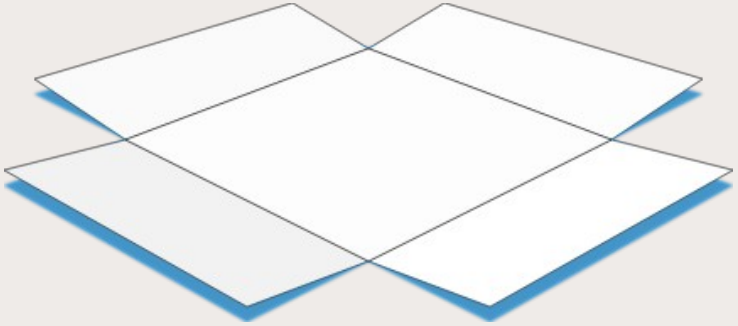
Utiliser Hyphe en sciences sociales

Benjamin Ooghe



Equipex DIME-SHS
ANR-10-EQPX-19-01

DIME Web

Des outils et méthodes numériques pour exploiter le Web comme terrain d'enquête

Utiliser Hyphe en sciences sociales

28 septembre 2018

Benjamin Ooghe-Tabanou, Sciences Po, médialab, Paris, France
DIME Web

SciencesPo
MÉDIALAB

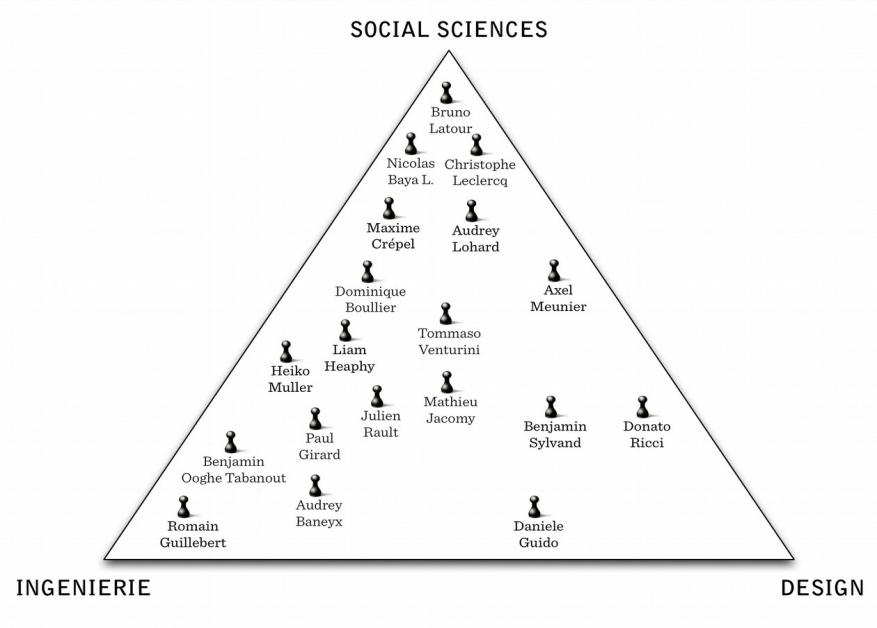
SciencesPo
DIME-SHS

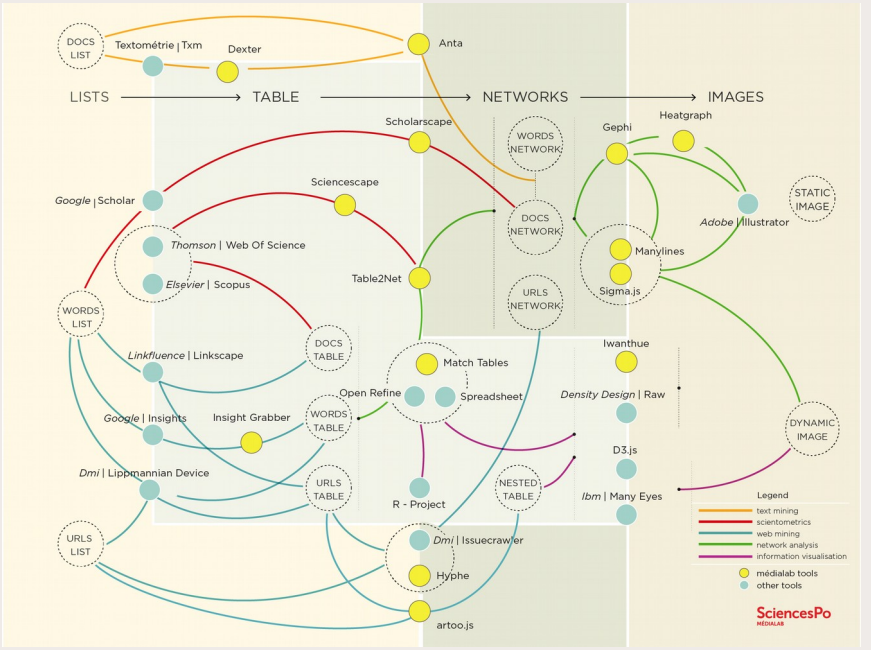
Le médialab de Sciences Po

https://medialab.sciencespo.fr

- Centre de recherche de Sciences Po, fondé par Bruno Latour en mai 2009, dirigé par Dominique Cardon depuis 2017
- Numérique, sciences sociales et design

  → Interdisciplinarité
- Articulation des méthodes quali & quanti
- Étude des traces numériques
- Un écosystème d'outils
  http://tools.medialab.sciences-po.fr
- Un atelier ouvert mensuel : le METAT
  http://www.medialab.sciences-po.fr/atelier/



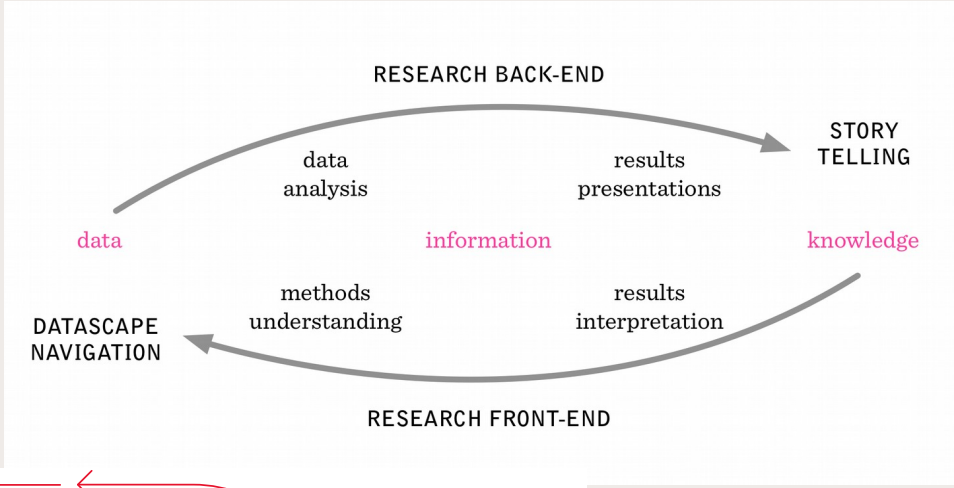


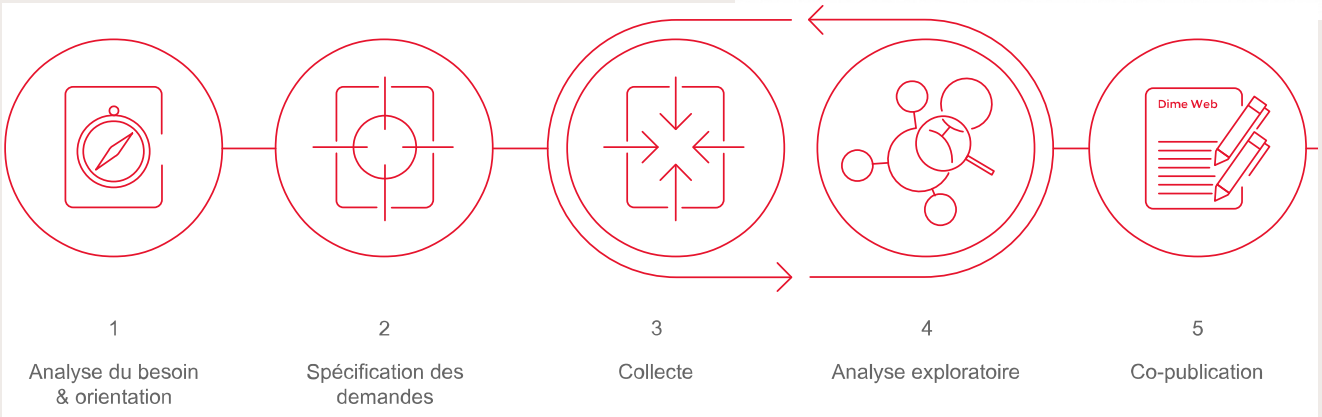
L'instrument DIME Web

http://dimeweb.dime-shs.sciences-po.fr

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
- Publication d'outils Open Source
- Méthodes numériques & itératives
  ≠ tout automatique





Le Web : une source de données « sales »

Collection de documents web (pages) sur une thématique

→ très grande hétérogénéïté (type de contenu & forme)

CRAWLING ≠ SCRAPING

(fouille systématique) (extraction ciblée)

contenus textuels & hyperliens → traitement automatique du langage

contenus textuels & hyperliens → analyse de réseau (effets de communauté)

