DPOE 2014: Survey Analysis

Nick Krabbenhoeft

02/02/15

Contents

Summary

This report analyzes the data collected from a survey conducted by the Digital Preservation Outreach and Education (DPOE)[1] program of the Library of Congress during July, August, and September of 2014. In total, 436 participants submitted responses to questions about their organizations, collections, and training preferences. The following analysis breaks down the data roughly according to those groups.

[1] The mission of DPOE is "to foster national outreach and education to encourage individuals and organizations to actively preserve their digital content, building on a collaborative network of instructors, contributors, and institutional partners."

Methodology

This is an exploratory report conducted on a volunteer basis by CodedCulture. As an exploratory report, there are no research questions to answer. Instead, the results are seeds for future questions and surveys to answer.

The dataset was provided by Barrie Howard of the DPOE Committee as a raw export from SurveyMonkey. Several steps were taken to clean and normalize the dataset, including[2]:

- removing potential personally identifiable information such as email address
- shortening column names for easier referencing during analysis
- normalizing values for responses to ranking questions

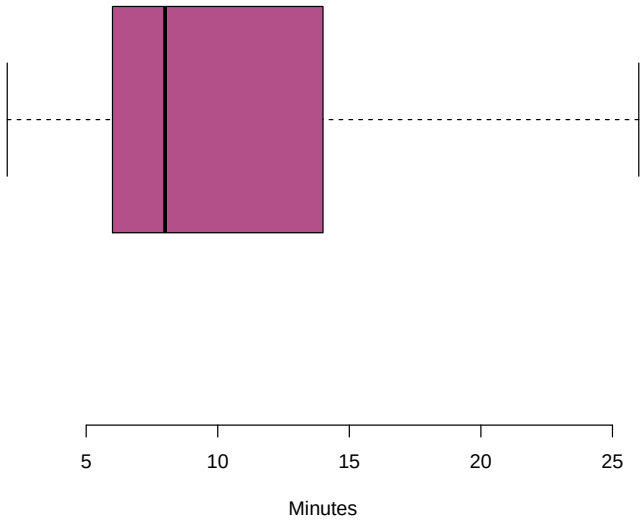
[2] Changes are documented more completely in the change log accompanying the dataset

The dataset was then loaded into R, an open-source statistical analysis and graphing program. A verbose version of this report with both the code and the results is available. All files used in the project were placed in a git repository, available on Github.

Three types of questions make up most of the survey, single-answer multiple-choice, multiple-answer multiple-choice, and forced ranking questions.

In the analysis, answers to multiple-choice questions are treated as categorical factors. Where a respondent chose multiple answers, the response is grouped with each category factor.

Forced ranking questions had respondents rank a set of choices with no ties were allowed. Respondents could enter an additional choice with a free text "other" field. As a result, responses to the same 5 choice question could be on a scale of 1-5 or 1-6. Combined with the small range of values, this made it difficult to interpret average rankings. All ranking data is presented as bar plots to better visualize the distribution of votes.




```
## Error in topSliceHist(df.rformat, df.rlength, "LengthRanking",
"Format", :  object 'df.rformat' not found
```

Time Analysis

SurveyMonkey records the start and end times for each completed survey, which can give a lot of insight into how the survey was administered and respondents interacted with it.

How long did it take to complete the survey?

Survey completion time varied greatly. Some respondents opened the survey in a tab and only completed it days later. Those outliers skew the average completion time to 38 minutes. Ignoring the outliers and looking at the distribution of the dataset shows that half of the respondents completed the survey in 8 minutes or less (see Figure 1). An average respondent likely took between 6 and 14 minutes to finish.

The average completion times are short enough that just the starting time alone acts as a rough estimate of completion rates.

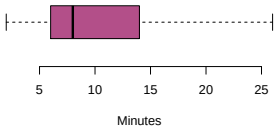


Figure 1: Boxplot of Completion Time

When were surveys started?

The survey was first released on Thursday, July 17 with announcements on several major community listservs. After an initial rush, the response rate tapered off quickly. Nearing the original deadline of Friday, August 15, weekly completion rates were in the single digits (see Figure 2). Then, the deadline was extend by another month, advertised with a second wave of listserv announcements.

