[Engaging the Collection: By the Numbers]

HathiTrust Growth and Usage in 2017

Angelina Zaytsev

February 2018

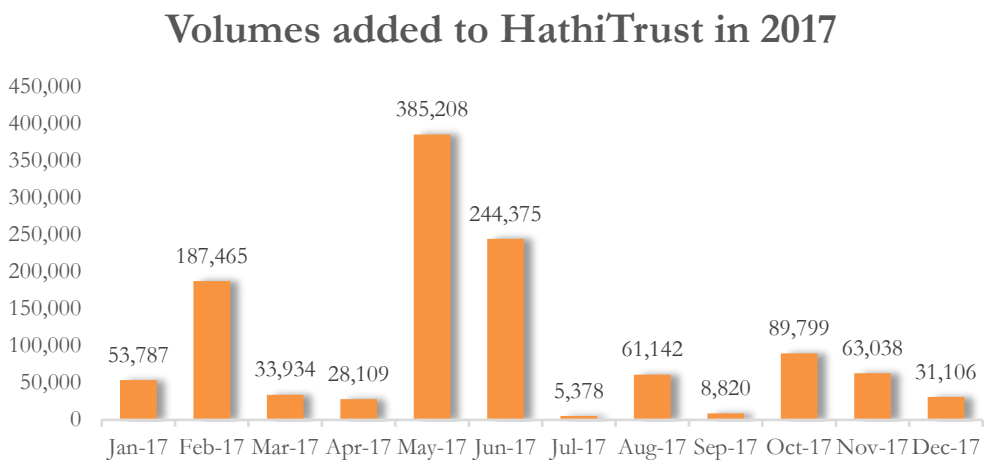
Contents

Growth of the HathiTrust Collection in 2017

We hit two big milestones in 2017, reaching 15 million works in February and 16 million in December. As of January 1, 2018, there were 16,008,348 works in the collection, and 1.2 million were added over the course of 2017.

1.2 million works were added to the collection

The chart below depicts the growth of the HathiTrust collection. As can be seen, the collection grows in irregular starts and stops throughout the year. Contributing libraries choose when and which volumes to contribute.



These were contributed by 37 HathiTrust partner libraries, 27 of which contributed over 100 works. The table below shows the top 10 contributing libraries.

37 members contributed to the growth

HathiTrust Member	Works contributed in 2017
University of Virginia	533,463
University of California	209,840
Northwestern University	140,580
University of Illinois, Urbana Champaign	82,427
University of Michigan	60,900
University of Minnesota	60,253
The Ohio State University	51,706
Michigan State University	11,118
New York Public Library	10,751
Getty Research Institute	5,817

The University of Virginia deposited a large collection of post-1922 works

The University of Virginia in particular deposited a significant collection, choosing to deposit content published after 1922. Of those 533,000 works, over 9,000 were manually reviewed for copyright status, and 4,000 or 45% of the reviewed works were determined to be in the public domain.

Opening Works in the Collection

In 2017, we continued the Copyright Review Program, through which staff at our member institutions volunteer to do manual copyright. The main focus in the last year has been on works published in the United States, either monographs published between 1923-1963 or publications of state and local governments. Volunteers also reviewed a smaller number of materials that were published in the United Kingdom, Canada and Australia. About 58% of works reviewed were determined to be in the public domain.

27,634 works were opened through the Copyright Review Program

	Determinations	Works opened
Published in the United States	43,096	24,419
Published in the United Kingdom, Canada, Australia	4,286	3,215
Total	47,382	27,634

