A Topological Data Analysis of Navigation Paths within Digital Libraries ⋆

Bayrem Kaabachi[1,2,*,†], Simon Dumas Primbault[1,3,4,†]

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

[2]Biomedical Data Science Center (BDSC), Centre Hospitalier Universitaire Vaudois (CHUV), CH-1002 Lausanne, Switzerland

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

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

Abstract

The digitization of library resources and services have opened up physical informational spaces to new dimensions by allowing users to access a wealth of documents in ways that differ from browsing bookshelves traditionally organized according to the "tree of knowledge". How do readers of digital library orient themselves within big corpora? What landmarks do they use to navigate masses of digital documents? Taking Gallica as a case study–the digital heritage platform of the French national library–, this paper presents an experimental research on the navigation practices of its users. Using methods from topological data analysis, we inferred from Gallica's server logs an informational space as it is roamed by readers. Coupled with user interviews, this mixed-methods study allowed us to identify a set of "regimes of navigation" characterizing how readers deploy various strategies to browse the digital library's corpus. From directed search to wandering to crawling, these regimes answer different needs and show that a single corpus can, in turns, be apprehended as a heritage collection, a database, a set of documents, and a mass of information.

Keywords

digital library, navigation practices, topological data analysis, information retrieval

1. Introduction

1.1. Research question

The birth and development of digital libraries–broadly understood as curated collections of electronic documents accessible online on dedicated platforms with tools for search and consultation [6]–have radically transformed research activities. From any computer with Internet access, the wealth of information available allows for the simultaneous consultation of a great variety of resources and for their continual rearrangement into renewed information

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France

*Corresponding author.

†These authors contributed equally.

bayrem.kaabachi@gmail.com (B. Kaabachi); simon.dumas-primbault@openedition.org (S. Dumas Primbault)

https://cv.hal.science/simon-dumas-primbault (S. Dumas Primbault)

https://orcid.org/0009-0002-7534-8493 (B. Kaabachi); https://orcid.org/0000-0002-0012-0550 (S. Dumas Primbault)

landscapes. Consequently, traditional practices observed in physical places of knowledge such as libraries and archives–searching catalogues, browsing through shelving, taking notes–have been supplemented with a series of digital practices developed by researchers to browse websites and databases–searching by keywords, filtering results, navigating through links.

While historically science was, and to a great extent still is, arranged into arborescent taxonomies, scholars have in practice always negotiated with this normative order of knowledge, resulting in a complex landscape much thicker than a tree-like structure. In contributing to the digitization of information practices, digital libraries are further redrawing this landscape and its relation to established orders of knowledge. The appropriation of digital libraries by their users opens the "tree of knowledge" to a multiplicity of other dimensions allowing for leaps from one "leaf" to another, as well as a continuous reconfiguration of branches of knowledge. Once we become aware that the practical landscape of knowledge is evolving into a much more dynamic arrangement, the use of spatial concepts to understand research practices from the point of view of scholars becomes essential.

1.2. Literature review

Research on digital information practices attracted data science in the late 2000s, focusing more specifically on "interaction traces" (understood as the sequential records of a user's interactions with one or across multiple platforms). Aiming at collecting, quantifying, and modelling online traces, the investigation of interaction traces helped emphasize the role of navigation, as opposed to search, in seeking information: on the one hand, "the perfect search engine is not enough" therefore prompting user to browse [22], while, on the other hand, existing recommendation algorithms do not favour navigability, even hindering it [14].

Yet, inheritors of a rather mechanical model supported by "information foraging theory", these studies are still too often dedicated to practical optimisation and problem-solving—aimed at enhancing the ranking of pages found through "post-query navigation" [5], [20], predicting the next page based on statistical regularities [19], [13], or suggesting the most beaten "trails" [24]. Furthermore, when users are brought to the fore by data science and UX studies, the emphasis is usually put on "directed search"–*i.e.*, shortest, directed, and local paths–, thereby neglecting a whole part of navigation strategies: crawling, exploratory searches, serendipity...

When eschewing this way more traditional ethnographic approaches based on interviews and on-site observations, navigation is deemed worthy of practice and study only for its end products or its "waypoints" [25], not as a process in itself, and it is assessed according to relevance only, rather than discovery or originality. Most quantitative studies on navigation therefore bypass altogether a practice that their authors nonetheless underlined as fundamental: search engines, however accurate and powerful, never fully satisfy users who tend to rely more on step-by-step contextual navigation. Only recently studies have been conducted on user navigation *per se*, especially on Wikipedia [18].

Within the realm of computational humanities, the recent development of topological data analysis (TDA) calls for a reappraisal of navigation through the use and honing of tools specifically devoted to the qualitative study of shape. The shape of the World Wide Web is an issue as old as the Internet itself and still widely debated to this day [12], [10]. Traditionally, the Web has been modelled as a graph. Although this has proven to be a powerful tool to study digital

communication, graphs are intrinsically limited to model pairwise interactions. The recent success of topological methods in studying data, and the parallel establishment of topological data analysis as a field [9], have confirmed the utility of viewing data through a higher-dimensional analogue of graphs. Providing the tools to model navigation as a rich and thick process, rather than as a problem to solve, TDA offers the possibility to understand it in a less systematic, more descriptive way, if coupled with humanistic methods addressing navigation in its lived practical thickness. Although TDA has been used to analyse scientific collaborations [17], nothing properly topological has been endeavoured for navigational practices yet.

1.3. Case study and methodology

to understand how digital library users navigate online content, we chose to study Gallica (https://gallica.bnf.fr), the online platform of the French national library (BnF). Gallica preserves and provides access to ten million documents in the public domain, freely available either for download or for consultation on the dedicated online reader. The collection gathers a wide variety of document types (printed books, press, manuscripts, musical scores, maps, images, videos, objects...) in more than ten languages, deriving either from the BnF preservation policies and digitisation campaigns, from other libraries, or from the *dépôt légal*.

Previously, Nouvellet *et al.* [16] already provided an exploration of Gallica 2016 server logs, focusing on the modeling of user sessions as Markov chains and providing first results about typical modes of engagement with resources, their provenance, or the mediation effect of Gallica's blog posts. More recently, Trabelsi [23] used off-the-shelf process mining techniques on these same logs to model users' paths between search pages, document viewing, blogs...

Our approach for the whole project extends Beaudouin *et al.*'s mixed methods [3], [4] by weaving together a socio-ethnography of users' practices (left of fig. 1)–based on semi-structured episodic interviews with seven Gallica users–and a computational ethnography (right of fig. 1)–a topological analysis of server logs understood as interaction traces. These two methods need to be cautiously dovetailed to yield interpretable results: the models used to reconstruct and cluster reading paths from the logs were based on users' testimonies, while the resulting visualisations have been submitted to interviewees for validation or as probes to incentivise discussion.

The main sociological results of this first study are presented in [8] and a semiotic analysis of the visualisations generated is presented in [7].

The present paper endeavours to shed light on the Python pipeline used to process Gallica server logs as part of a computational ethnography of users' navigation practices within digital library. The code is available under a GNU license on https://github.com/Kaabachi/TDA-Gallica.

