Data-driven Decision Making at L. Tom Perry Special Collections

RYAN K. LEE, CORY L. NIMER, J. GORDON DAINES III[1]

Brigham Young University

Abstract: Archivists are looking for new ways to identify materials to digitize in ways that meet the needs of their patrons. This case study examines efforts to improve digitization selection choices in the L. Tom Perry Special Collections at Brigham Young University. It describes a pilot project done in the department to utilize data-driven digitization and describes the sources of data used to drive those decisions.

Introduction

The concept of patron-driven digitization is not new to archives and special collections. However, strategies for making data-driven, objective decisions for digitization based on patron usage have been slow to emerge from the archival profession. Precisely what data to collect and how to analyze it are questions the profession continues to grapple with,[2] and options for data sources are continually increasing. Some institutions focus primarily on circulation statistics, which may or may not accurately reflect similar use by online audiences, where the digitized content resides. Studies of Web analytics, including Google Analytics, have emerged in recent literature, but have either focused on analysis of statistics to merely improve website interfaces, or on using one or two data points to aid in determining potential candidates for digitization. Many suggest the importance of coupling Web analytics with circulation statistics, but a study combining these different data sets has yet to be published. The following is a case study of how the L. Tom Perry Special Collections at Brigham Young University has attempted to bring both Web analytics and in-house use statistics together to show how more informed, data-driven decisions can be made.

Problem Statement

While recent literature has begun to address the issue of using data to drive decisions in archives and special collections, the focus seems to be primarily on online users and neglects in-house usage data. This study hopes to provide an example of combining both Web analytics and in-house use statistics to make more informed digitization decisions, and potentially other decisions in special collections processes.

A growing number of researchers are using Web analytic data to inform various decisions, including website design and usability;[3] increasing traffic and visibility of digitized materials;[4] and, prioritizing

[1] We would like to acknowledge the tireless efforts of Sam Dearden, our student research assistant, for all he did in compiling and helping to analyze the statistical information, and in helping create some of the graphs necessary for this project.

[2] For examples, see the Resource list at the end of this article.

[3] Christopher J. Prom, "Using Web Analytics to Improve Online Access to Archival Resources," *American Archivist* 74, no. 1 (2011): 158-84.

[4] Michael Szajewski, "Using Google Analytics Data to Expand Discovery and Use of Digital Archival Content," *Practical Technology for Archives* 1 (2013). Accessed July 8, 2014. http://practicaltechnologyforarchives.org/issue1_szajewski/

Published by Society of American Archivists, November 2015.

mass digitization.[5] Mark Custer specifically mentions the usefulness of Web analytics in making digitization decisions and in analyzing the impact of these decisions in his study of unique page views (UPVs) at East Carolina University.[6] This study follows Custer's justification for using Web analytics of finding aid use as a valid metric for determining collections to digitize, that:

> "…if an online finding aid has already attracted and sustained an audience, some portion of those digitized materials, if accessible from the finding aid, should receive the same level of visibility once digitized, if not more."[7]

While Web analytics are a valid metric for analyzing potential online use of digitized materials, there is evidence that coupling this with in-house use is beneficial to making the best, most accurate decisions for digitization. The same studies that focus on Web analytics in decision making also indicate that they are analyzing in-house use statistics in this process, but do not suggest how, or do not include this as part of their research results. Michael Szajewski actually gives greater weight to measuring online finding aid use when determining digitization, since "the user base for in-person patrons and online users varies greatly" and online users will likely be the ones using these digital surrogates. By studying online finding aid use, he suggests archivists might "discover hard data to measure the demand for digitized archival assets."[8] In contrast, Custer states that "gathering in-house reading room statistics in a consistent, unobtrusive manner can be correlated with online use" and mentions that reading room statistics were gathered and analyzed as part of his study, but were not included in the published report.[9] Christopher Prom gives the greatest credence to using the two data sets together by stating that while Web analytics do not substitute consulting in-house users, when used together and with other methods of studying user behavior, "it can force us to ask new questions about users and their information seeking behaviors."[10] We follow the same theory as Prom in our study, that there is benefit to analyzing the use of both audiences when determining digitization decisions.

