Journal of Western Archives

Volume 2 | Issue 1 Article 1

10-10-2011

Applying Web Analytics to Online Finding Aids: Page Views, Pathways, and Learning about Users

Mark R. O'English
Washington State University, markoe@wsu.edu

Applying Web Analytics to Online Finding Aids: Page Views, Pathways, and Learning about Users

Mark R. O'English

ABSTRACT

Online finding aids, internet search tools, and increased access to the World Wide Web have greatly changed how patrons find archival collections. Through analyzing eighteen months of access data collected via web analytics tools, this article examines how patrons discover archival materials. Contrasts are drawn between access from library catalogs and from online search engines, with the latter outweighing the former by an overwhelming margin, and argues whether archival description practices should change accordingly.

The past decade brought a revolution in intellectual access to archival collections through the emergence of the online finding aid. These tools, created in a multiplicity of formats (EAD, HTML, PDF, and others), have made it possible for researchers anywhere to use an Internet connection to find materials in archives. Since archivists first introduced this genre of digital documents to support unmediated discovery, they have been accompanied by questions about their use and usability. Tools exist to measure access counts and the pathways followed to these finding aids, but to date very little has been published to show whether these tools are being used for this purpose, or what the results are if they are being used.

This article will examine the experience of one repository, Manuscripts, Archives, and Special Collections (MASC) at the Washington State University Libraries, in collecting and analyzing data about the use of online finding aids. By 2008, MASC had well over eight hundred HTML finding aids online but no real numeric data

either on how often they were being accessed or on how they were found. In an effort to gather this data, MASC staff placed code from Google Analytics, a statistics-gathering Web service, on each finding aid webpage, as well as on other local gateway and digital collections pages. The intent was to explore the paths patrons follow to finding aids with the hope of improving access, as well as gaining patron-derived information that would help establish priorities for processing collections from the repository backlog and to determine which processed collections might benefit from further attention.

The results of the statistical analysis provide useful data on where and how MASC users access our finding aids. Prior to the age of digital finding aids, their paper predecessors were found through various methods, which were generally predictable and well-defined, including online and analog library catalogs, published listings, and mediation by archives staff in person or by telephone. As the digital age emerged, many archives channeled substantial resources into major conversion projects, digitizing finding aids and often creating access to them through existing or new MARC records in online library catalogs. Digital search tools have become increasingly powerful and widespread, and social media have emerged as another path to online finding aids. Archivists with quantitative and qualitative data about the use of online finding aids can make better-informed decisions about the most effective ways to support discovery. The findings in this study suggest that it may again be time to change focus, shifting resources away from catalogs and toward the finding aids themselves.

Background

Washington State University (WSU) is located in Pullman, Washington, in the southeastern corner of the state, and is Washington's land-grant university. Pullman has a population of 29,000; three other similarly sized towns are within thirty miles, but the nearest major city is Spokane, eighty miles north. MASC's holdings include the WSU archives, manuscript and photograph collections, rare books, maps, ephemera, and other special collections. These have a broad geographic and topical scope, including collections of national and international importance. Eastern Washington is especially well represented. As a result of WSU's fairly remote location, much of the access to MASC materials typically begins with contact from off -site researchers, either through the repository website, via e-mail, or by phone, making Web-based finding aids a vital part of accessing MASC's services.

The first online finding aids for MASC's collections were generated by digitizing and reformatting legacy paper documents into HTML files, beginning about the year 2000. By 2002, with very few exceptions, all legacy finding aids had been converted and were online. Encoded Archival Description (EAD) versions of 813 of these were generated between 2004 and 2007. These were not hosted on the MASC site, but were contributed to the Northwest Digital Archives (NWDA). The study described here is the first effort by MASC to measure and analyze online finding aid use.