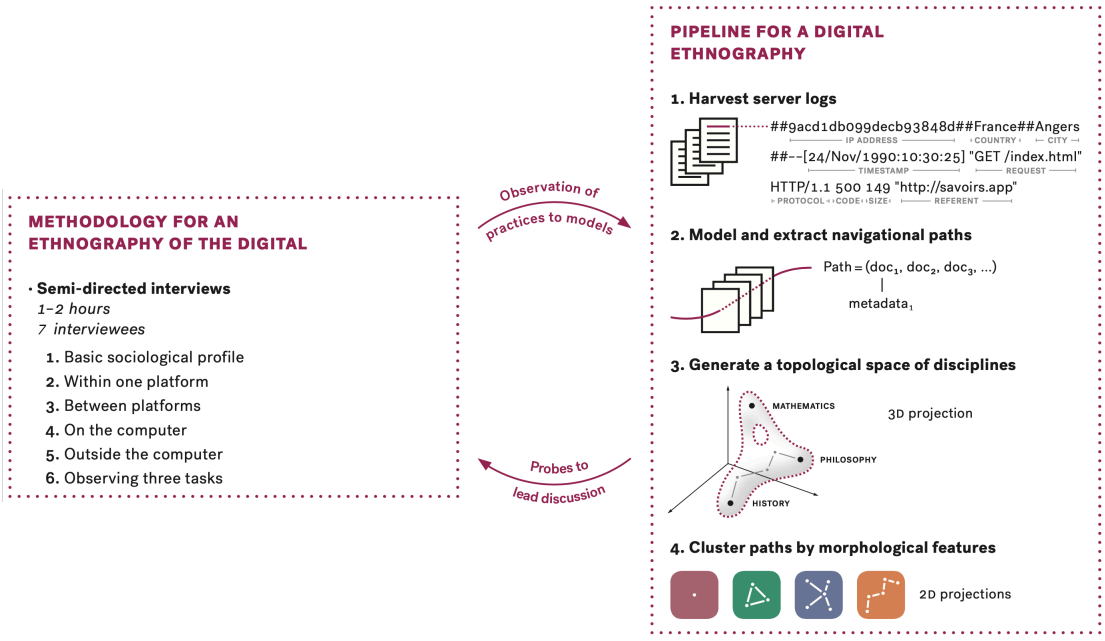


Figure 1: A Mixed-Methods Ethnography: Pipeline and Methodology

2. Dataset and Sessionization

In this study, we work with the browsing logs of Gallica over the month of April 2016.[1] A description of the features contained in the dataset can be seen in Table 1.

Table 1
Dataset Description

Feature	Description
Hashed IP Address	Anonymized IP addresses
Country	Country of requester
City	City of requester
Complete date	Date provided in day/month/year:hour:minutes:seconds format
Request	There are different types of requests, the user may either request an HTML static page or make web-design-related requests such as JavaScript/CSS... The ARK requests are another type of request, which would be the most important for us during this research.
Protocol	Communication protocol
Answer code	response status codes
Length	Length of request
Referring website	Indicates the website from which the user comes.

[1]Compliance with GDPR, both for the qualitative and the quantitative approach, was approved by decision HREC No. 056-2020 of EPFL ethical committee.

Notably, certain fields such as country, city, or referring website may lack specificity due to incomplete information in the requests. These fields are subsequently populated with a "null" value in cases of information absence.

The focus of the project is on exploring user interactions and navigation patterns within Gallica's digital platform. As such, the raw data requires augmentation with relevant document-specific details, considering the inherent limitations of the original dataset.

To enhance the data, we utilize ARK-type requests. These requests act as document identifiers within Gallica's system but do not provide semantically interpretable information about the respective documents. To address this, we systematically extract the unique ARK name associated with each request. Following this, we employ Gallica's dedicated service, designed to extract bibliographical information corresponding to a document, to query the digital platform, as illustrated in Figure 2.

To render our data analysis more informative, we supplement our dataset through the execution of these queries. This approach yields additional features of the accessed documents, such as the title, the year of publication, the primary language, and the prevailing theme. This enhancement of our dataset deepens our understanding of user interactions within Gallica.

In our effort to accurately model a user's journey, akin to navigating a traditional library, we place significant emphasis on the topic, or discipline, of each document accessed by the user. Gallica utilizes the Dewey Decimal Classification system, a widely accepted numerical taxonomy with established hierarchical categories. At a broad level, documents are classified into ten primary classes, based on the first digit, reflecting fundamental disciplines or fields of study. Delving deeper, the second digit enables a more nuanced categorization. For instance, 61$\underline{0}$ is assigned to general works on medicine and health, 61$\underline{1}$ is associated with human anatomy, 61$\underline{2}$ designates human physiology, and 61$\underline{3}$ corresponds to personal health and safety. This granular classification offers a comprehensive and systematic structure for navigating the array of documents.

Once we obtain the core data needed for the study, we separate requests made by distinct users to properly characterize their interactions with the digital platform. Previous work by Nouvellet *et al.* [16] introduced methods to properly distinguish users through the parsing of Gallica logs. In this work, we adopt a similar approach to analyzing users' behavior through the concept of "sessions". A session refers to a single, continuous period during which a user is actively engaging with Gallica's web platform. At the end of this process, we aim to filter the logs so that we only look at active human users[2].

The technique we used to achieve this relies on applying several filters and transformations to the dataset. We first introduced an inactivity threshold of 60 minutes; the absence of any user query after this duration is interpreted as session termination. Subsequent queries by the same user after this interval are considered distinct sessions. We then assign those queries different session numbers according to the hashed IP address linked to it. Finally, we drop the sessions that contain no ARKs at all as they contain no topical information. Note that this is a major limit of this study as only printed documents and prints have ARKs.

Following the aforementioned transformations, we analyze the duration of each session, specifically focusing on sessions that contain a minimum of three ARKs (Archival Resource

[2]Specifying the human part is important, as we want to avoid potential interference from internet bot crawler data.

Exemple : https://gallica.bnf.fr/services/OAIRecord?ark=bpt6k5738219s

```
<?xml version="1.0" encoding="UTF-8" ?>
<results countResults="1" resultType="LuceneOAIRecordSearch" searchTime="0:00:00.001">
  <notice>
    <record xmlns="http://www.openarchives.org/OAI/2.0/">
      <header>
        <identifier>oai:bnf.fr:gallica/ark:/12148/bpt6k5738219s</identifier>
        <datestamp>2012-01-27</datestamp>
        <setSpec>gallica:theme:8:84</setSpec>
        <setSpec>gallica:typedoc:monographies</setSpec>
      </header>
      <metadata>
        <oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/
          <dc:identifier>https://gallica.bnf.fr/ark:/12148/bpt6k5738219s</dc:identifier>
          <dc:title>La plage d'Etretat par l'auteur de "Monsieur X et Mme ***"</dc:title>
          <dc:publisher>Michel Levy (Paris)</dc:publisher>
          <dc:date>1868</dc:date>
          <dc:format>In-18</dc:format>
          <dc:language>fre</dc:language>
          <dc:relation>Notice du catalogue : http://catalogue.bnf.fr/ark:/12148/cb33539190h</dc:relation>
          <dc:type xml:lang="eng">text</dc:type>
          <dc:type xml:lang="fre">monographie imprimée</dc:type>
          <dc:type xml:lang="eng">printed monograph</dc:type>
          <dc:format>application/pdf</dc:format>
          <dc:source>Bibliothèque nationale de France, département Littérature et art, Y2-59413</dc:source>
          <dc:rights xml:lang="fre">domaine public</dc:rights>
          <dc:rights xml:lang="eng">public domain</dc:rights>
        </oai_dc:dc>
      </metadata>
    </record>
  </notice>
  <mode_indexation>text</mode_indexation>
  <nqamoyen>092.57</nqamoyen>
  <provenance>bnf.fr</provenance>
  <source>Bibliothèque nationale de France, département Littérature et art, Y2-59413</source>
  <typedoc>monographies</typedoc>
  <date>1868</date>
  <title>La plage d'Etretat par l'auteur de "Monsieur X et Mme ***"</title>
  <sdewey>84</sdewey>
</results>
```

Figure 2: Example of an OAI query in Gallica

Keys). A predominance of Gallica users tend to consult only a single document as indicated in prior research [16][23]. Consequently, sessions involving the consultation of just one document hold less relevance for our study, which is centered around exploring user transition behaviors. Therefore, these sessions are not included in our analysis. Additionally, we introduce an upper limit, filtering out sessions that incorporate more than 50 consecutive ARKs. This step ensures we manage extreme cases and outliers in our dataset that could potentially skew the analysis.