This study was undertaken, of course, realizing there are limitations to digitizing primarily based on patron demand or use. Citing these dangers, Charles Cullen stated, "In deciding on what to digitize, relying only on past use in a special collections library is wrong or impossible or foolish." He also mentioned that many archivists or special collections librarians can cite examples of items in their or other collections that were "hidden" for years but were used by a scholar in new ways, leading to new insights and interpretations of the item.[11] Some have observed that digitizing solely based on patron demand can lead to an assortment of images that do not adequately represent the diversity of collections held by a special collections library and archive.[12] It is recognized that factors beyond usage may also be considered when making digitization decisions, including conservation issues.[13] While such things are important to consider when digitizing our collections, one should also consider the following point made

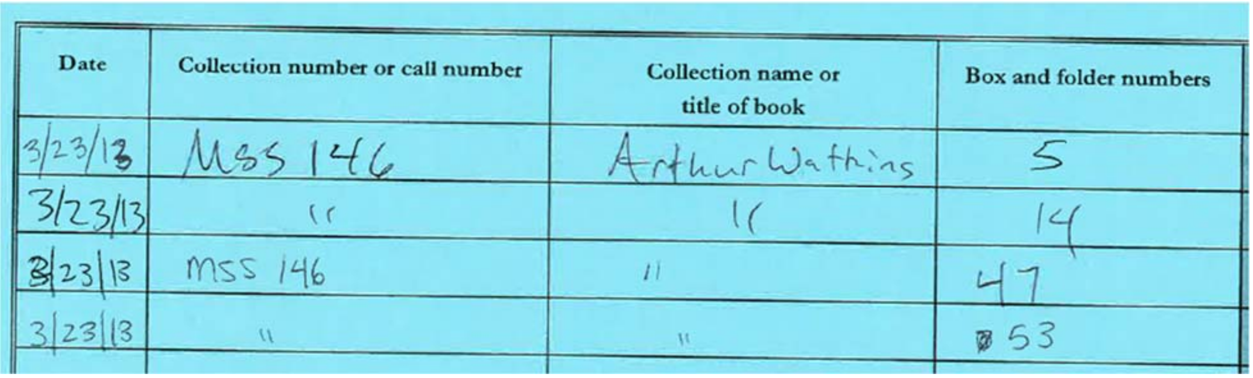
[5] Mark Custer, "Mass Representation Defined: A Study of Page Views at East Carolina University," *American Archivist* 76, no. 2 (2013): 481-501.
[6] Ibid., 483.
[7] Ibid., 489.
[8] Szajewski, "Using Google Analytics Data."
[9] Custer, "Mass Representation Defined," 489.
[10] Prom, "Using Web Analytics," 163.
[11] Charles T. Cullen, "Special collections libraries in the Digital Age: a scholarly perspective,*" Journal of Library Administration* 35, no. 3 (2001): 86.
[12] Elizabeth A. Novara, "Digitization and researcher demand: Digital imaging workflows at the University of Maryland Libraries." *OCLC Systems & Services* 26, no. 3 (2010): 174.
[13] In fact, L. Tom Perry Special Collections is in the process of creating a digitization threshold policy where conservation reasons are a consideration for determining if and when an item may be a candidate for digitization.

by Janet Gertz: "Digitizing and mounting materials publicly on the internet is a form of publishing, and success in publishing means knowing and targeting viewers."[14] Thus it is important to think first of users of library and archival materials when making digitization decisions. Of course, archival materials and their digital descriptions and surrogates may have high usage or views for a variety of reasons. Those that are not as highly used may be poorly described, and, thus not readily discoverable, but still worthy of digitization. Thus it behooves all administrators of special collections and archives to also think of ways that they can improve access and discoverability to ALL materials, and have a strategy for bringing those lesser known and used items to light through better description or improved discovery systems, or even better marketing of these "hidden collections" through social media or other avenues**.**

Methodology

To perform this study, three sources of statistical data were analyzed: two that related to the in-person use of archival and manuscript collections in the Perry Special Collections reading room and one related to the use of finding aids for archival and manuscript collections made available online through the department's finding aids database (http://findingaid.lib.byu.edu). The aim of the study was to compare the 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.

