



## Collecter et analyser des données du web et des réseaux sociaux pour les SHS

Benjamin Ooghe-Tabanou

### ► To cite this version:

Benjamin Ooghe-Tabanou. Collecter et analyser des données du web et des réseaux sociaux pour les SHS. Séminaire "Méthodologie de la recherche" du CIMEOS, Université de Bourgogne, Nov 2022, Dijon, France. <hal-03904254>

**HAL Id: hal-03904254**

**<https://sciencespo.hal.science/hal-03904254v1>**

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 CC BY-SA 4.0 - Attribution - ShareAlike - International License

# **Collecter et analyser des données du web et des réseaux sociaux pour les SHS**

Séminaire "Méthodologie de la recherche" du CIMEOS  
Université de Bourgogne – Dijon – 10 novembre 2022

Benjamin Ooghe-Tabanou (@boogheta)  
Sciences Po médialab (@medialab\_ScPo)

# Bruno Latour, fondateur du médialab



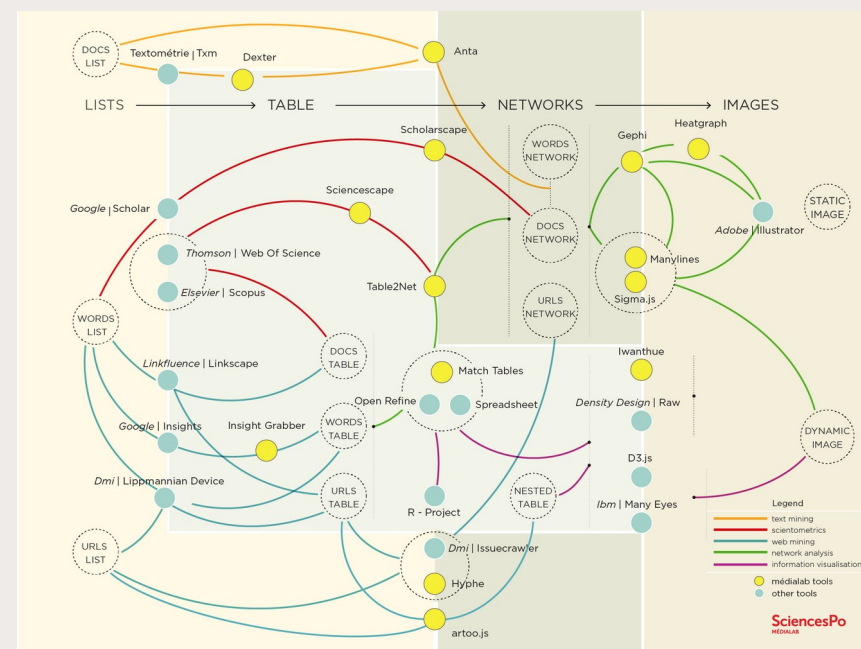
*« Google is nice,  
but we need  
something better »*

The Indian Express, 2011

# médialab @ Sciences Po

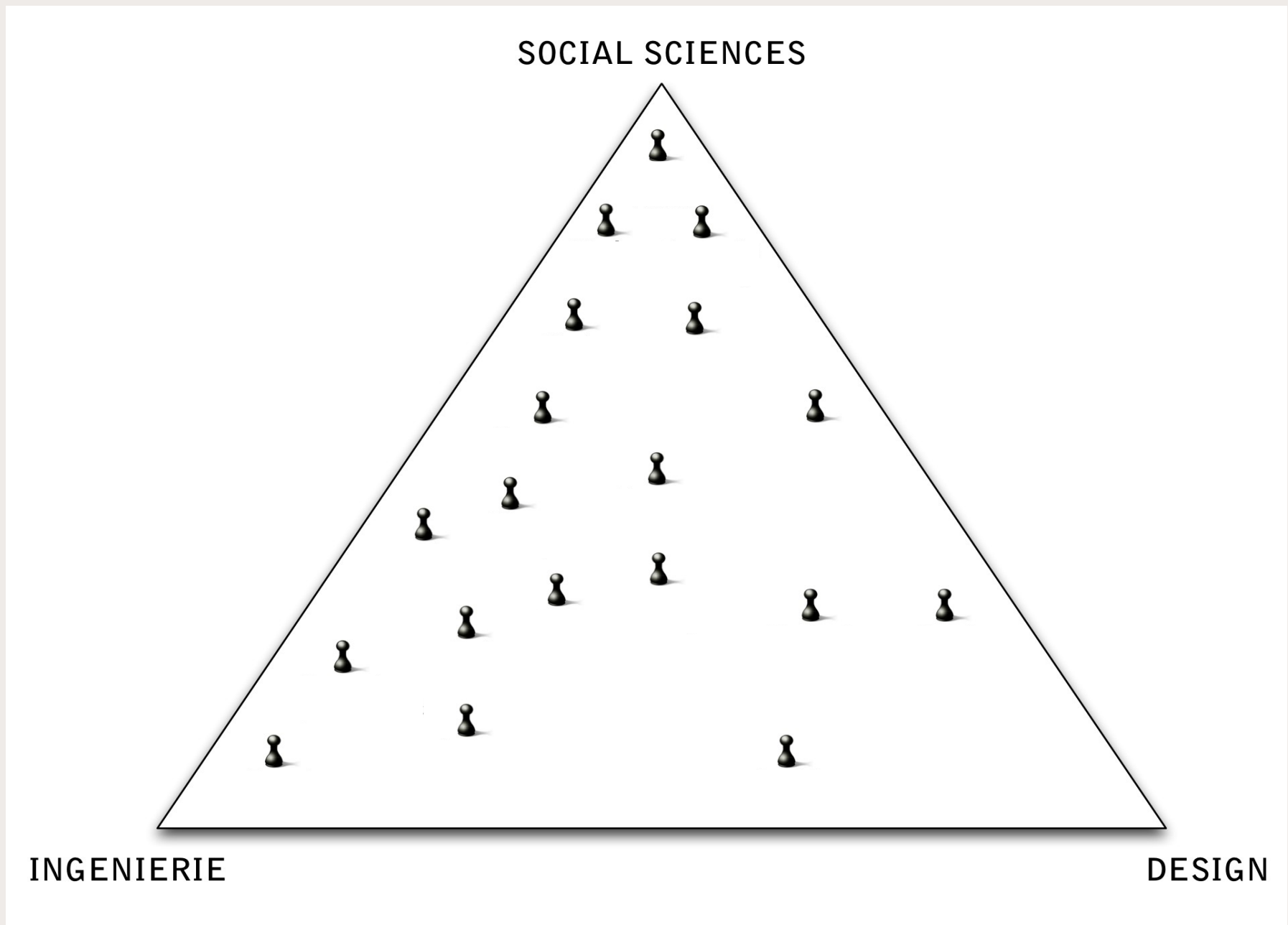
<https://medialab.sciencespo.fr>

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





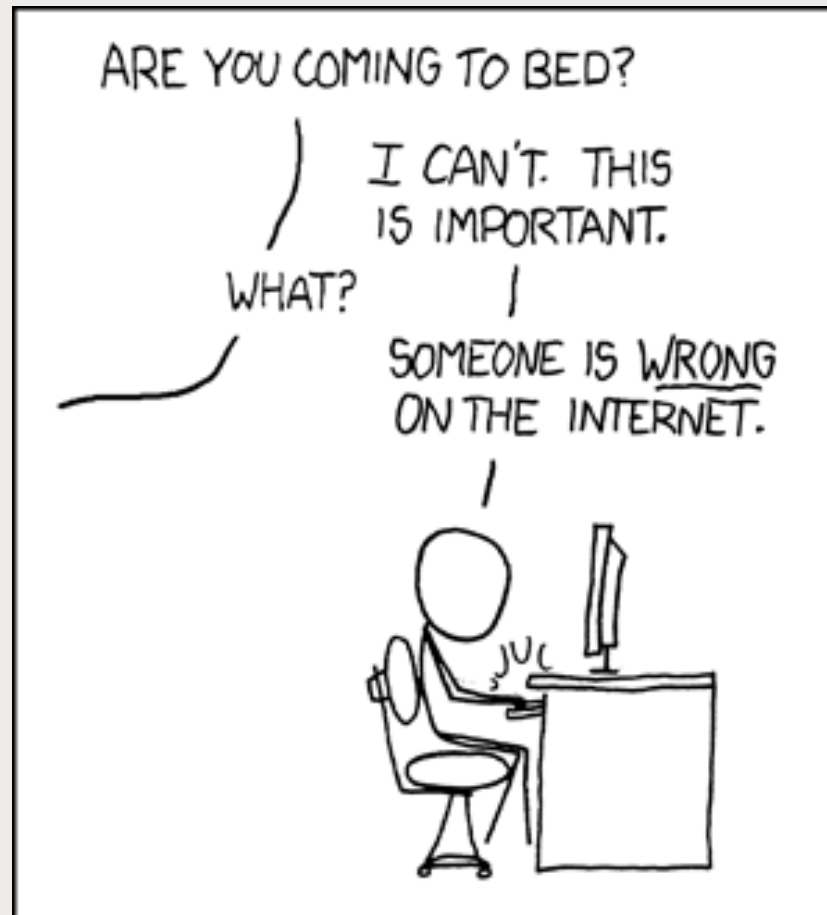
# Une équipe interdisciplinaire



# Exploiter le Web comme terrain d'enquêtes

Le Web : un espace de débats et de controverses

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

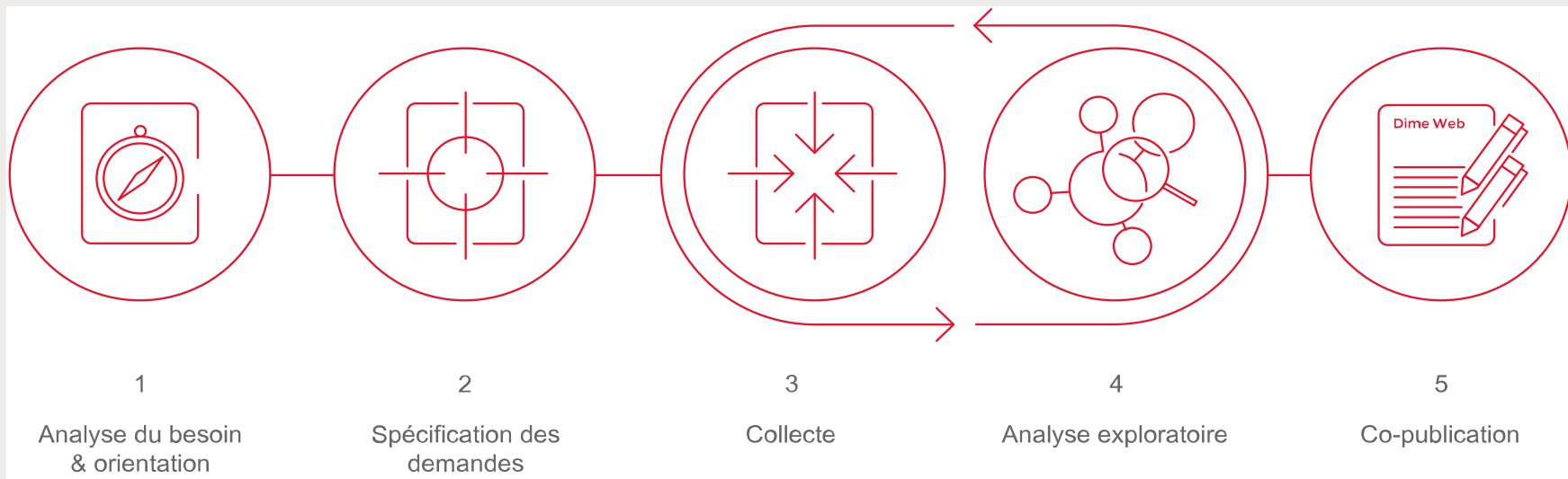


CC-BY-NC - Randall Munroe - XKCD

# Une méthodologie pour réaliser un terrain Web

Accompagnement technique et méthodologique à l'utilisation du Web comme nouveau terrain d'enquêtes

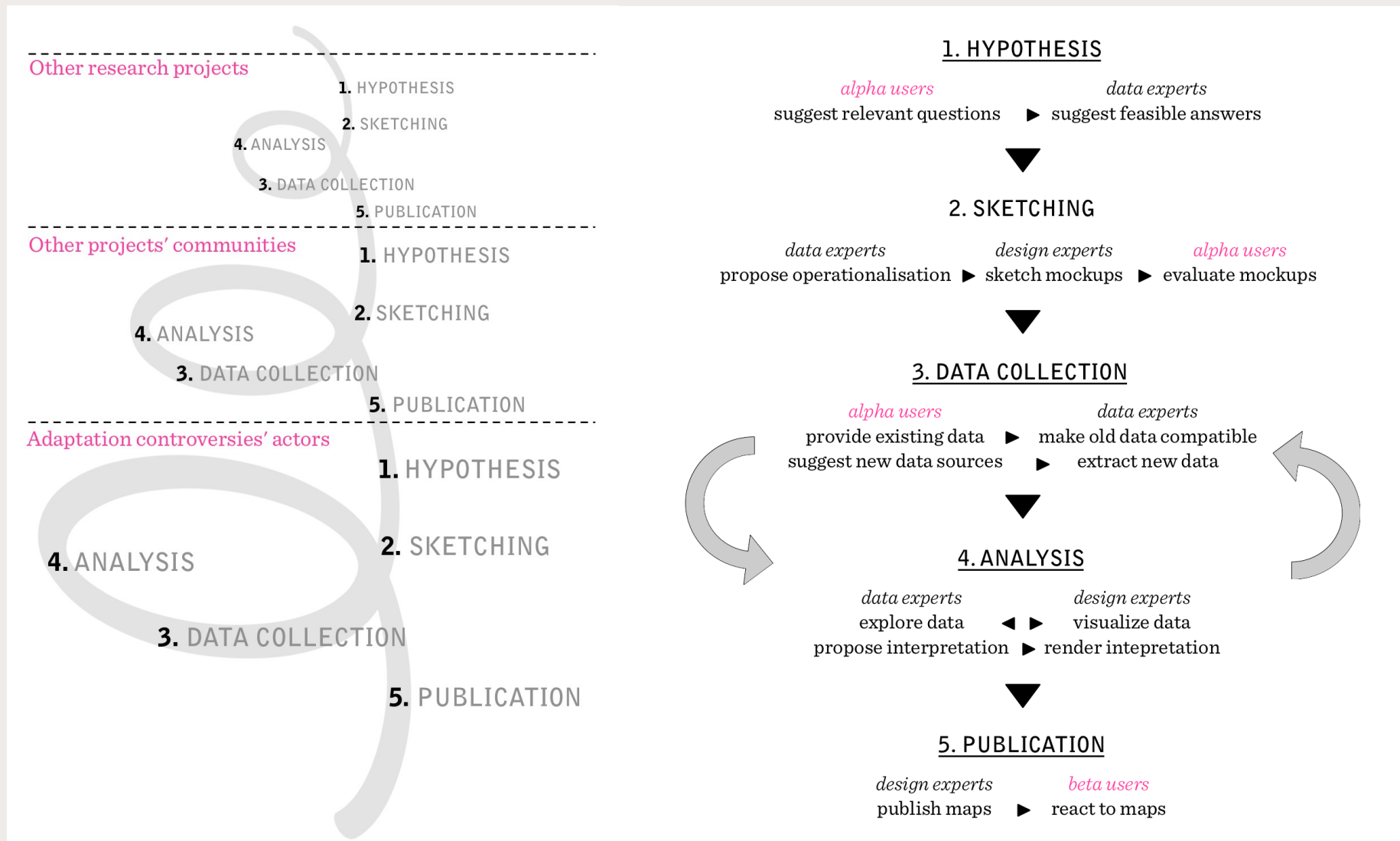
- Collecter, enrichir, nettoyer, visualiser et analyser des traces numériques
- Analyse de réseaux, archivage du web, analyse de controverses (ANT)
- Développement d'outils génériques
- Extraction ciblée de contenus
- Analyse Exploratoire de Données



# Une approche quali-quantitative exploratoire itérative

Numérique ≠ Magique

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



# Développer un écosystème d'outils OpenSource

<http://tools.medialab.sciences-po.fr>

- 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)
- **Documentation académique** et pratique des outils & méthodes  
(publications scientifiques, tutoriels, formations...)