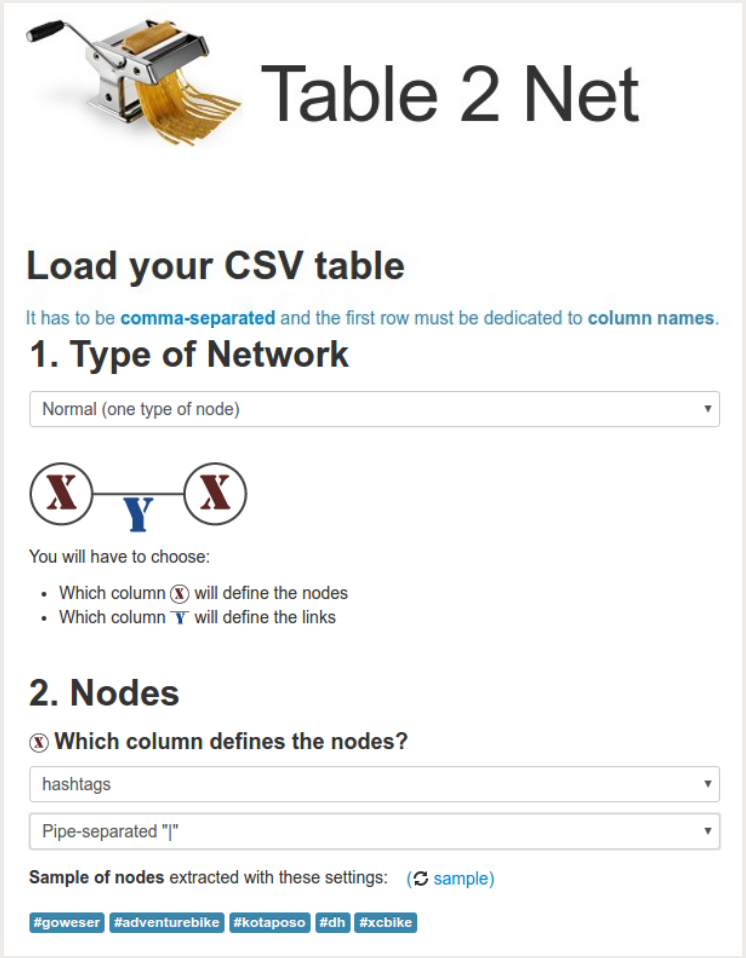
données structurées → méthodes quantitatives, statistiques…

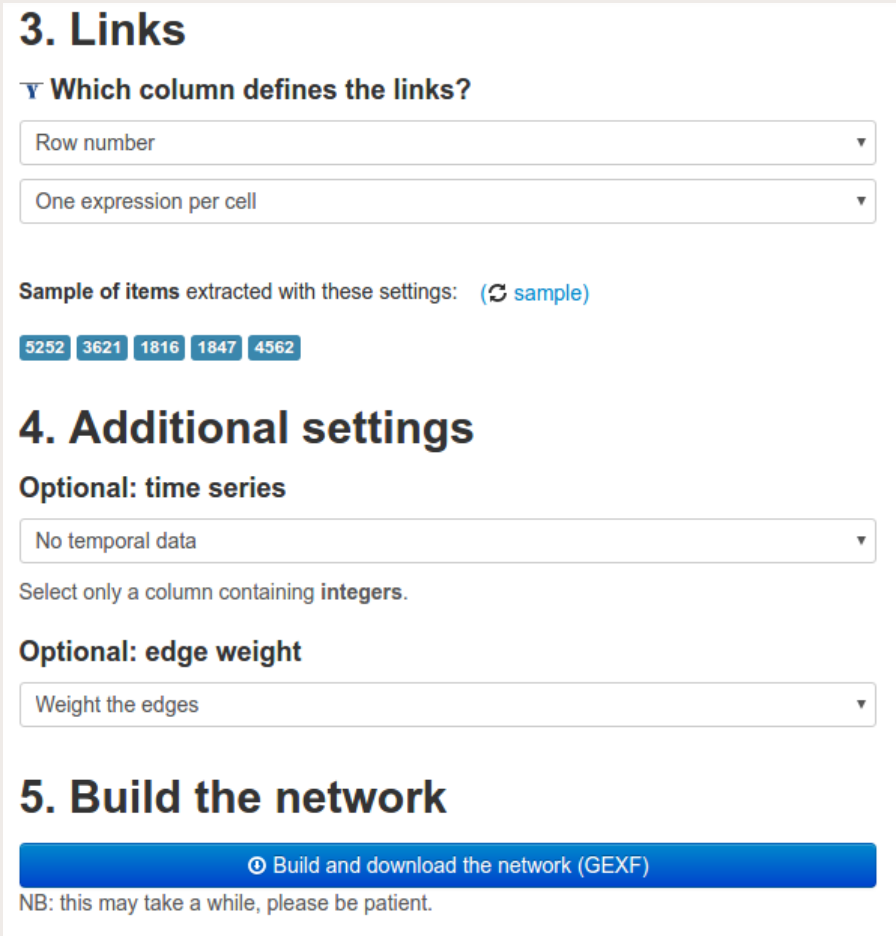
redirections, liens erronés, liens morts et sites disparus, encodage mal indiqué…

