École Polytechnique Fédérale de Lausanne

Semester Project in Data Science

Analysis of Researchers' Online Behavior

Author:
Bayrem Kaabachi

Supervisor:
Dr. Simon Dumas Primbault

June 11, 2021

Contents

1 Introduction

Over the past decades, the development of online scientific platforms like arXiv.org or JSTOR radically changed the way researchers access, browse, or read scientific articles. Yet, while observing the behavior of researchers in libraries and laboratories has become commonplace in the humanities, the computational study of digital research practices is only in its early days. This project aims at documenting the behavior of scientists on online platforms by making sense of the digital traces they generate while navigating.

In the context of this project, work has been done on the browsing logs of Gallica which is the digital library for online users of the Bibliothèque nationale de France and its partners.

The documents present in that website form an encyclopaedic and comprehensive library, representative of major French authors and of the various trends of reflection and research over the centuries. [13]

2 Problem Statement

Understanding researchers interactions within digital libraries platform has become more and more prevalent. As technology advances, it becomes easier for researchers of all fields to access documents necessary for their hunt for knowledge.

Studying libraries and library users has already been an important subject in the field of ethnography. The goal of Ethnography is to study the culture and social organization of a particular group or community. [3] In previous research, ethnographic methods have heavily relied on data collection through observation [7], where one would simply unobtrusively observe participants enter and wander through a library. [5] With modern technology, that kind of observation can now be easily recreated in Digital Libraries through user access logs to different documents.

As digital libraries become more and more adopted, it is thus more important than ever to understand researchers patterns that appear while utilising its resources.

3 Literature review

The study of users behaviour in conventional libraries has already been done in previous research. According to a survey [7], there has been a total of 81 studies that researched library users behaviours as of 2012. To support their claims, these studies relied on multiple types of methods. The researcher could either observe or watch subjects in their natural settings [5], conduct interviews with library users [6] or conduct fieldwork in order to contextualize research results. During this project, we are trying to implement a combination of these methods while applying them to a digital resource.

When it comes to digital resources, previous research did utilize the potential of user access logs to understand how users interact with digital libraries. One notable previous research would be Romain Deveaud who worked on the OpenEdition platform [12] with the goal of understanding users needs and interests. A sizeable part of that research considered users path through Journals and their relationship with citations. In our case, not all the documents present in Gallica have a citation link nor are they scientific papers.

Another interesting previous work has been the one done by Nouvellet et al. directly on Gallica's platform. [10] From this work we were able to understand users provenance and how they accessed the website. This research also highlighted the importance of users paths and how would perform different actions on Gallica platform (for example navigating from the home page to downloading a document). Although our research works on similar data, our goal during this project is to interpret users behaviours through paths that specifically rely on document themes and how one could go from a theme to another.

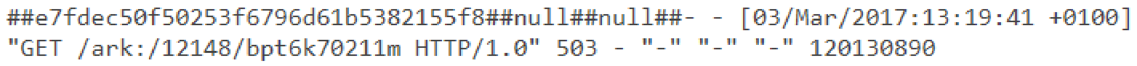
4 Data acquisition and description

Throughout this project, we worked with user browsing logs of Gallica over the month of April 2016. Each time a user accesses Gallica website, their navigation would logged and we have access to several information.

- Hashed IP Address: IP Addresses have been anonymized.
- Country: Country of requester

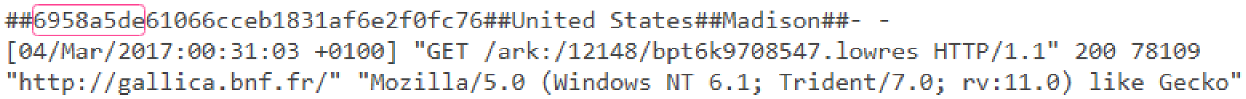
- City: City of requester
- Complete date: Date provided in day/month/year:hour:minutes:seconds format
- Request: There are different type of requests, the user may either request an HTML static page or do web-design related requests such as javascript/css... Another type of request, which would be the most important for us during this research are the **ARK requests**.
- Protocol: Communication protocol
- Answer code: response status codes.
- Length: Length of request.
- Referring website: Indicates the website from which the user comes.

It is worth noting that not every field has to be specified. Several requests do not contain information about either the country or city or the referring website. These fields would be filled with ”null”



```
##e7fdec50f50253f6796d61b5382155f8##null##null##- - [03/Mar/2017:13:19:41 +0100]
"GET /ark:/12148/bpt6k70211m HTTP/1.0" 503 - "-" "-" "-" 120130890
```

Figure 1: Example of unsuccessful HTTP get 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"
```

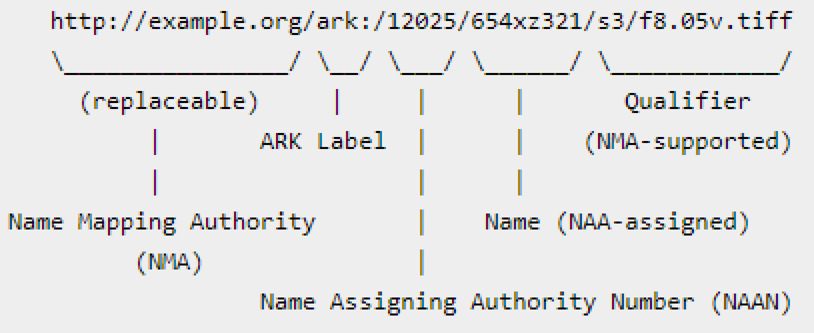
Figure 2: Example of successful HTTP get request

4.1 ARK (Archival Resource Key)

ARKs represent the way documents can be requested through Gallica’s website. Each ARK is an identifier for a document and does not contain any semantically recognizable information about the document itself.

The implementation of arks on Gallica's website is based on several principles. One of those principles makes it so that it is possible to collect metadata from an ARK.

The structure of the ARK can be represented as such: [11]



