



HAL
open science

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 Attribution - ShareAlike 4.0 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



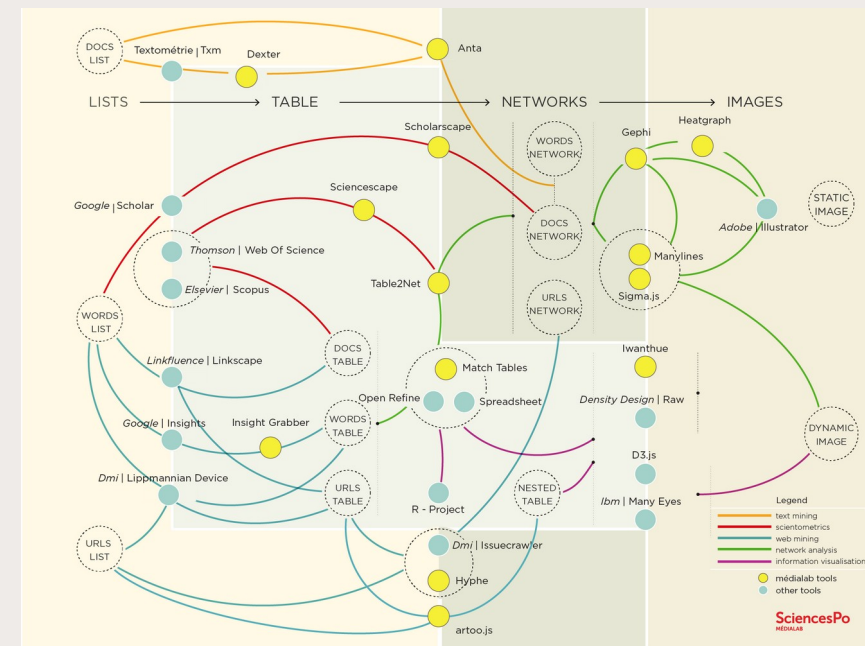
*« Google is nice,
but we need
something better »*

