"AND THE SURVEY SAYS...OK, BUT THE DATA SAYS!"

Cassandra Gallegos, George Blood Audio and Video

Introduction

Data. The mere mention of the term sends waves of quantitative hope down the spines of librarians. This is evident in the conference programming within the library and archives field. Every year at many library conferences there are multiple presentations on preservation surveys or data collection. The information is presented in its raw natural state, exactly as it came from the survey with little exploration into what the numbers may mean for not only their collection, but for the field as a whole. The data is often silenced before it has a chance to tell us anything of real consequence. Even worse, a librarian runs the risk of becoming incapacitated by the sheer quantity of data they receive. Upon contemplating this data dilemma, I wondered if there is another way to listen to our data and hear what it has to say, because if we do not our survey data can become nothing more than noise.

The preservation field's love affair with data causes unsure librarians to rely heavily on data collected from surveys. They are left to hope that this data will guide them towards making a definitive decision on any number of preservation concerns. What should I digitize first? How much of my collection has problems with mold? Is this object going to deteriorate soon? Unfortunately, data alone cannot tell you what to do and it is unlikely to tell you what you do not already instinctively know about your collection materials. This survey zeal hits fever pitch when it comes to audio and moving image materials. The idea that more data equals a better-protected collection is prevalent even amongst those of us who should know better! But surveys and data are not the enemy. In fact they can be quite helpful if we know what questions we want to ask of our data. The goal of my research is to find new ways of analyzing traditional American survey data along with an analysis of the data collected where I work.

Institutional collaboration

I am the Preservation Administrator for George Blood Audio and Video. We are an audio and moving image digitization vendor located in Philadelphia, PA. Our data collection was initially motivated by a need to provide such metadata to clients following a digitization project; however, I felt that there must be more we could do with this data. In preparing my analysis I worked with additional data sets from Columbia University (New York, NY), Harvard University (Cambridge, MA) and Indiana University (Bloomington, IN), each of which has created a unique survey tool.

Columbia University's data was collected using AVDb, a popular survey instrument created in 2008 by Columbia University with a generous grant from The Andrew W. Mellon Foundation. It is designed to aid in setting preservation priorities for unique and rare audio and moving-image materials. Surveyors are not expected to be experts in audio-visual media, but they are required to have familiarity with broad categories such as format identification. The accompanying users manual provides sufficient support to fill in any knowledge gaps. Applicable either to an item-by-item inventory or a random-sample survey, AVDb does not require collection of a great deal of detail about each item, yet has the ability to be a very granular tool. Data is collected purely through visual inspection. No playback of any medium is required.

http://library.columbia.edu/services/preservation

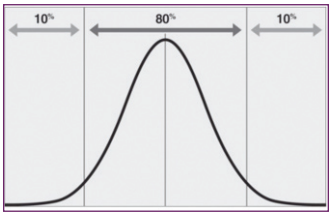
Harvard University created a similar tool called S.A.V.E. (Support for Audio-Visual Evaluation). This web-based tool enables the collection and aggregation of item-level condition assessment data for audio-visual materials that unites physical characteristics and condition data with curatorial information related to an item's research value. The survey taxonomy was designed to support the multi-use and item-level physical condition assessment survey of all audio-visual materials across the Harvard University campus.

http://preserve.harvard.edu/index.html

Indiana University Bloomington Campus created FACET (Field Audio Collection Evaluation Tool) to survey the various audio-visual collections on campus. FACET was created by Michael Casey in the Archives of Traditional Music. It is a point-based, open-source software tool that ranks audio field collections based on preservation condition, including the level of deterioration they exhibit and the degree of risk they carry. It was designed to assess the characteristics, preservation problems, and modes of deterioration of the audio formats historically associated with field recordings rather than those formats developed for commercial recordings. Similar to AVDb, FACET has an accompanying user manual to aid in the identification and assessment process.

http://research.iub.edu/communications/media_preservation/

Columbia, Harvard and Indiana Universities' survey tools have collected rather large quantities of data. The Law of Large Numbers tells us that large amounts of collected data have a common distribution pattern and that most observations hover around the average; the odds of a deviation declining faster and faster as you move away from the average, resulting in a standard Gaussian distribution. In other words, the more data you have the more likely it will assume a bell curve.



