Usages & productions scientifiques en SHS à partir de données massives

Benjamin Ooghe-Tabanou

▶ **To cite this version:**

Benjamin Ooghe-Tabanou. Usages & productions scientifiques en SHS à partir de données massives. École thématique. Accompagner les chercheurs et valoriser les collections dans le cadre de la recherche, Bry sur Marne, France. 2021, pp.48. hal-03904119

HAL Id: hal-03904119

https://sciencespo.hal.science/hal-03904119

Submitted on 16 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

SciencesPo
MÉDIALAB

Usages & productions scientifiques en SHS à partir de données ~~massives~~

Formation « *Accompagner les chercheurs et valoriser les collections dans le cadre de la recherche* »

INA - 24 mars 2021

Benjamin Ooghe-Tabanou

Sciences Po médialab (@medialab_ScPo)

CC BY SA

Bruno Latour, fondateur du médialab

« Google is nice, but we need something better »

The Indian Express, 2011

médialab @ Sciences Po

https://medialab.sciencespo.fr

- Laboratoire de recherche SHS
  fondé par Bruno Latour en mai 2009,
  dirigé par Dominique Cardon depuis 2017

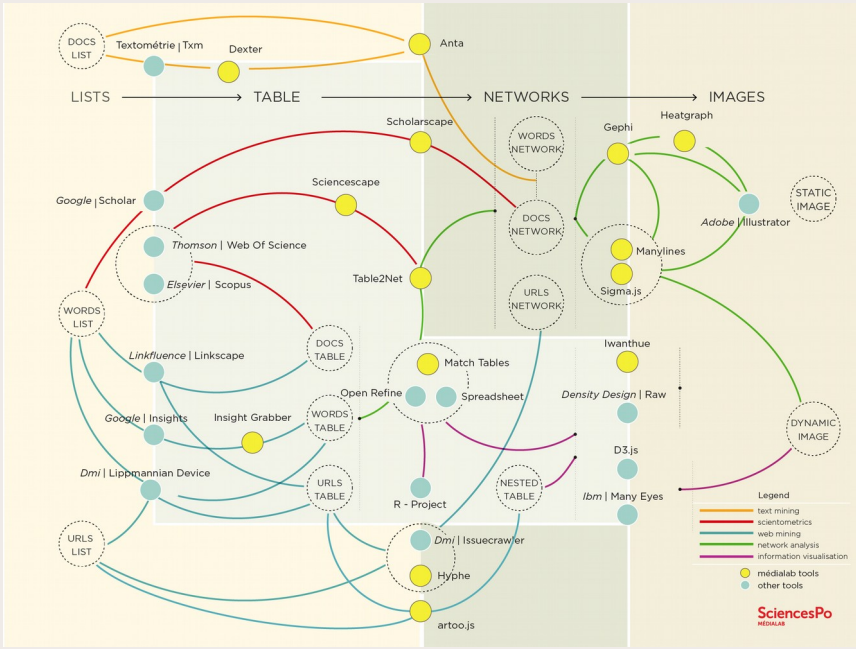
- Sciences Sociales, Ingénierie & Design
  → **interdisciplinarité**

- Articuler méthodes **quali & quanti**
  à travers une approche numérique

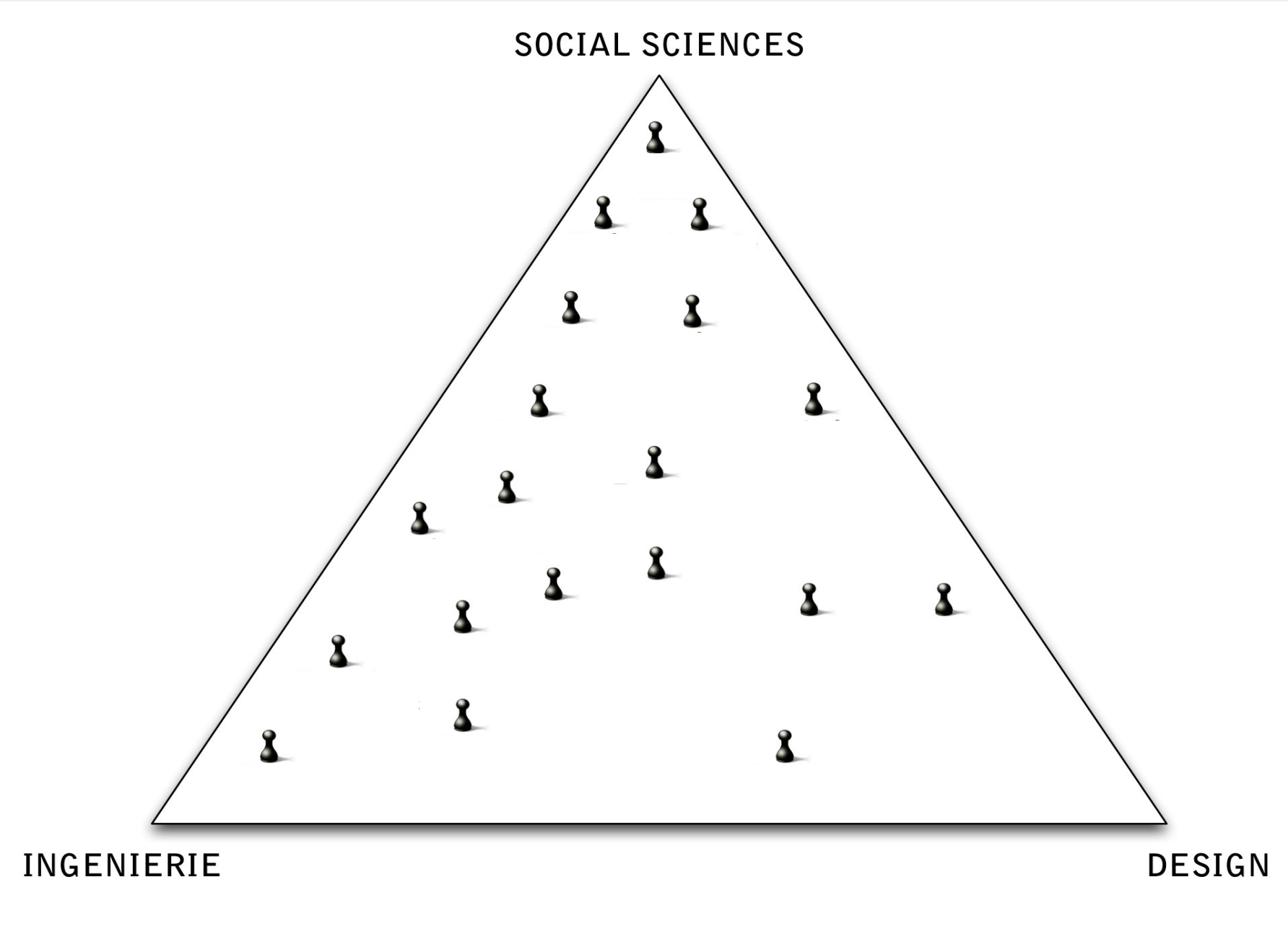
- Travailler avec les **traces numériques**

- Un écosystème d'**outils OpenSource**
  http://tools.medialab.sciences-po.fr

- Un atelier ouvert mensuel : le METAT
  https://www.sciencespo.fr/recherche/fr/content/metat-latelier-de-methodes



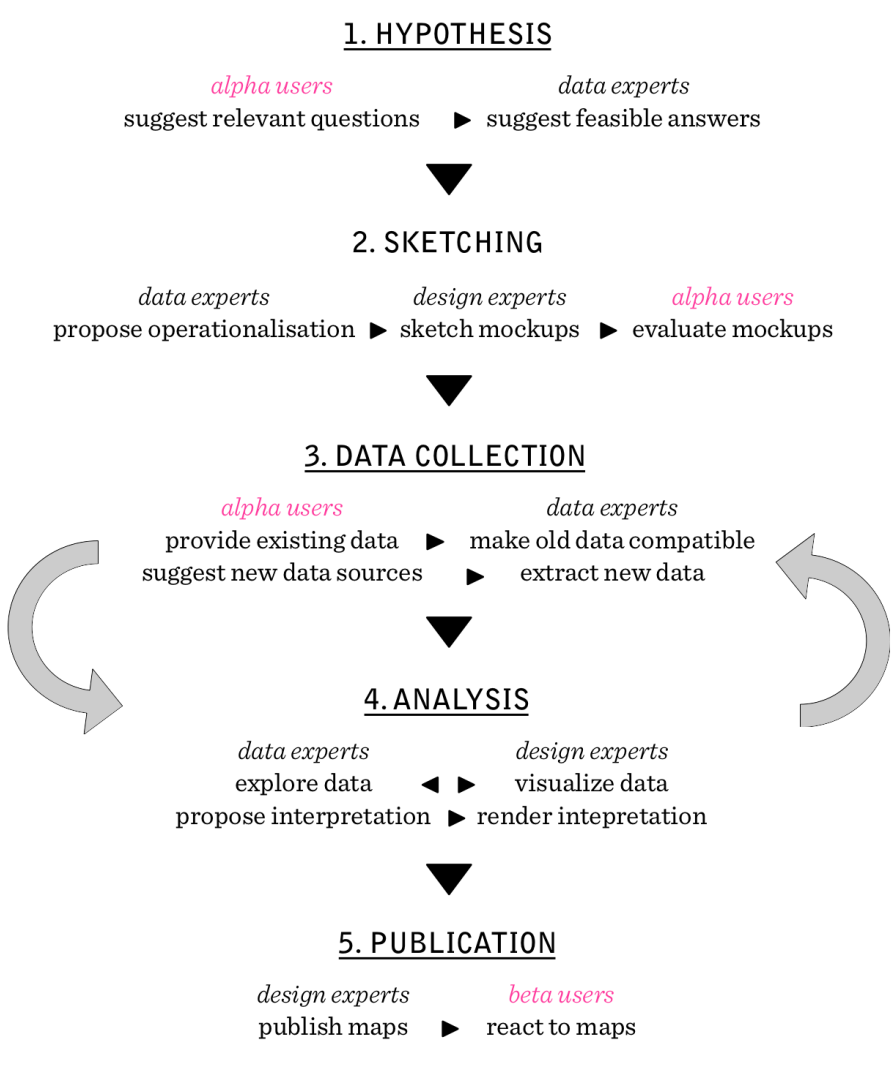
Une équipe interdisciplinaire



Une approche quali/quanti exploratoire itérative

Numérique ≠ Magique

→ toujours éviter l'entièrement automatique et garantir le retour aux sources



Les « datasprints »

Ateliers collaboratifs de travail exploratoire et analytique

→ centrés sur un ou quelques jeux de données

→ sur plusieurs jours, en petits groupes

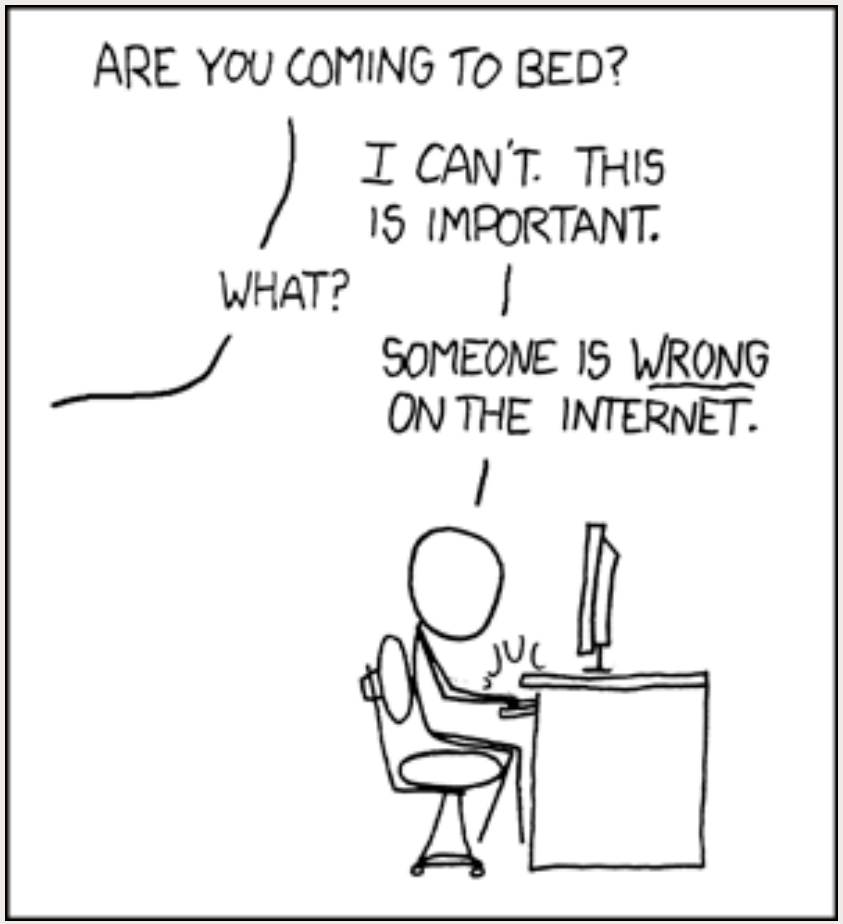
→ rassemblant une diversité de profils
(ingénieurs, designers, académiques, journalistes...)

→ confrontant les auteurs ou experts à leurs données

Exploiter le Web comme terrain d'enquêtes

Le Web : un espace de dialogue et de débats

Collecter, enrichir, nettoyer, visualiser & analyser les traces numériques



CC-BY-NC - Randall Munroe - XKCD

Différents modes de collecte de données web

2 approches bien distinctes aux cibles et résultats différents

CRAWLING **Vs.** **SCRAPING**

fouille systématique
(sources multiples hétérogènes)

contenus textuels & hyperliens

traitement
du langage

analyse de réseau
(effets de communauté)

extraction ciblée
(source unique ou ensemble cohérent)

données structurées

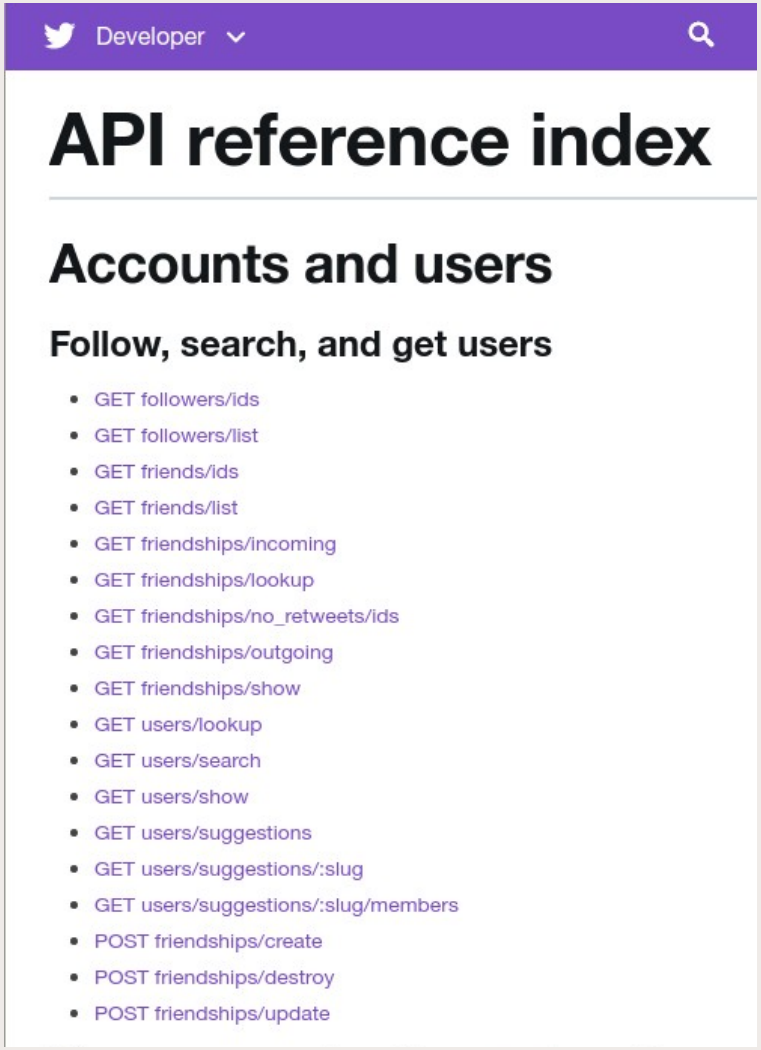
méthodes quantitatives,
statistiques…

Problèmes : des données « sales » et un important coût en maintenance

CC-BY-NC - Randall Munroe - XKCD

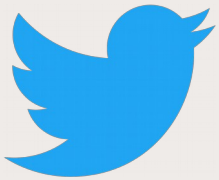
Accès contrôlé via les APIs des plateformes

- **« Application Programmation Interface »**
- **Avantages :**
  - données structurées « propres »
  - accès à de gros volume
  - relative complétude
  - accès à des informations d'usage
- **Problèmes :**
  - accès à de gros volumes
  - limitation des appels
  - « boîtes noires »
  - risques de refermeture
  - dépendance à la vision des plateformes



https://developer.twitter.com/en/docs/api-reference-index

De nombreuses métadonnées à exploiter



```
TWEET_FIELDS = [
  "id",                          # digital ID
  "time",                        # UNIX timestamp of creation
  "created_at",                  # ISO datetime of creation
  "from_user_name",              # author's user text ID (@user)
  "text",                        # message's text content
  "filter_level",                # internal TCAT field, ignorable
  "possibly_sensitive",          # whether a link present in the message might contain sensitive content according to Twitter
  "withheld_copyright",          # whether the tweet might be censored by Twitter following copyright requests, ignorable
  "withheld_scope",              # whether the content withheld is the "status" or a "user", ignorable
  "withheld_countries",          # list of ISO country codes in which the message is withheld, separated by |, ignorable
  "truncated",                   # whether the tweet is bigger than 140 characters, obsolete
  "retweet_count",               # number of retweets of the message (at collection time)
  "favorite_count",              # number of likes of the message (at collection time)
  "reply_count",                 # number of answers to the message, dropped by Twitter (since Oct 17, now charged), unreliable and ignorable
  "lang",                        # language of the message automatically identified by Twitter's algorithms (equals "und" when no language could be detected)
  "to_user_name",                # text ID of the user the message is answering to
  "to_user_id",                  # digital ID of the user the message is answering to
  "in_reply_to_status_id",       # digital ID of the tweet the message is answering to
  "source",                      # medium used by the user to post the message
  "source_name",                 # name of the medium used to post the message
  "source_url",                  # link to the medium used to post the message
  "location",                    # location declared in the user's profile (at collection time)
  "lat",                         # latitude of messages geolocalized
  "lng",                         # longitude of messages geolocalized
  "from_user_id",                # author's user digital ID
  "from_user_realname",          # author's detailed textual name (at collection time)
  "from_user_verified",          # whether the author's account is certified
  "from_user_description",       # description given in the author's profile (at collection time)
  "from_user_url",               # link to a website given in the author's profile (at collection time)
  "from_user_profile_image_url", # link to the image avatar of the author's profile (at collection time)
  "from_user_utcoffset",         # time offset due to the user's timezone, dropped by Twitter (since May 18), ignorable
  "from_user_timezone",          # timezone declared in the user's profile, dropped by Twitter (since May 18), ignorable
  "from_user_lang",              # language declared in the user's profile (at collection time)
  "from_user_tweetcount",        # number of tweets sent by the user (at collection time)
  "from_user_followercount",     # number of users following the author (at collection time)
  "from_user_friendcount",       # number of users the author is following (at collection time)
  "from_user_favourites_count",  # number of likes the author has expressed (at collection time)
  "from_user_listed",            # number of users lists the author has been included in (at collection time)
  "from_user_withheld_scope",    # whether the user content is withheld, ignorable
  "from_user_withheld_countries", # list of ISO country codes in which the user content is withheld, separated by |, ignorable
  "from_user_created_at",        # ISO datetime of creation of the author's account
  "collected_via_thread",        # whether the tweet was retrieved only as part of a thread including a tweet matching the desired query
  "retweeted_id",                # digital ID of the retweeted message
  "retweeted_user_name",         # text ID of the user who authored the retweeted message
  "retweeted_user_id",           # digital ID of the user who authoring the retweeted message
  "quoted_id",                   # digital ID of the retweeted message
  "quoted_user_name",            # text ID of the user who authored the retweeted message
  "quoted_user_id",              # digital ID of the user who authoring the retweeted message
  "links",                       # list of links included in the text content, with redirections resolved, separated by |
  "medias_urls",                 # list of links to images/videos embedded, separated by |
  "medias_files",                # list of filenames of images/videos embedded and downloaded, separated by |, ignorable when medias collections isn't enabled
  "mentioned_user_names",        # list of text IDs of users mentionned, separated by |
  "mentioned_user_ids",          # list of digital IDs of users mentionned, separated by |
  "hashtags"                     # list of hashtags used, lowercased, separated by |
]
```

« Big Data » ou données massives et SHS ?