The selection criteria of a minimum of three ARKs and a maximum of 50 ARKs in a session thus ensure a focused and meaningful investigation of user behavior patterns. Future work might consider different criteria based on the research question at hand or employ different

statistical approaches to handle sessions with varying numbers of ARKs.

$$session = [ark_1, ark_2, ark_3, ark_1, ark_4...] \tag{1}$$

As explained in the previous section, we extract the topics or disciplines from these ARK sequences by utilizing the OAI protocol. The findings reveal a sequence of actions within each session represented as a sequence of themes. This delineates a model for user behavior within a digital library, facilitating a comparison with user movement patterns within a physical library.

$$session = [DeweyClass_1, DeweyClass_2, DeweyClass_3, DeweyClass_1, DeweyClass_4...] \tag{2}$$

3. From word embedding to TDA: An Integrated Model

to model users' pathways through Gallica, we adopted three different approaches, each capturing a unique dimension of user interactions.

The first approach employed the word2vec algorithm, treating each Dewey class as a word. We created a corpus where the sentences were the transitions between classes in a user session. The concept behind this representation was to learn the relationships between classes akin to context relationships between words in natural language processing. While this approach was useful in understanding immediate relationships and similarities between themes, it didn't consider the global structure or topology of theme interactions and the chronological order of class visits.

The second method was a network approach where we constructed a graph with classes as vertices and transitions between classes as edges. Using betweenness centrality, we identified popular classes based on their position and frequency of appearance in users' journeys. This method addressed some limitations of the word2vec approach by incorporating the directionality of transitions between classes and providing a global view of class interactions. Nevertheless, it only captured one type of global structure and was limited in its ability to detect subtler topological patterns.

The third, and most novel approach, was to employ Topological Data Analysis (TDA). With TDA, we capture and quantify high-dimensional structural information about the users' journey and effectively map out a "topological fingerprint" of their interactions with classes. We use persistent homology, a tool in TDA, to track the creation and destruction of connected components, loops, and voids offering a unique multiscale perspective of the data. The TDA approach allows us to observe the existence of intricate patterns and structures that other methods might overlook. The combination of these three methods provides a comprehensive view of user behavior and class interactions in Gallica.

3.1. Word2vec representation

In this part, we aim to develop a metric that distinguishes one class from another, analogous to the organisation of a physical library.

The establishment of this metric relies on a word embedding representation, namely word2vec. We regard each session as a phrase, where every visited class represents a word.

A corpus is constructed with each sentence signifying the class transitions in a session, resembling the output produced by a Bag-Of-Words model applied to a document.

Corpus	
$Sentence_1$	$DeweyClass_1$, $DeweyClass_2$, $DeweyClass_3$, $DeweyClass_1$, $DeweyClass_4$...
$Sentence_2$	$DeweyClass_5$, $DeweyClass_2$, $DeweyClass_7$, $DeweyClass_8$, $DeweyClass_4$...
$Sentence_3$	$DeweyClass_6$, $DeweyClass_9$, $DeweyClass_7$, $DeweyClass_2$, $DeweyClass_1$...

Employing word2vec, we discern the relationships among the classes as it embeds words in a lower-dimensional vector space. The result is a collection of word vectors with similar meanings for vectors proximate in vector space, and dissimilar meanings for vectors distant in the space, based on their context. Our implementation uses the skip-gram model. This model selects word pairs by moving a window across the text data and trains a one-hidden-layer neural network. For a window of size c and a centered word r_i, we predict the context words (or in our case themes) $\{w_j\}, (i - c \leq j \leq i + c, j \neq i)$.

The cost function for one target word minimizes the negative log-likelihood of the target word vector given the associated predicted word, formulated as follows:

$$\mathscr{L}_{skipgram}(c, i) = \sum_{i-c \leq j \leq i+c, i \neq j} -\log P(w_j | r_i) \tag{3}$$

As a result, we identify classes closest to each others according to our constructed corpus. The top-N most similar keys are determined by computing the cosine similarity between a simple mean of the projection weight vectors of the given keys and the vectors for each key in the model. Positive keys contribute positively towards the similarity, negative keys contribute negatively.[3]

As an illustration, let us consider the classes that we found closest to ”Bible”:

Closest classes to ”Bible”	Similarity
History, geographic treatment, biography of Christianity	0.758
Other Literatures	0.611
Italian, Romanian and related languages	0.589.
Christian practice and observance	0.545
Earth sciences and geology	0.540
Religion	0.536
Epistemology	0.535

To conclude on this first approach, the use of word2vec embedding allowed us to construct a high-dimensionality metric space within which Dewey classes are represented closer to each other when they are more frequently consulted sequentially by users, and reciprocally. Later on, this space will allow us to draw navigational path and characterize them according to class interactions.

[3]To visualize the output of the word2vec model, we proceed with a projection of our vectors from a 200-dimensional space to a 3-dimensional one using t-SNE, a t-distributed stochastic neighbor embedding method (see Appendix C and https://kaabachi.github.io/TDA-Gallica/).

3.2. Network representation

To extract further insights from our data, we propose an alternative representation via a social network-like structure. All classes are added to the graph as vertices, with each transition from one class to another in sessions represented as edges.

From this representation, we employ graph theory to calculate the betweenness centrality of each class and thereby comprehend the ”influential” classes in the network. The betweenness centrality of a node v is given by the expression:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{4}$$

where σ_{st} is the total number of shortest paths from nodes s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v. This formula measures for a given node v its importance in terms of circulation through the network: the higher, the more the node connects other nodes and is used as a step to navigate from one class to another.

The most ”popular” classes we obtained can be seen in Table 2.

Table 2
Top 5 classes according to betweenness centrality metric

Classes	Betweenness centrality
History of Europe	0.027
French and related literatures	0.024
Latin and Italic literatures	0.0216
The arts	0.0214
News media, journalism, and publishing	0.0174

To visualize the graph, we positioned nodes using the Fruchterman-Reingold [11] force-directed algorithm. This algorithm arranges the nodes of a graph in two- or three-dimensional space such that all edges are roughly of equal length, and crossing edges are minimized (see fig. 3).

This second approach shows that the metric space of Dewey classes is not flat. Beyond the fact that some classes are very central and others peripheral due to being more or less consulted, the network approach also allowed us to start identifying which classes may act as ”pivots,” helping users to transition between dissimilar or farther classes. This inquiry was pursued parallel to the next subsection and is exposed in appendix B.

