Using Google Analytics, Voyant and Other Tools to Better Understand Use of Manuscript Collections at L. Tom Perry Special Collections

Ryan K. Lee
Cory L. Nimer
J. Gordon Daines, III
Shelise Rupp
Brigham Young University

Introduction

Developing strategies for making data-driven, objective decisions for digitization and value-added processing.[1] based on patron usage has been an important effort in the L. Tom Perry Special Collections (hereafter Perry Special Collections). In a previous study, the authors looked at how creating a matrix using both Web analytics and in-house use statistics could provide a solid basis for making decisions about which collections to digitize as well as which collections merited deeper description.[2] Along with providing this basis for decision making, the study also revealed some intriguing insights into how our collections were being used and raised some important questions about the impact of description on both digital and physical usage. We have continued analyzing the data from our first study and that data forms the basis of the current study. It is helpful to review the major outcomes of our previous study before looking at what we have learned in this deeper analysis. In the first study, we utilized three sources of statistical data to compare two distinct data points (in-house use and online finding aid use) and determine if there were any patterns or other information that would help

curators in the department make better decisions about the items or collections selected for digitization or value-added processing. To obtain our data points, we combined two data sources related to the in-person use of manuscript collections in the Perry Special Collections reading room and one related to the use of finding aids for manuscript collections made available online through the department's Finding Aid database (http://findingaid.lib.byu.edu/). We mapped the resulting data points into a four quadrant graph (see figure 1).

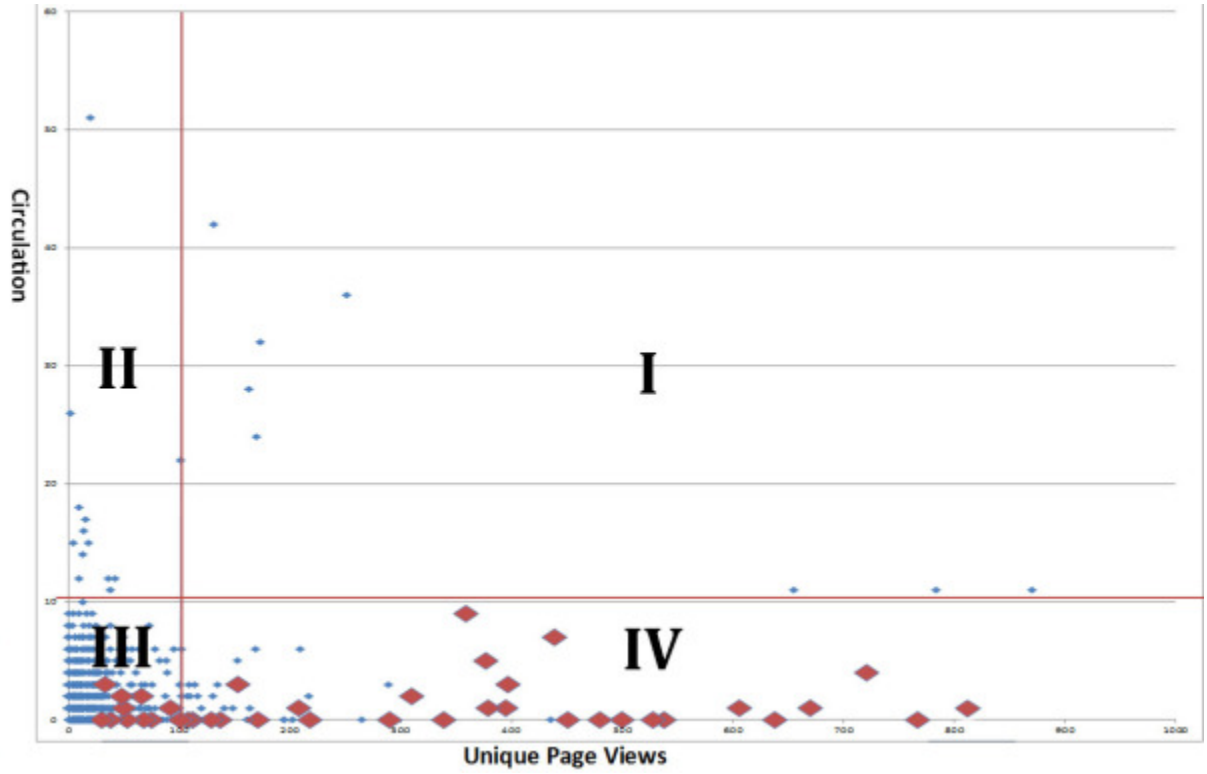

Figure 1. Plot graph from original research project showing quadrants where collections fell based on use

Quadrant I indicated high Web visibility and high reading room circulation, Quadrant II indicated high reading room circulation and low Web visibility, Quadrant III indicated low Web visibility and low reading room circulation, and Quadrant IV indicated high Web visibility and low reading room circulation. One of the most surprising results of our data mapping was the discovery that nearly 98 percent of collections fell in

Quadrant III. This was perplexing and a little concerning to us. We also wondered what drivers resulted in collections falling into Quadrants II and IV.