Big Data = trop gros pour être manipulé par un ordinateur (> To)
des données personnelles le plus souvent

→ extrêmement rare en Sciences Humaines

Développement d'outils au service des SHS

- Viser une large **Adoption :**
  - → **conception** d'outils dédiés aux besoins des chercheurs
  - → **design** d'interfaces centrées sur l'utilisateur
  - → **publication** d'outils web utilisables directement en ligne

- Assurer un maximum de **Réutilisabilité :**
  - → développement « **opportuniste** » de fonctionnalités
  - → diffusion en Logiciel Libre **Open Source**
    (téléchargeable, installable, vérifiable & modifiable)

Gazouilloire : extraction systématique de tweets

https://github.com/medialab/gazouilloire

- Collecter en direct en continu (et jusqu'à 7 jours en arrière)
  - des tweets par mots-clés, urls, utilisateurs, localisation, langue…
  - les conversations et médias associés
  - des profils d'utilisateurs

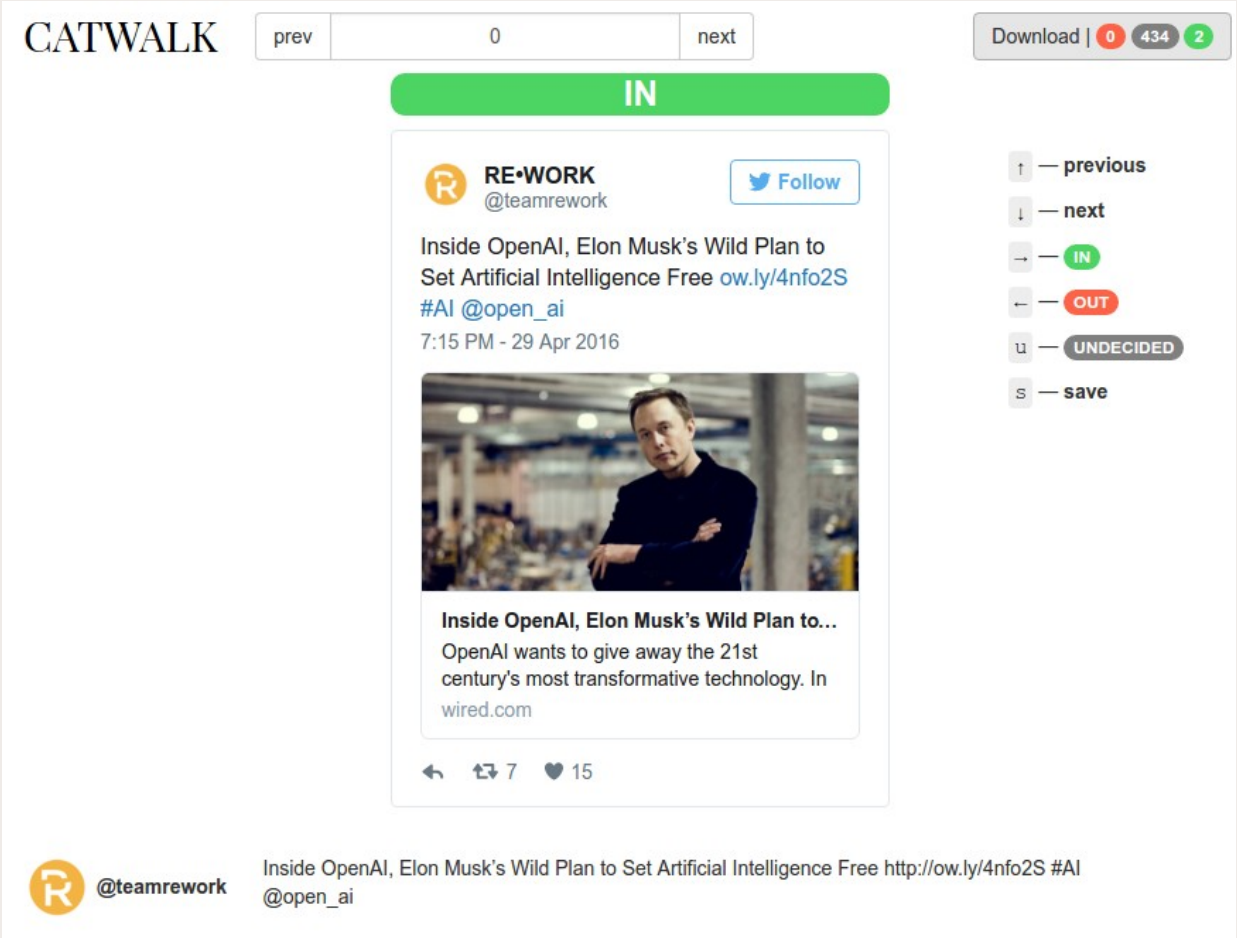
```
{
    "twitter": {
        "user": "Gazou_medialab2",
        "key": "[redacted]",
        "secret": "[redacted]",
        "oauth_token": "[redacted]",
        "oauth_secret": "[redacted]"
    },
    "mongo": {
        "host": "localhost",
        "port": 27017,
        "db": "tweets-naturpradi"
    },
    "keywords": [
        "écologique Paris",
        "végétation Paris",
        "verger Paris",
        "grenelle environnement Paris",
        "locavore Paris"
    ],
    "time_limited_keywords": {
    },
    "geolocalisation": null,
    "geolocalisation_type": "admin",
    "resolve_redirected_links": true,
    "grab_conversations": true,
    "download_medias": true,
    "medias_directory": "/store/tweets/naturpradi/media/",
    "timezone": "Europe/Paris",
    "debug": true
}
```

```
[2016-11-22 15:23:34.056196] DEBUG: Starting search queries with 328 remaining calls for the next 655 seconds
[2016-11-22 15:23:34.259849] DEBUG: [search] +1 tweets (agriculture%20Paris OR agricultures%20Paris OR agroforesterie%20Paris)
[2016-11-22 15:23:35.807085] DEBUG: Saved 1 tweets in MongoDB
[2016-11-22 15:23:37.358533] DEBUG: [search] +1 tweets (espaces%20verts%20Paris OR ferme%20Paris OR fermes%20Paris)
[2016-11-22 15:23:37.810930] DEBUG: Saved 1 tweets in MongoDB
[2016-11-22 15:23:45.049743] DEBUG: [stream] +1 tweet
[2016-11-22 15:23:45.821150] DEBUG: Saved 1 tweets in MongoDB
[2016-11-22 15:24:51.598045] DEBUG: [stream] +1 tweet
[2016-11-22 15:24:51.893009] DEBUG: Saved 1 tweets in MongoDB
[2016-11-22 15:24:52.401661] DEBUG: [medias] +1 files
[2016-11-22 15:24:58.073013] DEBUG: Starting search queries with 286 remaining calls for the next 571 seconds
[2016-11-22 15:25:00.383614] DEBUG: [stream] +1 tweet
[2016-11-22 15:25:01.905385] DEBUG: Saved 1 tweets in MongoDB
[2016-11-22 15:26:18.060840] DEBUG: Starting search queries with 246 remaining calls for the next 491 seconds
[2016-11-22 15:26:19.922864] DEBUG: [search] +1 tweets (compost%20Paris OR composts%20Paris OR compostage%20Paris)
[2016-11-22 15:26:19.989779] DEBUG: Saved 1 tweets in MongoDB
```

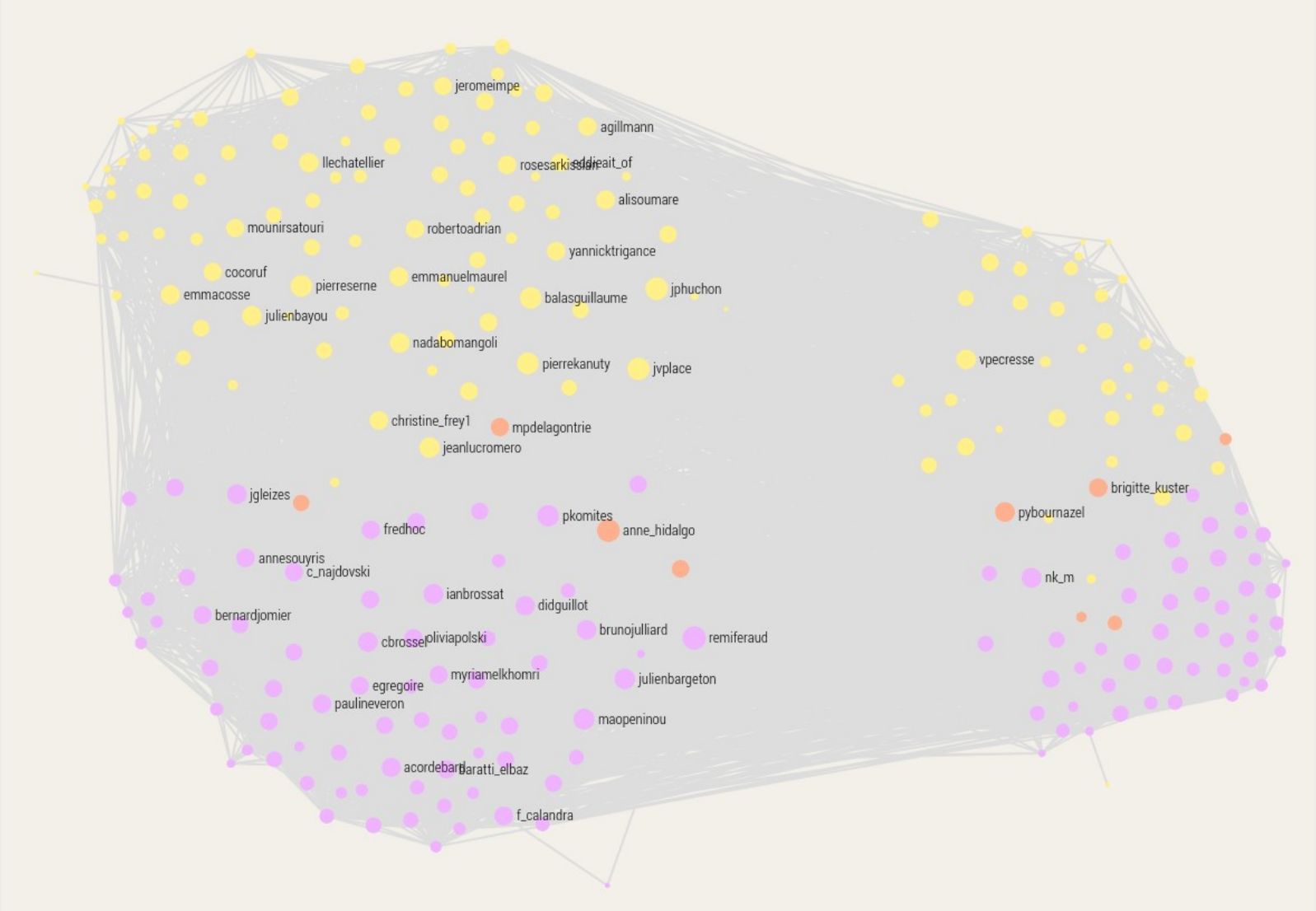
CatWalk : sélection qualitative de tweets

https://medialab.github.io/catwalk/

- Passer en revue rapidement « *à la Tinder* » tous les tweets d'un CSV pour décider de les inclure / exclure d'un corpus

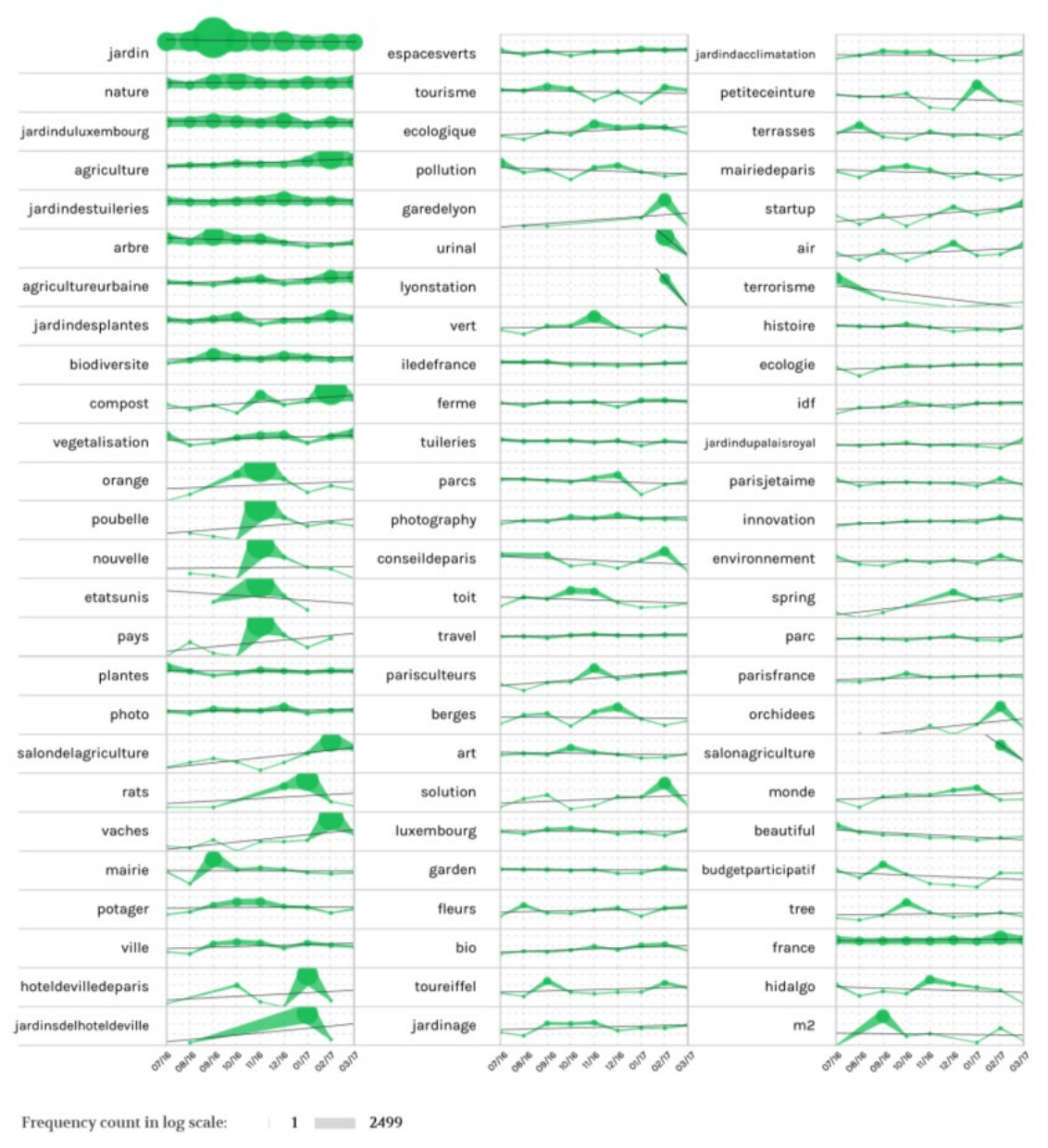


Explorer des réseaux de followers

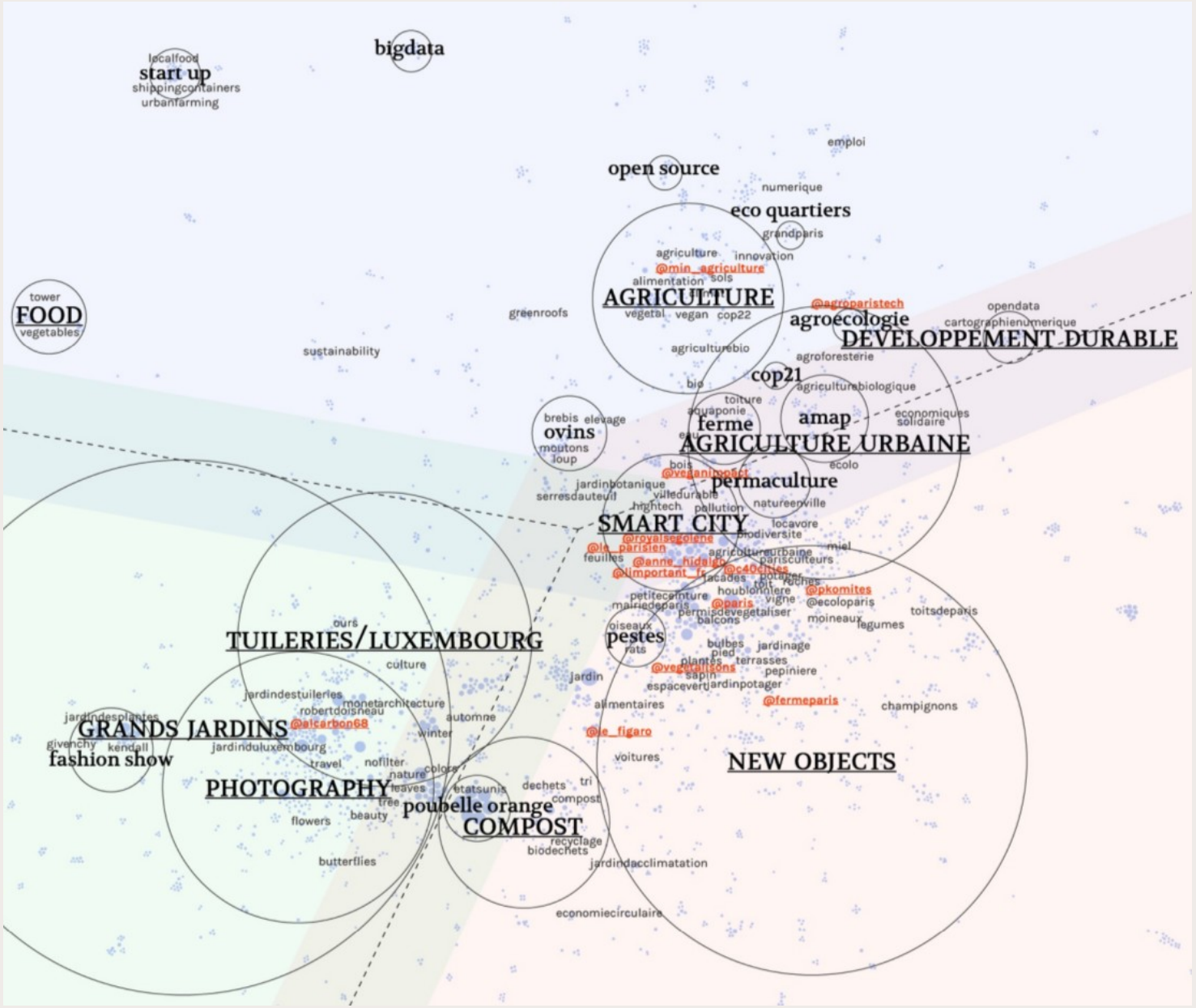


Liens de proximité Twitter entre les élus du Conseil Régional d'Île-de-France et du Conseil de Paris

