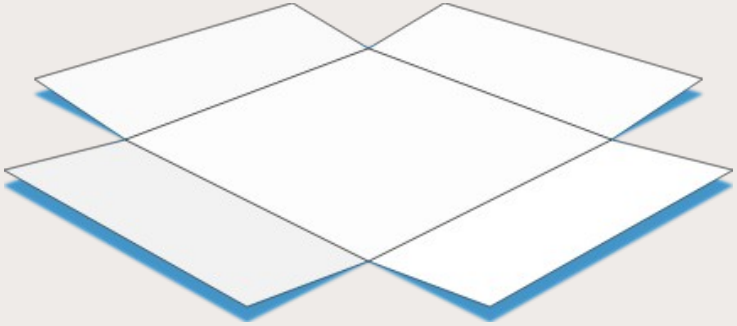
Web communities network mapping with Hyphe

Benjamin Ooghe-Tabanou



Web communities network mapping with Hyphe

NodeXL Academy '22
January 19th 2022

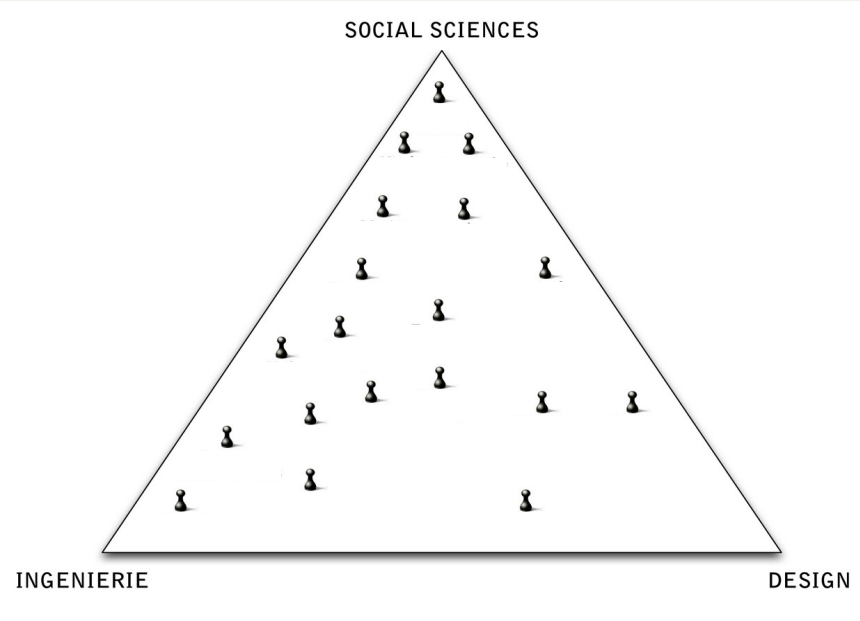
Benjamin Ooghe-Tabanou (@boogheta)
Sciences Po médialab (@medialab_ScPo)

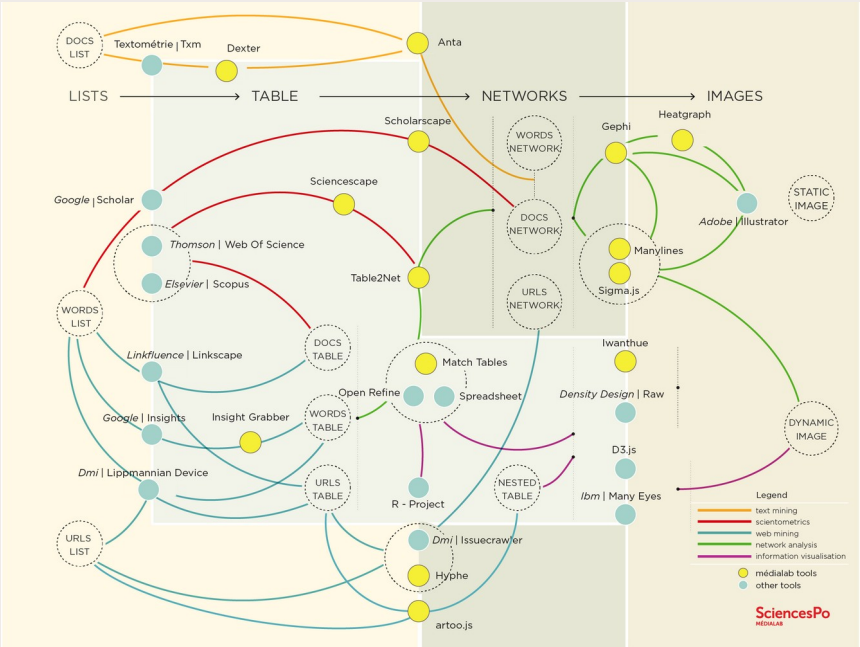
SciencesPo
MÉDIALAB

médialab @ Sciences Po

https://medialab.sciencespo.fr

- Pluridisciplinary Research Lab created by Bruno Latour in May 2009, led by Dominique Cardon since 2017
- Social Sciences, Engineering & Design
- Articulate qualitative & quantitative methods through a digital approach
- Work with digital traces
- Deploy an ecosystem of tools
  http://tools.medialab.sciences-po.fr
- METAT: a monthly Open Support Workshop
  https://www.sciencespo.fr/recherche/fr/content/metat-latelier-de-methodes





Bruno Latour, médialab founder

« Google is nice, but we need something better »

The Indian Express, 2011

Webcrawlers? What is that?

« Web crawling is the process of building a collection of webpages by starting with an initial set of URLs and recursively traversing the corresponding pages to find additional links. »

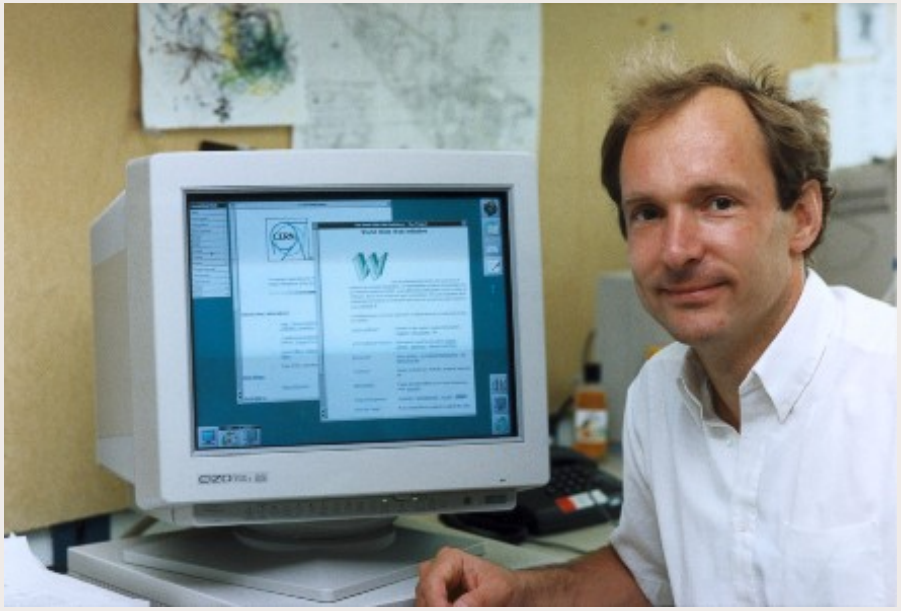
Gabe Ignatow & Rada Mihalcea, in: Text Mining: A Guidebook for the Social Sciences

But why would anyone want to do that?

- create an index for a search engine (like... you know...)
- extract contents or structured data (like scraping)
- perform some text mining
- study network communities through hyperlinks

Crawling the web for social sciences?