Methodologies

To track website use and access, MASC selected and installed Google Analytics into the repository's webpages. Google Analytics is a free service offered by Google, requiring only a Google account for use.[1] Upon registering for the service the user is assigned a unique piece of code that must be placed in each page to be counted by Google Analytics (hereafter, Analytics). Because an identical code string goes on every page, MASC staff automated the process of inserting the code by using a global search -and-replace program. Using the freeware program Replace Text, formerly known as BK ReplaceEM, we identified a piece of existing code that was common to every page and replaced that piece with itself plus the Analytics code.[2] In this instance, the existing common code used for the search-and-replace process was a universal footer. There are other options for this step; for example, in the absence of that footer we could have used the HTML </body> tag, which usually closes every HTML webpage. As a safeguard against introducing inadvertent errors during the process, we copied the directories of the webpages we wanted to edit and made the change on the duplicates. This precaution proved unnecessary but was still advisable. The coding was placed into approximately 900 pages, including all 841 HTML finding aids, plus other major pages for collection access: lists of unprocessed accessions, directories of university publications, and a few other pages functioning as pointers to the collections. This process required less than one hour of work. The replacement software showed how many changes were made in each file, and with that information we were able to identify the few files missed because of our own internal coding inconsistencies. We then manually updated these pages to include Analytics.

Once installed, Analytics tracks each time a tagged page is accessed, and that data is tabulated into an online portal, which is updated nightly (see fig. 1). The user who sets up the Analytics account can give others access to the data or assign others to give that access, meaning the data can be easily viewed by any authorized user. Analytics tabulates a variety of data; the most relevant data points for our purposes include usage counts, which websites generate access to our pages, and, for access through search engines, which search terms are being used to find the pages.

As time passes, the pool of data grows large enough for meaningful analysis. MASC installed Analytics on July 15, 2008. This study examines eighteen months of data, from August 1, 2008 to January 31, 2010, with some brief comparisons to later data.

It should be noted that Analytics is a JavaScript-based code and requires an actual visit to the page, and loading of the page, to activate. As a result, visits by Web-

1. The service is available at Google Analytics, http://www.google.com/analytics/ (accessed October 5, 2011).

2. The Replace Text software is available at http://www.ecobyte.com/replacetext/ (accessed October 5, 2011).

indexing spiders and bots are not counted in these statistical results. Another consideration is that MASC placed coding at the end of the pages, making it the last thing to load; it is at least possible that on the larger or more complex pages we have failed to count visits by people who realized immediately that they did not want what they found and left the page before it fully loaded. This could potentially artificially inflate the "time spent on page" statistics, but it is impossible to calculate that effect.

