Mass Representation Defined: A Study of Unique Page Views at East Carolina University[1]

Mark Custer

ABSTRACT

The following case study examines three years' worth of data gathered by East Carolina University's Special Collections Division. The analysis focuses on the number of Unique Page Views (UPVs) reported by Google Analytics. A straightforward method to analyze this data is proposed, which ranks the frequency of UPVs received per year for each collection-level web page. The results indicate that the overall distribution of UPVs is heavily skewed toward the top 20%—or, top quintile—of online finding aids, which account for over 70% of the views. The term *mass representation* is introduced to refer to the disproportionate amount of visibility attained by the top quintile.

KEY WORDS

Encoded Archival Description (EAD), Management, Metadata, Technology

George K. Zipf claimed that he could distill the findings from his book, *National Unity and Disunity: The Nation as a Bio-Social Organism*, into a single sentence: "In union there is strength, and in numbers there is strength; hence strength for the attainment of any objective lies in the organization of numbers."[2] Although certain of its veracity, Zipf qualified his summation by noting that even though unity can be utilized to achieve objectives, it would be fallacious to conclude that "in numbers and in organization there is *truth.*"[3] In as much as this case study also focuses on the organization of numbers, it will not attempt to prove, on its own, that statistical laws govern the use of archival collections, whether that use is conducted online, in an on-site reading room, or at a microfilm reader an ocean away.[4] Instead, this study serves to emphasize 1) the ease with which collection-use data can be analyzed, once the given metrics have been stipulated, so that archivists can make data-driven prioritizations, and 2) the need for analogous data sets to be collected, evaluated, and openly shared.

Over the years, numerous studies have been conducted on the use of archives, but generally these focused on users.[5] In these studies, archivists often acquire data sets about their users by means of questionnaires and reference logs. Consequently, few user-based studies contain quantifiable data about the comprehensive use of archival collections.[6] Of those that do, it is likely that no one has published the complete data sets that were the focus of their study.

An alternative method for studying the use of archives is to measure collection use, not users, which is a subtle but important distinction.[7] For instance, numerous collection-use studies have been undertaken within the field of librarianship; likely, this is due to that profession's more immediate and historical need to track the circulation of its materials. These types of studies were published frequently during the 1960s when the concept of a "no-growth library"[8] first originated, which was itself a direct result of the circulation studies made famous by Richard Trueswell.[9]

Trueswell's name is invoked in the OCLC report, "Shifting Gears: Gearing Up to Get into the Flow," which includes tips on how to scale and speed up the digitization of special collection materials. The authors assert, for example:

> [O]nce you've got all your collections represented on the web, you now have a basis upon which to determine where to apply further effort. Examining use is a great way to learn about researchers' needs. Trueswell's 80/20 observation suggests that 20% of library materials will satisfy 80% of requests. Iterate once you've identified that 20%.[10]

The following study can be read as an attempt to answer that challenge. Any repository with online finding aids can follow the procedure presented here to identify the top 20% of its collections.

In the context of this study, it is also necessary to note that Richard Trueswell relied upon a single metric, the Last Circulation Date (or LCD), to demonstrate the 80/20 rule within the field of library science.[11] Even though this study focuses on the use of online finding aids, primarily through the lens of a single metric, the same mode of inquiry could be applied to any manifestation of archives use. Only the data collected and the choice of metrics would need to change.

Collecting Data and Selecting a Metric

On June 9, 2008, the Digital Collections Department at East Carolina University (ECU) installed Google Analytics on Joyner Library's EAD-encoded finding aids.[12] Google Analytics is an easy-to-install, proprietary, and free software suite that reports a variety of predefined web metrics.[13] It uses JavaScript and first-party cookies to send information to Google servers, which then analyze the incoming web traffic. This software can be installed on any website by adding a few lines of JavaScript to each page. Because of this, Google Analytics and similar products such as Piwik[14] are sometimes referred to as "page tagging" analytics, in contrast to the web server logfile analysis approach provided by software packages like AWStats.[15] An excellent introduction to web analytics addressed to the archival profession can be found in Christopher Prom's paper, "Using Web Analytics to Improve Online Access to Archival Resources."[16]

Whereas Prom focused primarily on how repositories can use Google Analytics to improve their websites, an important objective in its own right, this case study avoids that topic altogether. Instead, this study extracts, explores, organizes, and interprets data from Google Analytics to answer questions that web analytics software is not intended to answer directly. For example, it provides archivists with a way to use web statistics to identify and prioritize collections for mass digitization. And, once collections are digitized, it helps archivists gauge, with the assistance of web statistics, the impact of their decisions.

The majority of this study focuses on a single metric, the unique page view (UPV), even though Google Analytics automatically reports over 30 different dimensions and metrics for any standard setup.[17] The UPV metric was chosen over the more generic page view metric because it represents a closer approximation of how Joyner Library gathers collection-use data in its reading room. William Jackson followed a similar approach in an earlier study of archives use; in it, he counted multiple uses of one collection—by a single user over the course of one day—as a single use.[18] Somewhat analogously, the UPV metric reported by Google Analytics registers only one UPV whenever a single web browser views a web page, regardless of how often that page is refreshed or revisited within a 30-minute timeframe. The value reported by the page view metric, on the other

hand, would continue to increase with each web page request.[19] In other words, when a user visits a URL 10 times within a 30-minute period, Google Analytics reports a total of 10 page views, but only one UPV.

Grouping Data

To begin the analysis of ECU's collection-level pages, 3 data sets were exported from Google Analytics' Top Content report:

- FY2008: 07/01/2008–06/30/2009
- FY2009: 07/01/2009–06/30/2010
- FY2010: 07/01/2010—06/30/2011

During this timeframe, every visit to the EAD website, including staff member use, was tracked.[20] Though staff-only web traffic cannot be isolated after it has been collected by Google Analytics, one year's worth of data was separated between on-campus and off-campus users by means of the network domain dimension.[21] The on-campus use of online finding aids for that year garnered nearly 31% of the total share, but the patterns of that use were nevertheless nearly identical between on-campus, off-campus, and overall use.[22] Therefore, the remainder of this study considers only the overall use of the EAD website, regardless of the type of user.