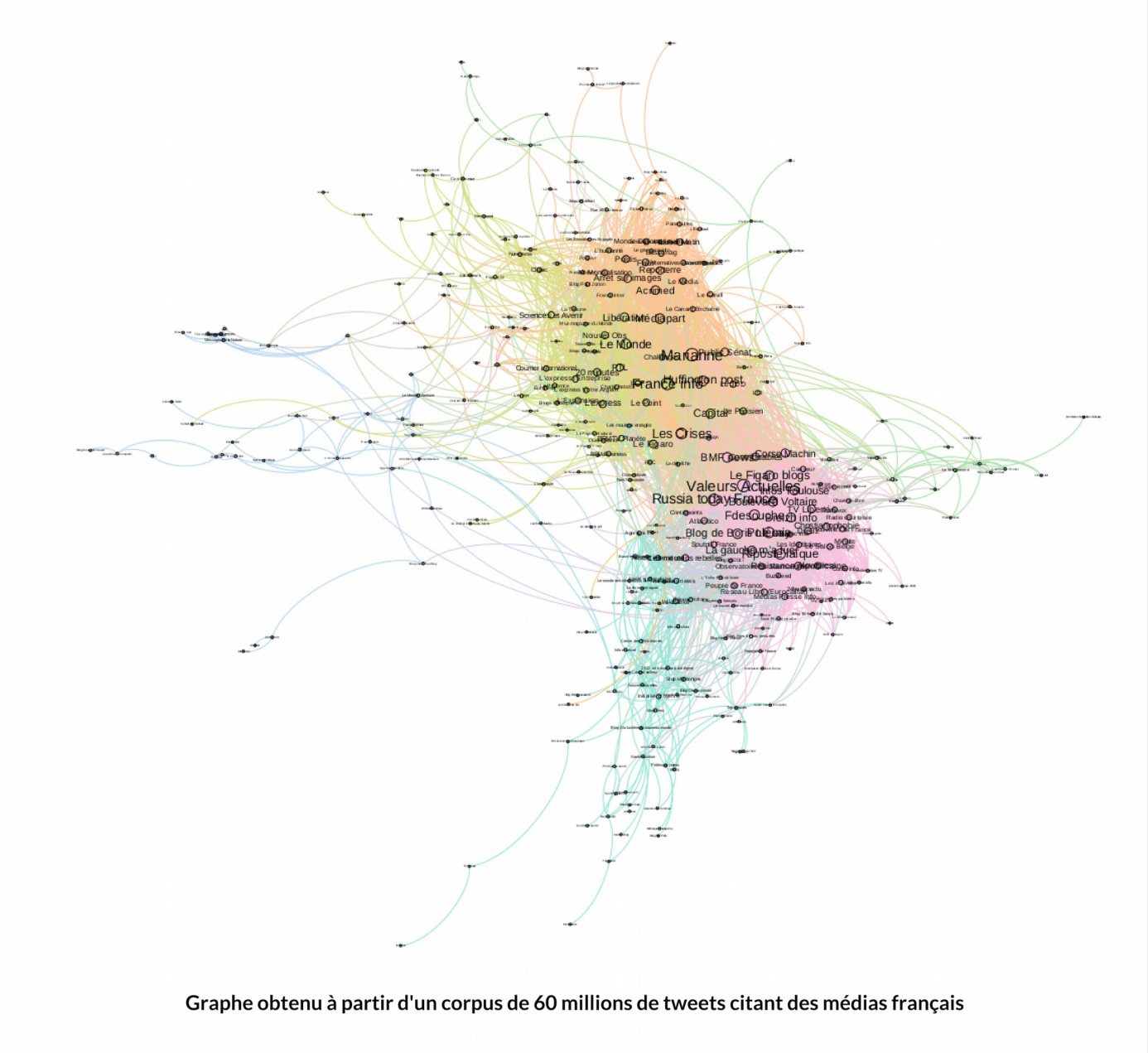
Explorer les dynamiques temporelles



Explorer des réseaux sémantiques



Explorer des réseaux de co-citation de sites web

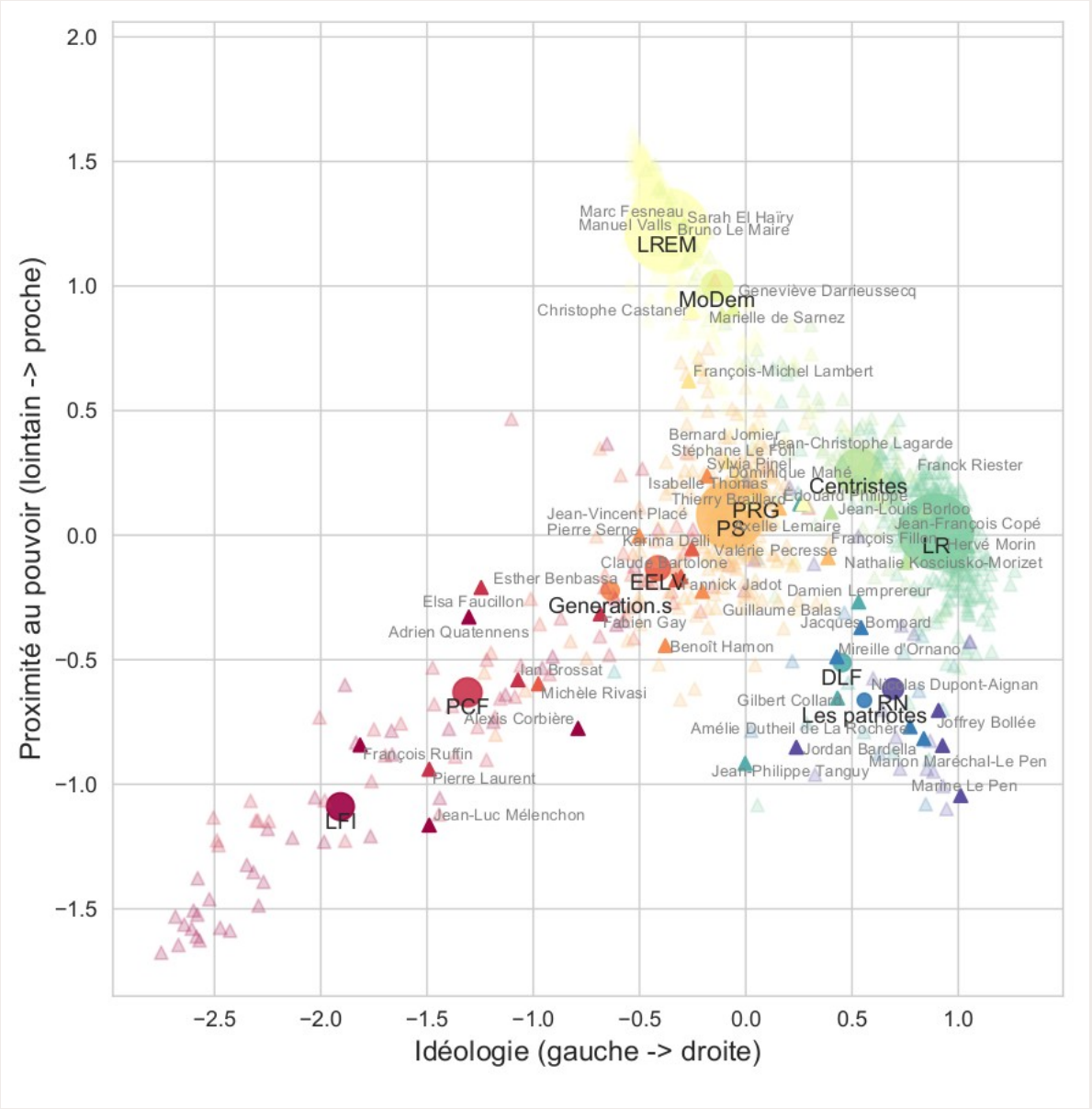


Graphe obtenu à partir d'un corpus de 60 millions de tweets citant des médias français

Explorer l'espace visuel d'un corpus

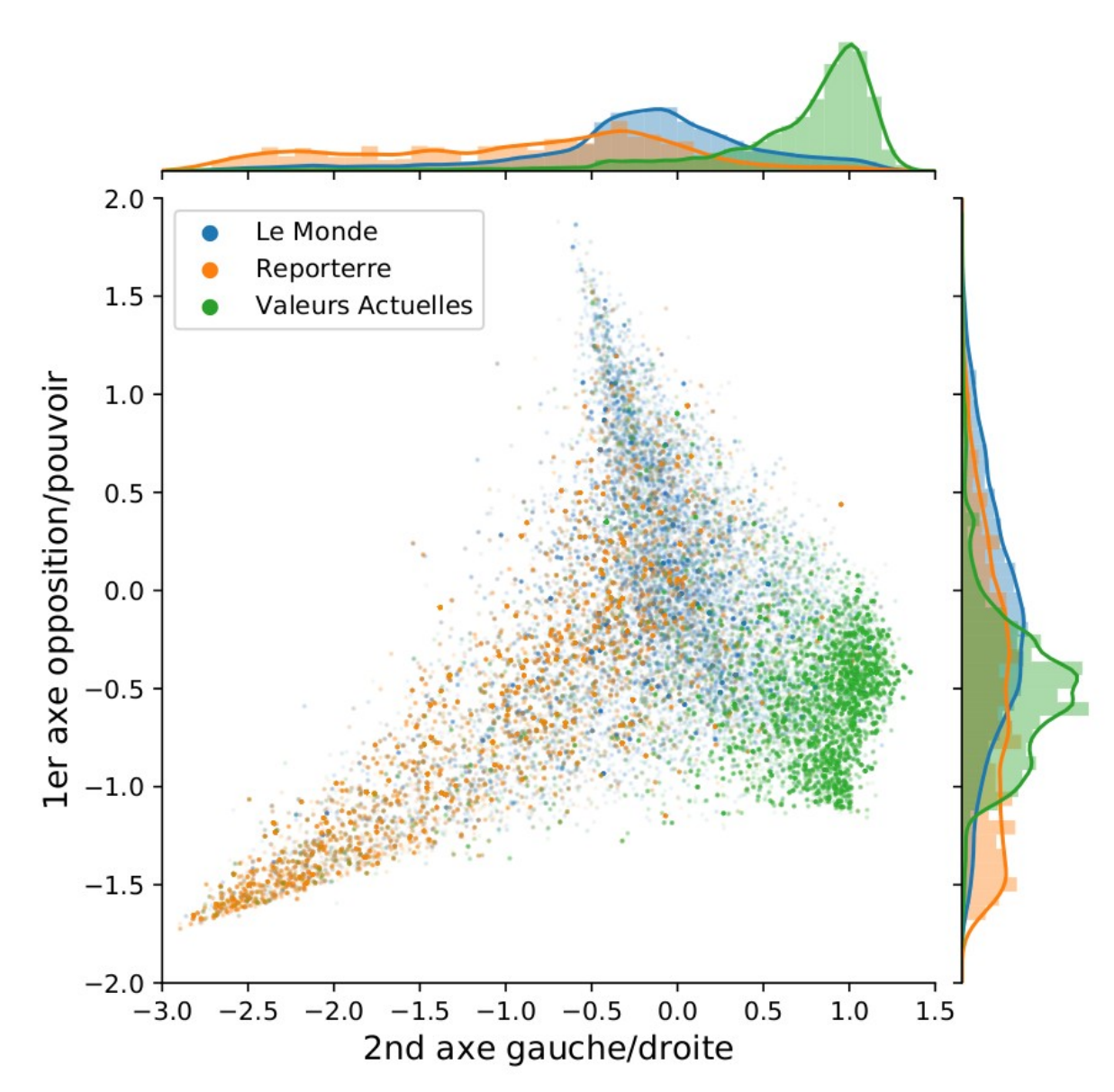
Étudier la propagation de l'info sur l'espace politique

Cartographie de l'espace politique français des utilisateurs de Twitter à partir des followers de parlementaires (méthode de Barbera)



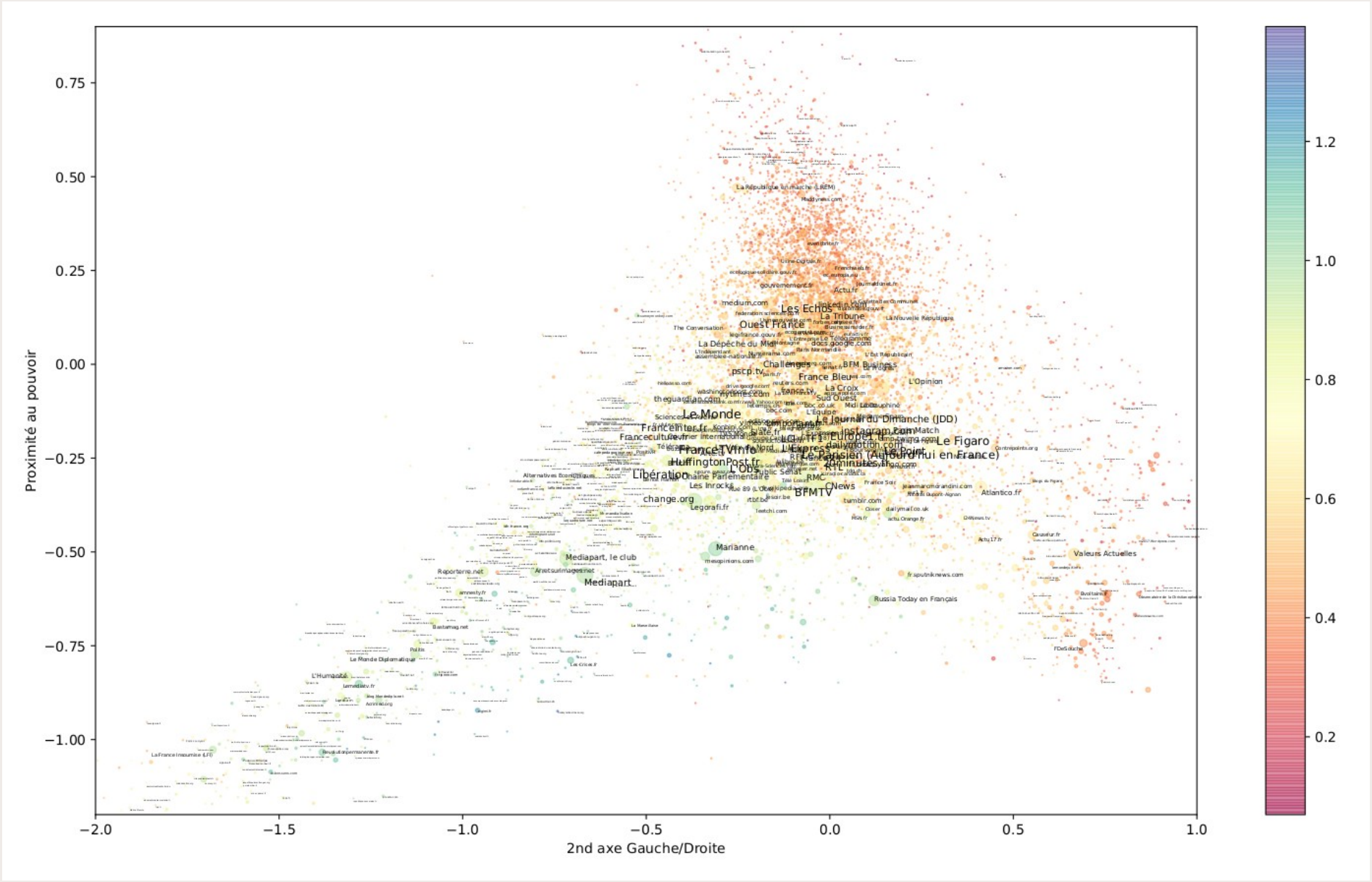
Étudier la propagation de l'info sur l'espace politique

Projection des articles de médias partagés par les utilisateurs de Twitter sur la cartographie de l'espace politique



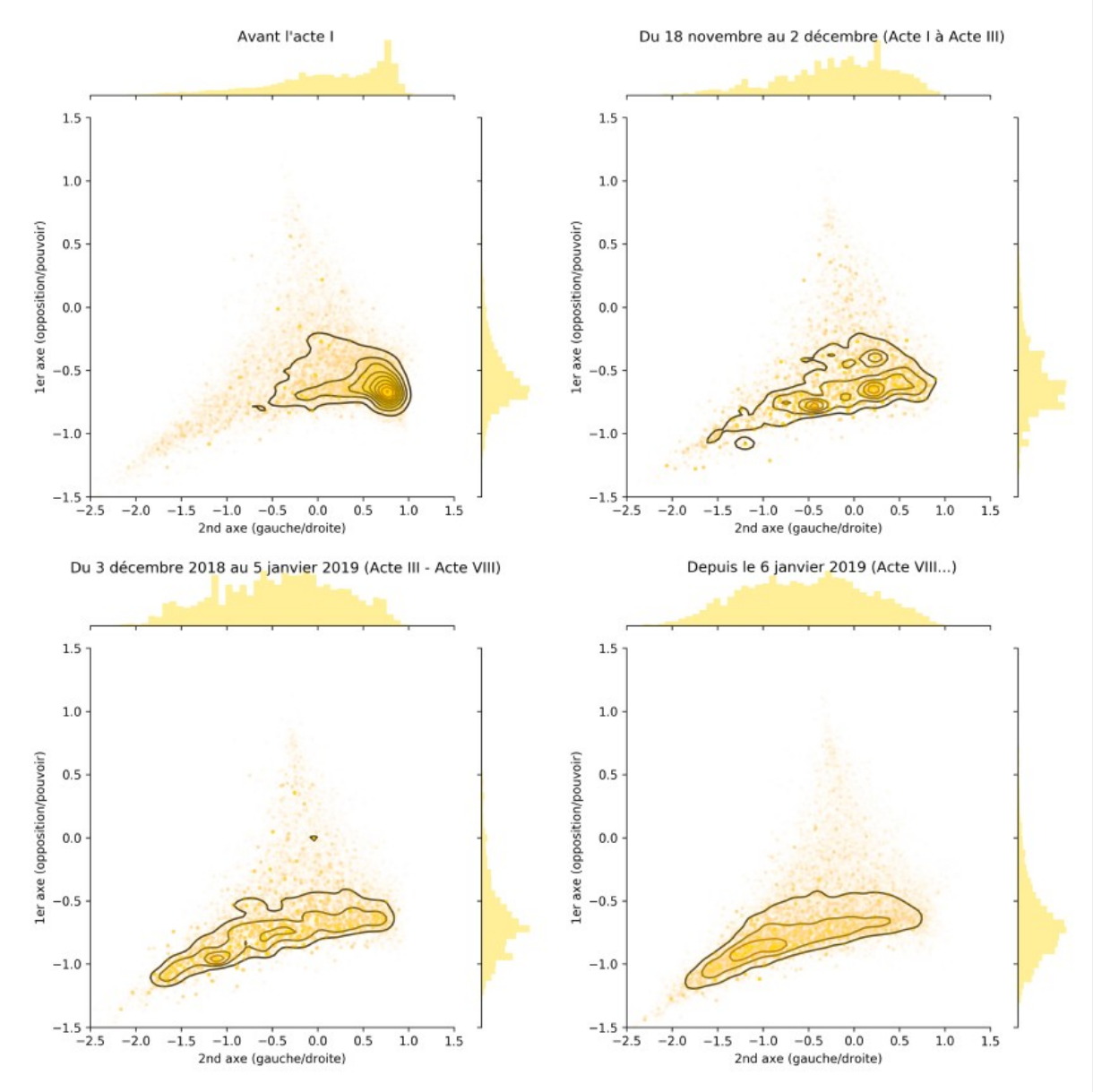
Étudier la propagation de l'info sur l'espace politique