« **Hyperlink »** is at the core of the Web's architecture
→ charged with meaning and 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**.*

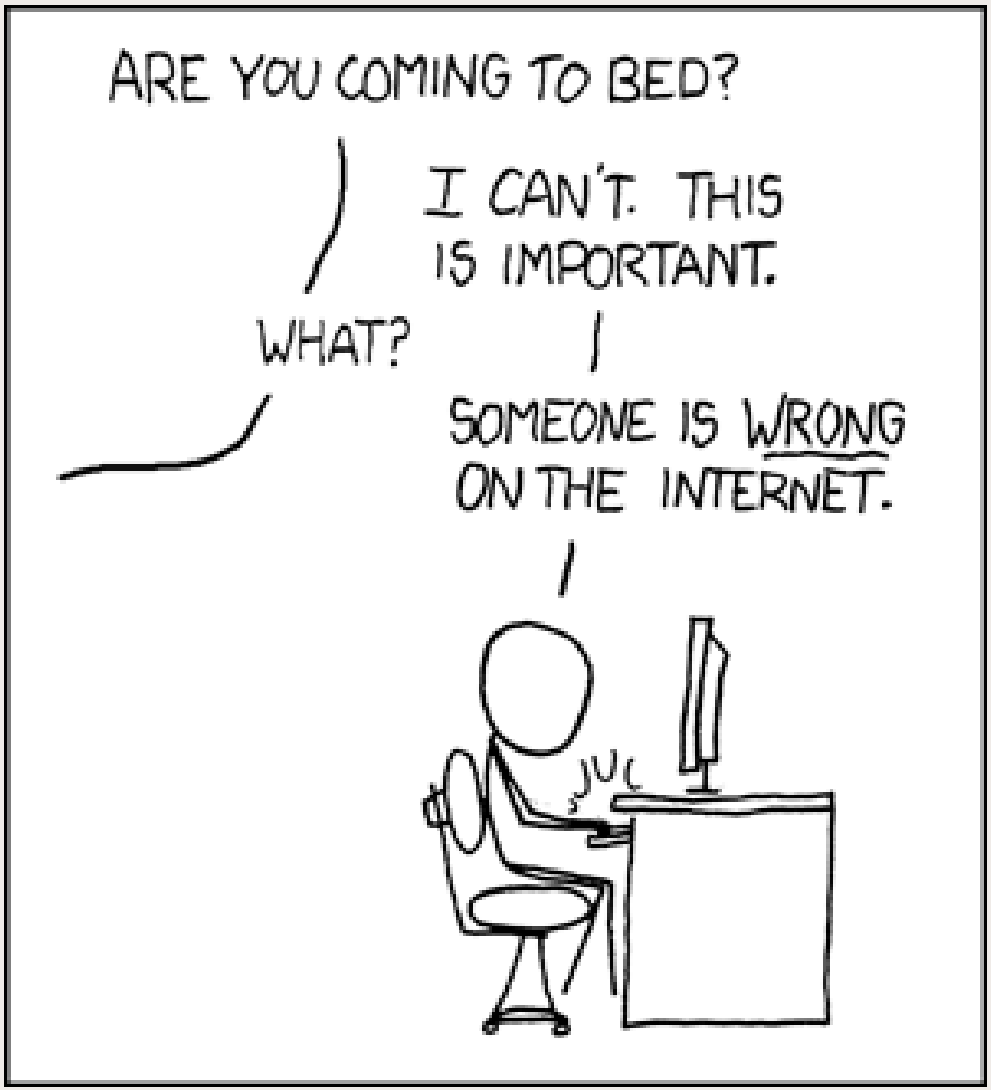
*The process of proceeding from node to node is called **navigation**. »*

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?

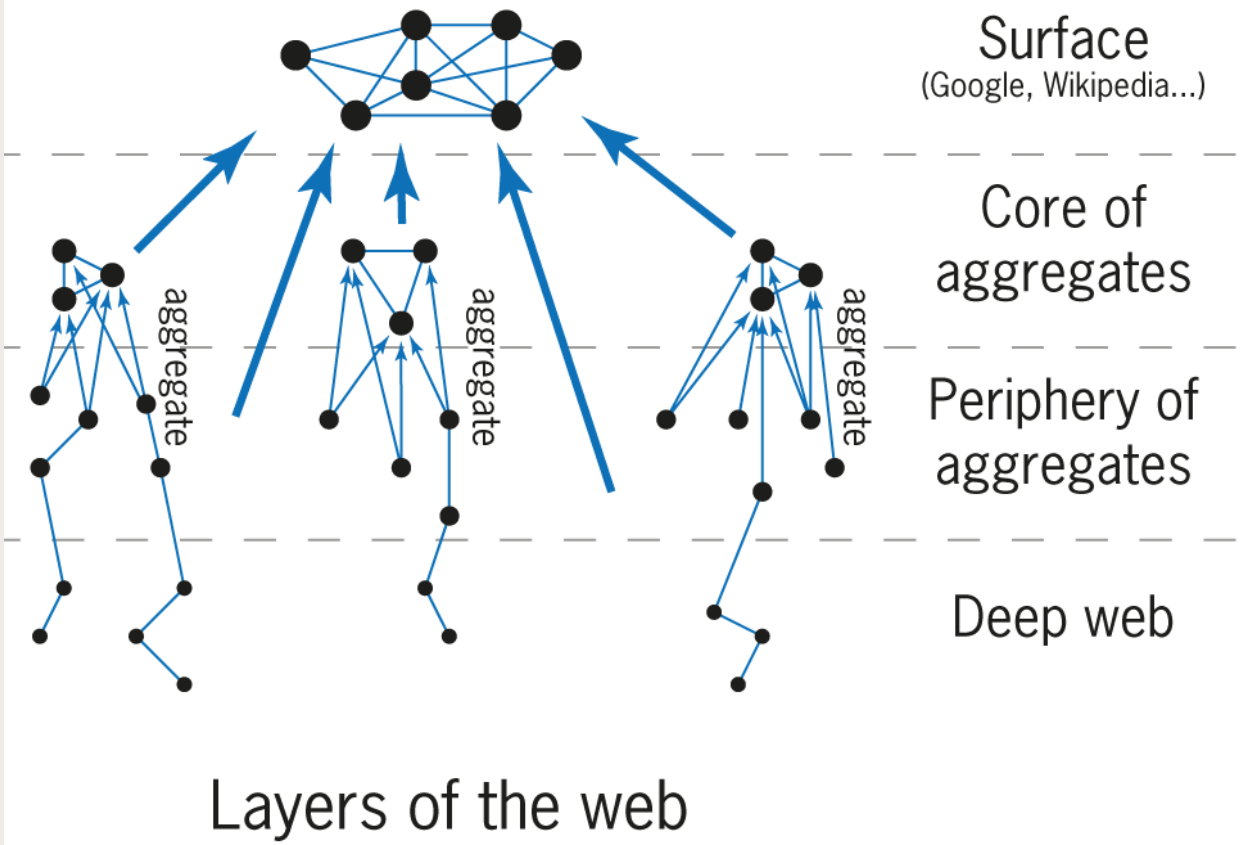
The Web: a place of controversies and debates



A bottom-up hierarchy emerged from hyperlinks

« Matthew effect » : preferential attachment

→ new web pages tend to cite the already most cited ones



Hyphe: a research directed crawler

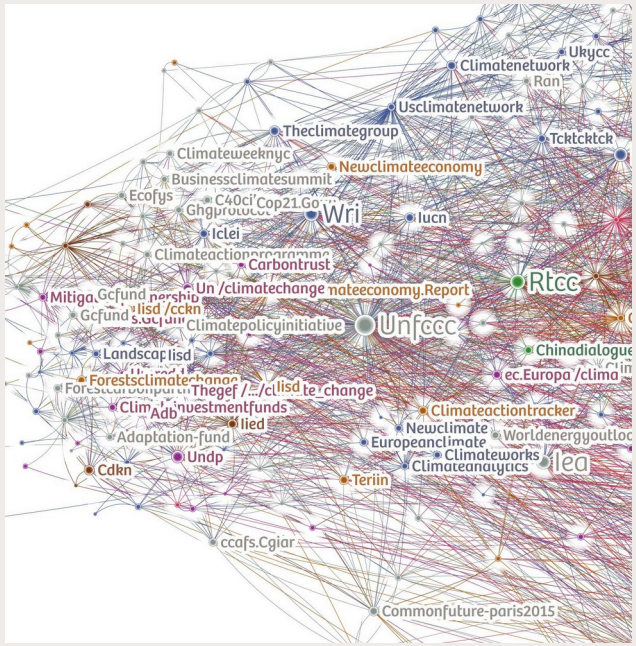
http://hyphe.medialab.sciences-po.fr/demo/

Build your own web documents corpus
to study social phenomena online

→ gather « web actors »

