View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Carolina Digital Repository

Lauren A Geiger. To Bot or Not To Bot: An Exploratory Usage Study of Digitized Materials from the North Carolina Collection. A Master's Paper for the M.S. in L.S degree. April, 2019. 36 pages. Advisor: Anne T. Gilliland

Usage statistics have consistently been an integral part of collection development for libraries and archives, as they allow the staff to understand what kinds and types of materials their patrons want. As more of these institutions turn to digitization to make their materials accessible, usage statistics become more important as digital materials can be costly. In this exploratory paper, I tracked the usage of recently digitized materials from the North Carolina Collection at Wilson Library that were uploaded to both the Internet Archive and the HathiTrust. As the study lasted for two months and one month, respectively, I can draw no strong conclusions from the data, but rather a baseline for further studies for this collection.

Headings:

Archives

Digital libraries -- Collection development

HathiTrust

Internet Archive

Use studies -- Digital collections

TO BOT OR NOT TO BOT: AN EXPLORATORY USAGE STUDY OF DIGITIZED MATERIALS FROM THE NORTH CAROLINA COLLECTION

by
Lauren A Geiger

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

April 2019

Approved by

_____

Anne T. Gilliland

Table of Contents

Introduction

Usages studies. Usages statistics. Web analytics. Web metrics. In the library science and archival fields, these are a few terms describing a central question: What are my patrons using? Knowing of specific materials and subject areas their patrons use creates part of the foundation for a library or archive's collection development policy. However, as the different terms suggest, there is no one way to track usage. Some, like Elise T. Freeman in her article "In the Eye of the Beholder: Archives Administration from the User's Point of View," draw the practice back to understanding who uses the libraries and archives. Others believe the internet holds the key, and tools like Google Analytics will help professionals understand their patron's information needs, once librarians and archivists understand how to use these tools (Prom, Perrin et al, and Farney, 2011 and 2017). The ultimate goal of papers focused on usage statistics is to find methods in which libraries and archives can better help their patrons obtain their information needs.

In the library system at the University of North Carolina at Chapel Hill (UNC-CH), this goal is at the forefront of their mission, with their first guiding principle stating: "Identify users' needs, measure the impact of library services on users, and use results to improve student, faculty, and library performance; (University of North Carolina at Chapel Hill University Libraries, n.d.)." Identifying what users use allows libraries and archives to know what types of materials are popular and which are not. Popularity of materials can be an indicator which materials should be digitized and placed online. For a

variety of reasons, not all materials can be digitized; however, usage data allows the staff to see which materials may benefit from being online.

For this paper, I will assess the accuracy of the usage statistics from the North Carolina Collection by analyzing the usage of newly digitized materials during a two-month window and comparing this data to their current physical usage statistics. From this data, the staff should be able to determine if their current usage statistics can predict future usage statistics. The information gathered from this should answer the following question and sub-question:

1. How do usage statistics for both digital and physical materials help predict what materials are digitized?
    a. What factors cause difference in usage data between materials found through UNC-CH catalog and other websites? If so, are the librarians accounting for this difference?

Literature Review

For the past two decades, archivists have heralded the inevitable arrival of digitization and digital collections. Subsequent calls followed for born digital material. Digital archiving was no longer a choice within the archiving profession. With the decision of the National Archives and Records Administration to only accept electronic records by the end of 2022, the distant watch turns into a clear warning. Digital is here to stay, whether institutions are ready or not.

Digitizing materials and maintaining them like professionals do with physical objects is a large commitment of both time and money. Therefore, selecting the right materials in which to invest these resources is crucial. Usage statistics is one way to obtain evidence of materials used. An analysis of usage data can help create digitization and collection development guidelines for an institution's materials. However, usage and user studies are often mixed together, with user studies emphasized in the paper. This literature review will examine the mix of usage and user studies, other factors in collection development, and how the Internet changes how usage data is collected to explain why refocusing on materials, instead of users, will help professionals create more meaningful digital collections.

Uses of Usage Data

Usage statistics can be gathered through a variety of means: click analysis, page views, downloads, surveys, reference conversations, and so forth. In digital

environments, a variety of usage statistics can be gathered through web analytic tools, which collect, analysis, and deliver reports to the organization about the usage of their website (Web Analytics Basic). Depending on the type of web analytic tool, the organization can program goals into the software and the website will report on whether the site is meeting or failing these goals. These are tasks that could be achieved by a person, but the bot performs them at faster pace. This type of software may sound more appropriate for a business than archivist and, in a fashion, it is. As Christopher Prom discusses in "Using Web Analytics to Improve Online Access to Archival Resources," "there is nothing specifically archival about Web analytics (Prom, 162, 2011)." This is a tool that was designed for businesses, but libraries, archives, and businesses do share a common purpose: providing a service. Therefore, it is understandable that some of practices will overlap.

Specifically relating toward use, web analytic tools can help librarians, archivist, or business persons determine if their uses are from real persons or Internet bots. These bots perform automated tasks on the Internet that could be done by real person, but the bots work at a faster pace. Huntington, Nicholas, and Jamali in "Web robot detection in the scholarly information environment," found that tracking bots can be difficult as they take on more human-like effect in their searching patterns (2008). Of course, not all bots are bad. Some bots take the form of web crawlers, which can take snapshots of pages on the Internet. This information provides changes on how the Internet has changed overtime. Adeel Anjum in "Aiding Web Crawlers: Projecting web page last medication," conducts a couple of experiments to demonstrate how content creators can improve the HTML and HTTP of their page to help web crawlers pick up information from their page

once it has changed. The main problem is to determine how much bots affect usage statistics and how to relate them to usage, especially if they are crawlers and placing the information in another site to be used.

