



# Comment utiliser le World Wide Web comme terrain d'enquête ?

Benjamin Ooghe, Paul Girard

## ► To cite this version:

Benjamin Ooghe, Paul Girard. Comment utiliser le World Wide Web comme terrain d'enquête ?. Action Nationale de Formation - Collecter et produire des données pour la recherche en SHS, Réseau MATE-SHS, Nov 2016, Fréjus, France. hal-03631542

**HAL Id: hal-03631542**

**<https://sciencespo.hal.science/hal-03631542>**

Submitted on 5 Apr 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

AXE 1 : De nouvelles sources de données, pourquoi faire ?

# Comment utiliser le World Wide Web comme terrain d'enquête ?

**Benjamin Ooghe-Tabanou - Paul Girard**

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

**SciencesPo**  
MÉDIALAB

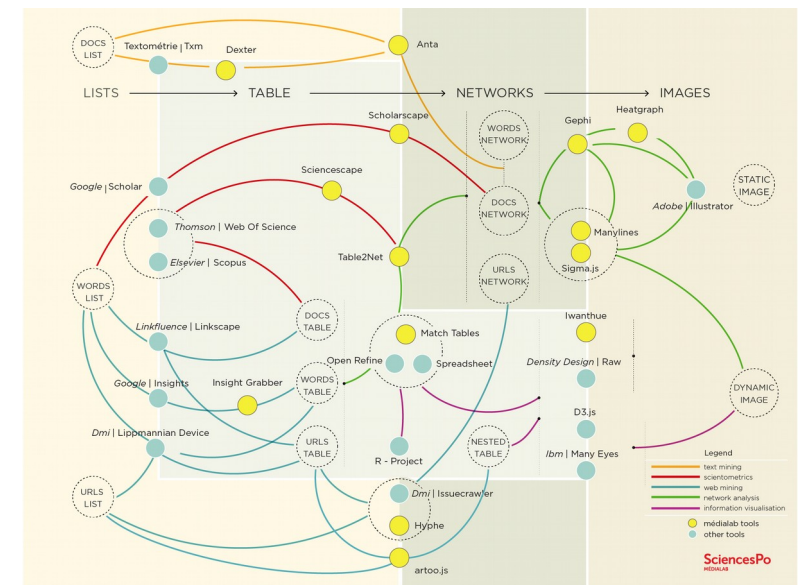
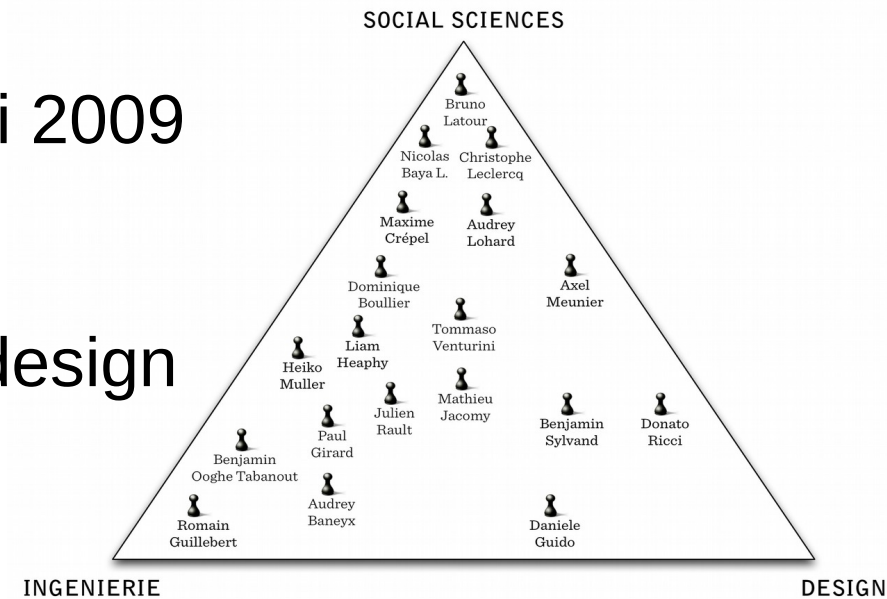