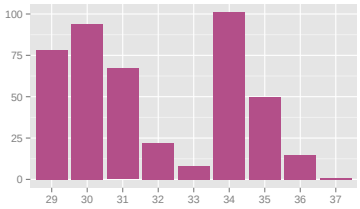


Figure 2: Survey Starts by Week

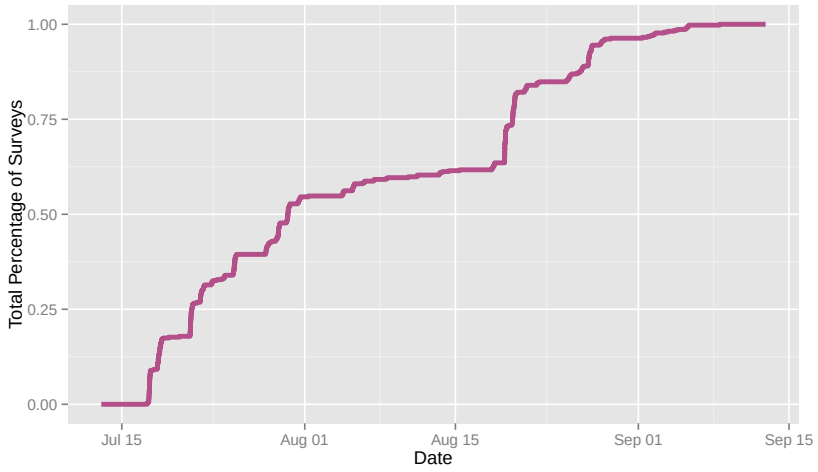


Figure 3: Completion Distribution of Surveys (July-August)

Nearly 40% of the total responses came after the extension of the deadline (see Figure 3). Like the first wave, the response rate dropped off quickly after the initial rush. Without records of where and when announcements were made, it's only speculation to say what drives survey completion. But, it wouldn't be surprising if there's a direct correlation between listserv announcments and survey completions.

How does work affect responses?

Looking at what day of the week, surprisingly, Tuesday and Wednesday were the most common days to take the survey (see Figure 4). Again, we would need more information about the announcements to give a reason for this. One potential is that filling out a survey is a nice break during the middle of the week. Unsurprisingly, nearly no one took the survey on the weekend.

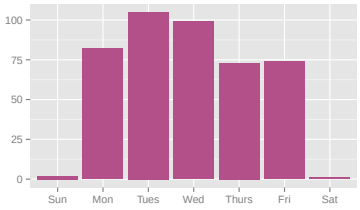


Figure 4: Survey Starts by Weekday

During the workday, responses are grouped into two rough time periods, before and after lunch (see Figure 5). Digital preservation starts early in the morning for many respondents, with a large number of responses coming between 6am and 9am. That amount continues growing until a sharp dip for lunch. The peak responses came

right after lunch at 1pm, but by 4pm the day is done.

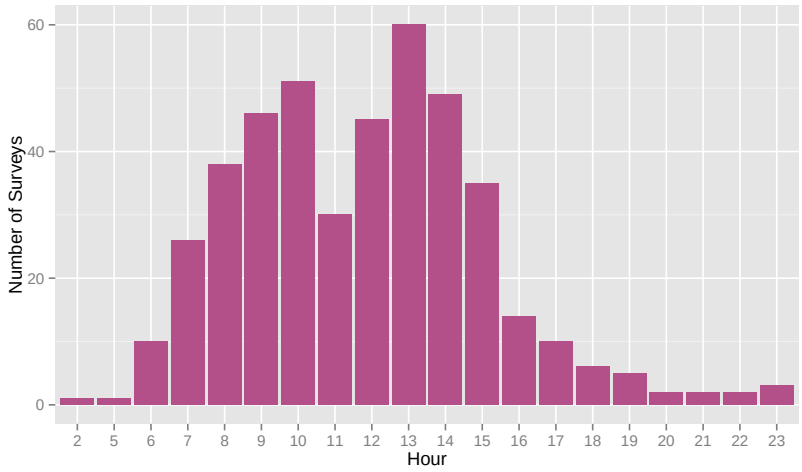


Figure 5: Survey Starts by Hour

However, exploding the time data by timezone reveals a strange skew to the dataset (see Figure 6). While the Eastern and Central timezones are well-represented, there are very few respondents from the western half of the US. This will be explored more in Respondents section.