→ explore hyperlinks between them

hyphe



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

COP 21

Vie privée

Extrême droite

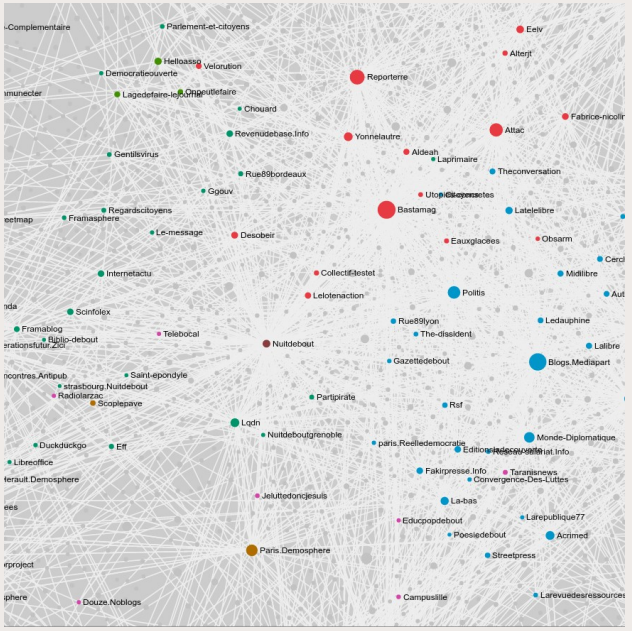
Tissu associatif

Produits laitiers

Cellules souches

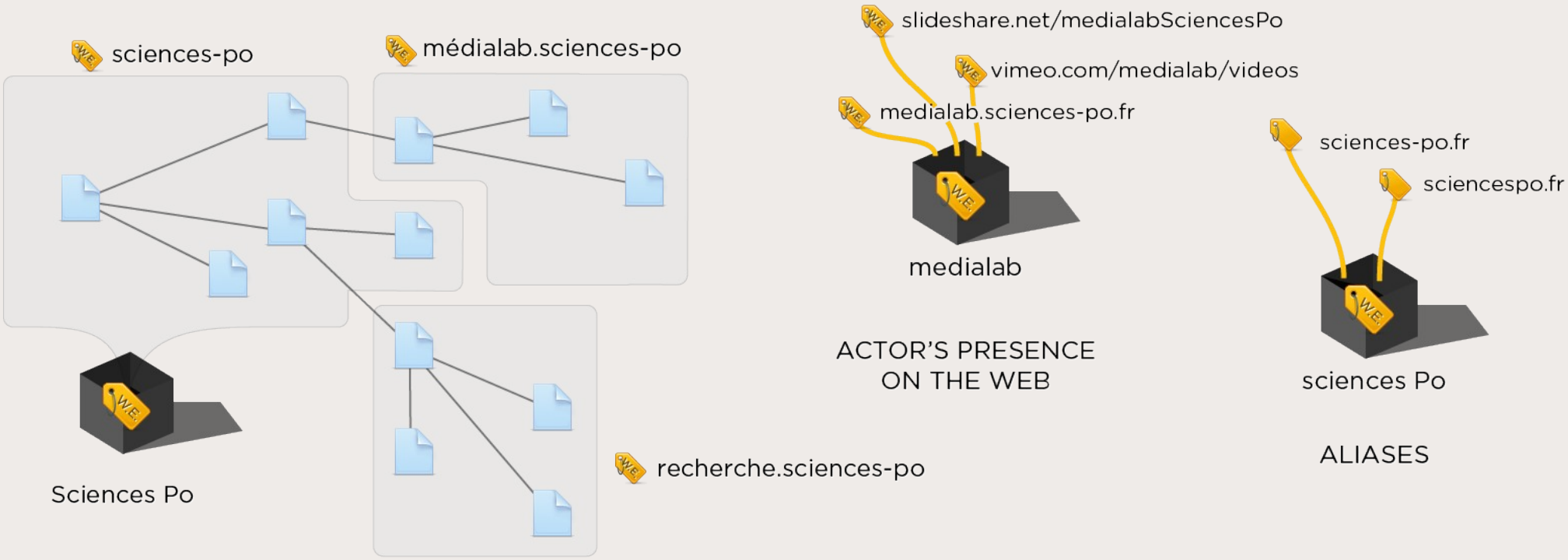
Administrations culturelles

…



http://utopies-concretes.org/

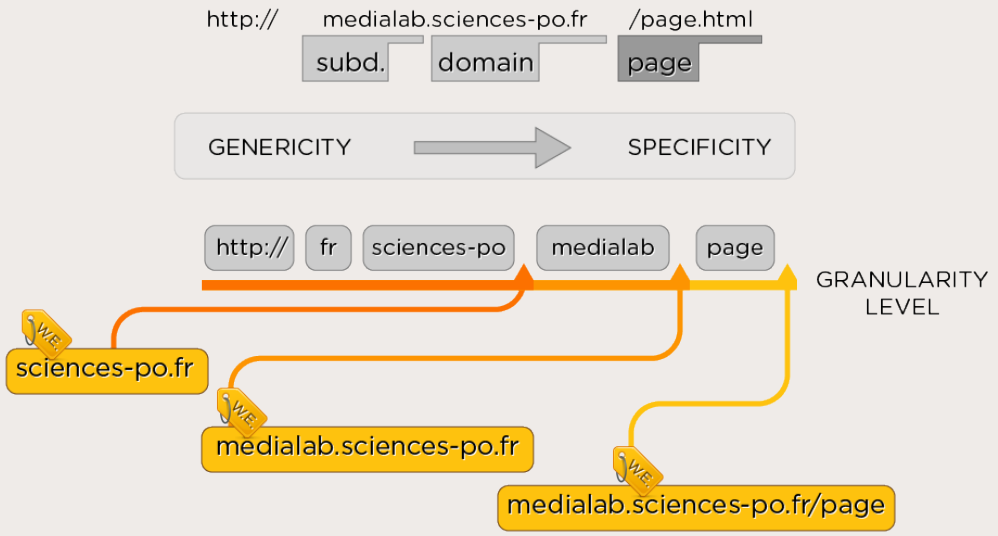
There is no such thing as a « website » !



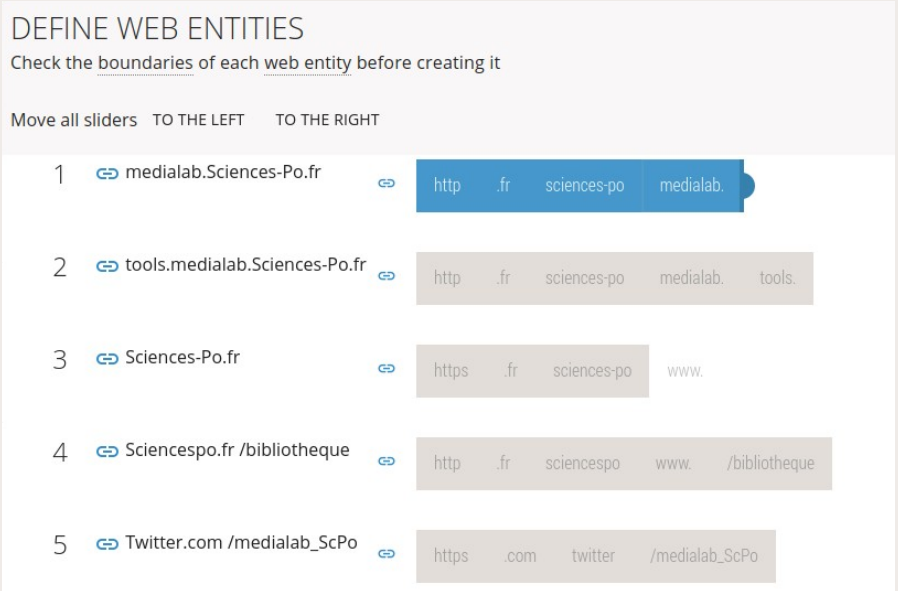
→ « **WebEntities** » : bundles of webpages aggregating coherent actors to answer a specific research question

= set of URL prefixes

Finely delimit the web territories of actors



Manually setup prefix patterns to adjust the cursor of « WebEntities »

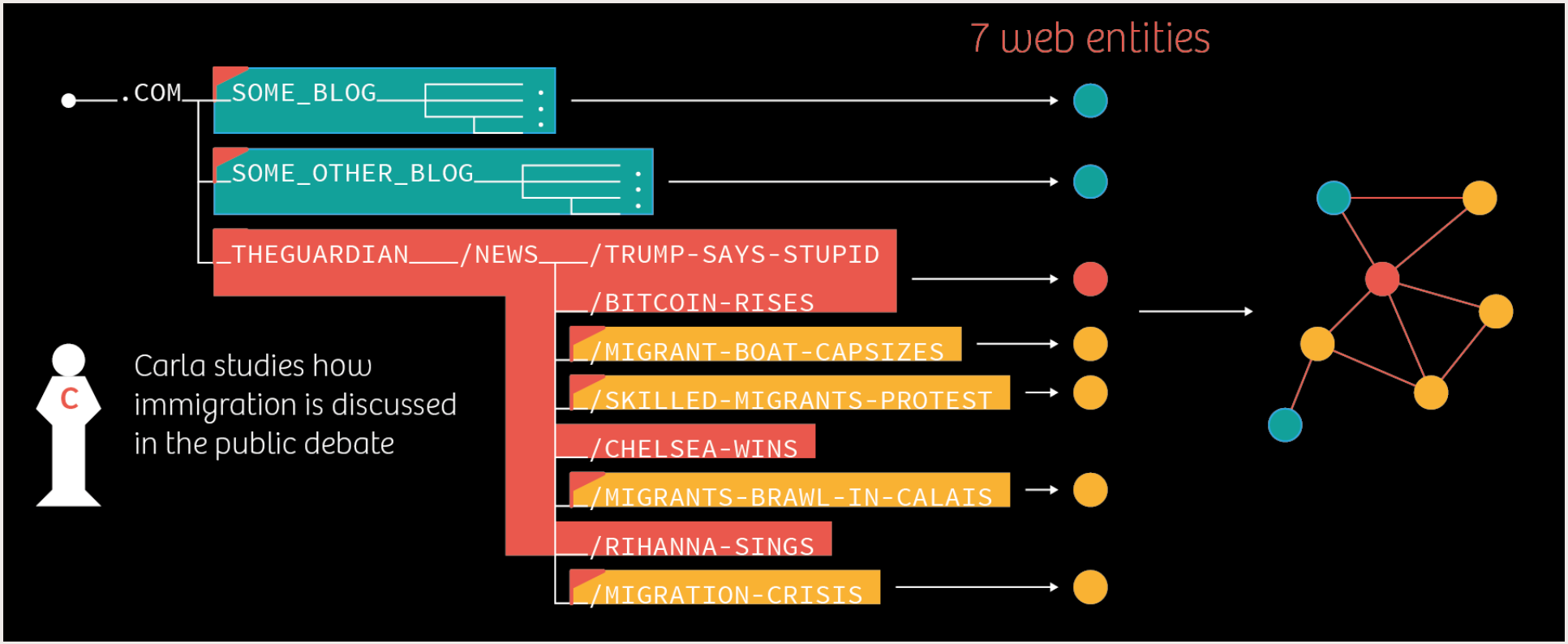


Which data structure to manage hypertexts?

