Quantifying the Need: A Survey of Existing Sound Recordings in Collections in the United States

Bertram Lyons, Senior Consultant
Rebecca Chandler, Consultant
Chris Lacinak, President

avpreserve

This report was written in collaboration with NEDCC and with support from The Andrew W. Mellon Foundation.

AVPreserve | Media Archiving & Data Management Consultants
253 36th Street, Suite C302 | Brooklyn, New York | 11232
917.475.9630 | www.avpreserve.com | info@avpreserve.com

Table of Contents

Executive Summary

In 2014, AVPreserve and the Northeast Document Conservation Center (NEDCC), with funding from The Andrew W. Mellon Foundation, undertook an in-depth, multi-faceted assessment to quantify the existing audio items held in institutional collections throughout the United States. This was performed in response to The Library of Congress National Recording Preservation Plan[1] and its call for the appraisal of collections, as well as to establish a foundation for articulating the current preservation need of sound recordings in collections nationwide.[2] Our goal was to acquire enough trustworthy data to be able to answer questions such as "How many sound recordings exist in *broadcast organizations* across the US?" or "How many sound recordings exist in archives throughout the US?" Moreover, we wanted to answer more complex questions such as "How many of such items are *preservation-worthy*?" or "How many have *already been digitized*?"

Our assessment consisted of three phases of data collection and analysis:

- Phase 1: Identify pre-existing cross-organizational surveys that report on audio holdings in the US.
- Phase 2: Conduct a targeted small-scale survey of a variety of organizations from across the US in order to establish a point of comparison for current and pre-existing data points, and to acquire information that would enable analysis by organization type (e.g., Special Libraries, Broadcast) and broad format type (e.g., grooved, magnetic, optical).
- Phase 3: Conduct a large-scale national survey of a variety of organizations in order to establish additional data points for comparison against data collected in Phases 1 and 2. This phase was also used to gather additional information on the preservation-worthiness of holdings and the percentage of items that have already been digitized.

In order to establish confidence in our numbers, we ran two statistical tests, using the Phase 1 data as a control. Based on the results of these two statistical tests, we feel confident that our data is as trustworthy as existing survey data (our control surveys) collected and published over that past 15 years regarding quantities of audio holdings in organizations throughout the US. With typical caveats of performing any type of extrapolation, we can use these numbers to project an estimated total quantity of audio holdings to be found in collections nationwide.

Our conclusions are as follows:

- There are over 537 million sound recordings in collection-holding organizations across the US.
- Academic libraries and archives/museums hold the highest quantities of sound recordings in the US.
- Grooved media and magnetic media are the most widely-held recorded sound media in collection-holding organizations in the US.
- Fifty-seven percent (57%) of audio holdings in US collections are either unique or rare.
- Seventeen percent (17%) of audio holdings in US collections have already been digitized to date.

1 The Library of Congress National Recording Preservation Plan / sponsored by the National Recording Preservation Board of the Library of Congress. Co-published by the Council on Library and Information Resources and the Library of Congress. 2012. Accessible at: http://www.loc.gov/programs/static/national-recording-preservation-plan/publications-and-reports/documents/NRPPLANCLIRpdfpub156.pdf.

2 Due to the resources available the scope of this study did not include private collectors. The value of including this group is seen as an important addition to future work and any existing data would be welcomed by the authors.

- Of the total existing sound recordings in US collection-holding organizations, over 250 million items are preservation-worthy and have not yet been digitized; of these, over 80 million (32%) will require a specialized audio preservation workflow.
- The estimated cost of digitizing all preservation-worthy items in audio collections in the US that have not yet been digitized is over $20 billion.

The National Recording Preservation Plan, published in 2012 states:

> "many analog audio recordings must be digitized within the next 15 to 20 years—before sound carrier degradation and the challenges of acquiring and maintaining playback equipment make the success of these efforts too expensive or unattainable."[3]

Mike Casey offers a detailed look into the threats posed by obsolescence and degradation in his paper *Why Media Preservation Can't Wait: the Gathering Storm*, published in the IASA journal in 2015.[4]

Looking at the scale of recordings in the US and the narrowing window of time within which to act before recordings are permanently lost brings the overarching challenge into sharp focus. The response required to address the totality of the challenge is so massive and complex that it appears effectively impossible. To avoid paralysis, we will be required to shift our thinking from the question of how we save everything to one that asks, what is it that we will save? Prioritization for digitization is as critical as both funding and timeliness. The foundation for action on all three of these fronts is trustworthy quantitative data. This paper aims to provide such data along with supporting information on the methodologies used in its generation.

Introduction

In 2010, the Council on Library and Information Resources (CLIR) and the Library of Congress released a publication, "The State of Recorded Sound Preservation in the United States: A National Legacy at Risk in the Digital Age," that noted the following three critical factors (among others):

- Public institutions, libraries, and archives hold an estimated 46 million recordings, but few institutions know the full extent of their holdings or their physical condition.
- Funding and advocacy for recorded sound preservation is decentralized and inadequate. Recorded sound preservation has been declared a national objective; however, without greater support as a matter of public policy, this objective will not be realized.
- Resources must be invested not only in rescuing specific collections but also in developing techniques and methodologies that will enable more institutions to afford to assume a share of the work.[5]

In 2012, the Library of Congress's National Recording Preservation Plan (NRPP) echoed these factors: "The nation's libraries, archives, and museums hold some 46 million sound recordings, millions of which are in need of preservation...Transitioning to digital audio preservation, however,

3 The Library of Congress National Recording Preservation Plan, 2012.
4 Casey, Mike. 2015. Why Media Preservation Can't Wait: the Gathering Storm. International Association of Sound & Audiovisual Archives Journal 44. Available at http://www.avpreserve.com/wp-content/uploads/2015/04/ casey_iasa_journal_44_part3.pdf.
5 CLIR, Library of Congress. "The State of Recorded Sound Preservation in the United States: A National Legacy at Risk in the Digital Age." CLIR. 2010. Pg 3-4. Accessible: http://www.clir.org/pubs/reports/pub148/pub148.pdf.