# Organiser des « 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

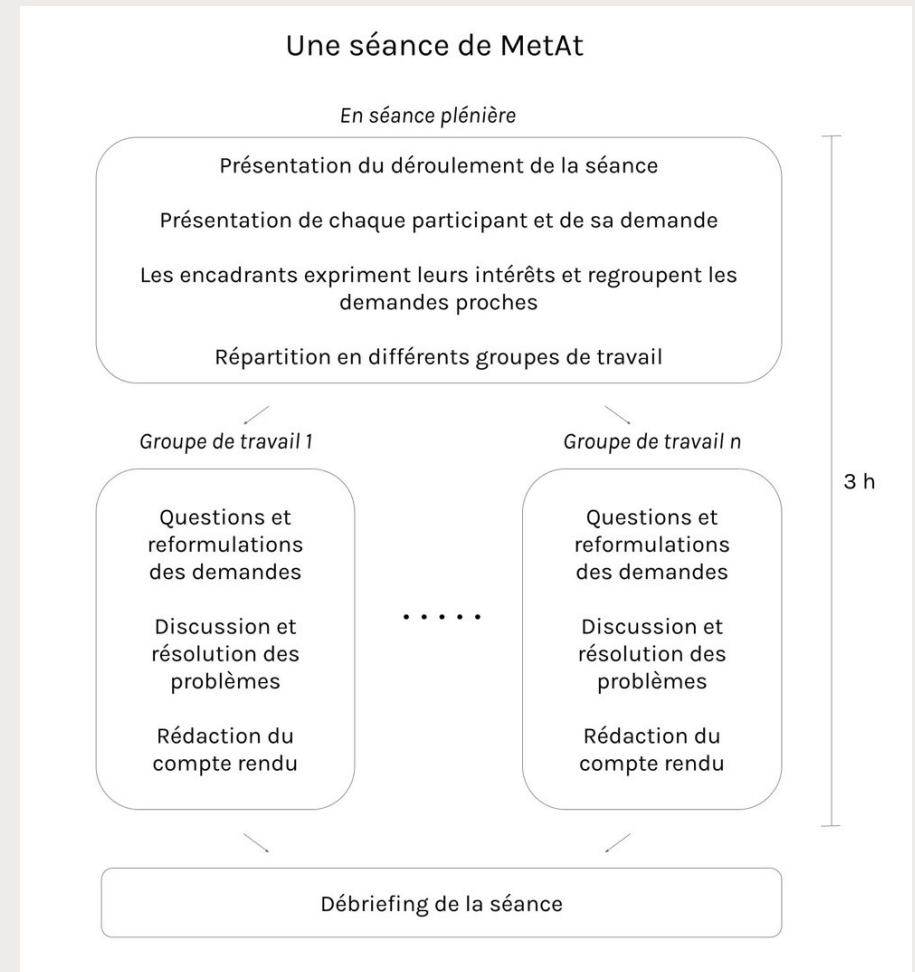




# Le MetAt, atelier de méthodes numériques

<https://www.sciencespo.fr/recherche/fr/content/metat-latelier-de-methodes>

- Demandes d'accompagnement :
  - discussion et conseil méthodologique
  - collecte & nettoyage de données
  - visualisation exploratoire
  - formation aux outils
  - ...
- Un mardi après-midi par mois
- Ouvert à tous, sur inscription préalable
- Initialement « atelier du médialab » :
  - canaliser les sollicitations
- Élargi à la communauté des ingénieurs de recherche de Sciences Po en 2017
- Contribue à l'autoformation continue



Diego Antolinos-Basso, Audrey Baneyx, Héloïse Théro, Benjamin Ooghe-Tabanou and Paul Girard, "L'atelier de méthodes de Sciences Po : apprendre, aider, rassembler", Humanités numériques, 5 | 2022, Online since 01 June 2022, connection on 09 November 2022.

<http://journals.openedition.org/revuehn/2799> DOI: <https://doi.org/10.4000/revuehn.2799>

# SHS et « Big Data » ?



Big Data = trop gros pour être manipulé par un ordinateur (> To)  
des données personnelles le plus souvent  
→ extrêmement rare en Sciences Humaines

# 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

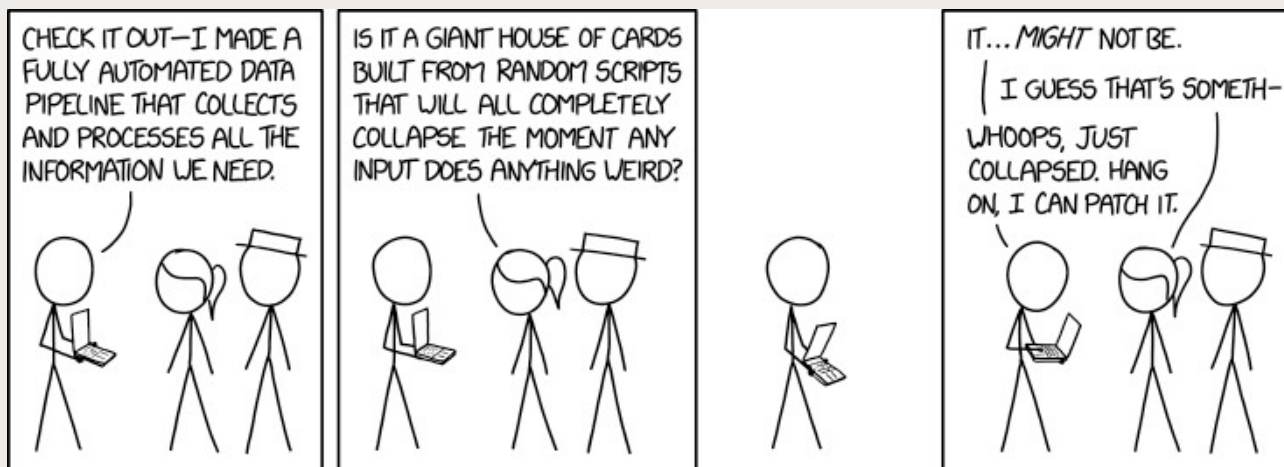
extraction ciblée  
(source unique ou ensemble cohérent)  
données structurées

traitement  
du langage

analyse de réseau  
(effets de communauté)

méthodes quantitatives,  
statistiques...

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

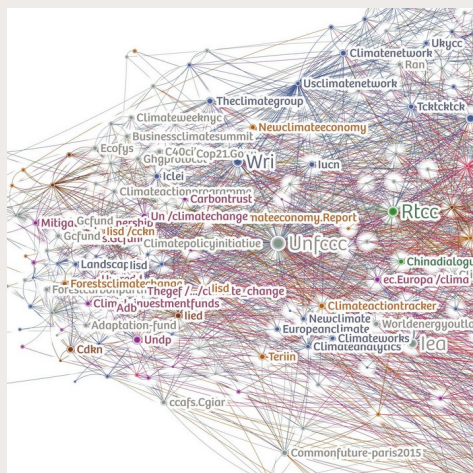


CC-BY-NC - Randall Munroe - XKCD

# Hyphe : un crawler orienté recherche

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

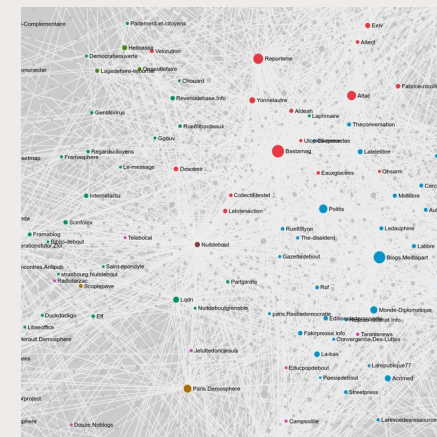
- Les liens hypertextes : nouveaux révélateurs de relations entre acteurs d'une thématique
- Créer un corpus documentaire
  - « acteurs web » & contenus textuels respectifs
  - liens hypertextes entre ces acteurs
- Études exploratoires ou de controverses dans tous les domaines



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

COP 21  
Vie privée  
Extrême droite  
Tissu associatif  
Produits laitiers  
Cellules souches  
Administrations culturelles

...



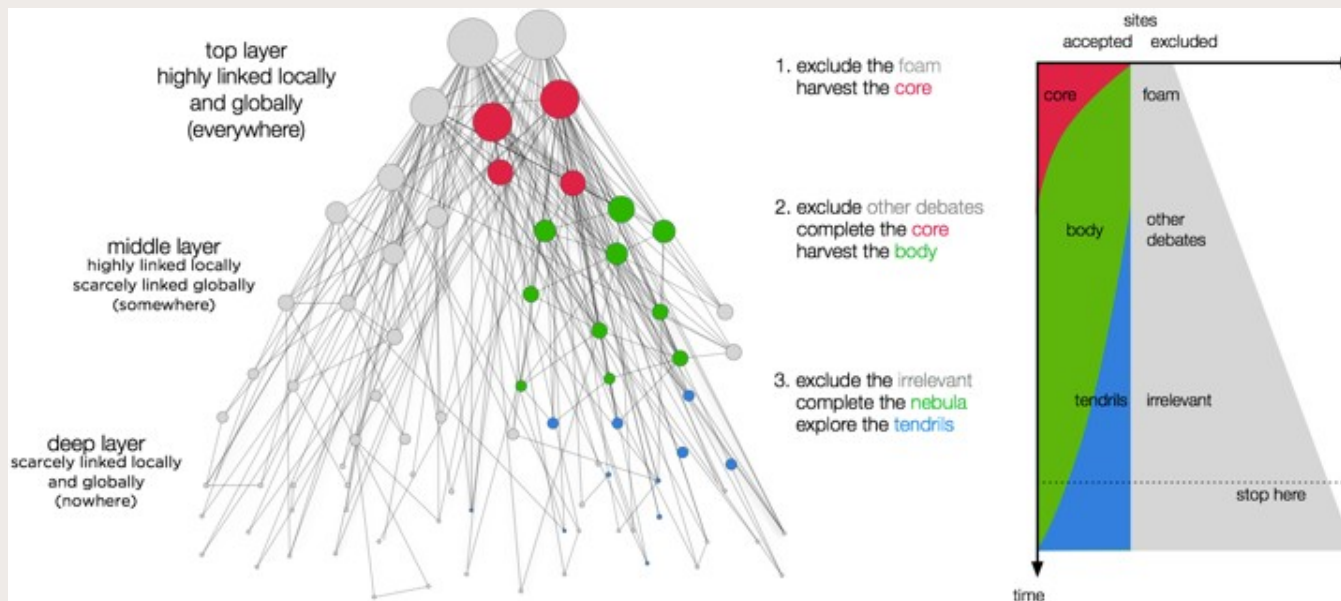
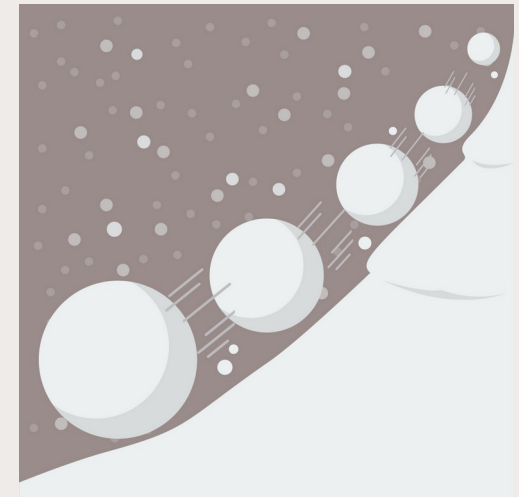
<http://utopies-concretes.org/>

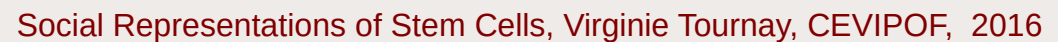
OOGHE-TABANOU, Benjamin, JACOMY, Mathieu, GIRARD, Paul & PLIQUE, Guillaume, "Hyperlink is not dead!", In Proceedings of the 2nd International Conference on Web Studies (WS.2 2018). ACM, New York, NY, USA, 12-18. DOI: <https://doi.org/10.1145/3240431.3240434>



