



HAL
open science

Web communities network mapping with Hyphe

Benjamin Ooghe-Tabanou

► **To cite this version:**

Benjamin Ooghe-Tabanou. Web communities network mapping with Hyphe. NodeXL Academy: Social Media Research Winter School 2022, Social Media Research Foundation, Jan 2022, online, France. hal-03904193

HAL Id: hal-03904193

<https://sciencespo.hal.science/hal-03904193v1>

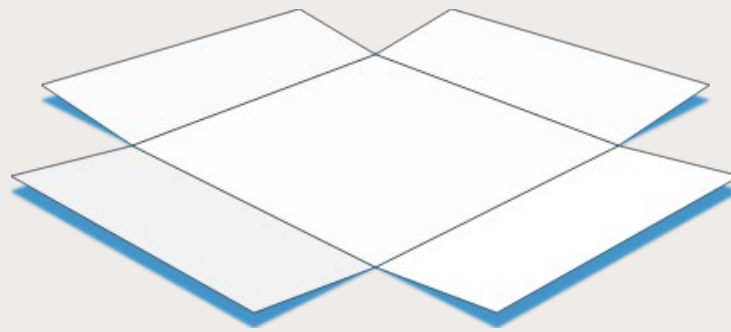
Submitted on 16 Dec 2022

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



Web communities network mapping with Hyphe

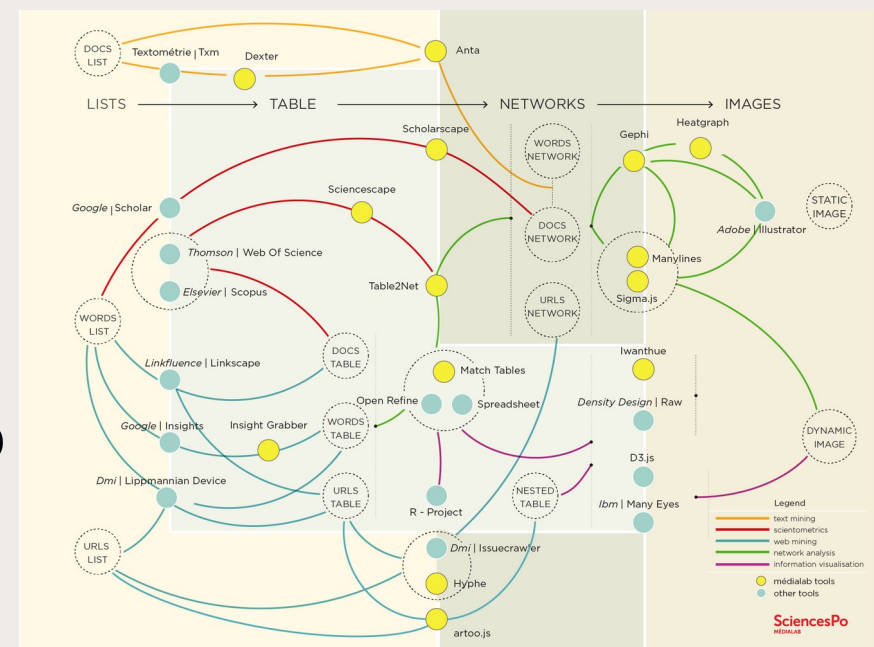
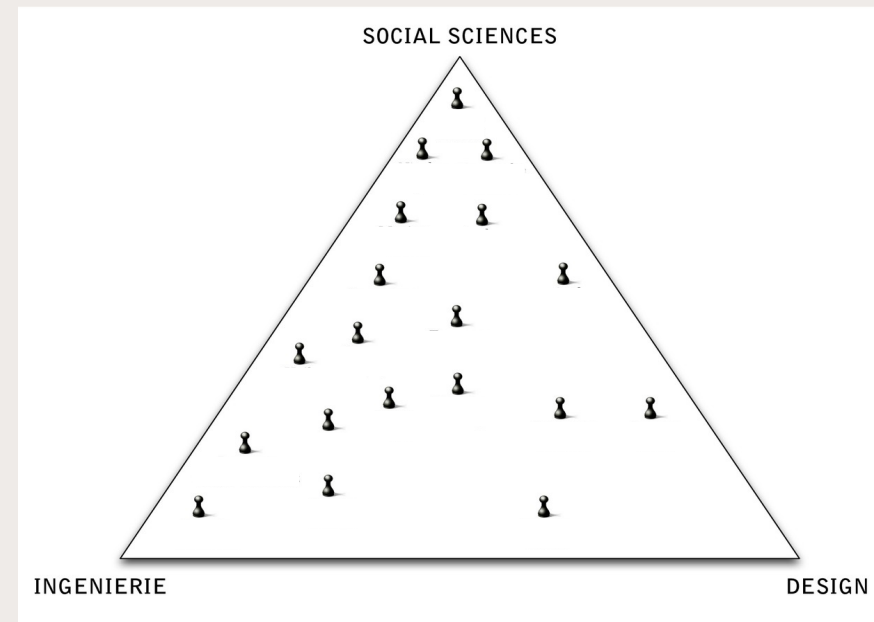
NodeXL Academy '22
January 19th 2022

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

SciencesPo
MÉDIALAB

médialab @ Sciences Po

- 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**.

