Joseph T. Gilmore. The Use of University of North Carolina Library Materials on Internet Archive. A Master's Paper for the M.S. in L.S degree. May, 2013. 32 pages. Advisor: Dr. Richard Marciano

An examination of rates of use of digitized materials uploaded to Internet Archive by UNC Libraries. Suggestions for improvement to the organization of those materials.

Headings:

Academic libraries

Digital libraries

# THE USE OF UNIVERSITY OF NORTH CAROLINA LIBRARY MATERIALS ON INTERNET ARCHIVE

by
Joseph T. Gilmore

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

May 2014

Approved by

_____

Dr. Richard Marciano

## Introduction

Special collections libraries are increasingly digitizing materials to improve their services to patrons. Since most special collections libraries are far too large to digitize more than a fraction of their collections, it is necessary for these institutions to be selective about what gets digitized. Several factors must be considered when making these decisions, including an item's condition, copyright status, and potential value to researchers. While condition and copyright issues must be addressed on a case-by-case basis, the potential value of materials can typically be generalized based on content. These generalizations allow libraries to create digital collections of thematically related items. The makeup of these collections is often left to curatorial discretion, at least in the early stages of digitization, but a subsequent analysis of the recorded use of digitized materials should inform future decisions about what gets digitized.

The various subdivisions of the Louis Round Wilson Special Collections Library at the University of North Carolina in Chapel Hill (and, to a lesser extent, other areas of the UNC Libraries system) undertake many digitization efforts on a variety of platforms. One of the largest of these efforts is the digitization of print materials for addition to Internet Archive. Tens of thousands of items have already been scanned and uploaded to Internet Archive, with more being added every day. Although there is plenty of anecdotal evidence of the value of these

digitally available resources to researchers, to date there has been no study of how much these materials are actually being used.

The present study is an analysis of the use of items digitized by UNC Libraries through their partnership with Internet Archive. This study has two interrelated goals: 1) to determine which of UNC's materials on Internet Archive have historically received the most use, and 2) to use this information to inform future digitization efforts. This will allow UNC Libraries to better serve their patrons by continuing to create and expand digital collections that have observable utility. UNC Libraries can thus make their partnership with Internet Archive as efficient and effective as possible.

## Literature Review

A review of the literature reveals that very little work has been done on use measurements of library-created digital collections. Probably due to budgetary reasons, there has been a greater focus on measuring the use of licensed digital resources.[1] Licensed e-resources command more of a library's budget than do freely available, library-created objects, so there is a greater demand for data-driven decision-making related to those resources. Some of the findings of these studies on licensed e-resources can be easily applied to items digitized by libraries and other institutions, but they differ from the present study in one important way. Regardless of whether they are licensed or free, it is often easy to see which individual digital resources are receiving the most use or what the total use of all digital resources is. As I will discuss in greater detail below, the present study hopes to determine which curated *collections* of digitized items are used most heavily in order to inform future development of these digitized collections.

There are several studies on the usability or discovery of library-created e-resources such as those on Internet Archive and Project Gutenberg. Each of these issues is clearly central to the use of digital resources, but they are outside the scope of this paper. Considering that each of the digital objects UNC uploads to Internet Archive uses the same interface and can be accessed in the same ways, issues of usability or discoverability will not affect which collections receive more use than others. These issues will therefore not have any impact on the curation of continued digitization initiatives.

Chmielewska & Wròbel endeavored to calculate the total use of their digital collections, but they did not conduct a detailed examination of which portions of their digital collections receive the most use and why.[2] Demonstrating that digital collections are being used is important, but continued development of the collections requires information on which collections are succeeding and which are not. This information can then be used to inform decisions on digitizing efforts that ought to be explored further.

User surveys examining how, why, and by whom e-resources are used are far more common than analysis of use statistics.[3] These are useful tools for collecting qualitative data, but the studied populations are often self-selecting, which leads to skewed input when examined as a whole. Use statistics, on the other hand, have the benefit of being impartial. For this reason, this latter sort of empirical data will be invaluable to making and defending decisions related to the creation and development of digitized collections.

As digitized materials become increasingly vital to research, libraries need to understand how they change the research practices of local and remote patrons. Sasser

examined the changing practices of academic research in the digital age, but library resources are widely used for research outside of the academy as well.[4] This must be kept in mind when evaluating use of digital library resources. For example, are digitized materials related to local or regional life likely to get much remote use? Similarly, are digital versions of relatively common published works as likely to be used remotely as digital copies of unique or rare items? We can speculate on the answers to these questions, but our theories would be greatly strengthened by empirical data.

An area of study that has received particular attention is the accessibility of e-resources. Some studies indicate that creating online catalog records increases the use of both licensed and free e-resources. The reasoning behind this is two-fold. First, the OPAC is the traditional place to search for and access library materials and is therefore a logical place to provide access. Second, linking to digitized materials that are already hosted by (and accessible from) another site adds a second point of access, thereby hopefully increasing its potential use.[5] McCracken discusses the ongoing discussion in librarianship about creating single or separate catalog records for analog and digital versions of items.[6] This literature on strategies for increasing the use of digital resources could be strengthened by empirical evidence – namely, examinations of use statistics like the one presented here.