Application à l'ensemble des médias partagés pour positionner ceux-ci sur la cartographie de l'espace politique



Étudier la propagation de l'info sur l'espace politique

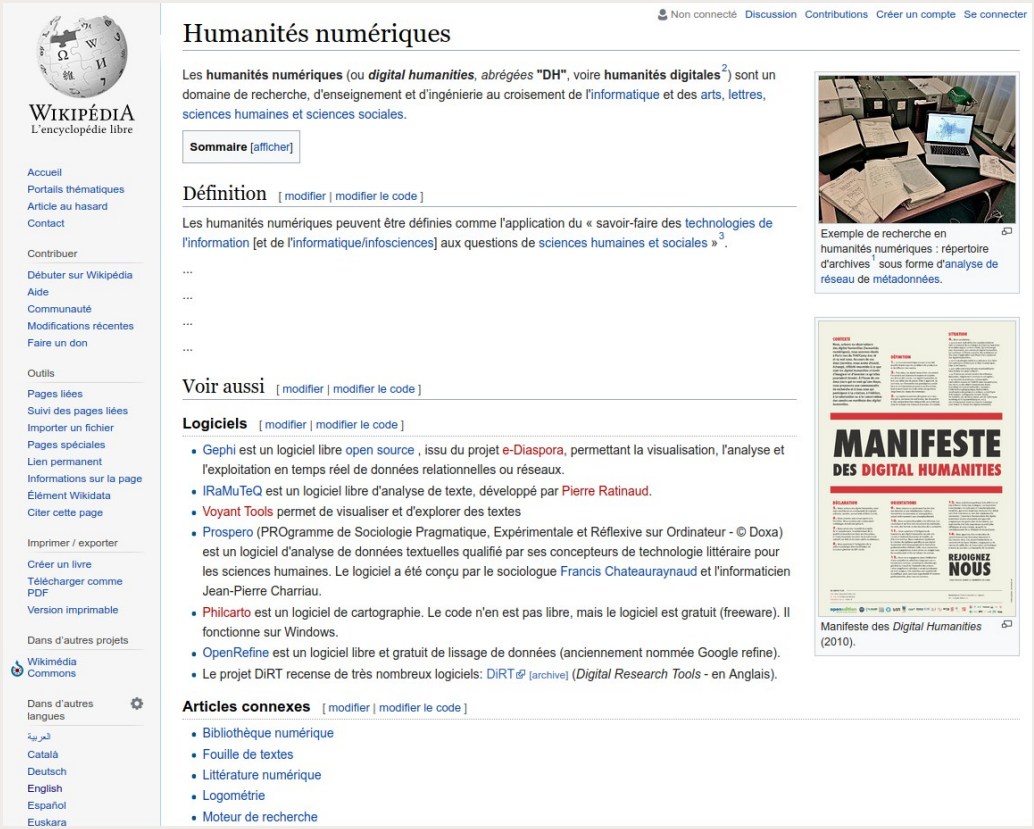
Analyse dynamique des médias partagés sur les groupes Facebook de Gilets Jaunes au fil du mouvement

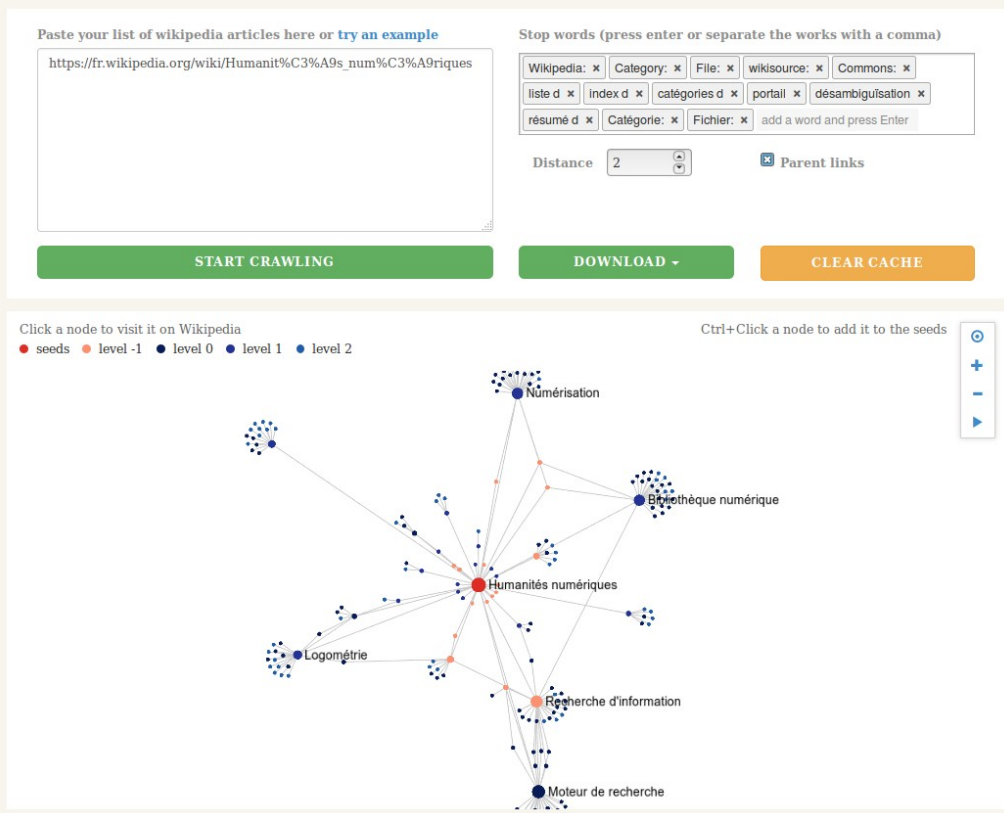


SeeAlsology : exploration sémantique rapide

http://tools.medialab.sciences-po.fr/seealsology/

Construire & explorer un réseau sémantique de liens entre concepts identifiés via les sections « Voir aussi » de Wikipedia

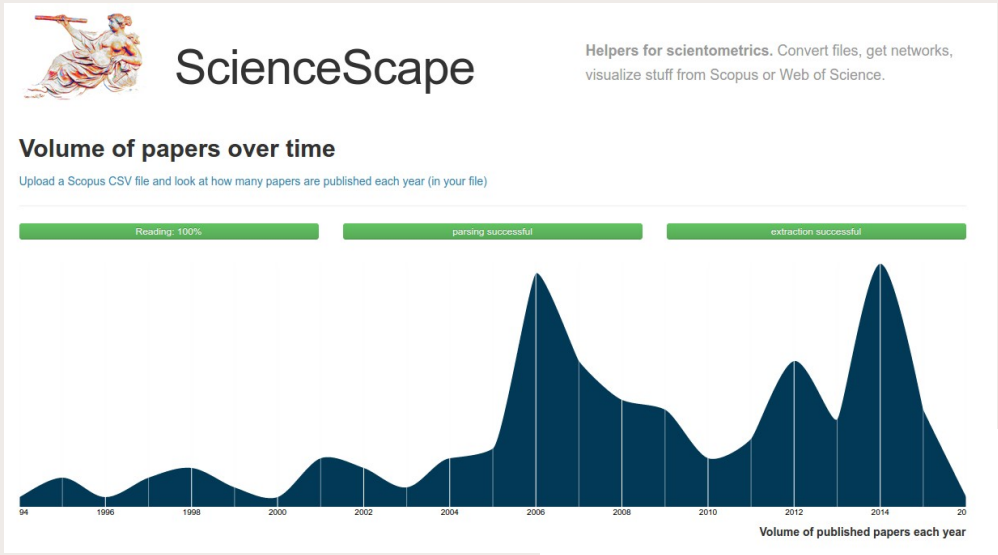




ScienceScape : scientométrie en quelques clics

http://tools.medialab.sciences-po.fr/sciencescape/

Explorer et analyser les auteurs, mots-clés et revues d'un corpus de publications issu de Scopus ou WebOfScience



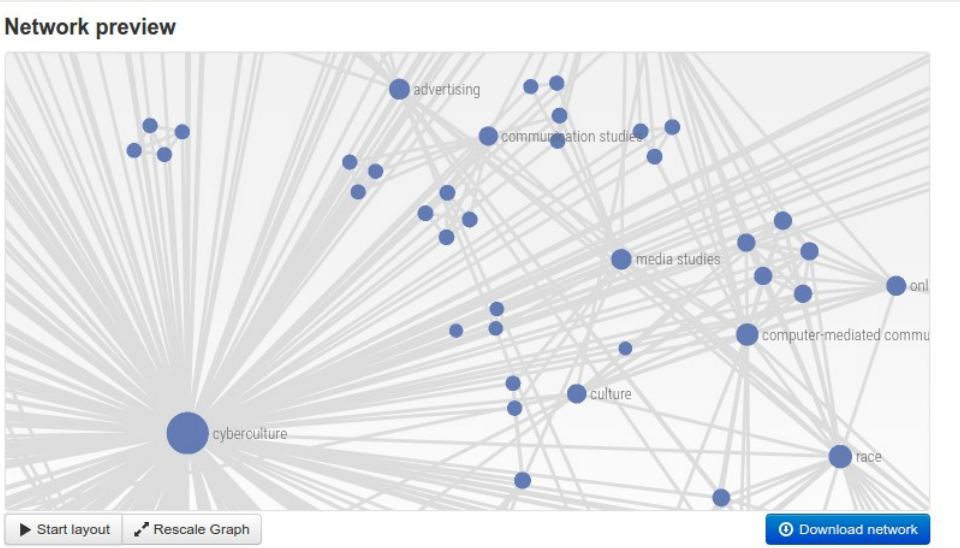
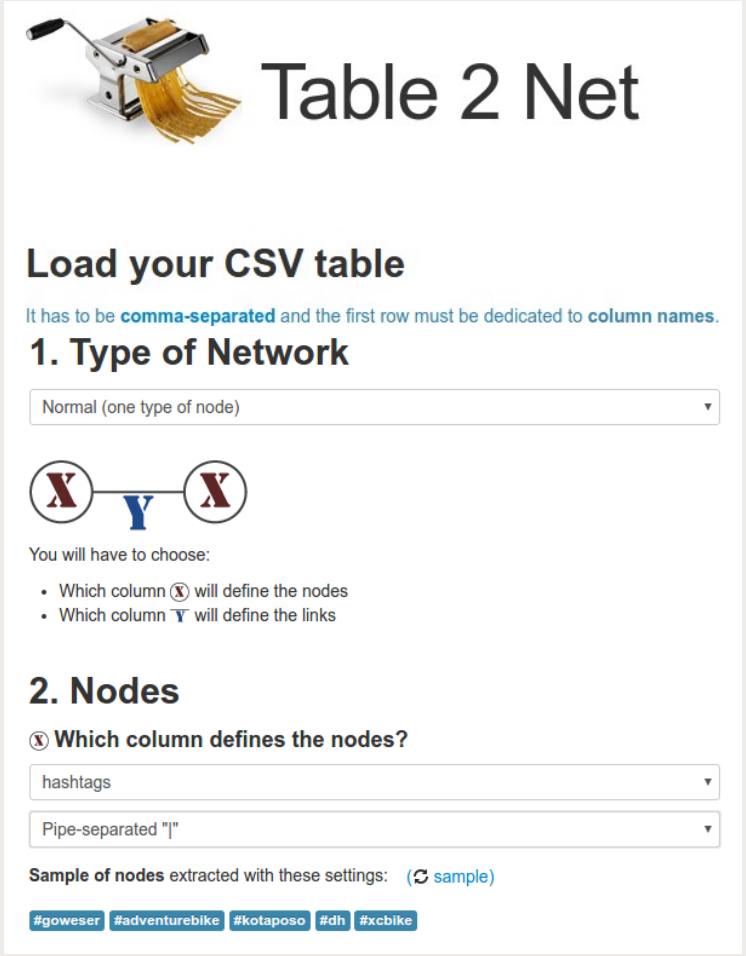


Table2Net : construire un réseau à partir d'un CSV

http://tools.medialab.sciences-po.fr/table2net/

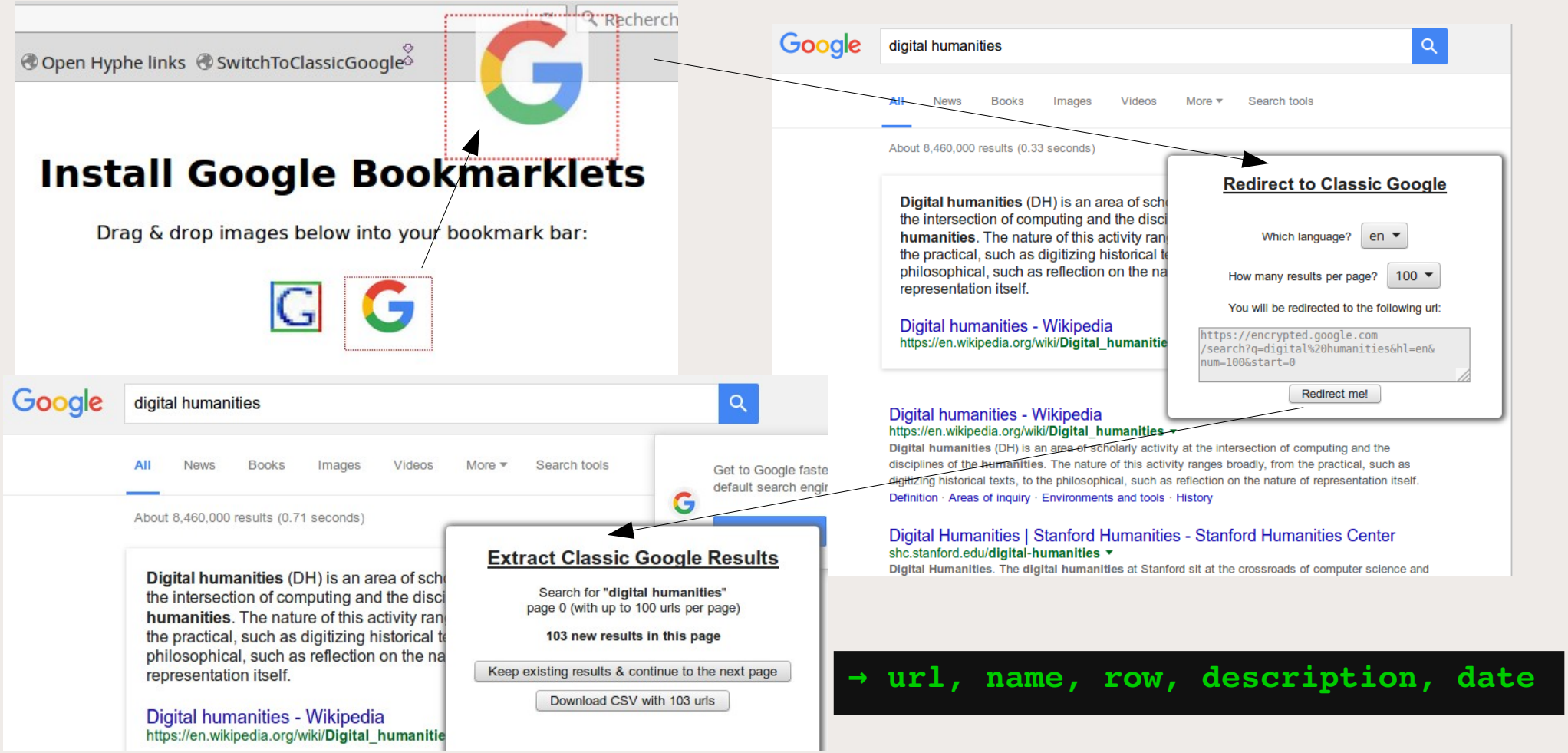
Générer un réseau de liens entre éléments à partir des données d'un fichier tableur



Google bookmarklets : résultats Google en data

https://medialab.github.io/google-bookmarklets/

Des petits boutons installables simplement dans les favoris du navigateur pour exporter simplement en tableur des résultats d'une recherche Google