3.3. Topological Representation

Although the previous approaches outlined in section 3.1 and 3.2 provided us valuable insights on how users interface with a digital library, the approaches were limited when it comes to providing a deeper understanding of the higher-level view of these sessions and how users navigate the website.

In this work, we use Topological Data Analysis (TDA) to reveal the topological features that might have been obscured in the original vector space. We supplement the analysis done using

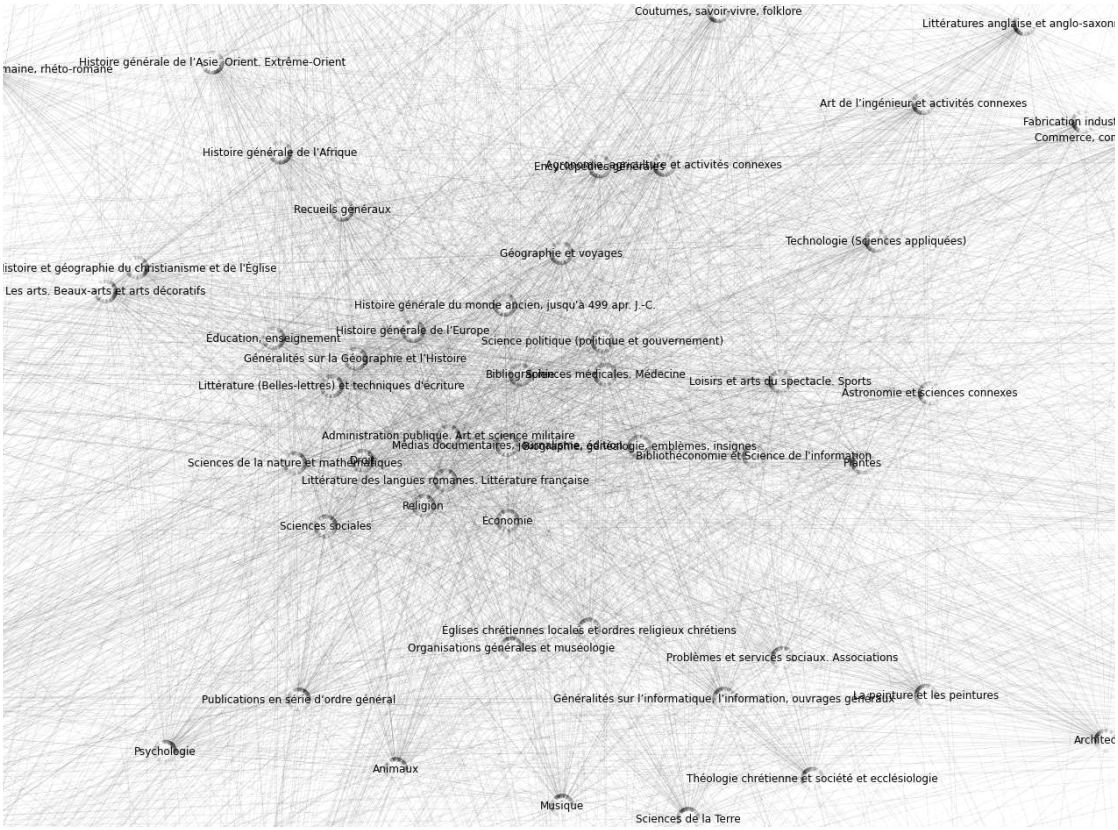


Figure 3: Zoom on the center of the network representation of Dewey classes according to Fruchterman-Reingold [11] force-directed algorithm

word2vec and combine both methods to provide a more comprehensive understanding of users' interactions with Gallica.

This novel approach borrows concepts from the realm of topological algebra to enable the extraction of deeper features that describe the topology of users' navigation paths in this digital library. The synergy and interaction between both the more traditional machine learning approach and the topological data analysis one can be seen in figure 4. The TDA applied in this study was executed using the giotto-tda library, an open-source Python toolbox designed for topological machine learning [21].

The input to this pipeline is a finite set of points called point clouds, where each point represents a Dewey class that a user has interacted with in a session, and the points are characterized by a metric that denotes the distance or similarity between them. The metric space that defines those points is the result of earlier word2vec transformations applied on the classes visited by users.

We denote a session as a point cloud, where each point is a pair of coordinates $[Class_i x, Class_i y]$, signifying the embeddings for each visited class in the session. The relation between these three components can be expressed as follows:

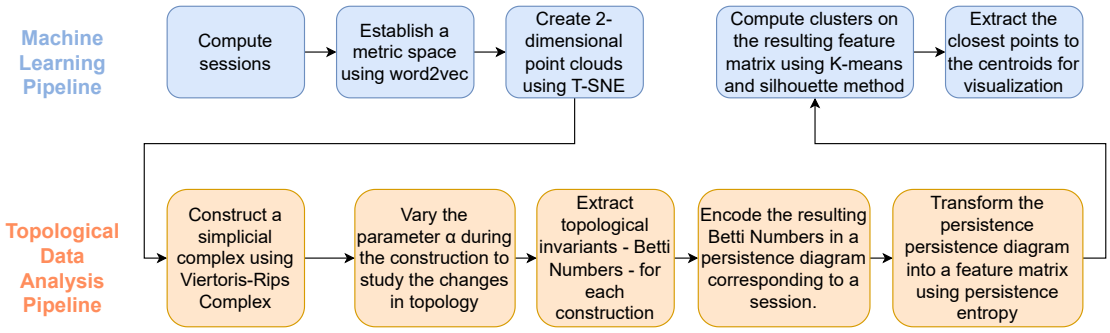


Figure 4: Schematic of the TDA pipeline inserted within a more traditional machine learning approach

$$\text{Session} \iff \text{Point Cloud} \iff (\text{Class1}_x, \text{Class1}_y)...(\text{Class}n_x, \text{Class}n_y) \tag{5}$$

where each pair $[\text{Class}_i x, \text{Class}_i y]$ corresponds to a specific class visited during a session.

The fundamental principle behind TDA [15]is to construct a continuous shape, also known as a simplicial complex atop the point cloud.[4] We construct simplicial complexes from our dataset using a specific method known as the Vietoris-Rips construction. The input set X of our metric space is derived from the result of a Word2Vec transformation on the document themes, producing a multidimensional vector space (M, d). For a non-negative real number α, we define the *Vietoris-Rips complex* $VR(\alpha)$ as the set of simplices $[x_0, \ldots, x_k]$ such that $d(x_i, x_j) \leq \alpha$ for all i, j.

Definition 1. *(Vietoris-Rips Complex). Let X be a subset of a metric space (M, d), where X is the result of a Word2Vec transformation applied to the document themes. For a non-negative real number α, the Vietoris-Rips complex $VR_\alpha(X)$ is defined as follows: The vertices are points in X, and for each subset $\{x_0, \ldots, x_k\}$ of X, we include a k-simplex $[x_0, \ldots, x_k]$ if and only if $d(x_i, x_j) \leq \alpha$ for all i, j.*

This construction has the effect of building higher-dimensional simplices (triangles, tetrahedra, and their higher-dimensional analogues) on top of the dataset whenever groups of points are closer to each other than the specified α distance. This method provides a versatile way to uncover the hidden geometric and topological structure embedded in the dataset.

Within this pipeline, we extract valuable topological information from the simplicial complex using a method called persistent homology. We vary the α parameter during the Vietoris-Rips construction to study the changes of topology in a scale-invariant perspective. We then extract the corresponding topological invariants - also called Betti Numbers. Those betti numbers are depicted in a persistence diagram that forms a ”barcode” as can be seen in figure 5 where the length of this barcode is indicative of the persistence of the homological feature.