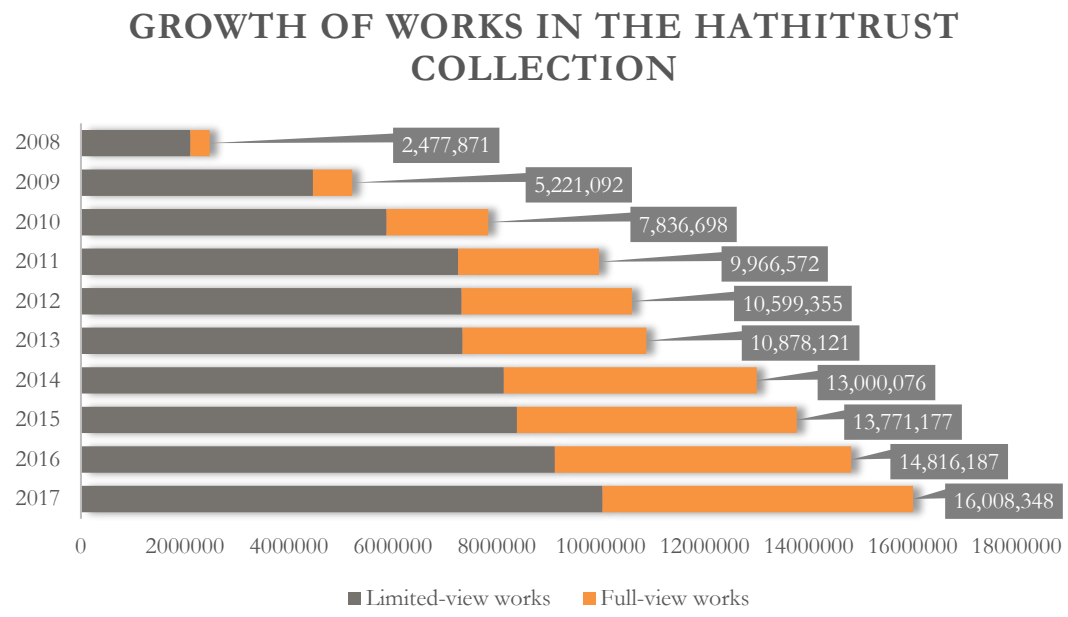
5,705 works were opened with permission of the author

In addition to manual copyright review, works may be opened to the public by permission of copyright holders. Over the course of 2017, HathiTrust staff processed 69 agreements. As a result, over 5,700 previously limited-view works were made available, 92% of which received a Creative Commons license that allows all users to download and reuse content.

Collection Changes Over Time

The following graph depicts the growth of works in the collection since 2008. Each year is broken down into limited-view works (in gray) and full-view works (in orange).

Comparison of volumes added versus access status



The percentage of open works has changed over time

The percentage of open works has fluctuated over time. This is as a result of changes in the collection characteristics (e.g., libraries may be more willing to digitize pre-1923 works), manual copyright review projects, and growing capabilities at HathiTrust to understand and handle complex copyright.

Usage of the HathiTrust Collection in 2017

This report looks at engagement of different user groups

In this report, we are primarily attempting to track various indicators of engagement in order to start to understand how well we are meeting the needs of users. Below we begin by looking at all users of the HathiTrust Digital Library and how those engagement metrics may vary based on different factors. Then we look into two separate subgroups, members and genealogists, and compare the activity of those users to all users.

One important qualifier about the following data: since the release of last year's report, we have discovered that HathiTrust regularly exceeds the number of permitted hits for Google Analytics Standard. The Standard version has a data limit of 10 million hits per month, and the HathiTrust account receives over 22 million hits per month. We can't use the Google Analytics data to get exact counts, but we can still rely on it for general trends and to understand user activity.

Indicators of high engagement may include:
- *Return visits*
- *Low bounce rates*
- *High numbers of pages viewed per session*
- *Longer session durations*

All Users[i]

Google Analytics provides several reports that help measure the variable engagement of users. One of these, the "New vs Returning" report, breaks users into "new visitors" who have never visited a site before and "returning visitors" who have visited a website on that specific computer or mobile device at any point in the past. This report tracks a few different metrics including bounce rate (i.e., how many users leave the site immediately after arriving), the average number of pages a user views in a session, and the average length of a session duration. The following table shows these numbers for all users to the HathiTrust website.

As can be seen below, new users leave the HathiTrust website immediately after arriving 40% of the time. This suggests a number of things that could be happening: new visitors find the content they are looking for immediately (which is what we're hoping for); they are bewildered by the HathiTrust interface; or they don't understand our access restrictions, particularly in the context of an Internet of freely available content.