Our initial hypothesis was that levels of description directly impacted use, and thus which quadrant a collection ended up in on our graph. However, we felt like we needed to better understand why certain collections fell where they did, partly to see if our initial research had any serious flaws. So, we decided to dig a little deeper into the data in order to answer several questions. For collections in Quadrants II and IV we asked the following questions:

- Were the parameters set to determine high use too high?
- Would additional description of collections in Quadrant II attract additional Web access, in addition to in-house use?
- Why are researchers (or the public) visiting the findings aids for materials that are not accessed in person, as charted in Quadrant IV, and what role does prior digitization have on these statistics?

These were important questions to answer to help us better understand the collections in these quadrants that were a mix of high and low use. However, the questions we really wanted to understand related to the collections in Quadrant III. We wanted to understand why these collections were not being used, by asking the following questions:

- What are the key characteristics of these low-use collections?
- Could low use be an effect of applying MPLP[3] principles, and should this approach be reconsidered? Would deeper description move the low use collections into other quadrants?
- Do the low use collections reflect patron interests and needs, or are these items being collected for other reasons? Has the niche or market for these items diminished, or even disappeared?

This study will show how we used free online tools and other means to dig deeper into our usage data to answer many of the questions posed in our initial study.

Methodology

To answer these questions, we determined to use some of the data we harvested from our initial research project to further assess and understand reasons for the use, or lack of use, of our collections. While it is noted that this data is now somewhat dated (since we were merely trying to assess the use of our collections at a point in time, as well as trying to answer questions that were derived from this same data) we determined that the use of this data from our prior research was still relevant for our purposes.

We first wanted to better understand the characteristics of the 98 percent of our collections that had low use over the two-year period, meaning 0-1 circulations or 0-10 unique pageviews[4] (hereafter UPVs), to see if we could determine reasons for these collections not being used. Using a research assistant, we analyzed different aspects of collections that fell in this category, including average use, level of description, digitization, restrictions, and size (see table 1).

Average circulations	0.45
Average UPVs	16.37
Average size (linear ft.)	3.79
Collection-level description	83%
No online finding aid	7.5%
Zero circulations + restricted	19%
Zero UPVs + no online finding aid	80%

Table 1. Characteristics of low-use collections over two-year period

The average circulations of collections in this group were 0.45 and average UPVs were 16.37 over two years. The average size was 3.79 linear feet. In sampling about 2300 collections, we discovered that 83% were only described at the collection level, and 7.5% did not even have an online finding aid. For those with zero circulations, which included 4,772 of 6,373 collections, we looked at how many had some sort of access restriction. After analyzing a sample of nearly 1,300 collections, we discovered that 19% were restricted. Most of these came from the University Archives, which makes sense since many of these records were access restricted to those within their respective administrative units. For those with zero UPVs, which included 566 of 6,373 collections, we wanted to look at their online presence, or lack thereof. Since this was a smaller number of collections, we did not need to sample. Only ten collections in this sub-category had been digitized, but none of these linked to the finding aid. Over 80% (456) did not have an online finding aid. In looking at smaller collections (less than one box), they seemed to have similar levels of online interest as other larger collections, yet they

circulated much less than larger collections. Finally, one interesting find was discovered when looking at all digitized collections in this category. We found twenty-three digitized collections that had no circulations, all of which had more than ten UPVs.

This analysis of low-use collections helped us realize the effects of level of description on use, with those described at lower levels (e.g., file or item level) likely having more use. Size also matters to some extent, at least when it comes to circulations. This is likely due to smaller collections, especially single items, having limited potential use due to the lower amount of information they provide on a particular topic. The obvious importance of online finding aids was also apparent, especially for online use, and digitization may also play a factor in such use, since of the hundreds of collections we have digitized, only a small percentage were in the low-use category. Digitization may also lower circulation.

To determine if subject matter was a factor in use, we decided to apply textual analysis to various collections in the Finding Aid database. Using another research assistant, we began by dividing the collections into separate groups based on use, using previously gathered data on UPVs and in-house circulation to do so (see table 2).

Category	# of circulations	# of UPVs
“Low”	1	1-10
“Medium”	2-4	11-49
“Moderately High”	5-9	50-99
“Very High”	10+	100+

Table 2. Groupings of collections based on use over a two-year period

Expanding the groupings from the previous research, the groups created were “No Use,” “Low,” “Medium,” “Moderately High,” and “Very High”. The “No Use” group (as the name implies) are those collections that had not been viewed online or circulated at all in the time period from which the data was gathered. The “Low” group are those collections which had either one in-house circulation, 1-10 UPVs, or both. The “Medium” group consisted of collections with 2-4 circulations, 11-49 UPVs, or both. “Moderately High” was all those with 5-9 circulations, 50-99 UPVs, or both. Finally, the “Very High” grouping was all of those collections that had reached 10+ circulations, 100+ UPVs, or both.

With these distinctions in place, we took samples from each group and began further analysis by looking first at the collections with the highest online and in-house use (Very High Use group). To do a textual analysis on finding aids in this category, we determined to use the Web-based research tool called Voyant[5] to provide an analysis of possible search terms or subjects pertinent to each collection and the groupings thereof. We hoped to find patterns that would demonstrate a common (and hopefully reproducible) factor across the board for high use collections.