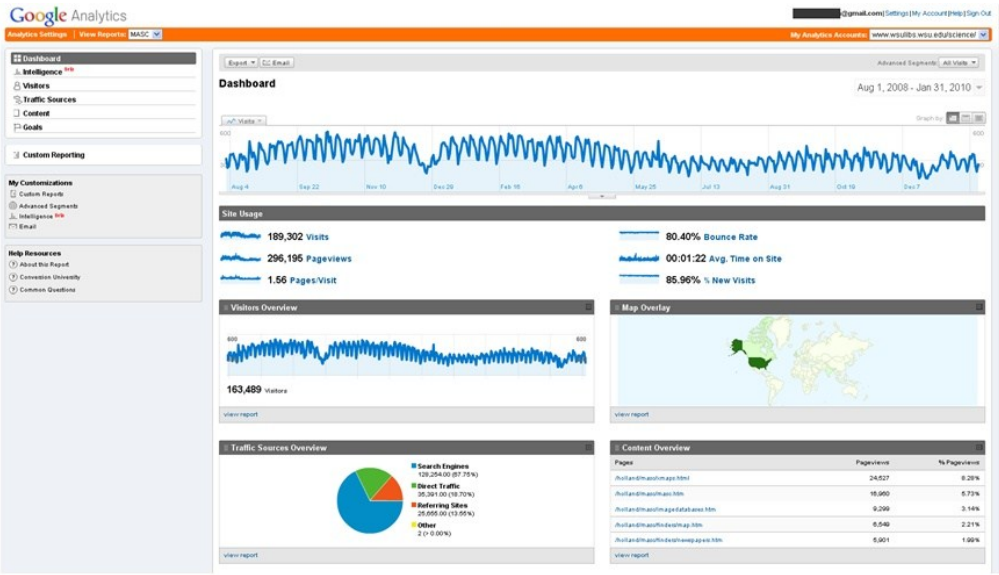

Figure 1. "Dashboard" view of Google Analytics

During this period, MASC's 841 online finding aids were accessed electronically 176,200 times (other site elements, most notably digital collections, bring the total count up to 296,000 as seen in fig. 1, but only the finding aid accesses are addressed here). In addition to being hosted on MASC's websites, the finding aids also exist in EAD formats in a consortial database known as the Northwest Digital Archives (NWDA). The NWDA also uses Analytics, and during this period the 813 WSU finding aids hosted there were accessed 45,600 times. Due to the complex system of filenames used by NWDA, this study will focus only on those accessed on MASC's site.

Results and Discussion

Examining the 176,200 views of MASC's finding aids, perhaps the most interesting data gathered is one of the simplest: where and how people find our finding aids. The following chart tallies total page accesses, based on how users are accessing the page, for the top fifteen referrers.

	Page Views	Unique Page Views	Average Seconds on Page	Bounce Rate
1. Google (including 4,400 from Google Images)	115,538	102,476	188.82	0.86
2. Direct access	26,000	23,669	176.78	0.82
3. Yahoo!	9,738	8,643	168.03	0.86
4. Bing/Live	5,522	4,867	130.83	0.82
5. MSN	3,083	2,716	139.72	0.84
6. AOL	2,865	2,504	126.69	0.83
7. Search	2,047	1,833	135.01	0.84
8. Ask	929	844	66.98	0.88
9. MASC's digital collections	790	646	171.99	0.35
10. Wikipedia	744	700	319.02	0.83
11. WSU's library catalog	533	459	117.49	0.42
12. NWDA	458	402	76.86	0.08
13. Dogpile	402	356	100.78	0.79
14. StumbleUpon	272	156	131.15	0.29
15. index.wsu.edu	255	217	116.94	0.34

Table 1. Finding aid Web traffic sources

The most common access points are search engines (the last, http://www.index.wsu.edu/, is an online campus directory whose only direct link to MASC's finding aids is through a home-built search engine). The 115,500 hits from Google alone comprise two-thirds of our site access. Tabulating all ten search engines listed in the top fifteen access points gives us 80 percent of all the access to MASC's finding

aids. The second most common point of access is direct access—entering a specific URL, as opposed to following a link or accessing via a search engine. Of the remainder of the top fifteen access points, Wikipedia, MASC's digital collections,[3] WSU's online catalog,[4] and the NWDA are, respectively, ninth, tenth, eleventh, and twelfth.

It is common practice for academic libraries to create collection-level MARC records for individual collections, which are incorporated into the library's electronic catalog; the practice dates back to the card catalog era and was long the primary means of collection discovery. Other guides existed, as significant finding aids were frequently published and distributed to other academic institutions, but even those guides needed to be discovered through MARC records in catalogs. WSU is no exception to this system, and its library catalog, Griffin, is at the eleventh position in table 1, with 533 access points from MARC records. During this time frame, the WSU libraries were beginning a move to a WorldCat discovery system; another forty hits came from various WorldCat access points, either national or local. This gives a total of 573 hits from our library catalog, or less than one-third of one percent of the total number of visits to the MASC website. In other words, for every one hit brought in via a MARC record, an online finding aid brought in 2,400 hits from search engines. In a 1998 study comparing finding aids to their MARC records, Rita Czeck found that between 23 and 41 percent of personal names, corporate names, chronological terms, and geographical terms found in finding aids were present in MARC records.[5] This implies that MARC records are useful for finding collections wherein a large portion of the materials relates to the patron's topic of interest, but less than useful for finding specific smaller pieces of a collection, a finding that is likely to surprise no one.

One of the problems with this type of data is that although it tells us that people visited the MASC website, it does not tell us if they found the visit of interest or useful. Another data point provided by Analytics is a tally of how much time users spent on the page—presumably if someone clicks into the page and leaves two seconds later, the page was of less use or interest than it was for someone who spends five minutes on the page. Table 2 takes those top fifteen referrers and reorganizes them by time spent on each page.