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

- **Tree** of urls
- **Graph** of hyperlinks
- Dynamic branches aggregates

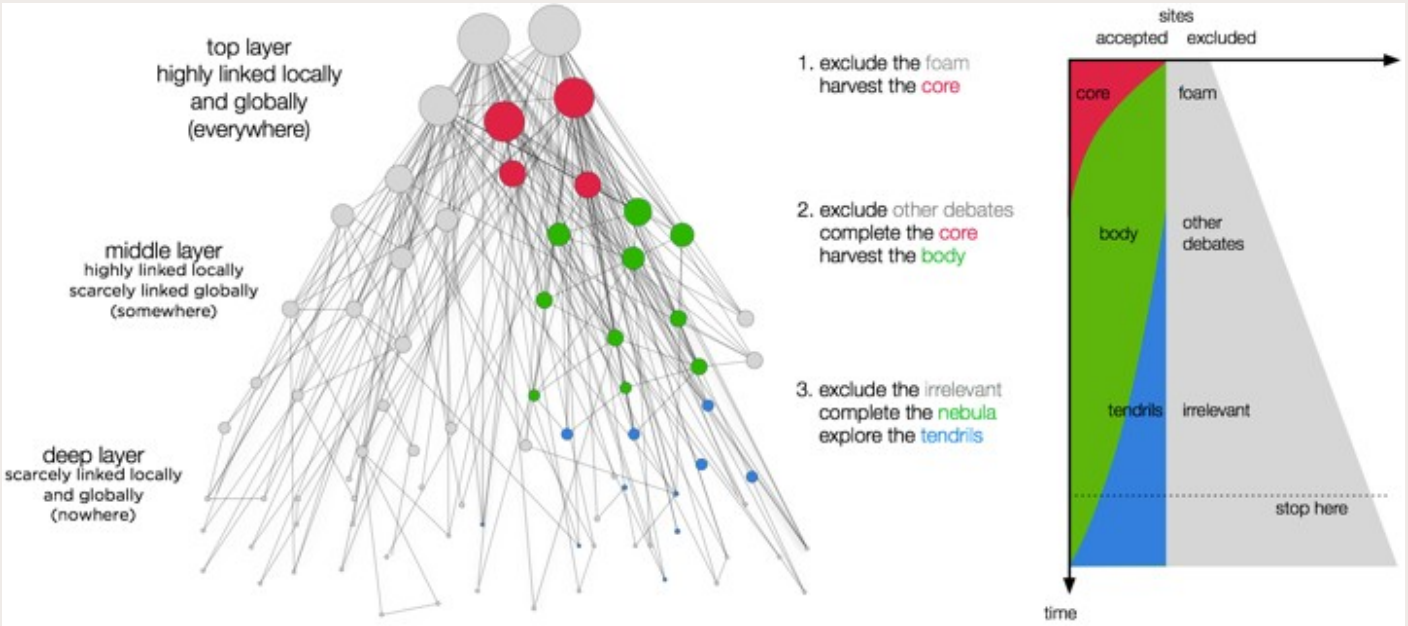
→ Hyphe's Traph



Hyphe's crawling strategy: leverage hyperlinks

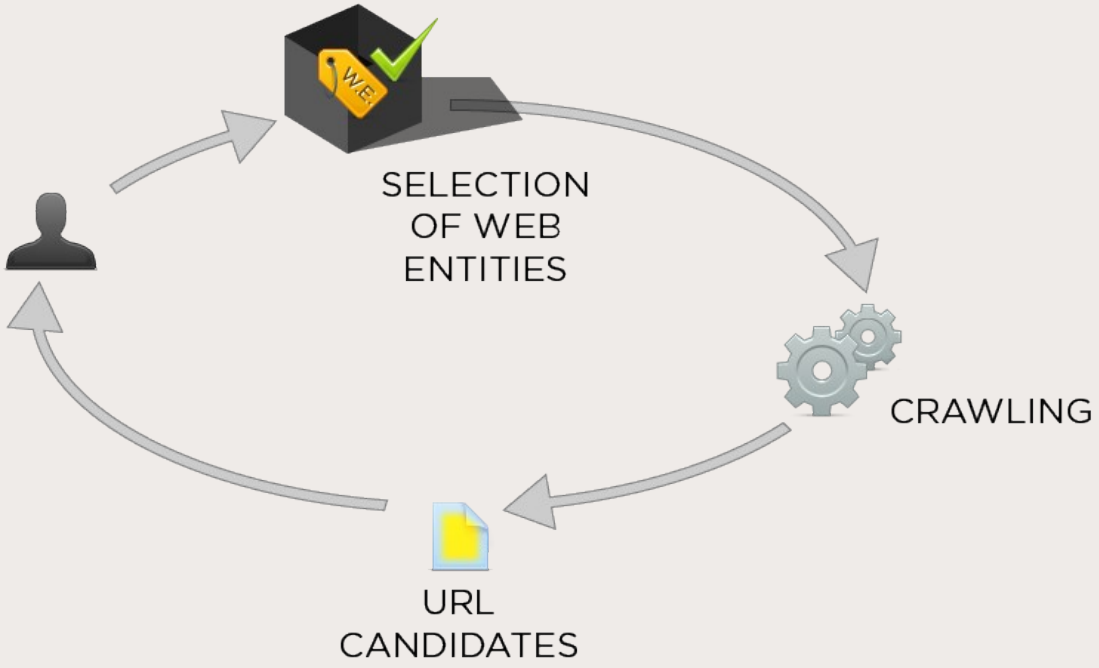
- Classical crawlers (DMI's IssueCrawler): Snowball
  → Top Layer attraction (Google, YouTube, Wikipedia…)
  → Topic drifts

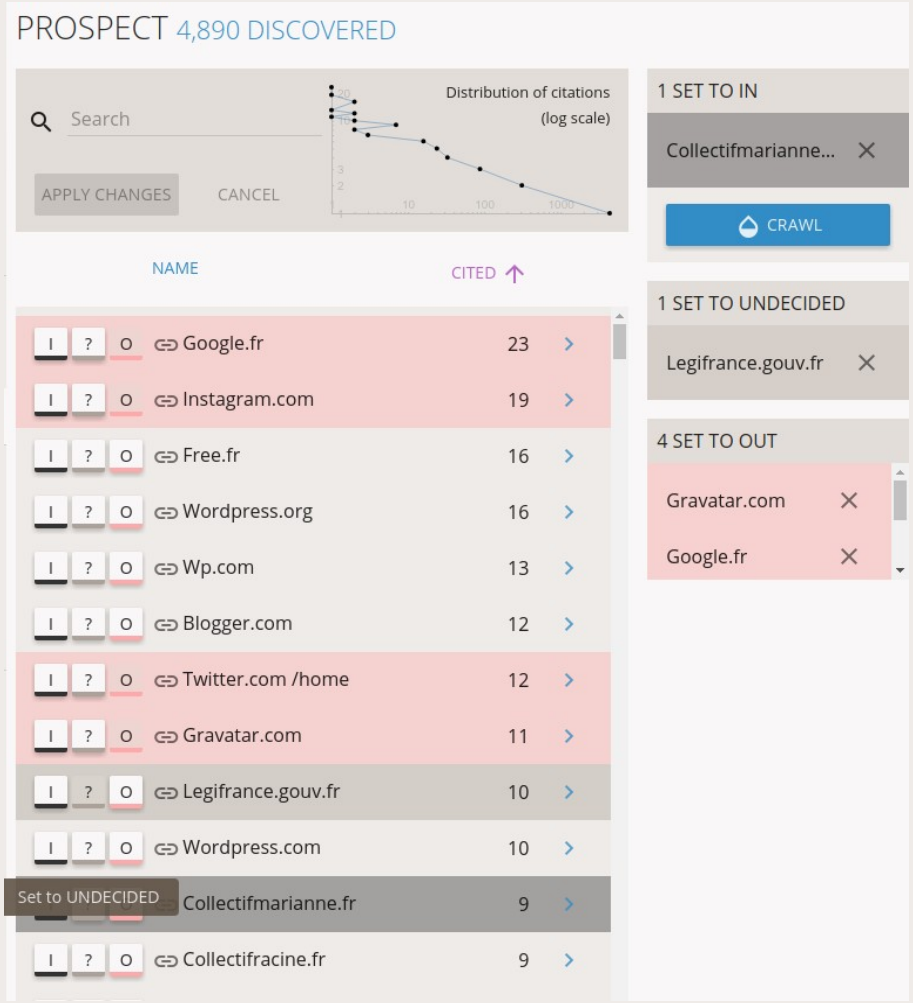
- Hyphe:
  - crawl exclusively pages within the chosen WebEntities
  - sort discovered entities by degree of citation
  - humanly select new entities to include and crawl