Table2Net : faire 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

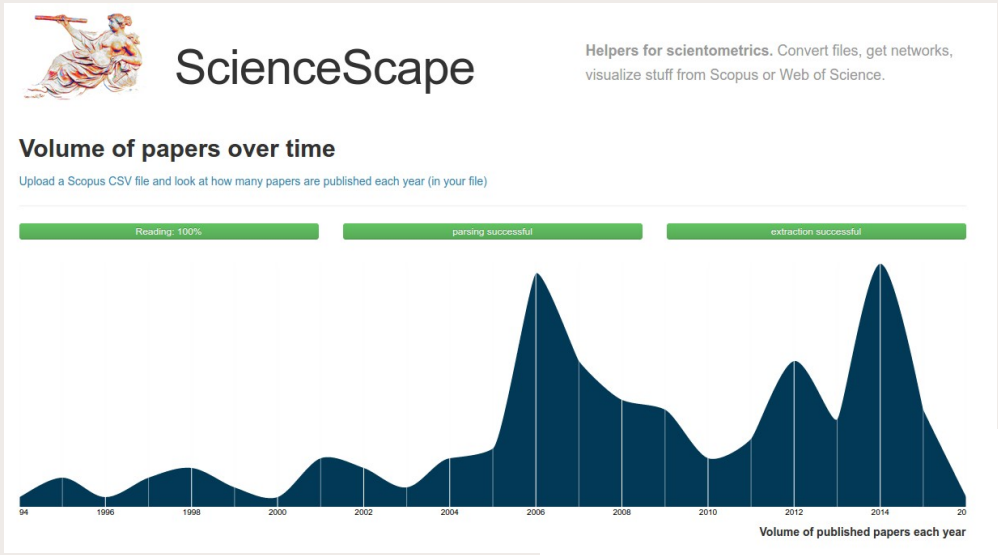


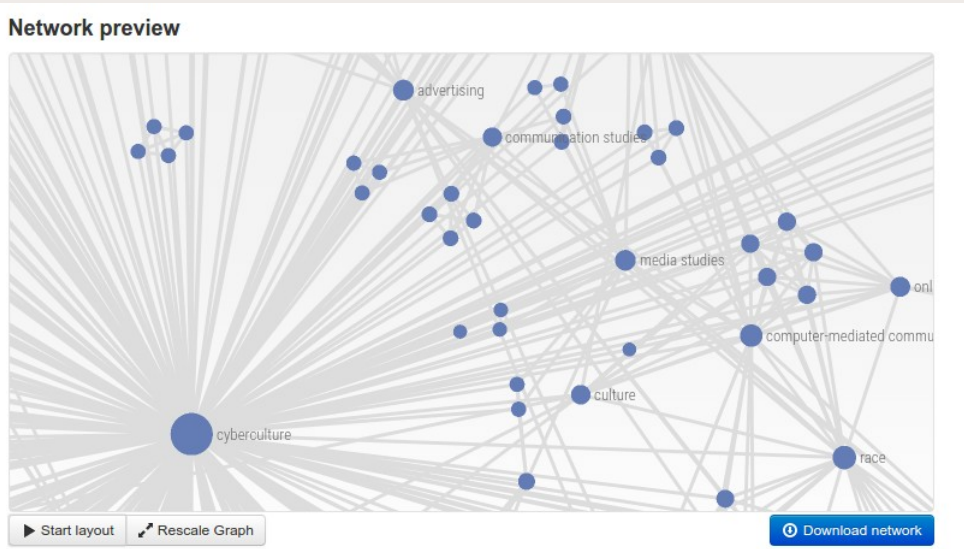


ScienceScape : scientométrie en quelques clics

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

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

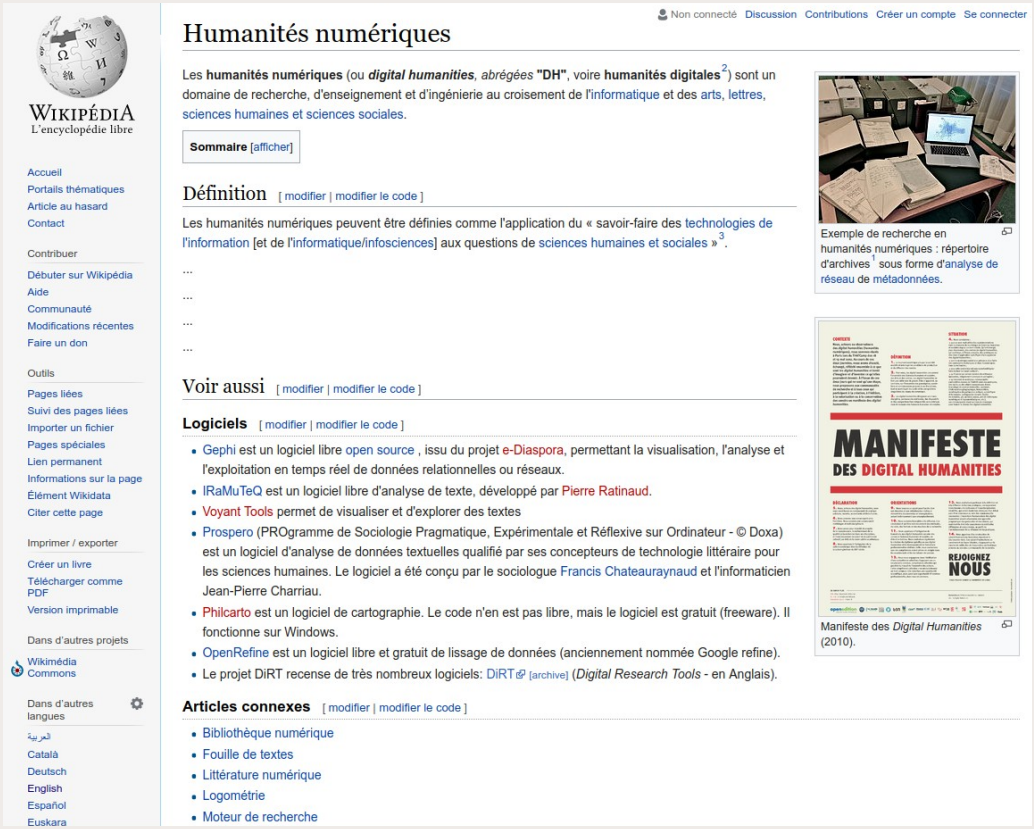


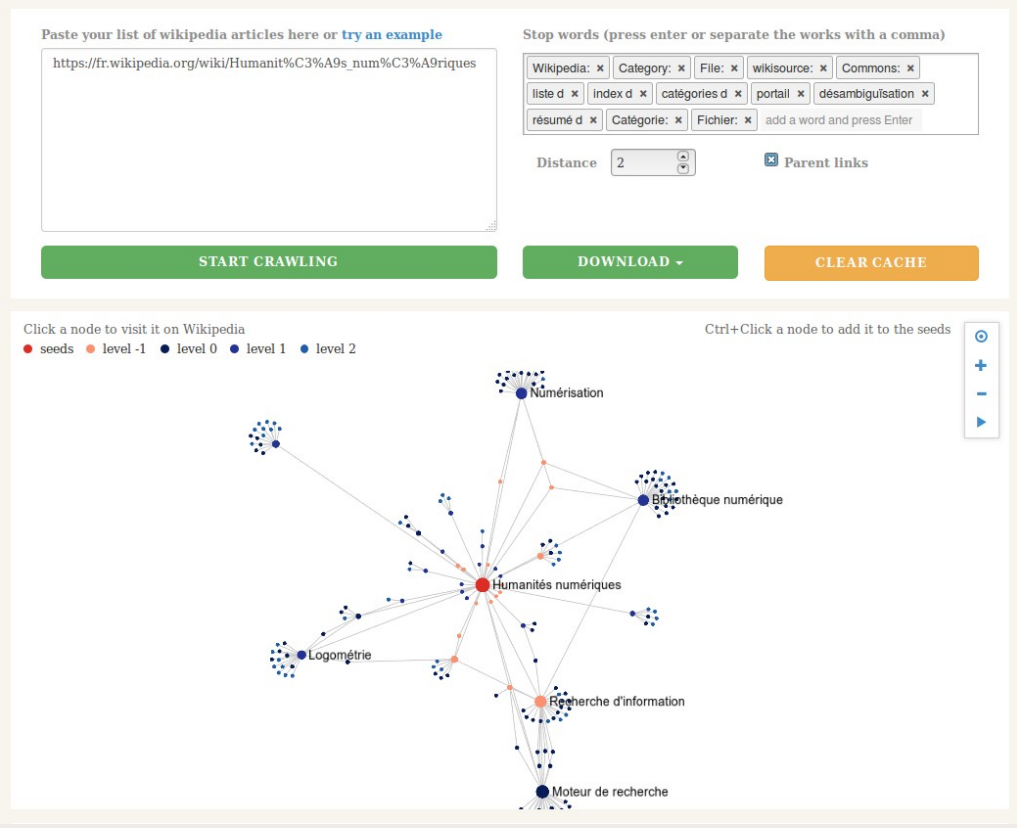


SeeAlsology : exploration sémantique rapide

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

- Explorer le réseau des liens présents dans les sections « Voir aussi », « Articles connexes » des pages Wikipedia