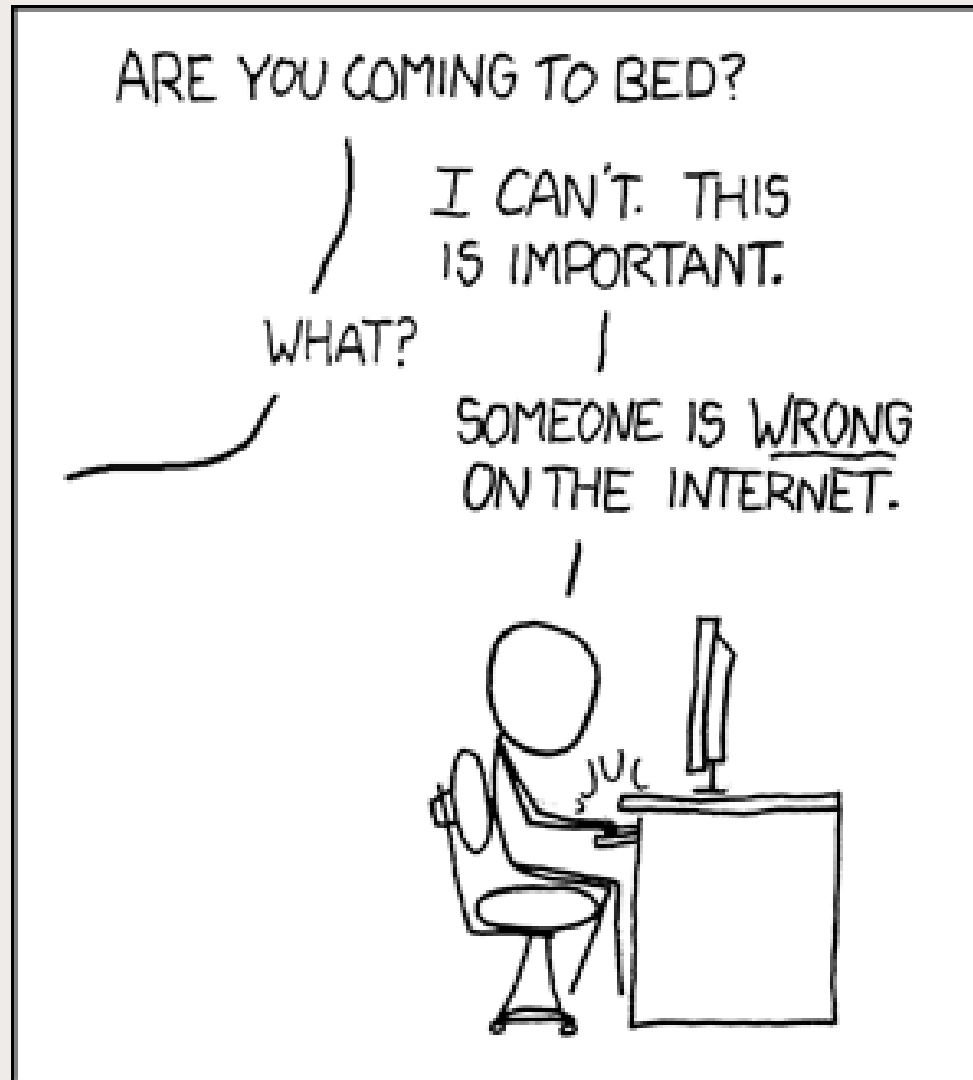
The process of proceeding from node to node is called **navigation**. »