has created significant technical, organizational, and funding challenges for those institutions responsible for preserving recorded sound history for future generations."[6]

This often-quoted 46 million items statistic comes from a study performed by Heritage Preservation in 2004 (published in 2005).[7] Although the HHI study is thorough and well-documented, the landscape of sound recording preservation has changed tremendously in ten years. Today, organizations across the US are more in tune with non-text-based collection materials because of the growing demand for access to these holdings by researchers and the general public. In order to overcome the funding and infrastructure challenges that will face US collection-holding organizations in the next twenty years as they work towards preserving the nation's audio heritage, it will be essential to have an accurate picture of what the need is today. Additionally, a total number of holdings is difficult to use in and of itself. More granular information enables greater insights and the ability to think more critically about the task at hand. The quantity and makeup of nationwide holdings is a foundational piece of information, serving as a cornerstone for many publications, initiatives, and efforts. Incorrect numbers will yield a response that is out of alignment and falls short of the need. Therefore it is critical to revisit and analyze the accuracy of this data.

We know that today and for the next twenty years, two challenges affecting the preservation of sound and audiovisual collections globally will be format obsolescence and degradation. Yet, as the NRPP and its predecessor, "The State of Recorded Sound Preservation in the United States," articulate, funding and advocacy for recorded sound preservation are currently inadequate. At this point the field has determined, with fairly unanimous agreement, the best methods and strategies to overcome obsolescence. Caring for physical collections is understood. Digitization practices are mature and an entire industry now offers both boutique and high-throughput digitization services for preservation. Learning from banks and other information-heavy organizational entities of the world, archives are equipped with the necessary role models for building, staffing, and sustaining digital repositories worthy of carrying our sound and audiovisual heritage into the future. We are not burdened with the ignorance of how we should proceed or what we should do to save our collections. Instead, one of our greatest challenges is determining how we will afford to do what we know we need to do to save our vanishing recordings. Funding for digitization and the necessary digital infrastructure to support the results of digitization (and born-digital) is the most pressing obstacle.

The field is in need of data and research that help quantify the reality of the situation so that we can, with greater clarity and accuracy, demonstrate to funding sources the task at hand and mobilize resources. To that end, we must quantify the needs of audiovisual preservation in business terms.

In 2014, AVPreserve and the Northeast Document Conservation Center (NEDCC), with funding from The Andrew W. Mellon Foundation, undertook an in-depth, multi-faceted assessment to quantify the existing audio items held in institutional collections throughout the United States. This was performed in response to The Library of Congress National Recording Preservation Plan[8] and its call for the appraisal of collections, as well as to establish a foundation for

6 CLIR, Library of Congress. "National Recording Preservation Plan." CLIR. 2012. Pg. 1. Accessible: http://www.loc.gov/programs/static/national-recording-preservation-plan/publications-and-reports/documents/NRPPLANCLIRpdfpub156.pdf.

7 Heritage Preservation, IMLS. "A Public Trust at Risk: The Heritage Health Index Report on the State of America's Collections." Heritage Preservation. 2005. Accessible at http://www.heritagepreservation.org/HHI/.

8 The Library of Congress National Recording Preservation Plan / sponsored by the National Recording Preservation Board of the Library of Congress. Co-published by the Council on Library and Information Resources and the Library of Congress. 2012. Accessible at: http://www.loc.gov/programs/static/national-

articulating the current preservation need of sound recordings in collections nationwide.[9] Our goal was to acquire enough trustworthy data to be able to answer questions such as "How many sound recordings exist in *broadcast organizations* across the US?" or "How many sound recordings exist in archives throughout the US?" Moreover, we wanted to answer more complex questions such as "How many of such items are *preservation-worthy*?" or "How many have *already been digitized*?"

The National Recording Preservation Plan, published in 2012 states:

> "many analog audio recordings must be digitized within the next 15 to 20 years—before sound carrier degradation and the challenges of acquiring and maintaining playback equipment make the success of these efforts too expensive or unattainable."[10]

Mike Casey offers a detailed look into the threats posed by obsolescence and degradation in his paper *Why Media Preservation Can't Wait: the Gathering Storm*, published in the IASA journal in 2015.[11]

Looking at the scale of recordings in the US and the narrowing window of time within which to act before recordings are permanently lost brings the overarching challenge into sharp focus. The response required to address the totality of the challenge is so massive and complex that it appears effectively impossible. To avoid paralysis, we will be required to shift our thinking from the question of how we save everything to one that asks, what is it that we will save? Prioritization for digitization is as critical as both funding and timeliness. The foundation for action on all three of these fronts is trustworthy quantitative data. This paper aims to provide such data along with supporting information on the methodologies used in its generation.

Survey Design & Data Availability

The feasibility of contacting and acquiring an inventory from every collection-holding organization in the US is low and would require tremendous resources and time. Therefore, our approach used a method of extrapolation based on averages and organization types (e.g., Special Libraries, Archives/Museums). Using data published by the American Library Association (ALA)[12] and GuideStar[13], we were able to classify types of collection-holding organizations in the US and to quantify the number of each type in existence within the US (Table 1).[14]

recording-preservation-plan/publications-and-reports/documents/NRPPLANCLIRpdfpub156.pdf.

9 Due to the resources available the scope of this study did not include private collectors. The value of including this group is seen as an important addition to future work and any existing data would be welcomed by the authors.

10 The Library of Congress National Recording Preservation Plan, 2012.

11 Casey, Mike. 2015. Why Media Preservation Can't Wait: the Gathering Storm. International Association of Sound & Audiovisual Archives Journal 44. Available at http://www.avpreserve.com/wp-content/uploads/2015/04/ casey_iasa_journal_44_part3.pdf.

12 We relied on ALA for our counts of the Number of Academic Libraries in the US and the Number of Special Libraries in the US. Accessible at http://www.ala.org/tools/libfactsheets/alalibraryfactsheet01.

13 GuideStar bases its data on IRS Subsection codes, which classify business types based on predetermined IRS codes. For the numbers in our survey, we used A34 (Radio), A32 (Television), A80 (Historical Societies and Related Activities), and A50 (Museum and Museum Activities). Accessible at http://www.guidestar.org/.