**DIME - SHS**

***Collecter et produire des données pour la recherche en SHS  
Fréjus, 15-18 novembre 2016***

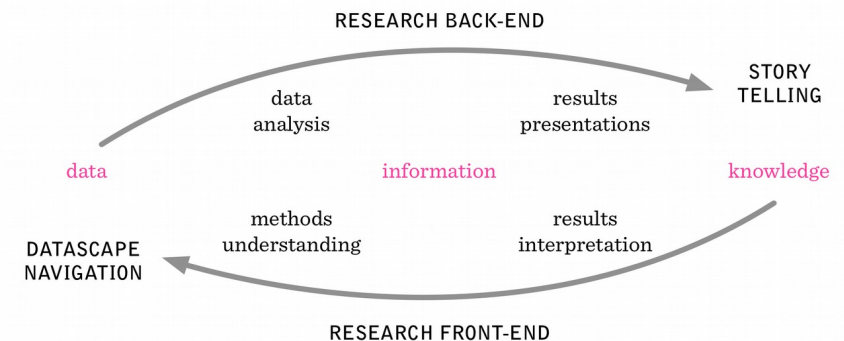
# Le médialab de Sciences Po

- Centre de recherche fondé en mai 2009 à Sciences Po par Bruno Latour
- 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>



# L'instrument DIME-Web

- 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



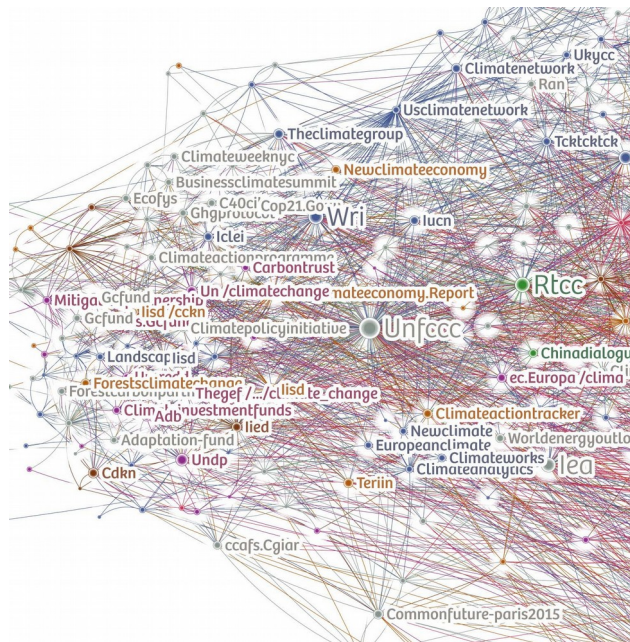
- Equipex (+ Ellips + beQuali = DIME-SHS)
  - 2 personnes (Mathieu Jacomy et moi-même)
  - Objectif ANR d'auto-financement
    - offre de service payant avec sélection
    - mutualisation (logiciels libres)





# Hyphe : un crawler orienté par la recherche

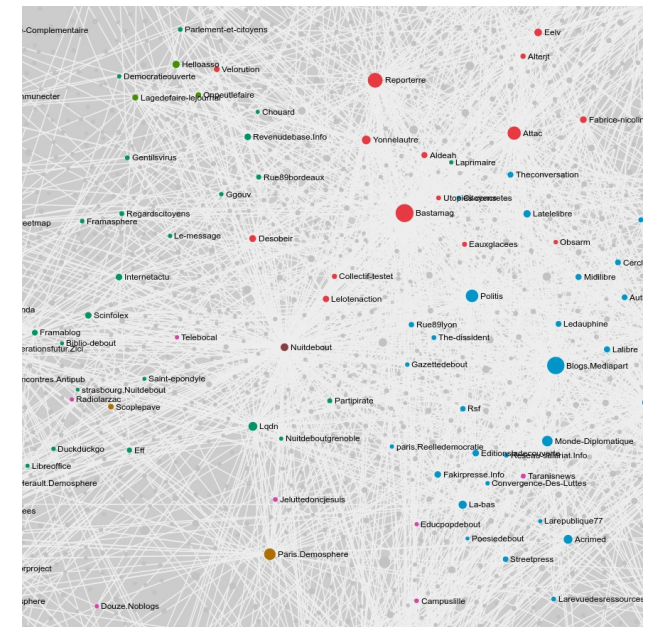
- Les liens hypertextes : nouveaux révélateurs de relations entre acteurs d'une thématique
- Créer un corpus documentaire
  - « acteurs web » & contenus textuels respectifs
  - liens hypertextes entre ces acteurs
- Études exploratoires ou de controverses dans tous les domaines



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

COP 21  
Vie privée  
Cellules souches  
Tissu associatif  
Fromages au lait cru  
Administrations culturelles  
Littérature

...

