

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/>

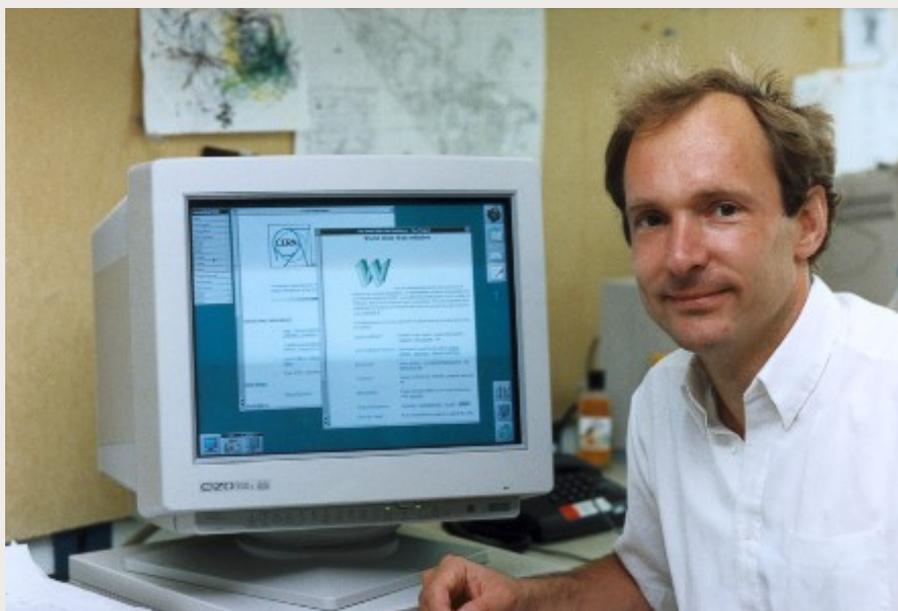
Des petits boutons installables simplement dans les favoris du navigateur pour exporter simplement en tableur des résultats d'une recherche Google

The image illustrates the workflow of using Google bookmarklets for targeted scraping. It consists of several key components:

- Installation Page:** A page titled "Install Google Bookmarklets" with instructions to "Drag & drop images below into your bookmark bar:" and two bookmarklet icons (a blue 'G' and a colorful 'G').
- Search Page:** A Google search page for "digital humanities" showing search results and navigation options.
- Redirect Dialog:** A "Redirect to Classic Google" dialog box with a language dropdown set to "en", a "How many results per page?" dropdown set to "100", and a "Redirect me!" button.
- Extract Dialog:** An "Extract Classic Google Results" dialog box showing search parameters: "Search for 'digital humanities' page 0 (with up to 100 urls per page)", "103 new results in this page", and buttons for "Keep existing results & continue to the next page" and "Download CSV with 103 urls".
- Output Format:** A black box with green text showing the output format: `→ 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*

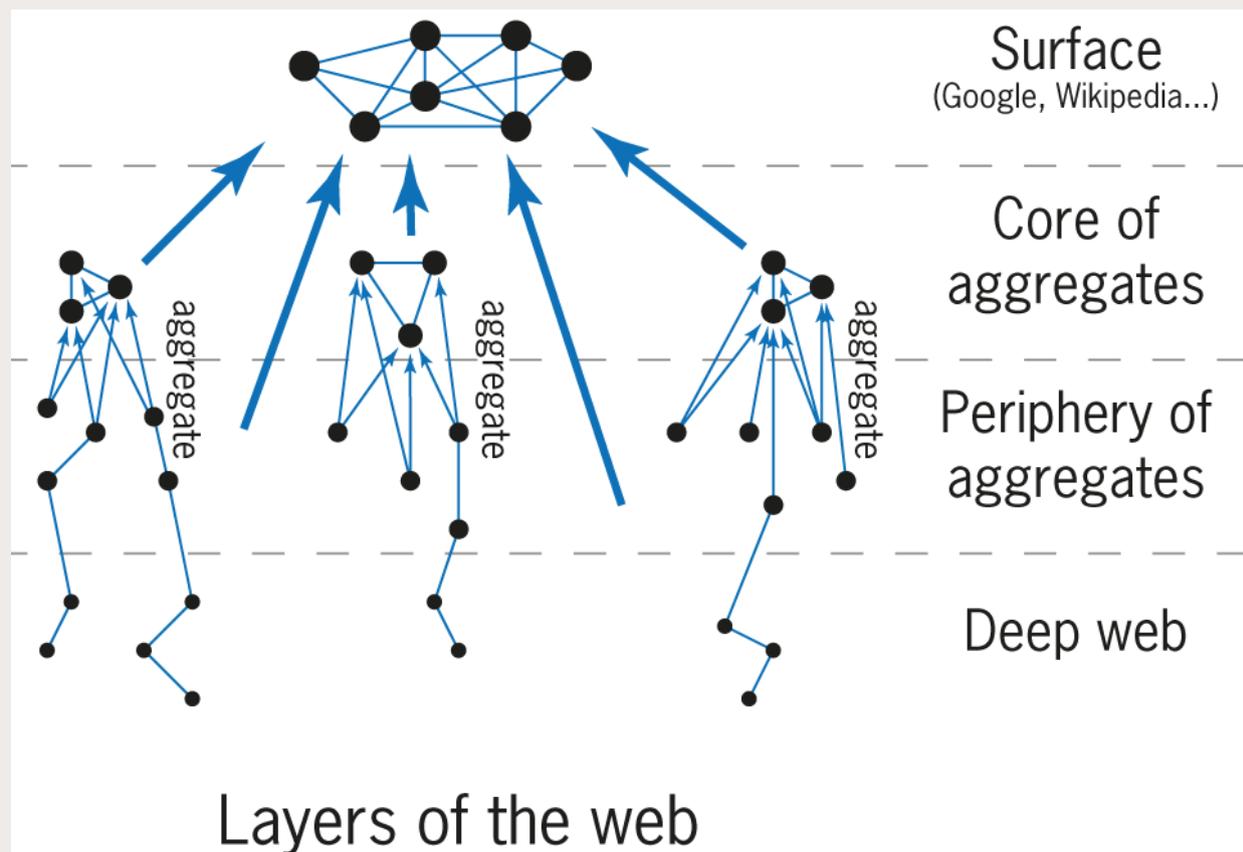
« A hyperlink is a **manifestation of intention**. By linking one page to another, one piece of text to another, **people intend** to do particular things. »