For the majority of UPVs collected during this study, the EAD website delivered two types of page views for each collection-level web page:

1. ***PV-1; or static page views.*** These account for the majority of UPVs, since most of the traffic (73.23%, over the 3-year period) is driven by search engines, which provide direct links to ECU's finding aids.
2. ***PV-2; or site-search page views.*** These views register a distinct URL that includes the search terms used to access a collection from within the website's internal information retrieval system.

Two additional types of page views were recorded after a website redesign in April of 2010:

3. ***PV-3; or notes page views***, represented by the "add/view" notes tab included on each finding aid, allow users to add comments or ask questions about a collection.
4. ***PV-4; or digitized-object page views***, represented by the "view digitized objects" tab, are present only on those finding aids with associated digitized materials.

The final type of page view was recorded after February of 2011:

5. ***PV-5; or request page views***, represented by the "request materials" tab, enable patrons and staff to request materials from a collection for on-site use.

On the one hand, this final type of page view will make future analyses even more valuable by connecting ECU's online finding aids directly to its reading room request system. On the other hand, and as a result of counting the last three types of page views, the frequency of UPVs within ECU's data sets are likely to become more biased toward those online finding aids that receive a higher degree of interaction.[23]

A collection-level web page was defined as any URL that contained a collection-level identifier that was delivered by the EAD website to a web browser. Three real-world examples, extracted from the FY2010 data set, are presented in Table 1.

Table 1. A Sample of UPV Data Reported by Google Analytics for a Single Collection

URL	UPV
/special/ead/findingaids/0001/	153
/special/ead/view.aspx?id=0001&show=request	33
/special/ead/view.aspx?id=0001&q=Civil+War	3

The three URLs presented in Table 1 contain the same collection-level identifier, "0001." These URLs were then normalized to their corresponding EAD filename; and, although only done in the case of the FY2010 data set, Table 2 demonstrates how the type of page view can also be retained.

Table 2. A Sample of UPV Data Reported by Google Analytics for a Single Collection

EAD ID	Page View Type	UPV
0001.XML	PV 1 (static)	153
0001.XML	PV 5 (request)	33
0001.XML	PV 2 (site-search)	3

Any rows of data not corresponding to an online finding aid were then removed.[24]

At this stage, repeated observations for the same collection-level pages were still present. The next step was to put the data into Microsoft Excel, or a similar tool, to take advantage of its pivot table functionality. Using a pivot table, all of the repeated entries for the same collection-level pages were merged together, allowing each collection that received at least one UPV to be represented in a single row of data. Continuing with this example, after all of the observations for collection 0001.XML were tallied, the final row of data from the FY2010 data set was reported, as seen in Table 3.

Table 3. A Complete Report of UPV Data from FY2010 for a Single Collection

EAD ID	Total	PV 1	PV 2	PV 3	PV 4	PV 5	On campus	Off campus
0001.XML	408	254	64	17	50	23	176	232

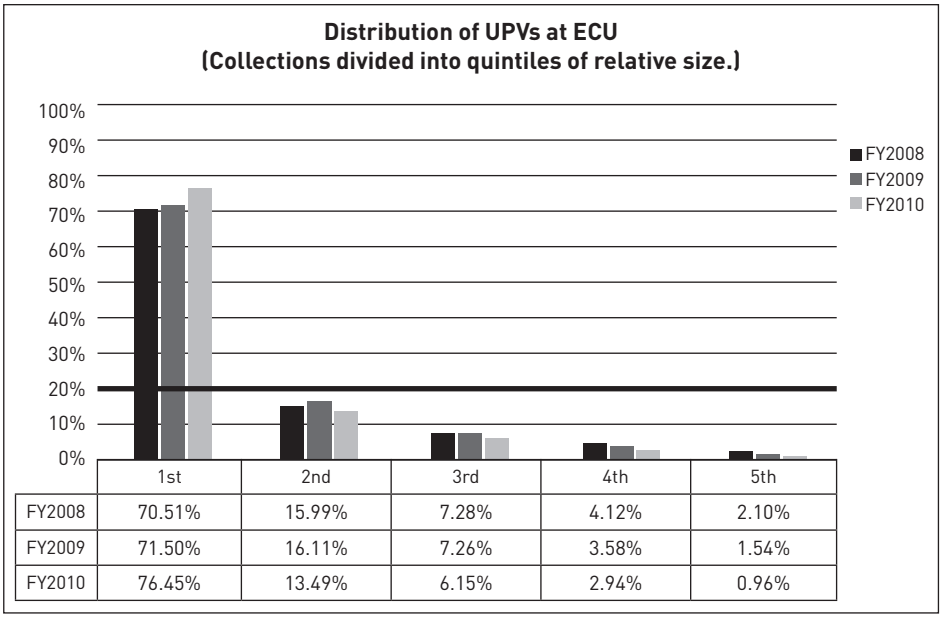
Lastly, one additional step, taken within each data set, determined whether there were any finding aids that received zero UPVs.

Though this process might seem time consuming, most of it is automated; in fact, one year's worth of data can be exported, analyzed, and verified in less than 20 minutes. Additionally, if both the data collection and presentation processes are tied to the delivery software, then the entire process may be automated and even displayed on the same website. For example, special packages created for delivery frameworks—like the eXtensible Text Framework (XTF), EADitor, Hydrangea, and more—may take advantage of an application programming interface (API) provided by Google Analytics or Piwik to report use data automatically, to revolutionize and standardize how those data sets are collected and made interoperable.

Defining Mass Representation and Its Boundaries

After the data have been exported and properly normalized, they can be analyzed in a variety of illuminating ways. To visualize the uneven distribution of UPVs received throughout ECU's EAD website, for example, one useful approach is to rank the collections according to the frequency of UPVs received and then to divide that range into 5 equal groups, or quintiles, so that each group contains approximately 20% of the total number of collection-level pages available (see Appendix A for more details).