Gazouilloire : extraction systématique de tweets

https://github.com/medialab/gazouilloire

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

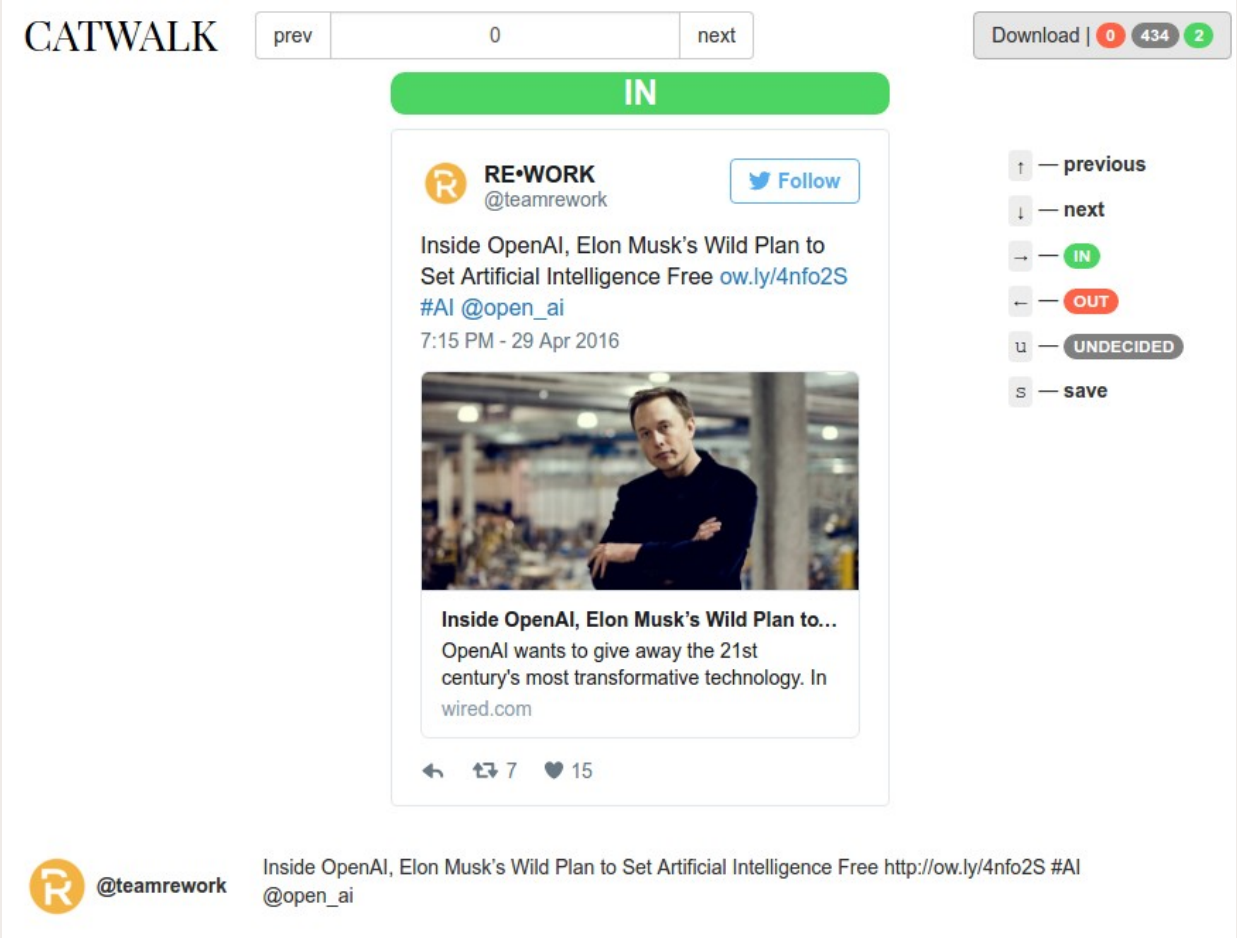
```
{
    "twitter": {
        "user": "Gazou_medialab2",
        "key": "[redacted]",
        "secret": "[redacted]",
        "oauth_token": "[redacted]",
        "oauth_secret": "[redacted]"
    },
    "mongo": {
        "host": "localhost",
        "port": 27017,
        "db": "tweets-naturpradi"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

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

CatWalk : sélection qualitative de tweets

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

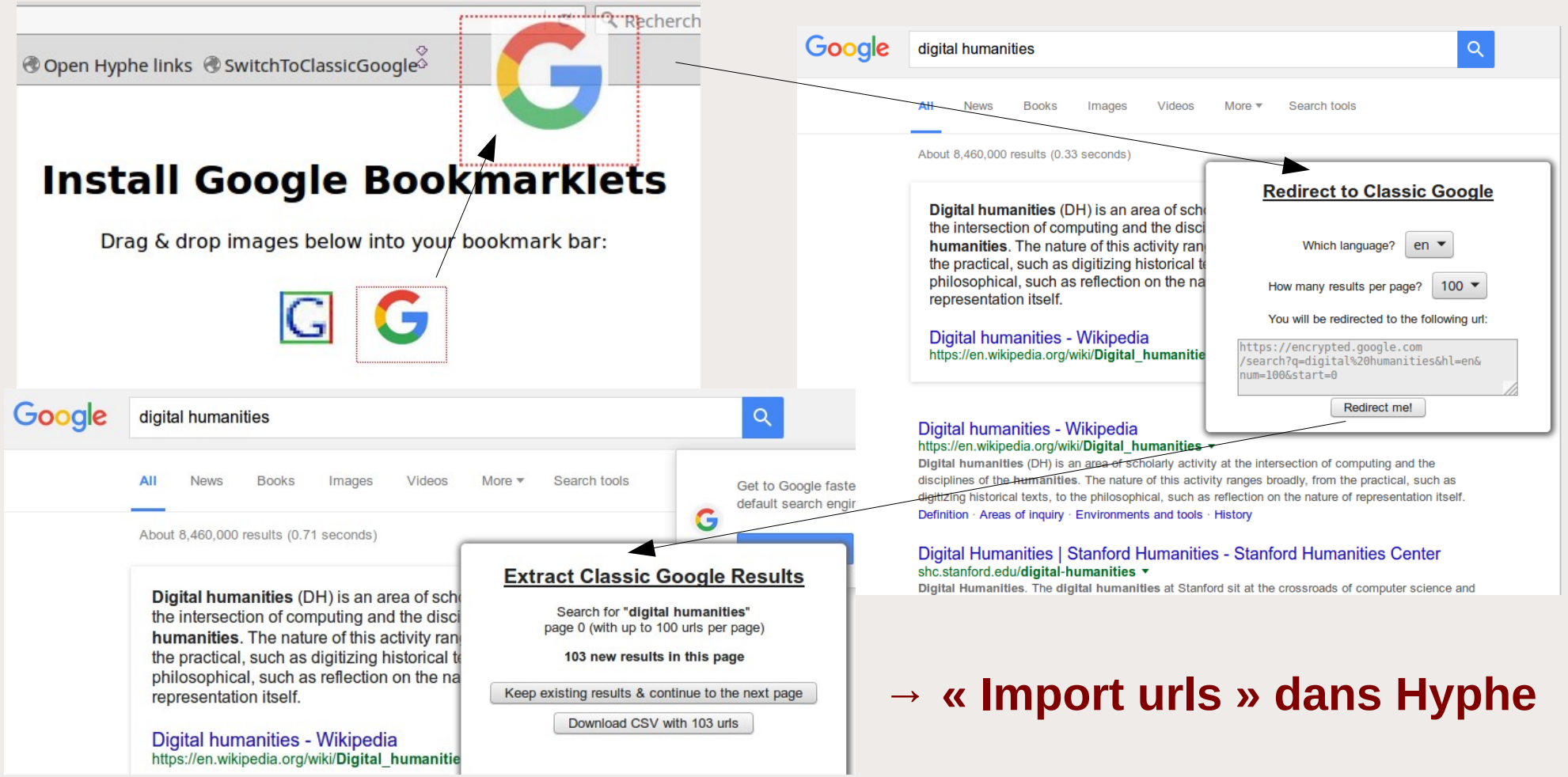
- Passer en revue rapidement « *à la Tinder* » tous les tweets d'un CSV pour décider de les inclure / exclure d'un corpus



Google bookmarklets : résultats Google en CSV

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

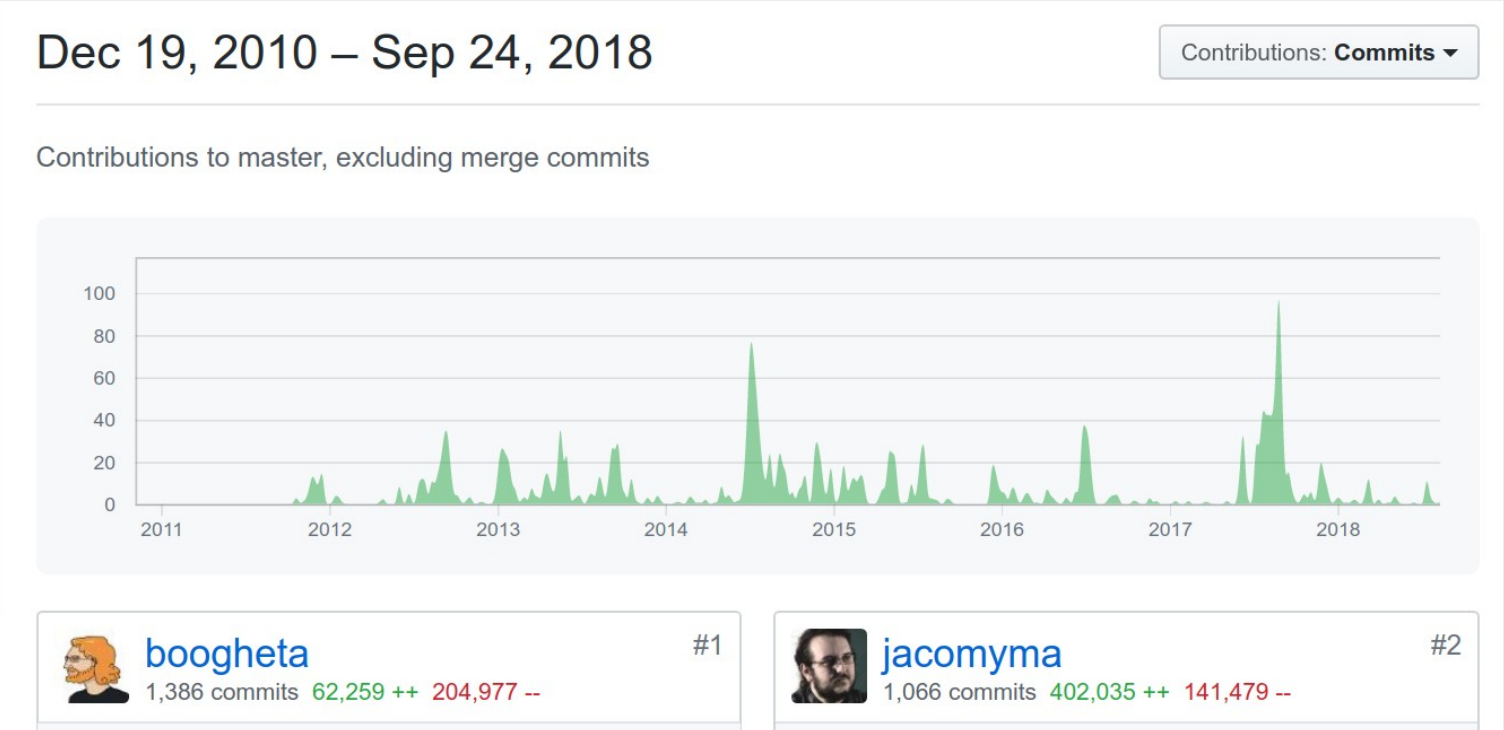
Des boutons dans vos favoris pour récupérer simplement au format tableur les résultats d'une recherche Google