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

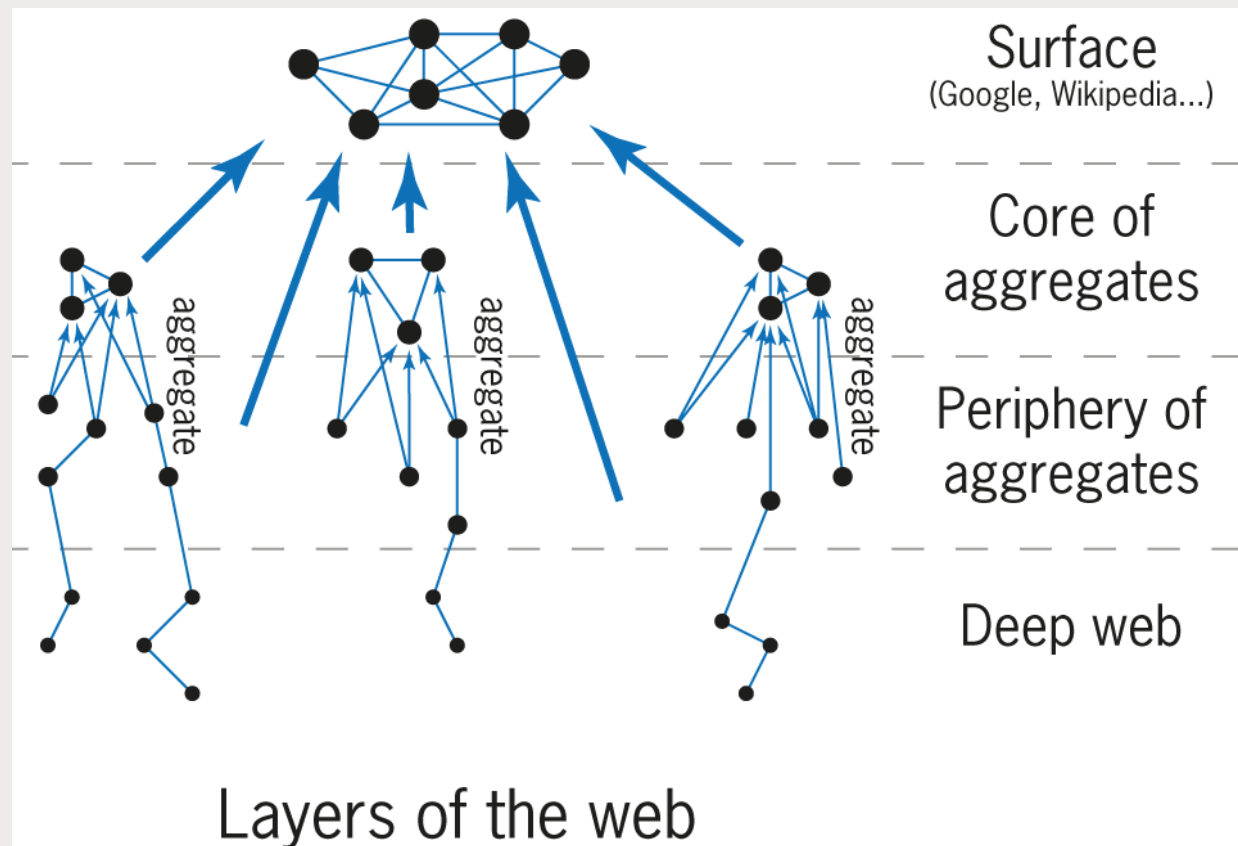


CC-BY-NC - Randall Munroe - XKCD

A bottom-up hierarchy emerged from hyperlinks

« Matthew effect » : preferential attachment

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

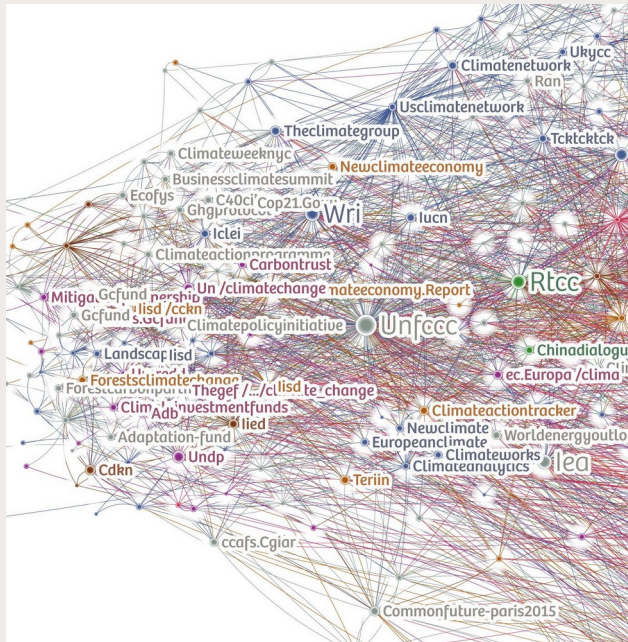


Hyphe: a research directed crawler

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

Build your own web documents corpus
to study social phenomena online

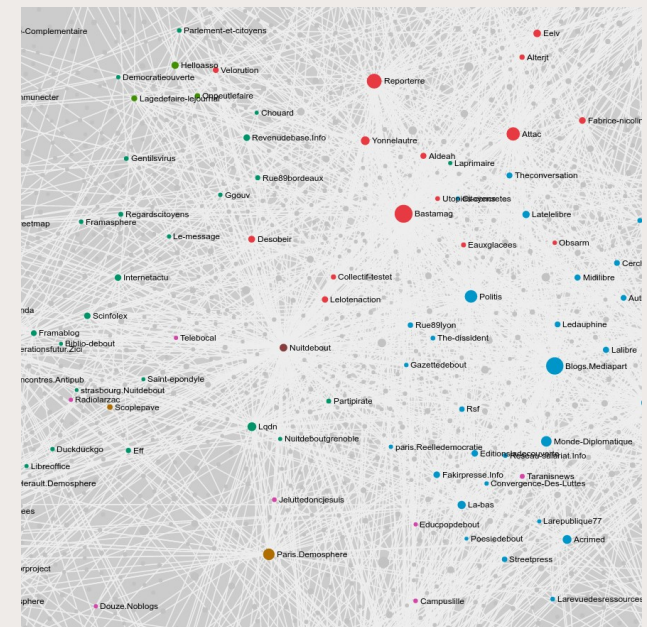
- gather « web actors »
- explore hyperlinks between them



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

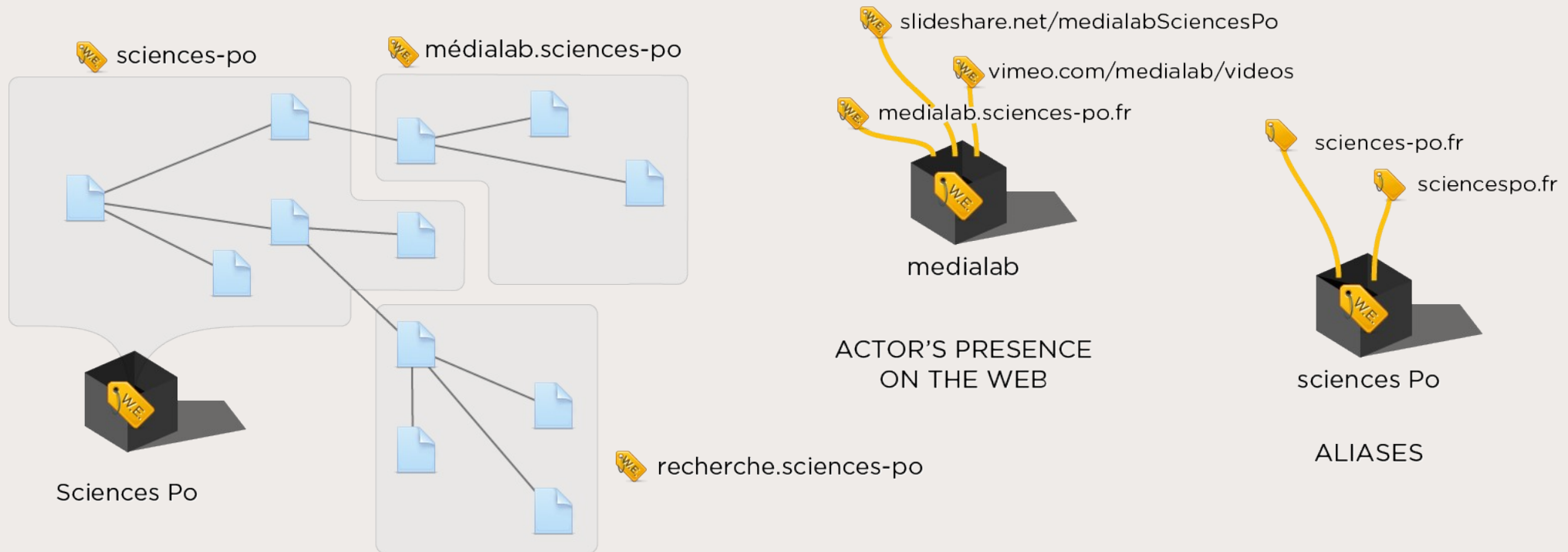
COP 21
Vie privée
Extrême droite
Tissu associatif
Produits laitiers
Cellules souches
Administrations culturelles

...



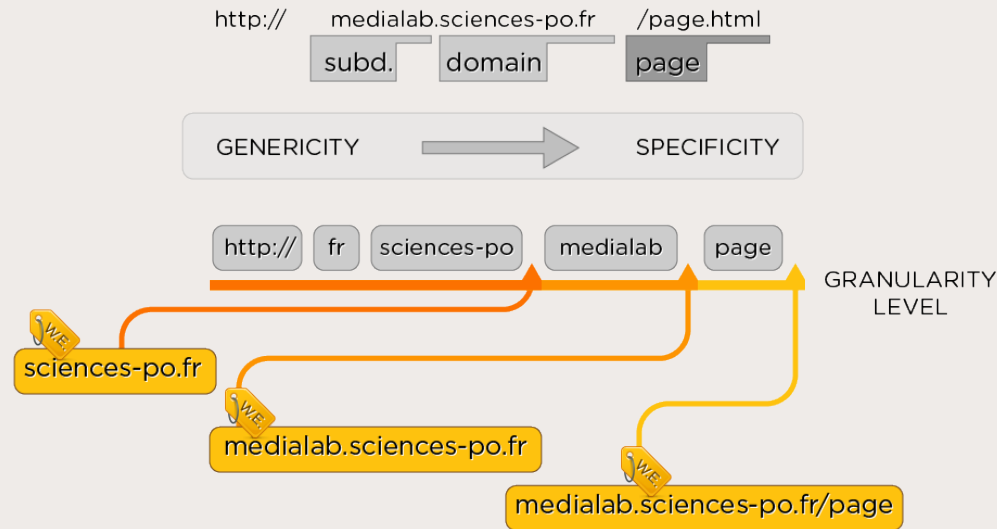
<http://utopies-concretes.org/>

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 »