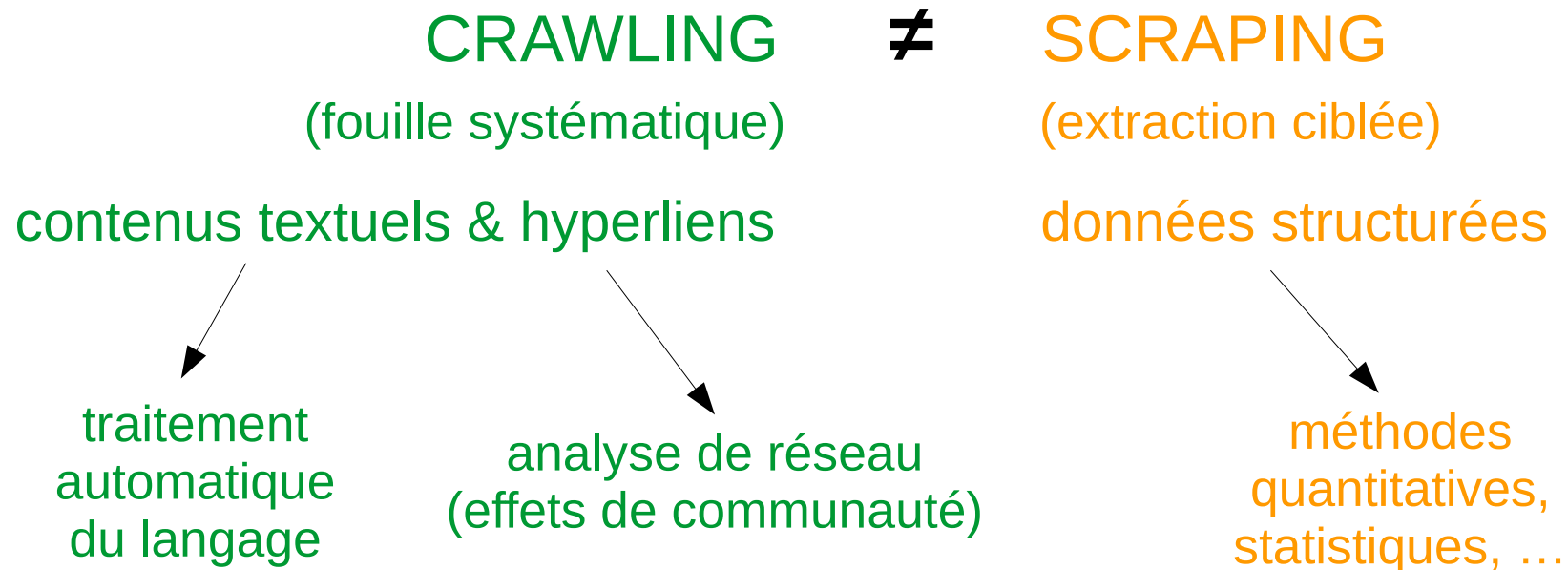


<http://utopies-concretes.org/>

# Le Web : une source de données « sales »

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

→ très grande hétérogénéité (type de contenu & forme)

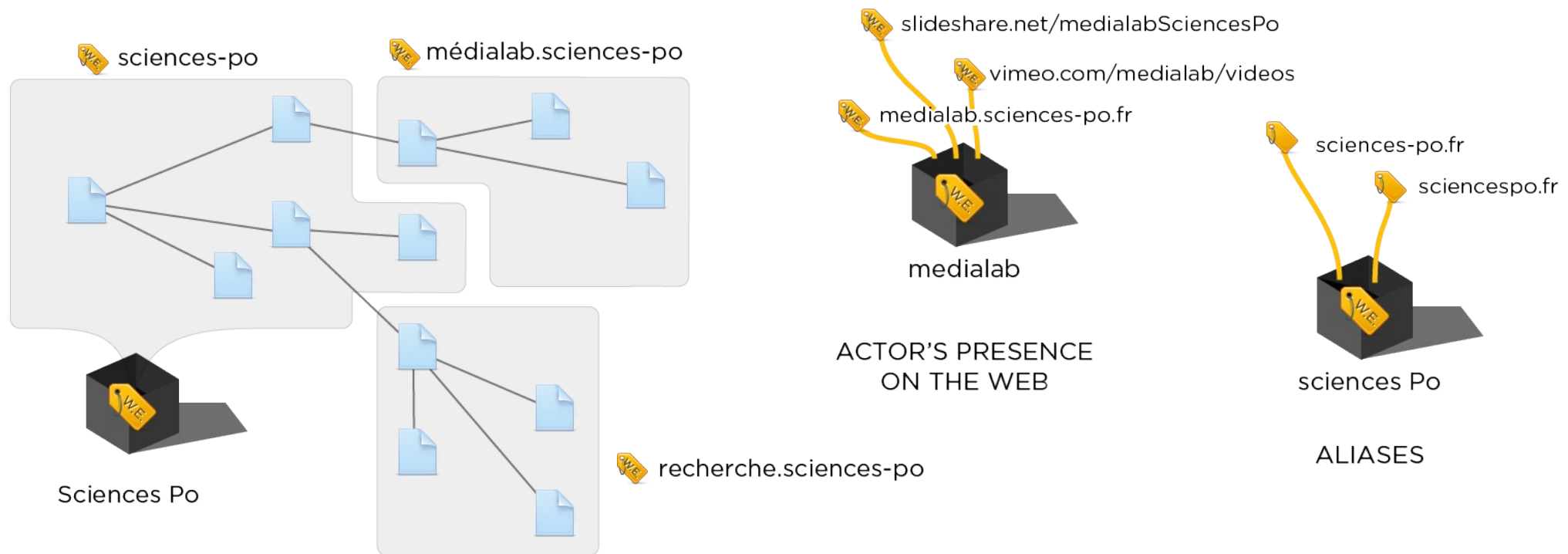


redirections, liens erronés, liens morts et sites disparus, encodage mal indiqué...