Web prospection loop: curating a corpus iteratively

- Step by step iterative expansion & curation of entities



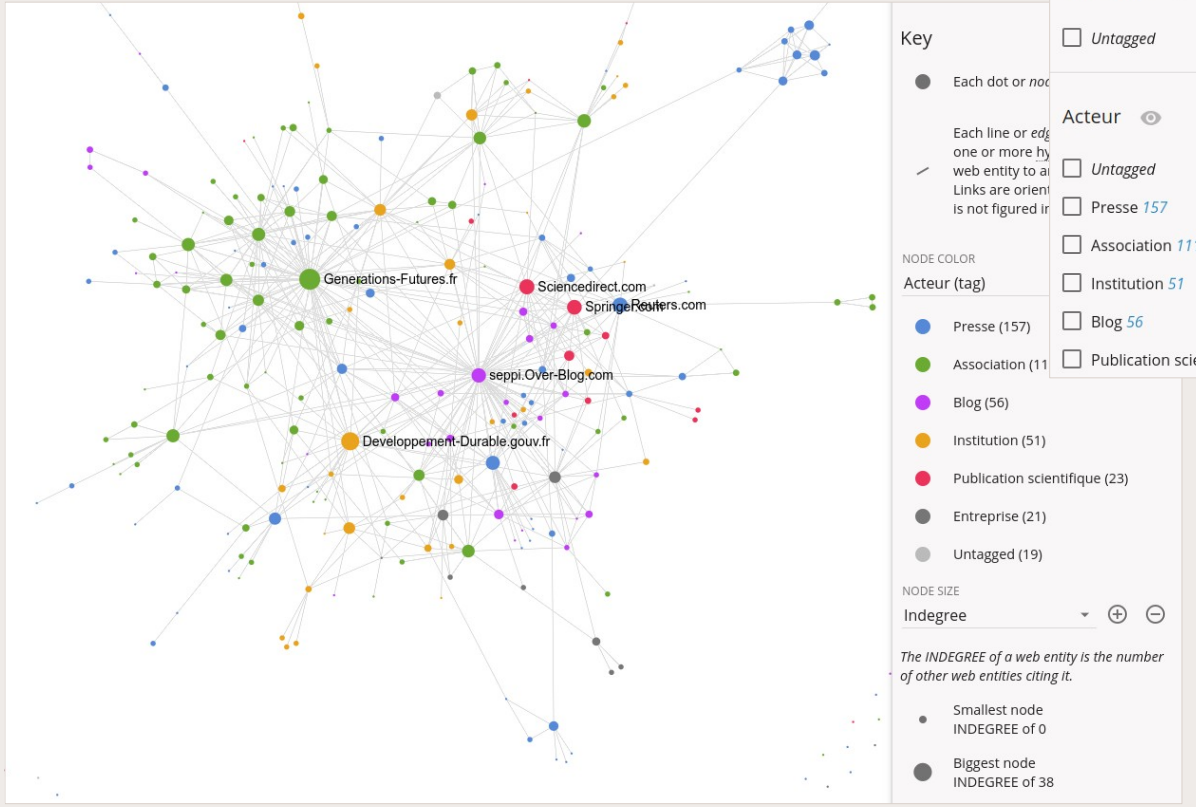


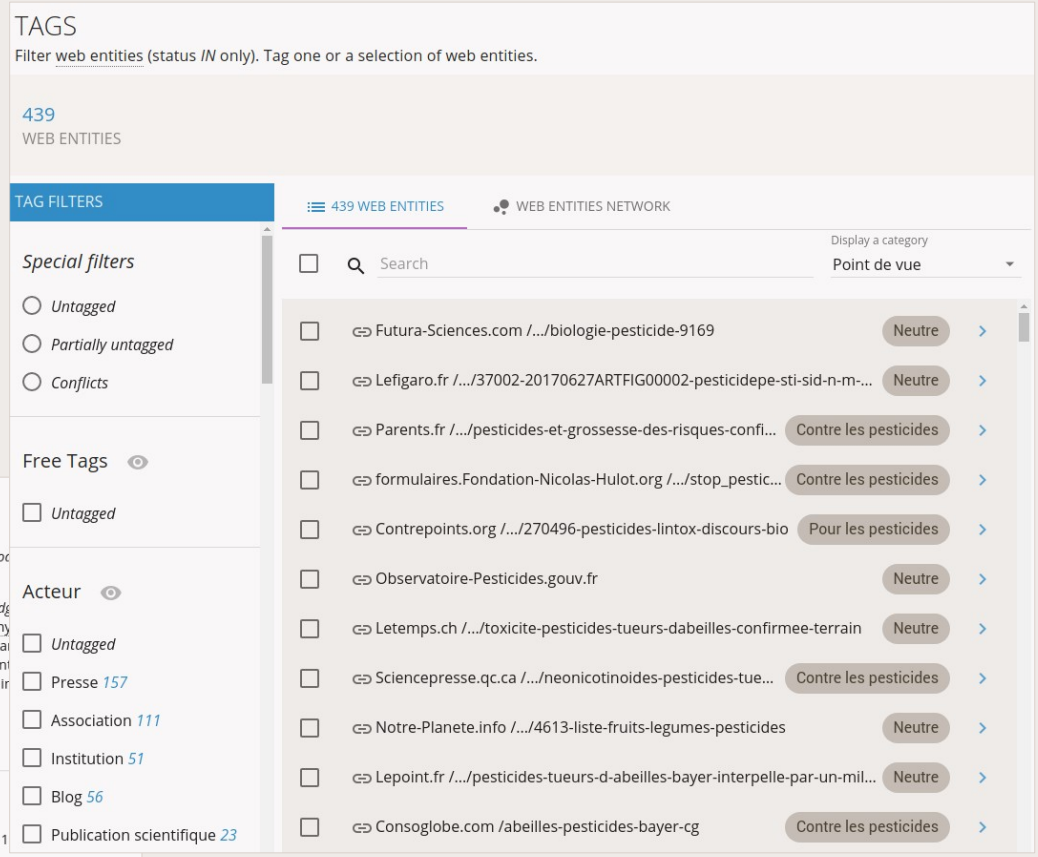
- Human/Time cost
- How to know when to stop?

  → hyperlink citations threshold

Qualify the corpus entities (tagging)