```
   http://example.org/ark:/12025/654xz321/s3/f8.05v.tiff
   \________________/ \__/ \___/ \______/ \____________/
     (replaceable)     |     |      |       Qualifier
          |       ARK Label  |      |    (NMA-supported)
          |                  |      |
Name Mapping Authority       |    Name (NAA-assigned)
         (NMA)               |
                  Name Assigning Authority Number (NAAN)
```

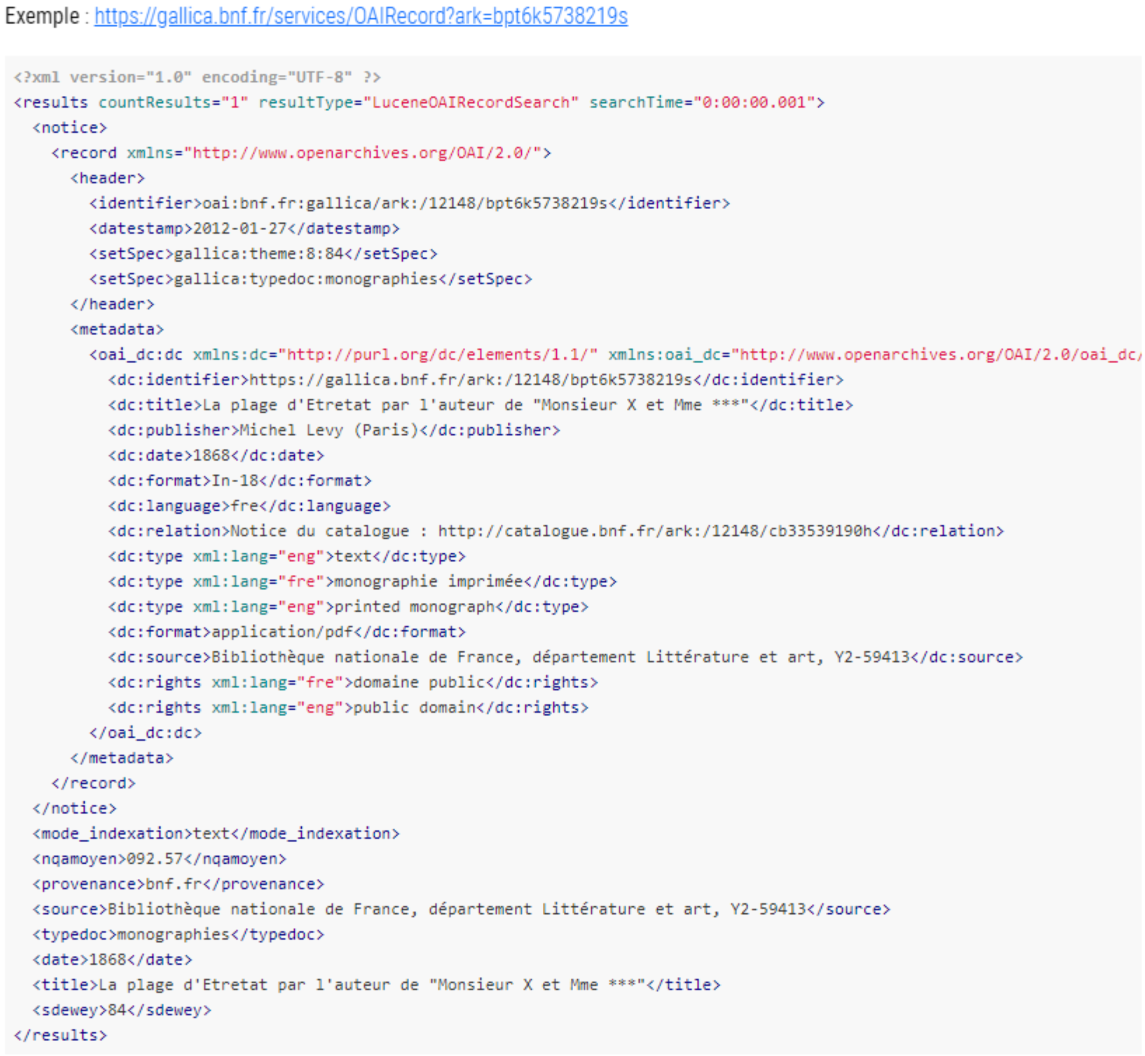
Figure 3: ARK structure

In our case, the name assigning authority number for Gallica's website is always the same 12148. Most of the work is then done through the NAA-assigned name to query the website and obtain metadata.

5 Data Enrichment

Since the project is about user behaviours and paths on Gallica's website, the raw data provided by itself is not enough as it does not contain any relevant information about the documents visited. It is thus important to enrich the preexisting data with meaningful information.

To gather this information, for each ARK type request we extract the meaningful ARK name then query Gallica's website using its Service for retrieving bibliographic information from a document.



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 4: Example of OAIRecord query and result

Using these queries we add features to each session such as the title of the document, year of publication, language and theme.

When it comes to themes, Gallica follows the Dewey Decimal Classification system which is a notation in numerals for categories with well-developed hierarchies. On a higher level, documents are classified with their first digit into 10 main classes which represent basic disciplines or field of studies. From the second digit it is possible to infer the division which is more precise. For example 610 is used for general works on medicine and health, 611 for human anatomy, 612 for human physiology, 613 for personal health and safety.

6 Tools used

6.1 Infrastructure

To deal with the large data provided in logs, we rely on EPFL IC cluster to do our calculations. The data is stored in a shared volume created specifically for this project.

To use that cluster effectively to do calculations and aggregations on data, we first create a custom docker image that we then use to initialize a Kubernetes pod running on the cluster. The shared volume containing the data is mounted during the initialization of the Kubernetes pod.

6.2 Framework

This project has been done on Python using jupyter notebooks. To enable remote work on the cluster, we connect by SSH into the pod, start a jupyter lab instance then forward its out port into a choosen address.

6.3 Libraries

For general data manipulation we relied on Pandas and Numpy. The Gallica queries and XML parsing were done using Beautifoulsoup. For word2vec representation of themes we used Gensim.

7 Analysis of researchers' behaviours

Researchers usage of Gallica website is varied and one interesting topic that was left unanswered is how researchers go from a theme to another. In physical libraries, that link from a theme to another can be done through the relative closeness of a section of the library to another. In this next section, through multiple analysis, we try to understand how that link is formed in a digital library.

