# Digital Trail Blazing

## Understanding and Remaking Intellectual Mobility on Online Research Platforms

A Quick Overview of the Quantitative Approach

Simon Dumas Primbault, LHST (EPFL)
Jérôme Baudry, LHST (EPFL)
Jean-François Bert, IHAR (UNIL)
Bayrem Kaabachi, Master student in data science (EPFL)

Context: In the past decades, the traditional documentary practices used in physical places of knowledge such as libraries and archives — searching catalogues and bibliographies, browsing through shelving, taking notes... — have been supplemented with a series of digital practices necessary for researchers to consult websites, online platforms, and databases — searching by keywords, filtering results, navigating through links... However, while cultural ethnography and the anthropology of knowledge have addressed matters of material practices in the making of science in libraries, laboratories, and archives, too few humanistic studies have yet endeavored to document the navigation practices of scholars within digital environments. Moreover, studies in data science and user experience have recently shown that search engines never fully satisfy users who tend to rely more on step-by-step contextual navigation, in other words on digital mobility.

In an effort to understand our digital culture, recent studies have demonstrated the importance of online mobility. Indeed, the investigation of "interaction traces" (understood as the sequential records of a user's interactions within one or across multiple platforms) shows that users favor personal navigation over search algorithms, steering their own personal course through an ever-growing cloud of data. Bringing ethnographical queries and concepts into the picture and drawing on recent efforts in data science to model different types of navigation practices, this project aims at furthering the inquiry on intellectual mobilities in the digital era.

**N.B.: the quantitative approach (digital ethnography) of this research project presented here is coupled with a qualitative approach (ethnography *of the* digital) made up of semi-structured interviews with users, structured observation of users' practices, analysis of the interface, discussions with the team that maintains the platform… Experience showed that the two approaches match well and that the shuttling back and forth is beneficial to both.**

## 1. Data

Gallica = online platform of the Bibliothèque nationale de France

- Almost 7M documents online
- Books, manuscripts, newspapers, photographs, posters, scores, etc.
- All disciplines, but mostly users in the human sciences
- Only research tools (no suggestion algorithm) *i.e.* studying user navigation is relevant and will not produce an artefact generated by a recommendation system

Data = one year (April 2016 – April 2017) of Gallica server logs

- Structure: Hashed IP – Country – City – Timestamp – HTTP request – Protocol – Answer code – Length – Referring website
- Example of a successful HTTP request:

```
##6958a5de61066cceb1831af6e2f0fc76##United States##Madison##- -
[04/Mar/2017:00:31:03 +0100] "GET /ark:/12148/bpt6k9708547.lowres HTTP/1.1" 200 78109
"http://gallica.bnf.fr/" "Mozilla/5.0 (Windows NT 6.1; Trident/7.0; rv:11.0) like Gecko"
```

- We only use 3 months for now

For future research:

Newer and bigger data from Gallica
Data directly from our interviewees
Expand to other platforms: Wikipedia (non researchers), arXiv (natural sciences)…

## 2. From Data to Corpus

Reading paths = series of 3 or more non-redundant documents consulted within 60mn from each other

- When two consecutive requests from the same IP are for the same document, it means the user zoomed or flicked pages, we merge these into one document only
- When 60mn have passed between two documents, we consider it is another path/user/session
- When more than 100 requests are made by minute by the same IP, we consider it is a crawler bot or a fixed IP from an institution (in interviews, users said using multiple tabs which can result in up to a few tens of requests within a minute)

Example: $Path_{example} = (Doc_A, t_1; Doc_B, t_2; Doc_C, t_3; Doc_A, t_4; Doc_D, t_5; \ldots)$

Enrich logs with ARK (Archival Resource Key) through Gallica's API

Metadata = Author, Title, Year, Discipline

Where Discipline = first two levels of the Dewey classification

E.g., 500 Natural sciences and mathematics => 510 Mathematics

All calculation is sent to IC cluster via a kubernetes pod

For future research:

Finer-grained discipline classification (e.g. 514 Topology)

