



**HAL**  
open science

## Utiliser Hyphe en sciences sociales

Benjamin Ooghe

► **To cite this version:**

Benjamin Ooghe. Utiliser Hyphe en sciences sociales. Colloque DIME-SHS "Des instruments au service de la recherche en sciences sociales", Sciences Po, Sep 2018, Paris, France. hal-03621701

**HAL Id: hal-03621701**

**<https://sciencespo.hal.science/hal-03621701v1>**

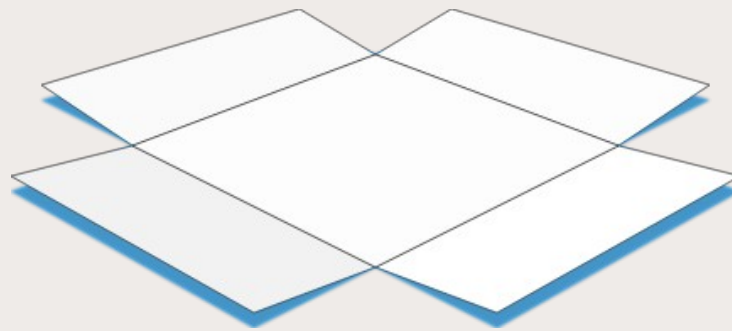
Submitted on 28 Mar 2022

**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



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

# DIME Web

Des outils et méthodes numériques pour  
exploiter le Web comme terrain d'enquête

## Utiliser Hyphe en sciences sociales

28 septembre 2018

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

DIME Web

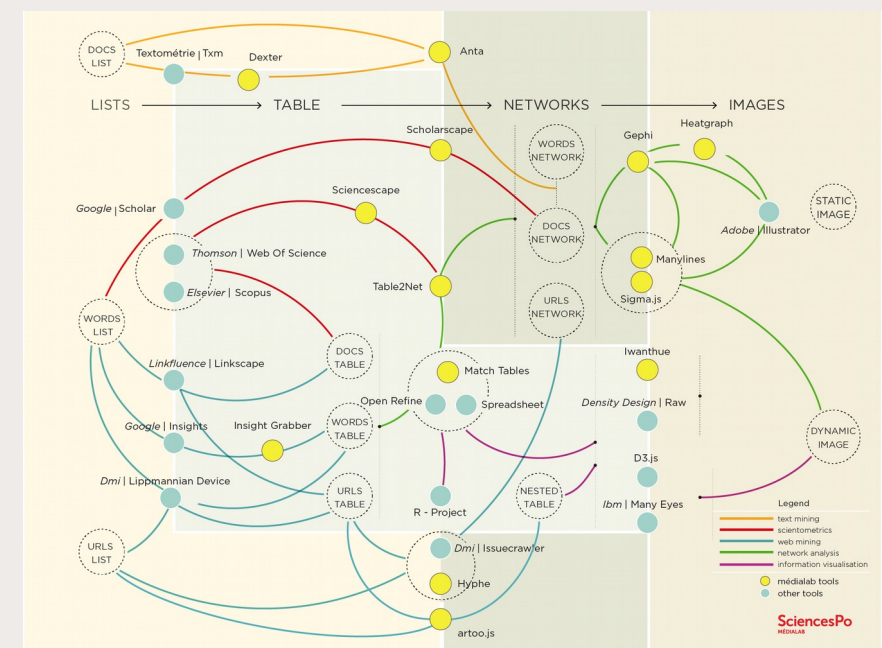
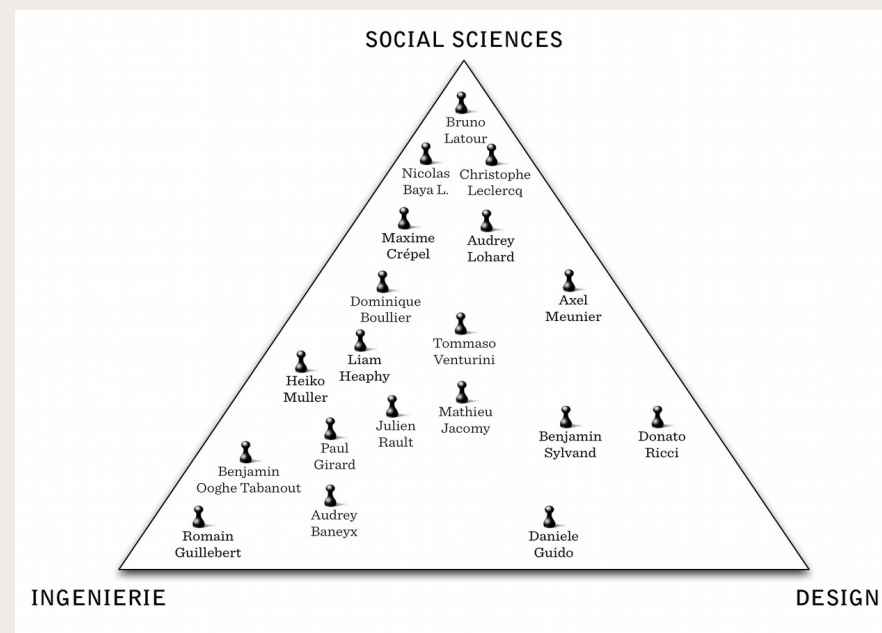
**SciencesPo**  
MÉDIALAB

**SciencesPo**  
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 : le METAT  
<http://www.medialab.sciences-po.fr/atelier/>