Figure 1 presents a quintile bar chart for all of the online use data analyzed in this study. It is readily apparent that the top quintile of finding aids—represented here by the first series of columns—receives a disproportionate amount of online visibility, year after year. The 20% line on the y-axis is emphasized to indicate a metaphorical watermark to which each quintile would ascend (or from which it would cease ascending, in the case of the first quintile) if the distributions were truly equal. But these distributions are clearly not equal. In fact, after the first quintile, each group receives significantly less coverage than the previous group. Therefore, the term "mass representation" refers to any percentage of values within a set range that far exceed the set range's proportionate share of coverage. In these examples, the top quintile is synonymous with both the quantitative value as well as the qualitative idea of its overall



	1st	2nd	3rd	4th	5th
FY2008	70.51%	15.99%	7.28%	4.12%	2.10%
FY2009	71.50%	16.11%	7.26%	3.58%	1.54%
FY2010	76.45%	13.49%	6.15%	2.94%	0.96%

FIGURE 1. Distribution of Unique Page Views. This illustrates the uneven amount of visibility attained over a 3-year period by those collections in the first quintile (i.e., the top quintile). Each quintile contains one-fifth of the total number of collections, after those collections were ranked according to the frequency of UPVs received. See Appendix A for more detail.

representation. Table 4 details mass representation as it is defined by the top quintile. In this definition, the percentage of UPVs identified is determined by a set percentage of collections (20%).

Focusing on the top 20% as a value might seem arbitrary. Or, it might even seem prohibitive, especially if a repository has limited funding or staffing. To provide additional perspectives of mass representation—which are as equally valid as the first—consider again that all of the collections have been ranked according to the frequency of UPVs received by their finding aids. Within the entirety of this ranked list, mass representation is definable in three ways: 1) the top quintile, which has already been illustrated; 2) the bare majority; and 3) the population mean. Mass representation by bare majority includes the least number of collections, in ranked order, that it takes to accumulate just over 50% of the total UPVs. In other words, it comprises the fewest number of

Table 4. Mass Representation by Top Quintile

	FY2008	FY2009	FY2010
Percentage of Highest-Ranked Collections	20%	20%	20%
Percentage of UPVs Received by Those Same Collections	70.51%	71.50%	76.45%

highest-ranked collections that receive the bare majority of UPVs. In this definition, which is illustrated in Table 5, the percentage of collections identified is determined by a set percentage of UPVs (>50%).

Table 5. Mass Representation by Bare Majority

	FY2008	FY2009	FY2010
Percentage of Highest-Ranked Collections	8.80%	8.31%	5.77%
Percentage of UPVs Received by Those Same Collections	50.07%	50.06%	50.09%

Mass representation by population mean, on the other hand, is the subset of collections that have a relative frequency of UPVs greater than the inverse of the total number of collections, where the total number of collections includes any collections that received zero UPVs. This definition comprises every collection that has received more than the average number of UPVs. This definition, which is illustrated in Table 6, is slightly different than the first two since neither the percentage of collections nor the percentage of UPVs will be a set figure.

Table 6. Mass Representation by Population Mean

	FY2008	FY2009	FY2010
Percentage of Highest-Ranked Collections	24.87%	24.64%	20.94%
Percentage of UPVs Received by Those Same Collections	76.04%	76.78%	77.39%

Depending on a repository's unique needs and objectives, any one of these definitions can be utilized. Oftentimes, though, little may separate the collections identified by each definition; in fact, all three definitions can potentially identify the exact same subset of collections.

These results also demonstrate that the concept of mass representation cannot be thought of as a simplistic, fixed value. Not only will interests wane and the zeitgeist change—the latter of which, arguably, drives the majority of potential use, via search engine traffic[25]—but if an acquisition program is successful, then new collections of high interest, once added, would jostle for position. With that in mind, one can begin to track the evolution of the top quintile. For instance:

- At the conclusion of FY2009, 82% of ECU's finding aids were present in the top quintile for both FY2008 and FY2009. By selection, though,

this percentage excludes 45 finding aids (2.5%) that first went online during FY2009.

- At the conclusion of FY2010, 76% of ECU's finding aids were present in the top quintile for both FY2009 and FY2010. Similarly, this figure excludes 47 finding aids (2.5%) that first went online in FY2010.
- At the conclusion of FY2010, 68% of ECU's finding aids were present in the top quintile for all 3 fiscal years. This figure excludes those 92 finding aids (5%) that were published during FY2009–FY2010.

It is useful to watch this particular trend, and others that emerge from collection use, to use the data to make intelligent, even creative, management decisions.

Additionally, gathering in-house reading room statistics in a consistent, unobtrusive manner can be correlated with online use. Toward this end, ECU also analyzed reading room statistics from FY2008 through FY2009. The reading room data are not reported here (although data capture improved significantly in February 2011, due to the installation of Aeon by Atlas Systems), but some interesting statistics were discovered in this early attempt to correlate on-site and online use. For example, 78% of the collections (282 of 361) in the top quintile for FY2008–FY2009 were also requested in the reading room during that same time period.

To understand additional implications of this statistic, consider the following, seemingly radical, suggestion that emerged from the Digitization Matters symposium[26] and the follow-up paper, entitled "Shifting Gears": namely, that "[a]s materials (whether a single item or a boxful) are requested for reading room use, circulation, reproduction, or interlibrary loan, digitize them and make the digital versions available to everyone."[27] ECU adds materials to its digital repository whenever a patron specifically requests those materials for digitization. At the end of FY2009, an impressive 36% of collections within the top quintile (129/361) had at least one digitized object. Had ECU followed the advice in the "Shifting Gears" report during that same period, however, and scanned samples of any materials requested for reading room use, it would have ended FY2009 with an even more impressive 78% of highly visible collections (282/361) with at least one digitized object available online.

Of course, even if an entire archival collection becomes digitized, that alone does not ensure that it will find an audience.[28] But it seems likely 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.

Putting Data to Use

The point in collecting these data is to use them in an effort to improve services and the overall management of collections. With that in mind, East Carolina University has already leveraged its UPV data in three different ways.