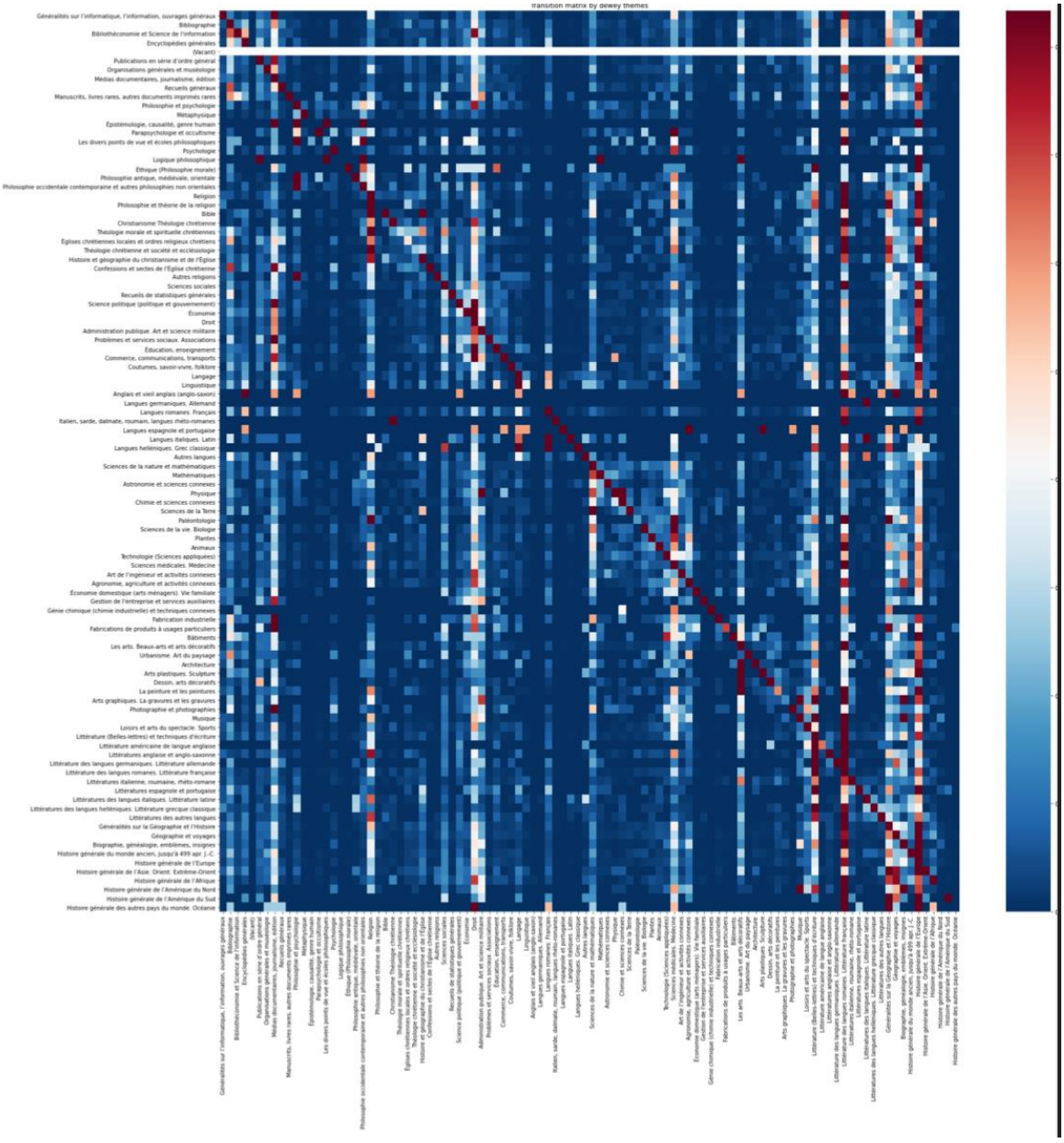
Rather than Dewey classification, infer topics from OCR of documents consulted?