We then transform the persistence diagrams, which depict the topological structure of the data set, into a feature matrix. The ensuing matrix, captures the entropy distribution of points

[4]Simplicial are sets of simplices that adhere to certain conditions, enabling a broad representation of geometric and topological properties of the data.

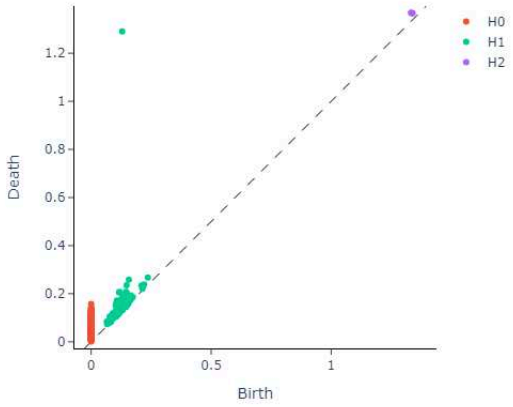


Figure 5: Example of a persistence diagram.

In our topological data analysis pipeline, the computation of Betti numbers plays a vital role as they are fundamental topological invariants. Each Betti number provides the rank of a certain homology group, hence quantifying the n-dimensional holes in our simplicial complex. The 0th Betti number represents the number of connected components, while the 1st Betti number signifies the number of one-dimensional or circular holes, and so forth. From the perspective of persistent homology, these Betti numbers evolve as we vary the parameter α in the Vietoris-Rips complex, generating one single topological signature per session, depending on the variation of homological features according to α. The birth and death of homological features for each session are encoded in a persistence diagram. The persistence diagram is thus a 'barcode' where the length of this barcode is indicative of the 'persistence' of the homological feature.

across each persistence diagram, thus summarising the underlying topology in a compact and insightful format. This is done using persistence entropy.

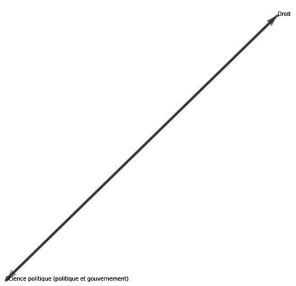
Finally, we subject resulting matrix to k-means clustering. We employ this unsupervised learning method to separate the transformed data into eight distinct clusters, each representing unique topological features discovered through TDA. The selection of eight clusters was guided by the examination of silhouette scores. These metrics provided a robust estimation of the clustering quality, with higher values (close to 1) indicating well-separated and cohesive clusters. The clustering process achieved an overall silhouette score of 0.70, indicating a reasonable separation between clusters.

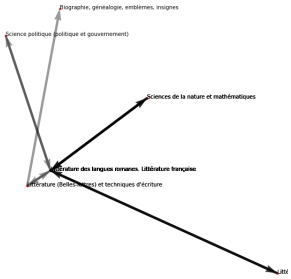
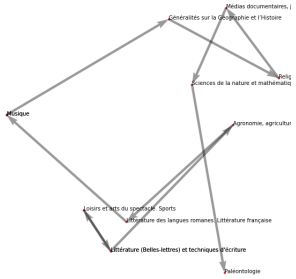
To summarize, the use of topological data analysis allowed us to cluster users' navigational paths according to their geometric features once drawn within the metric space of Dewey classes previously constructed. The core idea here was to identify typical user navigation behaviours on Gallica based on the shape of their path through the available corpus. The results are presentend in the following section.

4. Results

Clustering navigation paths according to their shape within a disciplinary space inferred from the embedding of Dewey classes allowed us to identify a number of "navigation regimes" depending on a number of variables and accounting for a variety of orienteering practices within the informational space that is Gallica's corpus. We identified four main regimes through clustering and a fifth one by a simple reflexive argument (see table below). Note that the regimes described here were interpreted by crossing users' testimonies with the shape and properties of the paths closest to the clusters' centroids. These regimes are ideal-typical in the sense that they represent archetypal user behaviour and are seldom observed in their purest form. Rather, users juggle between regimes depending on their needs, their skills, and their strategies, thereby exhibiting navigation patterns that lie farther from the centroids, albeit with dominant features pertaining to specific regimes.

- **Directed search**: very short paths spanning a limited number of Dewey classes, directed searches using very specific keywords and filters in the search engine are made by users who know precisely what they are looking for, either a specific reference, or a piece of information, and want to retrieve it through the quickest, shortest, and most efficient path–ideally a straight line from query to result;
- **Constitution or consultation of a corpus**: when a user tries to gather all available references and information on a given field that is well defined by a combination of variables including a time period, a geographical area, one or multiple topics, one or multiple types of documents..., they tend to consult more documents for a longer time to make an exhaustive survey within a very specific discipline–usually less than 2 Dewey classes;
- **Star-shaped radiation**: as was well described by our interviewees, some users start from one or two main Dewey classes and try to establish links with other perspectives on the same subject by foraying into adjacent classes–these paths tend to be even longer and to spread to more classes either in a comparative manner or in search for a diversity of views;
- **Wandering**: although some interviewees forget to mention this regime or actively discard it (and then contradict themselves), wandering through Gallica's corpus is quite a common practice that knows no disciplinary boundary and usually takes users along longer navigation paths spanning a greater number of Dewey classes, either for mere entertainment or to actively set the conditions for serendipity and the discovery of unexpected documents, information, ideas, or concepts;
- **Crawling**: finally, and although our method did not allow us to catch this kind of regime (due to the fact that we decided to catch only human behaviour), we did crawl Gallica's corpus with the use of an algorithm to retrieve, reliably and in significant quantities, the metadata of the requested documents.

		
Regime	Directed search	Constitution of a corpus
Duration	Very short	Long
Mean nb of doc.	6 to 7	7 to 10
Dewey reach	Limited (2 to 3)	Very limited (1 to 2)
Documentary unit	Information or document	Corpus, references
Values	Relevance, efficiency, findability	Exhaustivity

		
Regime	Star-shaped radiation	Wandering
Duration	Short to very long	Medium to long
Mean nb of doc.	20	More than 20
Dewey reach	Extended (10)	Very extended (more than 10)
Documentary unit	Topic or Field	Unexpected documents, ideas or concepts
Values	Comparison, diversity	Serendipity, discoverability

Regime	Crawling	Other regimes?
Duration	Long	
Mean nb of doc.	Unknown (supposedly very large)	
Dewey reach	Undefined	
Documentary unit	Metadata	
Values	Reliability, quantity	

Finally, we also identified paths of users getting lost and then finding the right path again. For example, one use looking for casts in the history of art may end up consulting a document on dental casts among the list of results. Further research, both quantitative and qualitative, will help us identify other meaningful navigation regimes.

5. Conclusion

The main conclusion of this study is a proof of concept of "navigation paths". Given that Gallica is first and foremost a service dedicated to providing digital documents to its users, mostly through the use of a single search engine (thereby fostering its consultation as a mere "database" to query), it was not obvious at first glance that readers would engage in navigating or browsing through the platform, from document to document. Yet, the wealth of spatial, mechanical, entertainment, play metaphors mobilised by our interviewees to try and objectify the way they use Gallica, coupled with the robustness and meaning of the navigation paths that we modeled based on their testimonies, we observed that a significant amount of users engage in longer sessions, consulting multiple documents, sometimes at length, in diverse classes. These informational practices are made possible by the size and diversity of Gallica's corpus that allows for comparisons, discoveries, serendipity...