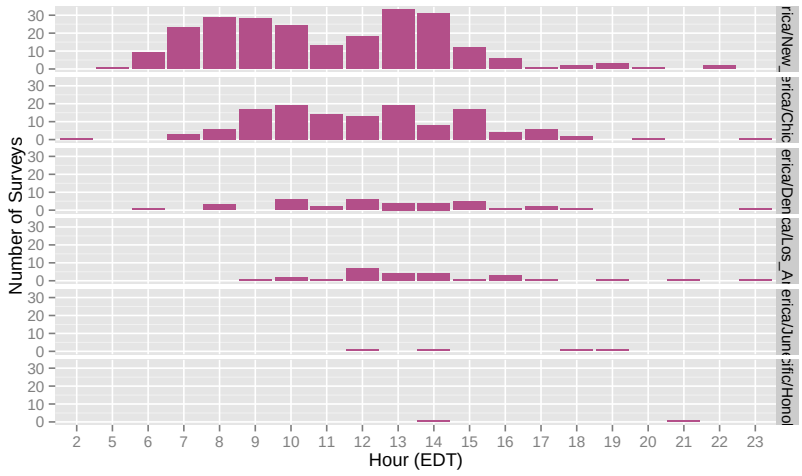


Figure 6: Survey Starts by Hour and Timezone

Recommendations

Many of the trends found in this dataset may seem trivial, they point to an important opportunity to strengthen future versions of the survey: tracking listserv announcements. SurveyMonkey offers several methods to track the source of inbound traffic such as Collector Options or Custom Survey Links. By tracking listserv announcements,

DPOE can evaluate and improve many aspects of the survey, including:

- how the timing of an announcement effects response
- how much response each announcement receives
- how quickly response to an announcement decays
- how listservs reinforce or cut across sections of the digital preservation community

The final question is one worth a lot of consideration. The digital preservation community is spread out. Someone can be the only digital preservation practicioner in their organization, city, or even region. The community relies on listservs and social networking to keep in contact, so measuring responses from them is important to improving engagement.

Respondents

The timezone analysis (see Figure 6) highlighted a strong eastern bias in the dataset. Understanding who is represented in the dataset bares strongly on how the rest of the responses can be interpreted.

Where are respondents from?

Comparing a state's portion of survey responses to its portion of the US population is a good, rough measure of how under- or over-represented it is.

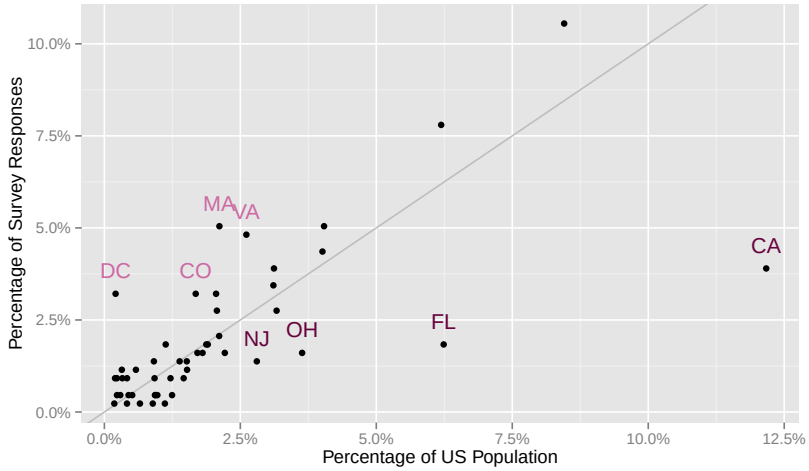


Figure 7: Survey Participation Compared to US Population by State

In Figure 7, the guide measures where the percentage of responses equals the percentage of the US population. States with less responses than expected are below the guide. Those with more are above. The further away from the guide, the more under- or over-represented a state is. In general, two proportions are correlated, but 8 states deviate from expectation by 50% or more: Washington DC[3], Colorado, Massachusetts, Virginia, New Jersey, Ohio, Florida, and California.

[3] Not a state, but it should be.

The first 4 are understandable. As regional or national cultural centers, DC, Colorado, Massachusetts, and Virginia have a higher number cultural heritage organizations than average. However, New Jersey, Ohio, Florida, and California are concerning. The last two represent nearly 20% of the nation's population and contributed only 6% of the responses. California has many large, well-established digital collections including the Internet Archive, Stanford Archives, California Digital Libaries, and Computer History Museum, and Florida hosted the Florida Center for Library Automation for 3 decades.

