



# Méthodes, outils et limites pour collecter et analyser le web et les réseaux sociaux

Benjamin Ooghe-Tabanou

## ► To cite this version:

Benjamin Ooghe-Tabanou. Méthodes, outils et limites pour collecter et analyser le web et les réseaux sociaux : Conduire des recherches à partir du web vivant ou comment faire face à la volatilité des contenus sur le web. Faire réseau autour des archives du web, usages et opportunités : journée de lancement du projet ResPaDon, Projet RESPADON, May 2021, Lille, France. hal-03904146

**HAL Id: hal-03904146**

**<https://sciencespo.hal.science/hal-03904146>**

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

# Méthodes, outils et limites pour collecter et analyser le web et les réseaux sociaux

Lancement projet ResPaDon

Session 4 « *Conduire des recherches à partir du web vivant ou  
comment faire face à la volatilité des contenus sur le web* »

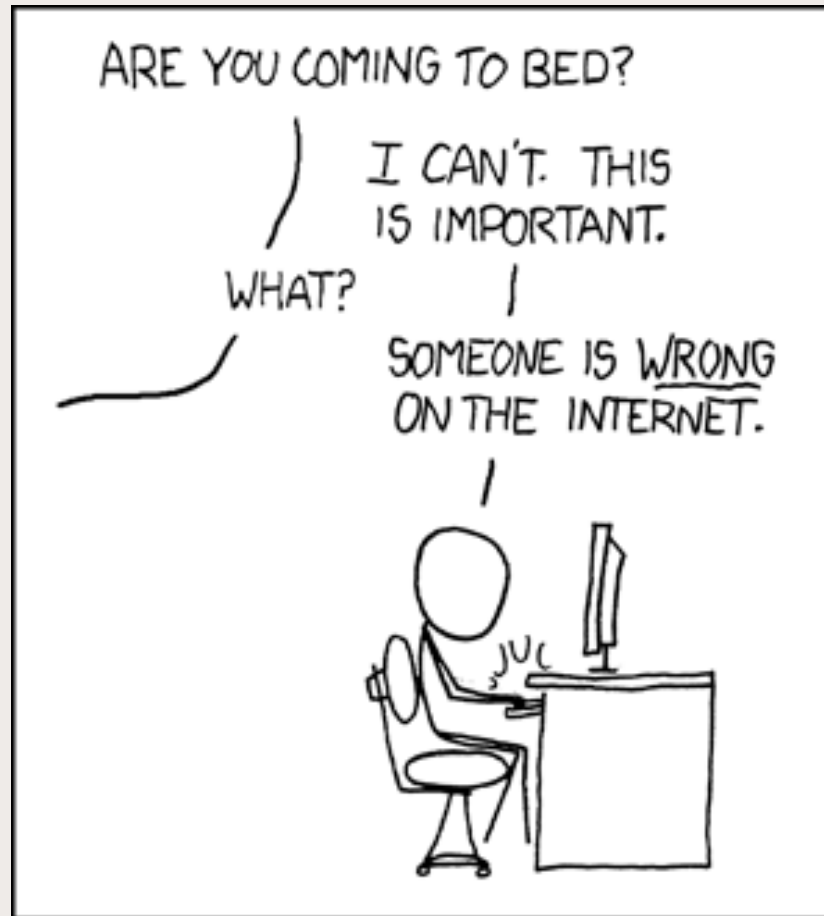
17 mai 2021

Benjamin Ooghe-Tabanou

Sciences Po médialab (@medialab\_ScPo)