→ « Import urls » dans Hyphe

Un projet logiciel phare : Hyphe

- 10 releases entre juillet 2013 (v.0.0.0) et août 2018 (v1.0.3)



- Plusieurs dizaines de présentations, sessions pratiques et formations assurées
- + de 50 déploiements externes à notre connaissance (France, UK, DK, USA...)
- + de 150 étudiants formés à l'analyse du web chaque année avec Hyphe
- + de 500 utilisateurs testant la démo chaque année

Hyphe : un crawler orienté par la recherche

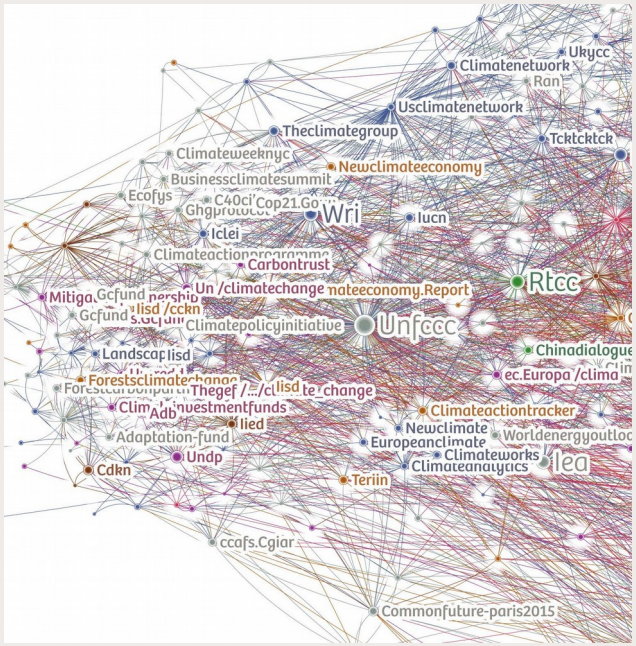
http://hyphe.medialab.sciences-po.fr/demo

- Les liens hypertextes : nouveaux révélateurs de relations entre acteurs d'une thématique
- Créer un corpus documentaire

  → « acteurs web » & contenus textuels respectifs

  → liens hypertextes entre ces acteurs
- Études exploratoires ou de controverses dans tous les domaines

hyphe



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

COP 21

Vie privée

Extrême droite

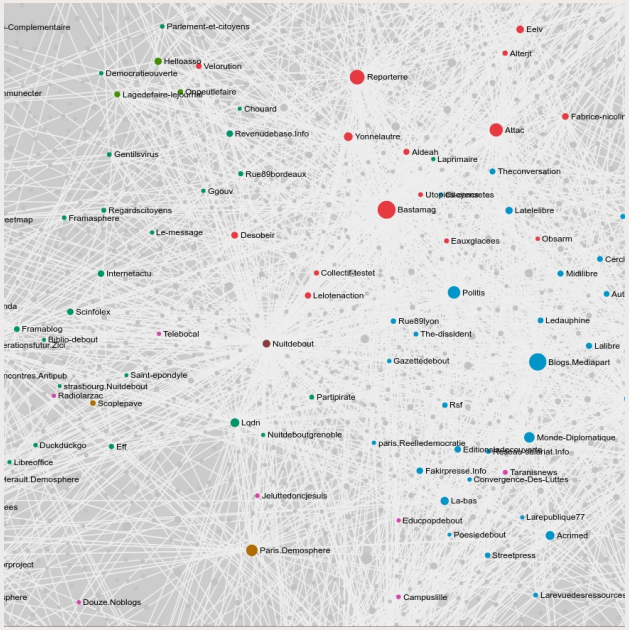
Tissu associatif

Produits laitiers