Like there are many types of businesses, there are many types of libraries and archives. Each one has its own goals and purposes, which drive what type of usage data they will need. In her 2011 master's paper, Molly Bragg conducted a survey to find out how libraries measured the use of digitized primary sources. Across forty-nine institutions there was no consistency in departments, positions, or usage statistics gathered to easily link a common framework analyzing usage data. An institution's uniqueness in materials and means complements the variety in web analytics. Dissonance in conformity allows these institutions to examine their own materials and the best way to maximize their digital presence.

Businesses, libraries, and archives also live or die by their budgets. The entire process of creating a digital collection, from initial starting costs to continual maintenance, can strain an institution's budget. Usage statistics help librarians with the budgeting process by showing them what types materials their users are attracted to, such as electronic books, videos, or sheet music. Usage statistics also help libraries look across the different mediums to see which themes or genres (History, War, German Literature, Agriculture) are sought after. In "Collection Development and Management: An Overview of the Literature, 2011-2012," author Kathleen Lehman notes how demand-driven acquisitions can decrease the costs of collection development and, depending on the method used, increase collaboration between institutions (2014). Concerning special collections, tracking which materials are being viewed and downloaded, the librarians

and archivists can make more accurate and precise choices for digitized materials and which subject areas deserve a higher priority for active development. Since usage statistics are an indirect form of observation, institutions do not have to worry whether a patron felt limited choices forced their use of a certain material (Prom, 2011). Therefore, the usage statistics that librarians and archivists receive from online materials may more accurately reflect what the patrons want, than usage statistics from in person interactions, where the staff may pull materials that the patron may not have wanted. This is not to say that in person interactions are less authentic than online ones, but rather that the patrons will not be guided by anyone other than themselves.

According to much of the literature, using views and downloads to measure usage is a weak form of analysis, as it does not give way to much interpretation. However, Tabatha Farney's article, "Click Analytics: Visualizing Website Use Data," demonstrates that basic information can be used to a solve a specific problem. Click analytics allowed librarians to re-design their website for greater usability. By using one specific tool from three different web analytic producers, Farney could "create easy to understand reports that instantly display[ed] where visitors are clicking on a webpage (147, 2011)." While this visualization tool helped their site, Farney did explain its weakness in determining the user's thought process while clicking on certain links. However, that weakness did not affect the strength of the overall report or the solution for the problem.

Some papers reflect on how using a simple or limiting usage statistic allow institutions to better observe outside factors affecting the usage of their collections. Midge Coats used page view data to determine what factors influenced user selection of digitized sheet music. With one method, she discovered that sheet music housed in larger

repositories had a greater chance of being selected and that the inclusion of an audio component did not influence a user's selection but cover art did (Coats, 2014). Melanie Schlosser and Brian Stamper examined page views to see if using a third-party site, Flickr and Knowledge Band, would increase their overall usage. Their data was too small to draw any significant conclusions, but they did learn that promotion influenced the items viewed (2012). Each paper acknowledged the weakness of the method, but like Farney's paper, proved that simple usage collection methods can provide substantial answers to specific problems. By keeping a narrow scope, these papers demonstrated the usefulness of targeted usage statistics.

The previous articles all discuss usage studies and their importance in helping librarians or archivists know what materials are used or how to better display the materials they possess. A better understanding of usage statistics help librarians and archivists with their collection development policies and overall general usage output. In the digital setting, collection development skills as critical as they librarian or archivist has to selective when choosing materials, because not everything needs to be digitized and the cost of maintaining these collections increase with the quantity and quality of said materials. Therefore, it is imperative that institutions understand which collections are receiving the most views and how users are finding these collections (Biswas and Marchesoni, 2016). Biswas and Marchesoni used Google Analytics to track their institution's materials to see which materials had the most views and what words the users were using to find these materials (2016, 24-5). They also discovered which search engines and sites their patrons used to find the materials (2016, 28). By knowing where a

user may find an institution's materials, that institution may use this information to understand how to present the materials on their own site, so they are more accessible.

Usage statistics can also alter how institutions view "popular" collections. Tali M. Beesley's Master paper, "Exploring Usage of Digital Collection via Web Analytic Tools," found that a collection's size does not automatically make it more popular. She discovered that one of the CDLA's (Carolina Digital Library and Archive) smallest sites was far more popular than its larger sites when comparing the number of views with how much material was presented in the collection (2012). This ratio of views to materials helped Beesley understand what types of which collections were used more than others, regardless of how many materials were in the collection. Of course, some sites may find that they are more interested in what is not being used and try to make those collections as strong as the popular ones (Waught, 2015). The authors of these papers utilize their usage statistics to help decide what materials should be digitized, further demonstrating how the Internet affects collection development policies.

Other Factors of Collection Development

Collection development is a major reason for usage studies. Usage statistics can help librarians shape their collections to best suite their users. However, usage statistics cannot hold final sway over what materials are put out into the world and which ones are not. As discussed in the literature, usage statistics are not perfect. If their influence is not balanced by other factors, they could cause a library to unintentionally favor one group over another, instead of looking at all of their users on equal footing (Mills, 2015). Other factors can include copyright and the institution's mission.