New users bounce 40% of the time

Returning visitors have more positive metrics: the bounce rate is significantly lower, they view more pages per session, and their average session duration is much longer than new users.

	Bounce Rate		Pages / Session		Avg. Session Duration	
	New Visitors	Returning Visitors	New Visitors	Returning Visitors	New Visitors	Returning Visitors
All Users	39.38%	16.42%	11.24	24.40	0:03:58	0:11:30

The metrics above can change depending on other circumstances. One of those circumstances is how users arrive at the HathiTrust site.

Users referred by search engines are less likely to engage

From search engines: Users arrive primarily from search engines, with 45% of session traffic arriving via searching at sites like Google, Bing, Yahoo, etc. However, this traffic noticeably has a higher bounce rate than any other traffic source, except for traffic that arrives via email links. Users arriving via web search engines also tend to view fewer pages in each session and stay on the site for shorter periods of time.

Users who are referred by other sites have higher engagement levels (perhaps because many come from library catalogs)

Referred from another website: Users who have been referred from websites (including library catalogs) or who arrive directly at the hathitrust.org website tend to have lower bounce rates, view more pages in a session, and have longer sessions. For these users, the low bounce rates may be explained by different expectations than users who arrive from search engines. We know that a large percentage of our referral traffic comes from library catalogs and academic websites, and users who are referred may already be expecting the familiar library search interface.

Direct Traffic: Users who arrive directly at the hathitrust.org have, in many cases, visited the site previously and also know what to expect, resulting in positive metrics that indicate higher engagement. These include users who type hathitrust.org into the address bar; have saved bookmarks to HathiTrust; or who click a link in a document that doesn't live online (e.g., a PDF or Word document).

How users arrive	% of Total Sessions	Bounce Rate	Pages / Session	Avg Session Duration
From search engines	45.77%	43.91%	14.10	0:05:32
Referred from another website	33.05%	13.17%	21.34	0:08:57
Direct traffic	19.79%	23.79%	16.56	0:08:36
Via social media links	1.40%	12.10%	16.12	0:05:15
From email links	0.00%	44.66%	13.64	0:04:21

The most popular books visited by our entire group of users are as follows. As can be seen below, these range across a variety of genres, topics, and formats and fully represent the diversity of our entire group of users

Top 10 Books for All Users in 2017
Quicksand, by Nella Larsen.
Handbook of marks on pottery & porcelain, by W. Burton and R. L. Hobson.
America is in the heart, a personal history, by Carlos Bulosan.
Representative men and old families of Rhode Island; genealogical records and historical sketches of prominent and representative citizens and of many of the old families, v.2.
Ḥayāt al-ḥayawān al-kubrá, by Damīrī, Muḥammad ibn Mūsá.
History of wages in the United States from Colonial times to 1928. Revision of Bulletin No. 499 with supplement, 1929-1933.
Representative men and old families of Rhode Island; genealogical records and historical sketches of prominent and representative citizens and of many of the old families, v.3.
The visitations of the county of Devon : Comprising the herald's visitations of 1531, 1564, & 1620 / With additions by Lieutenant-Colonel J. L. Vivian.
Return to life through contrology, by Joseph H. Pilates.
Peterson's magazine, v.99-100.

A wide assortment of titles

In the following two sections, we consider two subgroups, members and genealogists. The data above for all users establishes a baseline for usage and provides us with the opportunity to compare engagement against the average behavior for the HathiTrust website. Keep in mind that the data for members and genealogists is folded into the data for all users.

We can track member usage by looking at login data

Members

Members are our core constituents. Students, staff, faculty and alumni of partner institutions are able to log into HathiTrust to get access to member privileges, notably full pdf downloads of public domain books. The primary way to track member usage is to look at users who log in with their partner institution accounts. (For simplicity, this report uses "members" to mean "users that are affiliated with a HathiTrust partner institution.")

Members who log in have high levels of engagement

The goal in tracking login data isn't really about increasing logins - it's about increasing end users' awareness that they are eligible for member services and ultimately about meeting end users' research needs.

The table below compares member usage to all users. For members, we see high usage, as characterized by low bounce rates, high numbers of pages viewed in a session, and average session durations.

