



**HAL**  
open science

## Hyphe - Utiliser le Web comme terrain d'enquête

Benjamin Ooghe, Audrey Baneyx

► **To cite this version:**

Benjamin Ooghe, Audrey Baneyx. Hyphe - Utiliser le Web comme terrain d'enquête. Ecole d'été Cergy-Pontoise - Analyse du discours médiatique sur l'Europe, Université de Cergy-Pontoise, Jun 2018, Cergy-Pontoise, France. hal-03621737

**HAL Id: hal-03621737**

**<https://sciencespo.hal.science/hal-03621737>**

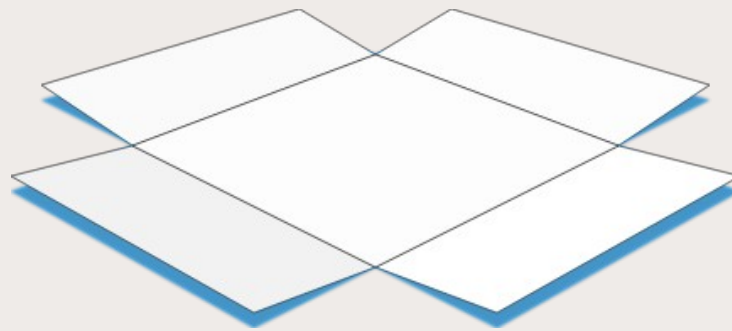
Submitted on 28 Mar 2022

**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 - NonCommercial - ShareAlike 4.0 International License



# hyphe

## Utiliser le Web comme terrain d'enquête

École d'été - Analyse du discours médiatique  
LEMEL – 21 juin 2018

**Benjamin Ooghe-Tabanou – Audrey Baneyx**

Sciences Po, médialab, Paris, France – DIME SHS Web

**SciencesPo**  
MÉDIALAB



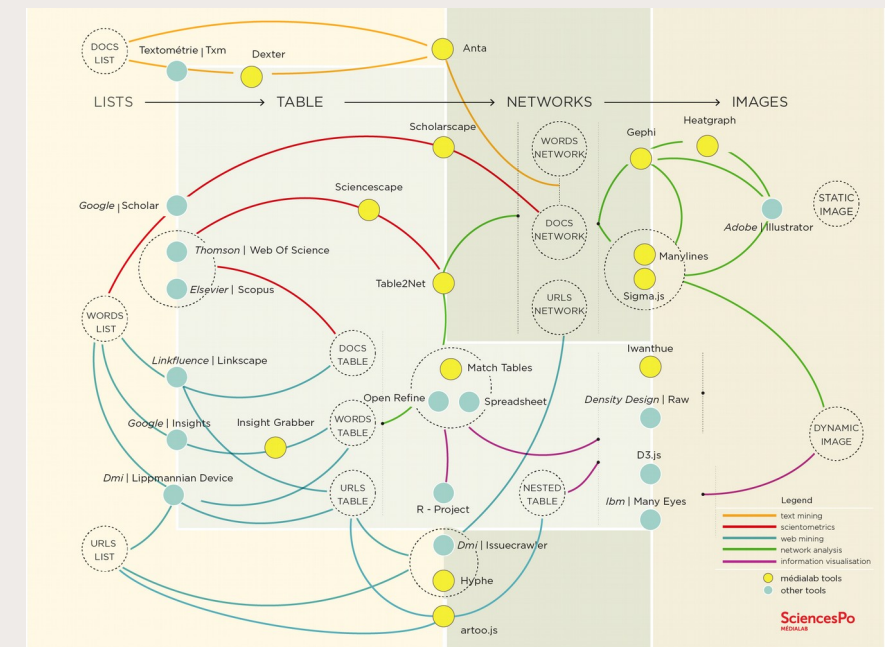
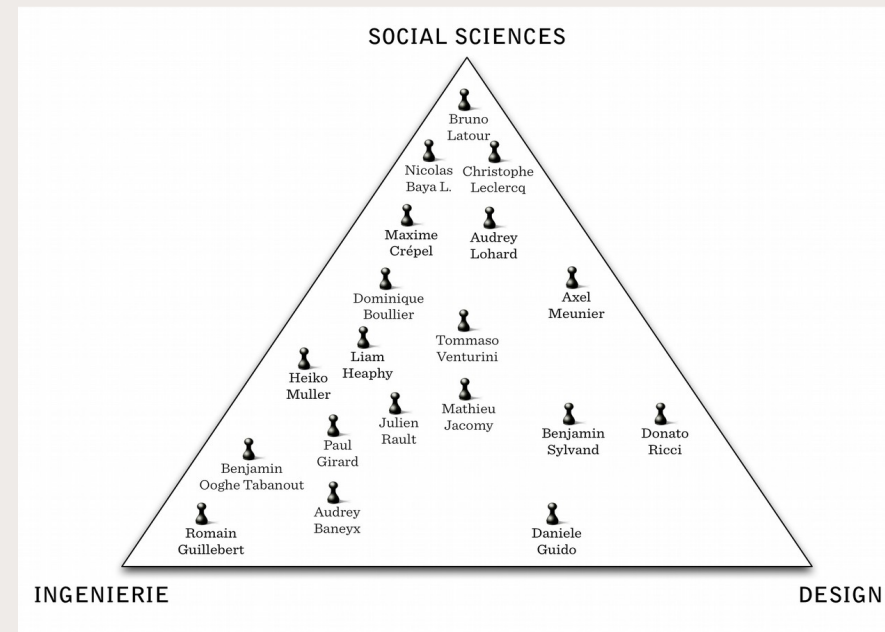
**DIME - SHS**