# Exploiter le Web comme terrain d'enquêtes SHS

Le Web : un espace de dialogue et de débats



CC-BY-NC - Randall Munroe - XKCD

→ Collecter, enrichir, nettoyer, visualiser & analyser des traces numériques

# Bruno Latour, fondateur du médialab



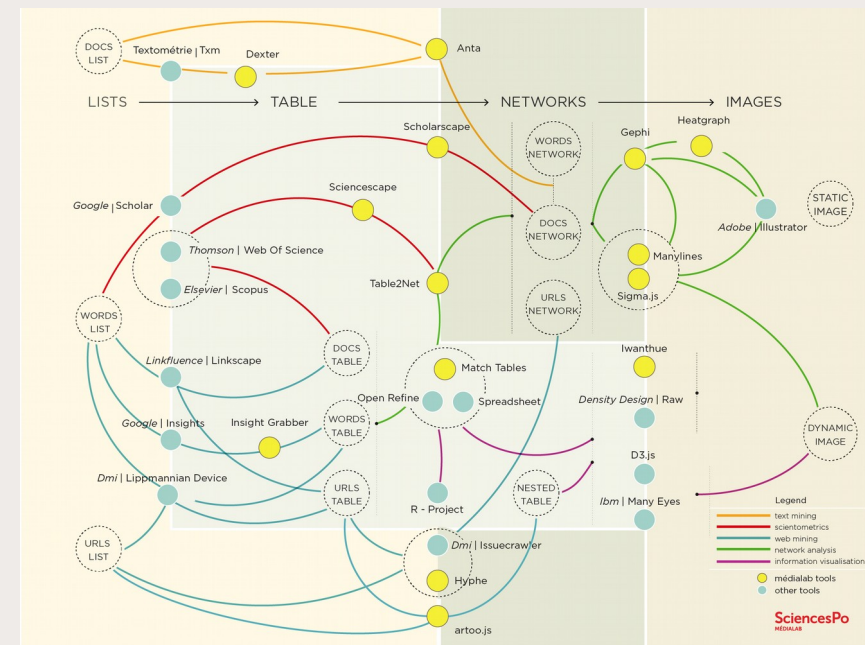
« *Google is nice,  
but we need  
something better* »

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

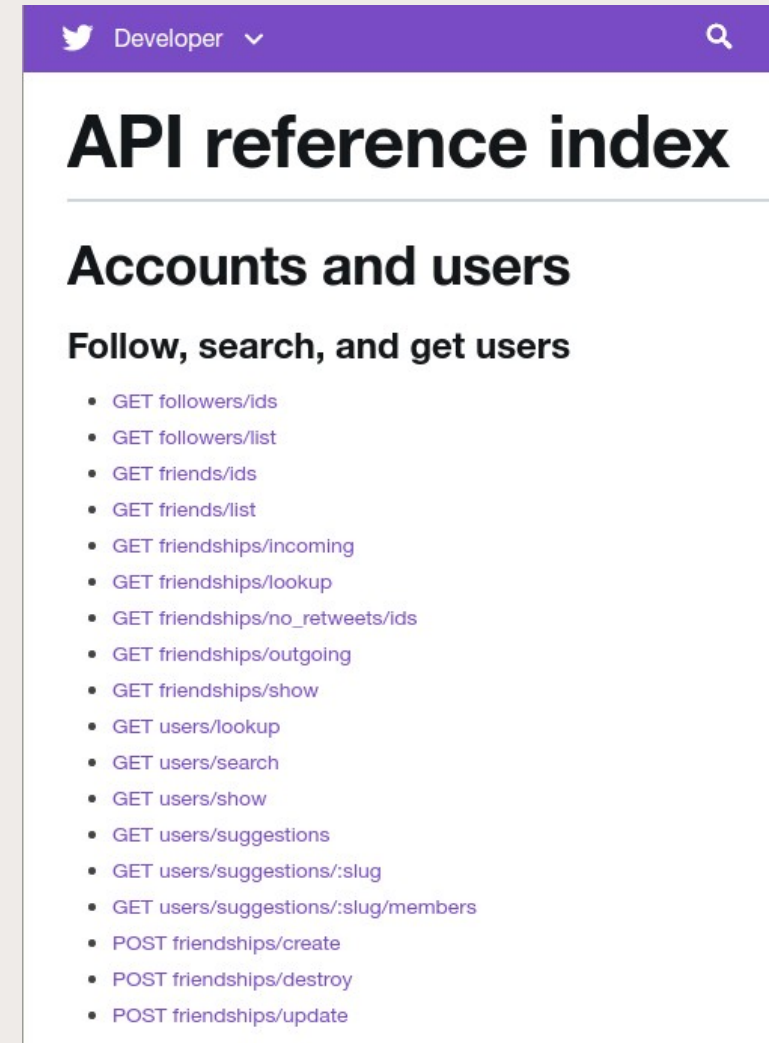


# Développement d'outils au service des SHS

- 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)
- **Documentation académique** des outils & méthodes

# Accès contrôlé via les APIs des plateformes

- « Application Programming Interface »
- Avantages :
  - données structurées « propres »
  - accès à de gros volumes
  - relative complétude
  - accès à des informations d'usage
- Problèmes :
  - accès à de gros volumes
  - limitation des appels
  - « boîtes noires »
  - risques de refermeture (ex : Twitter V2)
  - dépendance à la vision des plateformes