Cellules souches

Administrations culturelles

…

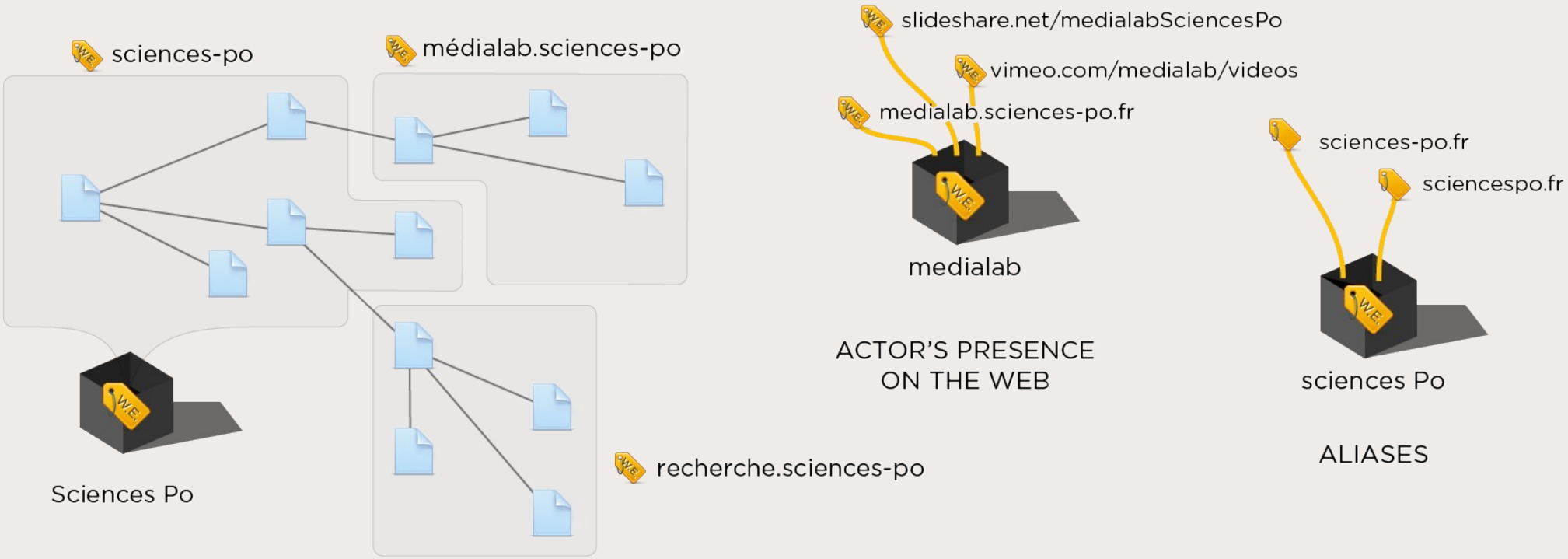


http://utopies-concretes.org/

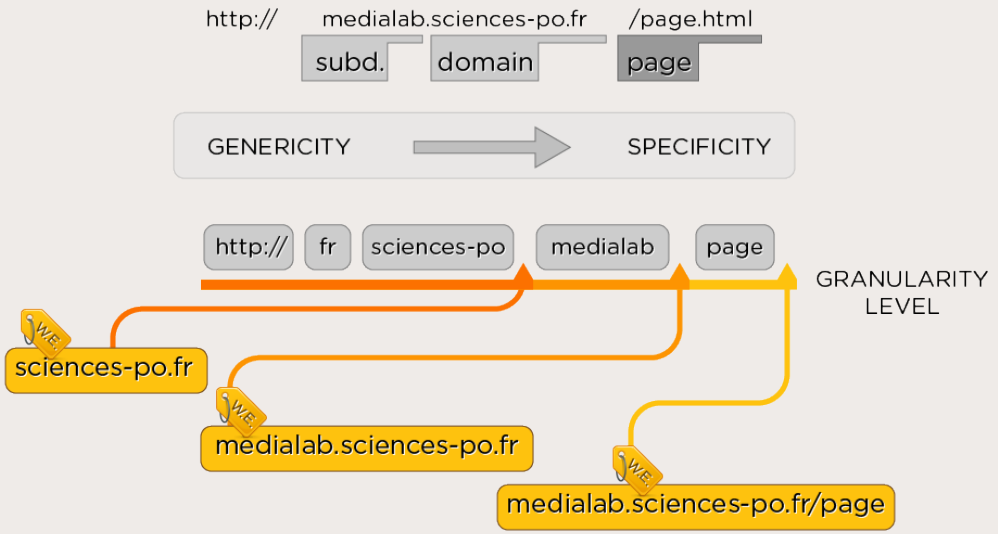
Principes méthodologiques : « WebEntités »

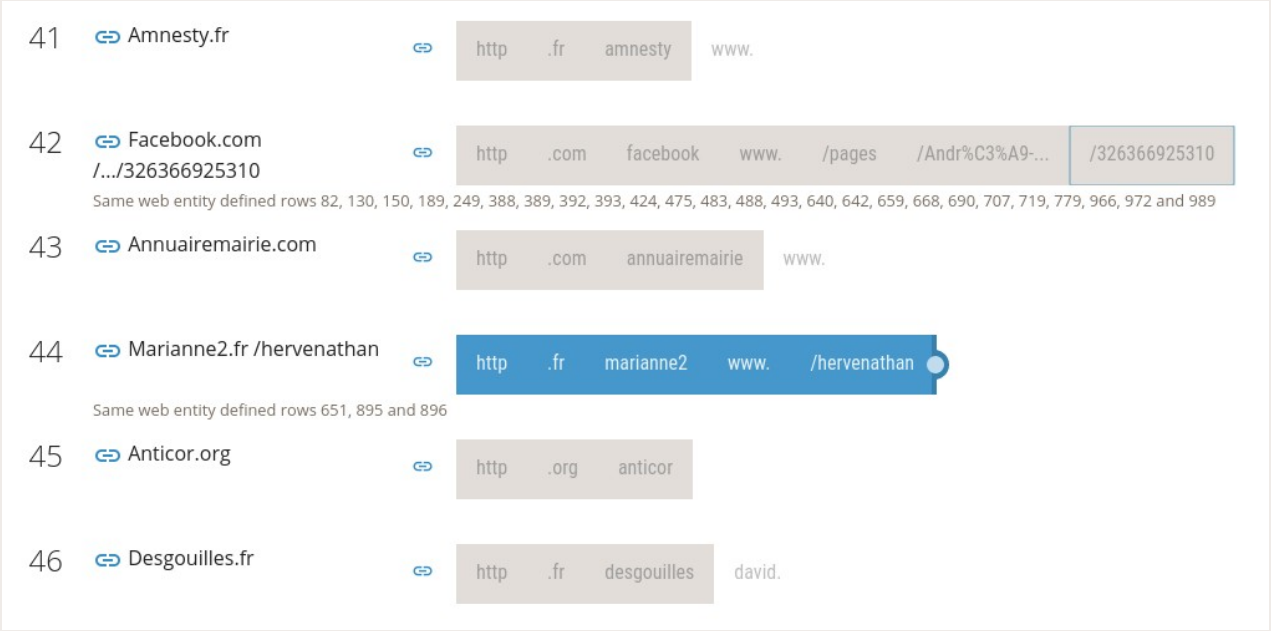
Comment gérer la diversité de granularité des sites web ?

→ « WebEntités » : agrégats reflétant des entités documentaires cohérentes du point de vue du chercheur



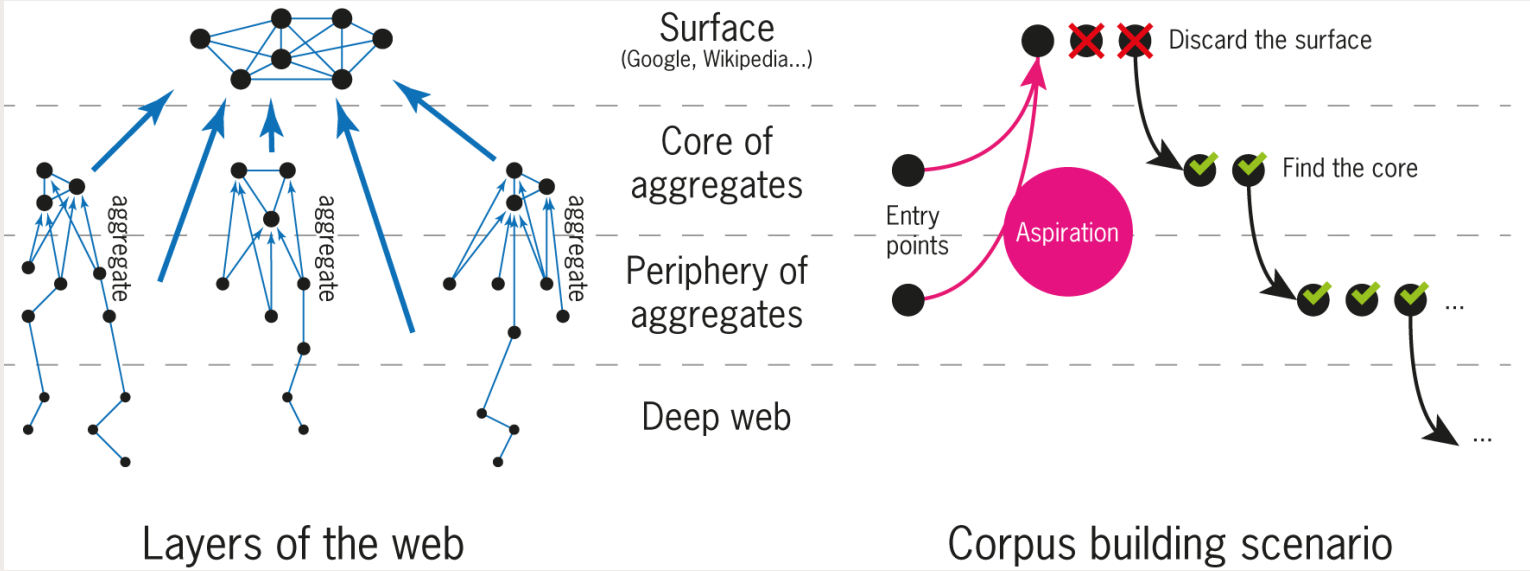
Principes méthodologiques : « WebEntités »





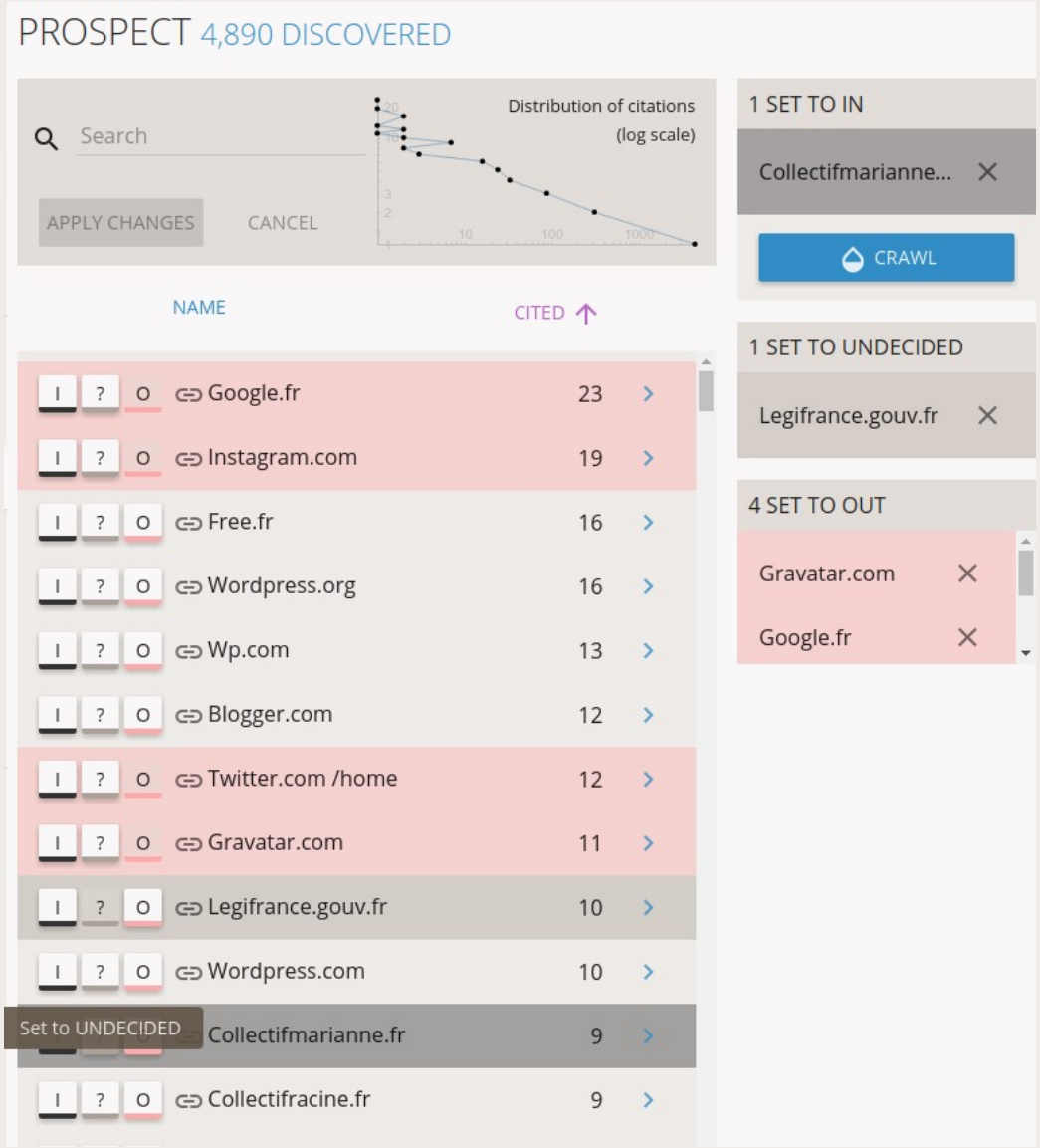
Principes méthodologiques : « Prospection »

- Démarrage : points d'entrées libres (recherche web qualitative, GoogleBookmarklets, annuaire, liste d'acteurs issue d'entretiens…)
- Crawler = robot qui fouille les pages web et clique sur les liens
  - Crawlers classiques : boule de neige (fouille systématique jusque N clics)
    → bruit de la couche haute du web (Google, YouTube, Wikipedia…)
  - Hyphe : crawl ciblé, uniquement les pages internes des WebEntités choisies
    → éditorialisation et contrôle de la construction thématique



