ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

DISCOVERABILITY ON A DIGITAL LIBRARY: A STUDY OF "RABBIT HOLES" IN GALLICA

SEMESTER PROJECT REPORT

Anne-Laure Tettoni
Supervisor : Dr. Simon Dumas Primbault

7th June 2024

Contents

ABSTRACT

Gallica, the digital repository of the French national library, offers its users a variety of digitalized documents, from books and articles to maps and objects. How do researchers navigate across this content ? Can we characterize navigation regimes that exhibit wandering-like qualities ? Do longer sessions lead to less visible content ? Using server logs as traces left by users, we created a sub-corpus of long and unstructured sessions, and inspected what makes them differ from ordinary sessions, using statistics and various metrics. We defined these "Rabbit holes" using thresholds and found a bundle of corroborating evidence showing their subtle differences from normal sessions.

CHAPTER 1

INTRODUCTION

1.1 CONTEXT

Gallica is the digital library of the French national library (the national repository of all publications in France) and its partners. Since 1997, it offers a variety of digitalized documents for consultation, and has been expanding since then, gathering over 10 million documents. These can vary from books, articles, images, to maps and objects. As researchers rely increasingly on digital resources, computational humanities have in turn developed studies on their research patterns and navigation practices on digital libraries. Gallica is a great source for research, as it contains a good diversity and variety of quality documents.

Gallica also contains a multitude of documents that are of interest to not only researchers but also everyday internet users. Just like Wikipedia, it can be used to browse for interesting information or to satisfy curiosity, and not just retrieving a particular piece of knowledge or finding references on a specific subject. Previously, Kaabachi and Dumas Primbault 2023 have conducted a mixed-methods study on Gallica user logs, identifying various "regimes of navigation", ranging from corpus constitution to wandering, demonstrating that the same corpus can be apprehended in various ways depending on the desire of the user.

Using the same logs, we were interested in further characterizing the long sessions. Piccardi, Gerlach and West 2022 also researched the long tail of reading sessions on Wikipedia, identifying so-called "Rabbit Hole" sessions. These are defined as long, exploratory sessions where the user may find unexpected articles (the term comes from Lewis Carroll's Alice in Wonderland). In Kaabachi and Dumas Primbault 2023, these would be part of the wandering sub-regime of navigation. Characterized by a high number of documents and a high diversity of themes across them, these sessions value serendipity and discoverability. The user is not focused on an end goal but rather on exploring and discovering new documents.

The main goal of this project will be characterizing and examining these long and diversified sessions, using data enrichment methods and statistics. This report will detail the data gathering and enrichment, the methods used for sessionization, the computed statistics and obtained results.

1.2 PROBLEM STATEMENT

Understanding users' research behaviour on digital libraries can lead to improvements on the platform, from search mechanisms to recommendation algorithms. On a digital library, server logs can enable reconstitution of user sessions and recreation of paths on the platform, similarly to a path in a physical library. From these, we can classify the sessions using various metrics. Our interest will be on long and diversified sessions that we will call "Rabbit Hole" sessions, as their length puts them in the long tail of the length of all sessions (Piccardi, Gerlach and West 2022).

These sessions do not serve a specific purpose like corpus generation or retrieval of a specific source but span across various types of documents and themes. Their length also indicates that the user may be more interested in the path than on the end goal. Dumas Primbault 2023 has conducted a mixed methods research based on quantitative data and testimonies of Gallica users, showing the wide range of navigation regimes and confirming the existence of sessions based on exploration and entertainment. From the characterization of these sessions, we want to understand what leads to them and how they differ from other navigation regimes.

Wikipedia and Gallica differ in their nature, as Gallica contains digitalized material and Wikipedia is born digital. The former contains scientific and historical documents that are in the public domain, and the latter is written by contributors based on varying sources, destined for a large public. The structure of Wikipedia makes it possible to navigate from article to article using hyperlinks, while in Gallica, each user must build their own path using the search tool.

Another aspect of both Wikipedia, Gallica and physical libraries is the visibility of articles, documents or books. Not all of them are equally popular and equally consulted by users. Arora, West and Gerlach 2024 have conducted a study on so-called "Orphan articles", which are articles that cannot be accessed from hyperlinks of other articles. By adding hyperlinks from other articles, they found that their pageview observes a statistically significant increase. On physical libraries, the human process of selecting and highlighting books by workers can serve the purpose of giving them more visibility. On Gallica, the home page highlights some documents, but there aren't any hyperlinks linking documents together. In 2016 (year of our source data), the home page also proposes to consult the newly arrived documents and recent blog posts.

The searching and finding of documents is then either done by the search tool or by a series of steps, from selecting type, theme or geographical area, to navigating to the desired document. Documents can be characterized by findability and discoverability. The former suggests that they are well indexed and easily found through the search tool, and in opposition, the latter correspond to a large, diverse, undefined corpus, where users wander, discover documents, and navigate without a precise goal in mind.

We would like to examine whether or not long, exploratory sessions lead to less visible documents. Can we find evidence that spending a long time on the platform wandering leads to its darker, less seen corners ? To do that, we need to build sessions, select "Rabbit hole" ones and examine them.

CHAPTER 2

LITERATURE REVIEW

2.1 GALLICA

The starting point of this project was understanding the Gallica logs and previous work conducted on them. Nouvellet and Beaudouin 2017 have published an extended study on Gallica logs, from description and enrichment to sessionization and various methods to analyze navigation paths, such as Markov models and clustering. This work was very helpful to have a global view of the data and the possible study methods. Kaabachi and Dumas Primbault 2023's work focused on identifying navigation regimes from clustering of traces left by users navigating across different themes. It was useful, to begin deciding which navigation regime to focus on more precisely.

2.2 WIKIPEDIA

As mentioned before, several studies have been conducted on Wikipedia. Arora, West and Gerlach 2024 examine orphan articles and the dark matter of Wikipedia, which led us to question the visibility of Gallica documents. Piccardi, Gerlach and West 2022 characterize the long tail of the sessions and call them "Rabbit Hole" sessions, which we transposed to Gallica to find similar sessions on that corpus.

They define a Rabbit hole as a session whose length is in the long tail of the length distribution, and whose navigation tree has a minimum depth of ten steps. Another important factor is the semantic diffusion form the origin, measured by comparing paths against a random walk. This ensures that the session leads to random topics and is not an exploration of a specific theme by the user.

From these sessions, they examine what is the most common entry point, the most common time of the day, the device used, the topic of the first article. For example, they find evidence that shows these sessions are more likely to happen at night.

2.3 RABBIT HOLES

The term "Rabbit hole" is not specific to Piccardi, Gerlach and West 2022 and Wikipedia. The metaphor is used in other works to define a state where users find themselves consumed by an activity and unable to stop engaging in it, or when they are lead into new beliefs by inadvertence. For example, the YouTube algorithm has been criticised for leading to echo chambers and promoting extremist content, as studied by Brown et al. 2022. Social media in general has been condemned for promoting content that generates more reaction, even when this reaction is outrage and that content is extremist or hateful. In this sense, the Rabbit hole is a path to radicalisation (Halfaker et al. 2015). It can also be a path to conspiracy beliefs, as studied in this Sutton and Douglas 2022 article.

Whether on extremist content or ordinary one, Rabbit holes are defined by the attention they demand from the user for an extended period of time. Woolley and Sharif 2022 have found that when in a Rabbit hole, users are more likely to choose to continue to engage with the same type of media or topic that they were previously engaging in, as opposed to doing a non-media activity or changing topic.