7.1 Description of researchers' usage of Gallica

Since Gallica is the digital library of Bibliothèque Nationale de France, we saw from the collected data that indeed most of the connections to the website have been from France, closely followed by the United States and other French speaking neighbouring countries.

Several statistics on user origin and on the consulted documents can be found on Appendix A.

7.2 Creating Sessions that describe users' paths

Since we are interested in understanding researchers' behaviours, we first start our data manipulation by creating sessions. The goal here is to assign a session to each request, where a session is a group of requests emanating from the same user in a constrained period of time.

7.2.1 Methodology

To obtain the sessions we first decide upon a set of rules. For example, a session would a series of request that a user made in Gallica website without hitting the inactivity threshold.

We defined our activity threshold to be 60 minutes long, where we consider that if a user does another request after that period then we create another session.

In that session we decide to filter the requests to keep only ARK requests since those are the ones that contain information about the documents. Since our goal is to understand transitions between themes we also get rid of duplicate consecutive arks in a session, where that would generally mean that someone is zooming through an image or refreshing a page.

At the end we would obtain a vector that represents a user path:

$$session = [ark1, ark2, ark3, ark1, ark4...]$$

Since each ark can be translated to the document's theme, the year of publication or the document title a session could also be rewritten as:

$$session = [Theme1, Theme2, Theme3, Theme1, Theme4...]$$

After doing those transformations, we look at the length of each session and decide to work on sessions that have at least 3 arks. Since most Gallica users only consult one document [10], it would not be interesting for our research that focuses on transitions to add those. We also filter the sessions that have more than 50 arks in a row.

From that vector we query the Gallica website for each ARK in the session, while relying on a caching system where each ARK is added to a dictionary after the first query in order to reduce the number of queries needed to obtain the data.

7.2.2 Identifying outliers

When extracting our sessions we keep track of the first connection date and the last one to try to identify outliers that may affect our analysis.

7.3 User paths as Markov chains

When we consider a transition from a theme to another in a session, we translate those transitions as a Markov model where each theme represents a state.

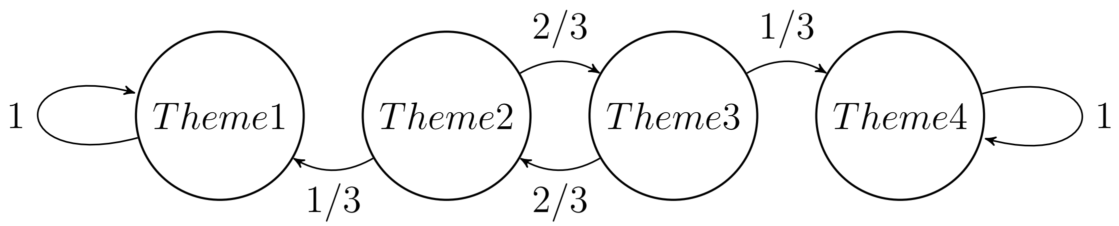
When considering a first-order Markov model we would have:

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

We estimate the transition probability from a theme to another from the overall statistics on our data where, if we let n_{ik} be the number of times that the process moved from state i to k, then:

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

A possible representation of our chain could be seen as such:



From this representation we are able to create a transition matrix from a theme to another that looks as such:

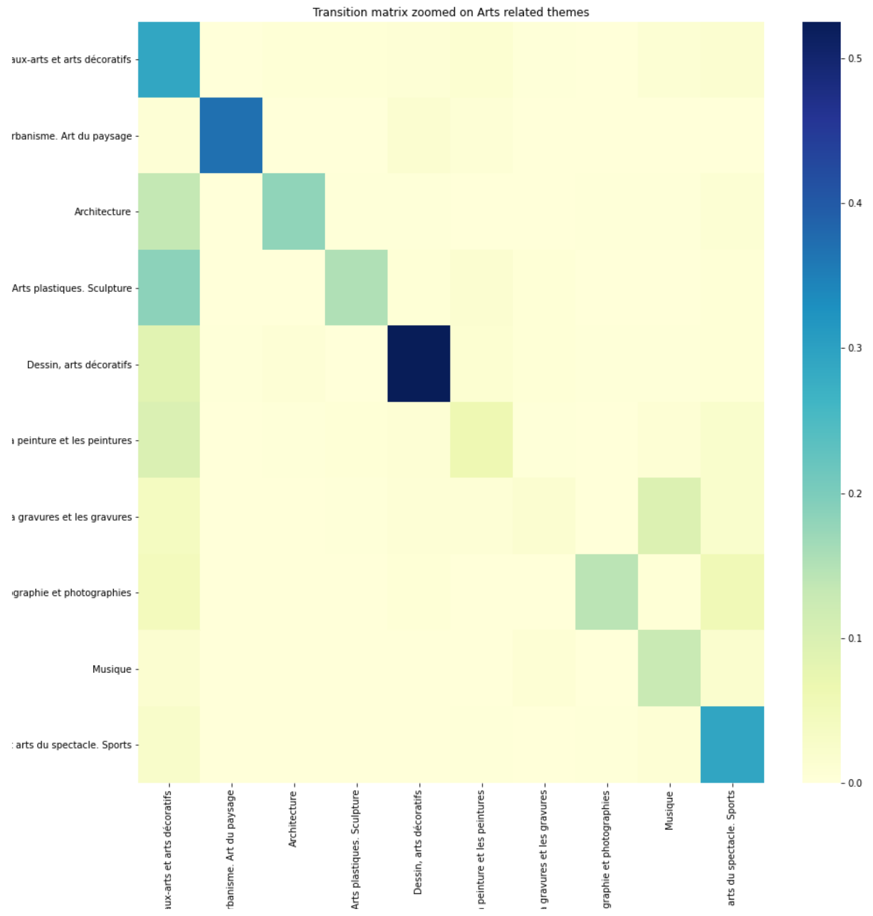


Figure 5: Transition matrix zoomed on Arts related themes

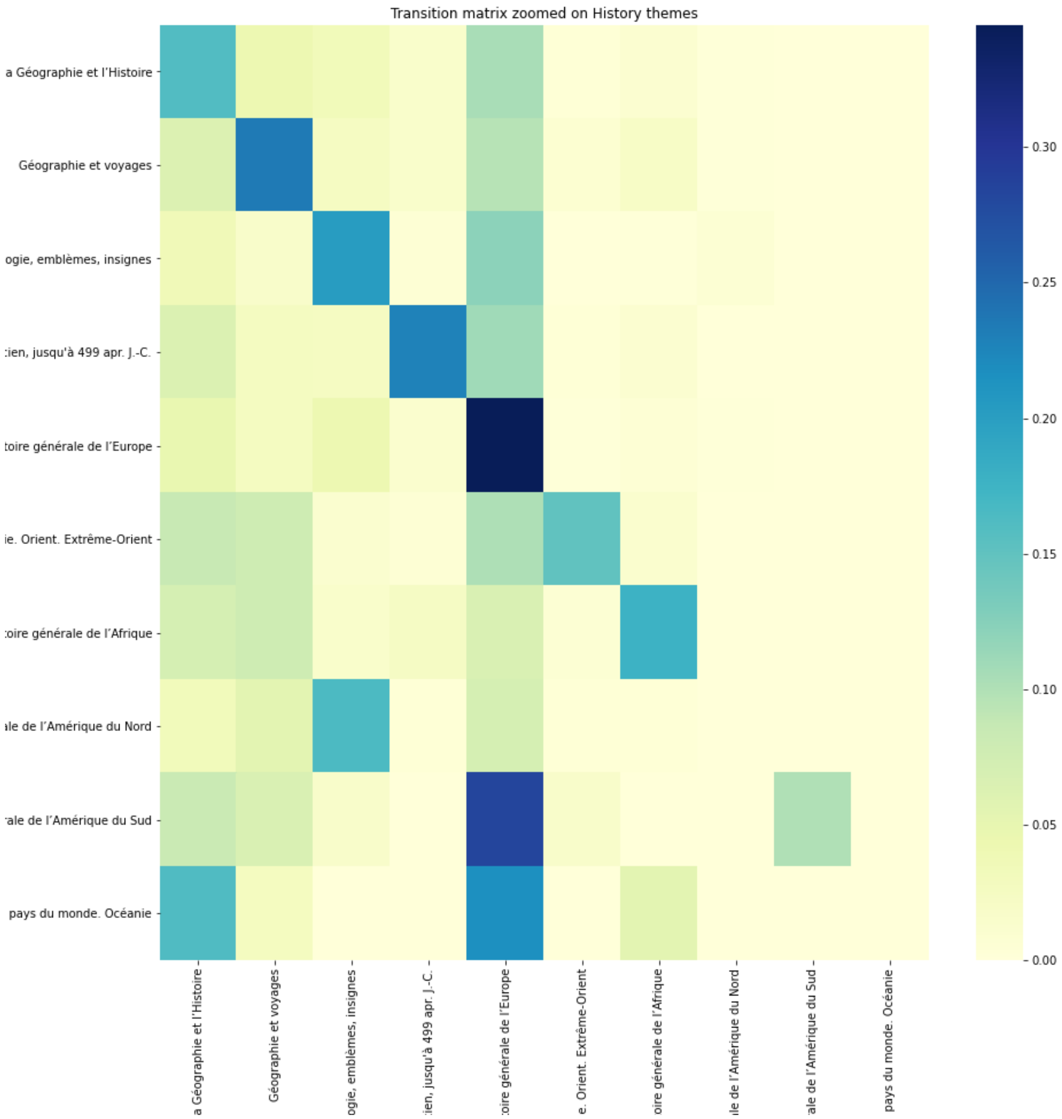


Figure 6: Transition matrix zoomed on History related themes

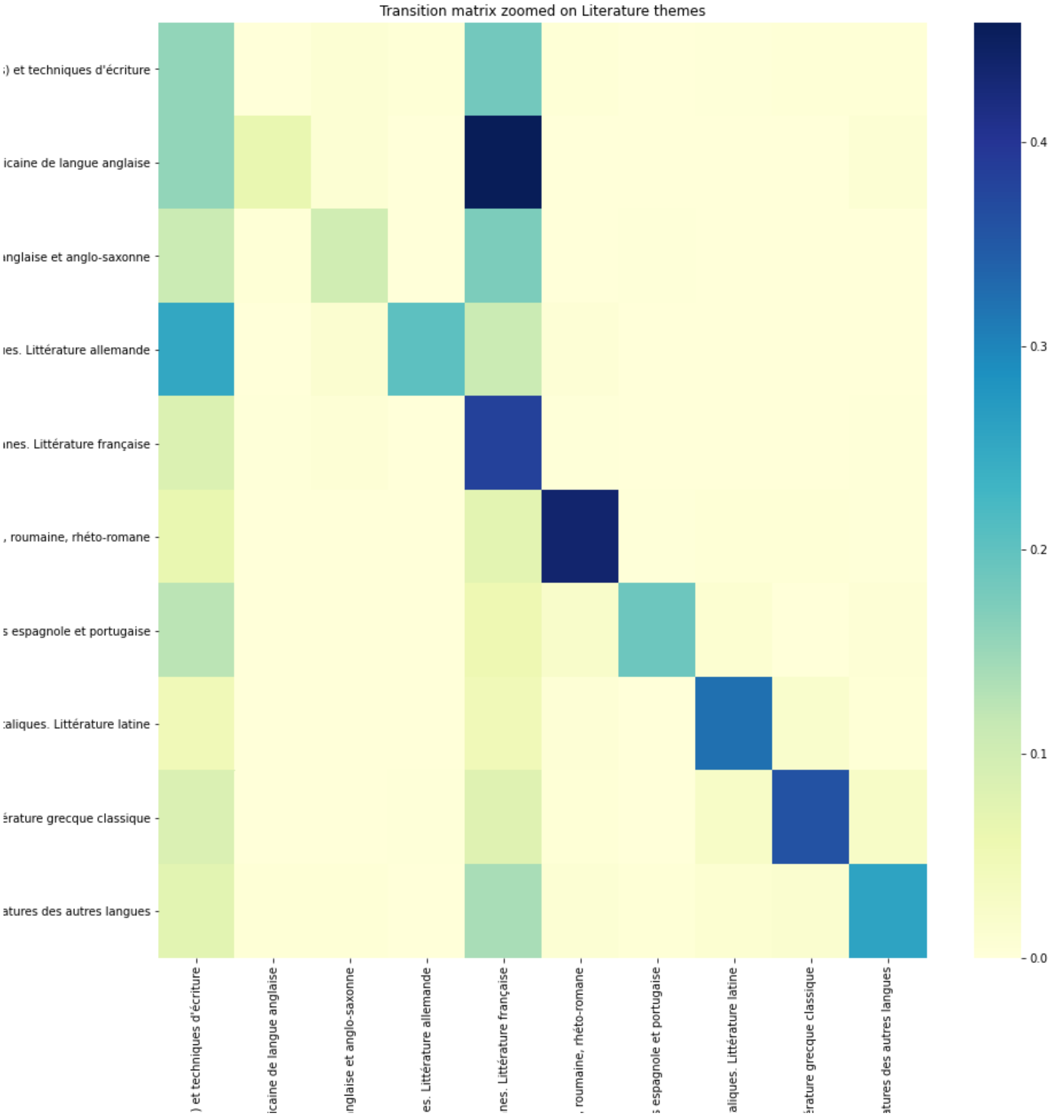


Figure 7: Transition matrix zoomed on Literature related themes

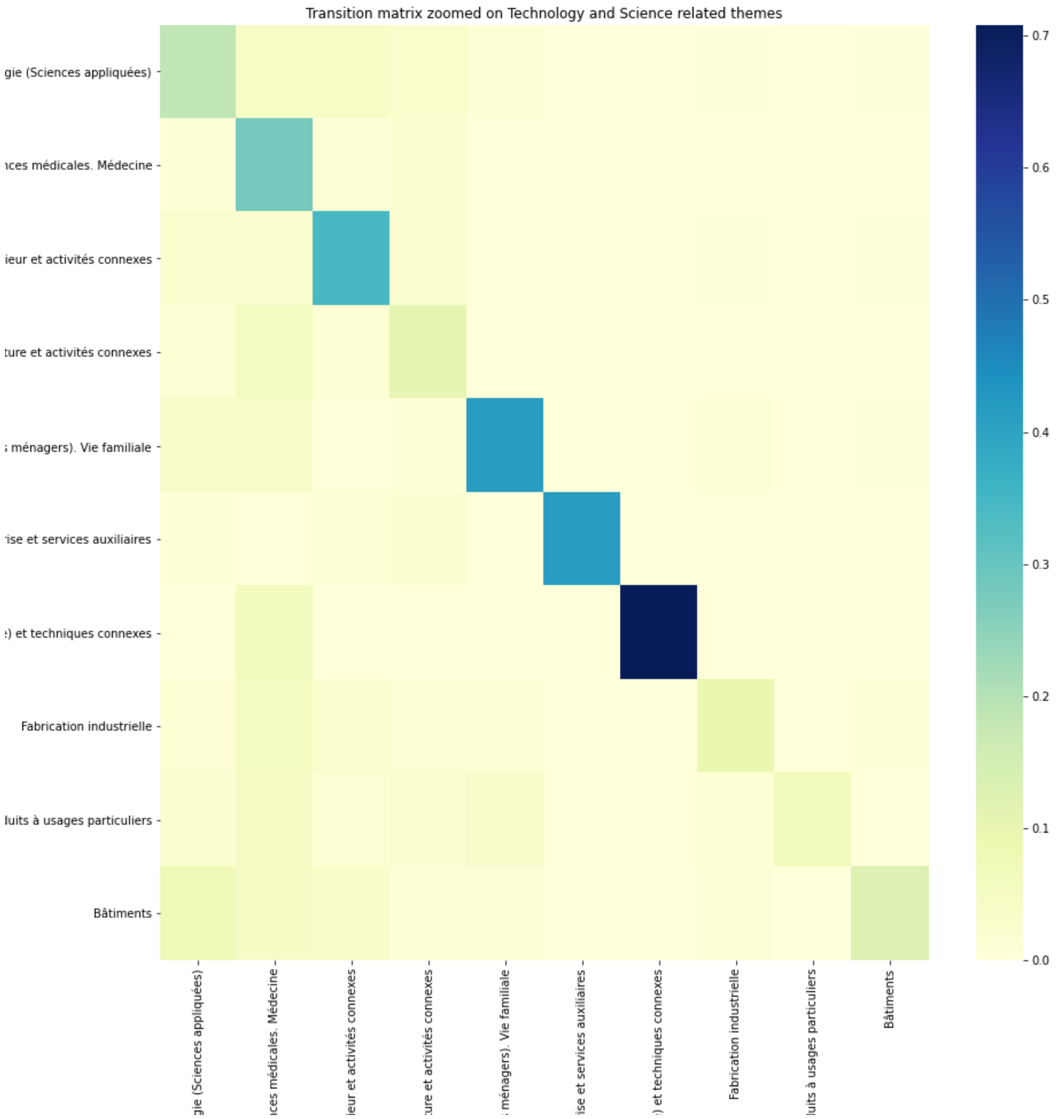


Figure 8: Transition matrix zoomed on Technology and Science related themes

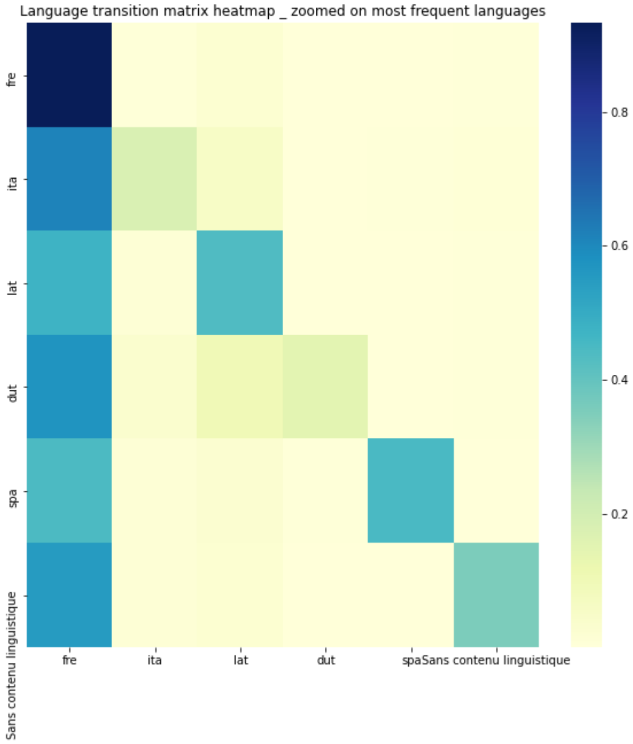


Figure 9: Transition matrix zoomed on most frequent languages

Since the themes heatmap is of dimension 100x100, one would need to zoom in to specific themes to understand the underlying structure. Nonetheless, it is possible to see a certain structure constructing itself. We see that the diagonal points in the matrix are the ones that have the highest probability. Which means that for each theme, our probability of staying inside that same theme is very high.

The complete transition matrices can be found in Appendix B.

It is also possible to see vertical lines that are darker than others, those are the themes that are more "in demand" and that we transition to the most. In the same way, looking at the horizontal lines gives us an idea of the states that we are more prone to transition from to other themes. Most of the times, those vertical lines represent the general themes.

When it comes to transitions between languages, we clearly see that most languages have a high probability of transition to French, which makes sense if we assume that the users of the platform come from France and have in common that language (see Appendix A, user origin).

Since the probability of staying in the same state is high, it is difficult to see other "interesting" transitions, we decided to use a robust heatmap that colors using quantiles to make those transitions stand out.

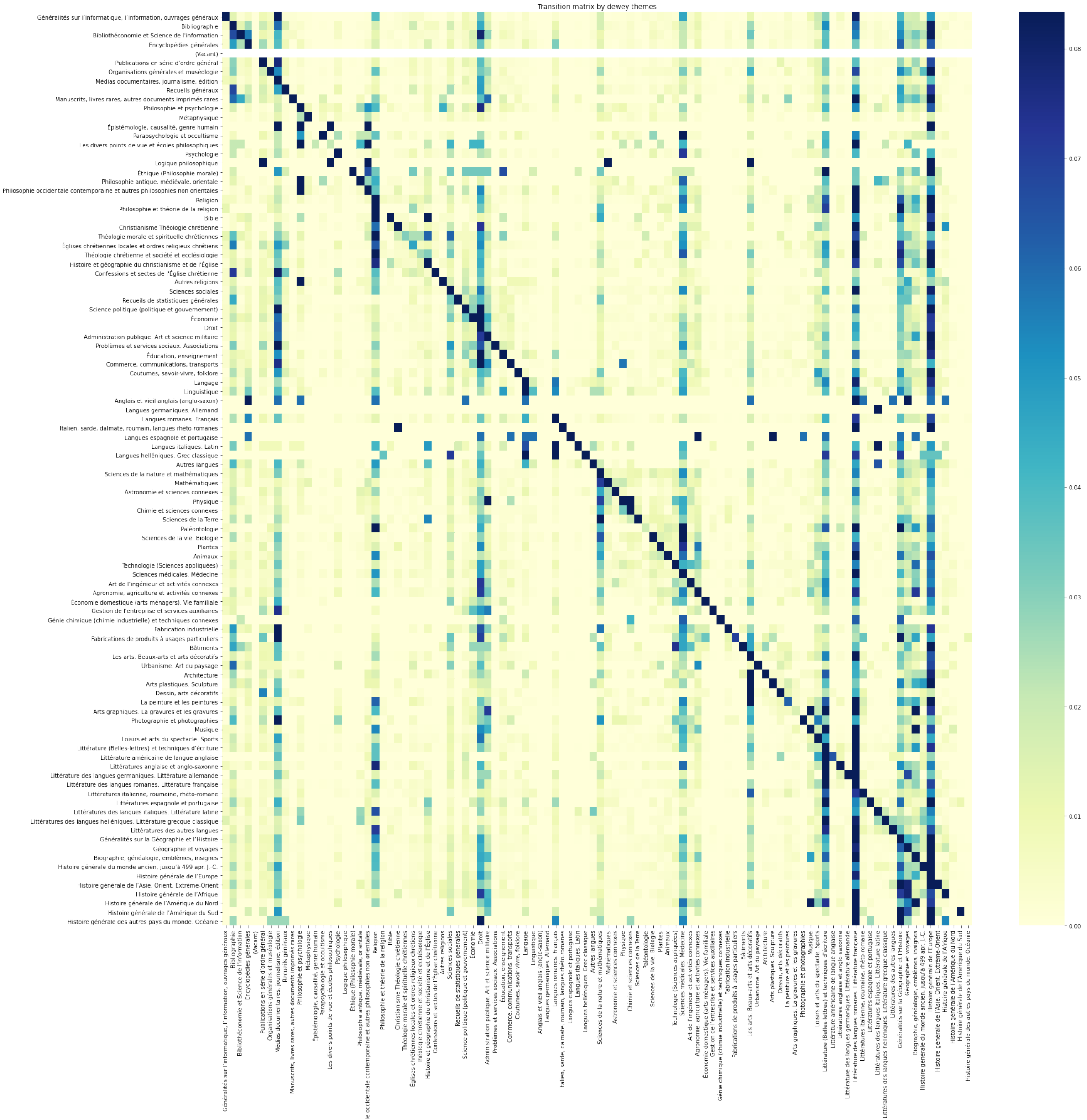


Figure 10: Heatmap of a transition matrix from a theme to another

One way to look at the themes that are more likely to lead to others would be to look at which theme users transitioned to the most from other states, i.e if we consider it as a vertex in a directed graph it would be the vertex with the most in-degree edges.

$$PopularTheme = argmax_j(\sum_{k=1}^{m} P_{Theme_k,Theme_j})$$

We found that the most transitioned-to theme was "Recueils généraux" 0.01% of probability of transition to on average.

To find the most transitioned-from theme, i.e the vertex with the most out-degree we applied the same principle:

$$PopularOutwardsTheme = argmax_j(\sum_{k=1}^{m} P_{Theme_j,Theme_k})$$

We found that the most transitioned-from theme was "Histoire générale de l'Europe" with 0.064% of probability of transition from on average.

7.4 Representation of paths in a word embedding form

One useful property that Data Science could provide to the sociology aspect of the project is how it is possible to visualise data not unlike what is usually done in ethnography. We then try to create a metric that separates a theme from another just like a physical library.