The first data point revealed the circulation patterns of collections in the Perry Special Collections reading room and is the combination of information from two data sources. The first data source was circulation information extracted from statistics manually compiled by reading room staff. Each time a patron requests a manuscript or archival collection for use in the reading room, reading room staff note the collection call number as well as the box number on a tracking form, known as "blue sheets." The data from tracking forms for a two-year period were manually entered into an Excel spreadsheet. While these sheets go back about a decade, for this project only sheets back to June 2012 and up to May 2014 were used, in order to match the parameters of the Web analytics data available to us. Then all of the entries for one call number were combined into a single entry to indicate total circulations. This was necessary because often circulations on the tracking form were noted as, for example, MSS 146 Box 5 and MSS 146 Box 14 (see Figure 1), which would appear to be two separate circulations. Since this represented requests for the same collection on the same day, however, it was counted only as one circulation.



Date	Collection number or call number	Collection name or title of book	Box and folder numbers
3/23/13	MSS 146	Arthur Watkins	5
3/23/13	"	"	14
3/23/13	MSS 146	"	47
3/23/13	"	"	53

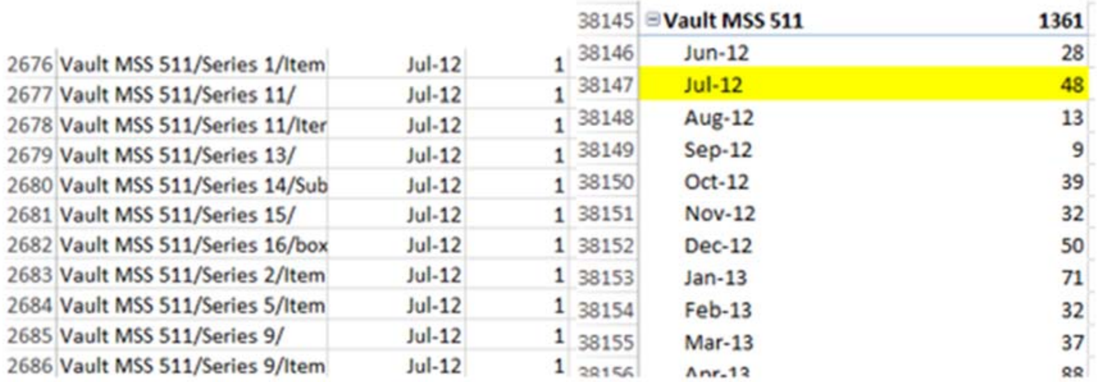
Figure 1. "Blue sheet" showing multiple entries for same collection on same day.

[14] Janet Gertz, "Should You? May You? Can You?: Factors in Selecting Rare Books and Special Collections for Digitization," *Computers in Libraries* 33, no. 2 (2013): (8)

Sorting in Excel indicated these entries as two separate circulating collections rather than as two circulations of one collection. Statistics were counted at the collection level, per researcher, per day. So if a researcher requested multiple boxes from the same collection on the same day, this was counted as one circulation for that collection. Requests for the same boxes for the same researcher the next day would count as another circulation, and so on. This was not always clear-cut and required some further analysis and consolidation, making it a very labor intensive step.

The second data source was circulation information pulled from the library catalog, which came in the form of an automated report from our cataloging system indicating items checked out in the Special Collections reading room. However, only collections with a barcode can be checked out through the library catalog and not all archival and manuscript collections have been barcoded. To get the most accurate count of items that were truly circulated in the reading room, the circulation data from the library catalog was compared with the data from the tracking forms and the two data sources were consolidated into a single data point representing reading room circulation.