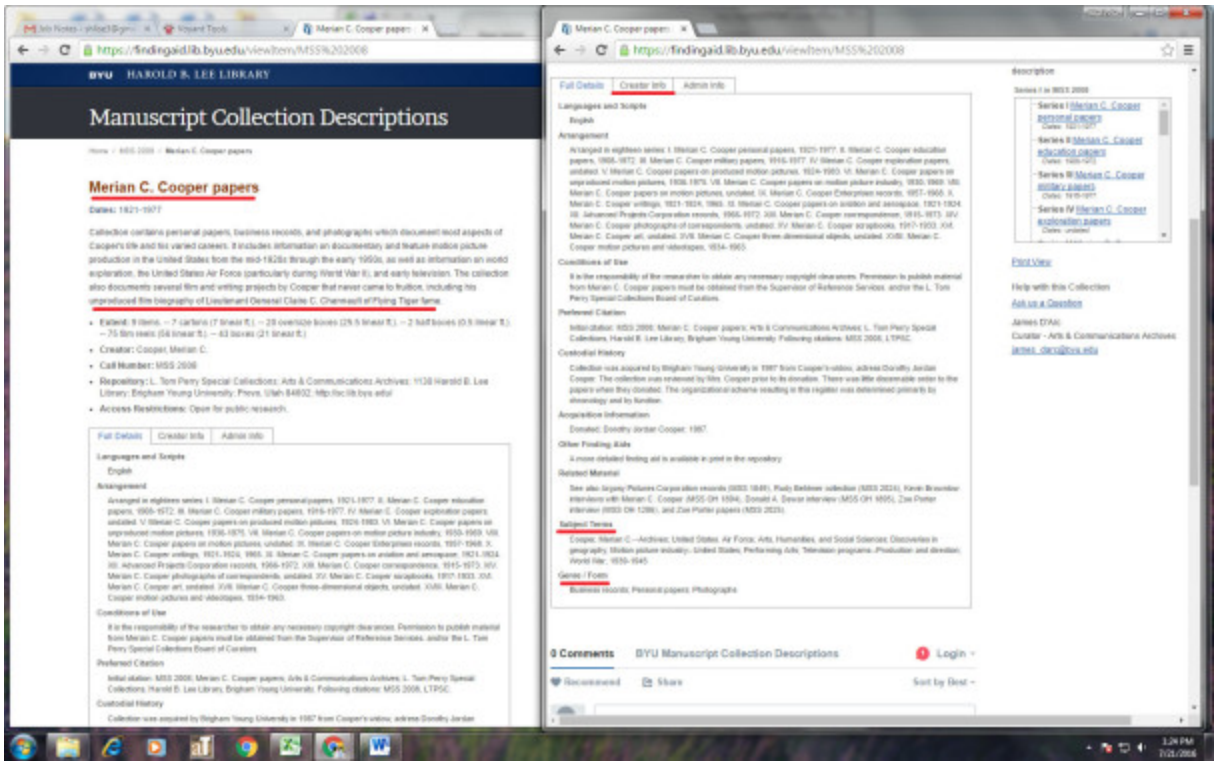

Figure 2. Example of finding aid, with highlights indicating sections of finding aid included in Voyant corpus

The following are the steps taken to use Voyant to achieve the above aim: First, we created the corpus[6] from the samples of the Very High 'both' group, a total of eight collections that had both 10+ circulations and 100+ UPVs. We located the online finding aid for each of these collections in our Finding Aid database and stripped each individual page included in the various online collections of title, description, subject, genre, and creator information (see figure 2), condensing them all into one Word document corpus. These fields were chosen to create the corpus because they are the fields that are searchable both within the Finding Aid database and via Google, where most people are coming to view our finding aids. Creating the corpus for these eight collections took a considerable amount of time because the finding aids in this category were, for the most part, large and complex, having lower levels of description which resulted in sometimes hundreds and thousands of unique pages[7].

Once we had the corpus with the text of each finding aid in one large document, the next step was to strip out any unwanted punctuation that would throw off the data analysis. The punctuation removed included the following: ! – “ ‘ ( ) & * # ^ / \ & : ; [ ] ?, and anything else questionable. Use of the Ctrl-F function while creating the corpus itself to find this punctuation saved a great deal of time, since we could effectively search and delete this data. When this corpus was finally completed, we went to the Voyant website at http://voyant-tools.org/ and uploaded the corpus (figure 3).