To create that metric we rely on a word embedding representation (word2vec), we consider each session as a phrase where each theme of ark visited is a word. We create a corpus where each sentence is the transitions between themes in a session, this would resemble the output of a Bag-Of-Words model applied on a document.

Corpus	
Sentence 1	Theme1, Theme2, Theme3, Theme1, Theme4...
Sentence 2	Theme5, Theme2, Theme7, Theme8, Theme4...
Sentence 3	Theme6, Theme9, Theme7, Theme2, Theme1...

Using word2vec [8], we learn the relationships between the themes. Word2Vec is a shallow neural network-based model that embeds words in a lower-dimensional vector space. The outcome is a collection of word-vectors with similar meanings based on context for vectors near together in vector space and different meanings for vectors far apart in vector space. Word2vec can either be implemented as SG - Skip-Grams or CBOW - Continuous-bag-of-words.

In our case we use the Skip-gram model. It takes a pair of words from moving a window across text data then it trains a 1-hidden-layer neural network as such:

For an index i and a window size c we predict the context words - or in our case themes - $\{w_j\}$, $(i - c \leq j \leq i + c, j \neq i)$ given the centered word r_i.

The cost function for one target word minimizes the negative log-likelihood of the target word vector given its corresponding predicted word. It is derived as follows:

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

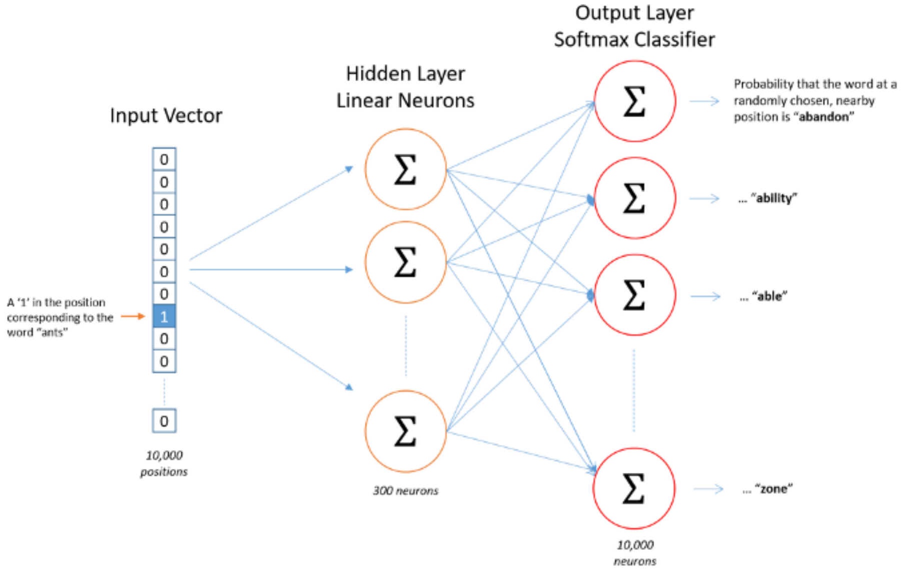


Figure 11: Architecture of skip-gram model.
Source : Chris McCormick

From word2vec, it is possible to find the themes that are most similar to other themes according to the corpus we constructed. To find the top-N most similar keys we compute 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 negatively.

Example - Similar Themes:

Similar Themes to ”Bible”	Similarity
Histoire et géographie du christianisme et de l’Église	0.758
Littératures des autres langues	0.611
Italien, sarde, dalmate, roumain, langues rhéto-romanes	0.589.
Théologie morale et spirituelle chrétiennes	0.545
Sciences de la Terre	0.540
Religion	0.536
Épistémologie, causalité, genre humain	0.535

As a way to visualize the output of the word2vec model, we proceed to do a projection of our vectors from 200D to a 2D dimension using t-SNE [4] which is the t-distributed stochastic neighbor embedding.

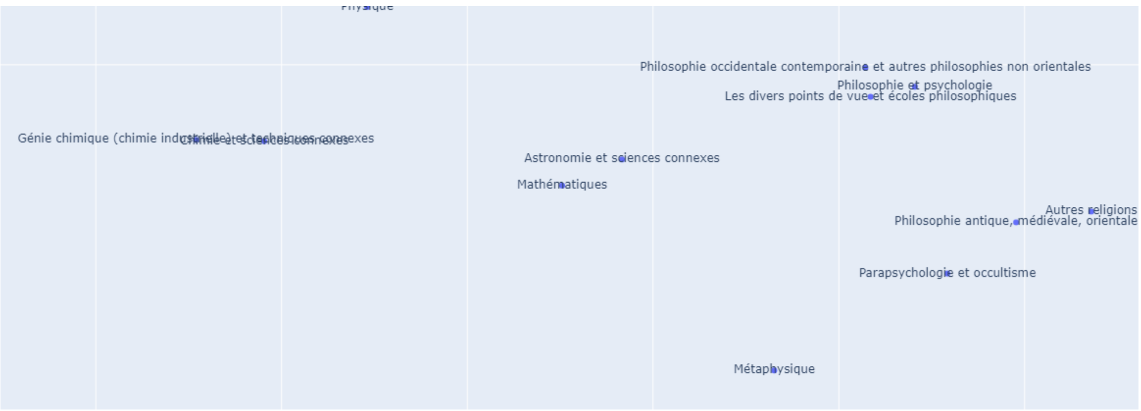


Figure 12: word2vec - zoom on themes similar to Mathématiques

7.5 Representing themes as a network

To further leverage interesting insights from our data, we propose another representation through a social network like structure. We add all themes in the graph as vertices where each transition from a theme to another in sessions is represented as edges.

From this representation, we make use of graph theory to calculate the betweenness centrality of each theme and then gain an understanding of "influential" themes in the network.

The betweenness centrality of a node v is given by the expression:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

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.

Since our graph is based on the transition matrix, we weight each edge by the probability of transition.

