



HAL
open science

Hyphe : utiliser le Web comme terrain d'enquête

Benjamin Ooghe, Maxime Crépel

► **To cite this version:**

Benjamin Ooghe, Maxime Crépel. Hyphe : utiliser le Web comme terrain d'enquête. Les journées data-shs: traiter et analyser ses données en sciences humaines et sociales, MESHS Lille, Dec 2017, Lille, France. hal-03582569

HAL Id: hal-03582569

<https://sciencespo.hal.science/hal-03582569v1>

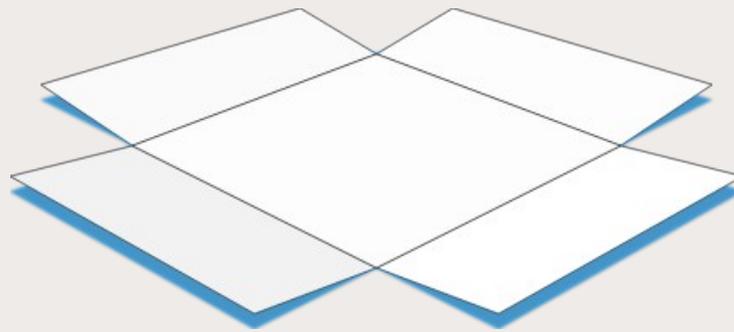
Submitted on 21 Feb 2022

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

Ateliers DATA SHS

MESHS Lille - 15 décembre 2017

Benjamin Ooghe-Tabanou, Sciences Po, médialab, Paris, France

DIME SHS Web

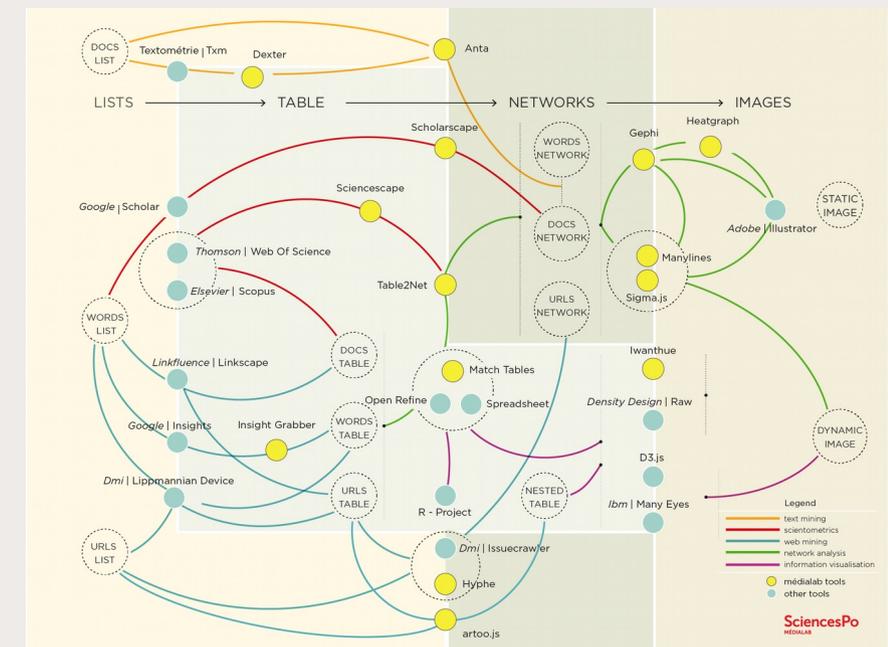
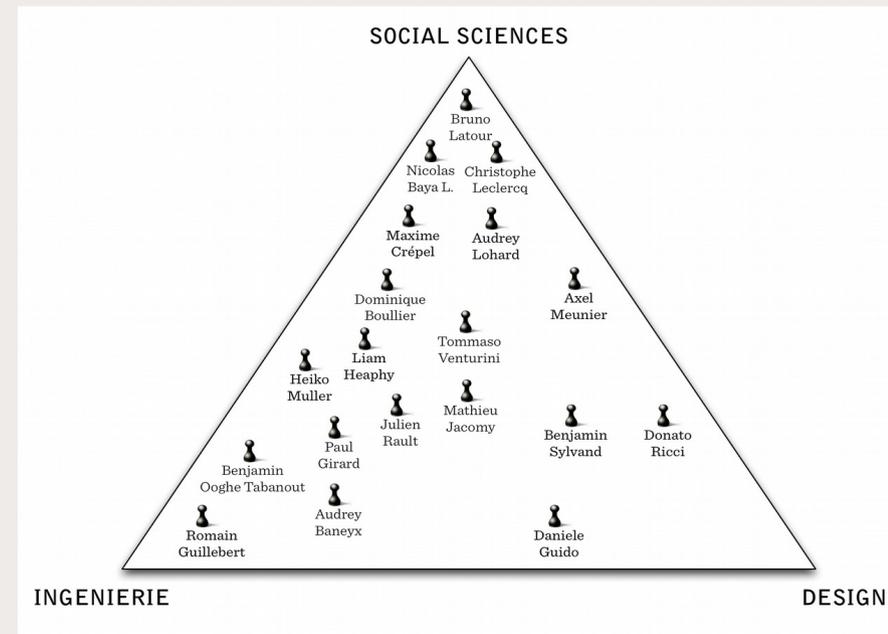
SciencesPo
MÉDIALAB