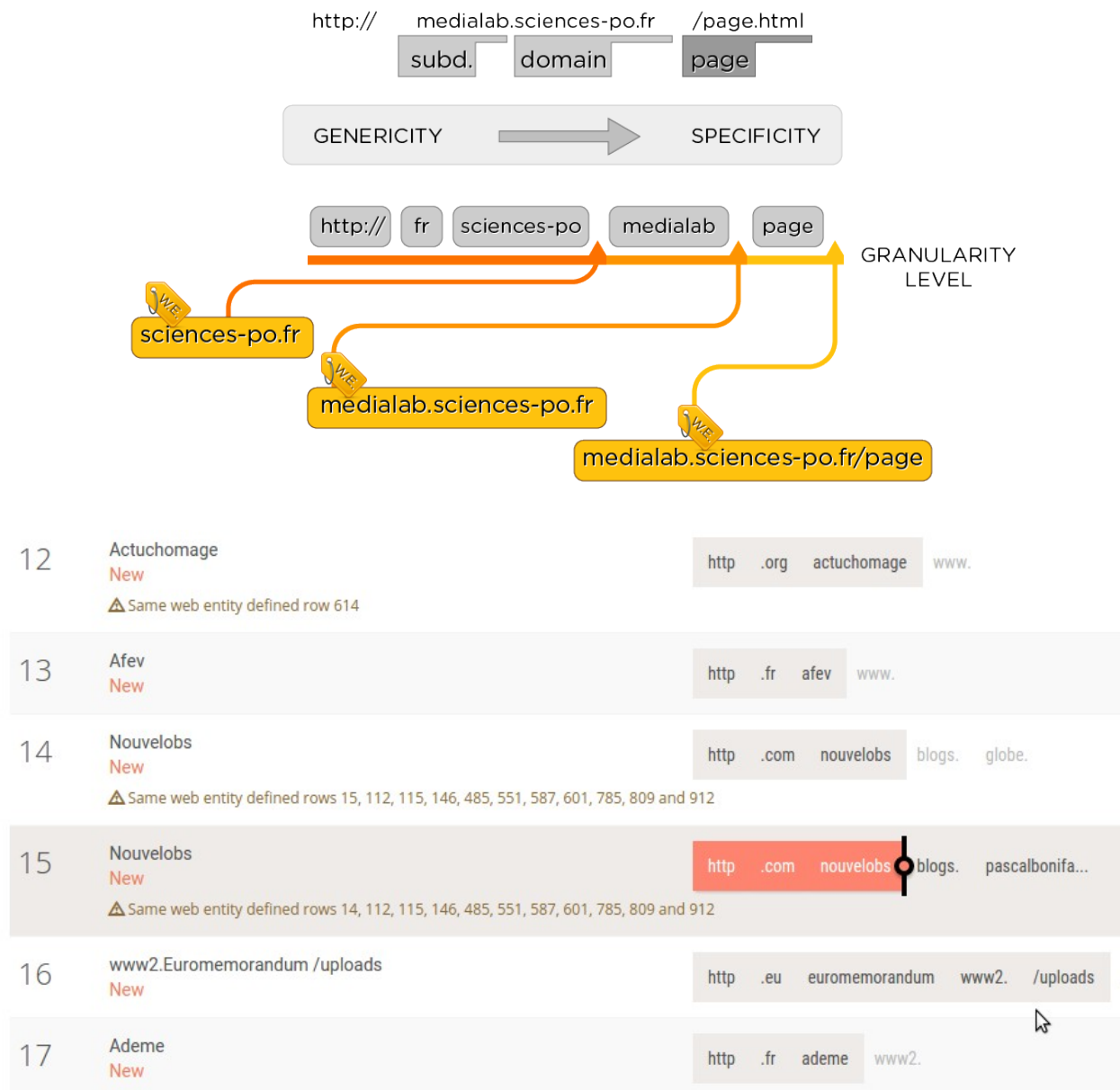
# Principes méthodologiques : « WebEntités »

Comment gérer la diversité de granularité des sites web ?