A node strength is then given by the sum of the weights of its adjacent edges. With a_{ij} and w_{ij} being adjacency and weight matrices between nodes i and j

$$s_i = \sum_{j=1}^{N} a_{ij} w_{ij}$$

Themes	Betweenness centrality
Histoire générale de l'Europe	0.027
Littérature des langues romanes. Littérature française	0.024
Littératures des langues italiques. Littérature latine	0.0216
Les arts. Beaux-arts et arts décoratifs	0.0214
Médias documentaires, journalisme, édition	0.0174

To visualize the graph, we positioned nodes using Fruchterman-Reingold [1] force-directed algorithm which positions the nodes of a graph in a two-dimensional or three-dimensional space so that all the edges are of more or less equal length and there are as few crossing edges as possible.

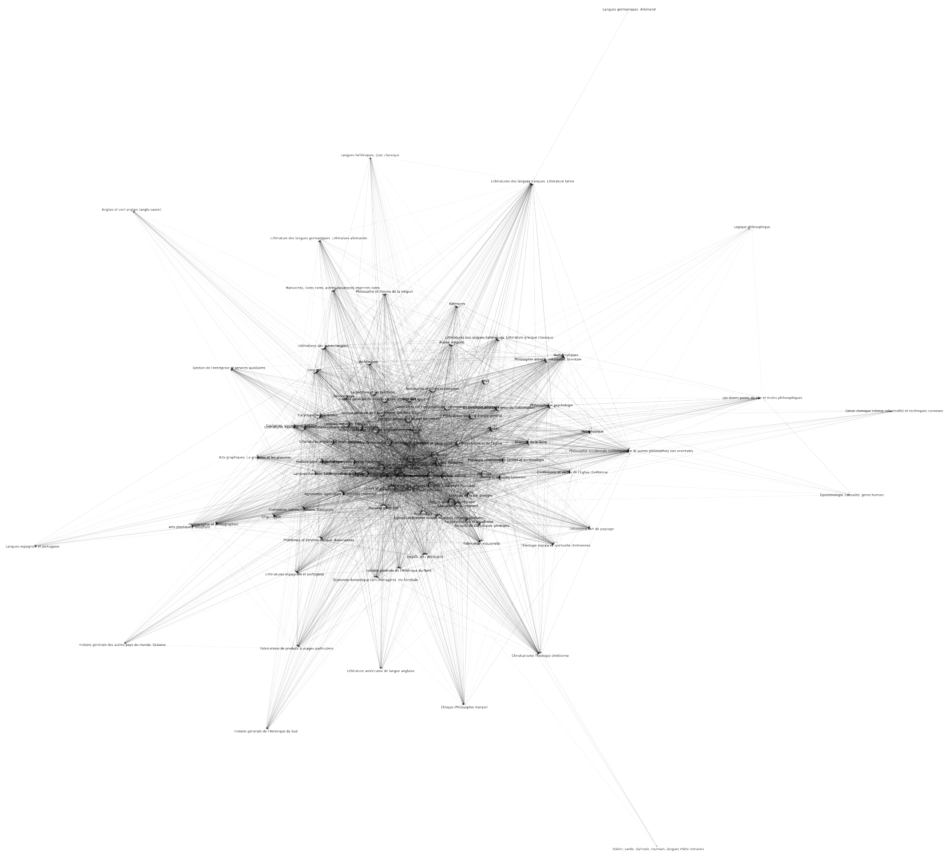

Langues germaniques. Allemand
Langues helléniques. Grec classique
Littératures des langues italiques. Littérature latine
Anglais et vieil anglais (anglo-saxon)
Logique philosophique
Littérature des langues germaniques. Littérature allemande
Manuscrits, livres rares, autres documents imprimés rares
Philosophie et théorie de la religion
Bâtiments
Littératures des langues helléniques. Littérature grecque classique
Autres religions
Littératures des autres langues
Mathématiques
Philosophie antique, médiévale, orientale
Architecture
Gestion de l'entreprise et services auxiliaires
Bible
Astronomie et sciences connexes
Les divers points de vue et écoles philosophiques
Génie chimique (chimie industrielle) et techniques connexes
Philosophie et psychologie
Métaphysique
Sciences de la Terre
Philosophie occidentale contemporaine et autres philosophies non orientales
Arts graphiques. La gravure et les gravures
Théologie chrétienne et société et ecclésiologie
Confessions et sectes de l'Église chrétienne
Épistémologie, causalité, genre humain
Langues espagnole et portugaise
Photographie et photographies
Arts plastiques. Sculpture
Urbanisme art du paysage
Problèmes et services sociaux. Associations
Fabrication industrielle
Théologie morale et spirituelle chrétiennes
Dessin, arts décoratifs
Littératures espagnole et portugaise
Histoire générale de l'Amérique du Nord
Économie domestique (arts ménagers). Vie familiale
Histoire générale des autres pays du monde. Océanie
Fabrications de produits à usages particuliers
Christianisme Théologie chrétienne
Littérature américaine de langue anglaise
Ethique (Philosophie morale)
Histoire générale de l'Amérique du Sud
Italien, sarde, dalmate, roumain, langues rhéto-romanes

Figure 13: Network of themes

8 Triangulation and qualitative research - An interview with a Gallica user

8.1 Introduction

During our research, I had the opportunity to assist during an interview of a Gallica user. The interview helped us connect with the grounded theory method, according to which we revisited the subject with deeper knowledge.

8.2 Participant observation

Participant observation, like other ethnographic approaches, is based on the classic methods employed by Malinowski and others in early anthropology to study specific people for years at a time while taking careful notes. [14]

Inductive participant observation is done as part of an exploratory research phase with the goal of creating hypotheses from the data.