The second data point revealed usage patterns of online finding aids for manuscript collections in Perry Special Collections. This information was gathered using Google Analytics generated for the same two-year period (June 2012-May 2014) as the information for the reading room circulation data point. To get the most accurate count, it was decided to focus on Unique Page Views (UPVs), as Custer did in his study.[15] Similar problems were encountered with the Google Analytics data as with the reading room circulation tracking forms because the finding aids are delivered as single-level (collection, series, file, item, etc.) descriptions, and so the Web analytics are delivered at the same level. This may result in two page views for one finding aid having differing URLs that made it appear as if they were separate collections. An Excel spreadsheet was used to consolidate all of the UPVs for one collection together so that it represented how many times a finding aid for one collection was viewed each month (see Figure 2).[16]



2676	Vault MSS 511/Series 1/Item	Jul-12	1
2677	Vault MSS 511/Series 11/	Jul-12	1
2678	Vault MSS 511/Series 11/Iter	Jul-12	1
2679	Vault MSS 511/Series 13/	Jul-12	1
2680	Vault MSS 511/Series 14/Sub	Jul-12	1
2681	Vault MSS 511/Series 15/	Jul-12	1
2682	Vault MSS 511/Series 16/box	Jul-12	1
2683	Vault MSS 511/Series 2/Item	Jul-12	1
2684	Vault MSS 511/Series 5/Item	Jul-12	1
2685	Vault MSS 511/Series 9/	Jul-12	1
2686	Vault MSS 511/Series 9/Item	Jul-12	1

38145	⊟Vault MSS 511	1361
38146	Jun-12	28
38147	Jul-12	48
38148	Aug-12	13
38149	Sep-12	9
38150	Oct-12	39
38151	Nov-12	32
38152	Dec-12	50
38153	Jan-13	71
38154	Feb-13	32
38155	Mar-13	37
38156	Apr-13	88

Figure 2. Comparison of expanded list of Unique Page Views and consolidated list

[15] Custer, "Mass Representation Defined," 481-482.

[16] UPVs for finding aids accessed internally were not filtered out of the results for this study. Many of these hits would have been generated by reference staff or curators in assisting patrons, and thus could still be considered relevant to this study. However, it is acknowledged that spikes in hits in some finding aids may be the result of staff in the digital lab accessing the finding aid in preparation for linking digital images. While it may be possible to filter out these hits, this was not done for this study.

The ultimate aim of the data gathering efforts was to create two data points that could be used to help make better decisions on what archival or manuscript collections to select for digitization. As described, the first two sources of data were combined to generate a data point representing the circulation of manuscript and archival materials in the reading room. This data was then compared with the Web analytics data point to see if there was any correlation between reading room usage and online usage of finding aids. Analyzing and comparing these data points revealed some very interesting and informative things pertaining to the usage of our collections.

Results

In total, 6,483 individual collections were analyzed. This represents the amount of collections that either had at least one UPV or one reading room use in the past two years. Of these, 5,917 had at least one UPV. Reviewing the statistics from Google Analytics, the collections were divided into five equal groups (quintiles) by the number of total UPVs. Doing so revealed a distribution similar to that reported by Custer,[17] with nearly 85 percent of our traffic over the two year period coming from the first 20 percent of the collections (see Figure 3). While seeing this concentration of online finding aid usage suggested that the first quintile collections should be considered for digitization, we also wanted to correlate the data with our circulation statistics to verify public interest in these materials.