# 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







# Cartographier le web autour d'un thème/débat

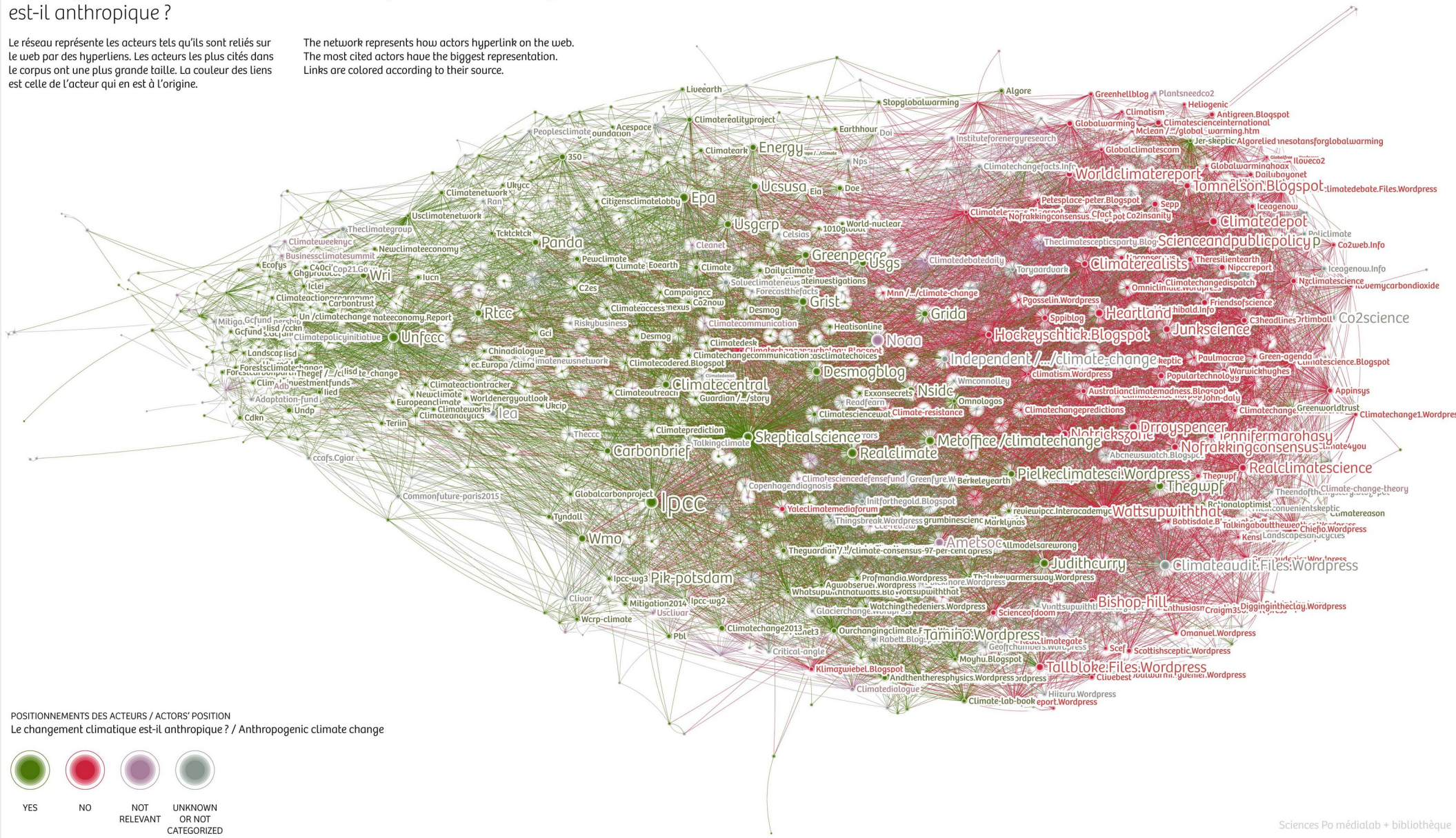
Corpus web sur le  
changement climatique :  
le changement climatique  
est-il anthropique ?

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.

<https://medialab.github.io/double-dating-data/#/>





# Cartographier le web par typologie d'acteurs

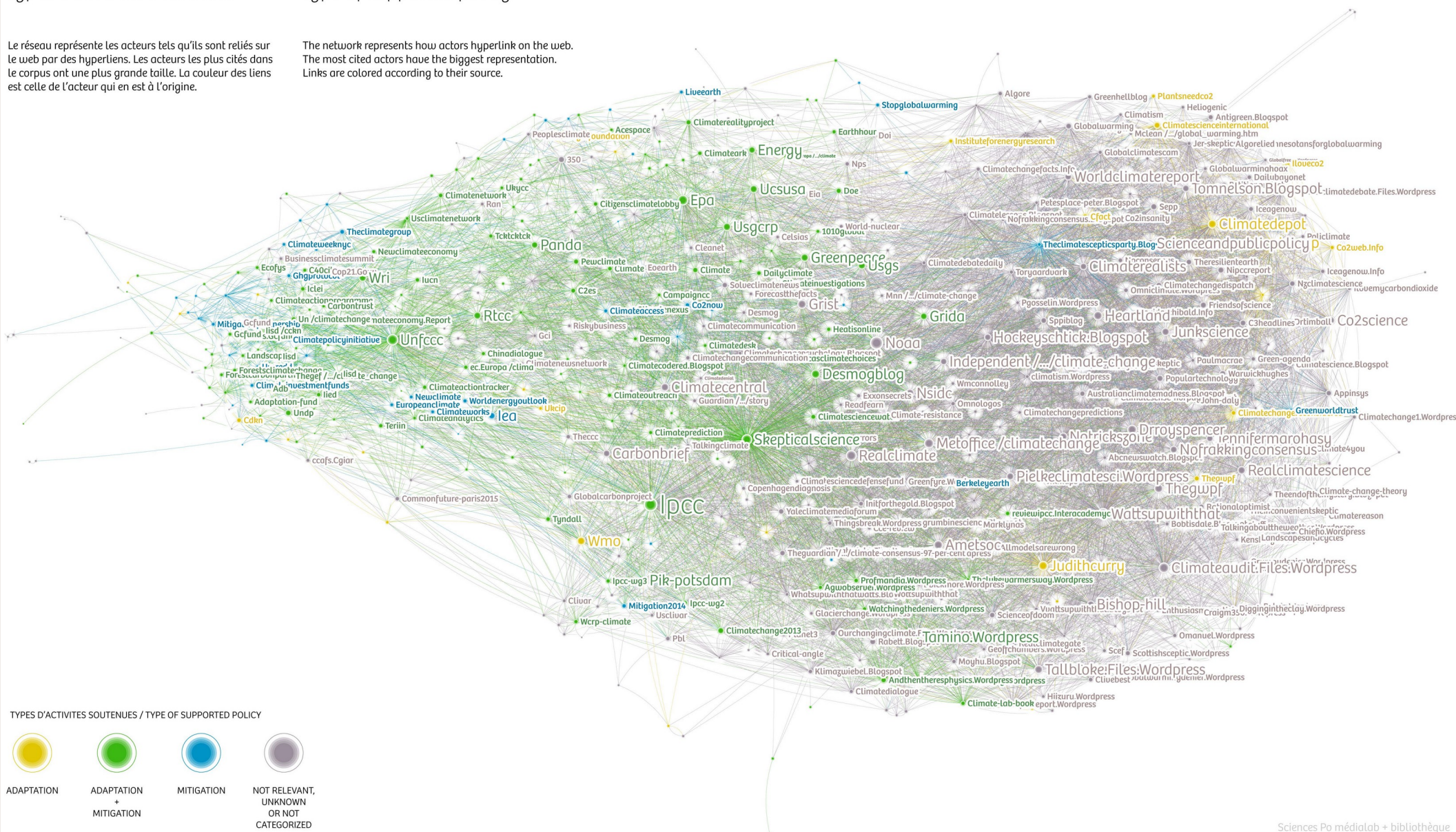
Corpus web sur le  
changement climatique :  
types d'activités soutenues

Web corpus  
on climate change:  
type of supported policy

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.

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

<https://medialab.github.io/double-dating-data/#/>





# Analyse de fond : traiter les contenus texte

PRIVACY WEB CORPUS

SciencesPo MÉDIALAB

AXA Research Fund Through Research, Protection

Data Innovation Lab

ABOUT

EXPLORE WEB ENTITIES

2,313 ENTITIES

7,549 entities represented as a cloud

Search

Q Apple FBI backdoor

nakedsecurity.Sophos

https://nakedsecurity.sophos.com/category/technologies/operating-s.. developers would rather quit than give FBI a backdoor A lead developer for the Tor Project said

Helpnetsecurity

https://www.helpnetsecurity.com/tag/backdoor/ encryption backdoors a bad idea March 4, 2016 backdoor cybercriminals encryption Apple and the FBI

nakedsecurity.Sophos

https://nakedsecurity.sophos.com/2016/03/23/tor-project-says-devel... developers would rather quit than give FBI a backdoor 23 Mar 2016 1 Apple , iOS , Law & order , Privacy

nakedsecurity.Sophos

https://nakedsecurity.sophos.com/2016/03/23/tor-project-says-devel... developers would rather quit than give FBI a backdoor 23 Mar 2016 1 Apple , iOS , Law & order , Privacy

Sidstamm

http://blog.sidstamm.com/2016\_02\_01\_archive.html their phones vulnerable is not the right approach. The current public discourse on the Apple vs. FBI "open

Laquadrature

https://mediakit.laquadrature.net/view.php?full=1&id=2374 20160219[AC] LCP Chiffrement Apple FBI Download : MP4 , WebM , How to embed ?

Laquadrature

https://mediakit.laquadrature.net/view.php?id=2374 20160219[AC] LCP Chiffrement Apple FBI Download : MP4 , WebM , How to embed ?

Laquadrature

https://mediakit.laquadrature.net/view.php?full=1&id=2374 20160219[AC] LCP Chiffrement Apple FBI Download : MP4 , We How to embed ?

Topics

Surveillance FR

Business & Media

Surveillance US

Cybersecurity

Big data & Analytics

Data Regulation FR

Cookies & Tracking

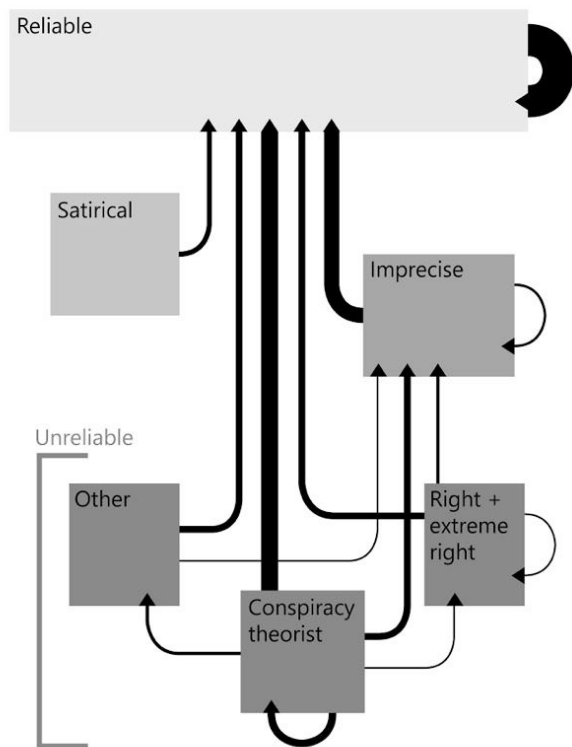
Telec Operators FR

Card and ID fraud

EXPLORE TOPICS

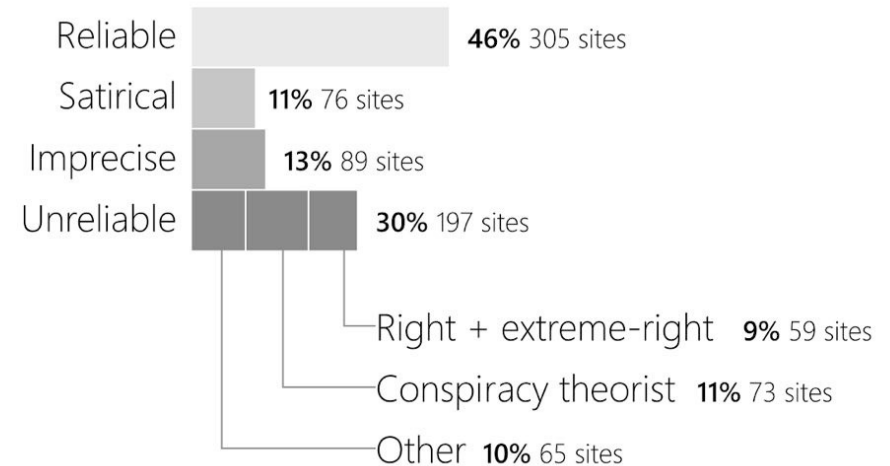
<http://tools.medialab.sciences-po.fr/privacy/>

# 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 is the same as the "Décodex".



The thickness is proportional to the normalized link density\*



\*Accounts for the different group sizes.  
Densities below 2‰ have been omitted.

Venturini, Tommaso & Jacomy, Mathieu & Bounegru, Liliana & Gray, Jonathan. (2018). **Visual Network Exploration for Data Journalists**.  
[https://www.researchgate.net/publication/320225750\\_Visual\\_Network\\_Exploration\\_for\\_Data\\_Journalists](https://www.researchgate.net/publication/320225750_Visual_Network_Exploration_for_Data_Journalists)

# Exemple de scraping ciblé : Google Bookmarklets

<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

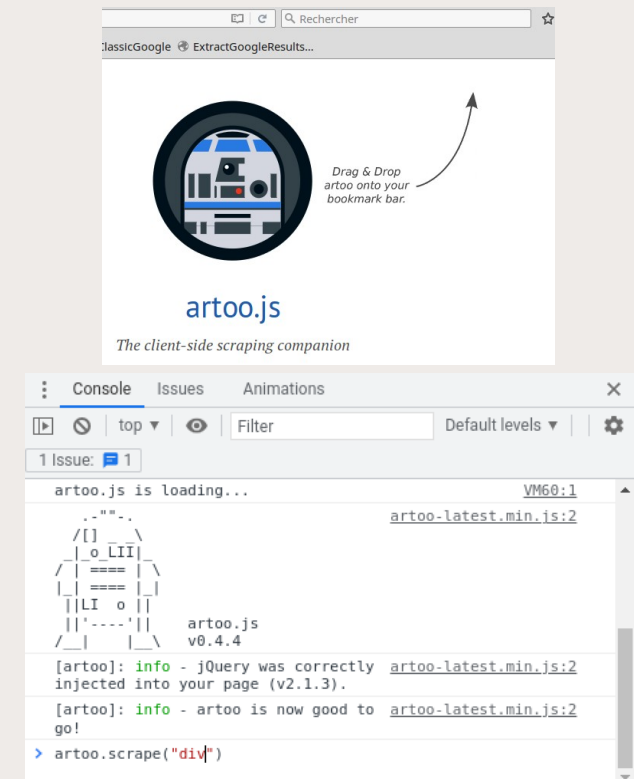
The image is a collage illustrating the Google Bookmarklets workflow for targeted scraping. It includes the following elements:

- Installation Page:** A screenshot of the 'Install Google Bookmarklets' page, showing instructions to drag and drop icons into the browser's bookmark bar.
- Google Search:** A screenshot of a Google search for 'digital humanities', showing the search bar, filters, and search results.
- Redirect Dialog:** A screenshot of the 'Redirect to Classic Google' dialog, which allows users to select a language and the number of results per page (set to 100).
- Extract Results Dialog:** A screenshot of the 'Extract Classic Google Results' dialog, which shows the search results and provides options to 'Keep existing results & continue to the next page' or 'Download CSV with 103 urls'.
- Output Format:** A black box at the bottom right displays the output format: `url, name, row, description, date`.