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



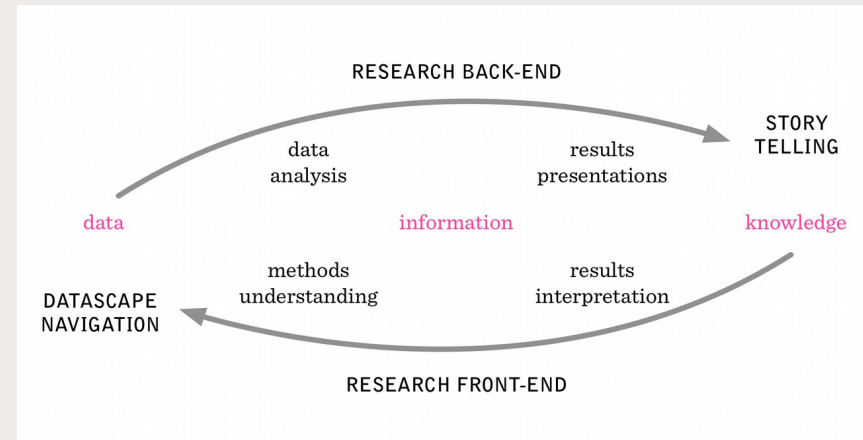
# Le médialab de Sciences Po

- 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 (Equipex DIME-SHS)

- Accompagnement technique et méthodologique à l'utilisation du Web comme nouveau terrain d'enquête
  - Support aux Sciences Humaines et Sociales
  - Extraction ciblée de contenus/discussions/traces
  - Création de corpus documentaire
  - Méthodes numériques, itératives  
≠ tout automatique



- Equipex (+ Ellips + beQuali = DIME-SHS)
  - 2 personnes (Mathieu Jacomy et moi-même)
  - Objectif ANR d'auto-financement
    - offre de service payant avec sélection
    - mutualisation (logiciels libres)