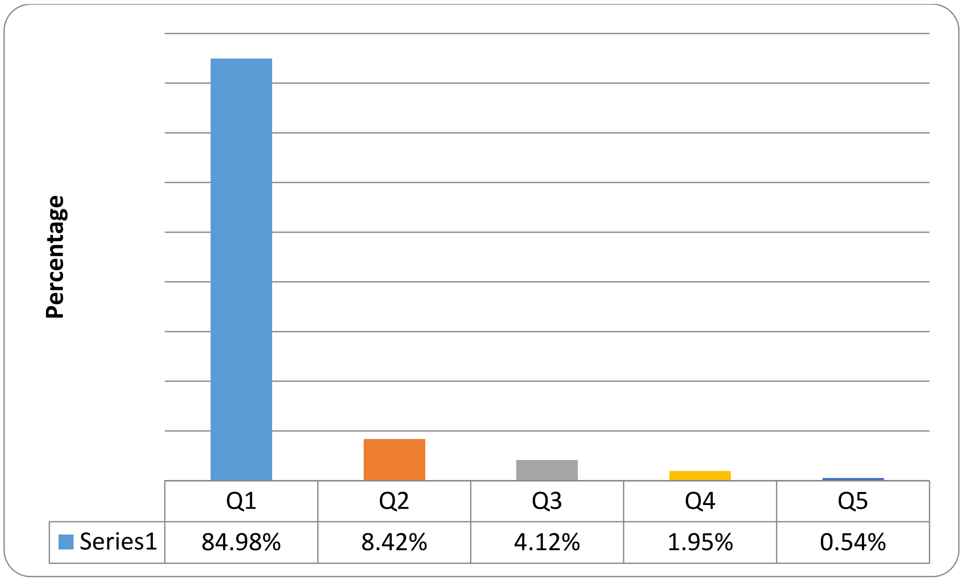
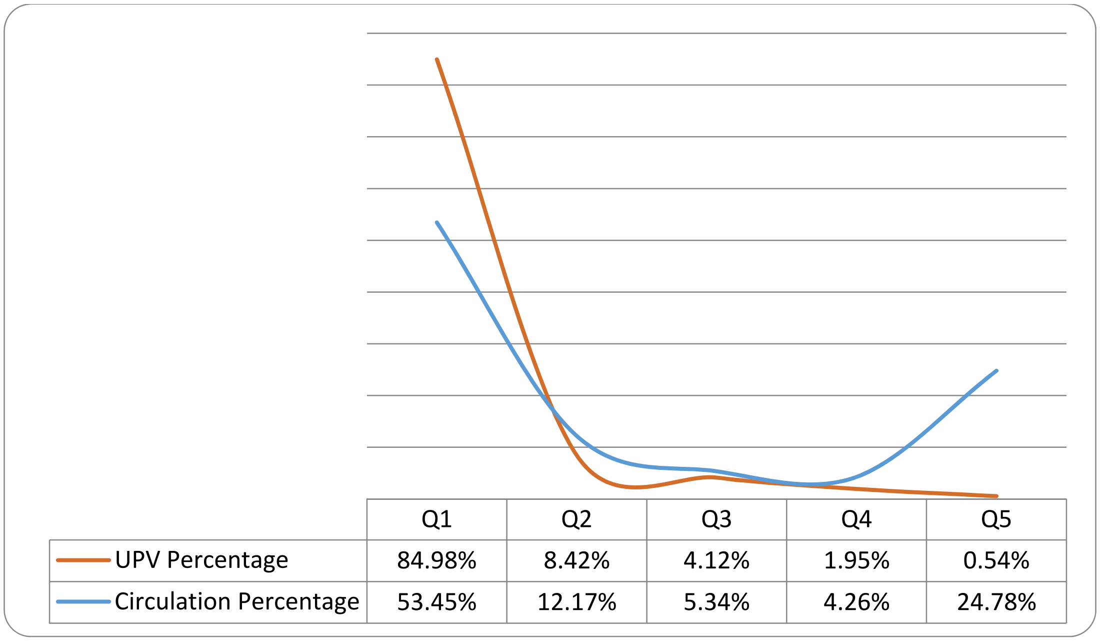


Figure 3. Distribution of Unique Page Views (UPV)

Yet in reviewing the aggregated circulation data for the collections that composed the UPV quintiles, it was found that online finding aid use did not appear to be a good predictor of in-person use of the materials (see Figure 4). This was particularly true for collections in Quintile 5 (Q5), which accounted for only 0.54 percent of the UPVs but nearly a quarter of the total material circulations.