Regardless of discoverability or accessibility, promotion is seen to be necessary for "unhiding" digital resources that are otherwise lost in the morass of online information. Being online (and therefore presumably indexed by Google) is not enough to guarantee discoverability or use. This is particularly true for very specialized

collections, which can be hard to search for even in a constrained environment like a library's OPAC. [7]

For all of the reasons outlined above, not to mention the present dearth of information on which types of digital resources get the most use, the following analysis of usage statistics for UNC's materials on Internet Archive hopes to shed light on the steps UNC's digitization efforts ought to take in the immediate future.

## Methodology

Usage statistics for all of the materials hosted by Internet Archive are freely available on the Internet Archive website. This is true for both individual items and entire collections of items, which I will discuss in greater detail below. Before discussing the specific methods and findings of this study, it is necessary to describe the data as it is made available by Internet Archive. This data has many peculiarities that shape its uses and in some cases even severely impair analysis.

The unit employed by Internet Archive to designate use is "download," but this term is construed broadly. It includes instances of an item being downloaded as a PDF or some other format, but it also encompasses types of use such as briefly opening an item in a web browser. This definition of "download," though vague, is sufficient for the purposes of this study, particularly given the amount of data in question. The download count supplied by Internet Archive may not give nuanced insight into how resources are being used or to what degree, but it at least indicates a minimum degree of use for a given item or collection.

It is not possible, from the public interface, to seamlessly cull data for all of the individual items uploaded to Internet Archive by UNC. While data is technically

available for each of these items, it is not possible to filter for only UNC items through Internet Archive's interface. Because of the huge number of digitized items uploaded to Internet Archive by UNC Libraries (roughly 70,000 items), it was therefore necessary to assess the use of materials at the collection level rather than the item level. All items on Internet Archive are organized into collections. Most of the collections created by UNC are organized thematically, so analyzing these collections provides generalized insight about the types of digitized materials that are being used most heavily. It is not possible to filter for only UNC collections either, but it is far easier to manually harvest data for UNC's sixty-eight collections than for its 70,000 individual items.

Internet Archive provides both monthly and total download statistics for every collection as well as the number of items that were in the collection at every point that the data was recorded, which happens once a month. Since there is unfortunately not a way to export this data from the Internet Archive site, it has to be copied and pasted from an internet browser before any meaningful analysis of the data can begin. This is extremely tedious, particularly since – as mentioned above – there is no way to sort this data by collection, contributor, etc. It was therefore necessary to identify specific collections to which UNC has contributed and copy – row by row – only the data pertaining to those collections.

The data for this study was gathered in June 2013. It spans the period from the earliest usage statistics made available by Internet Archive from January 2009 up to May 2013. Rather than harvesting and analyzing data for UNC's collections on Internet Archive for each month, a sample was taken from every fourth month of each year covered: January, May, and September.

For each of the sixty-eight collections to which UNC contributed, the number of monthly downloads was divided by the number of items that were in the collection at that time. Thus, the use of each collection was calculated as monthly downloads per item. These rates of use for each month were then averaged for each collection across their respective lifetimes. The rate of use was therefore measured as *average monthly downloads per item*. Using this figure, rather than the raw numbers provided by Internet Archive, normalized the rate of use across collections so that very small collections could be meaningfully compared to very large collections.

Once all of this data was compiled and normalized, it was analyzed according to basic characteristics to rule out possible correlations to use such as collection size and age. The purpose of this was to determine whether the collections to which UNC Libraries were devoting the most resources (i.e., the larger collections) or to which they had remained committed (i.e., the older collections) were being used commensurate with those resources and commitments.

The thinking behind this is twofold. First, larger collections cast a wider net and therefore may conceivably either attract more use or dilute use as it is dispersed over a greater number of items. To this end, either positive or negative correlations would be informative. Second, more library resources are allocated for the creation and maintenance of larger collections, which hopefully will correspond with greater utility for patrons. This will be particularly important if a collection's use is found to negatively correlate to its size, since this would raise questions about the benefits of creating larger collections.

Similarly, the possibility that collection use correlates to collection age would shed light on the timeframe that ought to be adopted for creating and promoting collections on Internet Archive. It may take time for a collection to attain a level of visibility that would bring in ideal rates of use, or a collection may see its use dwindle after a certain length of time (or both). Nevertheless, it was recognized that even if such correlations were found they would have to be explored in order to best interpret other, more fruitful relationships observed in the data.

Data exists for nearly four years of usage statistics for UNC materials on Internet Archive. This is not an incredibly long time, but it may be enough to observe changes in the use of materials over time. Ideally, the use will have steadily increased since UNC first started uploading digitized items to Internet Archive. Comparing the rate of use across UNC's collections from month to month, we can see whether this is the case, or whether any other trends can be spotted.

After looking at the patterns of use over time for all of UNC's collections on Internet Archive, patterns of use of individual collections over their respective lifetimes were compared. The purpose of this was to examine whether collections saw changes between their first trimester, second trimester, and so on, rather than between January 2010 and May 2010. (Trimesters are used here because, as mentioned above, data was only harvested every four months.) This would provide a more nuanced analysis of the relationship between collection use and collection age than calculating for a correlation coefficient, since periodic fluctuations might reveal themselves that would otherwise have been hidden.

Some of the collections to which UNC contributes are shared with other institutions. More will be said on these collections and the problems they create for analysis later, but it was important to examine whether shared collections are more or less used than UNC-only collections. Differences in the level of use between collections UNC Libraries share and those to which they alone contribute would be relevant for future digitization efforts. Nevertheless, it will be necessary to consider possible explanations for such observed differences.