DEFINE WEB ENTITIES

Check the boundaries of each web entity before creating it

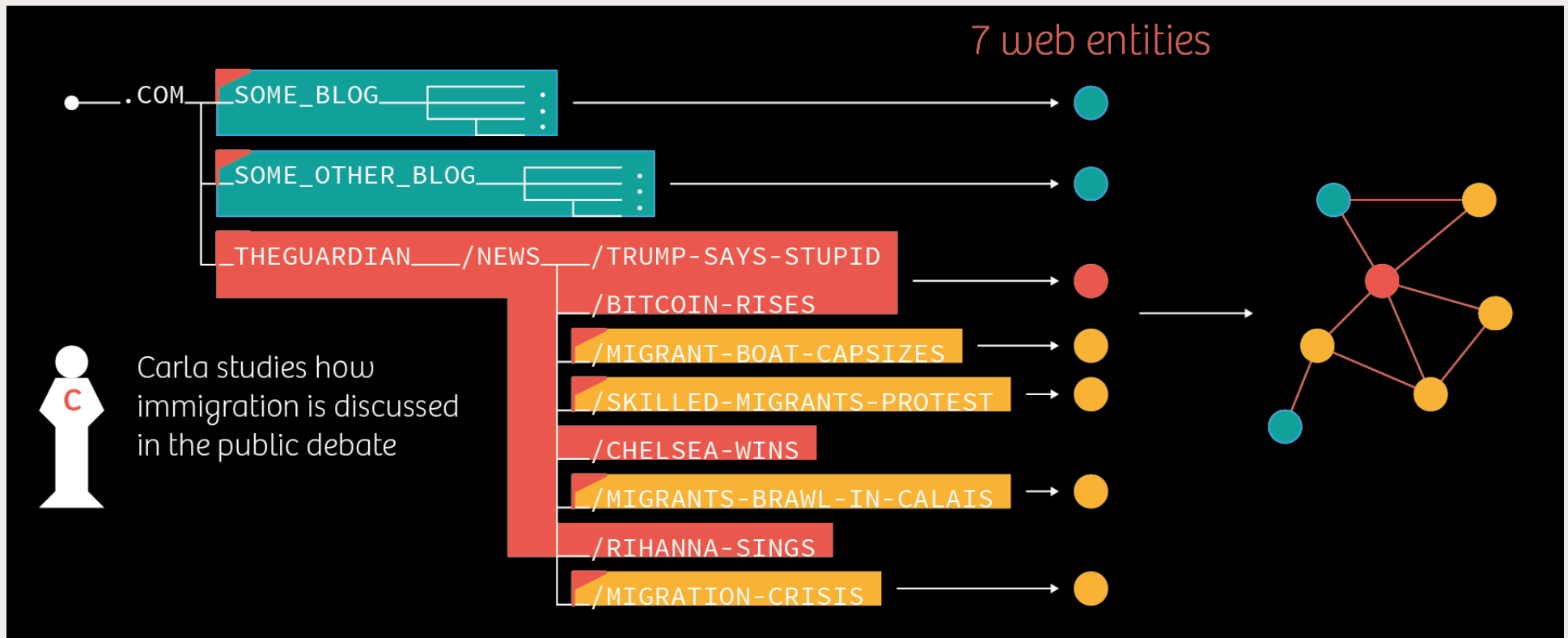
Move all sliders [TO THE LEFT](#) [TO THE RIGHT](#)

1	medialab.Sciences-Po.fr	http .fr sciences-po medialab.
2	tools.medialab.Sciences-Po.fr	http .fr sciences-po medialab. tools.
3	Sciences-Po.fr	https .fr sciences-po www.
4	Sciencespo.fr/bibliotheque	http .fr sciencespo www. /bibliotheque
5	Twitter.com /medialab_ScPo	https .com twitter /medialab_ScPo

Which data structure to manage hypertexts?

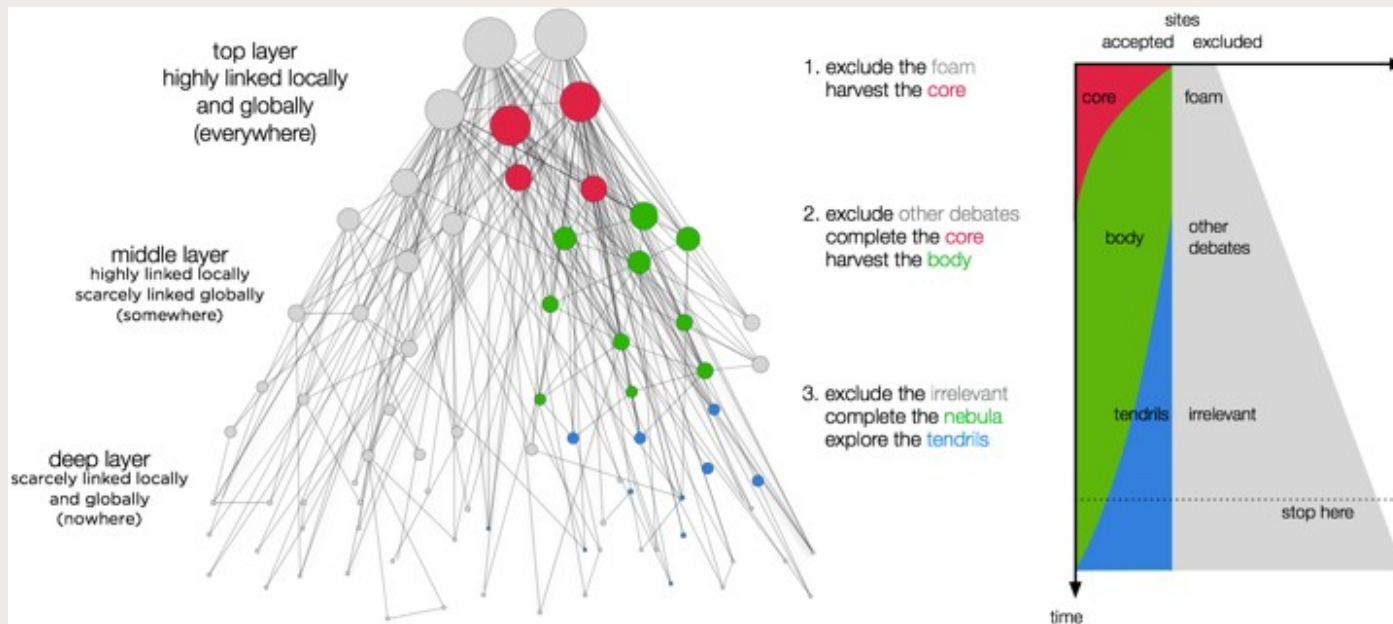
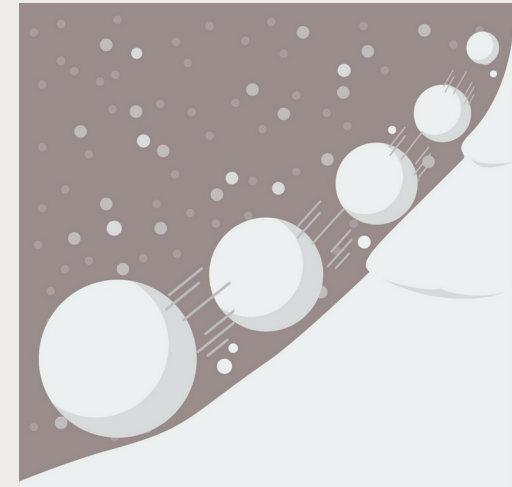
<https://medialab.github.io/hyphe-traph/fosdem2018/#/>

- **Tree** of urls
 - **Graph** of hyperlinks
 - Dynamic branches aggregates
- Hyphe's Traph



