# Faire l’analyse critique d’un *pipeline* d’analyse de données

## Un cas d’étude en humanités numériques

Simon Dumas Primbault

OpenEdition, CNRS UAR 2504

Associate Research LHST (EPFL)

Associate Researcher BnF (French National Library)

GdR ModMat, Banyuls-sur-Mer

22.08.24

## Outline of the presentation

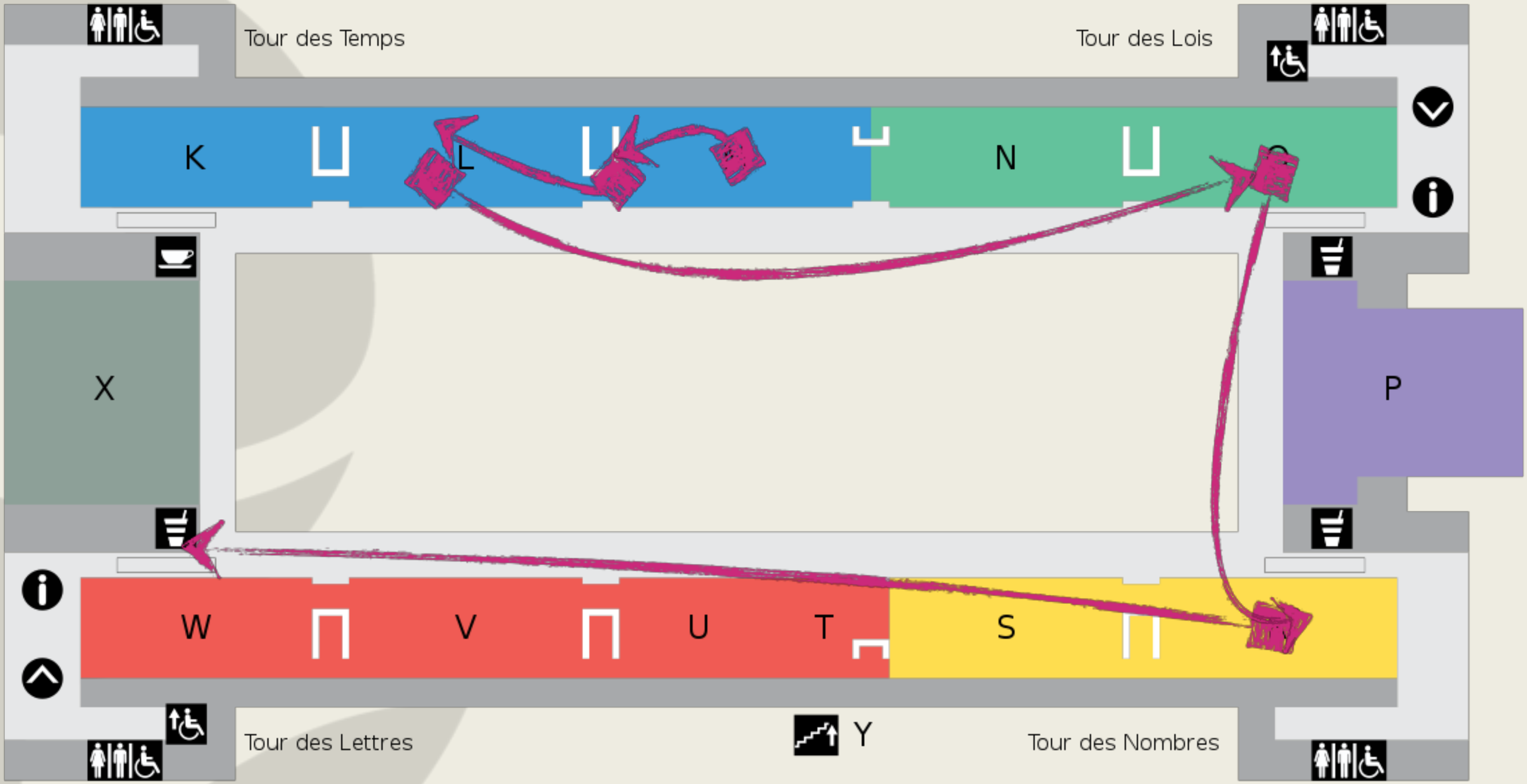
1. Introduction to the research project
   - Research questions
   - Mixed-method

2. Readers' *representation* of an informational space
   - General method
   - Quantitative analysis
   - Regimes of navigation
   - Critique of pipeline

3. Workshop

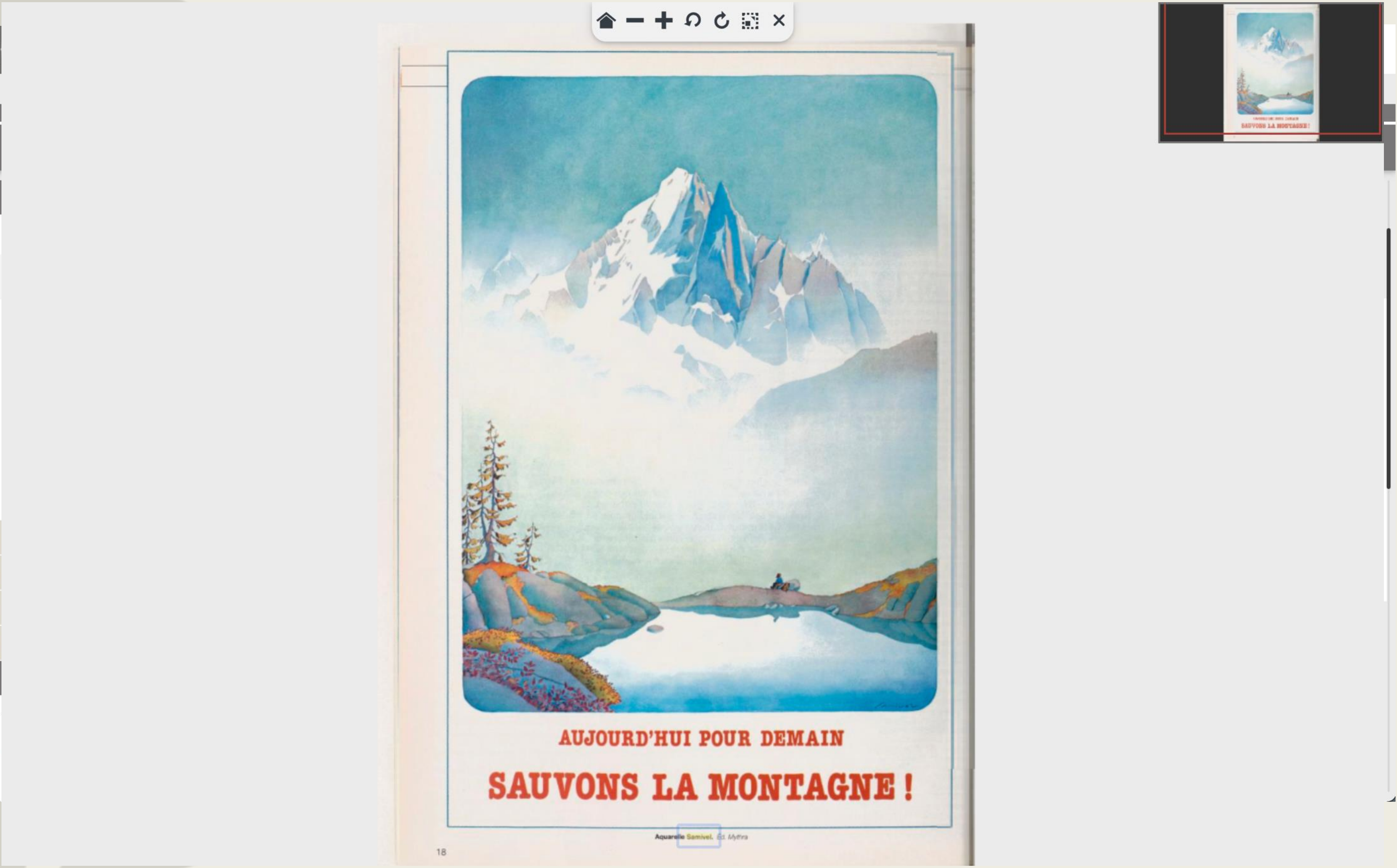
# 1. Introduction to the research project

## Research questions



# 1. Introduction to the research project

## Research questions



Aquarelle Samivel. Éd. Mythra

18

# 1. Introduction to the research project

## Research questions

BnF Gallica

the online digital library of the French National Library (BnF)

Today

- 10 million documents available freely online
- Books, manuscripts, press, photography, posters, musical scores, etc.
- Very few instruments to navigate this corpus
- Predominance of a (“dysfunctional”) search engine

# 1. Introduction to the research project

## Research questions

BnF Gallica

**Can we conceive of Gallica’s corpus as a documentary *milieu*?**

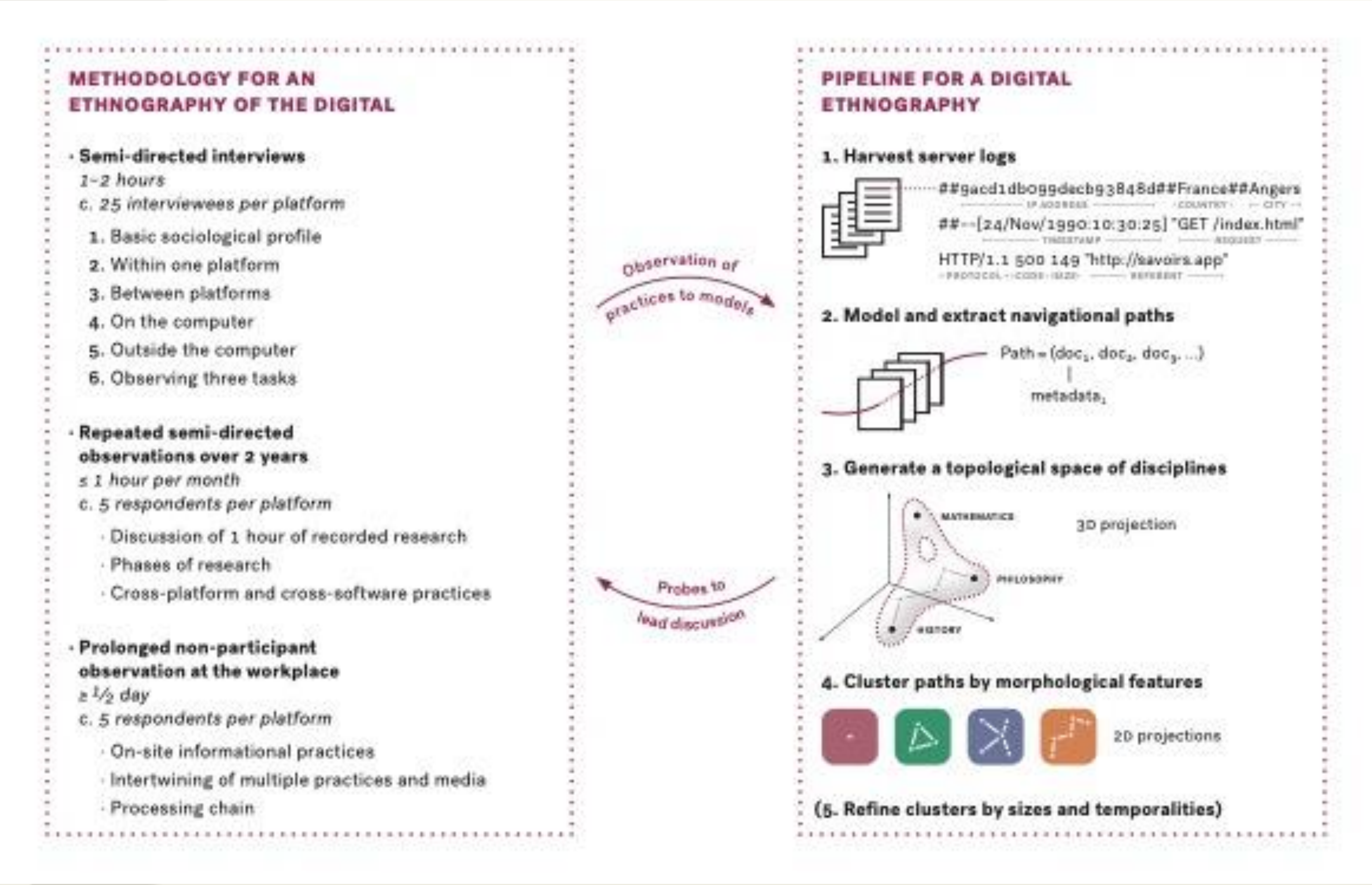
Question the notion of “digital library”