## 3. Random walks

1st model = first-order Markov chains

Transition probability from one discipline to another

Transition matrix from line-topic to column-topic (robust heatmap with quantile color scale)



Observations:

Although interviewees say they never use the Dewey classification, it still appears relevant to analyze their logs

Users tend to stay within one discipline

They only venture in adjacent fields

Some disciplines act as pivots (General topics, Encyclopedias, Dictionaries, Collections…)

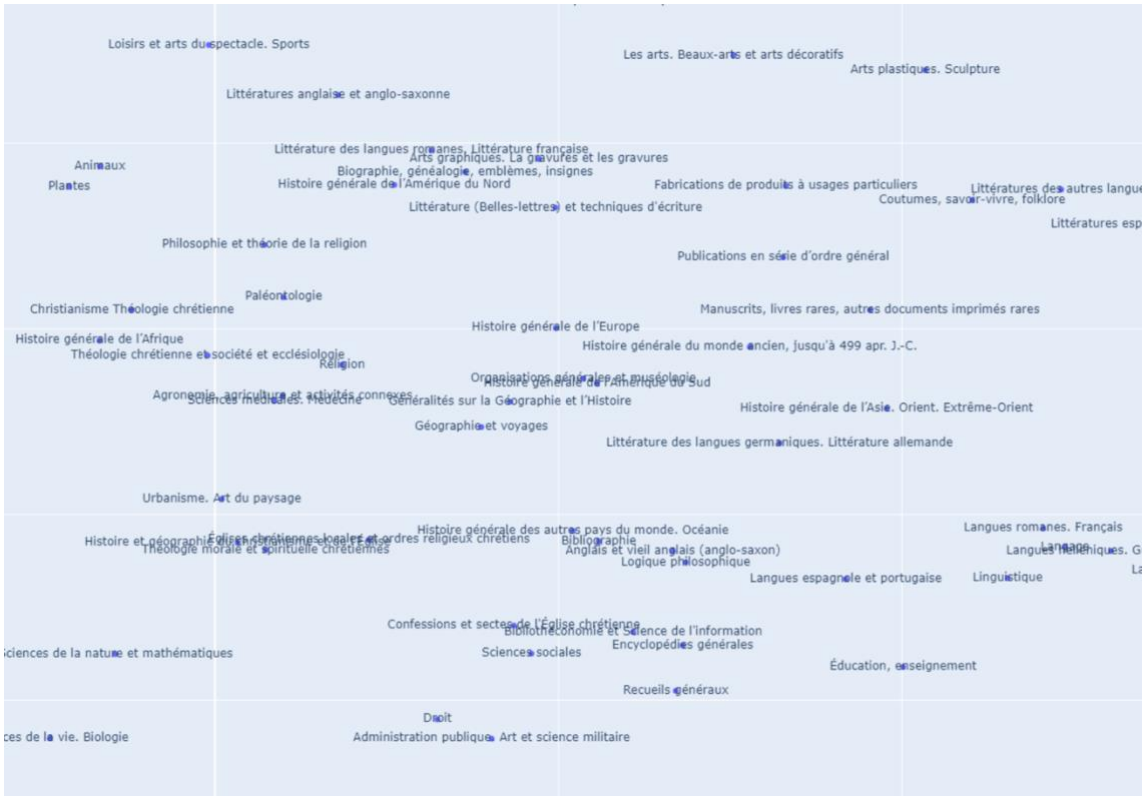
For future research:

Interviewees said they opened multiple tabs => 2nd- or 3rd-order Markov chains…

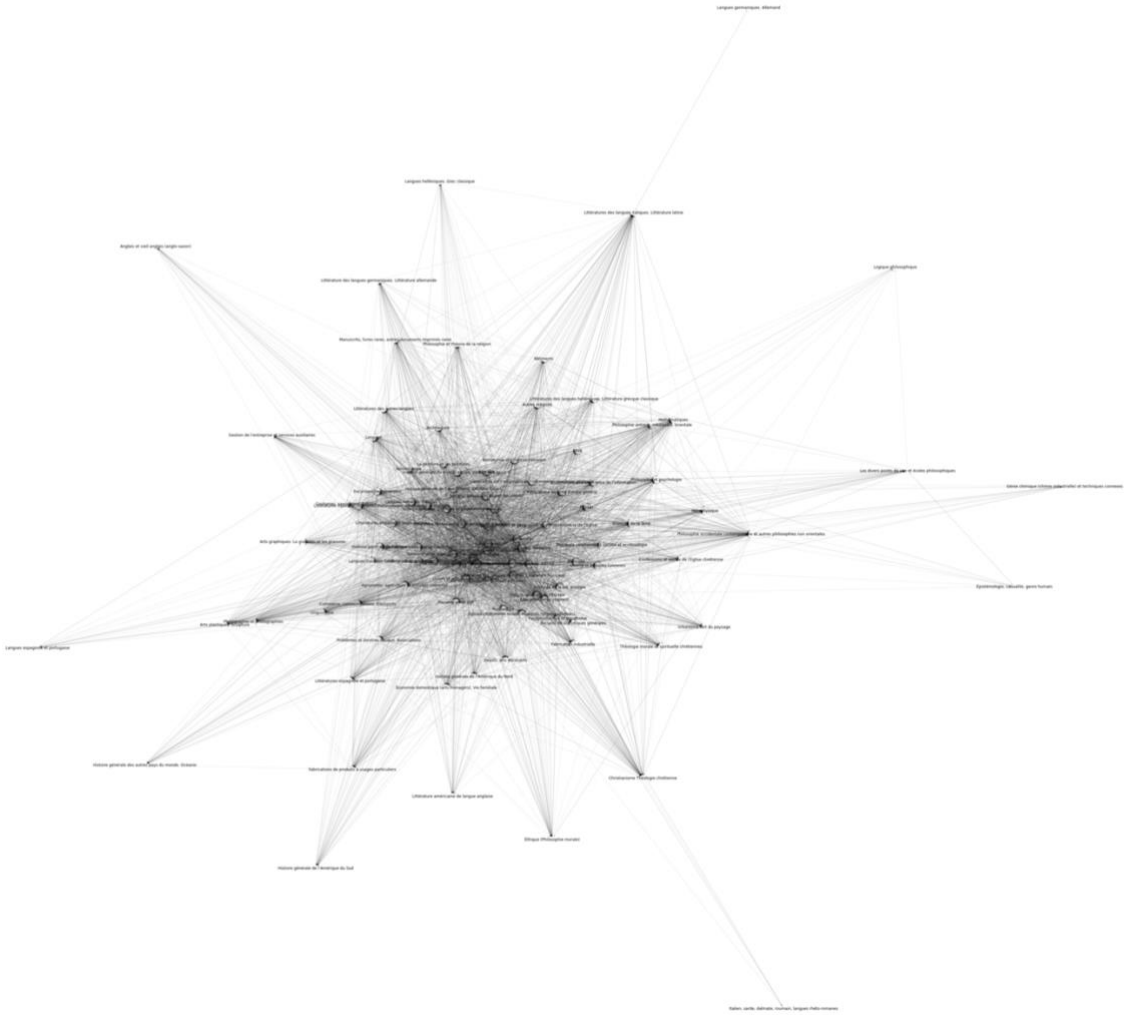
## 4. A Cartography of Knowledge

Big question 1: What metric space to embed and visualize paths? = What kind of cartography of knowledge to draw with this data?

- The tree structure of the Dewey classification is arbitrary and cannot be meaningful (alone)
- Model 1: word2vec embedding
  where "words" = disciplines and "sentences" = path of disciplines
  e.g. $Path_{disc}$ = ($Disc_1$; $Disc_2$; $Disc_3$; $Disc_2$; …)
  => pairwise distances between disciplines
  tSNE projection (only for matters of visualization)



- Model 2: network analysis (betweenness centrality)
  2D viz (Fruchterman-Reingold force-directed algorithm)



Observations:

Predominance of human and social sciences

2D viz seems relevant although tSNE does not seem to be the most appropriate projection method

Refining with the century of the document consulted only adds noise (twice same discipline from two different centuries are not separated enough to be meaningful)

We find the same pivot disciplines as the most "socially" linked in the network analysis