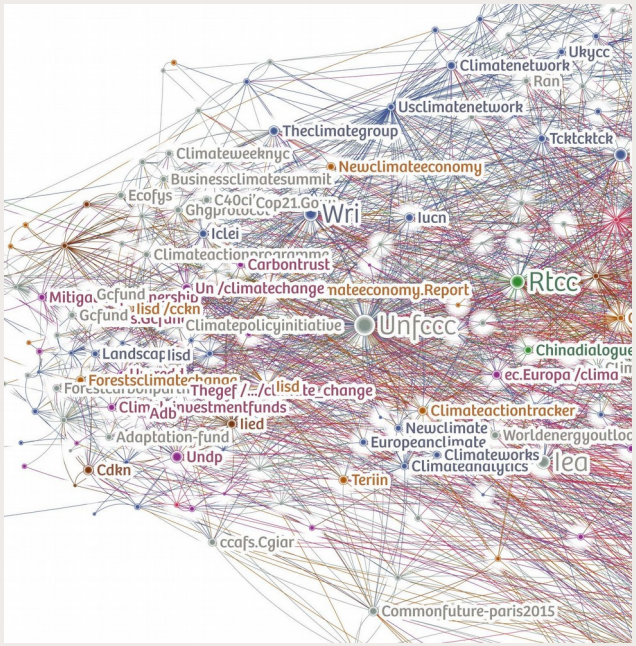
→ url, name, row, description, date

Hyphe : un crawler orienté par la recherche

http://hyphe.medialab.sciences-po.fr/demo/

- Les liens hypertextes : nouveaux révélateurs de relations entre acteurs d'une thématique
- Créer un corpus documentaire
  - « acteurs web » & contenus textuels respectifs
  - liens hypertextes entre ces acteurs
- Études exploratoires ou de controverses dans tous les domaines

hyphe



http://medialab.github.io/double-dating-data/

COP 21

Vie privée

Extrême droite

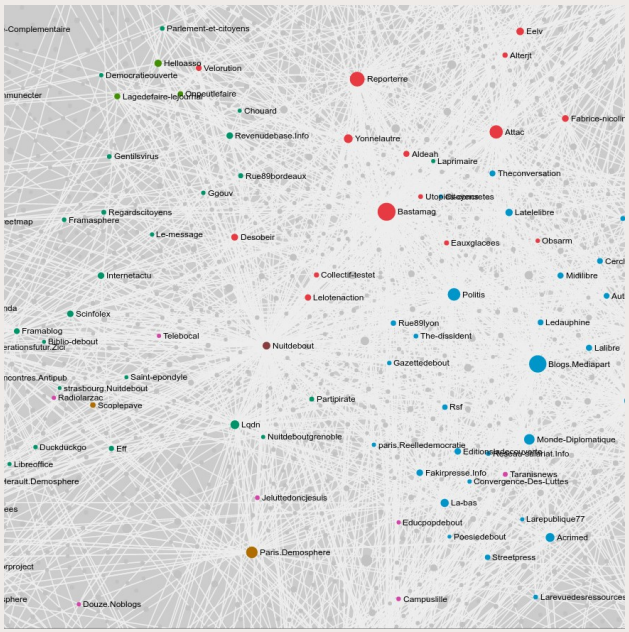
Tissu associatif

Produits laitiers

Cellules souches

Administrations culturelles

…

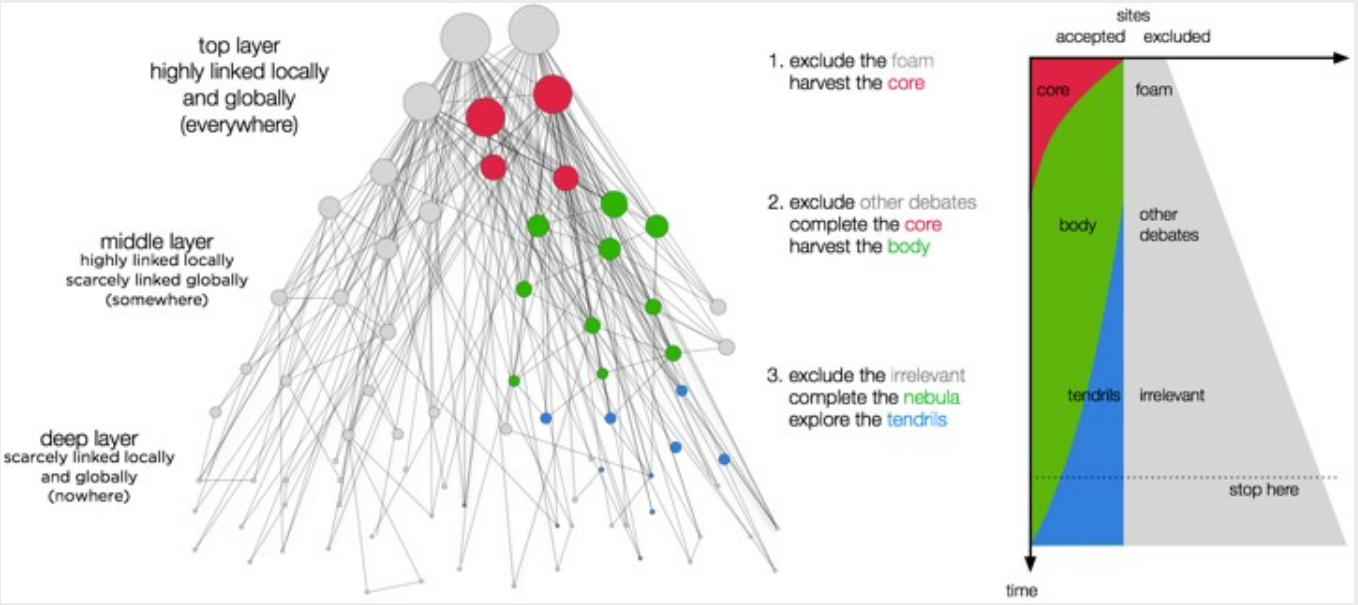


http://utopies-concretes.org/

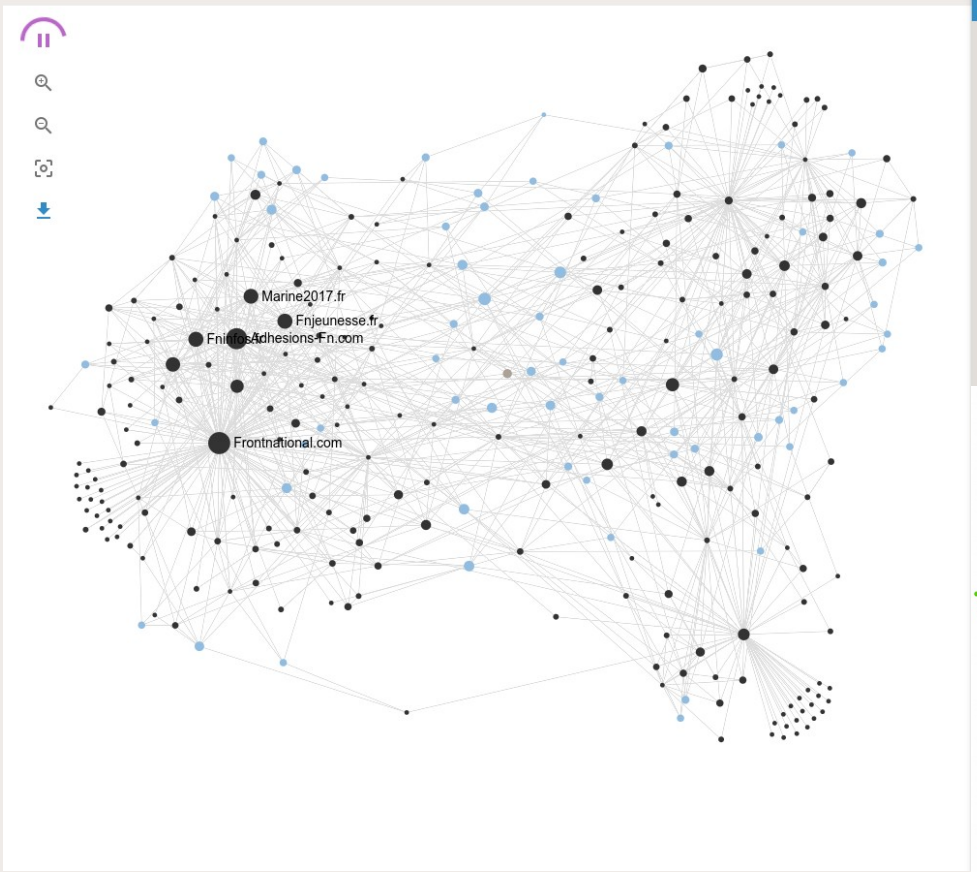
Hyphe : une stratégie de crawling contrôlé

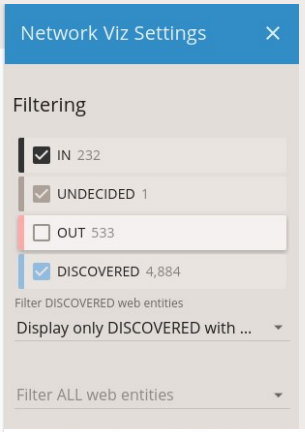
- Crawlers classiques : snowballing
  → Surreprésentation des couches hautes (Google, YouTube, Wikipedia…)
  → Dérive thématique rapide

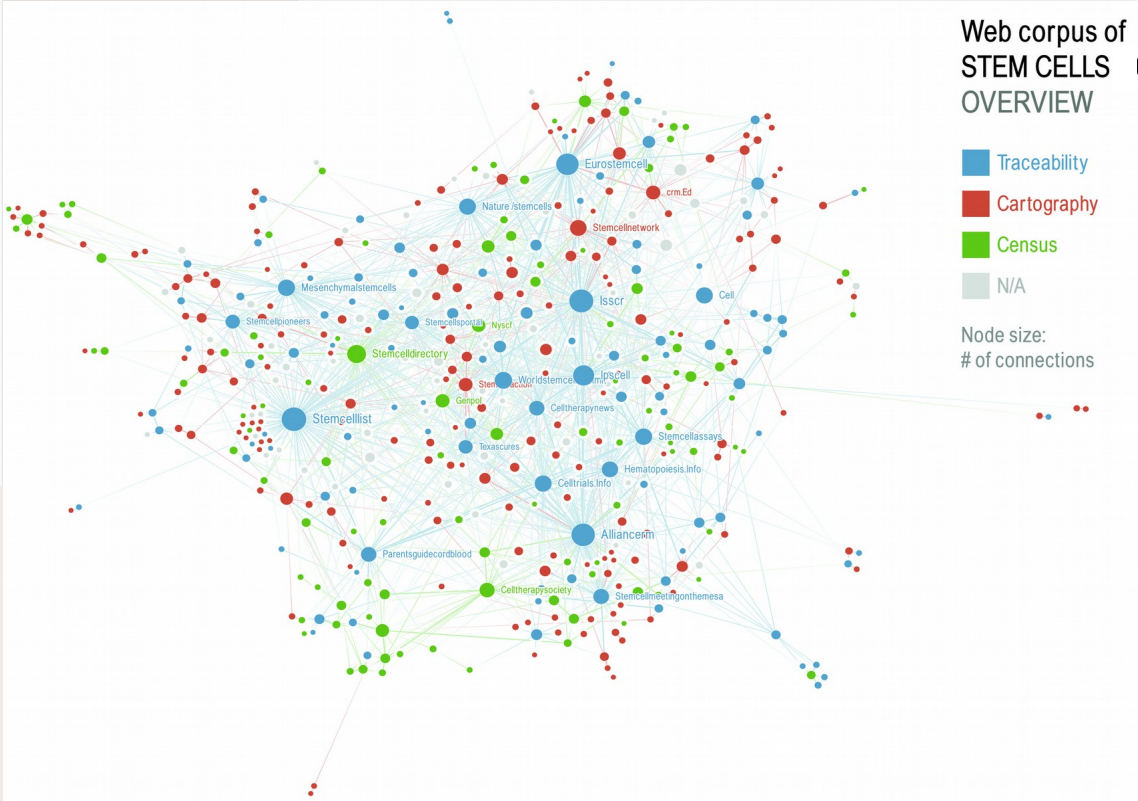
- Hyphe : crawling semi-automatique
  - Fouille systématique des pages des WebEntités choisies uniquement
  - Choix humain des autres WebEntités à crawler grâce au degré de citation



Analyse de réseau : clusters, oppositions & affinités







Social Representations of Stem Cells, Virginie Tournay, CEVIPOF, 2016

SciencesPo
MÉDIALAB

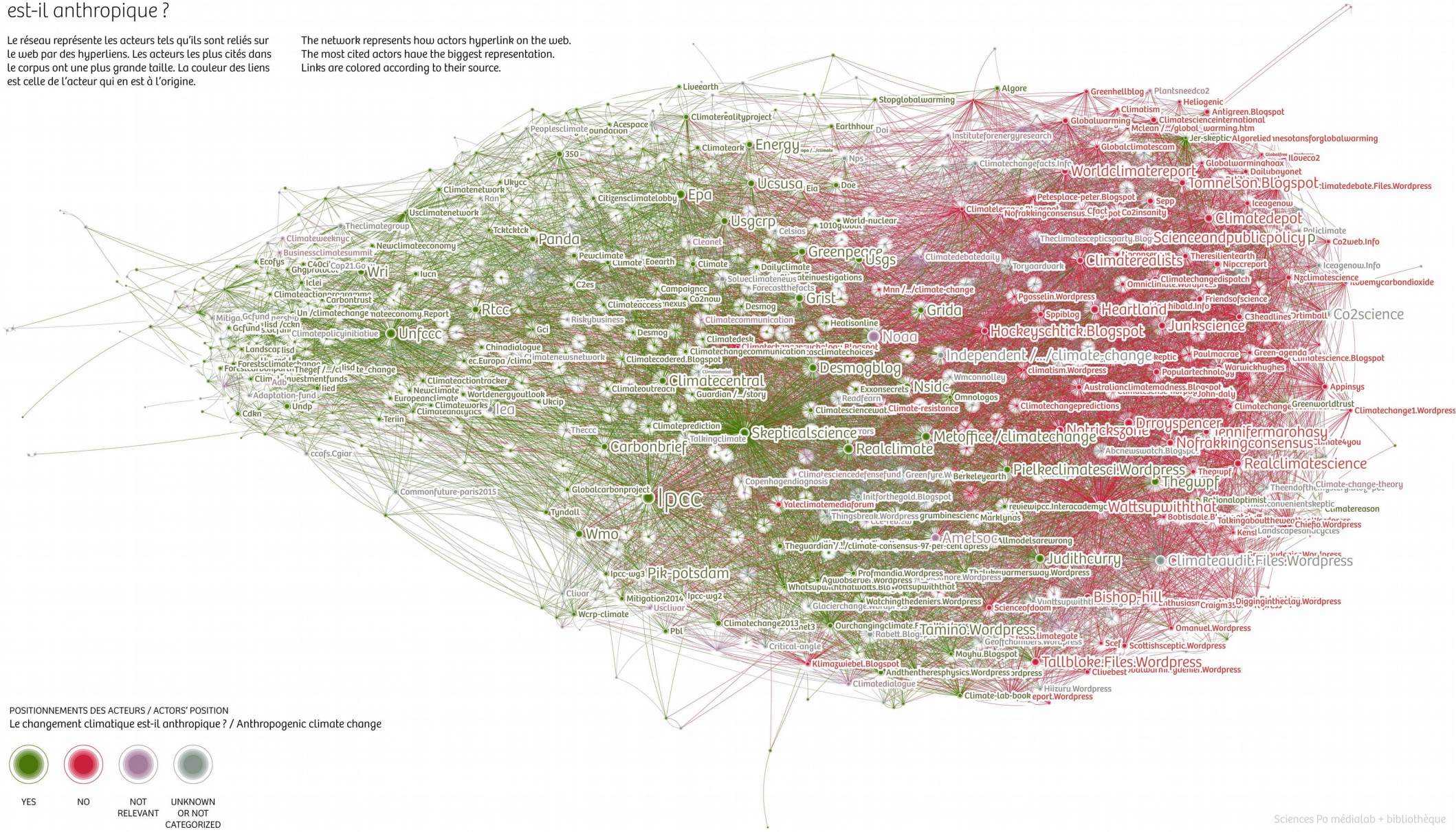
Cartographier le web autour d'un thème/débat

Corpus web sur le changement climatique : le changement climatique est-il anthropique ?

Le réseau représente les acteurs tels qu'ils sont reliés sur le web par des hyperliens. Les acteurs les plus cités dans le corpus ont une plus grande taille. La couleur des liens est celle de l'acteur qui en est à l'origine.

Web corpus on climate change: anthropogenic climate change

The network represents how actors hyperlink on the web. The most cited actors have the biggest representation. Links are colored according to their source.



POSITIONNEMENTS DES ACTEURS / ACTORS' POSITION
Le changement climatique est-il anthropique ? / Anthropogenic climate change

YES | NO | NOT RELEVANT | UNKNOWN OR NOT CATEGORIZED

SciencesPo
MÉDIALAB

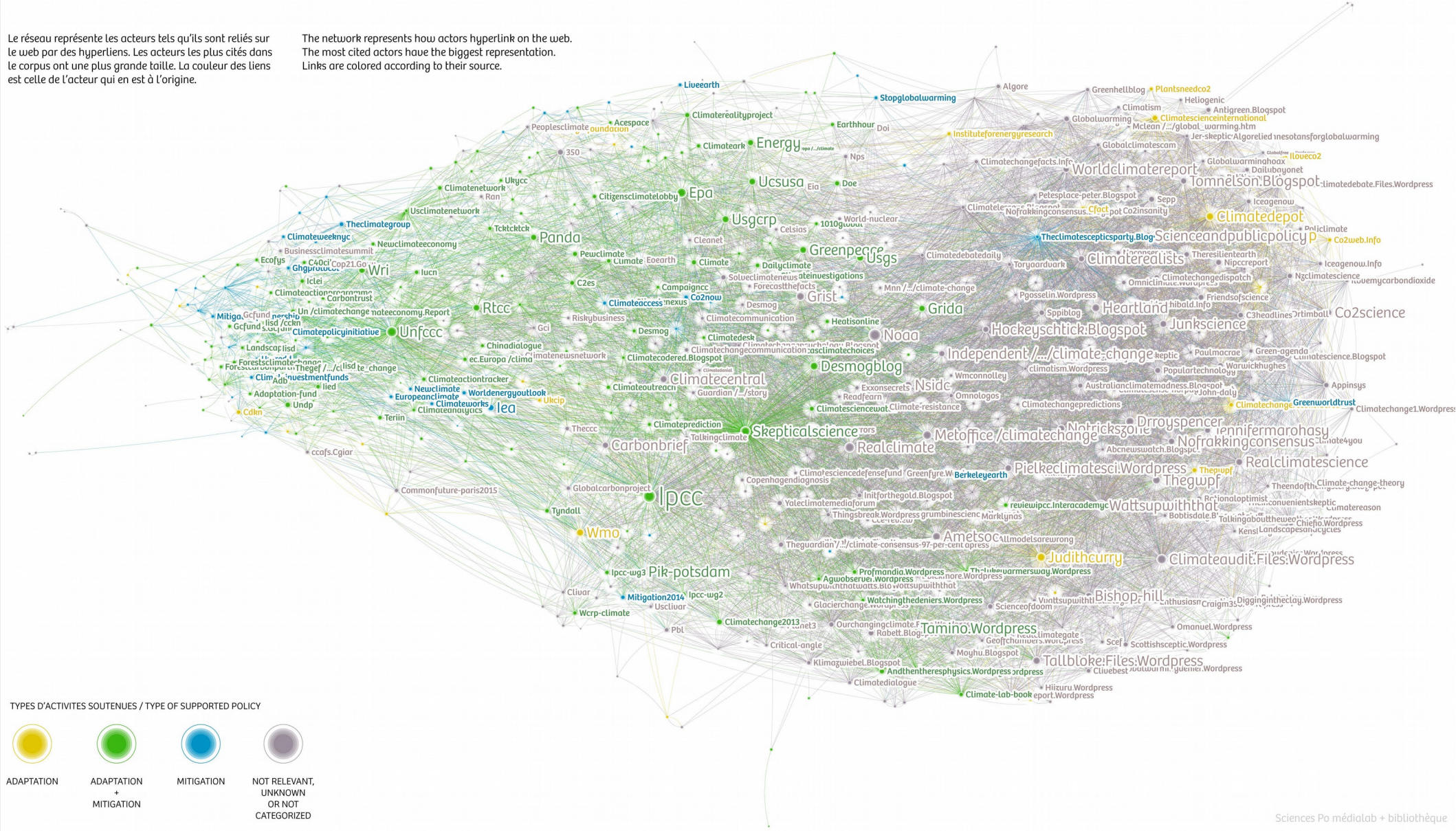
Cartographier le web par typologie d'acteurs

Corpus web sur le changement climatique : types d'activités soutenues

Le réseau représente les acteurs tels qu'ils sont reliés sur le web par des hyperliens. Les acteurs les plus cités dans le corpus ont une plus grande taille. La couleur des liens est celle de l'acteur qui en est à l'origine.

Web corpus on climate change: type of supported policy

The network represents how actors hyperlink on the web. The most cited actors have the biggest representation. Links are colored according to their source.