14 Based on results from GuideStar, there are over 10,000 registered Historical Societies in the US. Many of these societies, however, do not collect and preserve primary documents. For the purposes of this survey, we classified Historical Societies as those at the state level and those found in large cities. We estimated this number to be closer to 200 collecting organizations.

Organization Type	Quantity in the US
Academic Libraries	3,689
Special Libraries	7,616
Broadcast Organizations	1,469
State and Large City Historical Societies	200
Archives/Museums	3,569

Table 1. Number of collecting organizations in the US by Organization Type.

We used the same classification types identified in Table 1 to organize our survey results with the intention of extrapolating our findings towards a total picture of the audio holdings by organization type throughout the US.

The assessment consisted of three phases of data collection and analysis:

- Phase 1: Identify pre-existing cross-organizational surveys that report on audio holdings in the US.
- Phase 2: Conduct a targeted small-scale survey of a variety of organizations from across the US in order to establish a point of comparison for current and pre-existing data points, and to acquire information that would enable analysis by organization type (e.g. Special Libraries, Broadcast) and broad format type (e.g. grooved, magnetic, optical).
- Phase 3: Conduct a large-scale national survey of a variety of organizations in order to establish additional data points for comparison against data collected in Phase 1 and 2. This phase was also used to gather additional information on the preservation-worthiness of holdings and the percentage of items that have already been digitized.

Phase 1 Discussion

Attempts to quantify the number of audio holdings in library, archive, and museum collections throughout the United States have been undertaken sporadically since at least 1998 when the Association of Research Libraries (ARL) published a report on the results of its survey of special collections in ARL libraries.[15] Although this ARL report did not focus entirely on sound recordings, it did provide a count of such items as part of its findings. Of the eight extant survey reports that we identified, only two looked specifically at audio collections.

During Phase 1 we evaluated these eight pre-existing cross-organizational surveys that included data on extant audio collections and we extracted a subset of data from each.[16] The data we were most interested in acquiring were:

15 http://www.arl.org/storage/documents/publications/special-collections-arl-libraries.pdf

16 1998, Association of Research Libraries, "Special Collections in ARL Libraries," accessible at http://www.arl.org/storage/documents/publications/special-collections-arl-libraries.pdf; 2003, Image Permanence Institute, "The Preservation of Magnetic Tape Collections: A Perspective," accessible at https://www.imagepermanenceinstitute.org/webfm_send/303; 2004, Council on Library and Information Resources, "Survey on the State of Audio Collections in Academic Libraries," accessible at http://www.clir.org/pubs/reports/pub128/reports/pub128/pub128.pdf; 2005, Heritage Preservation, "Heritage Health Index," accessible at http://www.heritagepreservation.org/HHI/; 2007, Association of Research Libraries, Statement to the National Recording Preservation Board, accessible at http://www.clir.org/pubs/reports/pub128/reports/pub128/pub128.pdf; 2007, California Preservation Program, "Preserving the 20th Century: California Preservation Survey of Moving Image and Recorded Sound Collections," accessible at http://calpreservation.org/wp-content/uploads/2013/08/CPP-AV-survey-report-14oct07.pdf; 2010, Online Computer Library Center, "Taking our Pulse: The OCLC Research Survey of Special Collections and Archives," accessible at http://www.oclc.org/content/dam/research/publications/library/2010/2010-11.pdf?urlm=162945; 2012, Nancy McKay, US Oral History Survey Report, unpublished.

- Number of Organizations Represented in the Survey
- Organization Types Represented
- Year the Survey was Conducted
- Total Audio Items Identified
- Total Audio Items Identified by Broad Format
- Total Audio Items Identified by Specific Format

Initially, we imagined we could use these numbers to project the total amount of audio holdings in the US by combining the averages of the eight surveys with the number of organization types across the US. We quickly realized two obstacles: 1) There were not enough data in the cross-organizational surveys to match organization types to holdings (i.e., we could not say how many holdings were in broadcast organizations and how many holdings were in archives, or how many were in libraries); and 2) There were not enough data in the cross-organizational surveys to assess the granularity of formats held across the US (i.e., we would be unable to project how many grooved media existed vs. magnetic media vs. optical media). The only projections we could make with these existing numbers were gross estimations of the total audio holdings across the US.

Instead, we decided that these pre-existing cross-organizational surveys could serve as our control group (see Table 2). We would use these numbers as baselines for comparing the veracity of our own surveys. They would serve as a means to measure confidence in the data we would collect in Phases 2 and 3 of this assessment.

Source	Year of Survey	Total Number of Organizations	Total Number of Audio Items	Average Audio Items per Organization
Association of Research Libraries	1998	72	151,920	2,110
Image Permanence Institute	2003	17	150,000	8,824
Council on Library and Information Resources	2004	69	1,899,150	27,524
Heritage Preservation	2005	30,827	46,000,000	1,492
California Preservation Program	2007	32	604,770	18,899
Association of Research Libraries	2007	123	10,000,000	81,301
Online Computer Library Center	2010	92	3,000,000	32,609
US Oral History Survey	2012	246	337,505	1,372

Table 2. Control group of pre-existing cross-organizational surveys.

Phase 2 Discussion

Although there are few published cross-organizational surveys, there are many single-organization reports available. To add a ninth data point to our control group of eight (documented in Phase 1), we gathered data from a variety of targeted sources to create a new survey of audio holdings in collecting organizations in the US. In this phase, we did not send out a survey to collect a random sampling. Instead, we gathered data that we knew already existed and we contacted colleagues from whom we had confidence we could get accurate and current inventories. Sources included data from the National Folklore Archives Initiative; George Boston's 2003 UNESCO survey[17]; testimonies to the National Recording Preservation Board; the American Archive project; inventory and assessment data from past projects conducted by AVPreserve; and direct correspondence with organizations such as Indiana University, the Minnesota Historical Society, and the University of Kansas, among many others.[18]

In contrast to our results in Phase 1, we were able to acquire information on organization types and audio format types held within those organizations.

