Discoverability in a Digital Library: A Study of "Rabbit Holes" within Gallica's Corpus

Anne-Laure Tettoni, Simon Dumas Primbault

Discoverability in a Digital Library: A Study of "Rabbit Holes" within Gallica's Corpus

Anne-Laure Tettoni[1,*,†], Simon Dumas Primbault[1,2,3,†]

[1]Laboratory for the history of science and technology (LHST), Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland

[2]OpenEdition (UAR 2504, CNRS/EHESS/AMU/AU), 22 rue John Maynard Keynes, 13013 Marseille, France

[3]Bibliothèque nationale de France (BnF), Quai François Mauriac, 75706 Paris, France

Abstract

The phenomenon of aimless web navigation, often compared to falling "down the rabbit hole," brings to light significant aspects of the Internet's "long tail" concept. This research examines whether longer, non-goal-oriented web sessions genuinely lead users into the long tail of digital libraries, thereby exploring the discoverability of cultural heritage. The focus of this study is on Gallica, the French national library's online platform. This work aims to identify and characterize such sessions within Gallica, defining rabbit holes as long and diversified navigation sessions. The difficulty lies in identifying rabbit holes within server logs, which requires a mixed-methods approach involving interviews, qualitative studies, and simple statistical analyses. Despite Gallica's lack of hypertextual structure, we show that users do engage in rabbit hole-like behavior, navigating through keyword searches and filters. The study's findings align with user testimonies. A crucial conclusion is that rabbit holes in Gallica do not generally lead users to less-consulted content. This limitation is attributed to the search engine, which users must somewhat "hack" to navigate effectively. Enhancing Gallica's discoverability tools without compromising the existing user experience is essential for improving content accessibility.

Keywords

digital library, navigation practices, discoverability, rabbit holes, long tail

1. Introduction

1.1. Research Question

Who has never found themselves surfing almost aimlessly throughout the vastness of content on the Internet? Indeed, the hypertextual structure of most content on the World Wide Web allows users to *navigate* from page to page—either by curiosity, distraction, or mere boredom—, to the extent that, not unlike *Alice in Wonderland*, they may fall "down the rabbit hole", supposedly in the most unknown of places. This trope alludes to the two meanings of the Internet's "long tail". Initially deriving from the network's structure [12], the long tail denotes the mass of Internet pages with very few links pointing towards them—as opposed to the head of the network, composed of very few websites with a great numbers of inward links attracting the

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

*Corresponding author.

†These authors contributed equally.

ana.tettoni@gmail.com (A. Tettoni); simon.dumas-primbault@openedition.org (S. Dumas Primbault)

0000-0002-0877-7063 (A. Tettoni); 0000-0001-7116-9338 (S. Dumas Primbault)

major part of the traffic. This structural observation was later reconceptualized as a business opportunity specific to the Internet, claiming that markets should now focus on the vast diversity contained in the long tail [1].

Though, this view relies on the assumption that on top of this hypertextual structure, there is actual—and substantial—traffic in the long tail. This calls onto another meaning of the long tail in terms of user practices: while most Internet users make brief sessions to *retrieve* known content by querying search engines with relevant keywords, a small but significant portion of sessions are longer and seemingly not goal-oriented, endeavoured by users *navigating* contextually from place to place. But do these longer sessions really lead users in the long tail of the content? Do longer aimless sessions actually take readers into the "dark matter" of the Internet?

By delaminating the two meanings of the long tail on the Internet, this work-in-progress aims to address a broader issue: that of the actual accessibility of content, beyond mere access; and the consequent need for discoverability tools, beyond mere findability. Indeed, research has shown that content in open access may not actually be accessible due to indexation issues, but also a lack of infrastructure in specific areas, the need for different skills or abilities to simply grasp the content, or the absence of incentives to wander beyond the supposed relevance of search engines [5].

With the exponential growth of their corpora—Gallica now hosts ten million digitized documents—, digital libraries have become genuine informational *milieus* that users practically navigate, step-by-step and iteratively. In this context, the traditional retrieval and recommendation systems—search engines, aggregators, socials...—, if not reassessed, tend to reinforce cultural asymmetries—favouring certain authors, languages, media types, while invisibilizing the rest—thereby offering a certain perspective on history and the social world in the case of cultural heritage. Discoverability could be defined at the intersection between users and content, as the propensity to stumble upon some unexpected content that the user was not searching for and nonetheless deems relevant for their purpose [15, 24].

In that sense, discoverability does not only reflect the desire for readers to make new and unforeseen connections—serendipity—but it also fosters cognitive justice and values such as inclusion and diversity. Consequently, it would not be enough to ensure findability in information systems—*i.e.* to provide relevant results to detailed queries within a closed and known informational space. In addition to this, it would indeed be necessary to design and develop tools to foster discoverability—*i.e.* to promote lesser known or consulted content that users were not looking for but might nonetheless find useful.

1.2. Related Work

Previous studies were led mainly on social media. Indeed, due to their informational architecture built around algorithmic recommendations to scroll endlessly, most social media are suited to rabbit holes, even "doomscrolling", and the discovery of content. In general, they have been criticized for promoting content that generates more reaction, specifically outrage regarding hateful content. For example, the YouTube algorithm has been criticised for leading to echo chambers and promoting extremist content [3]. In this sense, rabbit holes can be understood as paths to radicalisation [11]. They can also foster the circulation of fake news and be a path

to conspiracy beliefs [22].

Extensive work was also led on Wikipedia. In an effort to understand how users either search in or navigate through Wikipedia, and how these two strategies relate [7, 6, 18, 21, 23], scholars have shown the reliance on the structure of the articles [14], the link structure of the corpus [8], or the citations [20]. More specifically, Tiziano Piccardi, Martin Gerlach, and Robert West [19], have shown by the computational analysis of Wikipedia server logs that they could extract a consistent and coherent subset of user sessions corresponding to a specific user navigation regime they call "rabbit hole". Rabbit holes are longer sessions, more likely to be undertaken on desktop computers during workday or on a mobile device at night, and they exhibit random navigation patterns. Let us emphasize that although such rabbit holes span a vast array of pages, they mostly remain in the same semantic or topical area as the first pages visited.

1.3. Case Study

While it has become common parlance to say that a library reader may serendipitously find something they were not looking for by roaming through bookshelves, there is surprisingly little work about rabbit holes in digital libraries. Studies have long shown that besides directed search, *navigation* is a common, if underrepresented, informational practice—e.g. "berrypicking" [2], browsing [4], "bouncing" [16]. The present work-in-progress intends to further research on navigation practices within Gallica by addressing rabbit holes as a kind of navigation.

