Interview with Ben Turkus and Genevieve Havemeyer-King of NYPL



Hey Ben, hey Gen! Introduce yourselves please.

(BT) Hi Ashley! I'm Ben Turkus; a long-time fan of MediaInfo/MediaConch, first-time interviewee. I'm the Assistant Manager of Audio and Moving Image Preservation at New York Public Library, and previously, I worked at the Bay Area Video Coalition in San Francisco, on projects similar to MediaConch (shoutout to QCTools). Rewinding even further, in what feels like a lifetime ago, I had a semi-illustrious career in the restaurant business (check out this other lame/hilarious/embarrassing interview if you dare).

With the support of the Andrew W. Mellon Foundation, NYPL is currently engaged in a major audiovisual digitization effort. Not only have we identified 230,000+ media objects to be ?high value, high risk,? and worthy of preservation, but we've begun working hard to actually reformat as many as possible, through a combination of in-house and outsourced digitization efforts. To date, we're about 75% of the way through an initial 60,000 project set for 2016-2017. In many ways, it's because of tools like MediaConch that we've been able to increase production without sacrificing quality.

When you're working with numbers like these, it can be easy lose sight of the content that you're striving to save and make available to the public, but I will say this: NYPL's collections are unbelievably rich and varied, and almost every single day we discover something incredible. It doesn't feel right to pinpoint specific collections, but just yesterday Gen and I had a fight over who would get to qc some 2-inch, 24 track reel-to-reel audio recordings from Arthur Russell. It got bloody.

(GHK) Yeah, right. Hi, I'm Genevieve Havemeyer-King. I've been the Media Preservation Assistant for the NYPL's Preservation of Audio and Moving Image Unit (PAMI) for about a year. As part of the initiative to preserve the Library's at-risk audiovisual research collections, my primary role so far has been to assist with coordination of mass-digitization for magnetic and optical media. This includes (but is not limited to) refining and documenting our technical specifications, implementing a robust quality control workflow for our high volume of deliverables, tracking digitization progress, and maintaining inventory of our vendor shipments. As the first of several gatekeepers for NYPL's media ingest system and digital repository, I also assist Ben with migrating and tracking assets on their journey towards long-term digital preservation.

What does your media ingest process look like? Does your media ingest process include any tests (manual or automated) on the incoming content? If so, what are the goals of those tests?

(GHK) I should start by mentioning that our unit primarily manages reformatting for audio and moving image (AMI) research collections; born-digital records, still images, and other collections are managed by our Archives Unit and Digital Imaging Unit colleagues (some of whom also use MediaConch!).

Before AMI deliverables reach NYPL's Media Ingest system, they are reviewed and tested using a combination of custom scripts, proprietary and open-source tools, and manual content inspection. We check fixity, technical specification conformance, metadata validity, signal quality, and adherence to our preservation policies. Our suite of tools includes bagit.py, JSON schema, ajv, MediaConch, MediaInfo, QCTools, Wavelab, and human eyes and ears. There are a lot of goals in running these checks, but namely we strive to: * Ensure the integrity of our bits;
* Ensure that metadata created for physical and digital assets is accurate and captures their preservation history to a reasonable degree;
* Achieve consistency between in-house and vendor produced deliverables; and
* Catch and communicate about errors as soon as possible, before they are out of our hands and on their way to our repository.

If they pass this review process, they are migrated to another pre-ingest staging area, where they undergo a similar series of automated tests to ensure they are safe to ingest.

Our metadata schema, specifications, and many of our customized tools are available on GitHub: ami-metadata, ami-specifications, and ami-tools.

Where do you use MediaConch? Do you use MediaConch primarily for file validation, for local policy checking, for in-house quality control, for quality testing for vendor files?

(GHK) MediaConch is a pretty integral part of our Quality Control workflow, and we make use of it for all of the above tasks. We created our own MediaConch policies based on the built-in and public policies, but because our specifications require that preservation master files retain many of the same characteristics as their physical source objects, we sometimes need to adjust or create new policies that are appropriate to particular media types as we encounter them. This means that ?Fail' results act more as flags for closer investigation, and that some specifications change slightly as needed.

We use the CLI for batch-checking entire shipments of deliverables, directly on the storage media on which they arrive (which can run anywhere from 400 to 6000 files, approximately 6TB per shipment, depending on the type of media on a given hard drive). Operating on a write-protected drive, we use multiple ?find' commands to simultaneously identify specific media types, apply a specific policy, and output the report as a .csv. We use the GUI for one-at-a-time testing of sample files and pilot projects, as well as investigating specific errors when an asset fails. We hope to create more complex tools that will integrate our metadata files (JSON) to inform exactly which policy is used for each file, for a more automated, streamlined system.

