Différents modes de collecte de données web

2 approches bien distinctes aux cibles et résultats différents

CRAWLING	≠	SCRAPING
sources multiples (pages web hétérogènes)		source unique (page seule ou ensemble cohérent)
fouille systématique		extraction ciblée
contenus textuels & hyperliens		données structurées

traitement du langage

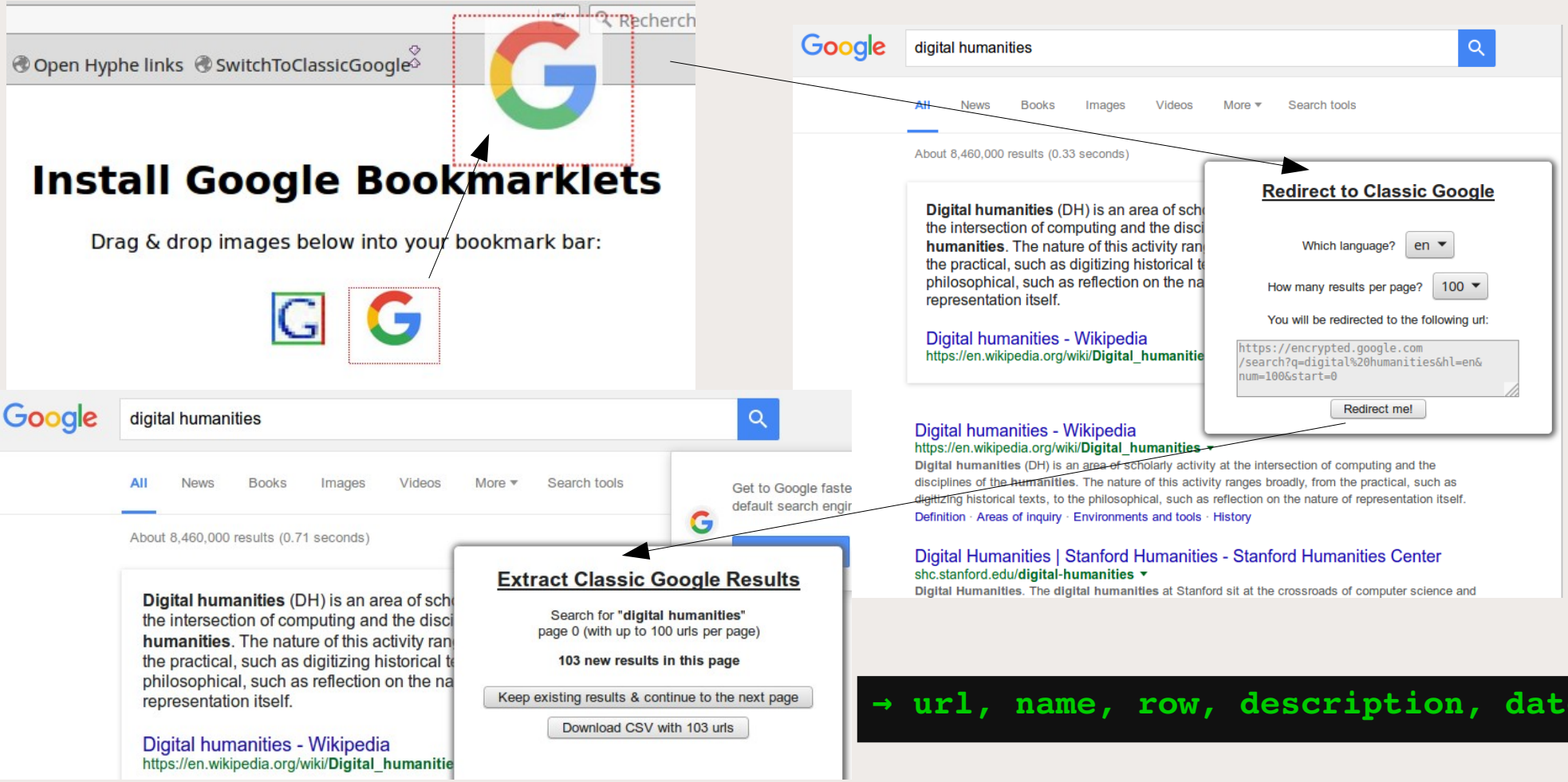
analyse de réseau (effets de communauté)

méthodes quantitatives, statistiques…

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



→ url, name, row, description, date

Le crawling : pour quoi faire ?

L' « **Hyperlien** » au cœur de la conception du Web
→ porteur de sens et de structure

*« The texts are **linked together** in a way that one can go from one concept to another to find the information one wants.*

*The network of links is called a **web**. […]*

*The texts are known as **nodes**.*

*The process of proceeding from node to node is called **navigation**. »*

Tim Berners-Lee, 1990, WorldWideWeb: Proposal for a HyperText Project

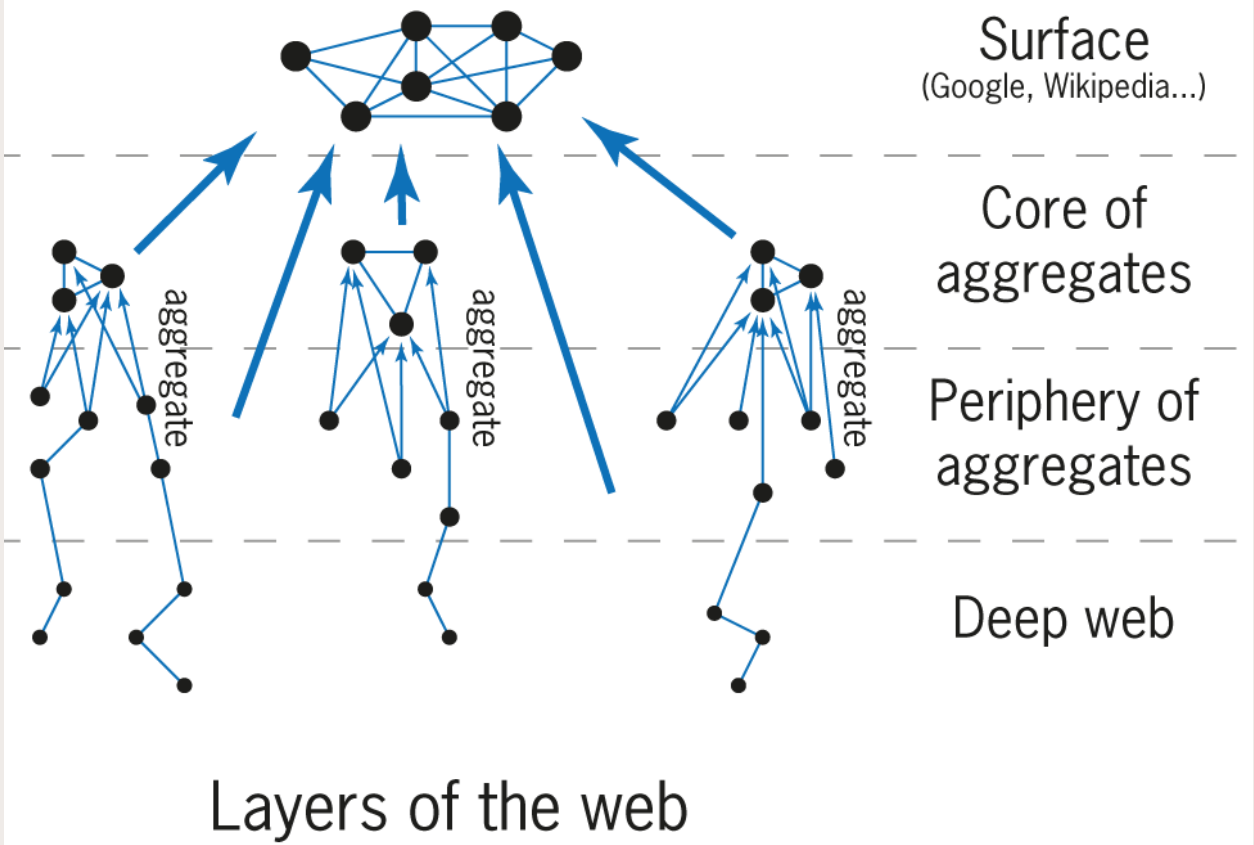
*« A hyperlink is a **manifestation of intention**. **By linking** one page to another, one piece of text to another, **people intend** to do particular things. »*

Ryfe, Mensing, & Kelley, 2016, What is the meaning of a news link?

Une hiérarchie « bottom-up » émergée des liens

Effet Matthew :

→ les nouvelles pages ont tendance à citer les pages préexistantes déjà les plus citées