First, since one of the Special Collections Division's goals was to digitize a few collections in their entirety, ECU needed a prioritized list of collections to consider for mass digitization. This process was initiated by looking at the top 50 collections, as determined by their UPV ranking. Any collections not requested for in-house use within the previous 2 years were cut from initial consideration. At the end of this process, ECU created a prioritized list of 42 collections, 5 of which were added later because it was agreed that their finding aids were not online long enough to make the initial cut. Second, ECU also created a list of its most-viewed collections that lack digitized materials. Staff members have consulted this list when selecting materials to digitize and feature in the department's blog.[29] Finally, Jonathan Dembo, special collections curator at ECU, has improved the descriptions for a subset of ECU's least-viewed collections, which he selected from the UPV-ranked list. Now, he can track and determine how their UPVs are affected relative to the time spent updating those descriptions.[30] For archivists to increase the overall visibility of an archival repository, it might prove more efficient for them to focus their attention on the least-viewed collections, rather than seek grants to process and digitize high-profile collections. More exploratory data analyses of use would be required, though, to be confident before proceeding with either assumption.

These data could be applied to many other objectives. As space is also a concern for many repositories, these same data sets could be analyzed to identify collections to move to an off-site storage facility, or even to provide support, when necessary, for deaccessioning collections.[31] Data can be collected and analyzed to help ensure the success of the objective, whatever it may be.

Toward a Multivariate Approach

Only one variable, the UPV, has been discussed in this study so far. Noah Huffman, archivist for metadata and encoding at Duke University, highlighted the importance of another variable: the amount of time that a user looks at an online finding aid.[32] In addition to analyzing the frequency with which users discover online finding aids, one could also analyze the overall duration that users view those resources. As Huffman demonstrated, it is useful to examine the duration of engagement—especially in combination with other dimensions reported by Google Analytics, such as the URL referral path—in an effort to identify meaningful discoveries.[33]

However, a critical caveat about web analytics must be understood: when a user leaves the domain of a website, there is no timestamp for the software to record. Therefore, the time recorded in those cases will be zero seconds, even if a user remains on a web page and reads it attentively for one hour.[34] For this reason, the "Avg. Time on Page" metric reported by Google Analytics cannot be understood outside of its relation to both the "Pageview" and "% Exit" metrics.

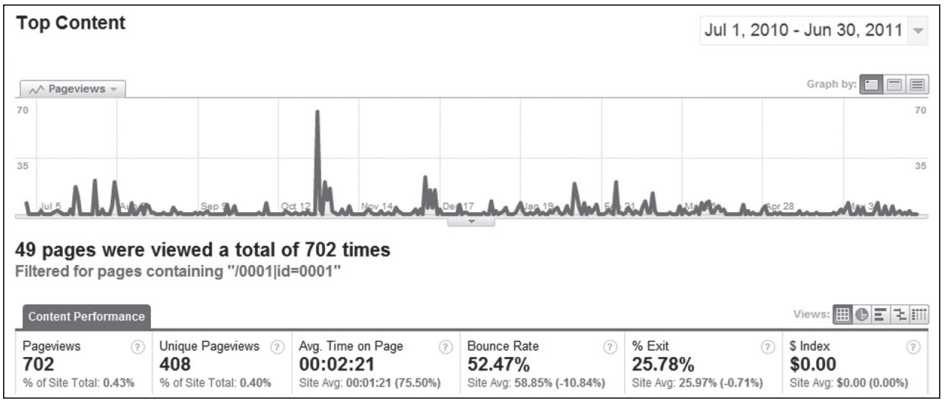


FIGURE 2. A summary of page views for a single collection, http://digital.lib.ecu.edu/special/ead/findingaids/0001/, during FY2010, presented by the Google Analytics user interface.

Figure 2, which contains one year's worth of metrics for a single collection-level page, illustrates this issue. The average time spent on this collection's pages is reported as 2 minutes and 21 seconds. However, since an exit percentage of 25.78% is also listed, it must be understood that this average time is not based on all 702 page views, but instead only on 521 page views.

New metrics can then be created by combining the default options provided by Google Analytics. For example, a metric for "estimated page view hours" (hereafter abbreviated as EPVH) can be created by multiplying the total number of page views per year by the average time spent on that collection-level page. Using the values from Figure 2, the EPVH for this collection is equal to 702 page views multiplied by 02:21, or roughly 27.5 EPVHs.[35] Additionally, a confidence level of 74.22% (the inverse of the exit percentage) could be applied to this particular measurement.[36] Not only did this finding aid receive 408 UPVs, then, it was viewed for an estimated total time of 27.5 hours that year, during which time Google Analytics recorded just over 20 hours of actual use.

Figure 3, which plots UPVs vs. EPVHs for FY2010, provides an example of how these two metrics can be correlated. Although the overall distribution of this data does not appear to be overwhelmingly linear, a trend line can nevertheless be added to isolate and examine a small number of collections in more detail. Regarding any outliers significantly above the trend line, a repository

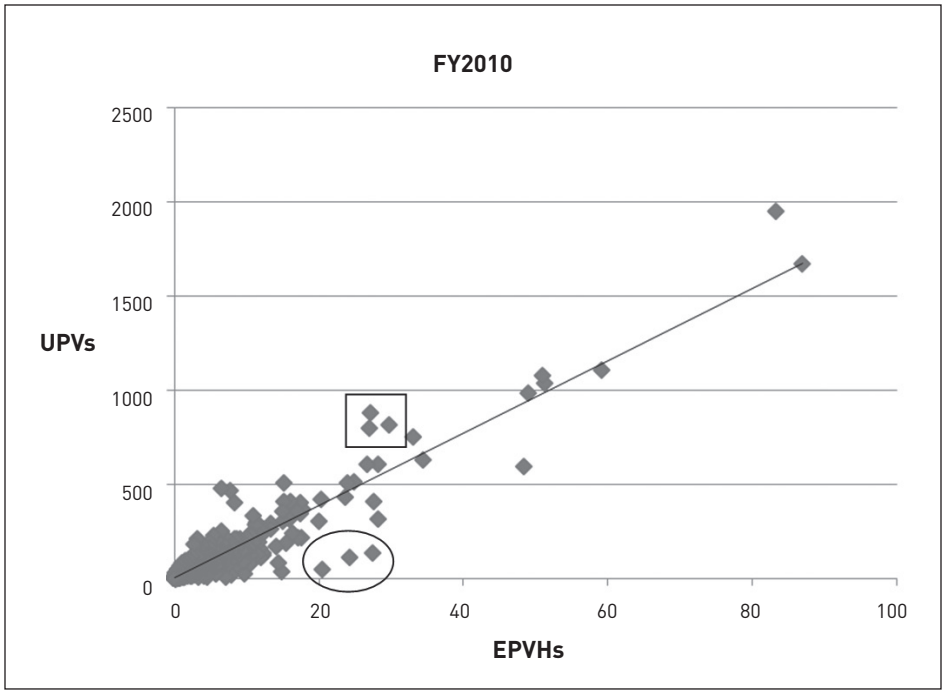


FIGURE 3. Estimated Page View Hours versus Unique Page Views for each collection-level web page.