	Bounce Rate		Pages / Session		Avg. Session Duration	
	New Visitors	Returning Visitors	New Visitors	Returning Visitors	New Visitors	Returning Visitors
All Users	38.63%	16.62%	11.42	24.77	0:03:58	0:11:27
Members	0.00%[ii]	0.35%	50.28	30.20	0:20:08	0:15:30

At the campus level, we can track 2 other groups of members:
1. Those who visit HathiTrust while on campus
2. Those who are referred from their university's websites

For HathiTrust members who wish to know more about their users, there are two other options for tracking usage, in addition to tracking logins as described above. We can also look at users who access hathitrust.org from an Internet network managed by their university, and we can track users who are referred to HathiTrust from a website managed by their university.[iii] These groups of users encompass the broader world of member-affiliated users who aren't being tracked in the login data above.

The following table shows some of the engagement metrics for four different campuses, University of California at Berkeley, University of Michigan, Harvard University, and George Mason University, as well as the numbers for all users.iv

	Bounce Rate	Pages / Session	Avg. Session Duration
All users	29.32%	17.01	0:07:16
University of California, Berkeley	17.10%	19.51	0:10:52
University of Michigan	6.04%	28.52	0:09:28
Harvard University	11.40%	17.55	0:07:22
George Mason University	21.79%	13.34	0:05:59

How many users log in when using their campus Internet?

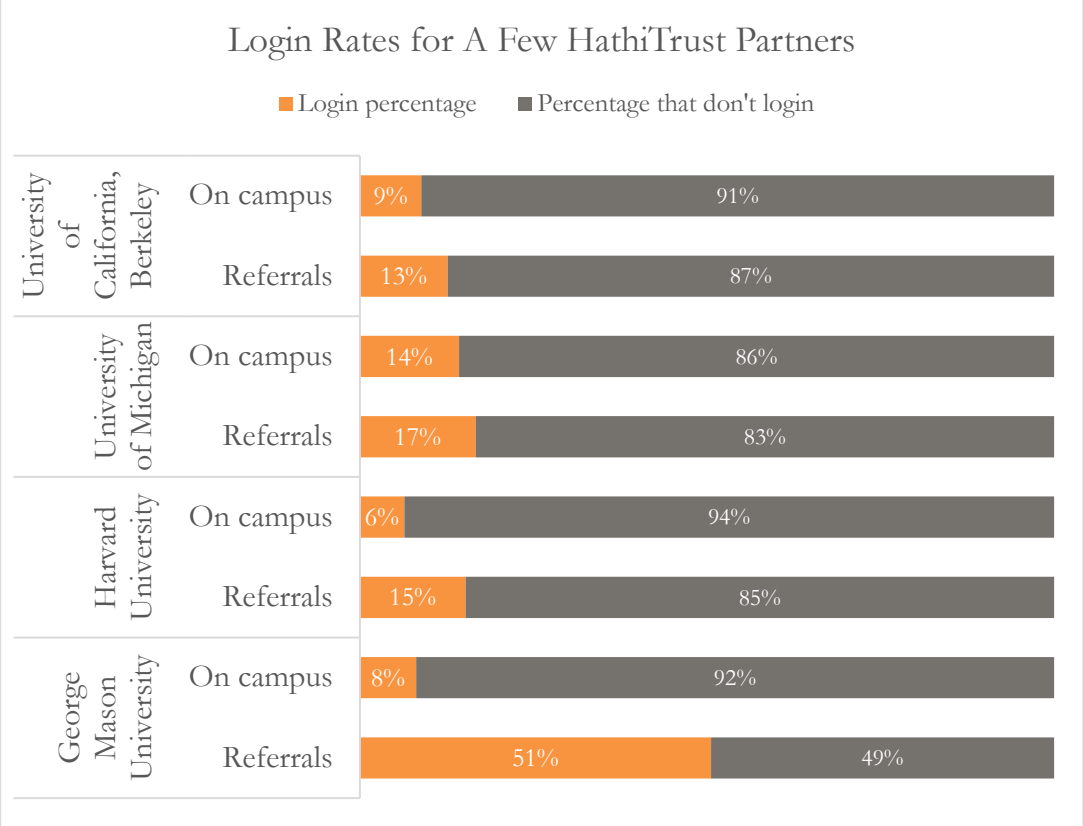
How many users log in after being referred from their library catalog?