Type of Organization	Total Number of Organizations	Total Number of Audio Items	Average Audio Items Per Organization
Academic Libraries	23	1,888,895	82,126
Non- Profit Broadcasting Organizations	91	757,454	8,324
State and Large City Historical Societies	8	16,754	2,094
Special Libraries	9	95,392	10,599
Archives/Museums	46	1,815,902	39,476
TOTAL	177	4,574,397	25,844

Table 3. Phase 2 targeted survey of sound recordings in US collections, organized by Organization Type.

Table 3 provides a high-level view of the results of our targeted survey. On average, the survey finds that Academic Libraries and Archives/Museums hold the highest quantities of sound recordings. Broadcasting Organizations and Special Libraries make up the middle, and Historical Societies hold the fewest sound recordings.

Because we had access to data from the American Archive project, the number of Broadcasting Organizations represented in our survey is skewed in relationship to the percentage of Broadcasting Organizations represented in the US. However, because we are interested in averages by organization type, this larger sample makes our estimations more statistically accurate regarding the average quantity of sound recordings held by Broadcasting Organizations.

We were unable to gather consistent information about specific format types, e.g., instantaneous disc, vinyl disc, cassette tape, cylinder, etc. However, we were able to acquire consistent information about broad formats, i.e., grooved, magnetic, optical. This is an improvement over the granularity found in the Phase 1 control surveys. Table 4 shows the total count of objects by format as found in the Phase 2 survey.

17 http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/programme_doc_survey_report.pdf
18 Raw data is available as Appendix A to this report.

Type of Organization	Total Number of Organizations	Total Number of Grooved Media	Total Number of Magnetic Media	Total Number of Optical Media
Academic Libraries	23	1,047,190	600,601	241,104
Non- Profit Broadcasting Organizations	91	11,450	541,075	204,929
State and Large City Historical Societies	8	2,039	13,829	886
Special Libraries	9	27,340	58,310	9,742
Archives/ Museums	46	927,037	727,475	161,391
TOTAL	177	2,015,056	1,941,290	618,052

Table 4. Phase 2 targeted survey of sound recordings in US collections by general format.

Although these numbers are only a sampling of the population of holdings that can be found across the US, the real value of this survey is that it identifies average counts per organization type. Table 5 shows average format type by organization type.

Type of Organization	Average Number of Grooved Media	Average Number of Magnetic Media	Average Number of Optical Media
Academic Libraries	45,530	26,113	10,483
Non- Profit Broadcasting Organizations	126	5,946	2,252
State and Large City Historical Societies	255	1,729	111
Special Libraries	3,038	6,479	1,082
Museum/Archives	20,153	15,815	3,508
TOTAL AVERAGE	11,384	10,968	3,492

Table 5. Phase 2 targeted survey of sound recordings in US collections: averages per organization type by general format.

An analysis of average holdings per organization shows that grooved media and magnetic media are the most widely held recorded sound media in collection-holding organizations in the US. They are found in almost equal proportions. Academic libraries and museums/archives seem to have the greatest numbers of holdings on average, especially in terms of grooved and magnetic media.

Comparison Against Control

Establishing Confidence in Phase 2 Survey

In order to establish confidence in our numbers, we ran two statistical tests against our Phase 2 numbers and the control numbers gathered in Phase 1.

Scatter Plot Analysis

First we employed a scatter plot in order to see the results of our survey in comparison to the results of the eight pre-existing surveys from Phase 1. A scatter plot uses two consistent variables for a given set of data points and then plots those variables on an x-y axis. The resulting graph shows visual relationships (if they exist) between the variables and the data points. We used nine data points in our scatter plot — eight pre-existing surveys from Phase 1 and the targeted survey from Phase 2. For each data point, we selected two variables (see Table 6): the total number of organizations represented in the survey and the total number of audio holdings identified in the survey.

Survey	Number of Organizations Represented in Sample	Number of Objects Counted Total
NEDCC Phase 2 targeted survey	177	4,574,397
OCLC	92	3,000,000
Image Permanence Institute	17	150,000
US Oral History Survey	246	337,505
California Preservation Program	32	604,770
Association of Research Libraries	72	151,920
Council on Library and Information Resources	69	1,899,150
Association of Research Libraries	123	10,000,000
Heritage Preservation (HHI)	30,827	46,000,000

Table 6. Scatter plot data points.

Because the Heritage Preservation survey (HHI) yielded numbers much higher than any of the other surveys, we removed the HHI from the sample so that the data set could be analyzed more closely, leaving us with a total of eight data points for the scatter plot (Table 7).

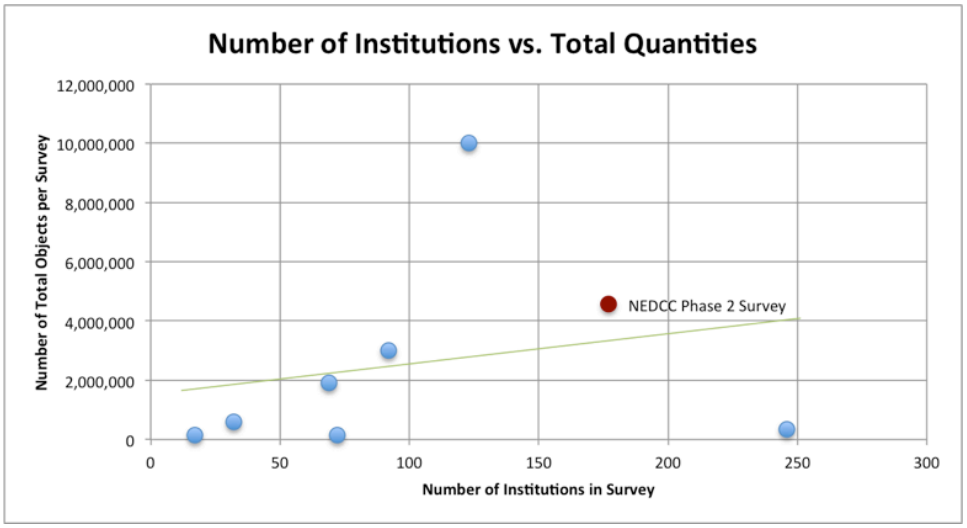

Table 7. Scatter plot of seven control groups and targeted survey (represented as the red data point).