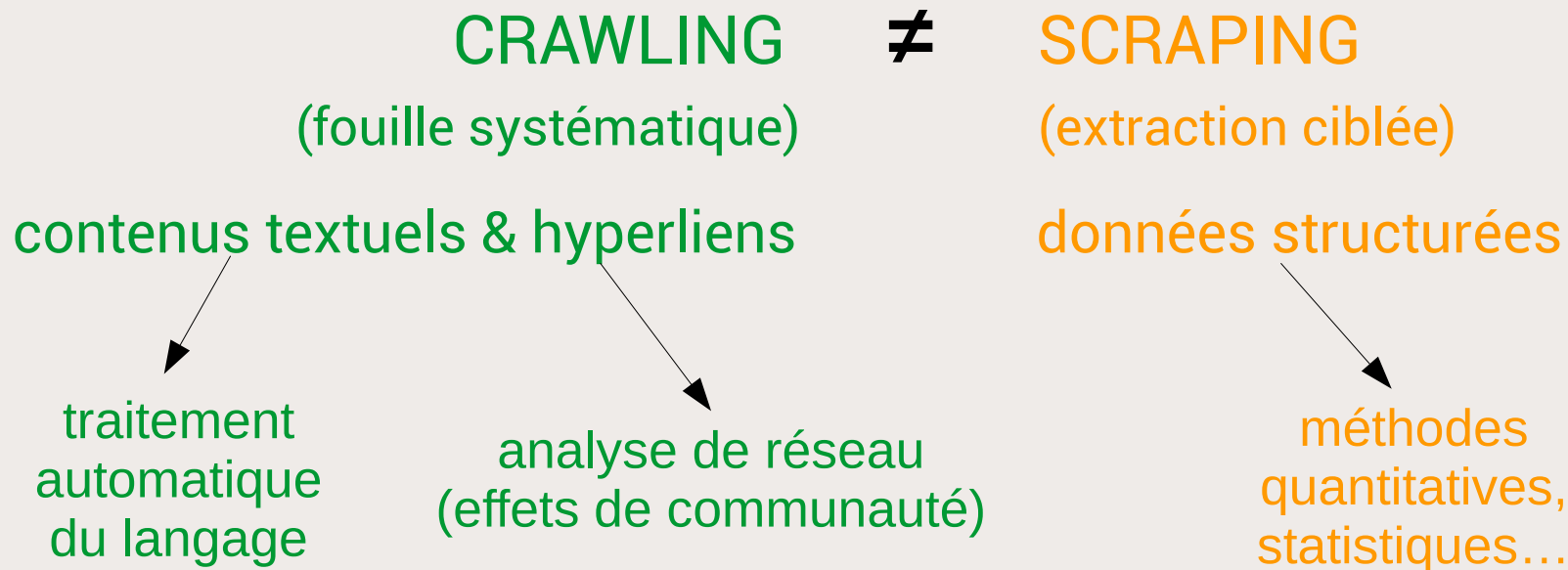


Equipex DIME-SHS  
ANR-10-EQPX-19-01

# Le Web : une source de données « sales »

Collection de documents web (pages) sur un sujet en SHS

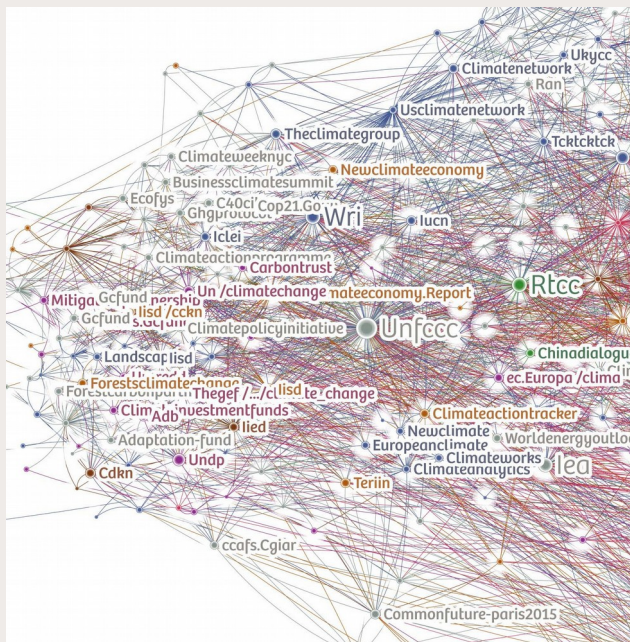
→ très grande hétérogénéité (type de contenu & forme)



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

# Hyphe : un crawler orienté par votre recherche

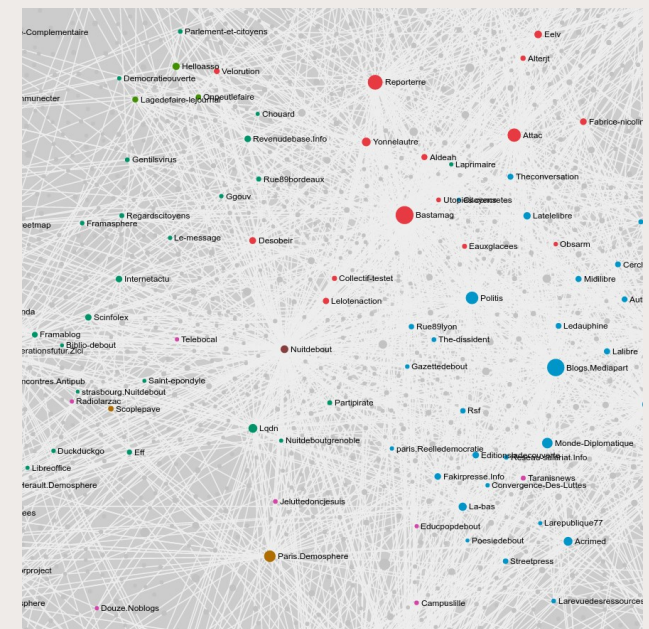
- Les liens hypertextes : nouveaux révélateurs de relations entre acteurs d'une thématique
- Créer un corpus documentaire
  - « acteurs web » & contenus textuels respectifs
  - liens hypertextes entre ces acteurs
- Études exploratoires ou de controverses dans tous les domaines



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

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

...