A cultural heritage repository, Gallica is the online platform of the French national library. As a consequence of a policy of mass digitization initiated in 2004, the digital library now hosts 10 million documents in the public domain, of 9 different types ranging from prints to manuscripts to maps and periodicals. After the seminal work of Nouvellet and Beaudouin [17], and using mixed methods at the intersection of ethnography (semi-directed interviews with users), semiotics (analysis of the interface and architecture), and digital humanities (analysis of server logs), we have identified several "navigation regimes" across the digital corpus [10, 13].

For the present study, it is important to note that the computational analysis of server logs was led after two campaigns of respectively 7 and 17 semi-structured interviews with users of Gallica, as shown in figure 1 (see also [10, 13]). Such interviews helped us point out that most users exhibited a specific navigation regime—unstructured exploratory browsing—whenever they could find the time to wander aimlessly throughout Gallica's corpus, that is either at the end of their long day, or during less intensive phases of their work. Nonetheless, these moments were considered highly important in their practice, leading them to make new and unforeseen connections between heterogeneous elements of their research; and semi-structured interviews helped us identify the properties of such regime (length, diversification) by modelling user sessions. Therefore, although rabbit holes are of low statistical significance, they are of paramount epistemological significance in practice. This decorrelation between the statistical significance of unstructured exploratory browsing and its epistemological importance in knowledge making is also due to the fact that directed search is technically favoured by the prevalence of the search engine and the interface that somewhat reinforces this common practice.

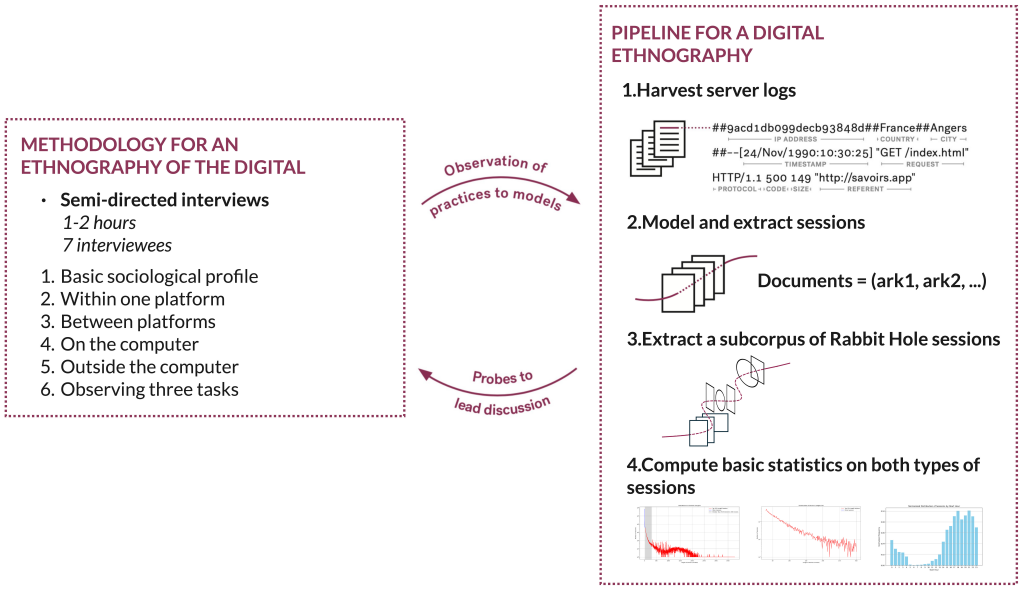


Figure 1: A Mixed-Methods Ethnography: Pipeline and Methodology

Note that contrary to Wikipedia, Gallica's corpus is not hypertextually structured: the documents are not linked in the database, and they are not clickable on the interface, neither are the metadata. Nonetheless, users have told us that they would engage in rabbit-hole like behaviours: e.g. one told us they would look for historical "lolcats" after a long day's work while another said Gallica was their own Candy Crush where they would waste time playing around. Indeed, another study showed us how, in order to navigate from document to document, users iterated keyword searches and filtering to "manage the noise" in the search results, thereby constructing step by step their own path throughout the corpus, sometimes even a rabbit hole [9]. The present work-in-progress aims at circumscribing a robust subcorpus of such sessions and characterizing it. We will define rabbit holes as a subregime of navigation, that is as a subset of user sessions that are long and diversified (Sec 3), and we will try to distinguish them using simple statistics in comparison with average sessions (Sec 4).

2. Data and Sessionization

We used Gallica server logs spanning the period from 31st January 2016, at 13h00 to 29th February, at 05h36. We chose to keep only one month to make computations easier to run. This still represents 319,344,032 log entries. The month of January is not particular in terms of academic deadlines or cycles, so it should not see an increase in students which would skew the data. Time-zones were not taken into account, as the vast majority of users are from French-speaking countries in Europe, located in the same timezone. Moreover, users could be using VPNs, so location of users is uncertain anyway. This research involves no sensitive data and

all personal data were definitively anonymized: IP addresses were hashed, and the hash table destroyed. It is compliant with GDPR and the Swiss Federal Act on Data Protection. It was approved by EPFL's Human Research Ethics Committee. For more details on the data, refer to appendix A or [13].

2.1. Sessionization

The pre-processing of this data involves enriching it by requesting individual Archival Resources Keys (ARKs) in order to provide additional information about the documents consulted. As mentioned before, the size of the data is considerable, so to make this step manageable, we ran the pre-processing and sessionization tasks in chunks. This leads to more sessions, as it severs sessions that may 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 merely 1.06%.

We also created a dictionary of document's "*visibilities*": every time a document is consulted by a user, we increment the visibility of the corresponding ARK.

From our enriched logs, we want to define user sessions. To do so, we start by grouping them by hashed-IP addressed and aggregating the features. Then, we want to find the difference of time between two requests, and if that time is higher than a given inactivity threshold, we consider it to be a new session. We compute the time differences for each hashed-IP address, then create session IDs. The use of time heuristics to define user sessions has been documented in [25].

We choose 60 minutes as the inactivity threshold. This choice relies on two factors. First, [11] 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, as shown in table 1.

Table 1
Number of sessions generated per inactivity threshold

threshold	45min	60min	75min	90min
number of sessions	32'858	32'610	32'478	32'329

From these session IDs, we separate the sessions from each other, 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.

3. Defining a Rabbit Hole

We broadly define a ”rabbit hole” as a session that is long and diversified. To characterize this diversity, we add a list of themes—Dewey classes—and types of documents—either prints, manuscripts, images, maps...—consulted during 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.