SciencesPo
MÉDIALAB

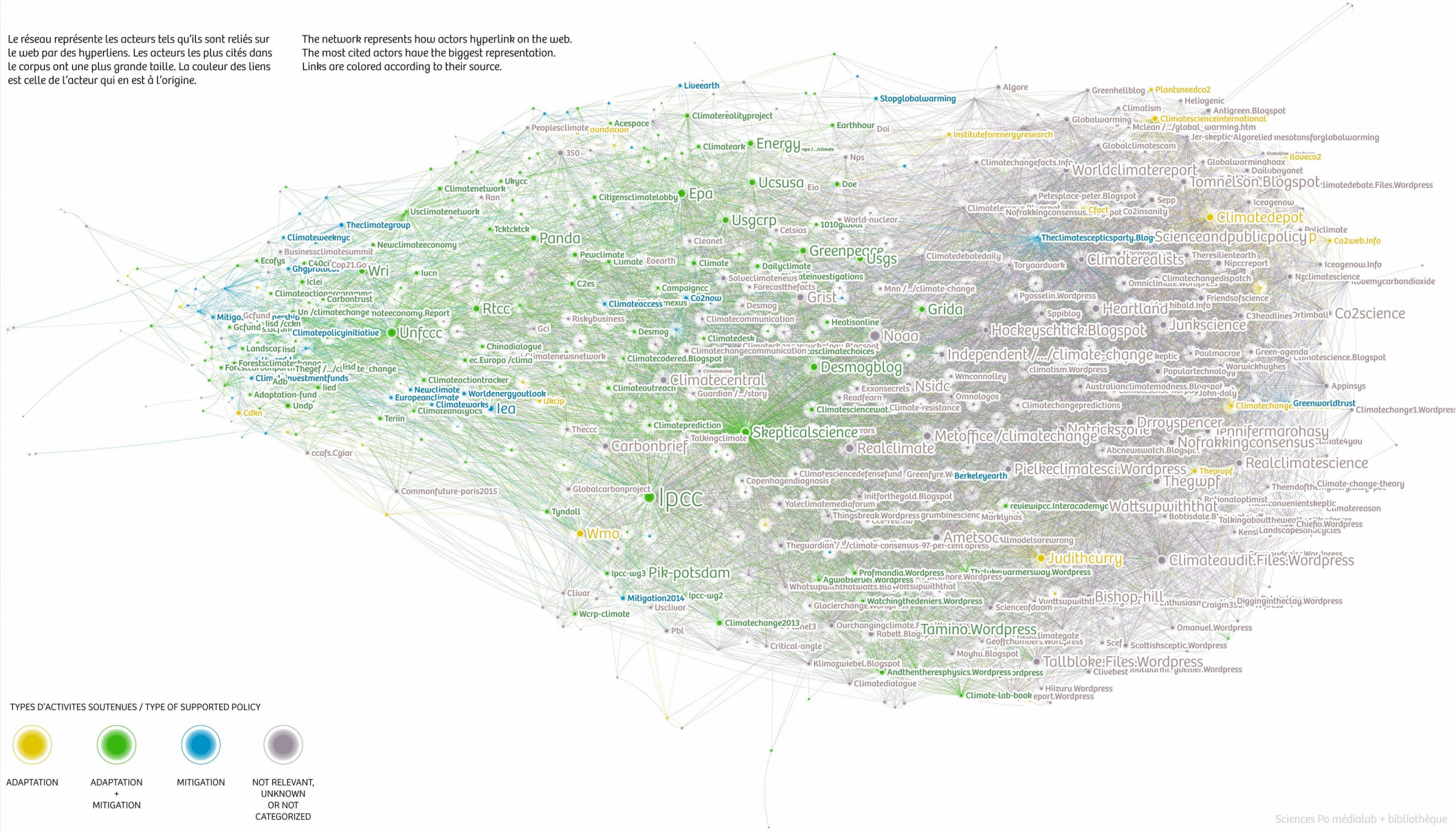
Cartographier le web par typologie d'acteurs

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.



TYPES D'ACTIVITES SOUTENUES / TYPE OF SUPPORTED POLICY

ADAPTATION

ADAPTATION + MITIGATION

MITIGATION

NOT RELEVANT, UNKNOWN OR NOT CATEGORIZED

SciencesPo
MÉDIALAB

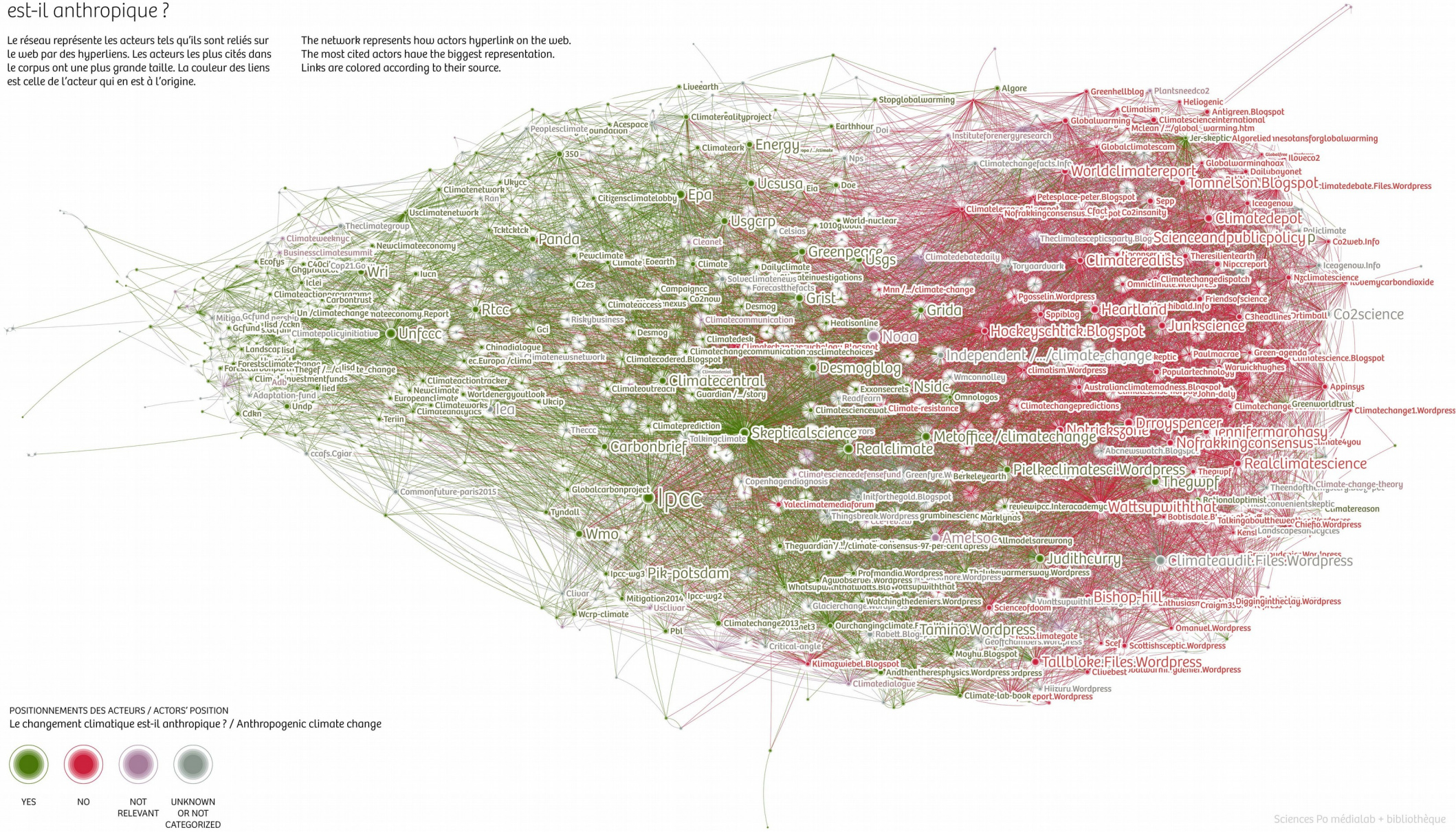
Cartographier le web autour d'un thème/débat

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

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

Web corpus on climate change: anthropogenic climate change

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

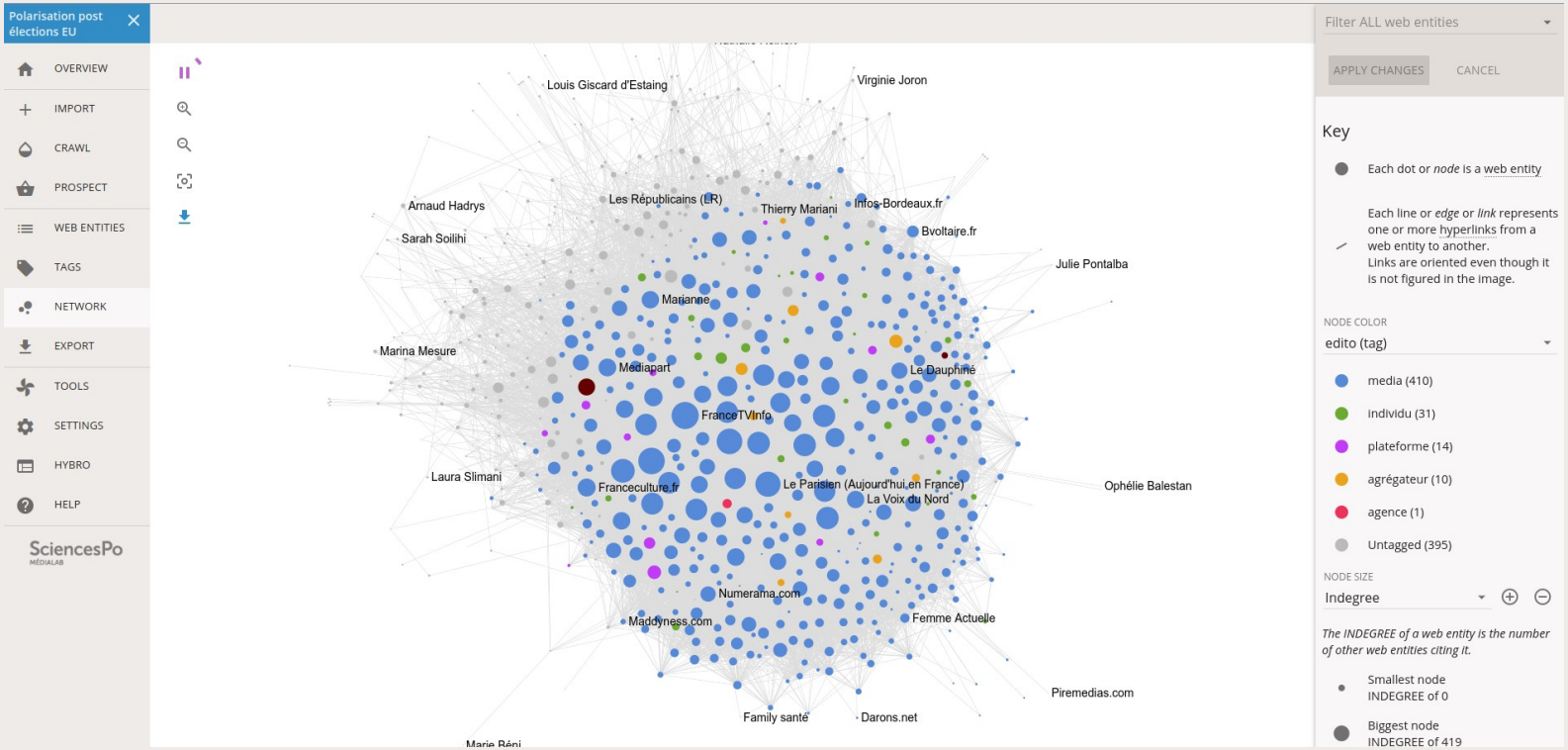


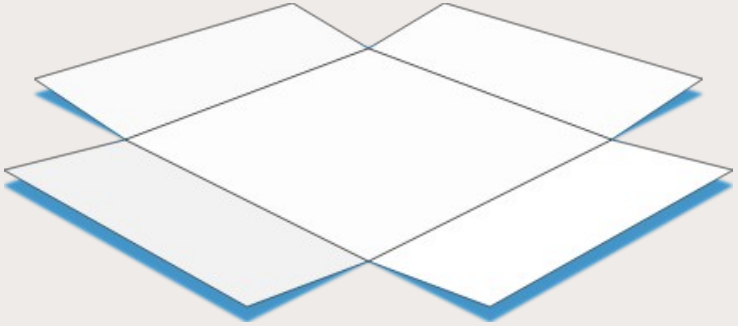
POSITIONNEMENTS DES ACTEURS / ACTORS' POSITION
Le changement climatique est-il anthropique ? / Anthropogenic climate change