CHAPTER 3

DATA

3.1 DESCRIPTION

For this project, we used Gallica logs over the month of February 2016. More precisely, they span over the period from the 31st of January 2016, at 13h00 to the 29th of February at 5h36. We chose to keep only one month to make computations easier to run. This still represents 319'344'032 log entries.

A log entry is a string from which we can extract the following meaningful data :

- A hashed IP address. This will be helpful for identifying requests coming from the same users.
- A country and a city from which the request comes from. We will not be using this field.
- The date of the request, in the following format : day/month/year:hour:minute:seconds time zone offset. We will use this, especially to find the time difference between two requests from the same user.
- HTTP request. These will be very useful as we will extract more information from them, such as the ARK.
- Protocol, which we won't use.
- Response number, also unused.
- Length of request, unused as well.
- Referrer : the website which the user comes from. We will use this to find which are the most frequent before starting a Rabbit hole.

Some of these fields may be empty, and then represented as either "null" or '-'. Here are two example of logs.

```
##e7fdec50f50253f6796d61b5382155f8##null##null##- - [31/Jan/2016:18:59:19
+0100] "GET /ark:/12148/bpt6k70211m HTTP/1.0" 200 24552 "-" "-" 48652
```

FIGURE 1
A log with missing information

```
##83bbb4ec83c93384666a2884238bbd55##United States##Menlo Park##- - [31/
Jan/2016:18:59:22 +0100] "GET /assets/static/javascripts/application/
layouts/achat_layout.js HTTP/1.1" 200 249 "-" "facebookexternalhit/1.1"
9439
```

FIGURE 2
A log with all information

3.2 ARK (ARCHIVAL RESOURCE KEY)

As mentioned, some requests contain ARKs that can be extracted from them, for example from figure 1, we could extract the ARK "bpt6k20211m". These Archival Resource Keys represent unique identifiers for each document, from which document metadata can be obtained by querying Gallica's website. These ARKs don't change over time.

Here is what an an ARK request looks like : (figure taken from Nouvellet and Beaudouin 2017)

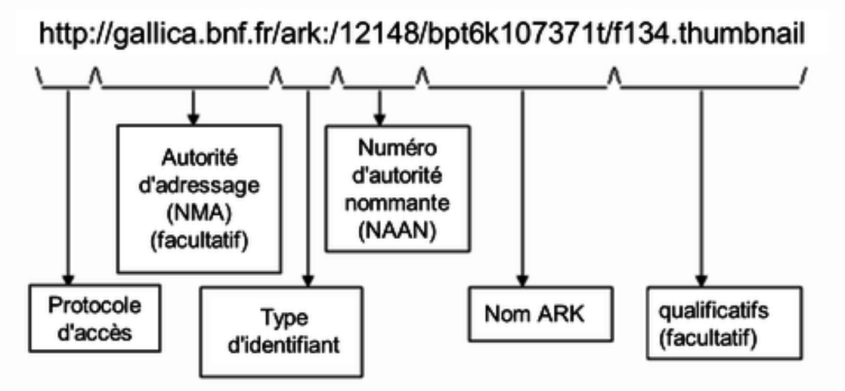


FIGURE 3
An ARK request to Gallica

The NAAN (name assigning authority number) will always be the same for Gallica, 12148. The NMA indicates the website that the resource can be accessed at. From a request like this, we can extract the ARK name, and then use it to request the metadata of the document.

3.3 ENRICHMENT

From our logs, we split them to extract the hashed IP address, the date, the request, the referrer and the ARK (obtained from parsing the request). From the request, we can also find the search

terms, if there were any, by checking if "search" is in the request then parsing the URL, finding the query parameters and using a regular expression to extract them.

From the ARKs, we want to obtain document metadata. What will interest us in our study of Rabbit holes is to qualify the diversity and semantic diffusion of a session. For that, we will want to obtain the type and theme of each document consulted. This is done by using Gallica's service for information retrieval. Here is an example of what the request "https://gallica.bnf.fr/services/OAIRecord?ark=btv1b6907077k" yields :


```
−<results ResultsGenerationSearchTime="0:00:00.119" countResults="1" resultType="CVOAIRecordSearchService" searchTime="">
  <visibility_rights>all</visibility_rights>
  −<notice>
    −<record>
      −<header>
        <identifier>oai:bnf.fr:gallica/ark:/12148/btv1b6907077k</identifier>
        <datestamp>2018-04-26</datestamp>
        <setSpec>gallica:typedoc:images:dessins</setSpec>
      </header>
      −<metadata>
        −<oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
          <dc:identifier>https://gallica.bnf.fr/ark:/12148/btv1b6907077k</dc:identifier>
          −<dc:title>
            [Pierre tombale sur laquelle est représenté un chevalier tenant une lance, dans un encadrement gothique] : [dessin]
          </dc:title>
          −<dc:description>
            Collectionneur : Gaignières, Roger de (1642-1715). Collectionneur
          </dc:description>
          <dc:description>Référence bibliographique : Gaignières, 3935</dc:description>
          <dc:format>Croquis à la sanguine</dc:format>
          −<dc:relation>
            Notice du catalogue : http://catalogue.bnf.fr/ark:/12148/cb40558009d
          </dc:relation>
          <dc:type xml:lang="fre">image fixe</dc:type>
          <dc:type xml:lang="eng">image</dc:type>
          <dc:type xml:lang="eng">still image</dc:type>
          <dc:type xml:lang="fre">dessin</dc:type>
          <dc:type xml:lang="eng">drawing</dc:type>
          −<dc:source>
            Bibliothèque nationale de France, BnF, Est. RESERVE Pe-4-Fol.
          </dc:source>
          <dc:rights xml:lang="fre">domaine public</dc:rights>
          <dc:rights xml:lang="eng">public domain</dc:rights>
          <dc:description>Appartient à l'ensemble documentaire : Des17Gaig</dc:description>
          <dc:format>image/jpeg</dc:format>
          <dc:format>Nombre total de vues : 1</dc:format>
        </oai_dc:dc>
      </metadata>
    </record>
  </notice>
  <provenance>bnf.fr</provenance>
  −<source>
    Bibliothèque nationale de France, BnF, Est. RESERVE Pe-4-Fol.
  </source>
  <typedoc>image</typedoc>
  <nqamoyen>0.0</nqamoyen>
  −<title>
    [Pierre tombale sur laquelle est représenté un chevalier tenant une lance, dans un encadrement gothique] : [dessin]
  </title>
  <date nbIssue="1"/>
  <first_indexation_date>05/12/2007</first_indexation_date>
  <streamable>false</streamable>
  −<listBibVirt>
    <label>gallica</label>
  </listBibVirt>
</results>
```


FIGURE 4
Result of an OAIRecord query

We can obtain the type of document under the **typedoc** field, here an image, and the theme, which is a Dewey class, under **sdewey**, not featured here as it is an optional field. This is a major limitation to finding the diversity of themes across a session, as only prints have a Dewey class.

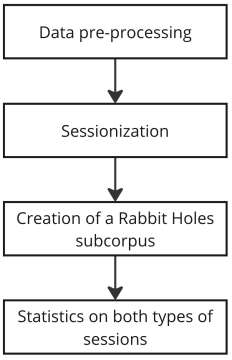
Lastly, we want to know how visible a document is, namely how many times it was consulted across all sessions of that month. To retrieve this information, we count unique occurrences of ARKs and IP address (a person can only contribute one view to a document), and store the ARK and its associated count for later use.

CHAPTER 4

METHODS

For this project, computations were run on the virtual machine of the LHST lab, and we implemented all the code using Python and jupyter notebooks. We mainly used Pandas for data processing and manipulation, and Beautifulsoup for HTML and XML parsing. The code is stored on GitHub : `https://github.com/ana571/rabbit-holes-gallica`.