might consider those collections as prime candidates for mass digitization. Consider, for example, the three finding aids within the rectangle in Figure 3. When compared to their peers, as determined by the correlation of UPVs and EPVHs, these finding aids should have received larger EPVH values. Based on past levels of online use, it is possible that embedding digitized content within these collection-level pages is the best means to increase their EPVHs. To assess the impact of this decision, an archivist could then compare the before and after values of the EPVH metric, especially in relationship to its movement along this graph for the next fiscal year. Conversely, collections below this trend line might make better candidates for finding aid revisions, or even reprocessing, since they received fewer UPVs than expected when compared to their peers. Therefore, it is probable that the three collections circled in Figure 3 do not have descriptions that adequately convey their scope. And so, it would be less wise to embed digitized content within these finding aids without first revisiting their descriptions. Both cases would need to be investigated more thoroughly, of course, but the point is that these two metrics, which summarize levels of use in the past, can be consulted to provide guidance in the present.

Conclusion

"We are in the computerized information area for the long pull, and wherever possible we should like to make it a professional effort instead of just an institutional project."[37]

—James B. Rhoads

James B. Rhoads, fifth archivist of the United States, made this remark in a speech delivered at the annual luncheon meeting of the Society of American Archivists and the Organization of American Historians on April 17, 1969. At the beginning of the talk, Rhoads referenced a paper that John Hope Franklin had delivered a few months earlier. In so doing, he rearranged the title of Franklin's paper, "Archival Odyssey: Taking Students to the Sources," admitting that he playfully considered titling his own paper "Archival Oddities: Taking Sources to the Students."

This inversion of subject and object—of researchers and collections—was more than just a playful act, however; it mirrored Rhoads's belief that historical research was changing and that the computer was acting as the agent of this inversion. And it was also the computer, Rhoads asserted, that would enable the archival profession to deliver more knowledge about its collections to the public than ever before.

In his talk, Rhoads detailed three areas in which archivists should use computers. Those areas are presented in the same order here, which is important, since each function builds upon the previous one. Rhoads believed the computer should be used to:

1. Manage collections in an inventory system
2. Analyze research use
3. Become a "cybernetic extension" of the researcher

Following this first suggestion, this entire case study would have been impossible if ECU did not already have its collection information in an inventory system. Without a system to apply the data (ECU's EAD database) and another to store them (in this case, Google Analytics), there would be nothing computable to analyze.

When Rhoads suggested that archives analyze research use, he proposed an experiment in which every finding aid from a repository was put "into a single data bank, and that the data bank be queried every time a question is asked about the records."[38] Jackson carried out a similar experiment in his study published in 1997.[39] This case study analyzes research use, by examining the visibility of online finding aids, from yet another perspective. To capitalize on these and similar studies, however, additional steps need to be taken to assemble, interpret, and—most importantly—to integrate the results back into our system of practice.

It is for this reason, this need, that Rhoads invoked "cybernetics" in his third suggestion about how archivists should use computers. Cybernetics, a scientific discipline that emerged not too long before the time of Rhoads's talk, can be defined as the practice of studying a system holistically and continuously to influence, or steer, that system's evolution toward optimum efficiency.[40] Consequently, it is also easy to see the influence of cybernetics in Trueswell's work[41] and the concept of no-growth libraries.

If archivists were to follow the lead set forth by cyberneticists, Rhoads mused that we "may in fact see the end of printed or published guides to or inventories of records."[42] Relatively static finding aids—like the online finding aids examined in this study—are produced within a system in which the bulk of information moves in one direction, from the archives outward. The cybernetic approach, alternatively, would provide an uninterrupted feedback loop between archivist and researcher. Rhoads provided one example for how this might manifest when he concluded that researchers should be able to retrieve and format archival information on their own terms to meet their individualized research needs.[43]

Because the results presented here, however, cannot be used—on their own, at least—to accomplish such an objective, this case study concludes with three recommendations. Together, these three recommendations provide a foundation upon which the cybernetic approach that Rhoads described might be realized.

First, the archival profession needs to specify quantitative-use metrics, specifically web metrics, to analyze research use.[44] This study offers UPVs and EPVHs as two metrics that can be used to analyze trends in the online use of archival finding aids and the digitized surrogates that they describe.

Second, we need to openly distribute the data that we collect, as long as they cannot be used to violate user privacy.[45] All of the data sets used in this study have been uploaded to East Carolina University's institutional repository to ensure that these data are freely available to anyone with Internet access.[46] Not only does this step help ensure that the results from this study are reproducible and verifiable, it also provides data that can be reused for additional studies. Furthermore, data sets collected with Google Analytics can easily be shared beyond institutional borders. Since these data are stored externally by Google, read-only access can be granted to anyone who has a Google account. At the time of this writing, for example, colleagues at three other institutions have access to all of the Google Analytics' reports from the Joyner Library Collection Guides websites.

Finally, we need to share and distribute any code utilized in the course of analyzing archives-use data. This study has not yet shared any of the code used to explore and analyze its data, which is a definite shortcoming. The process by

which the study was conducted, however, made that a difficult task, especially due to the unique architecture of the home-grown EAD website. Even if the pre-analysis stage of data preparation remains unique for many repositories due to the differing structures of their websites, it is still possible for the archival community to distribute and share packages of code for statistical analyses.[47] Though not accomplished with this project, future projects have an opportunity to follow the procedure presented here, improving and expanding its focus in the process.

This case study, then, which admittedly originated as an institutional project, should also be read as a modest attempt to transform archival collection-use studies into a collaborative and professional effort. Whatever our professional objectives—and regardless of whether they are as lofty as the intentions once espoused by cyberneticists—aggregating, analyzing, and sharing data sets in a unified effort might be the best means that we have to attain them.