Twenty-six of UNC's sixty-eight collections on Internet Archive – over a third – are primarily or exclusively devoted to materials in a language other than English. The majority of these collections contain Russian materials from UNC's Savine Collection or resources on individual countries in Latin America. Examining how much these materials are getting used compared to English language materials may shed some light on the global possibilities for UNC's materials on Internet Archive. Of course, this could also raise questions about the community UNC is or ought to be serving.

Perhaps the most important question considered in relation to this data was whether the collections to which UNC is actively adding materials are also the collections seeing the most use. It is important to know whether ongoing digitization efforts are being rewarded by high rates of use. The other side of this question is whether terminated projects ought to be revisited based on a sustained level of patron interest. Since the purpose of this study was partly to ensure that UNC Libraries are digitizing resources that are actually getting used, it was necessary to pay special attention to usage rates of ongoing digitization efforts of UNC Libraries.

## Findings

The condensed, normalized data for all sixty-eight collections to which UNC contributes is listed in Table 1. These collections are sorted in descending order by their rate of use (average monthly downloads per item). The average rate of use across all of these collections is 7.2 monthly downloads per item. This includes the troublesome "unclibraries" collection (6.3 monthly downloads per item), which I will discuss in much greater detail below. Removing this problematic collection raises the average rate of use less than .02 monthly downloads per item, from 7.168 to 7.181).

The most used collection on this list is the "annali" collection, whose 50.9 average monthly downloads per item are more than double that of the next collection on the list ("nchist," at 24.2 average monthly downloads per item). This is a small collection, consisting only of back issues of the UNC publication *Annali D'Italianistica*. Unlike most items on Internet Archive, these items are all from the past thirty years or so and are therefore not in the public domain. This fact surely accounts for at least part of the extraordinary rate of use of this collection. On the other end of the spectrum, items in the "populargovernmentunc" collection are only downloaded an average of 1.5 times per month.

The average size of a collection is 1,039 items, although this number is definitely problematic. As hinted above, the "unclibraries" collection skews this total collection data considerably. The "unclibraries" collection has 28,521 of the 70,681 total items picked up in this data. Removing this collection, the average size of a collection drops all the way to 629 items, though most of the collections are actually far smaller.

The first question considered after the data had been compiled and normalized was whether rates of use had changed significantly over the time period for which data is available. The short answer is that, yes, rates of use have changed considerably over that time. These changes have not been unidirectional, however, and the rates of use over time present themselves as a series of peaks and valleys (see Fig. 1 below). In January 2009, the first month for which there is data for materials on Internet Archive, the average rate of use across UNC's collections was 8.6 monthly downloads per item. This is the highest rate of use recorded for any of the months analyzed in this study. The following two Januaries, 2010 and 2011, each saw their respective rates of use plummet to a mere 3.8 monthly downloads per item. These are the two lowest rates of use, with January 2010 being slightly lower (3.76 monthly downloads per item versus 3.77 monthly downloads per item in January 2011). Since January 2011, use has seen a steady but not uninterrupted increase. The most recent rate of use recorded in this data, from May 2013, indicates that UNC items on Internet Archive were being downloaded on average roughly 6.6 times per month.

This picture is complicated by looking at how use of a collection changes, on average, from its first trimester of existence to its second trimester, and so on. Collections have seen, on average, a rate of use of 7.6 monthly downloads per item in their first trimester. This figure is followed by a steep drop-off over the next year which then gives way to a steady rise to 10.2 average monthly downloads per item in the fourth year – or fourteenth trimester – of use (see Fig. 2 below). It is important to keep in mind when comparing these figures that only eleven of the sixty-eight collections were in existence

in January 2009 (giving them the maximum – for this study – fourteen trimesters worth of data).

Furthermore, five of the top eleven collections in total average use date back to January 2009. This begs the question of whether highly used collections are creating the appearance of an upward trend in use over time, or whether their high rates of use are due to the length of time they have been in existence. A preliminary glance at the data is inconclusive. For example, the "keats" collection had rates of use of 12.2, 9.2, 14.6, 20.1, and 13.7 monthly downloads per item from January 2012 to May 2013. This seems to indicate a random rather than increasing distribution. The "savmil" collection, on the other hand, saw rates of 8.6, 8.2, 10.9, 12.7, and 12.5 monthly downloads per item over the same period. This latter series of numbers certainly adheres more closely to the theory that use steadily increases with a collection's age.

Interestingly, no statistical correlation was found to exist between collection use and collection age. These two variables had a Pearson's correlation coefficient of merely .074, with a probability of .547 (see Fig. 3 below). These figures do not take into account the possible ebb and flow of a collection's use, in particular a possible drop-off after an initially high rate of use. Nevertheless, a correlation coefficient so close to zero all but erases the possibility of there being any predictable change in use over a collection's lifetime. Thus, the nearly random distribution of rates of use for the "keats" collection is likely typical of the fluctuation in use that a collection sees over its lifetime.

After examining changes in use over time, collection use was compared to collection size to see whether any correlation existed between the two. Collection use and collection size were found to have a Pearson's correlation coefficient (-.065, with a

probability of .599) that was even lower than the correlation between use and age (see Fig. 4 below).