Now what if we wanted to compare on campus access and referred users to logins? Some users who access HathiTrust while on a campus network or are referred from their university's websites log in, but many don't. We need to understand the size of that gap because it points at two problems:

- Many members may not know that they need to log in to receive member privileges. (E.g., some users may assume if they visit HathiTrust while in a library building, they are eligible for full access privileges.)
- The HathiTrust collection itself may not be meeting the needs of member users due to gaps or unavailable content, so they don't have a reason to log in.

For each of the four example universities, we can look at the rates for users that log into HathiTrust versus users who don't. The following chart compares login rates for users

who access HathiTrust from their campus network (i.e., "on campus") against users who are referred to HathiTrust from their university's websites (i.e., "referrals").[v] For both characteristics, the percentage of users who logged in is displayed in orange, and the percentage of users who didn't log in is in gray. This covers the time period October 10,2017 through December 31, 2017[vi], in contrast to the data in the rest of this report.



Data covers the period Oct 10, 2017-Dec 31, 2017

Looking at data for 4 example partners

Users who are referred to HathiTrust have higher login rates

For the first three institutions, login rates range between 6 and 17 percent. Referrals from websites managed by the university tend to have a higher percentage of logins. This is likely because most referrals tend to come from library catalogs, when a user is in "research mode." George Mason University (one of our more recent partners that joined in 2016) has a surprisingly high login rate, with 51% of users logging in who are referred from gmu.edu websites. It is hard to know what is causing this high rate, but George Mason University should keep doing what it's doing!

Good job, George Mason University!

The books visited by our logged-in member users are listed as follows. As can be seen, there is no overlap with the above list of books accessed by all users.

Top 10 Books for Members in 2017
Pin money; a novel. By the authoress of "The manners of the day" ... v.1.
The World almanac and encyclopedia. 1916.
Pin money; a novel. By the authoress of "The manners of the day" ... v.3.
Area handbook for the Republic of Turkey [by] Thomas D. Roberts.
Über Land und Meer. : [Deutsche illustrierte Zeitung]. ... 1882 pt.1.
Über Land und Meer. : [Deutsche illustrierte Zeitung]. ... 1882 pt.2.
The Saturday evening post,. v.192:no.44-48 (1920).
L'Italia nei cento anni del secolo XIX (1801-1900) giorno per giorno illustrata ... [v.2:pt.1].
With cavalry in 1915, the British trooper in the trench line, through second battle of Ypres, by Coleman, Frederic Abernethy.
Encyclopedia of American Quaker genealogy, by William Wade Hinshaw. v.1.

Lots of serials in this list

Genealogists

Genealogists are heavy users of the HathiTrust collection

Another group of users that we know about and want to track as a separate group are genealogists. We know from frequent user feedback and referrals that HathiTrust is a source of data for some individuals who are researching their family histories.

We can identify some genealogists based on referral websites

One way to identify these users in Google Analytics is to track sessions where users are referred from genealogy websites. This method won't capture the activity of users who go directly to the HathiTrust website or arrive through other methods, but it will still allow us to glean some data about the behavior of some genealogists.

Genealogists tend to be more engaged users than our larger group of all users, in the three main categories that we can track this. Bounce rates are lower, genealogists view more pages per session, and the average duration tends to be longer as well. Notably, however, the numbers of pages viewed per session and the average session duration for returning genealogists and all users are pretty close, suggesting the most important category to track is whether a user returns or not.

	Bounce Rate		Pages / Session		Avg. Session Duration	
	New Visitors	Returning Visitors	New Visitors	Returning Visitors	New Visitors	Returning Visitors
All Users	38.63%	16.62%	11.42	24.77	0:03:58	0:11:27
Genealogists	14.84%	6.89%	21.34	28.28	0:06:54	0:11:44

Interestingly, there is minor overlap between the top titles accessed by genealogists and members. Both groups visited The Encyclopedia of American Quaker genealogy, by William Wade Hinshaw. This suggests, perhaps, that member users overlap with genealogists, and that members are logging in for a variety of reasons, including for personal research into their family histories.