Appendix A

This table summarizes the data presented in Figure 1. It also includes information for the combined data set of all three fiscal years, which is labeled as FY2008–FY2010:

	Number of Online Finding Aids	Average Number of Collections per Quintile	Number of Finding Aids with Zero Views	Maximum UPV	Minimum UPV	Total UPVs
FY2008	1,761	352	0	1,344	1	74,521
FY2009	1,806	361	6	1,816	0	75,351
FY2010	1,853	371	58	1,951	0	70,196
FY2008–FY2010	1,853	371	0	4,729	2	220,068

- Since the total number of finding aids cannot be evenly divided into quintiles (i.e., by 5), the average number of collections per quintile is reported in column 2. The few remainders present in these data sets, however, are evenly distributed throughout the quintiles. For clarity, that distribution is as follows:

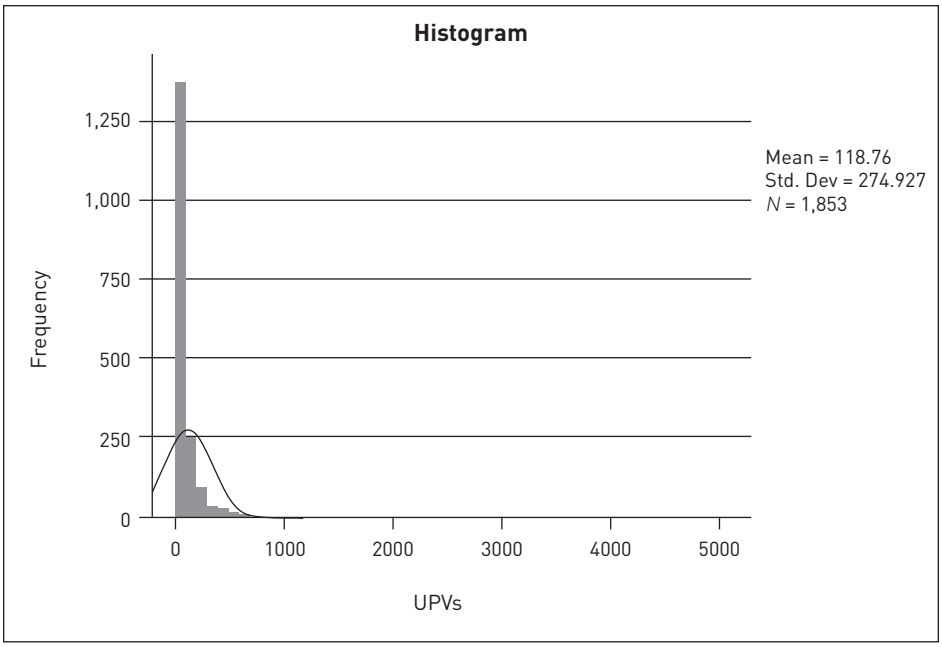
Number of Finding Aids per Quintile

Quintile	FY2008	FY2009	FY2010
1	352	361	*371
2	352	361	370
3	*353	*362	*371
4	352	361	370
5	352	361	*371

* The asterisks denote the presence of the remainders.

- Trend to watch: will the number of "zero-view" pages continue to increase as new finding aids are added each year?

Appendix B



- FY2008–FY2010, with UPVs binned by 100. Column 1 contains the number of finding aids with 1–100 UPVs; column 2 contains finding aids with 101–200 UPVs, etc.
- The data set is skewed to the right very heavily (i.e., it has a positive skew, such that most of the observations are less than the mean).
- The overlay curve represents a normal distribution. The distribution within the histogram is far from normal, however, due to the high number of finding aids with a small number of UPVs.
- About 73% of the finding aids (1,344/1,853) occur in the first column. The top quintile, on the other hand, occurs within 45% of the second column (113/251) plus all of the remaining columns to the right.

Notes

[1] This case study is an expansion of my paper, "Incorporating Patron Requests into Archival Workflows and Digital Repository Interfaces" (presented at the annual meeting of the Society of American Archivists, Austin, Texas, August 11–16, 2009), http://saa.archivists.org/Scripts/4Disapi.dll/4DCGI/events/eventdetail.html?Action=Events_Detail&InvID_W=1089.

[2] George K. Zipf, *National Unity and Disunity: The Nation as a Bio-Social Organism* (Bloomington, Ind.: Principia Press, 1941), 404.

[3] Zipf, *National Unity and Disunity*, 392–93.

[4] The inspiration to rank online finding aids based on the frequency of their use, as I have done throughout this case study, was based on Zipf's law, specifically as it is detailed in his book, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (1949; reprint, New York: Hafner Publishing Company, 1965). Zipf formulated this law, originally, in an effort to explain the distribution of word frequencies in textual communications, which correlates well with how search engines function, by matching queries to documents, directing traffic to online finding aids in the process. The data sets presented in this study cannot always be arranged into harmonic series, however, as faithful adherence to Zipf's law would require; instead, the bulk of observations present in the long tails of data appear to follow lognormal distributions. Nevertheless, one would need to analyze more data before drawing conclusions or attempting to formalize probability distributions on the use of archives.

[5] For a thorough literature review on archives-use studies, see Wendy M. Duff et al., "The Development, Testing, and Evaluation of the Archival Metrics Toolkits," *The American Archivist* 73 (Fall/Winter 2010): 569–99. Other articles of interest include Bruce W. Dearstyne, "What Is the *Use* of Archives? A Challenge for the Profession," *The American Archivist* 50 (Winter 1987): 76–87; William J. Jackson, "The 80/20 Archives: A Study of Use and Its Implications," *Archival Issues* 22, no. 2 (1997): 133–46; Christian Dupont and Elizabeth Yakel, "'What's So Special about Special Collections?' Or, Assessing the Value Special Collections Bring to Academic Libraries" (paper presented at the Library Assessment Conference, Baltimore, Maryland, October 26, 2010). The conference proceedings, which include this paper, are available online, http://libraryassessment.org/bm~doc/proceedings-lac-2010.pdf.