Given the calculation that 2,400 search engine hits occurred for every one library catalog hit, one might expect that most of those search engine hits would be false positives, that users would access and leave, while those accessing through a MARC record would already have an idea of what to expect, and a greater proportion of

3. The digital collections website is available at http://www.content.wsulibs.wsu.edu/ (accessed October 5, 2011).

4. The online catalog is available at http://www.griffin.wsu.edu/ (accessed October 5, 2011).

5. Rita Czeck, "Archival MARC Records and Finding Aids in the Context of End-User Subject Access to Archival Collections," *American Archivist* 61, no. 2 (1998): 426-40.

	Page Views	Unique Page Views	Average Seconds on Page	Bounce Rate
1. Wikipedia	744	700	319.02	0.83
2. Google (including 4,400 from Google Images)	115,538	102,476	188.82	0.86
3. Direct access	26,000	23,669	176.78	0.82
4. MASC's digital collections	790	646	171.99	0.35
5. Yahoo!	9,738	8,643	168.03	0.86
6. MSN	3,083	2,716	139.72	0.84
7. Search	2,047	1,833	135.01	0.84
8. StumbleUpon	272	156	131.01	0.29
9. Bing/Live	5,522	4,867	130.83	0.82
10. AOL	2,865	2,504	126.69	0.83
11. WSU's library catalog	533	459	117.49	0.42
12. index.wsu.edu	255	217	116.94	0.34
13. Dogpile	402	356	100.78	0.79
14. NWDA	458	402	76.86	0.08
15. Ask	929	844	66.98	0.88

Table 2. Finding aid Web traffic sources, by time on page

those users would find what they were looking for. Contrary to expectations, the opposite proves true: users accessing through the Griffin library catalog spent an average of just under two minutes with the finding aid while users accessing through Google spent fifty percent more time, an average of just under three minutes. However, one data point that may support the idea that MARC records are more

useful is the bounce rate, which is the number of users who came to a page and then left without looking at anything else on that site. "Local" sites—specifically the library catalog, MASC's digital collections, the WSU-specific search engine, and the NWDA—have a significantly lower bounce rate than all but one of the search engines.

One obvious difficulty in using time on page to evaluate whether patrons find what they need is determining why someone left. If a patron immediately finds what he or she needs, he or she may well leave quickly; however, it should be remembered that we're discussing finding aids that are guides to resources, not the actual resources, and someone who finds what he or she needs will likely have to spend enough time on the page to determine how to gain further access to the materials. If the page is so complex as to be intimidating, however, he or she may also leave quickly. In general it seems safe to assume that if a patron remains on a page for a longer period of time, then he or she has found something on that page related to his or her needs.

In 1991, Steven Hensen, one of those who successfully pushed for using cataloging records for archival use, contended that

> "while some of the more advanced current thinking envisions a day when such finding aids will be available online nationally and internationally (and indeed, even full text databases of archival documents themselves), it still seems likely that the pointers to such material will probably be structured catalog records, obviating the need for a more formal structure in the finding aids themselves. Thus the need for a highly structured approach to the finding aids themselves seems highly dubious at best."[6]

Hensen listed a number of reasons supporting this need for detailed catalog records, not the least of which was that "the length of most of the descriptions are beyond the record size capacities of most bibliographic systems into which they would be entered."[7]

Clearly Hensen's projected international availability of online finding aids became a reality, as the 2000s saw a massive movement within archives to implement online formats. Helen Tibbo and Lokman Meho reported that in February 2001 fewer than one in twelve repositories had more than four online finding aids.[8] By 2010, having over one thousand online finding aids at a single educational institution is not

6. Steven L. Hensen, "RAD, MAD, and APPM: The Search for Anglo-American Standards for Archival Description," *Archives and Museum Informatics* 5, no. 2 (1991): 4.