Data analyzing procedures

As a vendor, we are in a position to harvest the metadata we collect from a wider range of institutional situations rather than the data from any single institution such as is the case with university data. Additionally the difference between the data collected with AVDb for example and the data collected at our facility is that our data is post-digitization, so we can know much of these data fields with certainty. Survey data often includes a lot of guesswork. For example, an archivist must rely on notes from reel boxes and cassette shells for length, speed, track formation and other various important data fields. Often these notes are wrong or the box is mislabeled which in turn misleads the archivist and skews the survey data.

At the time of my initial data collection, metadata on 33,000 objects was available in our database. These factors, large number of institutions, wide range of collections, large number of data points about *actual* samples, allow me to consider our data set as representative of the general population.

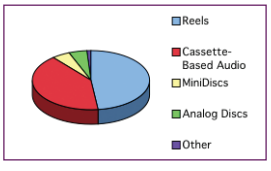
I have used this general population, the George Blood Audio and Video production data, to spot trends, then make hypotheses based on these trends. I tested these hypotheses by examining the similar data points from the Columbia, Harvard and Indiana surveys. In what ways do the surveys gather information that is similar to or different from the production data that we collect? Furthermore, I wondered if it was possible to extrapolate for unknown or uncertain data in surveys based on known information from our production data to help librarians and archivists make more informed guesses about their collection materials.

In analyzing the data I needed to control for a few factors. Some institutions collect their data on a different basis than others. For instance, does the "intellectual unit" or what I would consider a single record in a database, correspond to a single physical object or a single interview. Another challenge is that new records are constantly being added to our production database. For this reason I used 1%, or about 300 objects (about one week of our production capacity), as the margin of error in my analysis.

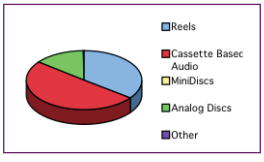
Additionally, each survey collected information in slightly different ways and I had various levels of access to each data set. For instance, Columbia's survey groups audiocassettes, microcassettes and DATs in one category. Therefore I had to do the same grouping with the other three surveys to make the data directly comparable. Columbia and Harvard collected data at the object level while Indiana collected information on collection level by format. Both Columbia and Indiana provided me with the actual database and data used during the survey. Harvard's system could only be accessed in-house, so I submitted queries directly to Harvard via email to obtain my results.

Results

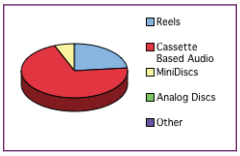
The first data I collected concerned format and length of audio. At George Blood Audio and Video we assessed 32,699 audio objects at the time of my data collection. Out of this total, 15,390 are reels and 13,399 are cassette-based audio formats. This accounts for nearly 90% of the audio we handle. After these two formats, numbers take a drastic downturn. 5% are MiniDiscs and 5% are analog discs. Cylinders, Wires, and CDs account for less than 1% each.



Similarly, Columbia's surveyed collections are 85% magnetic tape, either reels or cassette-based, with approximately 35% being reels and 50% cassettes out of 35,766 total audio objects.

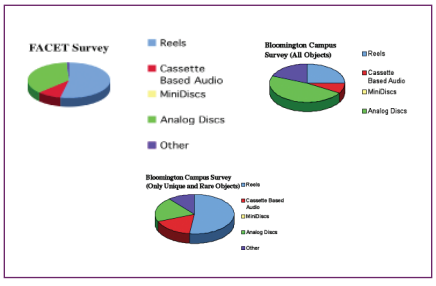


Harvard audio survey of 3,839 objects is 69% cassette and 22% open reel.