An analysis of this scatter plot shows a directional trend in the relationship between number of total objects per survey and the total number of organizations in the survey. As one might expect, as the number of organizations rise so do the number of objects. In this case, however, there is a linear trend in the ratio of organizations to objects. The Phase 2 survey, at 4.6 million items and 177 organizations, lands just above the trendline of this scatter plot, giving us confidence that there is nothing statistically different about the ratio of the number of organizations to the number of objects among the pre-existing surveys and the Phase 2 targeted survey.

The three outliers in the graph, upon analysis, are explainable. The ARL data point at 10,000,000 items and 123 organizations is taken from a written testimony without supporting evidence. With no numbers provided to support the claim, it is difficult to ascertain the veracity of quantities reported. This is apparent in comparison with the other data points on the scatter plot. The second ARL data point at 151,920 items and 72 organizations represents data collected in 1998 (the earliest survey in the control group). In 1998, efforts to consider the risks facing audio collections were just beginning to emerge and it is likely that organizations either did not have accurate numbers or that they under-reported their audio holdings due to lack of information. The third outlier, the data point from Nancy McKay at 337,505 items and 246 organizations represents an effort to survey only oral history holdings in organizations across the US. It is likely that the respondents to McKay's survey only provided numbers for audio holdings representing oral histories, as opposed to the entirety of their audio holdings. This would explain why the ratio of items to organizations is visibly different between McKay's survey and the trend of the other surveys in the control group. Due to the low number of data points in the sample, we have decided to leave these outliers in our data set.

Based on this analysis, there is a trend among the surveys in the ratio of the number of items documented and the number of organizations in the survey. Our Phase 2 targeted survey lies within the trend. This gives us confidence that there is nothing significantly different about the ratio of the numbers of items to organizations found in the Phase 2 survey — which is important because the goal of this assessment is to extrapolate from our numbers to reveal an accurate picture of the amount of audio holdings in collections nationwide.

T-Test Analysis

To test the targeted survey results against the control group from a different angle, we employed a statistical hypothesis test called a t-test which tests two sets of data to see if there is any significant difference between the two. The t-test tests against a null hypothesis that the means of two sets of data are equal. In this case, we established one data set as the targeted survey from Phase 2; the second data set was comprised of all eight of the control surveys. Specifically we wanted to test if there was any significant difference between the average number of items per organization found in the Phase 2 survey versus the average number of items per organization found in all eight of the control group surveys. To put it another way, our null hypothesis was that there is no significant difference between the mean values of the Phase 2 survey versus the combined mean values of all eight control surveys.

To test this, we needed to calculate the mean of the Phase 2 survey, which was 25,844 average items per organization. We also needed to calculate the mean of the control surveys, which turned out to be 21,766 average items per organization (see Table 6). Additionally we needed to calculate the standard deviation of both data sets. Standard deviation is a measure of the average variance from the mean across a set of data. Phase 2 survey's standard deviation is 78,012; the standard deviation of the control group surveys is 25,209.[19]

19 We are aware that our comparisons in this t-test are problematic because we are comparing raw data in one data set against compounded and averaged data in the other data set. We are aware that this will

Mean (Phase 2 Survey)	25,844	Mean (Control Group Surveys)	21,766
Variance (Phase 2 Survey)	6,085,809,107	Variance (Control Group Surveys) (average of squared differences from the mean)	635,517,526
Standard Deviation (Phase 2 Survey) (square root of the variance)	78,012	Standard Deviation (Control Group Surveys) (square root of the variance)	25,209

Table 8. Mean, Variance, and Standard Deviation of Phase 2 Survey vs. Control Group Surveys.

The equation used to perform a t-test is as follows:

$$t = (Mean1 - Mean2) \div \sqrt{((SD1*SD1)/Q1 + (SD2*SD2)/Q2)}$$

Mean1 is the mean number of items per organization found by the Phase 2 survey. Mean2 is the mean number of items per organization found in the sample of the control group surveys. SD1 and SD2 represent standard deviations for the Phase 2 survey and the control group surveys respectively. Q1 and Q2 represent the number of data points represented in the Phase 2 survey and the control surveys respectively.

For our study, the t-value was calculated as follows:

$$t = (25{,}844 - 21{,}766) \div \sqrt{((78{,}012 * 78{,}012)/177 + (25{,}209 * 25{,}209)/8)}$$
$$t = 0.382$$

After the t-value is computed, it must be compared against a standard table of t-values, given the size of the sample and a standard alpha level, or risk level, of a = .05. If the t-value that resulted from the test exceeds the standard value, the null hypothesis is rejected and the two samples are found to have significantly different means.[20] Then it is necessary to calculate the degrees of freedom (df) for the test. In the t-test, the degrees of freedom is the sum of the data points in both groups minus 2. In our case, the df was 177 + 8 - 2, which comes out to 183.

Given the alpha level, the df, and the t-value, we referenced the standard table of significance to determine whether our t-value was large enough to be significant. At a = .05, our t-value of 0.328, df=183, was not sufficient to reject the null hypothesis of no difference between the two samples, thus we conclude that there is no difference between the number of items per organization found by the Phase 2 survey and the control group.

Confidence Levels

Based on the results of these two statistical tests, we feel confident that the data acquired as part of the Phase 2 survey can be trusted, as much as we might trust existing survey data (our control surveys) collected and published over that past 15 years, as indicators of the quantities

compound any significant differences from the individual control group surveys. This certainly causes our use of the t-test to be "watered down." We postulate that this would have more effect if we found significant difference between the data sets. In our case, we found little difference between the data sets and we were unable to overturn our null hypothesis.

20 We used this standard table to look up our t-value: http://www.statisticsmentor.com/tables/table_t.htm.

of audio holdings in organizations throughout the US; and that we can use these numbers to project an estimated total quantity of audio holdings to be found in collections nationwide.