Copyright is arguably the largest factor determining if an item goes online, because copyright determines if an institution can legally allow others to use the material, or if they can make a copy at all. Archives, libraries, and special collections all hold a special exemption, Section 108 of the Copyright Act. This section states that since most copyright directly interferes with their mission of providing information to patrons free of monetary charge. Section 504 acts a safe harbor for libraries, archives, and special collections when they make fair use decisions. However, this doesn't mean that libraries are permitted to do as they please for all scenarios. They still have to observe copyright regulations and alert their patrons of the restrictions that occur when they digitize a copyrighted material for them . Preservation of materials also plays into this, as copyrighted materials can be digitized to help preserve them for future use (Hirtle, 2009). An institution's digital collection could fall under the fair use exemption, if their materials are considered low-risk, which means materials that the rights owner will be less likely to object to the materials being accessible for all without the user having to pay for it. High-risk materials can also be under the fair use exemption, if they are digitized for preservation or research, but archivists and librarians may need to prepare themselves to more actively defend the digitization as fair use.

Jean Dryden conducted a survey concerning the role of copyright in the selection of digitization and found that most institutions generally choose materials in the public domain (2014). By choosing to select low risk materials, they protect their users from being in violation of a copyright as they cannot control what users do with their digital materials. Dryden summarizes her interviewees' opinion on copyright in this short statement: "Access is our [archivist/librarians] business (68, 2014)." The interviewee's

answers allude that making materials accessible for their patrons is at the heart of their jobs. The importance placed on providing materials to users is heavily repeated throughout all of the literature in this review.

Returning to Freeman's call from 1984 to focus on users instead of materials, Freeman does mention that all institutions need to be wary of technology. While archivists and librarians understand that not all materials can be digitized for many of the reasons mentioned above, the allure of placing everything on the Internet does linger, as the Internet is deceptively all knowing and seemingly easy to use. Freeman warned her peers to "not become caught up in useless technologies or technologies that only make more quickly and expensively mistakes we have made manually (1984, 112)." One way to avoid this is to always remember why an institution exists and its goals and missions. These types of statements can save much time and frustration later on if an institution aligns its online presence with the major goals that were established for its physical collection. How an institution achieves their mission online may be different from how they interact in their physical space, but the values do not change.

Understanding copyright and an institution's original mission can help institutions create digital collections without the fear that they will be taken down. In "Should You? May You? Can You?," Janet Gertz describes the multitude of factors concerning why an object may be digitized and why it may not. Her first criteria for digitizing an object is its value, either on its own or in context. Her second reason for digitization is user demand for this material. Gertz argues that materials cannot be digitized solely because they will look pretty or because it may save them from a perceived use; people have to want to use them (2013, 8). Otherwise, they are just as safe and valuable sitting safely in the stacks.

Gertz follows Freedman's logic by believing that if an institution understands who its users are, then they will know what materials they are looking for (2013). However, in this digital age, it is hard to know who our users are as the Internet provides a thick blanket of anonymity. This is where usage statistic can balance out the value of the material, its copyright, and institution's mission.

Usage vs User Studies

Usage studies not only allow the staff of an institution to understand what their users are using, but if the staff themselves know their own collection. Laura Waught led an evaluation of the University of North Texas Digital Collections and Institutional Repository and found that graduate students using the university's digital library knew more about the collections than the staff and faculty at the university (2015). The study also found that across all disciplines or areas of study, the graduate students also preferred to work with digital materials (749, 2015). This type of information is vital for the Digital Collections, because if the staff does not understand the materials already in their digital collection, then it will be challenging for them to express what materials they want to use to the digital library's staff. This part of the study does lean toward the user vs the usage of the materials. However, Waught was able to gather some usage data as she learned which departments seemed more interested in the digital library than others. From this information, she and her team are planning on promoting the materials to those who are less interested in the digital library rather than those who are (749, 2015). Considering the cost and time of digitizing materials, it may seem strange that Waught would then target those who do not use the digital library. However, the goal of the research was outreach and educational opportunities (749, 2015). So, by focusing on

departments who do not use the digital library, Waugh can create new paths of collection development for her team, while establishing new relationships to gather these materials.

Waught's in-depth evaluation of the University of North Texas's digital library demonstrates why analyzing the data collected is important to the growth of an institution. However, not all usage studies complete this step. While developing a questionnaire for user-based evaluations of archives, Wend Duff, Jean Dryden, Carrie Limkilde, Joan Cherry, and Ellie Bogomazova noted that "archives gather data about their users from registration forms, informal conversations at the reference desk, and exit interviews, but archivists rarely analyze this data systematically to evaluate whether their services or systems meet the archives' goals and users' needs (145, 2008)." Rachel A. Fleming-May also noted that while illustrative information is good for quick answers, it usually lacks the substance to be used to create lasting change (2010). These studies are mainly focused on the usage of digital collection and gaining information from some of the web analytic tools mention in the Methods of Collecting Data section. Web analytics can't analyze the intentions behind why a researcher clicks on a link or why they would download it. A way to make usage data more understandable would be to interpret the usage statistics with other forms of metrics and analytics that could account for outlying factors (Perrin, 2017). The non-digital equivalent would be a reference conversation or a candid conversation about what the researcher is doing with the materials. Of course, that sentence implies that institutions with a higher rate of face-to-face interactions may have more accurate and in-depth usage statistics and use those to help establish a digital presence. However, Polona Vilar and Alenka Sauperl found that these types of institutions still had difficulties identifying what materials their patrons used to find

information (2015). They also had difficulty explaining what types of user were visiting their institution. This mix of "user" and "usage" was a problem I encountered while trying to examine the literature.

Even if the words "use" and "usage" are in the title of an article, the main focus was on identifying users and their habits. Much of the literature surrounding usage statistics and studies have a significant focus on figuring out the users of an institution and their research habits and material use to predict which materials they will use, and which format is the best way to present this information. The actual usage statistics of a collection comes in as a second or third point in collection development. Of course, this trend of mixing usage and users makes sense as one does not exist without the other. Understanding who uses a particular repository can help the staff understand what type of materials they may want and understanding what materials are being used helps the staff know in which area to focus their collections. OAIS (Open Archival Information System) crated a theory called Designated Communities, to help archives understand their user groups and how they could better help them find and use materials. This theory works well as long as archives understand that their designated communities must change as the archive changes. By not adjusting for changing communities, institutions run a risk of their materials not being used and their users being unable to find what they need.