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



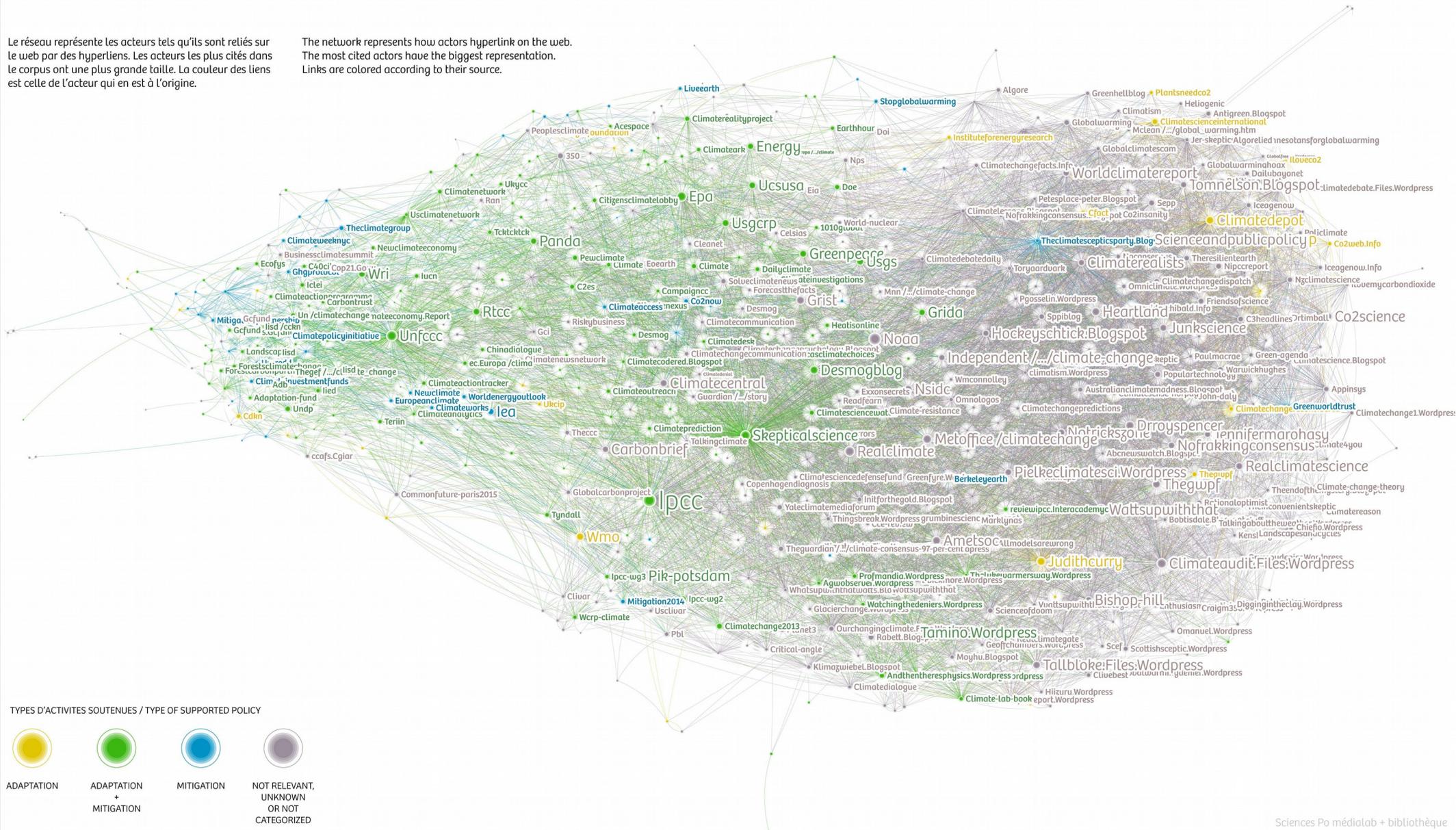
Cartographier le web par typologie d'acteurs

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

Web corpus on climate change: type of supported policy

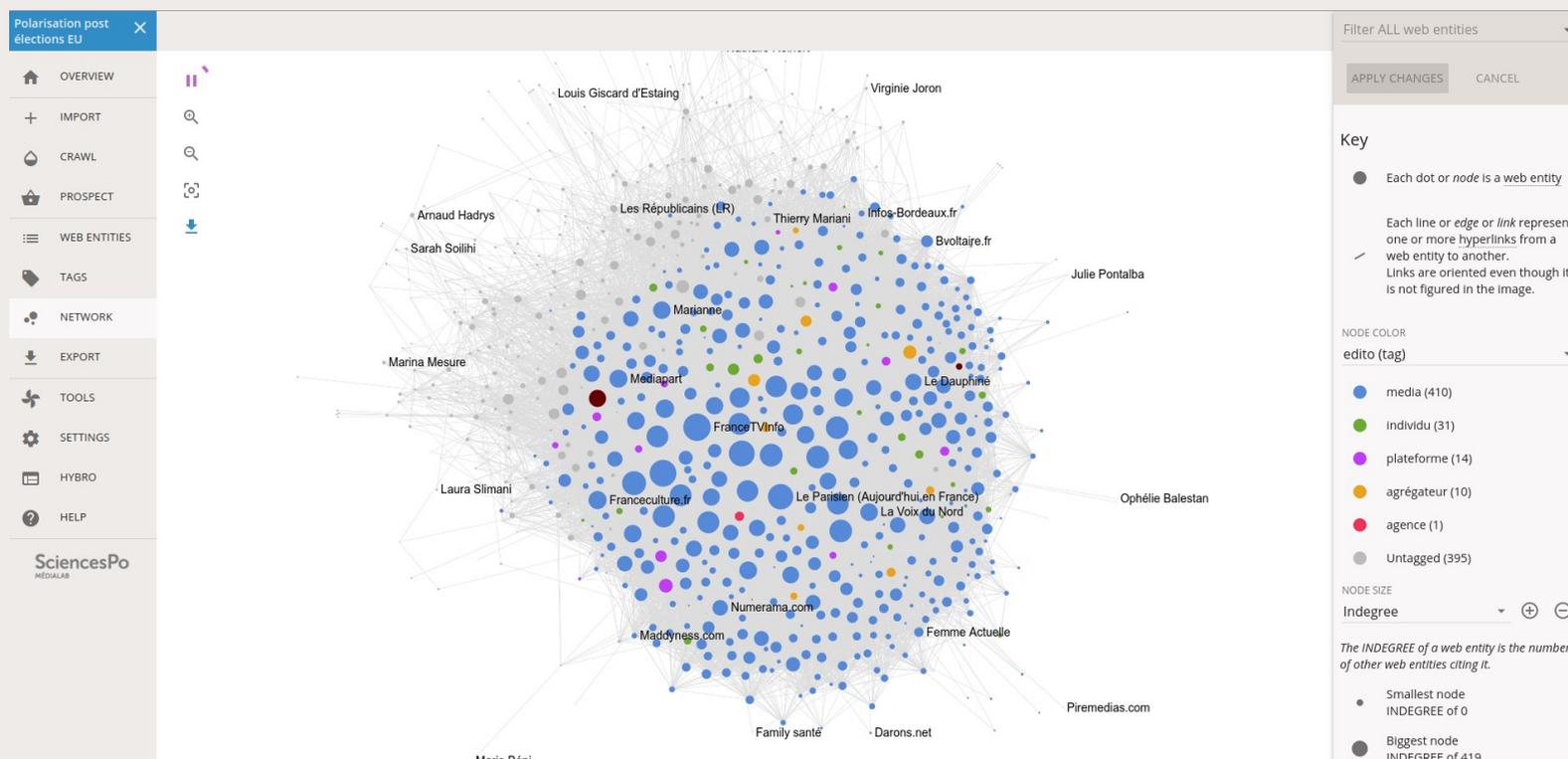
Le réseau représente les acteurs tels qu'ils sont reliés sur le web par des hyperliens. Les acteurs les plus cités dans le corpus ont une plus grande taille. La couleur des liens est celle de l'acteur qui est à l'origine.