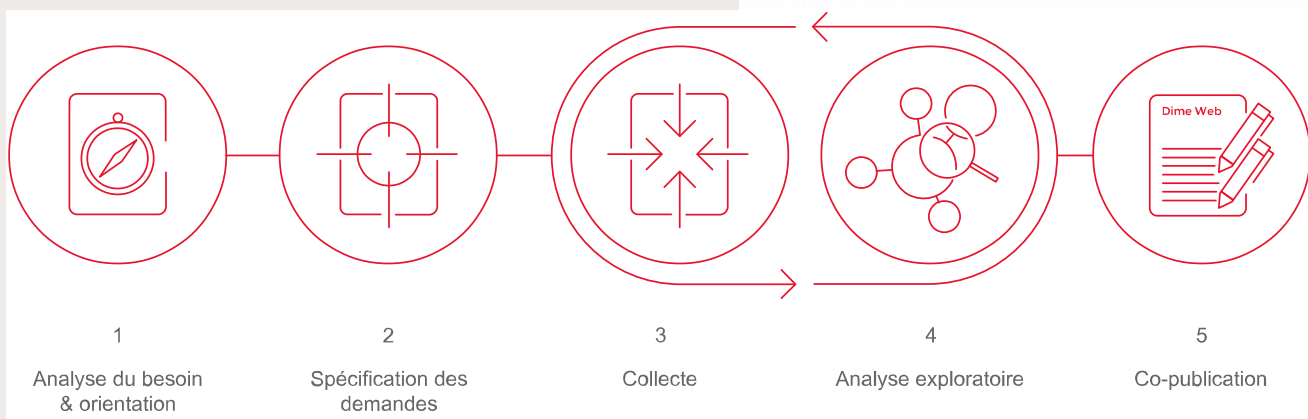
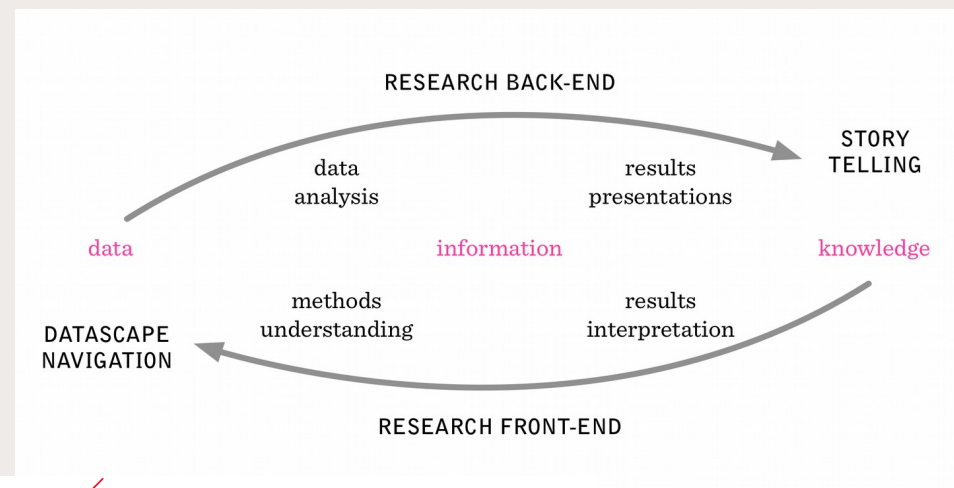


# L'instrument DIME Web

<http://dimeweb.dime-shs.sciences-po.fr>

Accompagnement technique et méthodologique à l'utilisation du Web comme nouveau terrain d'enquêtes

- Collecter, enrichir, nettoyer, visualiser et analyser des traces numériques
- Analyse de réseaux, archivage du web, analyse de controverses (ANT)
- Développement d'outils génériques
- Extraction ciblée de contenus
- Publication d'outils Open Source
- Méthodes numériques & itératives  
≠ tout automatique



# Le Web : une source de données « sales »

Collection de documents web (pages) sur une thématique

→ très grande hétérogénéité (type de contenu & forme)

**CRAWLING**

≠

**SCRAPING**

(fouille systématique)

(extraction ciblée)

contenus textuels & hyperliens

données structurées

traitement  
automatique  
du langage

analyse de réseau  
(effets de communauté)


méthodes  
quantitatives,  
statistiques...

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

# Table2Net : faire un réseau à partir d'un CSV

<http://tools.medialab.sciences-po.fr/table2net/>

- Générer un réseau de liens entre éléments à partir des données d'un fichier tableur




## Table 2 Net

### Load your CSV table

It has to be **comma-separated** and the first row must be dedicated to **column names**.

#### 1. Type of Network

Normal (one type of node)



You will have to choose:

- Which column **X** will define the nodes
- Which column **Y** will define the links

#### 2. Nodes

Which column defines the nodes?

hashtags

Pipe-separated "]"

Sample of nodes extracted with these settings: (sample)

#goweser #adventurebike #kotaposo #dh #xcbike

### 3. Links

Which column defines the links?

Row number

One expression per cell

Sample of items extracted with these settings: (sample)

5252 3621 1816 1847 4562

### 4. Additional settings

Optional: time series

No temporal data

Select only a column containing integers.