Therefore, tracking usage of materials may help an archive or library staff member better help their users. By looking at solely at materials and not on users, the staff can see what is being used without any preconceived notions affecting their judgement. The inherent difference between user and usage becomes more prevalent in digital collections as the Internet creates anonymity for the user and broadens the field of

users. Therefore, it is difficult to apply user studies information to their digital collections as they may with physical materials. Tracking and analyzing the usage data from digital collections is far more useful to professionals in the library science field, as they create or refine digital collection development policies.

Methodology

Wilson Special Collections Library is one of the major libraries on UNC-CH campus as it houses UNC-CH's rare materials. In 2004, Joe A. Hewitt, University Librarian Emeritus, gave a speech for Wilson Library's 75th Anniversary, citing how the library had adapted over the last three quarters of a century. The speech was entitled, "Louis Round Wilson Library: An Enduring Monument to Learning." This title summarizes the goals that Wilson Library's staff and faculty hold not only for their patrons, but for themselves.

Each of the five special collections housed within Wilson collects usage statistics to better understand what their patrons require. The staff at the North Carolina Collection gathers their usage statistics from multiple software programs that track their two systems and from reference interviews with patrons. From these statistics, the staff compiled eight main categories from which patrons collect information: World War 1, Women, Agriculture, Education, African Americans, The University of North Carolina at Chapel Hill, Chapel Hill the town itself, and Public Health/Mental Illness (S. Carrier, personal communications, August 29, 2018).

For this study, I compared current physical usage statistics from the North Carolina Collection to recent online usage studies to see where they do and do not align. To gather this data, I viewed the page views and downloads from January 14 to March 14, 2019 for materials digitized and contributed to the Internet Archive and the HathiTrust to commemorate Public Domain Day, which was the expansion of the public

domain on January 1, 2019. Web analytics from both repositories provides further context for these views and downloads. This information combined allows me to assess whether or not the usage statistics from the North Carolina Collection are accurate predictors for what materials should be digitized and placed in third-party repositories.

Currently, the North Carolina Collection gathers usage statistics from multiple sources as their online systems do not sync in a way that they can aggregate all of their data at one time. According to Sarah Carrier, a research and instructional librarian at Wilson Library, the North Carolina Collection gathers its usage information from: different web analytic tools on various webpages, recording information through Sierra, and from reference interactions with patrons (personal communication, August 29, 2018). While this data covers a broad spectrum, the multiple sources could cause an overlap in this data, which would create inaccuracies. By closely examining a set of materials and knowing where the overlaps exists, more accurate data can be gathered. Once analyzed, this data will then allow the staff of the North Carolina Collection to know if they need to adjust their strategies for digital collection development or if they should continue with current methods.

Procedure

The collection period for materials on the Internet Archive was eight weeks, from January 14 to March 11, while the collection period for the HathiTrust was four weeks, from February 14 to March 14. Ideally, these collection periods would have been the same. However, the digitization and ingest period for both repositories resulted in the materials being launched at different times. As stated in my literature review, several studies criticize using views and download counts for usage studies, but this method will

allow the staff of the North Carolina Collection to establish a baseline for later usage statistics concerning digital collections. Schlosser and Stamper define a "view" for a digital object as a person viewing the webpage associated with the object or as person viewing the object itself (2012).

The Internet Archive's policy for collecting views and downloads aligns with Schlosser and Stamper and narrows their point further. The Internet Archive only counts an item as being "viewed" if the user interacts with the media of the object; simply clicking on an item from a search page does not constitute a view. According to their blog post, New Views Stats for the New Year, engaging the media of an item includes "experiencing the media through the player in the item page (pressing play on a video or audio player, flipping pages in the online book reader, emulating software, etc.), downloading files, streaming files, or borrowing a book (2018)." They further expand on viewing text items by explaining that a user must interact or "flip" through the book twice for it be a view (2018). This narrow definition alleviates some of concerns mentioned in the literature review about how accurately a library can document views. By counting and clustering specific media engagement, the Internet Archive's collected data demonstrates a higher level of intentional user engagement. The Internet Archive also tracks which media engagements are conducted by bots crawling through their materials and which ones are non-bots, furthering the accuracy of intentional engagement.

The HathiTrust's measurement for view and downloads count leans toward the first portion of Schlosser's and Stamper's definition. A click on the item is considered a view. Whereas the Internet Archive counts media engagement separate from the webpage

it is placed on, the HathiTrust allows the user to access the catalog record or the material itself. From there, a user can scroll through the material and download a page, section, or the entire material. These distinctions are noted in their web analytics. While a broad definition for media engagement can result in less accuracy of user intention, it does allow a library to know how many times their materials are being seen. The specific download counts also further an institution's knowledge of which type of materials are being used or seen the most. HathiTrust's web analytics also tracks where a user came from to access their materials. This is important as it allows the repositories that place materials in HathiTrust to see if those who access them are using their personal catalogs or if they are finding it from another source.

The media engagement was collected weekly from the Internet Archive, as they provided the information from a landing page. Therefore, I did not add media engagement views when collecting my data. The bot and non-bot data were gathered at the end of the collection period. The bot and non-bot media engagement is viewed holistically so as to provide context on the views, which is why it was not gathered at the same time. The data from the HathiTrust was collected after the first two weeks and then after the last two weeks. This material was collected less frequently, because, in order to find the materials on HathiTrust's Google Analytic page, I had to go to each item and retrieve the record number. Therefore, I did add to the view count for the item. My views for this section were removed from analysis.