DIME-SHS

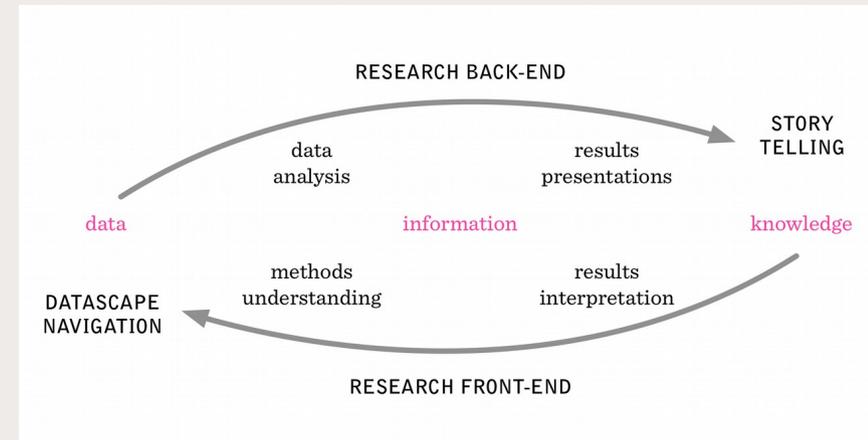
Le médialab de Sciences Po

- 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
<http://www.medialab.sciences-po.fr/atelier/>



L'instrument DIME-Web (Equipex DIME-SHS)

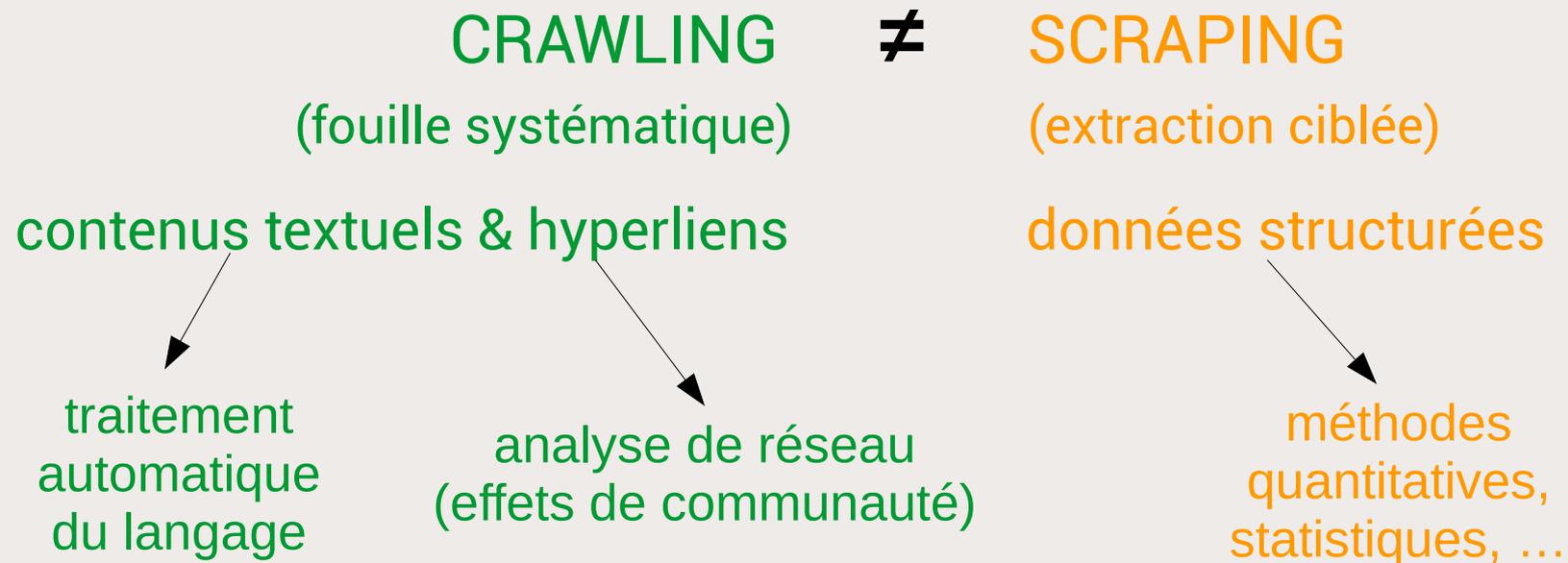
- Accompagnement technique et méthodologique à l'utilisation du Web comme nouveau terrain d'enquête
 - Support aux Sciences Humaines et Sociales
 - Extraction ciblée de contenus/discussions/traces
 - Création de corpus documentaire
 - Méthodes numériques, itératives
≠ tout automatique
- Equipex (+ Ellips + beQuali = DIME-SHS)
 - 2 personnes (Mathieu Jacomy et moi-même)
 - Objectif ANR d'auto-financement
 - offre de service payant avec sélection
 - mutualisation (logiciels libres)