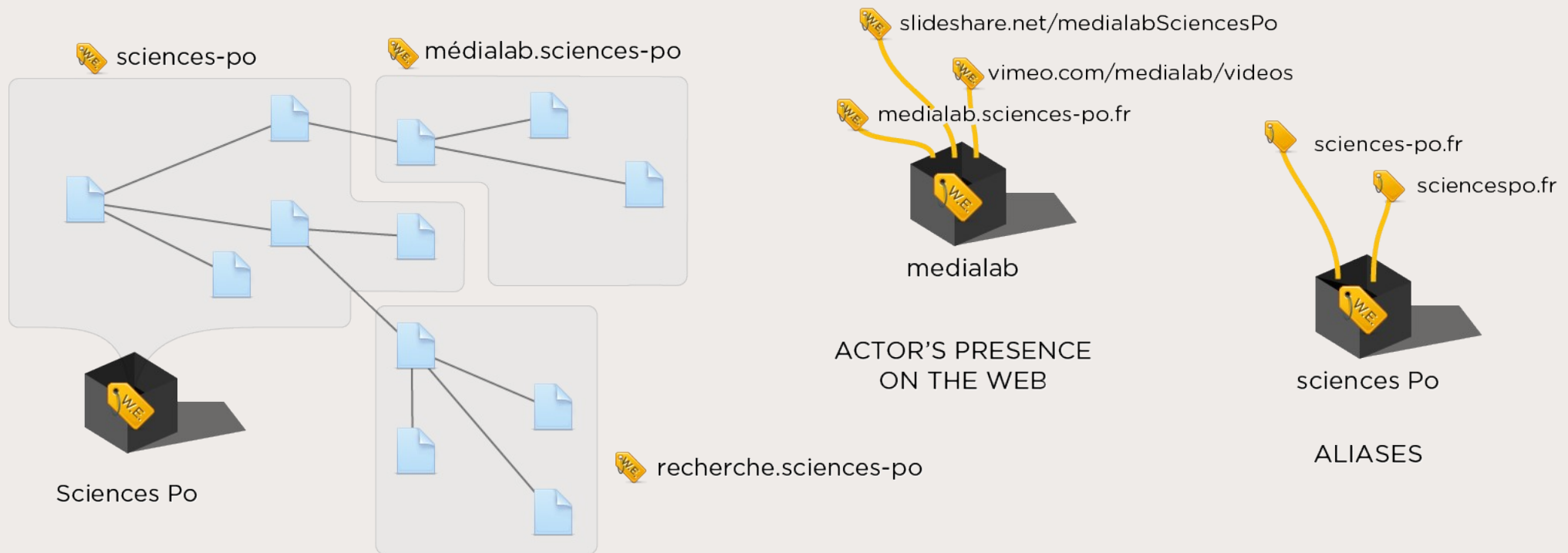


<http://utopies-concretes.org/>

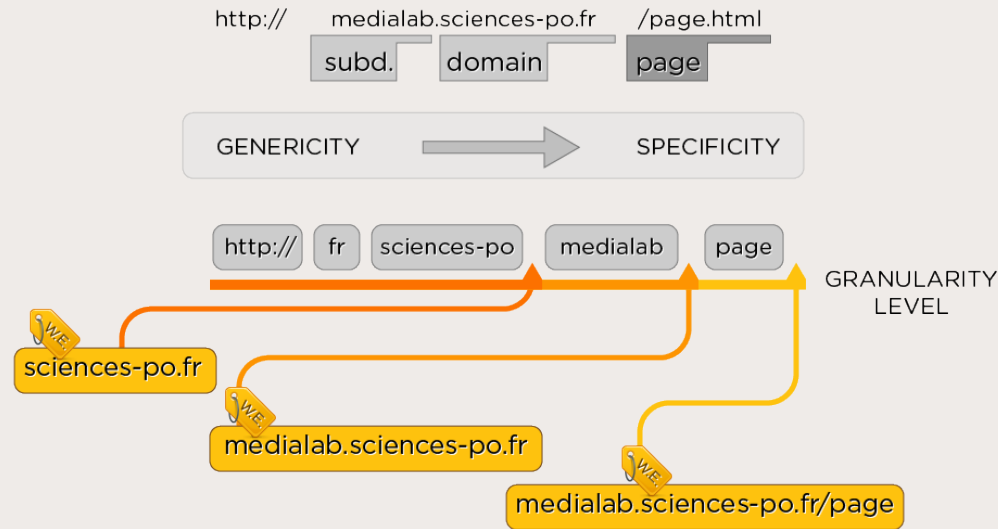
# Principes méthodologiques : « WebEntités »

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 »

