Modéliser les parcours de recherche en SHS

Une analyse quantitative des logs de visites sur Isidore.

Huma-Num

#RencontresHumanum

Pierre-Carl Langlais

ANR Numapresse

These : theses.fr/s96560

Twitter : @Dorialexander

Wikipédia : User:Alexander Doria

Github : github.com/Dorialexander

Hypotheses : http://scoms.hypotheses.org/

Les conditions de l’étude.

Présentation en trois temps.

- **Le public d’Isidore**
- **Cartographier les mots-clés**
- **Modéliser les requêtes**

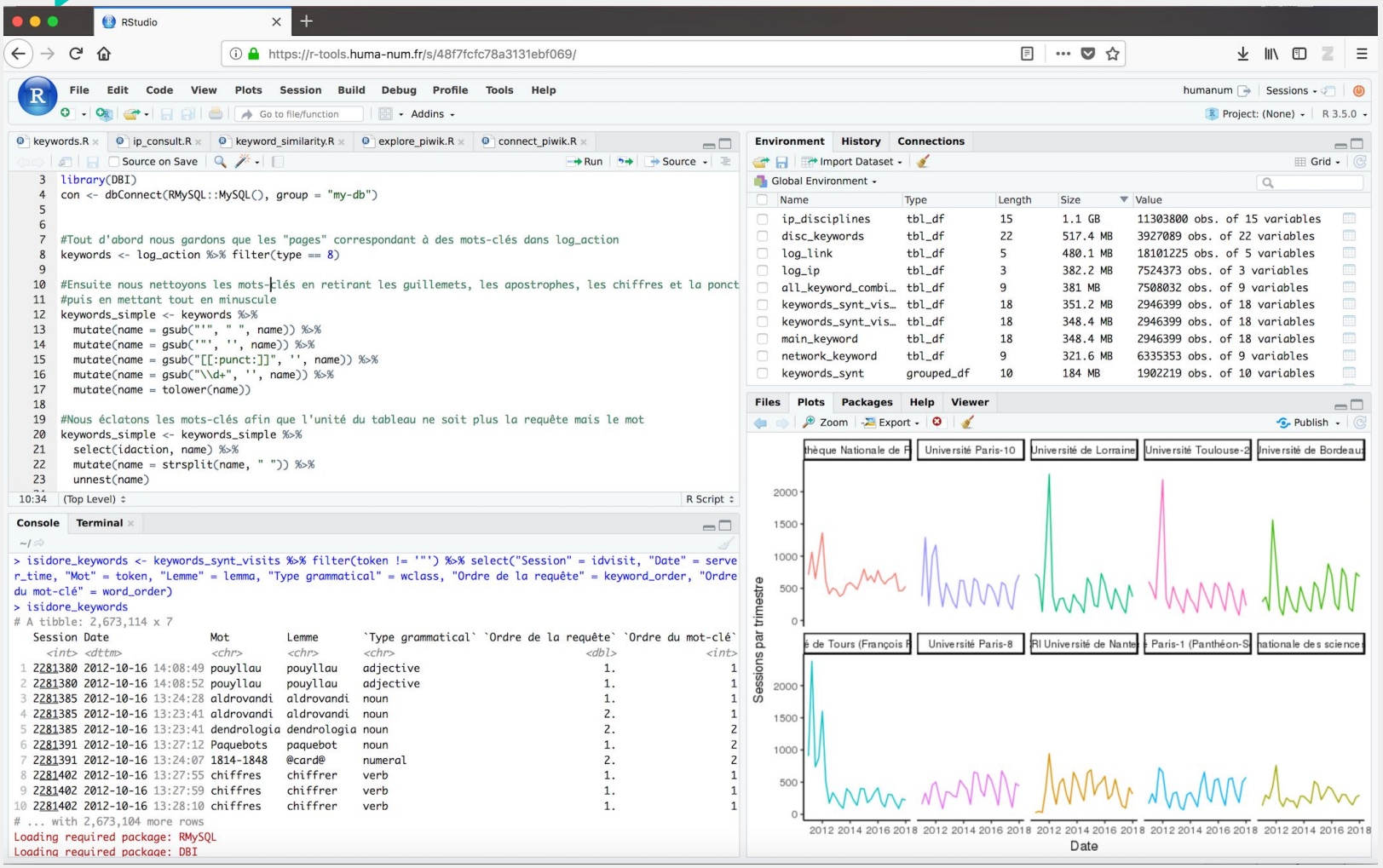
…et un prologue : **Les coulisses de l’analyse**

Les coulisses de l'analyse

Une étude entreprise en avril 2018 :

- **Base de données de 10 gos de logs de connexion** couvrant plus de 7 ans d'usages sur Isidore (janvier 2011-avril 2018)
- **Approche quantitative des pratiques de lecture et de consultation**, entre SHS et *data science* centrée notamment sur les 800 000 requêtes en mots-clés déposés par les utilisateurs d'Isidore.
- **Travail mené sur l'infrastructure d'Humanum**, avec l'interface serveur de R Studio
- **Analyse de données non personnelles** en particulier les adresses IP collectives des universités.

Les coulisses de l'analyse



L'interface idéale du « small big data » : R Studio

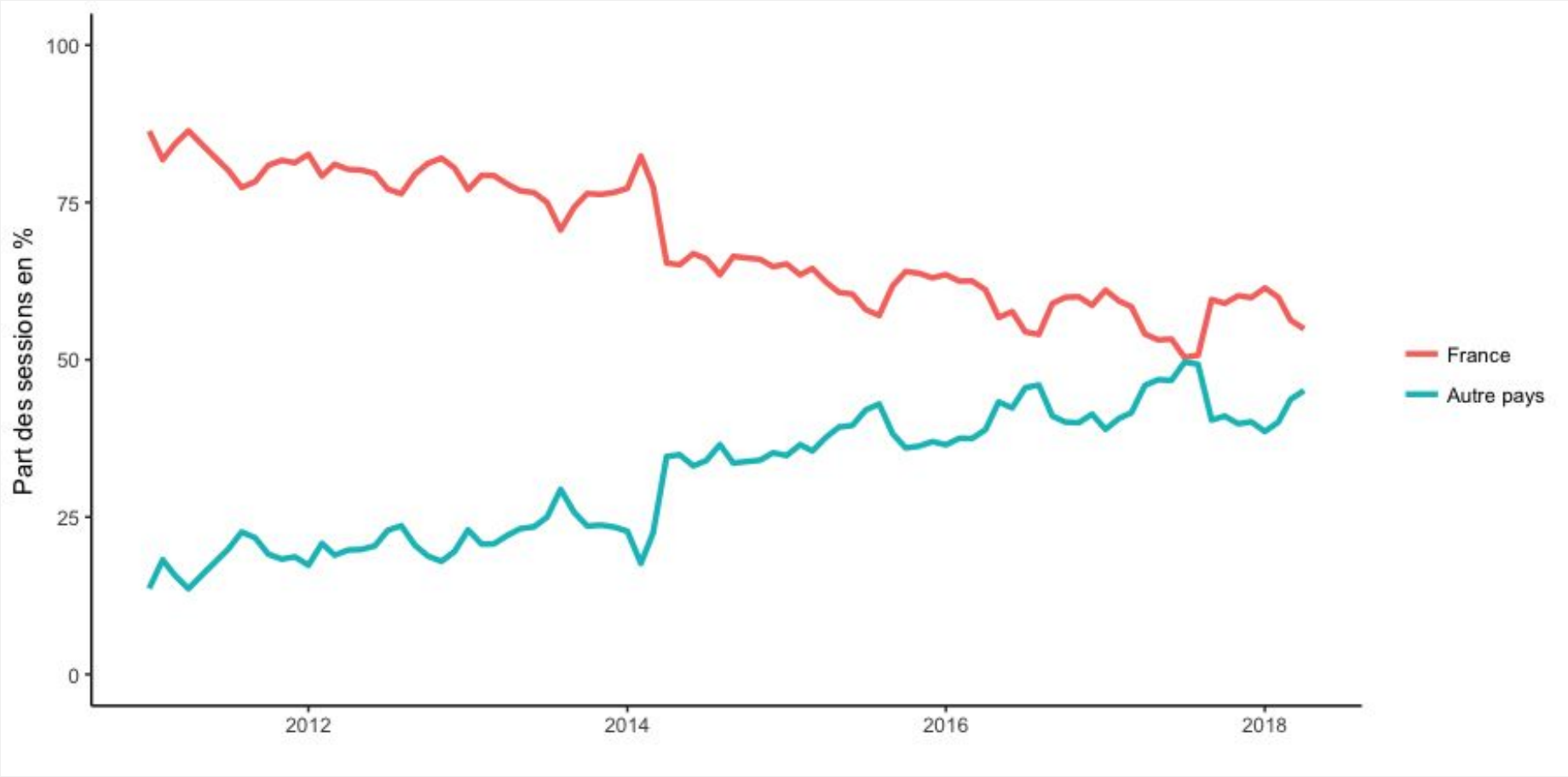
Les coulisses de l'analyse

```
> isidore_keywords <- keywords_synt_visits %>% filter(token != '"') %>% select("Session" = idvisit, "Date" = server_time, "M
ot" = token, "Lemme" = lemma, "Type grammatical" = wclass, "Ordre de la requête" = keyword_order, "Ordre du mot-clé" = word_
order)
> isidore_keywords
# A tibble: 2,673,114 x 7
   Session Date                Mot         Lemme       `Type grammatical` `Ordre de la requête` `Ordre du mot-clé`
     <int> <dttm>              <chr>       <chr>       <chr>                              <dbl>               <int>
 1 2281380 2012-10-16 14:08:49 pouyllau    pouyllau    adjective                             1.                   1
 2 2281380 2012-10-16 14:08:52 pouyllau    pouyllau    adjective                             1.                   1
 3 2281385 2012-10-16 13:24:28 aldrovandi  aldrovandi  noun                                  1.                   1
 4 2281385 2012-10-16 13:23:41 aldrovandi  aldrovandi  noun                                  2.                   1
 5 2281385 2012-10-16 13:23:41 dendrologia dendrologia noun                                  2.                   2
 6 2281391 2012-10-16 13:27:12 Paquebots   paquebot    noun                                  1.                   2
 7 2281391 2012-10-16 13:24:07 1814-1848   @card@      numeral                               2.                   2
 8 2281402 2012-10-16 13:27:55 chiffres    chiffrer    verb                                  1.                   1
 9 2281402 2012-10-16 13:27:59 chiffres    chiffrer    verb                                  1.                   1
10 2281402 2012-10-16 13:28:10 chiffres    chiffrer    verb                                  1.                   1
# ... with 2,673,104 more rows
>
```

Pour être analysables, les mots-clés doivent être soumis à un « nettoyage » en profondeur : retrait des mots-outils, lemmatisation, etc.