7. Ibid.

8. Helen Tibbo and Lokman Meho, "Finding Finding Aids on the World Wide Web," *American Archivist* 64, no. 1 (2001): 61-77.

uncommon. During that time period, the number of Internet users worldwide has increased from 361 million to approximately two billion. The increase in online finding aids and in online access seems to be having an effect on in-person access to archive collections; in that same period, while it is generally acknowledged that reference inquiries at libraries are slowly decreasing, overall reference use in archives (both online and in person) seems to be increasing. An unpublished internal survey at Washington State University found virtually no change in the number of reference questions handled in Manuscripts, Archives, and Special Collections between 1992 and 2001. However, from 2001 to 2011 reference questions increased by 300 percent, and time spent answering those questions increased by 700 percent.

Given these changes, few are surprised to discover that Hensen's pre-World Wide Web contention that the access point for locating finding aids must be catalog records has been disproven, but it seems likely that the sheer size of the difference between finding aid access and MARC record access (in this study, 2,400 finding aid accesses for every one MARC record access) may prove a surprise to many. As Hensen noted in 1991, "the chief motivating principle behind archival descriptive standards is the ease with which such standards make the sharing of descriptive information."[9] So, if the need for MARC record use within archives evolved out of the need to create access points to these finding aids, and if those access points are minimally used, does it not follow that it is time to question the usefulness of the MARC record in archives?

Search engines continue to evolve, and the last few years have seen the development of image searches. In fact, many of our finding aids contain sample images of the items in the collections, which can be found via tools like Google Images. While the Google Image data has been incorporated into the general Google search statistics for counting purposes above, it is worth looking at this data separately. Forty-four thousand page views of our finding aids came through Google Images, and the average time spent with finding aids found by Google Images was just sixty-three seconds, about one-third of the time spent through Google arrivals overall. Further research is necessary to determine why image searching is less useful, but in looking at individual page results, it is worth noting that the MASC finding aid most accessed by a Google Image search is not one of our photo collection pages, but rather a moderately obscure collection of political papers, the Joseph Baily Political Papers, 1845-1878, which is illustrated by several photographs of individual papers.[10] It could be theorized that to make image searching useful for finding aids, we might need to balance MASC's usual practice of displaying something eye-catching against showing a representative image of the type of materials contained within the collection. A patron looking for political materials apparently will follow an image of papers to the collection.

9. "RAD, MAD, and APPM," 4.

10. Washington State University Libraries, "Joseph Baily Political Papers, 1845-1878," Washington State University, http://www.wsulibs.wsu.edu/holland/masc/finders/cg670.htm (accessed October 5, 2011).

To determine the effectiveness of social media access, MASC established a Facebook account for promoting collections/items and engaging users; by the end of 2010 it reached just over 160 fans. During the period of this study MASC averaged 9.5 hits per page from Facebook pages; after the Facebook site was created, the number of hits per page rose slightly to twelve per month, but average time on site plummeted from forty-one seconds to just twelve seconds. While Facebook may have promotional or other uses, to date it has not been cost-effective as a tool for connecting people to the finding aids. One reason for this might be the small size of the audience, but it also seems clear that few people use Facebook in a research capacity— while one might be a Facebook fan of his or her favorite musician, he or should would not likely access Facebook for historical information on that artist's albums or performances. Similarly, people researching in MASC do not begin their search using Facebook, and promotional notes do not necessarily lead to immediate use of finding aids.

One qualifying factor that should be taken into account is that while MASC and NWDA's sites combine for 221,800 online hits on our finding aids, an informal, in-house count of in-person, phone, and e-mail reference visits estimates that MASC had about 3,750 visitors in an eighteen-month period (unfortunately no month-by-month data is retained for this period, so this is an extrapolation from data for the 2009-2010 fiscal year). This translates to approximately one visit for every sixty Web hits. Again, though this is an approximate figure, it is clear that only a small percentage of website hits results in real-world follow-up. Without further study, it is difficult to do more than speculate on what aspects of Web-based finding aids invited or discouraged follow-up by users, or, indeed, if any factors beyond the applicability of the content are in play here. Certainly research into this connection is merited.

Conclusions

Gathering real data about the number of page views for online finding aids and the electronic pathways used to connect to them is an important first step in learning about how finding aids are being used. Accessing an online finding aid requires an investment of time by users, so it is not just a matter of ensuring that finding aids can be located, but rather it is important to increase the likelihood that those finding aids are located only when they are appropriate to the users' needs and are presented in such a manner as to make them immediately understandable and usable, both in and of themselves and as primary access tools for the collections they describe. The unobtrusive measurement tool used for this study provides valuable information about researcher behavior that can be incorporated into the development of better professional practices.

