Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions

Anne-Laure Tettoni, Simon Dumas Primbault

Computational Humanities Research
Fifth conference | 4–6 December 2024
Aarhus, Denmark

Down the Rabbit Hole

Discoverability in a Digital Library

Anne-Laure Tettoni
Laboratory for the History of Science and Technology (LHST)
Swiss Federal Institute of Technology (EPFL)

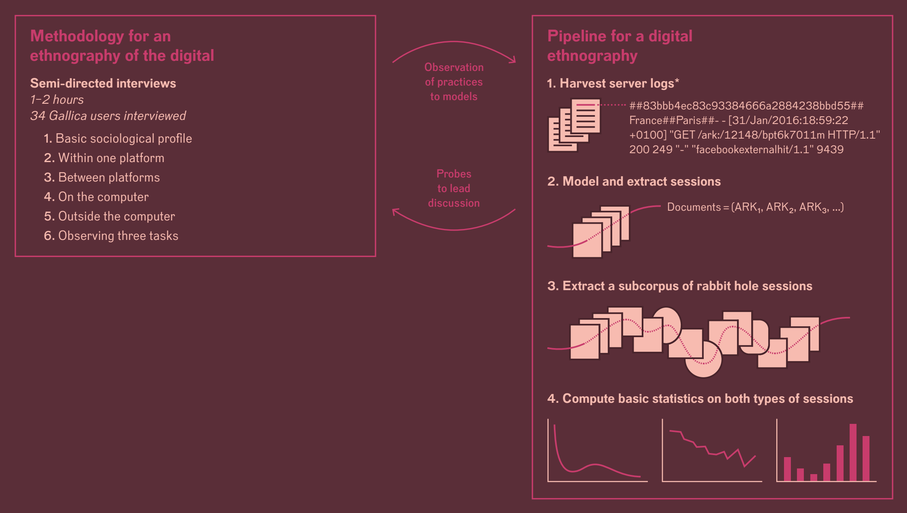
Simon Dumas Primbault
Laboratory for the History of Science and Technology (LHST)
Swiss Federal Institute of Technology (EPFL)
OpenEdition (UAR 2504, CNRS/EHESS/AMU/AU)
Bibliothèque nationale de France (BnF)

Who has never found themselves surfing almost aimlessly throughout the vastness of content on the Internet?

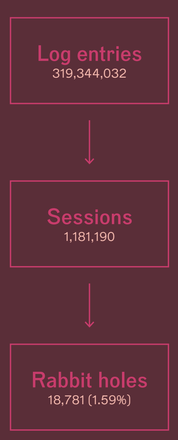
Indeed, the hypertextual structure of most content on the World Wide Web allows users to navigate from page to page—either by curiosity, distraction, or mere boredom—, to the extent that, not unlike Alice in Wonderland, they may fall "down the rabbit hole", supposedly in the most unknown of places. These 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?

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. Digital libraries such as Gallica, the French national library's online platform, grow exponentially, and become genuine informational milieus that users navigate step-by-step and iteratively. In this context, recommendation systems tend to direct users towards more popular content, at the expense of less visible one, which invisibles certain authors, languages or media types. An information system that promotes discoverability could give users access to less visible or seen content, but find useful nonetheless.

A Mixed-Methods Ethnography: Pipeline and Methodology



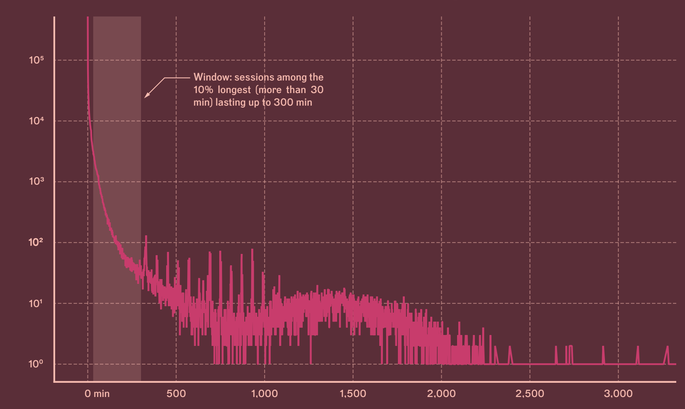
* **A log entry** is a string from which we can extract the following meaningful features: anonymised IP address, location of the request, date of the request, HTTP request information (such as the ARK), communication protocol, HTTP response code, length of the request, referrer (website the user comes from).



Sessionisation and extraction of rabbit holes

Gallica server logs span the period from 31st Jan. 2016, 13:00, to 29th Feb., 05:36. Logs are enriched by requesting individual Archival Resources Keys (ARKs) to provide additional information about the documents consulted, such as type and theme. User sessions are created by grouping logs by hashed-IP addresses and aggregating the features, then separating them using a given inactivity threshold of 60 minutes. Sessions are enriched 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. We have 1,181,190 sessions where at least one document was consulted.

A "rabbit hole" session is a session that is long and diversified. We use ARKs to define diversity by accessing themes and types of consulted documents, and take sessions whose length is in the top 10%, but no longer than 5 hours, to exclude robotic behaviour.



Results

Conclusion

→ It is possible to 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).

→ These sessions' differences from average sessions are subtle but significant.

→ Differences are most notable in the starting time of sessions in the day and referrers.

→ User oral testimonies align with computational observations.

→ The users falling in rabbit holes are the same users navigating the platform for average sessions.

→ Rabbit holes do not generally lead users towards less consulted content.

→ Gallica's somewhat "faulty" (in terms of relevance) search engine helps users "manage the noise" to their taste in the search results.

→ "Fixing" the search engine would harm content discoverability.

Selected references

T. Piccardi, M. Gerlach, R. West, "Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions" (2022). dl.acm.org/doi/10.1145/3487553.3524930

B. Kaabachi, S. Dumas Primbault, "A Topological Data Analysis of Navigation Paths within Digital Libraries" (2023). ceur-ws.org/Vol-3558/paper935.pdf

A. Nouvellet, V. Beaudouin, F. d'Alché-Buc, C. Prieur, F. Roueff, "Analyse des traces d'usage de Gallica" (2017). hal.science/hal-01709264

S. Dumas Primbault, "Naviguer dans les savoirs à l'ère numérique. Pour une ethnographie des pratiques informationnelles sur Gallica" (2023). journals.openedition.org/edc/16108

All the code produced for this project is open source and available on the following GitHub repository: github.com/ana571/rabbit-holes-gallica.

The related paper can be freely read or downloaded here: ceur-ws.org/Vol-3834/paper78.pdf.

This project was funded by the Laboratory for the History of Science and Technology (EPFL) and the 2022 Mark Pigott Grant for the Digital Humanities, awarded by the Bibliothèque nationale de France.

EPFL College of Humanities · Laboratory for the History of Science and Technology
BnF Gallica
OpenEdition Lab

Graphic design: C. B.