Interview with Julia Kim

Julia Kim at NYU (photo credit: Elena Olivo)

Hey Julia! Introduce yourself please.

Hi, I'm Julia Kim. I'm the Digital Assets Specialist at the American Folklife Center at the Library of Congress. I've been here for just about 2 years now. I get to work with a good mix of both digitized and born-digital multi-format collections. While we create collections documenting concerts and events, the heart of our collections are ethnographic materials from collectors, folklorists, and anthropologists. AFC highlights include the Lomax family archival collections, the indigenous Native American wax cylinders, StoryCorps, and ?Web Cultures,? but in-between, we get a little bit of everything that is evidence of ?folk? around the world. I'm considered a specialist, but I do a bit of everything in my day-to-day work.

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?

Great question. We have different workflows for in-house created versus externally produced, and vendor-digitized collections. The collections themselves also are processed to very different levels depending on factors like quality, difficulty, extent size and types, and staff. For now though, only several very special collections get MediaConch treatment and love, but this is all great preparatory work for a future division-wide time-based media digitization push AFC is in the very beginning stages of.

Any vendor digitized still image collection goes through technical assessors to check against file headers and specifications, similarly, we also have bulk processes in place to QC and sample still images. These checks have been integrated into our repository system and are available upon copying, verifying checksums, and running malware scans on content on ingest servers. Audiovisual and audio content (the bulk of our collections), however, generally runs through checks and processes that are external to our repository environment. This means a mix of tools and software like exiftool, mediainfo, bwf metaedit, sleuthkit, ftk, exact audio copy, exactly, and ? it can go on. The time-based media in our collections are a great challenge. Sometimes these tools come into play after the SIP is ingested to prepare the AIP, sometimes they are used before. Regardless, tools help identify, confirm, and even migrate content to give them a better chance at longterm preservation. Digital preservation as simply copying files to geographically dispersed and backed-up linear tape is no longer sufficient; our jobs are a lot harder now. While we have a few command-line scripts cobbled together and repository tools that work en bulk, I would say that we also rely on a lot of manual processes as well. So? it's a

bit of a smorgasbord that is collection-dependent.

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?

So far, I've primarily used MediaConch to create reports for new and incoming born-digital video from the Civil Rights History Project (Apple Pro Res 4:2:2, 10TB unique) and the DPX files (1536, 10bit, printing density, 20 TB unique) from digitizing celluloid film. Both of these collections share a few factors in common: they're really important to the department, they're extremely large in size, and as is, they present some technical difficulties.

Years ago, some of the first accessions of the CRHP collections were corrupted. In a post-ingest analysis, technical metadata fields created betrayed some indications such as truncated audio streams. While all the content was recovered, I decided to adapt workflows this with the new accession and advocated for creating some extremely granular reports as part of the archival package. The challenge now is to sit down and review the reports effectively.

With DPX, we didn't get checksums for some of the batches. With that baseline measure gone, I knew I needed to find something else to ensure that each of the easily 7,000 - 50,000 files per directory were at least mirroring each other's specifications. I reached out to the hive mind and MediaConch was highly recommended (thanks, Katharine!).

Initially, after creating the XML reports for each of the collections I was using, our staff used the GUI, but MediaConch would conk out when we tried to point to DPX directories; even the modest sub 10,000 files were simply too many. After several rounds of freezing my Mac and then uninstalling and reinstalling MediaConch, I realized I should just integrate a script. **It was much easier than I thought it would be to set-up and use right away.** Also, it's great to use MediaConch in all three ways in which the developers have made it available. I like the browser-based version for storing local policies and comparing local policies against the public policies other users have generously shared and made available. It's really useful for thinking about real-world time-based video specification, too. I was silly when I crashed my computers and had not downloaded my created policies for future re-use (fail!), so this is a great and easily accessible policy back-up. The GUI is also just incredibly easy and simple, too. I have trained staff to use it in minutes, which is not normal for implementing new software. Obviously though, with the previously unencountered numbers of files per film created when digitizing celluloid to DPX, I had to use the command line. At this point, I'm going to start reviewing the reports created and, again, I think that's when I need to really think about making good use of all the data created. While some of this is a ?store it and forget it? thing, I want it to be used much more actively as well. I'd be really curious to know how other people use and (re)package reports?

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

At the end, at least right now. That should change soon, but as a new staff member at AFC, I'm still catching up on various backlogs? although as I say that I think there will always be some sort of backlog to catch-up to! The collections are all actually already copied and checksummed on longterm servers long before I've used MediaConch with them? at least so far. My number one priority with any newly acquired hard drives is to get them backed up to tape and into our repository systems. We've also had a lot of division staff turnover with the first 2 digital technicians and 1 (hybrid) archivist leaving (all promotions), and the current digital technician I'm working with also leaving very shortly. So, excuses aside, I'm probably using MediaConch against my preconception of how I would have implemented it in workflows. But this is all in keeping with my focus this past year to start reevaluating older already ingested digital collections. AFC has been collecting and engaging in digital preservation for a long time, but MediaConch and tools like it had not existed before.

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

I use it for video and non-video (DPX), but once I'm through with the 2 collections mentioned earlier, I expect to expand its application. September is also my annual policy review month for me here, so I'm hoping that through update specifications for future vendor work and donors. I have started to create piles of migrated LPCM, so? I'm hopeful that I'll be playing with this more

and more.

Why do you think file validation is important?

File validation, specification checking, and checksumming verifying are probably the bedrocks of digital preservation, regardless of format. They all answer the questions of: what is this on the server? Is it what I think it is? Is it what I wanted? This is incredibly important, but it can be difficult to justify because of the time it can take in highly distributed workflows. Problems with collection quality and ingests often only become apparent with access. Given the wild world of audiovisual file specifications, MediaConch's work with FFV1 and Matroska is really amazing and forward thinking? and I'm excited for when I get to work with these them in the future.

Of course, file validation itself is still not enough for many file types. Many file types are often invalid, but knowing that a collection include invalid files is important for assessing preservation risks and understanding collection content. It can also help with creating clearer archival policies for supported versus less supported specifications - that gray area where we many donor-created digital collections fall into.

Anything else you'd like to add?

MediaConch has made my life better. I'm grateful to the stars behind the development of MediaConch! Thank you also to the European Commission for funding a tool that is critical to archival work today.