Studies looking at finding aid effectiveness have determined that patrons are likely to be confused by overly complex search tools and that fielded searching

options should be eliminated in favor of Google-like search options.[11] Again, this supports our findings—MARC searching is, at heart, a fielded search, as opposed to the simpler but less precise full-text searches found through online engines.

In an age when linking collections from Wikipedia proves more useful in bringing patrons to collections than linking them from a library catalog does, does it follow that we need to spend more time sharing collections on the greater World Wide Web? One advantage of social sites like Facebook or Wikipedia is that the patrons do the work of placing finding aids into what they feel is a relevant and useful context for themselves and others; the archivist needs do only the work involved in creating the finding aids. Libraries and archives have very recently expanded upon this practice: at the April 2011 Association of College and Research Libraries conference, the University of Houston reported a substantial increase in traffic to their collections after they intentionally created links to their sites from Wikipedia pages,[12] and the University of Washington found that up to 5 percent of the viewings in their digital images collections came from Wikipedia after a similar project.[13] It should be noted that in the case study here, the Wikipedia links to the WSU MASC pages have all been added by patrons, with no prompting by MASC. One might theorize that patron-added links would be more useful than those added by archivists, as the patrons are presumably acting because they found something useful, while the archives are acting, at least in part, to promote their own materials. Further study on the effectiveness of comparing patron-added and archives-added links is necessary to determine if this is true.

Clearly there are advantages to creating MARC records, not the least of which is the ability to collocate similar records in one place. The author would not yet contend that MARC records are not worth the time investment, though it seems possible that may be the case in the future. Streamlining the process of MARC record creation is already underway, as EAD-encoded finding aids are frequently constructed to allow export to MARC formats with little additional effort required. However, this still requires the creation of the just-as-complex EAD records, which raises the question of whether creating EAD finding aid records makes sense when compared to creating simple searchable HTML. What EAD encoding does well is allow interoperability with other institutions by requiring uniform standards for all finding aids and a framework for encoding into XML webpages. However, it is unclear whether institutions can be as effective at lesser cost by using EAD's overall best practices

11. Christopher J. Prom, "User Interactions with Electronic Finding Aids in a Controlled Setting," *American Archivist* 67 no. 2 (2004): 234-68.

12. Steve Kolowich, "News: Wielding Wikipedia—Inside Higher Ed," *Inside Higher Education*, April 5, 2007, http://www.insidehighered.com/news/2011/04/052/college_libraries_use_wikipedia_to_increase_exposure_of_their_collections (accessed June 27, 2011).

13. Ann Lally, "Using Wikipedia to Highlight Digital Collections at the University of Washington," *The Interactive Archivist*, May 18, 2009, http://interactivearchivist.archivists.org/case-studies/wikipedia-at-uw/ (accessed July 6, 2011).

without using any XML encoding and simply using HTML. A study of two parallel systems that compares the effectiveness and likelihood of the use of HTML finding aids using simple keyword searching against XML/EAD finding aids and their more advanced search capabilities might cast more light on any real gains toward faster, more effective searches, allowing archivists to evaluate those gains against the time and costs incurred.

As noted, MARC records and library catalogs are a benefit to academic libraries because they allow them to have all their records in one place. As academic access to many research materials currently occurs through a pool of databases that are frequently separate and non-interoperable, the absence of archival collections from the library catalog creates yet one more unique external site to be searched. However, as shown here, the existing MARC records bring only a very small number of users to the collections, again raising the question of cost-effectiveness.

What is clear is that we need to focus our cataloging attention on creating more effective collection guides rather than on writing MARC records for library catalogs. The creation of each record format involves a cost to the archive in terms of time and training, and most archives already have accession backlogs still in need of processing. With an increasing demand to satisfy repository priorities, and in an environment of scarce resources, careful cost-benefit analysis is essential. This study supports an increased focus on the use of the greater World Wide Web.