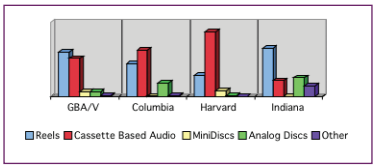
Indiana University's survey has a total of 364,867 audio objects across the entire Bloomington campus. FACET was used to survey objects within one collection; however, another campus-wide survey looks at the number of formats across all collections. I parsed both sets of data and came to the following conclusions: The FACET data shows that this one surveyed collection is approximately 53% reels, 36% analog discs and nearly 10% cassette-based formats. Up to this point in my research, it had become clear that the average collection consisted of mostly reels and cassette based media. Indiana consists of mostly reels and analog discs. I wanted to delve further into the Indiana data because I noticed that Indiana made a distinction between objects that were unique and rare and those that were commercially available. Unique and rare objects are more of a preservation concern than those that are commercially available. Unfortunately this information was not quantifiably available within FACET. There were various content notes and a point system dedicated towards the uniqueness of each subset of the collection. This data was not useful without intimate knowledge of the initial data collection process. I turned toward the inventory report for the entire Bloomington campus to shed some light on whether Indiana's collections consisted of mostly reels and analog discs that were unique/rare or if the inclusion of commercially available objects were skewing the data. This report extrapolated the type of data I was unable to process within FACET. The initial results within this report indicate that overall the collections on the Indian Bloomington campus consist of a large percentage of discs (47% analog discs) much like the FACET survey.

Upon further inspection of the objects indicated as either unique or rare, only 15% of the total collection is unique or rare discs that are not available commercially, 21% of the total audio objects are unique or rare reels that are not available commercially and 5% of the total audio objects are unique or rare cassettes. These results align more with my hypothesis that the average is mostly reels and cassettes.



It is noticeable that 21% of the total is devoted to various other audio formats. Taken as a whole this 21% appears to be a substantial part of the collection but within this 21%, 82% are optical discs only of which 14% are unique or rare and not commercially available. When you adjust the data to only include the unique or rare non-commercially available audio objects, you see that the total amount is really 6% unique and rare objects of various other formats.

With all of the data together it is clear that most of each collection is either reels or cassettes.



As far as reel preservation is concerned, I know anecdotally, from handling material, that many more reels suffer from poor tape pack rather than biological contamination or chemical degradation. I tested my experience against the data collected at Columbia. They assessed that 90.59% had a poor library wind. This information is critical for an archive thinking about creating an in-house audio digitization department: it may make sense to have a large-scale library wind project.

After spotting these specific trends in collection content, I considered what further questions could I ask of the data. Librarians and archivists in general are concerned with issues of degradation such as sticky shed syndrome and contaminants such as mold. I often receive questions on baking reels from archivists — how to do it, when to do it, and if it should be done at all. I would not be surprised if there are portions of collections out there that were accidentally destroyed by a well meaning archivist.

I wanted to know, from our production data, exactly how often stickiness occurs in reels. Surprisingly I cannot tell you! At our studio it has been our experience that most reels that arrive exhibiting signs of sticky shed syndrome will "get better" after spending 6-8 weeks in our temperature and humidity controlled vault.[10] The time a tape needs in the proper storage environment to diminish sticky shed varies based upon the width of the tape. We have experienced that a reel must remain in our vault for 6-8 weeks per each ¼ inch in width. This data is merely anecdotal and will require further study. The reels that are still sticky after their stay in our vault are reels that require baking. While I cannot tell you how often sticky shed syndrome occurs within the general population, I can tell you, within the general population, how often reels will still require baking even after stored in a proper, low humidity environment.

Around 6% of all the reels we received required baking. And this is highly concentrated within collections. If kept in the correct conditions for longer, who is to say this may not "cure" more of the sticky reels. Specifically it seems that tightly controlled relative humidity will help these very sticky reels. The data also tells us that sticky shed is not a large problem within the general population but rather a concentrated problem within collections. If you find one sticky reel in a collection, there is a higher likelihood that you will find more sticky reels within the collection. This is something that a random survey may miss. Furthermore, proper storage conditions should be the primary goal for several reasons. First, there is likely only a small chance of encountering a bad case of sticky shed syndrome. Second, since this chance is indeed so small it is more cost effective and safer to send these few reels out to a professional rather than risk these materials by baking the reels yourself.

As for mold, less than 1% of reels in our production data had mold. Mold occurs only rarely on other media. Jim Riley at the Image Permanence Institute speculates this is due to the nature of the exposure of the media. Magnetic media does not contain anything that mold will live on. Mold is growing on the biofilm that develops on objects when exposed to the environment and human handling. Since reels are more exposed, more biofilm develops and therefore supports the growth of mold. So while mold and sticky shed syndrome should be taken seriously, they do not occur very often in the general population. Worry and effort should be placed elsewhere. Be aware of these issues but only so much as to know to seek information and help when these problems arise.