# artoo.js : extraire des données du web (avancé)

<https://medialab.github.io/artoo/>

- Un bookmarklet à ajouter dans la barre de favoris du navigateur
- Une librairie JavaScript de fonctions utiles pour le scraping depuis la console du navigateur (F12)



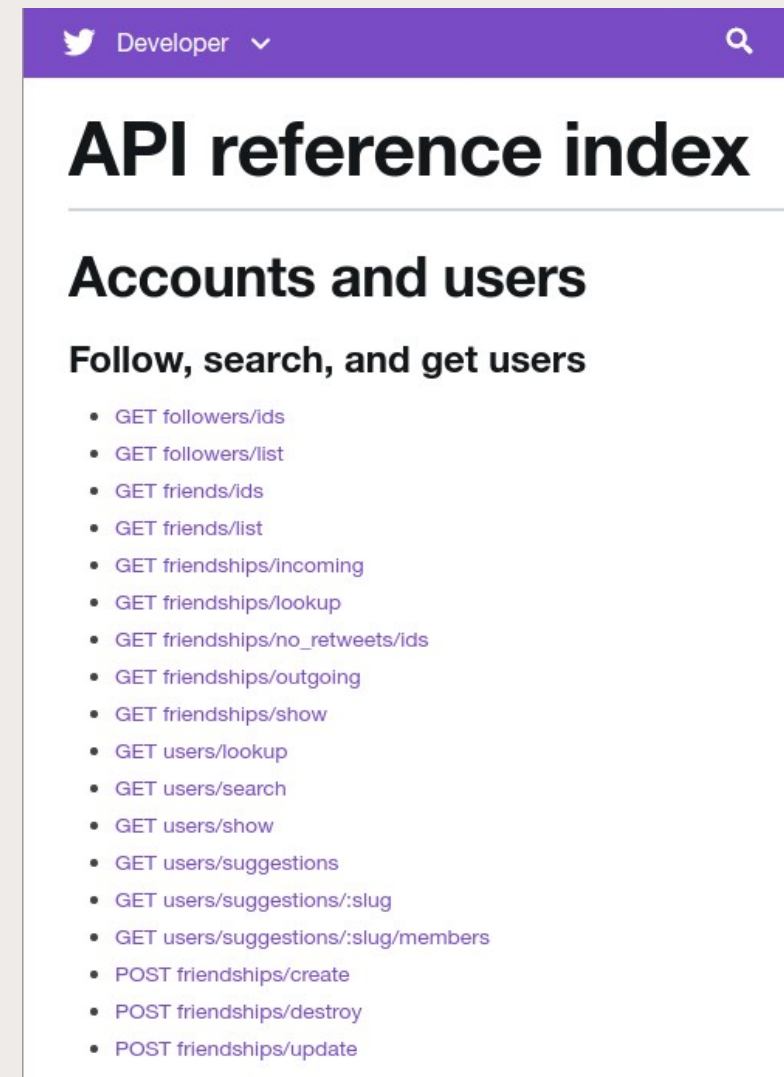
→ exemple : [https://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_carbon\\_dioxide\\_emissions](https://en.wikipedia.org/wiki/List_of_countries_by_carbon_dioxide_emissions)

```
> var data = artoo.scrapeTable( ".wikitable", {headers: 'th'} );
undefined
> data.length;
49
> data[0];
Object {Country: " World", C02 emissions (kt) in 2014[2]: "35,669,000", " % C02
Emissions by Country": "100%", Emission per capita (t) in 2014[3]: "5.0"}
> artoo.saveCsv(data, "C02-world-emissions.csv");
undefined
```



# Accès contrôlé via les APIs des plateformes

- « Application Programming 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



Pour mieux comprendre qui sont les producteurs de connaissances sur la Russie 🇷🇺 en France entre 1980 et 2020, le projet "Russia, made in France" de V. Lepinay et E. Lezean propose trois bases documentaires, en partie accessible en ligne. A découvrir sur



Russia, made in France : des connaissances sur la Russie | médialab Sciences ...  
Comment parler de la Russie et qui définit l'ordre du jour de la conversation sur la Russie soviétique et post-soviétique ? Initié en 2017, Russia, made in France...  
🔗 medialab.sciencespo.fr

11:13 AM · 18 mars 2021 · TweetDeck

12 Retweets 2 Tweets cités 31 J'aime

```
TWEET_FIELDS = [
    "id",
    "time",
    "created_at",
    "from_user_name",
    "text",
    "filter_level",
    "possibly_sensitive",
    "withheld_copyright",
    "withheld_scope",
    "withheld_countries",
    "truncated",
    "retweet_count",
    "favorite_count",
    "reply_count",
    "lang",
    "to_user_name",
    "to_user_id",
    "in_reply_to_status_id",
    "source",
    "source_name",
    "source_url",
    "location",
    "lat",
    "lng",
    "from_user_id",
    "from_user_realname",
    "from_user_verified",
    "from_user_description",
    "from_user_url",
    "from_user_profile_image_url",
    "from_user_utcoffset",
    "from_user_timezone",
    "from_user_lang",
    "from_user_tweetcount",
    "from_user_followercount",
    "from_user_friendcount",
    "from_user_favourites_count",
    "from_user_listed",
    "from_user_withheld_scope",
    "from_user_withheld_countries",
    "from_user_created_at",
    "collected_via_thread",
    "retweeted_id",
    "retweeted_user_name",
    "retweeted_user_id",
    "quoted_id",
    "quoted_user_name",
    "quoted_user_id",
    "links",
    "medias_urls",
    "medias_files",
    "mentioned_user_names",
    "mentioned_user_ids",
    "hashtags"

    # digital ID
    # UNIX timestamp of creation
    # ISO datetime of creation
    # author's user text ID (@user)
    # message's text content
    # internal TCAT field, ignorable
    # whether a link present in the message might contain sensitive content according to Twitter
    # whether the tweet might be censored by Twitter following copyright requests, ignorable
    # whether the content withheld is the "status" or a "user", ignorable
    # list of ISO country codes in which the message is withheld, separated by |, ignorable
    # whether the tweet is bigger than 140 characters, obsolete
    # number of retweets of the message (at collection time)
    # number of likes of the message (at collection time)
    # number of answers to the message, dropped by Twitter (since Oct 17, now charged), unreliable and ignorable
    # language of the message automatically identified by Twitter's algorithms (equals "und" when no language could be detected)
    # text ID of the user the message is answering to
    # digital ID of the user the message is answering to
    # digital ID of the tweet the message is answering to
    # medium used by the user to post the message
    # name of the medium used to post the message
    # link to the medium used to post the message
    # location declared in the user's profile (at collection time)
    # latitude of messages geolocalized
    # longitude of messages geolocalized
    # author's user digital ID
    # author's detailed textual name (at collection time)
    # whether the author's account is certified
    # description given in the author's profile (at collection time)
    # link to a website given in the author's profile (at collection time)
    # link to the image avatar of the author's profile (at collection time)
    # time offset due to the user's timezone, dropped by Twitter (since May 18), ignorable
    # timezone declared in the user's profile, dropped by Twitter (since May 18), ignorable
    # language declared in the user's profile (at collection time)
    # number of tweets sent by the user (at collection time)
    # number of users following the author (at collection time)
    # number of users the author is following (at collection time)
    # number of likes the author has expressed (at collection time)
    # number of users lists the author has been included in (at collection time)
    # whether the user content is withheld, ignorable
    # list of ISO country codes in which the user content is withheld, separated by |, ignorable
    # ISO datetime of creation of the author's account
    # whether the tweet was retrieved only as part of a thread including a tweet matching the desired query
    # digital ID of the retweeted message
    # text ID of the user who authored the retweeted message
    # digital ID of the user who authored the retweeted message
    # digital ID of the retweeted message
    # text ID of the user who authored the retweeted message
    # digital ID of the user who authored the retweeted message
    # list of links included in the text content, with redirections resolved, separated by |
    # list of links to images/videos embedded, separated by |
    # list of filenames of images/videos embedded and downloaded, separated by |, ignorable when medias collections isn't enabled
    # list of text IDs of users mentioned, separated by |
    # list of digital IDs of users mentioned, separated by |
    # list of hashtags used, lowercased, separated by |
]
```

# Risques de refermeture des APIs

- Twitter API v2.0 « *designed for academics needs* »  
... or not...

	scraping web	search API v1.1	search API v2 standard	search API v2 académique
time coverage	all time	8 days back	8 days back	all time
# tweets limit	~ 50 M / month	~ 180 M / month	0.2 M / month	10 M / month
retweets included	no	yes	yes	yes
extra metadata	yes	no	yes	yes

task	API version	results / query	app API / 15 min	user API / 15 min	queries / heure	results / heure	max queries / jour	queries / jour	max tweets / mois	results / jour
user followers/friends	1.1	5000	15	15	120	600000		2880		14400000
users followers/friends (max si 10000 / user)	1.1	0,5	15	15	120	60		2880		1440
users metas	1.1	100	300	900	4800	480000		115200		11520000
tweets by ids	1.1	100	300	900	4800	480000		115200		11520000
tweets by ids	2.0	100	300	900	4800	480000		115200	500000	500000
user tweets	1.1	200	1500	900	9600	1920000	100000	100000		20000000
users tweets (max si 3250 tweets / user)	1.1	0,05882352941	1500	900	9600	564	100000	100000		5882
tweet retweets	1.1	100	300	75	1500	150000		36000		3600000
tweets by search	1.1	100	450	180	2520	252000		60480		6048000
tweets by search	2.0	100	450	180	2520	252000		60480	500000	500000
tweets by search	2.0 académique	500	300	0	1200	600000	86400	28800	10000000	10000000

# Creuser le web avec Minet

<https://github.com/medialab/minet>



- Pour utilisateurs semi-avancés (no-code, mais ligne de commande)
- Génériciser nos pratiques et expertises de webmining
- Autonomiser chercheurs et doctorants sur la collecte de données
- Extraire des contenus textes, des liens, des images, etc. à partir de listes d'URLs ou mots-clés
- SCRAPING :
  - Facebook posts, pages & groups
  - Twitter search
  - Instagram en cours d'ajout (v0.65.0!)
  - Telegram en cours d'ajout
  - TikTok un jour ?
- APIs :
  - Crowdtangle (métriques Facebook)
  - Twitter friends/followers/search
  - YouTube videos & channels
  - Reddit à venir

```
# Yomgui at mbp-de-plique-1.home in ~/code/minet on git:master ✖ [15:51:23]
→ ./ftest/fetch.sh
```

# 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": " ",
    "secret": " ",
    "oauth_token": " ",
    "oauth_secret": " "
  },
  "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
}
```

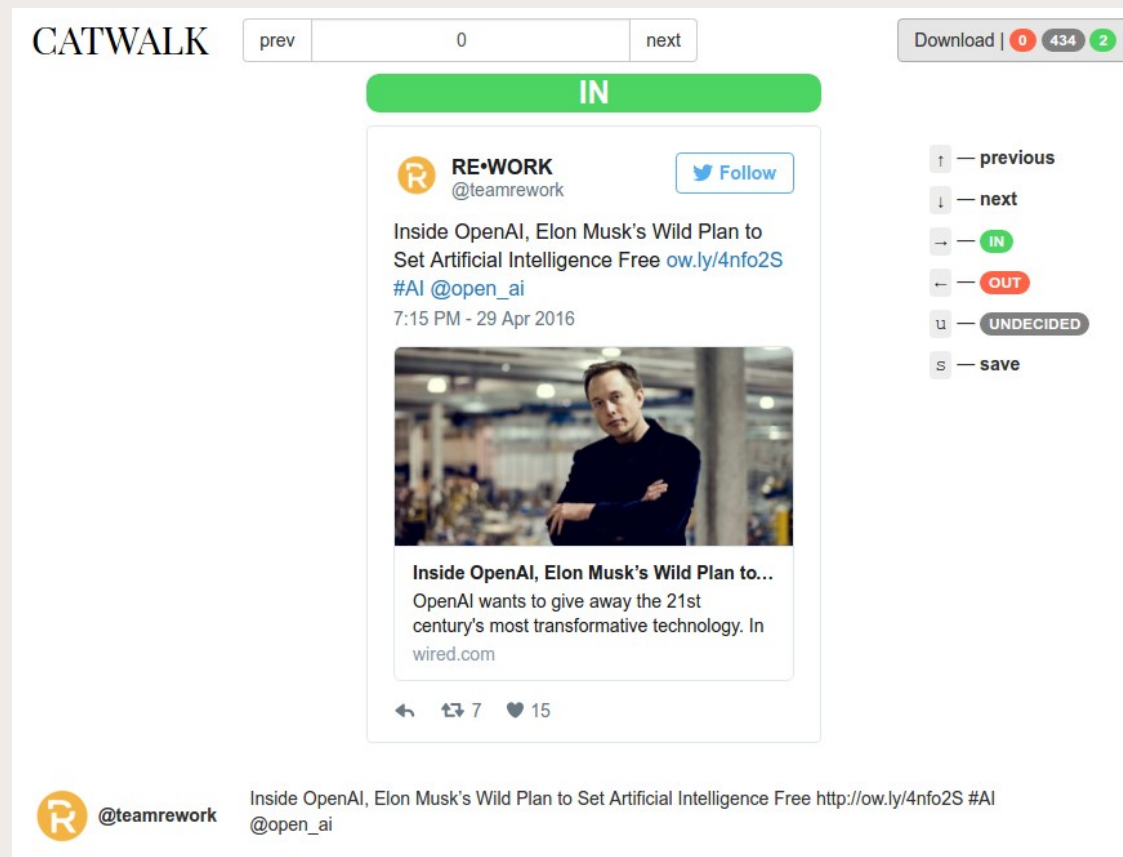
```
[2016-11-22 15:23:34.056196] DEBUG: Starting search queries with 328 remaining calls for the next 655 seconds
[2016-11-22 15:23:34.259849] DEBUG: [search] +1 tweets (agriculture%20Paris OR agricultures%20Paris OR agroforesterie%20Paris)
[2016-11-22 15:23:35.807085] DEBUG: Saved 1 tweets in MongoDB
[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/>

