AAPB-NDSR Final Assessment

Host Mentors: Dave Rice & Catriona Schlosser
Organization: City University of New York Television (New York, NY)
Resident: Andrew Weaver
Date: July 12, 2017

Original Project Plan

The original plan was developed in order to be responsive and flexible to the skills and objectives of the resident. The plan started with the resident analyzing the organization's archival practices and evaluating them against the NDSA Levels. The resident was then able to identify areas where that assessment may be improved and proceed to propose and enact changes, so that a second NDSA Level assessment at the end of the project could demonstrate the impact. The project plan also theorized that perceptual hashing practices could be implemented within the archive's processing and that the resident could experiment with usage of perceptual hashing in order to facilitate more opportunities for collection management and discovery.

Changes to the Original Plan

The plan left many of the details of the plan up to the resident, based on their assessments and judgements, so overall there was not a feeling of change as the plan proceeded, but the plan became more clear as the resident determined what NDSA Levels could be addressed and how.

In regards to the perceptual hashing aspect of the project, at the time the original plan was written it was unclear what tools could be used to meet the goals. At the time the resident started it was unclear if this could be done efficiently. After some experimentation, Andrew adjusted the plan to use audio perceptual hashing rather than video perceptual hashing. During this time, he integrated tools to assess audio, produce hashes, and manage them in a database. At some point in this project, an FFmpeg contributor provided a draft of a video perceptual hash tool and Andrew shifted his work from the audio perceptual hasher to the draft video perceptual hasher. This FFmpeg filter stayed in experimental status for several months, while Andrew managed development with the coordination of a stable release of FFmpeg and a patched version to support perceptual hashing. After some nudging to the development community, the review work was completed and FFmpeg now integrates a video perceptual hashing filter, called 'signature'. Andrew integrated this hashing into our processing workflows and created tools to search for video likenesses against our databases. The objectives were met but the process and timeline were impacted by the availability of tools and coordination with development communities.

Project Milestones & Deliverables

NDSA Levels Assessment
Andrew's first project was the NDSA levels assessment. He accomplished this task by observing CUNY TV workflows, speaking with staff and directly asking us questions about our digital assets. With this information, Andrew recommended the areas where CUNY TV could improve. He also created a preliminary document with a summary of his findings, information about his methodology and an annotated table with the NDSA Levels of Digital Preservation. The table was particularly helpful in showing our strengths and weaknesses. With this information, Andrew made suggestions on which categories we can improve. Following the preliminary assessment, Andrew then drafted a post NDSR assessment in May 2017. This evaluation showed that CUNY TV gained 3 points (his documentation describes his methodology) and gives us a better understanding of our overall digital preservation workflows. The post evaluation will help shape which NDSA categories we will focus on in the future. The pre and post evaluations are both on the CUNY TV institutional Github page, so we are able to reference it when needed.

Premis Database
Following Andrew's NDSA levels assessment, he started work on a mysql database called PremisDB. This database was incorporated into the CUNY TV microservices and arranges our metadata in a more readable fashion. In our project proposal, we wrote that we have a large amount of metadata, yet we did not have it represented in a centralized database. The PremisDB gives us even more information about our files such as when it is ingested, when an access derivative is created or when it is written to LTO. It allows us to track a file from its creation up until it is written to LTO tape. This added fixity information helped CUNY TV gain another NDSA level. Along with the actual database, Andrew also wrote up documentation outlining the function of the database, how to install it and what information it provides. Since this database lives on an open platform, other archivists can use it and incorporate it into their workflow as well. The code and documentation can be found on the mediamicroservices Github page: https://github.com/mediamicroservices/mm.

Perceptual Hashing
As mentioned above, Andrew used the FFmpeg hashing filter 'signature' to fulfill the perceptual hashing aspect of the project. Although the timeline shifted as a result of the limited technology, Andrew created a script called 'makefingerprint' that generates a video perceptual hash for an input. That script can also be found on the mediamicroservices Github page: https://github.com/mediamicroservices/mm. This script creates a folder called 'fingerprints' and these hashes are uploaded into the Premis database. We hope to use this information to connect raw footage to the master and in the process, make cataloging much faster and more efficient.

Project Impact

Overall

As the project and work of the resident were working directly with the workflows and procedures employed by the archive in nearly all processing, we would see the impact of the project clearly as it progressed. Andrew added 267 commits to our primary microservice repository with impact to over 2,500 lines of code. In vrecord, our recording utility, he added 59 commits and added 649 lines of code and in our LTO handling utilities he added 54 commits and added 486 lines of code. These efforts would be reviewed, merged into our work, and updated across our computers as releases were marked.

Audio Handling

The archive has gradually improved its audio handling over time. For example the archive may receive all sorts of production content with different audio channel configurations and loudness where we'll have to programatically assess and adjust audio in order to provide broadcast access copies that have the required audio channel mix and loudness metrics. Prior to Andrew's project we had in place methods to downmix audio to stereo (conditionally depending on the arrangement of the production file) and adjust the overall volume to achieve a target loudness. Andrew improved on this by integrating the loudnorm filter from FFmpeg. With this improvement, our broadcast access copies would be adjusted to match target loudness range and peak levels in addition to target volumes. Andrew's work also identified and repaired prior bugs and inefficiencies in our audio handling workflows.

In Andrew's research time, he also developed an audio capture utility called https://github.com/amiaopensource/audiorecorder/ which we may likely use later this year in an audio digitization project.

Community

Andrew contributed significant amounts of work in support of public and open source projects. He has been an active contributor to ffmprovisor, FLAC, XFR Collective, and helped make CUNY Television's microservices and utilities more accessible to the public.