The strength of participant observation is its ability to describe depth (thick description) and to help understand human behaviour.

While observing the user scroll through the website, I quickly came to the conclusion that he does not rely on Gallica's Theme classification but mostly looks at documents from the research bar. The "links" between documents and the "transitions" came to be thanks to the logical links that can be found inside the document in itself.

8.3 Participant interview

When presented with the opportunity, I asked several questions during the interview which had the goal of further understanding the transitions. For example, the user has in general 3 tabs open on Gallica's website, and while browsing a document his mind may still linger on an older document that he visited and he creates the link that way.

This may be a good motivation to try other Markov models like second or third order that may try to emulate the user "memory" and open tabs.

9 Conclusion

Our observation shows that the type of consultations of documents are diverse but it is still possible to see the trend where researchers either have an idea of what they are looking for and stay within the same theme or they don't and they transition from a theme to another mostly by coming back to the more general themes in between.

This discussion raises the question of how a digital platform could assist complex research tasks by creating efficient and usable interfaces and structures for its documents. One such way would be recommendations based on the document consulted which would emulate the behaviour of a librarian.

10 Future Research

10.1 Distances between sessions using Word Mover's distance

One possible way to create a distance between different sessions would be using Word Mover's Distance that leverages the previously defined word embedding. [9] The WMD distance measures the dissimilarity between two text documents as the minimum amount of distance that the embedded words of one document need to ”travel” to reach the embedded words of another document. From this distance it would be possible to perform clustering on those documents.

10.2 Clustering on sessions and identifying trends

Future work may include performing clustering on different user sessions to identify sessions that are similar to each other. Spectral clustering could be done on the previously created graph to identify communities of nodes based on the edges connecting them. [2]

Acknowledgement

I would like to express my deep and sincere gratitude to my research supervisors, Dr. Simon DUMAS PRIMBAULT. I am really grateful for his patience, his availability and wisdom throughout this research process. He never hesitated to find a time slot among their ongoing projects to answer my questions. My understanding of the vast field of sociology was non-existent at the start of this project but through several discussions, I became fascinated by the field and its intricacies.

References

[1] Thomas M. J. Fruchterman and Edward M. Reingold. “Graph drawing by force-directed placement”. In: *Software: Practice and Experience* 21.11 (1991), pp. 1129–1164. DOI: https://doi.org/10.1002/spe.4380211102. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/spe.4380211102. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.4380211102.