Phase 3 Discussion

In March and April of 2014, NEDCC and AVPreserve conducted a nationwide online survey to assess the need for audio digitization services among collection-holding organizations in the US. The survey was sent to a variety of listservs and mailing lists in order to reach the widest US audience possible.[21] Although the intention of this survey was less about quantifying holdings across the US and more about understanding the nature of the field's demand for audio digitization services and its satisfaction with current options, we can extract a few data points that can be used to inform the Phase 2 survey projections. We can also use some of the results in our confidence tests to strengthen our trust in the Phase 2 results.

We are aware that the respondents to the survey are likely to be those who are most aware of the needs of audio holdings in their collections. This self-selection may bias the results of the Phase 3 survey. Although we cannot control the sampling of respondents from the greater population, we feel it is important to note that we are aware of this possibility.

Survey Question 27: Organization Type

There were 221 respondents to the Phase 3 survey. Question 27 of the survey asked respondents to identify the type of organization they represent based on the classifications established at the beginning of this study. 181 respondents answered Question 27. Table 9 shows the number of respondents (the sample) by organization type and then compares those percentages against percentages of total organizations in the US (the population) as gathered in the beginning of the project.

	Quantity established at project beginning	Percentage against total established at project beginning	Quantity from Phase 3 survey	Percentage against total Phase 3 quantity
Number of Academic Libraries in the US	3,689	22%	63	35%
Number of Special Libraries in the US	7,616	46%	22	12%
Number of nonprofit Broadcast Organizations in the US	1,469	9%	7	4%

21 Among others, the survey was sent to listservs for New England Archivists (NEA), Preservation and Access Special Interest Group (PASIG), American Association for State and Local History (AASLH), Society of American Archivists (SAA), Association of Tribal Archives Libraries and Museums (ATALM), New England Museum Association (NEMA), Association of Recorded Sound Collections (ARSC), Association of Moving Image Archivists (AMIA), Oral History Association (OHA), American Folklore Society (AFS) and to mailing lists for the Council of State Archivists (CoSA) and the Presidential Libraries.

Number of State and Large City Historical Societies in the US	200	1%	14	8%
Number of Archives/ Museums in the US	3,569	22%	75	41%
TOTALS	16,543		181	

Table 9. Analysis of types of organizations represented in Phase 3 survey.

The main outlier here is that the representation of special libraries is very low, which may either be a factor of the data gathered in the beginning of the project or an actual low representation in the survey.

Survey Question 1: Total Audio Holdings

Question 1 asked the respondents about the size of their total audio holdings. 221 respondents answered the question (although 12 respondents selected the "I Don't Know" option). The question did not allow respondents to enter a specific number. Instead, it offered a set of five quantity ranges from which a respondent could select one.

Table 10 shows that in order to estimate the total quantity of sound recordings held by respondents of the survey we had to create averages for each quantity range. We then multiplied the averages by the number of respondents to arrive at our estimated total of 7,632,500 audio holdings represented by the respondents of the Phase 3 survey.

Size of Collection	Average	Count	Total
0-5,000	2,500	128	320,000
5,000-25,000	17,500	45	787,500
25,000-100,000	75,000	23	1,725,000
100,000-500,000	300,000	11	3,300,000
more than 500,000	750,000	2	1,500,000
	TOTALS	209	7,632,500

Table 10. Estimated total quantity of sound recordings held by respondents of the survey.

One use of this data point is to add it to our scatter plot of cross-organizational surveys to see how it corresponds to the control groups and to the Phase 2 survey. Table 11 shows the results of the Phase 3 survey added to the previous scatter plot data set.

Survey	Number of Organizations Represented in Sample	Number of Objects Counted Total
Phase 2 targeted survey	177	4,574,397
Phase 3 survey	209	7,632,500
Online Computer Library Center	92	3,000,000
Image Permanence Institute	17	150,000
US Oral History Survey	246	337,505
California Preservation Program	32	604,770
Association of Research Libraries	72	151,920
Council on Library and Information Resources	69	1,899,150
Association of Research Libraries	123	10,000,000
Heritage Preservation (HHI)	30,827	46,000,000

Table 11. Scatter plot data points with Phase 3 survey added.

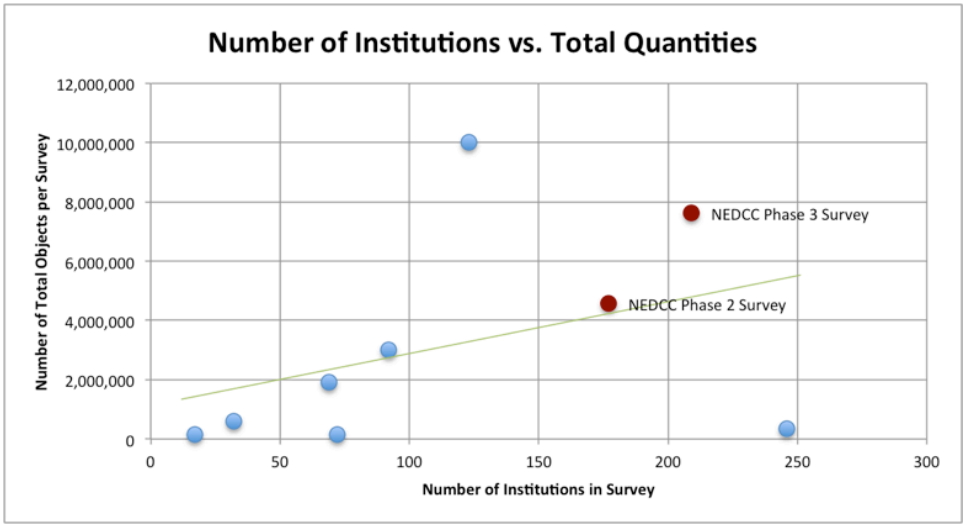

Table 12. Scatter plot of seven control groups, Phase 2 targeted survey, and Phase 3 survey (both represented as red data points).

At 7,632,500 items and 209 organizations, Phase 3 survey falls a little high, but in the general direction of the ratio we noticed among the surveys between number of organizations and number of total audio holdings in a survey. The three outliers are the same outliers discussed previously in this report. We notice that with the addition of the Phase 3 numbers to the plot, the Phase 2 survey is even closer to lying directly on the trendline.