Optional: edge weight

Weight the edges

### 5. Build the network

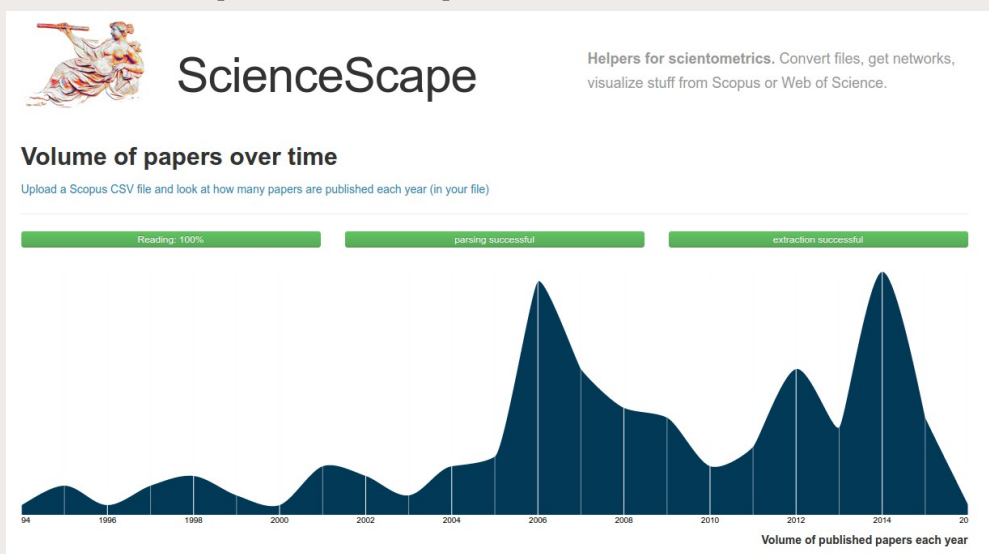
Build and download the network (GEXF)

NB: this may take a while, please be patient.

# ScienceScape : scientométrie en quelques clics

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

- Explorer et analyser les auteurs, mots-clés et revues d'un corpus de publications issu de Scopus ou WebOfScience

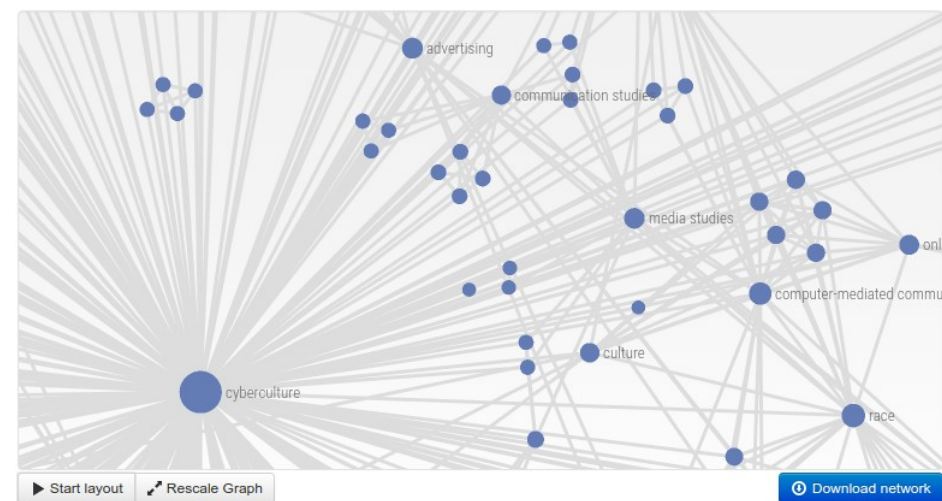


**Settings**

Type of network

Filtering

## Network preview



# SeeAlsology : exploration sémantique rapide

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

- Explorer le réseau des liens présents dans les sections « Voir aussi », « Articles connexes » des pages Wikipedia

Non connecté Discussion Contributions Créer un compte Se connecter

## Humanités numériques

Les **humanités numériques** (ou *digital humanities*, abrégées "DH", voire **humanités digitales**<sup>2</sup>) sont un domaine de recherche, d'enseignement et d'ingénierie au croisement de l'informatique et des arts, lettres, sciences humaines et sciences sociales.

**Sommaire** [afficher]

### Définition

Les humanités numériques peuvent être définies comme l'application du « savoir-faire des technologies de l'information [et de l'informatique/infosciences] aux questions de sciences humaines et sociales »<sup>3</sup>.

### Voir aussi

### Logiciels

- Gephi est un logiciel libre *open source*, issu du projet *e-Diaspora*, permettant la visualisation, l'analyse et l'exploitation en temps réel de données relationnelles ou réseaux.
- IRaMuTeQ est un logiciel libre d'analyse de texte, développé par **Pierre Ratinaud**.
- **Voyant Tools** permet de visualiser et d'explorer des textes
- Prospero (PROgramme de Sociologie Pragmatique, Expérimentale et Réflexive sur Ordinateur - © Doxa) est un logiciel d'analyse de données textuelles qualifié par ses concepteurs de technologie littéraire pour les sciences humaines. Le logiciel a été conçu par le sociologue Francis Chateauraynaud et l'informaticien Jean-Pierre Charriau.
- **Phlcarto** est un logiciel de cartographie. Le code n'en est pas libre, mais le logiciel est gratuit (freeware). Il fonctionne sur Windows.
- OpenRefine est un logiciel libre et gratuit de lissage de données (anciennement nommée Google refine).
- Le projet DIRT recense de très nombreux logiciels: DIRT@ [archive] (*Digital Research Tools - en Anglais*).