Principes méthodologiques : « Prospection »

- Exploitation de la nature hypertextuelle du web
- Identification des acteurs web liés potentiellement pertinents
- Travail de terrain (numérique)
  → exclure ou inclure
- Décisions éditoriales classiques de type gestion documentaire

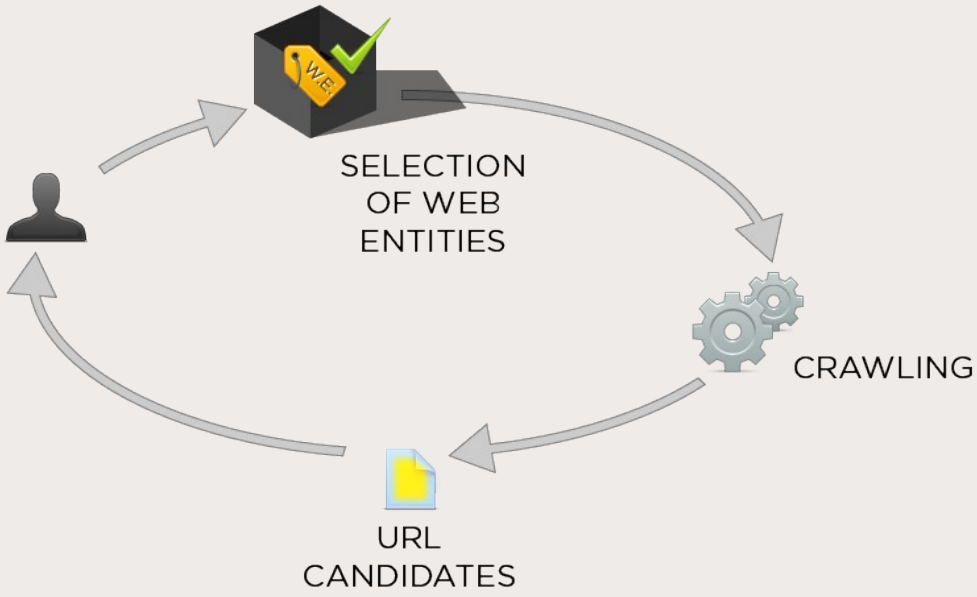


Principes méthodologiques : « Itérations quali »

- Expansion éditorialisée et itérative du corpus
- Coût en temps humain : travail de curation itératif
  « crawler orienté par la recherche »
- La liste des WebEntités découvertes s'allonge exponentiellement

  → Quand s'arrêter ?

  → Seuil de citation



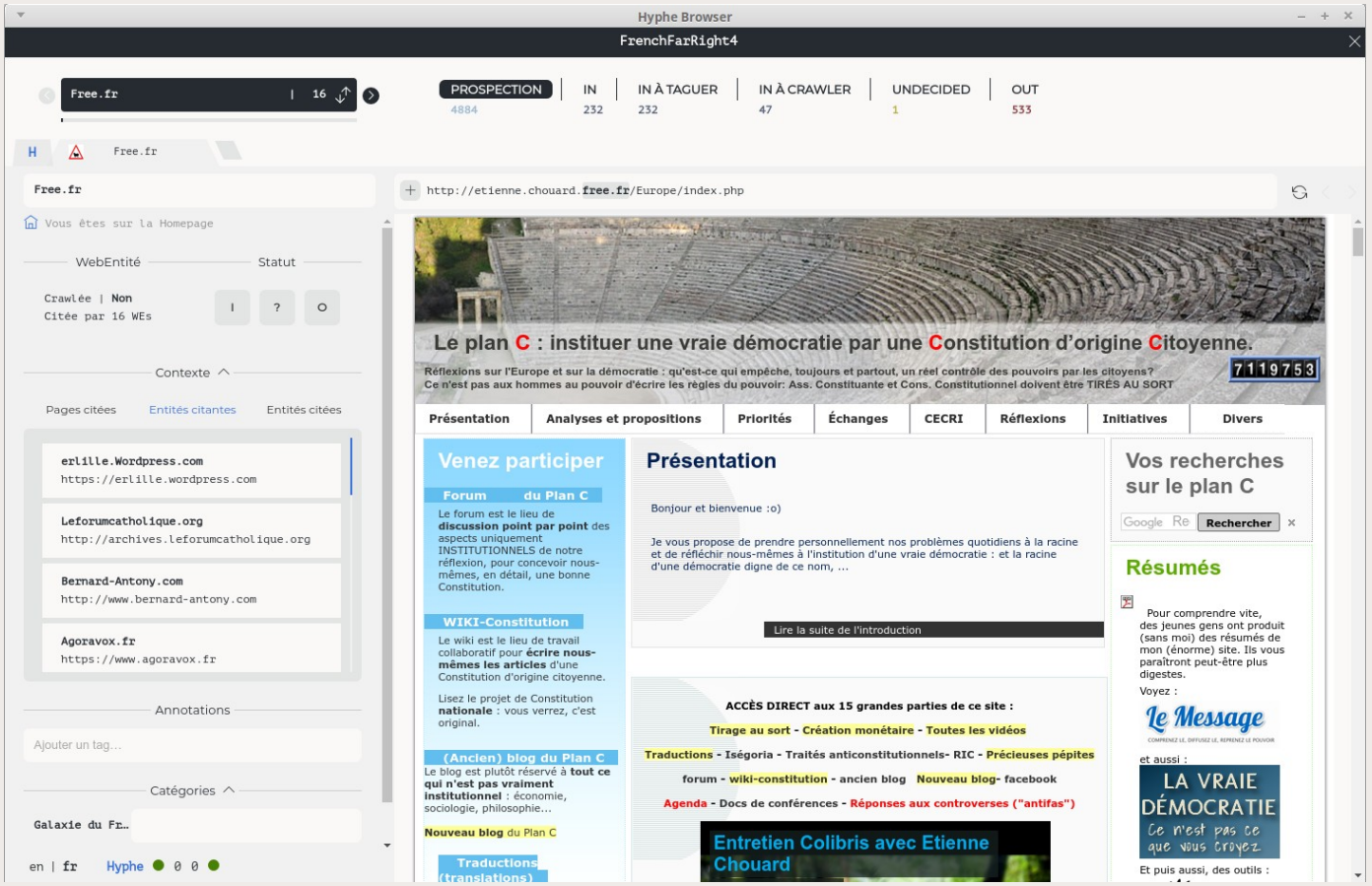
HyBro : un browser dédié à la curation de corpus

https://github.com/medialab/hyphe-browser/releases/

Héritier du « NaviCrawler » : un navigateur web connecté à Hyphe :

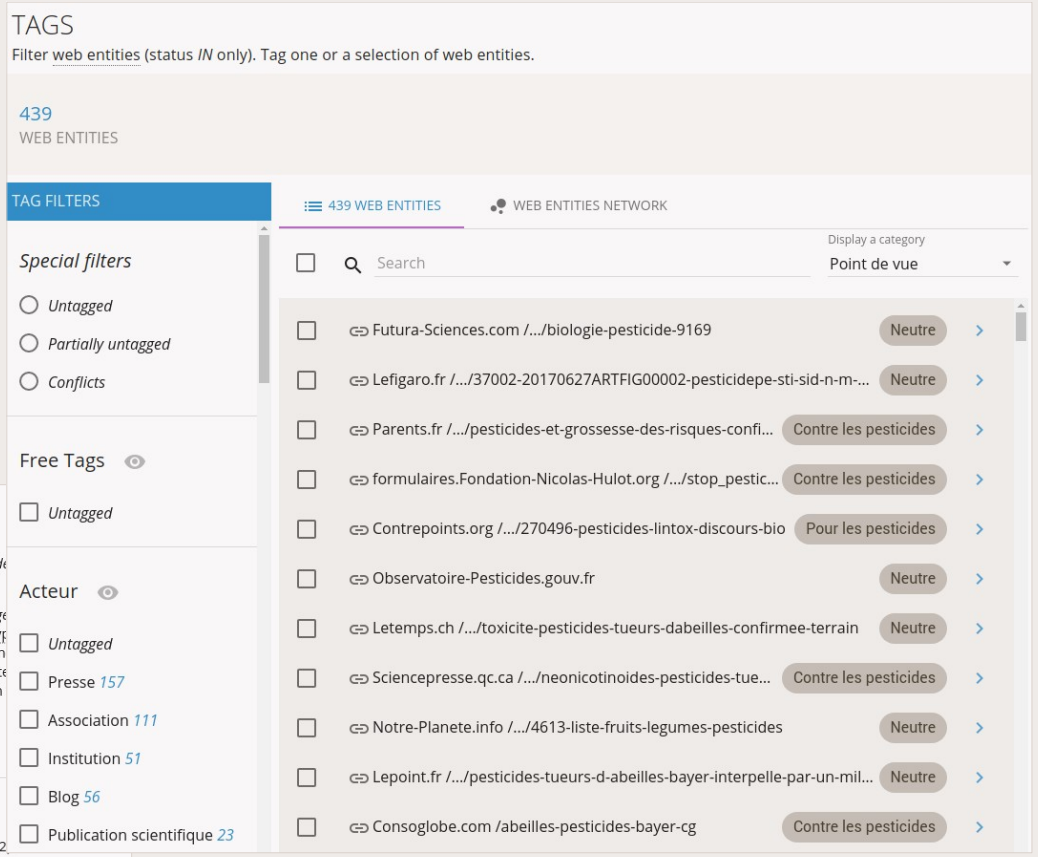
→ prospection et catégorisation in-situ (terrain numérique)

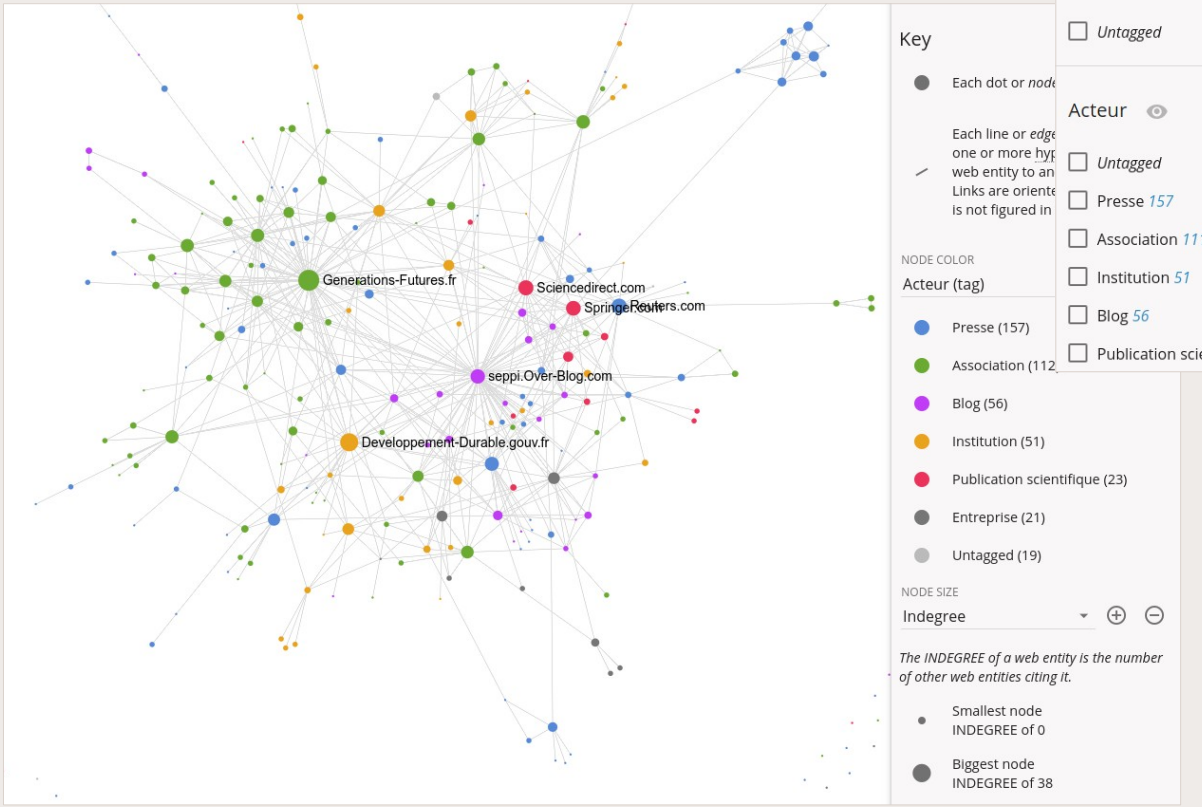
→ formation d'étudiants et lycéens au web (IDEFI FORCCAST)