**How do readers orient themselves within Gallica’s documentary mass?**

Question the centrality of the search engine

# 2. Users' *representation* of an informational space

## General method



### METHODOLOGY FOR AN ETHNOGRAPHY OF THE DIGITAL

- **Semi-directed interviews**
  *1–2 hours*
  *c. 25 interviewees per platform*
  1. Basic sociological profile
  2. Within one platform
  3. Between platforms
  4. On the computer
  5. Outside the computer
  6. Observing three tasks

- **Repeated semi-directed observations over 2 years**
  *≤ 1 hour per month*
  *c. 5 respondents per platform*
  - Discussion of 1 hour of recorded research
  - Phases of research
  - Cross-platform and cross-software practices

- **Prolonged non-participant observation at the workplace**
  *≥ ½ day*
  *c. 5 respondents per platform*
  - On-site informational practices
  - Intertwining of multiple practices and media
  - Processing chain

Observation of practices to models

Probes to lead discussion

### PIPELINE FOR A DIGITAL ETHNOGRAPHY

**1. Harvest server logs**

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

**2. Model and extract navigational paths**

$\text{Path} = (\text{doc}_1, \text{doc}_2, \text{doc}_3, \dots)$

$\text{metadata}_1$

**3. Generate a topological space of disciplines**

3D projection

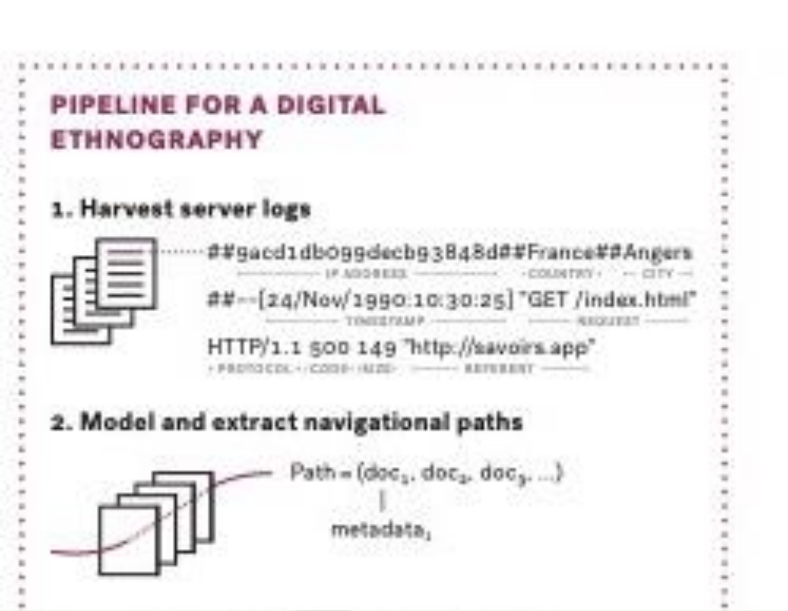
**4. Cluster paths by morphological features**

2D projections

**(5. Refine clusters by sizes and temporalities)**

# 2. Users’ *representation* of an informational space

Quantitative analysis



PIPELINE FOR A DIGITAL ETHNOGRAPHY

1. Harvest server logs

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

2. Model and extract navigational paths

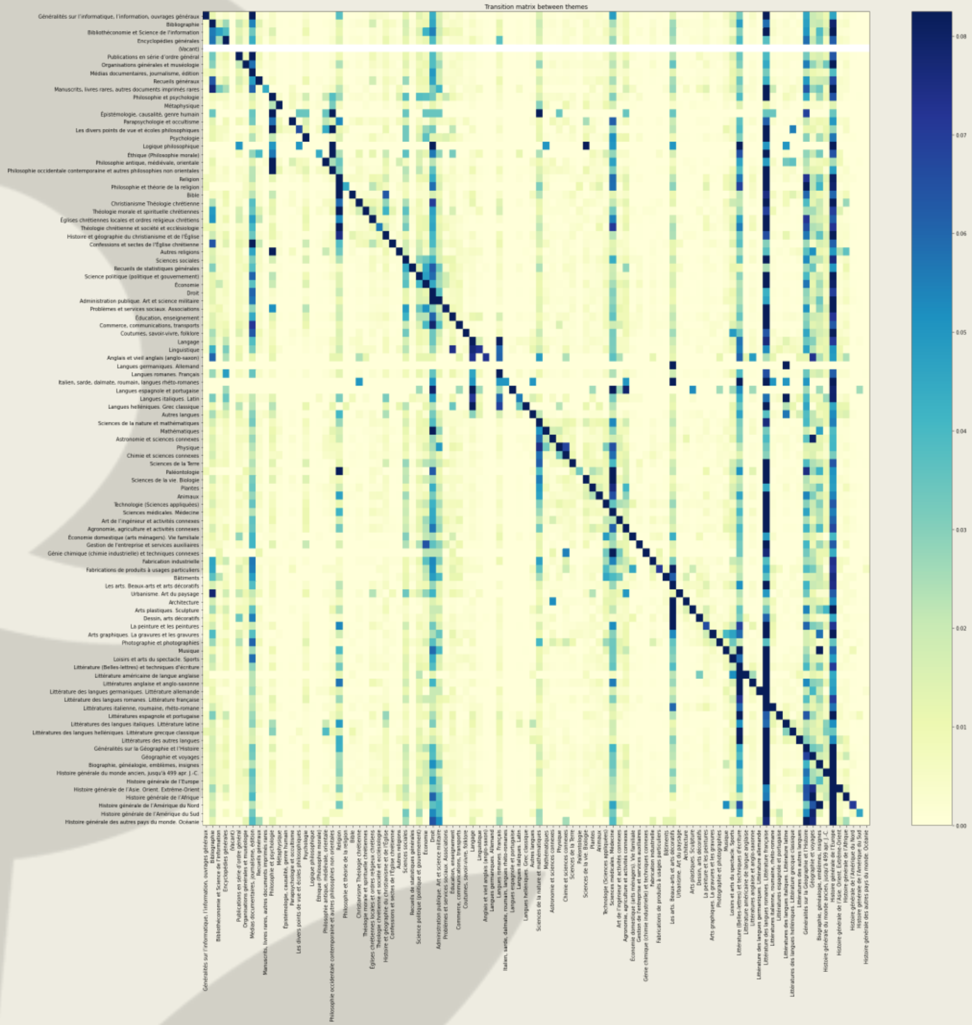
$\text{Path} = (\text{doc}_1, \text{doc}_2, \text{doc}_3, \ldots)$

$\text{metadata}_i$

+ data enrichment
+ exploratory analysis

# 2. Users' *representation* of an informational space

## Quantitative analysis



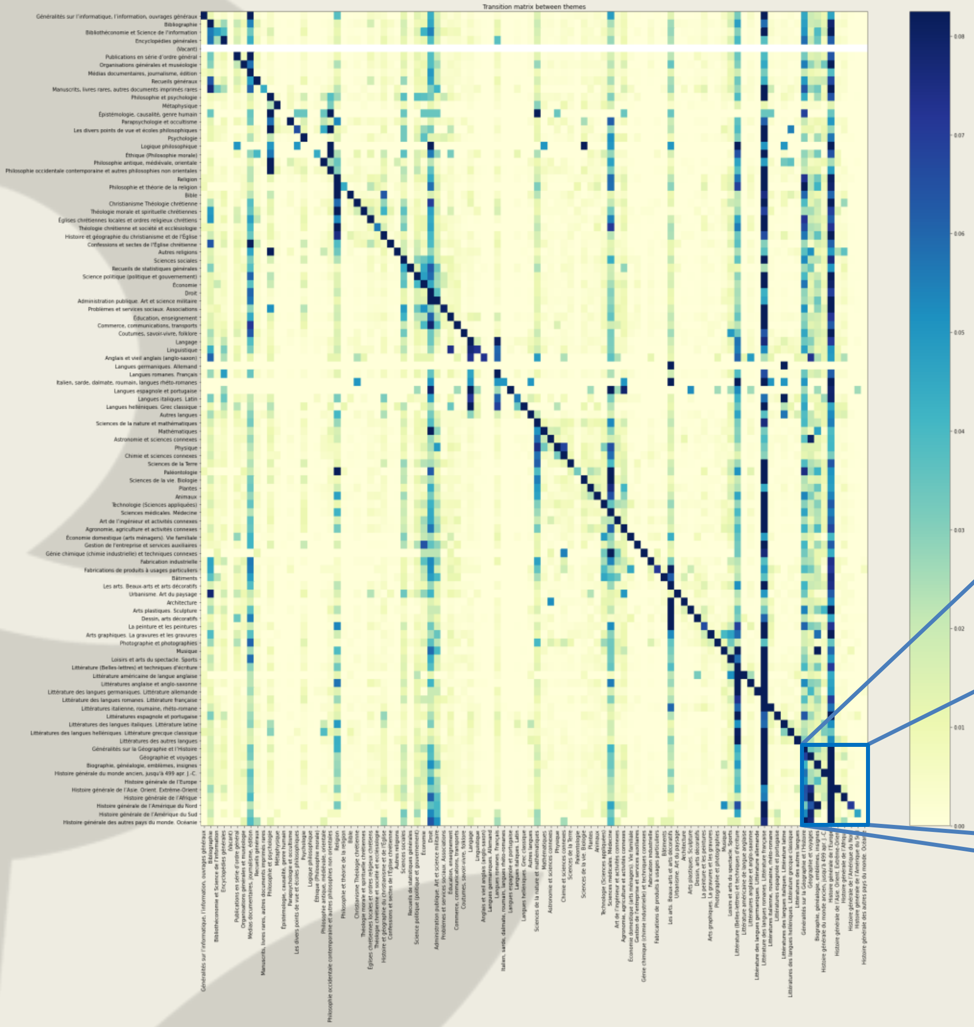
Dewey classes structure navigation paths…

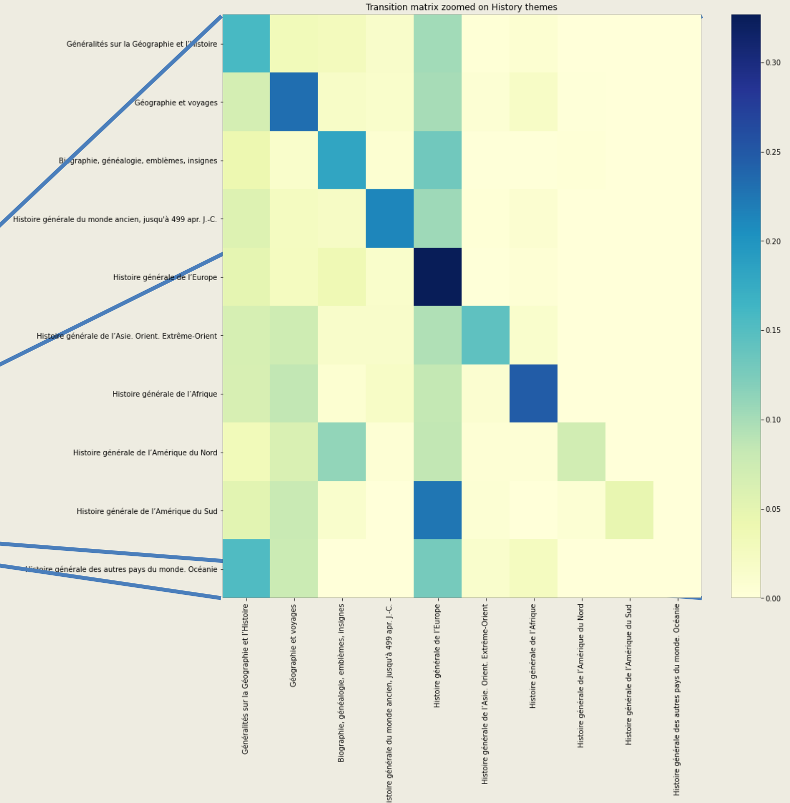
…despite users

- « *Enclosures* »
- « Pivot literature »