[6] Two examples are Fredric Miller, "Use, Appraisal, and Research: A Case Study of Social History," *The American Archivist* 49 (Fall 1986): 371–92, which studied the academic use of archival materials by means of a citation analysis; and Joyce Chapman, "Special Collections Physical Materials Usage Patterns" (October 20, 2010), NCSU Libraries, http://www.lib.ncsu.edu/dli/projects/dataviz/visscrcphysical, which studied the overall use of archival materials in one repository's reading room.

[7] Paul Conway, "Facts and Frameworks: An Approach to Studying the Users of Archives," *The American Archivist* 49 (Fall 1986): 393–407, for example, not only provided a theoretical discussion about the need for and use of user studies, it also provided a concrete example on how to conduct such assessments. Even though Conway's framework provided some structure to record collection-use data, its primary focus is nevertheless on the researcher, and—even more important, in relation to the current case study—the data collection procedure that emerged from that framework requires the *active* involvement of the researcher as well as the archivist. The data collection procedure that will be presented in this case study, however, could best be described as *passive* involvement since it does not require the intervention of the researcher or archivist once it has been initiated (i.e., the only user involvement required is the act of his or her use).

[8] For an introduction to this concept, see the following anthology of essays: Daniel Gore, ed., *Farewell to Alexandria: Solutions to Space, Growth, and Performance Problems of Libraries* (Westport, Conn.: Greenwood Press, 1976).

[9] The literature concerning collection use within libraries is vast, but a few highlights include Herman H. Fussler and Julian L. Simon, *Patterns in the Use of Books in Large Research Libraries* (Chicago: University of Chicago Library, 1961), with a re-edited version published by the University of Chicago Press in 1969; Frederick G. Kilgour, "Recorded Use of Books in the Yale Medical Library," *American Documentation* 12 (October 1961): 266–69; Richard Trueswell, "A Quantitative Measure of User Circulation Requirements and Its Possible Effect on Stack Thinning and Multiple Copy

Determination," *American Documentation* 16 (January 1965); A. K. Jain, "Sampling and Short-Period Usage in the Purdue Library," *College and Research Libraries* 27 (May 1966): 211–18; Quentin L. Burrell and Violet R. Cane, "The Analysis of Library Data," *Journal of the Royal Statistical Society* 145, no. 4 (1982): 439–71.

[10] Ricky Erway and Jennifer Schaffner, "Shifting Gears: Gearing Up to Get into the Flow" (Dublin, Ohio: OCLC Research, 2007), 4–5, http://www.oclc.org/research/publications/library/2007/2007-02.pdf.

[11] Richard W. Trueswell, "Some Behavioral Patterns of Library Users: The 80/20 Rule," *Wilson Library Bulletin* 43 (1969): 458–61.

[12] East Carolina University, "Joyner Library Collection Guides," http://digital.lib.ecu.edu/special/ead/.

[13] Google Analytics is a free service as long as the website utilizing it receives less than ten million page views per month. For more information, see "Google Analytics Terms of Service," http://www.google.com/analytics/terms/us.html.

[14] Piwik is an open source alternative to Google Analytics that follows the same approach in that it relies on JavaScript and first-party cookies. Unlike Google Analytics, the Piwik software is installed on a locally hosted web server, and the data can be stored in a local database. Piwik also reports a unique page view (UPV) metric calculated in the same way as is done in Google Analytics. The software is available at http://piwik.org/.

[15] The AWStats logfile web metrics software is available at http://awstats.sourceforge.net/.

[16] Christopher J. Prom, "Using Web Analytics to Improve Online Access to Archival Resources," *The American Archivist* 74 (Spring/Summer 2011): 158–84. See pages 161–69, specifically, for the introduction to web analytics and Google Analytics.

[17] For more information on all of the data that Google Analytics reports, see Google Developers, "Dimensions and Metrics Reference," http://code.google.com/apis/analytics/docs/gdata/gdataReferenceDimensionsMetrics.html.

[18] Jackson "used the technique resulting in Richard Trueswell's '80/20 Rule' to see if 80 percent of the use involves only 20 percent of the collection at the UW–Milwaukee Archives." Jackson, "The 80/20 Archives," 133–46.

[19] One definition of a UPV, in this case taken from the Google Analytics documentation, reads as follows: "A *unique pageview*, as seen in the *Top Content* report, aggregates pageviews that are generated by the same user during the same session. A *unique pageview* represents the number of sessions during which that page was viewed one or more times." For more information, as well as this definition, see Google Analytics, "The difference between clicks, visits, visitors, entrances, pageviews, and unique pageviews," http://support.google.com/analytics/bin/answer.py?hl=en&answer=1257084.

[20] It is possible, however, when using Google Analytics, to predefine collection filters to ignore website traffic over a range of Internet protocol addresses. Information regarding the utilization of various filters is available at Google Analytics, "Exclude Internal Traffic," http://support.google.com/googleanalytics/bin/answer.py?hl=en&answer=55481.

[21] Though Google Analytics does not report Internet protocol addresses, website traffic can be segmented by the network domain dimension. For the FY2010 data set, on-campus use is defined as those views that originated when the service provider is reported as "East Carolina University."

[22] To cite only one example, the Thomas Sparrow Papers (#0001)—which will be used as a recurring example in this section—had a FY2010 on-campus ranking of 10 (out of 1,853 total), an off-campus ranking of 35, and an overall ranking of 24. The highest and lowest of these rankings are within less than 1% of each other when compared to the total range of collections. Similar patterns exist for the majority of online finding aids, so it does not appear that the population of users had any dramatic impact on the probability of collection rank (at least not in this single data set).

[23] As will be evident in the next section, ECU's FY2010 data set shows strong indications that this bias occurs in those collections where there is more intense interaction. It should be noted, though, that this type of bias was intentionally built into the website redesign.

[24] This step can also be accomplished from within Google Analytics by filtering the list of URLs present in the Top Content report. For instance, if you know the pattern of URLs that you want to analyze, you can use regular expression patterns to limit to those URL strings, or to exclude others.

Then, you can export the specific data set that you have just isolated. If you merely attempt to increase the export limit by using the options available from within the web interface, though, then you will only be able to export 500 rows at a time. However, even without making use of the API for data exports, you can currently export up to 20k rows of data at a time (this limit was previously set as high as 50k) by appending an extra query parameter to the Google Analytics URL. For more information, see Google Analytics, "To export more than 500 rows of data," http://support.google.com/analytics/bin/answer.py?hl=en&answer=1038573.