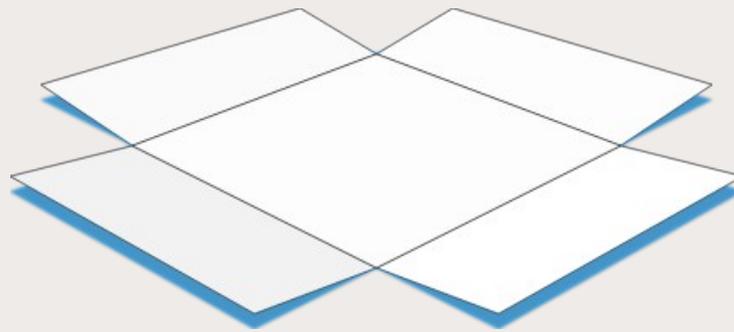
The network represents how actors hyperlink on the web. The most cited actors have the biggest representation. Links are colored according to their source.



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

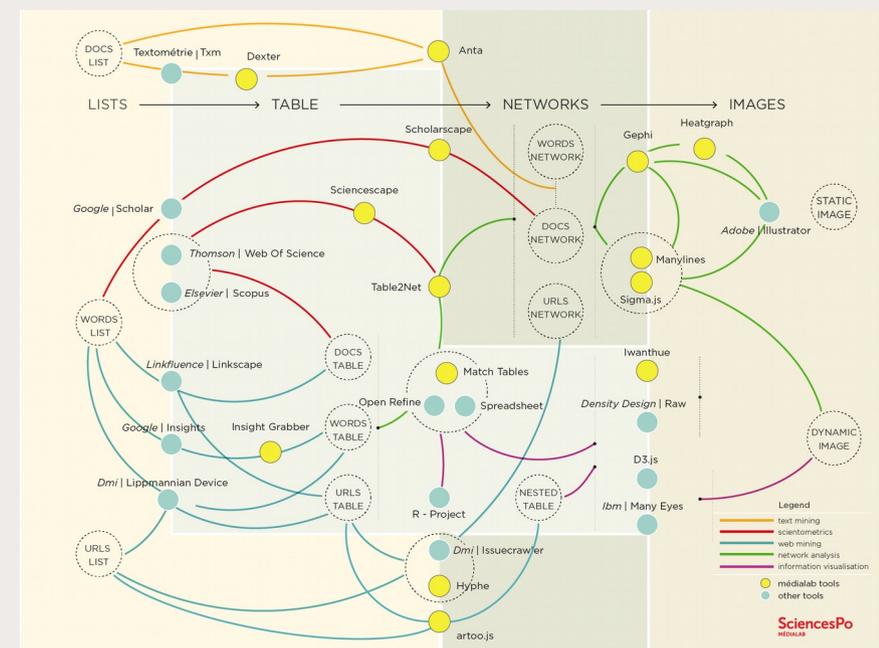
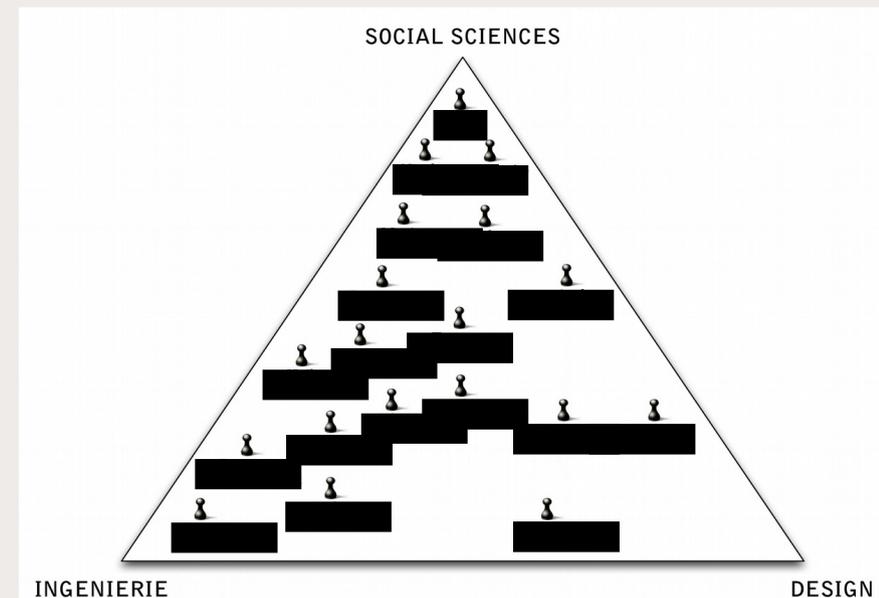


DIME - SHS

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



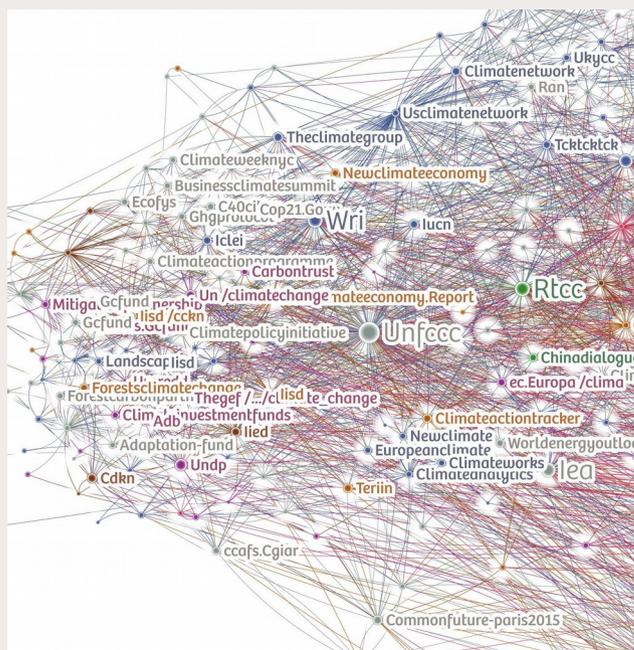
[CC-BY-SA - Rob Hille on Wikimedia Commons](#)