Le Web : une source de données « sales »

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

→ très grande hétérogénéité (type de contenu & forme)

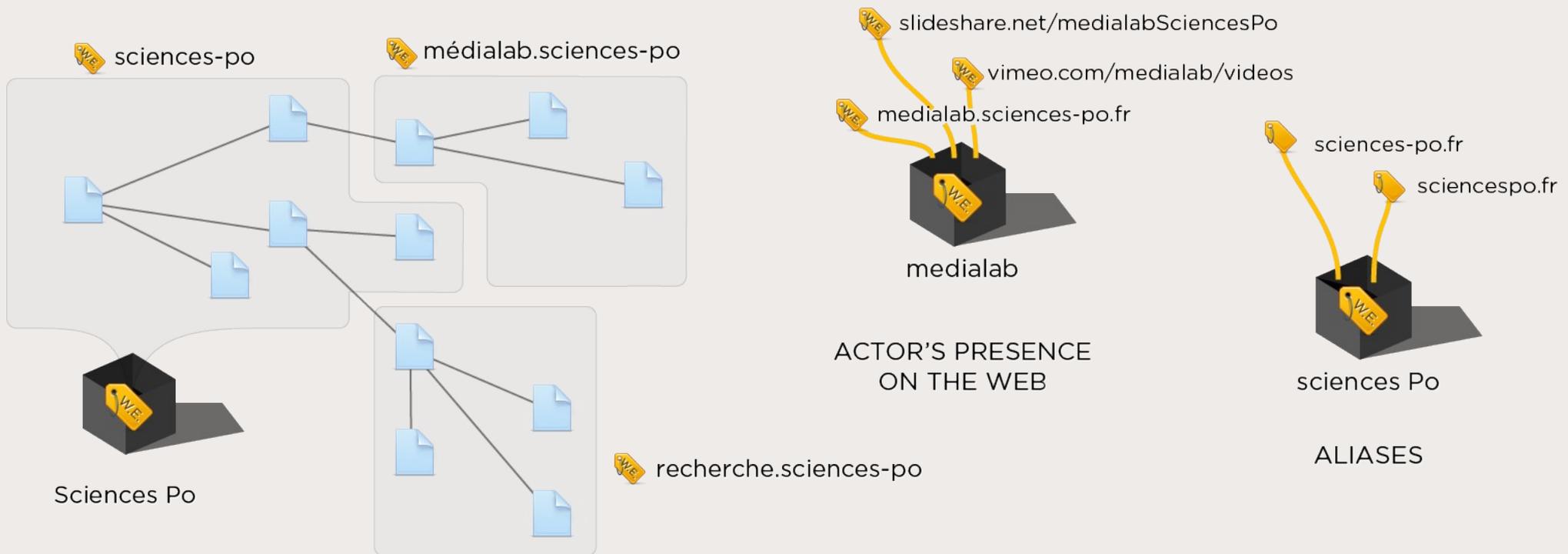


redirections, liens erronés, liens morts et sites disparus, encodage mal indiqué...