YES | NO | NOT RELEVANT | UNKNOWN OR NOT CATEGORIZED

Différents crawlers pour les Sciences Sociales

- IssueCrawler (DMI – UVA)
- SocSciBot (Statistical Cybermetrics Research Group)
- Navicrawler (WebAtlas) *(obsolète)*
- Hyphe + Hyphe-Browser (médialab – Sciences Po)





Fouiller son terrain en explorant le Web avec Hyphe

Explo SHS
La Rochelle – 15 octobre 2020

Benjamin Ooghe-Tabanou
Sciences Po médialab – DIME SHS Web

SciencesPo MÉDIALAB

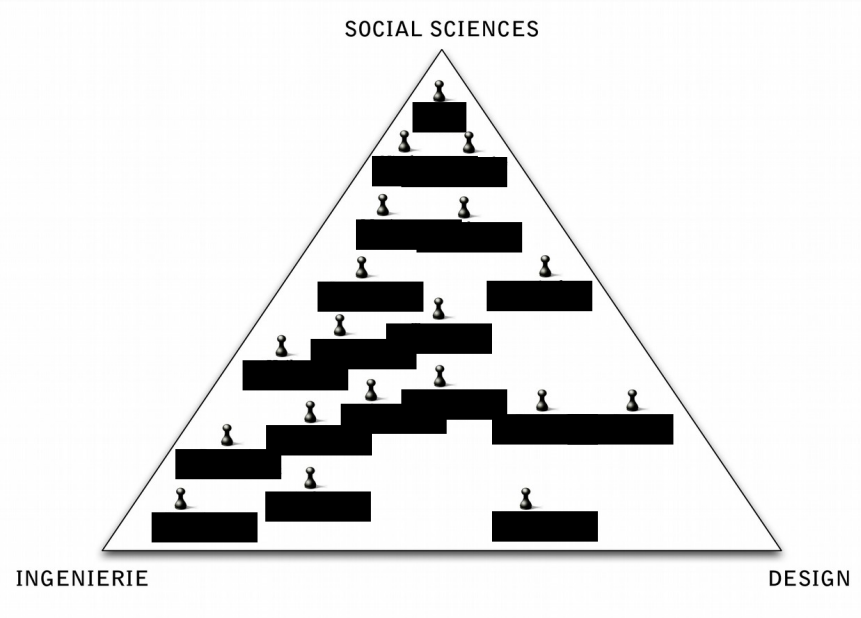
INVESTISSEMENTS D'AVENIR

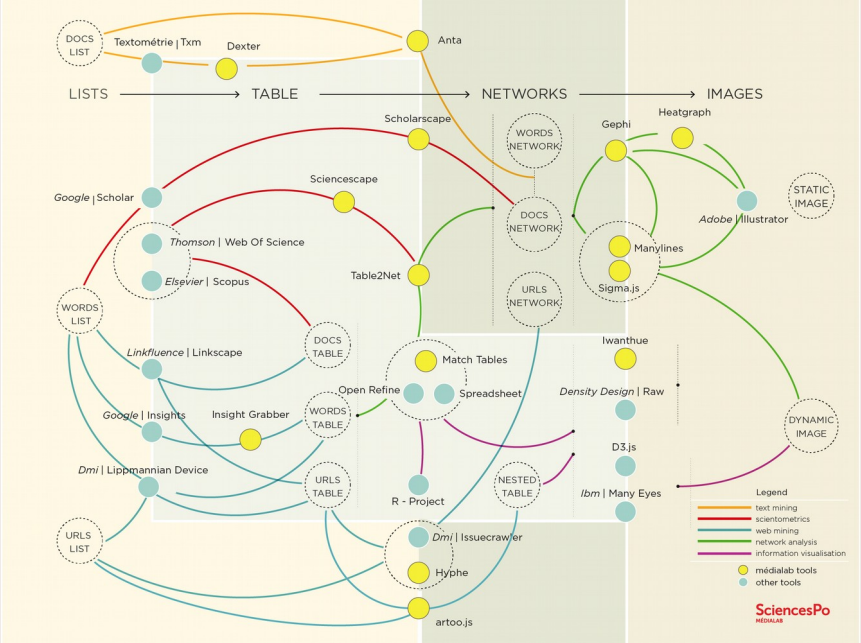
DIME-SHS

médialab @ Sciences Po

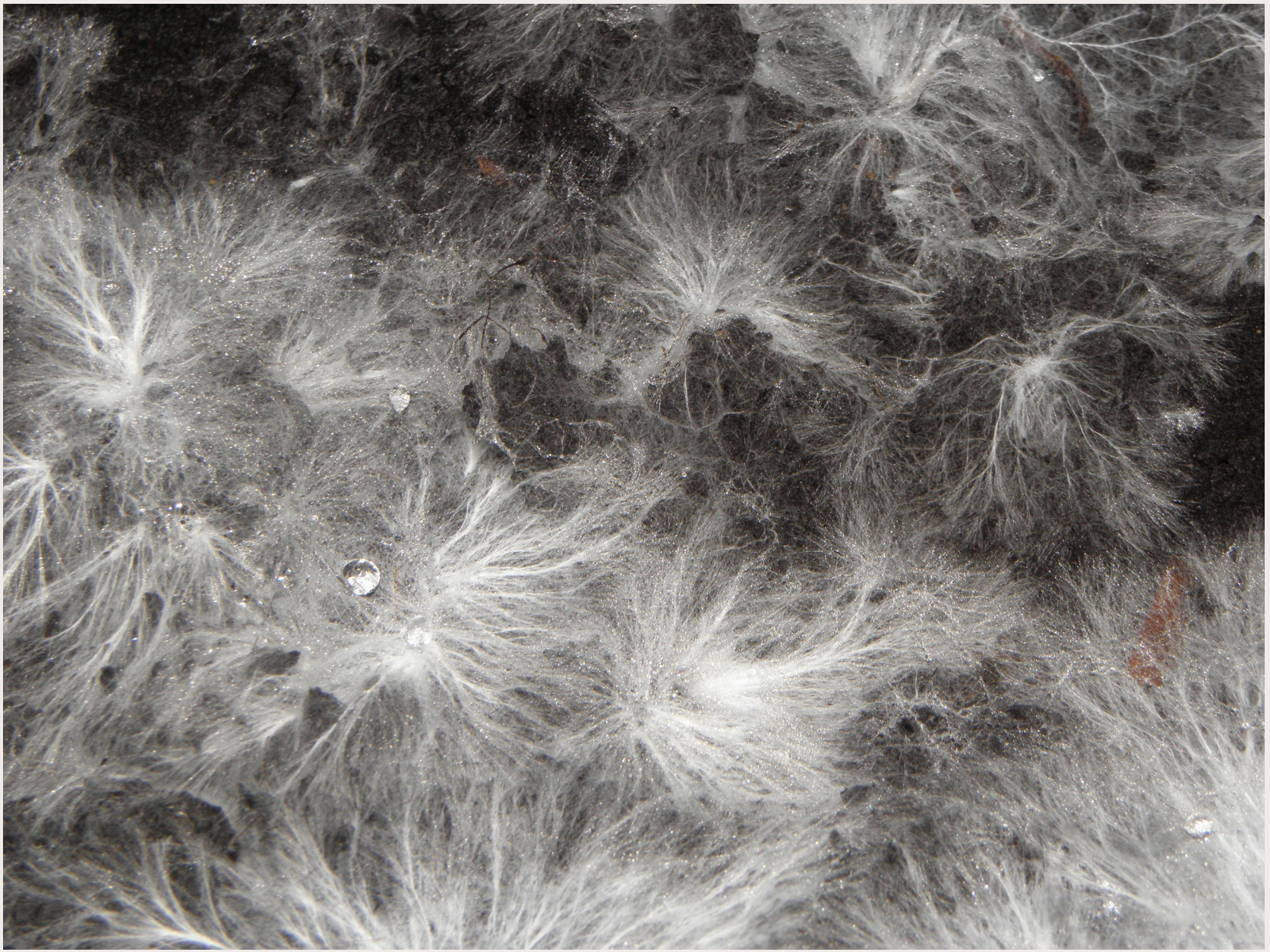
https://medialab.sciencespo.fr

- Laboratoire de recherche pluridisciplinaire fondé par Bruno Latour en Mai 2009, dirigé par Dominique Cardon depuis 2017
- Sciences Sociales, Ingénierie & Design
- Articuler méthodes quali & quanti à travers une approche numérique
- Travailler avec les traces numériques
- Un écosystème d'outils OpenSource
  http://tools.medialab.sciences-po.fr
- METAT : atelier de support ouvert mensuel
  https://www.sciencespo.fr/recherche/fr/content/metat-latelier-de-methodes





Le mycellium : un réseau d'hyphes



Hyphe : un crawler orienté par la recherche

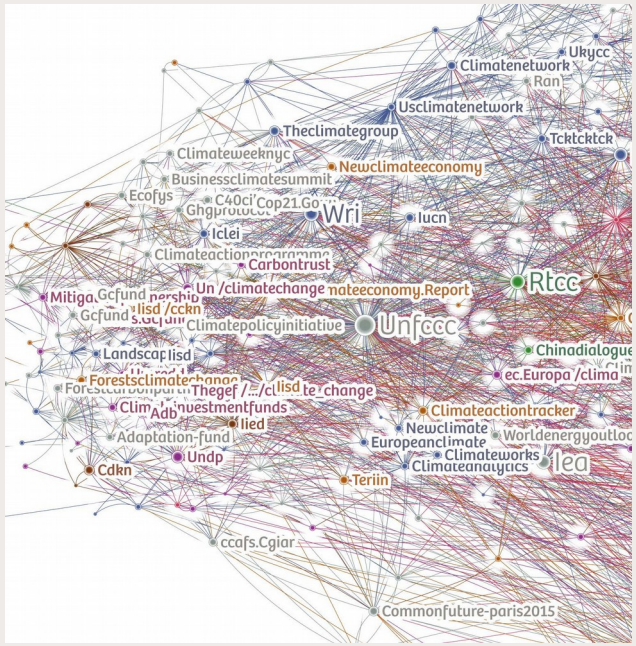
http://hyphe.medialab.sciences-po.fr/demo/