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

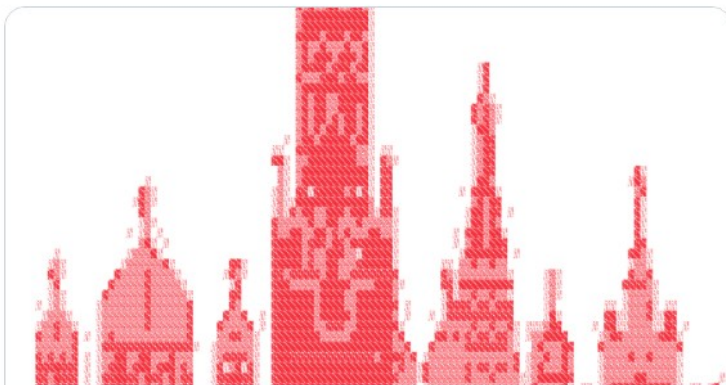


# De nombreuses métadonnées à exploiter



ScPo médialab  
@medialab\_ScPo

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 mentionned, separated by |
    # list of digital IDs of users mentionned, separated by |
    # list of hashtags used, lowercased, separated by |
]
```



# 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, etc.
  - les conversations et médias associés
  - des utilisateurs et leurs mentions

```
{
  "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
```

# Différents modes d'extraction de données web

2 approches bien distinctes aux cibles et résultats différents

## CRAWLING Vs. SCRAPING

fouille systématique  
(sources multiples hétérogènes)  
contenus textuels & hyperliens

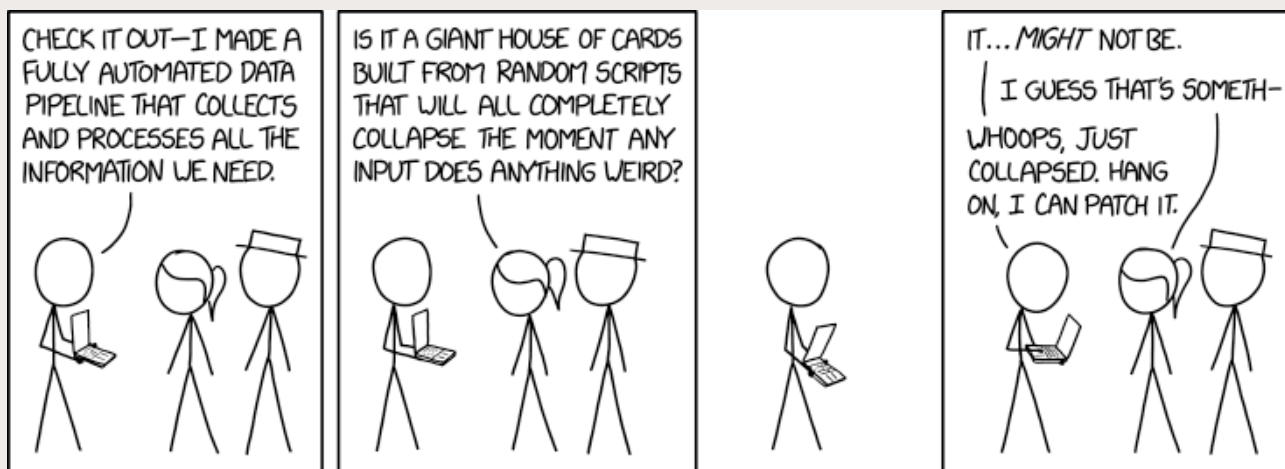
traitement  
du langage

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

# 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

**Install Google Bookmarklets**  
Drag & drop images below into your bookmark bar:

**Redirect to Classic Google**  
Which language?   
How many results per page?   
You will be redirected to the following url:  
`https://encrypted.google.com/search?q=digital%20humanities&hl=en&num=100&start=0`  
Redirect me!

**Extract Classic Google Results**  
Search for "digital humanities"  
page 0 (with up to 100 urls per page)  
103 new results in this page  