TYPES D'ACTIVITES SOUTENUES / TYPE OF SUPPORTED POLICY

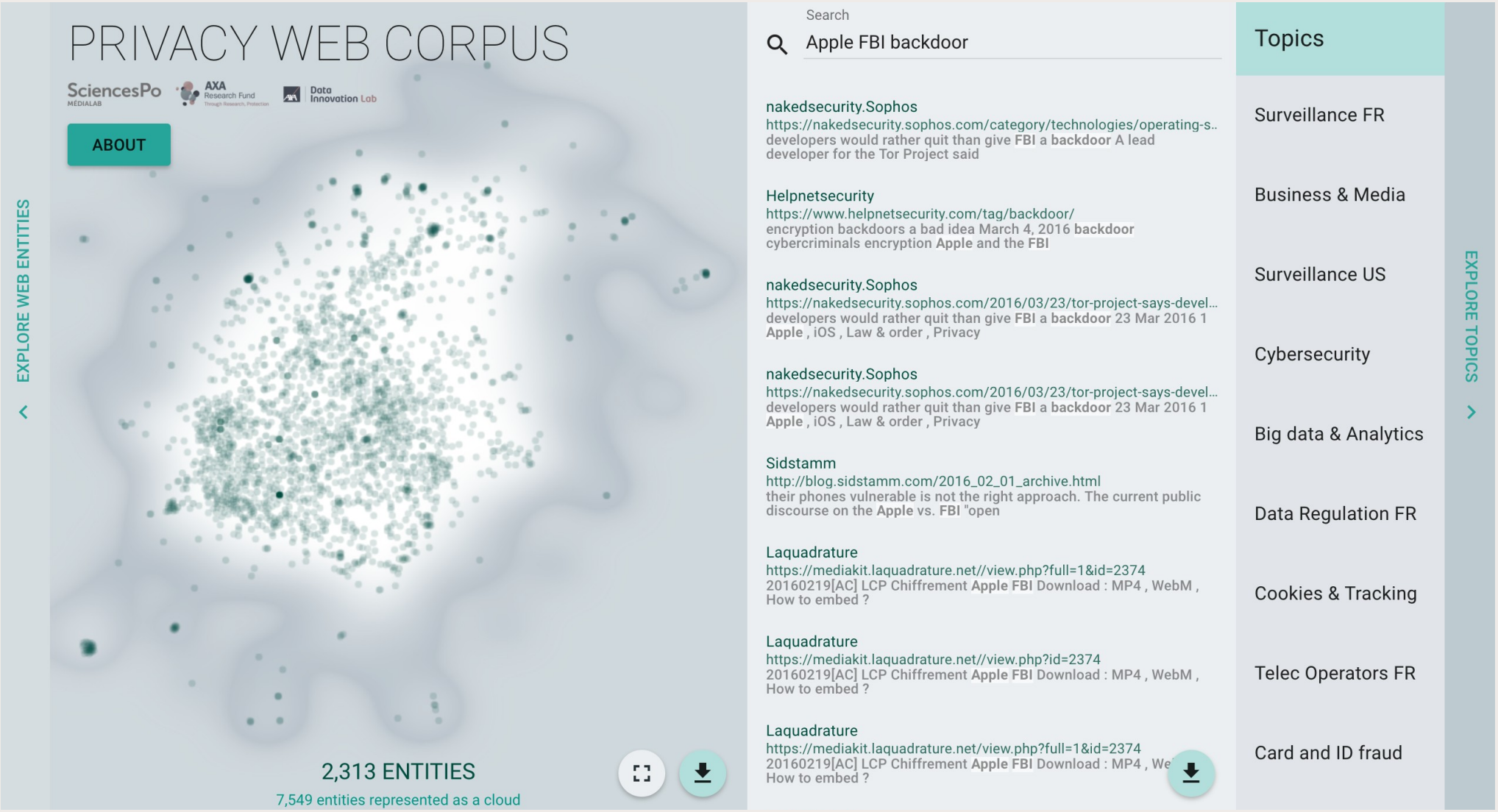
ADAPTATION

ADAPTATION + MITIGATION

MITIGATION

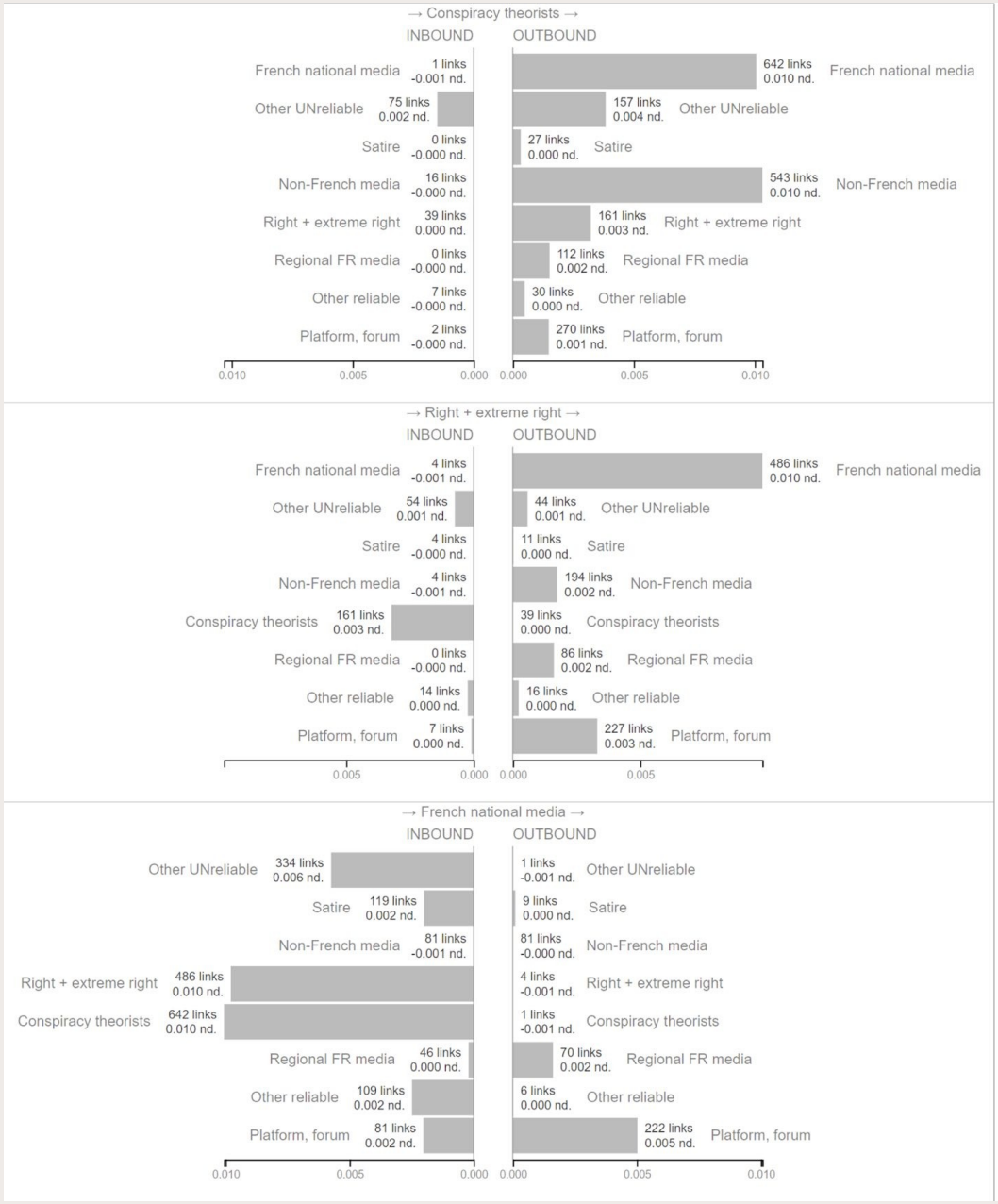
NOT RELEVANT, UNKNOWN OR NOT CATEGORIZED

Analyse de fond : traiter les contenus texte



http://tools.medialab.sciences-po.fr/privacy/

Analyser la circulation & la structuration des liens



Conspiracy theorists
- Not reliable
- Not cited
- Cite many reliable sources

other

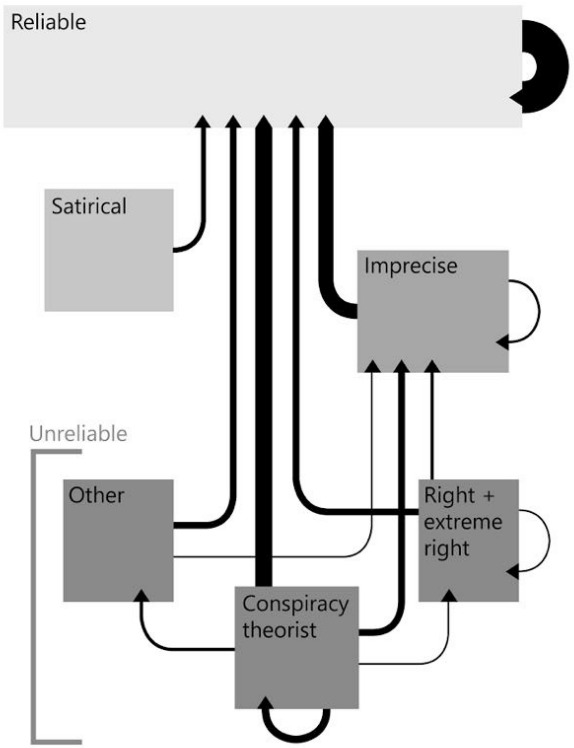
Right and extreme right
- Not reliable
- Cited only by conspirationists
- Cite many reliable sources

other

other

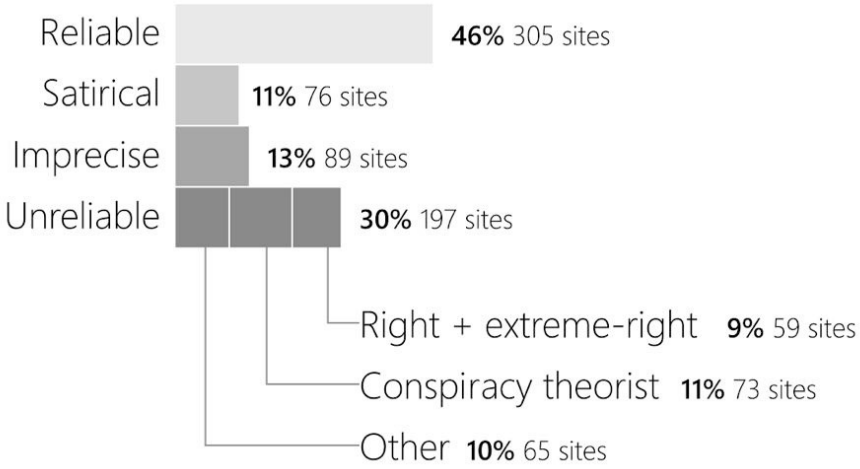
French national media
- Reliable *(*in this qualification)*
- Cited by everyone
- Cite only platforms and regional media

Exploiter la directionnalité des hyperliens

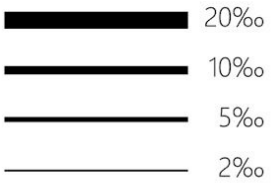


Most hyperlinks stem from the unreliable and aim at the reliable resources

Each bloc's surface is proportional to the count of websites. The color code si the same as the "Décodex".



The thickness si proportional to the normalized link density*

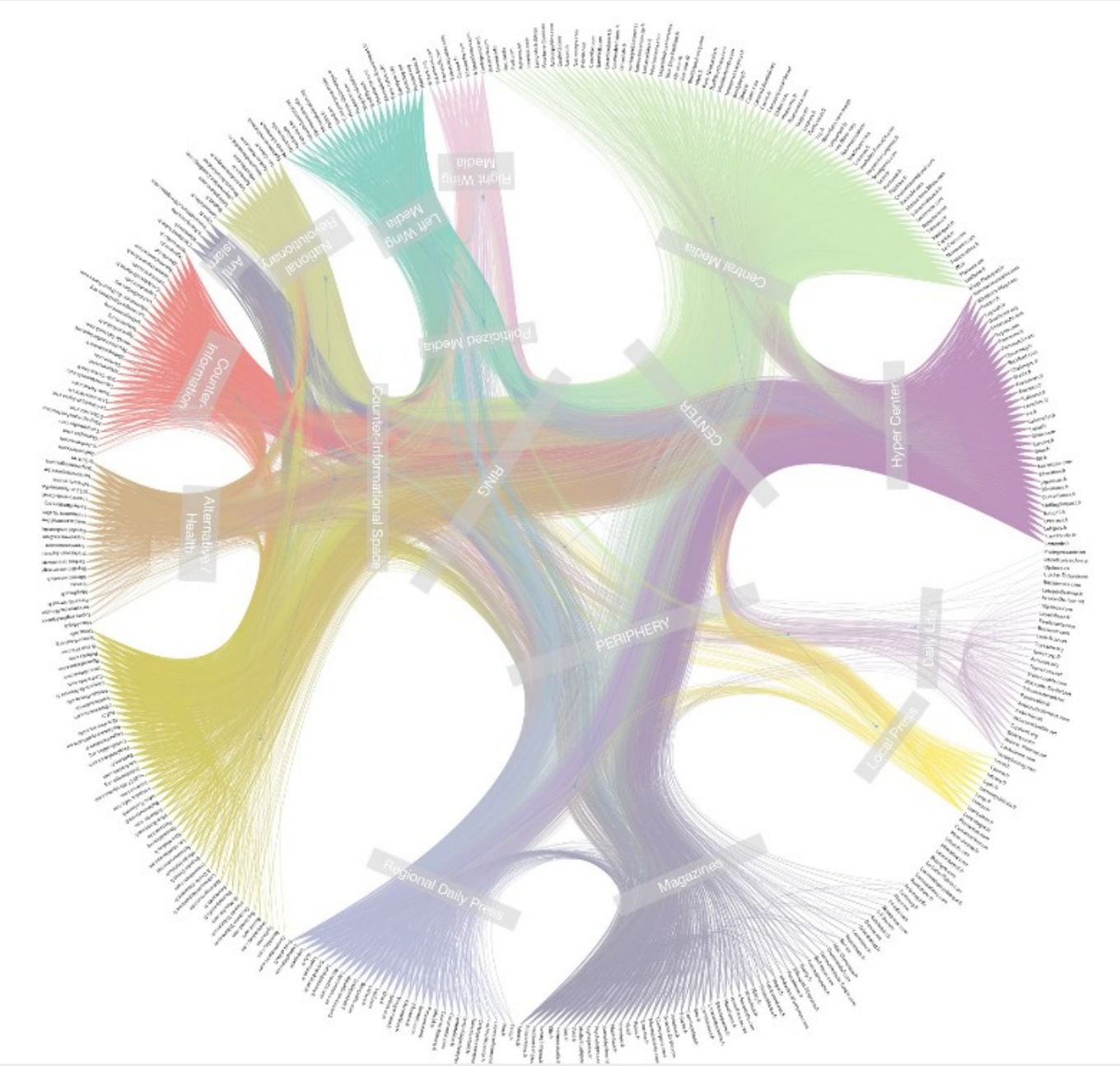


*Accounts for the different group sizes.

Densities below 2‰ have been omitted.

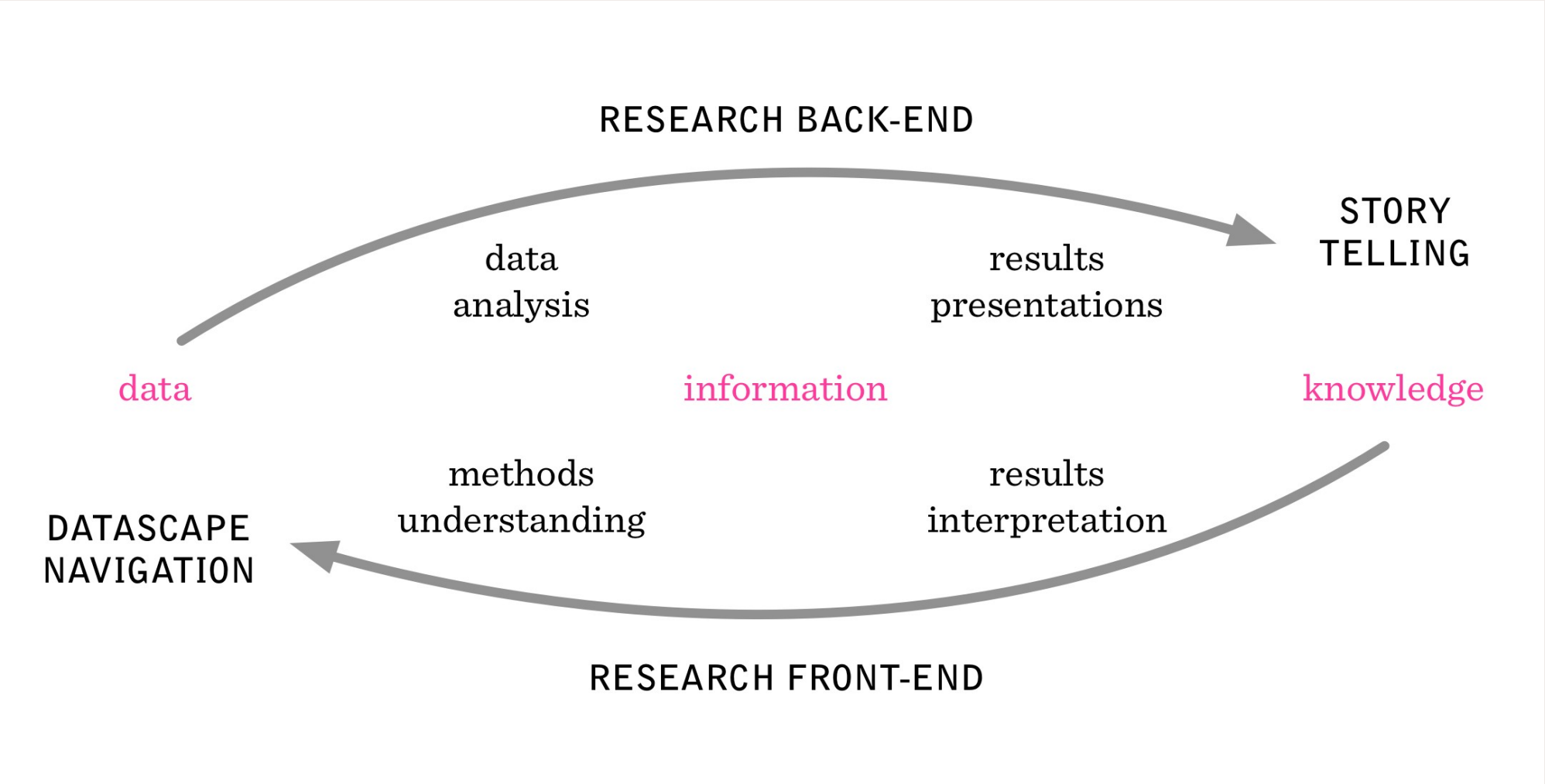
https://www.researchgate.net/publication/320225750_Visual_Network_Exploration_for_Data_Journalists

Explorer les dynamiques de polarisation



Les « datascapes » : zoomer/dézoomer

Construire des interfaces interactives d'exploration d'un jeu de données à différentes granularité