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



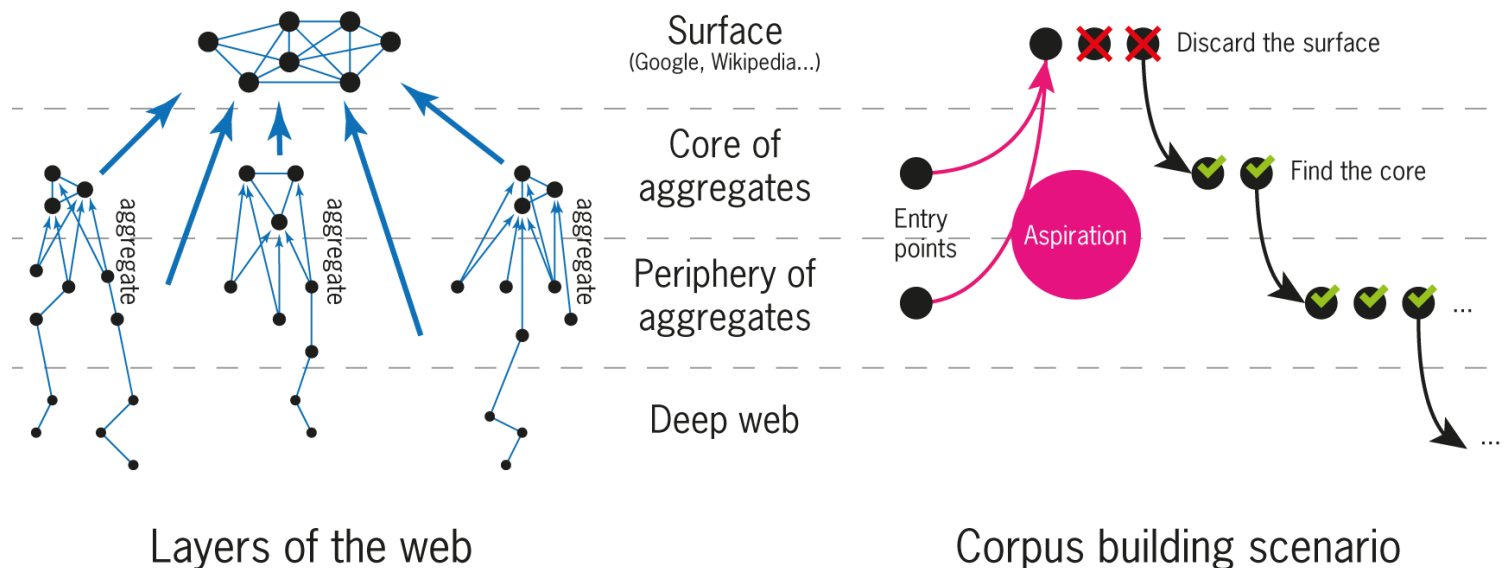
# Principes méthodologiques : « WebEntités »





# Principes méthodologiques : « Prospection »

- Démarrage : points d'entrées libres (recherche web qualitative, annuaire, liste de sites d'acteurs issue d'entretiens...)
- Crawler = robot qui fouille les pages web et clique sur les liens
  - Crawlers classiques : boule de neige (fouille systématique jusqu'à 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

**FILTERING** ▼

🔍 Type a query

**DISCOVERED** 2006 DISCOVERED WEB ENTITIES

	Name	In Cited
IN UND. OUT	Google.com	18
IN UND. OUT	Facebook.com /sharer.php	9
IN UND. OUT	plus.Google.com /share	9
IN UND. OUT	Environnement.brussels	7
IN UND. OUT	Facebook.com /unsupportedbrowser	7
IN UND. OUT	Fgov.be	7
IN UND. OUT	Google.be	7
IN UND. OUT	Instagram.com	7
IN UND. OUT	Vimeo.com	7
IN UND. OUT	Apple.com	6
IN UND. OUT	Bruxellesenvironnement.be	6
IN UND. OUT	Cim.be	6
IN UND. OUT	leb.be	6
IN UND. OUT	Twitter.com /intent	6
IN UND. OUT	Twitter.com /share	6
IN UND. OUT	Facebook.com	5
IN UND. OUT	Facebook.com /events	5
IN UND. OUT	Fian.be	5