Top 10 Books for Genealogists in 2017
Aldermans in America, by Parker, William Alderman.
Pennsylvania archives, edited by Thomas Lynch Montgomery under the direction of the Secretary of the Commonwealth. ... v.6.
Patent rolls of the reign of Henry III. Preserved in the Public record office. Printed under the superintendence of the Deputy keeper of the records. Pub. by authority of His Majesty's principal secretary of state for the Home department, v.5.
Encyclopedia of American Quaker genealogy, by William Wade Hinshaw, v.1.
A history of the Seymour family : descendants of Richard Seymour of Hartford, Connecticut, for six generations, compiled and arranged for publication under the direction of George Dudley Seymour, by Donald Lines Jacobus
Encyclopedia of American Quaker genealogy, by William Wade Hinshaw, v.6.
Calendar of inquisitions miscellaneous, Chancery, preserved in the Public Record Office, prepared under the superintendence of the Deputy Keeper of the Records, v.3.
Calendar of inquisitions miscellaneous, Chancery, preserved in the Public Record Office, prepared under the superintendence of the Deputy Keeper of the Records, v.1.
Roster of the Confederate soldiers of Georgia, 1861-1865, v.1.
Names of foreigners who took the oath of allegiance to the province and state of Pennsylvania, 1727-1775, with the foreign arrivals, 1786-1808. Edited by William Henry Egle, M.D.

Unsurprisingly, family histories and records are the top titles for genealogists

Return Visits for All Users, Members and Genealogists

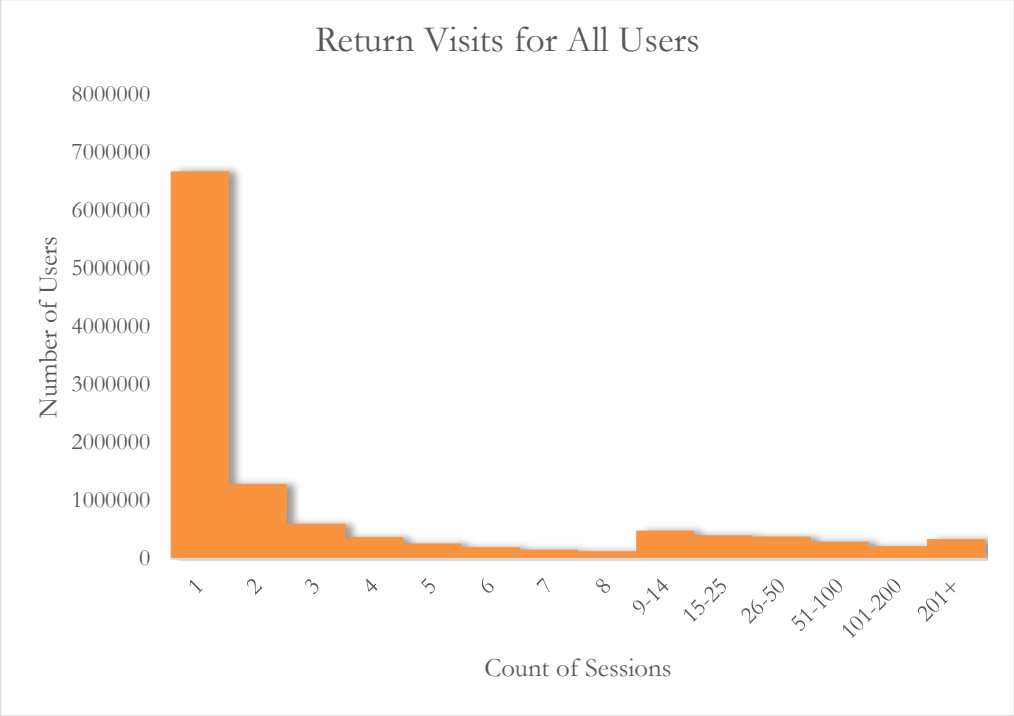
A final metric that is interesting to compare across the three groups above (all users, members, and genealogists) is return visits to the HathiTrust website. The charts below depict the different curves for each user group. For the "All Users" group as well as genealogists, the largest subgroups visit the website only once. However, for members, the largest subgroup is users who visit the website twice.