Hyphe : un crawler orienté par la recherche

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

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

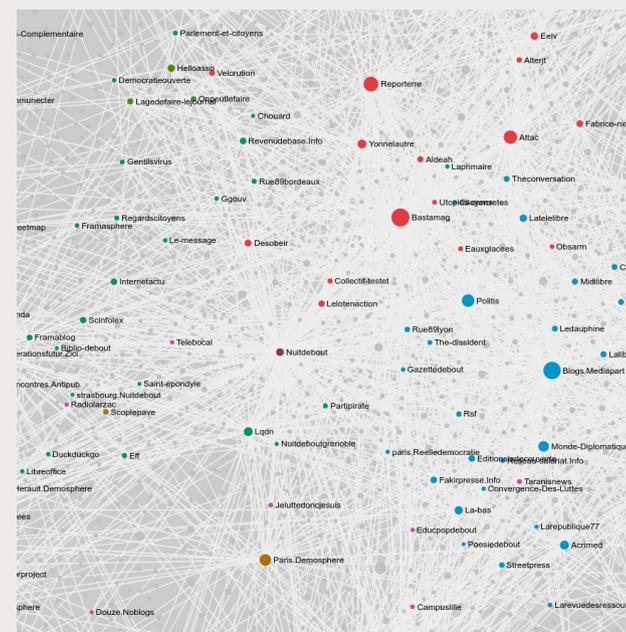
- identifier et rassembler des « acteurs web »
- explorer les liens entre leurs sites



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

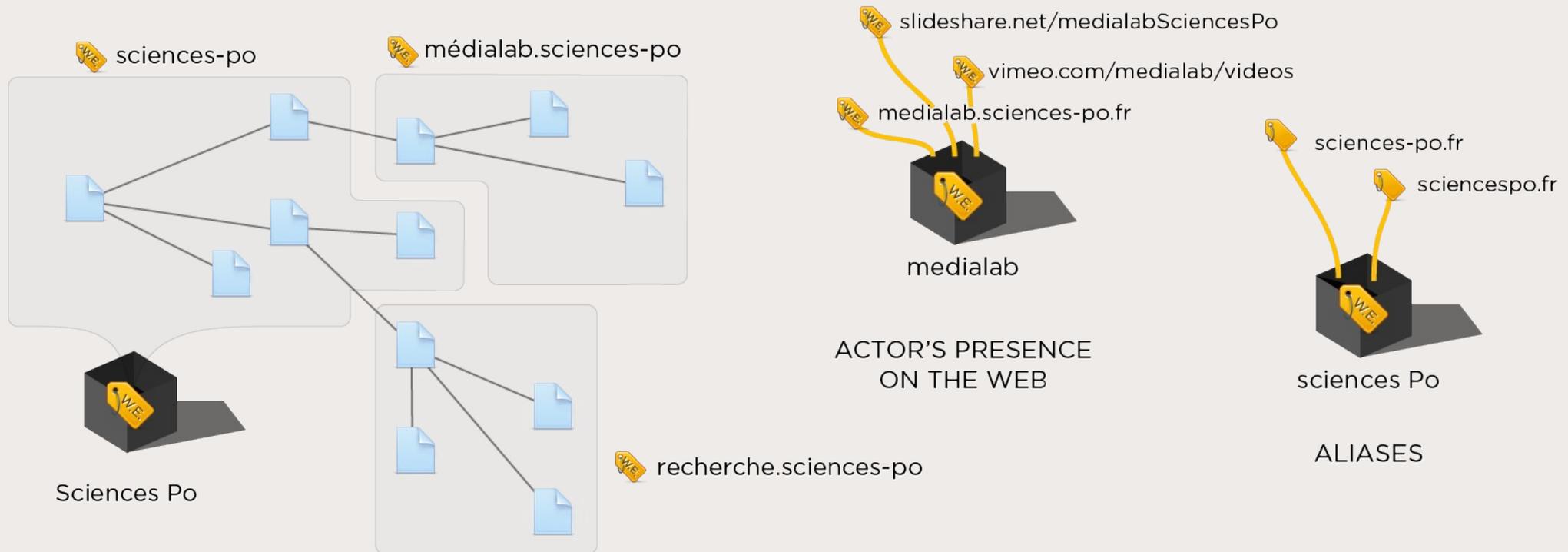
COP 21
Vie privée
Extrême droite
Tissu associatif
Produits laitiers
Cellules souches
Administrations culturelles

...



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

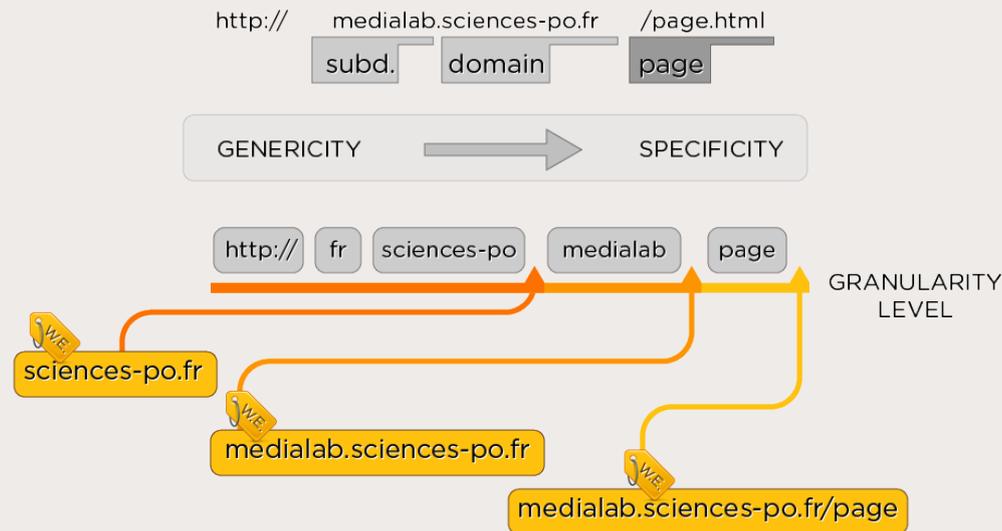
Mais qu'est-ce donc qu'un « site web » ?