10 The set points are 65°F and 35% R.H.

I asked further questions of my data concerning reels. I wanted to know how many of each stock brand required baking. Is there a correlation here? First off, the two tape stocks we encounter most often are Scotch and Ampex, with Scotch coming out as the clear leader; however, we only baked 2% of the Scotch while we baked 19% of the Ampex.

Amongst compact audiocassettes I found that both Columbia's and the George Blood Audio's cassettes were approximately 92% Type 1, which is not overwhelmingly surprising as they were the most affordable type of cassette. What is surprising is that 50% of the cassettes we received needed their record tabs removed. Removing record tabs is the cheapest and easiest precaution you can take towards safeguarding a cassette collection, yet half of all cassettes were left without write protection. In some cases, whole collections were in danger while other collections only had one or two cassettes with record tabs not removed. Additionally we know that replacing the shell solves most cassette playback problems. We replaced cassette shells on about 3% of all cassettes.

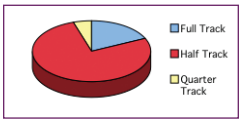
Although we do not see much mold on cassettes, I did notice that, while assessing a large batch of thousands of cassettes from a collection from Puerto Rico, that there seemed to be a correlation between mold and the opacity of the cassette shell. I pulled all the data from all of the cassette assessments done at George Blood Audio and Video. Over 90% of the moldy cassettes were from this Puerto Rican project, so I decided to concentrate on these 1,385 cassettes. Puerto Rico is an ideal location for mold growth with high temperature and high humidity. Out of all of these cassettes, 804 are clear and 581 are opaque; 92% of the moldy cassettes had opaque shells. I also observed that the mold on the tapes housed in the opaque shells was not present in the small clear window that exposed the tape to sunlight. Despite this data you must remember that correlation does not necessarily indicate causation. Further scientific study is needed to prove that cassettes with opaque shells are more likely to be moldy.

Surveys can only collect known information. Object duration is one important datum that is often not known at the time of the survey. Many surveys allow you to collect "maximum duration" based on the possible duration of the audio object. For instance, you assume a C-90 cassette will be full of 90 minutes of audio. Our production data can answer the questions of average and distribution of durations of audio objects. Here are the average durations of various audio formats:

Audiocassette – 53 minutes
Reel – 50 minutes
MiniDisc – 60 minutes
Analog Disc – 6 minutes
Microcassette – 30 minutes
DAT – 69 minutes
Cylinder – 2.4 minutes
Wire – 43 minutes

Track configuration of reels is also not often known until time of transfer, unless the archivist owns a magnetic viewer and has an inordinate amount of time to discover the track configurations of all of the reels within the collection. Even with a magnetic viewer, it is generally not possible to discern two track mono from two track stereo simply by looking. I found that a reel will be full track around 18% of the time and quarter track about 5% of the time. The remaining 77 % are half track.

The distribution of speeds among reels varied greatly:



7 ½ ips (inches per second): 45%
15/16 ips: 2%
1 7/8 ips: 15%
3 3/4 ips: 20%
15 ips: 16%
30 ips: 2%
Cassettes did not vary as much in speed, 99% were 1 7/8 ips.

Conclusion

Surveys tools can be quite helpful if you know how to use them and do not rely too heavily on the data to form conclusions. A more accurate snapshot of a typical collection can be gained by looking at production data that comes from a digitization project. What I have learned is that there is a good chance that the typical American audio collection will consist of cassette-based and reel-to-reel tapes. The reels are not in great danger from sticky shed as long as temperature, and more importantly relative humidity are controlled. If sticky shed is found during a random sample survey then it is likely many other reels within the collection will also be sticky. Mold is a relatively small problem for reels and even smaller for cassettes but there may be a correlation between cassette shell opacity and mold. Additional data points and sample collections will shed more light on these and other issues in audio preservation.

Acknowledgements

I would like to thank individuals and their institutions for sharing their survey data and help preparing this report:

Janet Gertz, Director for Preservation, Columbia University Libraries; Elizabeth Walters, Program Officer for Audiovisual Materials, Weissman Preservation Center, Harvard University; Mike Casey, Associate Director for Recording Services, Archive of Traditional Music, Indiana University; George Blood; Annie Peterson; Melissa Hernandez; Michael Mullin.