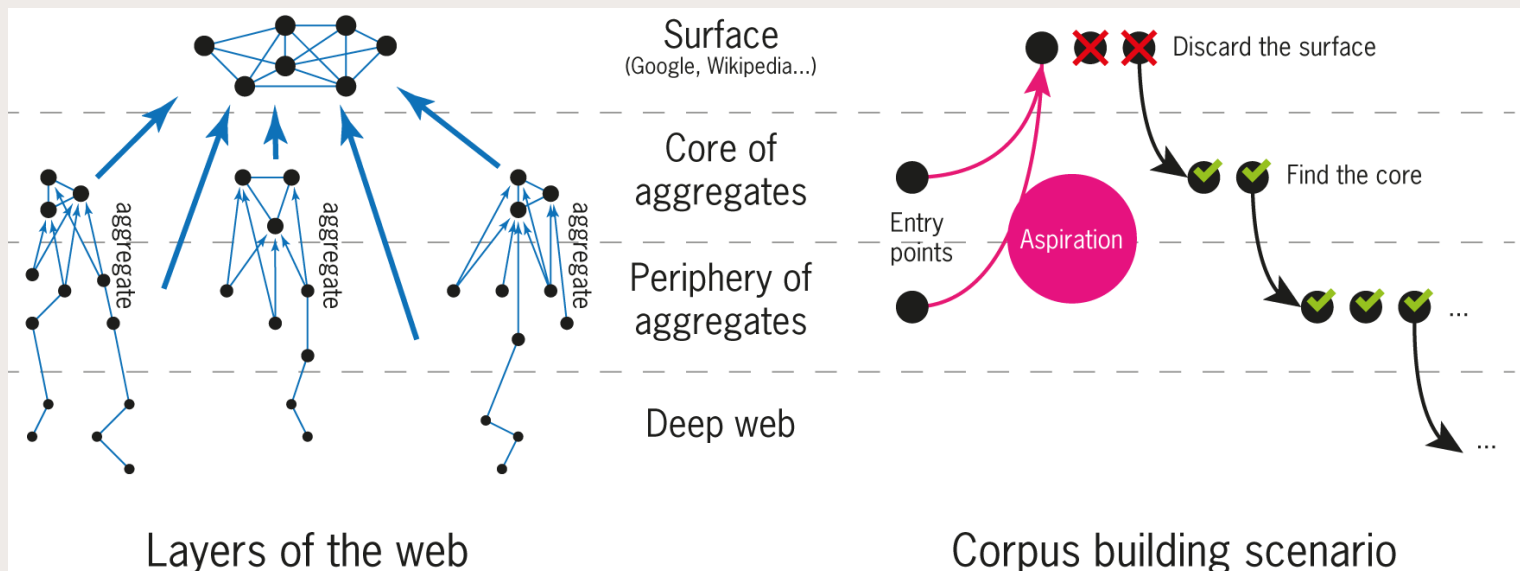


41	<a href="#">Amnesty.fr</a>	<a href="#">http</a> <a href="#">.fr</a> <a href="#">amnesty</a> <a href="#">www.</a>
42	<a href="#">Facebook.com /.../326366925310</a>	<a href="#">http</a> <a href="#">.com</a> <a href="#">facebook</a> <a href="#">www.</a> <a href="#">/pages</a> <a href="#">/Andr%C3%A9-...</a> <a href="#">/326366925310</a> Same web entity defined rows 82, 130, 150, 189, 249, 388, 389, 392, 393, 424, 475, 483, 488, 493, 640, 642, 659, 668, 690, 707, 719, 779, 966, 972 and 989
43	<a href="#">Annuairemairie.com</a>	<a href="#">http</a> <a href="#">.com</a> <a href="#">annuairemairie</a> <a href="#">www.</a>
44	<a href="#">Marianne2.fr /hervenathan</a>	<a href="#">http</a> <a href="#">.fr</a> <a href="#">marianne2</a> <a href="#">www.</a> <a href="#">/hervenathan</a> Same web entity defined rows 651, 895 and 896
45	<a href="#">Anticor.org</a>	<a href="#">http</a> <a href="#">.org</a> <a href="#">anticor</a>
46	<a href="#">Desgouilles.fr</a>	<a href="#">http</a> <a href="#">.fr</a> <a href="#">desgouilles</a> <a href="#">david.</a>



# 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

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

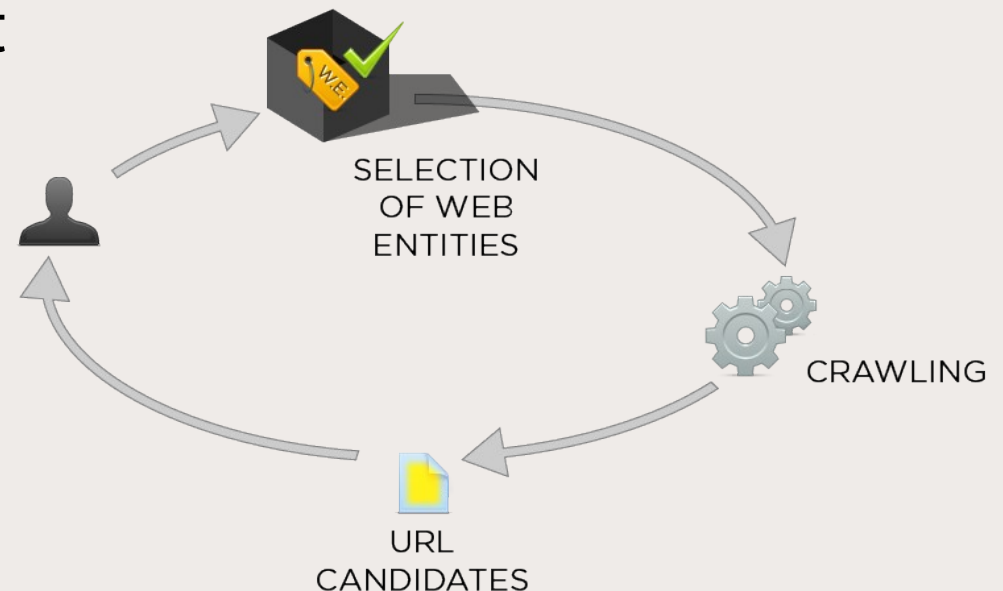
4 SET TO OUT

Gravatar.com X

Google.fr X

# Principes méthodologiques : « Prospection »

- 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 pour prospecter in situ

- Hype-Browser : héritier du « NaviCrawler »
- Un navigateur web connecté à Hype

The screenshot shows the Hype Browser interface. At the top, the browser title is 'Hype Browser' and the address bar shows 'FrenchFarRight4'. Below the address bar, there's a search bar with 'Free.fr' and a magnifying glass icon. To the right of the search bar, there are several filters: 'PROSPECTION' (4884), 'IN' (232), 'IN À TAGUER' (232), 'IN À CRAWLER' (47), 'UNDECIDED' (1), and 'OUT' (533). The main content area displays a webpage for 'Free.fr' with the URL 'http://etienne.chouard.free.fr/Europe/index.php'. The page features a large image of an ancient amphitheater and a headline: 'Le plan C : instituer une vraie démocratie par une Constitution d'origine Citoyenne.' Below the headline, there's a paragraph of text and a search bar with the number '7119753'. The page is organized into several sections: 'Présentation', 'Analyses et propositions', 'Priorités', 'Échanges', 'CECRI', 'Réflexions', 'Initiatives', and 'Divers'. On the left side, there's a sidebar with 'Pages citées' and 'Entités citantes'. The sidebar lists several websites: 'erlille.wordpress.com', 'Leforumcatholique.org', 'Bernard-Antony.com', and 'Agoravox.fr'. At the bottom of the sidebar, there are 'Annotations', 'Ajouter un tag...', and 'Catégories'.