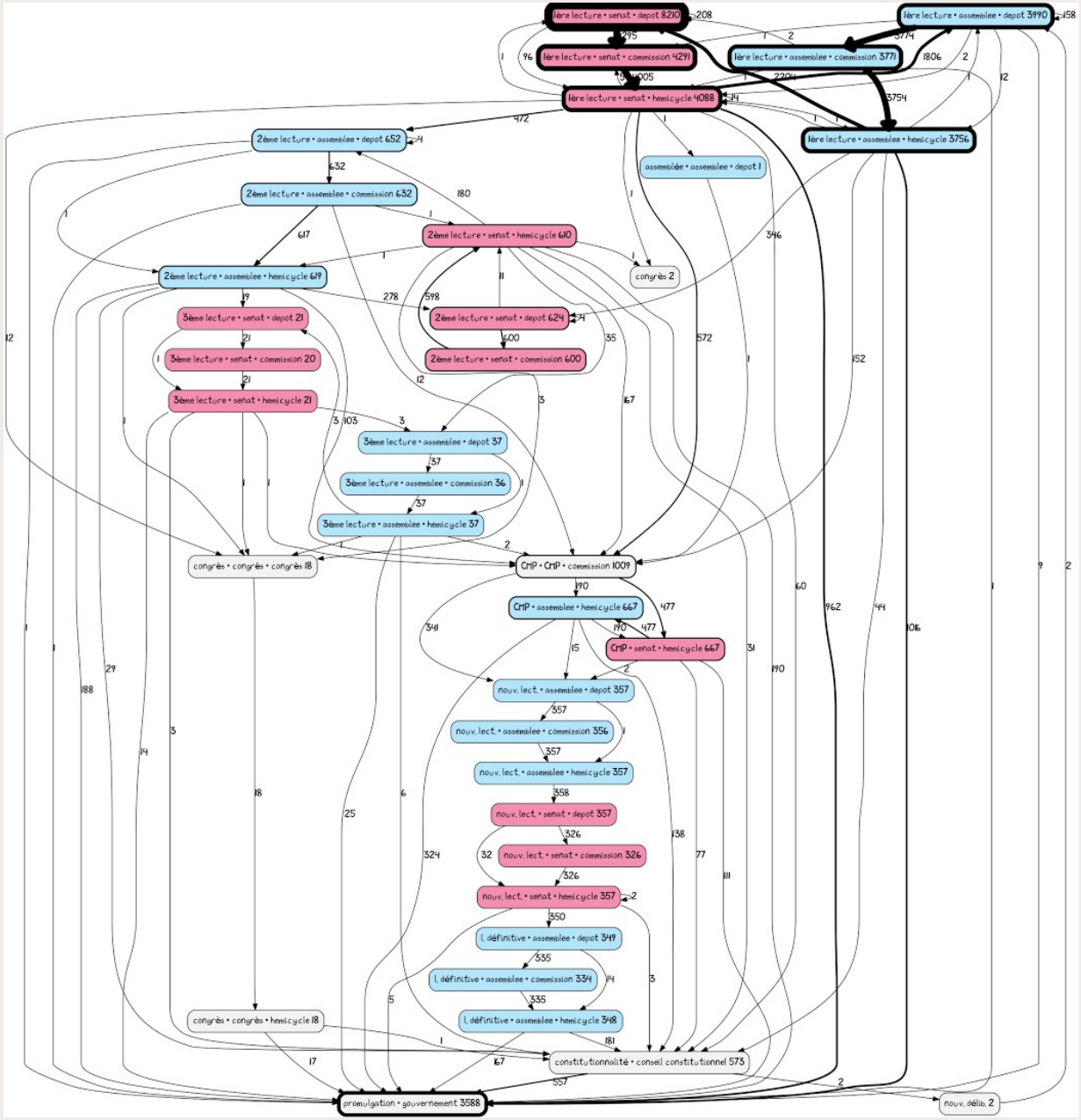


RICardo : visualiser les échanges au 19ème siècle

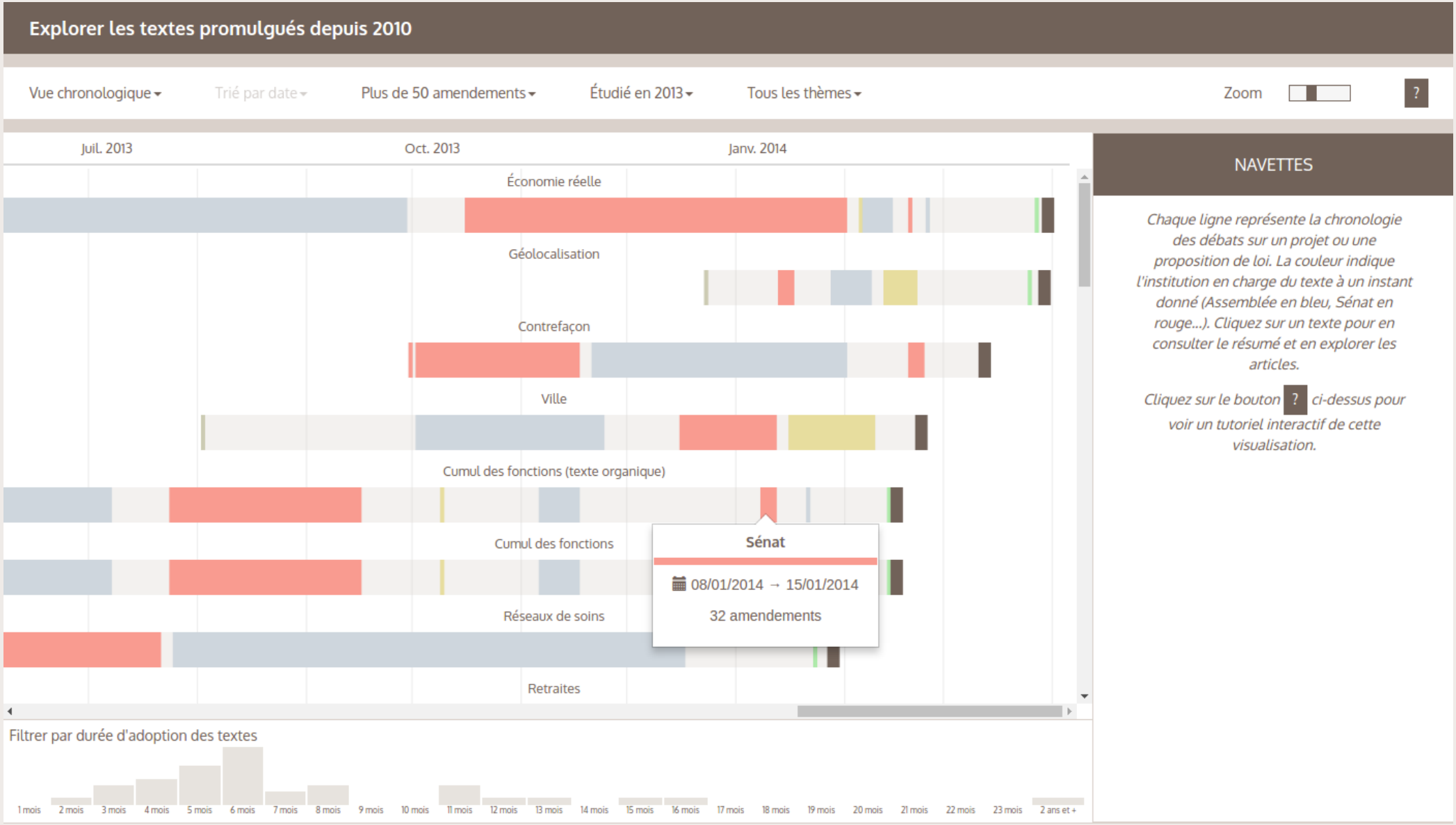
http://ricardo.medialab.sciences-po.fr

LaFabriqueDeLaLoi : explorer la complexité législative

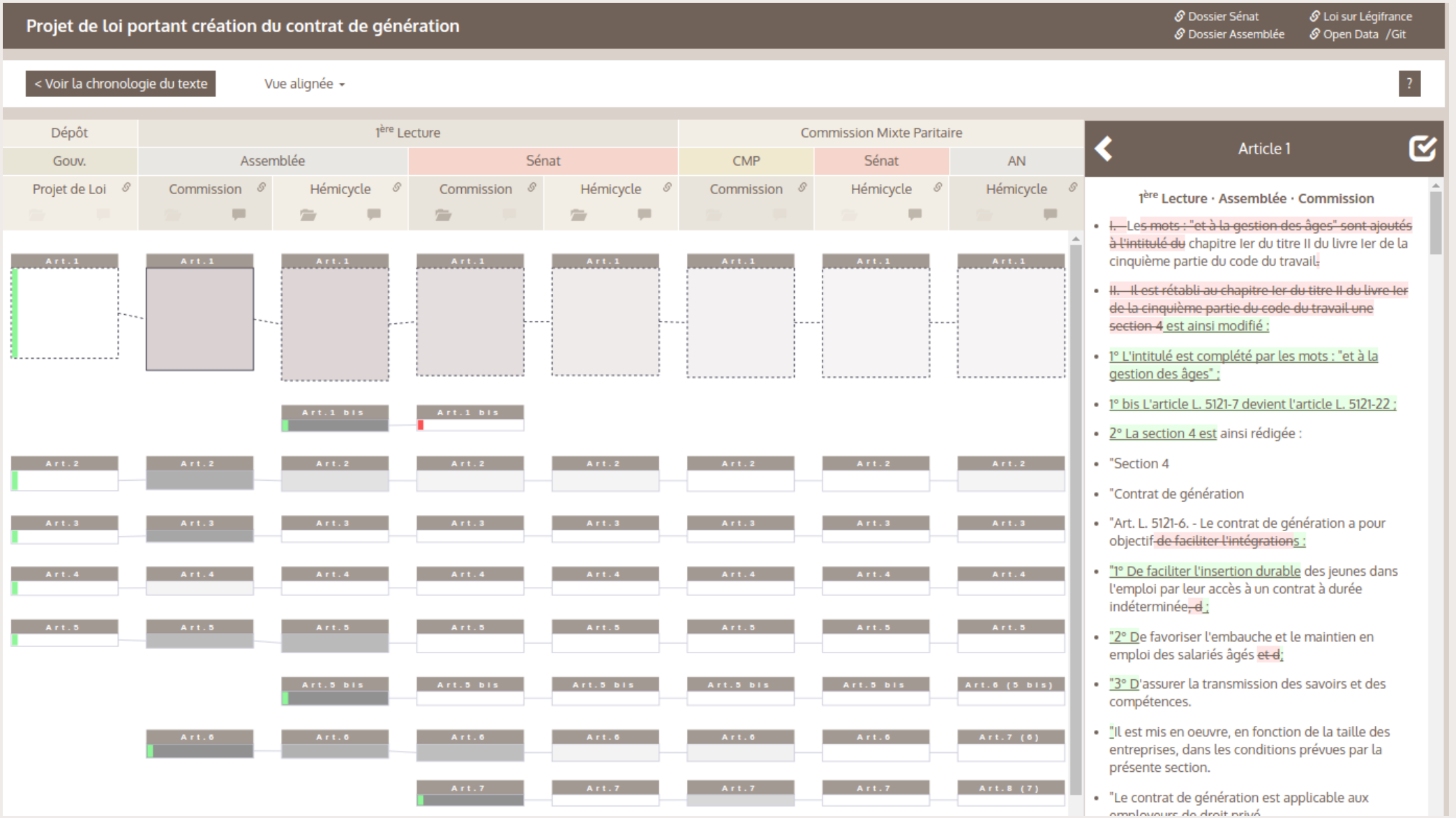
LA FABRIQUE DE LA LOI.FR



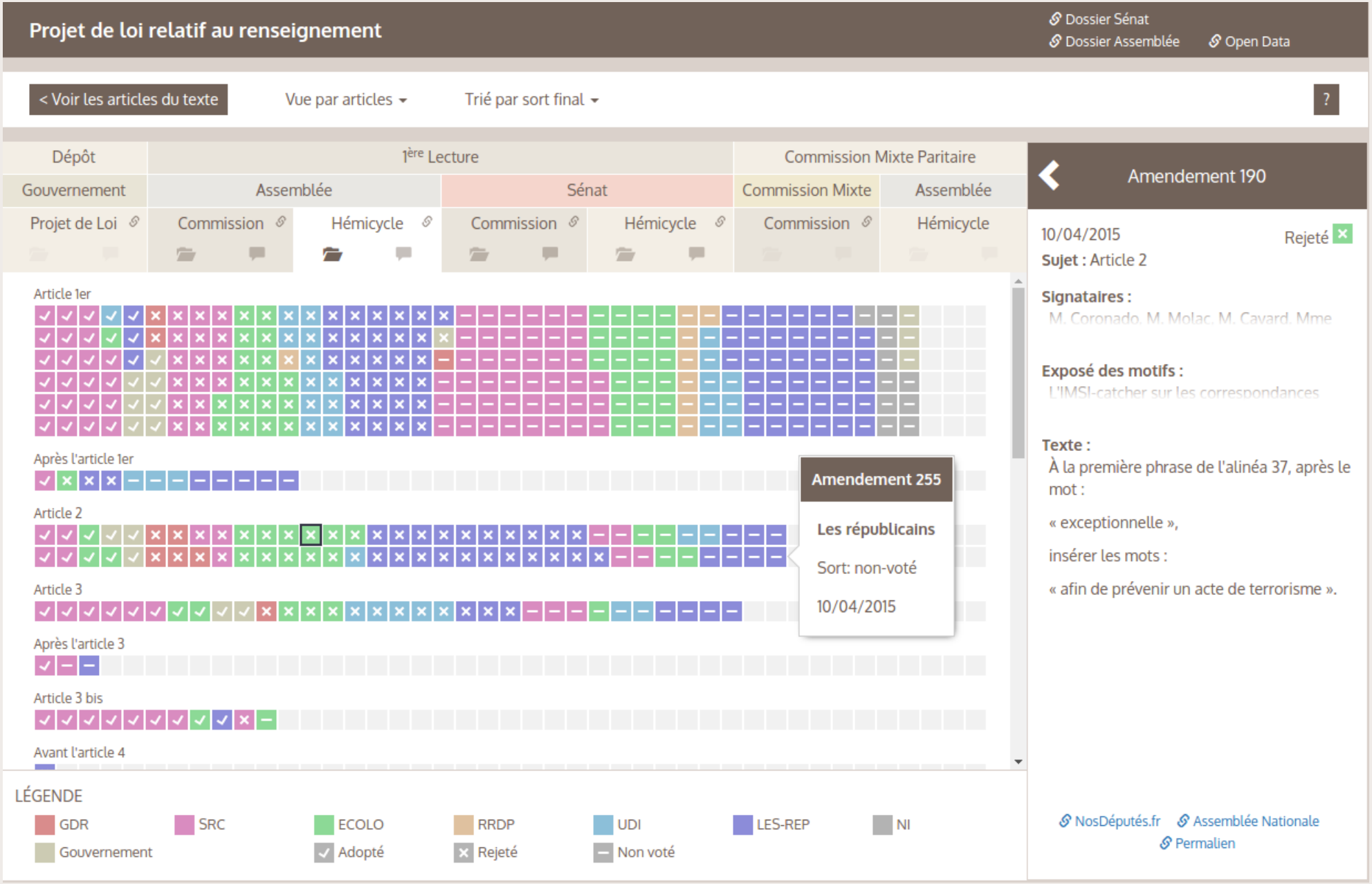
LFDLL : plus de 1000 lois promulguées ou en cours



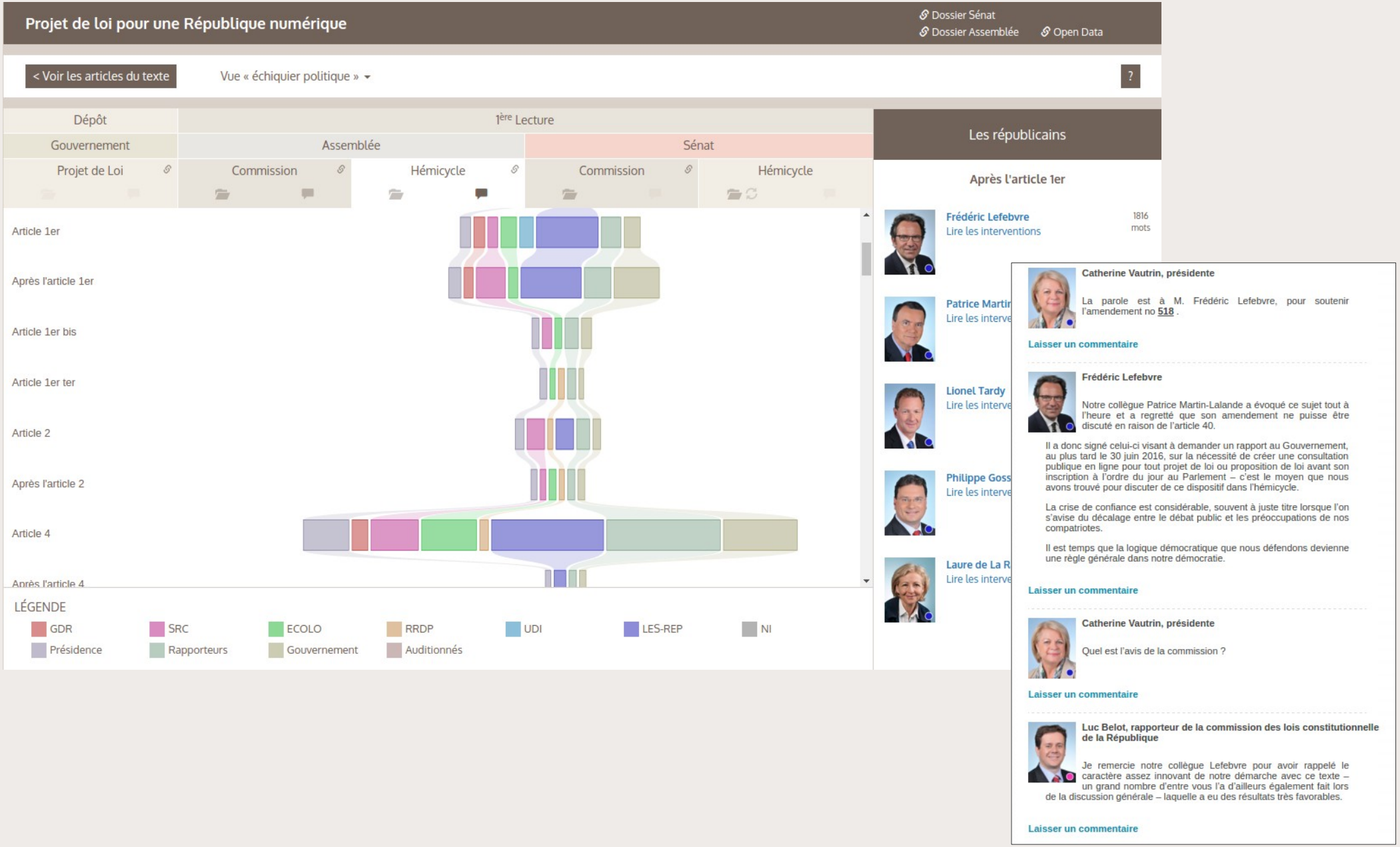
LFDLL : retracer l'évolution du texte à chaque étape



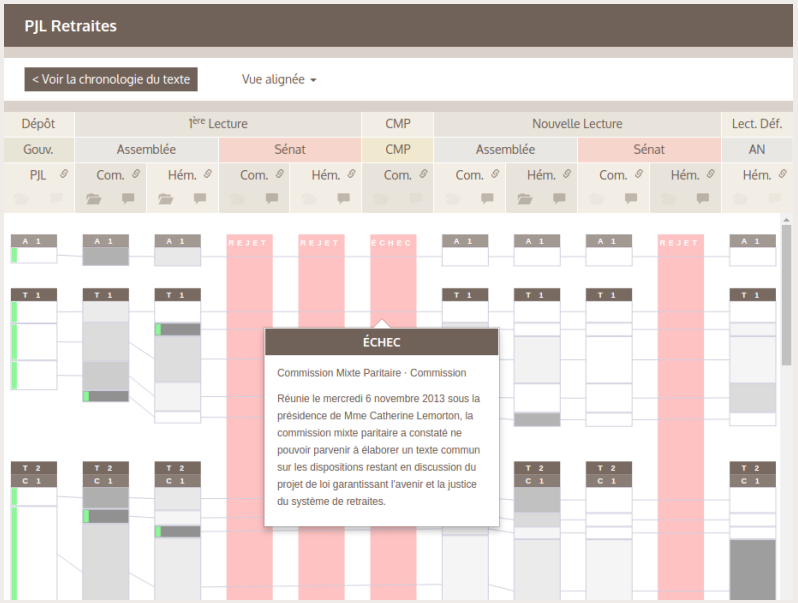
LFDLL : retrouver les amendements discutés et votés