Survey Question 2: Percentage of Unique or Rare Audio Holdings

In our Phase 2 survey, we were able to gather detailed numbers about the quantity of audio holdings by general format (i.e., grooved, magnetic, and optical). However, we were unable to quantify what percentage of these holdings were considered unique or preservation-worthy. We think this is an important distinction that should be accounted for as we attempt to quantify the number of sound recordings in the US in need of digitization. Question 2 in the Phase 3 survey asked respondents to identify what percentage of their total audio collections would be digitization candidates based on rarity or uniqueness. 201 respondents answered this question. Similar to Question 1, this question offered a set of five percentage ranges from which the respondents could select one.

In order to calculate an average percentage of unique or rare audio holdings in the collections of respondents to the Phase 3 survey, we calculated high and low outcomes based on the percentage ranges offered in the survey. For each high and low number we multiplied that number by the quantity of respondents and divided the total value by the total number of respondents to get a total unique percentage of 47% (low) and 67% (high). For calculation purposes in this report, we use an average of these two numbers, 57%, to represent the midpoint between high and low possible outcomes.

Percentage Range	Low Value	Quantity of Respondents	
0 - 20%	0%	40	
20 - 40%	20%	25	
40 - 60%	40%	33	
60 - 80%	60%	26	
80 - 100%	80%	77	Percentage Unique
		201	47%

Percentage Range	High Value	Quantity of Respondents	
0 - 20%	20%	40	
20 - 40%	40%	25	
40 - 60%	60%	33	
60 - 80%	80%	26	
80 - 100%	100%	77	Percentage Unique
		201	67%

Table 13. Low and high total percentages of unique or rare audio holdings in collections of survey respondents.

As we project the total number of audio holdings at collecting organizations in the US, it will be of interest to apply this factor of 57% to give us a clearer picture of the amount of audio holdings that organizations will consider worthy of preservation, and therefore digitization, in order to help assess the potential opportunity in combination with other factors and considerations.

Our experience working with collections has demonstrated that different formats inherently have different rates of uniqueness — usually this is based on the history of certain formats that were used as mass-produced commercial products (e.g., vinyl discs) versus those that were

developed for individual use and independent recording technologies (e.g., open-reel tape). However, the survey did not break this question down into media type (i.e., grooved, magnetic, and optical). Therefore, we applied the 57% assumed preservation-worthy items across the board; using the data we have, we felt it was the most prudent option at this stage.

Survey Question 3: Percentage of Digitized Audio Holdings

Efforts to digitize audio accelerated in the early 2000s. The pace has quickened in the past five years as collection-holding organizations and preservation advocates have increasingly made the case that the clock is ticking for physical analog and digital sound media. Recognizing this, a final factor of interest to our projections is the quantity of sound recordings in collections that have already been digitized and are therefore not relevant to our assessment. Question 3 asked respondents to identify the percentage of their audio holdings that have already been digitized. Like Question 2, respondents selected a percentage range. 221 respondents answered the question (13 answered "I Don't Know").

Percentage Range	High Value	Quantity of Respondents	
0 - 20%	0%	168	
20 - 40%	20%	20	
40 - 60%	40%	11	
60 - 80%	60%	3	
80 - 100%	80%	6	Percentage Digitized
		208	7%

Percentage Range	High Value	Quantity of Respondents	
0 - 20%	0%	168	
20 - 40%	20%	20	
40 - 60%	40%	11	
60 - 80%	60%	3	
80 - 100%	80%	6	Percentage Digitized
		208	27%

Table 14. Low and high total percentages of audio holdings that have already been digitized in collections of survey respondents.

From the respondents' answers to Question 3, we calculate that a range of 7% to 27% of audio holdings have been digitized to-date. For the purpose of this study, we use the average of this range, 17%, to estimate the total percentage of audio holdings that have been digitized already in collection-holding organizations. As we project the total number of audio holdings in collections across the US, it will be of interest to factor this percentage of sound recordings that will not be relevant to the assessment.

Findings

Our intention with this assessment was to acquire a dataset reliable enough to allow us to project the total quantity of preservation-worthy, not-yet-digitized audio holdings in collection-holding organizations throughout the US. We also hoped to acquire enough detailed information to be able to estimate the quantities of formats at granular levels. Based on our efforts with Phase 1 and Phase 2 data aggregation and our use of statistical analyses to evaluate the integrity of our data, we feel comfortable offering the following projections:

	Average Number of Items per Organization	Estimated Number of Like Organizations in the US	Estimated Number of Audio Items Held by Like Organizations in the US	Estimated Number of Preservation-worthy Audio Items Held by Like Organizations in the US	Estimated Number of Preservation-worthy Audio Items Not Already Digitized by Like Organizations In the US
	x	y	x*y	(x*y)*57%	((x*y)*57%)*(1-17%)
Academic Libraries/ Archives	82,126	3,689	302,962,333	172,688,530	143,331,480
Broadcast	8,324	1,469	12,227,472	6,969,659	5,784,817
Historical Societies	2,094	200	418,850	238,745	198,158
Special Libraries	10,599	7,616	80,722,830	46,012,013	38,189,971
Museums/ Archives	39,476	3,569	140,890,310	80,307,476	66,655,205
		Total Estimated Number of Items in the US Held by these Organization Types	537,221,794	306,216,423	254,159,631

Table 15. Projected number of preservation-worthy, not-yet-digitized audio holdings nationwide, as of December 2014.

Based on the per-organization-type averages from our Phase 2 survey and the total estimated number of organizations in the US by organization type, we are able to project an estimated 537,221,794 sound recordings held by collecting-organizations in the US. Based on our Phase 3 survey, we can refine this projection by considering the number of “preservation-worthy” and “already digitized” items. In the Phase 3 survey we found that 57% of sound recordings are considered preservation-worthy (based on rarity and/or uniqueness) by collection-holding

organizations in the US. Factoring this percentage against our projected total we find there to be an estimated 306,216,423 preservation-worthy sound recordings in collections in the US. Finally, knowing that some percentage of these recordings have already been digitized, we factor in the Phase 3 finding that an average of 17% of holdings have already been digitized. This provides our final projection that there are an estimated 254,159,631 preservation-worthy items in US collections that have not yet been digitized as of December 2014.