Is there less of a need for digital preservation education in these states? Are there fewer but larger organizations there? With 1 year of data, it may just be an anomaly. The state-by-state participation rate will need continued measuring to establish a trend.

What kinds of organizations do respondents work for?

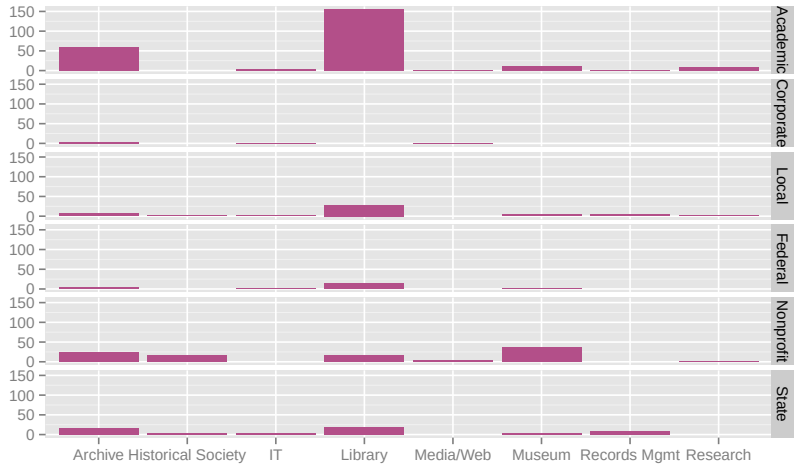


Figure 8: Responses by Organization Type and Parent Organzation (Count)

The survey asked respondents to characterize their organization by the following categories:

Parent Organization Academic, Corporate, Nonprofit, Federal, Tribal, State, or Local

Organization Type Archives, Historical Society, IT firm, Library, Media/Web firm, Museum, Records Management firm, or Research Center

Staff Size 1-12, 13-25, 26-50, 51-200, 201-500, 500+

Looking at the distribution of parent organization and organization types, the most common parent organizations are Academic, Nonprofit, and State, and the most common organization types are Libraries, Archives, and Museum (see Figure 8). The rest of respondant analysis defaults to analyzing just these categories.[4]

Academic libraries are represented by over a third of the responses, and their weight can overwhelm clear comparison in the following analysis. As a result, many of the following graphs calculate the percentage of responses per category instead of the pure count of responses.[5]

[4] Using the union of these two sets, not the intersection. For example, a federal library is included in the list of libraries and an academic research center is included in the list of academic organizations.

[5] This is noted in each graph title with a (Count) or (Percentage) label.

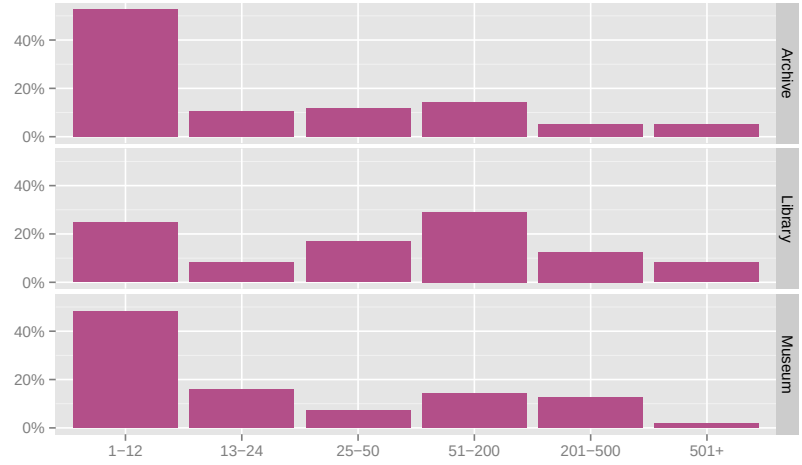


Figure 9: Staff Size by Organization Type (Percentage)

The next question is how large these organizations are (see Figure 9). A majority of archives and museums have staffs of 12 or less. In contrast, the distribution of libraries staff sizes is more spread out. Digging further into the parent organizations, most libraries resembles museum and archives in terms of staffing, but academic and state libraries are typically much larger (see Figure 10). For example, the 155 academic libraries themselves average an estimated miniumum staff of 98 per library. With staffs that large, management and bureaucratic challenges probably have a strong impact on digital preservation in these organizations.