→ « **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

DEFINE WEB ENTITIES

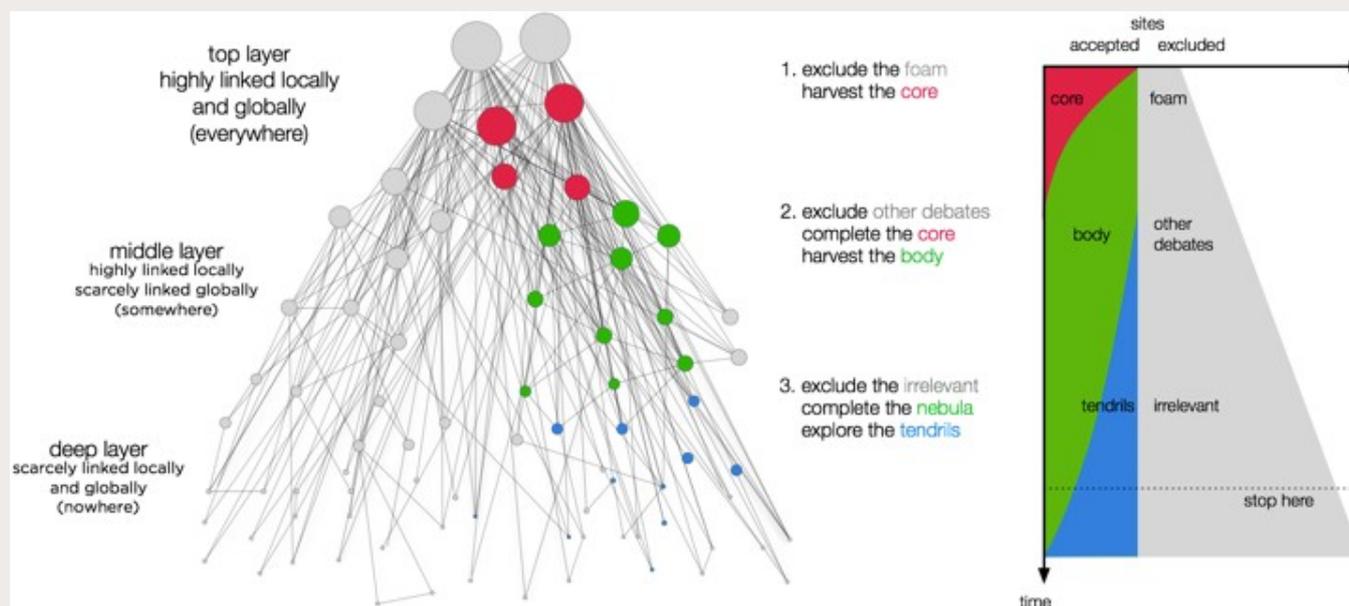
Check the boundaries of each web entity before creating it

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](#)

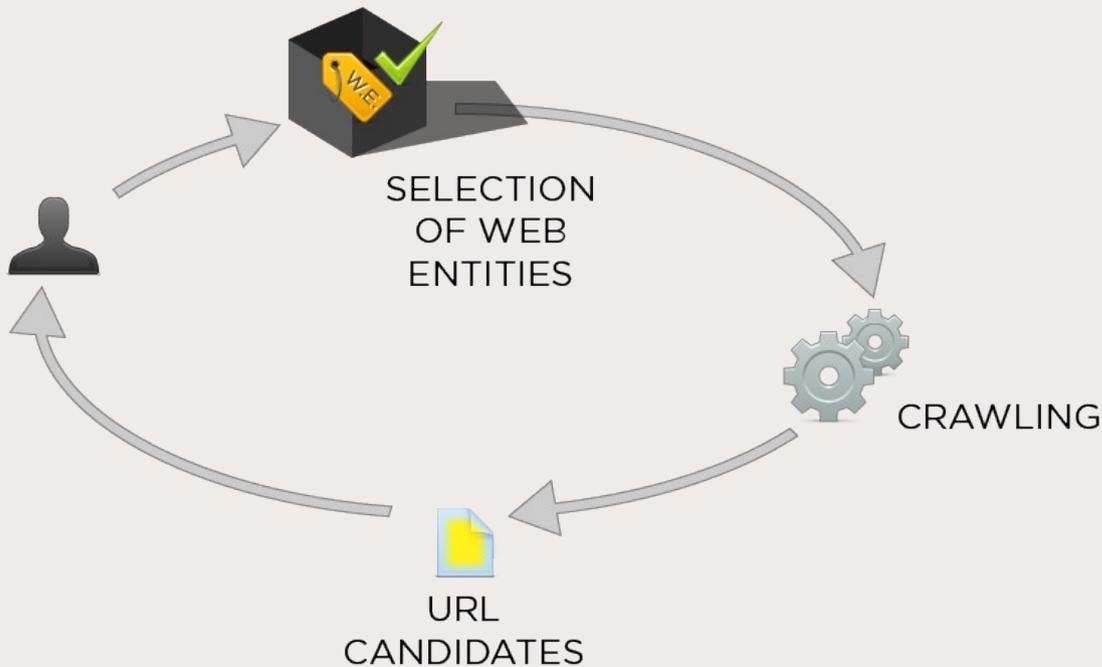
Hyphe : une stratégie de crawling contrôlé

- 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

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

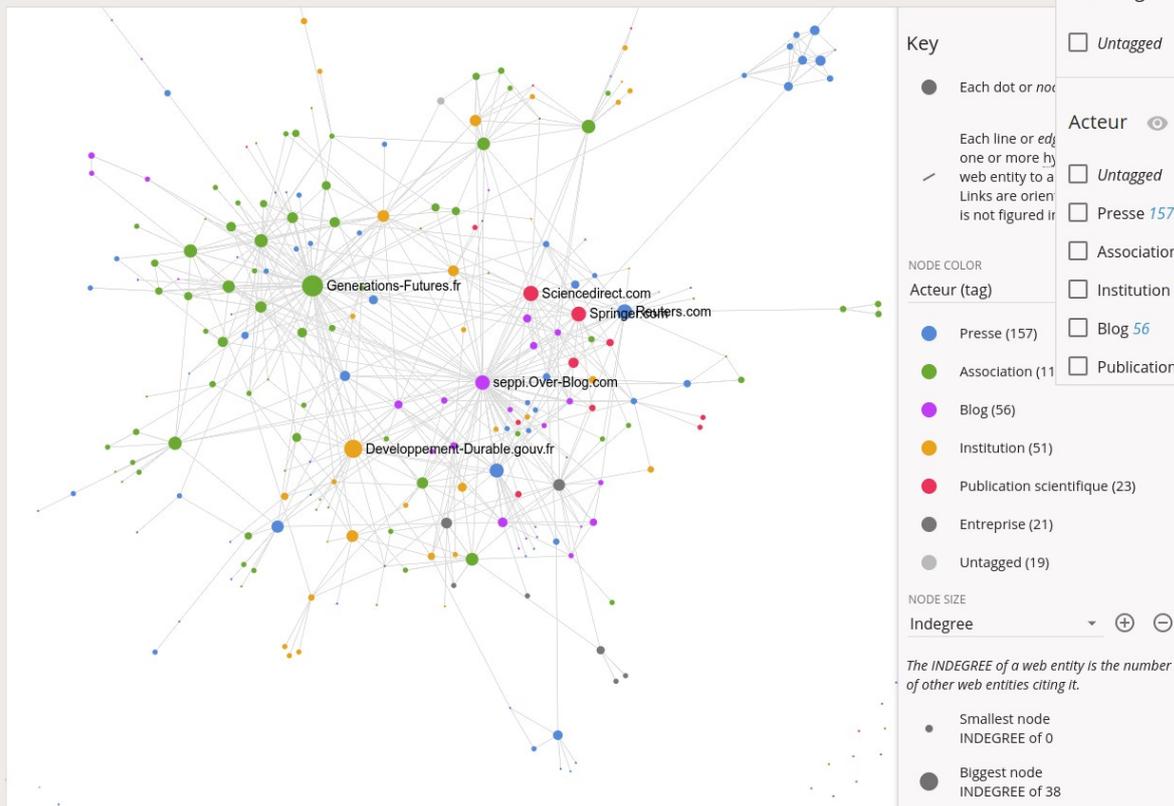
CRAWL

1 SET TO UNDECIDED
Legifrance.gouv.fr X

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

Qualifier les acteurs du corpus (tags)

