Building the plane as we're landing it
The final migration of the American Religious Sounds Project

by Kate Topham

INTRODUCTION

I have difficulty sometimes talking to people who don't race sailboats.

So begins a brief speech in season 3, episode 4 of *The West Wing.*

It comes from Bruno Gianelli, and he's talking about how minute choices can change the course of a presidential election. In a sailboat race, when a piece of kelp gets stuck to the hull, you can't shove a stick in the water—it will cause more drag than then kelp ever did.

See, what you gotta do is you gotta drop it in and let the water lift it out in a windmill motion. Drop it in, and let the water take it by the kelp and lift it out. In, and out. In, and out, till you got it.

Digital preservation in digital humanities functions the same way. While preservation is absolutely critical for the success of the project, it won't work if it slows the *work of* the project.

For the last four years, I have served as the Digital Archivist for the American Religious Sounds Project, a multi-year endeavor to document the diversity of American religious life through sound. I boarded this sailboat partway through its journey, given the task of (metaphorically) scraping the kelp off the hull so we could finish the race. Thousands of recordings, hundreds of ephemera, but major gaps in metadata and few standards stood in our way, and I had to fix it without disrupting the progress of the project itself.

PROJECT HISTORY

The American Religious Sounds Project (ARSP) began with a simple question: "What does American religion sound like?" Co-directors Amy DeRogatis and Isaac Weiner sent undergraduate students into the world to capture the broadest possible portrait of American religion.

Eight years later, the project spanned two institutions (Michigan State University and Ohio State University), dozens of curated exhibits, and nearly a terabyte of data. The collection includes recordings of Serbian Orthodox fish fries, Theravada Buddhist chanting, Wiccan Yule rituals, shape note singing, and everything in between. A resource this comprehensive requires a solid backbone of source material, quality control, and rich metadata. However, despite efforts to control information flows and enforce standards, many recordings lacked contextual information that was missing or difficult to find. Recordings were misplaced as they were moved from folder to folder.

The project team also needed a plan for these recordings once the project ended. The G. Robert Vincent Voice Library (VVL) at Michigan State University offered a long-term home for the audio recordings, but they would need metadata formatted to their standards, a considerable effort even if the ARSP materials had consistent metadata. No one on the American Religious Sounds Project team or at the Vincent Voice Library had the bandwidth for a project of this magnitude.

Kate Topham is the Digital Humanities Archivist at Michigan State University.

In 2018, the project applied for and received a four-year grant from the Henry Luce Foundation to fund the next phases of the project. I joined the project as the Digital Archivist to facilitate the Archival Preservation initiative of the grant. This paper reports on the processes, challenges, and lessons learned over my four years preparing and migrating this collection.

BUILDING THE PLANE AS WE FLEW IT: PROCESS, PRAXIS, PHILOSOPHIES

Many institutions struggle to implement digital preservation best practices due to a variety of factors, including technology, cost, and expertise (Conway 2010, Kastellec 2012). As a grant-funded archivist with half time on this single project, I had more resources to devote towards preservation and migration than many large institutions.

In his book, *The Theory and Craft of Digital Preservation,* Trevor Owens lists sixteen axioms for digital preservation. While nearly all are applicable to my work on this project, three are the most relevant to the larger themes of this migration, which I have divided into three movements.

Movement I: Process

Axiom 4, "Nothing *has been* preserved, there are only things *being* preserved" (Owens 2018), emphasizes that archiving is an active process, a constant commitment to keep materials accessible. This is never more relevant when preserving materials that are part of an active digital humanities project that needed to adapt as the project developed.

