop.com/). Pinedrop is an interesting service option for scholars and researchers in the arts and humanities looking to develop custom websites and research tools for their projects.

Mukurtu (http://mukurtu.org/) is a free open source platform that is built with indigenous communities to manage and share digital cultural heritage, with a first priority of fostering a relationship of respect and trust. Murkurtu is managed by the Center for Digital Scholarship and Curation at Washington State University and funded in part by the National Endowment for the Humanities and the Institute of Museum and Library Services. Mukurtu is developed for digital content that primarily requires cultural protocols for rights and permissions, grassroots consensus and community control. It uses traditional knowledge metadata alongside the Dublin Core schema. Although Mukurtu is the most appropriate choice for some content types, it may not address the general case.

Databrary (https://nyu.databrary.org/) is a video data library specializing in storing, managing, preserving, analyzing, and sharing video and other temporally dense streams of data for developmental science. The project is based at New York University and Penn State with grant support from the U.S. National Science Foundation (NSF) and the National Institutes of Health (NIH). Databrary features integration with the open source cross-platform video coding and data visualization tool Datavyu (http://datavyu.org/), a feature rich interface with sophisticated search/browse functionality and support for a wide range of formats. Although Datavyu is a social science tool, the functionality it offers is also applicable to other research contexts, such as the ability to simultaneously view multiple videos for comparison purposes, or to input time-stamped video annotations.

The presentation at the Literary Audio Symposium raised questions that we continue to think about. From a software architecture point of view, what conditions are needed to facilitate the development of a sustainable feature-rich access platform for audio/video content? Does audio/video content need to be made accessible in a separate system, or can it be accommodated in the same digital asset management system as all other special collections formats? Given that there is an inherent problem of lack of awareness around digital preservation issues, how do we promote the development of sustainable and responsible preservation planning for audio and video? Should we continue to build a wide variety of niche repositories or aim towards a strategy of using centralized repositories? Governments and cultural institutions (libraries, archives, and museums) have the responsibility of preserving digital research and cultural content and making it accessible.