Return visits for these 3 user groups show different patterns of return

All three groups show an artificial bump in the middle of their series, where the numbers of visits start to be grouped together by Google Analytics (e.g., the number of visits goes

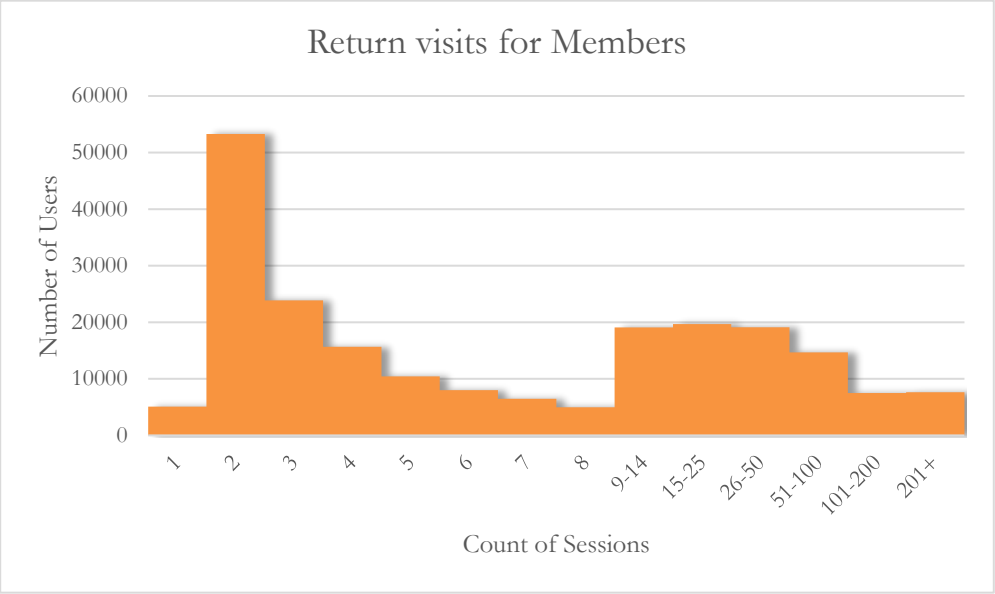
from 8 to 9-14 visits). What's interesting to note here is how drastic some of those patterns are.

The "All Users" group shows a less remarkable increase and a continued gradual decrease.



In the data for all users, making one visit to the HathiTrust website is most common

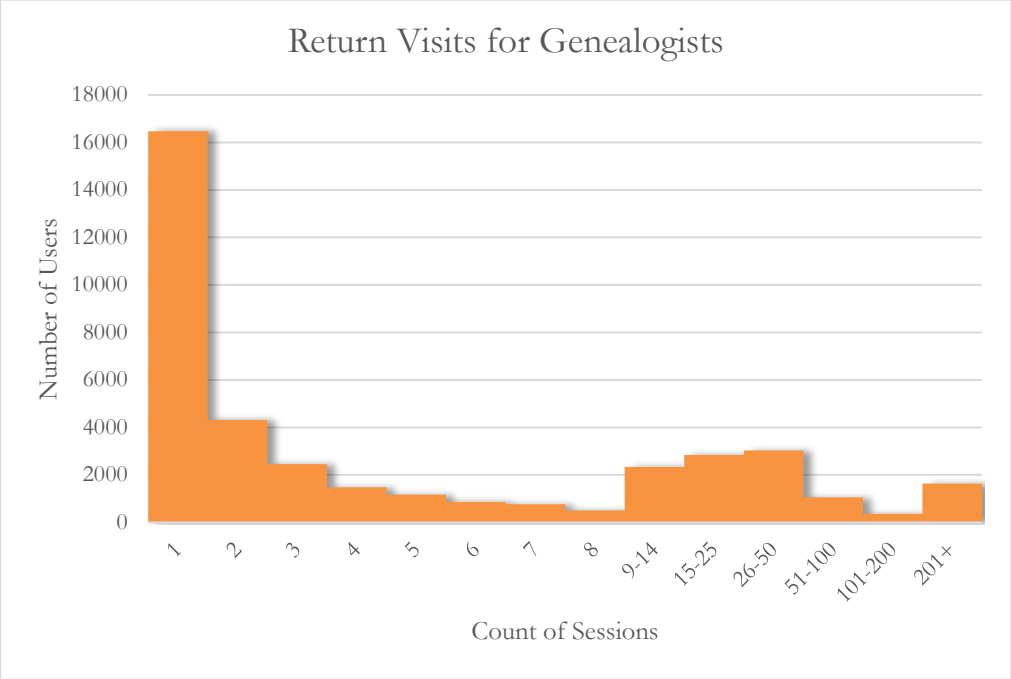
Members, however, have a large leap in the 9-14 visits category that continues on strongly through the 15-25, 26-50 and 51-100 visits categories, indicating that larger percentages of member users are repeat visitors as a whole.