When I joined the team, the project had turned from a sporadic, coursework-based model to a more continual cycle of recording and publishing clips.
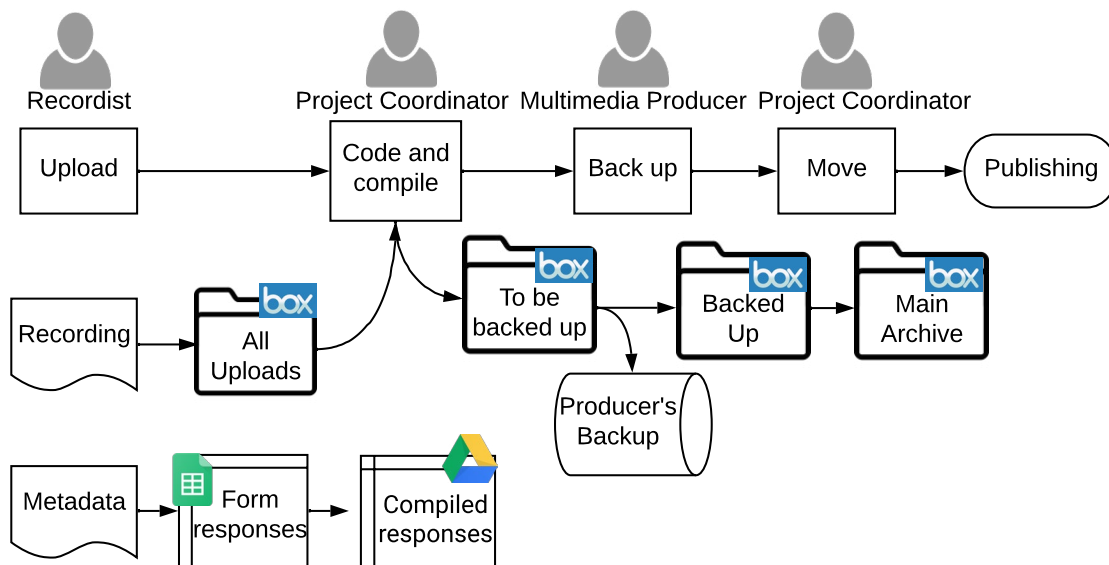

Figure 1: pre-publication workflow for the American Religious Sounds Project

In the pre-publication workflow (Fig. 1), recordists uploaded new recordings to Box, then entered metadata into a Google Form. The Project Coordinator edited the name of the file and moved

it to another folder. At that point, the Multimedia Producer would download the recording, make a backup, and create an edited clip. The finalized clip was uploaded to a different folder on Box, at which point the Project Coordinator would move it to its proper place in the archive.

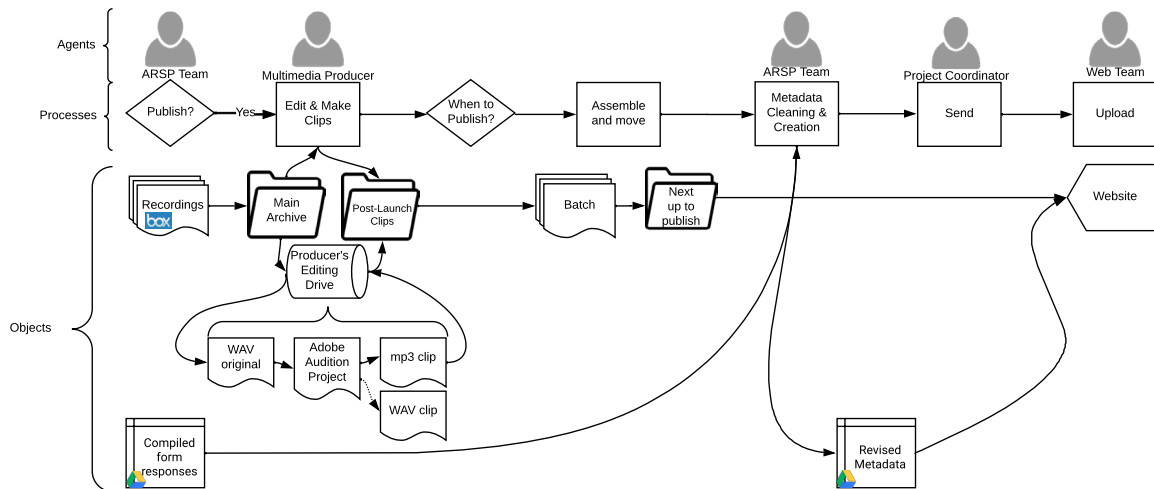Edited clips were selected and published to the ARSP website in batches.