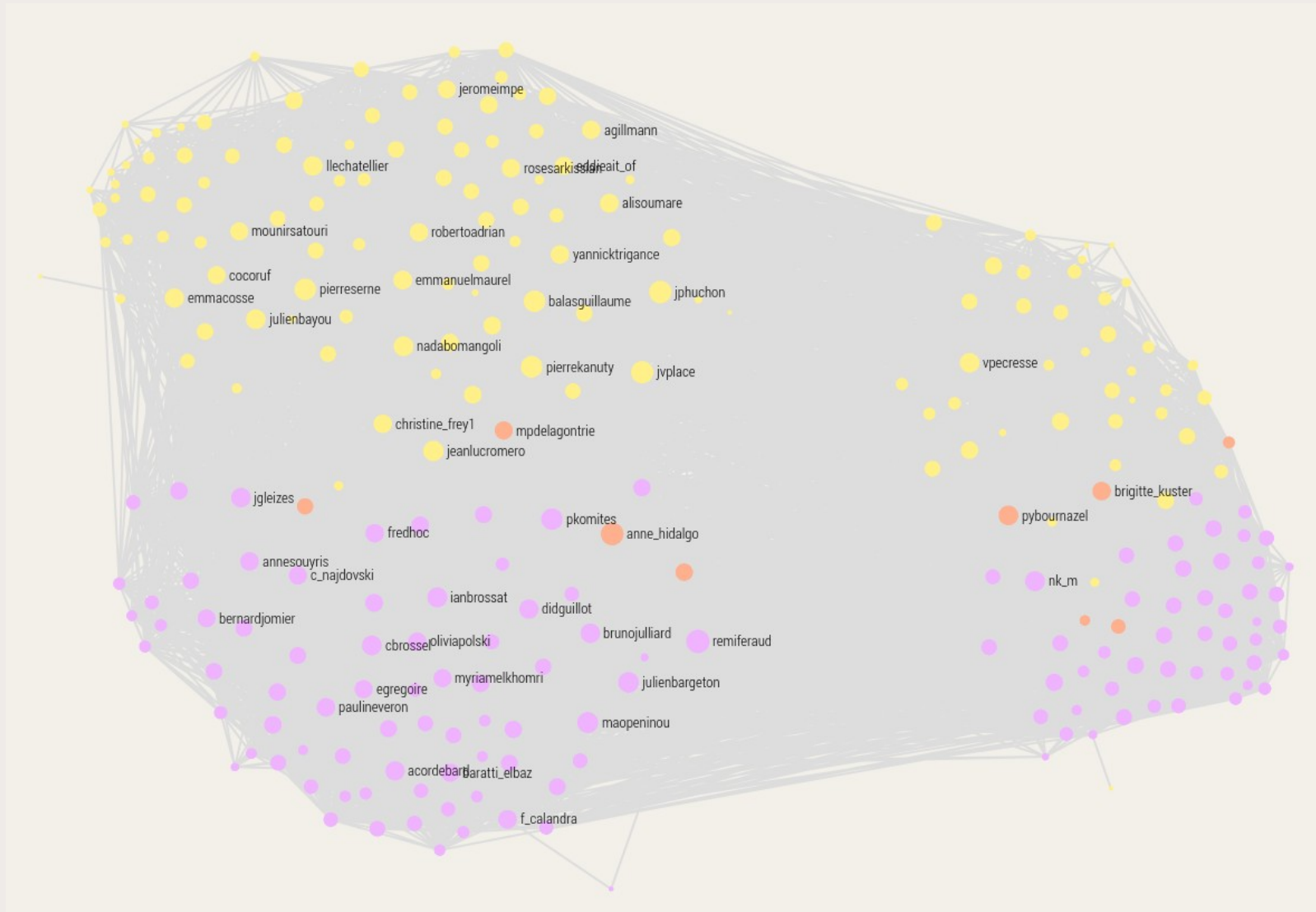
Passer en revue rapidement « *à la Tinder* » tous les tweets d'un CSV pour décider de les inclure / exclure d'un corpus



→ V2 en approche !

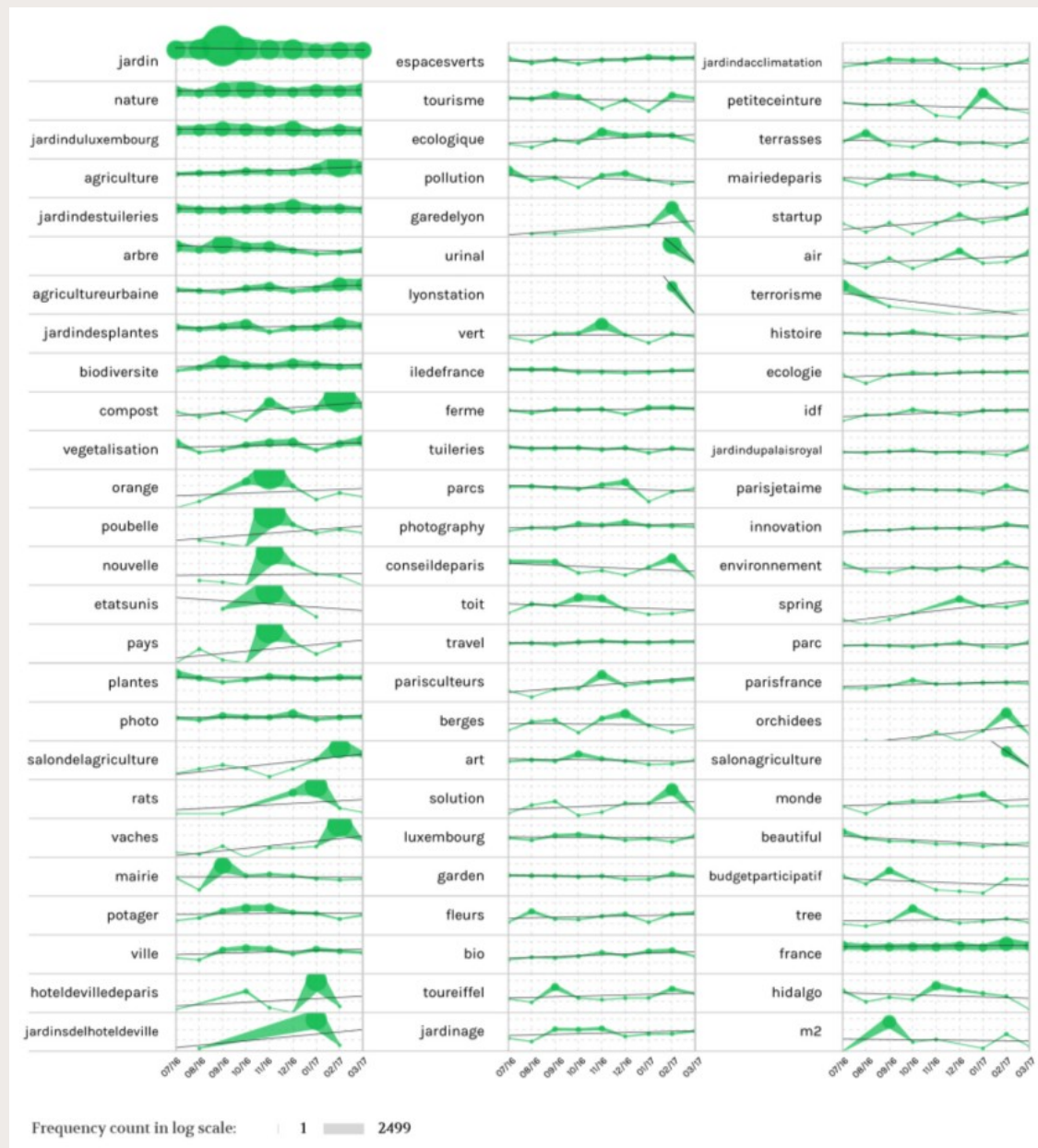


# Explorer des réseaux de followers

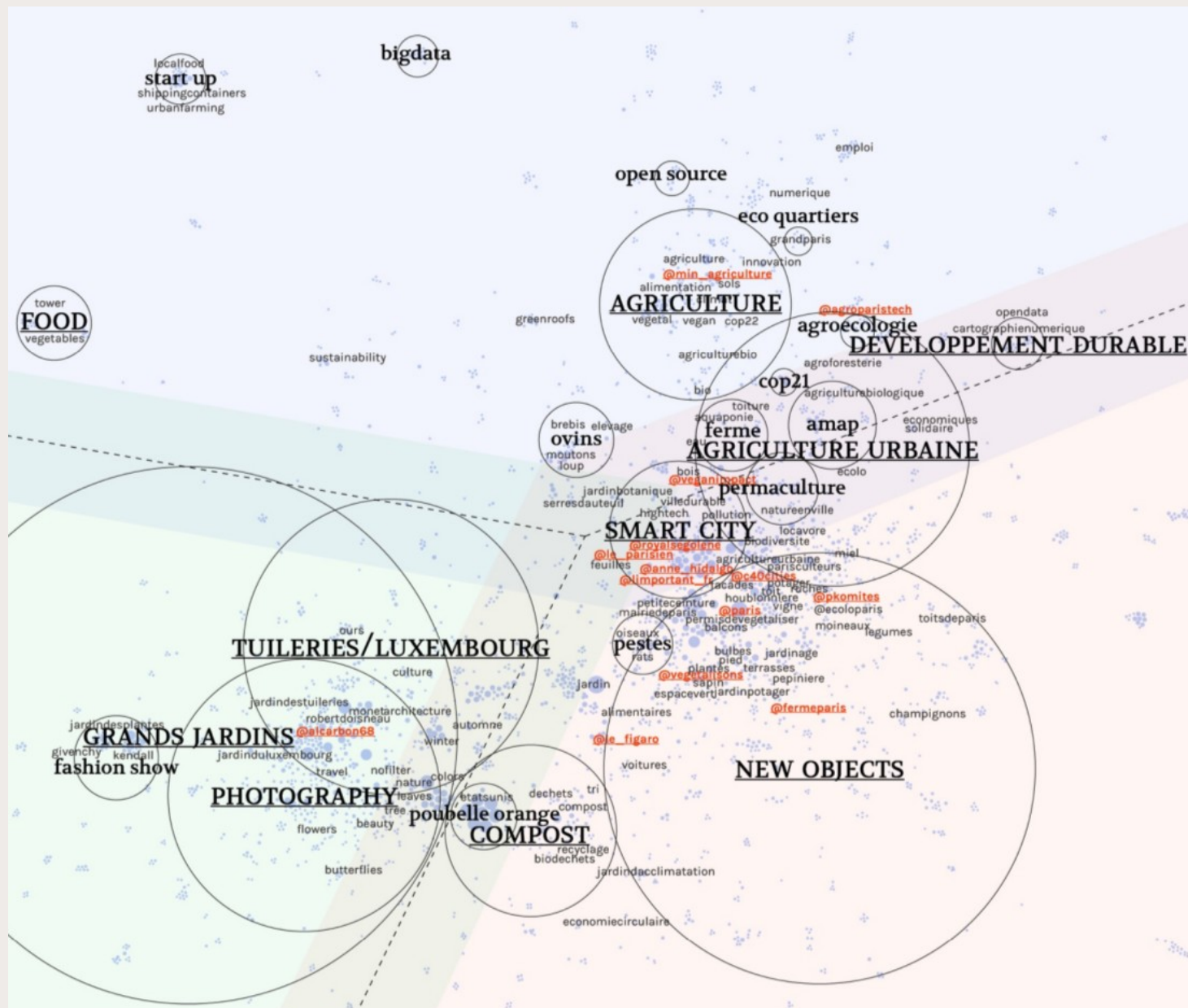


Liens de proximité Twitter entre les élus du Conseil Régional d'Île-de-France et du Conseil de Paris

# Explorer les dynamiques temporelles



# Explorer des réseaux sémantiques



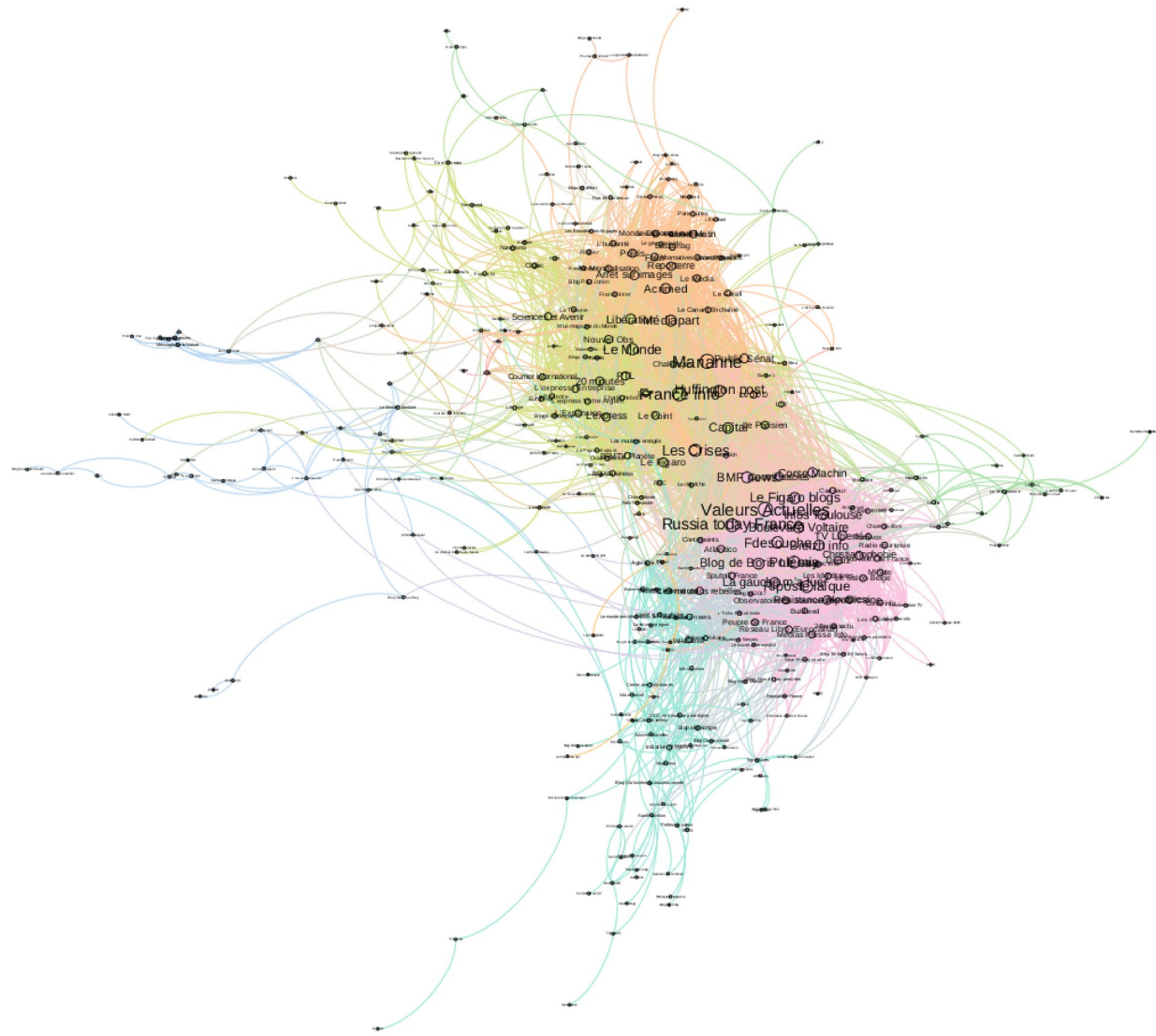
# Explorer l'espace visuel d'un corpus



RICCI, Donato, COLOMBO, Gabriele, MEUNIER, Axel, et al. **Designing Digital Methods to monitor and inform Urban Policy. The case of Paris and its Urban Nature initiative.** In: 3rd International Conference on Public Policy (ICPP3)-Panel T10P6 Session 1 Digital Methods for Public Policy.  
[https://re.public.polimi.it/bitstream/11311/1038509/1/IPPA\\_Ricci-Colombo-Meunier-Brilli.pdf](https://re.public.polimi.it/bitstream/11311/1038509/1/IPPA_Ricci-Colombo-Meunier-Brilli.pdf)