Indeed, further ongoing study shows that these users somewhat "hack" the principles of the search engine by iterating queries in a manner that allows them to manage the noise present in the list of results. Thereby, using the main tool at their disposal, they construct their navigation path step-by-step according to different strategies either taking the scenic route or radiating from one or several poles. Nonetheless, all our interviewees explained that they were almost always working "in parallel", *i.e.* juggling between tabs, platforms, and physical content. If this issue was partly addressed by using a word2vec embedding that takes into account the non-linearity of navigation paths, we have to acknowledge that the paths studied are but a narrow view on much thicker and complex research practices, amounting the "digital workflow" [2] to a kind of "bricolage" [1].

From a methodological point of view, this study shows how a qualitative and a quantitative approaches can be precisely dovetailed so as to enrich each other. Indeed, the users' discourses were extremely important in defining the model used to extract navigation paths, as well as they played the role of safeguards in the manipulation of data. Reciprocally, the results and visualisations obtained through quantitative analysis allowed us to reinterpret the users' interviews and better understand their practices. Furthermore, the use of tools from topological data analysis, which is quite novel in the field of computational humanities and social sciences, proved particularly relevant to document and enact the informational space through which digital library users navigate.

Finally, these results will help digital libraries take into account the experience of their users in very meaningful ways. First, the well-known imbalance of the Dewey classification is now better documented with network analysis and an understanding of pivotal literature that can help build new taxonomies, or even folksonomies, e.g. made of tags selected by users. Second, the navigation regimes can help librarians design new navigation tools that could foster the discoverability of lesser known content, thereby facilitating serendipity. Eventually, a whole

new set of information architectures could be devised and co-designed with users depending on their needs and their research practices.

6. Acknowledgments

This exploratory study was funded by a CROSS grant awarded by EPFL and Unil.

The second phase of the research began in October 2022 thanks to a Mark Pigott Fellowship in Digital Humanities, awarded by BnF.

We would like to thank Jeanne Fernandez of the DHLab for giving us a masterclass on TDA at the onset of this project.

References

[1] S. Antonijevic and E. S. Cahoy. "Researcher as Bricoleur: Contextualizing humanists' digital workflows". In: *DH Quarterly* 12.3 (2018). URL: http://www.digitalhumanities.org/dhq/vol/12/3/000399/000399.html.

[2] S. Antonijević. "Digital Workflow in the Humanities and Social Sciences: A Data Ethnography". In: *Anthropological Data in the Digital Age*. Ed. by J. W. Crowder, M. Fortun, R. Besara, and L. Poirier. London: Palgrave Macmillan, 2020, pp. 59–83. DOI: 10.1007/978-3-030-24925-0_4.

[3] V. Beaudouin and J. Denis. *Observer et évaluer les usages de Gallica. Réflexion épistémologique et stratégique*. Research Report. Paris: Télécom ParisTech; BnF; Labex Obvil, 2014. URL: https://hal.archives-ouvertes.fr/halshs-01078530.

[4] V. Beaudouin, I. Garron, and N. Rollet. *'Je pars d'un sujet, je rebondis sur un autre' : Pratiques et usages des publics de Gallica*. Research Report. Paris: Télécom ParisTech; BnF; Labex Obvil, 2016. URL: https://hal.archives-ouvertes.fr/hal-01709238.

[5] M. Bilenko and R. W. White. "Mining the Search Trails of Surfing Crowds: Identifying Relevant Websites From User Activity". In: *Www 2008*. 2008, pp. 51–60. DOI: 10.1145/1367497.1367505.

[6] C. Borgman. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, MA: MIT Press, 2000. DOI: 10.7551/mitpress/3131.001.0001.

[7] S. Dumas Primbault. "Documenter la navigation en bibliothèque numérique. Retour sur un chassé-croisé méthodologique entre qualitatif et quantitatif". In: (forthcoming 2024).

[8] S. Dumas Primbault. "Naviguer dans les savoirs à l'ère numérique. Pour une ethnographie des pratiques informationnelles sur Gallica". In: *Études de communication* 61 (2023).

[9] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. Providence: American Mathematical Society, 2010. URL: https://www.maths.ed.ac.uk/~v1ranick/papers/edelcomp.pdf.

[10] F. Ghitalla. *Qu'est-ce que la cartographie du web ? Expéditions scientifiques dans l'univers des données numériques et des réseaux*. Marseille: OpenEdition Press, 2021. DOI: 10.4000/books.oep.15358.

[11] *Graph Drawing by Force-directed Placement - Fruchterman - 1991 - Software: Practice and Experience - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/10.1002/spe.4380211102.

[12] B. A. Huberman. *The Laws of the Web: Patterns in the Ecology of Information*. Cambridge, MA: MIT Press, 2001. DOI: 10.7551/mitpress/4150.001.0001.

[13] T. Koopmann, A. Dallmann, L. Hettinger, T. Niebler, and A. Hotho. "On the right track! Analysing and Predicting Navigation Success in Wikipedia". In: *Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT '19)*. 2019, pp. 143–152. DOI: 10.1145/3342220.3343650.

[14] D. Lamprecht, F. Geigl, and T. Karas. "Improving recommender system navigability through diversification: A case study of IMDb". In: *i-KNOW '15*. 2015. DOI: 10.1145/2809563.2809603.

[15] J. Murugan and D. Robertson. *An Introduction to Topological Data Analysis for Physicists: From LGM to FRBs*. 2019. arXiv: 1904.11044 [astro-ph, physics:hep-th].

[16] A. Nouvellet, V. Beaudouin, F. D'Alché-Buc, C. Prieur, and F. Roueff. *Analyse des traces d'usage de Gallica: Une étude à partir des logs de connexions au site Gallica*. Research Report. Paris: Télécom ParisTech; BnF; Labex Obvil, 2017. URL: https://hal.archives-ouvertes.fr/hal-01709264.

[17] A. Patania, G. Petri, and F. Vaccarino. "The shape of collaborations". In: *EPJ Data Science* 6.18 (2017). DOI: 10.1140/epjds/s13688-017-0114-8.

[18] T. Piccardi, M. Gerlach, and R. West. "Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions". In: *WikiWorkshop '22 Proc. of World Wide Web Conference (Companion)*. 2022. DOI: 10.48550/arXiv.2203.06932.

[19] R. Sen and M. H. Hansen. "Predicting. Web Users' Next Access Based on Log Data". In: *Journal of Computational and Graphical Statistics* 12.1 (2003). DOI: 10.1198/106186003127575.

[20] A. Singla, R. W. White, and J. Huang. "Studying trailfinding algorithms for enhanced web search". In: *Sigir'10*. 2010, pp. 443–450. DOI: 10.1145/1835449.1835524.

[21] G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, W. Reise, A. Medina-Mardones, A. Dassatti, and K. Hess. *Giotto-Tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration*. 2021. DOI: 10.48550/arXiv.2004.02551. arXiv: 2004.02551 [cs, math, stat].

[22] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. "The perfect search engine is not enough: a study of orienteering behavior in directed search". In: *Chi 2004*. 2004, pp. 415–422. DOI: 10.1145/985692.985745.

[23] M. Trabelsi. “Modélisation des processus utilisateurs à partir des traces d’exécution, application aux systèmes d’information faiblement structurés”. PhD thesis. La Rochelle: La Rochelle Université, 2022.