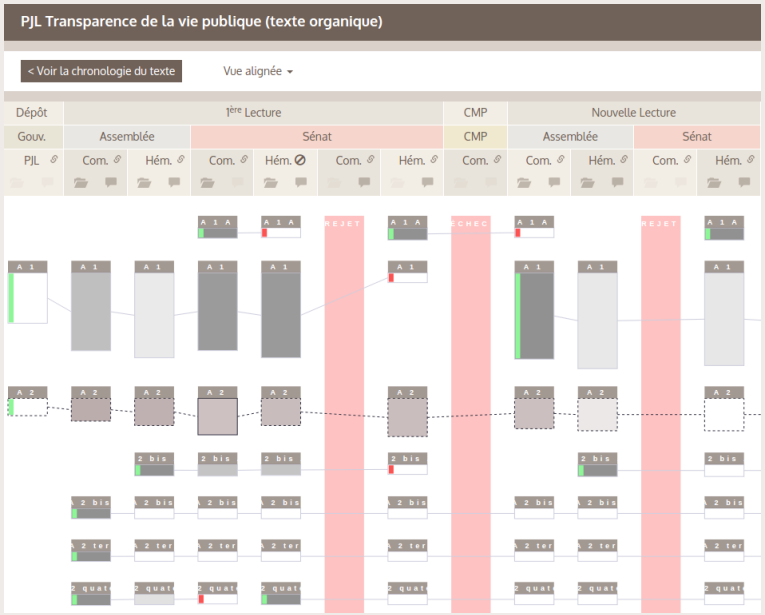


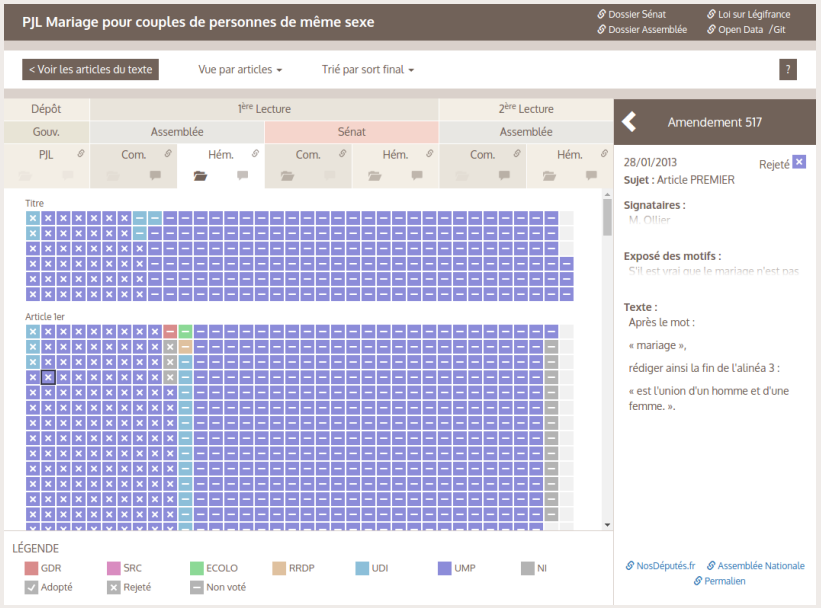
LFDLL : remonter aux débats parlementaires

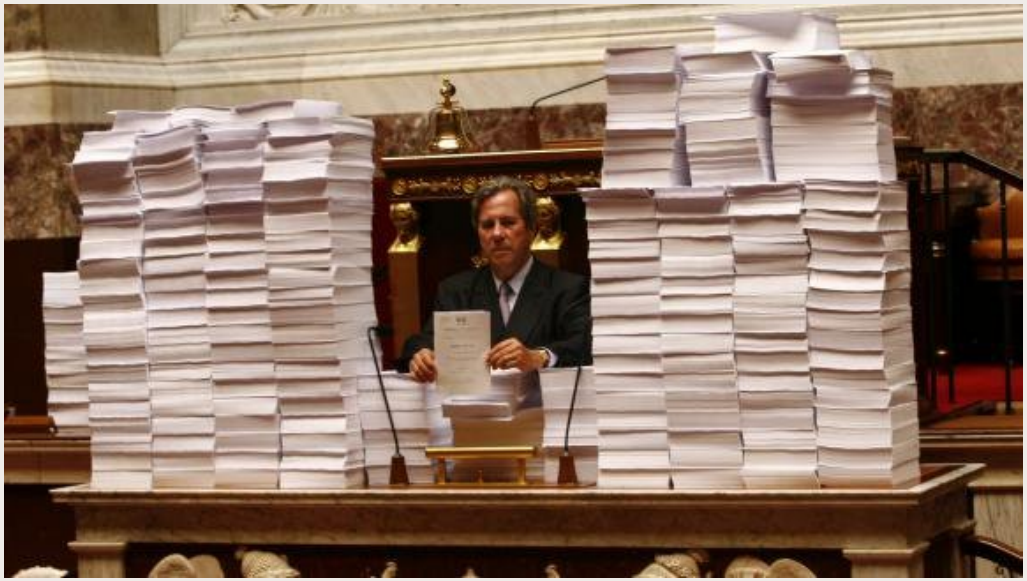


LFDLL : révéler les techniques d'obstruction

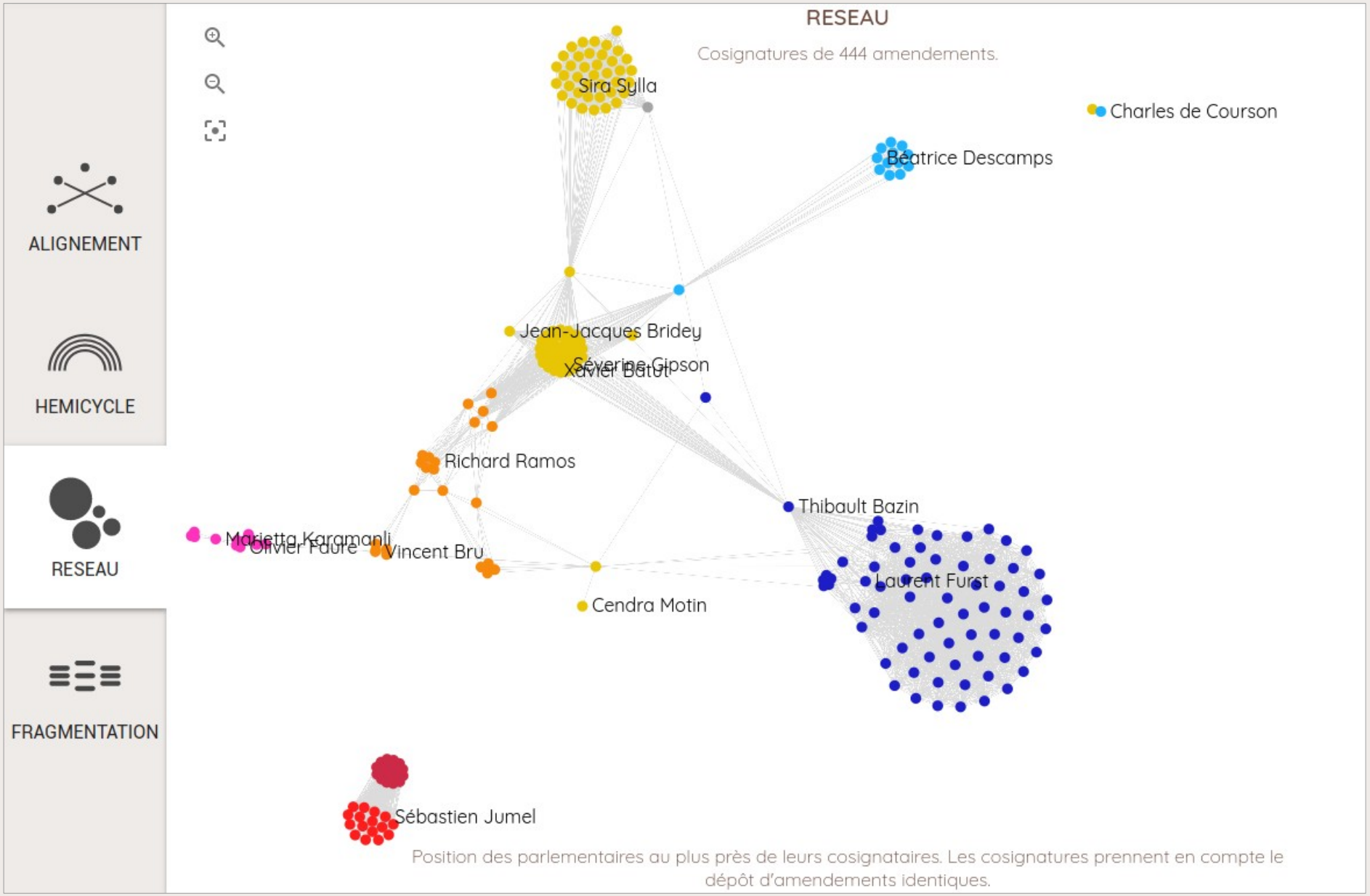




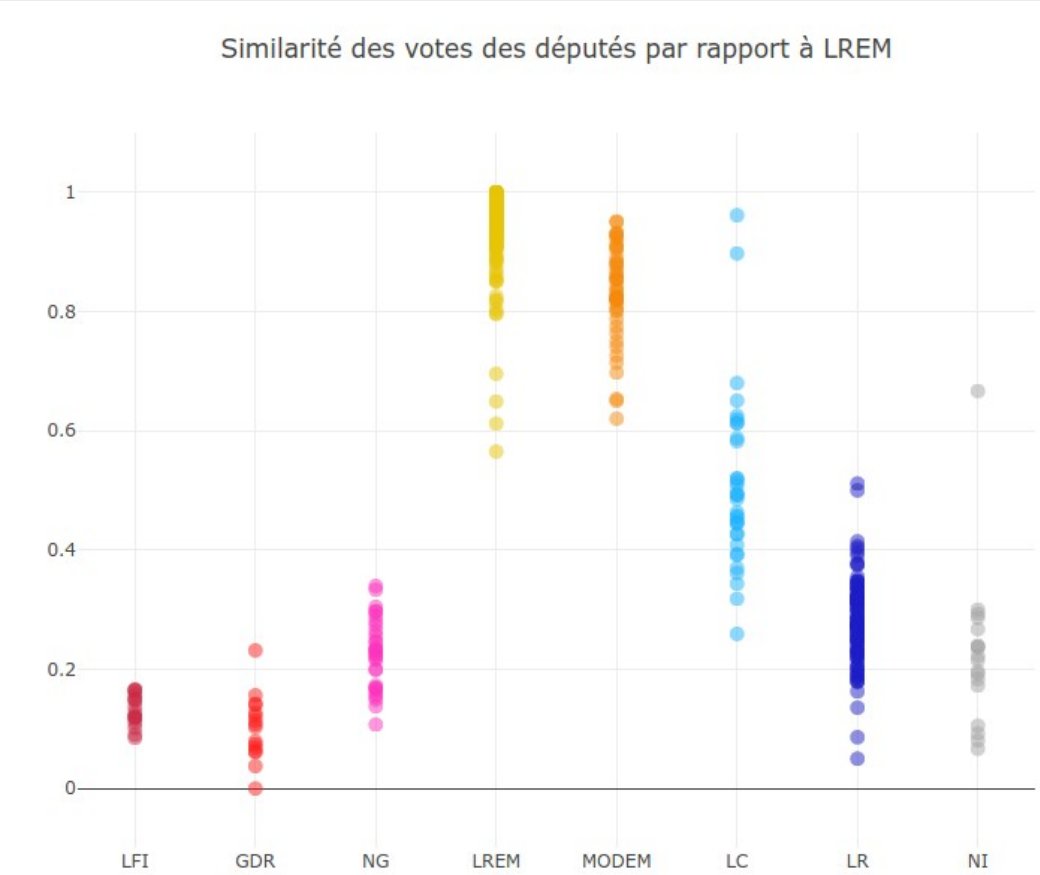


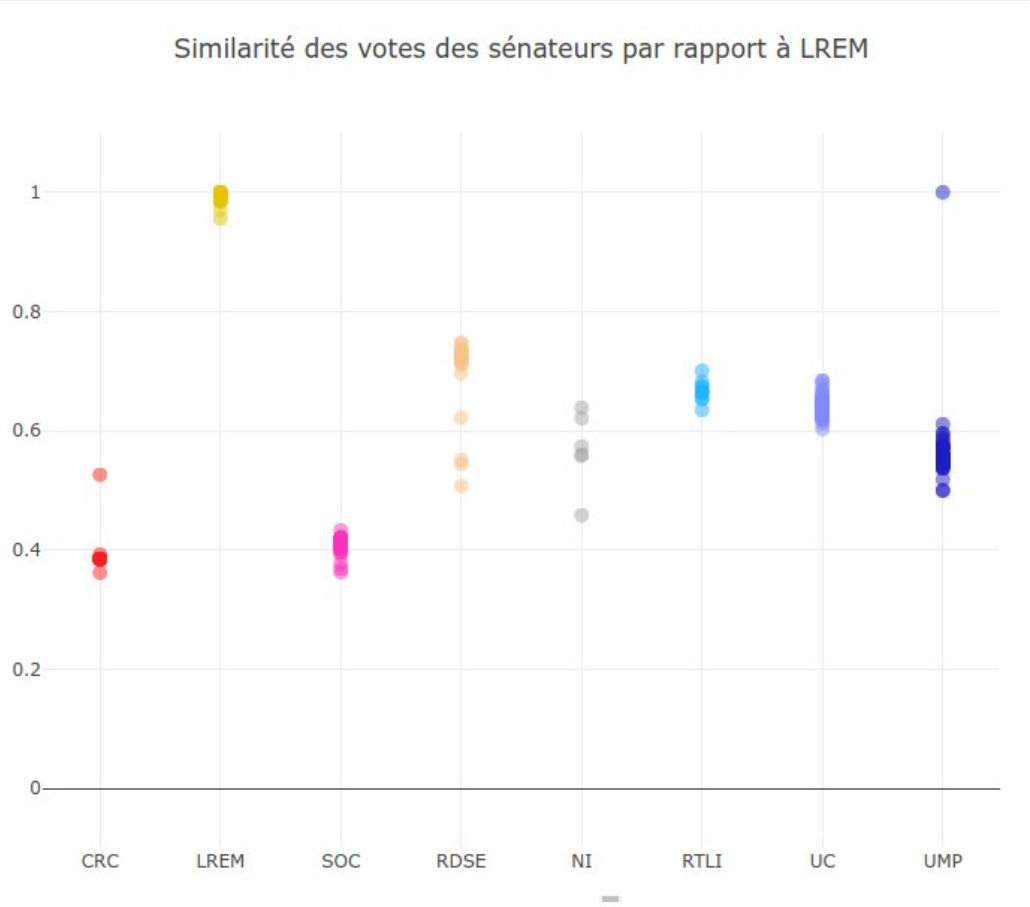


LFDLL : l'alignement et la fragmentation des partis

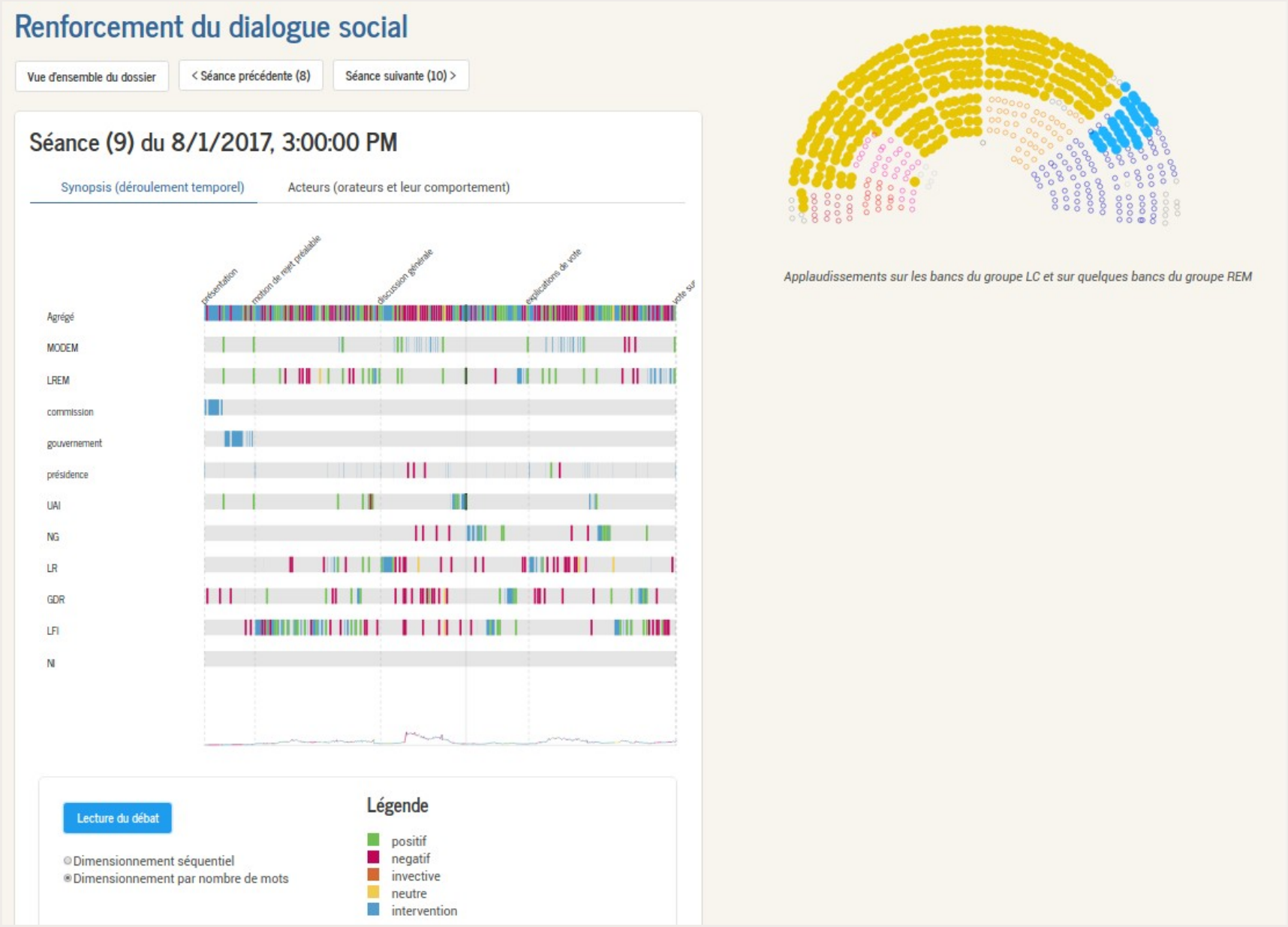


LFDLL : la loyauté politique au prisme des votes





LFDLL : visualiser la dynamique théâtrale des débats



Applaudissements sur les bancs du groupe LC et sur quelques bancs du groupe REM

À vos questions !

https://medialab.sciencespo.fr

benjamin.ooghe@sciencespo.fr

@boogheta **@medialab_ScPo**