Here is an outline of the steps undertaken.



4.1 PRE-PROCESSING

The pre-processing involves first the data enrichment steps described above. Then, we request separately the unique ARKs, using the list of ARKs associated with their counts. This process takes a long time and is subject to failure, for different reasons : either the ARK was not processed properly, or it corresponds to a document that is not available anymore on Gallica, yielding an error. Out of the 1'939'409 ARKs detected across all February logs , we obtained metadata for 1'584'653 of them, which represents 81%. This process was run in chunks. It is also useful to note that among those 1'939'409 ARKs, not all of them are actually ARKs but a small number are the result of parsing errors, and among the requested ones, not all of them have associated data.

As mentioned before, the size of the data is considerable, so to make the processing manageable, we ran it and the sessionization in chunks. This could lead to more sessions, as it separates a session that could run over multiple chunks. To have an estimate of how many sessions were

added, we ran the process on a chunk and then on the same chunk divided in two. We found that it added 797 sessions, out of the 74'917 ones found without chunking. This represents an increase of 1.06%, which is not significant.

As it was run in chunks, we also had many dataframes of ARKs associated with counts. We concatenated them and summed the counts for each ARK, then created a dictionary for quicker lookup. This allows the creation of a list of visibilities of documents consulted for each session.

4.2 SESSIONIZATION

From our enriched logs, we want to define user sessions. To do so, we start by grouping them by IP address and aggregating the features. Then, we want to find the difference of time between two requests, and if that time is higher than a set inactivity threshold, we will consider it to be a new session. We compute the time differences for each IP address, then create session IDs. The use of time heuristics to define user sessions has been documented in Zhang and Ghorbani 2004.

We choose 60 minutes as the inactivity threshold. This choice relies on two factors. First, Halfaker et al. 2015 have found that a 60 minute inactivity threshold is a good rule-of-thumb, and second, we tested with other thresholds (45, 75 and 90) on a chunk representing 24 hours and found that it didn't drastically change the number of sessions.

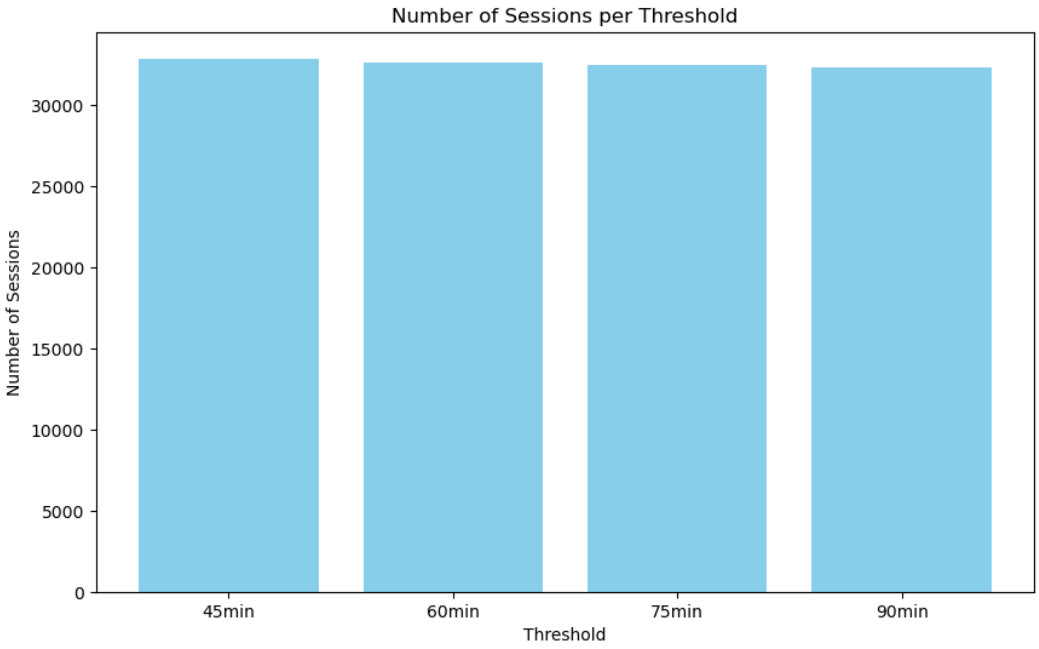


FIGURE 5
Number of sessions per threshold

From these session IDs, we find the sessions and enrich them by adding features, such as the length in minutes, a list of visibilities associated with each consulted ARK, and the first referrer.

We also removed the sessions with no ARK visited, which removed about 16.6% of the sessions. In total over the month, we have 1'181'190 sessions where at least one document was consulted.

From the list of visibilities, we create new features that will enable us to evaluate the evolution of the visibility of documents across a session. First, the mean and minimum visibility of a document, then the mean and minimum of the first and last three documents, and finally the variation of the mean and minimum visibility of the first and last documents. From this, we will be able to tell whether or not the session lead to more popular and more visited documents or the opposite.

4.3 RABBIT HOLES DEFINITION

A Rabbit hole is a session that is long and diversified. To characterize this diversity, we add a list of themes and types of visited documents to each session, using the metadata from the requested ARKs. To characterize the length, we create features that indicate if the session is in the top 10% and top 5% of length in minutes. To be in the top 10% longest sessions, it must be over 30 minutes long, and over 60 minutes long in the top 5%. We also add a feature that indicates the number of visited documents, and another if this number of documents is above 10.

Here is the distribution of sessions for each session length. The curve is red above the 10% threshold.

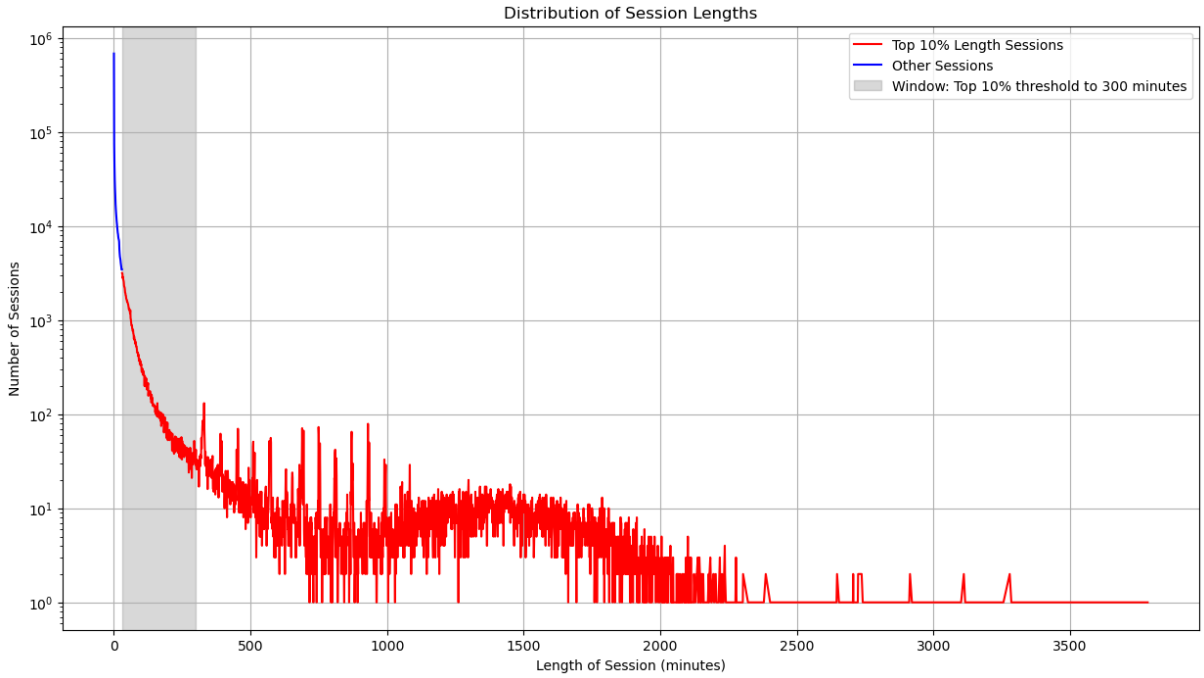


FIGURE 6
Distribution of sessions depending on length

Our Rabbit holes will correspond to the long tail of this distribution, in the highlighted gray area. We will remove sessions over 5 hours as they do not correspond to human behaviour.