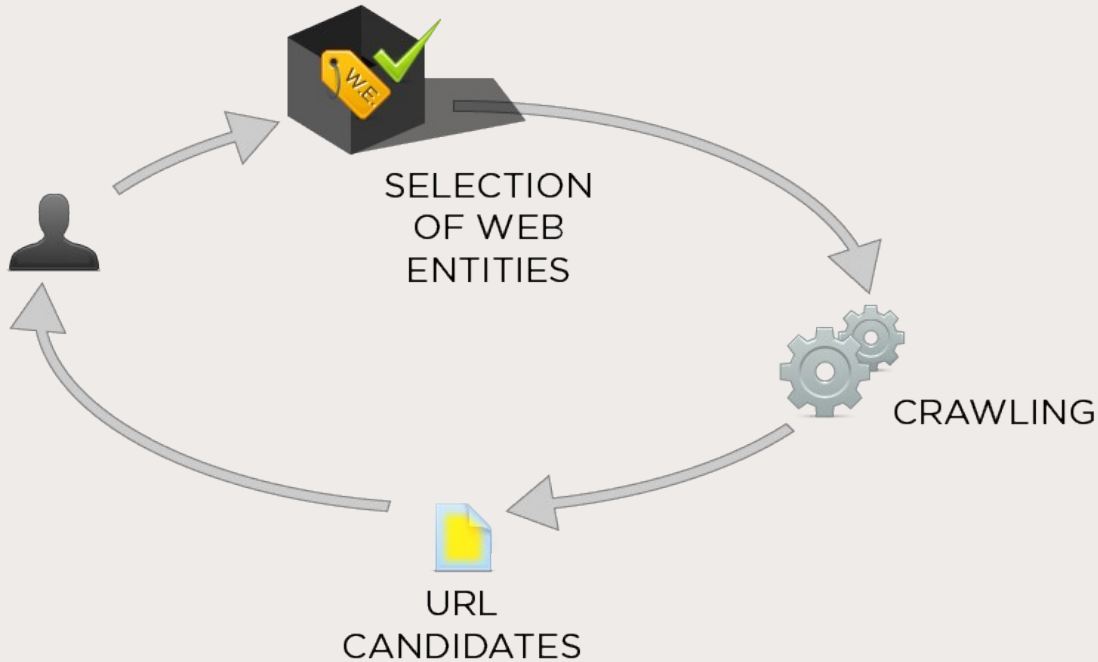
Hyphe's crawling strategy: leverage hyperlinks

- 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

PROSPECT 4,890 DISCOVERED

Search APPLY CHANGES CANCEL

Distribution of citations (log scale)

NAME	CITED ↑
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Google.fr	23
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Instagram.com	19
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Free.fr	16
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Wordpress.org	16
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Wp.com	13
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Blogger.com	12
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Twitter.com /home	12
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Gravatar.com	11
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Legifrance.gouv.fr	10
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Wordpress.com	10
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Collectifmarianne.fr	9
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Collectifracine.fr	9

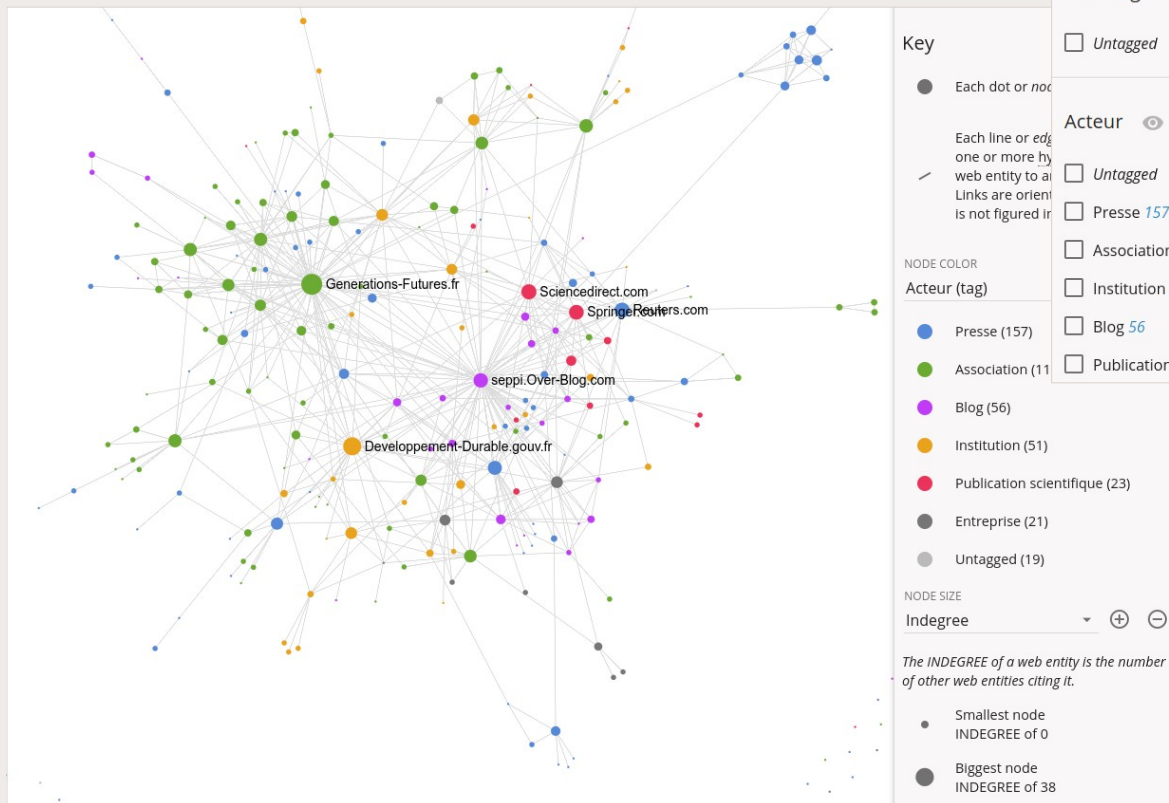
1 SET TO IN
Collectifmarianne... X

1 SET TO UNDECIDED
Legifrance.gouv.fr X

4 SET TO OUT
Gravatar.com X
Google.fr X

Qualify the corpus entities (tagging)

- Free notes
- Categories



TAGS

Filter web entities (status *IN* only). Tag one or a selection of web entities.