The Indian Express, 2011

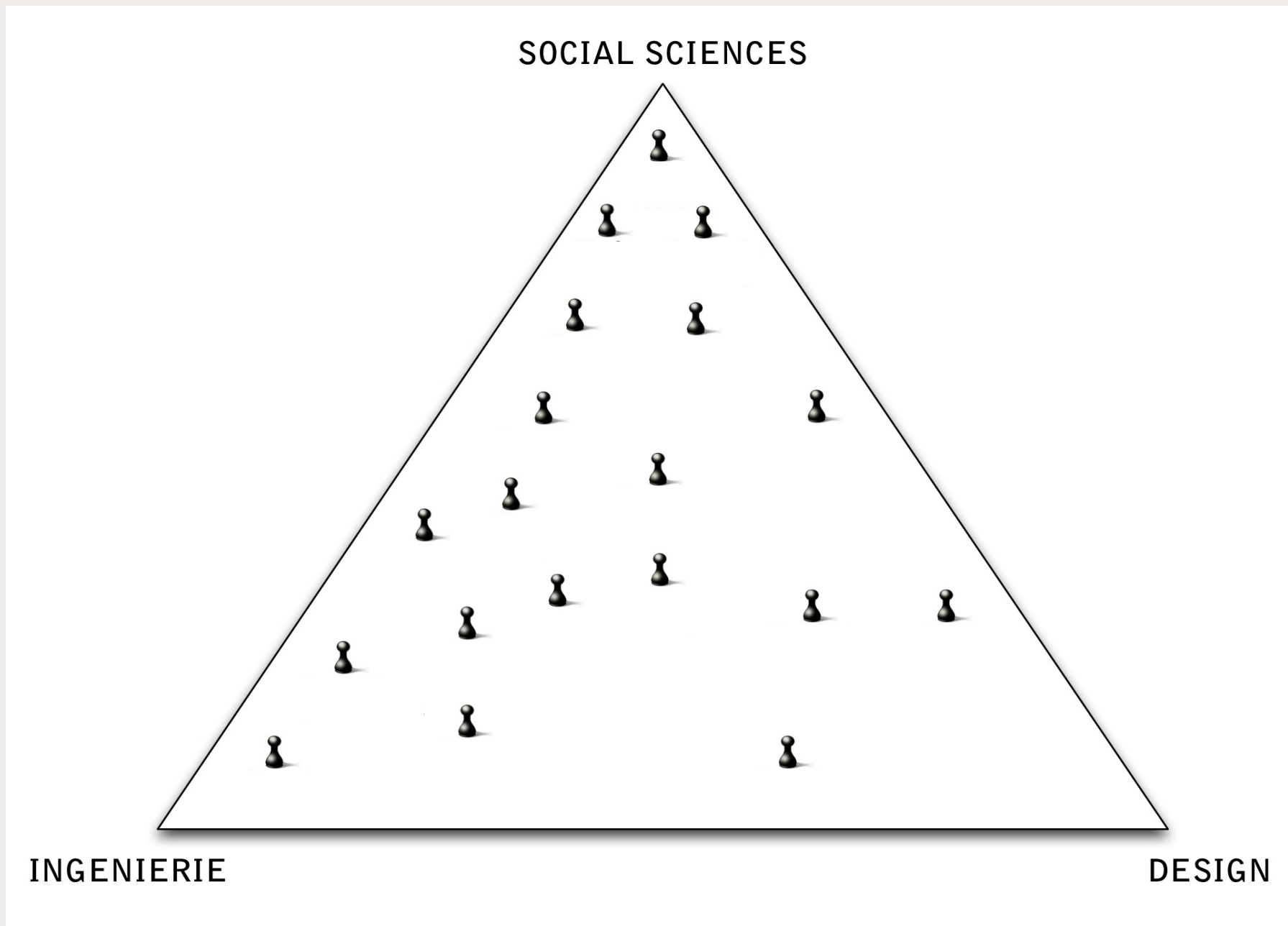
médialab @ Sciences Po

<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>



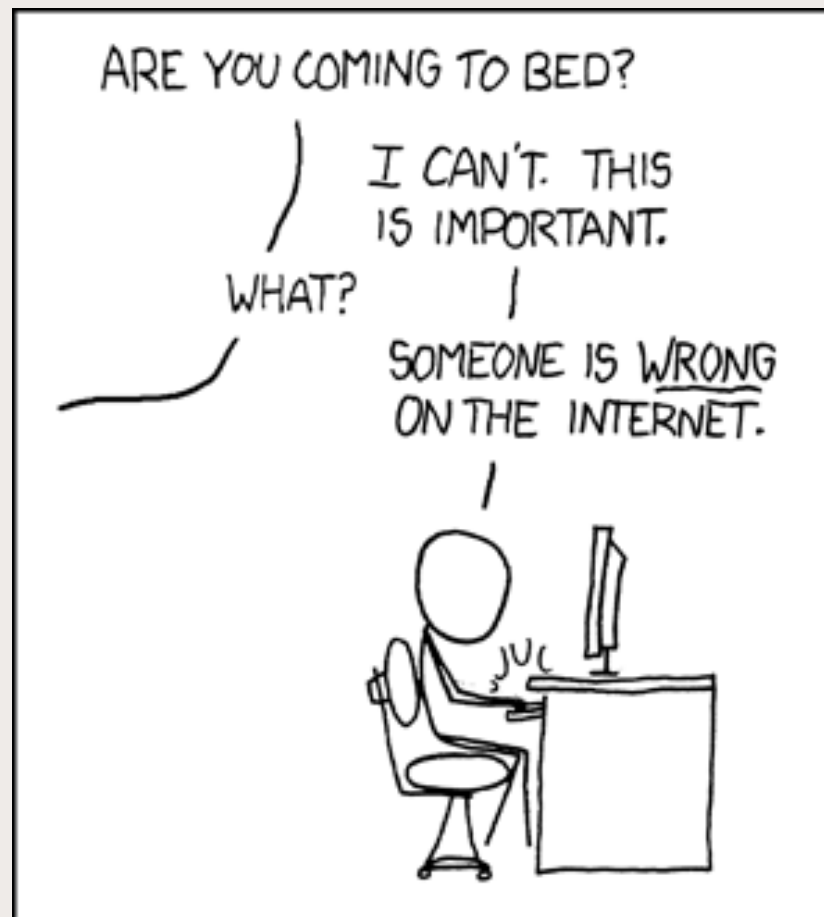
Une équipe interdisciplinaire



Exploiter le Web comme terrain d'enquêtes

Le Web : un espace de débats et de controverses

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

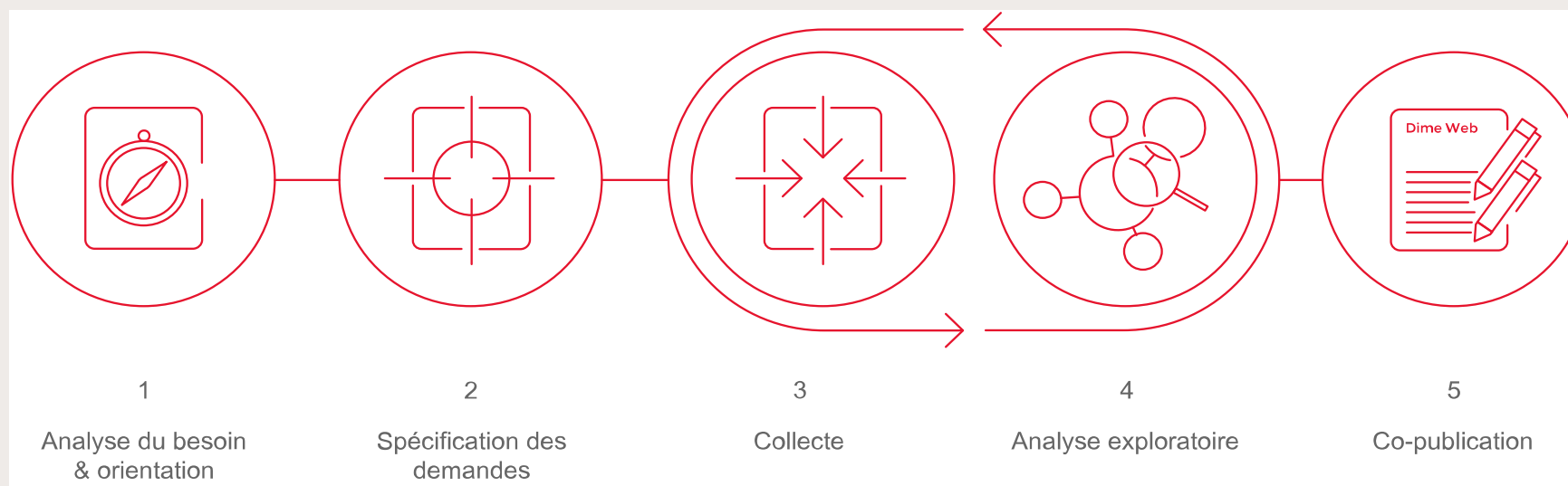


CC-BY-NC - Randall Munroe - XKCD

Une méthodologie pour réaliser un terrain Web

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

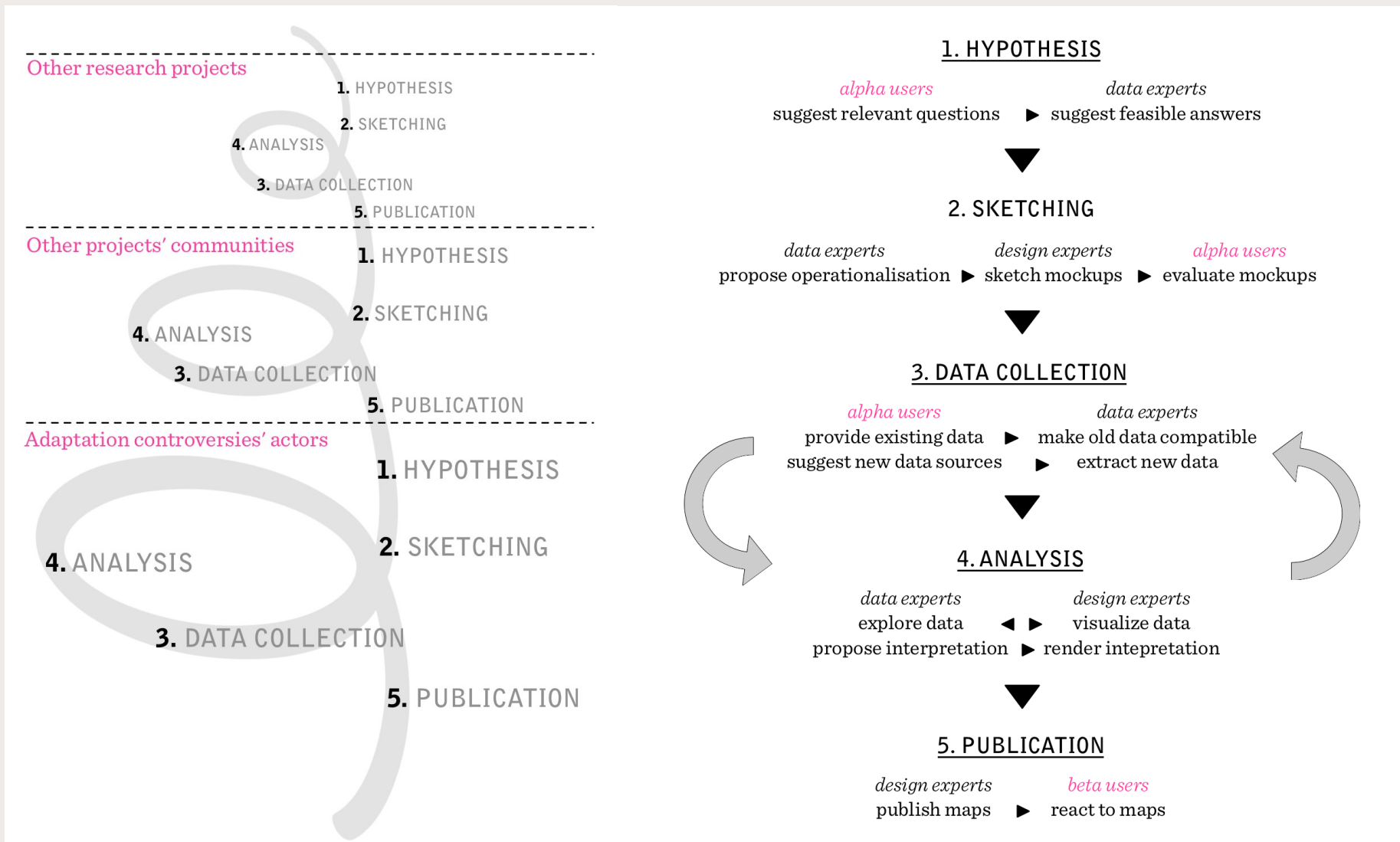
- Collecter, enrichir, nettoyer, visualiser et analyser des traces numériques
- Analyse de réseaux, archivage du web, analyse de controverses (ANT)
- Développement d'outils génériques
- Extraction ciblée de contenus
- Analyse Exploratoire de Données



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

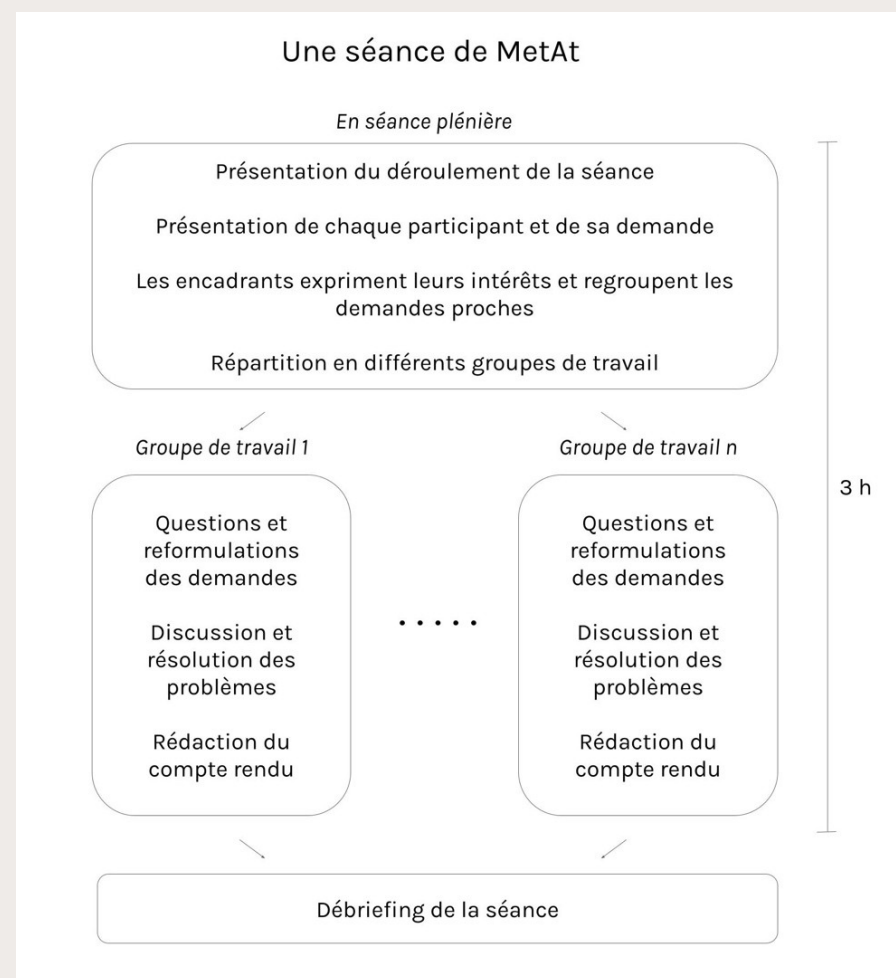
- centrés sur un ou quelques jeux de données
- sur plusieurs jours, en petits groupes
- rassemblant une diversité de profils
(ingénieurs, designers, académiques, journalistes...)
- confrontant les auteurs ou experts à leurs données



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

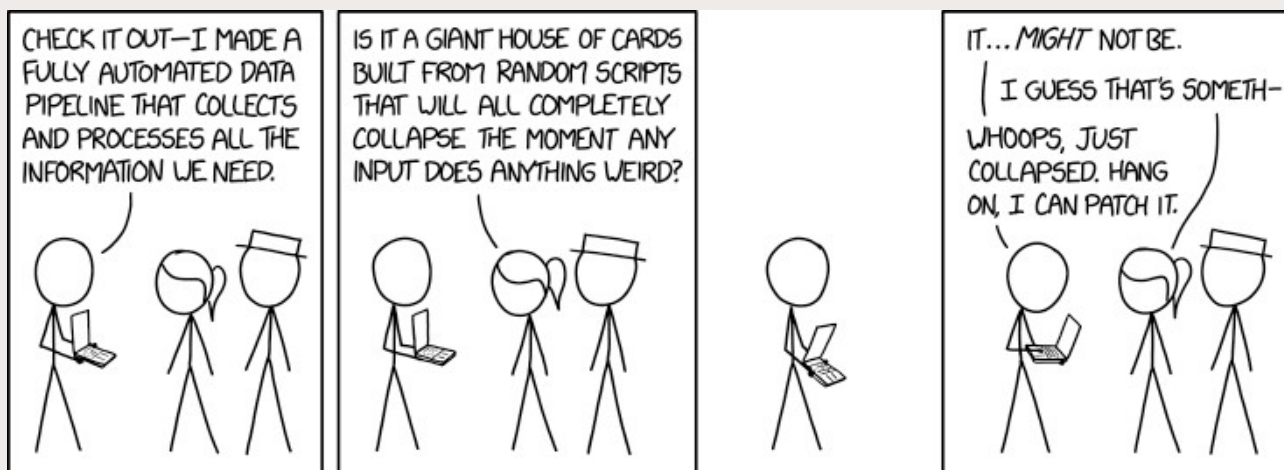
extraction ciblée
(source unique ou ensemble cohérent)
données structurées

traitement
du langage

analyse de réseau
(effets de communauté)

méthodes quantitatives,
statistiques...

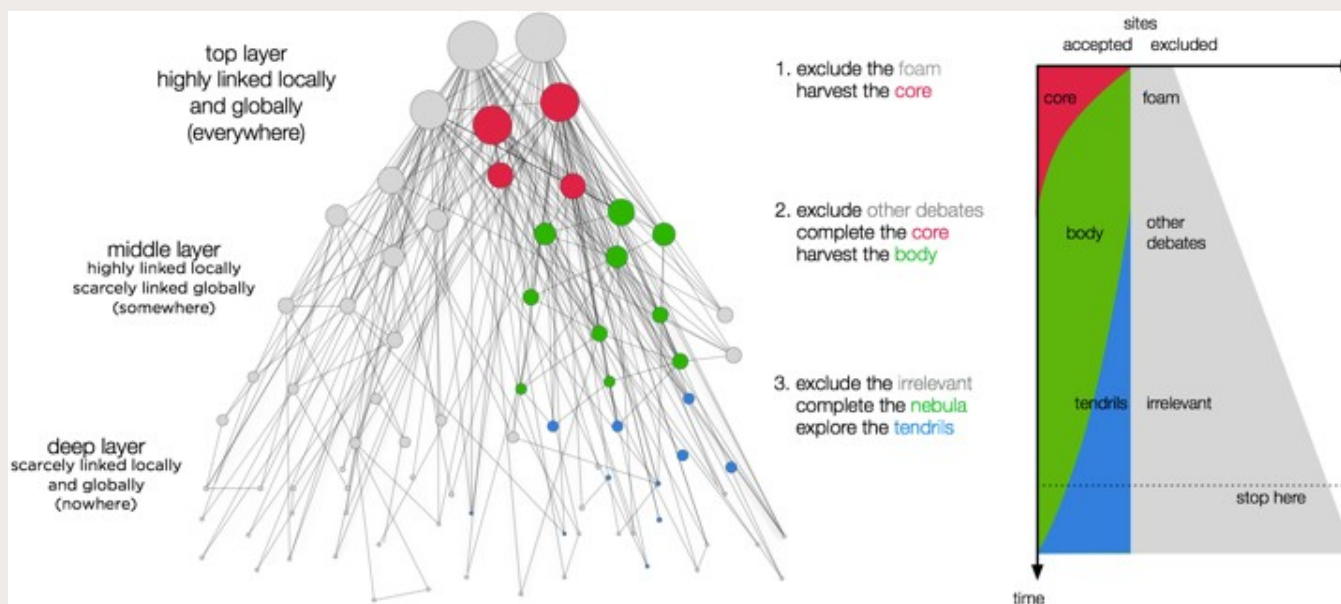
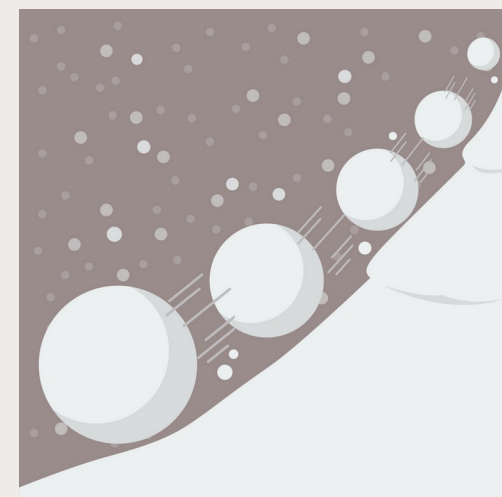
Problèmes : des données « sales » et un important coût en maintenance