### Articles connexes

- Bibliothèque numérique
- Fouille de textes
- Littérature numérique
- Logométrie
- Moteur de recherche

Paste your list of wikipedia articles here or [try an example](#)

[https://fr.wikipedia.org/wiki/Humanit%C3%A9s\\_num%C3%A9riques](https://fr.wikipedia.org/wiki/Humanit%C3%A9s_num%C3%A9riques)

Stop words (press enter or separate the works with a comma)

Wikipedia: x Category: x File: x wikisource: x Commons: x  
 liste d x index d x catégories d x portail x désambiguisation x  
 résumé d x Catégorie: x Fichier: x add a word and press Enter

Distance  Parent links

**START CRAWLING** **DOWNLOAD** **CLEAR CACHE**

Click a node to visit it on Wikipedia

● seeds ● level -1 ● level 0 ● level 1 ● level 2

Ctrl+Click a node to add it to the seeds





# CatWalk : sélection qualitative de tweets

<https://medialab.github.io/catwalk/>

- Passer en revue rapidement « à la Tinder » tous les tweets d'un CSV pour décider de les inclure / exclure d'un corpus

The screenshot displays the CatWalk web interface. At the top left, the text 'CATWALK' is visible. To its right are navigation buttons: 'prev', a central input field containing '0', and 'next'. Further right is a 'Download' button with a counter showing '0', '434', and '2'. A prominent green bar with the text 'IN' is centered above the main content. The main content area features a tweet from 'RE-WORK @teamrework'. The tweet text reads: 'Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free [ow.ly/4nfo2S](http://ow.ly/4nfo2S) #AI @open\_ai' with a timestamp of '7:15 PM - 29 Apr 2016'. Below the text is a photo of Elon Musk. The tweet's content is partially truncated, showing 'Inside OpenAI, Elon Musk's Wild Plan to...' and 'OpenAI wants to give away the 21st century's most transformative technology. In wired.com'. At the bottom of the tweet are icons for reply, retweet (7), and like (15). To the right of the tweet is a vertical sidebar with five buttons: 'previous', 'next', 'IN' (highlighted in green), 'OUT' (highlighted in red), 'UNDECIDED' (highlighted in grey), and 'save'. At the bottom of the interface, there is a footer area with the profile picture and name '@teamrework' on the left, and the full tweet text and URL on the right.

# 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

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

**Redirect to Classic Google**  
Which language? en  
How many results per page? 100  
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 uris per page)  
103 new results in this page  
Keep existing results & continue to the next page  
Download CSV with 103 uris

→ « Import urls » dans Hyphe

# Un projet logiciel phare : Hyphe

- 10 releases entre juillet 2013 (v.0.0.0) et août 2018 (v1.0.3)

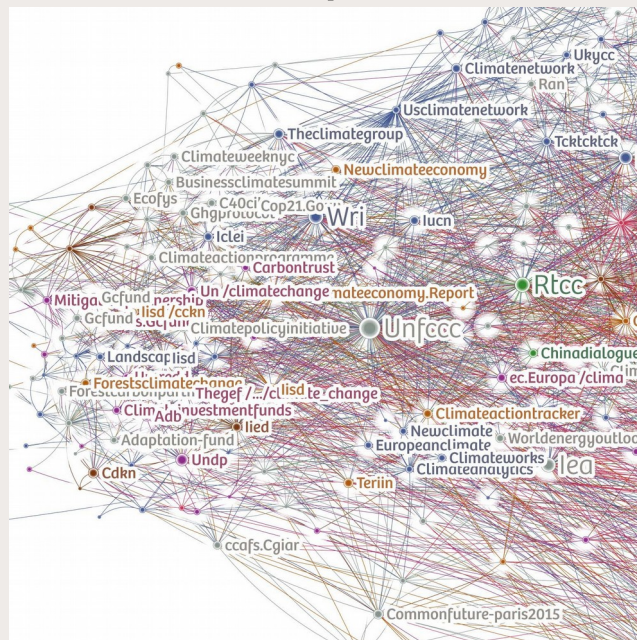


- Plusieurs dizaines de présentations, sessions pratiques et formations assurées
- + de 50 déploiements externes à notre connaissance (France, UK, DK, USA...)
- + de 150 étudiants formés à l'analyse du web chaque année avec Hyphe
- + de 500 utilisateurs testant la démo chaque année

# Hyphe : un crawler orienté par la recherche

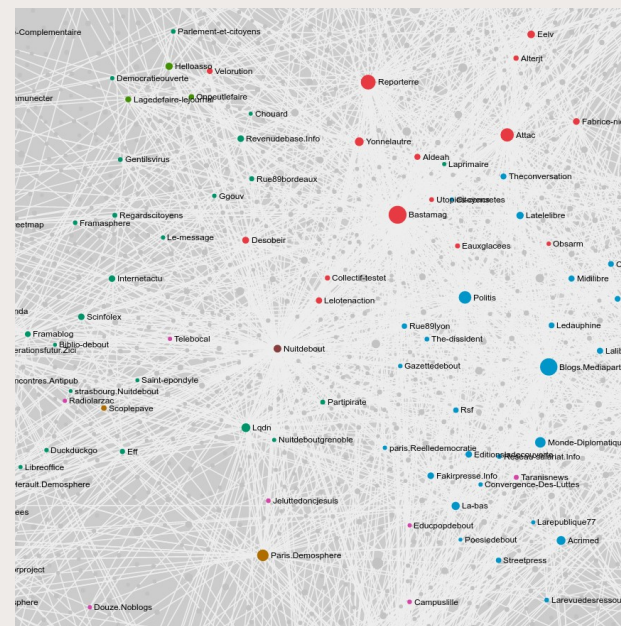
<http://hyphe.medialab.sciences-po.fr/demo>