- Tags catégoriels (clé/valeur)
- Annotations



TAGS

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

439
WEB ENTITIES

TAG FILTERS

Special filters

- Untagged
- Partially untagged
- Conflicts

Free Tags

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

439 WEB ENTITIES

WEB ENTITIES NETWORK

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 : un navigateur web conçu pour Hyphe

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

The screenshot displays the Hyphe Browser interface. At the top, the address bar shows 'Free.fr' and a search bar. Below the address bar, there are several tabs and filters: 'PROSPECTION' (4884), 'IN' (232), 'IN À TAGUER' (232), 'IN À CRAWLER' (47), 'UNDECIDED' (1), and 'OUT' (533). The main content area shows a page from 'Free.fr' with a title 'Le plan C : instituer une vraie démocratie par une Constitution d'origine Citoyenne.' and a sub-header 'Réflexions sur l'Europe et sur la démocratie : qu'est-ce qui empêche, toujours et partout, un réel contrôle des pouvoirs par les citoyens? Ce n'est pas aux hommes au pouvoir d'écrire les règles du pouvoir: Ass. Constituante et Cons. Constitutionnel doivent être TIRES AU SORT'. The page content includes a 'Présentation' section, a 'Forum du Plan C' section, and a 'WIKI-Constitution' section. On the right side, there is a 'Vos recherches sur le plan C' section with a search bar and a 'Résumés' section with a 'Le Message' logo. The bottom of the browser shows the language 'fr' and the 'Hyphe' logo.

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

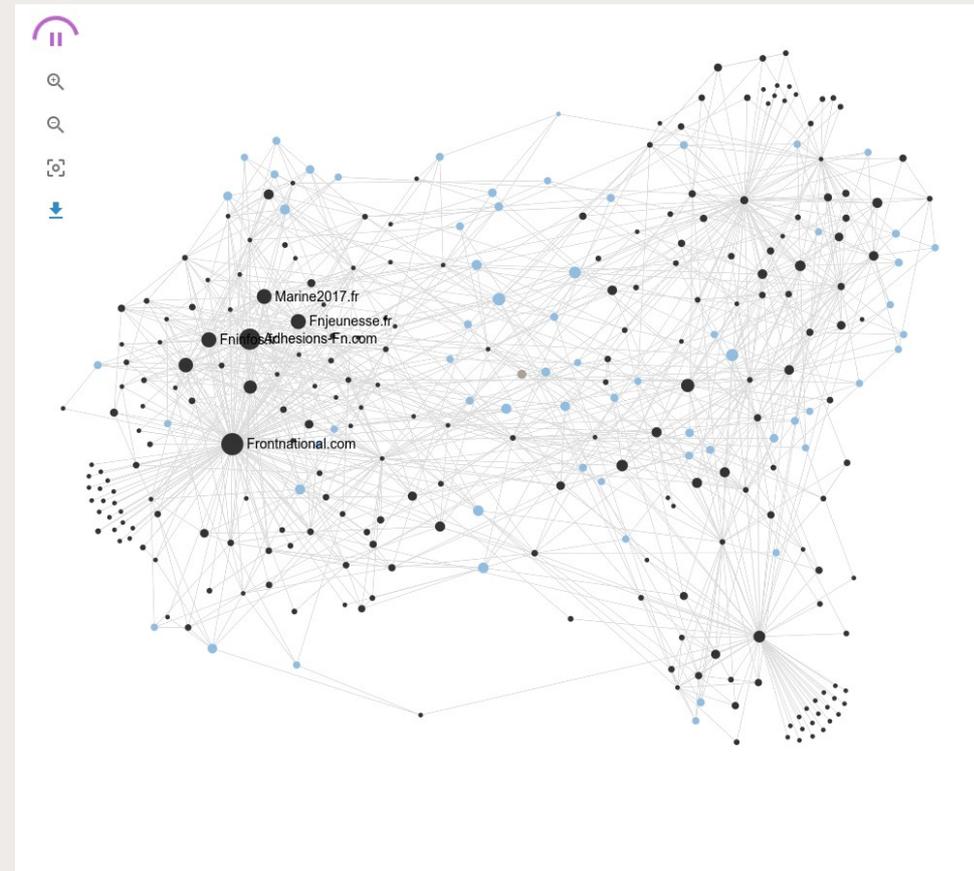
→ prospection et tagging « in-situ » (terrain numérique)

→ 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



Network Viz Settings

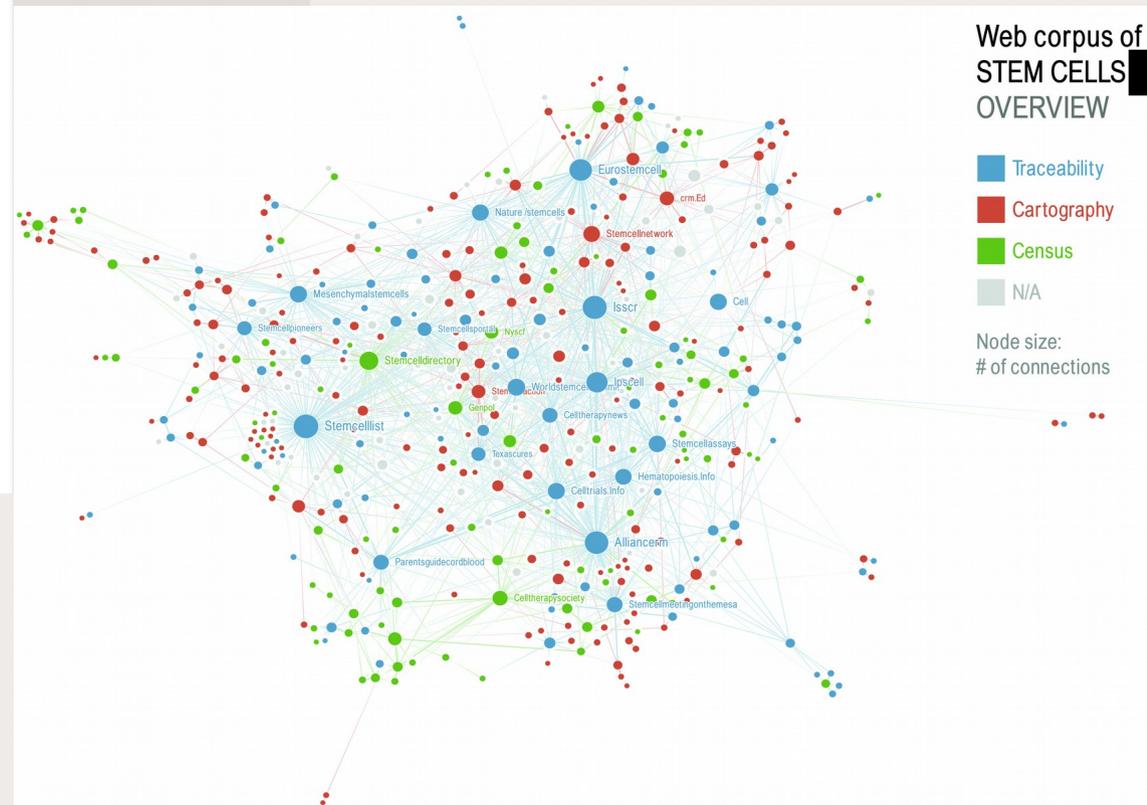
Filtering

- IN 232
- UNDECIDED 1
- OUT 533
- DISCOVERED 4,884

Filter DISCOVERED web entities

Display only DISCOVERED with ...