**2 SET TO IN** ▼

Bruxellesenvironnement.be ✕

Environnement.brussels ✕

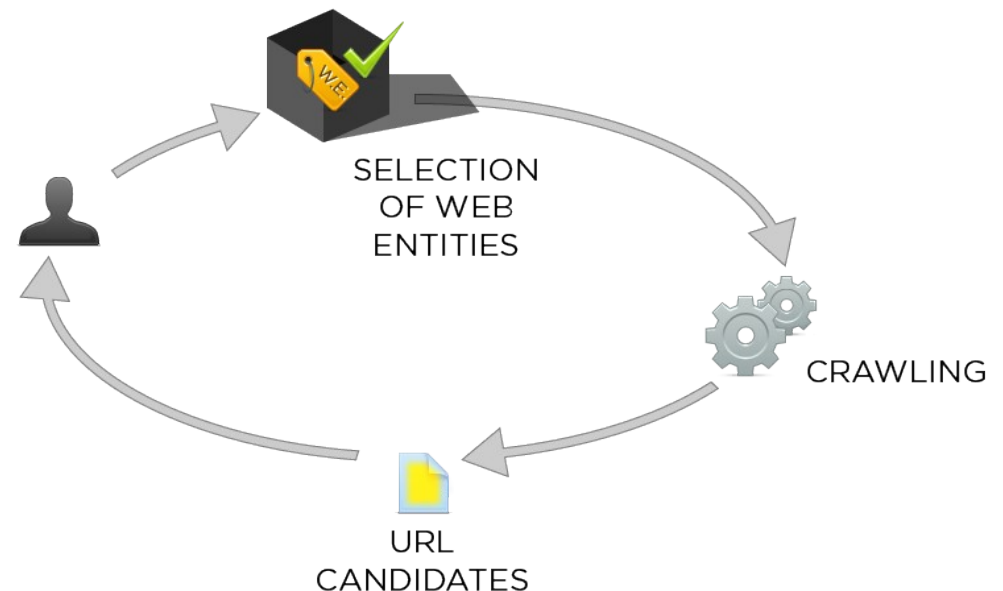
**CRAWL**

**3 SET TO UNDECIDED** ➤

**12 SET TO OUT** ➤

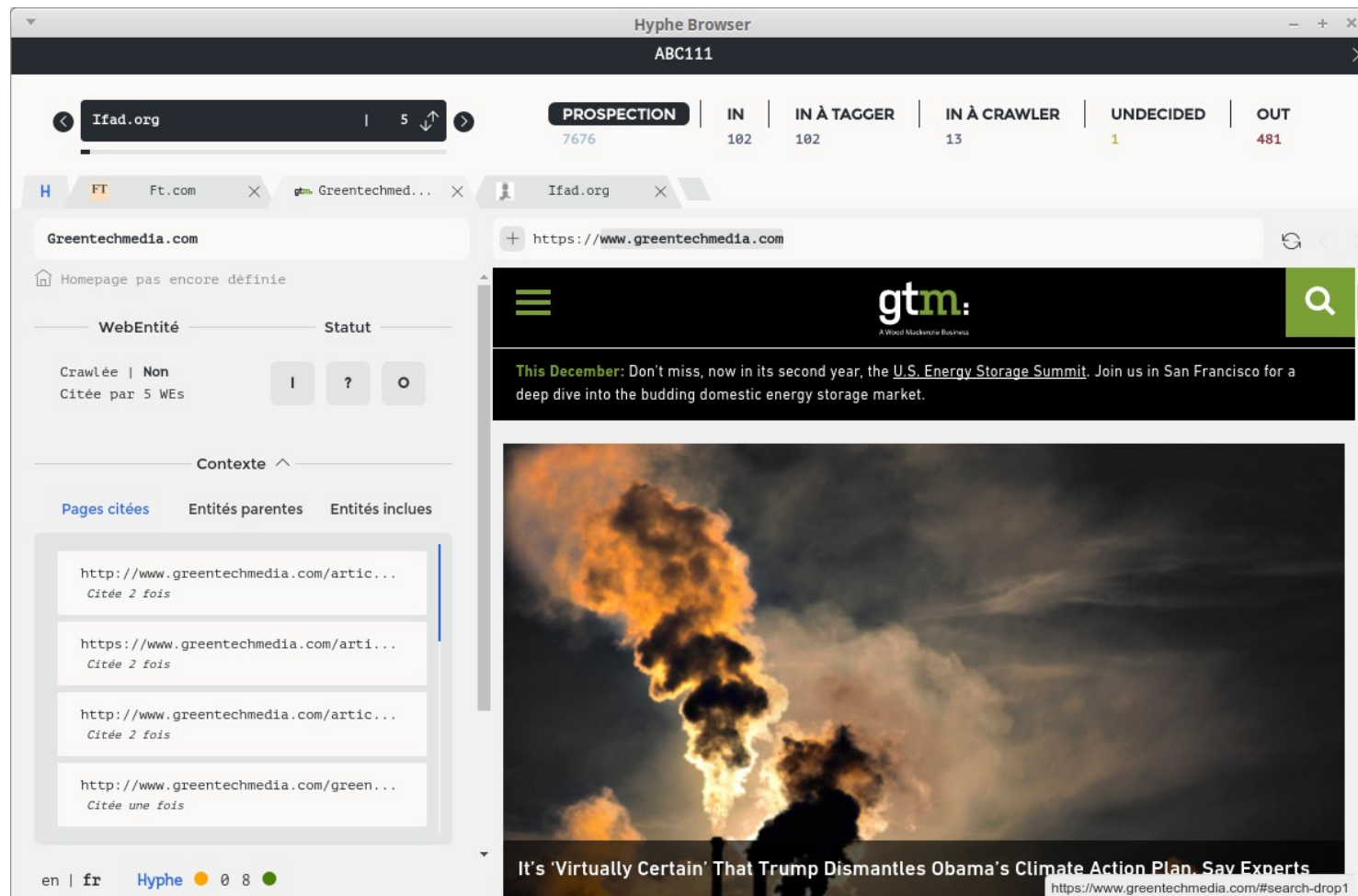
# Principes méthodologiques : « Prospection »