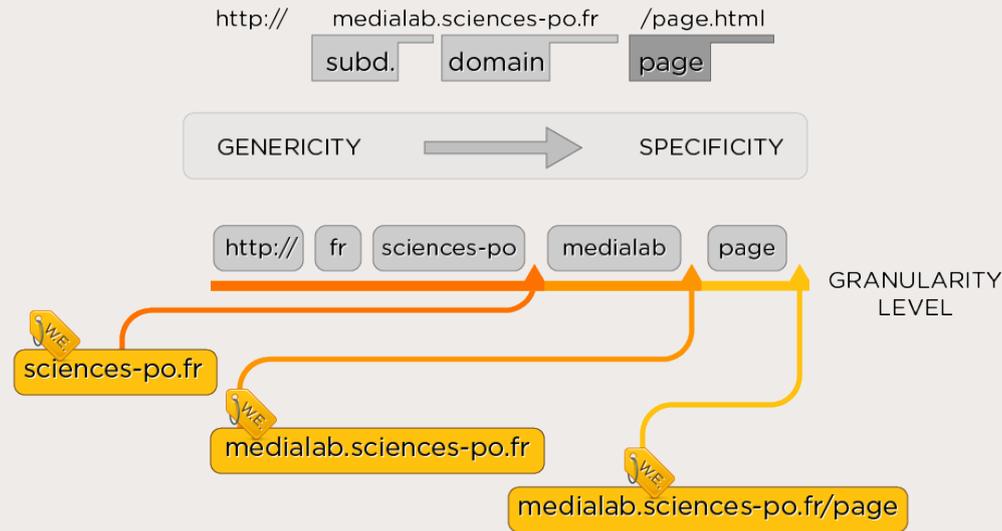
Principes méthodologiques : « WebEntités »

Comment gérer la diversité de granularité des sites web ?

→ « WebEntités » : agrégats reflétant des entités documentaires cohérentes du point de vue du chercheur



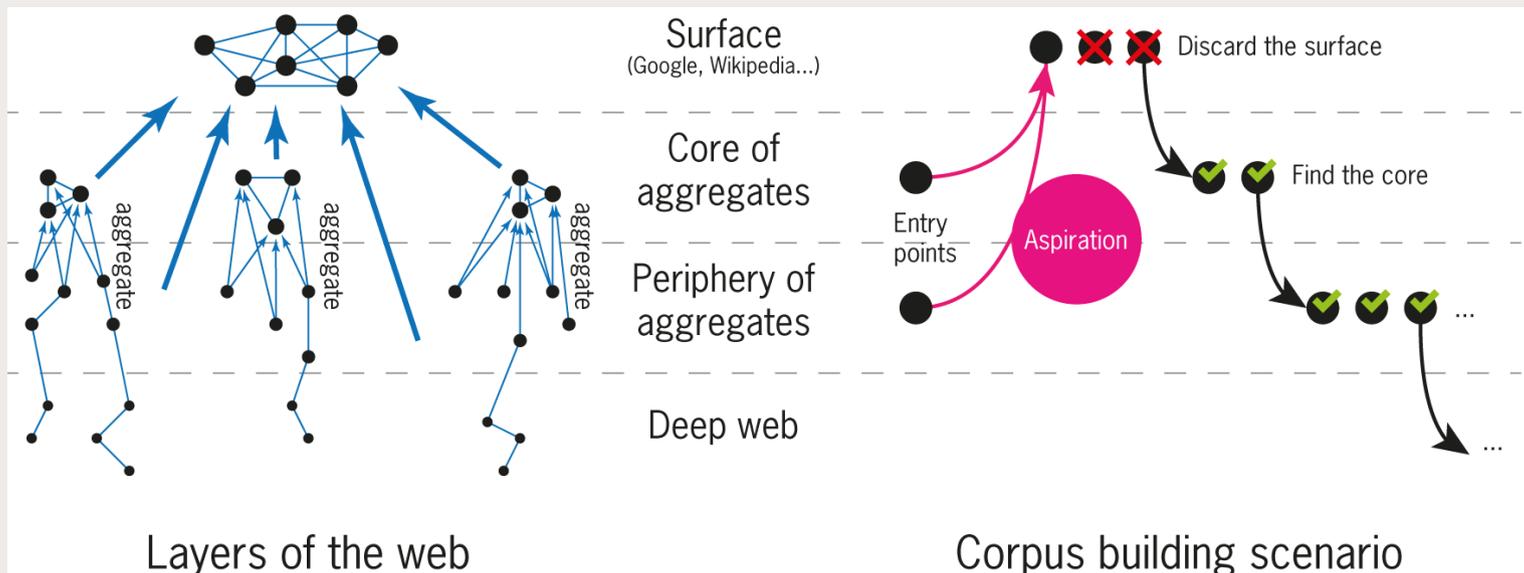
Principes méthodologiques : « WebEntités »