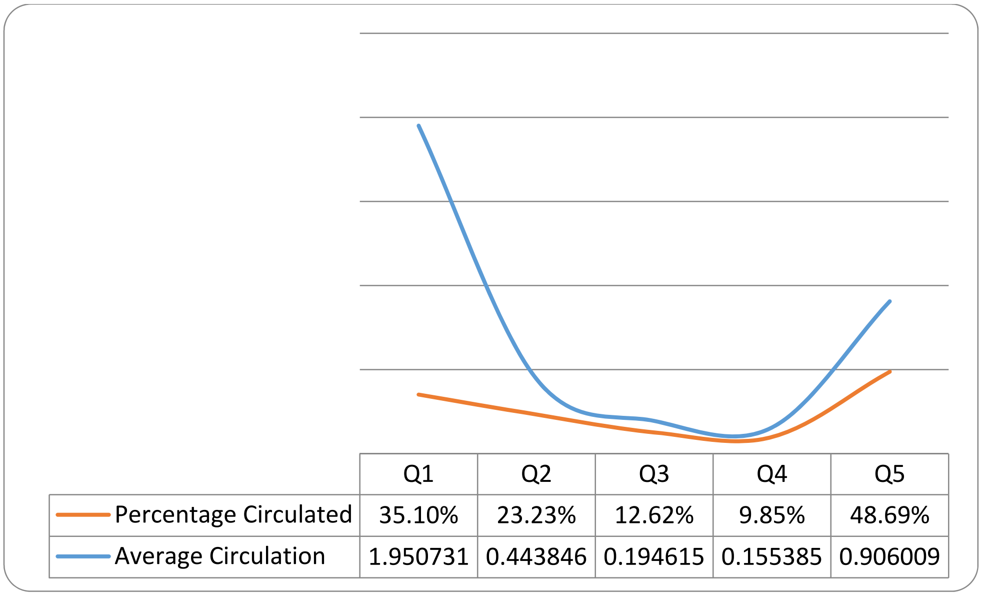
[17] Ibid., 482-83.



	Q1	Q2	Q3	Q4	Q5
UPV Percentage	84.98%	8.42%	4.12%	1.95%	0.54%
Circulation Percentage	53.45%	12.17%	5.34%	4.26%	24.78%

Figure 4. UPV/Circulation Comparison

This divergence between UPV and circulation rates was even more exaggerated when looking at absolute circulation numbers within the quintiles (see Figure 5). In the first quintile, for example, collections had on average circulated approximately twice, though only 35 percent of the collections had been circulated at all. The fifth quintile, on the other hand, had circulated on average once, with nearly 49 percent of its collection accessed in the reading room.



	Q1	Q2	Q3	Q4	Q5
Percentage Circulated	35.10%	23.23%	12.62%	9.85%	48.69%
Average Circulation	1.950731	0.443846	0.194615	0.155385	0.906009

Figure 5. Circulation rates per UPV quintile

These divergences reinforced the fact that in-house use might provide a useful secondary criteria for determining the suitability of materials for digitization. In order to facilitate the review of the data, a visualization of the two criteria was created. After removing some outliers, including collections with more than 1,000 unique Web views, the remaining collections were charted as a scatter plot (see Figure 6). Digitization work that had been completed or was underway was also reviewed, and those collections with available digital content were marked in red in the scatter plot chart, to see where these collections in particular ended up on the graph.