At what point in the archival process do you use MediaConch?

(GHK) We use it exclusively in our pre-Ingest quality control workflow, which we carry out as soon as we receive a shipment from a vendor, and also on in-house deliverables before they're migrated to our server, where they are then staged for Ingest.

(BT) We also to use MediaConch's ?system? policies, and other organizations' public policies, as targets/inspiration when drafting or refining our own technical specifications. In this way, MediaConch is there from the very beginning (or, in some cases, maybe it should have been). As with QCTools, there's this educational side to MediaConch that, for me, is absolutely essential. On numerous occasions, we've used MediaConch to work backwards, referring to either the FFV1/MKV implementation checker, or the ?Matroska is well described? or ?TN2162 compliant?? system policies to gain a better understanding of the files that we're creating/having created for us.

I could go on and on about this, but in short: MediaConch has helped us right some of the self-descriptive wrongs that are presented

by various capture hardware/software configurations. By clueing us into the ways that requesting or creating ?uncompressed video in Quicktime? is not really sufficient, MediaConch has pushed us to rectify issues either during a transcode, or by learning to use cool tools like MKVToolnix.

Do you use MediaConch for MKV/FFV1/LPCM video files, for other video files, for non-video files, or something else?

(GHK) We use it for both video and audio. We began by checking our Quicktime-wrapped, 10-Bit Uncompressed video preservation master files, and have now switched to MKV/FFV1/LPCM, and we also use it to check our MPEG-4 service copies. For audio, we use it to check our Broadcast Wave preservation masters and edit masters.

(BT) Recently, we've also been identifying and flagging outlier formats that may need more in-depth analysis, such as early digital audio formats that were recorded on videotape, digital audio in general, HDV, and DV-family video. These formats may not have consistent characteristics that are easily checked against a standard ?policy?, so we're still exploring how best to approach them at-scale.

Why do you think file validation is important?

(BT) There's a baseline of quality and conformance that we'd like to adhere to. Beyond that, some nuanced aspects of validation require some fluidity; there is no consensus on whether certain specifications are essential. We try to ensure that files are as self-descriptive as possible, with the understanding that preservation is a process that involves many stakeholders, and that fluctuating resources and capabilities, as well as the complex nature of audiovisual media, impact our ability to ensure and enforce certain requirements.

Anything else you'd like to add?

(GHK) As a practical tool, MediaConch has made a big difference in our ability to manage large-scale digitization. We've caught many errors that may have been overlooked if not for the ability to check a high-volume of media all at once. Some things we've caught include random 8-bit preservation master files in the mix, inconsistency in audio channel configuration among access copies, and files for which exceptions to our specs had to be made but weren't communicated right away. By catching these early, and using the accompanying metadata to help diagnose where they may have originated, we've been able to identify and prevent several issues from being replicated throughout an entire preservation project.

Using it has also continuously underlined how complex and diverse audiovisual formats are, and how a nuanced approach to preservation can sometimes lead to a rabbit hole of requirements and specifications. So, it's helped us rethink some of our own processes - how we can keep simplifying things to balance our requirements with our scale of production - and has inspired more workflow development for ideal automated QC processes. The tool, its developers, and its user base continue to provoke much-needed dialogue about format and codec standardization in the preservation community, which our whole field really benefits from.

(BT) I think we've found that engaging in the practice of conformance checking, and participating in the development of tools to support this practice, can influence wider discussion about language, standardization, and compliance for all kinds of pre-existing ?standards.? Again, returning to the ?uncompressed video in Quicktime? question, if a vendor is incapable of creating files that adhere to the parameters set forth in Apple's Technical Notes (TN2162), and if that vendor can't capture directly to FFV1/MKV, we're presented with an interesting challenge. How can we ensure that vendors are creating the ?right? kind of Quicktime files for transcoding to MKV/FFV1; what are the risks and compromises involved in this approach?

There is a large amount of trust required in our vendors, that they are reformatting at the specifications appropriate to particular formats (i.e. audio sampling rate); but while MediaConch will ensure that a file may pass our specifications, it cannot ensure that the specification for a given format is what was appropriate for that particular object. For us, this makes MediaConch an excellent tool for supplementing manual quality control with automated systems to speed up what can often be a slow process.