[24] R. W. White and J. Huang. “Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs”. In: *Sigir’10*. 2010, pp. 587–594. DOI: 10.1145/1835449.1835548.

[25] R. W. White and A. Singla. “Finding our way on the web: exploring the role of waypoints in search interaction”. In: *Www 2011*. 2011, pp. 147–148. DOI: 10.1145/1963192.1963267.

A. ARK (Archival Resource Key)

The implementation of ARKs on Gallica’s website is based on several principles, one of which allows for the collection of metadata from an ARK. The structure of an ARK can be depicted as follows:

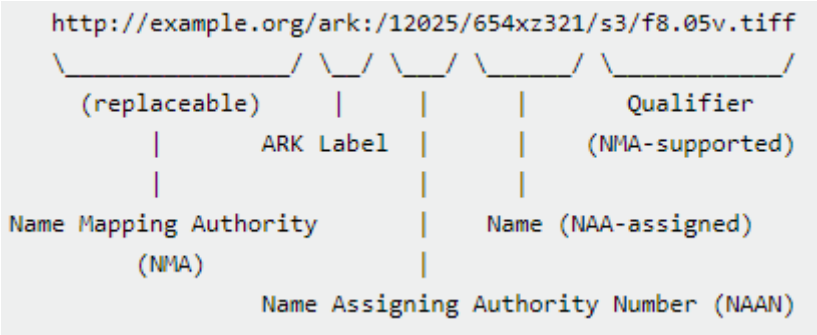


Figure 6: ARK example from https://bnf.fr/fr/lidentifiant-ark-archival-resource-key

In our specific case, the assigning authority number for Gallica’s website is consistently 12148. Consequently, the majority of the work revolves around utilizing the NAA-assigned name to query the website and obtain the desired metadata.

The subpipeline used to enrich the user sessions with additional metadata by querying Gallica’s API with the corresponding ARKs is detailed in figure 7 below.

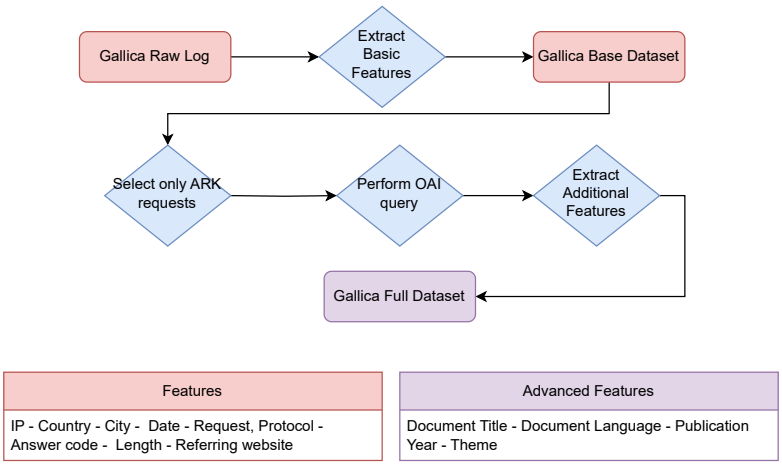

Gallica Raw Log
Extract Basic Features
Gallica Base Dataset
Select only ARK requests
Perform OAI query
Extract Additional Features
Gallica Full Dataset
Features
IP - Country - City - Date - Request, Protocol - Answer code - Length - Referring website
Advanced Features
Document Title - Document Language - Publication Year - Theme

Figure 7: ETL

B. Pivotal literature

to better understand how users actually navigate between Dewey classes, we endeavoured to model sessions as Markov chains. This allowed us to identify the type of literature mobilized by readers to jump from branches to branches within the tree of knowledge.

B.1. Sessions as Markov chains

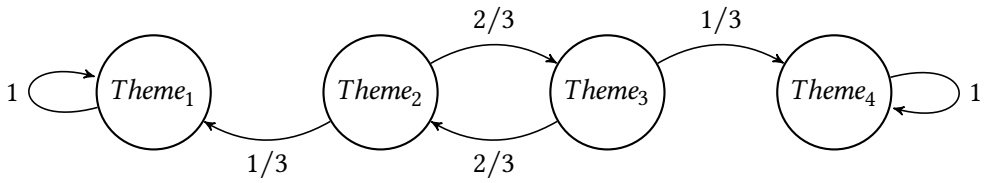
Sessions were modeled as Markov chains, a proven technique in literature to understand human navigation. The key characteristic leveraged here is the memoryless property of Markov chains, providing a robust framework for comprehending the structural mechanisms driving a user's journey in a digital library where each theme equates to a state. For a first-order Markov model, the transition probability from one theme (state) to another is defined as:

$$P_{(Theme_i, Theme_j)} = P(X_{n+1} = Theme_j | X_n = Theme_i) \tag{6}$$

The above equation signifies that the probability of transitioning to theme 'j' from theme 'i' is dependent solely on the present theme 'i' and not on any previous themes. The transition probability is estimated from the overall statistics of our dataset. If we let n_{ik} denote the number of transitions from state 'i' to state 'k', then the estimated transition probability is given by:

$$\hat{P}_{Theme_i, Theme_j} = \frac{n_{Theme_i, Theme_j}}{\sum_{k=1}^{m} n_{Theme_i, Theme_k}} \tag{7}$$

This equation indicates that the estimated probability of transitioning from theme 'i' to theme 'j' is the ratio of the count of transitions from 'i' to 'j' to the total number of transitions from 'i' to any other theme.



B.2. Results

This model sheds light on the relation of users to the practical architecture of information in Gallica. In the introduction, we were wondering whether the digitization of libraries actually allowed users to jumped from one branch, or one leaf, of the "tree of knowledge" to another, thereby facilitating interdisciplinarity and a diversity of approaches. Indeed, the main descriptors available at the French national library, hence on Gallica, to know about the topic of a document are the Dewey classes. They are arranged into an arborescent structure, that is they comprise a set of general classes making up for a partition of all possible topics, and are ramified into multiple levels of sub-classes. In physical libraries, the Dewey classification is frequently used to order the books along the shelves, thereby structuring the whole informational space.

On Gallica, most interviewees told us they would not use the Dewey classes for their searches, sometimes explicitly despising them for various reasons. Yet, a quantitative analysis of navigation paths modeled as Markov chains depicts a different situation. As can be seen on fig. 8, the probabilities to transition from discipline A to discipline B is significant only if B is equal to A, or if B is contained within the sub-classes of the general class of discipline B. This means that overall, users tend to stay within the disciplinary boundaries of a single Dewey classes and some of its sub-classes. This can be explained by a variety or factors, ranging from the users disciplinary habits, to the majority of "directed searches" (see below), to the search algorithm which must favour results from close classes.