The distribution of all sessions, irrespective of their length or diversity, is shown in figure 2. The curve is red above the 10% threshold.

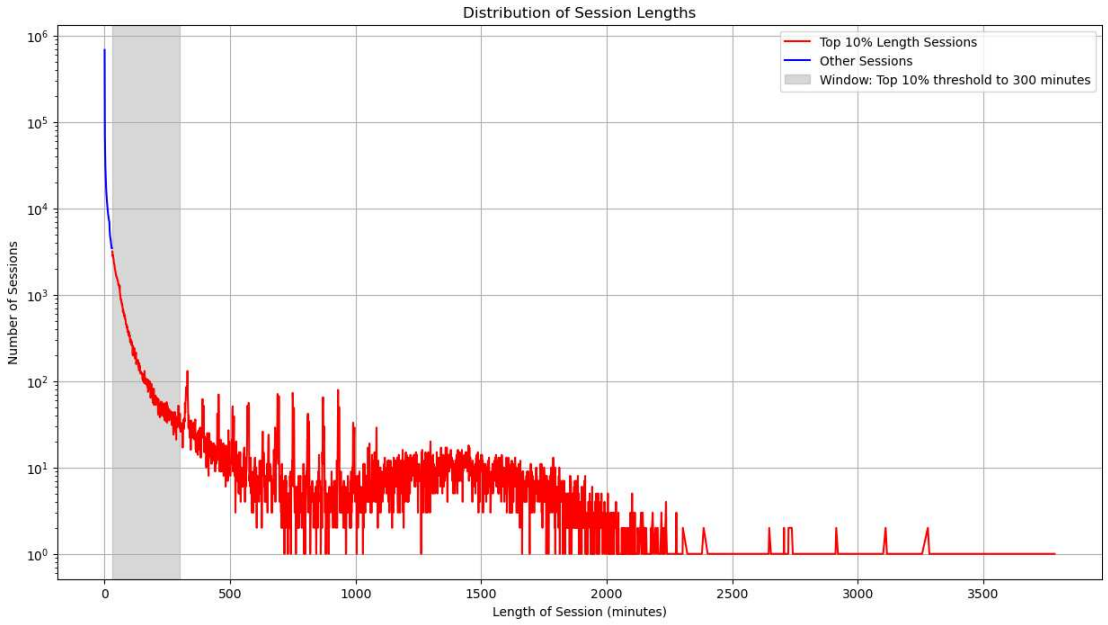


Figure 2: Distribution of sessions depending on length

Note that this curve exhibits an interesting behaviour above 300 minutes *i.e.* 5 hours: regular spikes emerge from the noise roughly every hour, possibly denoting a surplus of timed sessions due to bots programmed to stop after a certain amount of hours. Furthermore, the longer tail exhibits a bump above 780 minutes *i.e.* 13 hours: these very long sessions cannot be considered human either. 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. Therefore, in order to set aside these robotic behaviours, our selected rabbit holes correspond to the long tail of this distribution, in the highlighted gray area.

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 consulted span more than 2 types or more than 2 themes. Then, we filter the sessions to extract the rabbit holes. We start by taking only sessions in the top 10% threshold of length in minutes (long sessions), 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 apply 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 hole 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.

Table 2 contains a summary of the various added features and their definition.

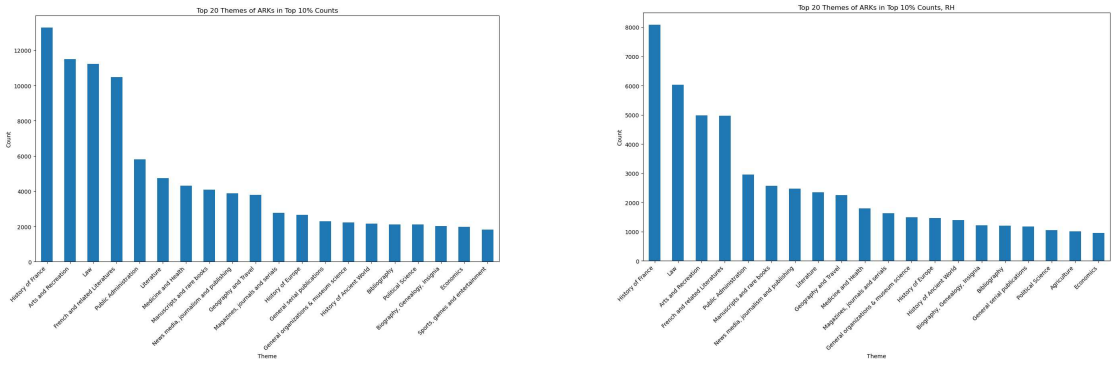
Table 2
Features added to the sessions

Feature	Description
`first_referrer`	The website from which the session started
`length_minutes`	(last_timestamp - first_timestamp) in minutes
`visibility`	A list of the visibility associated with each ARK
`min_visibility`	Smallest non-zero visibility
`mean_visibility`	Average of all visibilities
`min_first_3`	Minimum visibility of the first 3 documents
`mean_first_3`	Average visibility of the first 3 documents
`min_last_3`	Minimum visibility of the last 3 documents
`mean_last_3`	Average visibility of the last 3 documents
`variation_min_vis`, `variation_mean_vis`	Difference between the last three documents' min/mean visibility and the first three
`themes`, `types`	Lists of themes and types associated with each ARK
`nb_themes`, `nb_types`	Number of unique themes and unique types
`nb_docs`	Number of accessed ARKs
`over_10_docs`	`nb_docs` ≥ 10
`top_10%_length`	True if `length_minutes` ≥ 60
`top_5%_length`	True if `length_minutes` ≥ 30
`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

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 requested. We also remove the sessions above 5 hours, as they don't represent human behaviour, and this removes an additional 12% of longer sessions. We will now compute a variety of statistics on them to find how they differ from average sessions and see if they lead to less visible content.

4. Provisional Results

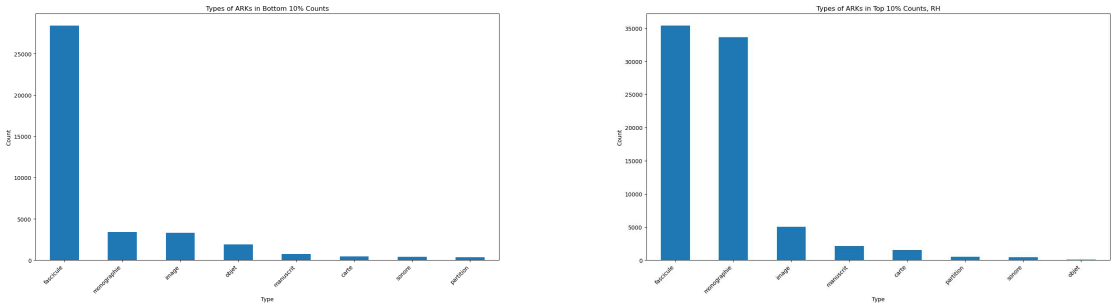
We begin by examining the least and most consulted themes and types of documents across both types of sessions, as well as the top 10% most consulted themes and types on the least visited documents, in both types of sessions.