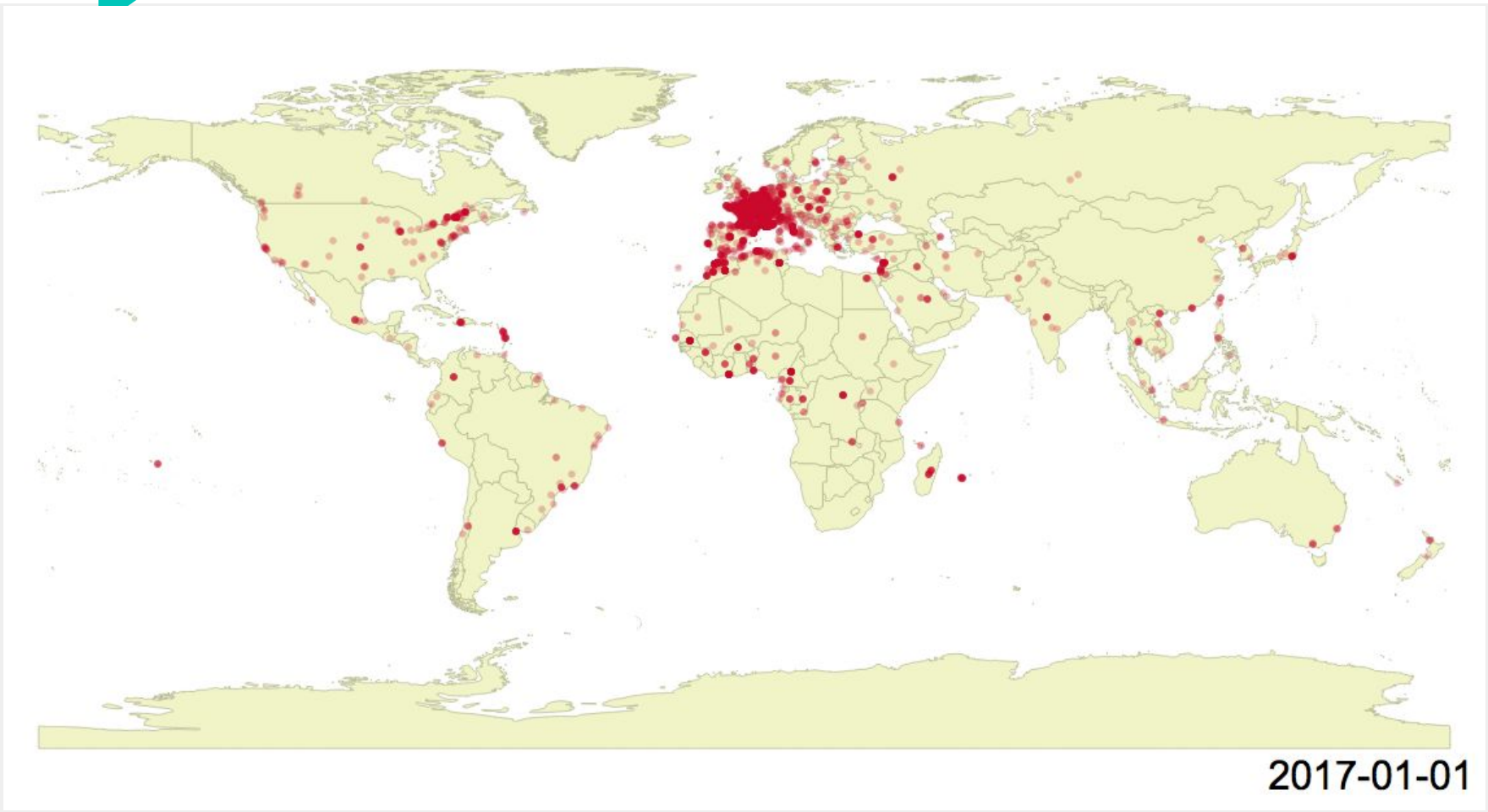
1.
Le public d’Isidore

Une audience internationale



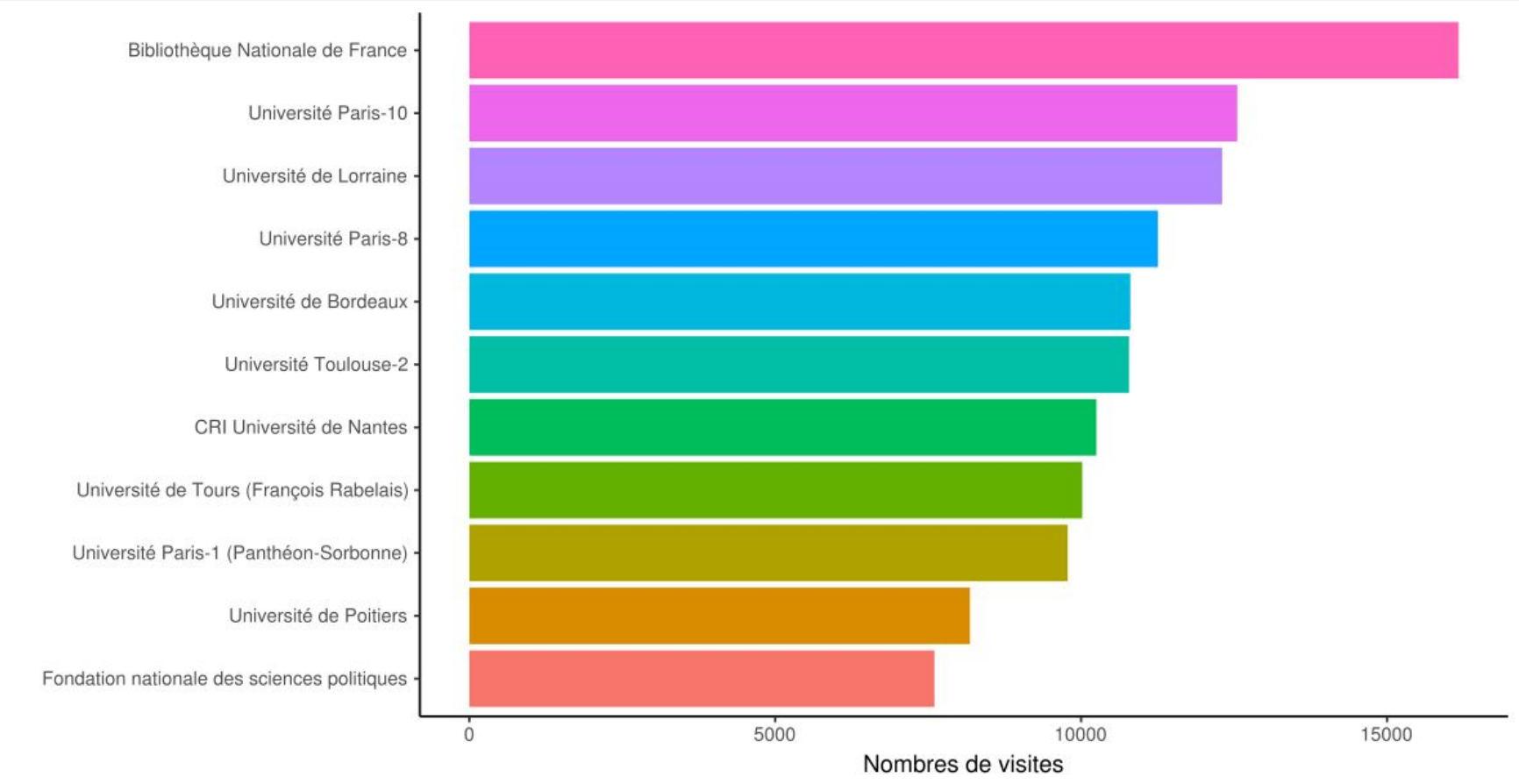
La dynamique d'internationalisation remonte à la mise en place d'une interface anglophone puis multilingue

Une audience internationale



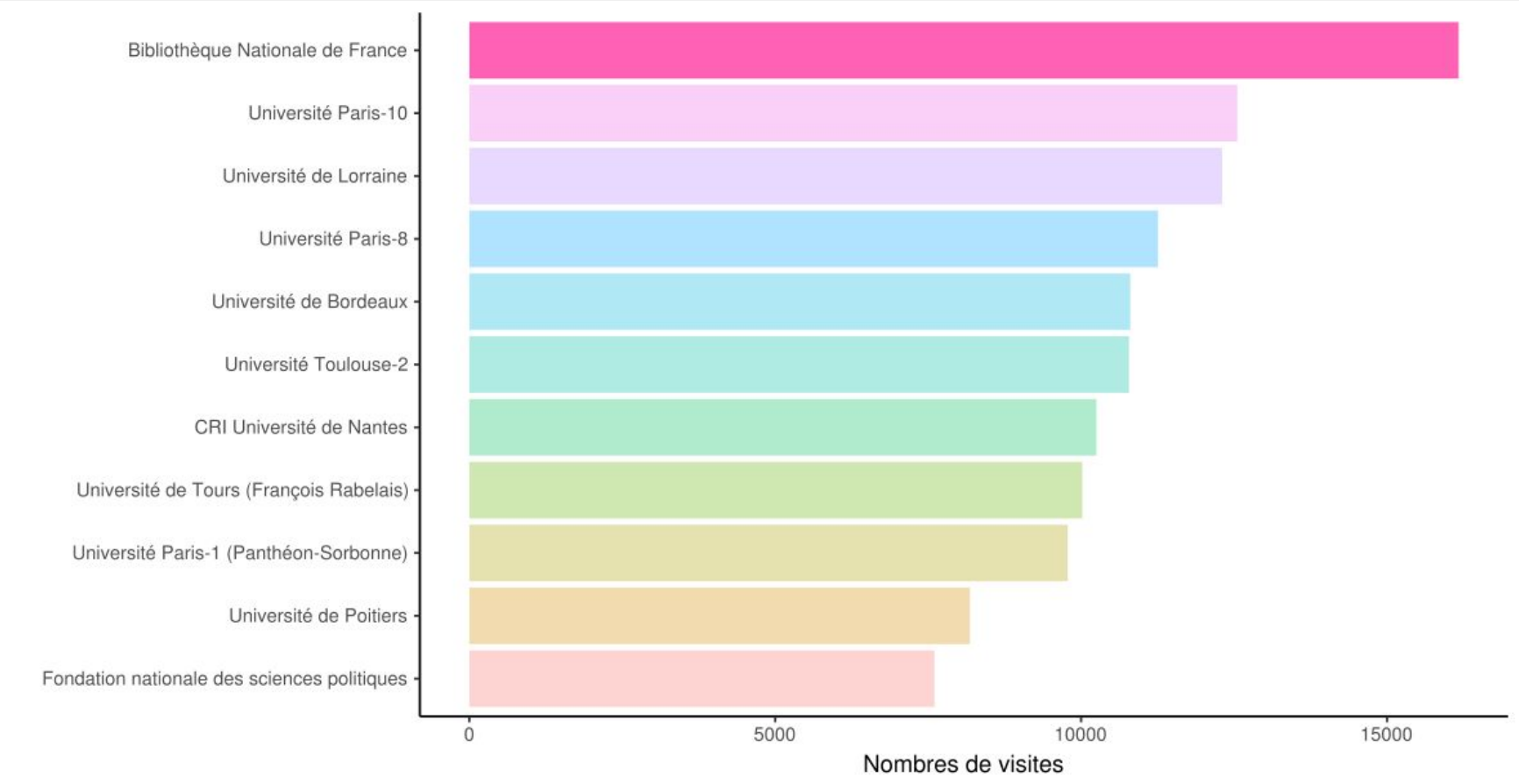
Un an d’Isidore en version animée : des internautes très présents en Europe, mais aussi en Amérique du nord et en Afrique.