41	Amnesty.fr	http .fr amnesty www.
42	Facebook.com /.../326366925310	http .com facebook www. /pages /Andr%C3%A9-... /326366925310 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	Annuairemairie.com	http .com annuairemairie www.
44	Marianne2.fr /hervenathan	http .fr marianne2 www. /hervenathan Same web entity defined rows 651, 895 and 896
45	Anticor.org	http .org anticor
46	Desgouilles.fr	http .fr desgouilles david.

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 (virtuel)
→ 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

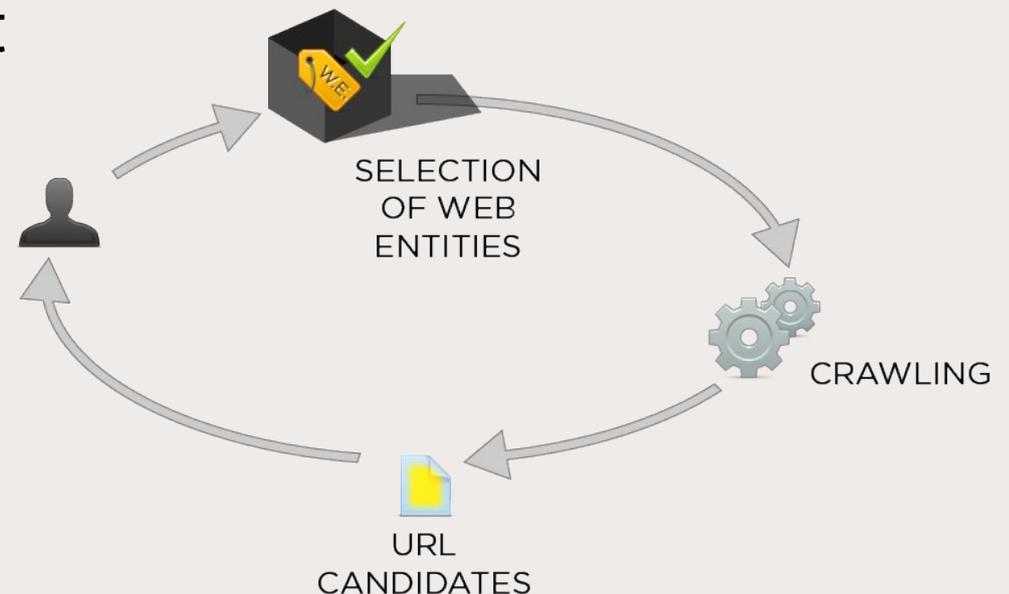
Gravatar.com X

Google.fr X

Set to UNDECIDED

Principes méthodologiques : « Prospection »

- Expansion éditorialisée et itérative du corpus
- Coût en temps humain : travail de curation répétitif
« 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

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

The screenshot shows the Hyphe Browser interface. At the top, the browser title is 'Hyphe Browser' and the address bar shows 'FrenchFarRight4'. Below the address bar, there is 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 page from '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 is a paragraph of text and a search bar with the number '7119753'. The page is divided into several sections: 'Présentation', 'Analyses et propositions', 'Priorités', 'Échanges', 'CECRI', 'Réflexions', 'Initiatives', and 'Divers'. On the left side, there is a sidebar with a list of related websites: 'erlille.wordpress.com', 'Leforumcatholique.org', 'Bernard-Antony.com', and 'Agoravox.fr'. At the bottom of the page, there is a section for 'Entretien Colibris avec Etienne Chouard'.