We also count the number of different themes and different document types across the session. Finally, from this information we can create the diversity metrics. We consider a session to be diverse if the documents span across more than 2 types or more than 2 themes. We then filter the sessions to find the Rabbit hole ones. We start by taking only sessions in the top 10% threshold of length in minutes, then only those that are diversified, which represents 28.85% of the long sessions. From them, we select the ones with over 10 documents consulted. This leaves 1.59% of the overall sessions.

To check if this percentage is reasonable, we create three other diversity metrics and do the same filtering process with them. First, a restrictive diversity metric : we need to have 2 types or more and 2 themes or more. With this one, the Rabbit holes sessions amount to 1.16%. Second, an augmented metric, where we need to have 5 types or more or 5 themes or more. This represents 1.02% of sessions. Lastly, a augmented and restrictive metric, where we need 5 types or more and 5 themes or more. With this one, Rabbit hole sessions are only 0.11% of the sessions. We conclude that around 1% is a reasonable percentage.

Here is a summary of the various added features and their definition.

`first_referrer`	The website from which the session started
`length_minutes`	(last date - first date) in minutes
`visibility`	A list of the visibility associated with each ARK
`min_visibility`, `mean_visibility`	Smallest non-zero visibility, average of all visibilities
`min_first_3`, `mean_first_3`, `min_last_3`, `mean_last_3`	For the three first or last documents, the minimum non-zero visibility and the average of the 3 visibilities
`variation_min_vis`, `variation_mean_vis`	The difference between the last three document's min/mean visibility and the first three
`themes`, `types`	Lists of themes and types associated with each ARK
`nb_themes`, `nb_types`	Length of unique themes and unique types
`nb_docs`	Length of the list of accessed ARKs
`over_10_docs`	`nb_docs` ≥ 10
`top_10%_length`, `top_5%_length`	True if the `length_minutes` is $\geq$ 60 / 30 minutes
`diversified`	`nb_themes` ≥ 2 or `nb_types` ≥ 2
`div_restrictive`	`nb_themes` ≥ 2 and `nb_types` ≥ 2
`diversified_5`	`nb_themes` ≥ 5 or `nb_types` ≥ 5
`div_restrictive_5`	`nb_themes` ≥ 5 and `nb_types` ≥ 5

TABLE 1
Features added to the sessions

We now have filtered sessions that correspond to Rabbit holes. These are sessions in the top 10% of session lengths, so over 30 minutes, they are diversified, meaning over two types of documents or two different themes were consulted, and more than 10 documents were accessed. We also remove the sessions above 5 hours, as they don't represent human behaviour, and this

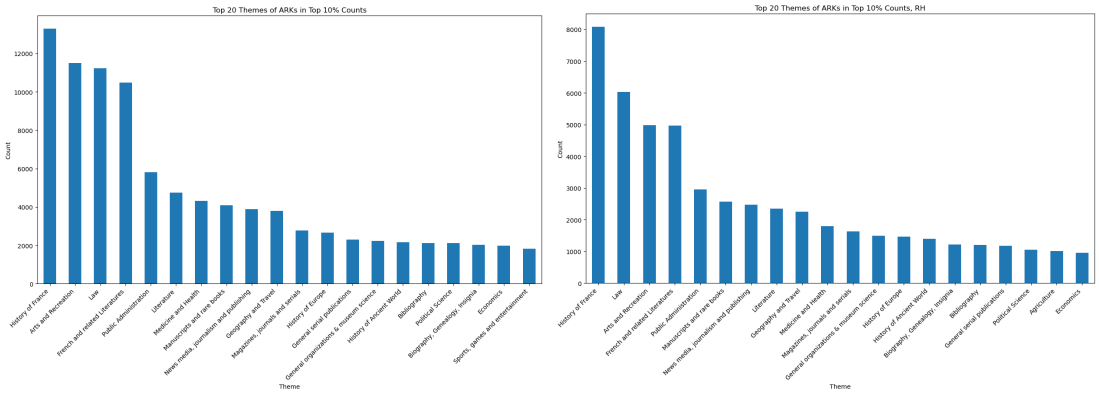
removes an additional 12% of the Rabbit holes. We will now compute a variety of statistics on them to find how they differ from normal sessions and see if they lead to less visible content.

CHAPTER 5

RESULTS

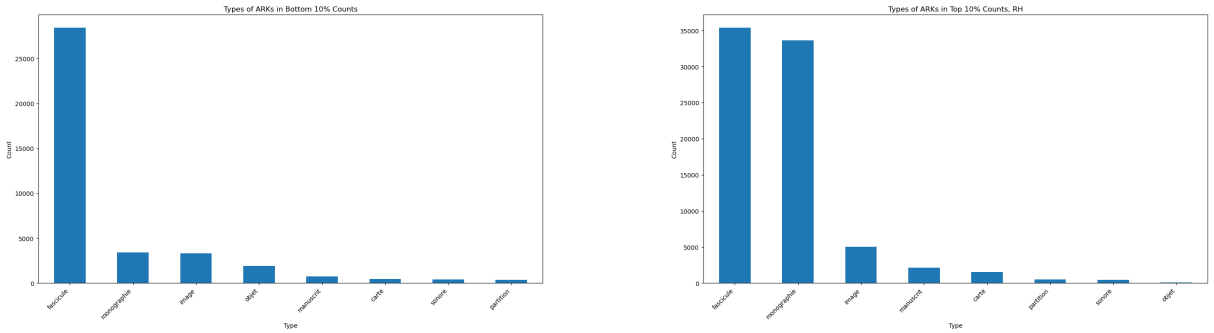
We start by running a few statistics on the logs and the arks. We find that there are about 16% of referrers missing and that only about 2% of requests are searches. We then look at the top 10% of most and least consulted documents over both types of sessions, ordinary ones and Rabbit hole ones, to find if there are differences.

Here are the top 10% most consulted themes and types on Rabbit holes and other sessions.