Once collected, the data for each repository was sorted for analysis. Both repository's data were analyzed for total media engagement growth over their eight- and four-weeks period. The Internet Archive's total media engagement is then broken down

into bot and non-bot engagement. The materials will be broken into theme categories to keep in line with the categories provided by the North Carolina Collection, and compared to both the total media engagement and bot and non-bot engagement. The HathiTrust's data was analyzed differently as their algorithms for collecting engagement are different. The views and download counts for all materials were counted against one another to provide a better picture of intentional use. Only the downloads of complete materials, and not sections or pages, are used so the data is more comparable to the Internet Archive. Finally, how the users found the materials is analyzed to see if they came from a UNC-CH related page or another site. Together, the data from these analyses demonstrates if the materials usage aligns or differs with the physical usage results from the North Carolina Collection.

Ethics and Limitations

There are a few assumptions made in this study for analyzing and collecting the data. To summarize, they are:

- Non-bot interactions with the materials are being made consciously by a person
- The time period gathering this data will affect the data's accuracy

Currently, there is no way to know how many of the assumptions are true or false. However, these assumptions should not comprise the integrity of the data as page views and download counts are objective and the web analytics can identify bots verses non-bot interactions. How I will analyze them to find out what materials are being used is subjective and will contain some flaws. If I had a longer testing period and more time to analyze all of the ways a person could view these materials, then I could lessen my

assumptions and give a more thorough analyze. For this study, the limits are necessary for the scope of my question and the time I have to write this paper.

Findings and Discussion

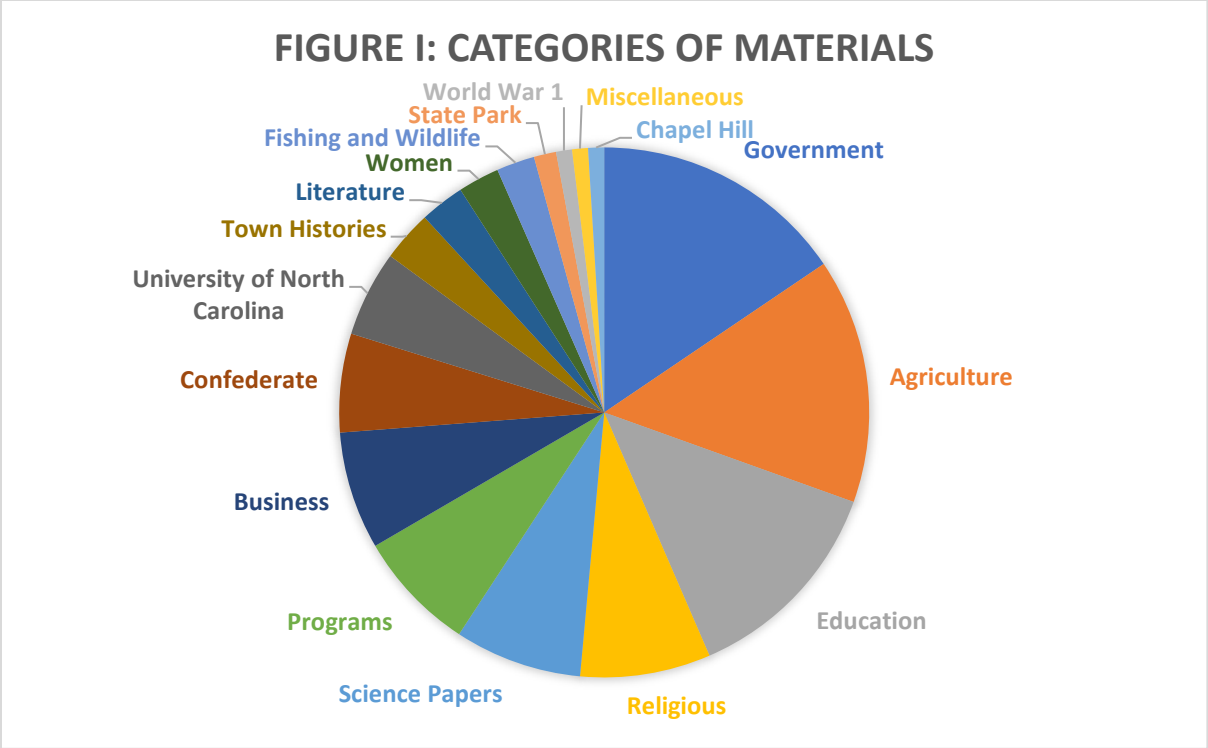
The materials in the Internet Archive received a total of 544 media engagement (views/downloads) for all 95 items uploaded. The total of non-bot media engagement is 297, which is 54.6% of the total. Just over half of the media engagement with the materials comes from non-bots. However, this does not mean that the bot views are not valid. As stated in the methodology, the Internet Archive only tracks engagement with the materials and not just people or bots clicking on the item. In their blog post, "New Views Stats for the New Year," the Internet Archive describes the bots as crawling through their materials, which might imply that these bots are web-crawlers. The blog post discusses how search engines like Google or Bing use these to ensure that their results are accurate and that they are displaying the correct materials.

However, search engines are not the only ones with crawlers. The Internet Archive itself uses web-crawlers on the Internet to gather materials for the Wayback Machine, which captures the history of the Internet. These crawlers may capture the main page of a website or explore several layers, depending on how it is programmed. This adds a layer of uncertainty as to where the intentional engagement lies. If these bots stem search engines, then their intentional engagement would be considered lower, as these bots search all of the internet. If they are from a specific institution that is tracking certain information, then this would imply a higher level of intentional engagement as the site wants to find this information and use the data collected. Without finding the source of the bots, it is difficult to tell where these engagements lie.