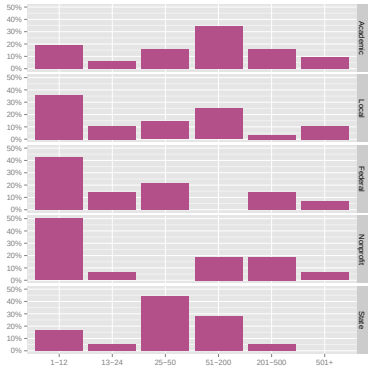


Figure 10: Staff Size at Libraries by Parent Organization (Percentage)

Who does the work of digital preservation?

The size of the overall organization effects the staffing for digital preservation responsibilities (see Figure 11). As the total staff size increases, organizations tend to hire more dedicated staff and add digital preservation responsiblities to existing staff. The smallest organizations are more likely to use volunteers (20%) or to have no staff (35%).

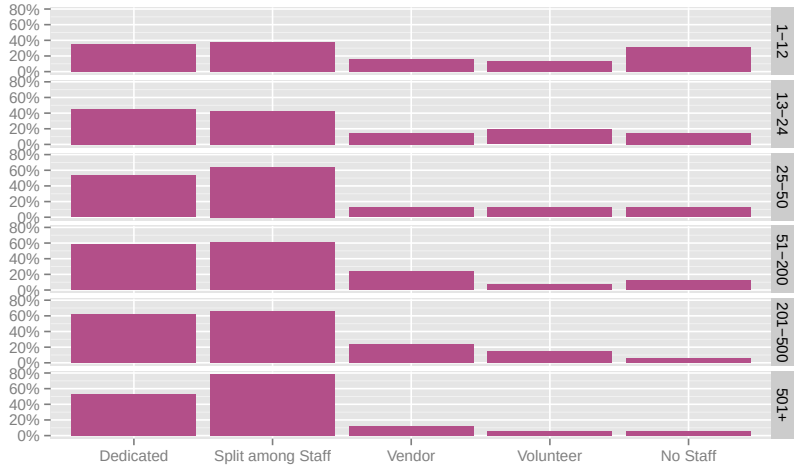


Figure 11: Responsibility by Staff Size

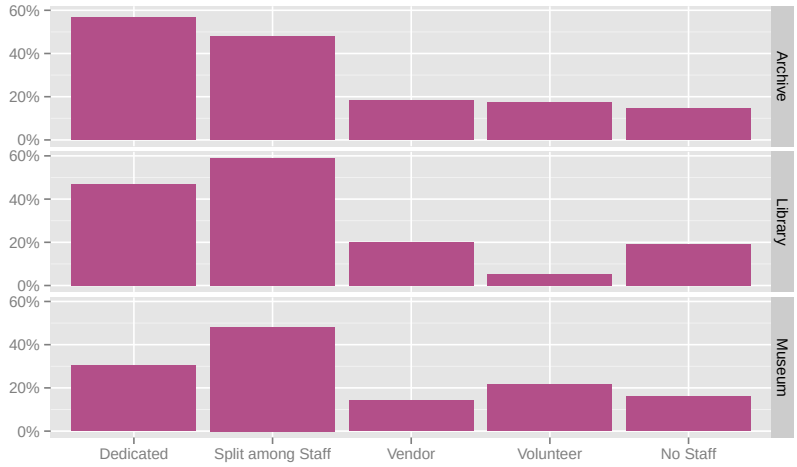


Figure 12: Responsibility by Organization Type

Comparing digital preservation responsibilities across archives and libraries and museums, shows a couple important trends. Archives and museums tend to use more volunteers; unsurprising, since they also tend to have very small staffs. There is also a clear difference in the hiring of dedicated staff: archives (60%), libraries (45%), then museum (30%). It's tempting to draw a connection between those

numbers and the variety of potential job titles for digital preservation staff in each group:[6]

[6] Drawn from job postings.

Archives Digital Archivist, Digital Project Archivist

Libraries Digital Projects Librarian, Digital Initiatives Librarian, Data Curation Librarian, Metadata Librarian

Museum Digital Asset Manager, Data Curator, Digital Projects Manager

What is being preserved?

One of the most interesting points of analysis is how much digital collections are alike in terms of their content. In terms of both:

Content type CAD, AV, Images, GIS, Office files, PDF, Research data, Websites

Content format Born-digital, Deposited materials, Licensed, Digitized

organization struggle with the same issues accross parent organizations, organzation types, and staff sizes (see Figure 13).