(a) Average sessions

(b) RH sessions

Figure 3: Top themes in most visited ARKs



(a) Average sessions

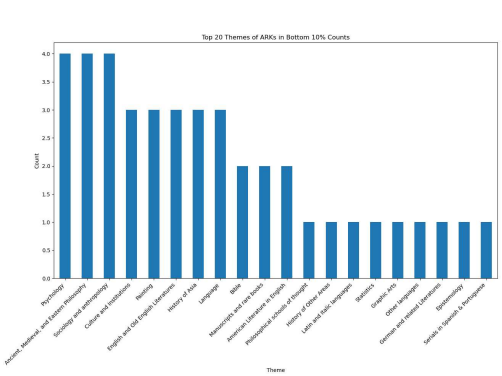
(b) RH sessions

Figure 4: Top types in most visited ARKs

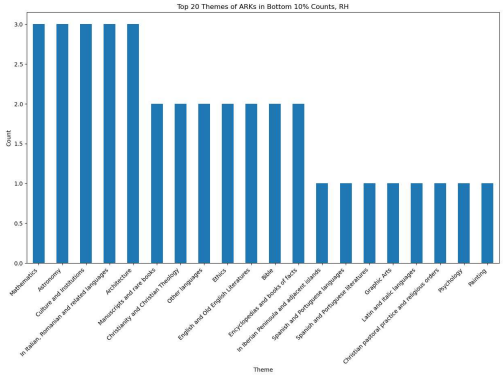
The top 5 themes most consulted are the same, although in different orders. These correspond the most popular themes on Gallica in general and to the interests of its audience mainly made up of historians and genealogists (figure 3).

Figure 4 shows the most popular type is "*fascicule*", which corresponds to an issue of a periodical publication, such as a journal or a newspaper, was also 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. The same examination is done for least popular types and themes (figures 5 and 6).

We observe that in rabbit holes, books (type "*monographie*") are much more popular than in average sessions. Indeed, longer sessions give more time to dive into longer documents.

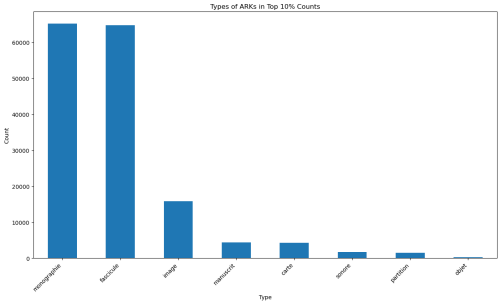


(a) Average sessions

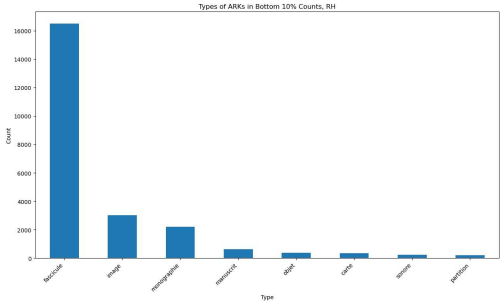


(b) RH sessions

Figure 5: Top themes in least visited ARKs



(a) Average sessions



(b) RH sessions

Figure 6: Top types in least visited ARKs

Next, we examined the most searched terms in rabbit holes and average sessions (figure 7 and 8).

In average 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”, maybe indicating users that are not looking for something in particular.

We also compute the correlation between the length of a session and the minimum and mean visibility of the documents. The results are summarized in table 3. The correlations are not

Table 3
Correlations of visibilities and length

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

statistically significant, but we see that they are higher in absolute value for RH sessions. A

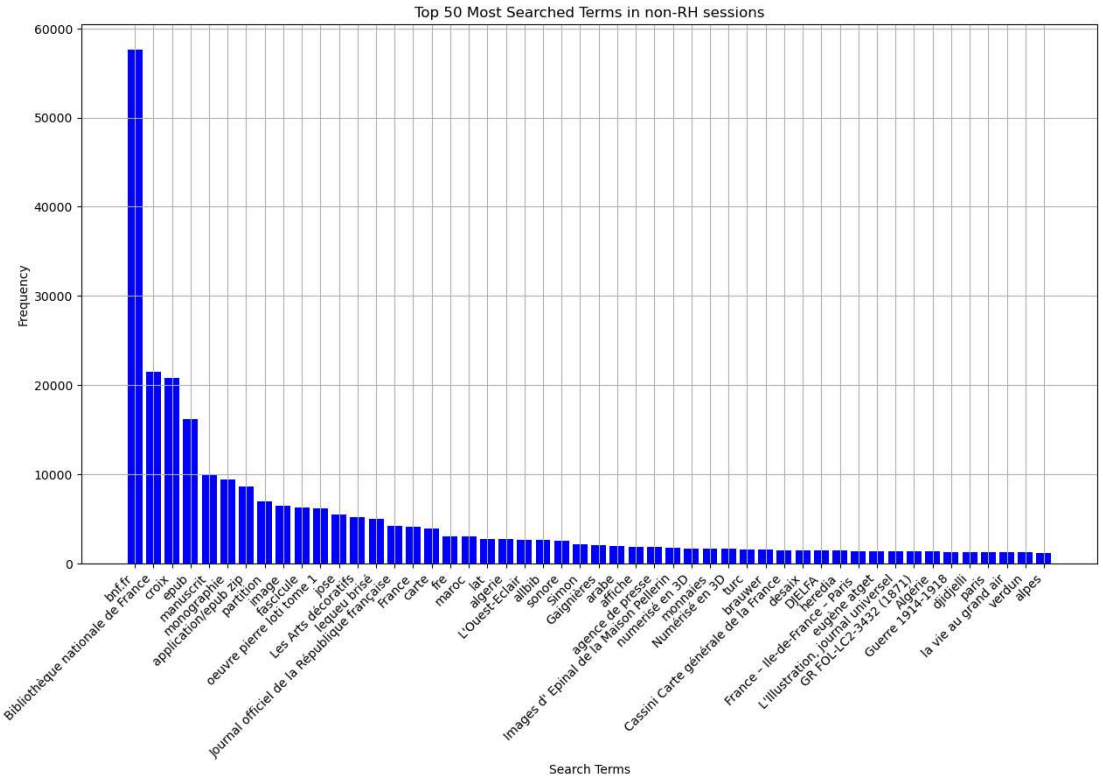


Figure 7: Top 50 search terms in average sessions

longer session sees a very slight decrease in the mean and minimum visibility of its documents. This suggests that when engaging in a RH session, the user is very marginally more likely to visit less popular documents. Though, we cannot conclude that, in general, rabbit hole sessions lead users to the ”dark matter” of Gallica.