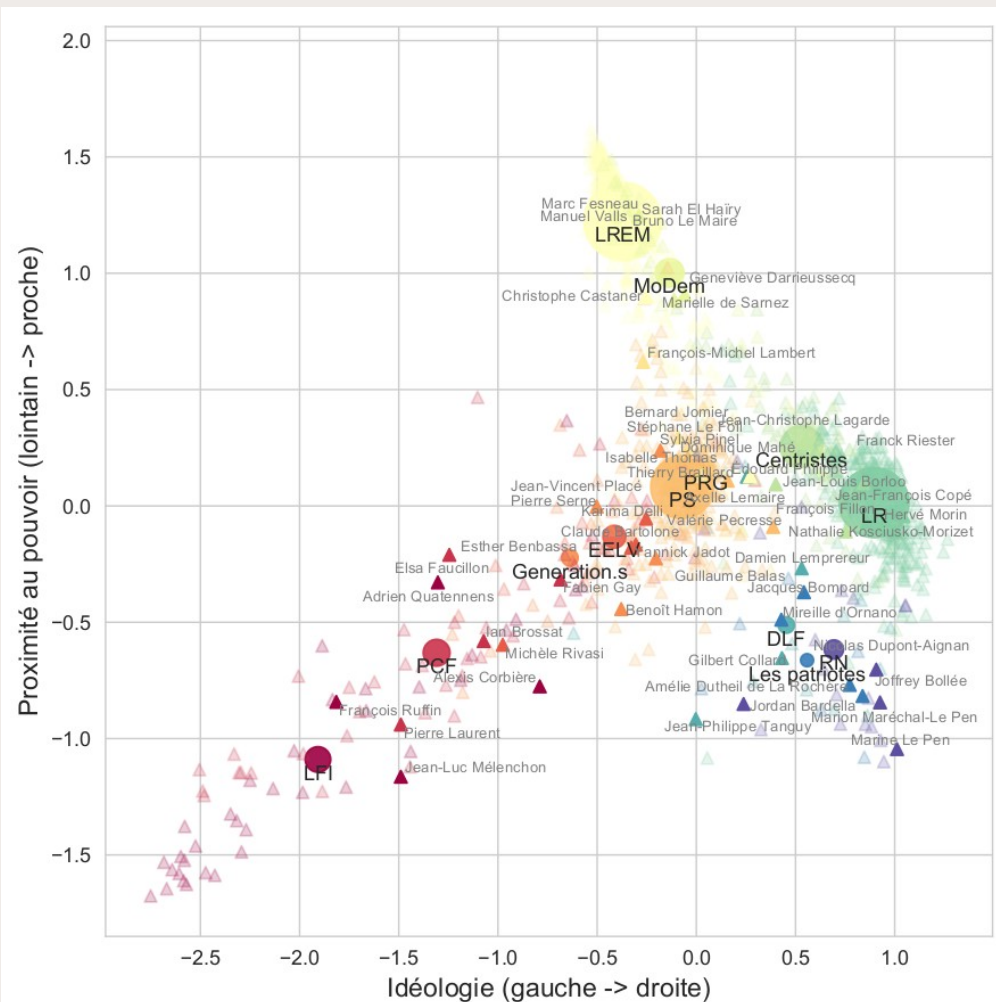
# Explorer des réseaux de co-citation de sites web



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

# 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)

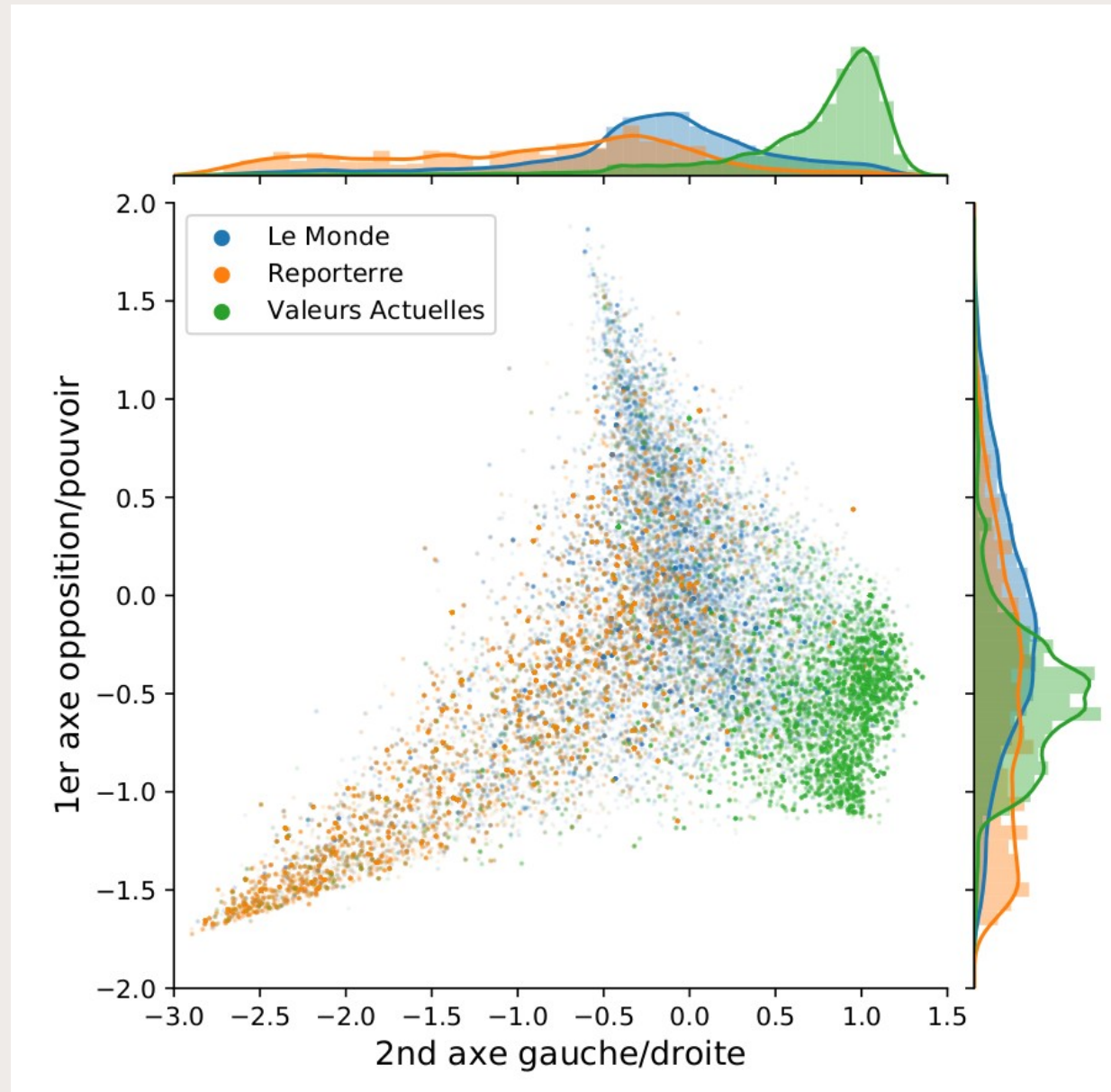


Jean-Philippe Cointet, Pedro Ramaciotti Morales, Dominique Cardon, Caterina Froio, Benjamin Ooghe, et al.. **De quelle(s) couleur(s) sont les Gilets jaunes ? Plonger des posts Facebook dans un espace idéologique latent.** Statistique et Société, Société française de statistique, 2021.  
<https://hal.archives-ouvertes.fr/hal-03700083>



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

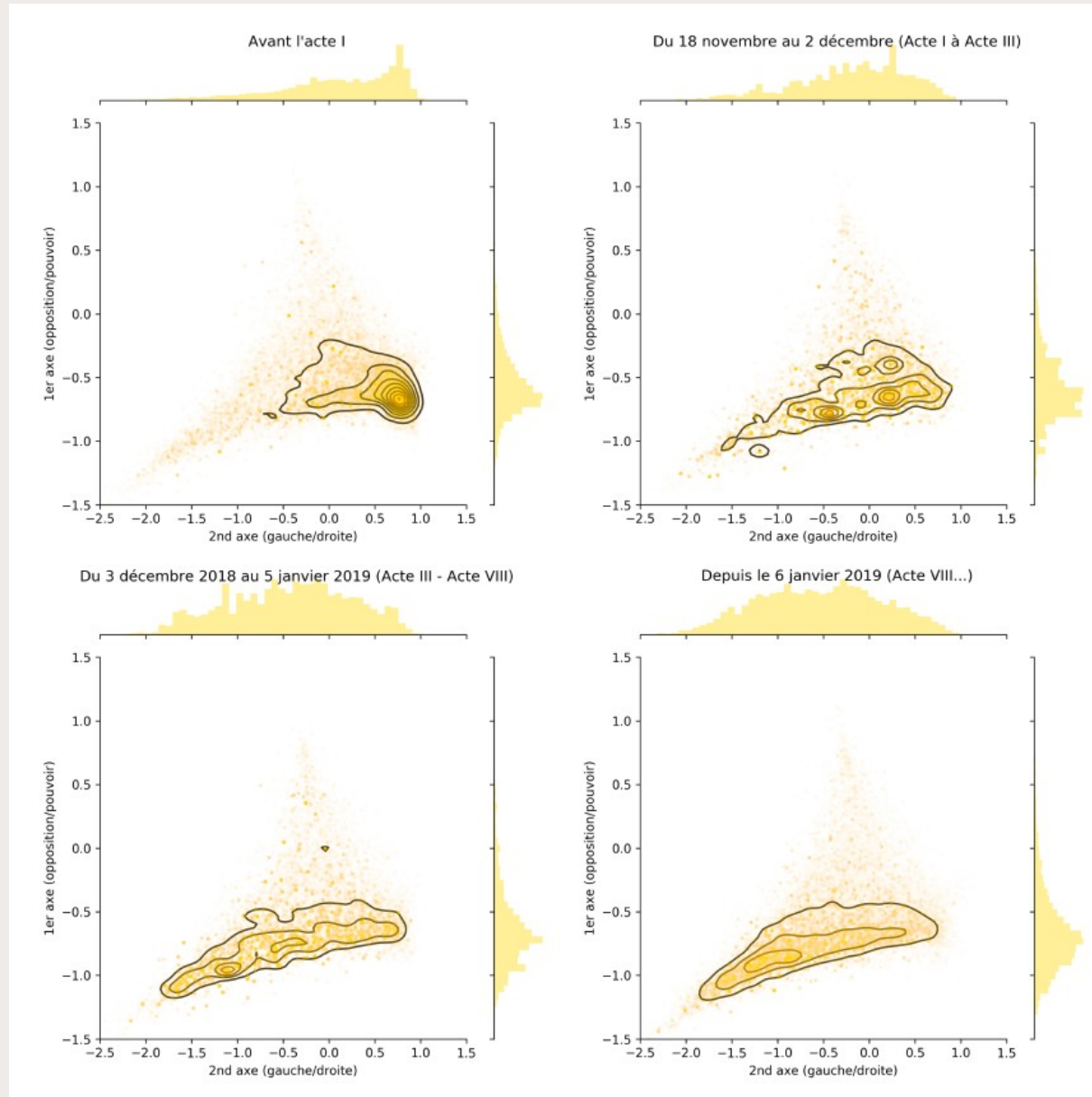
Projection des articles de médias partagés par les utilisateurs de Twitter sur la cartographie de l'espace politique