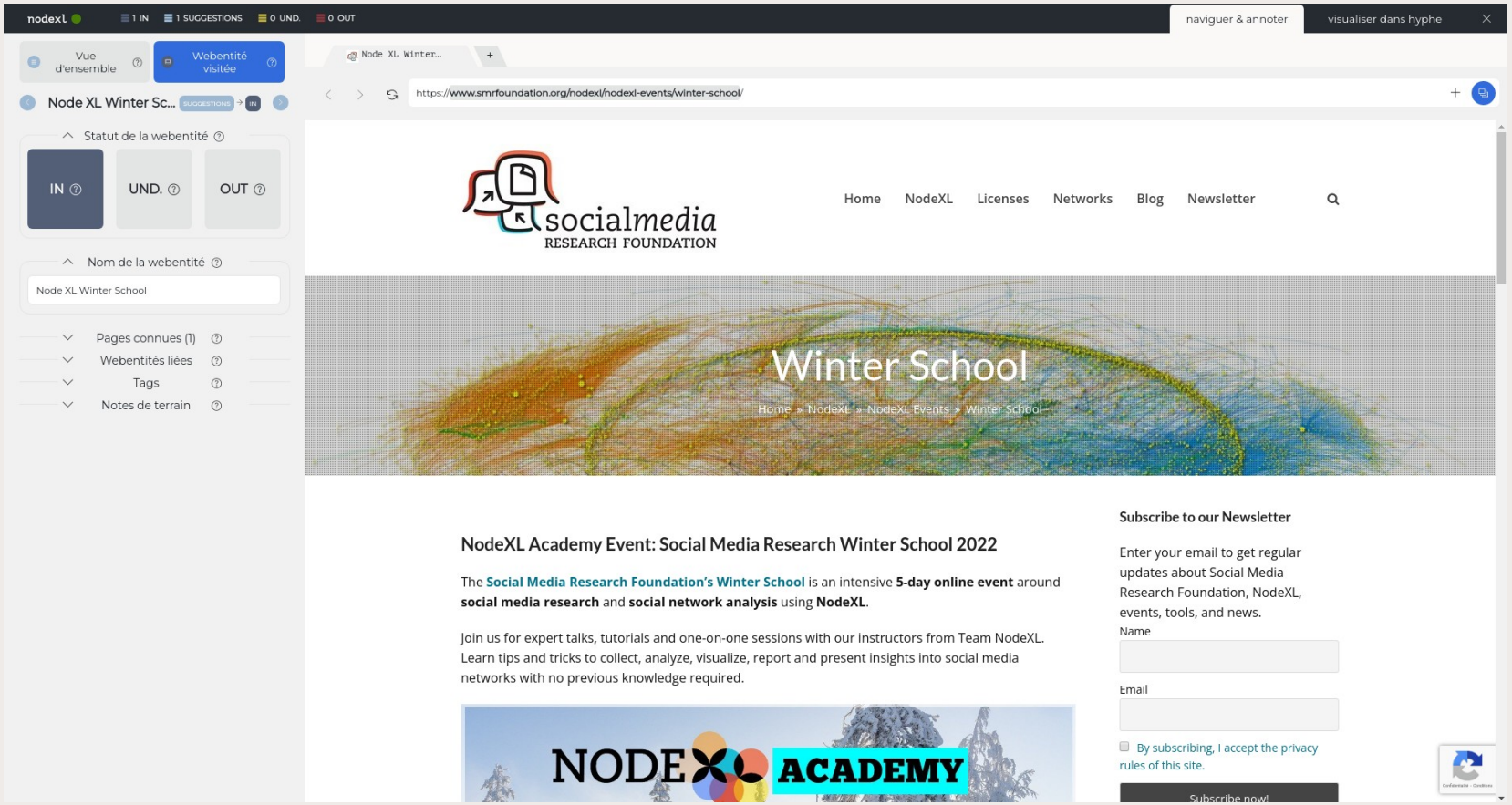
- Free notes
- Categories





HyBro: a web browser designed for corpus curation

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



« NaviCrawler » heritage : build a Hyphe web corpus while browsing

→ « in-situ » prospection & tagging (digital field work)

→ teach the web to students

Many ways to use Hyphe

- Complete methodology includes:

  sourcing, automatized collection, iterative corpus building
  qualitative categorization, exploratory analysis,
  network visualization, quantitative statistical analysis

- Diverse audiences:
  - Research: help social scientists work on digital fields
  - Pedagogy: teach students what the web is beyong Google & Facebook

- Possible small & large scale analyses:
  - a website's internal structure
  - a theme's ensemble of actors and their ties
  - a controversy's alliances & oppositions
  - etc.

SciencesPo
MÉDIALAB

From above: clusters, opposition & affinity

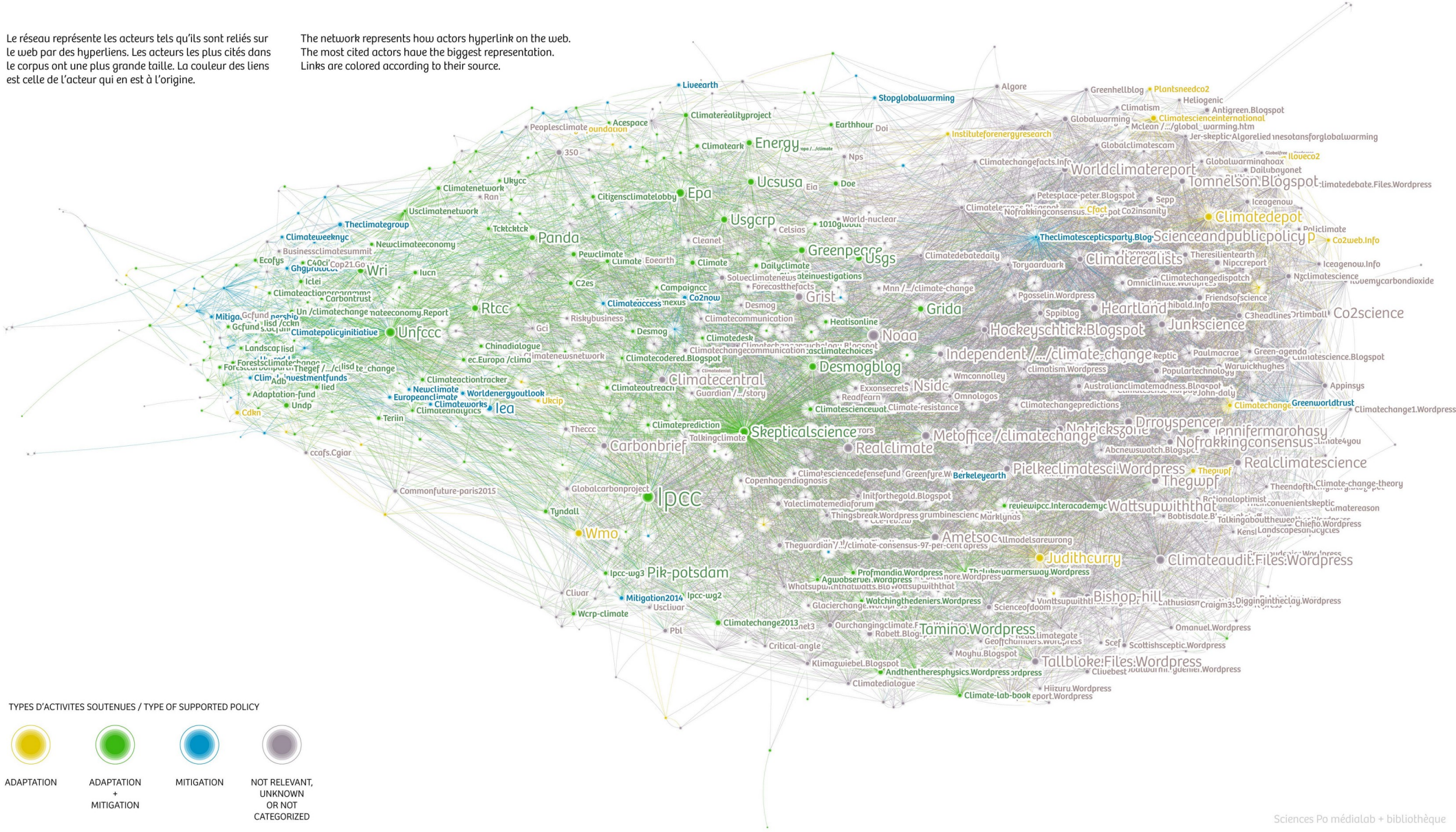
Corpus web sur le changement climatique : types d'activités soutenues