The absence of a correlation to collection size or age indicates that other factors contribute to a collection's use. There is a multitude of other factors that could be considered here, but there are a couple that immediately present themselves from the data at hand. First, over a third of the collections on Internet Archive to which UNC contributes are devoted primarily to foreign language materials. These collections see an average rate of use well above the total UNC average (8.8 monthly downloads per item), but this figure is skewed dramatically by the presence of the aforementioned "annali" collection (see Table 4 below). This high rate of use predictably plummets to 7.1 average monthly downloads per item, which is just a shade under the average rate of use for all of UNC's materials on Internet Archive. Adjusting for the extreme use of the "annali" collection indicates that UNC's foreign language materials on Internet Archive are used at a comparable rate to its English language materials.

Another trend that is evident from t data is that the large shared collections to which UNC contributes, such as "civilwardocuments" and "ncreligion," are almost exclusively among the least used collections on the list in Table 1. Indeed, the average rate of use for these collections is 5.1 monthly downloads per item, a mark that is well below the average mentioned above. Of the seven collections that UNC shares with other institutions, only "worldwartwodocuments" sees an above average rate of use at 8.3 average monthly downloads per item (see Table 5 below).

This data should be taken with a grain of salt, of course. As mentioned above, since UNC does not contribute all or even most of the items to these collections, their

rates of use cannot be ascribed exclusively to UNC's contributions. These shared collections have an average of over 3500 items each, of which UNC has contributed only a fraction. Unfortunately, UNC's materials cannot be parsed out of the broader data on these collections, and this is as nuanced an analysis as can be reasonably made of these shared collections.

Finally and most importantly, the collections to which UNC is actively adding materials were examined separately in order to determine whether those ongoing efforts are being rewarded by the highest rates of use. Fifteen collection to which UNC Libraries contribute grew between January 2013 and May 2013 (see Table 6 below). On average, these collections see far less use than the totality of UNC's collections, at a paltry 5.1 monthly downloads per item. Two of the collections on this list, "civilwardocuments" and "statelibrarynorthcarolina," are shared collections which were added to by institutions other than UNC. Removing these collections from consideration actually brings the rate of use for active UNC collections down a hair farther, from 5.13 monthly downloads per item to 5.11 monthly downloads per item.

These numbers indicate that UNC is not devoting its resources to developing highly used collections on Internet Archive and that it has ceased to develop its collection which have historically received the most use. Indeed, only one of these active collections, "unccubanhistorical," has seen above average use over its lifetime at 9.3 average monthly downloads per item. This collection only had one item added in the three months between the last two points at which data was gathered for UNC's materials on Internet Archive.

## Limitations

As can be partly seen from the above analysis, the data that Internet Archive maintains on its content is very messy. Far and away the largest limitation of the data I will be analyzing is that collections on Internet Archive are not exclusive. Individual items can belong to any number of collections simultaneously. This has an obvious effect on the data. For an item listed as part of more than one collection, every download count adds to the total of each of the associated collections. In other words, if a collection that gets otherwise low use contains items also listed as part of a heavily used collection, it is likely that this would skew the data for one or both collections. This is only one of many imaginable scenarios.

An item, upon its addition to Internet Archive, is assigned one or more existing collection codes for organizational purposes. "Collection" is therefore a bit of a misnomer. Rather than thinking of items as being *in* a collection, it is perhaps more helpful to think of them as being tagged with one or more terms (i.e., collection codes). Thus, each collection is not a bounded grouping of like items, but rather a loosely connected cluster of items with shared tags. I will therefore sometimes refer below to "collection codes," the labels given to collections, rather than to collections as such. This system is not necessarily a bad thing. Indeed, the capacity for existing "in" more than one collection at once is a useful affordance that is unique to digital media.

Nevertheless, as I will discuss in greater detail below, cognizance of this organizational schema ought to shape UNC's organization of its materials on Internet Archive. Otherwise, extricating any single collection (or collection code) from other collections is all but impossible – as is, by extension, any useful analysis resulting from

such isolation of collections. It is possible that items tagged with multiple collection codes see increased use by virtue of having more than one access point, but we cannot tell based on the available data.

This is the largest obstacle to a helpful analysis of collection use. It would be problematic enough even on its own, but it is exacerbated for the purposes of this study by the creation in 2010 of the "unclibraries" collection code, already mentioned in passing above. Every UNC item digitized by Internet Archive from that point on has been tagged with the "unclibraries" code along with whatever other collection code is most appropriate for it. This means that individual items contributed to Internet Archive are *guaranteed* to add at least two data points rather than one to both item and download counts. Unfortunately, "unclibraries" represents well under half of the total items in the data under present analysis, so simply excising it from the data would not solve the problems its presence creates.

In addition to this frustrating characteristic of UNC's materials on Internet Archive, a handful of the collection codes used by UNC are also used by other institutions. This further muddies the data, since the rate of use for a shared collection may or may not reflect UNC's contributions to that collection. Examples include "ncreligion," "worldwaronedocuments," and "ncgovdocs." These joint collections (or shared collection codes) are often part of grant-funded partnerships between UNC and other institutions. For example, UNC, Duke University, and Wake Forest University are each responsible for adding items to the "Religion in North Carolina Digital Collection," which is designated by the "ncreligion" collection code. Parsing out the respective

contributions of each institution to the collection is all but impossible because of the filtering issues for Internet Archive's data that have already been mentioned.

The opposite is also true, of course, which only makes matters worse. Since non-UNC items cannot be filtered out of the data, items added to Internet Archive by other institutions or even private individuals to collections used by UNC are inevitably swept up in UNC's data.