# 2. Users’ *representation* of an informational space

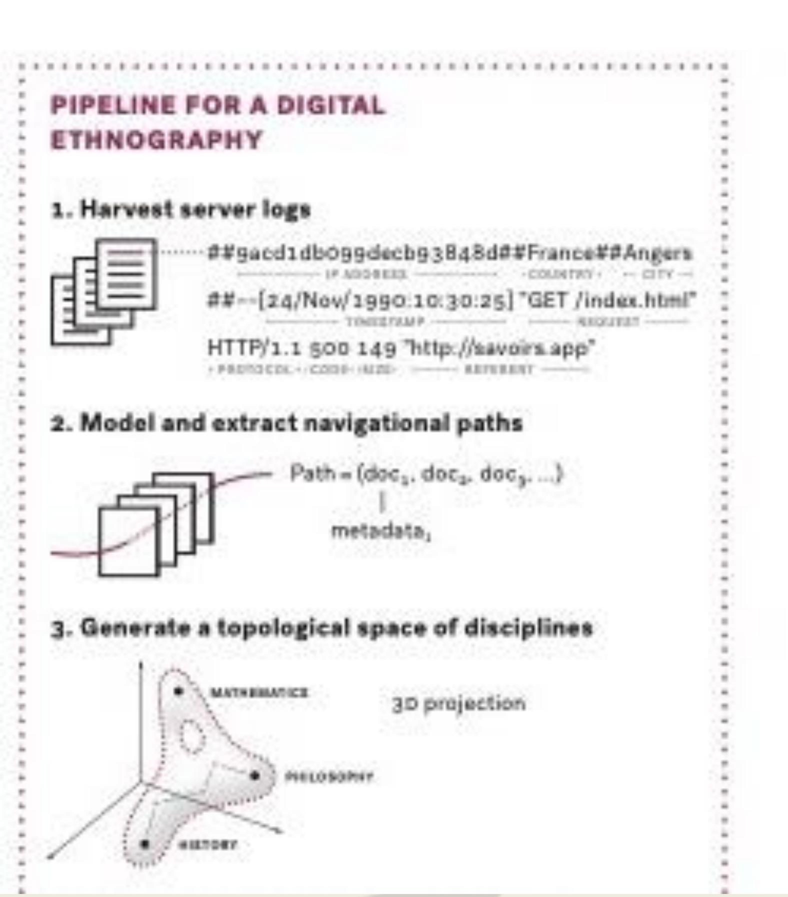
## Quantitative analysis





# 2. Users’ *representation* of an informational space

## Quantitative analysis



PIPELINE FOR A DIGITAL ETHNOGRAPHY

1. Harvest server logs

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

2. Model and extract navigational paths

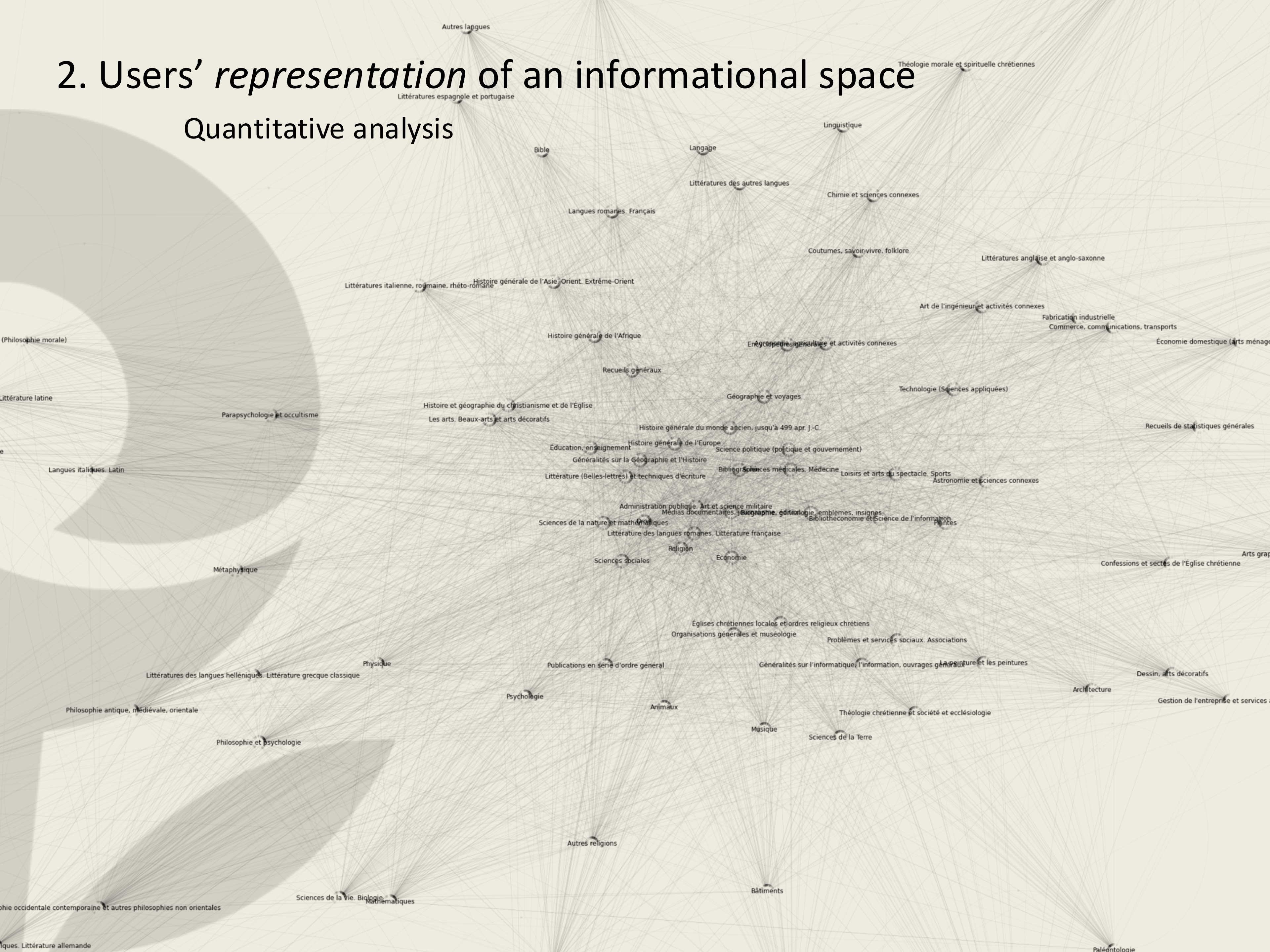
Path = ($doc_1$, $doc_2$, $doc_3$, …)

$metadata_1$

3. Generate a topological space of disciplines

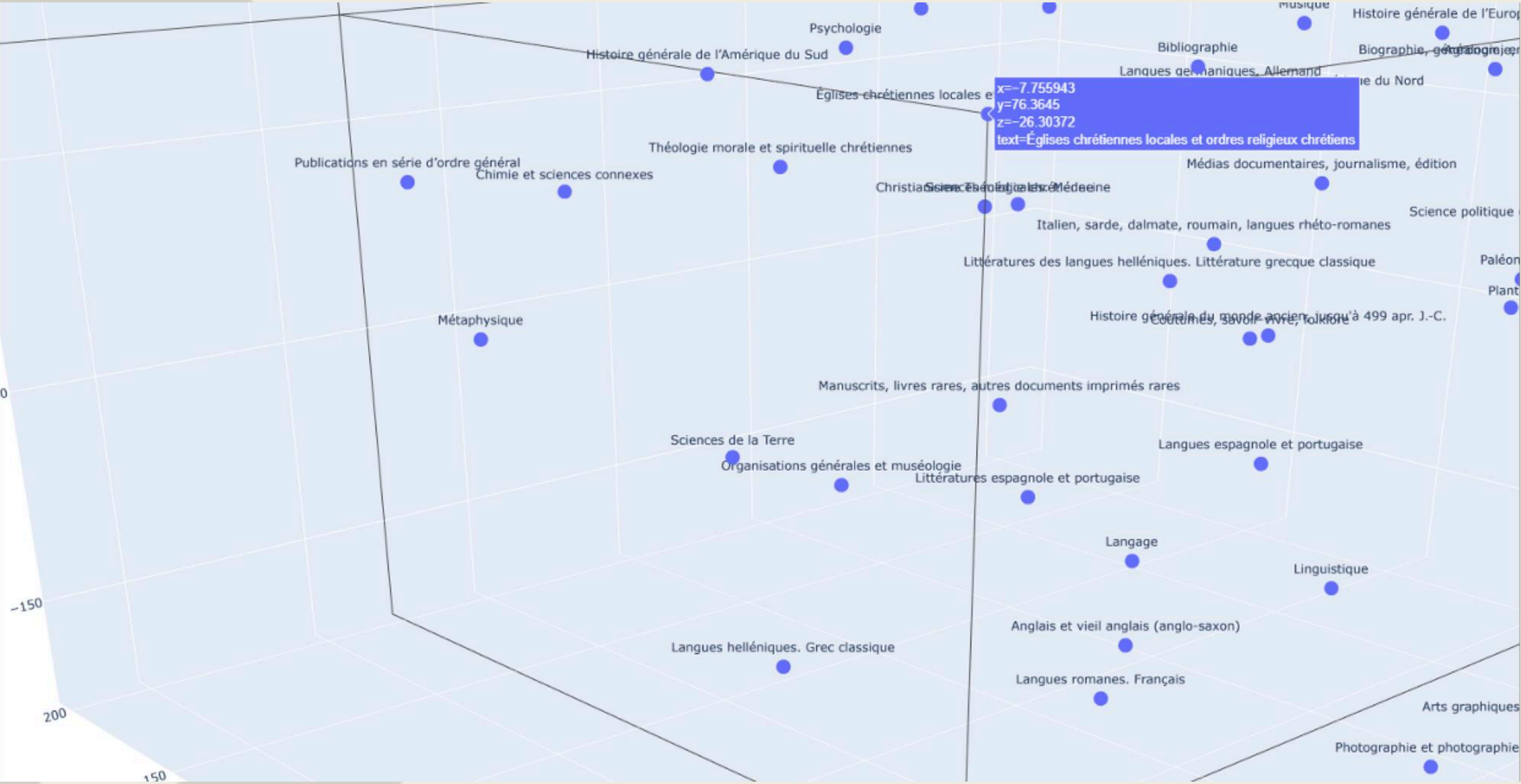
# 2. Users' *representation* of an informational space

## Quantitative analysis


Autres langues
Théologie morale et spirituelle chrétiennes
Littératures espagnole et portugaise
Linguistique
Bible
Langage
Littératures des autres langues
Chimie et sciences connexes
Langues romanes. Français
Coutumes, savoir-vivre, folklore
Littératures anglaise et anglo-saxonne
Littératures italienne, roumaine, rhéto-romane
Histoire générale de l'Asie. Orient. Extrême-Orient
Art de l'ingénieur et activités connexes
Fabrication industrielle
Commerce, communications, transports
(Philosophie morale)
Histoire générale de l'Afrique
Encyclopédies générales
Agronomie, agriculture et activités connexes
Économie domestique (arts ménage
Recueils généraux
Littérature latine
Géographie et voyages
Technologie (Sciences appliquées)
Parapsychologie et occultisme
Histoire et géographie du christianisme et de l'Église
Les arts. Beaux-arts et arts décoratifs
Histoire générale du monde ancien, jusqu'à 499 apr. J.-C.
Recueils de statistiques générales
Éducation, enseignement
Histoire générale de l'Europe
Science politique (politique et gouvernement)
Généralités sur la Géographie et l'Histoire
Langues italiques. Latin
Bibliographie
Sciences médicales. Médecine
Loisirs et arts du spectacle. Sports
Littérature (Belles-lettres) et techniques d'écriture
Astronomie et sciences connexes
Administration publique. Art et science militaire
Médias documentaires, journalisme, édition
Biographie, généalogie, emblèmes, insignes
Sciences de la nature et mathématiques
Droit
Bibliothéconomie et Science de l'information
Plantes
Littérature des langues romanes. Littérature française
Religion
Économie
Sciences sociales
Arts grap
Métaphysique
Confessions et sectes de l'Église chrétienne
Églises chrétiennes locales et ordres religieux chrétiens
Organisations générales et muséologie
Problèmes et services sociaux. Associations
Physique
Littératures des langues helléniques. Littérature grecque classique
Publications en série d'ordre général
Généralités sur l'informatique, l'information, ouvrages généraux
La peinture et les peintures
Dessin, arts décoratifs
Architecture
Psychologie
Gestion de l'entreprise et services a
Philosophie antique, médiévale, orientale
Animaux
Théologie chrétienne et société et ecclésiologie
Musique
Sciences de la Terre
Philosophie et psychologie
Autres religions
Bâtiments
Sciences de la vie. Biologie
Mathématiques
phie occidentale contemporaine et autres philosophies non orientales
iques. Littérature allemande
Paléontologie