Construire un corpus de documents web
pour étudier un phenomène social en ligne

→ identifier et rassembler des « acteurs web »

→ explorer les liens entre leurs sites

hyphe



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

COP 21

Vie privée

Extrême droite

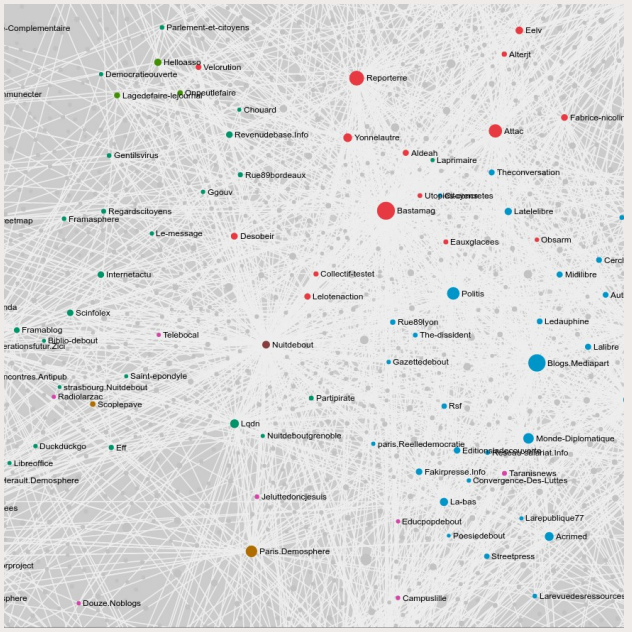
Tissu associatif

Produits laitiers

Cellules souches

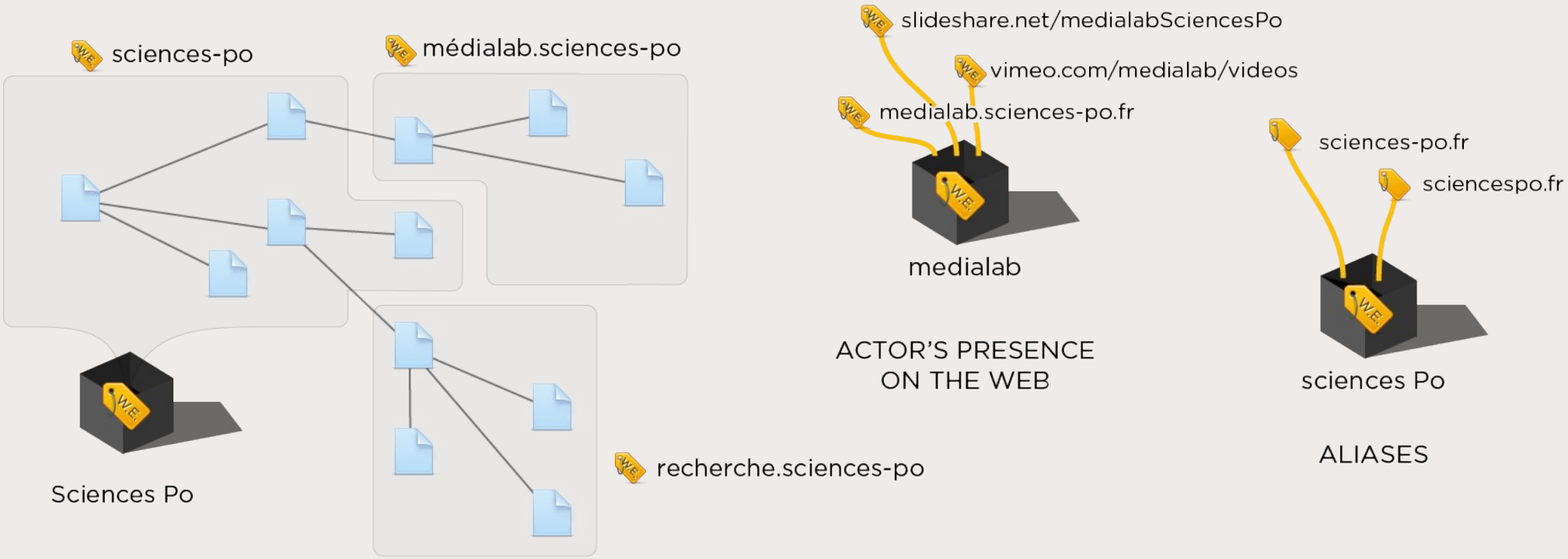
Administrations culturelles

…



http://utopies-concretes.org/

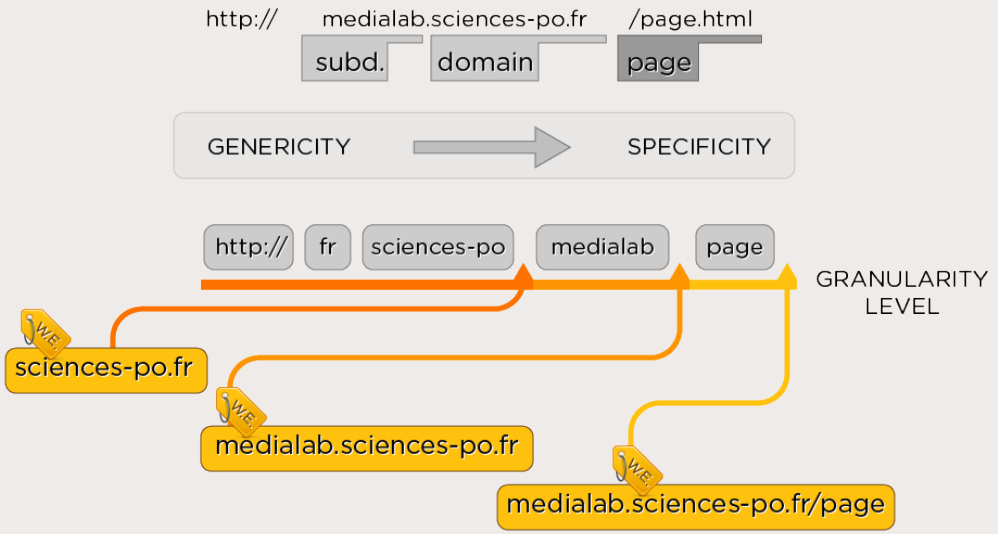
Mais qu'est-ce donc qu'un « site web » ?



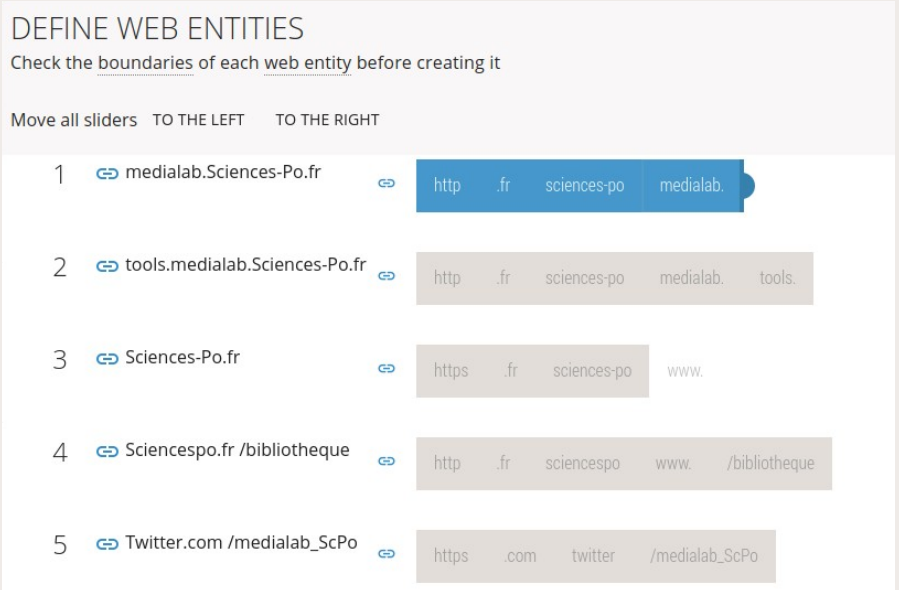
→ « **WebEntité** » : ensemble de pages web agrégées pour rassembler l'incarnation précise d'un acteur sur le web au sens d'une question de recherche spécifique

=> ensemble de préfixes d'URLs

Définir finement les frontières de nos acteurs



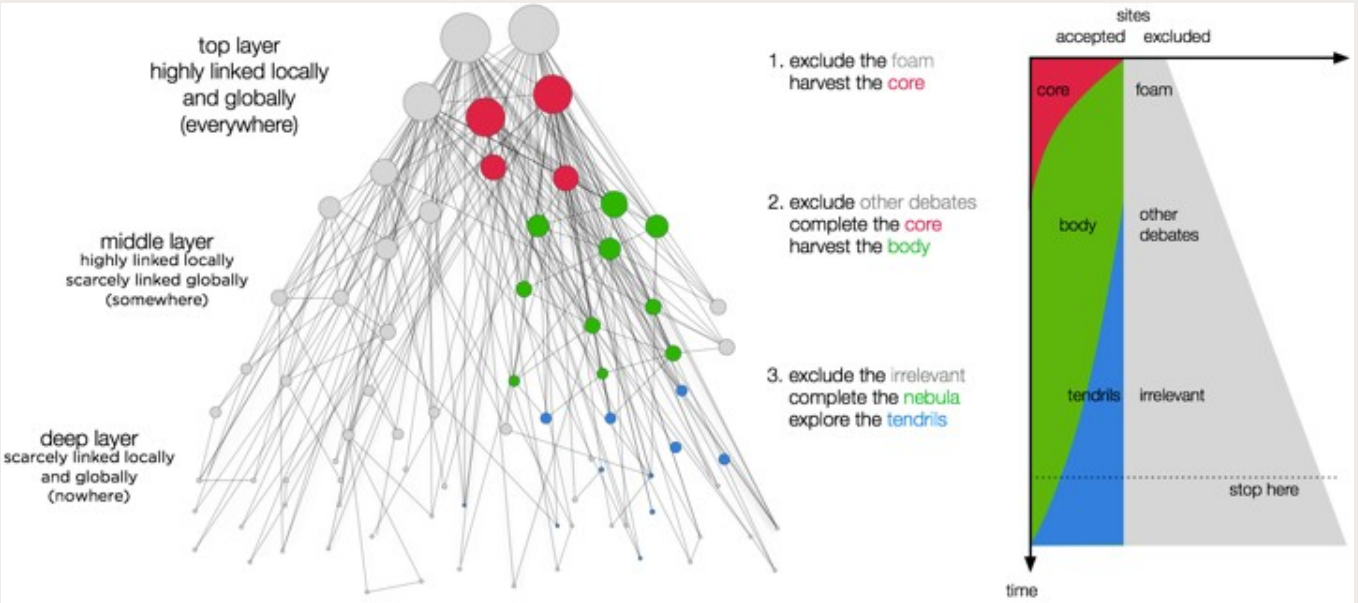
Ajustement manuel des limites des WebEntités par le chercheur en choisissant un niveau de préfixe