<u>For future research:</u>

Local projections

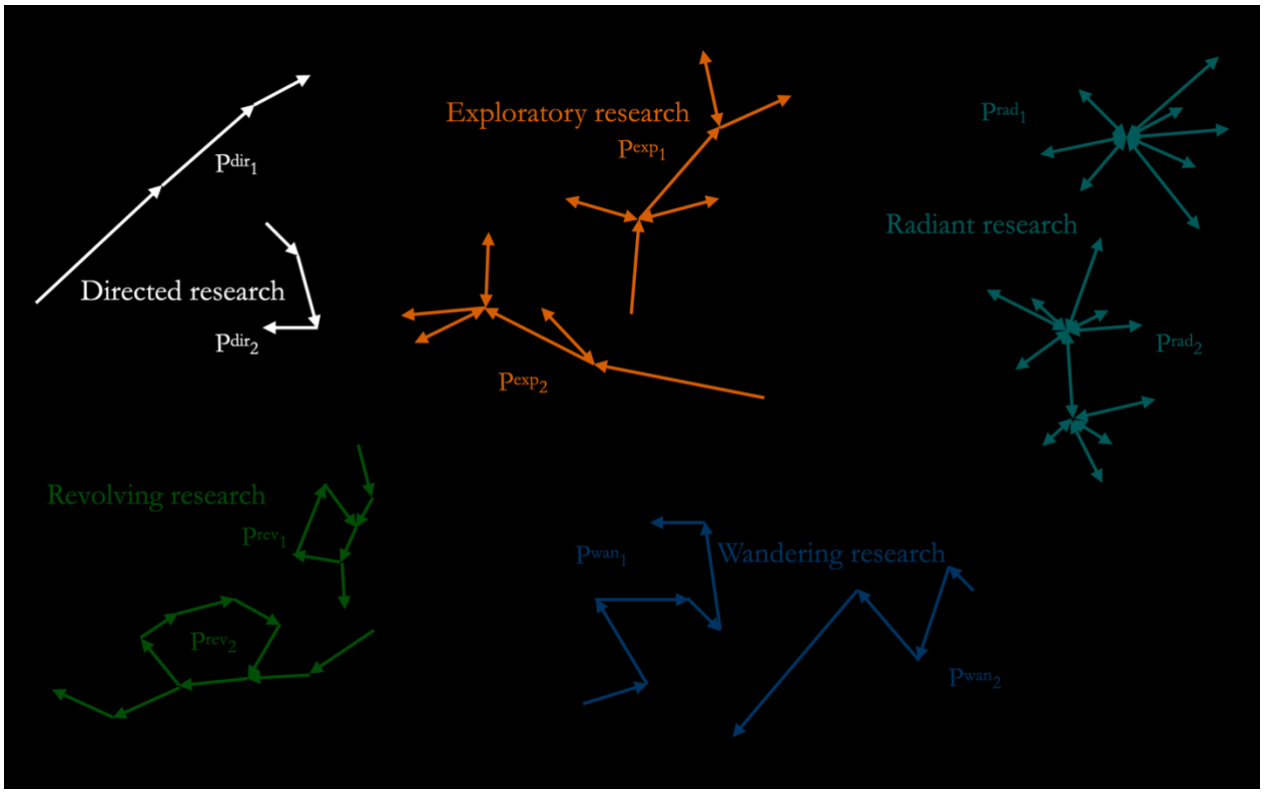
Integrate timestamp in path disciplines: two disciplines are further apart if it takes more time to jump from one to another

Rather than just the usual 2D map, possibility to infer the topology of a manifold from simplicial complex generated by all paths?
+ diachronic evolution of fields over the years = visualize the dynamic of knowledge reconfiguration

## 5. Taxonomy of navigation

Big question 2: Can we cluster paths according to their topology? = Can we build a taxonomy of researchers' online behaviour based on navigation practices only?