The highest category for members is users that make at least two visits

Genealogists are also strong repeat visitors as compared to all users, although most genealogists do visit only once.

The genealogist curve shares similarities with all users as well as members



Conclusion

Segmenting users into groups reveals more useful data

In this version of the HathiTrust Collection Growth and Usage report, we have begun to investigate the behavior of different groups. Because the number of visitors that we get to our website is so high, if we looked at the averages for the entire world of users, we would find little information that is useful. By segmenting our users into known groups, we can begin to understand engagement levels of our users and a little bit of their activities.

It is becoming clear that users who arrive at our website without additional context for HathiTrust leave quickly without taking the time to discover what we offer. Users who arrive via referrals from other websites tend to stay longer and browse more p ages.

Referrals result in more engagement overall

The data above also revealed that there is a large gap between when users who are affiliated with member institutions know that they are eligible for member services. This is an opportunity that we need to work on in conjunction with librarians at our member institutions.

For users who understand what HathiTrust offers, however, we see a high degree of loyalty. Members and genealogists stay on the website, they view more pages per visit than the average user, and have repeat visits. **Once a user understands what HathiTrust offers them, they keep coming back.**

Stay in touch with us!

In order to receive alerts about development and improvement of our services, please subscribe to our newsletter at http://eepurl.com/cxjNWT. Contact us at feedback@issues.hathitrust.org with any questions.

Notes

[i] For more detailed information about all users, please see last year's report "14 Million Books & 6 Million Visitors". That report includes the following information about general users: location, browser language, how users arrive at the HathiTrust website, and top books accessed.

[ii] The 0% bounce rate is a technical anomaly as opposed to representing real usage patterns. A "bounce" occurs when a user views a single page and then leaves. However, when a member arrives at the HathiTrust web site, typically they browse around to find the items of interest, and then they log in to download a pdf or add an item to a collection that they need to log in. This means that when they have logged in, they have been on the website for some time.

[iii] There are some assumptions implicit in tracking usage for these two scenarios. We assume that users using a campus Internet provider are affiliated with that university in some way, as in most cases this is true. Likewise, we assume that users who are referred to HathiTrust from a university website, such as the library catalog, are affiliated with that university. These two scenarios are used to approximate the larger category of all users who are affiliated with a partner institution but may not be logging in. Even with this approach, we are missing some users. There is no way to identify the user who goes to HathiTrust through a search engine or other source while using a non-campus Internet provider and who never logs in during their session.

[iv] For the purposes of simplicity, this table is collapsed from "new visitors" and "returning visitors" into total counts.

[v] There are overlaps between these two characteristics: some users may be on their university's Internet *and* are referred to hathitrust.org from their university's websites. However, combining Internet access, referrals and logins for this short time period results in a very small group of users. We'll be able to do this comparison once we have more data for a larger time frame.

[vi] On October 9, 2017, we made a change in our tracking code that enabled us to get more granular data about logins.