(A) Normal sessions (B) RH sessions

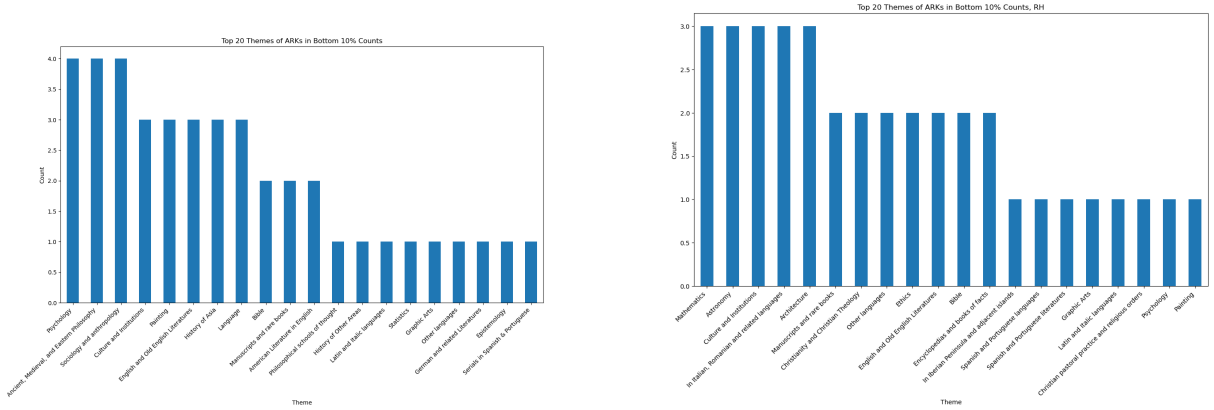
FIGURE 7
Top themes in most visited ARKs



(A) Normal sessions (B) RH sessions

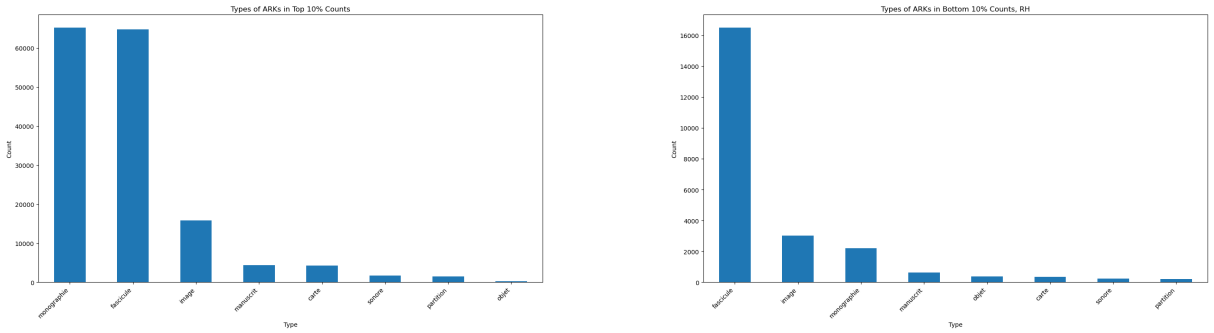
FIGURE 8
Top types in most visited ARKs

And the top 10% most consulted themes and types on the least visited documents, in both types of sessions.



(A) Normal sessions (B) RH sessions

FIGURE 9
Top themes in least visited ARKs



(A) Normal sessions (B) RH sessions

FIGURE 10
Top types in least visited ARKs

The type "fascicule" corresponds to an issue of a periodical publication, such as a magazine or a press article.

Next, we examined the most searched terms in Rabbit hole and normal sessions.