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

Catégoriser les WebEntités avec HyBro

The screenshot shows the HyBro interface with the following elements:

- Browser Title:** Hyphe Browser ABC111
- Navigation:** Back, Forward, and address bar showing `blogs.ei.Columbia.edu /.../u-s-drought-risk-wi...`
- Page Content:**
 - Navigation menu: AGRICULTURE CLIMATE EARTH SCIENCES ECOLOGY ENERGY HEALTH SUSTAINABILITY URBANIZATION WATER
 - Article Title: **WATER** U.S. Drought Risk Wider than Previously Thought
 - Author: BY LAKIS POLYCARPOU | MAY 4, 2015
 - Text: 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.
- Right Sidebar:**
 - Education News: Alumna Planting 'Seeds' for Sustainable Education in Africa
 - FROM THE FIELD: Image of a field.
- Left Panel (WebEntity Management):**
 - Statut: Crawlée | ✓, Citée par 1 WE
 - Contexte: Dropdown menu
 - Annotations: Input field with "water", "Créer le tag: 'water'", and "Ajouter une catégorie +"
 - Lang: en
 - type: blog
- Top Bar (Status):**
 - PROSPECTION: 8367
 - IN: 105
 - IN À TAGGER: 104
 - IN À CRAWLER: 7
 - UNDECIDED: 1
 - OUT: 497
- Footer:** en | fr Hyphe 0 3