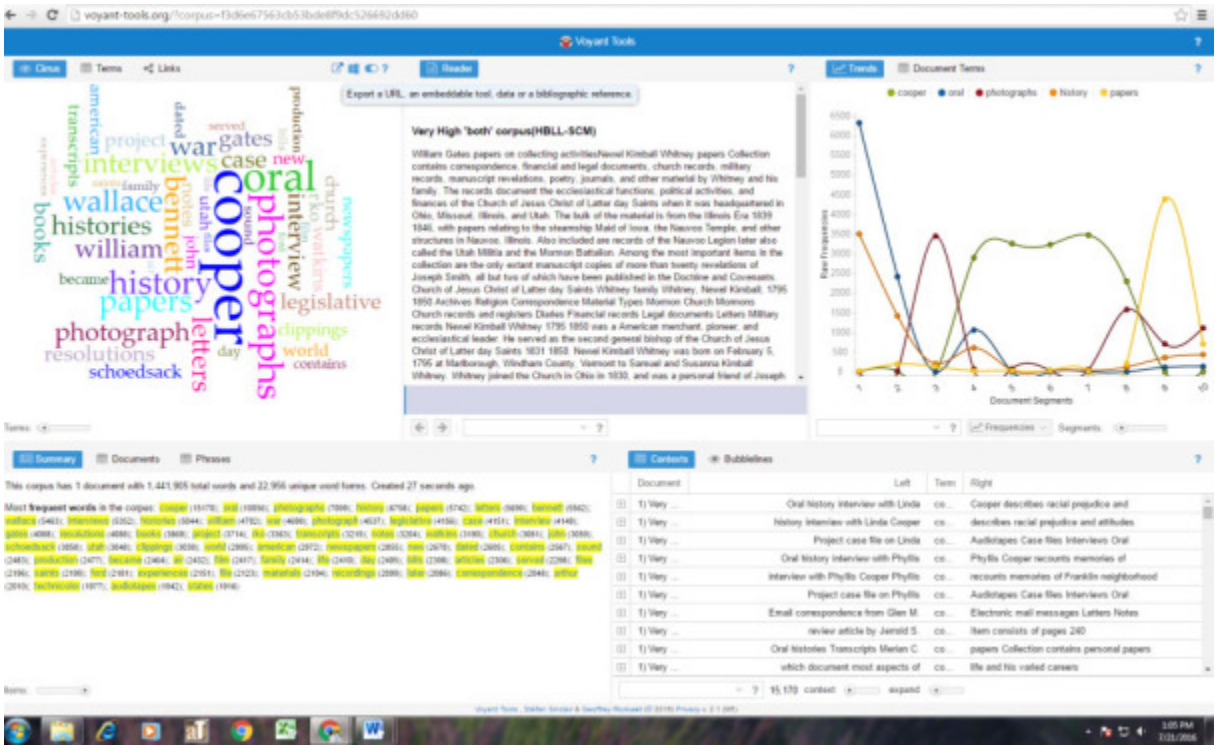

Figure 3. Screenshot of uploaded Voyant corpus

The word cloud in the top left corner held the information we wanted, so we exported the word cloud to isolate it, producing the screen shown in figure 4.