Keep existing results & continue to the next page  
Download CSV with 103 urls

**Digital humanities - Wikipedia**  
[https://en.wikipedia.org/wiki/Digital\\_humanities](https://en.wikipedia.org/wiki/Digital_humanities)

**Digital Humanities | Stanford Humanities - Stanford Humanities Center**  
[shc.stanford.edu/digital-humanities](https://shc.stanford.edu/digital-humanities)

→ url, name, row, description, date

# Creuser le web avec Minet

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



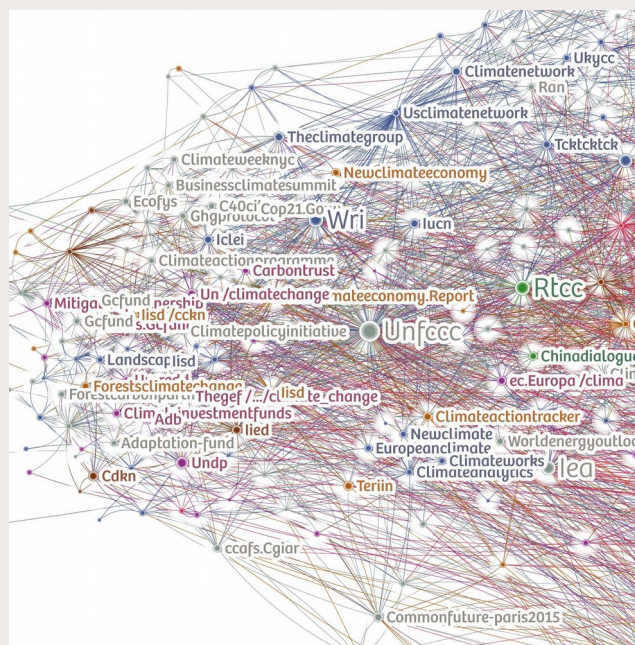
- Outil pour utilisateurs avancés (ligne de commande)
- Extraire des contenus textes, des liens, des images, etc. à partir de listes d'URLs
- Scraping :
  - posts & groupes Facebook
  - recherche Twitter
- APIs :
  - Crowdtangle (métriques Facebook)
  - friends/followers Twitter
  - métadonnées vidéos Youtube



# Hyphe : un crawler orienté par la recherche

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

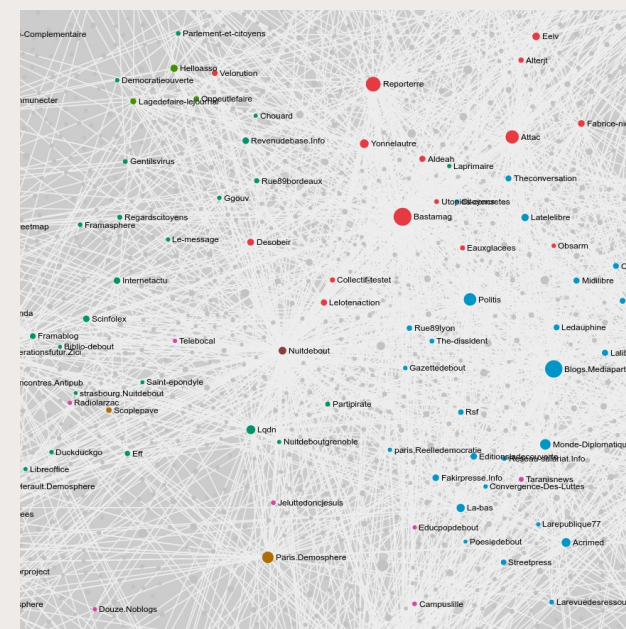
- Les liens hypertextes : nouveaux révélateurs de relations entre acteurs d'une thématique
- Créer un corpus documentaire
  - « acteurs web » & contenus textuels respectifs
  - liens hypertextes entre ces acteurs
- Études exploratoires ou de controverses dans tous les domaines



<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/>



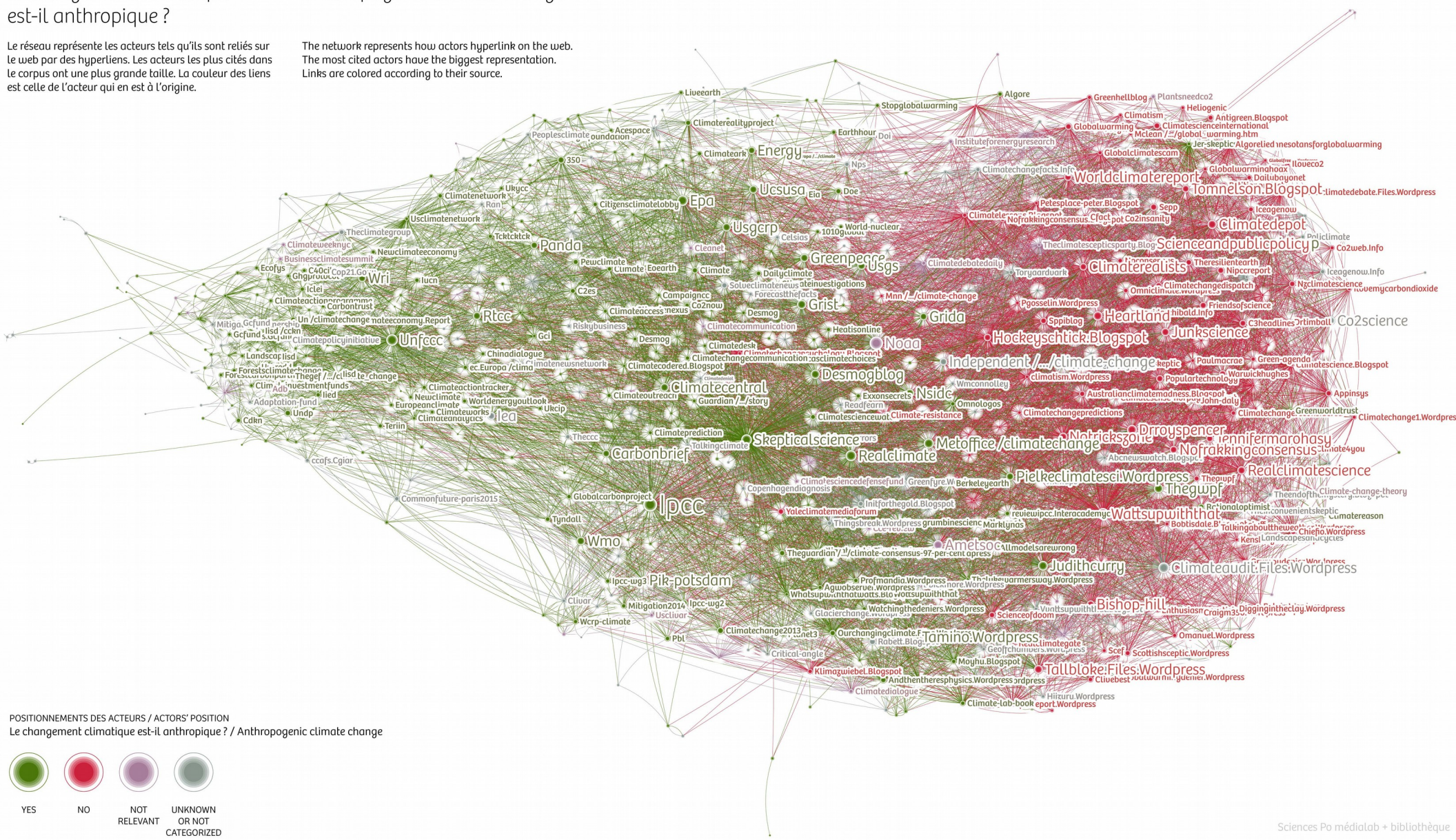
# Cartographier le web autour d'un thème/débat

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.