Fictional clusters based on users' interviews




Current TDA pipeline:

1. Draw individual paths in the space generated by the word2vec pairwise distances
2. Compute persistence diagram for each path
3. Compute pairwise Wasserstein (or Bottleneck?) distances between paths
4. Clustering (exact method still to be refined)

Observations:

While some users say they search Gallica only to find very specific documents quickly (directed research), others say they browse/navigate the corpus through associations of ideas, references, curiosity… (they mention "games," "starlike navigation," "ring roads," and "slip roads" etc.)

Minimal path length for non-trivial topology = 5
Refine the metric
Compute the metric on one time period and the consequent PDs on the following

**Work in progress**

For future research:

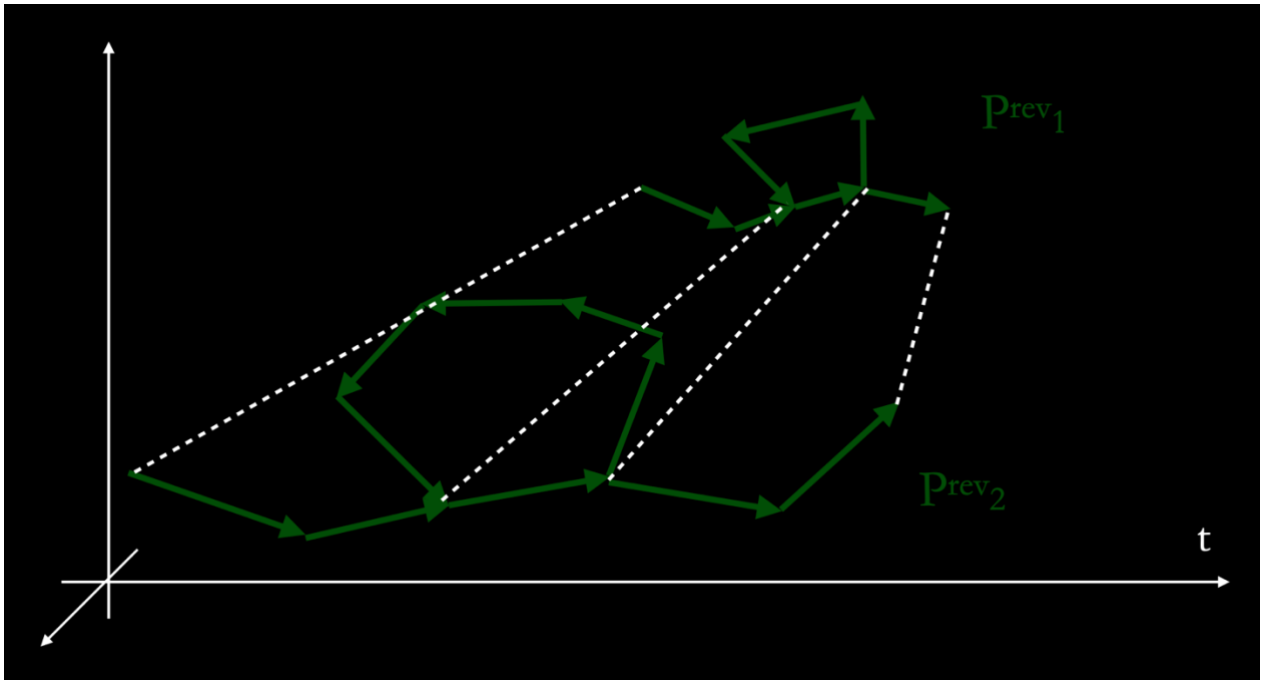
Random walks on simplicial complexes (*i.e.* on the approximate manifold generated at the previous step as a cartography of knowledge)?

**N.B.: The following is only speculation for now.**

## 6. Finer-grained taxonomy

Big question 3: Is It possible to refine the previous taxonomy by geometric comparison of paths within the same cluster?

Dynamic Time Warp?




## 7. Path-Suggestion

Big question 4: Can we predict paths according to certain criteria? = Can we recommend readings either along mean paths or, on the contrary, off the beaten track?

! Not to capture attention

Rather, either for newcomers, following the trodden path (that of an introductory course to TDA through multiple readings), or for others, try new associations, new links, new navigation/mobilities (blaze a new trail).

For further research:

Simplicial neural networks for trajectory prediction?