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

Analyse du discours médiatique sur l'Europe - Hype

# Catégoriser les WebEntités avec HyBro

The screenshot displays the HyBro interface for a web page. At the top, the browser title is 'Hyphe Browser' and the page ID is 'ABC111'. A navigation bar shows the following categories and counts: PROSPECTION (8367), IN (105), IN À TAGGER (104), IN À CRAWLER (7), UNDECIDED (1), and OUT (497). The address bar shows the URL: 'blogs.ei.Columbia.edu /.../u-s-drought-risk-wider-than-previously-thought#1'. The left sidebar contains a 'WebEntité' section with 'Statut' (Crawlée | ✓, Citée par 1 WE) and 'Statut' buttons (I, ?, O). Below it are 'Contexte' and 'Annotations' sections. The 'Annotations' section has a search box with 'water' and a button 'Créer le tag: "water"'. There are also 'Lang' (en) and 'type' (blog) dropdowns, and a button 'Ajouter une catégorie'. The main content area shows a navigation bar with categories: AGRICULTURE, CLIMATE, EARTH SCIENCES, ECOLOGY, ENERGY, HEALTH, SUSTAINABILITY, URBANIZATION, WATER. The article title is 'U.S. Drought Risk Wider than Previously Thought' with a 'WATER' tag. The author is 'LAKIS POLYCARPOU' and the date is 'MAY 4, 2015'. There are social media sharing icons and a 'Comments' button. The article text starts with 'New project research conducted as part of the Columbia Water Center's "America's Water Initiative" suggests that many more places in the United States are at risk of drought-induced water stress than is commonly thought, including dense metropolitan regions in the mid-Atlantic and Northeast such as New York City and Washington, D.C.' The right sidebar has a search bar, an 'Education News' section with a photo of a woman and the text 'Alumna Planting "Seeds" for Sustainable Education in Africa', and a 'FROM THE FIELD' section with a photo of a field.