439
WEB ENTITIES

TAG FILTERS 439 WEB ENTITIES WEB ENTITIES NETWORK

Special filters

- Untagged
- Partially untagged
- Conflicts

Free Tags

- Untagged
- Acteur
- Untagged
- Presse 157
- Association 111
- Institution 51
- Blog 56
- Publication scientifique 23

Display a category
Point de vue

Search

- Futura-Sciences.com /.../biologie-pesticide-9169 Neutre
- Lefigaro.fr /.../37002-20170627ARTFIG00002-pesticidepe-sti-sid-n-m... Neutre
- Parents.fr /.../pesticides-et-grossesse-des-risques-confi... Contre les pesticides
- formulaires.Fondation-Nicolas-Hulot.org /.../stop_pestic... Contre les pesticides
- Contrepoints.org /.../270496-pesticides-lintox-discours-bio Pour les pesticides
- Observatoire-Pesticides.gouv.fr Neutre
- Letemps.ch /.../toxicite-pesticides-tueurs-dabeilles-confirmee-terrain Neutre
- Sciencepresse.qc.ca /.../neonicotinoides-pesticides-tue... Contre les pesticides
- Notre-Planete.info /.../4613-liste-fruits-legumes-pesticides Neutre
- Lepoint.fr /.../pesticides-tueurs-d-abeilles-bayer-interpelle-par-un-mil... Neutre
- Consoglobe.com /abeilles-pesticides-bayer-cg Contre les pesticides

HyBro: a web browser designed for corpus curation

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

The screenshot shows the HyBro web browser interface. The browser's address bar displays the URL <https://www.smrfoundation.org/nodexl/nodexl-events/winter-school/>. The page content includes the Social Media Research Foundation logo, a navigation menu (Home, NodeXL, Licenses, Networks, Blog, Newsletter), and a large banner for the 'Winter School' event. Below the banner, there is a section titled 'NodeXL Academy Event: Social Media Research Winter School 2022' with a description of the event and a 'Subscribe to our Newsletter' form. The sidebar on the left contains various controls for corpus curation, including 'Statut de la webentité' (IN, UND, OUT) and 'Nom de la webentité' (NodeXL Winter School).

- « 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 beyond 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.

From above: clusters, opposition & affinity

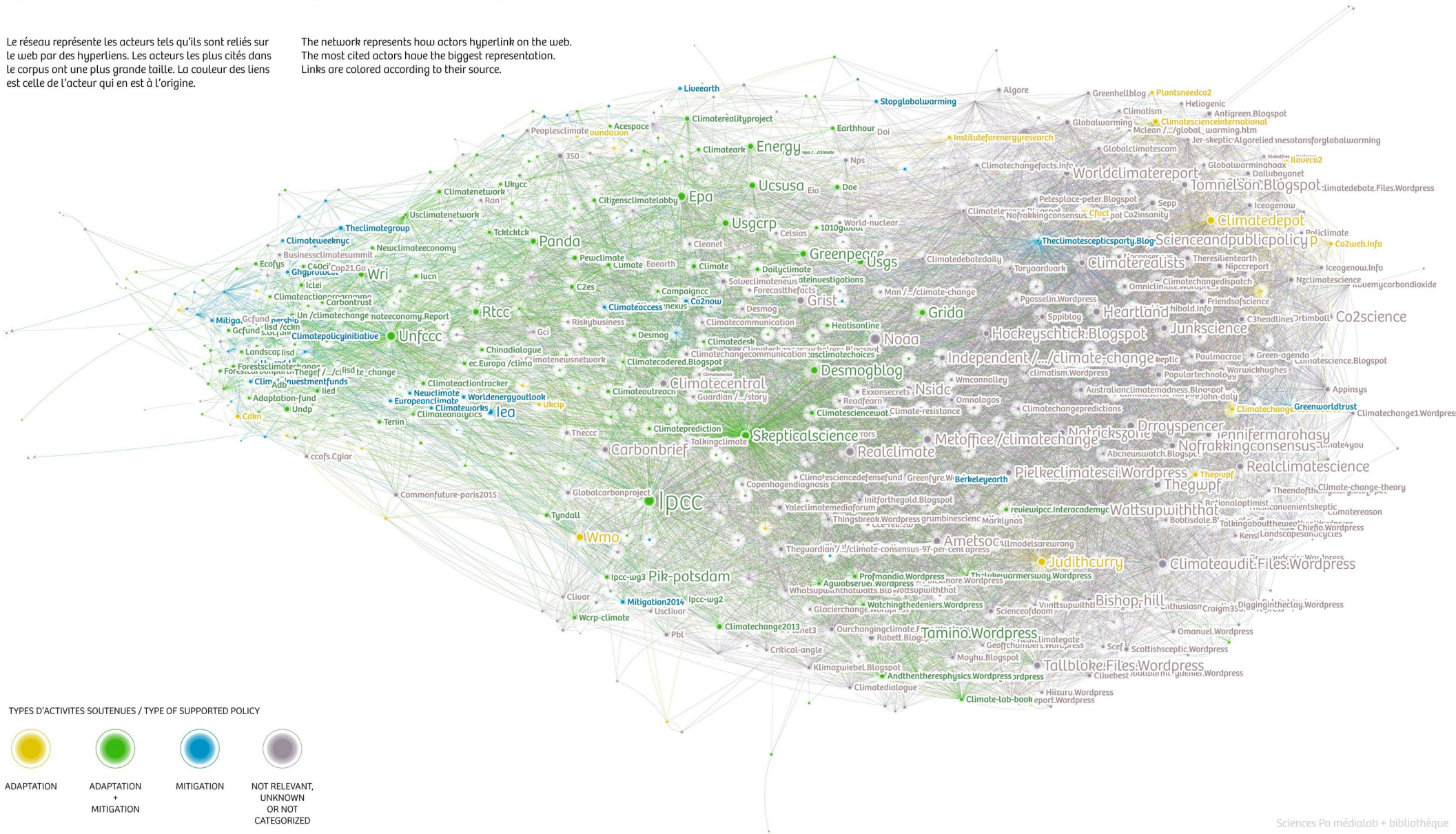
Corpus web sur le changement climatique : types d'activités soutenues