- 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

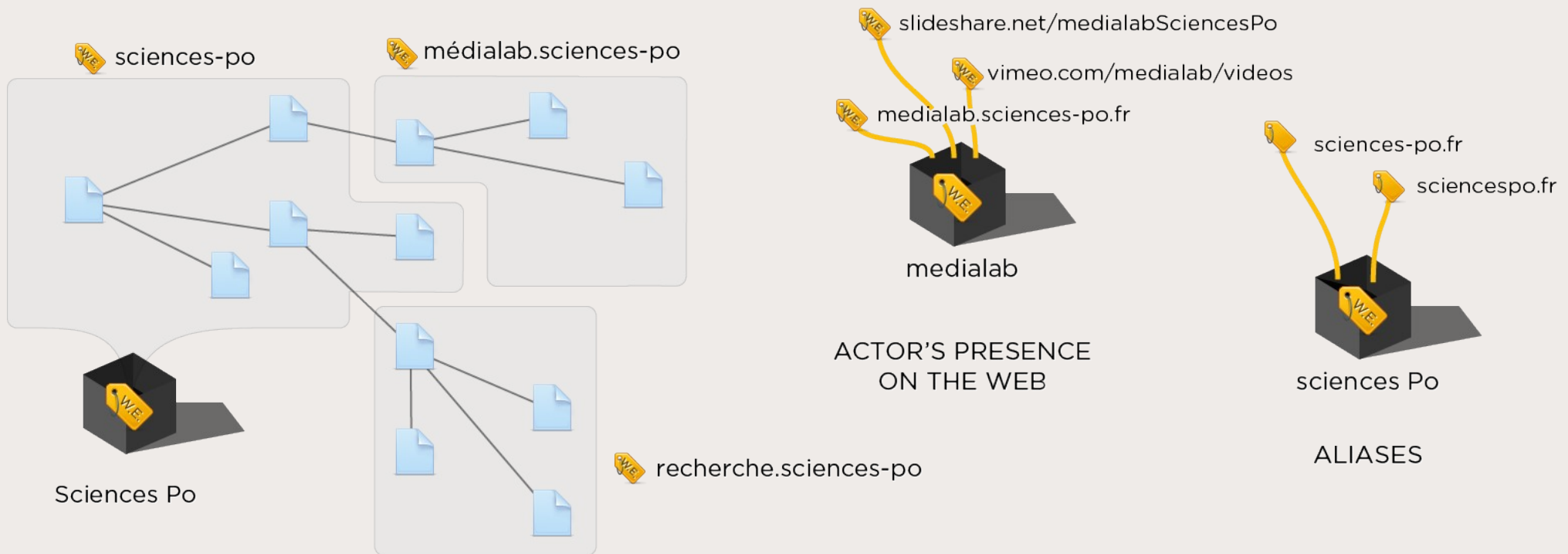


<http://utopies-concretes.org/>

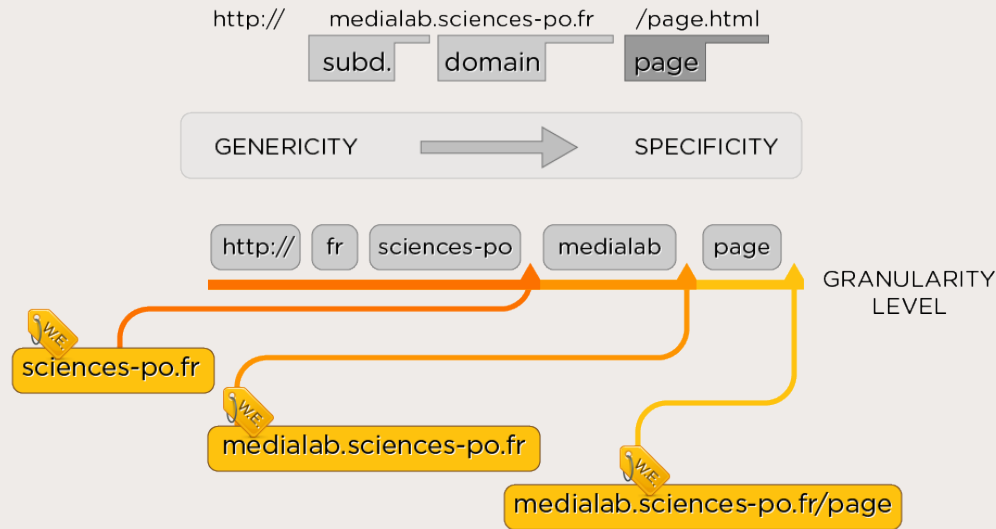
# Principes méthodologiques : « WebEntités »

Comment gérer la diversité de granularité des sites web ?

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



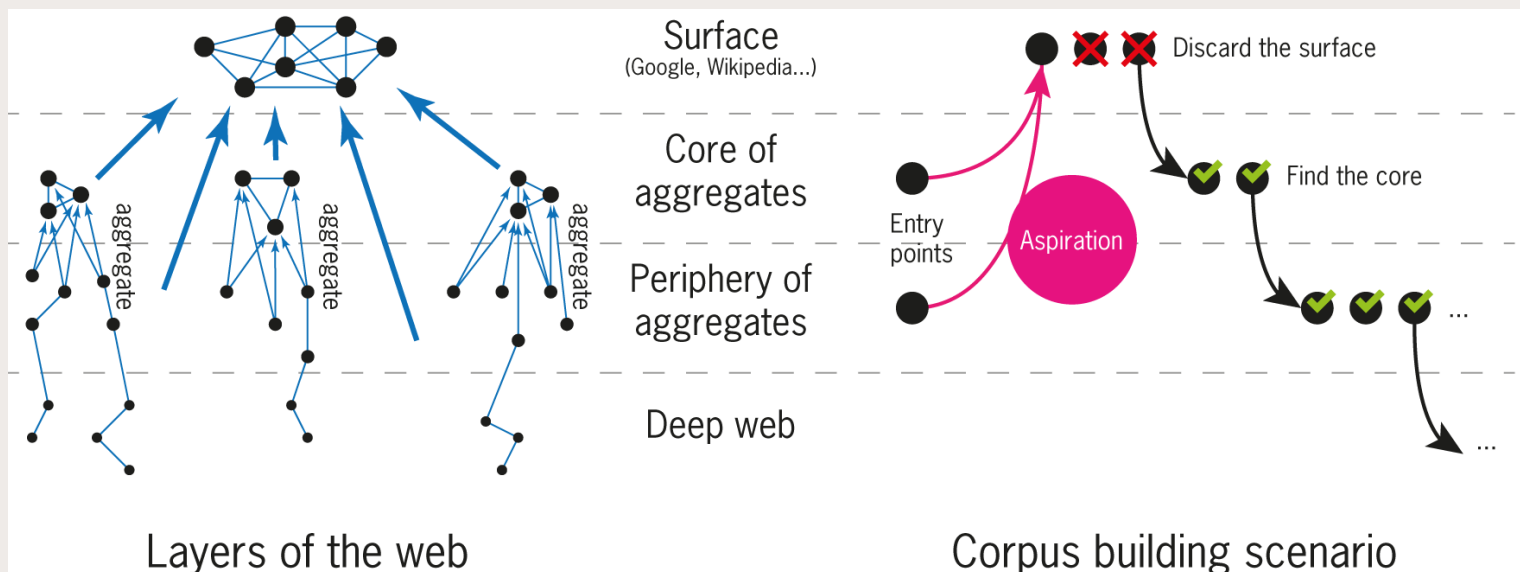
# Principes méthodologiques : « WebEntités »