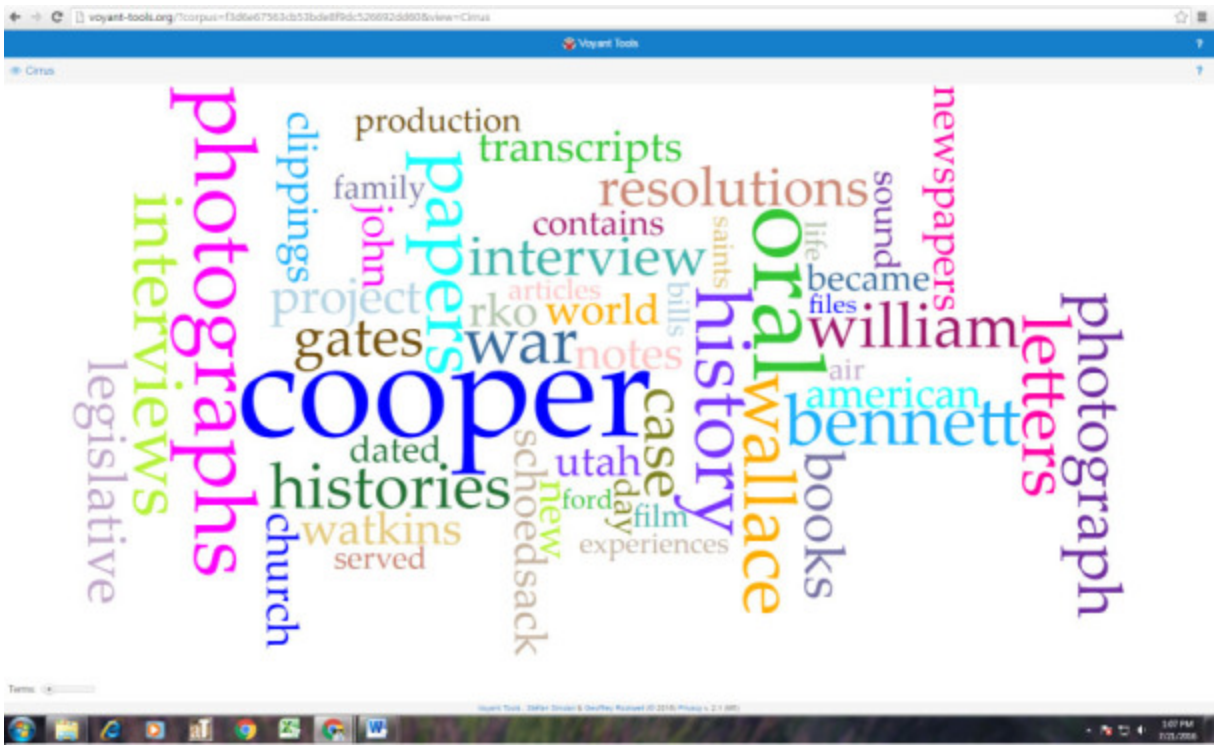

Figure 4. Voyant word cloud for sample of "Very High" use collections

As you can see in the resulting word cloud, the largest word (meaning it was the most frequent) is "cooper." This came from the repetition of a biographical summary on each page in a very large collection (Merian C. Cooper papers). Since repeating the biographical note in each component is not our normal standard for description, we realized this one finding aid was skewing our data, and we needed to narrow our corpora. We also realized that if a sample in any category included an unusually large finding aid that repeated titles with the same term at each level, we would get a similar result. We determined that if we were going to accurately portray patterns of potential search terms in each category, we should focus only on the subject and genre terms of each collection, since these would be more balanced in terms of quantity across collections, thus hopefully avoiding one large finding aid dominating the resultant word cloud.

So we repeated the process of creating a corpus, this time using only subject and genre terms stripped from each online page of the same collections. This was a much more manageable process, and with the time saved we were able to include the entire set of collections in the Very High 'both' usage group. Following the same process as outlined above, we plugged the new corpus into Voyant, and our resultant word cloud (figure 5) was much cleaner.

Figure 5. Voyant word cloud for "Very High" use collections, including only subject terms

Voyant allows you to include a variable number of terms in the word cloud in order of frequency. The word cloud shown here is a compilation of the top 55 words found in our "Very High" use collection. Though the name "cooper" remained, likely because this name is also included in subject terms in each component of the Merian C. Cooper papers, many other names were weeded out, leaving us with more of a focus on the subject terms, as we had hoped.

We next repeated this same process with each of the Low, Medium, and Moderately High 'both' groups, creating similar corpora and the same set of word clouds for each respective group. In order to remain consistent across each group, we used the same process each time to create a sample from which to extract the corpus. To do this, we used an online sample size calculator[8] to determine the number of collections for the sample, plugging in the population (the number of collections in each group) with the "confidence level" consistently at 95% and the "margin of error" consistently at 5%. This online tool would then tell us what size our sample needed to be in order to be accurate across the board. For example, when we plugged in a population of 4,317, the size of our "Low Use" group, it calculated that we needed 353 collections to make an accurate sampling of that group. Once we had this number for the sample size, we plugged the complete list of call numbers for the collections in that group into another online tool for creating randomized lists[9]. This tool reorganized the list of call numbers randomly, and we took the necessary collections for the sample (using the earlier example, 353 collections) from the top of the new list and created the corpus from them and their subject terms. These corpora were then plugged into Voyant to create the word clouds described above.

Analyzing these different corpora, we were able to deduce some things related to our collections that helped us answer a few of the questions we had from our previous research. One question we were interested in was if the low-use collections were made up of collections that did not fit our collecting policies for manuscripts, and if this was a factor in their low use. The premise was that if they did not fit our collecting policies, most of which related the topics of Mormonism (i.e., The Church of Jesus Christ of Latter-day Saints), Utah, and the American West, possibly patrons did not know we had them, since they did not think to look in our repository for such collections. What we found was that the majority of the collections do indeed fit our collecting policies. In every single one of the word clouds that we created, the words "Utah," "history," "church," and "states" were not only present but in the top 50 of every corpus. Among

the other words that were in all but one or two of the clouds were “saints,” “Mormon,” “Jesus,” “Christ,” and even “American.” All of these words are tied to the collecting policies of the majority of the manuscripts we collect, and serve to demonstrate that the collections are, at least on average, in line with these policies. This holds true across the various usage groups, regardless of how high- or low-use they are. The subject matter, then, does not seem to have a notable effect on the popularity of the various collections.