[25] Barbara L. Craig, "Old Myths in New Clothes: Expectations of Archives Users," *Archivaria* 45 (Spring 1998): 118–26, for instance, argued that online use, no matter how accidental we might think it to be, is an opportunity for outreach.

[26] OCLC Research, Digitization Matters, Chicago, Illinois, August 29, 2007, http://www.oclc.org/research/events/2007/08-29.html.

[27] Erway and Schaffner, "Shifting Gears," 4.

[28] Nick Poole, "What Audience? The Death of Mass-Digitisation and the Rise of the Market Economy" (paper presented at the Digitaal Erfgoed Conferentie in Rotterdam, the Netherlands, December 12–13, 2007). Here, Nick Poole issued the sobering reminder that "[a]ccess is not sufficient to grow audiences," http://www.slideshare.net/DEconferentie/k2poole.

[29] East Carolina University, "Special Collections Staff Picks," http://digital.lib.ecu.edu/staffpick/.

[30] Jonathan Dembo and Mark Custer, "An Experiment to Increase Online Archival Accessibility: Using Unique Page Views to Measure Online Efficiency," *North Carolina Libraries* 68 (Fall/Winter 2010): 2–11, http://www.ncl.ecu.edu/index.php/NCL/article/viewFile/325/408.

[31] Space and facilities topped the list of "most challenging issues" facing special collections as reported in OCLC's "Taking Our Pulse" survey. Jackie M. Dooley and Katherine Luce, "Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives" (Dublin, Ohio: OCLC Research, 2011), http://www.oclc.org/research/publications/library/2010/2010-11.pdf.

[32] Noah Huffman, "More than Just Linking: Integrating MARC and EAD in a Single Discovery Interface at Duke, UNC–Chapel Hill, and NCSU," *Journal for the Society of North Carolina Archivists* 8 (Spring 2011): 2–17.

[33] Huffman, "More than Just Linking," 11–14. Specifically, Huffman discovered that users tend to spend more time on Duke University's online finding aids when they had discovered those resources in the library's catalog rather than the open web.

[34] As discussed in Google Analytics' Dimension and Metrics Reference document, the "time on page" metric is "[c]alculated by subtracting the initial view time for a particular page from the initial view time for a subsequent page. Thus, this metric does not apply to exit pages for your website." For complete documentation, see Google Developers, "Dimensions and Metrics Reference," http://code.google.com/apis/analytics/docs/gdata/gdataReferenceDimensionsMetrics.html. Though strategies exist to ameliorate this issue—by recording timestamps during custom events, for instance—no such techniques were implemented during this study.

[35] Conducting this same calculation outside of Google Analytics results in a value of 27.6 hours, not 27.5. The slight variation is because, within its interface, Google Analytics provides a rounded value of 141 seconds, whereas the calculated value upon export equals approximately 141.5 seconds.

[36] Shawn Purtell, "Time on Page and Time on Site—How Confident Are You?," *The ROI Revolution Blog*, May 29, 2008, http://www.roirevolution.com/blog/2008/05/time_on_page_and_time_on_site_how_confident_are_yo.php. If the confidence level were 0% (that is, if every view for a specific collection was also an exit from the site), then an explanation for how the average time was estimated for that collection would also need to be provided. An estimate could be based on historic data for the same collection, or it could be based on a current data set for collections that have similar characteristics but a higher level of confidence. In the case of the FY2010 data set, 62 different collections have no time reported, since all of their recorded views were also exits from the site. In this study, however, no EVPHs have been estimated; instead, those 62 collections have been left out of the correlation chart provided in Figure 3.

[37] James B. Rhoads, "The Historian and the New Technology," *The American Archivist* 32 (July 1969): 213.

[38] Rhoads, "The Historian and the New Technology," 211.

[39] Jackson, "The 80/20 Archives," 133–46.

[40] For a fascinating history and analysis of cybernetics, see N. Katherine Hayles, *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics* (Chicago: The University of Chicago Press, 1999).

[41] See, for example, Trueswell, "Some Behavioral Patterns of Library Users," 458–61.

[42] Rhoads, "The Historian and the New Technology," 213.

[43] Rhoads, "The Historian and the New Technology," 213.

[44] This has also been the goal of the Archival Metrics research project, http://archivalmetrics.cms.si.umich.edu/. However, this project has worked on defining metrics for user-based studies in the form of surveys, whereas this case study focuses exclusively on archival collections. Both types of study are vital, but little to no attention has been applied so far to the quantitative metrics that are derivable from archival collections and their finding aids, such as the UPV, which can be collected continuously, easily, and unobtrusively.

[45] As more data sets become publicly available, privacy issues become an increasing concern. For instance, Netflix released a subset of its reviews data in 2006 to be used in its Netflix Prize contest in an effort to improve the accuracy of its recommendation service, Netflix, http://www.netflixprize.com/. The winning entry was submitted on July 26, 2009, and the prize was awarded on September 21, 2009. Arvind Narayanan and Vitaly Shmatikov demonstrated potential privacy risks when these data are supplemented with freely available auxiliary data sets in Cornell University, "How to Break Anonymity of the Netflix Prize Dataset" (February 5, 2008), arXiv:cs/0610105v2.

[46] Mark Custer, "EAD Collection-level Web Statistics for FY2008–FY2010" (July 2011), East Carolina University, http://hdl.handle.net/10342/3606.

[47] For example, this could be accomplished if archival data sets were analyzed with the R programming language, and the code libraries developed were shared via the "Comprehensive R Archive Network" (CRAN) website, "Contributed Packages," http://cran.r-project.org/web/packages/.

ABOUT THE AUTHOR

Mark Custer is an archivist/metadata coordinator at the Beinecke Rare Book and Manuscript Library at Yale University. From 2011 to 2012, he worked as an EAD manager at the Smithsonian Institution. Previously, he worked at East Carolina University's Joyner Library for four years, during which time he helped manage its EAD finding aid database. He earned a BA in English literature from Indiana University, Bloomington, and an MLIS from Syracuse University.