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

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

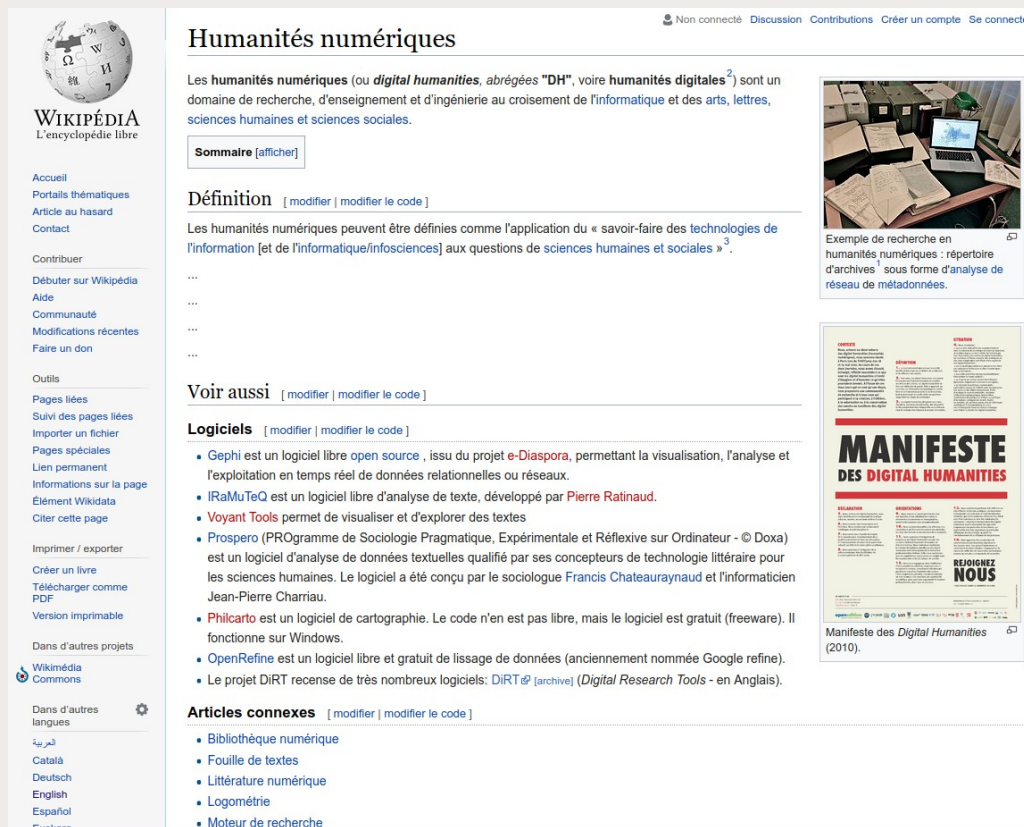




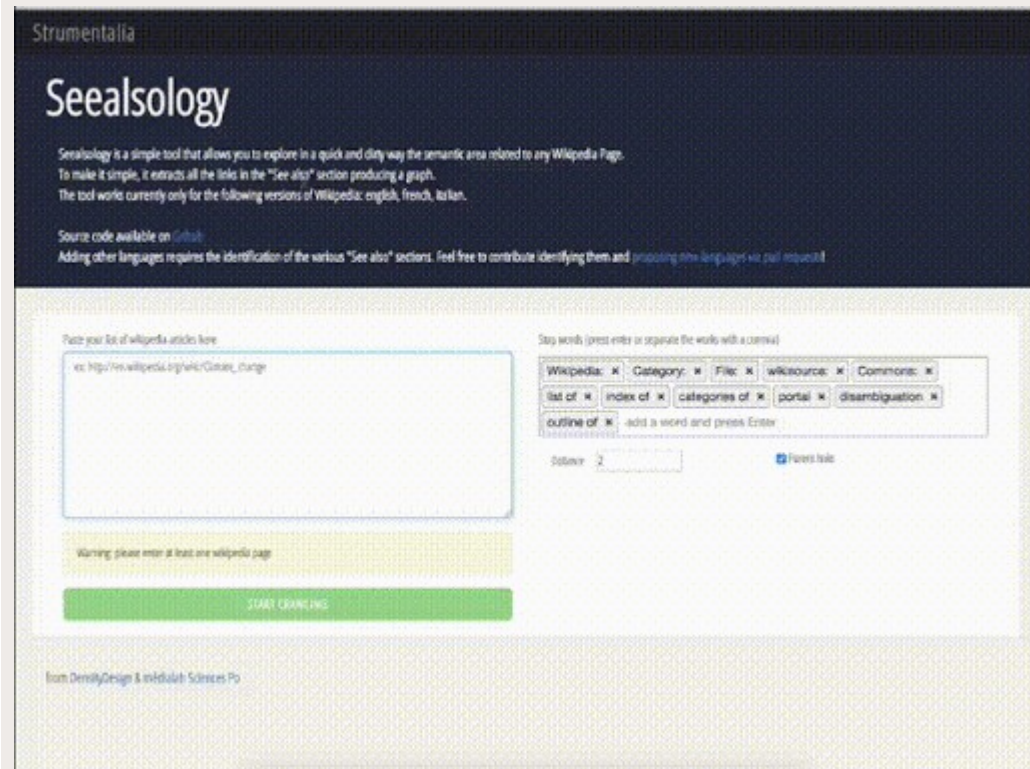
# SeeAlsology : exploration sémantique rapide

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

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



The screenshot shows the Wikipedia page for 'Humanités numériques'. The page is in French and includes a sidebar with navigation links, a main content area with a definition and a list of software tools, and a 'Voir aussi' (See also) section. The definition states that digital humanities (DH) is a domain of research, teaching, and engineering at the intersection of computer science and arts, letters, and social sciences. The 'Logiciels' (Software) section lists tools like Gephi, IRaMuTeQ, Voyant Tools, Prospero, Philcarto, OpenRefine, and DIRT. The 'Voir aussi' section lists related topics like 'Bibliothèque numérique', 'Fouille de textes', 'Littérature numérique', 'Logométrie', and 'Moteur de recherche'.




The screenshot shows the Seealsology tool interface. It has a dark header with the title 'Seealsology' and a description: 'Seealsology is a simple tool that allows you to explore in a quick and dirty way the semantic area related to any Wikipedia Page. To make it simple, it extracts all the links in the "See also" section producing a graph. The tool works currently only for the following versions of Wikipedia: english, french, italian.' Below the description, there is a text input field for pasting a list of Wikipedia articles, a dropdown menu for selecting the language (English, French, Italian), and a 'START CRAWLING' button. The footer mentions 'from DervidDesign & medialab Sciences Po'.

# 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




## Table 2 Net

### Load your CSV table

#### 1. Type of Network

Normal (one type of node) ▼



You will have to choose:

- Which column **X** will define the nodes
- Which column **Y** will define the links

#### 2. Nodes

**X** Which column defines the nodes?

hashtags ▼

Pipe-separated "|" ▼

Sample of nodes extracted with these settings: (↻ sample)

#goweser #adventurebike #kotaposo #dh #xcbike

### 3. Links

**Y** Which column defines the links?

Row number ▼

One expression per cell ▼

Sample of items extracted with these settings: (↻ sample)

5252 3621 1816 1847 4562

### 4. Additional settings

Optional: time series

No temporal data ▼

Select only a column containing integers.

Optional: edge weight

Weight the edges ▼

### 5. Build the network

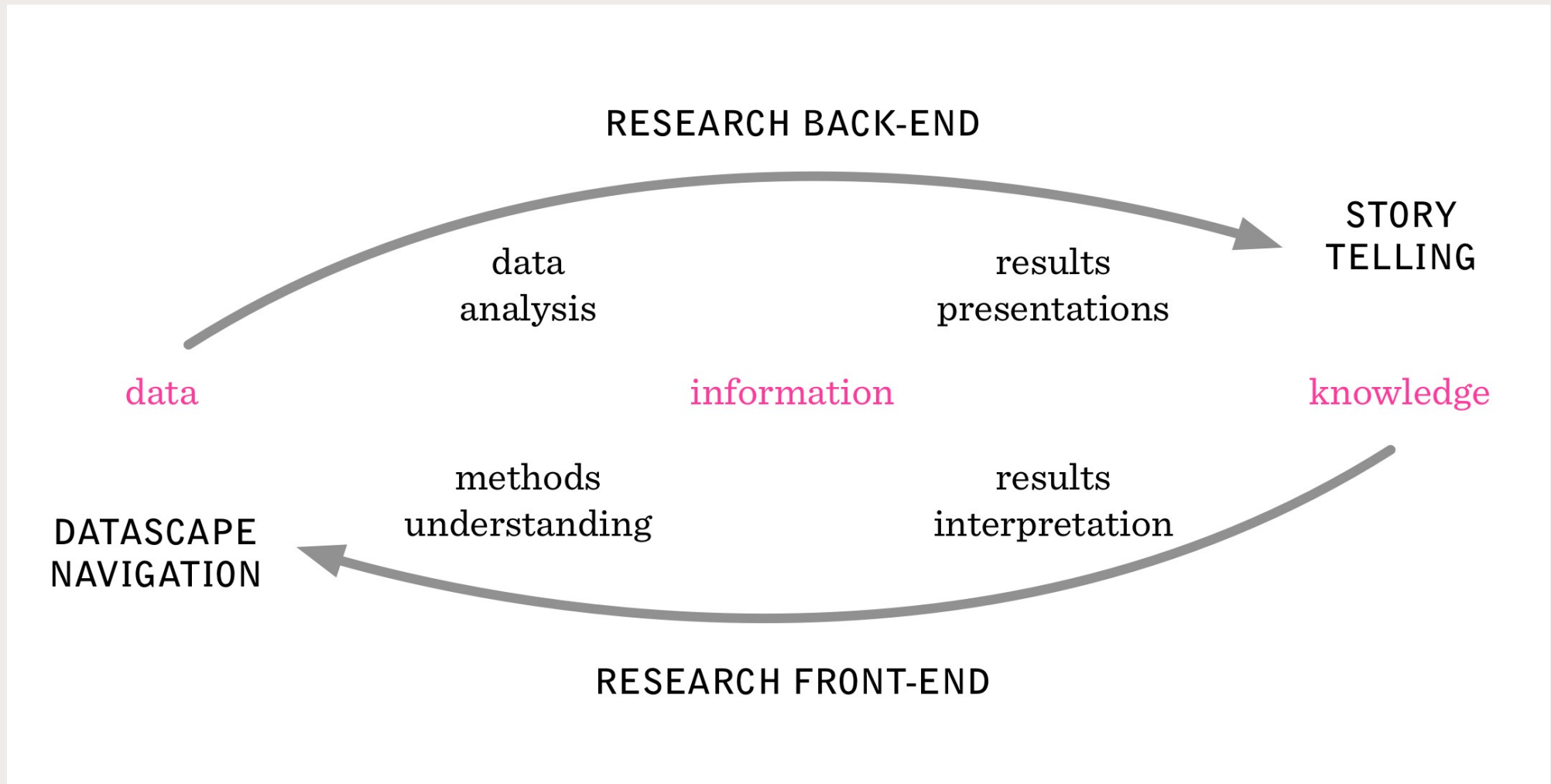
Ⓞ Build and download the network (GEXF)

NB: this may take a while, please be patient.