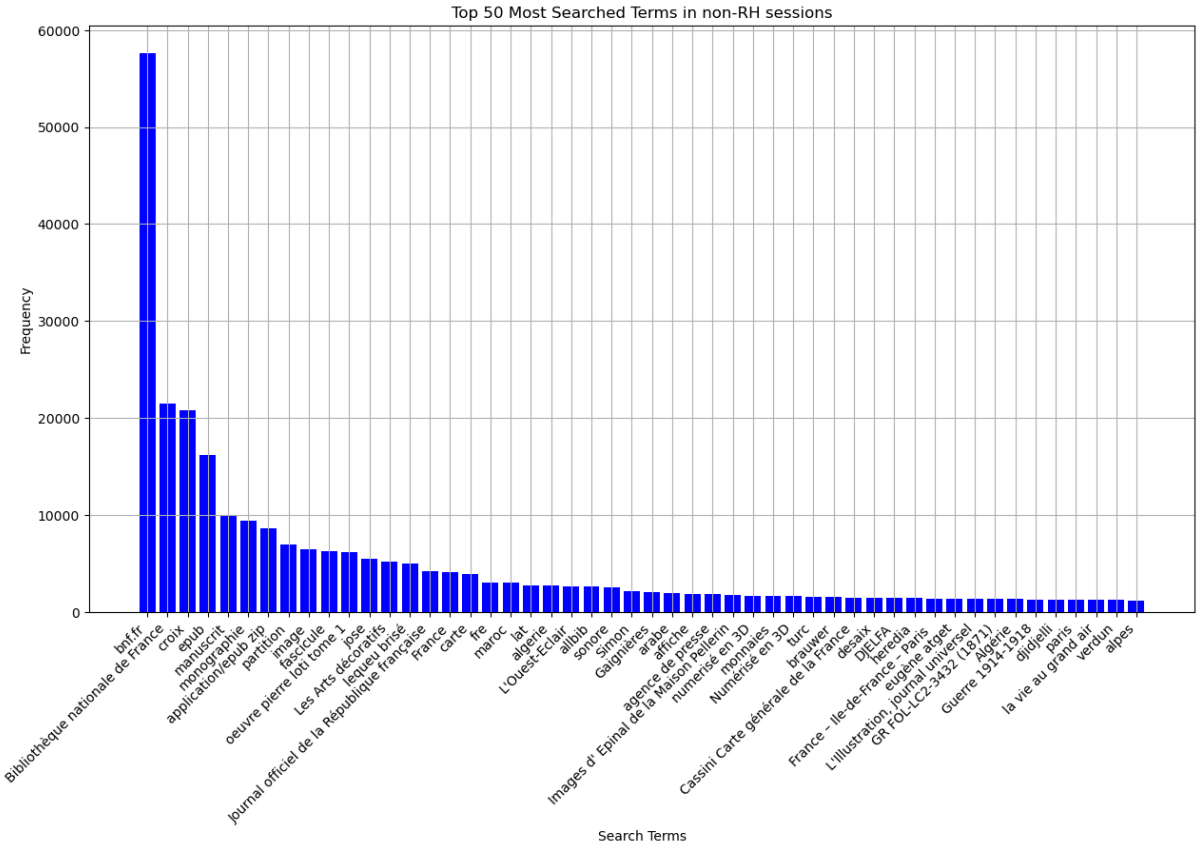


FIGURE 11
Top 50 search terms in normal sessions

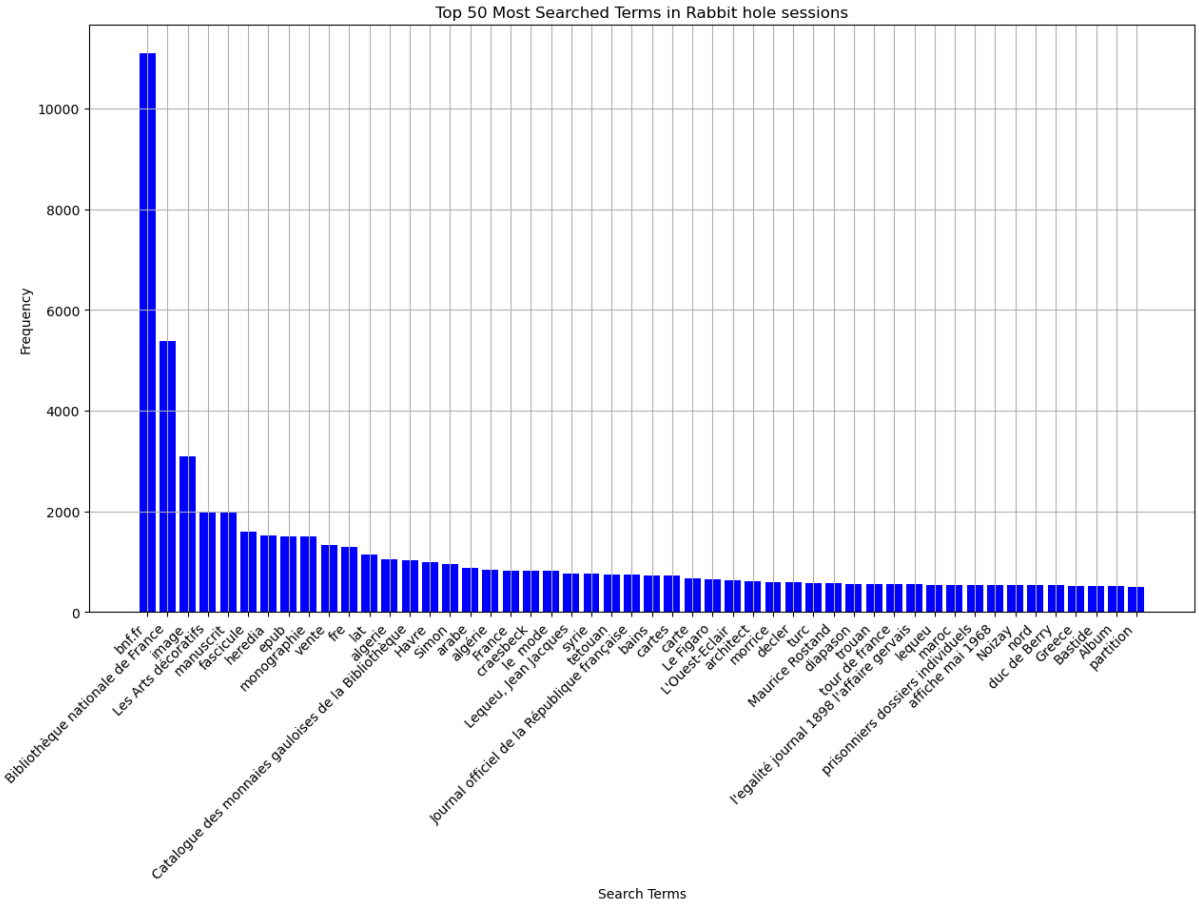


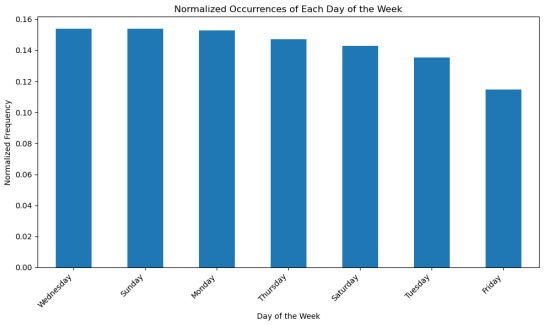
FIGURE 12
Top 50 search terms in RH sessions

We also compute the correlation between the length of a session and the minimum and mean visibility of the documents. Here are the results summarized in a table.

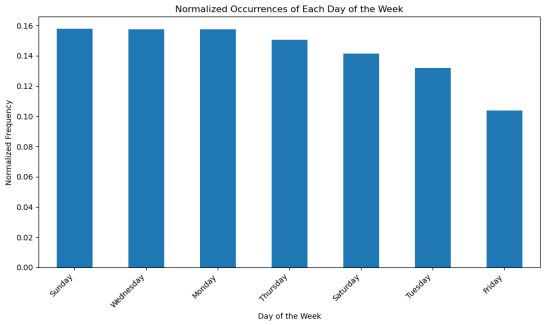
	length-minimum	length-mean
Normal sessions	-0.0005	0.0027
RH sessions	-0.0109	-0.0358

TABLE 2
Correlations of visibilities and length

Next, we want to examine when the Rabbit hole sessions happen, in terms of at which hour and which day of the week they begin.

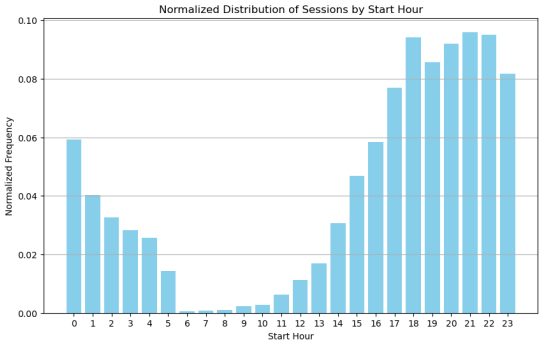


(A) Normal sessions

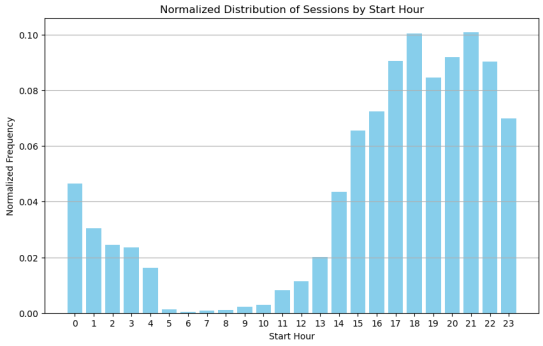


(B) RH sessions

FIGURE 13
Day of beginning of session



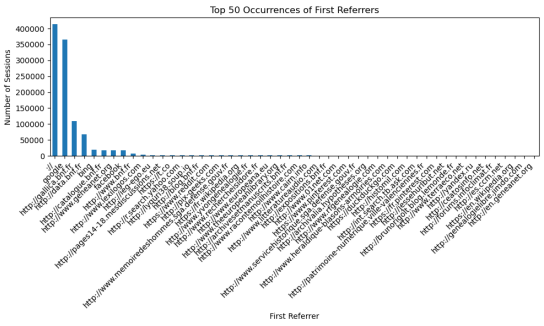
(A) Normal sessions



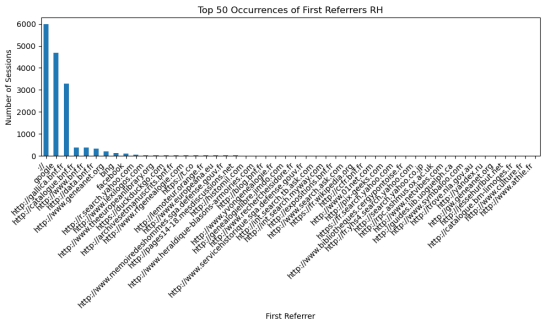
(B) RH sessions

FIGURE 14
Hour of beginning of session

We also want to plot the most common referrer to begin a session with.



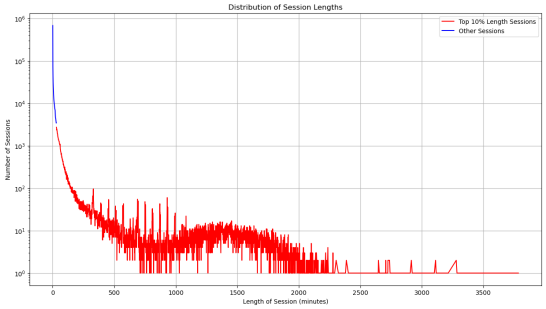
(A) Normal sessions



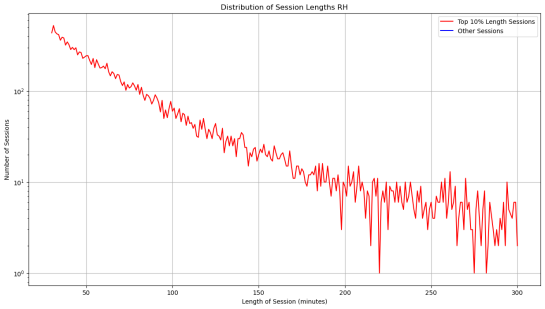
(B) RH sessions

FIGURE 15
Most common referrer

And to show the difference between normal and Rabbit hole sessions, we plot their differences : length, number of themes, number of types and number of documents.