Hyphe : une stratégie de crawling contrôlé

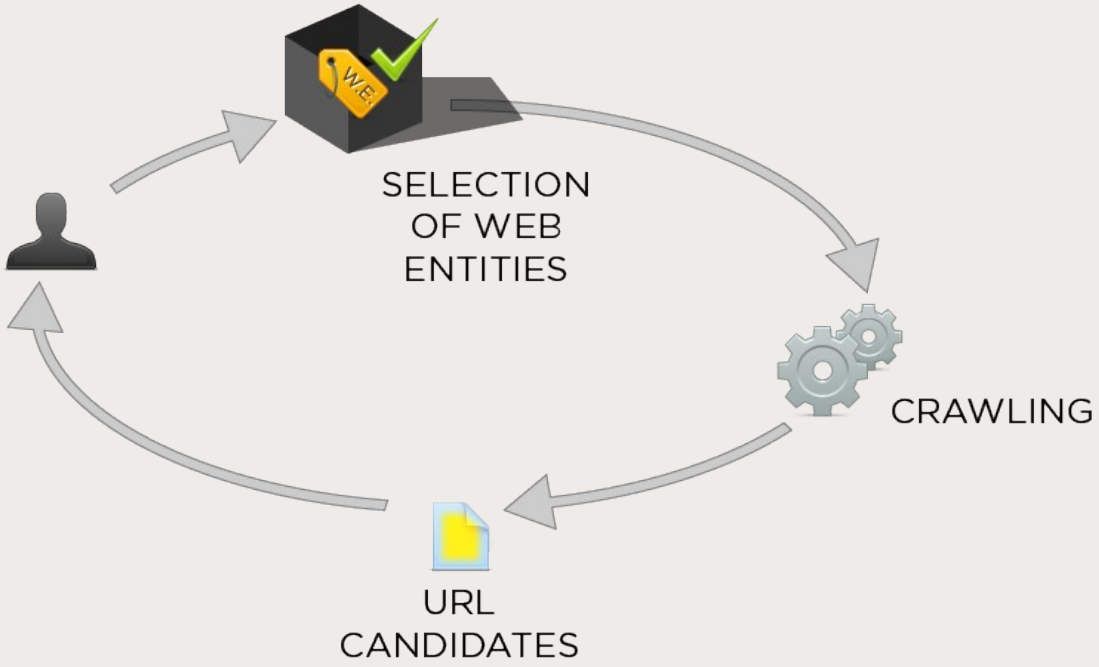
- Crawlers classiques : snowballing
  - → Surreprésentation des couches hautes (Google, YouTube, Wikipedia…)
  - → Dérive thématique rapide

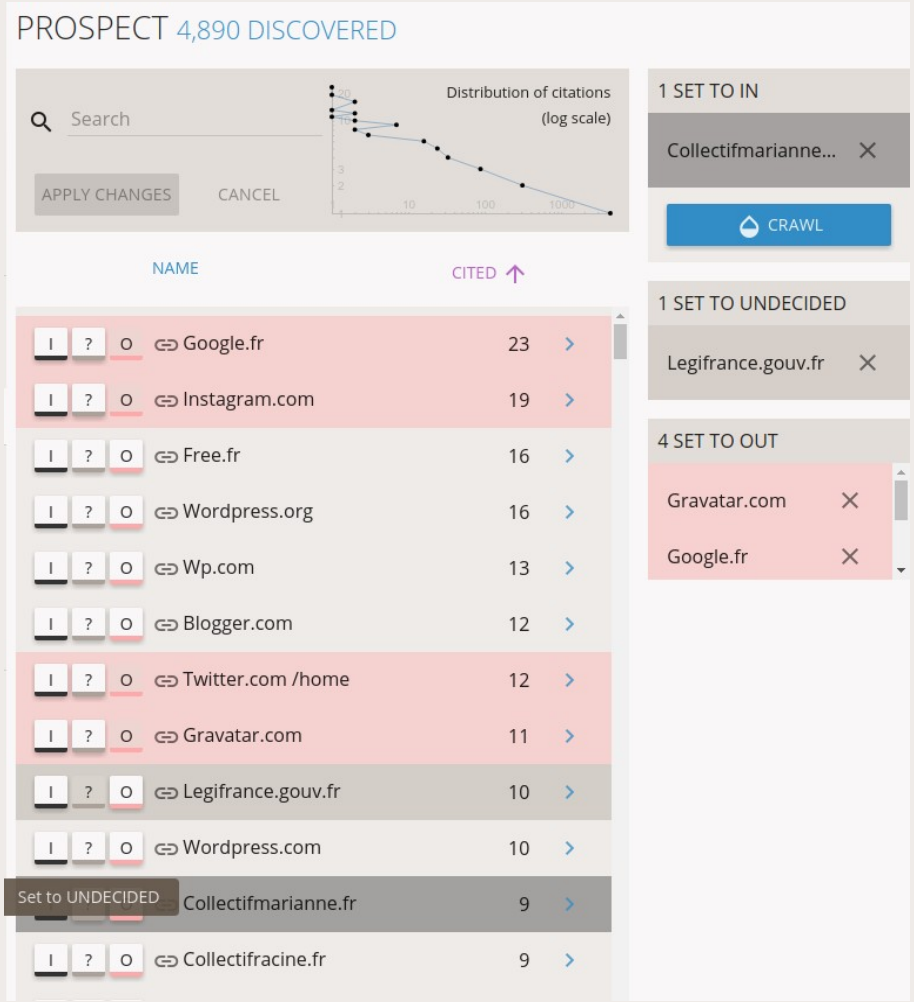
- Hyphe : crawling semi-automatique
  - Fouille systématique des pages des WebEntités choisies uniquement
  - Choix humain des autres WebEntités à crawler grâce au degré de citation