Total Audio Holdings by Format

For the purpose of this study, we are also interested in understanding the breakdown by format of preservation-worthy items in the US. Because none of the data from Phase 1, 2, or 3 provided enough evidence to establish firm numbers, we took a close look at past AVPreserve collection assessments, paying close attention to the breakdown of preservation-worthy grooved, magnetic, and optical media. From these assessments, we calculated the percentage of all preservation-worthy items for each media type and applied those percentages to 254,159,631, the total number of preservation-worthy media holdings (Table 16).

Media Type	% of Total Items	Number of Preservation-worthy Items
Grooved	20%	50,831,926
Magnetic	70%	177,911,742
Optical	10%	25,415,963

Table 16: Total number of preservation worthy items in US collections by media type.

Costs and Workflows (Specialized vs. Non-Specialized)

To calculate an average cost for audio digitization per item, we drew on vendor quotes received over the past few years, as well as our knowledge of the market. Table 17 outlines these costs and averages them. We split the vendor costs into two categories: items requiring a specialized workflow and those that do not. The distinction here is that some items, because of their physical condition or because of the nature of the content they contain, cannot be handled in batch and, therefore, cannot benefit from pricing at scale. On average, the items requiring a specialized workflow cost twice as much as those that do not. It is notable that current market prices for digitization are extremely variable across vendors and volatile overall. Today's prices represent all-time historically low costs for digitization. Many project that prices will rise again as it becomes more resource intensive to manage operations that depend on obsolete equipment and expertise. The pricing below represents averages over a range of time and are intended to estimate a rough order of magnitude. The reader is strongly discouraged from using this information as a pricing guide or in any other way than the intended use within this study.

Format	Non-Specialized Cost per Item	Specialized Cost per Item
1/4 Inch Open Reel Audio - Acetate & Polyester	$75.00	$150.00
CD, CD-R	$15.00	$30.00
Compact Audiocassette	$45.00	$90.00
DAT	$45.00	$90.00
Flexi-disc	$100.00	$200.00
Lacquer Disc	$75.00	$150.00
LP	$45.00	$90.00

Metal Master	$100.00	$200.00
Microcassette	$45.00	$90.00
Minicassette	$45.00	$90.00
Minidisc	$45.00	$90.00
Pressed 45 RPM Disc	$45.00	$90.00
Pressed 78 RPM Disc	$45.00	$90.00
Wax Cylinder	$100.00	$200.00
Wire Recording	$75.00	$150.00
Average Cost per Item	$60.00	$120.00

Table 17. Recent vendor quotes of audio carrier costs by format with the average carrier cost.

In order for this information to be of use to this study, we need to estimate the quantity of the total market (see Table 16) of media items that will require specialized or non-specialized workflows. To determine the number of magnetic media requiring a specialized workflow, we assumed that of the 177,911,742 total magnetic items calculated above, 70% are 1/4 inch open reels, 25% are cassettes, and 5% are DAT tapes. These numbers are based on the likelihood that after narrowing down the items to preservation-worthy items, most cassettes and DAT tapes will have been eliminated due to the prevalence of commercial recordings or duplicate copies. Then, in order to calculate the percentage of media that is suitable for specialized vs. non-specialized workflows, we applied the same allocations used by Indiana University when deciding how to route assets by format through the digitization process. This results in the conclusion that 19% of magnetic media will require a specialized workflow, or 33,803,231 items. Based on our experience, we estimated 90% of preservation worthy grooved media (45,748,734) and 3% of preservation worthy optical media (762,479) will require a specialized workflow. Therefore, 80,314,443 of the total number of preservation-worthy items, or 32%, will require a specialized workflow.

Conclusions

Our findings from this needs assessment demonstrate that there are over 537 million audio items held in collecting organizations across the US, mostly in academic libraries and archives/ museums. Knowing that an approximate 57% of holdings are unique or rare and 17% of holdings have already been digitized, using extrapolations based on quantities of organizations in the US, we are comfortable estimating the following numbers of preservation- worthy items that have not been digitized to-date:

Total Grooved Media	50,831,926
Total Magnetic Media	177,911,742
Total Optical Media	25,415,963
Total Audio Items	254,159,631

Table 18: Total number of not-yet-digitized preservation worthy audio holdings by media type in collections in the US.

To quantify the market, using average costs for specialized and non-specialized workflows, we can estimate the following numbers:

Media Type	% of Total Items	Number of Preservation Worthy Items	% Requiring Specialized Workflow	Number of PWI Requiring SW	Number of PWI not Requiring SW
Grooved	20%	50,831,926	90%	45,748,734	5,083,193
Magnetic	70%	177,911,742	19%	33,803,231	144,108,511
Optical	10%	25,415,963	3%	762,479	24,653,484
Total	100%	254,159,631		80,314,443	173,845,188
			Market Costs	$9,637,733,203	$10,430,711,251

Table 19: Total number of items requiring a specialized and non-specialized workflow by media type.

Based on these numbers, the effort to digitize preservation-worthy sound recordings in collection-holding organizations nationwide is estimated to cost over $20 billion, which does not include the costs that will be associated with preparing materials for digitization, describing materials for access, ongoing storage of digital files for preservation and access, and general organizational overhead. This is the reality of the challenge ahead to preserve our nation's audio heritage.

In order to overcome the funding and infrastructure barriers that US collection-holding organizations will face in meeting this challenge over the next twenty years, it will be essential to have an accurate picture of what the need is today. The quantity and makeup of nationwide holdings is a foundational piece of information, serving as a cornerstone for many publications, initiatives, and efforts. Incorrect numbers yield a response that is out of alignment and falls short of the need. Therefore it is critical to have accurate data in hand. When coupled with other business-oriented analysis, such as the Cost of Inaction (https://coi.avpreserve.com/), and quantitative data we can begin to think critically and pragmatically about the available options and their implications. The challenges and tough choices ahead are best faced with good data in hand in order to make well-informed decisions.