Un public spécialisé



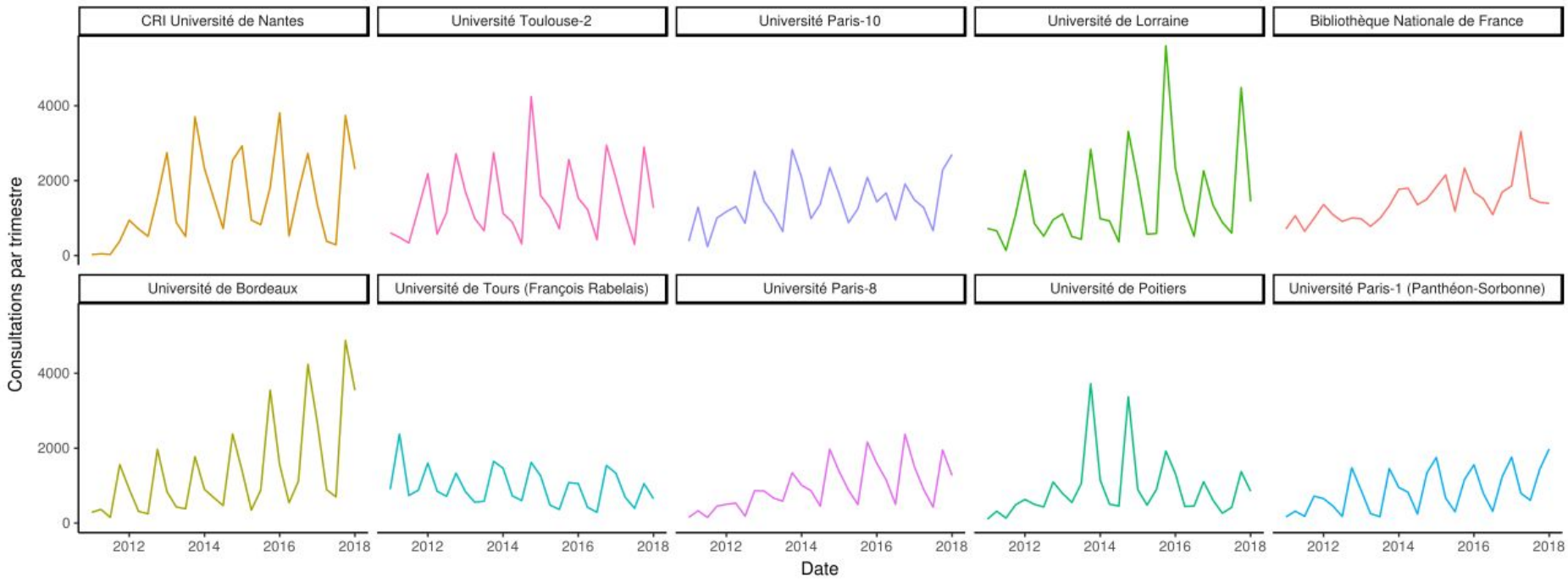
Au moins une consultation sur 10 provient d'une institution universitaire. La principale source est… la BNF.

Un public spécialisé



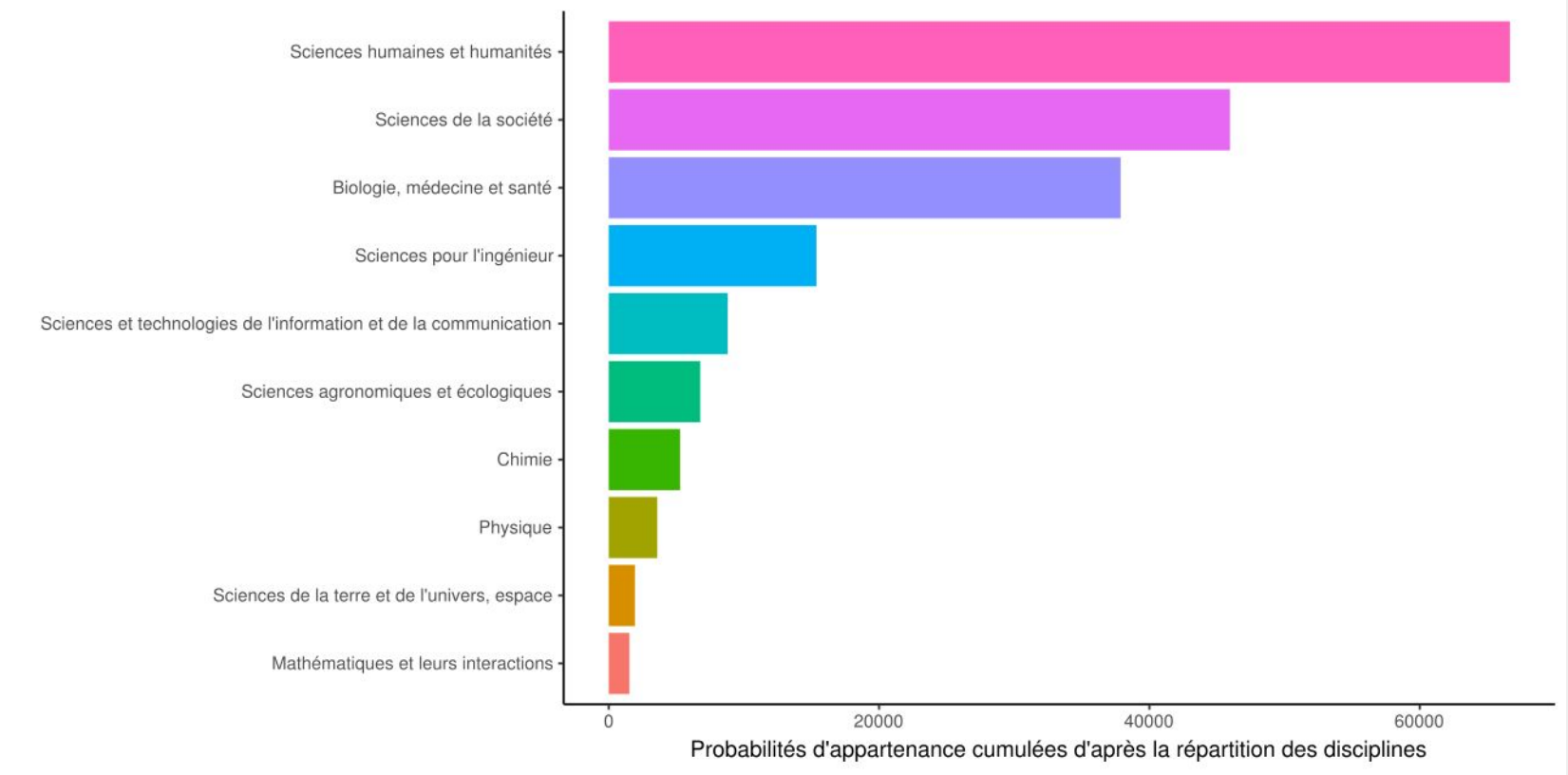
Au moins une consultation sur 10 provient d'une institution universitaire. La principale source est… la BNF.

Un public spécialisé



Les consultations varient significativement, notamment à la faveur du cycle universitaire sauf… à la BNF