# 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/>

# Web ouvert ou fermé : quid de l'inaccessible ?

- Problème :

Comment étudier certaines parties du web moins visibles ?

web « profond », sites verrouillés derrière mot de passe,  
pages requérant un compte utilisateur autorisé, etc.

- Exemple :

Groupes Facebook non publics

- Possibilités :

Travail de terrain en « immersion web »

→ obtention d'un accès (demande / « implication directe »)

→ réutilisation des cookies du navigateur (par exemple avec minet)

# Pérennité du web : comment gérer l'éphémère ?

- Problème : comment exploiter les contenus web disparus ?  
domaines non renouvelés, contenus retirés, sites refondus ou non maintenus, messages supprimés, etc.  
→ nécessité d'accéder au passé (et donc de le conserver)
- Exemple Twitter :  
accès au passé via scraping (sans RTs) ou via l'accès dédié aux chercheurs à l'API v2, mais...  $\approx 20\%$  des tweets supprimés en un an  
→ privilégier la collecte live au plus tôt (par exemple avec gazouilloire)
- Exemple web :  
corpus Hyphe de liens hypertextes renvoyant vers des pages disparues  
→ exploitation des archives ([BnF](#), [INA DL-Web](#), [Internet Archive](#), etc.)  
→ connecter Hyphe à ces archives (WP4 ResPaDon)

# À vos questions !

---

<https://medialab.sciencespo.fr>

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

[@boogheta](#) [@medialab\\_ScPo](#)