- 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



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

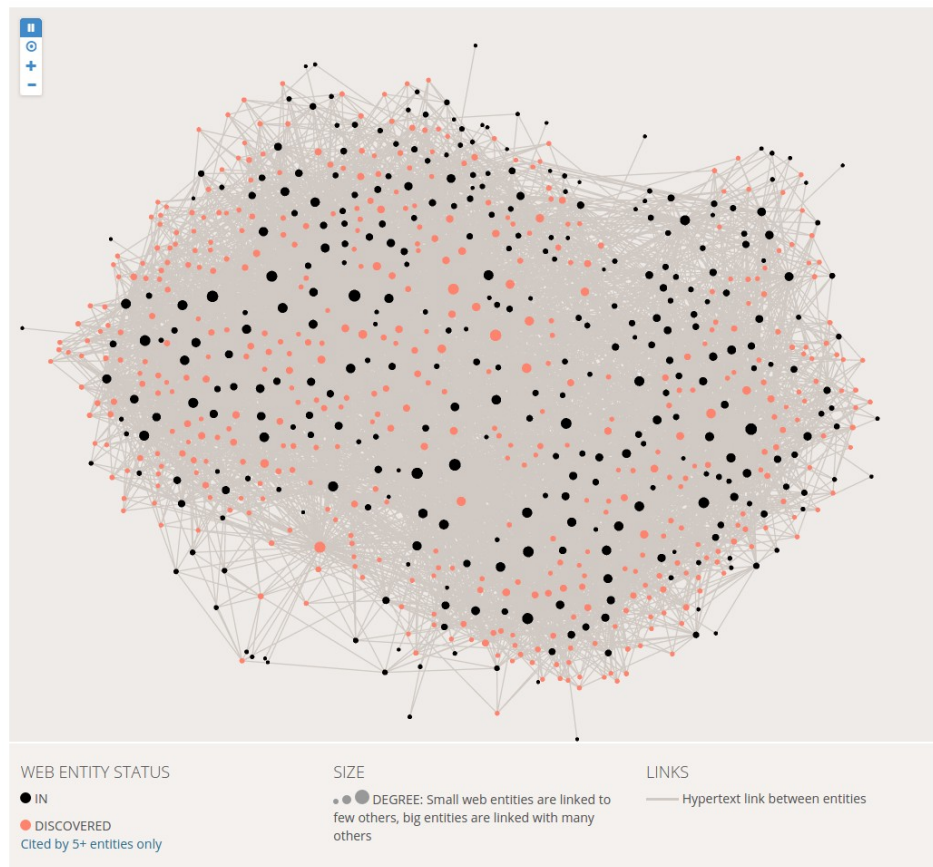
# Catégoriser les WebEntités avec HyBro

The screenshot displays the HyBro web browser interface. The top navigation bar includes a search bar with the URL 'blogs.ei.Columbia.edu /.....' and a status bar showing 'PROSPECTION 8367', 'IN 105', 'IN À TAGGER 104', 'IN À CRAWLER 7', 'UNDECIDED 1', and 'OUT 497'. The main content area shows a news article titled 'U.S. Drought Risk Wider than Previously Thought' by LAKIS POLYCARPOU, dated MAY 4, 2015. The article is categorized under 'WATER' and 'FROM THE FIELD'. The left sidebar contains a 'WebEntité' section with a 'Statut' dropdown (set to 'Crawlée'), a 'Contexte' dropdown, and an 'Annotations' section with a search bar containing 'water'. The bottom of the sidebar shows a language selector (en | fr) and a 'Hyphe' status bar (0 3 ●).