# 2. Users’ *representation* of an informational space

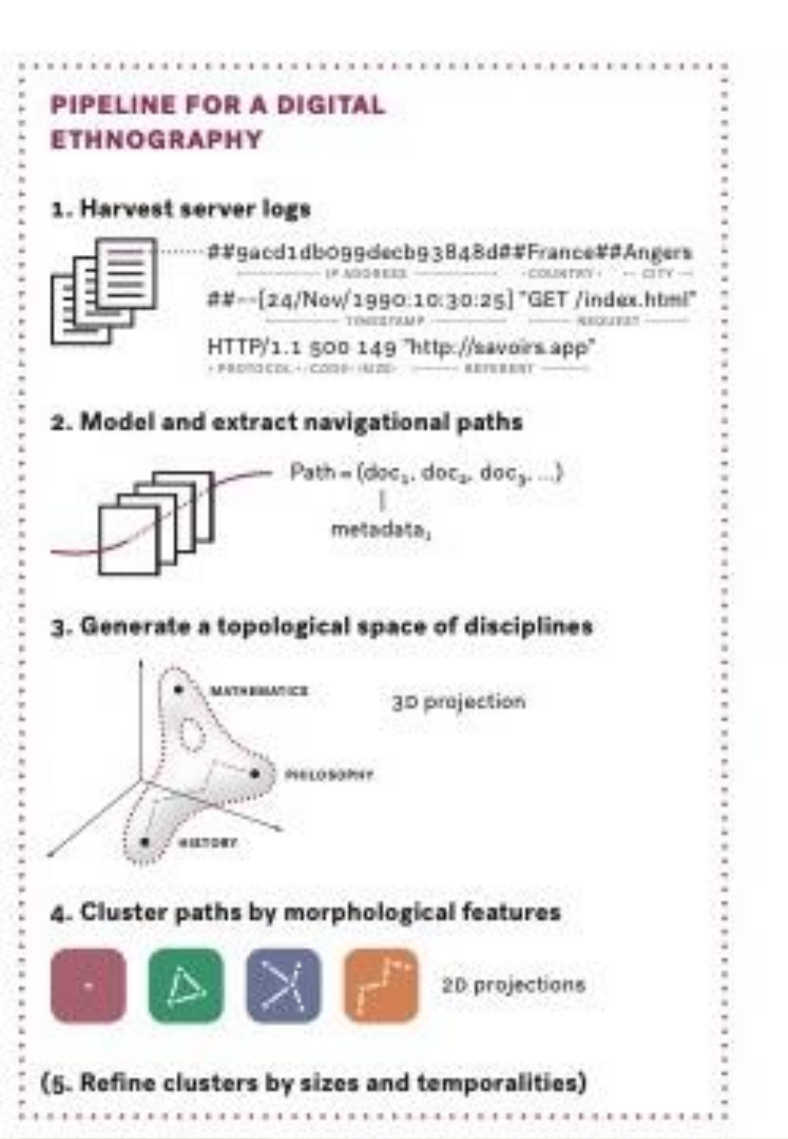
## Quantitative analysis



x=−7.755943
y=76.3645
z=−26.30372
text=Églises chrétiennes locales et ordres religieux chrétiens

# 2. Users' *representation* of an informational space

Quantitative analysis



PIPELINE FOR A DIGITAL ETHNOGRAPHY

1. Harvest server logs

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

2. Model and extract navigational paths

Path = ($doc_1$, $doc_2$, $doc_3$, ...)
|
$metadata_i$

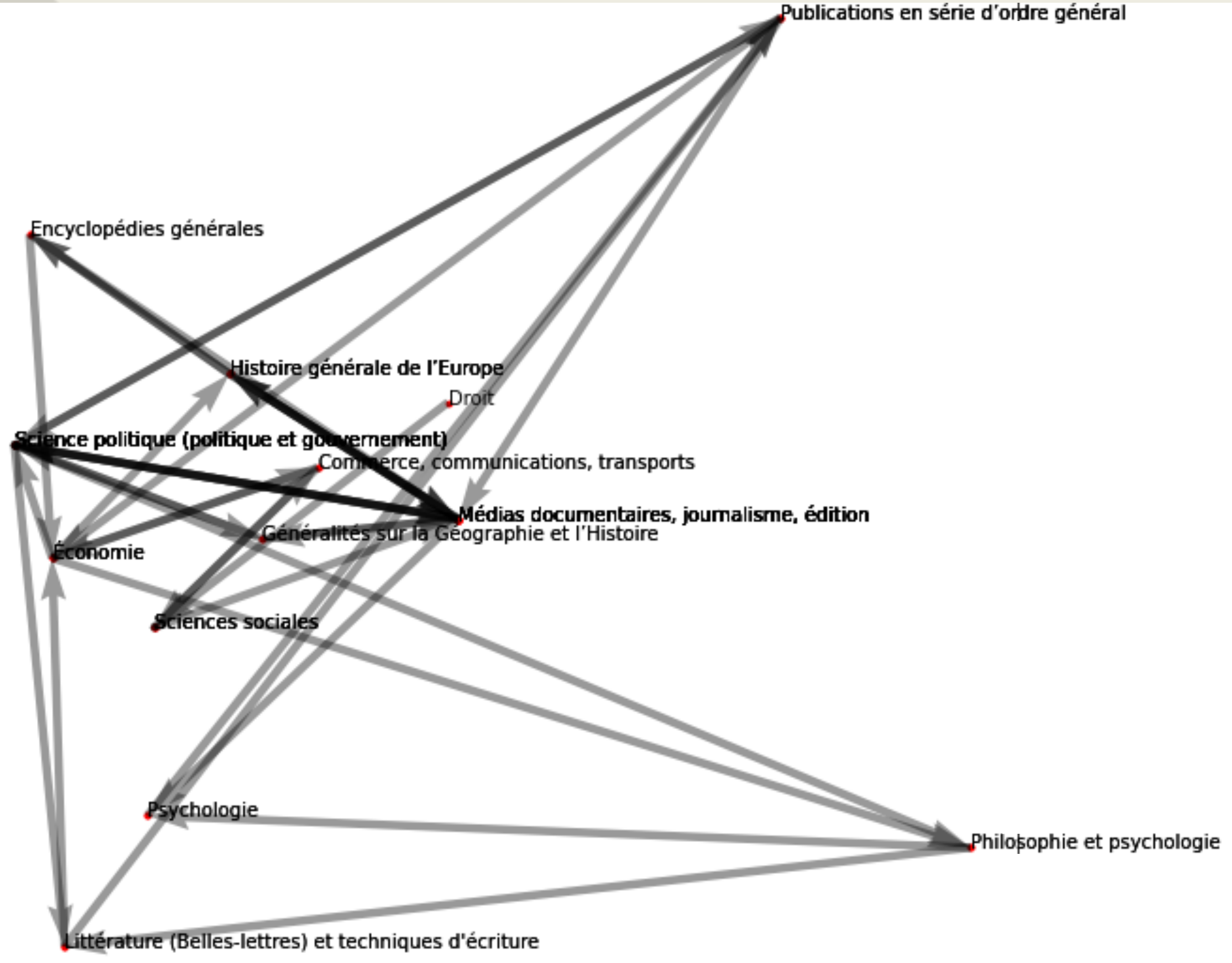
3. Generate a topological space of disciplines

4. Cluster paths by morphological features

(5. Refine clusters by sizes and temporalities)

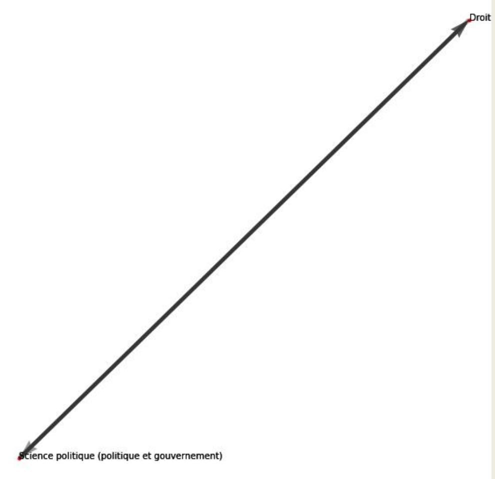
# 2. Users' *representation* of an informational space

## Quantitative analysis


Publications en série d'ordre général
Encyclopédies générales
Histoire générale de l'Europe
Droit
Science politique (politique et gouvernement)
Commerce, communications, transports
Médias documentaires, journalisme, édition
Généralités sur la Géographie et l'Histoire
Économie
Sciences sociales
Psychologie
Philosophie et psychologie
Littérature (Belles-lettres) et techniques d'écriture

# 2. Users' *representation* of an informational space

## Regimes of navigation



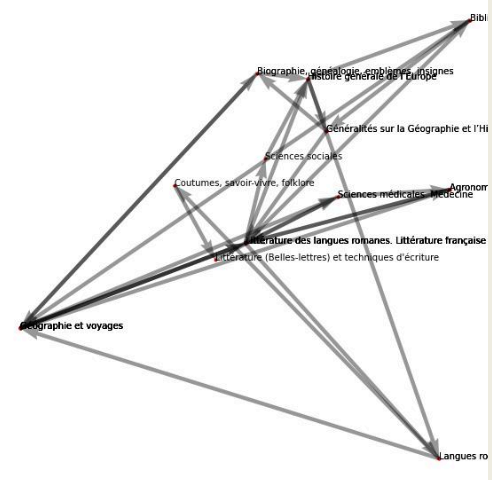
| Regime | Directed search |
|---|---|
| Duration | Very short |
| Mean nb. of docs | 6 to 7 |
| Dewey reach | Limited (2 to 3) |
| Documentary unit | Information or document |
| Values | Relevance, efficiency, findability |
| ~~Proportions~~ | ~~~12%~~ |

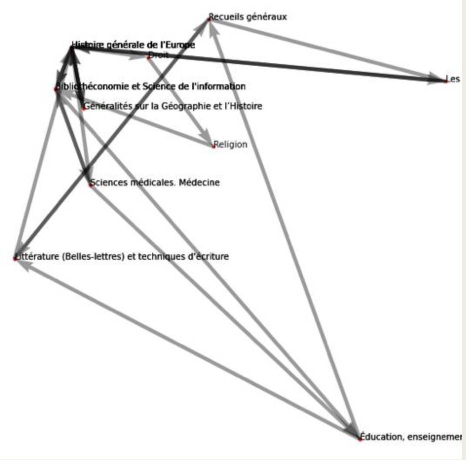
Langues romanes. Français

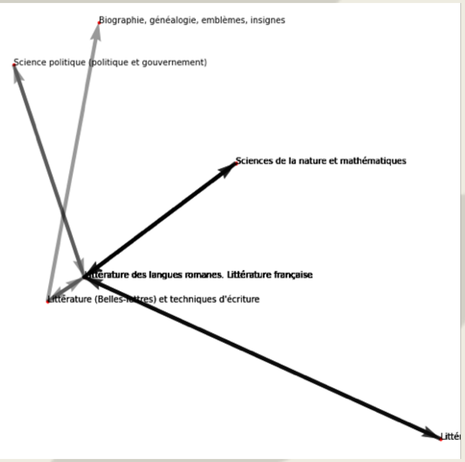
| Regime | Consultation / constitution of a corpus |
|---|---|
| Duration | Long |
| Mean nb. of docs | 7 to 10 |
| Dewey reach | Very limited (1 to 2) |
| Documentary unit | Documents, references, corpus |
| Value | Exhaustivity |