Filter ALL web entities



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

Analyse du fond : traiter les contenus texte

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

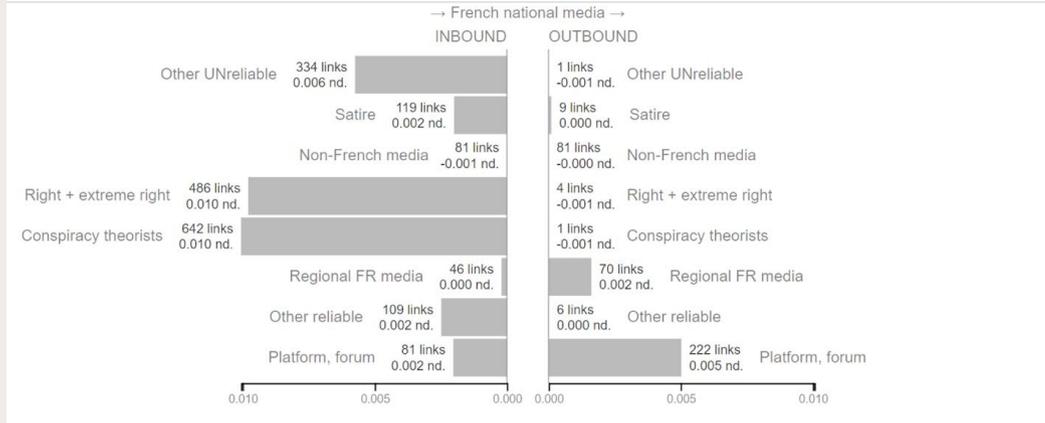
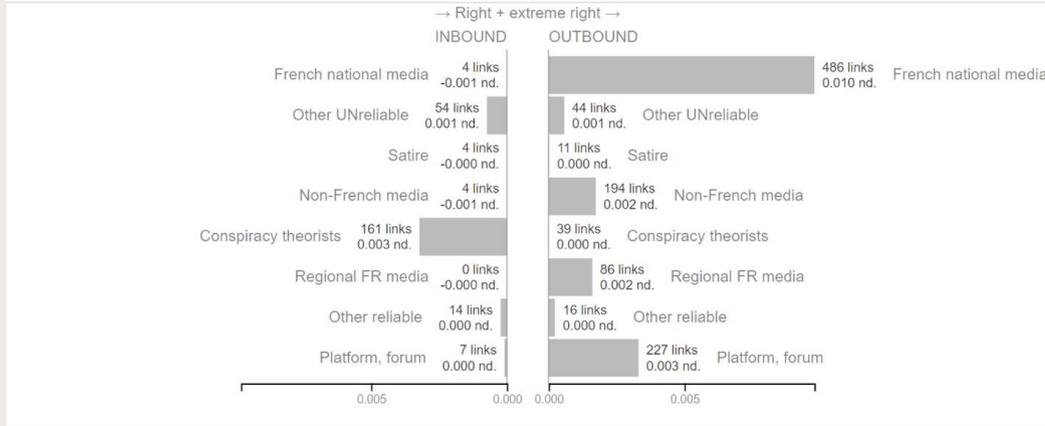
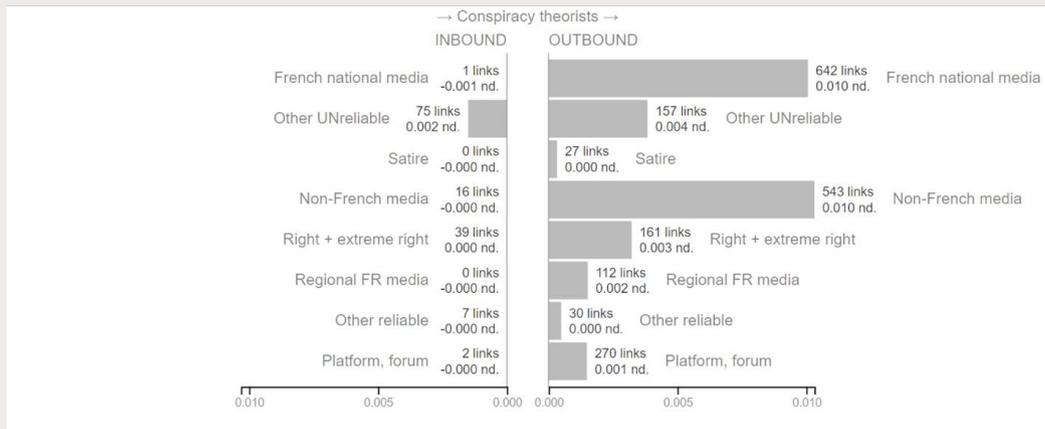
Telec Operators FR

Card and ID fraud

EXPLORE TOPICS

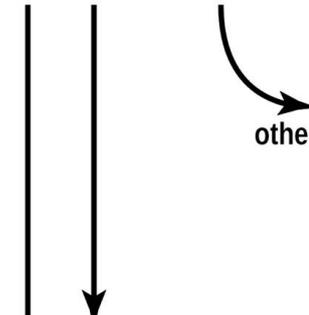
<http://tools.medialab.sciences-po.fr/privacy/>

Analyser la circulation & la structuration des liens



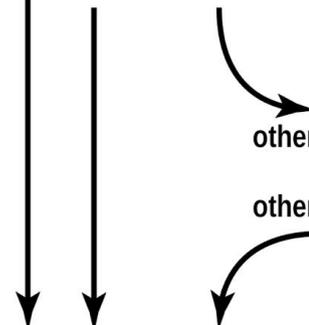
Conspiracy theorists

- Not reliable
- Not cited
- Cite many reliable sources



Right and extreme right

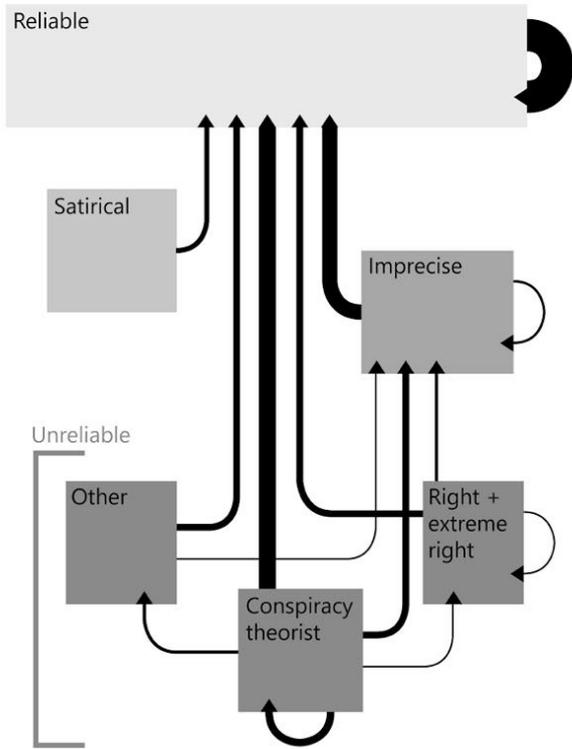
- Not reliable
- Cited only by conspirationists
- Cite many reliable sources