Besides these problems inherent in the way Internet Archive makes its data available, there was a major flaw in the design of this study. In particular, gathering data from only every fourth month is problematic given UNC is an academic institution whose calendar is already broken into fall, spring, and summer semesters. As such, data for this study was inadvertently gathered from only the first month of each of these semesters. It is unclear the extent to which this affects the overall rates of use recorded, but it almost certainly affects the periodic changes observed in that use.

## Recommendations

UNC Libraries' digitization efforts in partnership with Internet Archive are already substantial. It is now time to take stock of these efforts to ensure that they remain efficient even as they continue to grow. A general, empirical look at the recorded use of UNC's digitized collections was a logical starting place. The data gathered in this study should provide some support for UNC Libraries' future decisions on digitizing materials that would otherwise have to rely on conjecture or anecdotal evidence. Furthermore, it provides a base on which future studies can now build. By turning to empirical analyses of its digitization efforts, UNC can begin to better extend its impressive resources to scholars around the world in the most effective and relevant ways possible.

Such empirical analysis requires good, accessible, sensible data. Indeed, the overarching takeaway from this otherwise rather abortive study is that the current state of the data on UNC's materials digitized through Internet Archive is so messy as to render meaningful analysis nearly impossible. This is partly due to the way Internet Archive keeps and makes available data on the content it hosts, but it is also due to UNC Libraries' organizational practices in relation to their materials on Internet Archive. It is my hope that this study, in spite of its several limitations, provides a catalyst for UNC Libraries to seriously examine its partnership with Internet Archive with an eye towards better organization of its digital collections. Only once a comprehensive approach to the organization of these materials is undertaken can the meaningful appraisal of their collective utility begin.

The biggest step toward cleaning up the data would be a reevaluation of the collection codes and their optimal uses. The problems created by overlapping collections described above create far too much confusion for the evaluation of use. Not only are problems created by individual items being assigned to multiple collections, but there is also no clear distinction between several of the collections ("nchist" and "ncgen," for example). UNC Libraries staff should get together to consider ways that the organization and terminology can be improved. Such improvements would not only raise the potential for analysis of the collections – a minor thing, after all – they would also make patron searching and access much easier.

Bearing the above limitations in mind, there are several avenues UNC Libraries might pursue in order to bolster the present study and mitigate the limitations of the data provided by Internet Archive. It would be especially helpful to attain a richer

understanding of the means through which UNC's materials on Internet Archive are accessed.

Multiple access points exist for all of the items digitized by UNC and uploaded to Internet Archive. Everything on Internet Archive is available directly from their website. Moreover, since every item digitized by UNC Libraries is already a part of UNC's holdings, a catalog record already exists for it before it is digitized. Once an item has been digitized and put online, a link to the digital version is added to its record in the OPAC. Besides this minimum of two access points, many digitized items or collections are linked to from various other parts of the UNC Libraries websites, including online research guides and pages related to digital humanities projects. As mentioned above, analyzing how different resources are accessed would shed light on a debate that is already ongoing within the digital library community.

The above analysis of how much collections are used should be supplemented by an examination of how UNC promotes and provides access to the materials digitized for Internet Archive. It is not enough to rely on decontextualized numerical data to interpret the utility of respective collections on Internet Archive. One analysis that would be informative is determining whether patrons are most often accessing these items from UNC's OPAC or directly from the Internet Archive site (or even from some other source).

It will also be necessary to analyze what UNC Libraries or the parties responsible for the collections are doing to ensure the continued use of the materials. If a collection is not linked in any way to the UNC Libraries website, it would not make sense to cite its low use as evidence of patron disinterest. The relationship of each individual collection to

the UNC Libraries web presence must be seen as integral to its use. Nevertheless, as mentioned above, a digitized collection being visible and accessible on UNC's website does not guarantee its use.

Although this portion of the project would not be as empirical as the analysis of the raw usage data, it would still be possible to quantify certain aspects of the promotion of UNC's digitized collections. For example, using Google Analytics, it would be possible to see what websites link to specific collections or even individual items on Internet Archive. In cases where there are links to these collections or items from UNC web resources, data should be available from UNC Libraries' ITS Department on the traffic seen by those links. This information will theoretically shed light on the degree of success of specific instances of promotion of UNC's digitized materials.

If any collection is linked to or otherwise promoted by any other agency or institution, this should be seen as particularly relevant to the study at hand. Not only could we speculate that such a scenario would lead to greater traffic for that collection, but we should also consider why that collection was deemed worthy of special mention by another party.

Furthermore, observing which of UNC's digitized materials receive more or less use is only part of the process. In order to give these figures more objective weight, they should be compared against usage rates for materials digitized by peer institutions through Internet Archive, as well as Internet Archive use statistics as a whole. This would indicate how the use of UNC's digitized materials rates in relation to the broader academic community. A general analysis of another institution's use of Internet Archive, comparable to the one undertaken here for UNC, would shed light on whether UNC's

materials on Internet Archive are being used more or less than should be reasonably expected. With such an end in mind, the aggregate average monthly downloads per item for all of an institution's digitized materials would probably suffice as a point of comparison (rather than the rates of use for each collection to which the institution contributes). The same applies to data for the entirety of the materials on Internet Archive.