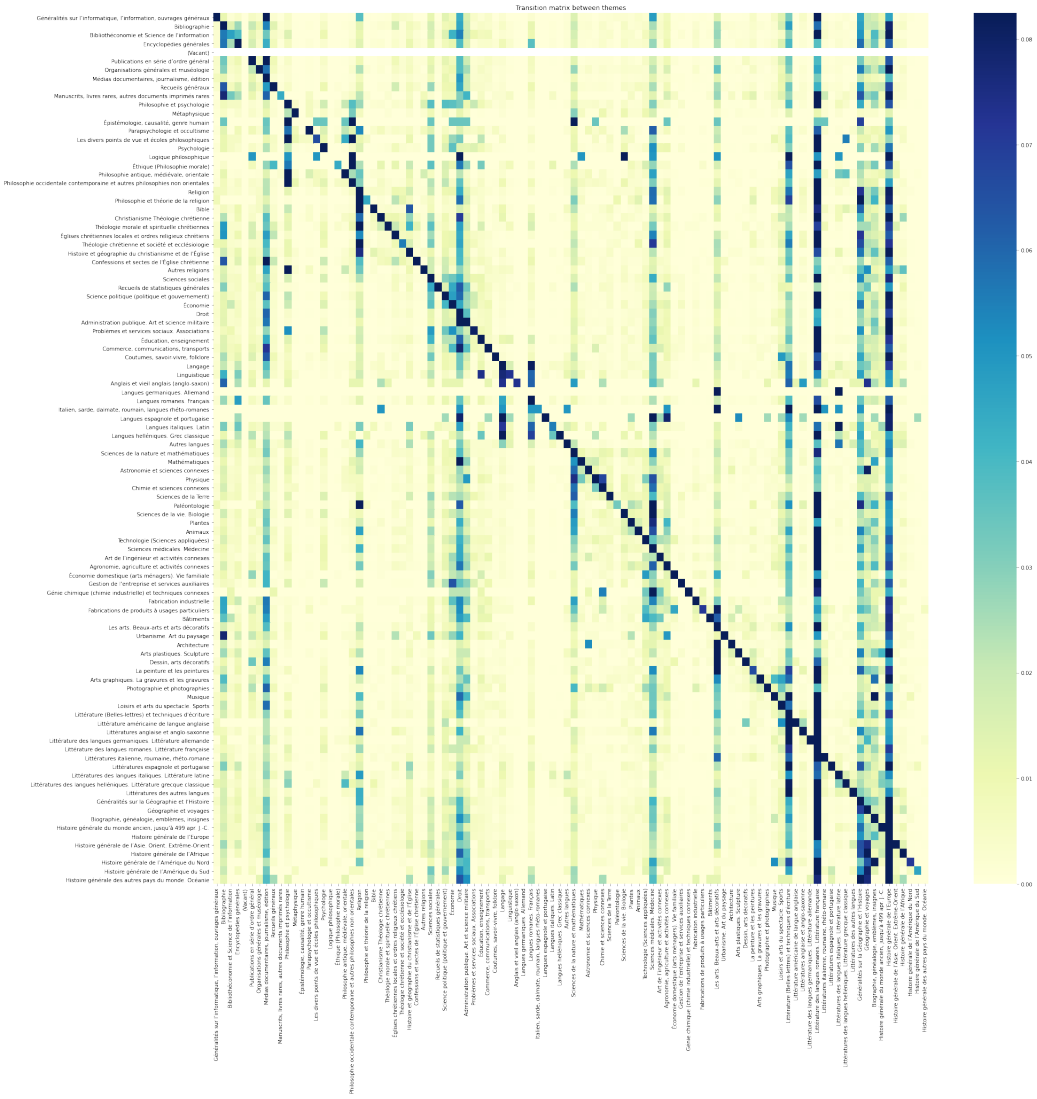


Figure 8: Transition matrix between Dewey classes (lines are normalized)

Besides the diagonal and dribbles around it, we can also identify some Dewey classes which

are more likely to be either transitioned-from or transitioned-to. Classes most transitioned-to (like 340 Law, 610 Medecine and health, or 840 French and related literatures) denote the imbalance of the Dewey classification: they are all sub-classes and yet very important and extensive fields. They are classes towards which some users converge and, then, tend to remain within since they cover ample informational space.

Most transitioned-from classes (such as 900 History and geography, 940 History of Europe, or 070 News media, journalism and publishing) cover general topics, as well as news, collections, or encyclopedias. They act as hinges or junctions between diverse other Dewey classes and allow users to navigate from one branch of the classification to another. They are what we called ”pivotal literature”.

C. Word2vec projection in 3 dimensions

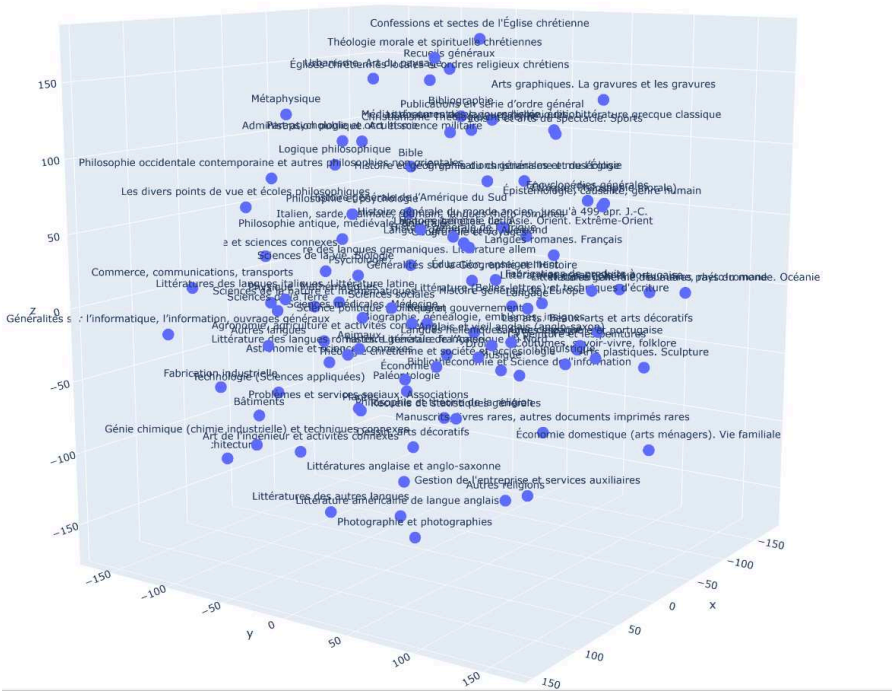


Figure 9: Visualization of the word2vec model output, we proceed with a projection of our vectors from a 200-dimensional space to a 3-dimensional one using t-SNE. This is a snapshot from the interactive viewer located in https://kaabachi.github.io/TDA-Gallica/)

D. Ongoing and future work

An ongoing research led by one of the authors furthers and widens the exploratory study presented in this paper. The same mixed-methods framework is currently deployed to document users' practices with an emphasis on action. Parallel to the present enquiry about how users imagine or picture their navigation through Gallica, said research tries to shed light on how users seize which elements of the interface and the search tools to concretely construct their navigation paths. Qualitatively, the interviews realised with about 20 users emphasise how they practically build their navigation paths step by step, by performing simple actions on the interface – iterating queries, filtering results, tweaking parameters... Quantitatively, the analysis of the server logs, rather than relying on sequence of Dewey classes, is built upon a tree of possible actions ranging from running a simple search, to turning the page of a document, to downloading it... The confrontation between both studies shows how users wanting to wander through Gallica's corpus have to "hack" its search engine by iterating slightly differing queries, thereby constructing their own scenic route step-by-step.

Future work should focus on generating other descriptors than Dewey classes to qualify with the relevant granularity the topics or disciplines of the documents consulted on the platform. This could be done by doing topic modeling on the metadata and OCR of the corpus. Furthermore, it would be interesting to have access to the navigation data of our interviewees so as to use as probes their own navigation paths. Finally, it would be ideal to be able to follow users across multiple tabs and websites to shed light on their cross-platform navigation patterns.