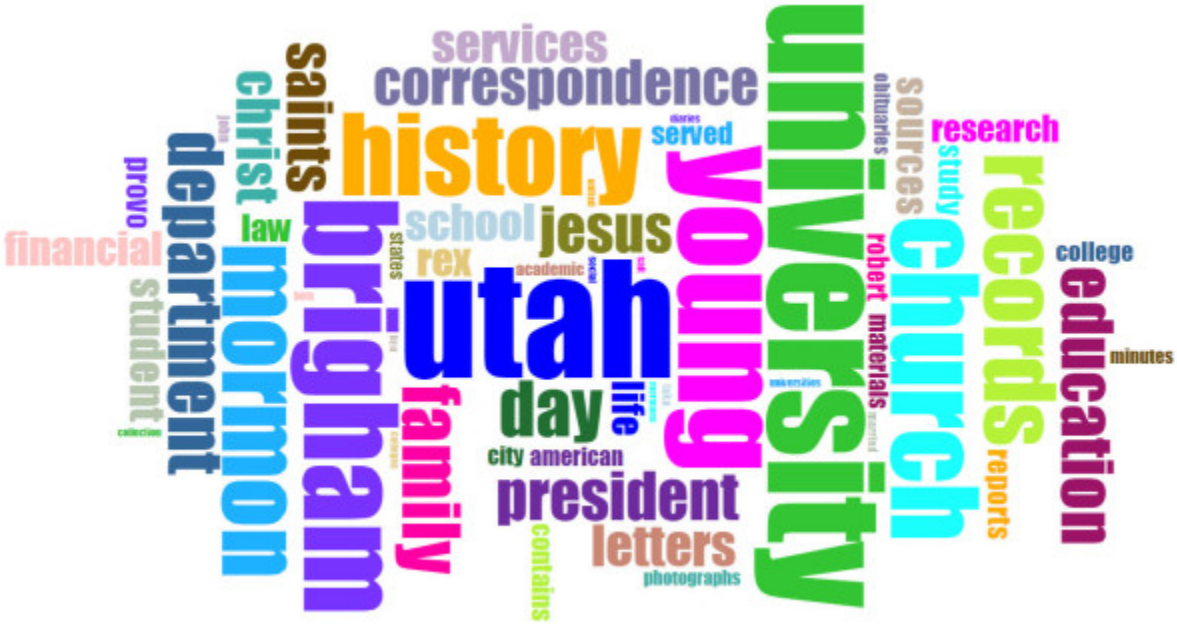

Figure 6. Voyant word cloud of subject terms in low-use collections

In looking more closely at the corpora for the low-use collections specifically (see figure 6), we did notice a few interesting things. First was the heavy presence of the terms in “Brigham Young University” The fact that these are low-use collections may or may not be surprising. University records are not normally very heavily researched by our patrons, and this is in part due to restrictions on many of them unless you are a member of the creating entity. We were a little surprised that so many still had terms related to the Mormon Church and Utah. One explanation for this is the fact that we have many small, single item collections in this area, from one letter to one or two

volumes of a journal or autobiography. It should be remembered that these are not "zero use" collections, but did have at least one use over two years. For items so specific to have one use is not necessarily a concern. What we cannot determine is how many of these represent larger collections that are getting very little use, which would be more concerning.

We took this analysis one step further and sampled low- and no-use collections from our collecting areas related to Mormonism[10]. Initially our plan for these samples was to use Voyant to create word clouds to compare with a word cloud of subjects found in recent scholarship related to Mormonism to see how the topics reflected in both groups matched up. However, the word clouds from the "Low Use" and "No Use" collections were so generic (with few specific subjects beyond general terms for the Mormon Church and history) that we determined the results would be inconclusive, and we abandoned this idea. Instead, we took these same samples and looked more specifically at the subjects of each collection to see how many, if any, within the samples did not fit within the scopes defined in the collecting policies, also known as "orphaned" collections, and if this was a potential factor in these collections having low or no use. The results showed that the number of "orphaned" collections in either group ("Low Use" and "No Use") was about the same in each group, with "Low Use" only having 22% orphaned and "No Use" only having 21%. While this is not a terribly high number, with one in five collections within each group being potentially "orphaned" and thus not part of a major collecting area, this could potentially be a factor for their low use.

The next issue we attempted to tackle was the effect of size on use – both size of the physical collection and size of the finding aid (i.e., level of description). In order to investigate this hypothesis, we took a sample of each of the different distinctive usage groupings (Very High, Moderately High, Medium, Low, and No Use) and found the average box size of the collections in each. We created a sample list of collections using the same tools mentioned previously for creating word clouds. The physical size

constantly increased in connection with the use of the group: “Very High” use collections had an average size of 38.93 linear feet; “Moderately High” had an average of 16.2 linear feet; “Medium” had 6.36 linear feet; “Low” had 4.4 linear feet; while collections with no use had an average of 1.16 linear feet (see table 3). So the parallel trends of patron use and size reaffirm that size is an important factor.

Grouping	Size (linear ft.)
No Use	1.16
“Low”	4.4
“Medium”	6.36
“Moderately High”	16.2
“Very High”	38.93

Table 3. Average size of collections per grouping

To further test this, we then looked at the size of the online finding aids within each of the categories. To do this, we first found the number of individual pages online that made up each finding aid in each sample. Our Finding Aid database displays each component of the finding aid as a unique page, so the lower the level of description (file and item level), the more components a finding aid will have, and thus more pages. Once we had the number of pages per finding aid, we then took the number of UPVs determined via Google Analytics from our previous research, and divided it by the number of individual pages. This gave us the UPVs per page, which allowed us to determine if there was any connection between online size and online use (see table 4).

Grouping	UPV per page
No Use	0
“Low”	5.06
“Medium”	12.92
“Moderately High”	23.1
“Very High”	42.21

Table 4. Average UPV per page of online finding aid, per grouping

For the “Very High” use collections, the average was 42.21 UPVs for each online page. For “Moderately High,” the average was 23.1 UPVs per page, while the average was 12.92 for “Medium” and 5.06 for “Low.” “No Use” was zero, having no UPVs and very little presence online. Again, the average uses per page paralleled the amount of use, just like with physical size. This again reaffirmed that the size of the finding aid, as well as the size of the collection, is a factor in predicting use of a collection.