Figure 2: publication workflow for the American Religious Sounds Project

In the publication workflow (Fig. 2), the team would divide up the selected clips and each create metadata to be used on the website collection. This information was kept in yet another spreadsheet on Google Drive. When it was complete, it was sent to the website team to update the website.

There were some file naming standards, but they were used inconsistently. Metadata was not always updated when a file name was changed. We directed recordists to implement a new file-naming convention themselves so that the Project Coordinator would not need to re-name the files, or at least would make only minor changes.

I worked with the Project Coordinator to structure the Box archive such that each site code had its own folder and sub-folders for each kind of file. The Producer and Project Coordinator spent less time moving files between folders, and thus fewer files were lost. I also implemented standards for file formats to minimize loss over time and upon migration.

The greatest problem in this workflow was the metadata. Metadata for field recordings was kept in a spreadsheet in Google Drive while the actual files were kept in an Ohio State University instance of Box. This meant the recordist often forgot to complete one or the other. The form also collected little to no metadata about photos and release forms, resulting in many of those files being lost or left without context.

The team did not think of the clip descriptions as metadata for archival management even though in many cases this was the only metadata that we had. Keeping the metadata separate meant team members had to dig through other spreadsheets or rely on memory to create descriptions. I insisted that we not only store all metadata together, but that it needed to be stored on Box with the files in question. Luckily, the team fully recognized that this system was problematic, so they had plans for a solution.

Movement II: Praxis

Axiom 9: "Digital preservation is about making the best use of your resources to mitigate the most pressing preservation threats and risks" (Owens 2018).

Once policies were in place to ensure new recordings would be covered, I could turn to the most critical threat to the collection: the metadata. Alison Furlong, our Project Coordinator, built AURA, or Application for Uploading Religious Audio, a custom Rails app designed to solve three major problems:

1) unify the upload process for files and metadata,
2) verify location data, and
3) export data for the website.

With AURA, metadata and files would no longer be stored separately and files would no longer need to be moved around on Box. I saw further potential for this tool to facilitate the migration to the Vincent Voice Library, to have an all-in-one tool for preserving this collection.

However, for AURA to work in the way I hoped, we needed *all* our legacy data. There were eight spreadsheets of metadata for field recordings and two for edited clips, each with a different set of fields and vocabularies. Using OpenRefine, I converted each spreadsheet to match the most recent metadata spreadsheet and controlled vocabularies. I extracted terms for sounds we were interested in, such as instruments, singing, speaking, and bells, to further enrich the sound tags in our metadata. This was particularly useful in extracting names of speakers, which the ARSP never collected, but which was important for the Vincent Voice Library.

The problem remained that many of the file names in these records did not match what was on Box. I created a pair of python scripts to query the Box API for each file name listed in the metadata and. If it found a match, it associated the new file name with the metadata and added it to a compiled csv file. File names that could not be matched were looked for manually, but most of them were never found.

In our conversations about the legacy metadata and AURA, Alison and I kept running up against the question of how to model this data. ARSP had been using a sort of collection-level metadata. Recordings were described in groups based on a particular event – sometimes a particular service, a meeting, or an event hosted by a community; but also included ambient or incidental recordings not tied to a specific event.

In OpenRefine, I created "event" identifiers for field recordings and clips based on their site code and date. This enabled me to create metadata for field recordings that did not have metadata themselves but did have an associated clip with metadata, and vice versa. I then split the metadata into three tables: one for events, one for field recordings, and one for clips, all linked by their event identifier.

Alison implemented this model in AURA. Each time a recordist went out to record, they would create an "Event" record, fill out contextual information, and then create metadata for each file they uploaded.

I approached migration to AURA in batches to match the recording and publishing workflows. Going site by site, I extracted metadata from OpenRefine and sent it to Alison for batch upload to AURA. Working in pieces allowed us to keep track of errors and ensure nothing was left behind.