<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

439 WEB ENTITIES WEB ENTITIES NETWORK

Display a category
Point de vue

Special filters

- Untagged
- Partially untagged
- Conflicts

Free Tags

- Untagged

Acteur

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

Search

Neutre

Contre les pesticides

Pour les pesticides

Neutre

Neutre

Contre les pesticides

Contre les pesticides

Neutre

Neutre

Contre les pesticides

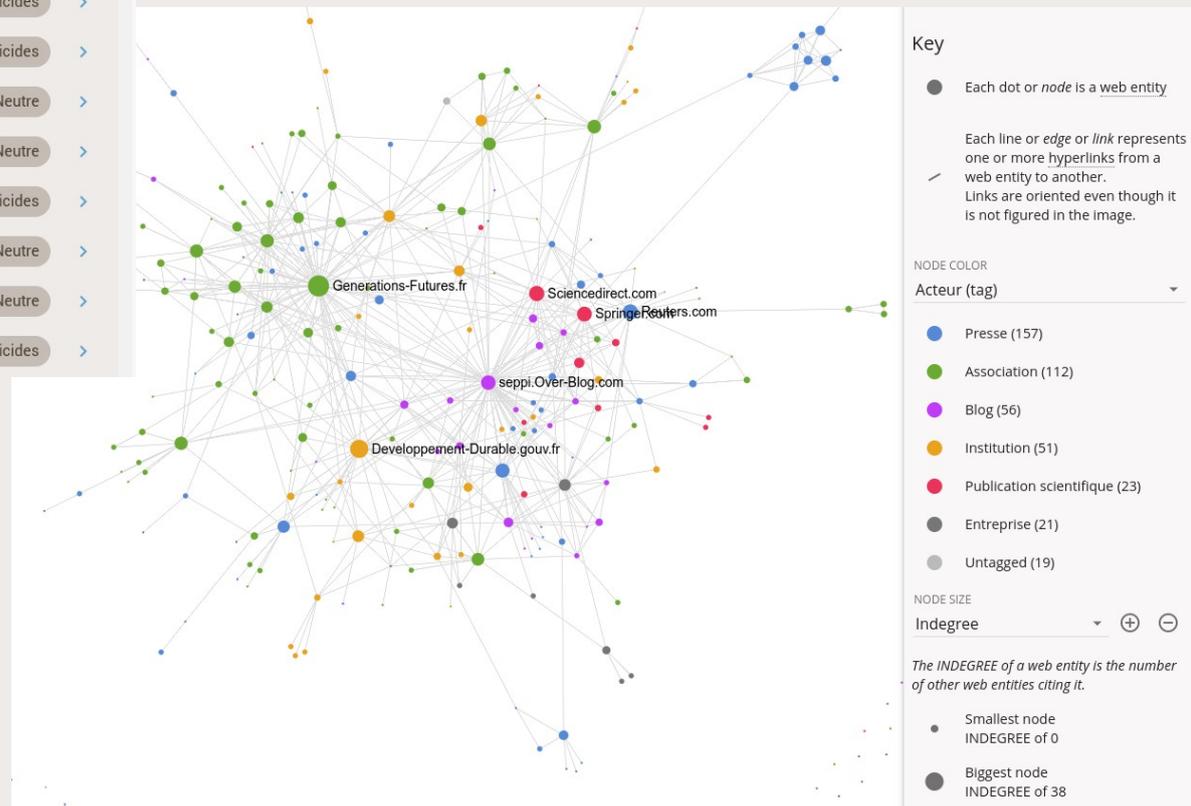
Neutre

Neutre

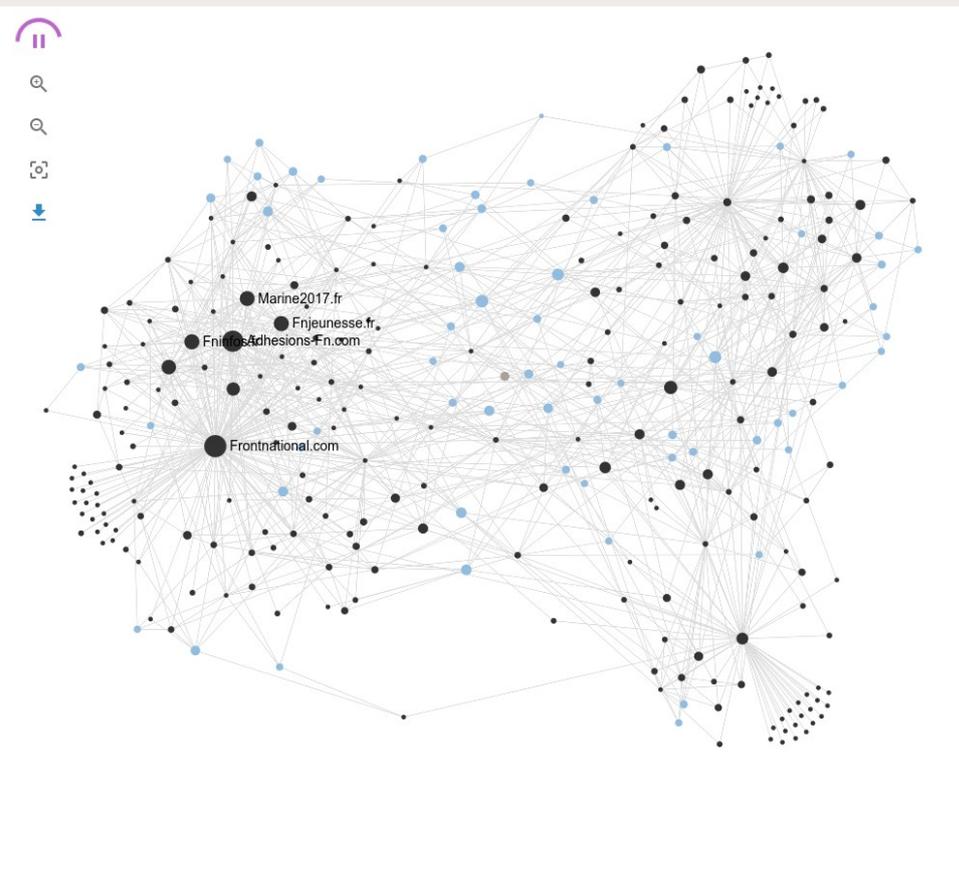
Contre les pesticides

Neutre

Contre les pesticides



Explorer le réseau des liens entre acteurs