The next issue we tackled with this analysis was determining if the criteria we used initially for determining high use (10+ circulations and 100+ UPVs over a two year period) was valid. This was especially a concern because 98% of our collections could not meet this standard, and we wondered if we set the bar too high. So, we wanted to see how many collections would fall short if we cut the factors in half, going from 10+ circulations and 100 + UPVs to 5+ circulations and 50+ UPVs. The result was only a 2% shift, where instead of 98% of the collections being below high-use, we now had 96%.

This showed us that wherever we set the bar, highly-used collections remained firmly in the minority.

For our final analysis, we went back to the quadrants from our initial research project (refer back to figure 1) to analyze some other aspects of the high and medium use collections. We wanted to compare finding aids within different quadrants, so again we turned to Voyant and word clouds. Word clouds from a sample of collections in each quadrant were made using the same process as had been used previously, focusing on subject terms. One initial finding when doing a comparison across all word clouds, regardless of quadrant and thus their usage statistics, was that the words "Utah," "history," "church," and "states" were not only present but in the top 50 of every word cloud. Among the other words that were in the top 50 of all of the clouds except one or two were "saints," "Mormon," "Jesus," "Christ," and even "American." All of these words are tied to the majority of Perry Special Collections' collecting policies, and serve to demonstrate that the collections are, at least on average, in line with them.

With word clouds from every quadrant, we next compared word clouds from finding aids in Quadrants I and II, along with an analysis of the level of description of finding aids within each quadrant. Quadrant I includes collections with the highest use collections both online and in-house, while Quadrant II consists of those with high in-house use but low online use. We hoped to determine if level of description and terminology explained why these collections fell into Quadrant II. The number of collections in these quadrants was small enough that we did not sample, so we looked at every collection. When examined, we found that Quadrant I had two collections that were described at only the collection level, one at the series level, one at the subseries level, three at item level, and nine at file level. In comparison, Quadrant II had twelve collections described only at the collection level, and four at series level. No collections in Quadrant II were described at the file or item level. This confirmed a correlation between level of description and online use.

However, it did not explain why the collections in Quadrant II had high in-house use. To determine this, we turned to the word clouds to compare subjects and terminology. We used the word cloud of the top 55 subjects already created for the Very High use collections, because this was essentially the same group of collections as Quadrant I. When comparing this with a word cloud of the top 55 subjects found in Quadrant II finding aids, again our findings were inconclusive (see figure 7).

Quadrant I subject word cloud

Quadrant II subject word cloud

Figure 7. Word clouds of subject terms found in collections in Quadrant I and Quadrant II

Quadrant I terms were primarily names and genre terms, like photographs and letters, basically showing that most of these finding aids were likely item-level descriptions, since our internal standard is to include names in titles and genre terms at item-level. Quadrant II terms were primarily general terms, such as “United States,” “Mormon,” “church,” and “history,” with “correspondence” being the major genre term. The presence of the term “20th century” potentially indicates that the majority of collections in this quadrant are from our 20th Century Mormon and Western American collection. The presence of names of U.S. Presidents like Warren Harding and Herbert

Hoover, as well as terms related to politics and government, may give some indication as to why they are highly-used collections in-house but not online, with many being some of our major political collections that patrons are aware of, but for which we do not yet have substantial online finding aids. These are usually large collections that warrant a lot of use, and are full of information for scholars of both Mormon and American history. But their size and complexity is often a deterrent for deep description because of the amount of resources required. While not entirely conclusive, this information was helpful in determining potential reasons for these collections falling on our graph as they did.

Our final step was to do a deeper analysis of Quadrant IV collections—those with high online but low in-house use. From our previous research, we knew that many, but not all, of the collections found in Quadrant IV had been digitized. In fact, forty-four of the seventy-nine collections (55.7%) had not yet been digitized, and six additional collections were only partially digitized. To see what might influence patrons to view these collections online, even though they did not include much or any digital content, we looked at each collection in this group that had did not have links to digital content to see if their finding aids contained other external links that could lead patrons either to or from some of its pages. We found that out of the forty-four collections without any links to digital content, ten contained a link to a legacy finding aid. So, thirty-four finding aids (43.0%) were still being used heavily online, although they did not have links to digital content or external finding aids. In comparison, 75% of Quadrant I, 82.4% of Quadrant II, and 20% of Quadrant III finding aids have links to digital content and/or external finding aids (see table 5).

Quadrant	%
I	75%
II	82.4%
III	20%
IV	57%

Table 5. Percentage of finding aids with external links, per quadrant

Going from links to level of description, of these thirty-four finding aids, fourteen (41.2%) were described at the file or item level; seven were described at the series or subseries level; eight were single items; and, five were described only at the collection level. Even among those finding aids with external links to legacy finding aids but no digital content, five are described at the series or subseries level, and five are at the file or item level. This makes for 43.2% of the finding aids that were not digitized having file or item level descriptions. This is further evidence that, while the existence of digital content and external links may be a factor in use online, level of description is even a greater factor.