Web corpus on climate change: type of supported policy

Double Dating Data: Climate change on the web

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

From above: clusters, opposition & affinity

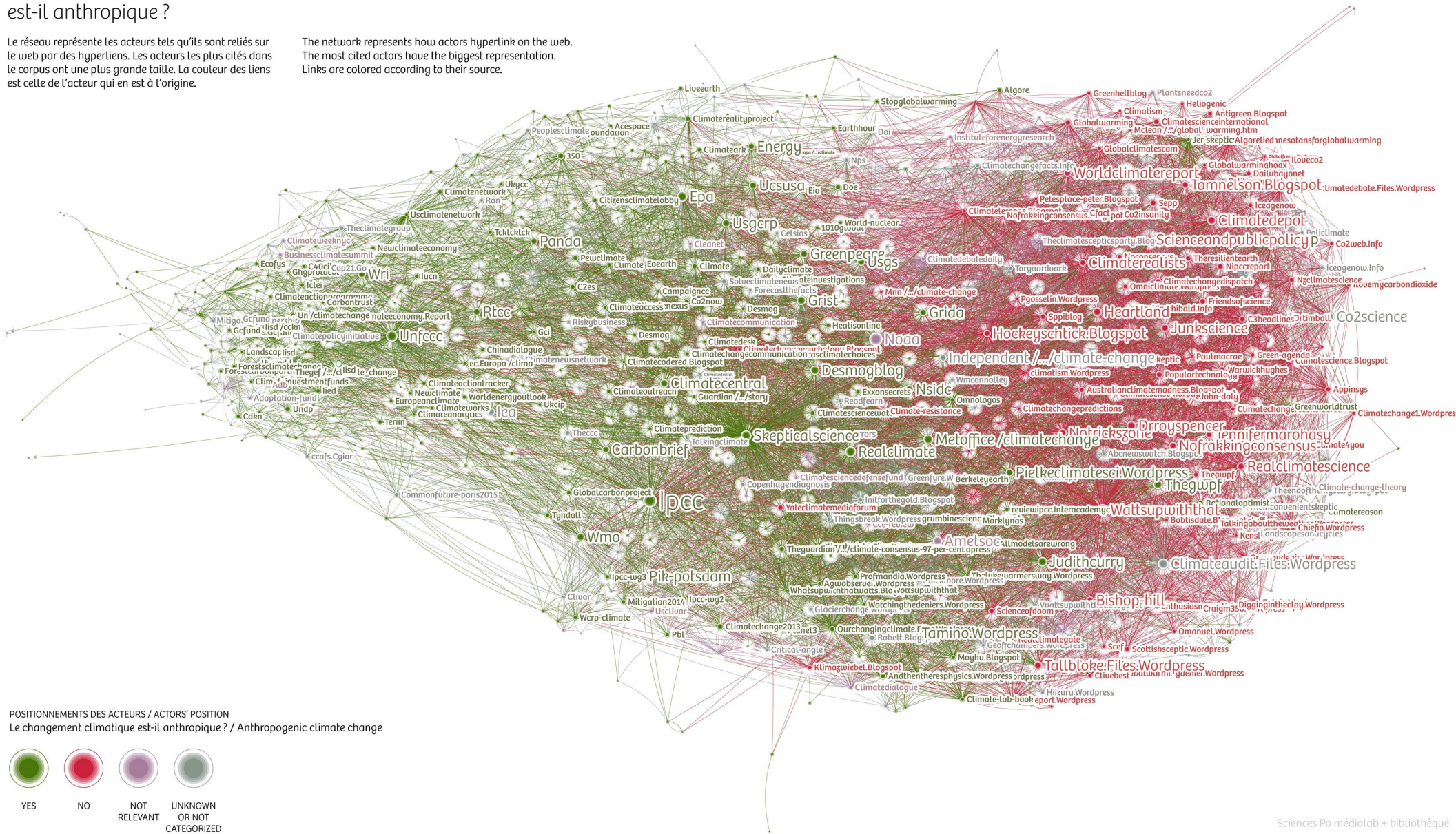
Corpus web sur le changement climatique : le changement climatique est-il anthropique ?

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



From within: explore webpages contents

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 ?