Un public spécialisé

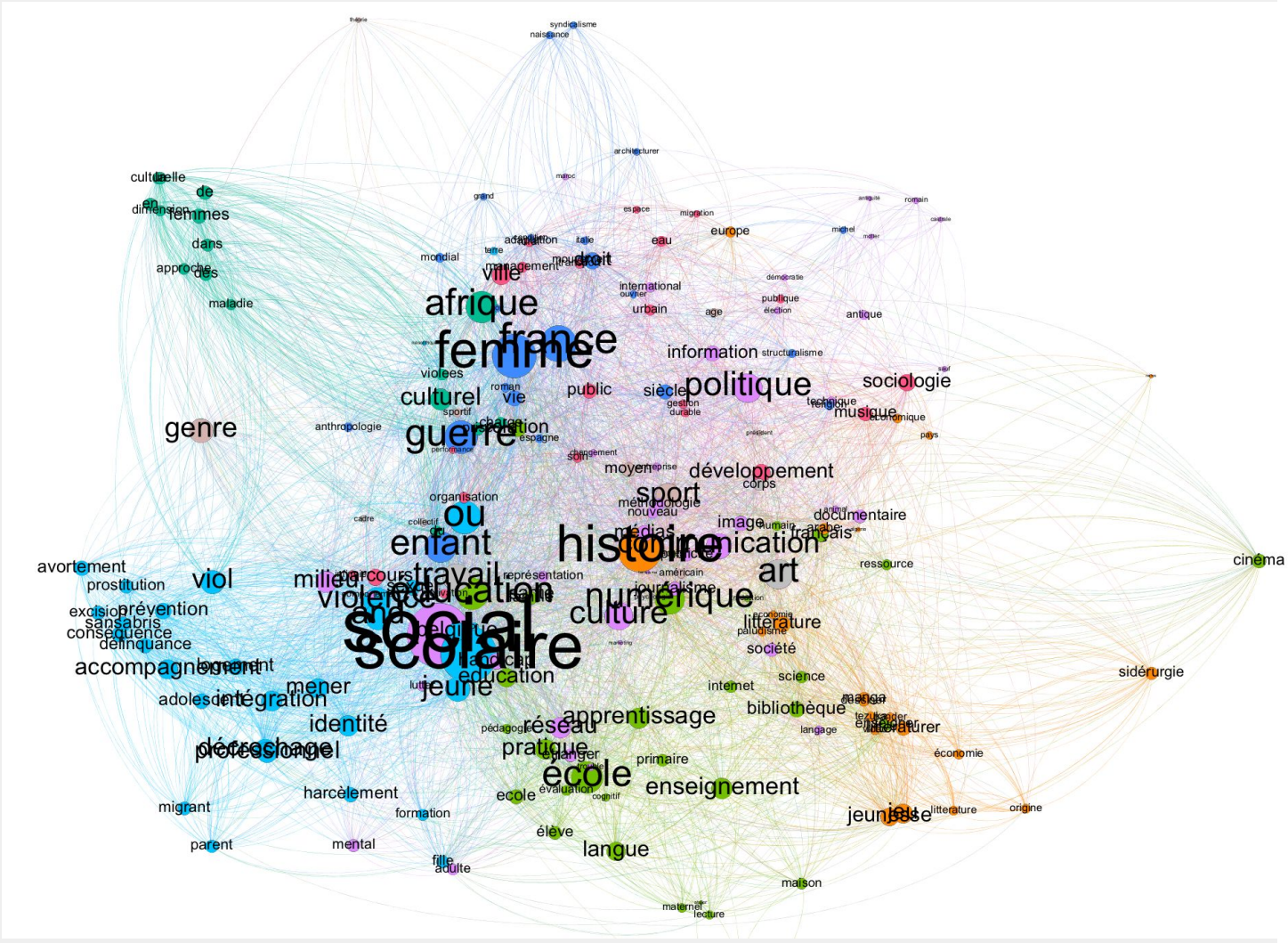


Les données de provenance des institutions universitaires permettent également par contrecoup de repérer l'environnement disciplinaire des consultations

2.
Cartographier les mots-clés

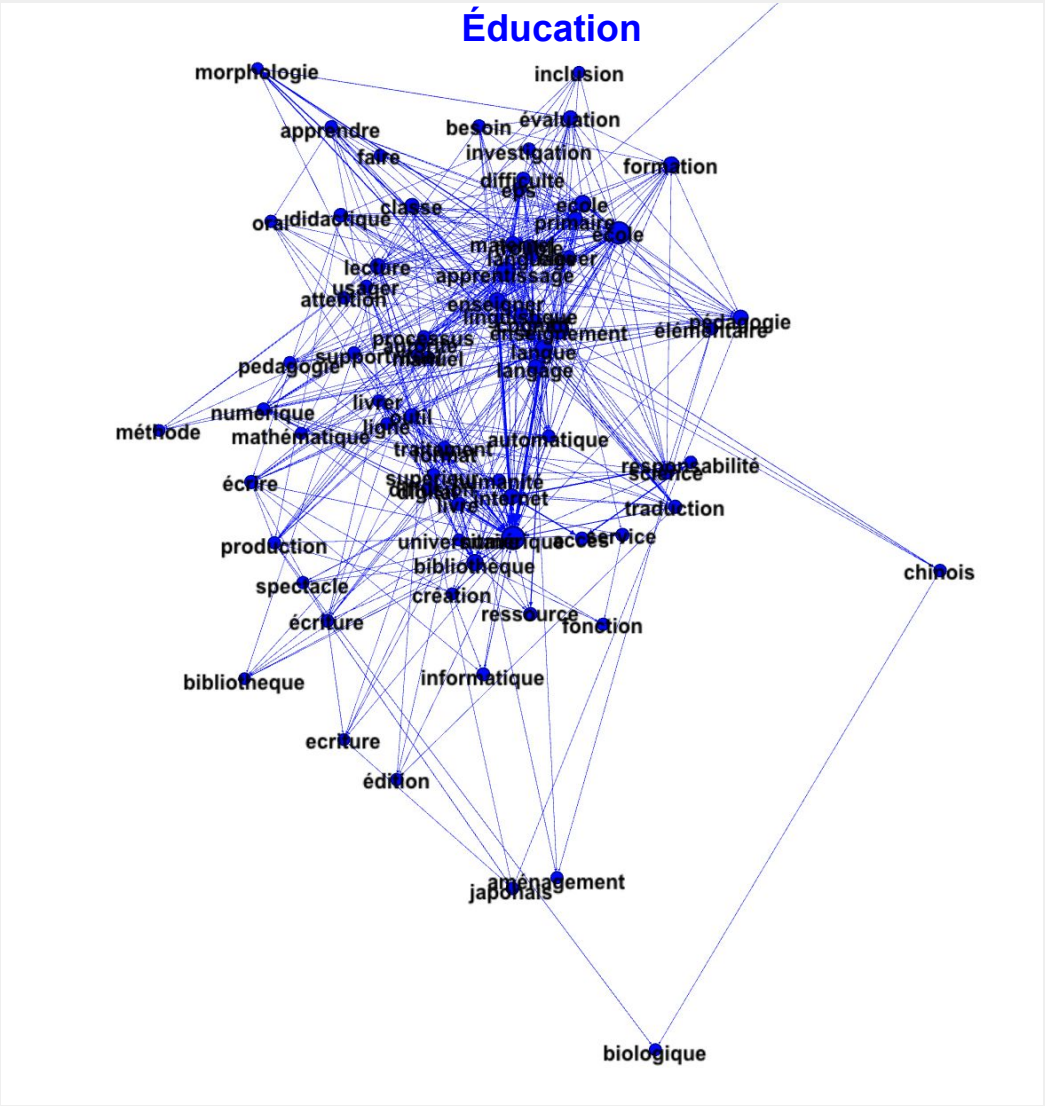
Des réseaux de mots-clés.

Les liens entre les mots au sein d'une seule requête s'inscrivent des réseaux sémantiques et disciplinaires sous-jacents.



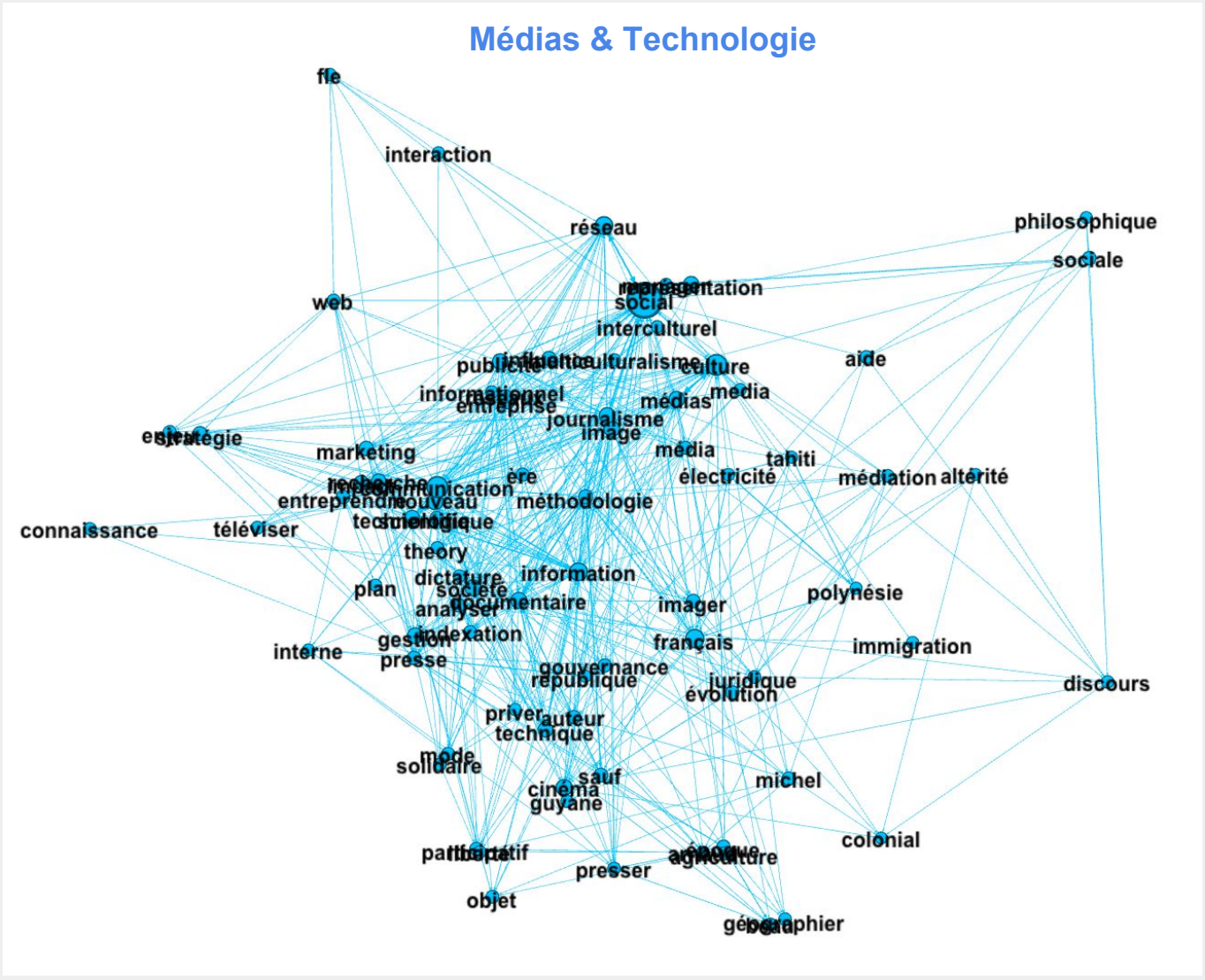
Du réseau à la classification

Chaque région du réseau correspond à des ensembles thématiques plus ou moins cohérent que nous pouvons « baptiser » a posteriori