Next, we want to examine when rabbit hole sessions are most likely to happen in the day or the week.

For the beginning of the sessions, the distribution in terms of day of the week is similar across all sessions (figure 10). For the hour, the distribution looks similar, but figure 9 shows rabbit holes have peaks at 6pm and 9pm, indicating sessions that happen after working hours but before night. Indeed, they are also less likely to happen late at night or early in the morning (from midnight to 5 am) than other sessions. This suggests that users falling into rabbit holes are not exceptional—e.g. insomniacs or early risers. Rather, they might be just the average kind of users that otherwise populate the library for other purposes.

We also plot the most common referrer to begin a session with (figure 11).

Google and Gallica itself rank first for any type of session. Through previous user interviews, it was found that it is easier to retrieve content from Gallica through Google than with their search engine. The prevalence of Google for average sessions and the big gap separating it from Gallica could indicate users that are searching for something in particular, while for RH

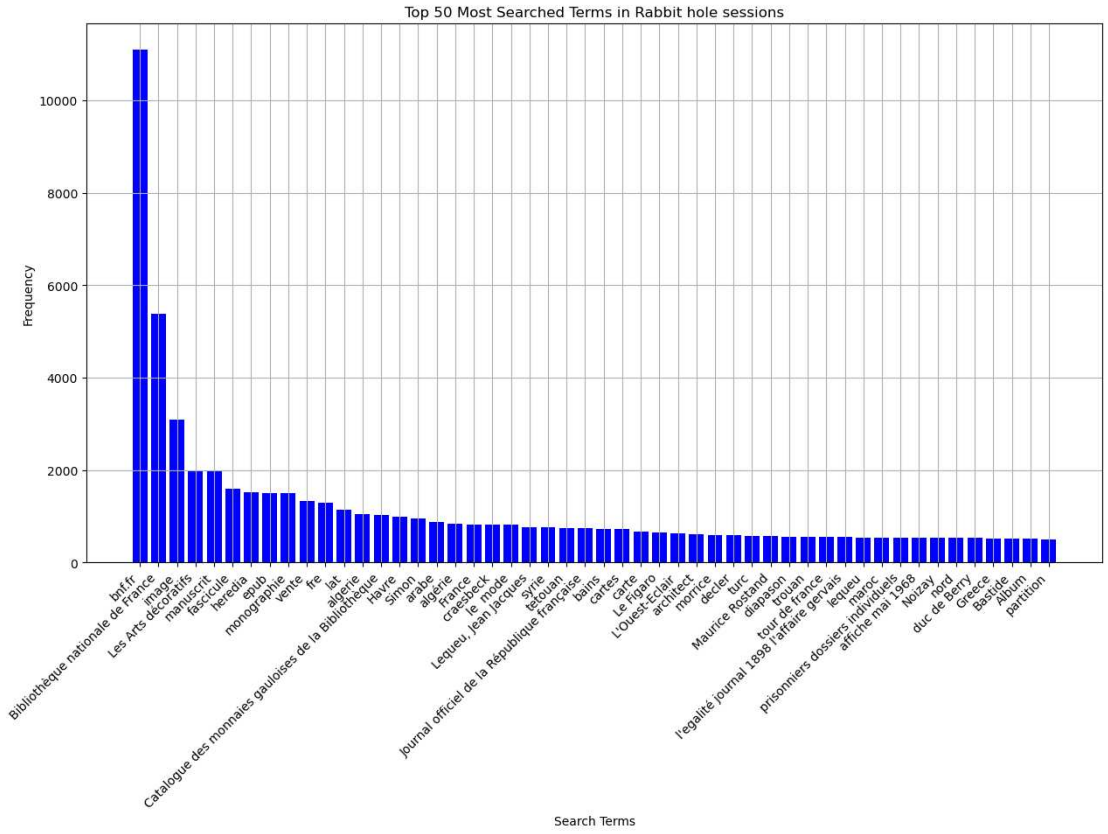
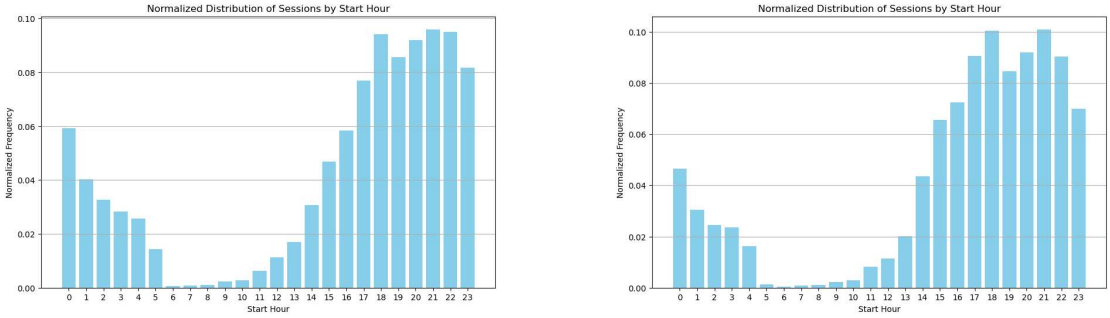


Figure 8: Top 50 search terms in RH sessions



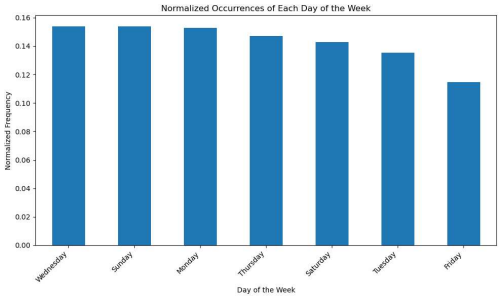
(a) Average sessions

(b) RH sessions

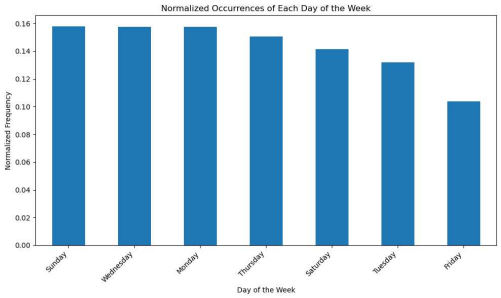
Figure 9: Hour of beginning of session

sessions, Gallica is much closer. Gallica's faulty search feature could then be used by users wandering and less concerned by accuracy.