CC-BY-NC - Randall Munroe - XKCD

Hyphe : 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

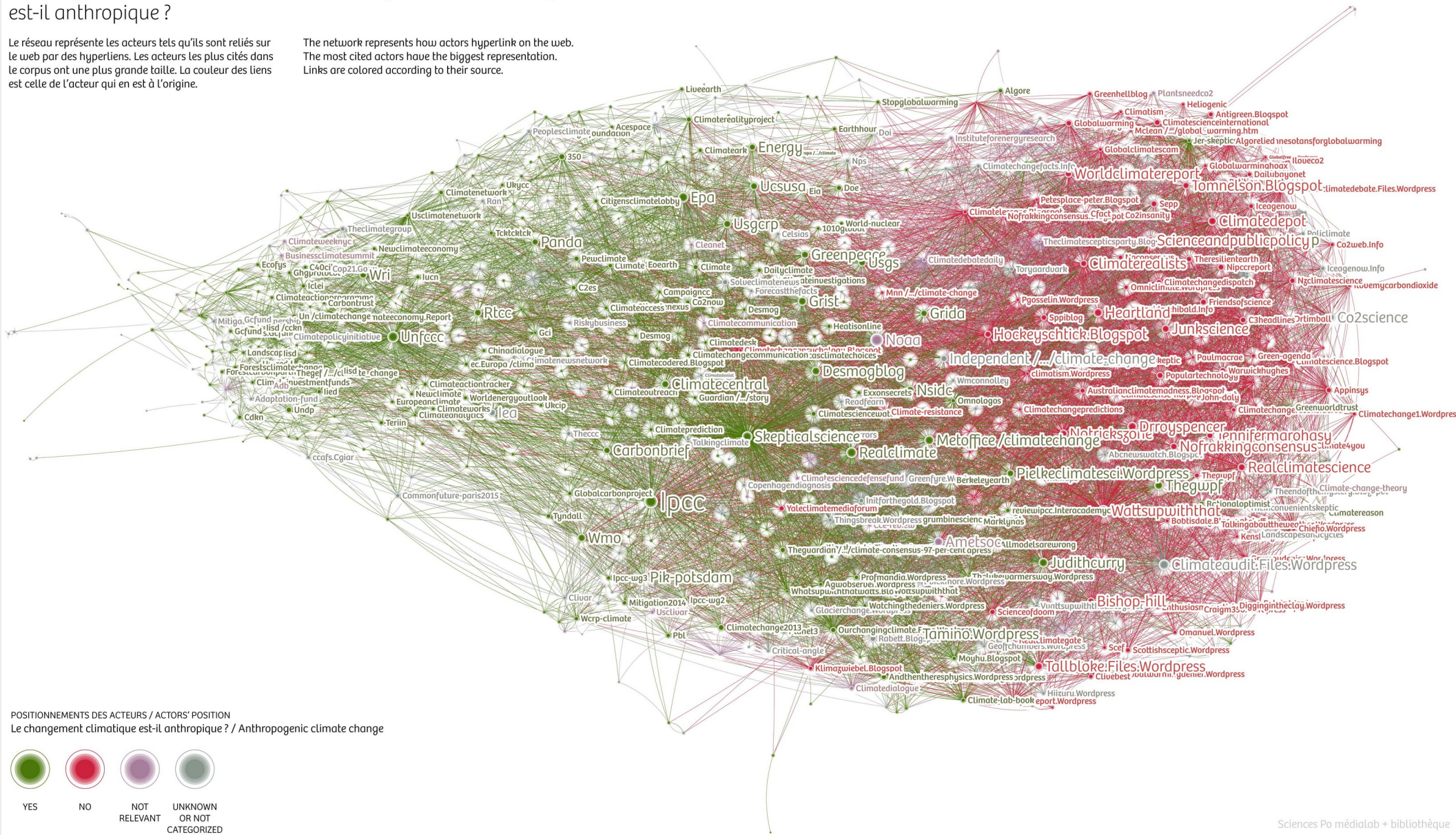
Corpus web sur le changement climatique : le changement climatique est-il anthropique ?

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

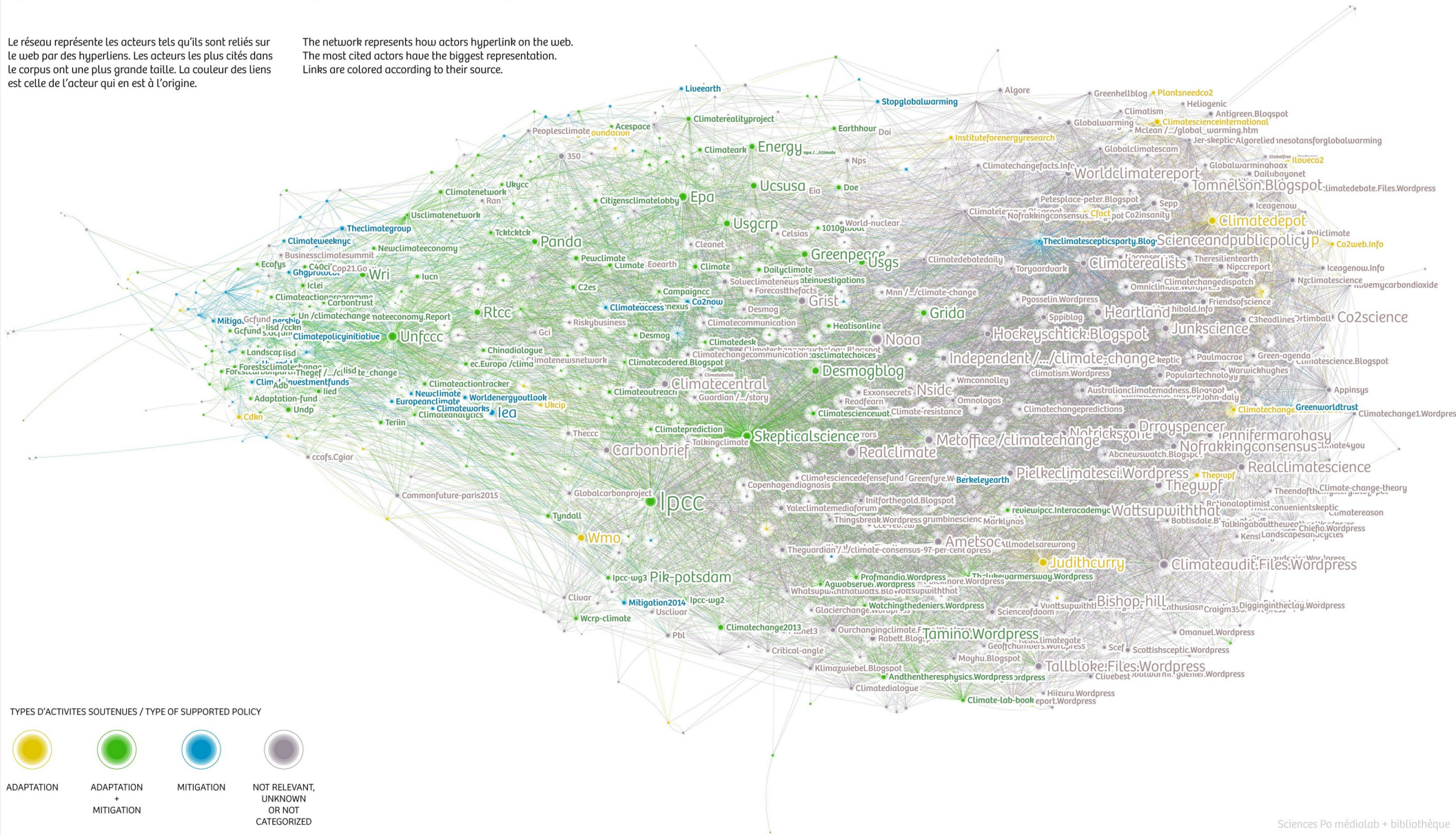
Corpus web sur le changement climatique : types d'activités soutenues

Web corpus on climate change: type of supported policy

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

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 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.



Analyse de fond : traiter les contenus texte

PRIVACY WEB CORPUS

SciencesPo MÉDIALAB, AXA Research Fund, 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 , WebM , How to embed ?

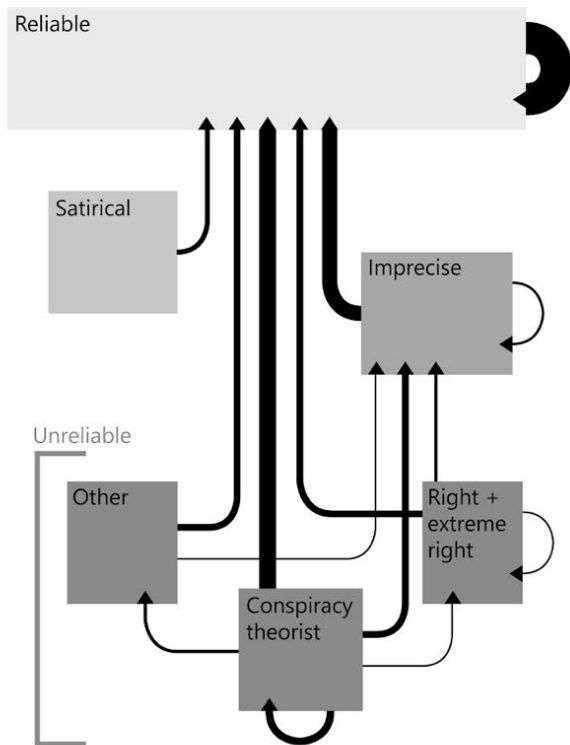
Topics

- Surveillance FR
- Business & Media
- Surveillance US
- Cybersecurity
- Big data & Analytics
- Data Regulation FR
- Cookies & Tracking
- Telec Operators FR
- Card and ID fraud

EXPLORE TOPICS

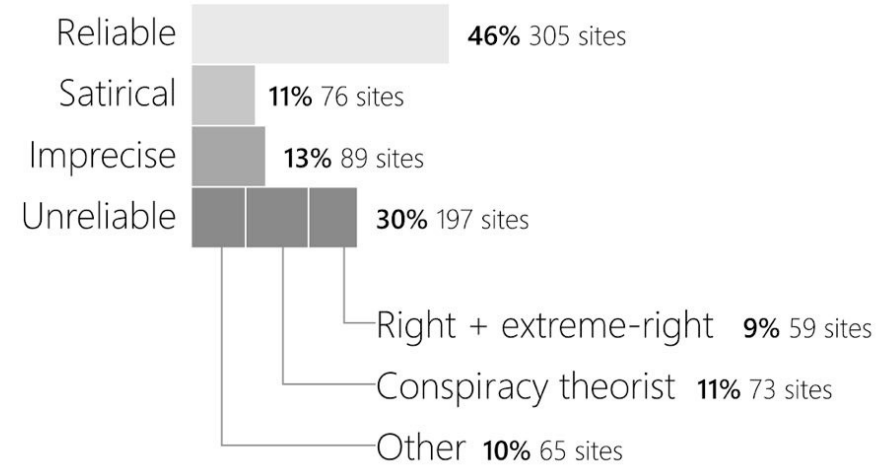
<http://tools.medialab.sciences-po.fr/privacy/>

Exploiter la directionnalité des hyperliens

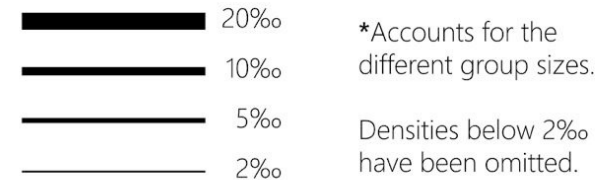


Most hyperlinks stem from the unreliable and aim at the reliable resources

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*



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

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 of screenshots illustrating the workflow of using Google bookmarklets for data scraping. It consists of several overlapping windows:

- Installation Guide:** A window titled "Install Google Bookmarklets" with the instruction "Drag & drop images below into your bookmark bar:" and two small Google logo icons.
- Search Results:** A Google search page for "digital humanities" showing search filters (All, News, Books, Images, Videos, More) and search tools. The results list includes "Digital humanities - Wikipedia" and "Digital Humanities | Stanford Humanities - Stanford Humanities Center".
- Redirect Dialog:** A "Redirect to Classic Google" dialog box with a language dropdown set to "en", a "How many results per page?" dropdown set to "100", and a "Redirect me!" button.
- Extract Dialog:** An "Extract Classic Google Results" dialog box showing search parameters: "Search for 'digital humanities' page 0 (with up to 100 urls per page)", "103 new results in this page", and buttons for "Keep existing results & continue to the next page" and "Download CSV with 103 urls".

Arrows indicate the flow of the process: from the installation guide to the search results, then to the redirect dialog, and finally to the extract dialog.