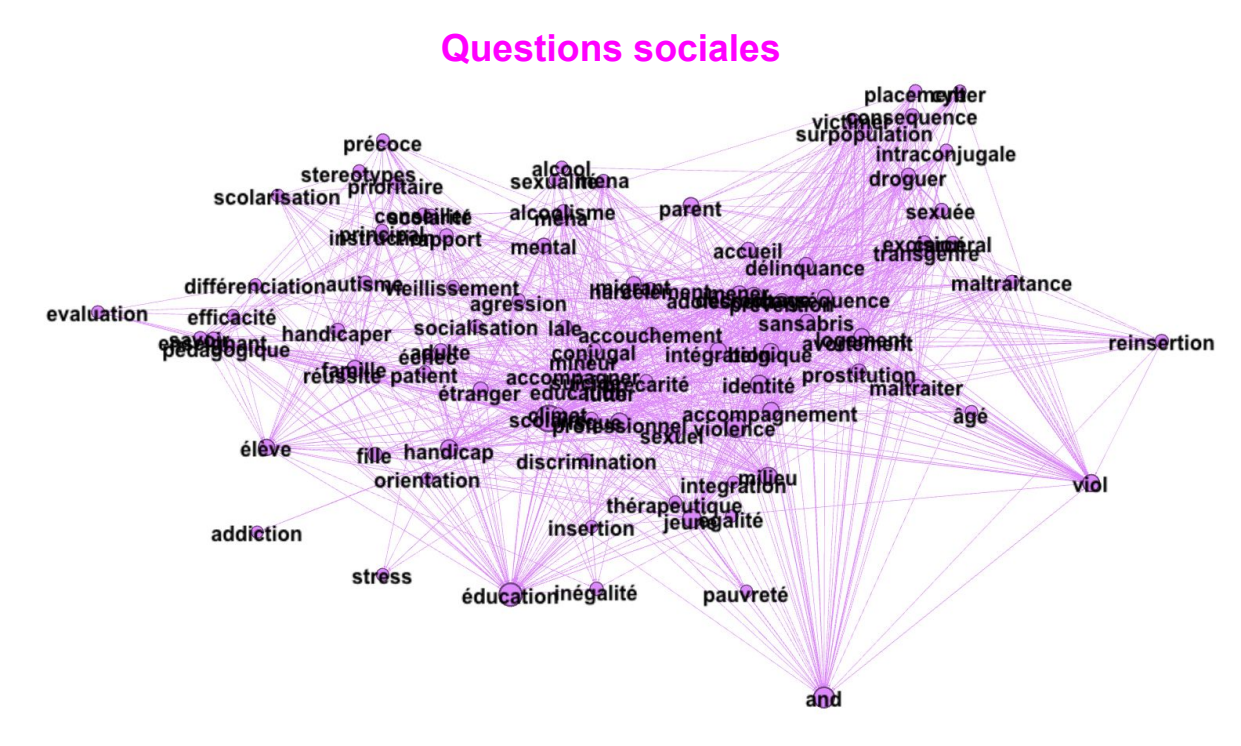
Du réseau à la classification

Chaque région du réseau correspond à des ensembles thématiques plus ou moins cohérent que nous pouvons « baptiser » a posteriori

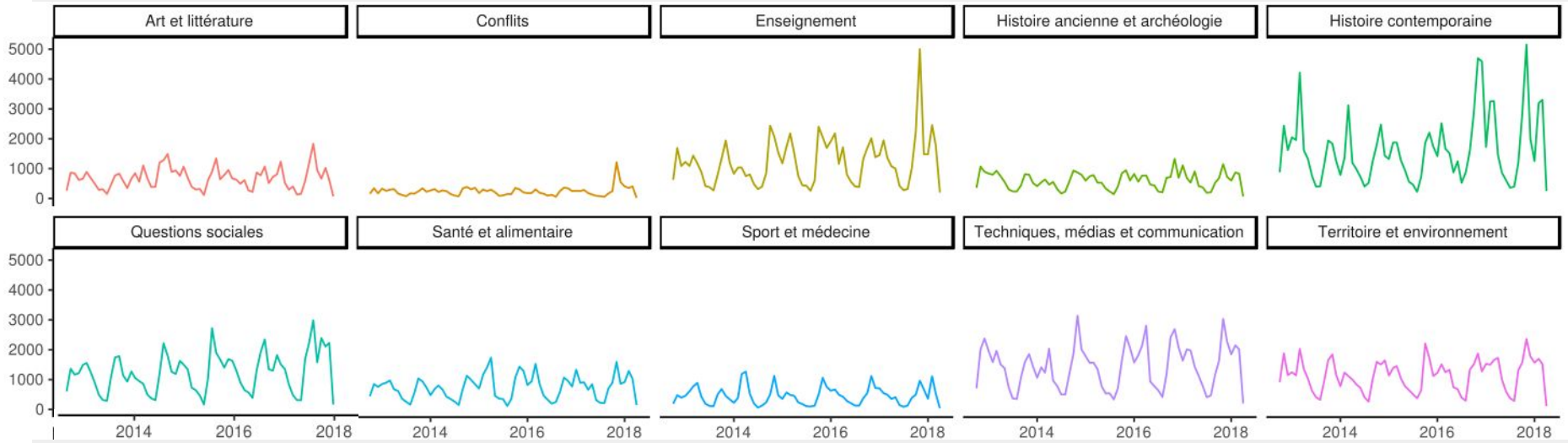


Du réseau à la classification

Chaque région du réseau correspond à des ensembles thématiques plus ou moins cohérent que nous pouvons « baptiser » a posteriori

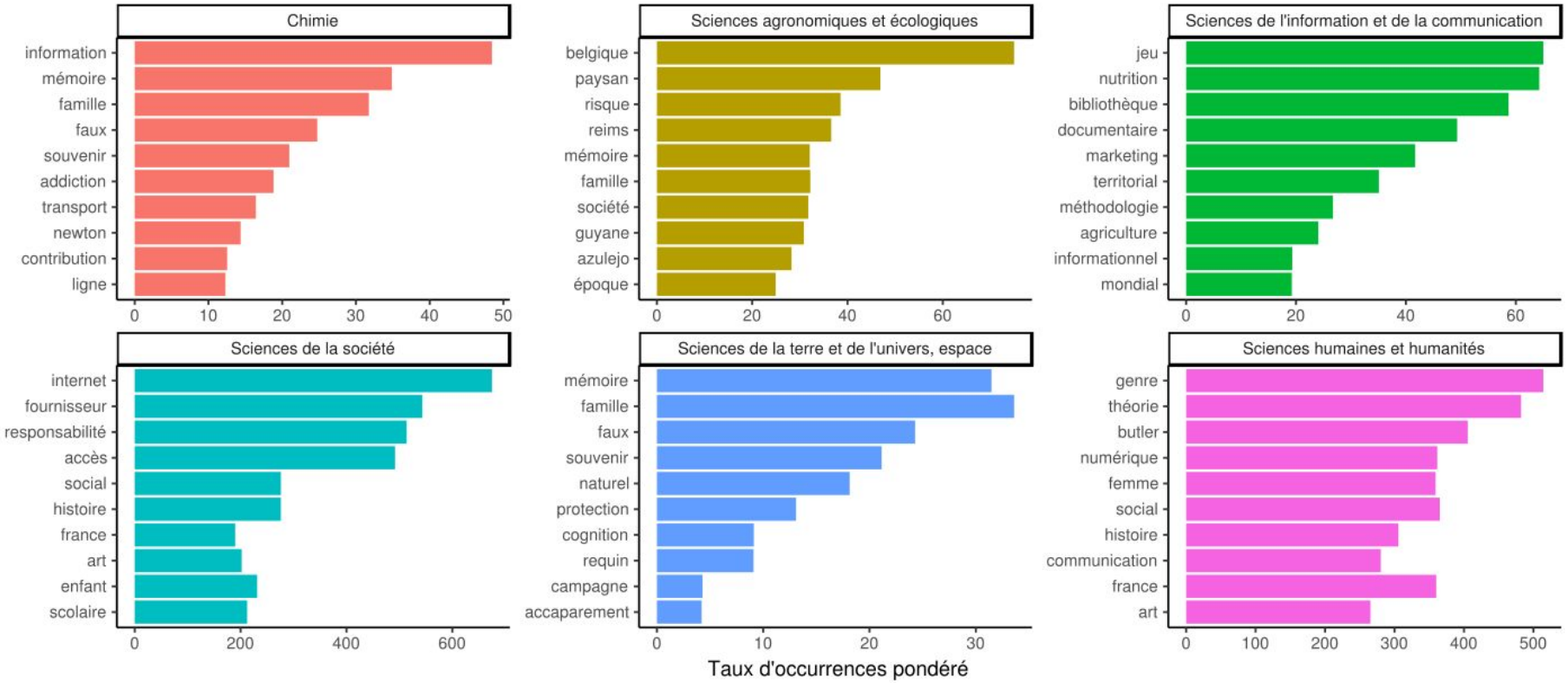


Du réseau à la classification.



La classification des mots clés constitue un moyen approximatif de repérer l'évolution des intérêts du public d'Isidore.

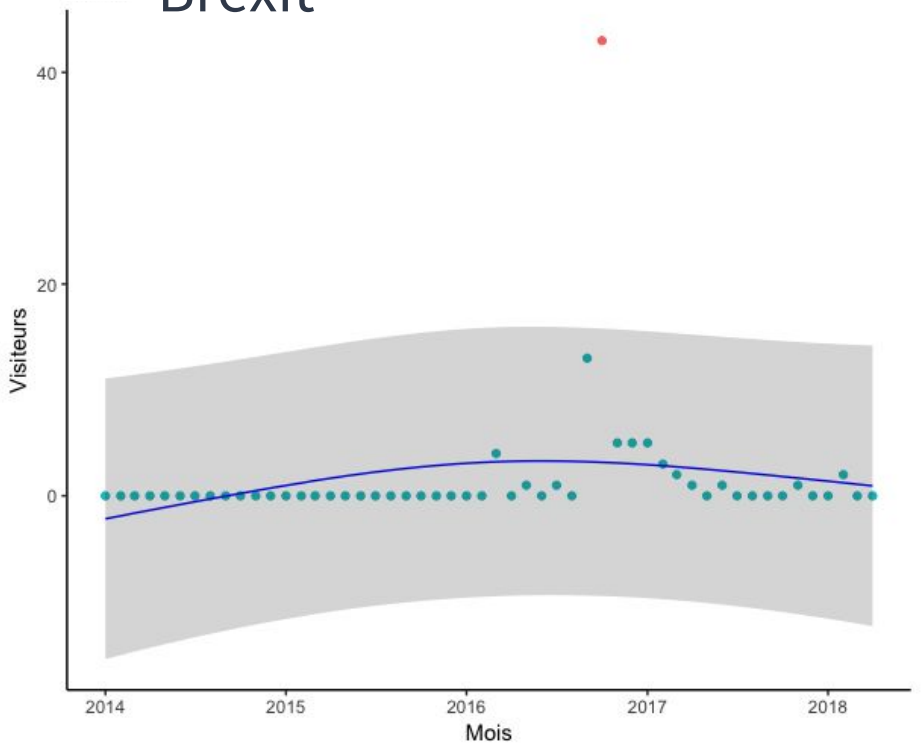
Du réseau à la classification.