Network Viz Settings

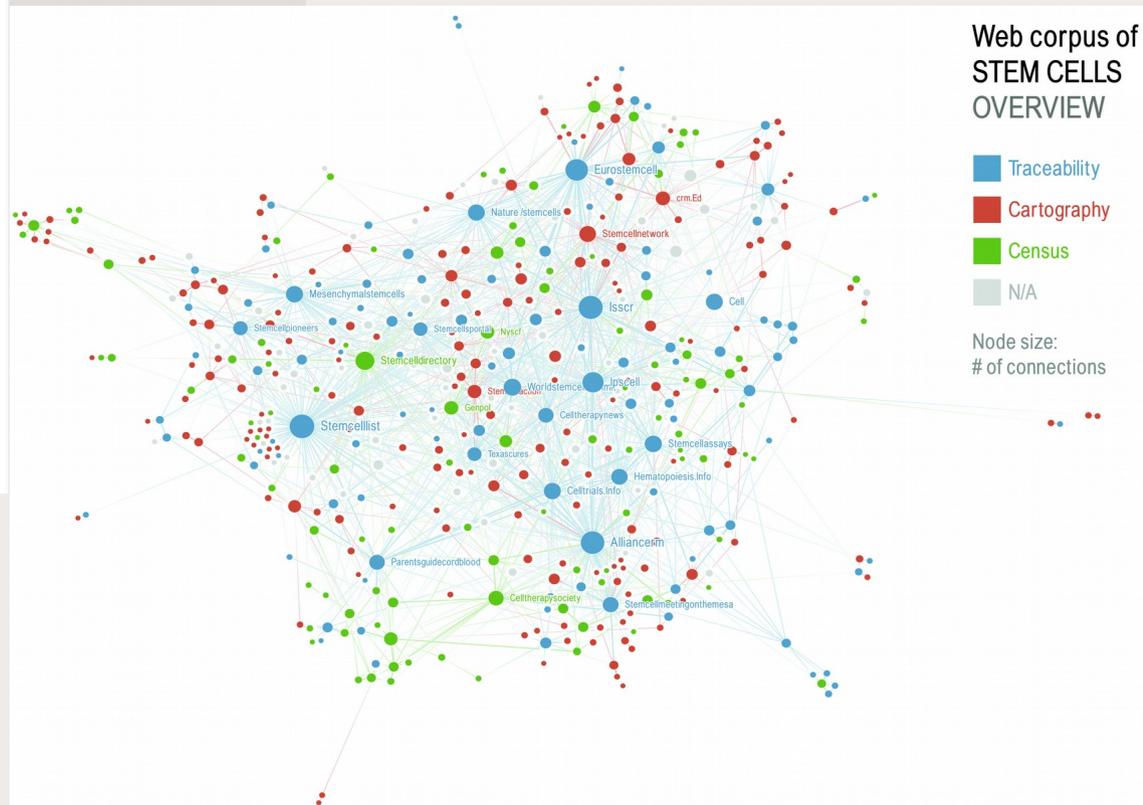
Filtering

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

Filter DISCOVERED web entities

Display only DISCOVERED with ...

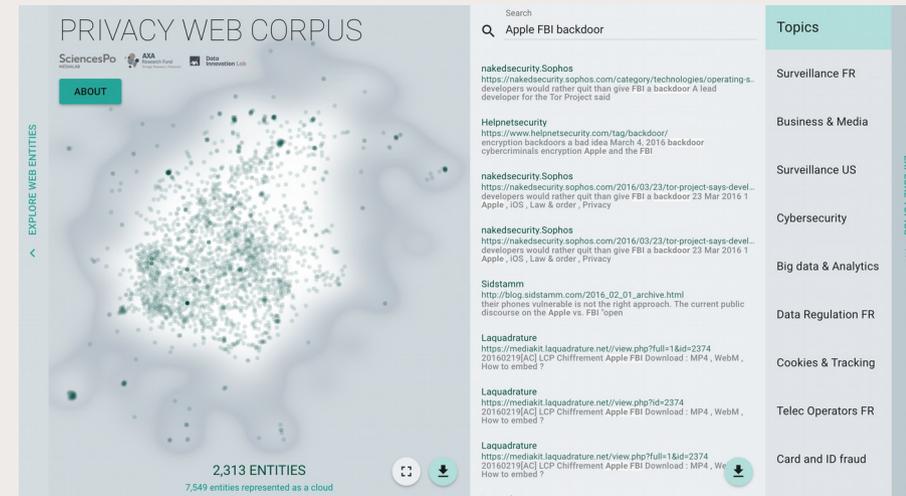
Filter ALL web entities



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
- Stabiliser PhantomJS pour le crawl browser-like (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

Merci de votre attention !

Et maintenant, à vous de jouer !

SciencesPo
MÉDIALAB

[@medialab_ScPo](#)

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 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