→ 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)

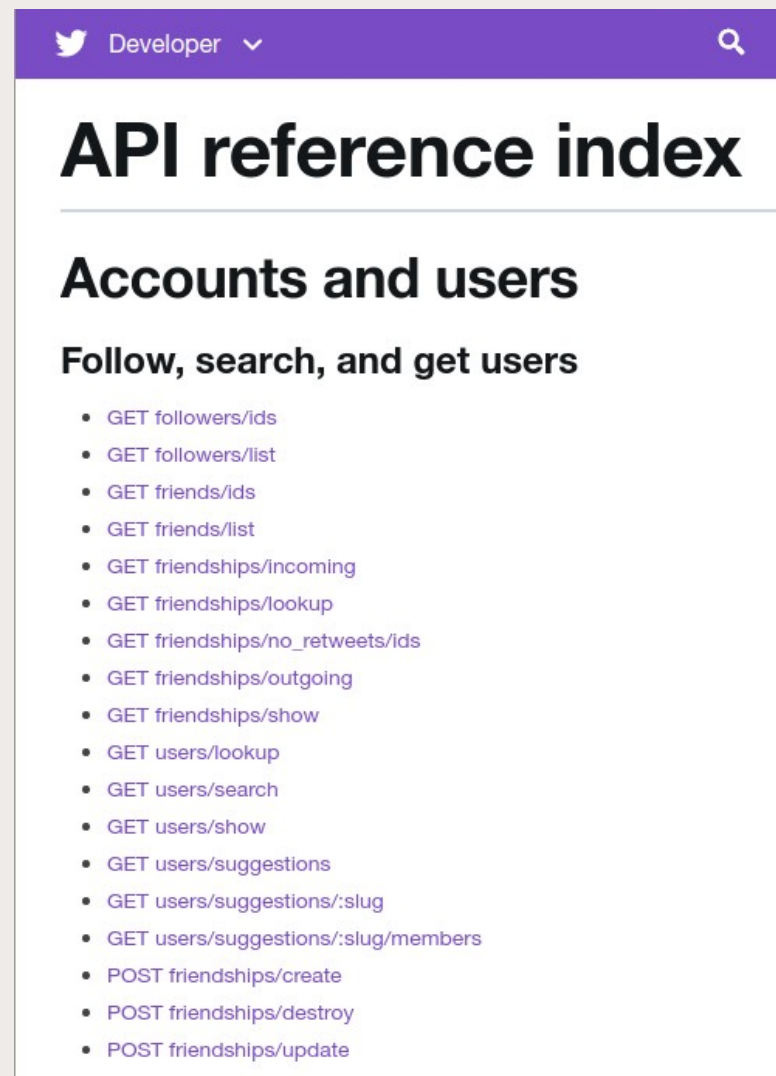


→ exemple : https://en.wikipedia.org/wiki/List_of_countries_by_carbon_dioxide_emissions

```
> var data = artoo.scrapeTable( ".wikitable", {headers: 'th'} );
undefined
> data.length;
49
> data[0];
Object {Country: " World", CO2 emissions (kt) in 2014[2]: "35,669,000", " % CO2
Emissions by Country": "100%", Emission per capita (t) in 2014[3]: "5.0"}
> artoo.saveCsv(data, "CO2-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", # 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 authored 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 authored 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 mentioned, separated by |
    "mentioned_user_ids", # list of digital IDs of users mentioned, separated by |
    "hashtags" # 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 couverture	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
```


CatWalk : sélection qualitative de tweets

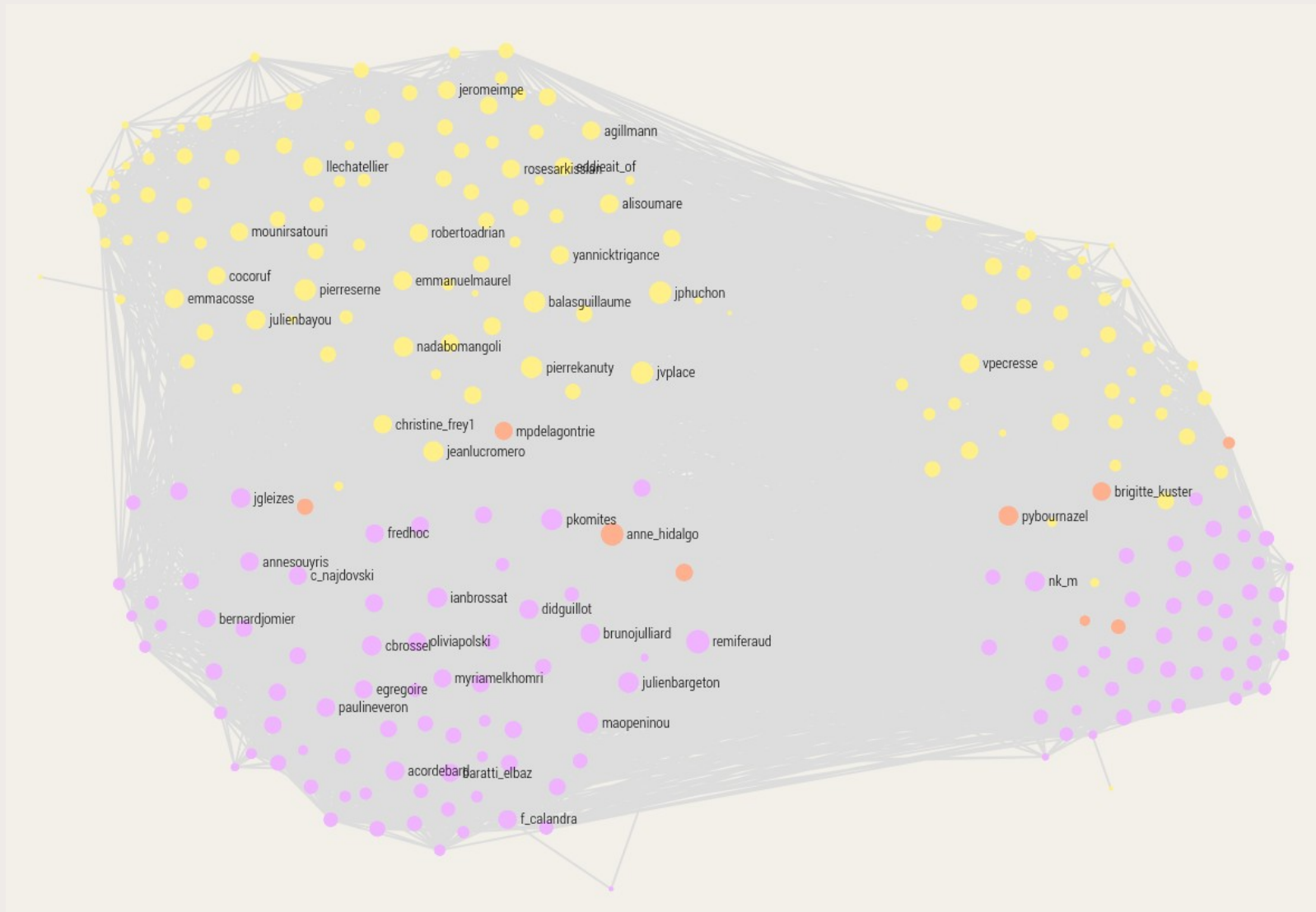
<https://medialab.github.io/catwalk/>

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

The screenshot displays the CatWalk web interface. At the top left, the text 'CATWALK' is visible. To its right are navigation buttons: 'prev', a central input field containing '0', and 'next'. Further right is a 'Download' button with a red circle containing '0' and a green circle containing '2'. Below these is a prominent green button labeled 'IN'. The main content area features a tweet from 'RE•WORK @teamrework' with a 'Follow' button. The tweet text reads: 'Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free ow.ly/4nfo2S #AI @open_ai' followed by the timestamp '7:15 PM - 29 Apr 2016'. Below the text is a photo of Elon Musk. Underneath the photo is a preview of the linked article: 'Inside OpenAI, Elon Musk's Wild Plan to...' with a sub-headline 'OpenAI wants to give away the 21st century's most transformative technology. In wired.com' and engagement icons for reply, retweet (7), and like (15). To the right of the tweet is a vertical sidebar with selection controls: 'previous', 'next', 'IN' (highlighted in green), 'OUT' (highlighted in red), 'UNDECIDED' (highlighted in grey), and 'save'. At the bottom of the interface, there is a footer area with the user profile '@teamrework' and a truncated version of the tweet text.

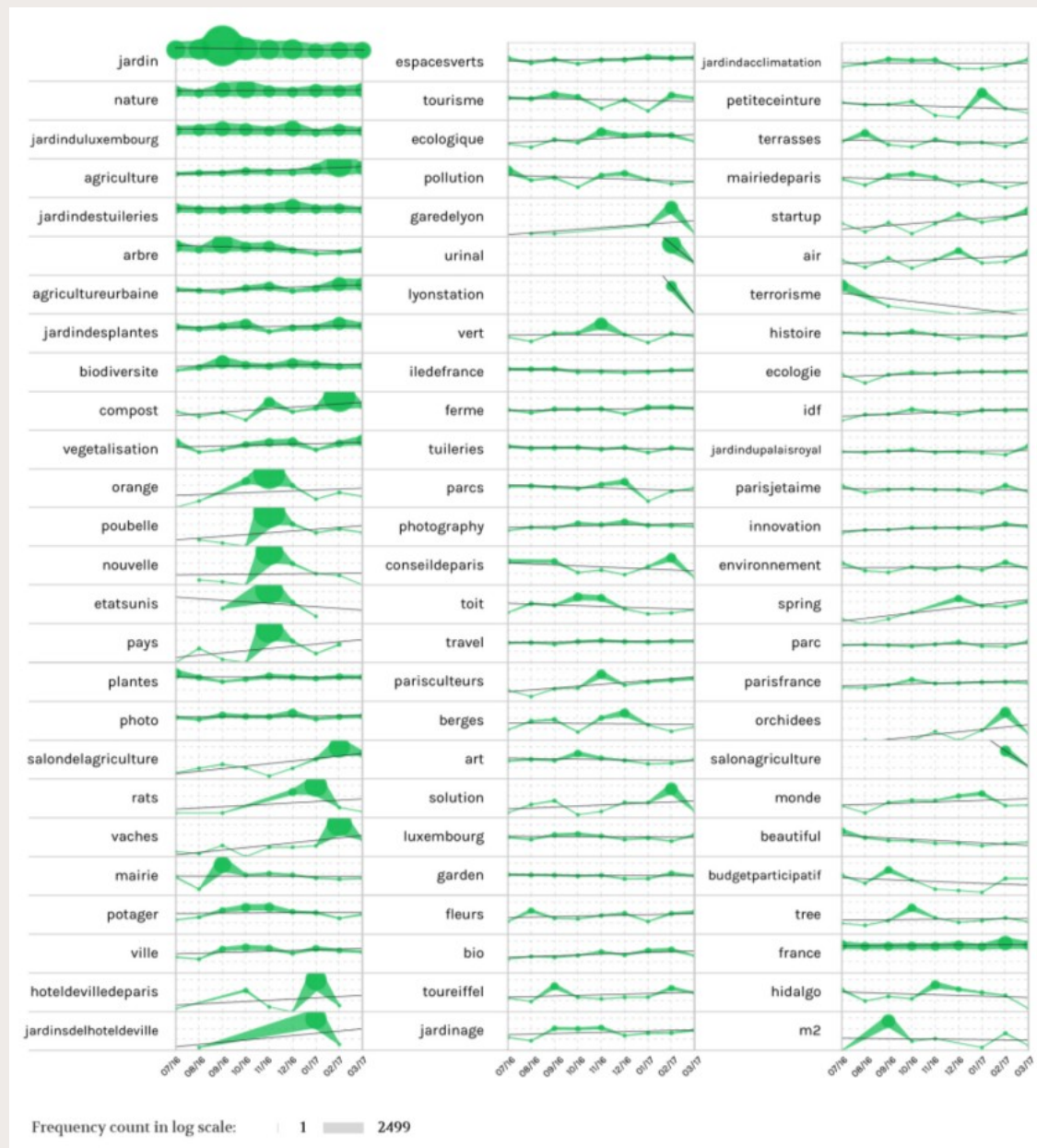
→ V2 en approche !