While the migration was ongoing, I turned my attention to the files that did not have any associated record. I combed through the Box files, checking everything against the metadata and noting files that had no metadata as I went.

In doing this work I discovered yet another place the project kept metadata. Before the Google Form system, "metadata" was collected as part of course work. Students filled out a Word document form that included some standard information as well as field notes, with questions shifting to suit the needs of that specific course. I copied the information from these Word documents as best I could into our Google Form so that it would match the metadata standards in AURA.

When I finished combing through the archive, I had over 1,400 sound files that still had no metadata. I knew that many of these files might be unusable due to lack of context or low sound quality, but I couldn't take on that assessment alone.

Movement III: Philosophies

Axiom 15, "the scale and inherent structures of digital information suggest working more with a shovel than with tweezers" (Owens 2018), encourages digital practitioners to work at scale as much as possible to perform actions at scale. I struggled with this principle throughout my work on this project, especially with regard to assessment. Migration is universally a slow and difficult process, but audio collections in particular require more effort: specialized knowledge, more storage space, and more metadata (Blewer 2020). Being the premier archive of *religious* sound also required extra care in the handling of the materials.

The ARSP team had continual, constant conversations about our project values, placing respect for the communities represented by these recordings above all. Our collection includes recordings of marginalized religions, of people who are vulnerable due to their identity, their immigration status, and/or their socioeconomic position. Decisions about whether to keep and how to describe a given recording required careful thought and often a group conversation.

Knowing that this was the mode in which the ARSP team was most comfortable, I broke the assessment process into small pieces to allow us to integrate it in our regular workflows. First, I sat down with the ARSP team to go over the remaining orphan files. We identified the recordists and reconstructed metadata where we could, but there were still hundreds of recordings left. Many of these recordings were made by students who were long gone or were low-quality recordings from the early days of the project, yet represented unique communities within our collection.

I set a goal for the summer of 2021 to assess all the remaining recordings. I met with Amy and Isaac every other week in conjunction with their other ARSP meetings. I sent out a list of files to listen to and asked each of them to listen to them and determine if they were worth salvaging. Once again, matching the preservation workflow to the overall project workflows proved fruitful. Initially we struggled with letting anything go, but by breaking it down to five or six recordings, it was much easier to decide that something wasn't worth saving. We also identified several recordings to publish on the website. By throwing "more product, less process" to the wayside, we enriched the diversity of locations, denominations, and sounds in our collection

BUILDING THE PLANE AS WE'RE LANDING IT

Scholarship only moves forward. Digital preservation is critical for the success of a digital humanities project, but not if it stops the flow of work. Keeping preservation involved in the larger project

workflow ensures that it remains a priority. The more I was aware of what was going on in the other aspects of the project, the better sense I had of what we needed and what it meant for preservation. It helped me find our "good practice" when "best practice" wasn't possible or necessary, focusing on what was most important about the collection and the project as a whole.

As much as I preached the value of letting go of recordings that could not serve the project, I find myself reluctant to cede control to the Voice Library. An external hard drive sits on my desk, waiting for the final description edits and metadata embedding. But if you think I'm going to miss even one opportunity to future-proof this collection, I have to tell you, you're absolutely out of your mind.

BIBLIOGRAPHY

Barclay, Paris, dir. *The West Wing*, Season 3 episode 7, "The Indians in the Lobby." Aired November 21, 2001, on NBC.

Blewer, Ashley. 2020. "Pragmatic Audiovisual Preservation." Digital Preservation Coalition. http://doi.org/10.7207/twr20-10

Conway, Paul. 2010. "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas." The Library Quarterly. University of Chicago Press. https://doi.org/10.1086/648463.

Kastellec, Mike. 2012. "Practical Limits to the Scope of Digital Preservation". *Information Technology and Libraries* 31 (2), 63-71. https://doi.org/10.6017/ital.v31i2.2167.

Owens, Trevor. 2018. *The Theory and Craft of Digital Preservation*. Baltimore, Maryland: Johns Hopkins University Press.