# 2. Users' *representation* of an informational space

## Regimes of navigation


Bibl
Biographie, généalogie, emblèmes, insignes
Histoire générale de l'Europe
Généralités sur la Géographie et l'Hi
Sciences sociales
Coutumes, savoir-vivre, folklore
Agronom
Sciences médicales. Médecine
Littérature des langues romanes. Littérature française
Littérature (Belles-lettres) et techniques d'écriture
Géographie et voyages
Langues ro


Recueils généraux
Histoire générale de l'Europe
Droit
Les
Bibliothéconomie et Science de l'information
Généralités sur la Géographie et l'Histoire
Religion
Sciences médicales. Médecine
Littérature (Belles-lettres) et techniques d'écriture
Éducation, enseignemen


Biographie, généalogie, emblèmes, insignes
Science politique (politique et gouvernement)
Sciences de la nature et mathématiques
Littérature des langues romanes. Littérature française
Littérature (Belles-lettres) et techniques d'écriture
Littér

| **Regime** | **Star-shaped (structured exploratory research)** |
|---|---|
| Duration | Short to very long |
| Mean nb. of docs | 20 |
| Dewey reach | Extended (10) |
| Documentary unit | Topic or field |
| Values | Comparison, diversity |

# 2. Users' *representation* of an informational space

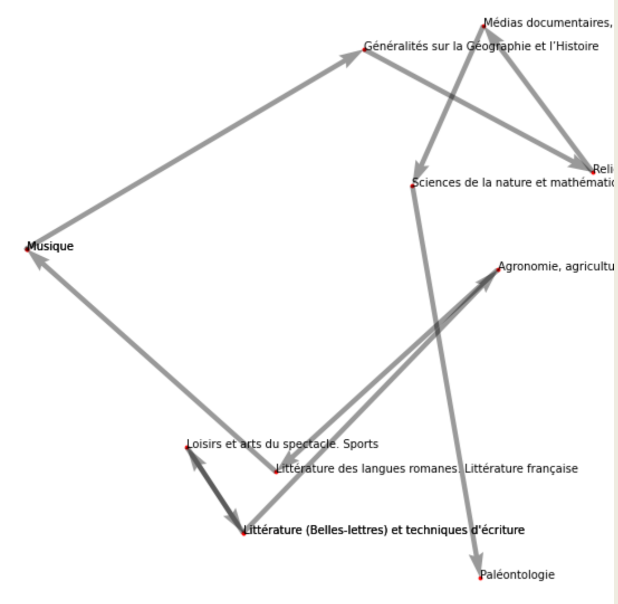
## Regimes of navigation

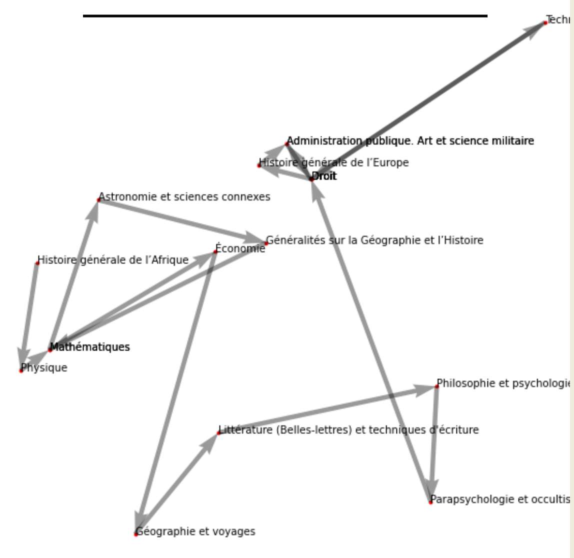
?

| Regime | Crawling (indexing bots or DH algorithms) |
|---|---|
| Duration | Long |
| Mean nb. of docs | Unknown (supposedly bery large) |
| Dewey reach | Unknown |
| Documentary unit | Metadata |
| Values | Reliability, quantity |

# 2. Users' *representation* of an informational space

## Regimes of navigation

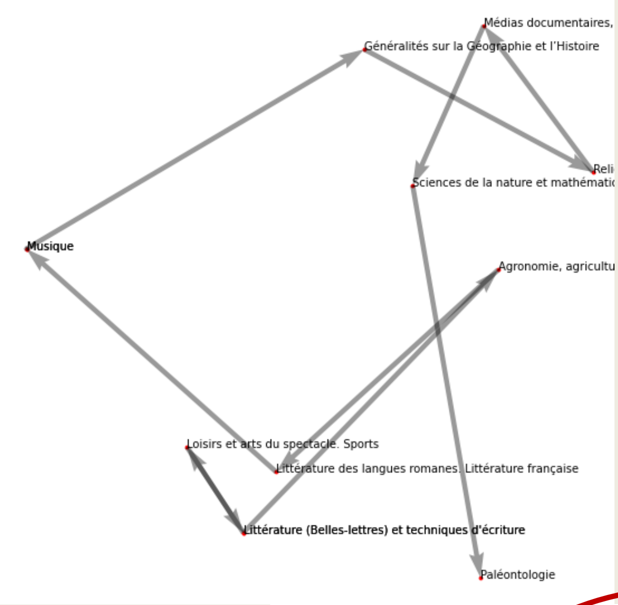

Médias documentaires,
Généralités sur la Géographie et l'Histoire
Reli
Sciences de la nature et mathématic
Musique
Agronomie, agricultu
Loisirs et arts du spectacle. Sports
Littérature des langues romanes. Littérature française
Littérature (Belles-lettres) et techniques d'écriture
Paléontologie


Techn
Administration publique. Art et science militaire
Histoire générale de l'Europe
Droit
Astronomie et sciences connexes
Généralités sur la Géographie et l'Histoire
Économie
Histoire générale de l'Afrique
Mathématiques
Physique
Philosophie et psychologie
Littérature (Belles-lettres) et techniques d'écriture
Parapsychologie et occultis
Géographie et voyages

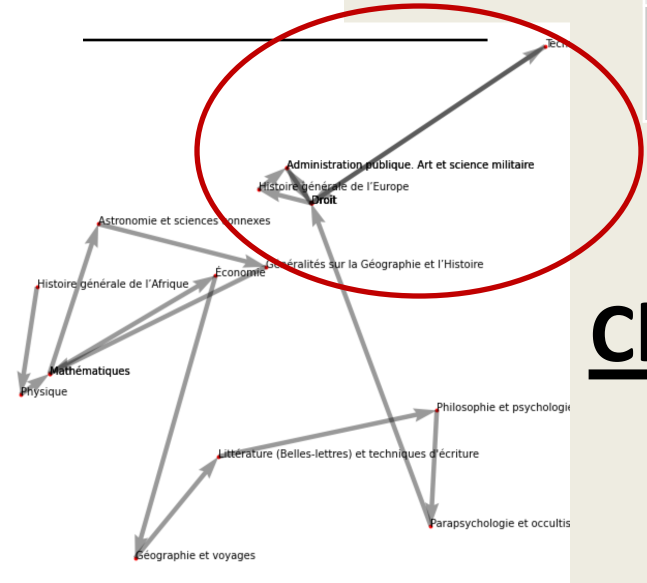
| Regime | Roaming or wandering (unstructured exploratory research) |
|---|---|
| Duration | Medium to long |
| Mean nb. of docs | More than 20 |
| Dewey reach | Very extended (more than 10) |
| Documentary unit | Unexpected documents, ideas, concepts |
| Values | Serendipity, discoverability |

# 2. Users' *representation* of an informational space

## Regimes of navigation



Médias documentaires,
Généralités sur la Géographie et l'Histoire
Reli
Sciences de la nature et mathémati
Musique
Agronomie, agricultu
Loisirs et arts du spectacle. Sports
Littérature des langues romanes. Littérature française
Littérature (Belles-lettres) et techniques d'écriture
Paléontologie



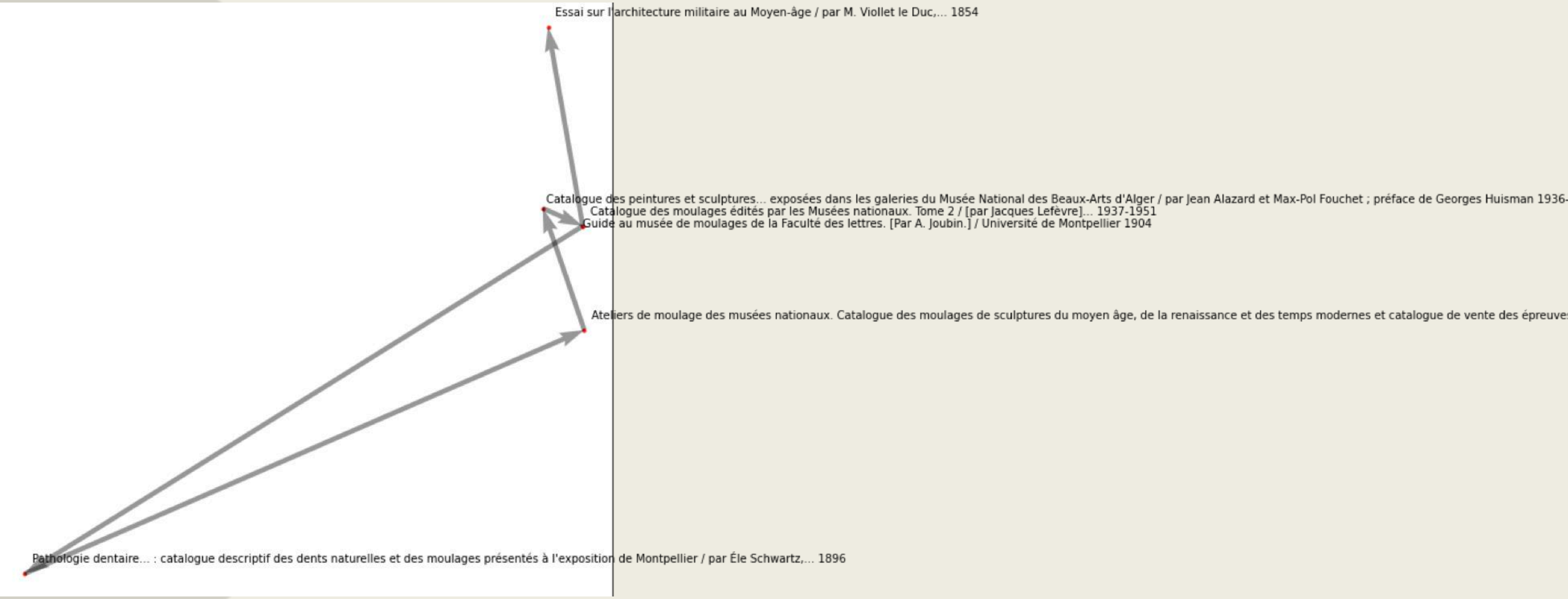
Tech
Administration publique. Art et science militaire
Histoire générale de l'Europe
Droit
Astronomie et sciences connexes
Généralités sur la Géographie et l'Histoire
Économie
Histoire générale de l'Afrique
Mathématiques
Physique
Philosophie et psychologie
Littérature (Belles-lettres) et techniques d'écriture
Parapsychologie et occultis
Géographie et voyages

| **Regime** | **Roaming or wandering (unstructured exploratory research)** |
|---|---|
| Duration | Medium to long |
| Mean nb. of docs | More than 20 |
| Dewey reach | Very extended (more than 10) |
| Documentary unit | Unexpected documents, ideas, concepts |
| Values | Serendipity, discoverability |