Finally, to show the difference between average and rabbit hole sessions, we plot their differences: length, number of themes, number of types and number of documents (figures 12, 13,

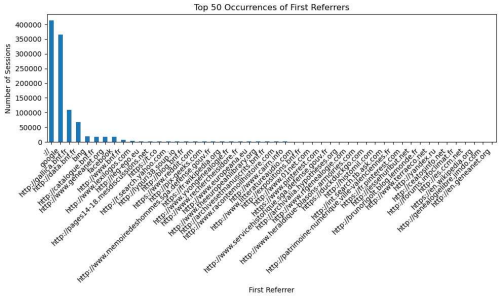


(a) Average sessions

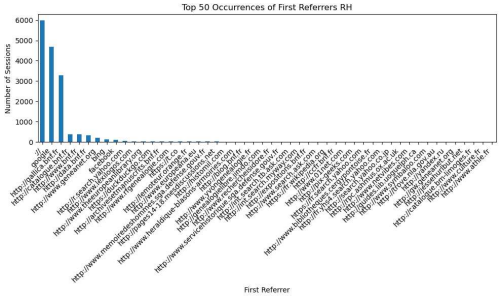


(b) RH sessions

Figure 10: Day of beginning of session



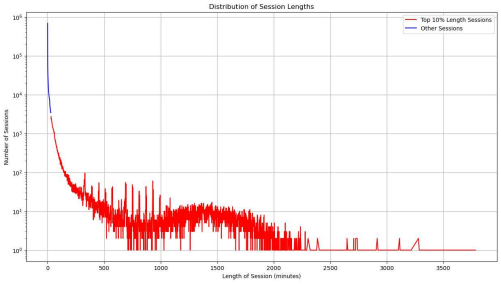
(a) Average sessions



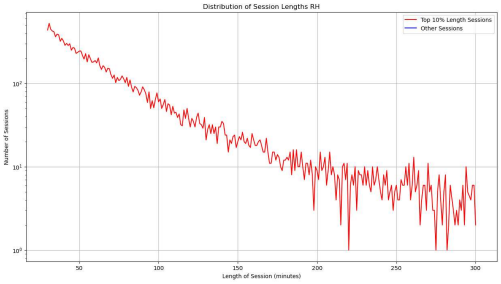
(b) RH sessions

Figure 11: Most common referrer

14 and 15).



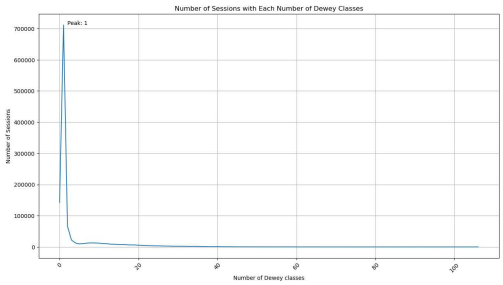
(a) Average sessions



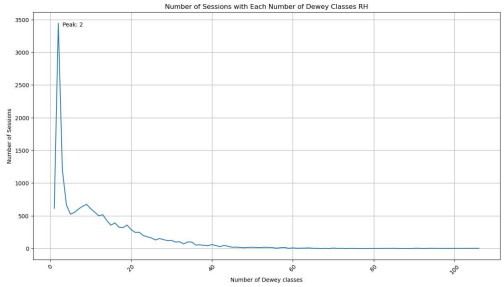
(b) RH sessions

Figure 12: Length of a session

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

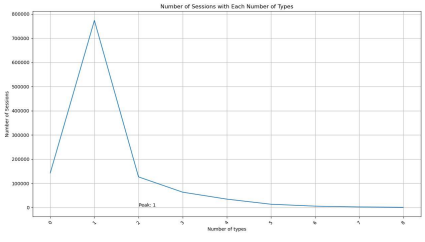


(a) Average sessions

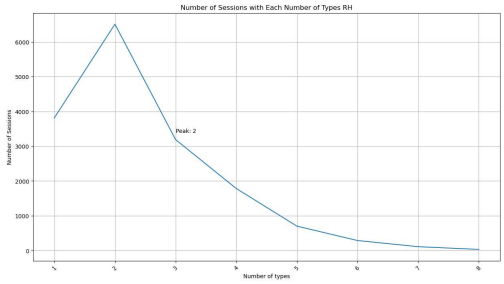


(b) RH sessions

Figure 13: Number of themes

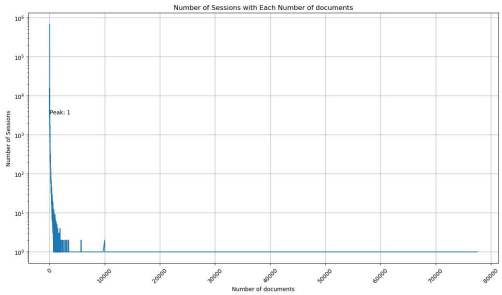


(a) Average sessions

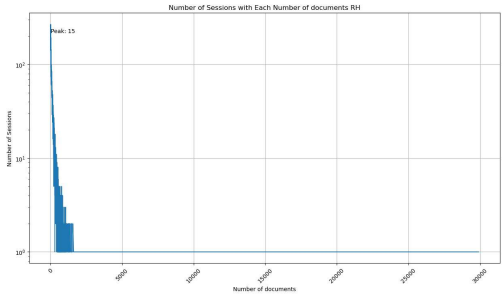


(b) RH sessions

Figure 14: Number of types



(a) Average sessions



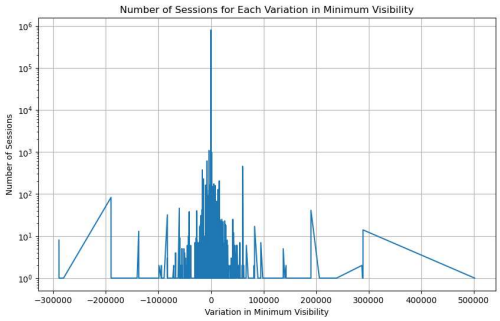
(b) RH sessions

Figure 15: Number of documents

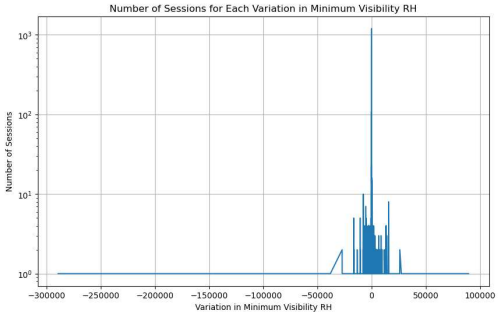
are systematically more documents.