→ Visualisation de réseaux avec **Gephi** / **Nansi** / **MiniVan**

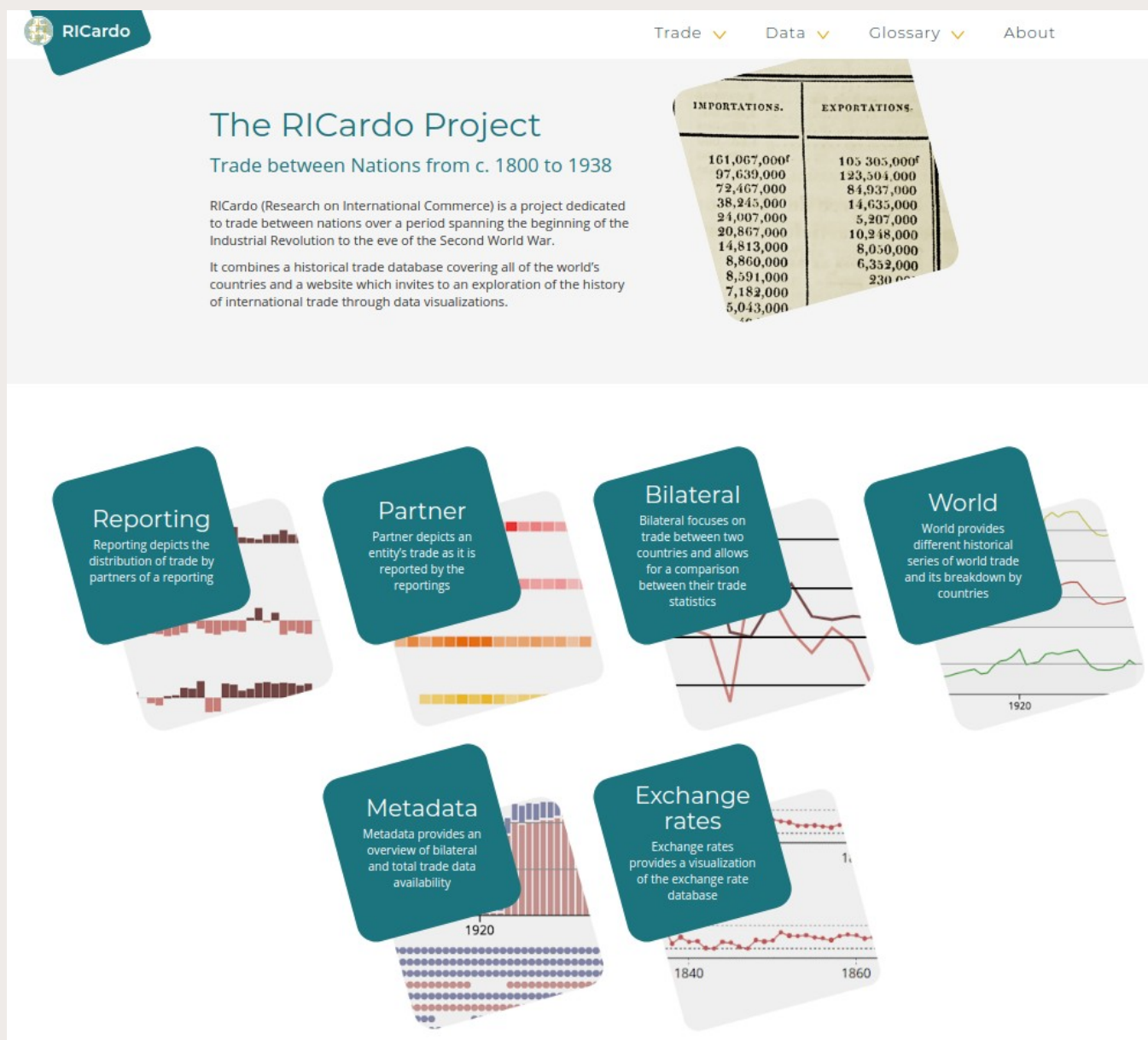
# Les « datascapes » : zoomer/dézoomer

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



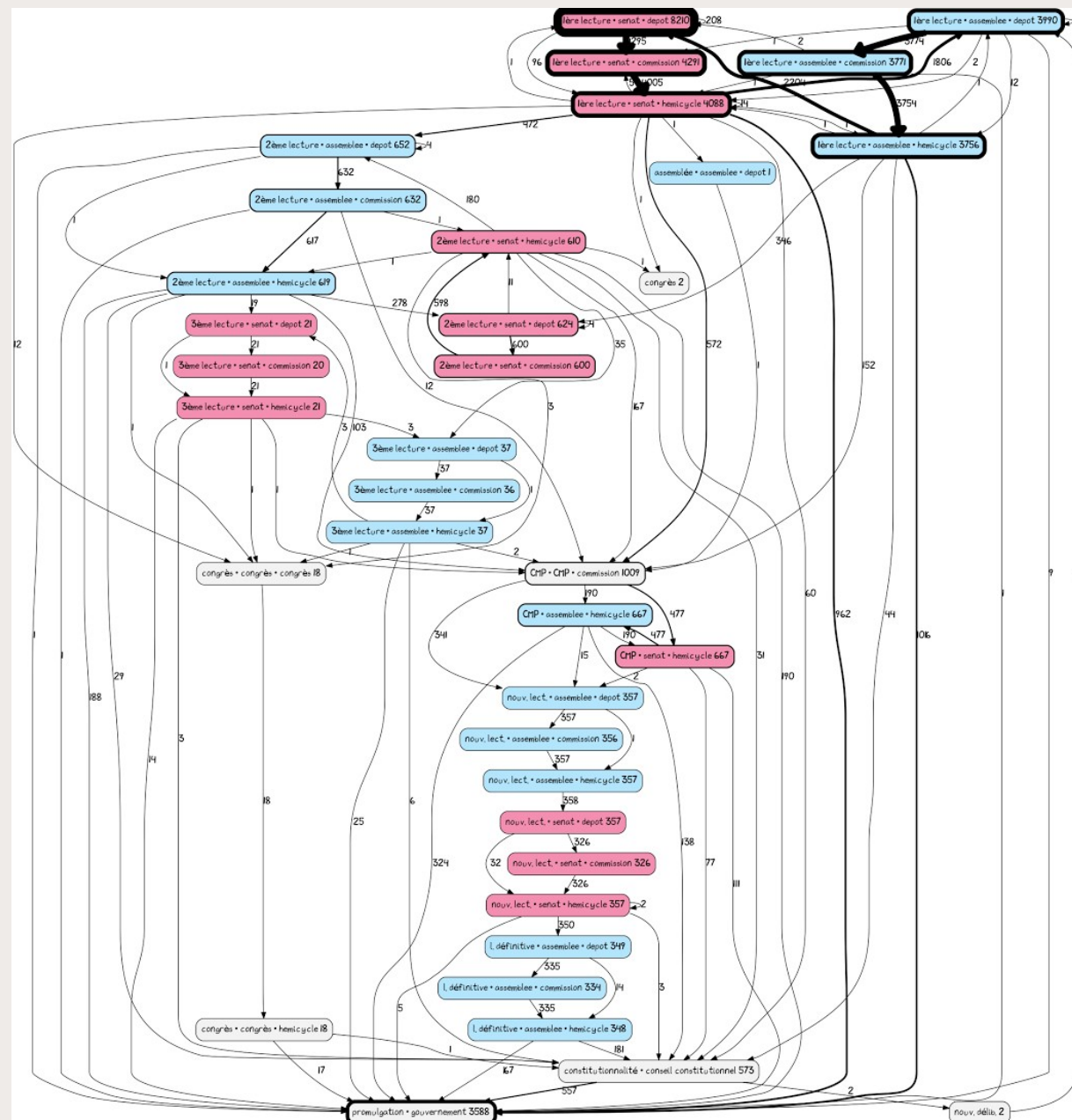


# RICardo : visualiser les échanges au 19ème siècle



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

# LaFabriqueDeLaLoi : explorer la complexité législative



# LFDLL : plus de 1000 lois promulguées ou en cours

Explorer les textes promulgués depuis 2010

Vue chronologique ▾

Trié par date ▾

Plus de 50 amendements ▾

Étudié en 2013 ▾

Tous les thèmes ▾

Zoom



Juil. 2013

Oct. 2013

Janv. 2014

Économie réelle

Géolocalisation

Contrefaçon

Ville

Cumul des fonctions (texte organique)

Cumul des fonctions

Réseaux de soins

Retraites

NAVETTES

Chaque ligne représente la chronologie des débats sur un projet ou une proposition de loi. La couleur indique l'institution en charge du texte à un instant donné (Assemblée en bleu, Sénat en rouge...). Cliquez sur un texte pour en consulter le résumé et en explorer les articles.

Cliquez sur le bouton ? ci-dessus pour voir un tutoriel interactif de cette visualisation.

Sénat

08/01/2014 → 15/01/2014

32 amendements

Filtrer par durée d'adoption des textes



# LFDLL : retracer l'évolution du texte à chaque étape

## Projet de loi portant création du contrat de génération

[Dossier Sénat](#)
[Loi sur Légifrance](#)  
[Dossier Assemblée](#)
[Open Data /Git](#)

< Voir la chronologie du texte

Vue alignée ▾

?

Dépôt	1 <sup>ère</sup> Lecture				Commission Mixte Paritaire			
Gouv.	Assemblée		Sénat		CMP	Sénat	AN	
Projet de Loi	Commission	Hémicycle	Commission	Hémicycle	Commission	Hémicycle	Hémicycle	Article 1
Art. 1	Art. 1	Art. 1	Art. 1	Art. 1	Art. 1	Art. 1	Art. 1	<p>1<sup>ère</sup> Lecture · Assemblée · Commission</p> <ul style="list-style-type: none"> <li>I. – Les mots : "et à la gestion des âges" sont ajoutés à l'intitulé du chapitre Ier du titre II du livre Ier de la cinquième partie du code du travail ;</li> <li>II. – Il est rétabli au chapitre Ier du titre II du livre Ier de la cinquième partie du code du travail une section 4 est ainsi modifiée :</li> <li>1<sup>o</sup> L'intitulé est complété par les mots : "et à la gestion des âges" ;</li> <li>1<sup>o</sup> bis L'article L. 5121-7 devient l'article L. 5121-22 ;</li> <li>2<sup>o</sup> La section 4 est ainsi rédigée :</li> <li>"Section 4</li> <li>"Contrat de génération</li> <li>"Art. L. 5121-6. - Le contrat de génération a pour objectif de faciliter l'intégrations ;</li> <li>"1<sup>o</sup> De faciliter l'insertion durable des jeunes dans l'emploi par leur accès à un contrat à durée indéterminée ;</li> <li>"2<sup>o</sup> De favoriser l'embauche et le maintien en emploi des salariés âgés et ;</li> <li>"3<sup>o</sup> D'assurer la transmission des savoirs et des compétences.</li> <li>"Il est mis en oeuvre, en fonction de la taille des entreprises, dans les conditions prévues par la présente section.</li> <li>"Le contrat de génération est applicable aux employeurs de droit privé.</li> </ul>
Art. 2	Art. 2	Art. 2	Art. 2	Art. 2	Art. 2	Art. 2	Art. 2	
Art. 3	Art. 3	Art. 3	Art. 3	Art. 3	Art. 3	Art. 3	Art. 3	
Art. 4	Art. 4	Art. 4	Art. 4	Art. 4	Art. 4	Art. 4	Art. 4	
Art. 5	Art. 5	Art. 5	Art. 5	Art. 5	Art. 5	Art. 5	Art. 5	
Art. 5 bis	Art. 5 bis	Art. 5 bis	Art. 5 bis	Art. 5 bis	Art. 5 bis	Art. 5 bis	Art. 6 (5 bis)	
Art. 6	Art. 6	Art. 6	Art. 6	Art. 6	Art. 6	Art. 6	Art. 7 (5)	
Art. 7	Art. 7	Art. 7	Art. 7	Art. 7	Art. 7	Art. 7	Art. 8 (7)	

10/11/22 - CIMEOS	Collecter et analyser des données du web et des réseaux sociaux pour les SHS	44
-------------------	--	----



# LFDLL : remonter aux débats parlementaires

Projet de loi pour une République numérique

[Dossier Sénat](#)
[Dossier Assemblée](#)
[Open Data](#)

[< Voir les articles du texte](#)
Vue « échiquier politique »

Dépôt	1 <sup>ère</sup> Lecture			
Gouvernement	Assemblée		Sénat	
Projet de Loi	Commission	Hémicycle	Commission	Hémicycle
Article 1er				
Après l'article 1er				
Article 1er bis				
Article 1er ter				
Article 2				
Après l'article 2				
Article 4				
Après l'article 4				

**LÉGENDE**

GDR	SRC	ECOLO	RRDP	UDI	LES-REP	NI
Présidence	Rapporteurs	Gouvernement	Auditionnés			

Les républicains

Après l'article 1er

**Frédéric Lefebvre**  
[Lire les interventions](#)
1816 mots

**Patrice Martin**  
[Lire les interventions](#)

**Lionel Tardy**  
[Lire les interventions](#)

**Philippe Gosselin**  
[Lire les interventions](#)

**Laure de La Rota**  
[Lire les interventions](#)

**Catherine Vautrin, présidente**  
 La parole est à M. Frédéric Lefebvre, pour soutenir l'amendement no 518.

[Laisser un commentaire](#)

**Frédéric Lefebvre**  
 Notre collègue Patrice Martin-Lalande a évoqué ce sujet tout à l'heure et a regretté que son amendement ne puisse être discuté en raison de l'article 40.  
 Il a donc signé celui-ci visant à demander un rapport au Gouvernement, au plus tard le 30 juin 2016, sur la nécessité de créer une consultation publique en ligne pour tout projet de loi ou proposition de loi avant son inscription à l'ordre du jour au Parlement – c'est le moyen que nous avons trouvé pour discuter de ce dispositif dans l'hémicycle.  
 La crise de confiance est considérable, souvent à juste titre lorsque l'on s'avise du décalage entre le débat public et les préoccupations de nos compatriotes.  
 Il est temps que la logique démocratique que nous défendons devienne une règle générale dans notre démocratie.

[Laisser un commentaire](#)

**Catherine Vautrin, présidente**  
 Quel est l'avis de la commission ?

[Laisser un commentaire](#)

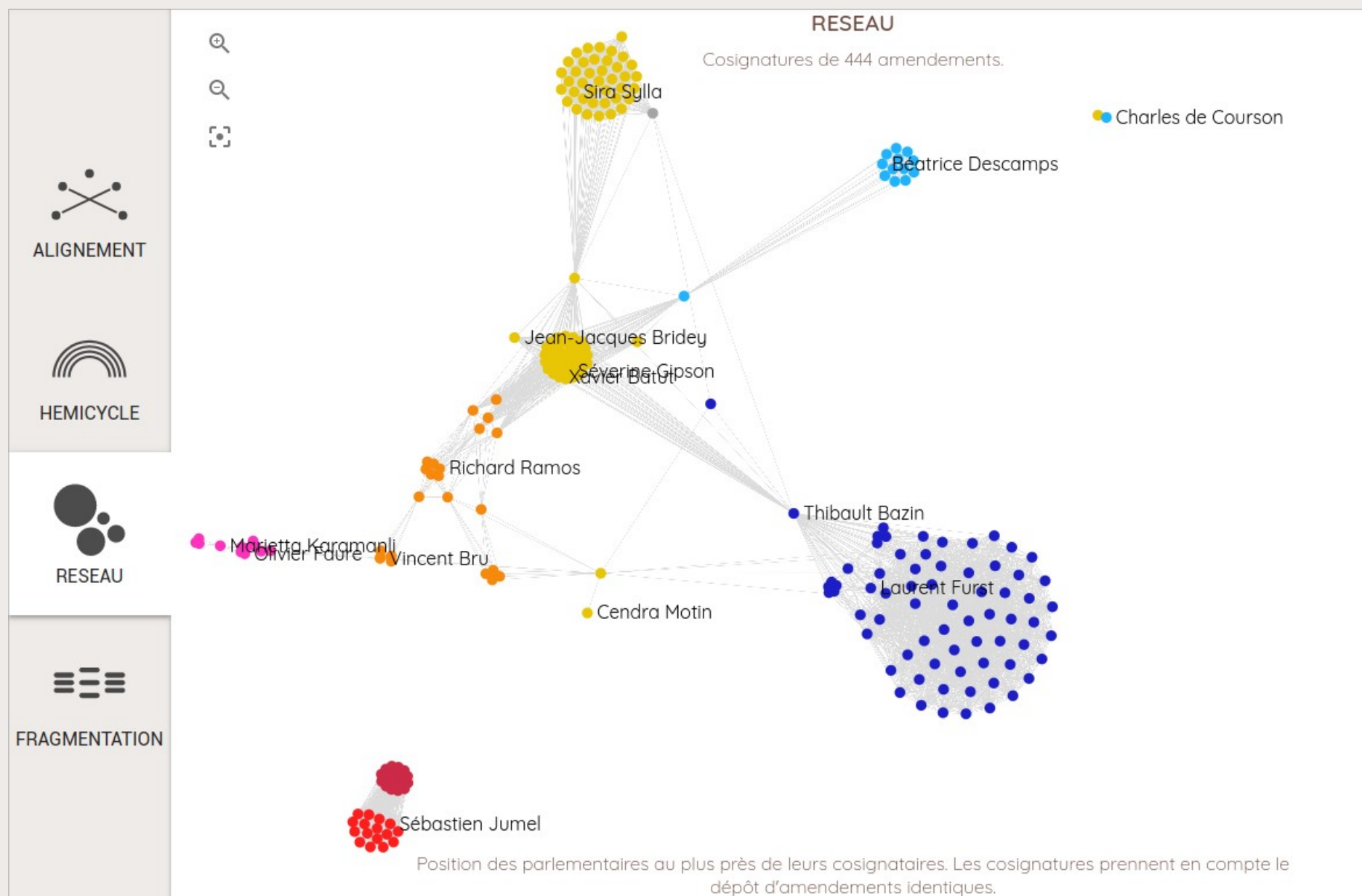
**Luc Belot, rapporteur de la commission des lois constitutionnelle de la République**  
 Je remercie notre collègue Lefebvre pour avoir rappelé le caractère assez innovant de notre démarche avec ce texte – un grand nombre d'entre vous l'a d'ailleurs également fait lors de la discussion générale – laquelle a eu des résultats très favorables.

[Laisser un commentaire](#)





# LFDLL : l'alignement et la fragmentation des partis



# À vos questions !

---

<https://medialab.sciencespo.fr>

[benjamin.ooghe@sciencespo.fr](mailto:benjamin.ooghe@sciencespo.fr)

[@boogheta@paille.fr](mailto:@boogheta@paille.fr)   [@boogheta](#)   [@medialab\\_ScPo](#)