From Chaos to Confirmed Chaos

Jump In (3) Report: Colorado State Archives
Participants: Aly Jabrocki (Audio Archivist) and Kathy McCardwell (Archivist)

Participants Aly Jabrocki and Kathy McCardwell are the two newest staff members at the Colorado State Archives. Aly and Kathy had been with the State Archives for about six weeks when the Jump In 3 cycle started, and they decided, with the support of upper management, to take advantage of this opportunity to start being more proactive with reference to digital media created by and deposited with the Colorado State Archives.

Aly and Kathy work with very different collections at the Archives, so have created their inventories separately, collecting info that is most relevant to those collections. Aly works primarily with the Legislative Audio project, which involves the creation of large amounts of digital objects as historic media (tapes) are digitized, with the resulting files made available to the public in lieu of risking damage to the historic tapes themselves through repeated playback. Kathy has responsibilities to the records management program that operates out of the State Archives, so she works with lots of depositing agencies who deposit security copies of permanent records for physical storage at the State Archives. At some point in the past, many depositing agencies decided to start depositing CDs or DVDs as “permanent” security copies, instead of microfilm, which was the previous standard. Both Aly’s and Kathy’s findings will be addressed in this essay.

Prior to the beginning of this project, there had been limited standardization and planning for digital resources, whether created in-house or deposited by outside agencies. As in so many organizations, the transition from analog to digital was largely organic, rather than planned in advance. By the time this project started, lots of content had already been created by the audio project and deposited with the records management program, but there was limited standardization. Digital content created by the audio team did have a standardized set of file formats, but where and how files were stored was not standardized, nor was there a policy in place regarding if and when to combine audio from various tapes (e.g., when all associated with the same legislation). The records management had not promulgated standards required for deposit of digital materials, so deposits came in a number of formats and on a variety of media. There was also very limited collections management in place during much of this time, which further complicated the issue.

For Aly and the audio project, it was obvious where to start: she chose to survey digital audio files, including files on CDs, external hard drives, and the server. The goal in beginning here was to establish a baseline of what has been done, how it is being stewarded, and what remains to be done and/or improved in the workflow process for the audio conversation and stewardship. It also enabled her to consolidate files that might have been stored elsewhere on a designated server, so that all files can be accessed from one location.

For Kathy and the records management team, there were many collections that could have been test cases for this project, but she chose a large collection that she believed included some of the oldest digital media on deposit with the State Archives. (Since she doesn’t want to embarrass anyone, she does not include the name of the agency or any other identifying information, either in this report or in the spreadsheet shared as

part of the project, although of course she collected that information for internal use.) She believed that this would be a good test case because it would enable the Archives to begin establishing an inventory procedure that will enable us to properly inform depositing agencies on their current holdings.

Together, these still represent only a small fraction of the digital collections, but the project established a procedure that we will be able to follow with remaining collections.

Aly's process began with the CDs and then moved to the hard drives and files already on the server. Aly focused on ensuring that all files related to a given bill were included (and if not, noting what needed to be added before a bill could be considered "complete"), ensuring that access and preservation copies existed in the appropriate formats, and assessing file sizes.

Kathy's process involved creating an item-level inventory of the CDs and other media types she encountered, focusing on capturing variables that might be useful for assessing preservation to date (e.g., date, media label, whether or not the media was still playable and readable). Unfortunately, because so few of the media were labeled, her process became more time-intensive than anticipated, which resulted in her getting through fewer items than anticipated. Initially, for each item for which we had a player, she inserted the object to assess playback, file types, and date-modified, which she used as a proxy for a date of creation for the media. Eventually, she shifted to inserting only a handful of media from each box, in an effort to speed up her process.

Aly and Kathy both met with substantial challenges during the course of the project. Time turned out to be the biggest challenge: though we anticipated that this would be a time-intensive project, we did not expect that staff turnover and pipe leakage in our storage facility would eat up so much of our time during the four months of this project. Additionally, we found that word-of-mouth information, extant fragmentary listings, and estimates of holdings were not terribly accurate. Because collections had been selected and workflows established with reference to these estimates, this faulty information impacted the timely completion of the project. Aly found less data, in terms of absolute data, than anticipated, but it was more widely dispersed than expected, so pulling it together took longer than originally thought. Similarly, because there was no adequate box list for deposited materials, Kathy found that some boxes were absent (presumably withdrawn by the depositing agencies), and that many boxes had been deposited without adequate notation in box lists. The additional twenty boxes (!!!) of CDs found in the final stages of the project essentially guaranteed that she would not be able to finish the inventory, so her piece of the inventory is ongoing within the same collection.

By the end of the project, Aly had found that office hearsay regarding the percentage of bills completely digitized substantially overestimated the percentage of the collection that was truly "done." While portions of 770 bills had been digitized, only 310 of these were complete and ready for public access via ContentDM. She was also able to determine that not all bills had access copies on CDs, but, even more alarmingly, found that many bills only existed on CD and not on the server. Checks will later be done to see if the data on the CDs is still playable, in the hopes that we do not have to digitize the fragile original analog tapes again. She also found that initial estimates of data from audio conversion to date were significantly

exaggerated (3TB projected, 756 GB in actuality). This is an important finding for us, as it will enable us to plan more strategically when to augment server space.

Kathy's findings were not particularly surprising: as anticipated, a number of the old media were unreadable, primarily due to hardware obsolescence and proprietary (and dated) file formats. To date, the records management inventory includes 753.55 GB of information. The "take home" message from her inventory was the need for more outreach and education for depositing agencies are needed. We knew this going in; however, the hard data with regard to the sort of files being submitted on electronic media, and the frequency with which they were unreadable, provide strong arguments that we can share with depositing agencies. While, again, this is not new information per se, being able to reference this inventory in discussion will provide us with compelling evidence that we are already seeing what has been predicted in the literature, in terms of the degradation and loss of data from the transition to the digital age.

While in many ways our findings were worse than expected, in that there is so much data at risk, some things were actually better than expected. We did not identify any catastrophic data loss. While some media were unreadable, this was identified less frequently than we had thought possible, given the lack of initial planning. We did, of course, expect to find some unreadable files, but we were surprised to find so many file formats we had never even heard of, and that we couldn't identify even after significant searching.

Our next steps are hardly surprising: the audio staff is going to establish more systematic workflows to ensure that files are appropriately backed up and filed in findable places. Audio will also restructure chains-of-command such that one person is responsible for ensuring standardization and preventing gaps, rather than letting the collection continue to grow organically, and for ensuring that all files are regularly backed up. We believe that this process improvement will be applicable to setting up the workflow by which we manage digital surrogates of our static collections, once we begin digitizing those holdings in earnest.

The records management team is going to continue working with depositing agencies to educate them about options for digital preservation--but also to make abundantly clear that putting your (proprietary) files on a CD and putting that in storage for the next 20 years is not adequate for preservation of permanent-retention materials. Records management staff had already been in the process of drafting new policies explicitly forbidding the deposit of electronic/digital media for permanent-retention records storage; the inventory will provide documentary evidence of why "the way we've done it in the past" is not a good strategy for the future.

Finally, both Kathy and Aly anticipate continuing their inventories to reflect, respectively, additional collections and newfound materials, and newly-created files. We anticipate using the data quantity estimates to help us strategically plan server space and associated budgets, and to advocate for the Archives and its digital records responsibilities to the broader government and community.