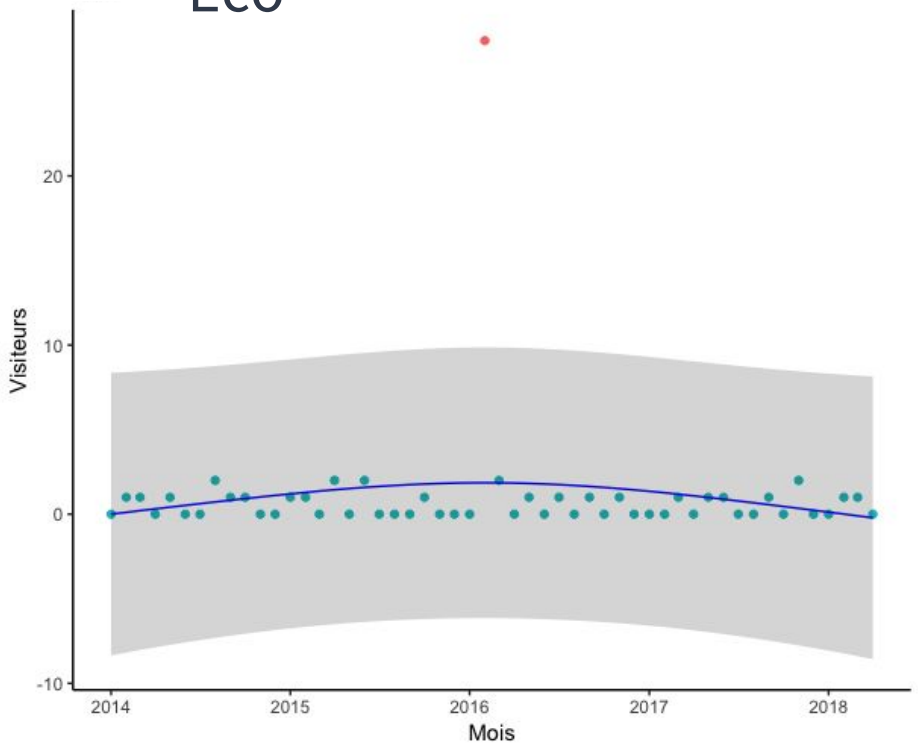
Une autre manière d’observer les « mots » des disciplines : croiser avec les données de provenances pour les visites universitaires

Des anomalies

Brexit

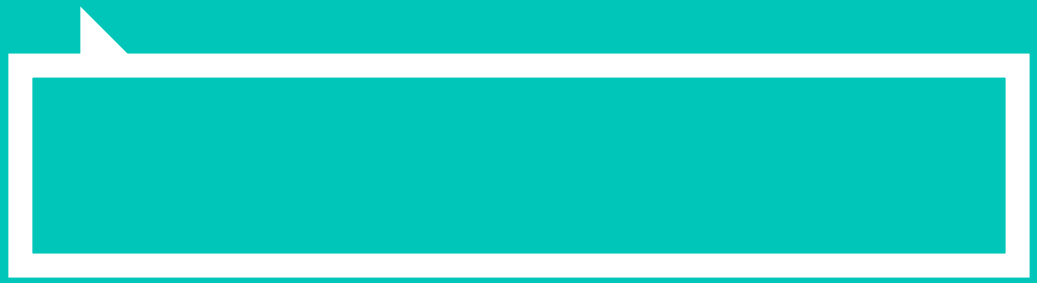


Eco

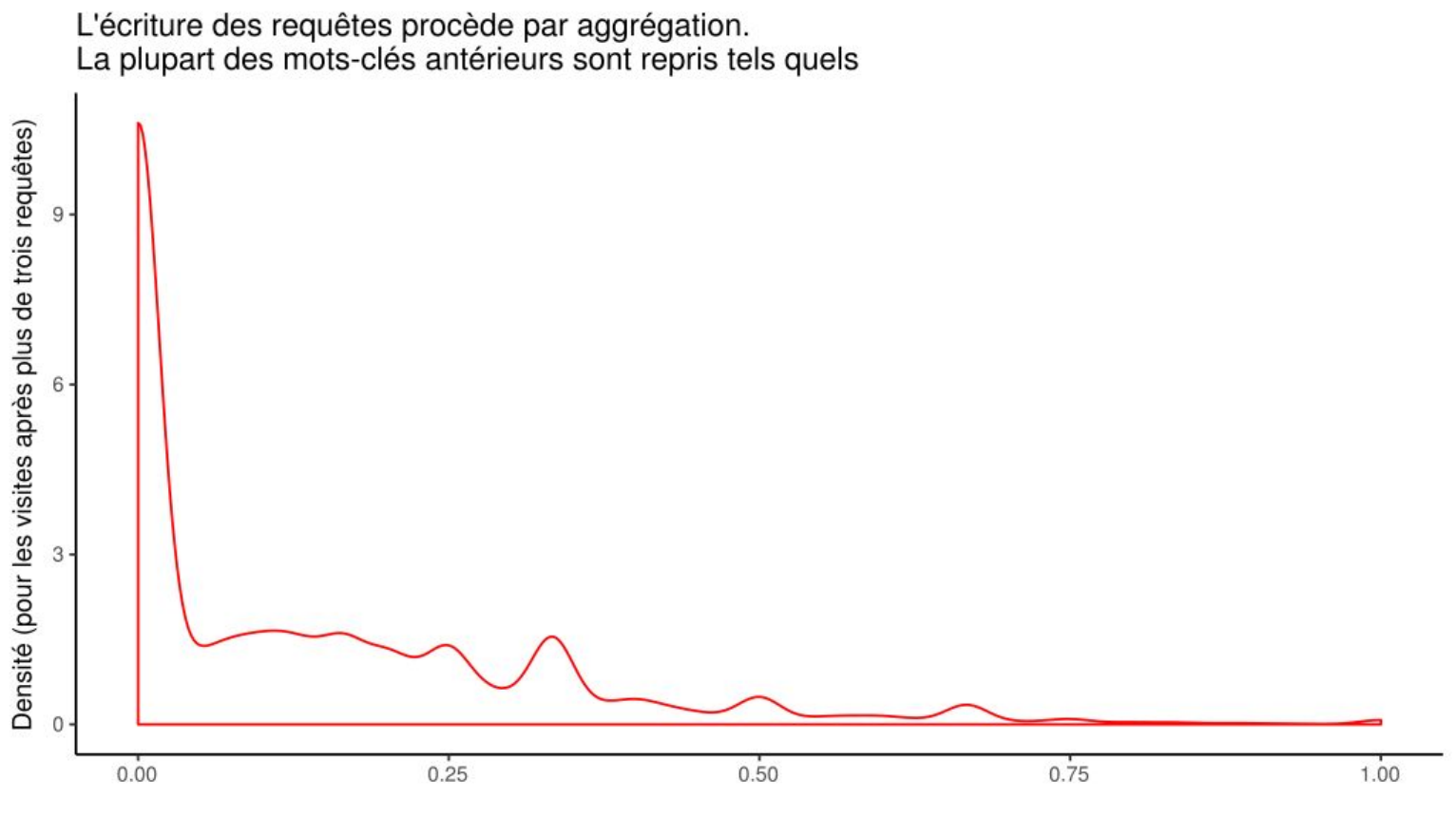


Au-delà des tendances structurelles : des effets de mode soudain pour certains mots.