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: type of supported policy

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

Double Dating Data: Climate change on the web



TYPES D'ACTIVITES SOUTENUES / TYPE OF SUPPORTED POLICY

ADAPTATION

ADAPTATION + MITIGATION

MITIGATION

NOT RELEVANT, UNKNOWN OR NOT CATEGORIZED

SciencesPo
MÉDIALAB

From above: clusters, opposition & affinity

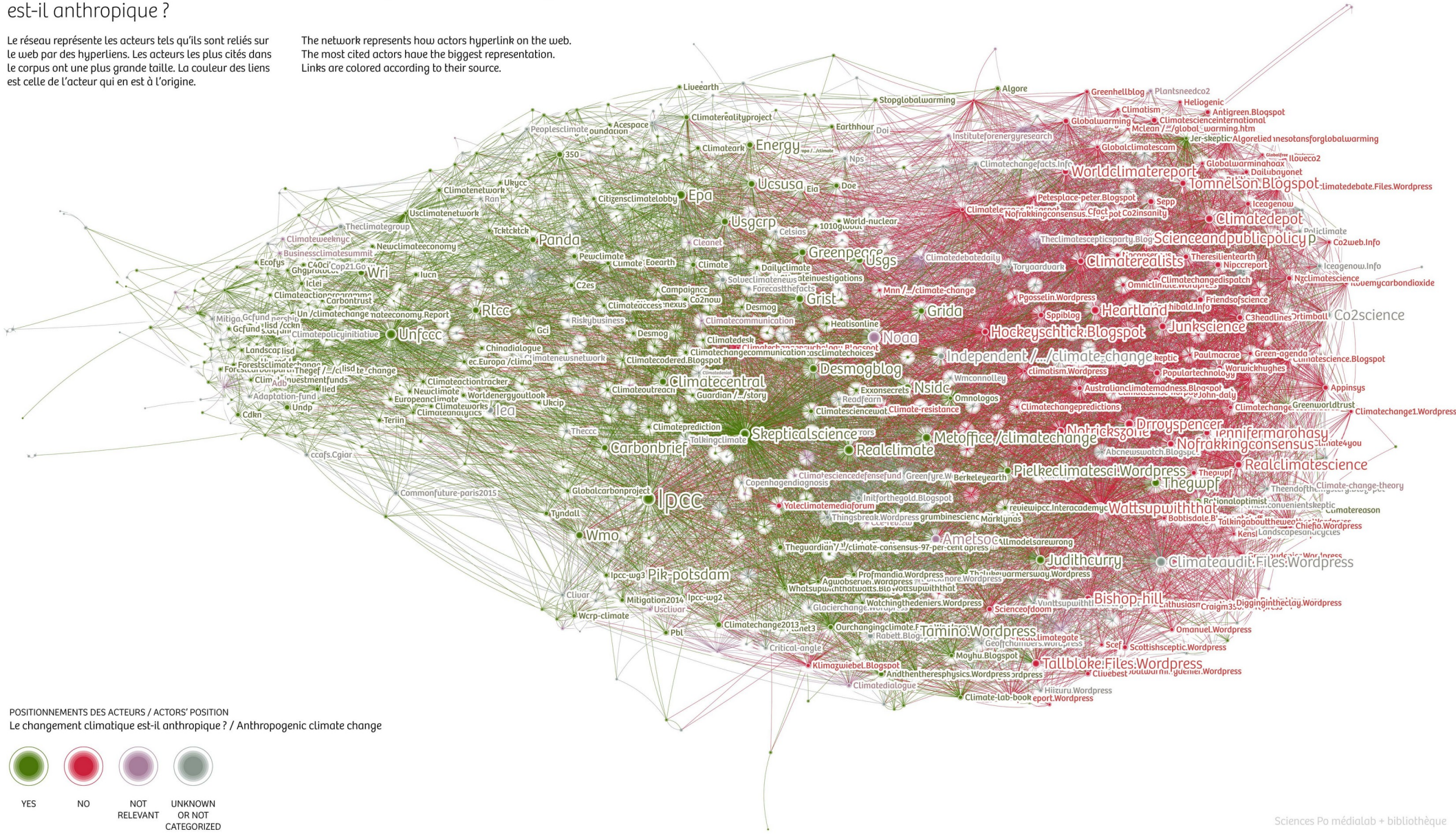
Corpus web sur le changement climatique : le changement climatique est-il anthropique ?

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.

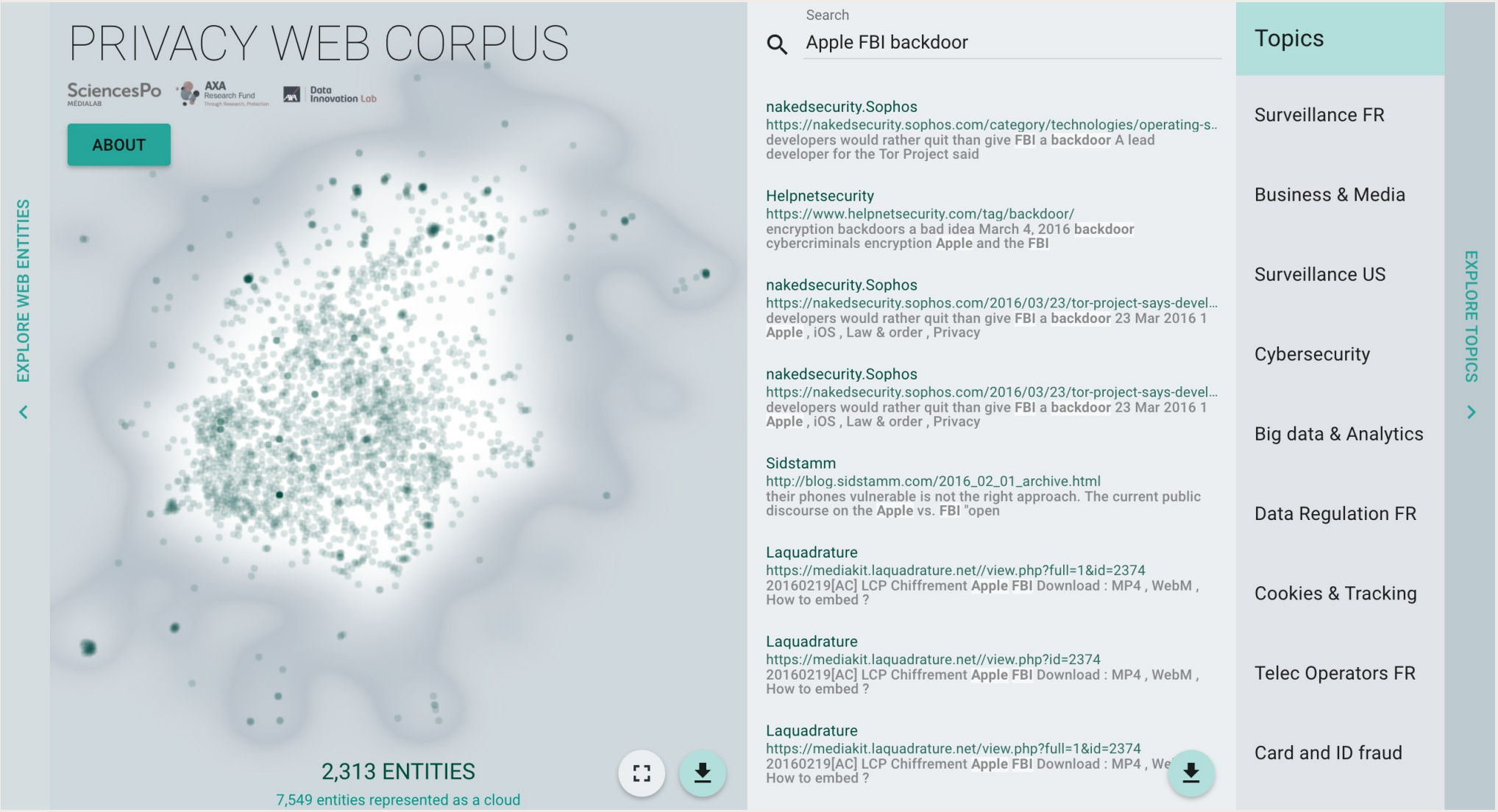
Double Dating Data: Climate change on the web



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

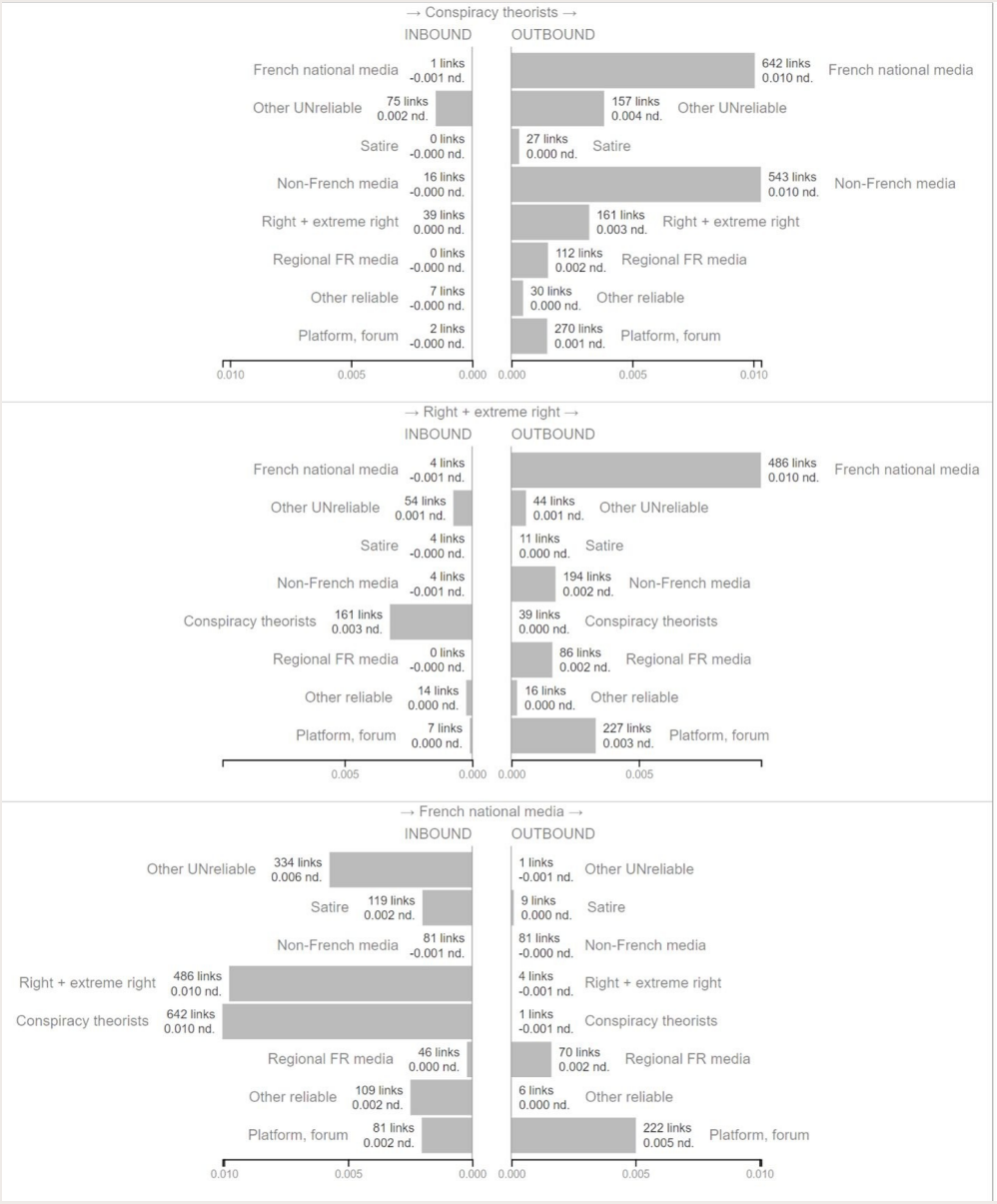
YES | NO | NOT RELEVANT | UNKNOWN OR NOT CATEGORIZED