Boucle de prospection Web : une curation itérative

- Extension pas à pas du corpus en sélectionnant les acteurs web



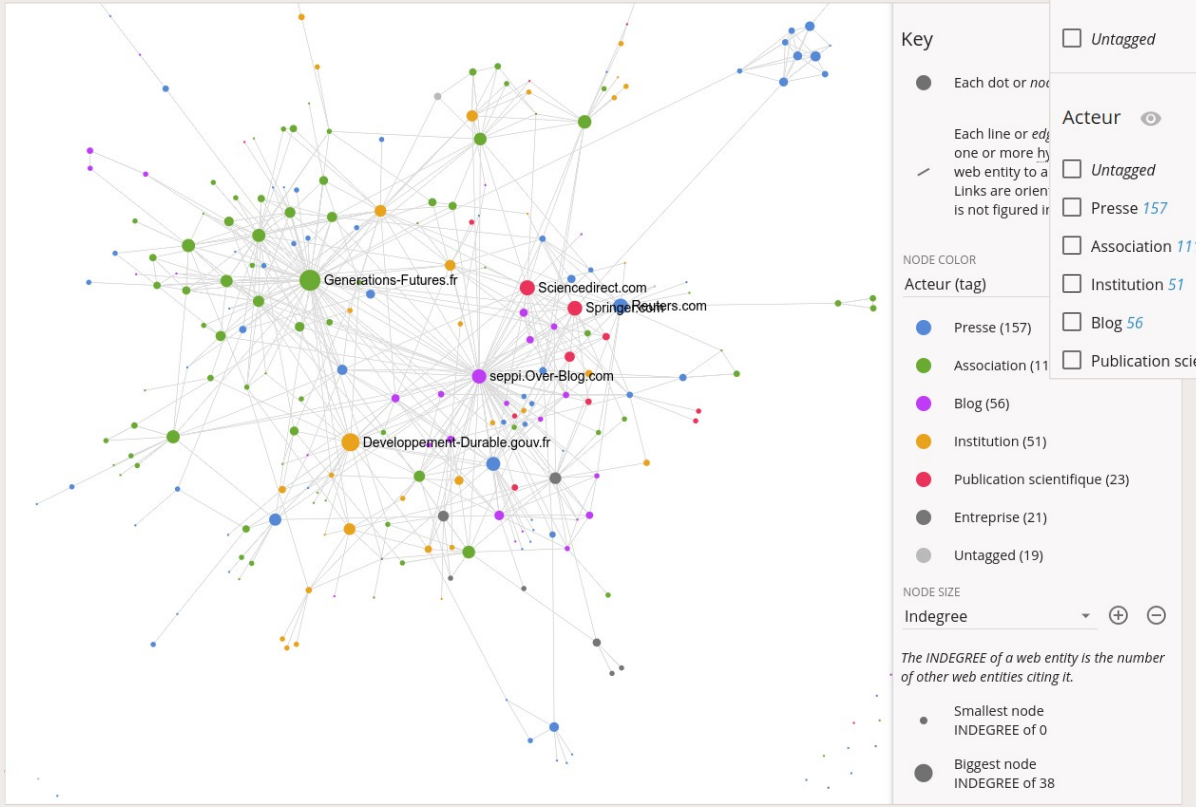


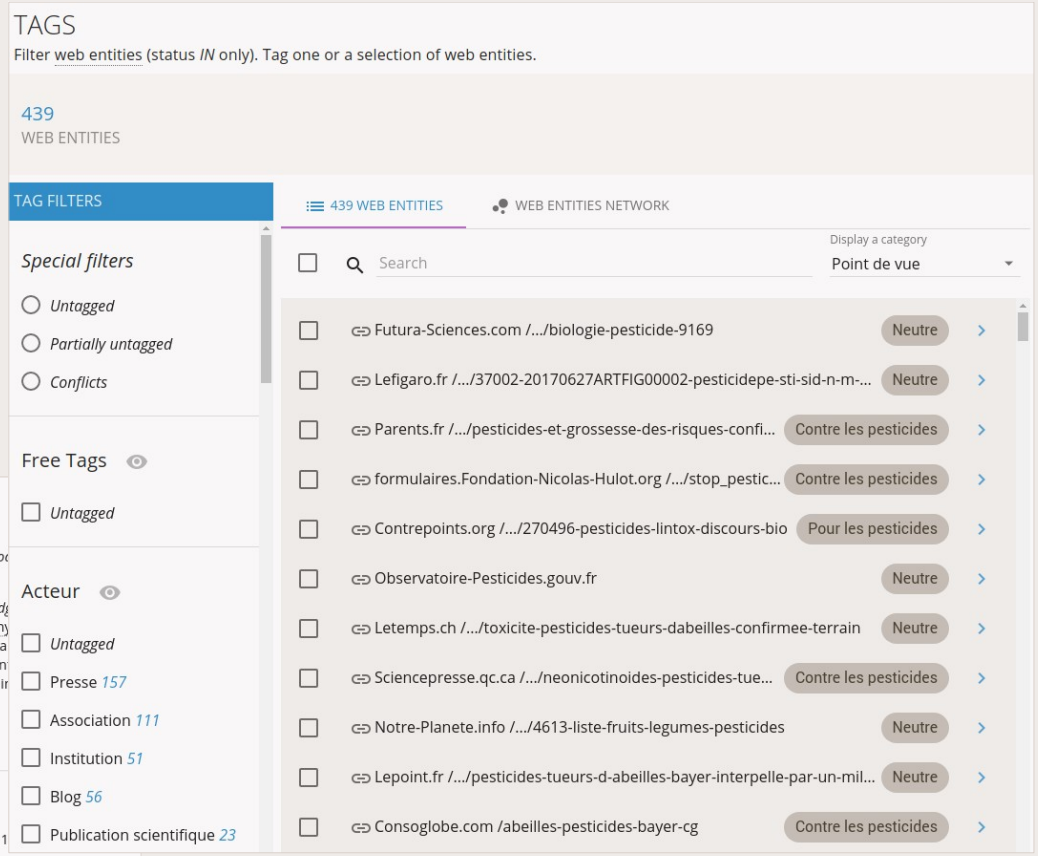
- Coût humain et temporel
- Quand décider d'arrêter ?

  → seuil de degré de citation

Qualifier les acteurs du corpus (tags)

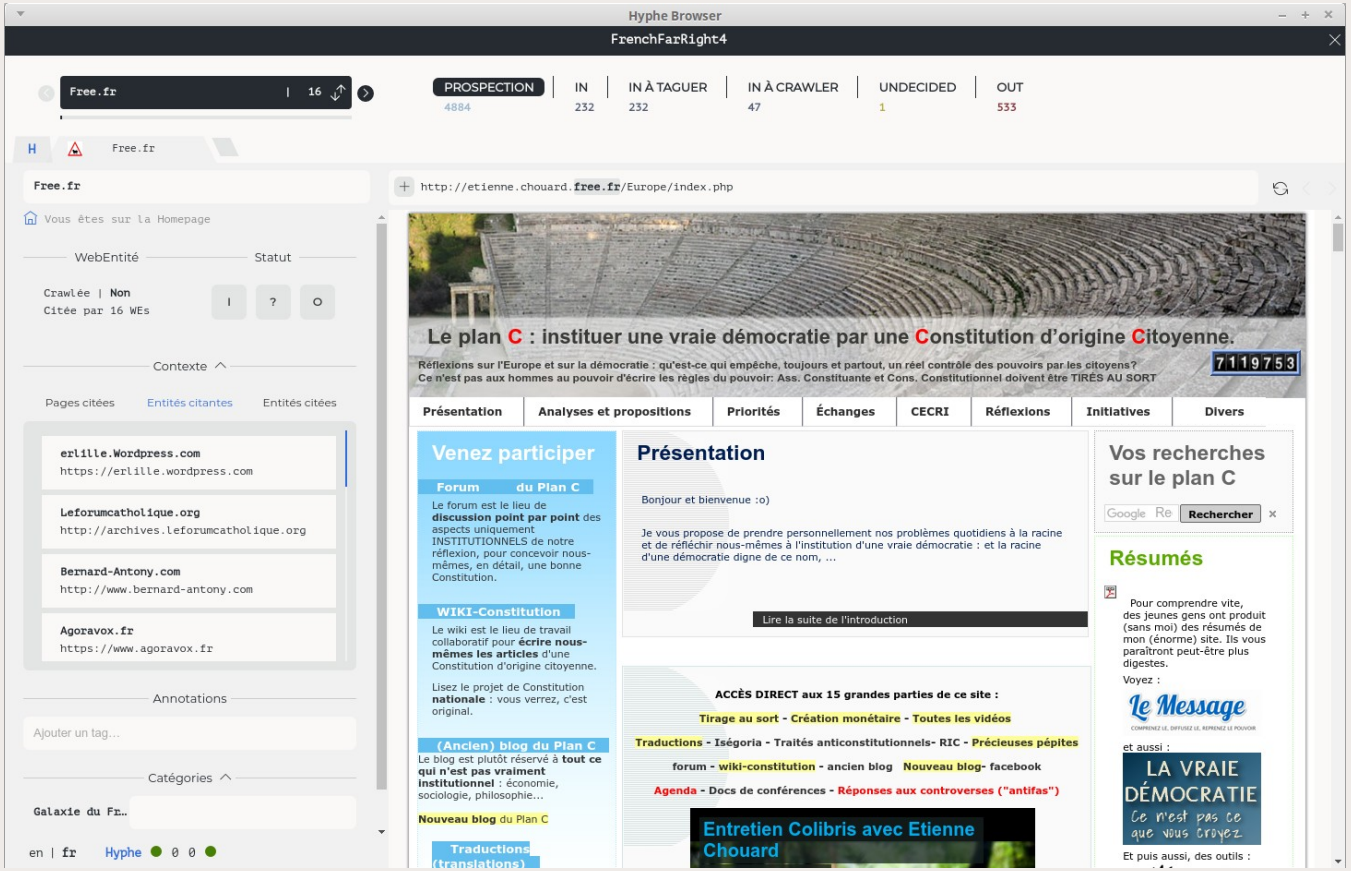
- Tags catégoriels (clé/valeur)
- Annotations





HyBro : un navigateur web conçu pour Hyphe

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



Héritage du « NaviCrawler » : construire un corpus web en visitant les pages

→ prospection et tagging « in-situ » (terrain numérique)

→ enseigner le web aux étudiants (pédagogie innovante FORCCAST)

Différents usages de Hyphe

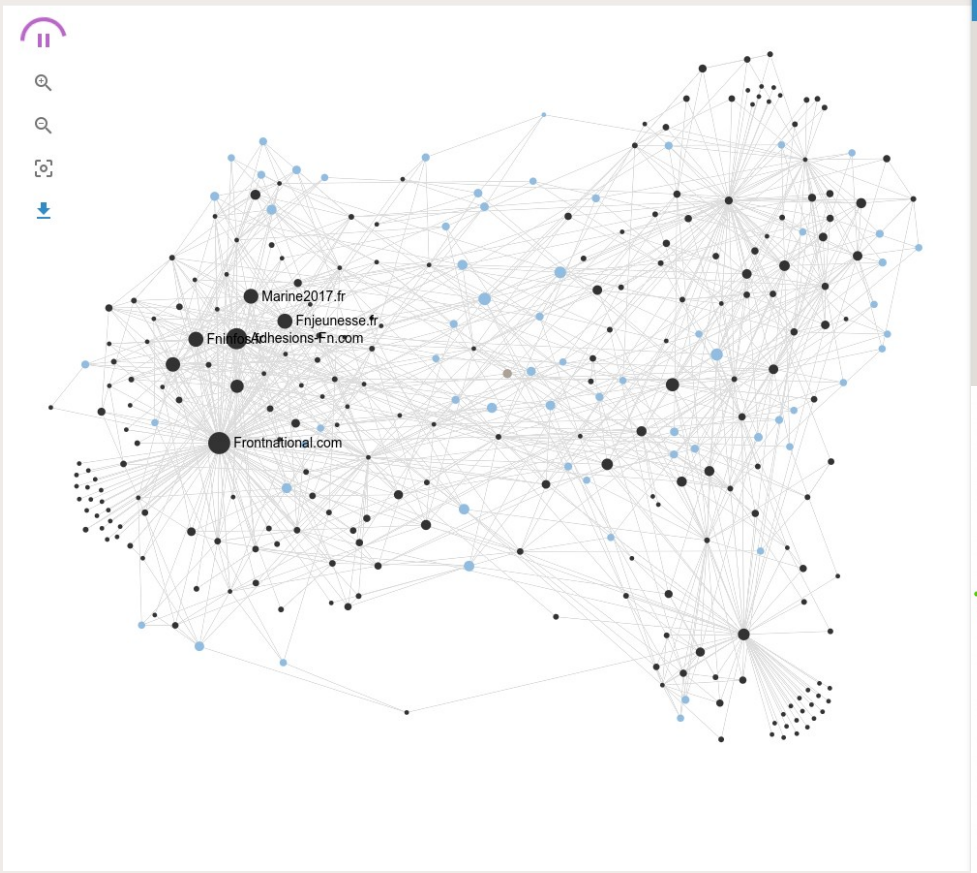
- Une méthodologie complète :

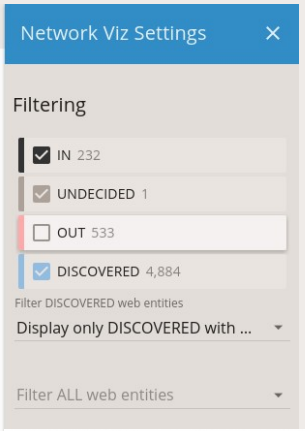
  sourcing, curation semi-automatisée, construction itérative,
  analyse exploratoire, catégorisation qualitative
  visualisation de réseaux, analyse statistique quantitative