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

# Gérer ses catégorisations (tags)

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

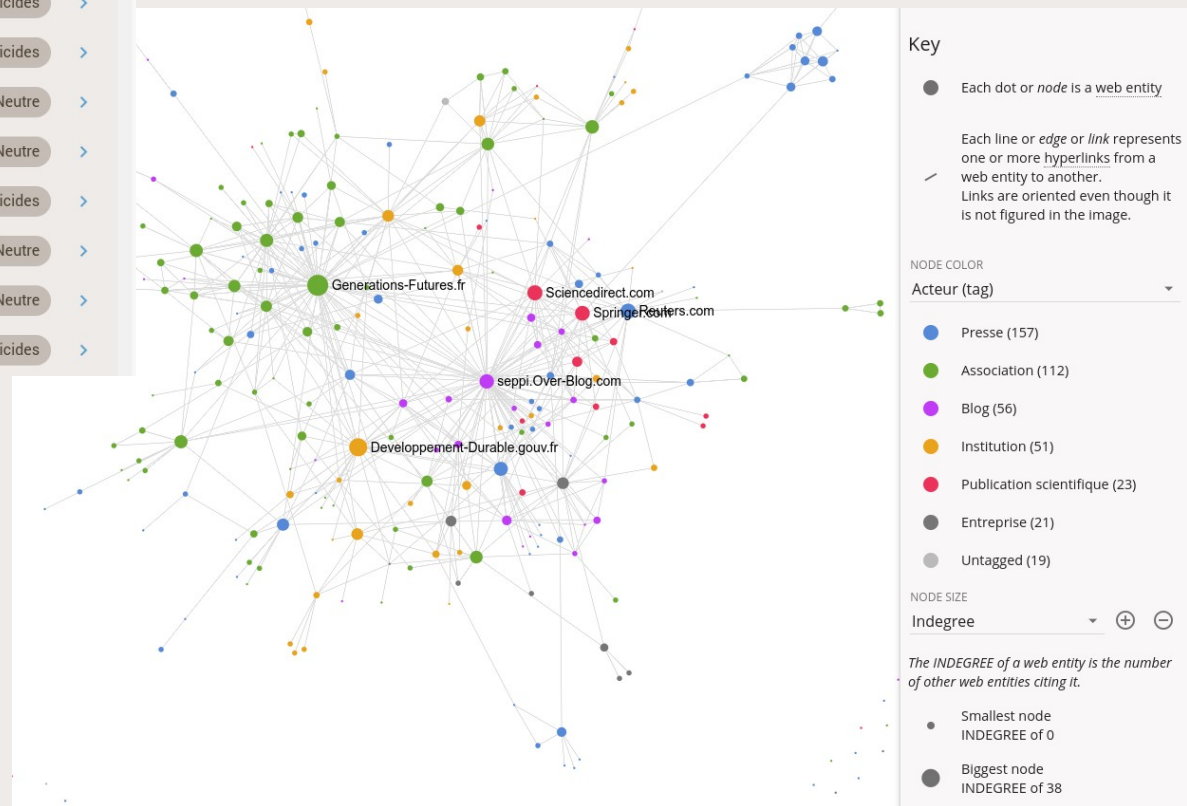
WEB ENTITIES NETWORK

Display a category

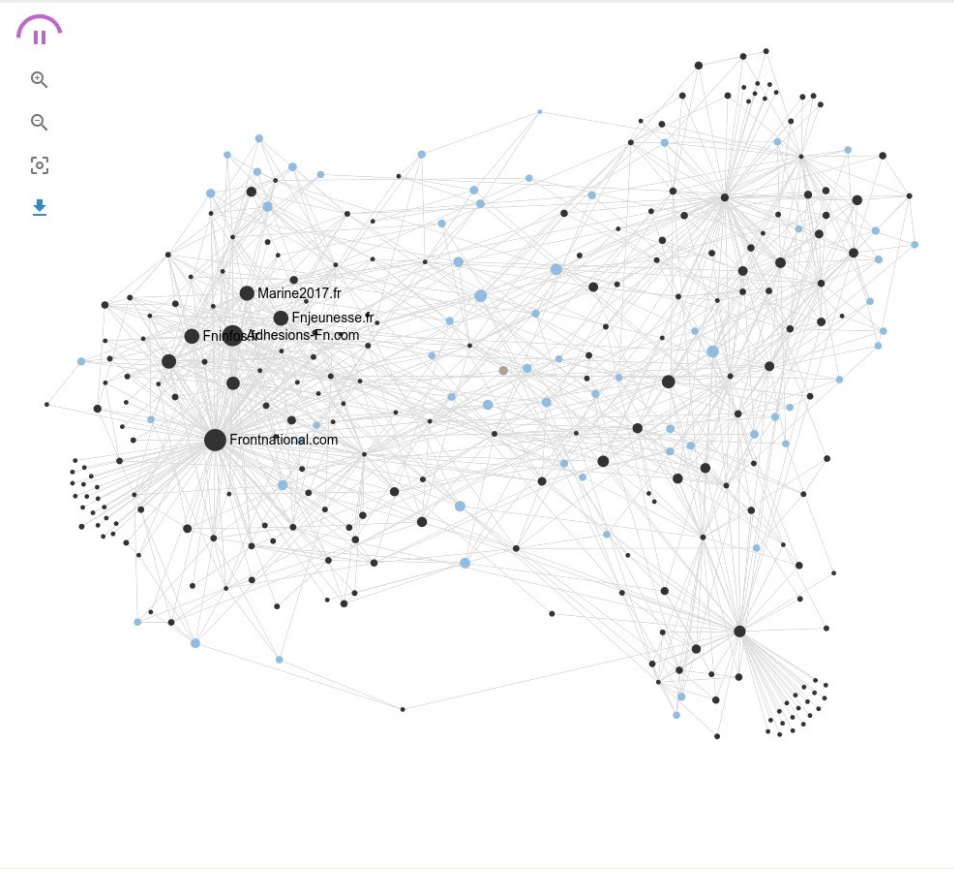
Point de vue

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



# Explorer le réseau des liens entre acteurs



Network Viz Settings

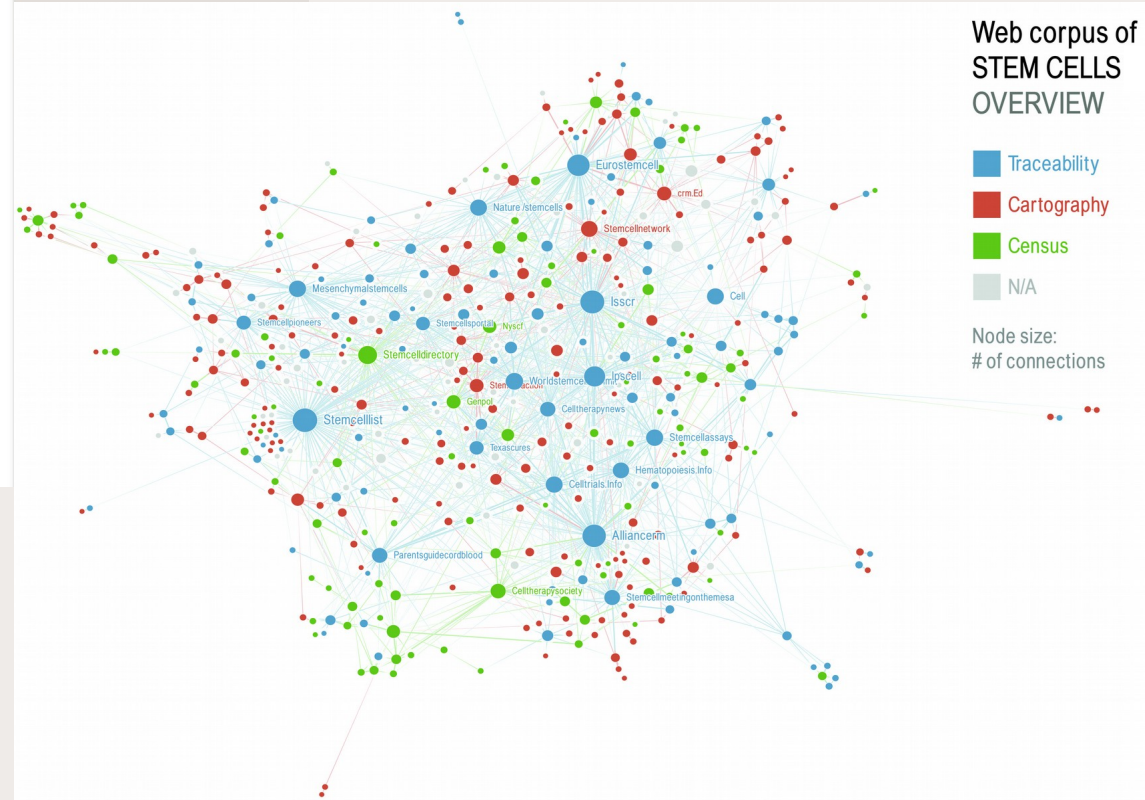
Filtering

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

Filter DISCOVERED web entities

Display only DISCOVERED with ...

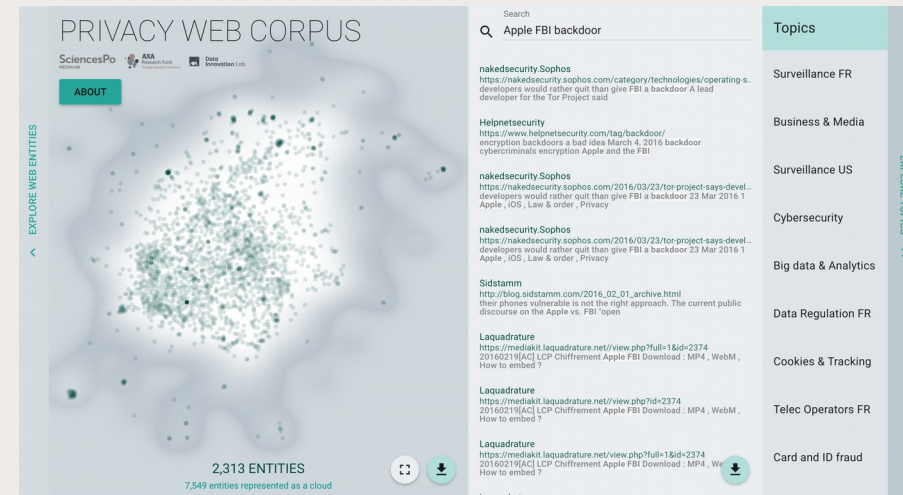
Filter ALL web entities



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

# 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.)
- Contrôle qualité des crawls et du corpus
- Outil d'archivage et présentation des corpus finalisés
- Hyphe embarqué sur clé USB





# Bibliographie & liens divers

- Concepts et explications :  
<http://hyphe.medialab.sciences-po.fr/>
- Instance de démo (restreinte) en libre accès :  
<http://hyphe.medialab.sciences-po.fr/demo/>
- Publications associées :
  - 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, Allemagne.  
<https://spire.sciencespo.fr/hdl:/2441/6obemb2hsj9pboj9bbvc7sftne>
  - 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): e98679.  
doi:10.1371/journal.pone.0098679.  
<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679>