**Clusters are ideal-types!**

# 2. Users’ *representation* of an informational space

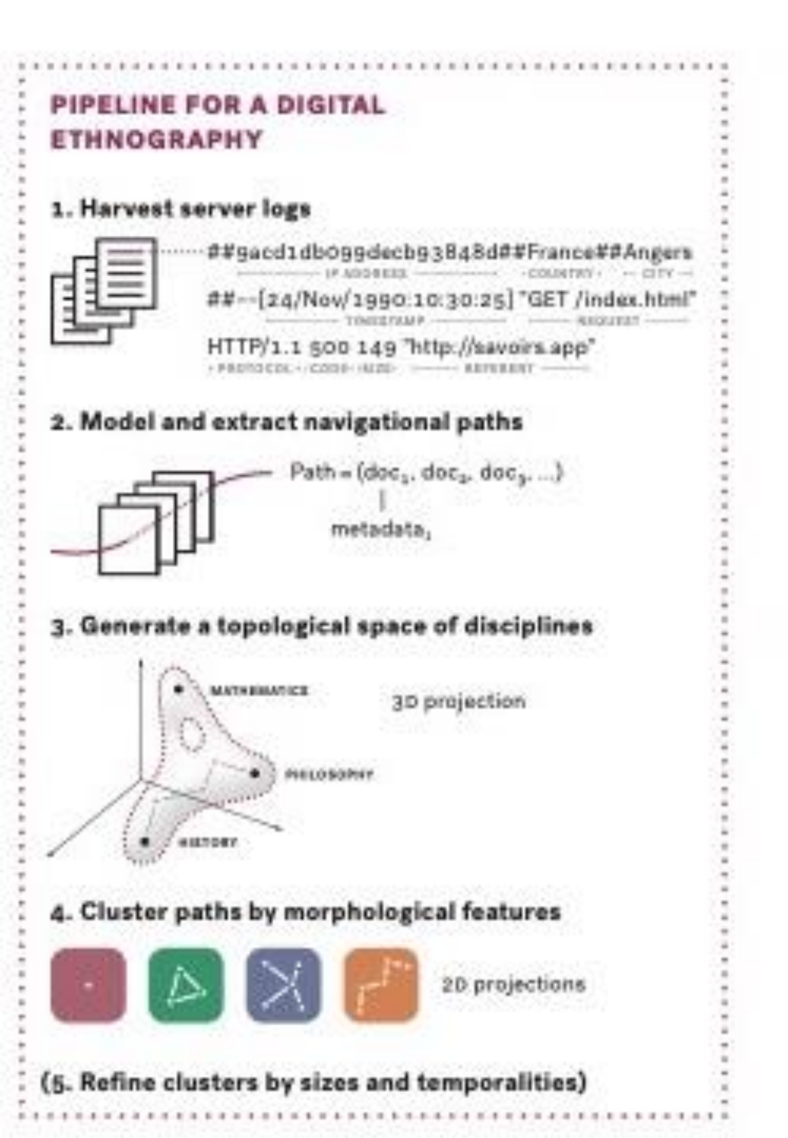
## Regimes of navigation


Essai sur l'architecture militaire au Moyen-âge / par M. Viollet le Duc,... 1854
Catalogue des peintures et sculptures... exposées dans les galeries du Musée National des Beaux-Arts d'Alger / par Jean Alazard et Max-Pol Fouchet ; préface de Georges Huisman 1936-
Catalogue des moulages édités par les Musées nationaux. Tome 2 / [par Jacques Lefèvre]... 1937-1951
Guide au musée de moulages de la Faculté des lettres. [Par A. Joubin.] / Université de Montpellier 1904
Ateliers de moulage des musées nationaux. Catalogue des moulages de sculptures du moyen âge, de la renaissance et des temps modernes et catalogue de vente des épreuve
Pathologie dentaire... : catalogue descriptif des dents naturelles et des moulages présentés à l'exposition de Montpellier / par Éle Schwartz,... 1896

**Significant outliers: e.g. getting lost!**

# 2. Users’ *representation* of an informational space

Critique of pipeline



PIPELINE FOR A DIGITAL ETHNOGRAPHY

1. Harvest server logs

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

2. Model and extract navigational paths

$\text{Path} = (doc_1, doc_2, doc_3, \ldots)$

$\text{metadata}_i$

3. Generate a topological space of disciplines

3D projection

4. Cluster paths by morphological features

2D projections

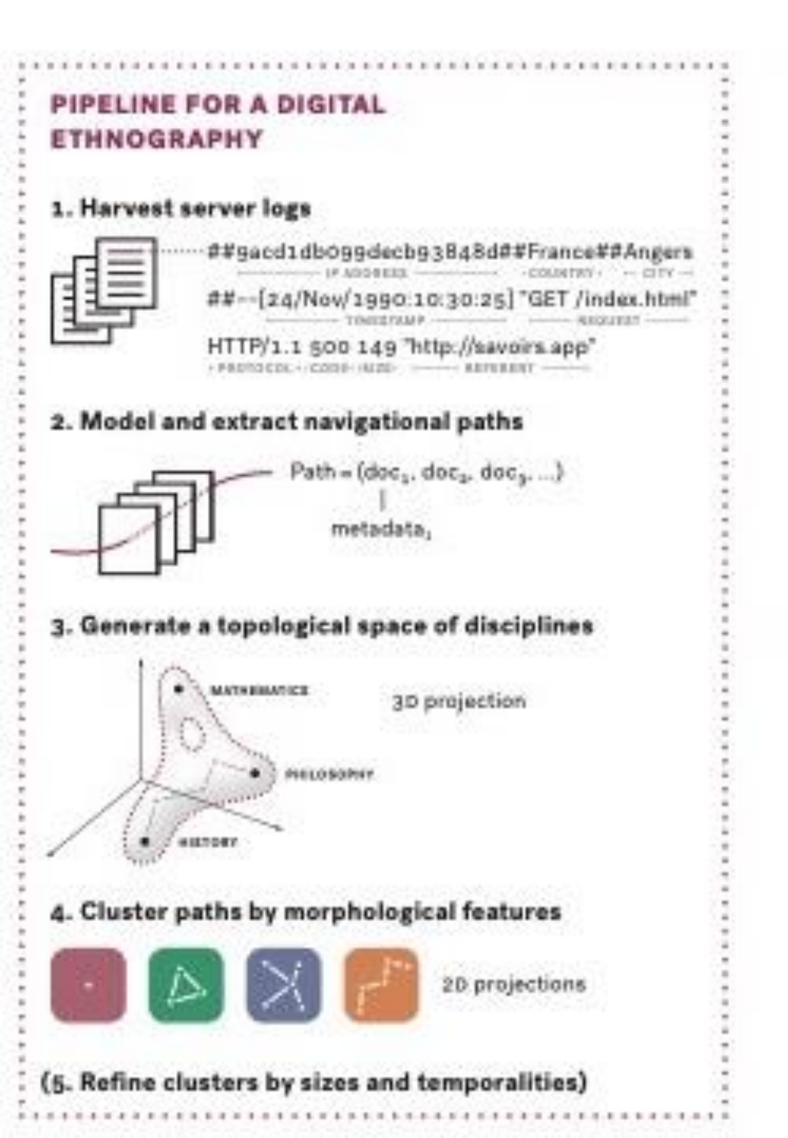
(5. Refine clusters by sizes and temporalities)

Issues raised:

Trace / Clue

# 2. Users' *representation* of an informational space

## Critique of pipeline



PIPELINE FOR A DIGITAL ETHNOGRAPHY

1. Harvest server logs

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

2. Model and extract navigational paths

$\text{Path} = (\text{doc}_1, \text{doc}_2, \text{doc}_3, \ldots)$

3. Generate a topological space of disciplines

4. Cluster paths by morphological features

(5. Refine clusters by sizes and temporalities)

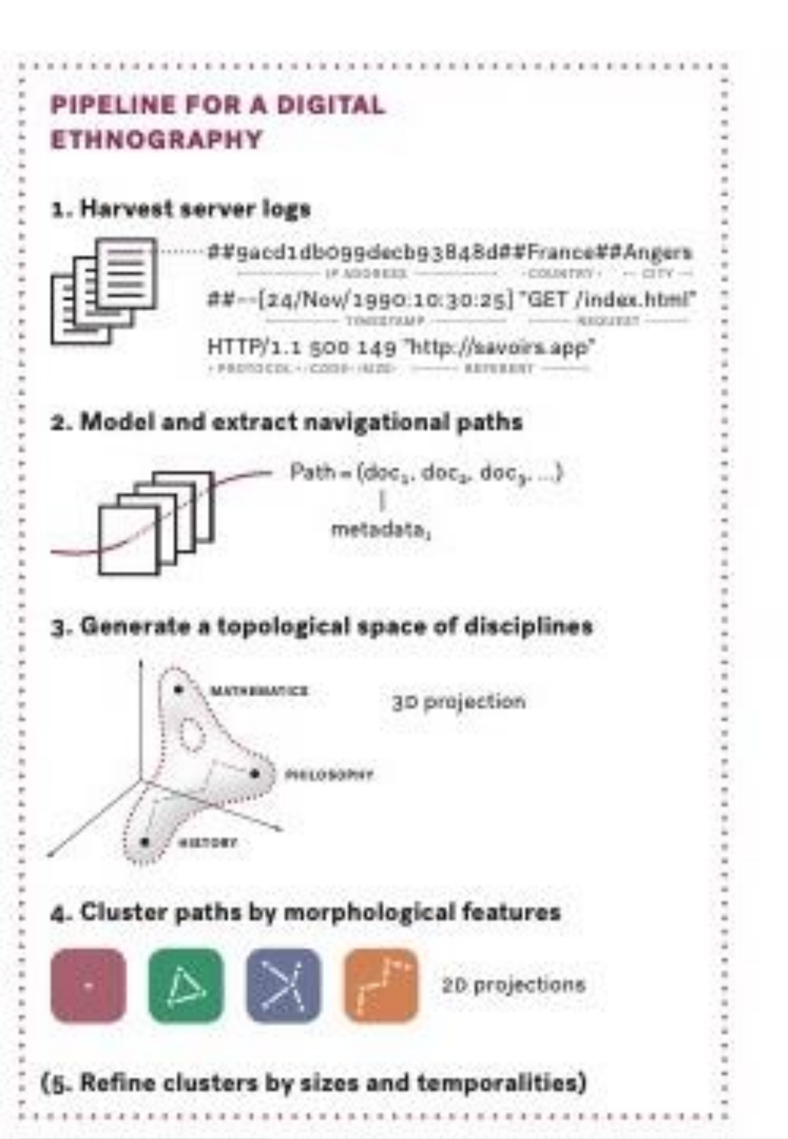
Issues raised:

## Trace / Clue

Turn *machine-machine-interaction traces* into *clues* of human behaviour by carefully *selecting* and *structuring* them

# 2. Users' *representation* of an informational space

Critique of pipeline



**PIPELINE FOR A DIGITAL ETHNOGRAPHY**

**1. Harvest server logs**

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

**2. Model and extract navigational paths**

Path = ($doc_1$, $doc_2$, $doc_3$, ...)