[2] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. “On Spectral Clustering: Analysis and an Algorithm”. In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. NIPS’01. Vancouver, British Columbia, Canada: MIT Press, 2001, pp. 849–856.

[3] Craig Calhoun. *Dictionary of the social sciences*. Oxford University Press, 2002.

[4] Laurens van der Maaten and Geoffrey Hinton. “Viualizing data using t-SNE”. In: *Journal of Machine Learning Research* 9 (Nov. 2008), pp. 2579–2605.

[5] Lauren H. Mandel. “Toward an understanding of library patron wayfinding: Observing patrons’ entry routes in a public library”. In: *Library Information Science Research* 32.2 (2010), pp. 116–130. ISSN: 0740-8188. DOI: https://doi.org/10.1016/j.lisr.2009.12.004. URL: https://www.sciencedirect.com/science/article/pii/S0740818810000034.

[6] Kylie Bailin. “Changes in Academic Library Space: A Case Study at The University of New South Wales”. In: *Australian Academic & Research Libraries* 42.4 (2011), pp. 342–359. DOI: 10.1080/00048623.2011.10722245. eprint: https://doi.org/10.1080/00048623.2011.10722245. URL: https://doi.org/10.1080/00048623.2011.10722245.

[7] Michael Khoo, Lily Rozaklis, and Catherine Hall. “A survey of the use of ethnographic methods in the study of libraries and library users”. In: *Library Information Science Research* 34 (Apr. 2012), pp. 82–91. DOI: 10.1016/j.lisr.2011.07.010.

[8] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. *Exploiting Similarities among Languages for Machine Translation*. 2013. arXiv: 1309.4168 [cs.CL].

[9] Matt J. Kusner et al. “From Word Embeddings to Document Distances”. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML’15. Lille, France: JMLR.org, 2015, pp. 957–966.

[10] Adrien Nouvellet et al. “Analyse des traces d’usage de Gallica : Une étude à partir des logs de connexions au site Gallica.” In: (2017).

[11] *Archival Resource Key (ARK) Identifiers*. URL: `https://n2t.net/e/ark_ids.html`.

[12] Romain Deveaud. “Modalités d’accès au savoir ouvert sur les plateformes d’OpenEdition”. In: ().

[13] *Gallica – The BnF digital library*. URL: `http://www.bnf.fr/en/gallica-bnf-digital-library`.

[14] Stuart Hannabuss. *How to... Use ethnographic methods and participant observation*. URL: `https://www.emeraldgrouppublishing.com/how-to/observation/use-ethnographic-methods-participant-observation`.

11 Appendix A

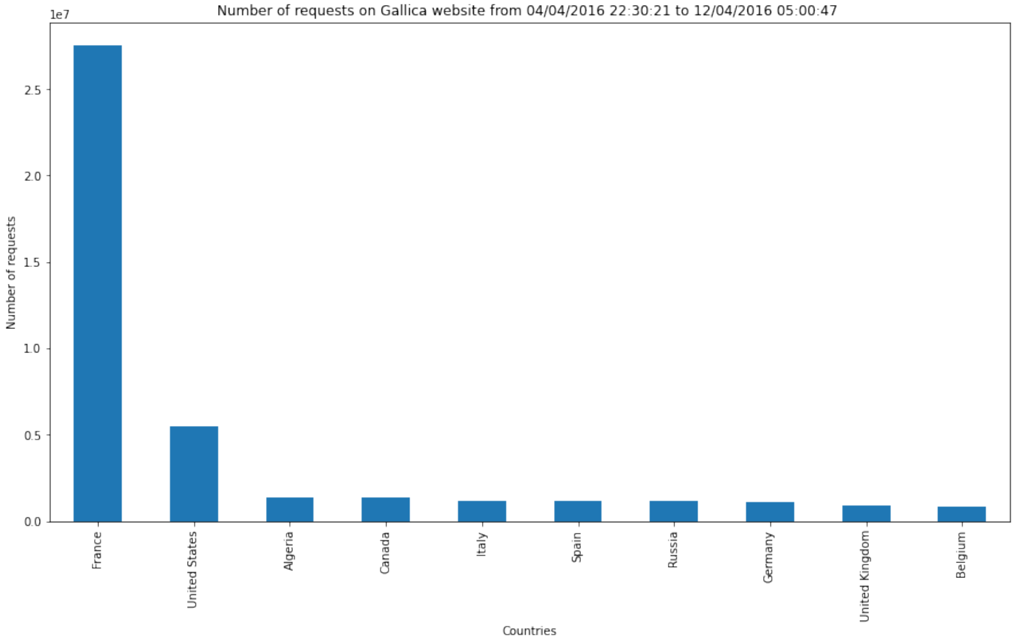


Figure 14: User origin: countries

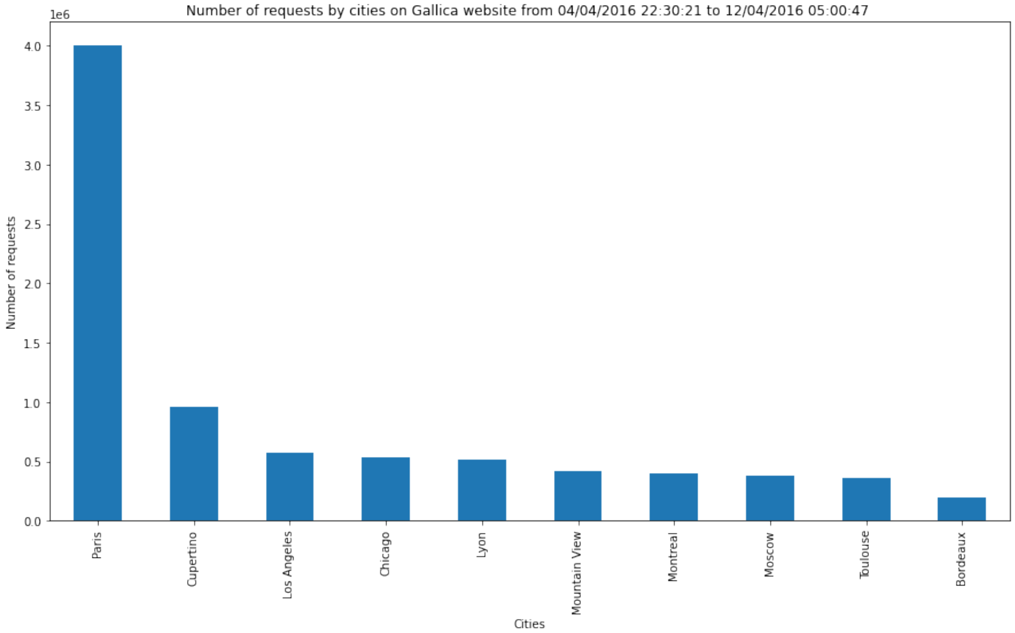


Figure 15: User origin: Cities

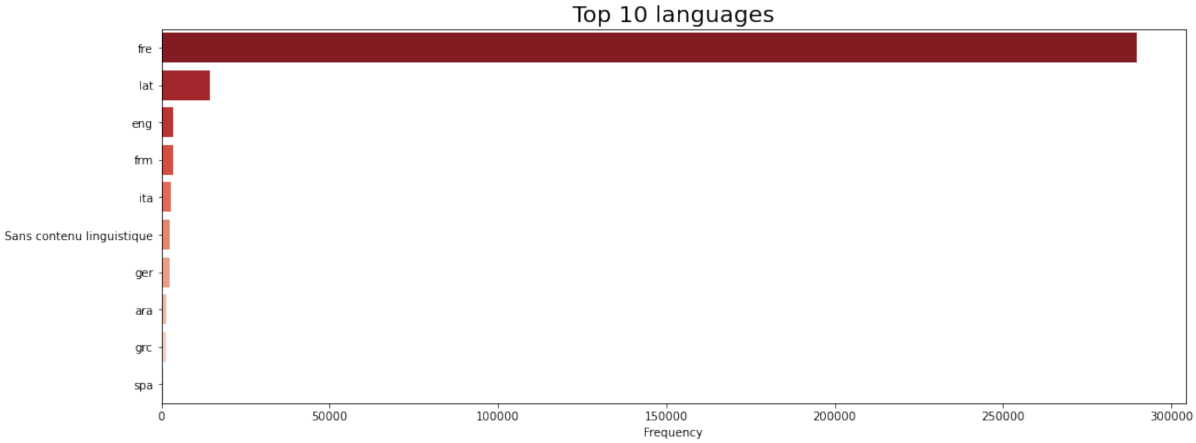


Figure 16: Document type: language

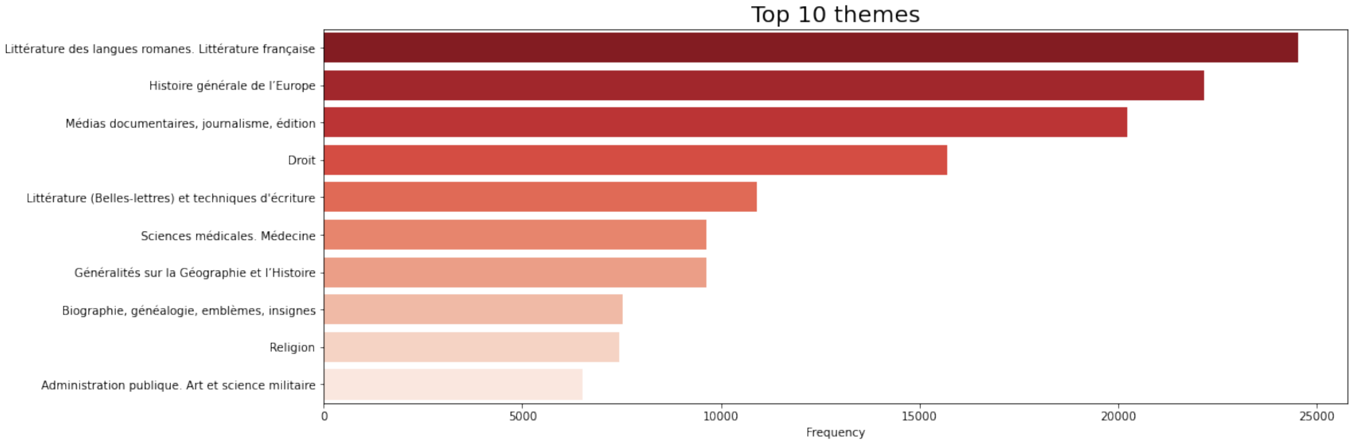


Figure 17: Document type: themes

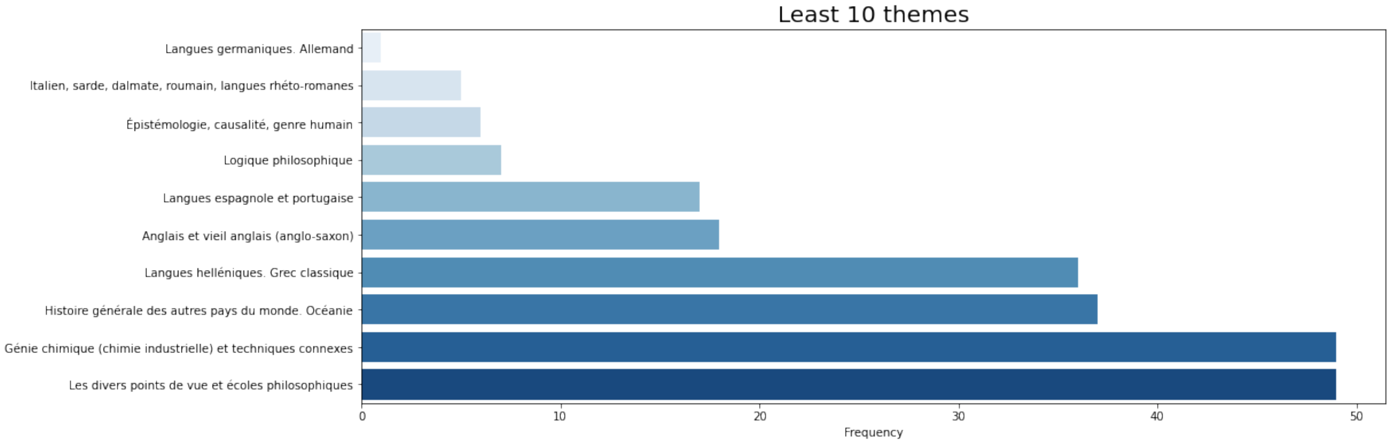


Figure 18: Document type: themes

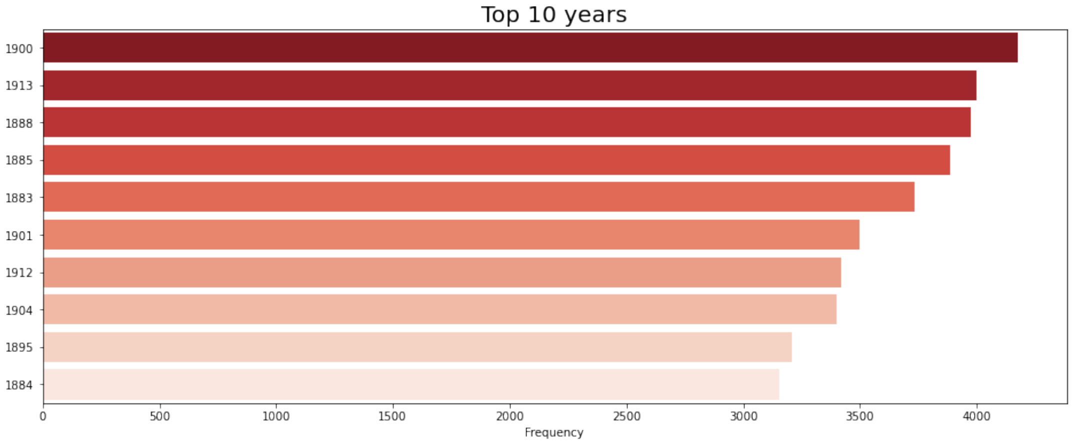


Figure 19: Document type: year of publication

12 Appendix B

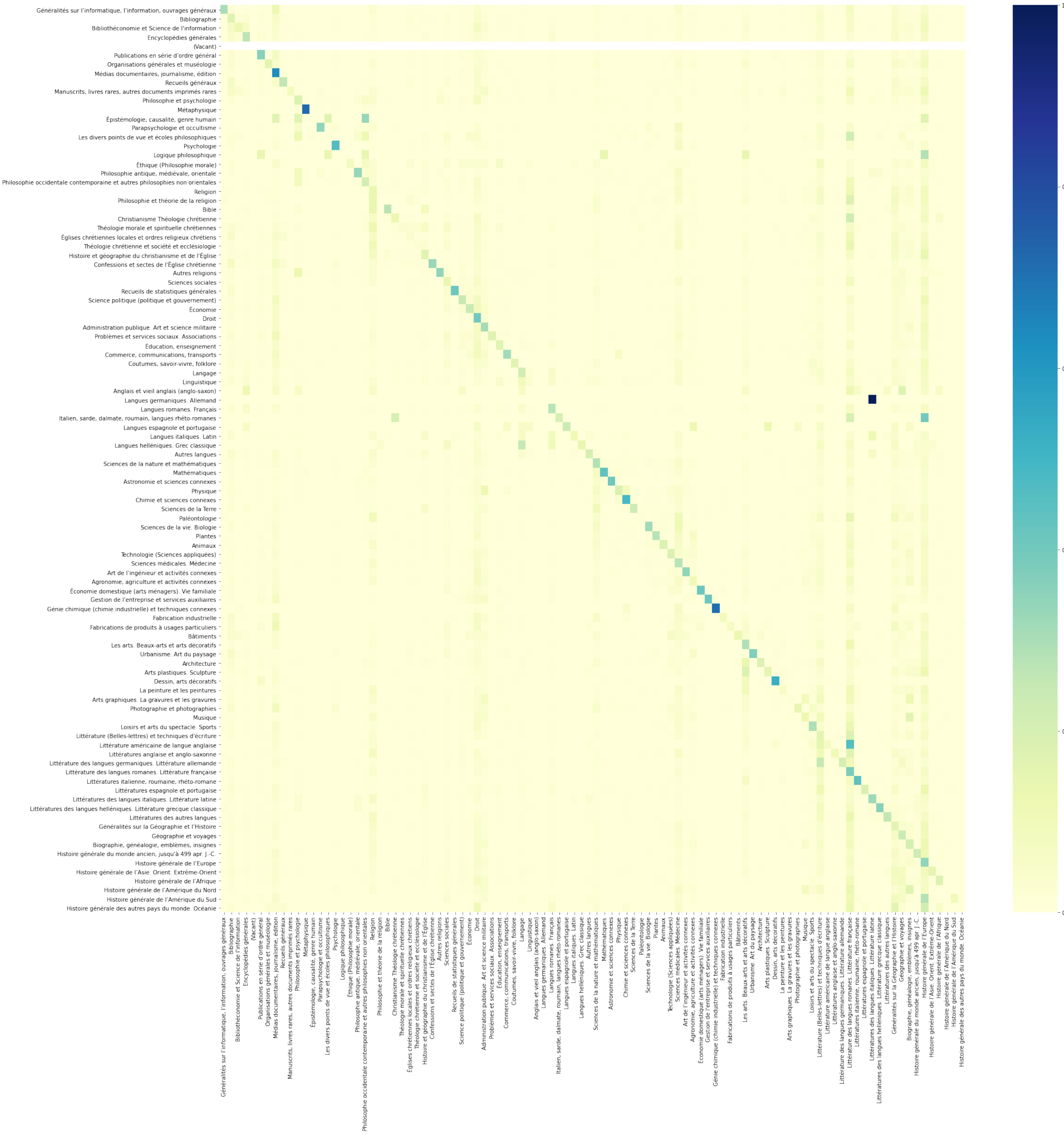


Figure 20: Heatmap of a transition matrix from a theme to another

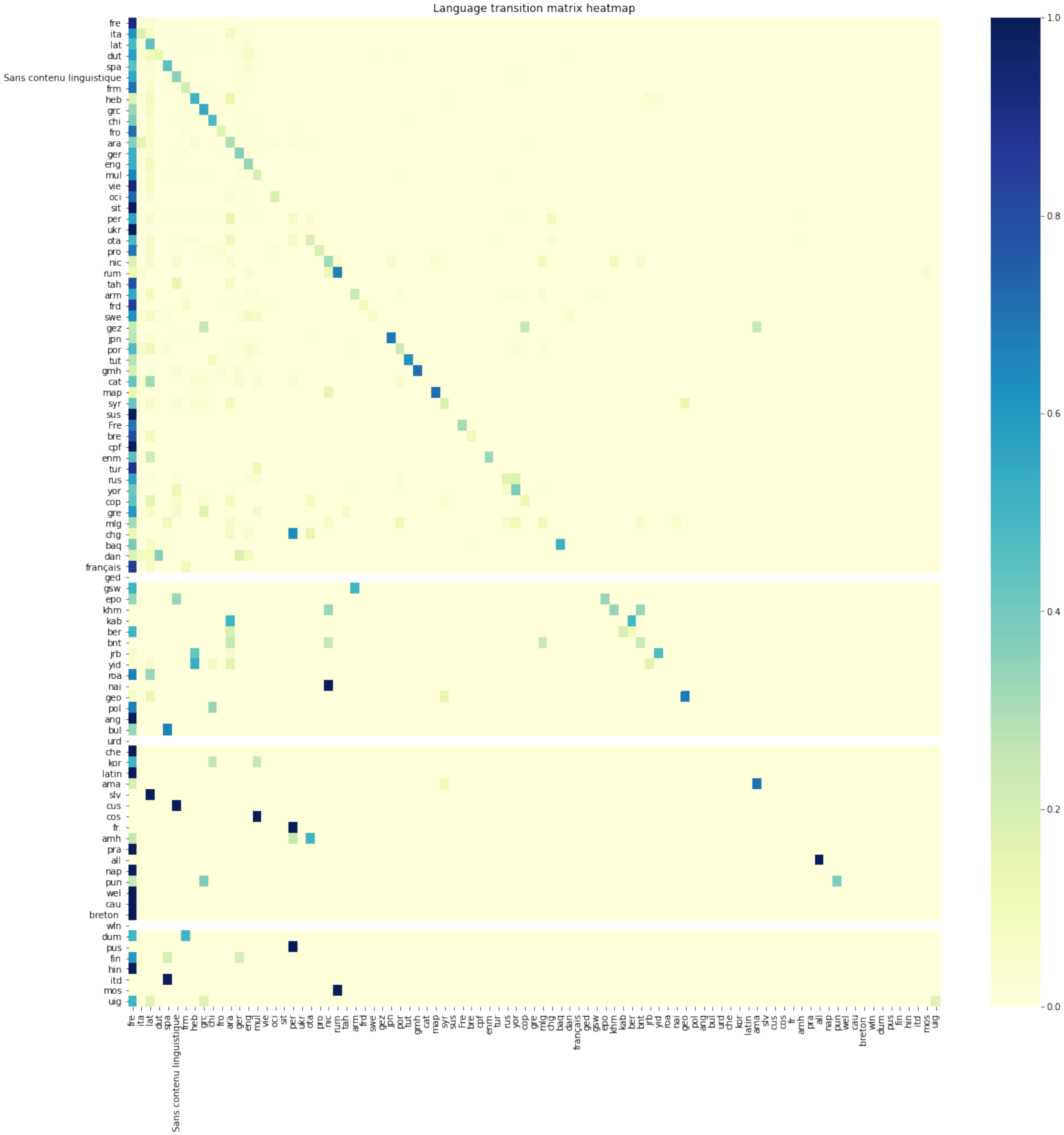


Figure 21: Heatmap of a transition matrix from a language to another