Conclusions and Further Directions

While our initial research using Google Analytics provided valuable insights into the local and online usage of our repository's holdings, our analysis left us with a number of questions about both our own methodology and the meaning of the results. However, further analysis of the dataset demonstrated that the usage level thresholds established in the original study were reasonable. While this second round of review

expanded the testing categories, adding a level for collections of moderate usage, adjusting these numbers did not significantly impact the distribution and characteristics of results across the original quadrants.

The closer analysis of these categories did provide some additional insights into characteristics of the collections in each quadrant that might have affected their usage. On average, materials that were listed in Quadrant II (high circulation, low online use) were found to be described in lower detail than those in Quadrant I (high circulation, high online use), suggesting that they could see additional use if the finding aid were expanded to include lower hierarchical levels. The results in each of the quadrants showed a general correlation between the level of description and online usage of collections, though in-person use was reduced in cases where the materials had been digitized.

This relationship between descriptive level and usage was particularly clear for collections in Quadrant III. While these materials included a range of collections from the university archives that were subject to restrictions that might have reduced use, this quadrant was also found to have the least granular (or even missing) descriptions. In some instances this low descriptive level appears to have been the result of MPLP-based processing procedures, raising some question as to the impact of this approach to archival description on use.

Although these study results of Quadrant III descriptions were not encouraging, through the use of Voyant our text analysis found that the content of these collections did not vary significantly from those with higher use. This alignment with high-use collections, as well as with collection development policies, reinforced our conclusions regarding descriptive levels and suggested that the lack of use was not a result of out-of-scope collecting, but instead was a consequence of limited descriptions.

This study also indicated that descriptive standards have a significant impact on the discoverability of materials, particularly as collections are processed at deeper hierarchical levels. The provisions of Describing Archives: A Content Standard for devised titles heavily favor recording creator names and genre/form terms, weighting search results in these directions[11]. In our own study, this ultimately led us to exclude titles from our corpus despite the fact that this would remove subject terms when they were included. This effect was compounded with the repetition of creator information in single-level displays, where inherited information is displayed in each component description.

One method of reducing this repetition would be the implementation of authority-based, linked record approaches to display. In such systems, information about creators or scope notes for vocabulary terms is stored separately from the archival description, reducing duplication across the displayed record. At the same time, linked data systems improve the interoperability of descriptive data and potentially simplify reporting and analysis.

Google Analytics we were able to identify usage trends for online finding aids, allowing for segmentation and deeper analysis of collection characteristics. Voyant provided an additional means of examining the descriptions of archival records in our holdings to identify common themes and subjects. By recognizing these themes we will be better able to select similar content for digitization and deeper levels of description. This methodology might also be used for exploring researcher interests to better align collecting and description with user needs. While we were encouraged by these results of textual analysis, additional experimentation is needed to understand the impact of archival descriptive standards on the findability of materials for researcher groups.

NOTES

[1] Value-added processing in the L. Tom Perry Special Collections refers to processing at either the file or item level.

[2] Ryan K. Lee, Cory L. Nimer and J. Gordon Daines III, "Data-driven Decision Making at L. Tom Perry Special Collections," 2014 Society of American Archivists Research Forum Proceedings, accessed August 26, 2016, http://files.archivists.org/pubs/proceedings/ResearchForum/2014/papers/LeeNimerDaines-ResearchForumPaper2014.pdf

[3] Mark A. Greene and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing," American Archivist 68 (2005): 208-263.

[4] This is a term specific to Google Analytics, a tool which was used in our previous research. Google defines unique pageviews as "the number of sessions during which that page was viewed one or more times," as opposed to just a page view, which is "a view of a page on your site that is being tracked by the Analytics tracking code. If a user clicks reload after reaching the page, this is counted as an additional pageview. If a user navigates to a different page and then returns to the original page, a second pageview is recorded as well." "The difference between AdWords Clicks, and Sessions, Users, Entrances, Pageviews, and Unique Pageviews in Analytics," Google Analytics Help, accessed September 9, 2016, https://support.google.com/analytics/answer/1257084?hl=en#pageviews_vs_unique_views.

[5] https://voyant-tools.org/. Voyant was considered an ideal tool for textual analysis because of its ability to produce word clouds and other reports to show patterns of terms within our finding aids. It would allow us to quickly see what terms were most common within each group of finding aids.

[6] The "corpus" is a collection of text that you provide Voyant to read and analyze, in a simple ASCII text format.

[7] This is due to the fact that our Finding Aid database (https://findingaid.lib.byu.edu/) displays each component of the finding aid as a unique page. So, if a collection is described at the file or item level, it will have more pages, since it will have more components. See the following for more information on the design of our database: J. Gordon Daines III and Cory L. Nimer, "Re-imagining Archival Display: Creating User-Friendly Finding Aids," Journal of Archival Organization, 9, no. 1 (2011): 4-31.

[8] Found at https://www.surveymonkey.com/mp/sample-size-calculator/.

[9] Found at https://www.random.org/lists/?mode=advanced.

[10] This includes 19th, 20th, and 21st Century manuscripts, Mormon literary/Mormon author collections, and Photograph collections.

[11] Society of American Archivists. Describing Archives: A Content Standard, 2nd ed. (Chicago: Society of American Archivists, 2013), 17-23.