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</a> /.../326366925310	<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</a> /hervenathan	<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

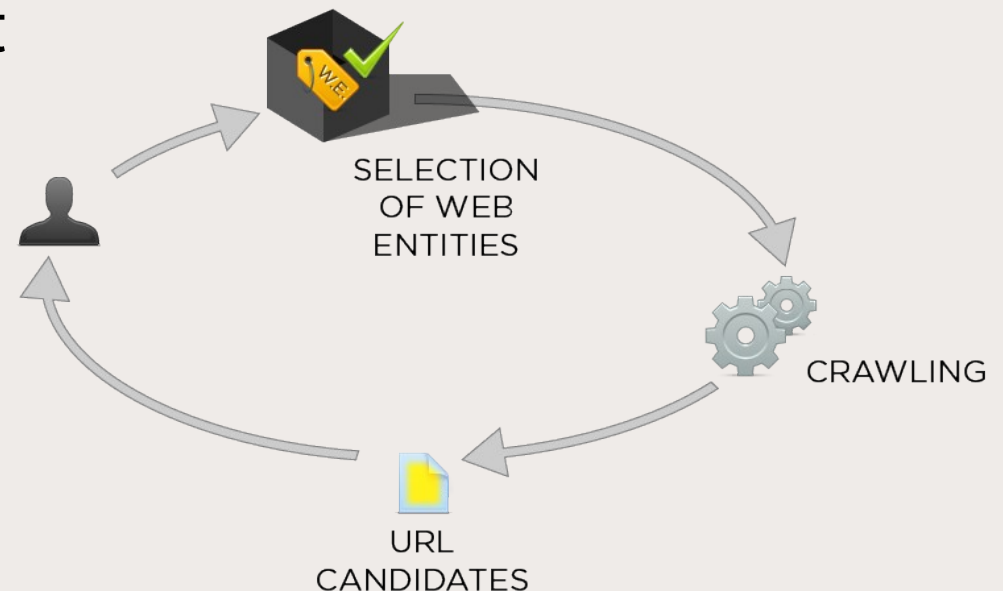
Set to UNDECIDED

# Principes méthodologiques : « Itérations quali »

- 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 dédié à la curation de corpus

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

Héritier du « NaviCrawler » : un navigateur web connecté à Hyphe :

→ prospection et catégorisation in-situ (terrain numérique)

→ formation d'étudiants et lycéens au web (IDEFI FORCCAST)

The screenshot shows the Hyphe Browser interface. At the top, there's a search bar with 'Free.fr' and a status bar with 'PROSPECTION' and other filters. Below the search bar, there's a sidebar with 'Free.fr' and various filters like 'WebEntité', 'Statut', 'Crawlée', and 'Citée'. The main content area displays a page from 'http://etienne.chouard.free.fr/Europe/index.php' with a large image of an amphitheater. The page title is 'Le plan C : instituer une vraie démocratie par une Constitution d'origine Citoyenne.' Below the title, there's a navigation menu with tabs like 'Présentation', 'Analyses et propositions', 'Priorités', 'Échanges', 'CECRI', 'Réflexions', 'Initiatives', and 'Divers'. The main content area has sections for 'Venez participer', 'Présentation', and 'Vos recherches sur le plan C'. The 'Venez participer' section includes links to 'Forum du Plan C', 'WIKI-Constitution', and '(Ancien) blog du Plan C'. The 'Présentation' section has a 'Bonjour et bienvenue :)' message and a 'Lire la suite de l'introduction' button. The 'Vos recherches' section has a search bar and a 'Résumés' section with a 'Le Message' logo and a 'LA VRAIE DÉMOCRATIE' logo.

# Gérer ses catégorisations (tagging)

- Annotations libres
- Catégories

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

Key

- Each dot or node
- Each line or edge
- ↔ one or more hyperlinks
- ↔ web entity to another
- ↔ Links are oriented
- ↔ is not figured in

NODE COLOR

Acteur (tag)

- Presse (157)
- Association (112)
- Blog (56)
- Institution (51)
- Publication scientifique (23)
- Entreprise (21)
- Untagged (19)