French national media

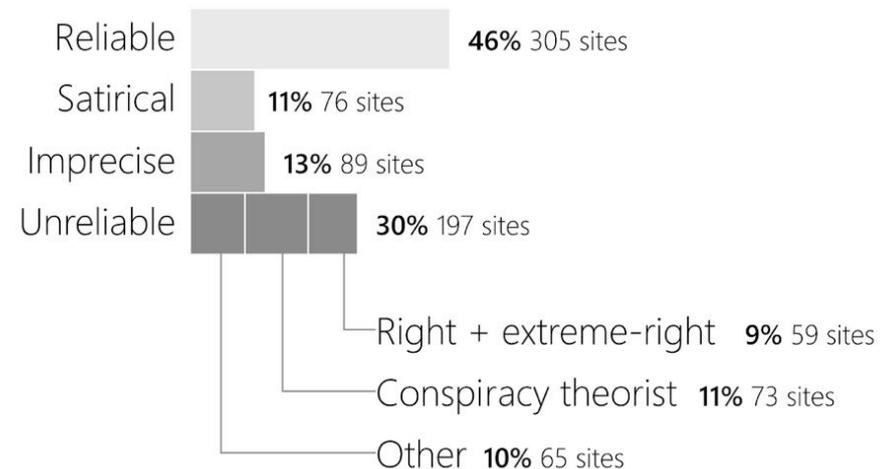
- Reliable (**in this qualification*)
- Cited by everyone
- Cite only platforms and regional media

Exploiter la directionnalité des hyperliens

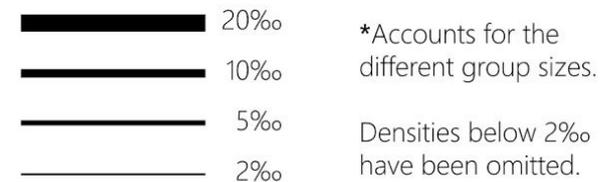


Most hyperlinks stem from the unreliable and aim at the reliable resources

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

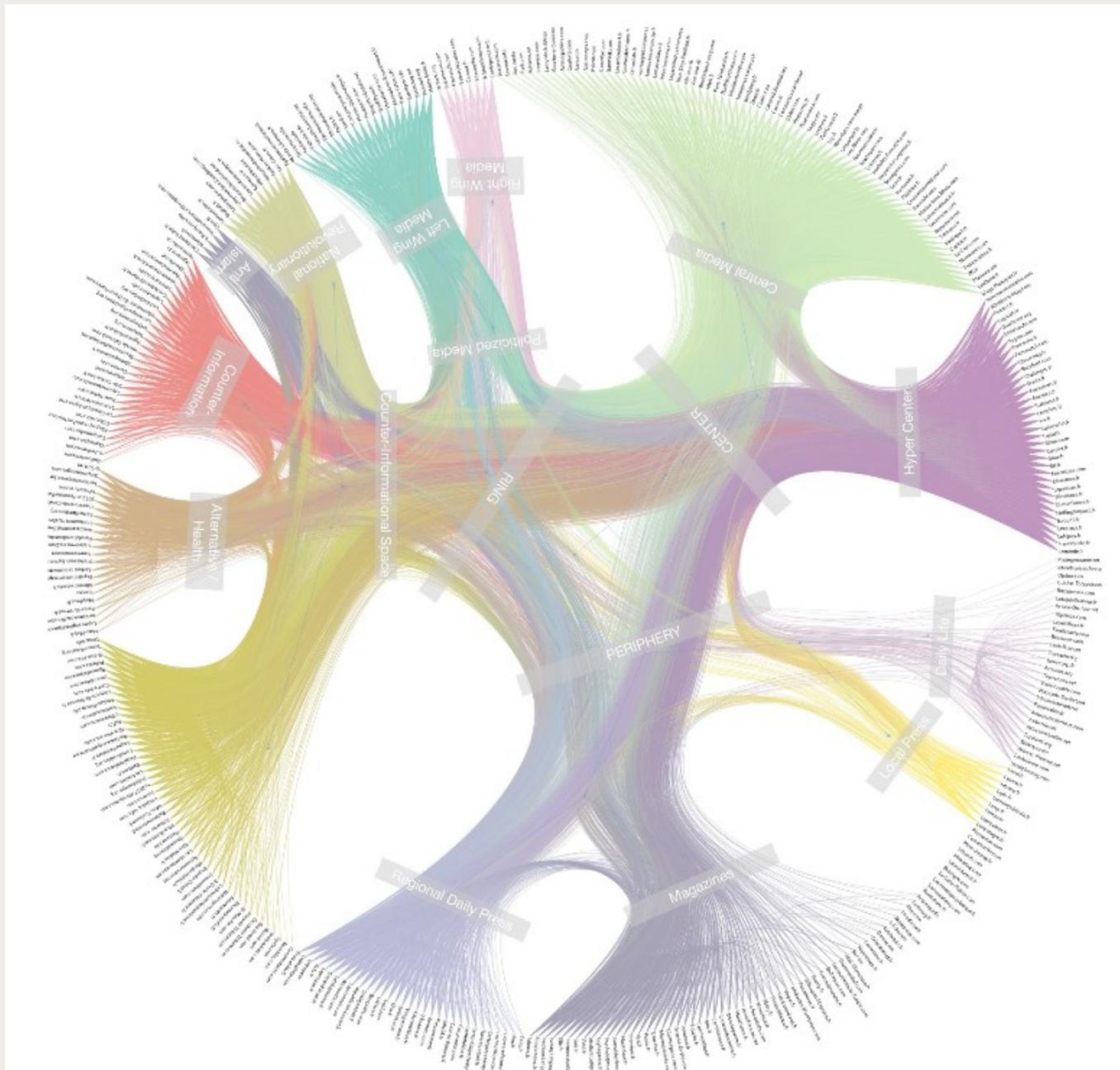


The thickness is proportional to the normalized link density*



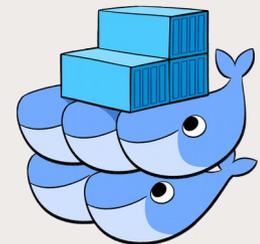
https://www.researchgate.net/publication/320225750_Visual_Network_Exploration_for_Data_Journalists

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>