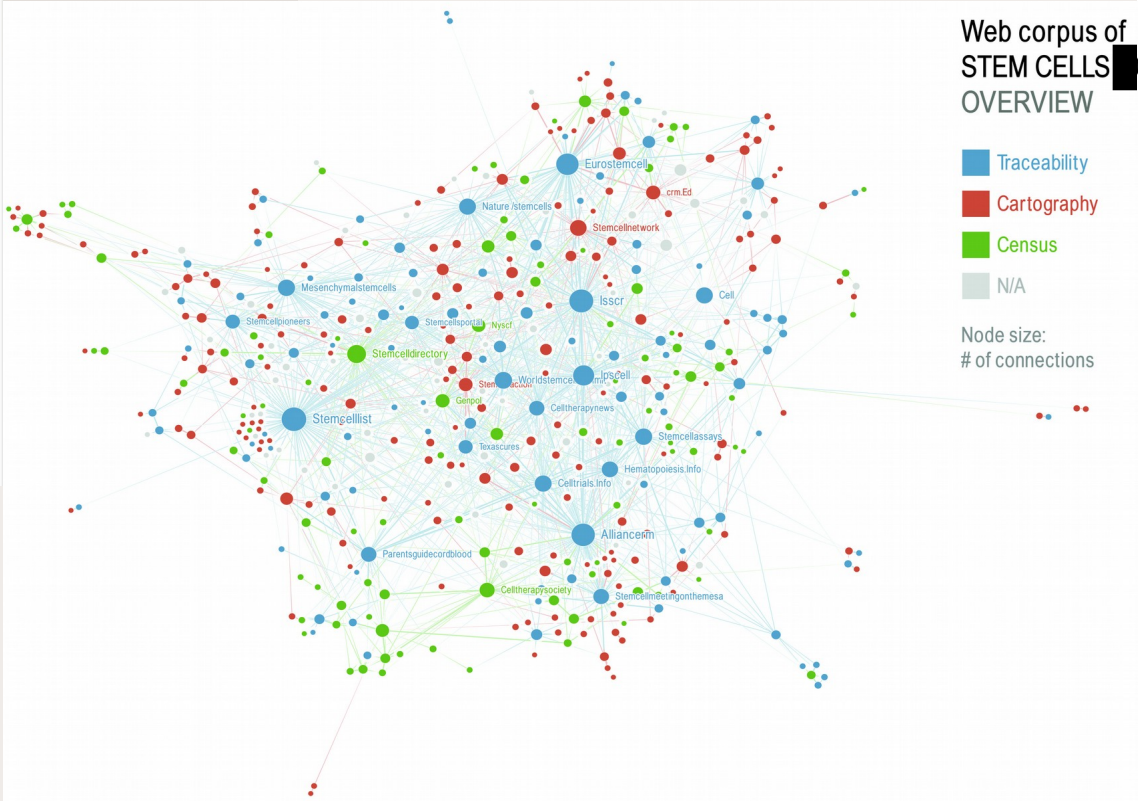
- Divers publics-cibles :
  - Recherche : équiper les chercheurs en SHS pour réaliser un terrain web
  - Pédagogie : enseigner aux élèves le web au-delà de Google & Facebook

- Des analyses de nature et d'ampleur diverses :
  - la structure interne d'un site web
  - les liens entre un ensemble d'acteurs d'une thématique
  - les alliances et les oppositions entre les acteurs d'une controverse
  - etc.

Analyse de réseau : clusters, oppositions & affinités

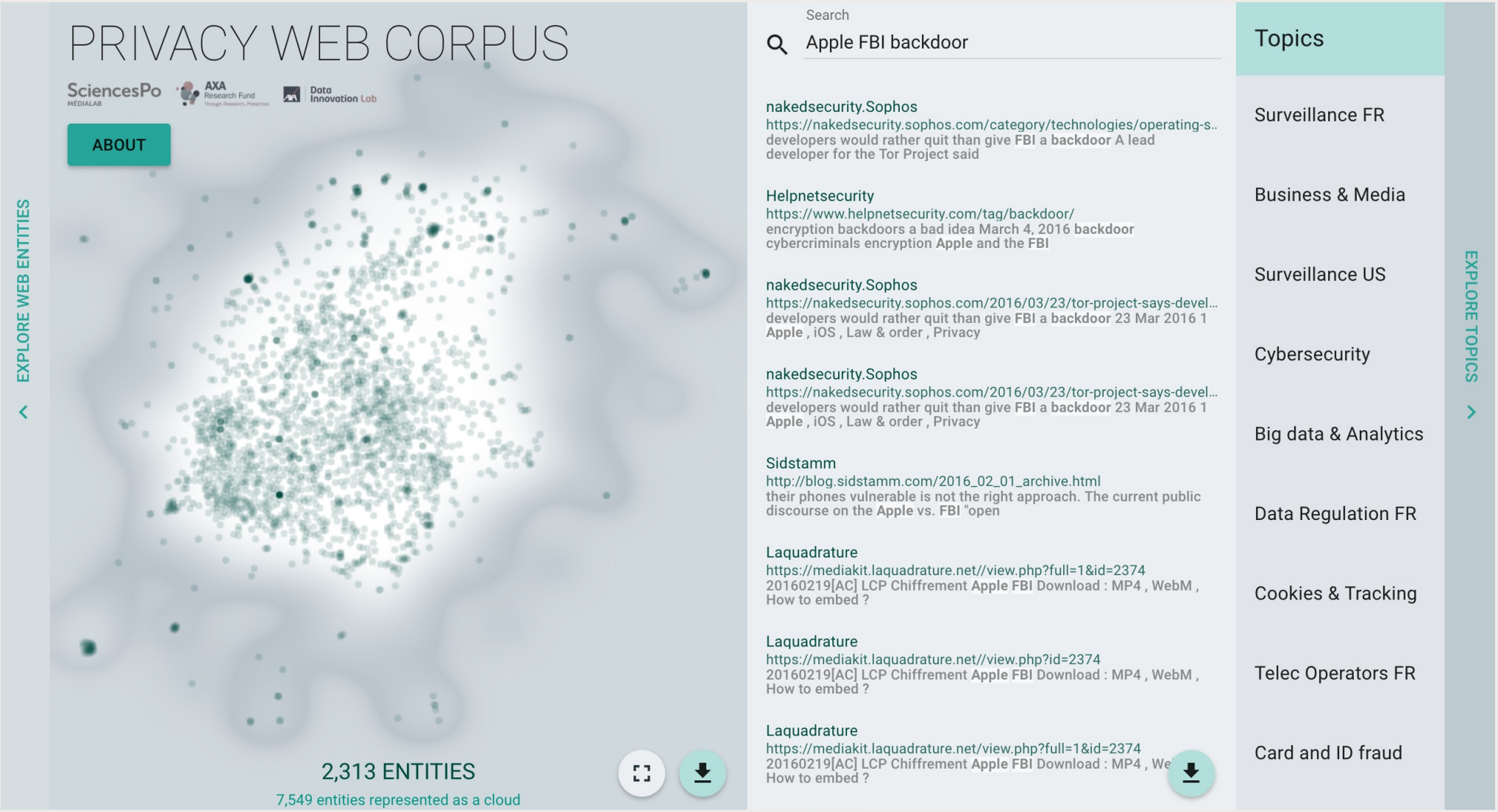






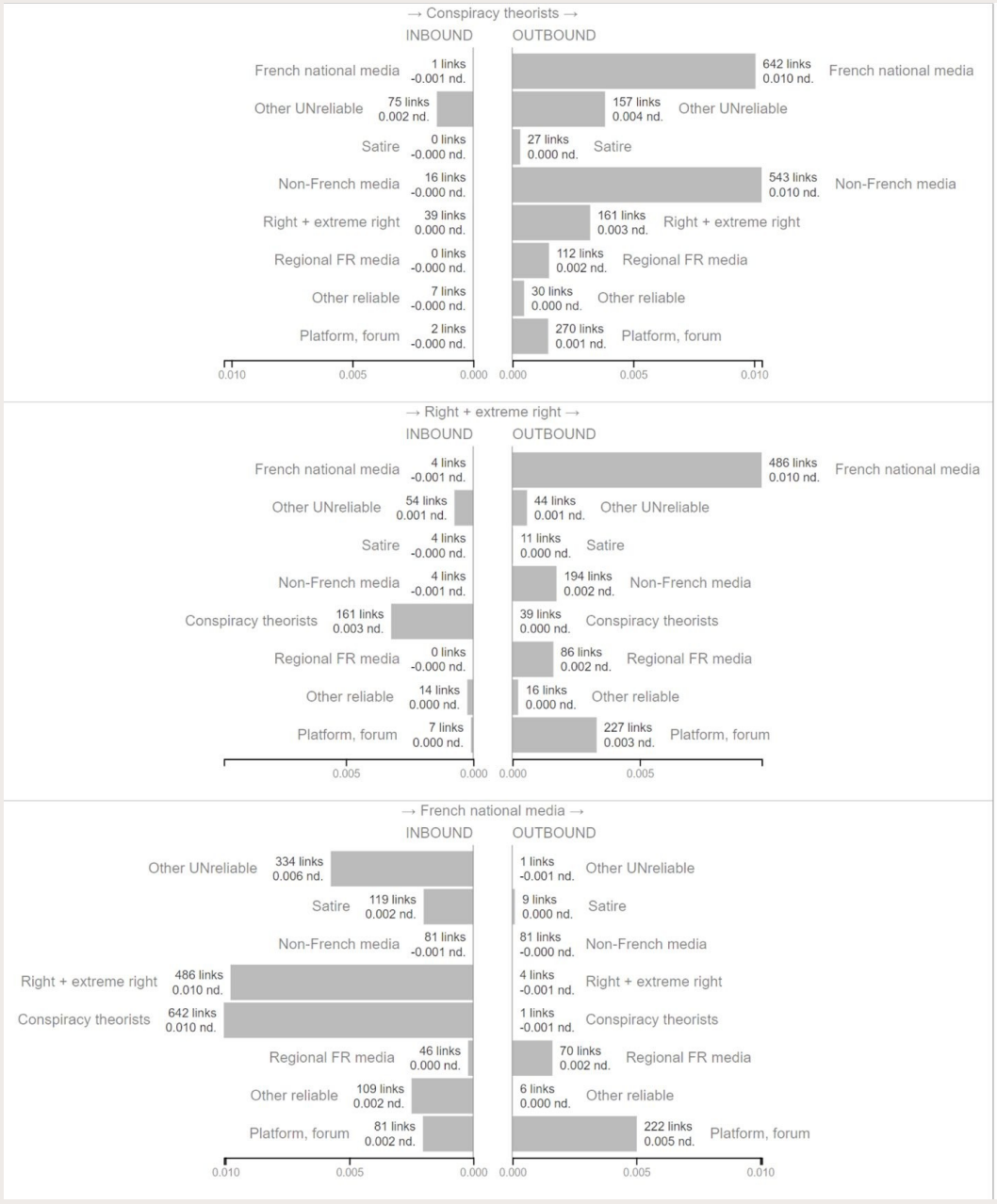
Social Representations of Stem Cells, Virginie Tournay, CEVIPOF, 2016