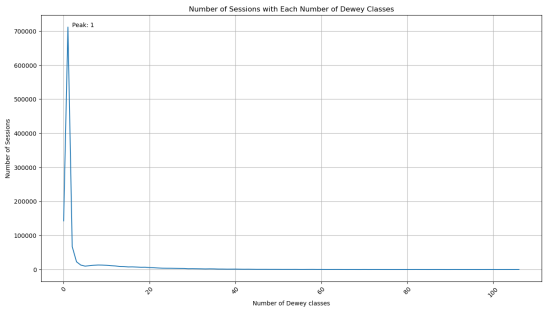


(A) Normal sessions

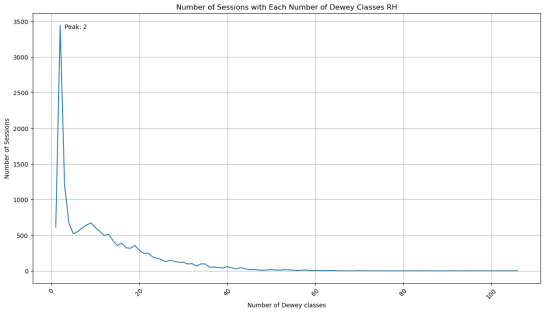


(B) RH sessions

FIGURE 16
Length of a session

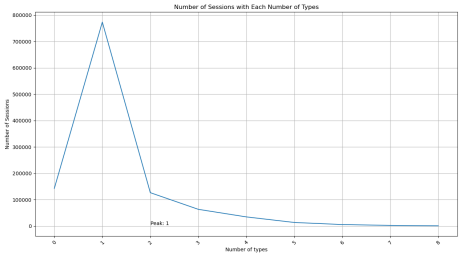


(A) Normal sessions

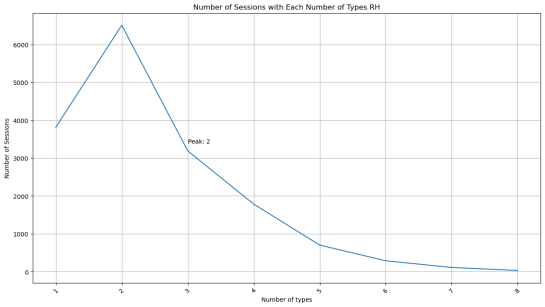


(B) RH sessions

FIGURE 17
Number of themes

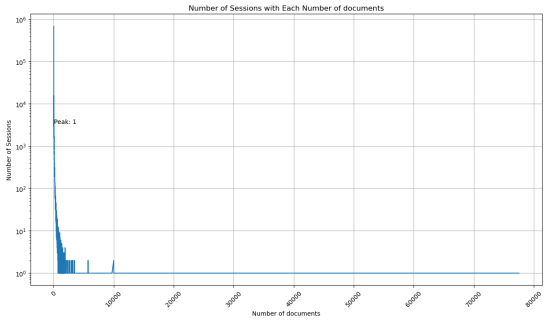


(A) Normal sessions

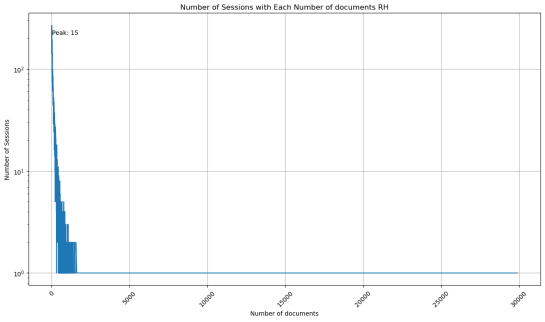


(B) RH sessions

FIGURE 18
Number of types



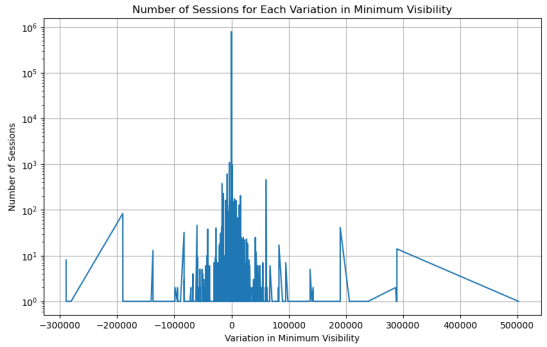
(A) Normal sessions



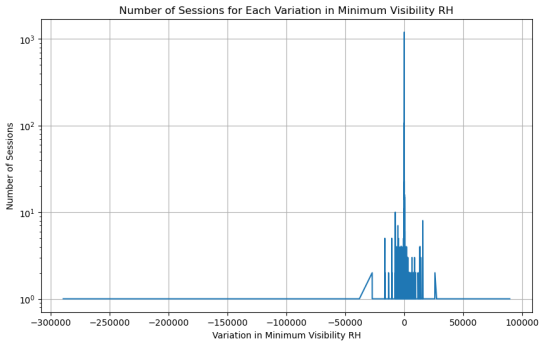
(B) RH sessions

FIGURE 19
Number of documents

Lastly we want to examine the variation in minimum and mean visibility on both types of sessions. Here are the plots of these variations across both types of sessions.

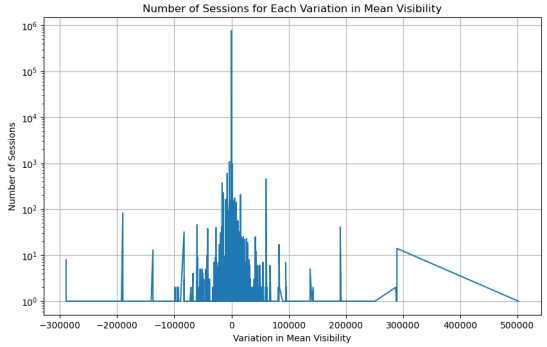


(A) Normal sessions

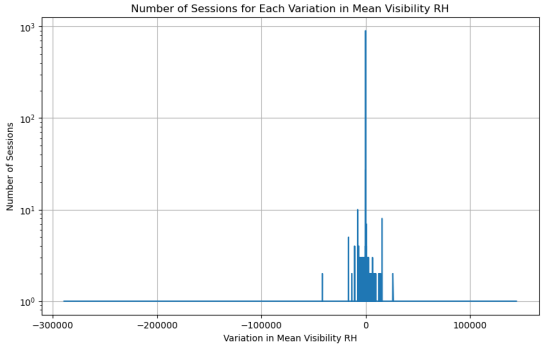


(B) RH sessions

FIGURE 12
Variation of minimum visibility



(A) Normal sessions



(B) RH sessions

FIGURE 21
Variation of mean visibility

CHAPTER 6

DISCUSSION

What can we interpret from these results ? First, the differences between ordinary sessions and Rabbit hole sessions are subtle. The top 5 themes most consulted are the same, although in different orders. These correspond the most popular themes on Gallica in general.

The most popular type is "fascicule", which corresponds to the main type of documents on Gallica at the time. Out of the approximately 3.6 million documents at the time, there are about 1.7 million of these, with the next main type being images, at almost 1 million.