NODE SIZE

Indegree

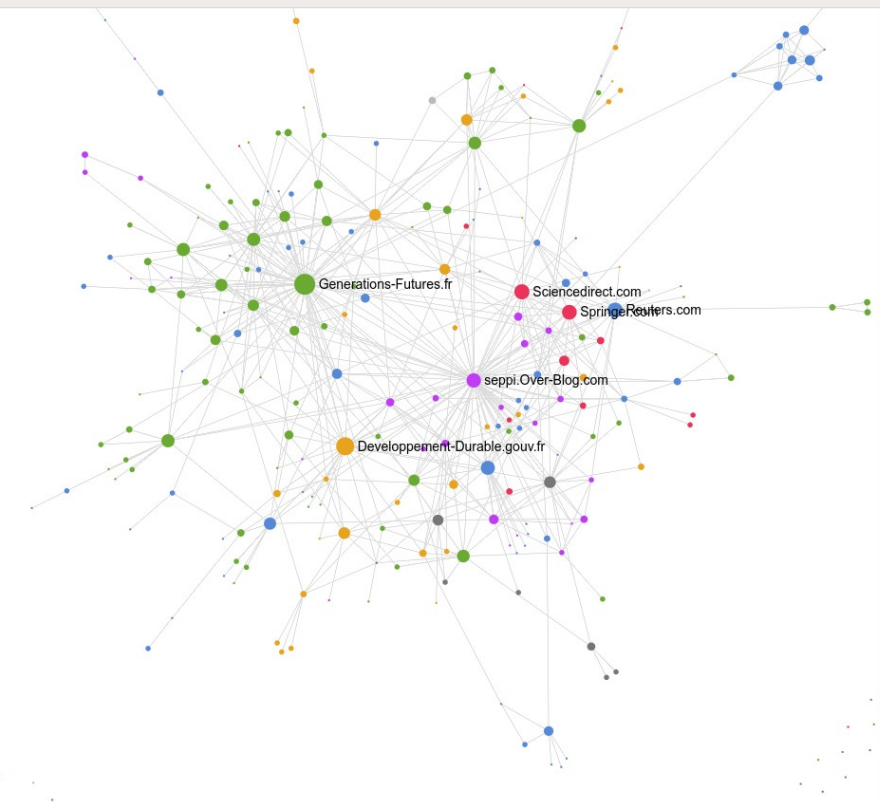
The INDEGREE of a web entity is the number of other web entities citing it.