Explorer des réseaux de followers

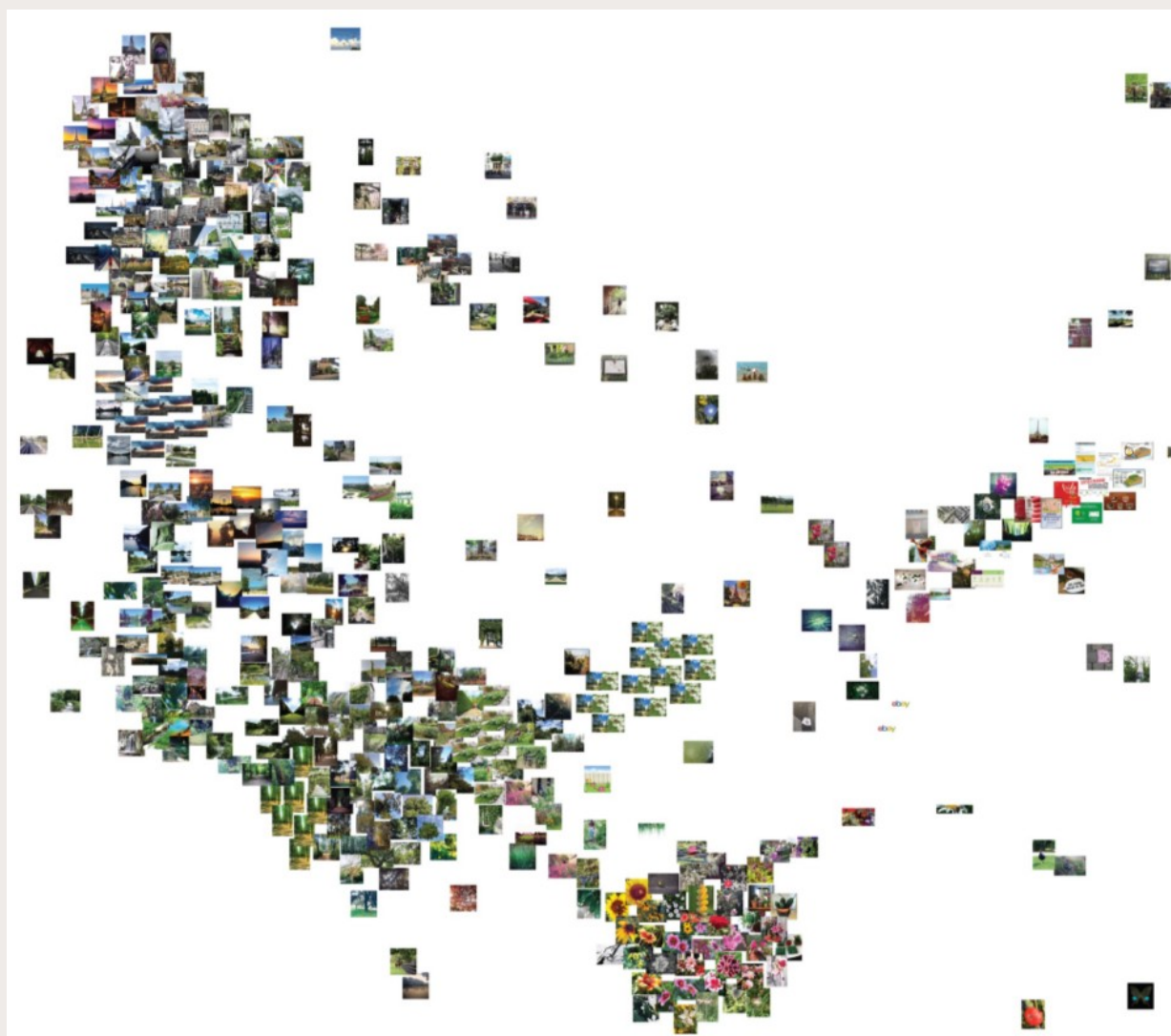


Liens de proximité Twitter entre les élus du Conseil Régional d'Île-de-France et du Conseil de Paris

Explorer les dynamiques temporelles

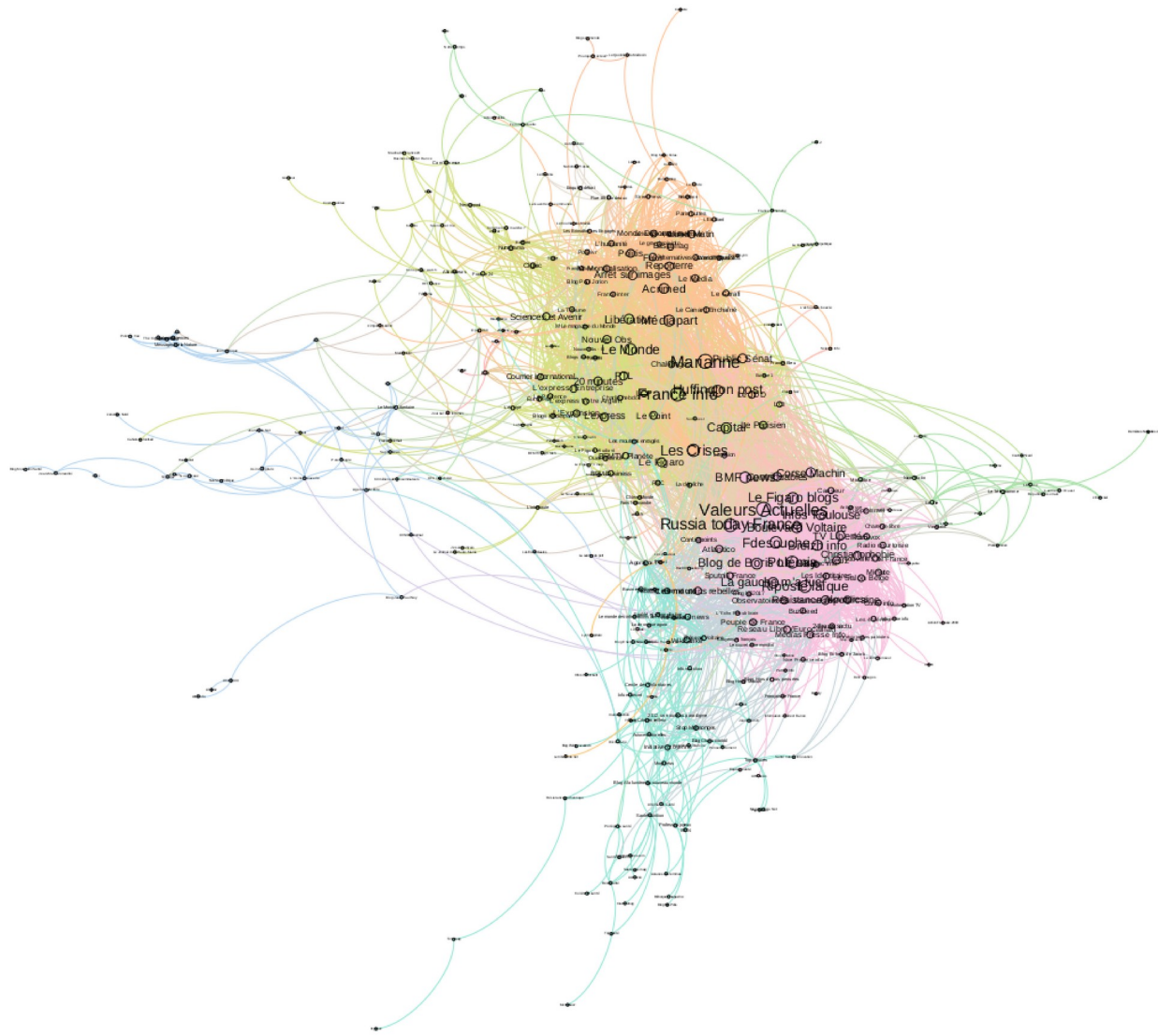


Explorer l'espace visuel d'un corpus



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

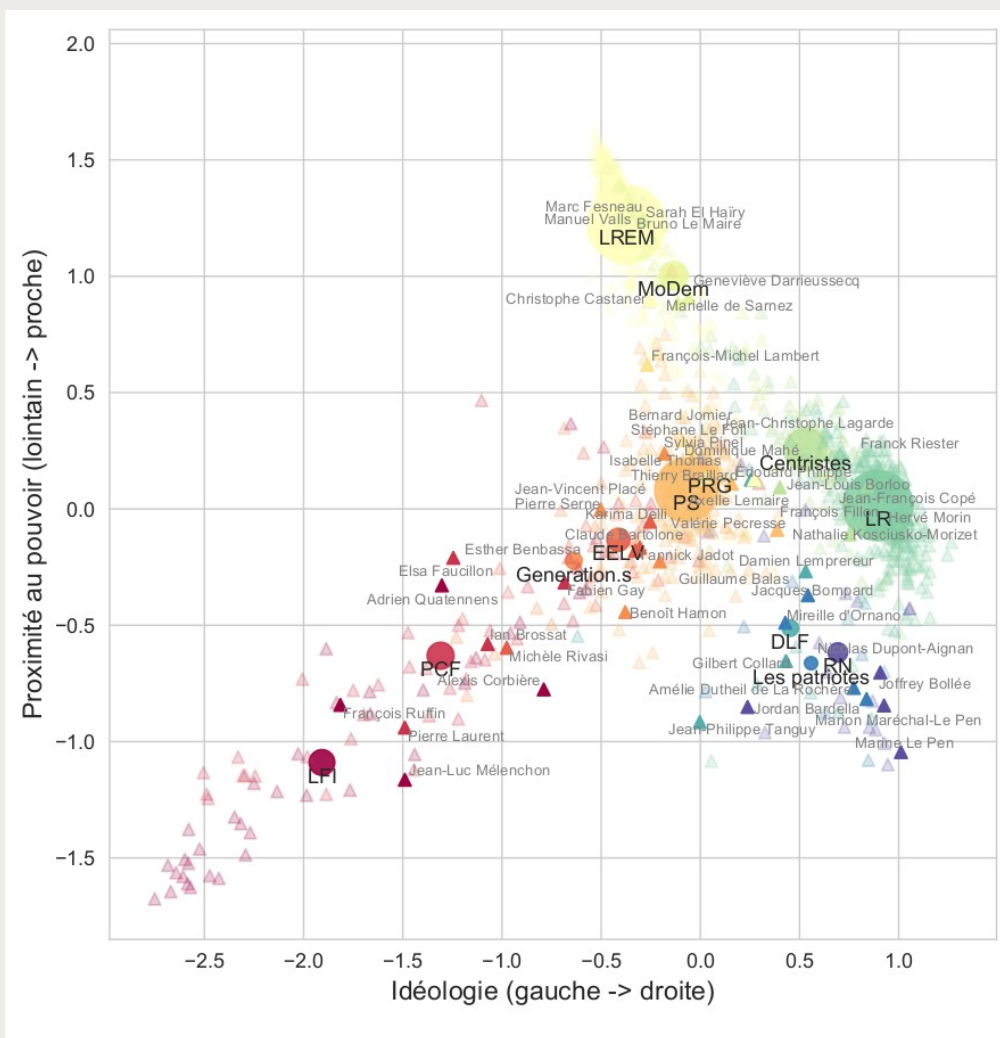
Explorer des réseaux de co-citation de sites web