Figure 13: Content Types and Formats by Parent Organization, Organization Type, and Staff Size (Percentage)

There are several differences worth noting. For example the importance of GIS and research data grows as the size of the organizations,

with GIS data particularly important to state and academic libraries. Licensed material is primarily important to academic libraries and larger organizations. These trends make sense given the specialized needs of these organizations's audiences (e.g. property registries and research labs).

However, outside of those areas, the similarity between each series of plots is remarkable. When it comes to the bits and bytes, everyone is working with very similar collections and probably facing the same problems.

What's the training budget?

In final organizational characteristic, funding available for training, nearly a third of respondents did not know how much funding was available (see Figure 14). These answers, along with the neglible number of responses with more than $3,000 of funding are not included in the following graphs.

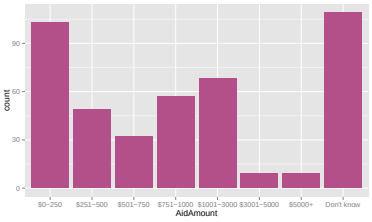


Figure 14: Available Training Budget (Count)

In regards to staff position, organizations with staff for digital preservation lie on extremes, equally able to fund thousands of dollars worth of training or none at all (see Figure 15). Comparing the organization size to funding may explain this split. Most organzations with 50 or less staff provide $500 or less for training. With more than 50 staff, training budgets expand to $750 or more (see Figure 16).

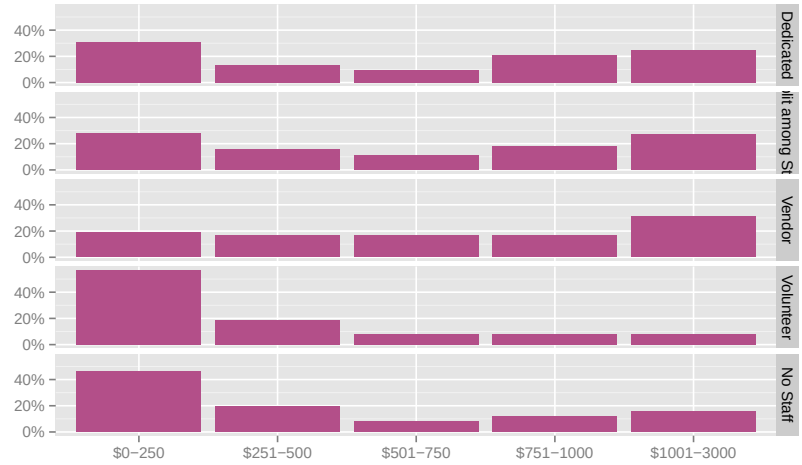


Figure 15: Responsibility by Training Budget (Percentage)

Recommendations

Respondents represent 2 distinct groups of organization, small and large. Digital preservation gains greater institutional buy-in as staffs

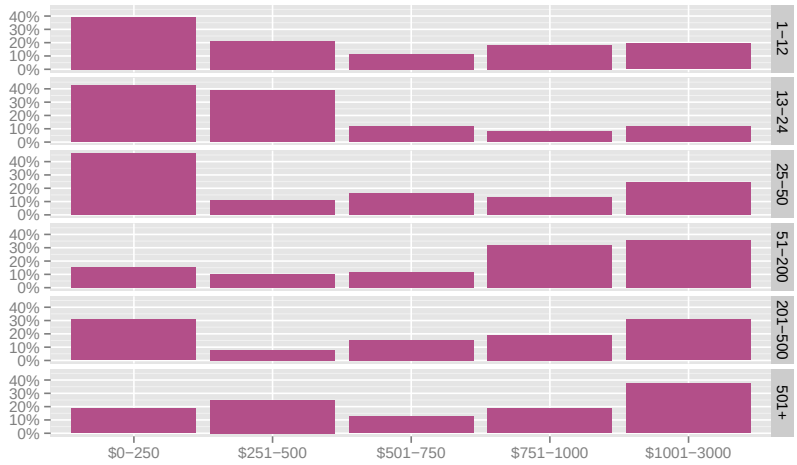


Figure 16: Training Budget by Staff Size (Percentage)

grow, as does budget for digital preservation training. This suggests a potential funding strategy, using training for larger organizations to subsidize training for smaller organizations. When it comes to training content, topics such as strategy and management will depend on the size of a participants organzation, but the fundamentals regarding collection content can be shared.