- Smallest node INDEGREE of 0
- Biggest node INDEGREE of 38

Display a category

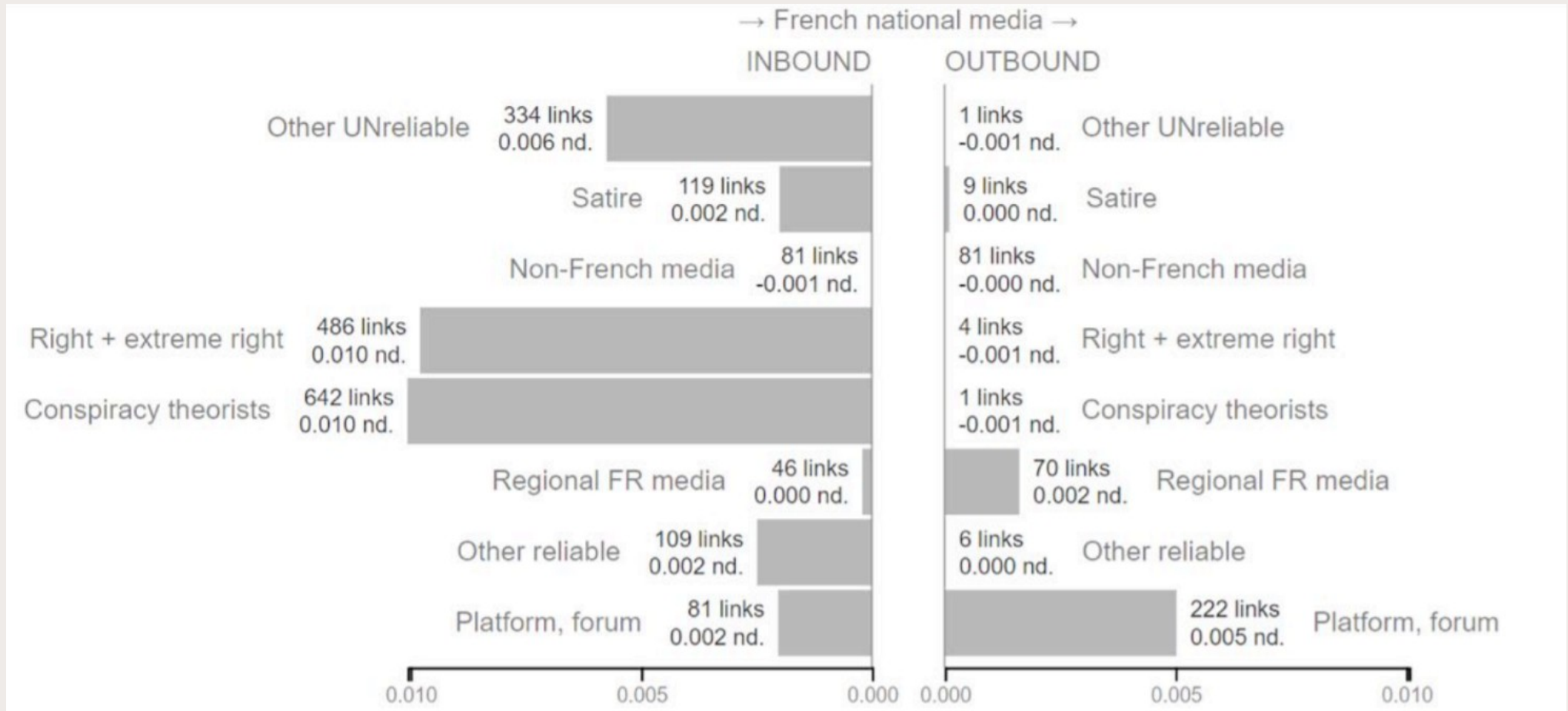
Point de vue

- 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

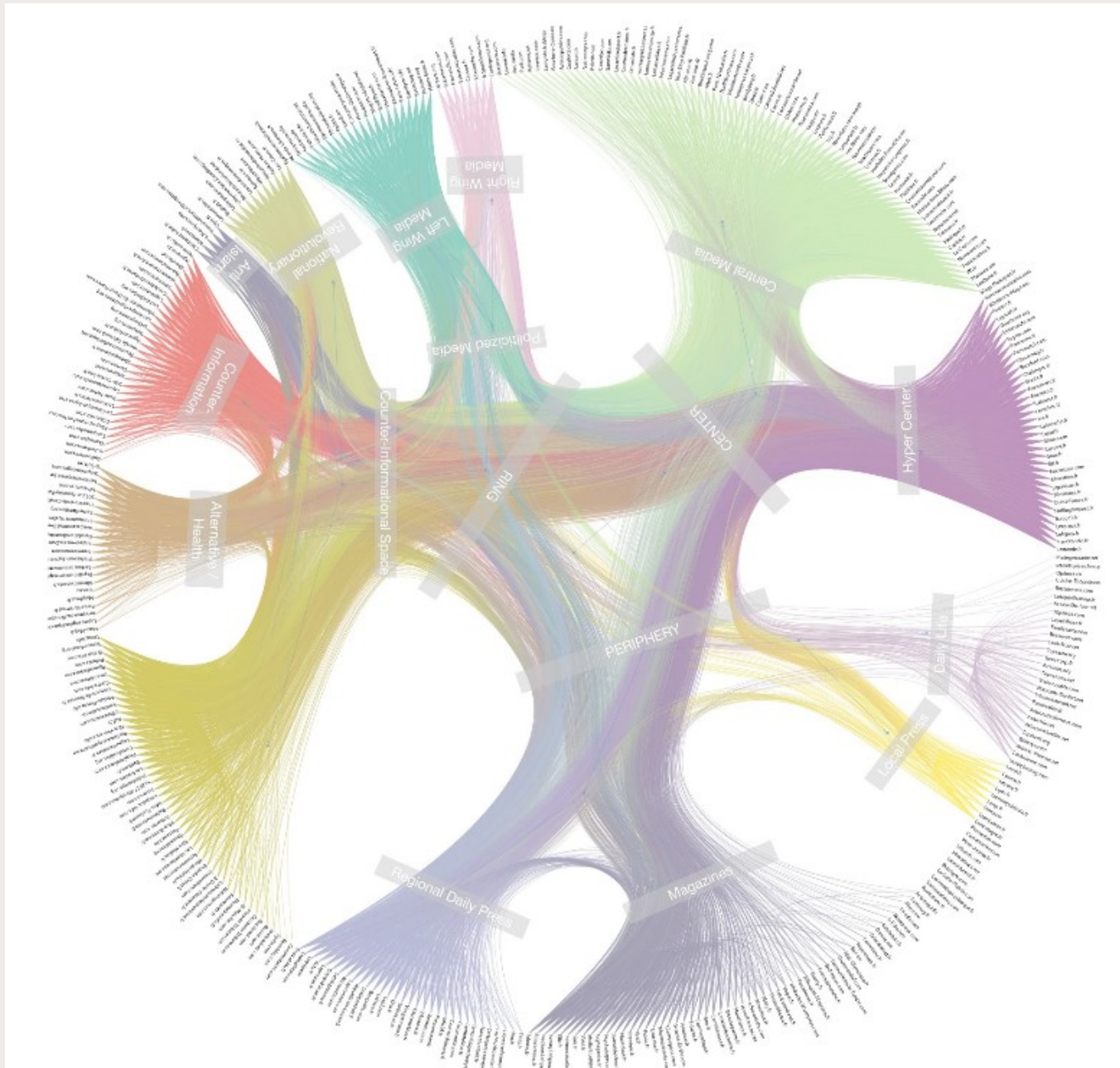




# Analyser la structuration interne des corpus



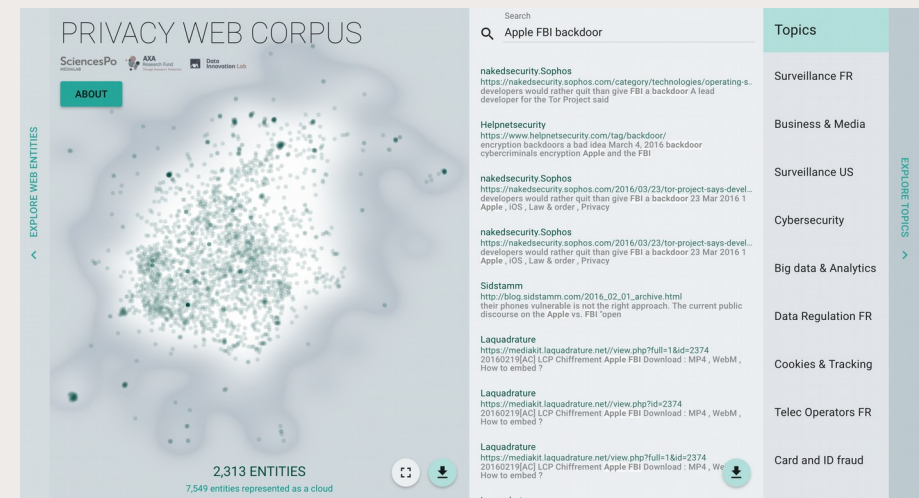
# Explorer les dynamiques de polarisation



# 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.)



- Outils de contrôle qualité des crawls et du corpus
- Outils d'archivage et présentation des corpus finalisés
- Mise à disposition à la demande automatisée (SAAS)