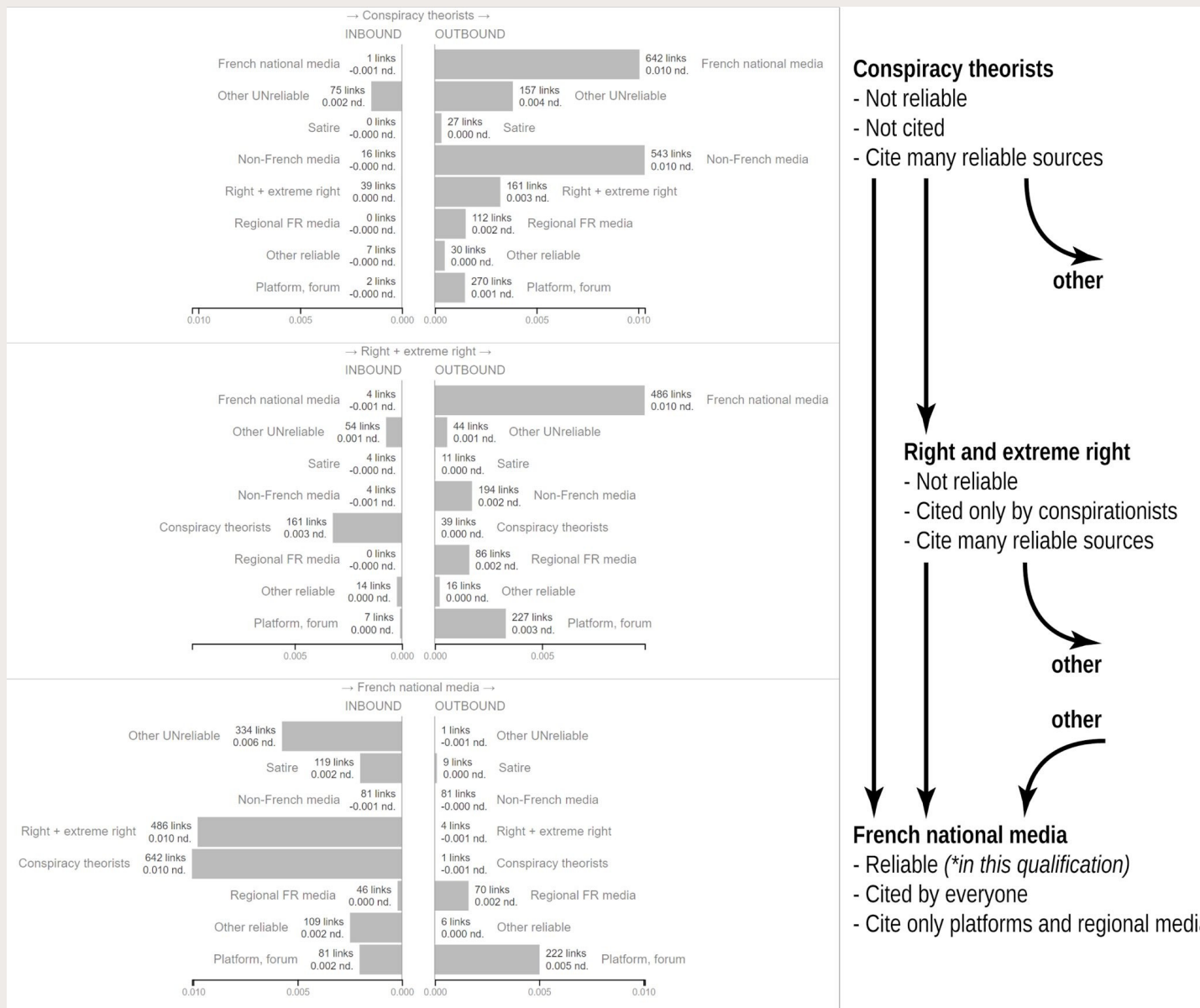
Topics

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

EXPLORE TOPICS

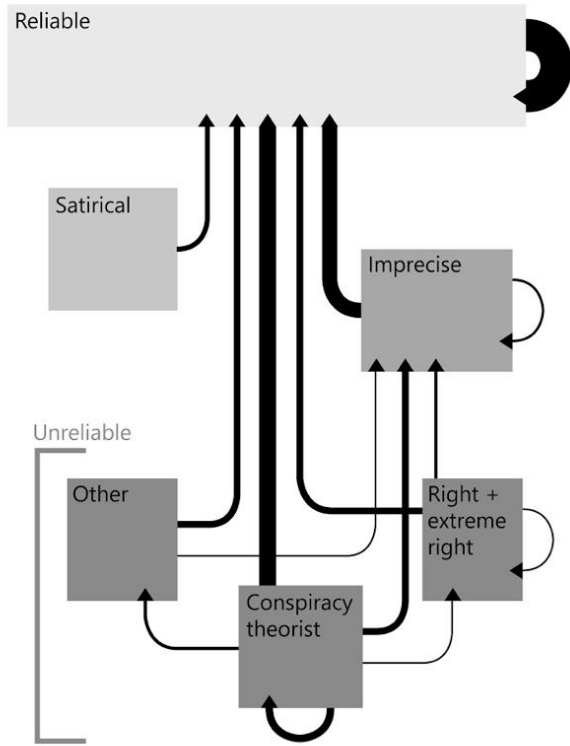
Privacy Web Corpus Datascape

From the sides: a hierarchy of directed hyperlinks



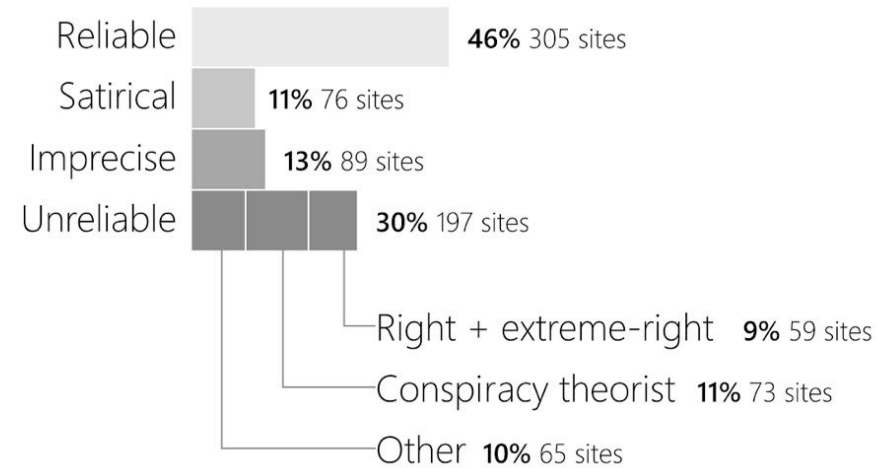
Visual Network Exploration for Data Journalists

Hyperlinks directionality: a bottom-up hierarchy

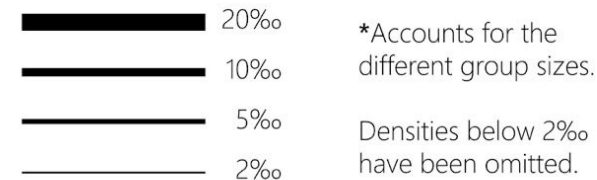


Most hyperlinks stem from the unreliable and aim at the reliable resources

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

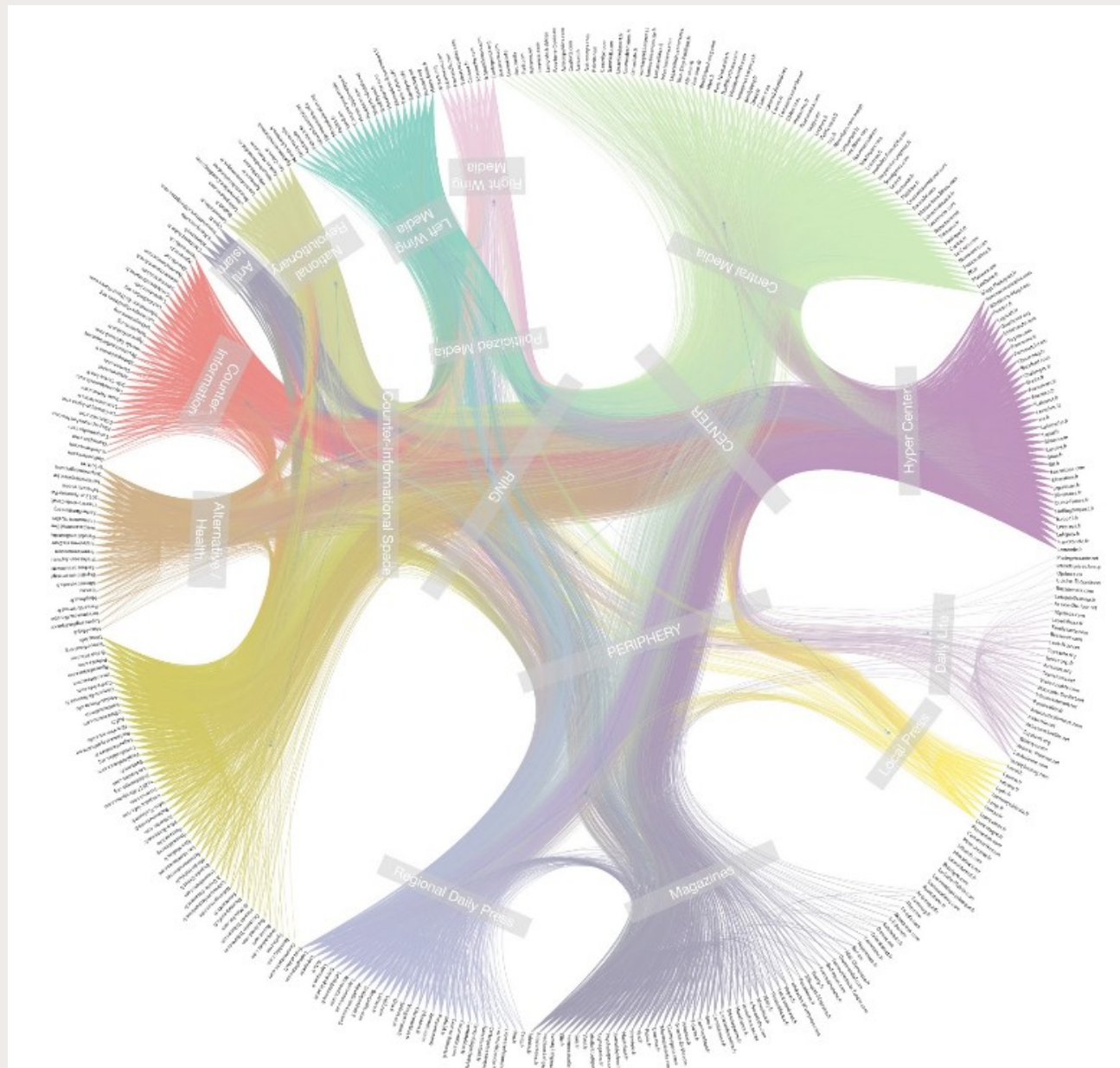


The thickness is proportional to the normalized link density*



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](#)