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



# Explorer le réseau des liens entre acteurs



**FILTERS**

Presets

CORPUS ONLY PROSPECTION FULL DATA

☒ IN

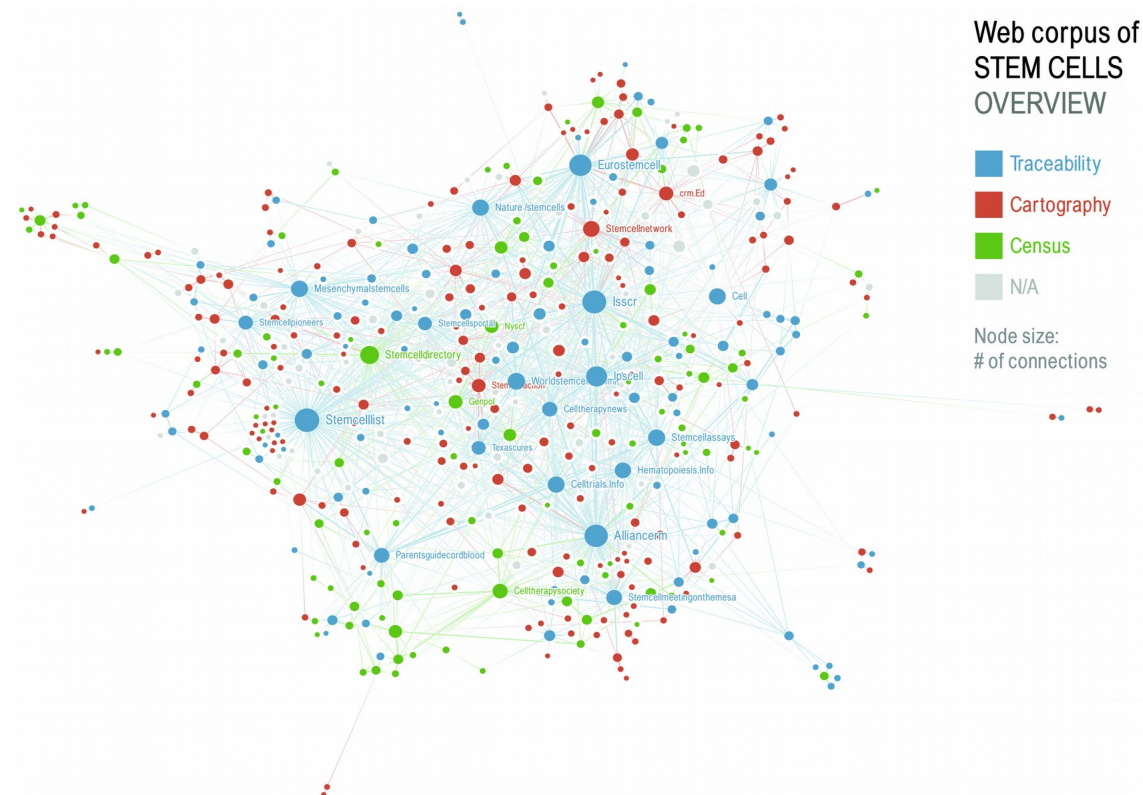
☐ UNDECIDED

☐ OUT

☒ DISCOVERED Cited by 5+ entities only

**DOWNLOAD NETWORK**

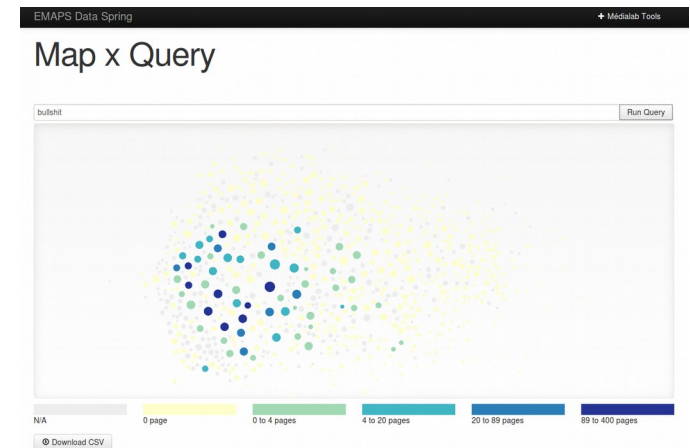
ⓐ DOWNLOAD GEXF



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

# Et pour la suite ?

- Import / export de listes de webentités ou de corpus
  - duplication, reproduction
  - exploration longitudinale dans le temps
- Interface de catégorisation avancée (tags)
- Exploitation des contenus textuels dans les pages crawlées et analyse automatique du langage
- Contrôle qualité des crawls et du corpus
- Stabiliser PhantomJS pour le crawl browser-like (Facebook, etc.)
- 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 !

---

Venez mieux découvrir et essayer Hyphe  
et les autres outils du médialab lors de nos  
2 ateliers pratiques (B1 & B2) mercredi à 18h30 !

**SciencesPo**  
MÉDIALAB

[@medialab\\_ScPo](https://twitter.com/medialab_ScPo)

benjamin.ooghe@sciencespo.fr