Of course, the non-bot engagement also causes problems. By saying non-bot, the Internet Archive means an engagement not created by a bot. This should imply that a real person is looking at these materials. However, this may not be the case. Today's bots are becoming smarter and more human-like in their movements, to the point where they can trick counters into thinking they are human and not bots. The Internet Archive cannot assure its users that their software is accurate enough to detect a bot acting like a human, so they use the term non-bot. For this paper, I am assuming that these non-bots are human and that their engagement with the materials are intentional. However, I will continue to call them non-bots to minimize any confusion over terminology and to maintain consistency with the Internet Archive.

Category Selection

The 95 materials are broken up into 17 categories: Government, Agriculture, Education, Religious, Science Papers, Programs, Business, Confederate, University of North Carolina, Town Histories, Literature, Women, Fishing and Wildlife, State Park, World War I, Miscellaneous, and Chapel Hill. The categories here are listed by the number of media engagements they received, from largest (80) to smallest (5). These categories were selected based off the type of material and the main subject. This mixing of types was necessary based both upon the materials digitized and to minimize the number of categories that would only contain one item. Figure I below visualizes the categories based upon media engagement.



Of the eight categories listed by Wilson as the most popular, only six of them appear here: Women, World War I, University of North Carolina, Agriculture, Education, and Chapel Hill. When sorting the materials in to categories, I based my selection on either the main theme running through the materials or by the type of material it was. If I were to base my selection on a secondary theme, then some of these materials could have fit into African American or Public/Mental Health, but it would not have been the most accurate description.

Media Engagement

The total media engagement for the Wilson categories are 191 engagements. Ninety-ninety of them, or 51.83%, are from non-bots. This ratio is slightly less than the total material ratio at 54.6% non-bot media engagement. The Wilson categories comprise 35.11% of the total media engagements and 33.33% of the total non-bot engagements.

Thirty-nine items comprise these six categories, which is 36.84% of the total 95 items. These categories were not consistent in their media engagement numbers, with Agriculture and Education gaining over double the total engagements of the remaining four categories. However, this is only one way of looking at the data. If one were to look at the ratio of total media engagement to non-bot engagement, then World War I, Women, and Chapel Hill have higher percentages of non-bot engagement, even though they had some of the lowest total engagements of all of the categories.

Table 1: Media Engagement

Category	Number of Items	Total Media Engagement	Non-Bot Engagement	Percentage of Difference Between Total Media Engagement and Non-Bot	Percentage of Total Media Engagement	Percentage of Total Non-Bot Engagement
Religious	7	41	22	53.66%	7.54%	7.41%
Confederate	4	31	21	67.74%	5.70%	7.07%
Town Histories	4	16	7	43.75%	2.94%	2.36%
Literature	7	46	28	60.87%	8.46%	9.43%
Government	15	80	35	43.75%	14.71%	11.78%
Business	6	37	22	59.46%	6.80%	7.41%
State Park	1	7	4	57.14%	1.29%	1.35%
Education	13	67	37	55.22%	12.32%	12.45%
Agriculture	17	77	32	41.56%	14.15%	10.77%
Programs	5	38	25	65.79%	6.99%	8.42%
University of North Carolina	5	24	14	58.33%	4.41%	4.71%
World War I	1	5	4	80%	0.92%	1.35%
Women	2	13	9	69.23%	2.39%	3.03%
Miscellaneous	1	5	3	60%	0.92%	1.01%
Fishing and Wildlife	2	12	9	75%	2.21%	3.03%
Chapel Hill	1	5	3	60%	0.92%	1.01%
Science Papers	4	40	22	55%	7.35%	7.41%
Complete Totals		**544**	**297**	**54.60%**		

When looking at total media engagement, the categories that gained the most views are Government, Agriculture, and Education with each 80, 77, and 67 engagements, with Literature, Science Papers, and Programs following with 46, 40, and 38 engagements. With only two of Wilson's pre-selected categories appearing, it could be argued that their current usage statistics do not work for the third-party platforms that have world-wide audiences. However, the total engagement contains engagements from bots. When looking at the percentage non-bot engagements to the total, then the results are different. The categories with the largest non-bot views are World War I, Fishing and Wildlife, Women, Confederate, and Programs. Agriculture and Government have less than 50% of their total engagements from non-bots, which completely flips the script.

A reason for this dramatic change is the number of items within each category. Government, Agriculture, and Education have the most items within them, 17, 15, 13, respectively, while World War I, Fishing and Wildlife, and Women have 1, 2, and 2, respectively. By averaging out the views based upon the items in the categories, the categories become far more even, with a majority averaging 5 engagements per item. The outliers for this viewing method are Science Papers (10 average), Programs (8 average), and Confederate (8 average).

Even with averaging out the materials, the bot engagements still have to be taken into account for. On a whole, the items were split about 50-50, with a few outliers. The outliers for an item with disproportionately high non-bot views were in Science Papers (19 total to 12 non-bot), Programs (15 to 11), Business (13 to 10). Government and Agriculture had some of the most disproportionately high bot views, with one item in Government having 11 total engagements and 4 non-bot engagements and two items in

Agriculture having 2 and 3 total engagements with no non-bot engagements. About half of the categories with more than one item had one or two items that gathered a large number of bot to non-bot engagement, such as 5 total to 1 non-bot or 4 total to 1 non-bot.