Lastly we want to examine the variation in minimum and mean visibility on both types of sessions.

Variation of the minimum visibility in average 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 (figure 16). In RH sessions, it is slightly more likely to have

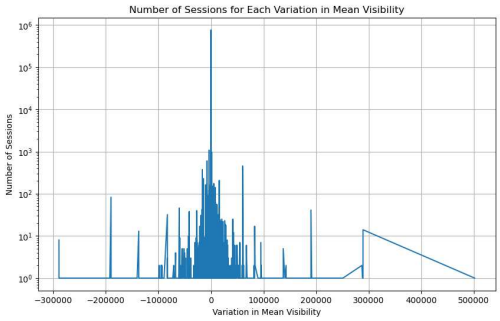


(a) Average sessions

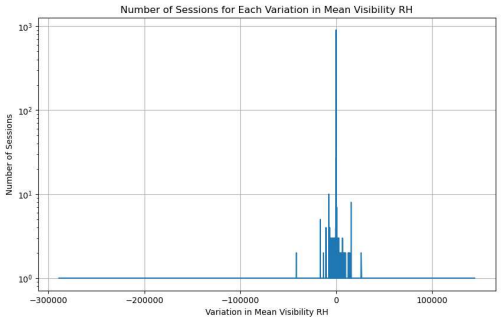


(b) RH sessions

Figure 16: Variation of minimum visibility



(a) Average sessions



(b) RH sessions

Figure 17: Variation of mean visibility

a negative variation, which indicates a session that leads to less visible content. The same behaviour is observed on the mean variation (figure 17).

5. Conclusion

This work-in-progress shows that we can indeed circumscribe a robust subcorpus of long and diversified navigation sessions that we call ”rabbit holes”. They can be considered a subregime of what we previously identified as ”exploratory non-structured navigation” and represent about 1% of overall sessions (including non-human sessions). Interestingly enough, from a merely statistical point of view, these sessions do not appear too different from average sessions and can only be distinguished when looking at significant but subtle details such as their prevalence after working hours or the importance of monographs.

These observations are in line both with users’ testimonies—e.g. one interviewee told us ”I’m a modernist, but as a historian, requesting something about Joan of Arc is already entertainment for me, a rabbit hole in sum”—, as with the digital library’s information architecture—a weakly structured database of rather homogeneous documents in content but not in type. And

indeed, we identified in previous work that the condition for Gallica's corpus to be actually navigable, rather than just retrievable, is that its impressive mass is both heterogeneous in types (prints, manuscripts, musical scores...) and periods (from Antiquity to Modern Times), but rather homogeneous in content (historical documents in the public domain). Thereby, it attracts a rather homogeneous audience, at least a community of practice which it is difficult to segment as said practices do not differ much from one another.

Searching for rabbit holes within Gallica's server logs amounts to looking for a needle in a haystack! Therefore, rather than employing cutting-edge algorithms straightaway, it was necessary to criss-cross multiple perspectives on the phenomenon—preparatory TDA work, interviews, qualitative study of the interface and architecture, extracting a subcorpus, computing simple statistics—in order to build a body of small corroborating evidence. This critical appraisal will now allow us to deploy slightly more complex methods to characterize rabbits holes, notably Markov chains, sequential pattern mining, and topological data analysis, in the hope of articulating more clearly this strand of research with our previous findings [13].

Turning back to matters of accessibility and discoverability, one important conclusion of this study is that rabbit holes on Gallica do not generally lead users towards less consulted content. Longer and more diversified sessions do not wander in the long tail of the corpus. This, we think, is due to the fact that the central tool to access documents in Gallica is a search engine that users have to hack and bypass if they want to build themselves navigation paths. Indeed, we showed elsewhere that precisely because this search engine is somewhat "faulty"—according merely to relevance criteria—it helps users "manage the noise" to their taste in the search results [9]. Due to policies of mass digitization initiated in the 2000s, Gallica can today boast an impressive corpus of over 10 million documents but it still lacks proper *navigation* tools; and we believe that "fixing" the search engine might do more harm than good to the discoverability of its content.

Acknowledgments

This exploratory study was led in collaboration between EPFL and OpenEdition. It was supported by the BnF thanks to a Mark Pigott Fellowship in Digital Humanities.

References

[1] C. Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*. New York: Hyperion, 2006.

[2] M. J. Bates. "The design of browsing and berrypicking techniques for the online search interface". In: *Online Review* 13.5 (1989), pp. 407–424. URL: https://pages.gseis.ucla.edu/faculty/bates/articles/berrypicking.pdf.

[3] M. Brown, J. Bisbee, A. Lai, R. Bonneau, J. Nagler, and J. A. Tucker. "Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users". In: *SSRN Electronic Journal* (2022). DOI: 10.2139/ssrn.4114905. URL: https://www.ssrn.com/abstract=4114905.

[4] D. O. Case. *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*. Academic Press, 2002.

[5] L. Chan, ed. *Contextualizing Openness: Situating Open Science*. Ottawa: University of Ottawa Press, 2019.

[6] D. Dimitrov, F. Lemmerich, F. Flöck, and M. Strohmaier. "Different topic, different trafic: How search and navigation interplay on wikipedia". In: *The Journal of Web Science* 1 (2019). DOI: https://doi.org/10.34962/jws-71.

[7] D. Dimitrov, F. Lemmerich, F. Flöck, and M. Strohmaier. "Query for Architecture, Click through Military: Comparing the Roles of Search and Navigation on Wikipedia". In: *Proceedings of the Conference on Web Science (WebSci), 2018.* Conference on Web Science (WebSci), 2018. 2018. DOI: https://doi.org/10.1145/3201064.3201092.

[8] D. Dimitrov, P. Singer, F. Lemmerich, and M. Strohmaier. "What Makes a Link Successful on Wikipedia?" In: *Proceedings of the International World Wide Web Conference (WWW)*. International World Wide Web Conference (WWW). 2017. DOI: https://doi.org/10.48550/arXiv.1611.02508.