3.
Analyser les requêtes

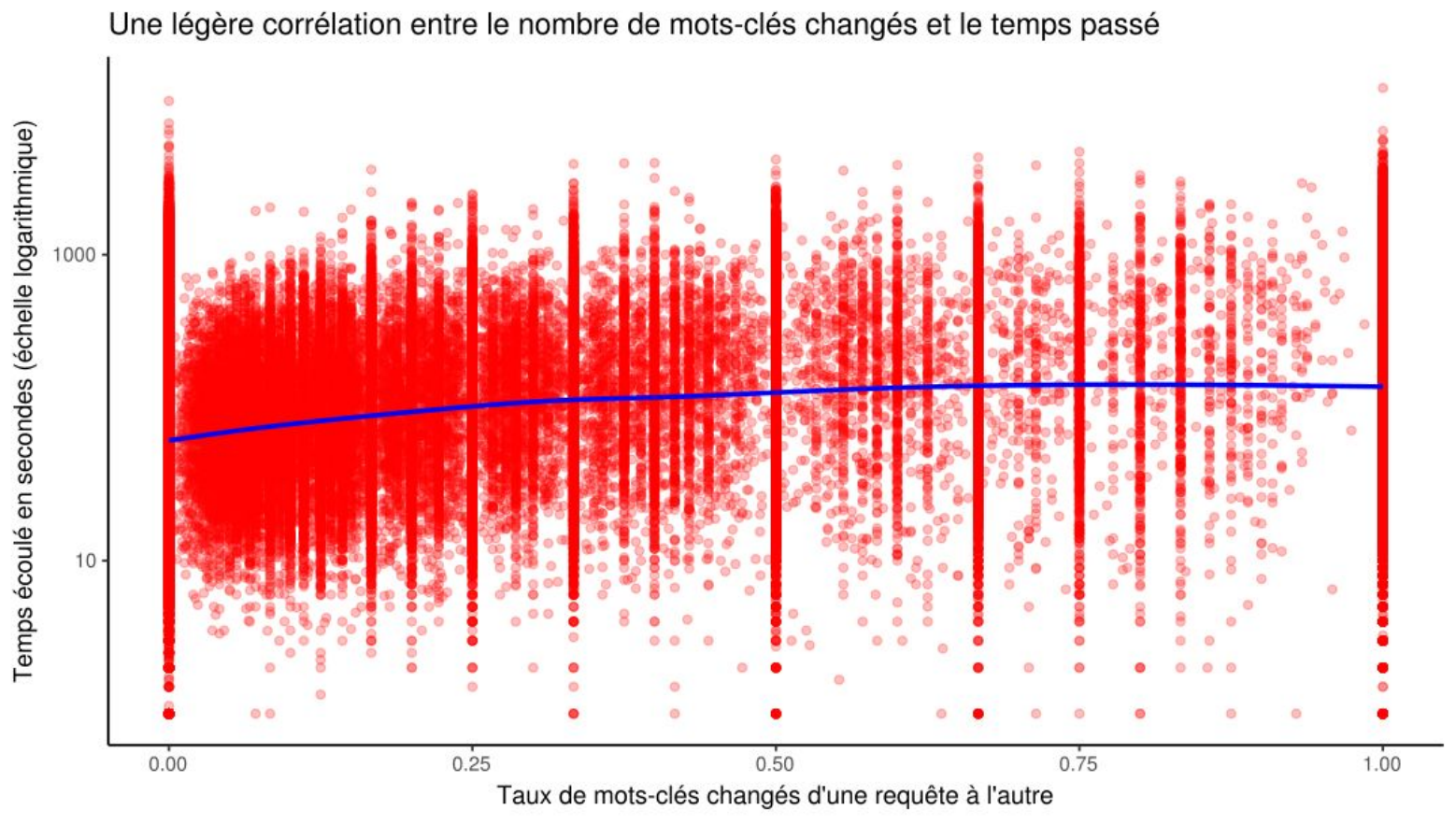


Un art de la reprise



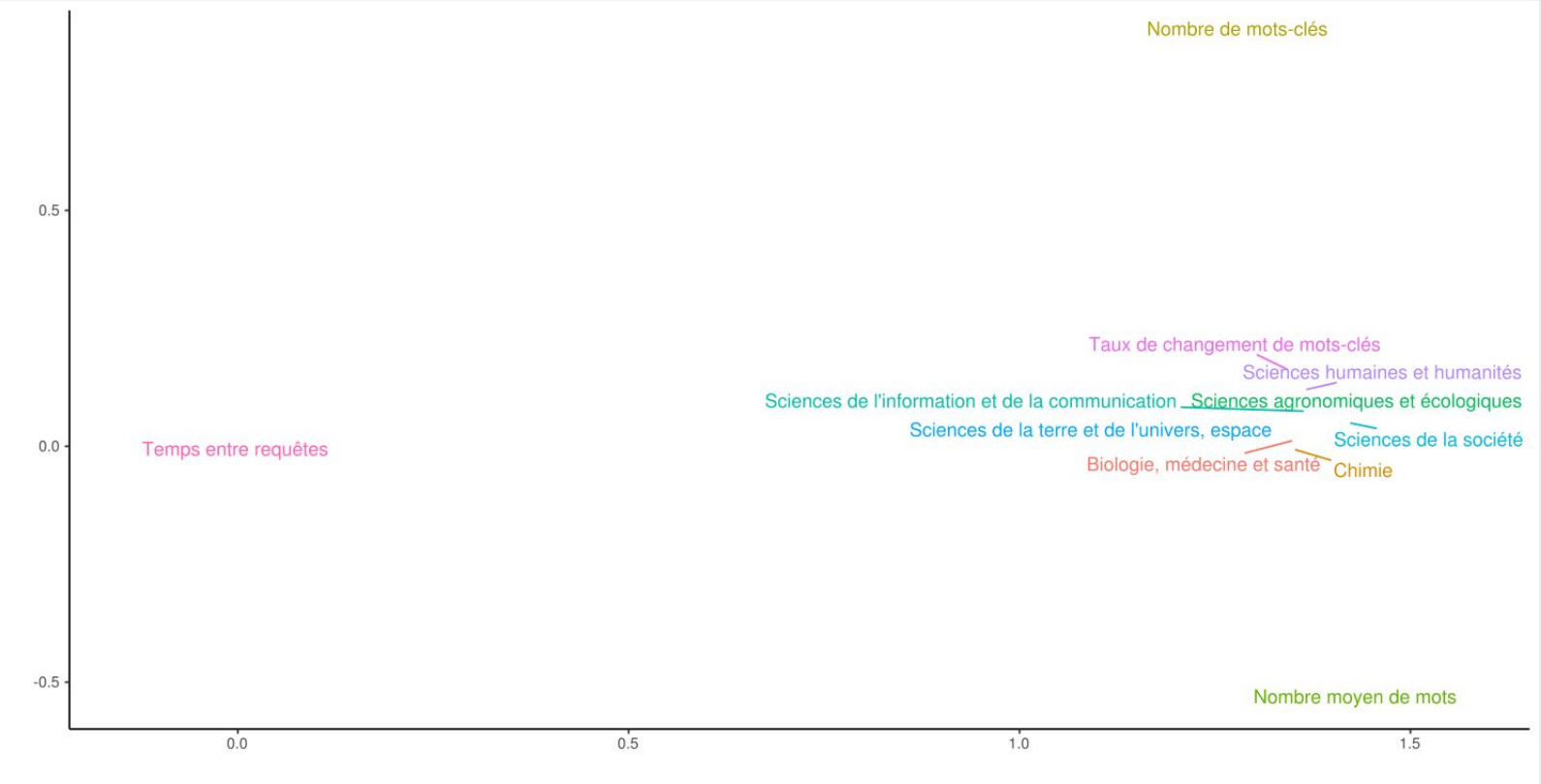
Notre indicateur montre que la plupart des requêtes sont des reprises partielles de requêtes antérieures.

Du réseau à la classification.



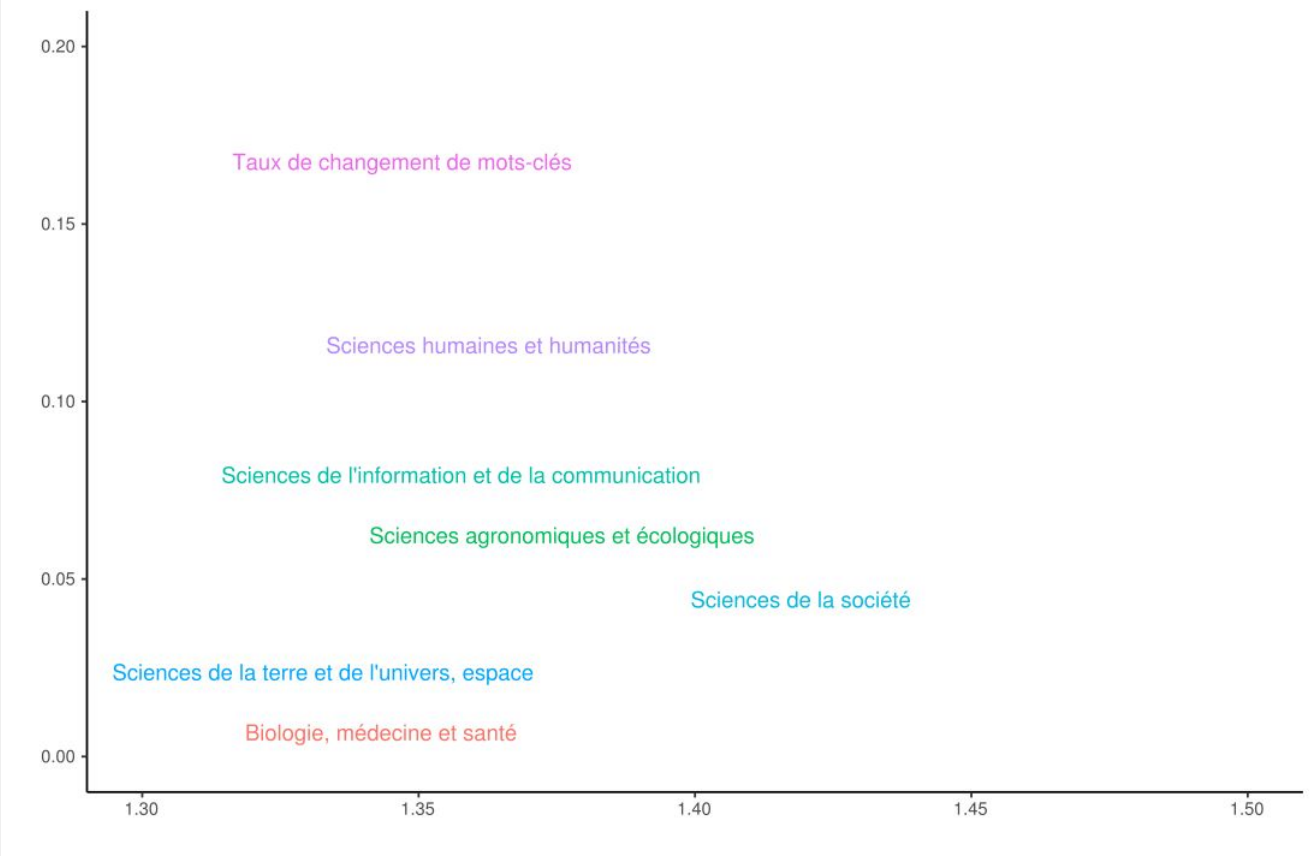
Assez logiquement, les requêtes avec des changements de vocabulaire fréquents prennent un peu plus de temps.

Du réseau à la classification.