  - Venturini T., Jacomy M., Pereira D. (2015), **Visual Network Analysis: the Example of the Rio+20 Online Debate**, Working paper.  
[http://www.medialab.sciences-po.fr/wp-content/uploads/2015/06/VisualNetwork\\_Paper-10.pdf](http://www.medialab.sciences-po.fr/wp-content/uploads/2015/06/VisualNetwork_Paper-10.pdf)

# Merci de votre attention !

---

Et maintenant, à vous de jouer !

**SciencesPo**  
MÉDIALAB

[@medialab\\_ScPo](#)

[benjamin.ooghe@sciencespo.fr](mailto:benjamin.ooghe@sciencespo.fr)

# 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

The image is a collage of screenshots illustrating the workflow of the Google bookmarklets. It shows the installation page, a search for 'digital humanities', a 'Redirect to Classic Google' dialog, and an 'Extract Classic Google Results' dialog. Arrows indicate the flow from the search page to the dialog boxes.

**Install Google Bookmarklets**  
Drag & drop images below into your bookmark bar:

**Redirect to Classic Google**  
Which language?   
How many results per page?   
You will be redirected to the following url:  
`https://encrypted.google.com/search?q=digital%20humanities&hl=en&num=100&start=0`  
Redirect me!

**Extract Classic Google Results**  
Search for "digital humanities"  
page 0 (with up to 100 urls per page)  
103 new results in this page  
Keep existing results & continue to the next page  
Download CSV with 103 urls

→ « Import urls » dans Hyphe