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

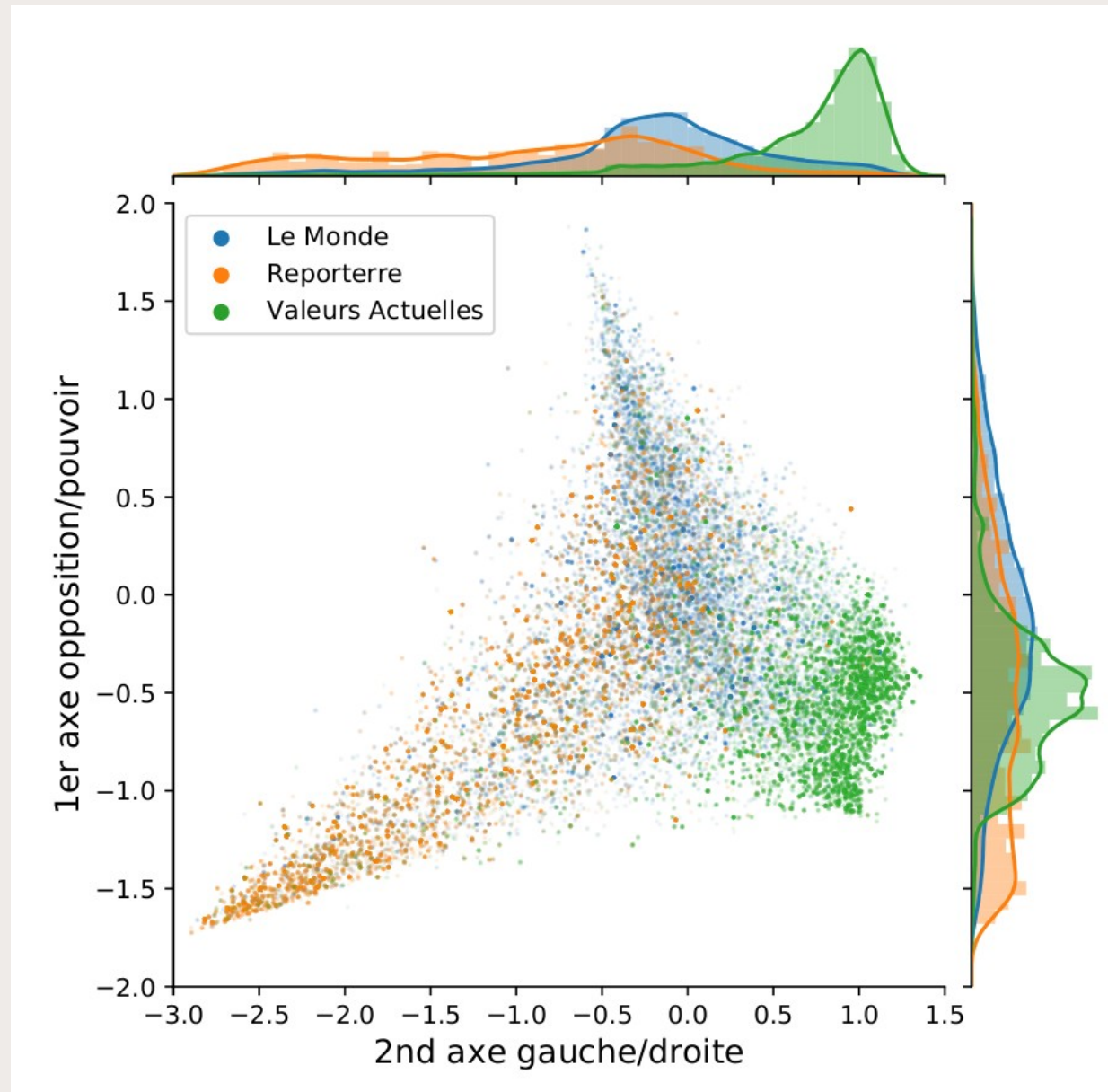
Cartographie de l'espace politique français des utilisateurs de Twitter à partir des followers de parlementaires (méthode de Barbera)



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>

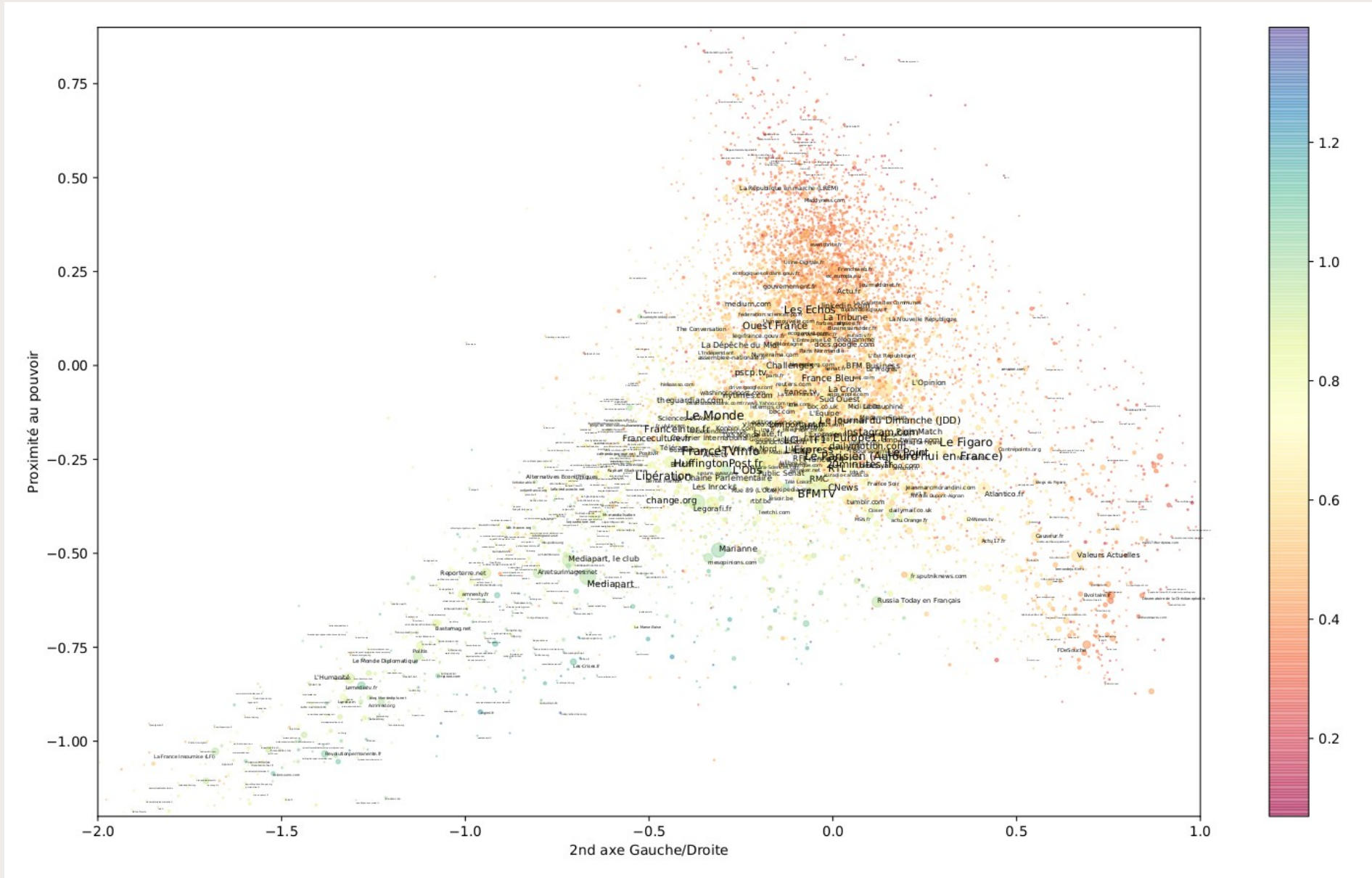
Étudier la propagation de l'info sur l'espace politique

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



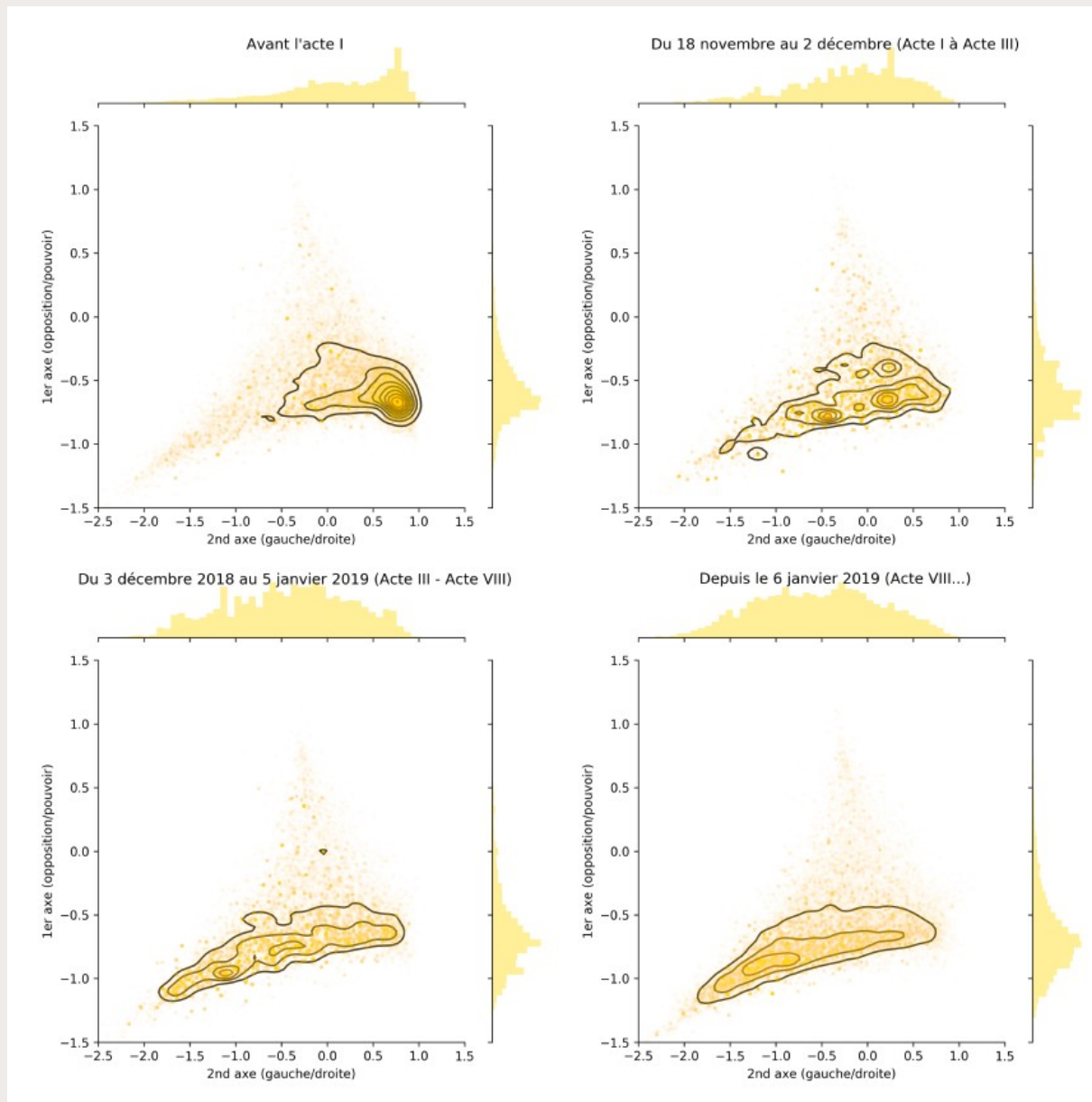
Étudier la propagation de l'info sur l'espace politique

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

Humanités numériques [modifier | modifier le code]

Les **humanités numériques** (ou *digital humanities*, abrégées "DH", voire **humanités digitales**²) sont un domaine de recherche, d'enseignement et d'ingénierie au croisement de l'informatique et des arts, lettres, sciences humaines et sciences sociales.

Sommaire [afficher]

Définition [modifier | modifier le code]

Les humanités numériques peuvent être définies comme l'application du « savoir-faire des technologies de l'information [et de l'informatique/infosciences] aux questions de sciences humaines et sociales »³.

Voir aussi [modifier | modifier le code]

Logiciels [modifier | modifier le code]

- Gephi est un logiciel libre open source , issu du projet e-Diaspora, permettant la visualisation, l'analyse et l'exploitation en temps réel de données relationnelles ou réseaux.
- IRaMuTeQ est un logiciel libre d'analyse de texte, développé par Pierre Ratinaud.
- Voyant Tools permet de visualiser et d'explorer des textes
- Prospero (PROgramme de Sociologie Pragmatique, Expérimentale et Réflexive sur Ordinateur - © Doxa) est un logiciel d'analyse de données textuelles qualifié par ses concepteurs de technologie littéraire pour les sciences humaines. Le logiciel a été conçu par le sociologue Francis Chateauraynaud et l'informaticien Jean-Pierre Charriau.
- Philcarto est un logiciel de cartographie. Le code n'en est pas libre, mais le logiciel est gratuit (freeware). Il fonctionne sur Windows.
- OpenRefine est un logiciel libre et gratuit de lissage de données (anciennement nommée Google refine).
- Le projet DIRT recense de très nombreux logiciels: DIRT@ [archive] (*Digital Research Tools* - en Anglais).