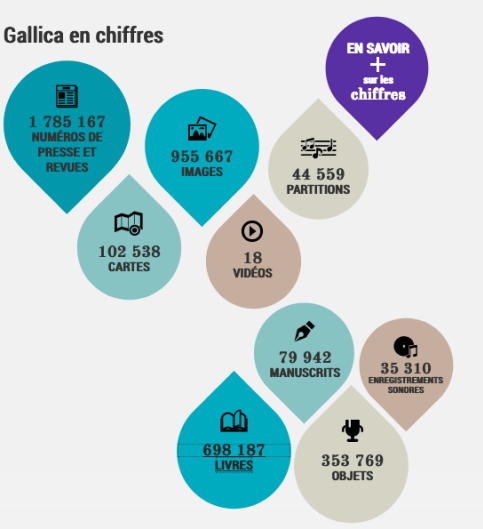


FIGURE 21
Distribution of types of documents on Gallica in 2016

We observe that in Rabbit holes, books (type "monographie") are much more popular than in normal sessions. This could be explained by the fact that longer sessions give more time to dive into longer documents.

For the search terms, in normal sessions, the most searched terms are ’bnf.fr’ and ’Bibliothèque de France’. This indicates users that are not familiar with the website. This is followed closely by the term "croix", and "epub", a popular format for ebooks. In RH sessions, the third most popular term is "image", indicating users that are not looking for something in particular.

The correlations are not significant, but we see that they are higher in absolute value for RH sessions. A longer session sees a slight decrease in the mean and minimum visibility of its documents. This suggests that when engaging in a RH session, the user is more likely to visit less popular documents.

For the beginning of the sessions, the distribution in terms of day of the week is similar across all sessions. This indicates that the users falling into Rabbit holes are the same users that use the platform for other purposes. For the hour, the distribution looks similar, but Rabbit holes have peaks at 6pm and 9pm, indicating that these sessions happen after working hours. They are also less likely to happen late at night or early in the morning (from midnight to 5 am) than other sessions.

The most common referrers for any sessions are Google and Gallica itself. Through previous user interviews, it was found that it is easier to find Gallica's content through Google than with their search bar. The prevalence of Google for normal sessions and the big gap separating it from Gallica could indicate users that are searching for something in particular, while for RH sessions, Gallica is much closer. Gallica's faulty search feature could then be used by users wandering and less concerned by accuracy.

Through plots of the length of sessions and the number of types, we show that the sub-corpus of Rabbit hole sessions is not representative of all sessions. The length is by construction in the top 10% of lengths, the number of different Dewey classes observes a peak at 2 and they are overall more diversified. Similarly, the peak of number of themes is at 2 instead of 1, and there are systematically more documents.

The plot of normal sessions observes peaks at regular intervals after 750 minutes. This length indicates non-human users, regularly querying the website, and not declared as a robot. An example of such a user is someone gathering the metadata of a list of ARKs, which is what we did in our data enrichment step.

Variation of the minimum visibility in normal sessions has a wider range than in RH sessions, and while it is aggregated around zero, it is mostly positive, which corresponds to sessions that lead to more popular documents. In RH sessions, it is slightly more likely to have a negative variation, which indicates a session that leads to less visible content. The same behaviour is observed on the mean variation.

6.1 LIMITATIONS

The chunking of the sessions for ease of computation adds about 1% of sessions, which may truncate sessions that would have been Rabbit holes otherwise. The requesting of ARKs was not complete, due to the time the process took, and that may lower the number of different types or themes for sessions that would have been qualified as Rabbit holes.

6.2 FUTURE RESEARCH

From this qualification of Rabbit holes, further methods could be applied to the sub-corpus to expand the study, such as modelling user paths as Markov chains or clustering.

User interviews could be conducted to further interpret the results. This sub-corpus of sessions exhibits an accumulation of weak signals, indicating a regime that is not out of the ordinary but another type of navigation undertaken by the same users. From this question of visibility, further work could be done to examine how to enhance the discoverability of documents on Gallica and how to promote curiosity among users.

CHAPTER 7

CONCLUSION

Using thresholds to define a sub-corpus of Rabbit holes, we examined them by building new features, creating visualizations on the data, and from that interpreted a new navigation regime. We found that this sub-corpus exhibits subtle differences from the main one, and found a bundle of clues showing these differences. These suggest a navigation regime that is slightly more likely to lead to less visible documents.

We outlined paths for further research, such as examining this sub-corpus with other techniques like clustering or Markov models. The evidence of robot behaviour after a certain session length could also lead to other research on this behaviour, by creating a sub-corpus of sessions where requests are fast and regular.

BIBLIOGRAPHY

Kaabachi, Bayrem and Simon Dumas Primbault (6th Dec. 2023). 'A Topological Data Analysis of Navigation Paths within Digital Libraries '. In: URL: https://ceur-ws.org/Vol-3558/paper935.pdf.

Piccardi, Tiziano, Martin Gerlach and Robert West (25th Apr. 2022). 'Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions'. In: *Companion Proceedings of the Web Conference 2022*. WWW '22: The ACM Web Conference 2022. Virtual Event, Lyon France: ACM, pp. 1324–1330. ISBN: 978-1-4503-9130-6. DOI: 10.1145/3487553.3524930. URL: https://dl.acm.org/doi/10.1145/3487553.3524930 (visited on 4th June 2024).

Dumas Primbault, Simon (18th Dec. 2023). 'Naviguer dans les savoirs à l'ère numérique. Pour une ethnographie des pratiques informationnelles sur Gallica'. In: *Études de communication* 61, pp. 61–89. ISSN: 1270-6841, 2101-0366. DOI: 10.4000/edc.16108. URL: http://journals.openedition.org/edc/16108 (visited on 4th June 2024).

Arora, Akhil, Robert West and Martin Gerlach (28th May 2024). 'Orphan Articles: The Dark Matter of Wikipedia'. In: *Proceedings of the International AAAI Conference on Web and Social Media* 18, pp. 100–112. ISSN: 2334-0770, 2162-3449. DOI: 10.1609/icwsm.v18i1.31300. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/31300 (visited on 4th June 2024).

Nouvellet, Adrien and Valérie Beaudouin (Nov. 2017). 'Analyse des traces d'usage de Gallica'. In.

Brown, Megan et al. (2022). 'Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users'. In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.4114905. URL: https://www.ssrn.com/abstract=4114905 (visited on 4th June 2024).

Halfaker, Aaron et al. (18th May 2015). 'User Session Identification Based on Strong Regularities in Inter-activity Time'. In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15: 24th International World Wide Web Conference. Florence Italy: International World Wide Web Conferences Steering Committee, pp. 410–418. ISBN: 978-1-4503-3469-3. DOI: 10.1145/2736277.2741117. URL: https://dl.acm.org/doi/10.1145/2736277.2741117 (visited on 4th June 2024).

Sutton, Robbie M. and Karen M. Douglas (Dec. 2022). 'Rabbit Hole Syndrome: Inadvertent, accelerating, and entrenched commitment to conspiracy beliefs'. In: *Current Opinion in Psychology* 48, p. 101462. ISSN: 2352250X. DOI: 10.1016/j.copsyc.2022.101462. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352250X2200183X (visited on 4th June 2024).

Woolley, Kaitlin and Marissa A. Sharif (June 2022). 'Down a Rabbit Hole: How Prior Media Consumption Shapes Subsequent Media Consumption'. In: *Journal of Marketing Research* 59.3, pp. 453–471. ISSN: 0022-2437, 1547-7193. DOI: 10.1177/00222437211055403. URL: http://journals.sagepub.com/doi/10.1177/00222437211055403 (visited on 4th June 2024).

Zhang, J. and A.A. Ghorbani (May 2004). 'The reconstruction of user sessions from a server log using improved time-oriented heuristics'. In: *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.* Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004. Pp. 315–322. DOI: 10.1109/DNSR.2004.1344744. URL: https://ieeexplore.ieee.org/document/1344744 (visited on 6th June 2024).