When looking across the different data manipulations, there are two categories that consistently appear: Science Papers and Programs. These two categories each contain four and five items within them, have at least four total and two non-bot engagements, and contain the two items with the most total and non-bot engagements. These two categories can be considered the most popular concerning non-bot engagements, followed by Women, Wildlife and Fish, Confederate, and World War I.

Looking at some of the individual items within these categories may shed some light on why the categories as a whole were popular. The Science Papers category contained two works by William Chambers Coker, who was a well-known biologist and mycologist, and a Kenan professor at UNC-CH during the early 1900s. The two works in this category, *The Clavarias of the United States and Canada,* and *The Saprolegniaceae: With Notes on Other Water Molds*, are standard reference texts for their respective fields. These popularity and necessity of these texts could account for their large media engagements. The most viewed item for Programs is *The Shawnee Trail Program: An Historical Pageant presented at Clarksburg, West Virginia, June 13 and 15, 1923*. This program was the final for Clarksburg's first year of community service. The Shawnee Trail was also a major trade and emigrant route across Missouri, Kansas, Oklahoma, and Texas and people may have confused this program with more historical information about the actual trail, or they might have been interested in Clarksburg, West Virginia.

Without speaking to the users, it is difficult to guess why they were viewing a specific item.

As for the other categories, World War I and Women were already identified by Wilson staff as major areas of research, so this accounts for part of their presence. Discussion over protection for wildlife and the United States' national parks has increased over the last few years. This could explain why the Wildlife and Fishing category obtained more media engagements than other categories. The discussion over the place and rights of Confederate monuments has also increased over the last few years. This conversation continues to be very loud at UNC-CH, where students pulled down the Confederate monument known as "Silent Sam" just before the start of the Fall 2018 semester. The debate on these monuments and the publications relation to UNC-CH could explain why this category has many media engagements. Again, without speaking to those who viewed these items, it is difficult to gather why these materials were popular.

Weekly Growth

The Internet Archive's eight-week growth period demonstrates that it takes time for materials to gain traction. I did not see statistically viable materials for the majority of the collection until Week 3. Six of the ninety-five items did have views during the first two weeks, but the remaining eighty-nine did not see results until Week 3 or Week 4. Table 1 shows the weekly growth of the materials broken into their categories, the number of items within each category, the total weekly growth of the categories, and the percentage of growth from Week 3 to Week 8.

Table 2: Weekly Growth

Category	Items in Category	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Total Growth
Religious	7	0	0	18	22	32	32	37	41	128%
Confederate	4	2	2	8	16	21	22	27	31	287.50%
Town Histories	4	0	0	2	3	7	10	14	16	700%
Literature	7	0	0	23	26	34	36	44	46	100%
Governmet	15	1	1	46	50	67	67	79	80	73.90%
Business	6	0	0	15	18	28	30	35	37	147%
State Park	1	0	0	1	3	5	6	7	7	600%
Education	13	2	2	39	44	60	61	67	67	72%
Agriculture	17	0	0	45	52	70	74	75	77	71%
Programs	5	0	0	13	20	30	33	37	38	192%
Univeristy of North Carolina	5	4	4	13	15	22	22	22	24	85%
World War I	1	0	0	3	4	5	5	5	5	67%
Women	2	0	0	5	7	10	10	12	13	160%
Miscellaneous	1	0	0	1	1	2	2	4	5	400%
Fishing and Wildlife	2	0	0	7	9	11	11	11	12	71%
Chapel Hill	1	0	0	2	2	3	3	5	5	150%
Science Papers	4	8	8	24	27	32	35	38	40	67%

Table 2 shows that five of the seventeen categories did see growth during Week 1 and Week 2, but these numbers are outliers compared to the rest of data. The Science Papers category is the greatest of these as two of the six individual items fall into this category, which partly explains why their numbers are at least double the other outliers for Week 1 and Week 2. The percentage of growth for each category is large, because the numbers for each category are small. As these numbers were collected after eight weeks, they are preliminary results. Examining materials in another two or four months, or one year should produce different results as promotional materials for these items will be created, which may affect their media engagement count.

Week 3 for this study fell from January 28 to February 4. The Association of Research Libraries, which is a nonprofit organization of research and university libraries

in the United States and Canada, published a list of public domain materials digitized to take advantage of the public domain expansion a few weeks earlier, but UNC-CH had not listed any items on it during the time of my study. Therefore, it is unlikely the ARL's promotional material affected the jump in media engagement. No other significant event surrounding the publishing of public domain materials happened around this time. As all materials did have bots interacting with the material, it could explain why these material's media engagement jumped at the same time. When gathering bot and non-bot engagement, a bot engagement was generally listed first. I added views to some of the materials during Week 6 and 7 as I was finding images to be used for promotional materials. I did count these media engagements within all of my results, because if I did not look at these materials then another person from UNC-CH would have. Therefore, my media engagements do not taint my data.

The materials uploaded to the HathiTrust, an international digital library, did not produce any results as there were no views for the items uploaded during the four-week observation period. If a user were able to access HathiTrust's Google Analytics, then they would find a view or two on the catalog records. Those views were made by me when I collected the record number for the materials, so they do not count. While not having any viable data is disappointing, it is not unexpected. The materials from the Internet Archive did not gain real, statistically usable data until Week Three. Therefore, the materials not having any views aligns with the pattern of items taking several weeks to be noticed. Also, UNC-CH's library system social media team had not promoted the material's digitization during the collection period, so these results potentially reflect how long it takes for materials to be discovered.