Analyse du fond : traiter les contenus texte



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

Analyser la circulation & la structuration des liens



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

other

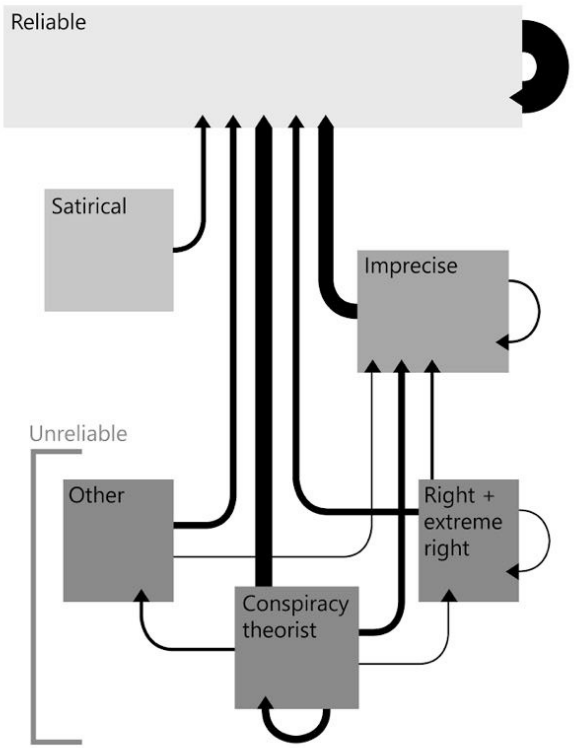
Right and extreme right
- Not reliable
- Cited only by conspirationists
- Cite many reliable sources

other

other

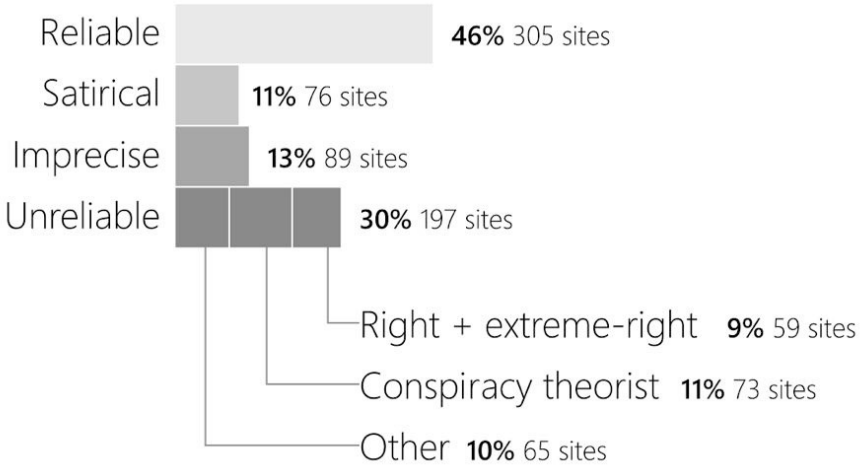
French national media
- Reliable *(*in this qualification)*
- Cited by everyone
- Cite only platforms and regional media

Exploiter la directionnalité des hyperliens

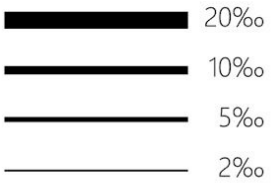


Most hyperlinks stem from the unreliable and aim at the reliable resources

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



The thickness si proportional to the normalized link density*

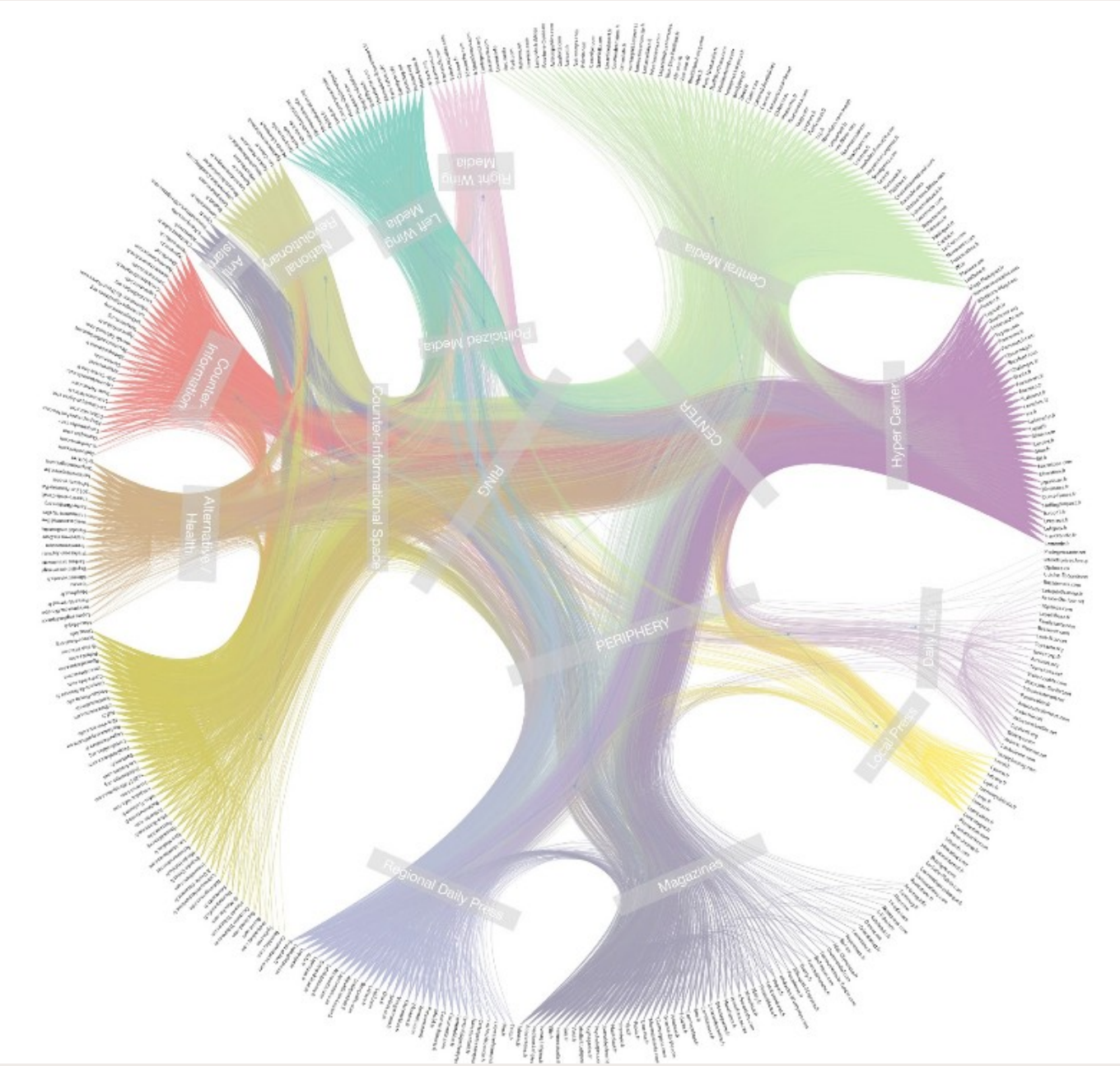


*Accounts for the different group sizes.

Densities below 2‰ have been omitted.

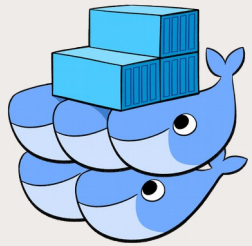
https://www.researchgate.net/publication/320225750_Visual_Network_Exploration_for_Data_Journalists

Explorer les dynamiques de polarisation



Quoi de neuf à l'avenir ?

- Importer / exporter des corpus ou lists de WebEntités & crawls :
  - → duplication, reproduction
  - → exploration longitudinale dans le temps
- Analyse de contenus textes (TAL) intégrée
- Utilisation des technologies web modernes pour gérer les sites récents réalisés entièrement en JavaScript (Facebook, applications React, etc.)
- Outils de contrôle qualité des crawls
- Outils d'archivage & d'exploration pour publier les corpus finalisés
- Outil de déploiement simplifié de Hyphe chez des hébergeurs sur le cloud (OVH, CityCloud, etc.)



Et maintenant, à vous de jouer !

Hyphe est un Logiciel Libre et Open Source :

https://github.com/medialab/hyphe

Une version de démonstration limitée est accessible gratuitement :

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

Le faire installer sur les serveurs d'une université est assez simple.

Le déployer temporairement sur une infrastructure cloud
est accessible à un coût relativement modeste.

Des questions?

benjamin.ooghe@sciencespo.fr

@boogheta **@medialab_ScPo**

Bibliographie

Publications de référence :

- Jacomy M., Venturini T., Heymann S., Bastian M. (2014), **ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software**, PLos ONE, 9(6), 1-18

  https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679

- Jacomy M., Girard P., Ooghe-Tabanou B., Venturini T. (2016), **Hyphe, a Curation-Oriented Approach to Web Crawling for the Social Sciences**, ICWSM 2016, Cologne

  https://spire.sciencespo.fr/hdl:/2441/6obemb2hsj9pboj9bbvc7sftne

- Plique G., Jacomy M., Ooghe-Tabanou B., Girard P. (2018), **It's a Tree... It's a Graph... It's a Traph! Designing an on-file multi-level graph index for the Hyphe web crawler**, FOSDEM 2018, Bruxelles

  https://medialab.github.io/hyphe-traph/fosdem2018/#/

- Ooghe-Tabanou B., Girard P., Jacomy M., Plique G. (2018), **Hyperlink is not dead!,** ACM Proceedings of the 2nd International Conference on Web Studies (WS.2 2018) Paris.

  http://hyphe.medialab.sciences-po.fr/docs/20181004-ACM-WebStudies-HyperlinkIsNotDead.pdf