[9] S. Dumas Primbault. ""Managing the Noise": Users' Hacks of Gallica's Search Engine as a Navigational Practice". In: (forthcoming 2025).

[10] S. Dumas Primbault. "Naviguer dans les savoirs à l'ère numérique. Pour une ethnographie des pratiques informationnelles sur Gallica". In: *Études de communication* 61 (Dec. 18, 2023), pp. 61–89. DOI: 10.4000/edc.16108. URL: http://journals.openedition.org/edc/16108.

[11] A. Halfaker, O. Keyes, D. Kluver, J. Thebault-Spieker, T. Nguyen, K. Grandprey-Shores, A. Uduwage, and M. Warncke-Wang. "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, May 18, 2015, pp. 410–418. DOI: 10.1145/2736277.2741117. URL: https://dl.acm.org/doi/10.1145/2736277.2741117.

[12] B. H. Huberman. *The Laws of the Web: Patterns in the Ecology of Information*. Cambridge, MA: MIT Press, 2001.

[13] B. Kaabachi and S. Dumas Primbault. "A Topological Data Analysis of Navigation Paths within Digital Libraries ". In: (2023). URL: https://ceur-ws.org/Vol-3558/paper935.pdf.

[14] D. Lamprecht, K. Lerman, D. Helic, and M. Strohmaier. "How the structure of Wikipedia articles influences user navigation". In: *New Review of Hypermedia and Multimedia* 23.1 (2017), pp. 29–50. DOI: https://doi.org/10.1080/13614568.2016.1179798.

[15] L. Magnani. *Discoverability: The Urgent Need of an Ecology of Human Creativity*. Springer, 2022.

[16] D. Nicholas and D. Clark. ""Reading" in the Digital Environment". In: *Learned Publishing* 25.2 (2012), pp. 93–98. DOI: https://doi.org/10.1087/20120203.

[17] A. Nouvellet and V. Beaudouin. “Analyse des traces d’usage de Gallica”. In: *[Rapport de recherche] Télécom ParisTech; Bibliothèque nationale de France* (2017). URL: https://hal.science/hal-01709264.

[18] T. Piccardi, M. Gerlach, A. Arora, and R. West. “A Large-Scale Characterization of How Readers Browse Wikipedia”. In: *ACM Transactions on the Web* 17.2 (2023), pp. 1–22. DOI: 10.1145/3580318. URL: http://dx.doi.org/10.1145/3580318.

[19] T. Piccardi, M. Gerlach, and R. West. “Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions”. In: *Companion Proceedings of the Web Conference*. WWW ’22: The ACM Web Conference. Virtual Event, Lyon France: Acm, Apr. 25, 2022, pp. 1324–1330. DOI: 10.1145/3487553.3524930. URL: https://dl.acm.org/doi/10.1145/3487553.3524930.

[20] T. Piccardi, M. Redi, G. Colavizza, and R. West. “Quantifying engagement with citations on Wikipedia”. In: *Proceedings of the International World Wide Web Conference (WWW)*. International World Wide Web Conference (WWW). 2020. DOI: https://doi.org/10.48550/arXiv.2001.08614.

[21] G. C. Rodi, V. Loreto, and F. Tria. “Search strategies of Wikipedia readers”. In: *PloS one* 12.2 (2017). DOI: https://doi.org/10.1371/journal.pone.0170746.

[22] R. M. Sutton and K. M. Douglas. “Rabbit Hole Syndrome: Inadvertent, accelerating, and entrenched commitment to conspiracy beliefs”. In: *Current Opinion in Psychology* 48 (Dec. 2022), p. 101462. DOI: 10.1016/j.copsyc.2022.101462. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352250X2200183X.

[23] R. West and J. Leskovec. “Human Wayfinding in Information Networks”. In: *Proceedings of the International World Wide Web Conference (WWW)*. International World Wide Web Conference (WWW). 2012. DOI: https://doi.org/10.1145/2187836.2187920.

[24] L. Woolcott and A. Shiri, eds. *Discoverability in Digital Repositories Systems, Perspectives, and User Studies*. Routledge, 2023.

[25] J. Zhang and A. Ghorbani. “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. May 2004, pp. 315–322. DOI: 10.1109/dnsr.2004.1344744. URL: https://ieeexplore.ieee.org/document/1344744.

A. Data and pre-processing

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

Some of these fields may be empty, and then represented as either ”null” or ’-’. Two example of logs are shown in figures 18 and 19.

A.1. ARK (Archival Resource Key)

As mentioned, some requests contain ARKs that can be extracted from them, for example from figure 18, we could extract the ARK ”bpt6k20211m”. These Archival Resource Keys repre-

Table 4
Features description

Feature	Description
Hashed IP address	Anonymized IP address
Country and City	Location of the request
Date of the request	Format: day/month/year:hour:minute time zone offset
HTTP request	Contains additional information such as the ARK
Protocol	Communication protocol
Response number	HTTP response code
Length	Length of the request
Referrer	Website the user comes from

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

Figure 18: 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 19: A log with all information

sent unique identifiers for each document, from which document metadata can be obtained by querying Gallica's website. These ARKs don't change over time.

An ARK request is shows in figure 20: (figure taken from [17]) The NAAN (name assigning

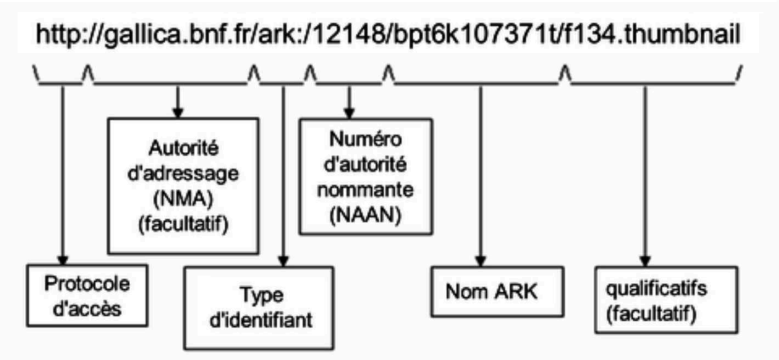


Figure 20: An ARK request to Gallica

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.

A.2. Enrichment

From our logs, we extract meaningful features and the ARK. 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. An example of what the request "https://gallica.bnf.fr/services/OAIRecord?ark=btv1b6907077k" yields is shown in figure 21. We can obtain the type of document under the **typedoc** field, here an image, and the theme,

```
−<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 21: Result of an OAIRecord query

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 hashed-IP address (a person can only contribute one view to a document), and store

the ARK and its associated count for later use.