$metadata_1$

**3. Generate a topological space of disciplines**

**4. Cluster paths by morphological features**

**(5. Refine clusters by sizes and temporalities)**

Issues raised:

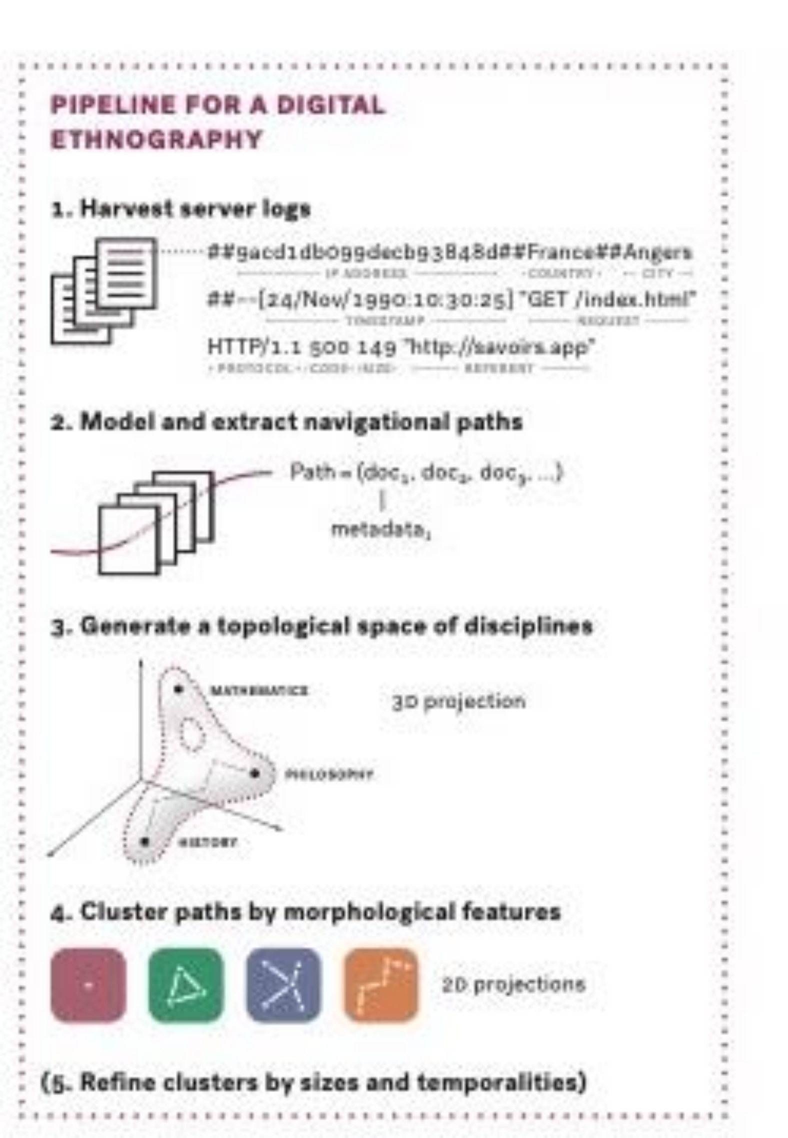
## Trace / Clue

Turn *machine-machine-interaction traces* into *clues* of human behaviour by carefully *selecting* and *structuring* them

## Models

# 2. Users' *representation* of an informational space

Critique of pipeline



**PIPELINE FOR A DIGITAL ETHNOGRAPHY**

**1. Harvest server logs**

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

**2. Model and extract navigational paths**

Path = (doc$_1$, doc$_2$, doc$_3$, ...)

metadata$_1$

**3. Generate a topological space of disciplines**

**4. Cluster paths by morphological features**

**(5. Refine clusters by sizes and temporalities)**

Issues raised:

## Trace / Clue

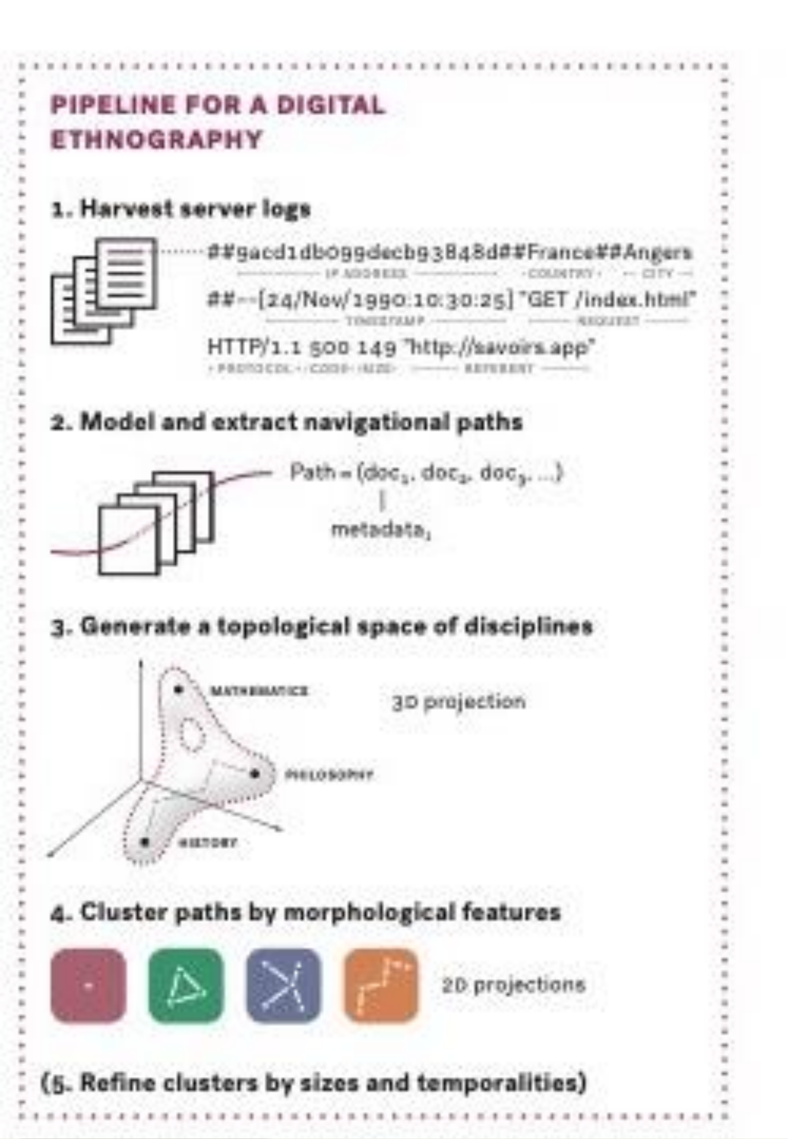
Turn *machine-machine-interaction traces* into *clues* of human behaviour by carefully *selecting* and *structuring* them

## Models

Recover complex structures from sequential logs by modelling multiple behaviors recounted by users (cognitive mechanisms and interaction with interface)

# 2. Users' *representation* of an informational space

Critique of pipeline



PIPELINE FOR A DIGITAL ETHNOGRAPHY

1. Harvest server logs

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

2. Model and extract navigational paths

Path = ($doc_1$, $doc_2$, $doc_3$, ...)

3. Generate a topological space of disciplines

4. Cluster paths by morphological features

(5. Refine clusters by sizes and temporalities)

Issues raised:

## Trace / Clue

Turn *machine-machine-interaction traces* into *clues* of human behaviour by carefully *selecting* and *structuring* them

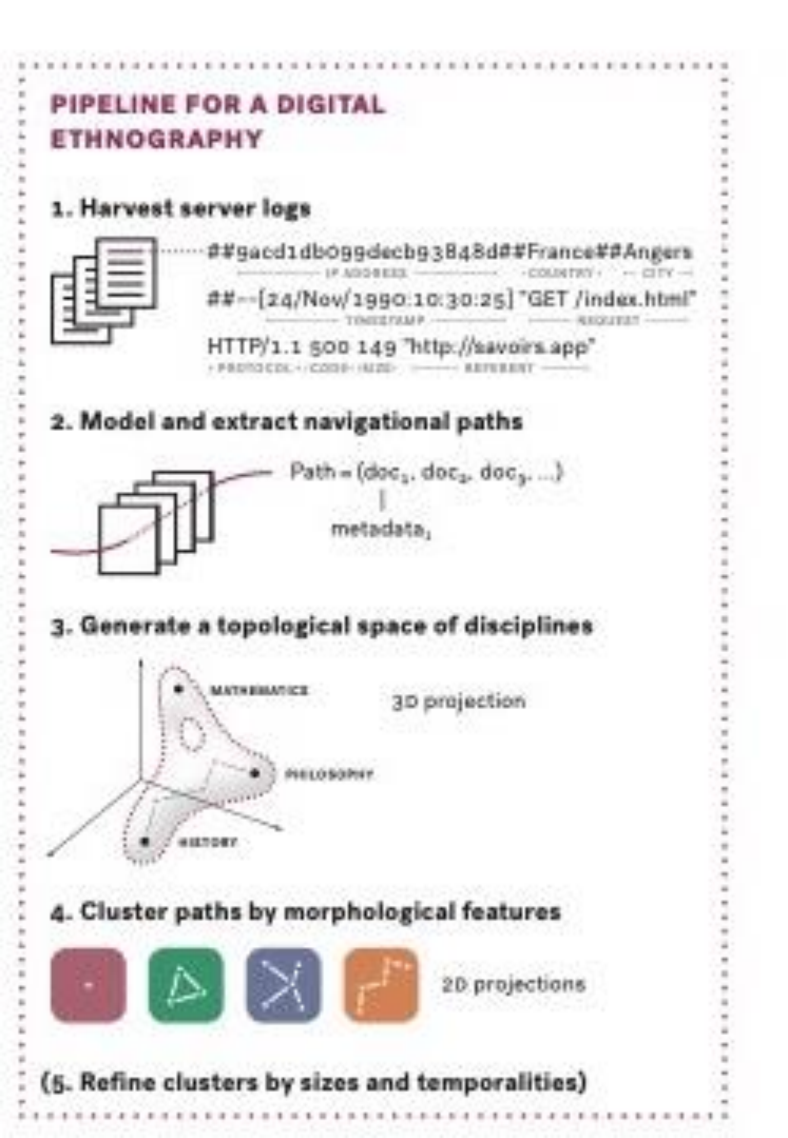
## Models

Recover complex structures from sequential logs by modelling multiple behaviors recounted by users (cognitive mechanisms and interaction with interface)

## Representation

# 2. Users' *representation* of an informational space

## Critique of pipeline



PIPELINE FOR A DIGITAL ETHNOGRAPHY

1. Harvest server logs

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

2. Model and extract navigational paths

3. Generate a topological space of disciplines

4. Cluster paths by morphological features

(5. Refine clusters by sizes and temporalities)

Issues raised:

### Trace / Clue

Turn *machine-machine-interaction traces* into *clues* of human behaviour by carefully *selecting* and *structuring* them

### Models

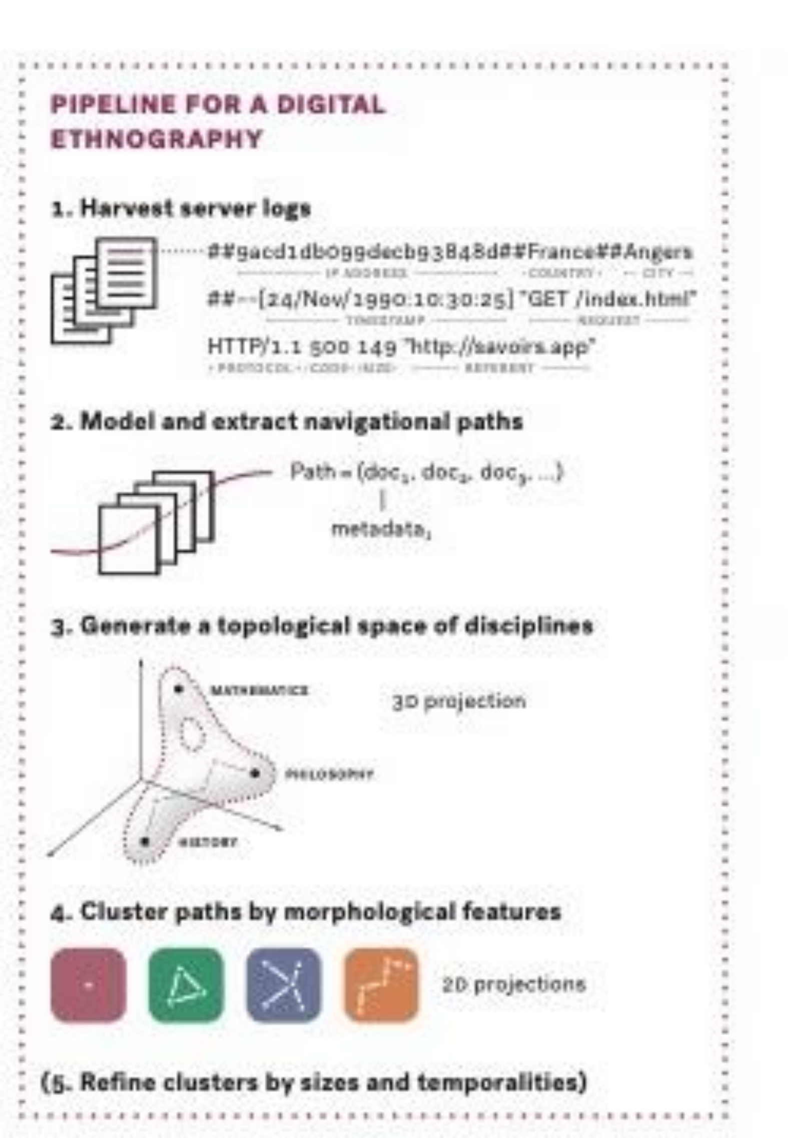
Recover complex structures from sequential logs by modelling multiple behaviors recounted by users (cognitive mechanisms and interaction with interface)