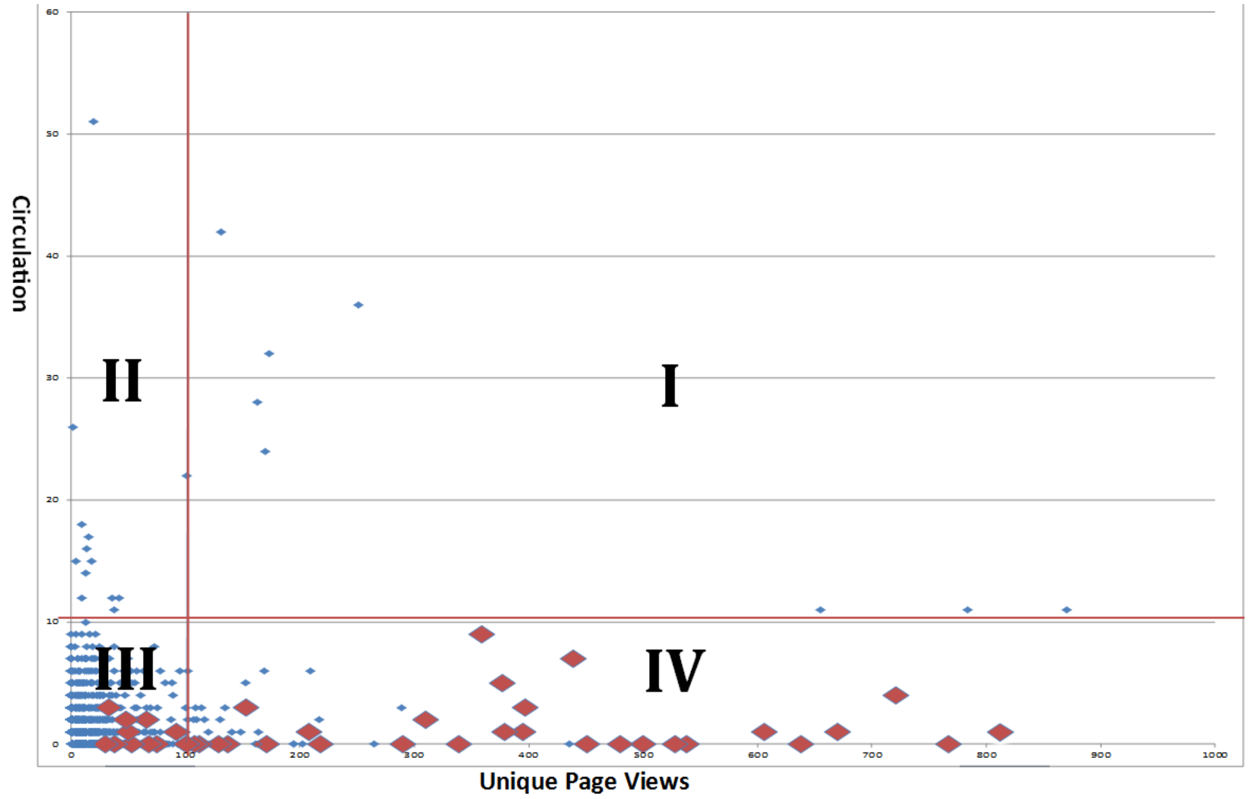


Figure 6. Unique Page View/Circulation comparison

Examining the visualization, the materials were divided into four quadrants—placing minimal criteria for potential digitization at 100 unique Web views, and at ten circulations. This placed 98 percent (6,373) of collections in Quadrant III (low Web visibility and low circulation), with a short list of materials in each of the other quadrants (Quadrant I: 15; Quadrant II: 16; and, Quadrant IV: 79), with Quadrant I including the outliers. Reviewing the content of these quadrants, it was found that most of the digitized content currently available through our finding aids database was from collections in either Quadrant IV (collections with high Web visibility and low circulation) or in Quadrant III (low Web visibility and low circulation). It was also discovered that, excluding the outliers, no materials from the collections in Quadrant I (high Web visibility and high circulation) had been digitized and made available to our patrons online.

Findings

Taken separately, using either Unique Page Views or circulation statistics could be seen as reasonable metrics for recommending digitization of archival materials. Page views provide a sense of general interest, while circulation statistics suggest personal engagement with the materials themselves. Using these metrics in conjunction to rank materials provides a more accurate sense of the usefulness of the collections. However, when comparing UPVs with circulation (as in Figures 4 and 5), page views do not necessarily predict of in-person use.

Looking at the finding aids available for materials in Quadrant I, many of these collections had item-level calendars available for their contents. This also seemed to be true for materials with a high number of page views, as given in Quadrant IV. Materials in Quadrant II, on the other hand, tended to have only collection or series-level descriptions.

Having identified the collections in Quadrant I as being of greater interest, it then becomes incumbent to determine what this comparison means for the accessibility of materials in other areas:

- 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 finding aids for materials that are not accessed in person, as charted in Quadrant IV, and what role does prior digitization have on these statistics?
- What needs to be done to improve access to the materials in Quadrant III, which currently are not being accessed either remotely or in the reading room?