Unfortunately, the limitations of the data provided by Internet Archive render even such a general analysis as the one proposed all but impossible. Indeed, the present study was only possible because of institutional records kept by UNC Libraries on which Internet Archive collections they had contributed to. Such a study would therefore require the participation of the institution whose materials would be the object of analysis.

In general, the data as it currently exists is far too noisy to draw definitive conclusions about the thematic content of the materials being used, which is arguably the most useful factor in predicting the success of future digitization efforts. Before such nuanced analysis can begin, it will be necessary to clean up the data that UNC Libraries have control over, such as collection codes. This will require an institutional reevaluation of the organization of materials on Internet Archive. Only then can continued analysis of UNC's materials on Internet Archive, with an eye towards improving their value to researchers, influence future decisions about UNC Libraries' digitization projects.

## Notes

1. Schlosser, M. and Stamper, B. “Learning to Share: Measuring Use of a Digitized Collection on Flickr and in the IR” (2012); Stewart, C. “Keeping Track of It All: The Challenge of Measuring Digital Resource Usage” (2011)

2. Chmielewska, B. and Wròbel, A. “Providing Access to Historical Documents through Digitization” (2013)

3. Chmielewska & Wròbel (2013); Sasser, P. “Sounds of Silence: Investigating Institutional Knowledge of the Use and Users of Online Music Collections” (2009)

4. Sasser (2009)

5. Chmielewska & Wròbel; Hill, H. and Bossaller, J. “Public Library Use of Free E-Resources” (2013)

6. McCracken, E. “Description of and Access to Electronic Resources (ER): Transitioning into the Digital Age” (2007)

7. Court, N. “When and Why Is a Collection ‘Hidden’? Awakening Interest in the Hornung Papers at West Sussex Record Office” (2013); Schlosser & Stamper (2012)

# Appendix

**Table 1: Collections**

| Collection Code | Use* | Age (Trimesters) | Items (as of May 2013) |
|---|---|---|---|
| annali | 50.9 | 9 | 21 |
| nchist | 24.2 | 13 | 32 |
| keats | 14.5 | 14 | 14 |
| savcos | 13.7 | 12 | 24 |
| uncsils | 12.4 | 12 | 1 |
| vargas | 12.2 | 14 | 110 |
| mesda | 11.9 | 12 | 63 |
| rbcwb | 11.5 | 14 | 14 |
| ncna | 10.6 | 14 | 69 |
| uncmexicanhistorical | 10.4 | 4 | 38 |
| savmil | 9.8 | 14 | 471 |
| savfw | 9.4 | 13 | 144 |
| unccubanhistorical | 9.3 | 4 | 38 |
| uncill | 8.6 | 10 | 93 |
| adamsem | 8.4 | 9 | 2 |
| worldwartwodocuments | 8.3 | 7 | 607 |
| savsov | 8.2 | 12 | 7 |
| savlit | 8.1 | 12 | 67 |
| savedu | 8.0 | 6 | 23 |
| savref | 8.0 | 12 | 4 |
| rbctrv | 8.0 | 14 | 75 |
| rbcyeats | 7.9 | 13 | 2 |
| uncargentinianhistorical | 7.9 | 3 | 67 |
| savdp | 7.6 | 6 | 2 |
| savpol | 7.3 | 8 | 31 |
| prscr | 7.0 | 13 | 765 |
| savkad | 6.7 | 12 | 81 |
| uncvenezuelanhistorical | 6.4 | 4 | 167 |
| savmon | 6.3 | 8 | 36 |
| **unclibraries** | **6.3** | **11** | **28521** |
| civilwardocuments | 6.1 | 7 | 3990 |
| unchs | 5.9 | 12 | 1081 |

**Table 1 (cont'd.)**

| Collection Code | Use* | Age (Trimesters) | Items (as of May 2013) |
|---|---|---|---|
| savatq | 5.8 | 13 | 97 |
| worldwaronedocuments | 5.7 | 7 | 12227 |
| ncbio | 5.5 | 14 | 257 |
| civilwarbooks | 5.4 | 7 | 172 |
| unclsce | 5.4 | 12 | 27 |
| rbcgen | 5.3 | 12 | 27 |
| uncmus | 5.1 | 13 | 1507 |
| sirwalterraleighbooks | 5.0 | 7 | 20 |
| savjuv | 4.8 | 12 | 11 |
| savrel | 4.8 | 11 | 15 |
| juvenilehistoricalcollection | 4.7 | 5 | 1086 |
| ncfic | 4.6 | 14 | 106 |
| spandr | 4.6 | 14 | 6105 |
| ncral | 4.6 | 14 | 22 |
| savcin | 4.4 | 13 | 1 |
| statelibrarynorthcarolina | 4.4 | 7 | 3375 |
| ncreligion | 4.3 | 3 | 2180 |
| docsouth | 4.3 | 4 | 606 |
| asgii | 4.1 | 10 | 1146 |
| southernfolklifecollection | 4.0 | 2 | 21 |
| uncchileanhistoricalcollection | 3.9 | 2 | 62 |
| uncuruguayanhistorical | 3.9 | 1 | 15 |
| northcarolinarailroads | 3.8 | 5 | 163 |
| nclhof | 3.6 | 12 | 8 |
| universityofnorthcarolinalaw | 3.5 | 12 | 708 |
| uncmp | 3.3 | 12 | 4 |
| ncgen | 3.3 | 12 | 1298 |
| ncrel | 3.3 | 14 | 343 |
| ncgovdocs | 3.3 | 12 | 1292 |
| unclscps | 3.2 | 12 | 27 |
| rbccw | 3.1 | 13 | 399 |
| iassistquarterly | 3.0 | 9 | 84 |
| rbcshaw | 2.3 | 13 | 1 |
| mazarin | 2.2 | 13 | 18 |
| uncsog | 2.2 | 12 | 119 |
| populargovernmentunc | 1.5 | 6 | 472 |
| **AVG.** | **7.1681** | **10** | **1039.4** |