Articles connexes [modifier | modifier le code]

- Bibliothèque numérique
- Fouille de textes
- Littérature numérique
- Logométrie
- Moteur de recherche



Seealsology

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.

Source code available on [Github](#). Adding other languages requires the identification of the various "See also" sections. Feel free to contribute identifying them and [proposing new languages we pull request!](#)

Place your list of wikipedia articles here
ex: <http://en.wikipedia.org/wiki/Gutenberg>

Stop words (press enter or separate the words with a comma)

Wikipedia: Category: File: wikisource: Commons:

list of index of categories of portal disambiguation

outline of add a word and press Enter

Output:

START CRAWLING

from [DerevyDesign](#) & [médialab 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

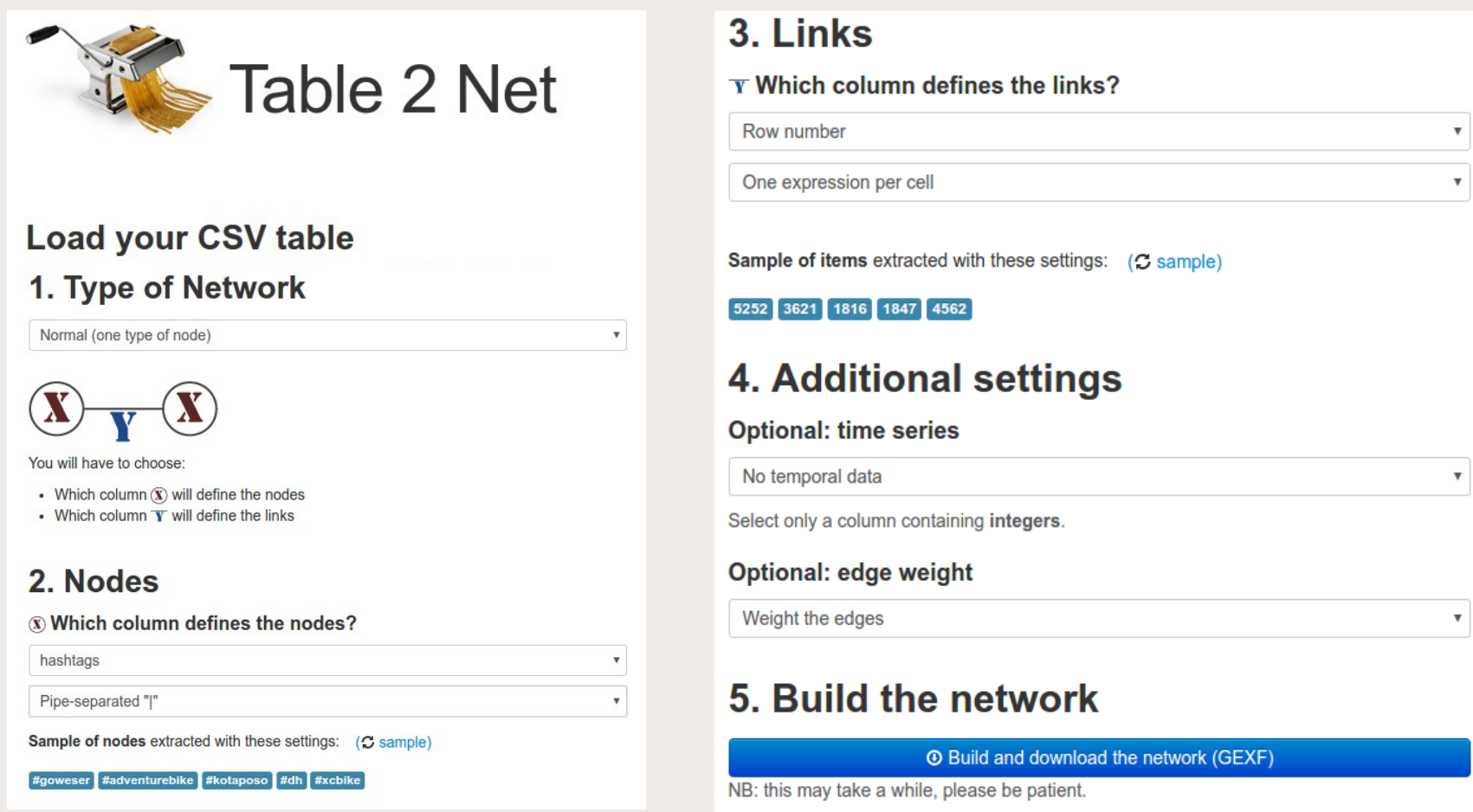


Table 2 Net

Load your CSV table

1. Type of Network

Normal (one type of node)

2. Nodes

Which column defines the nodes? (X)

hashtags

Pipe-separated "|"

Sample of nodes extracted with these settings: (sample)

#goweser #adventurebike #kotaposo #dh #xcbike

3. Links

Which column defines the links? (Y)

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

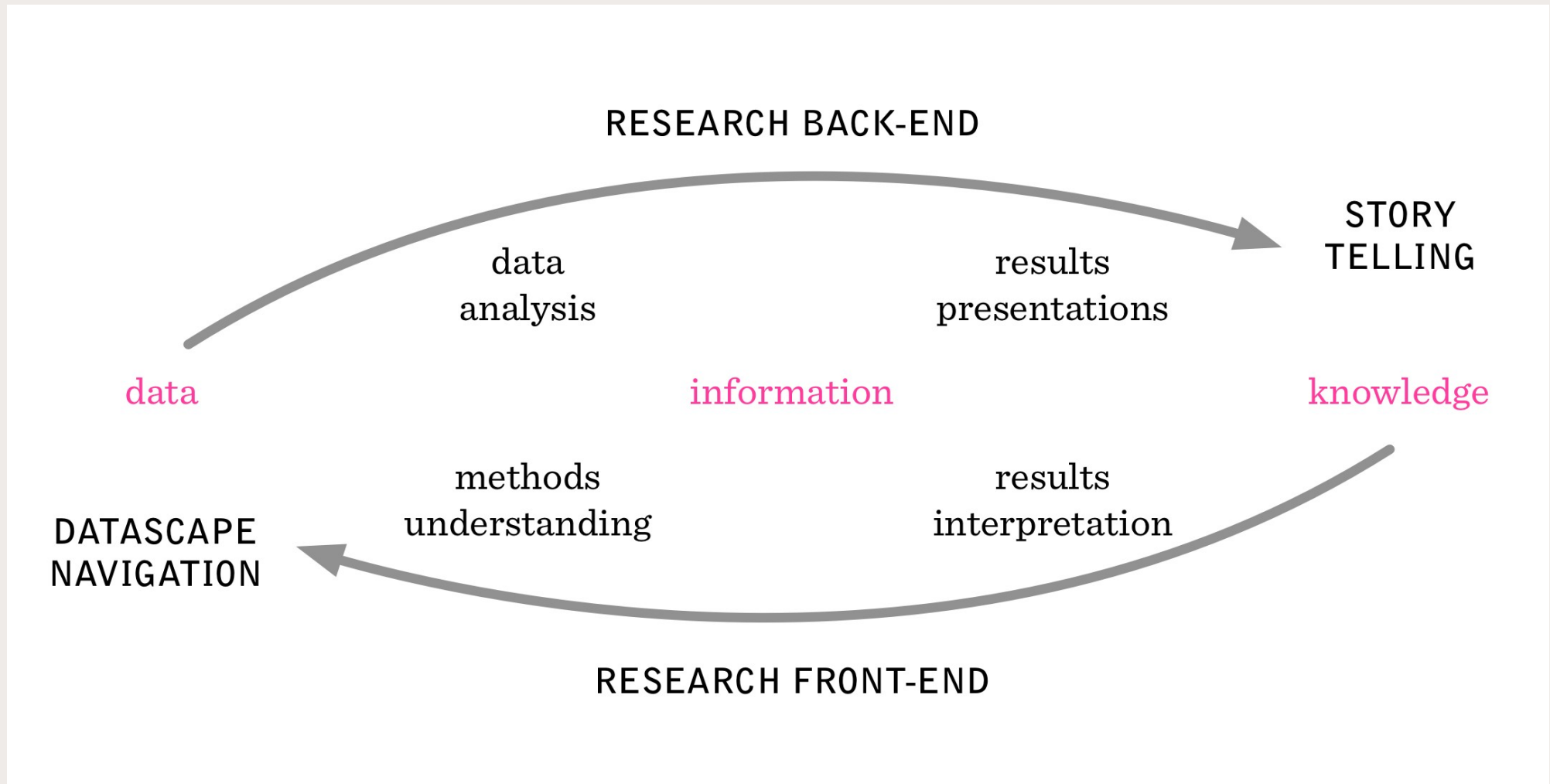
Build and download the network (GEXF)

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

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

Les « datascares » : zoomer/dézoomer

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



RICardo : visualiser les échanges au 19ème siècle

The screenshot shows the RICardo website with a navigation bar containing 'Trade', 'Data', 'Glossary', and 'About'. The main content area features the title 'The RICardo Project' and a sub-heading 'Trade between Nations from c. 1800 to 1938'. Below this is a descriptive paragraph about the project's scope and a small image of a historical trade table.

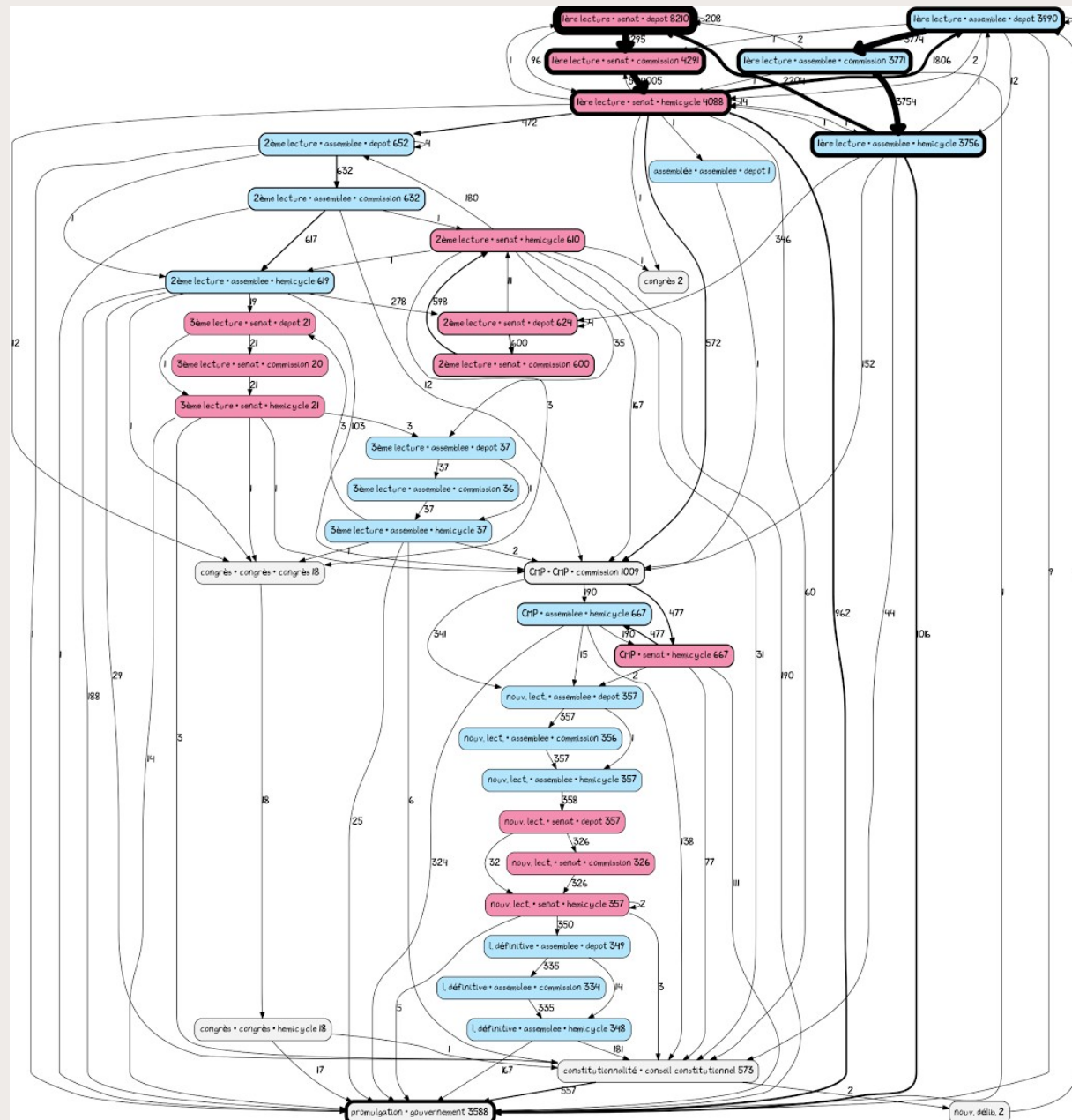
IMPORTATIONS.	EXPORTATIONS.
161,067,000f	105,305,000f
97,639,000	123,504,000
72,467,000	84,937,000
38,245,000	14,635,000
24,007,000	5,207,000
20,867,000	10,248,000
14,813,000	8,050,000
8,860,000	6,352,000
8,591,000	230,000
7,182,000	
5,043,000	