### Representation

A relative and relational space representing navigation practices, not a pre-existing territory

# 2. Users’ *representation* of an informational space

Critique of pipeline



**PIPELINE FOR A DIGITAL ETHNOGRAPHY**

**1. Harvest server logs**

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

**2. Model and extract navigational paths**

**3. Generate a topological space of disciplines**

**4. Cluster paths by morphological features**

**(5. Refine clusters by sizes and temporalities)**

Issues raised:

## Trace / Clue

Turn *machine-machine-interaction traces* into *clues* of human behaviour by carefully *selecting* and *structuring* them

## Models

Recover complex structures from sequential logs by modelling multiple behaviors recounted by users (cognitive mechanisms and interaction with interface)

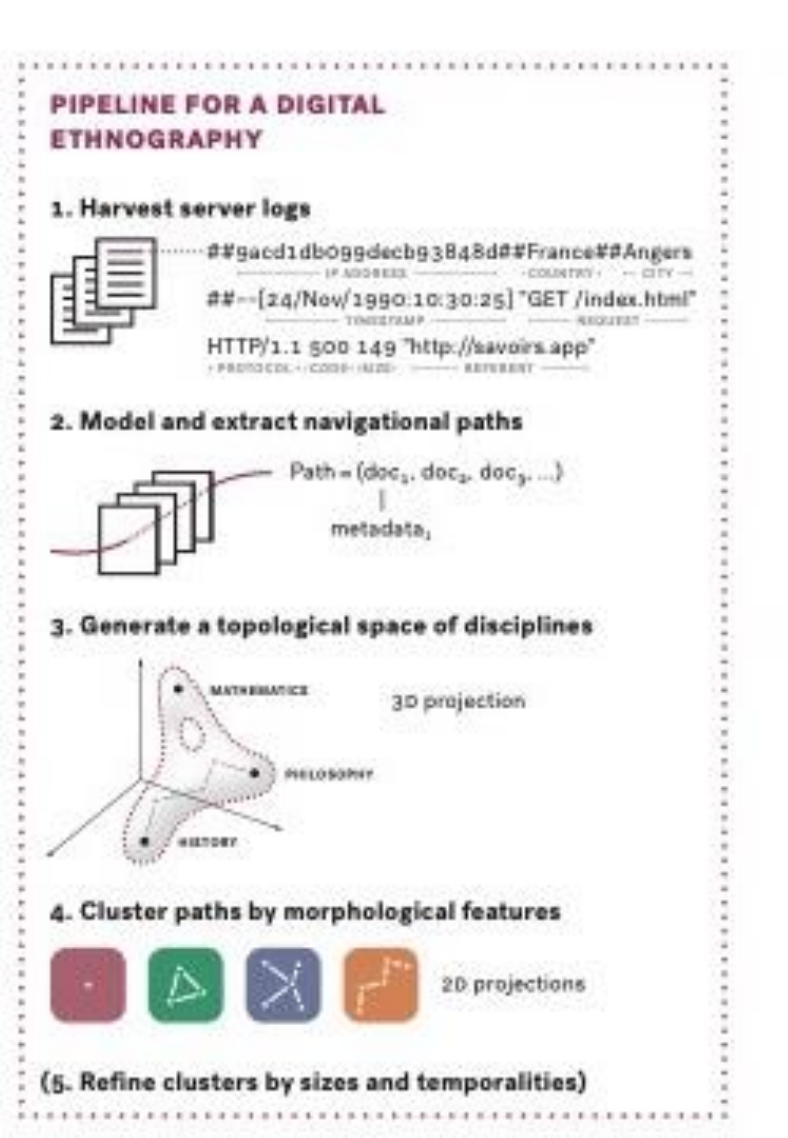
## Representation

A relative and relational space representing navigation practices, not a pre-existing territory

## Point of view

# 2. Users' *representation* of an informational space

Critique of pipeline



PIPELINE FOR A DIGITAL ETHNOGRAPHY

1. Harvest server logs

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

2. Model and extract navigational paths

Path = (doc$_1$, doc$_2$, doc$_3$, ...)

metadata$_i$

3. Generate a topological space of disciplines

4. Cluster paths by morphological features

(5. Refine clusters by sizes and temporalities)

Issues raised:

## Trace / Clue

Turn *machine-machine-interaction traces* into *clues* of human behaviour by carefully *selecting* and *structuring* them

## Models

Recover complex structures from sequential logs by modelling multiple behaviors recounted by users (cognitive mechanisms and interaction with interface)

## Representation

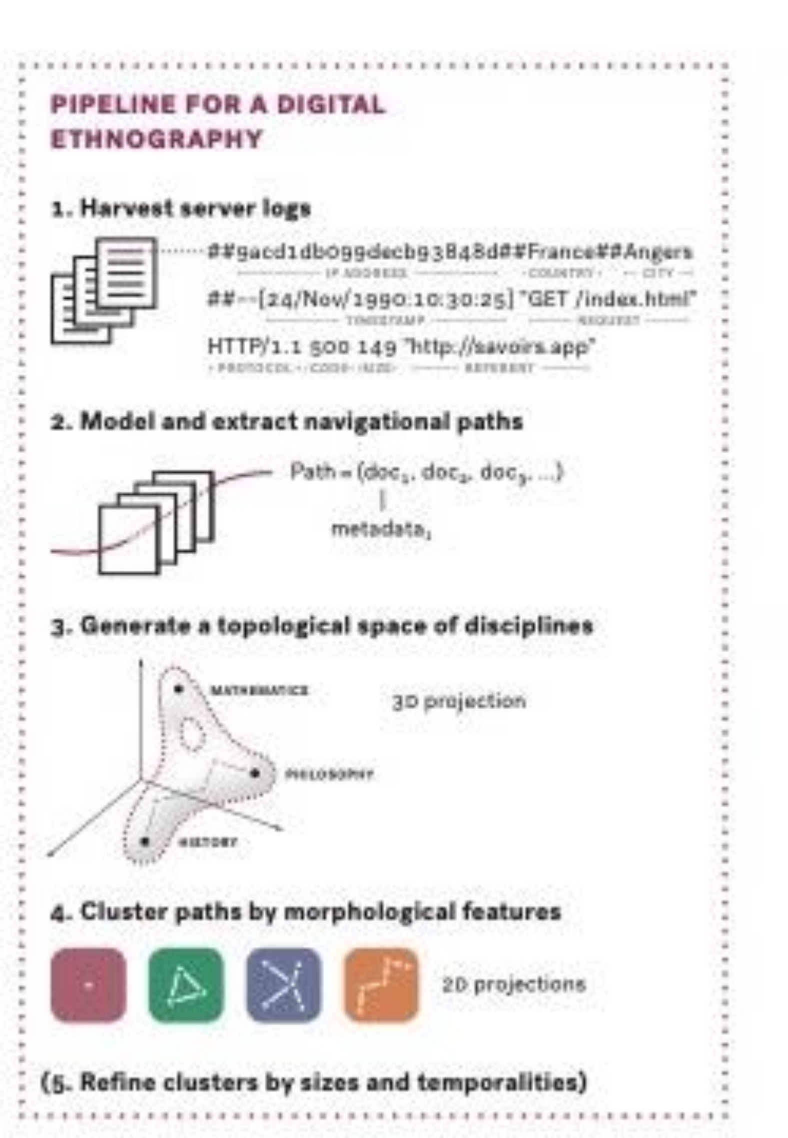
A relative and relational space representing navigation practices, not a pre-existing territory

## Point of view

Evaluate the effects of the projection method on the images produced

# 3. Workshop

**Preliminary step:** reconstruct your pipeline as a sequence of operations



PIPELINE FOR A DIGITAL ETHNOGRAPHY

1. Harvest server logs

```
##9acd1db099decb93848d##France##Angers
##--[24/Nov/1990:10:30:25] "GET /index.html"
HTTP/1.1 500 149 "http://savoirs.app"
```

2. Model and extract navigational paths

3. Generate a topological space of disciplines

4. Cluster paths by morphological features

(5. Refine clusters by sizes and temporalities)

## Trace / Clue

What is your dataset? Where does it come from? How was it produced? How is it formatted? What is NOT in the dataset?

## Models

What models/theories do you implicitly/explicitly use to make sense of your dataset? How do you justify this use? What are its limits?

## Representation

What kind of representation do you use (graph, historgram, vector space…)? For what reasons? What was lost/gained in dimensionality reduction/projection?

## Further improvements

How could you refine this approach? What is NOT shown? How could another approach (data-driven or not) complement this one?

# Thank you!

simon.dumas-primbault@openedition.org

OpenEdition Lab

# About Spaces

| Closed space | Open space | Guided space | Reticular space |
|---|---|---|---|
| Corpus exhaustively known | Unknown in detail | | |
| Perfectly indexed | Incomplete metadata | Subjective description | |
| Rigid (taxonomies) | Flexible (tags and folksonomies) | Adaptable (curation, recommendation) | |
| Findability (directed search) | Discoverability (browsing) | Guided navigation | Mobility along edges |
| Piece of information, one document | Ideas, unexpected references | | |
| Efficiency, relevance | Serendipity, diversity (comparatism, interdisciplinarity) | Sociabilities, transmission | |

# Publications, logiciels et exposition de données

## Publications

- Simon Dumas Primbault, « Naviguer dans les savoirs à l'ère numérique. Pour une ethnographie des pratiques informationnelles sur Gallica », *Études de communication*, no 61, 2023. **DOI ?**

  Étude socio-ethnographique quali-quanti complète des pratiques de navigation des usagers de Gallica (étude des parcours de lecture 2021-2022).

- Bayrem Kaabachi et Simon Dumas Primbault, « A Topological Data Analysis of Navigation Paths within Digital Libraries », *Computational Humanities Research*, CHR 2023. https://ceur-ws.org/Vol-3558/paper935.pdf

  Software paper de présentation détaillée du pipeline d'analyse quanti des logs serveurs (étude topologique 2021-2022).

- Simon Dumas Primbault, « La bibliothèque numérique comme espace à arpenter ? Retour réflexif sur un chassé-croisé méthodologique dans l'étude de la navigation dans un milieu documentaire numérique », à paraître 2024.

  Retour réflexif sur la méthodologie quali-quanti employée.

  Pre-print disponible sur demande.

# Publications, logiciels et exposition de données

## Logiciels et rapports de recherche

- Bayrem Kaabachi et Simon Dumas Primbault, TDA-Gallica, *Github*, 2022, https://github.com/Kaabachi/TDA-Gallica

  Analyse topologique des logs (2021-2022)

- Mohamed Aziz Ben Chaabane, Ahmed Nour Achiche  et Simon Dumas Primbault, gallica-seq-mining, *Github*, 2023, https://github.com/LHST-EPFL/gallica-seq-mining

  Analyse des logs en séquences d'actions (2023)

## Poster

- Simon Dumas Primbault, « How to Orient Oneself in Thinking in the Digital Era: A mixed-methods ethnography of researchers' navigational practices on Gallica », DARIAH-CH Study Day, Oct 2022, Mendrisio, Suisse. https://hal.science/hal-03840530

# Publications, logiciels et exposition de données

## Exposition de données

- Simon Dumas Primbault, « Transcriptions de sept entretiens semi-directifs réalisés avec des usagers de Gallica au sujet de leurs pratiques informationnelles » [Survey data] NAKALA. https://doi.org/10.34847/nkl.320eu55q

  Transcriptions anonymisées des 7 entretiens de l'étude 2021-2022, avec trame d'entretien et formulaire d'information.