*Use is average monthly downloads per item, as defined above

**Table 2: Total Use Over Time**

| Month | Total Use* | "unclibraries" Use** |
|---|---|---|
| January 2009 | 8.6 | |
| May 2009 | 8.4 | |
| September 2009 | 6.9 | |
| January 2010 | 3.8 | 4.2 |
| May 2010 | 5.9 | 7.3 |
| September 2010 | 5.9 | 6.9 |
| January 2011 | 3.8 | 4.6 |
| May 2011 | 4.8 | 6.1 |
| September 2011 | 6.2 | 6.8 |
| January 2012 | 5.2 | 5.7 |
| May 2012 | 5.2 | 5.7 |
| September 2012 | 6.2 | 7.0 |
| January 2013 | 6.8 | 7.8 |
| May 2013 | 6.6 | 7.2 |

*Use is average monthly downloads per item, as defined above
** "unclibraries" was created in January 2010

**Table 3: Collection Use by Age (In Trimesters)**

| Trimester | Total Use* | unclibraries Use** |
|---|---|---|
| 1st | 7.6 | 4.2 |
| 2nd | 6.3 | 7.3 |
| 3rd | 6.6 | 6.9 |
| 4th | 5.9 | 4.6 |
| 5th | 5.8 | 6.1 |
| 6th | 6.5 | 6.8 |
| 7th | 7.2 | 5.7 |
| 8th | 8.7 | 5.7 |
| 9th | 9.1 | 7.0 |
| 10th | 7.6 | 7.8 |
| 11th | 8.3 | 7.2 |
| 12th | 8.6 | |
| 13th | 10.2 | |
| 14th | 10.2 | |

*Use is average monthly downloads per item, as defined above
** "unclibraries" was created in January 2010

**Table 4: Foreign Language Collections**

| Collection Code | Use* |
|---|---|
| annali | 50.9 |
| savcos | 13.7 |
| vargas | 12.2 |
| uncmexicanhistorical | 10.4 |
| savmil | 9.8 |
| savfw | 9.4 |
| unccubanhistorical | 9.3 |
| savsov | 8.2 |
| savlit | 8.1 |
| savedu | 8.0 |
| savref | 8.0 |
| uncargentinianhistorical | 7.9 |
| savdp | 7.6 |
| savpol | 7.3 |
| savkad | 6.7 |
| uncvenezuelanhistorical | 6.4 |
| savmon | 6.3 |
| savatq | 5.8 |
| savjuv | 4.8 |
| savrel | 4.8 |
| spandr | 4.6 |
| savcin | 4.4 |
| uncchileanhistoricalcollection | 3.9 |
| uncuruguayanhistorical | 3.9 |
| northcarolinarailroads | 3.8 |
| mazarin | 2.2 |
| **AVG.** | **8.8** |

*Use is average monthly downloads per item, as defined above

**Table 5: Shared Collections**

| Collection Code | Use* | Items (as of May 2013) |
|---|---|---|
| worldwartwodocuments | 8.3 | 607 |
| civilwardocuments | 6.1 | 3990 |
| worldwaronedocuments | 5.7 | 12227 |
| statelibrarynorthcarolina | 4.4 | 3375 |
| ncreligion | 4.3 | 2180 |
| asgii | 4.1 | 1146 |
| ncgovdocs | 3.3 | 1292 |
| **AVG.** | **5.1** | **3545.3** |

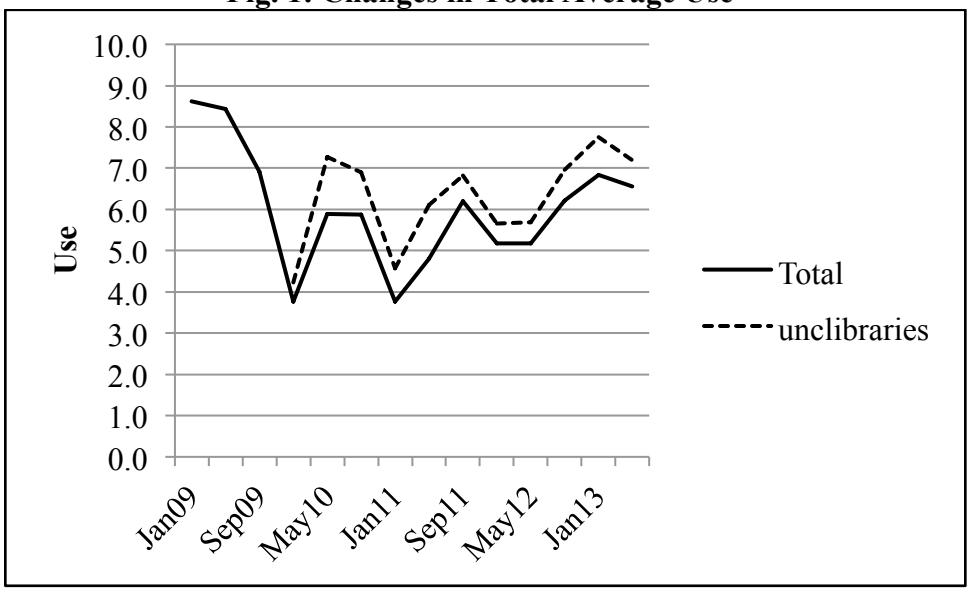
*Use is average monthly downloads per item, as defined above