Below the overview, six interactive cards are displayed, each with a title, a brief description, and a representative visualization:

- Reporting**: Reporting depicts the distribution of trade by partners of a reporting. (Visualization: Stacked bar chart)
- Partner**: Partner depicts an entity's trade as it is reported by the reportings. (Visualization: Horizontal bar chart)
- Bilateral**: Bilateral focuses on trade between two countries and allows for a comparison between their trade statistics. (Visualization: Line chart with two series)
- World**: World provides different historical series of world trade and its breakdown by countries. (Visualization: Line chart with multiple series, labeled 1920)
- Metadata**: Metadata provides an overview of bilateral and total trade data availability. (Visualization: Stacked bar chart, labeled 1920)
- Exchange rates**: Exchange rates provides a visualization of the exchange rate database. (Visualization: Line chart with multiple series, labeled 1840 and 1860)

<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

Sénat

08/01/2014 → 15/01/2014

32 amendements

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.

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

LFDLL : retrouver les amendements discutés et votés

Projet de loi relatif au renseignement

Dossier Sénat Dossier Assemblée Open Data

< Voir les articles du texte Vue par articles Trié par sort final ?

Dépôt	1 ^{ère} Lecture				Commission Mixte Paritaire	
	Assemblée		Sénat		Commission Mixte	Assemblée
Gouvernement	Commission	Hémicycle	Commission	Hémicycle	Commission	Hémicycle
Projet de Loi	Commission	Hémicycle	Commission	Hémicycle	Commission	Hémicycle

Amendement 190

10/04/2015 Rejeté

Sujet : Article 2

Signataires :
M. Coronado, M. Molac, M. Cavard, Mme

Exposé des motifs :
L'IMSI-catcher sur les correspondances

Texte :
À la première phrase de l'alinéa 37, après le mot :
« exceptionnelle »,
insérer les mots :
« afin de prévenir un acte de terrorisme ».

Amendement 255
Les républicains
Sort: non-voté
10/04/2015

LÉGENDE

GDR	SRC	ECOLO	RRDP	UDI	LES-REP	NI
Gouvernement	Adopté	Rejeté	Non voté			

NosDéputés.fr Assemblée Nationale Permalien

LFDLL : remonter aux débats parlementaires

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

Projet de loi pour une République numérique

< Voir les articles du texte
Vue « échiquier politique » ▾
?

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

LÉGENDE

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

Les républicains

Après l'article 1er

Frédéric Lefebvre
Lire les interventions

Patrice Martin
Lire les interventions

Lionel Tardy
Lire les interventions

Philippe Gosselin
Lire les interventions

Laure de La Roubin
Lire les interventions

1816 mots

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](#)

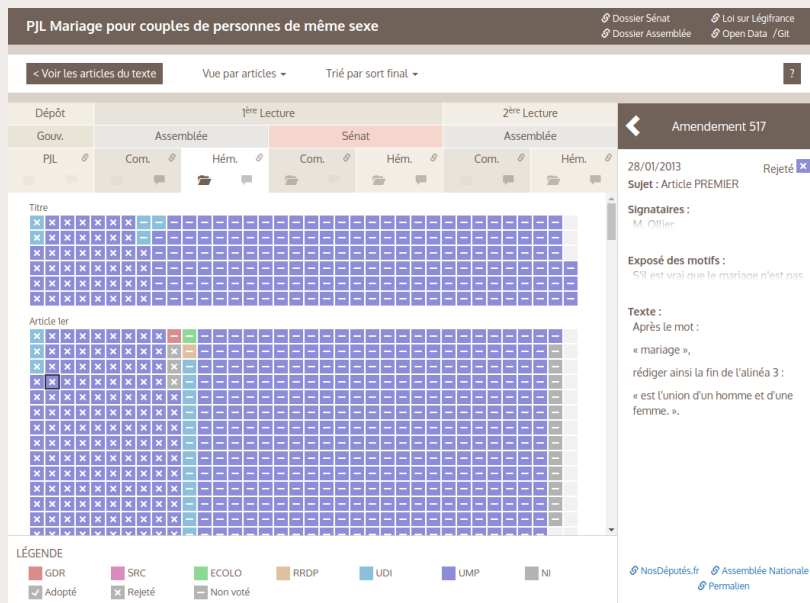
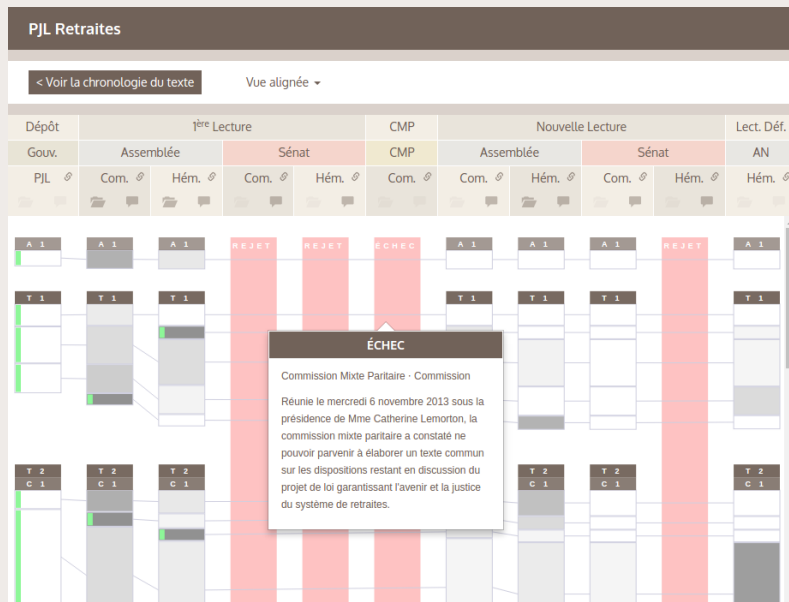
Catherine Vautrin, présidente
Quel est l'avis de la commission ?

[Laisser un commentaire](#)

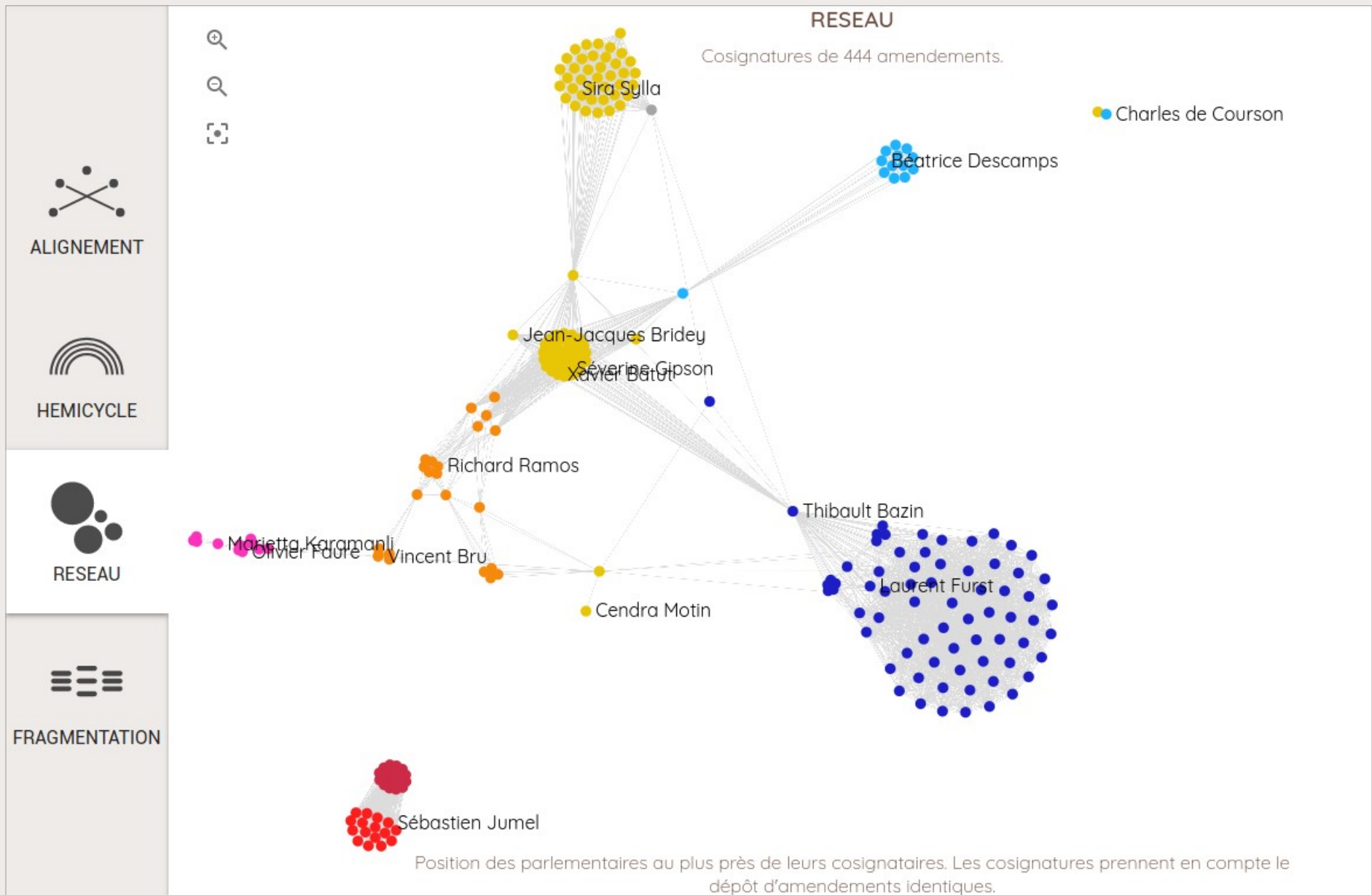
Luc Belot, rapporteur de la commission des lois constitutionnelle de la République
Je remercie notre collègue Lefebvre pour avoir rappelé le caractère assez innovant de notre démarche avec ce texte – un grand nombre d'entre vous l'a d'ailleurs également fait lors de la discussion générale – laquelle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

[Laisser un commentaire](#)

LFDLL : révéler les techniques d'obstruction



LFDLL : l'alignement et la fragmentation des partis



À vos questions !

<https://medialab.sciencespo.fr>

benjamin.ooghe@sciencespo.fr

@boogheta@paille.fr [@boogheta](https://twitter.com/boogheta) [@medialab_ScPo](https://twitter.com/medialab_ScPo)