From within: explore webpages contents



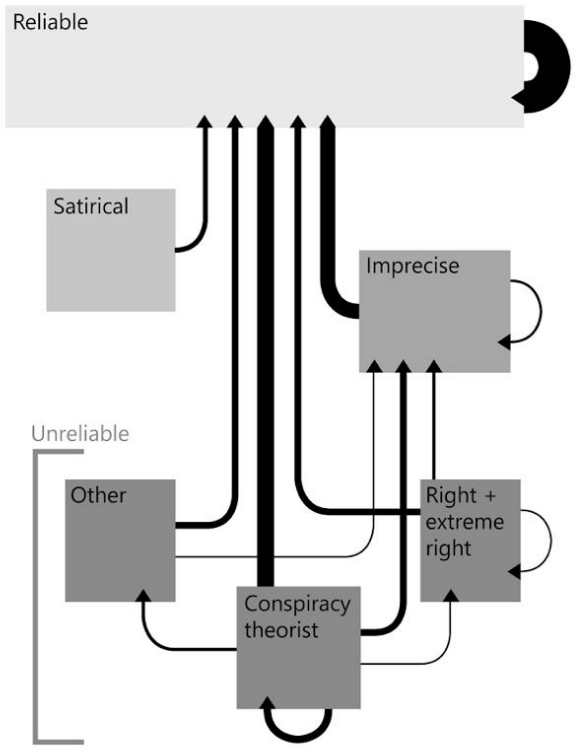
Privacy Web Corpus Datascape

From the sides: a hierarchy of directed hyperlinks



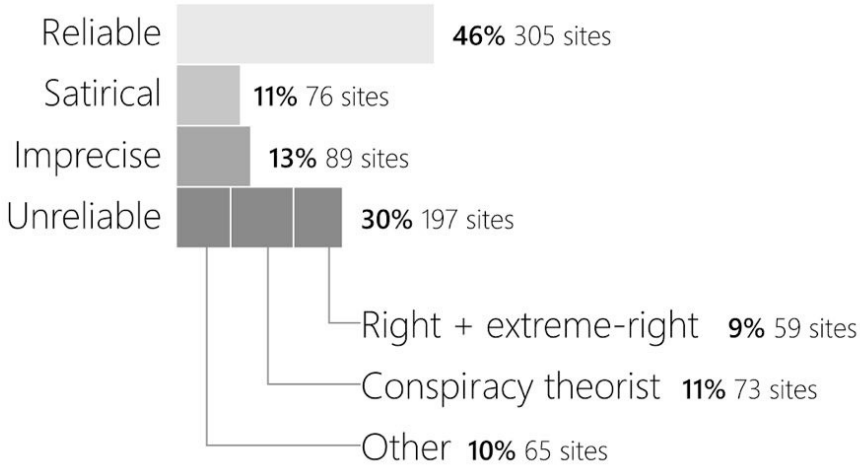
Visual Network Exploration for Data Journalists

Hyperlinks directionality: a bottom-up hierarchy

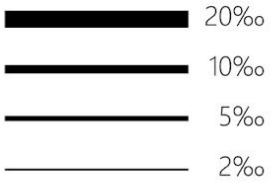


Most hyperlinks stem from the unreliable and aim at the reliable resources

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*

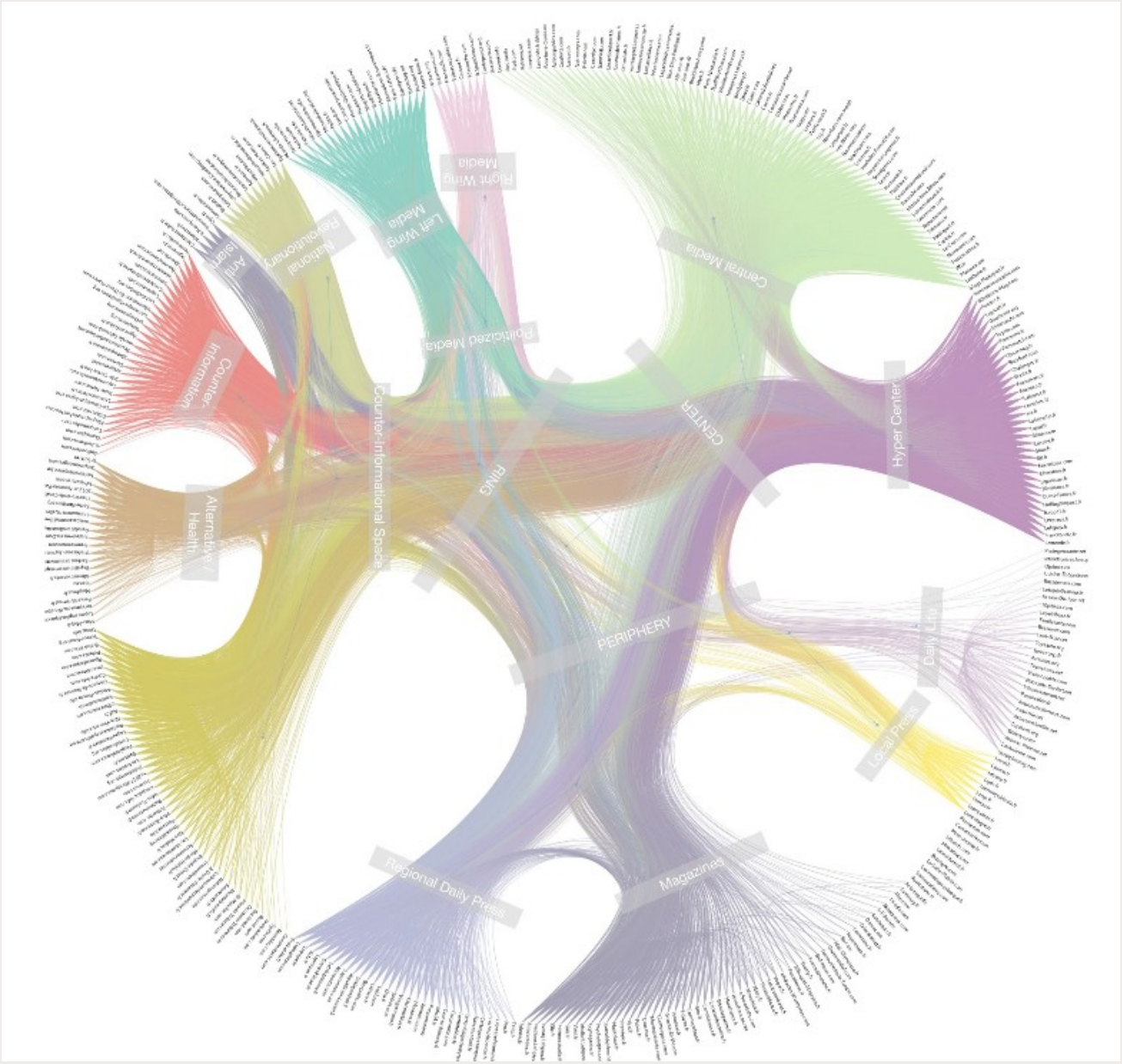


*Accounts for the different group sizes.

Densities below 2‰ have been omitted.

Visual Network Exploration for Data Journalists

Explore polarization dynamics



Unfolding the multi-layered structure of the French Mediascape

Bibliography

Reference publications:

- 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

Now let’s play!

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

https://github.com/medialab/hyphe

Any question first?

benjamin.ooghe@sciencespo.fr

@boogheta **@medialab_ScPo**