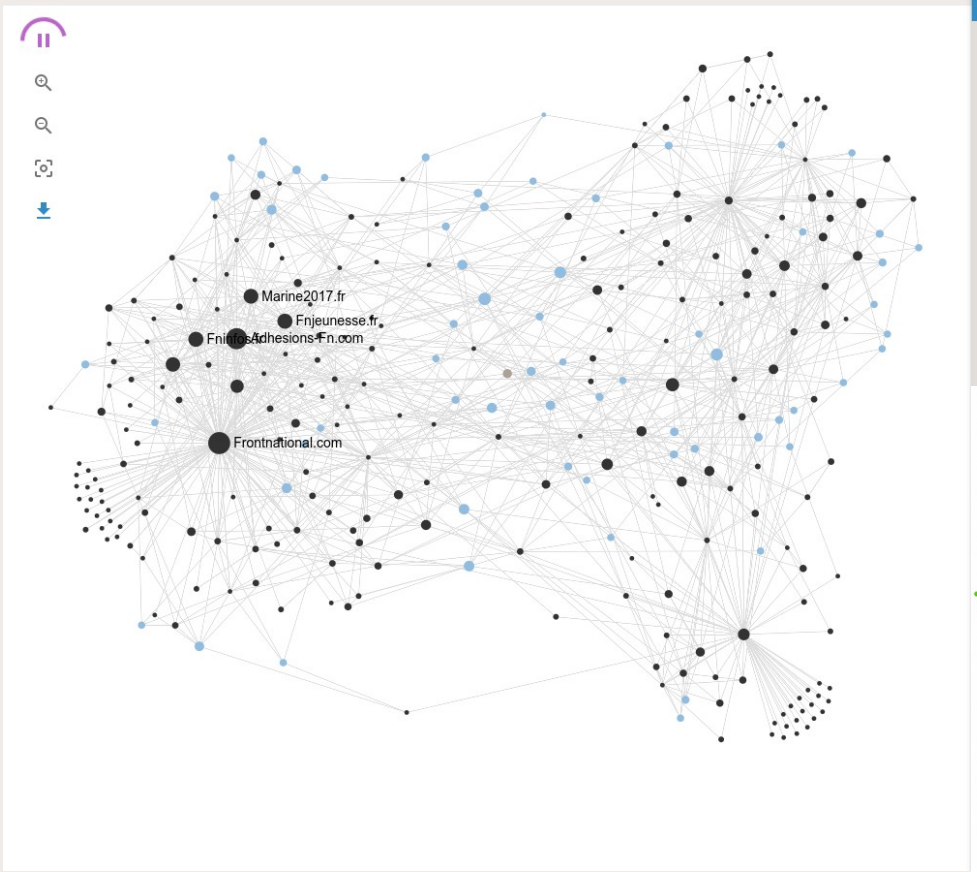
Gérer ses catégorisations (tagging)

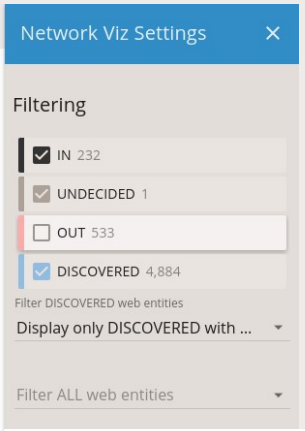
- Annotations libres
- Catégories

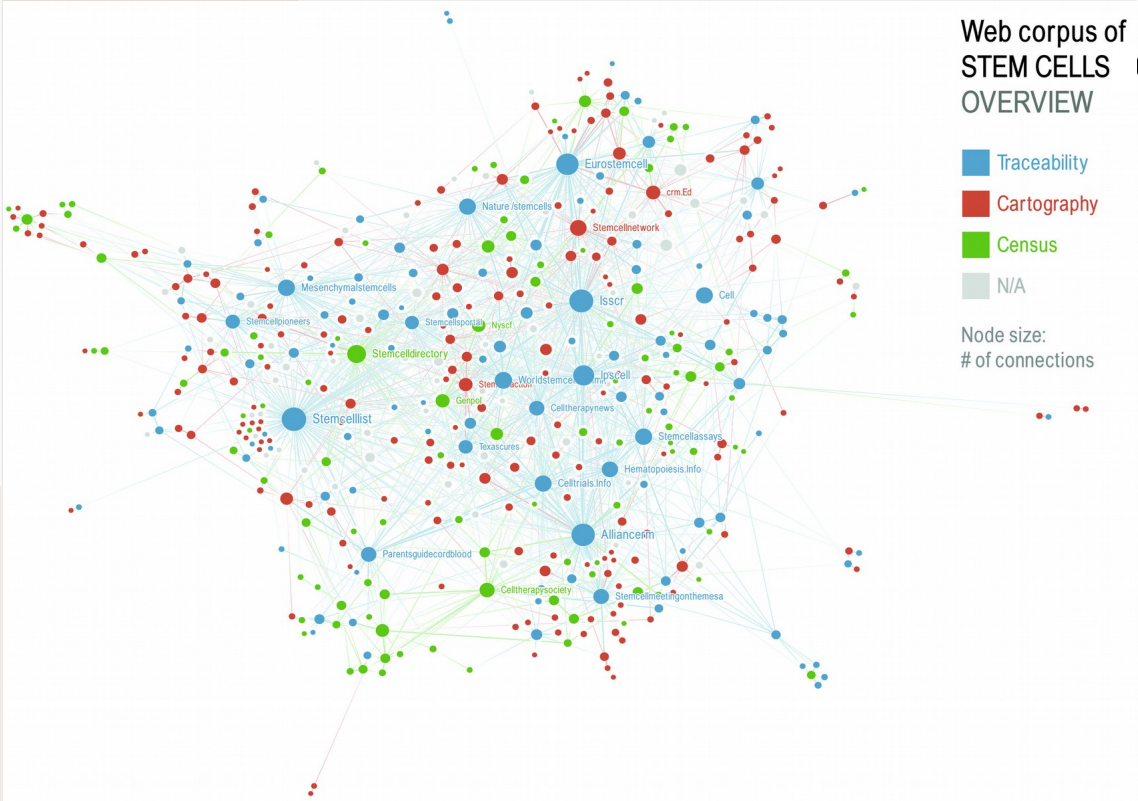




Explorer le réseau des liens entre acteurs

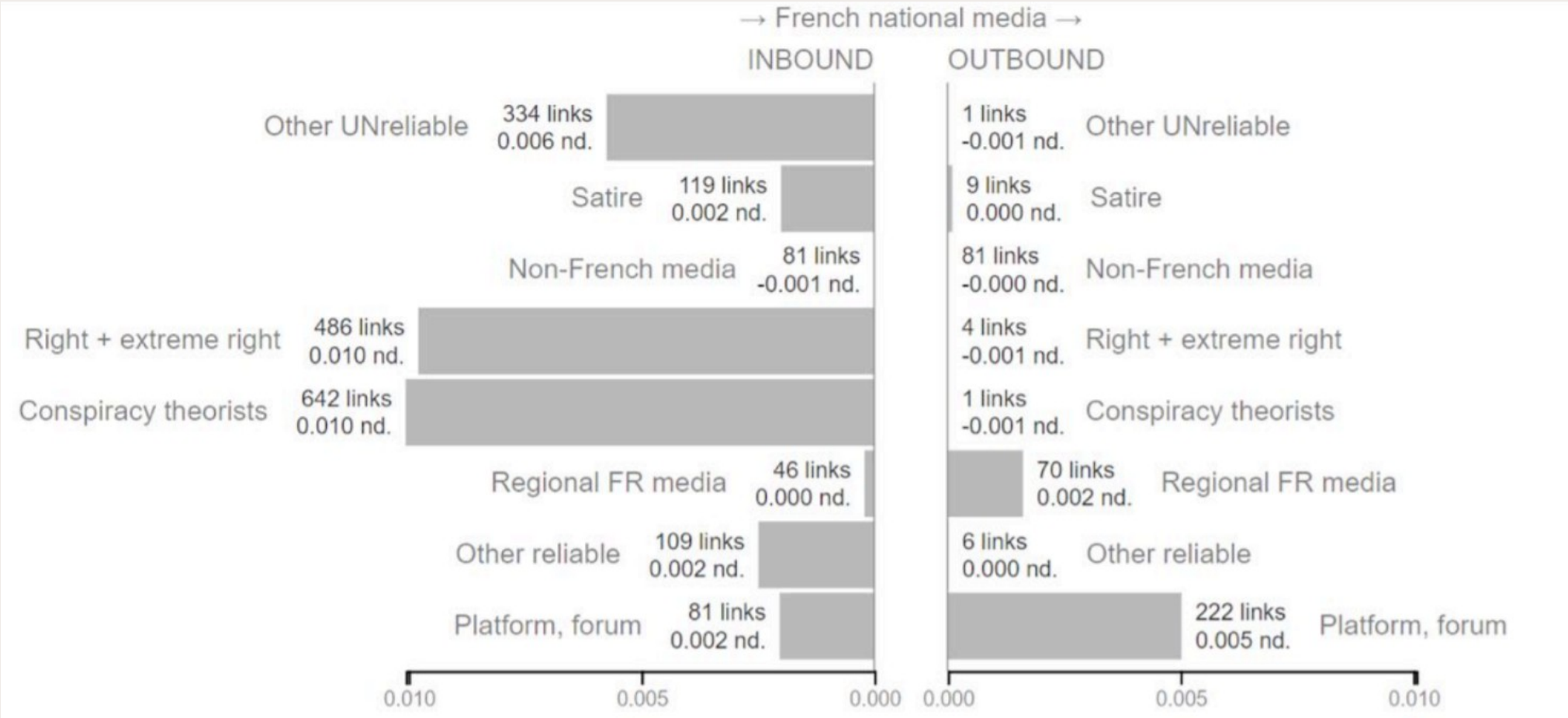




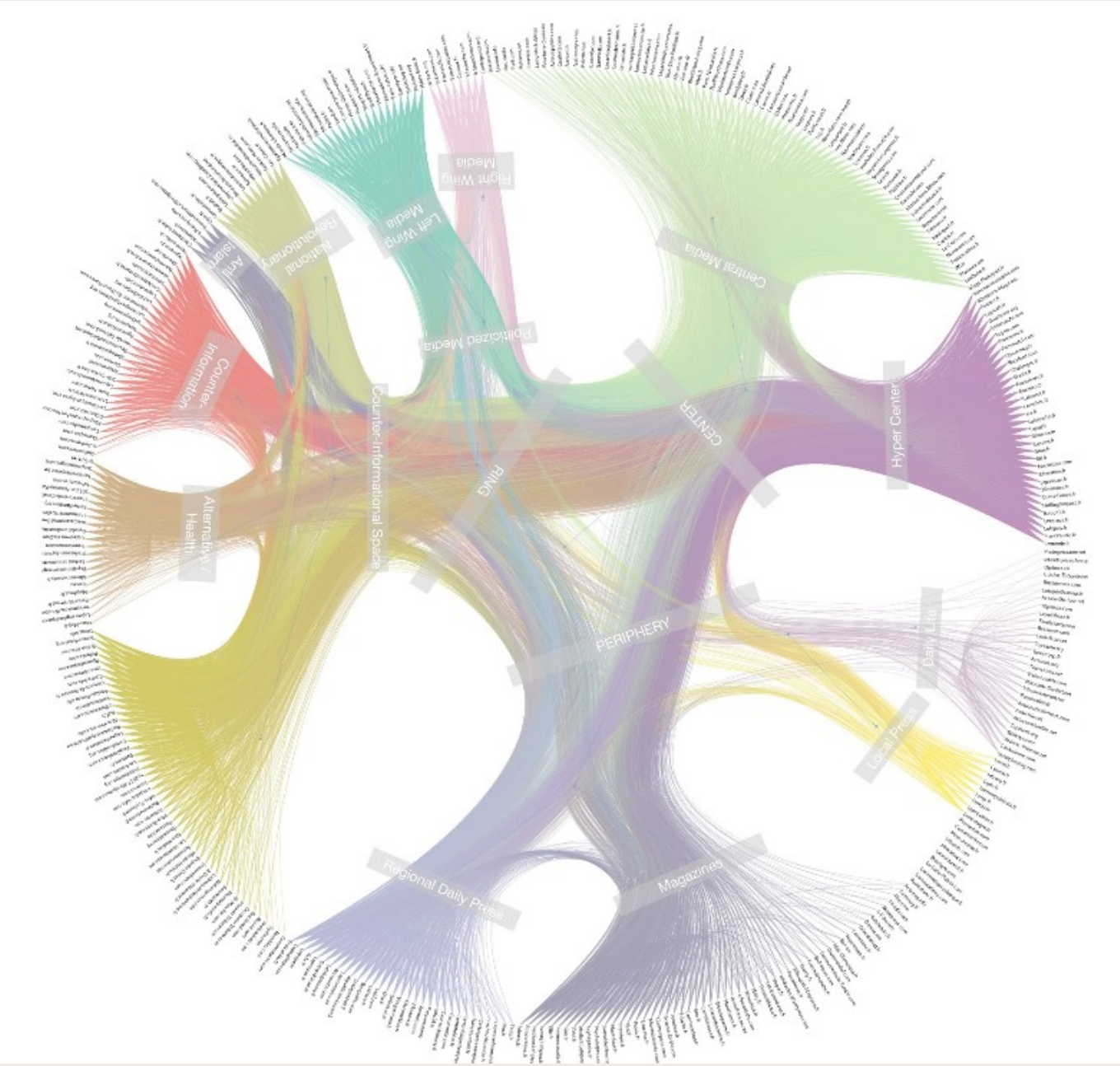


Social Representations of Stem Cells, Virginie Tournay, CEVIPOF, 2016

Analyser la structuration interne des corpus



Explorer les dynamiques de polarisation



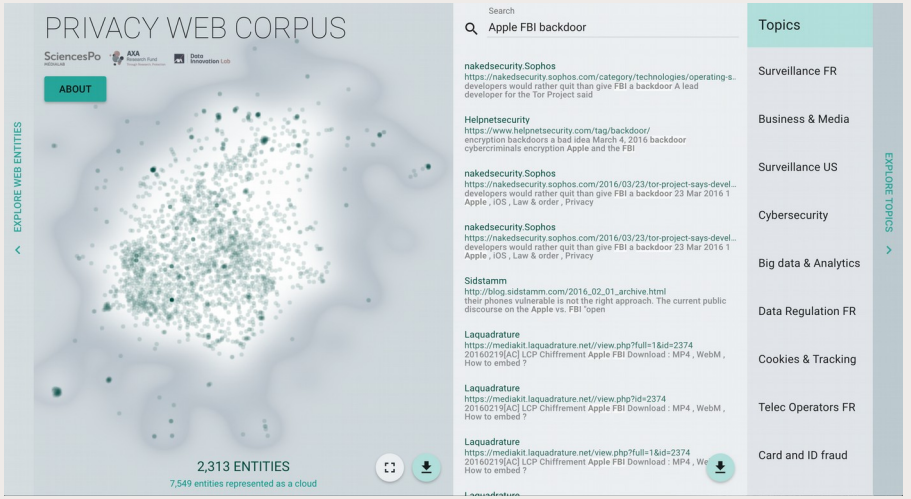
Et pour la suite ?

- Import / export de listes de webentités et crawls ou de corpus :

  → duplication, reproduction

  → exploration longitudinale dans le temps

- Exploitation intégrée des contenus textuels issus des pages crawlées et analyse automatique du langage



- Utiliser les technologies modernes pour crawler les sites en JavaScript (Facebook, etc.)

- Outils de contrôle qualité des crawls et du corpus

- Outils d'archivage et présentation des corpus finalisés

- Mise à disposition à la demande automatisée (SAAS)