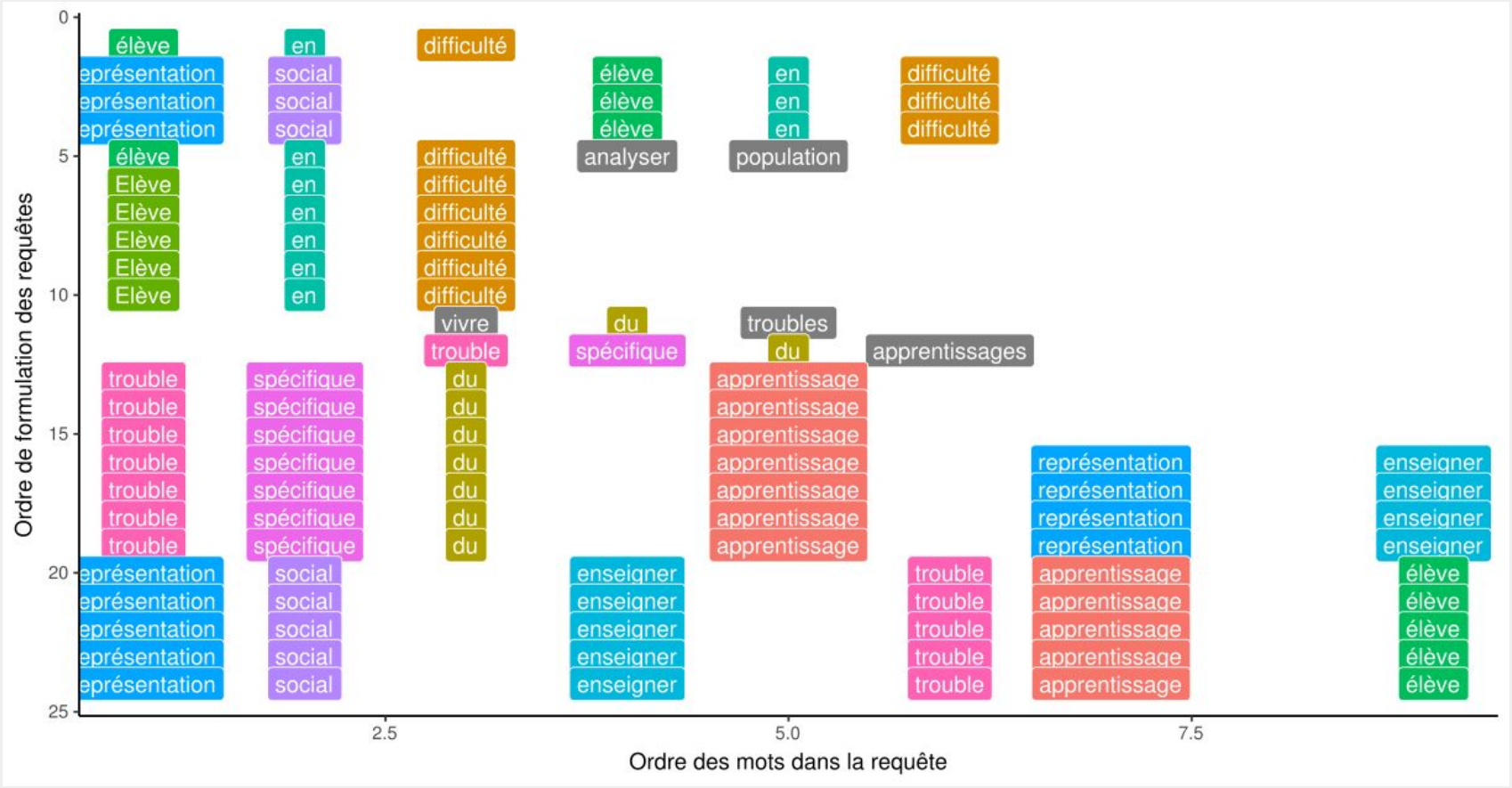
Ces pratiques se différencient (légèrement) selon les disciplines : les Sciences humaines et humanités tendent à varier davantage les mots-clés d'une requête à l'autre

Du réseau à la classification.



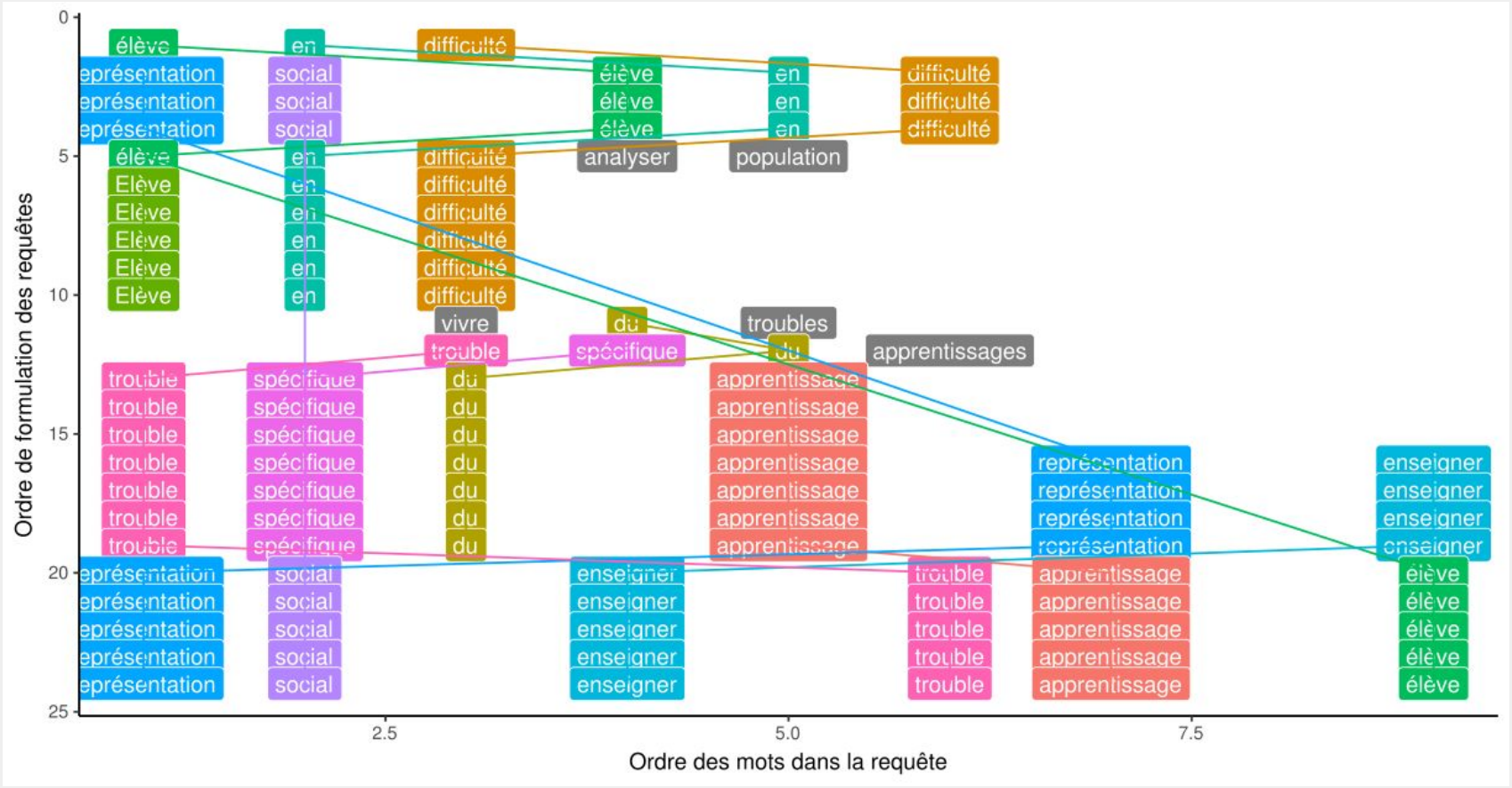
Ces pratiques se différencient (légèrement) selon les disciplines : les Sciences humaines et humanités tendent à varier davantage les mots-clés d'une requête à l'autre

Du réseau à la classification.



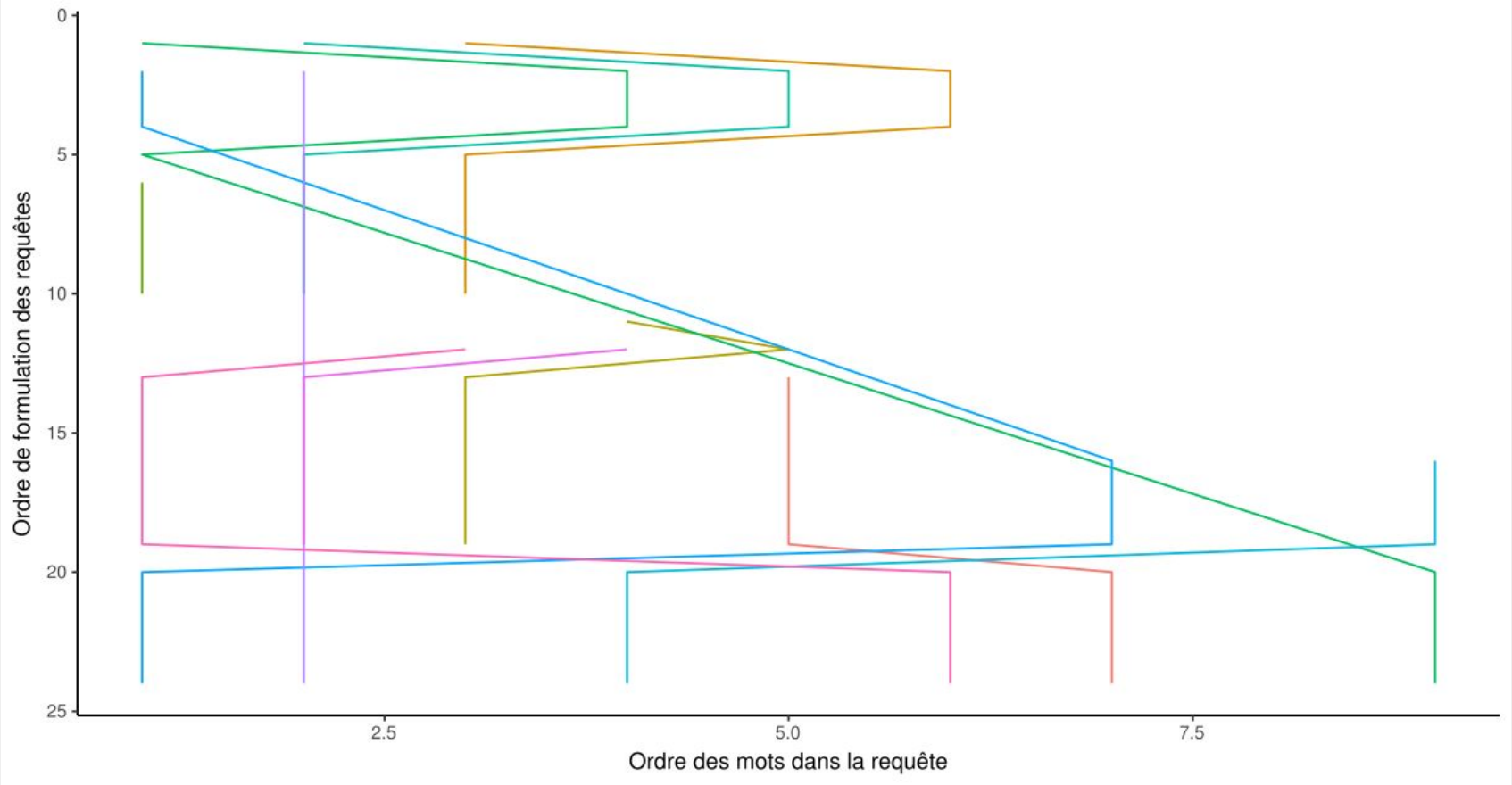
Ce que l'on n'arrive pas à mesurer / modéliser : les opérations d'aggrégations et de re-combination d'une requête à l'autre.

Du réseau à la classification.



Ce que l'on n'arrive pas à mesurer / modéliser : les opérations d'aggrégations et de re-combination d'une requête à l'autre.

Du réseau à la classification.



Ce que l’on n’arrive pas à mesurer / modéliser : les opérations d’aggrégations et de re-combination d’une requête à l’autre.

Conclusion (et perspective)