**Table 6: Active Collections**

| Collection Code | Use* | Items Added** |
|---|---|---|
| unccubanhistorical | 9.3 | 1 |
| prscr | 7.0 | 65 |
| civilwardocuments | 6.1 | 480 |
| ncbio | 5.5 | 2 |
| rbcgen | 5.3 | 5 |
| uncmus | 5.1 | 182 |
| savjuv | 4.8 | 1 |
| savrel | 4.8 | 1 |
| juvenilehistoricalcollection | 4.7 | 36 |
| spandr | 4.6 | 607 |
| statelibrarynorthcarolina | 4.4 | 58 |
| ncreligion | 4.3 | 1036 |
| uncchileanhistoricalcollection | 3.9 | 47 |
| uncuruguayanhistorical | 3.9 | 15 |
| ncgen | 3.3 | 228 |
| **AVG.** | **5.1** | |

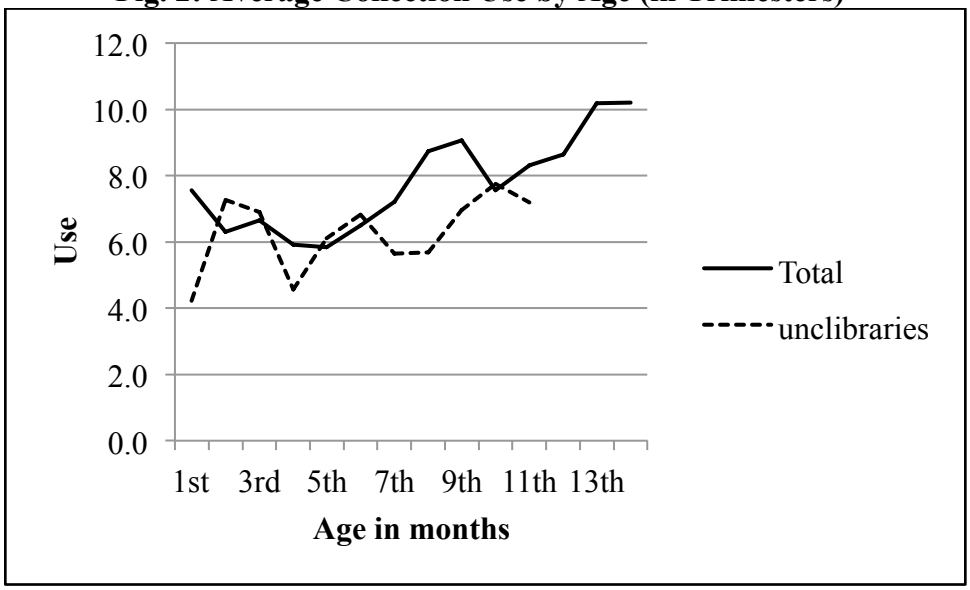
*Use is average monthly downloads per item, as defined above
**Items added between January 2013 and May 2013

**Fig. 1: Changes in Total Average Use***




*Use is average monthly downloads per item, as defined above

**Fig. 2: Average Collection Use by Age (in Trimesters)***




*Use is average monthly downloads per item, as defined above

**Fig. 3: Correlation of Collection Age to Collection Use***

| | Age (in Trimesters) | |
|---|---|---|
| Use | Pearson Correlation | .074 |
| | Sig. (2-tailed) | .547 |
| | N | 68 |

*Use is average monthly downloads per item, as defined above

**Fig. 4: Correlation of Collection Size to Collection Use***

| | Items | |
|---|---|---|
| Use | Pearson Correlation | -.065 |
| | Sig. (2-tailed) | .599 |
| | N | 68 |

*Use is average monthly downloads per item, as defined above

## Bibliography

Chmielewska, B. and Wròbel, A. "Providing Access to Historical Documents through Digitization." *Library Management*. 34:4/5 (2013). 324-334.

Court, N. "When and Why Is a Collection 'Hidden'? Awakening Interest in the Hornung Papers at West Sussex Record Office." *African Research & Documentation*. No. 121 (2013). 21-33.

Hill, H. and Bossaller, J. "Public Library Use of Free E-Resources." *Journal of Librarianship and Information Science*. 45:2 (2012). 103-112.

McCracken, E. "Description of and Access to Electronic Resources (ER): Transitioning into the Digital Age." *Collection Management*. 32:3-4 (2007). 259-275.

Sasser, P. "Sounds of Silence: Investigating Institutional Knowledge of the Use and Users of Online Music Collections." *Music Reference Services Quarterly*. 12 (2009). 93-108.

Schlosser, M. and Stamper, B. "Learning to Share: Measuring Use of a Digitized Collection on Flickr and in the IR." *Information Technologies and Libraries*. 31:3 (2012). 85-93.

Stewart, C. "Keeping Track of It All: The Challenge of Measuring Digital Resource Usage." *The Journal of Academic Librarianship*. 37:2 (2011). 174-176.