Conclusion

The last question is probably the most profound result of this study, since Quadrant III represents the fact that 98 percent of collections analyzed were considered low use (less than ten circulations and less than one hundred Unique Page Views over a two-year period). This raises some serious questions to consider, and will require another study to come up with answers:

- What is the cause of this low use? Is it lack of interest in the content, or is this simply a result of poor description?
- Could this be an effect of applying MPLP principles, and should this approach be reconsidered?
- Were the parameters set to determine high use too high?
- Do the 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?
- What are possible solutions to moving more of these collections into other quadrants? Deeper description, or even deaccessioning?

These questions are definitely worth asking, although the answers may be difficult for some to admit. Answering these questions would require a deeper analysis of the collections, their content and descriptions, as well as the current audience for these materials. Yet doing such an analysis of collections and their use would help improve many areas of an archival institution, including collection development and arrangement and description practices, resulting in the institution better serving the needs of their patrons, including library and university administrators.

Beyond answering these questions, there are other interesting factors that can be derived from this study. First, further work must be done to make gathering and analyzing statistics less labor-intensive. One goal of this study was to be able to provide curators with data that can help them make digitization decisions. While this was an eventual result, the process for getting this result was not conducive to being repeated often. It is recommended that more tools be developed that would allow archivists and special collections librarians to gather and analyze both circulation and Web analytical data in a more efficient and systematic matter.

Furthermore, one interesting, and possibly unexpected, point this study has brought to light is the impact that digitization may have on reference services in special collections. The findings of this study reflect much of what Peter Hirtle suggests would happen as we digitize more and more of our special collections materials, when he stated that “[e]lectronic access will replace most uses of printed, paper copies, [and]… [t]he use of paper originals will decrease."[18] Through analyzing and comparing online finding aid use and in-house use, it was found that digitization has often significantly reduced the use of originals in Perry Special Collections. This is likely primarily due to an internal policy of pointing patrons to the digitized version when they request the item in the reading room or contact the library via email or phone. However, it is something to seriously consider and continue to track as it may impact future decisions and directions for the library, or special collections in general.

While this impact is likely similar in other special collections libraries, administrators should not fear digitization, but instead should embrace it as the way of the future. Hirtle suggests that some impacts may bring into question the very existence of special collections, but then implores that "[r]ather than bemoaning the fact that people are happy using surrogates, [special collections librarians] should accept it." After all, special collections can still contribute by doing more to "emphasize those elements in their holdings that are truly unique."[19] Special collections libraries all have materials that are important to the advancement of knowledge and information, and most would agree that digitization is one way to provide increased access to this unique information. This study provides one example of how digitization programs can be improved and more relevant to all patrons, both online and in-house, by taking into account use statistics.

Resources

Cullen, Charles T. "Special collections libraries in the Digital Age: a scholarly perspective." *Journal of Library Administration* 35, no. 3 (2001): 79-91.

Custer, Mark. "Mass Representation Defined: A Study of Page Views at East Carolina University." *American Archivist* 76, no. 2 (2013): 481-501.

Gertz, Janet. "Should You? May You? Can You?: Factors in Selecting Rare Books and Special Collections for Digitization." *Computers in Libraries* 33, no. 2 (2013): 6-11.

Hirtle, Peter B. "The Impact of Digitization on Special Collections in Libraries." *Libraries & Culture* 37, no. 1 (2002): 42-52.

Novara, Elizabeth A. "Digitization and researcher demand: Digital imaging workflows at the University of Maryland Libraries." *OCLC Systems & Services* 26, no. 3 (2010): 166-76.

Prom, Christopher J. "Using Web Analytics to Improve Online Access to Archival Resources." *American Archivist* 74, no. 1 (2011): 158-84.

Szajewski, Michael. "Using Google Analytics Data to Expand Discovery and Use of Digital Archival Content." *Practical Technology for Archives* 1 (2013). Accessed July 8, 2014. http://practicaltechnologyforarchives.org/issue1_szajewski/

[18] Hirtle, Peter B. "The Impact of Digitization on Special Collections in Libraries," *Libraries & Culture* 37, no. 1 (2002): 45-46.
[19] Ibid., 49.