Conclusion

Overall, this preliminary study of usage for materials from the North Carolina Collection at Wilson Library identified baseline numbers and possible trends. More information could be gathered from speaking to users about why they chose to look at these materials. However, that is not always possible with the Internet. Therefore, librarians and archivists need to know how to interpret the data without user input. Finding out how many views are from bots and how many are not appears to be a good place to start, if we assume this preliminary data's trends will continue to be accurate. When viewing the media engagement for the 1923 materials uploaded to the Internet Archive and the HathiTrust, there are not many conclusions that can be made. Eight and four weeks is not long enough to make real, lasting conclusions for the data collected. However, I can use the data to loosely predict how these materials may continue to grow.

The Google Analytics for the HathiTrust did not reveal any media engagement during the four weeks it was live. Based upon the media engagement timeline for the same materials uploaded to the Internet Archive, these materials should start to gain views over the next few weeks. The HathiTrust contains millions of materials from partners around the world, so 95 new items that were just released to the public may take some time to catch one. Although, with the links to these items live in the UNC-CH library's catalog, this may change sooner than later.

If I assume these initial patterns to be correct, then the Science Papers and Program categories in the Internet Archive should continue to gain the most non-bot

media engagement, followed by World War I, Women, Fishing and Wildlife, and Confederate. If the overall materials gather more non-bot than bot views, then this trend could be disrupted by Government, Agriculture, Education, or Literature taking over as the most non-bot media engagements. These materials should be viewed again in another 4 months, and then 6 months to track where they are after half a year and year. Of course, this all depends on if the bot engagement is treated as a less intentional view than one of a non-bot. If these non-bots are web-crawlers that are depositing these materials or their location into a repository where others can access them, then their intentional engagement may increase. All of these ideas are excellent starting points for furthering this research and gathering more credible data.

Bibliography

Anjum, Adeel & Anjum, Adnan (2002, December). Aiding web crawlers: projecting web page last medication. *2012 15th International Mulitopic Conference*, 245-52. https://doi.org/10.1109/INMIC.2012.6511443

Beesley, Tali M. (2012, April). *Exploring Usage of Digital Collections via Web Analytics Tools*. University of North Carolina at Chapel Hill.

Bragg, M. (2011, December). *"There is Always More That Can be Done": A Survey Investigating Libraries' Measurement of Digitized Primary Source Use*. University of North Carolina at Chapel Hill.

Biswas, P., & Marchesoni, J. (2016). Analyzing Digital Collections Entrances: What Gets Used and Why It Matters. *Information Technology and Libraries*, 35(4), 19-34. https://doi.org/10.6017/ital.v35i4.9446

Boyle, J. (2006). *The Public Domain: Enclosing the Commons of the Mind*. New Haven: Yale University Press. Retrieved from http://ebookcentral.proquest.com/lib/unc/detail.action?docID=3420630

Coates, M. (2014). Because You're You: factors influencing item selection in a digital sheet music collection. *The Electronic Library*, 32(6), 884–97. https://doi.org/10.1108/EL-09-2012-0116

Dryden, J. (2014). The Role of Copyright in Selection for Digitization. *The American Archivist*, 77(1), 64–95.

Duff, W., Dryden, J., Limkilde, C., Cherry, J., & Bogomazova, E. (2008). Archivists' Views of User-based Evaluation: Benefits, Barriers, and Requirements. *The American Archivist*, 71(1), 144–66. https://doi.org/10.17723/aarc.71.1.y70837374478t146

Farney, T. A. (2011). Click Analytics: Visualizing Website Use Data. *Information Technology and Libraries*; Chicago, 30(3), 141–48.

Fleming-May, R. A., & Grogg, J. E. (2010). Assessing Use and Usage. *Library Technology Reports;* Chicago, 46(6), 5–10.

Freeman, E. T. (1984). In the Eye of the Beholder: Archives Administration from the User's Point of View. *The American Archivist*, 47(2), 111–23.

Gertz, J. (2013). Should You? May You? Can You? *Computers in Libraries*, 33(2), 6–11.

Hirtle, P. B., Hudson, E., & Kenyon, A. T. (2009). *Copyright and cultural institutions: guidelines for digitization for U.S. libraries, archives, and museums*. Ithaca, N.Y: Cornell University Library.

Hunting, Paul, Nicholas, David, & Jamali, Hamid R. (2008, October). Web robot detection in the scholarly information environment. Journal of Information Science 34(5), 726-41.

Lehman, K. A. (2014). Collection Development and Management: An Overview of the Literature, 2011-12. *Library Resources & Technical Services*; Chicago, 58(3), 169–77.

Mills, A. (2015). User Impact on Selection, Digitization, and the Development of Digital Special Collections. *New Review of Academic Librarianship*, 21(2), 160–69. https://doi.org/10.1080/13614533.2015.1042117

Mission Statement. Retrieved from https://library.unc.edu/about/mission/

Perrin, J. M., Yang, L., Barba, S., & Winkler, H. (2017). All that glitters isn't gold: The complexities of use statistics as an assessment tool for digital libraries. *The Electronic Library*, 35(1), 185–197. https://doi.org/10.1108/EL-09-2015-0179

Prom, C. J. (2011). Using Web Analytics to Improve Online Access to Archival Resources. *The American Archivist*, 74(1), 158–84.

Vilar, P., & Šauperl, A. (2015). Archives, Quo Vadis et Cum Quibus ?: Archivists' self-perceptions and perceptions of users of contemporary archives. *International Journal of Information Management*, 35(5), 551–60. https://doi.org/10.1016/j.ijinfomgt.2015.06.001

Waugh, L., Hamner, J., Klein, J., & Brannon, S. (2015). Evaluating the University of North Texas' Digital Collections and Institutional Repository: An Exploratory Assessment of Stakeholder Perceptions and Use. *The Journal of Academic Librarianship*, 41(6), 744–50. https://doi.org/10.1016/j.acalib.2015.08.007