



**HAL**  
open science

## Outils et méthodes pour créer, traiter et analyser des corpus web

Benjamin Ooghe

### ► To cite this version:

Benjamin Ooghe. Outils et méthodes pour créer, traiter et analyser des corpus web. Ateliers du Dépôt légal du web - Saison 6, atelier 3: Qu'est ce qu'un corpus web?, Institut National de l'Audiovisuel (INA), Apr 2015, Paris, France. hal-03631536

**HAL Id: hal-03631536**

**<https://sciencespo.hal.science/hal-03631536>**

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



**SciencesPo.**

médialab

# Outils & méthodes pour créer, traiter et analyser des corpus web

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

[medialab.sciences-po.fr](http://medialab.sciences-po.fr)

Atelier INA #3 Saison 6 - 17/04/15

# I Le médialab

- Fondé en mai 2009

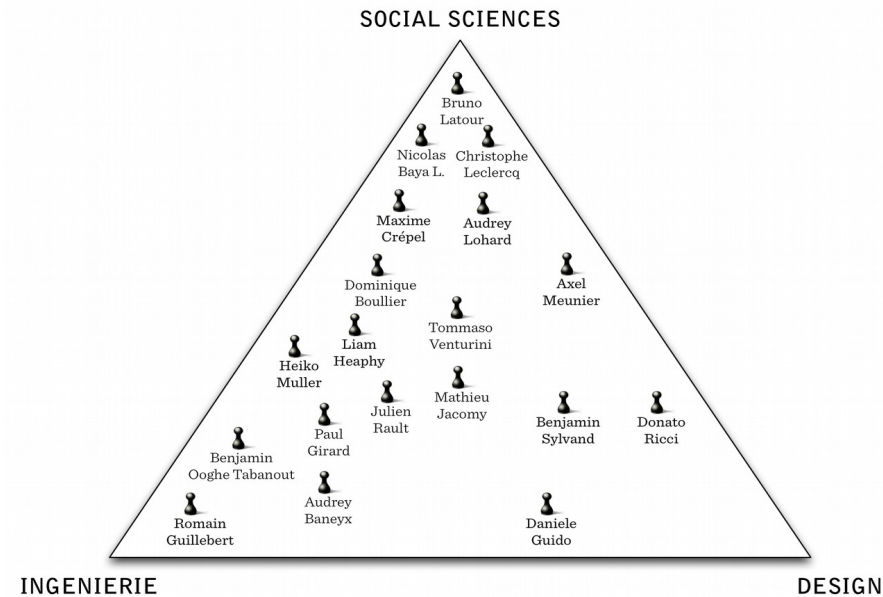
Centre de recherche numérique  
au service de SciencesPo et des

- sciences sociales.

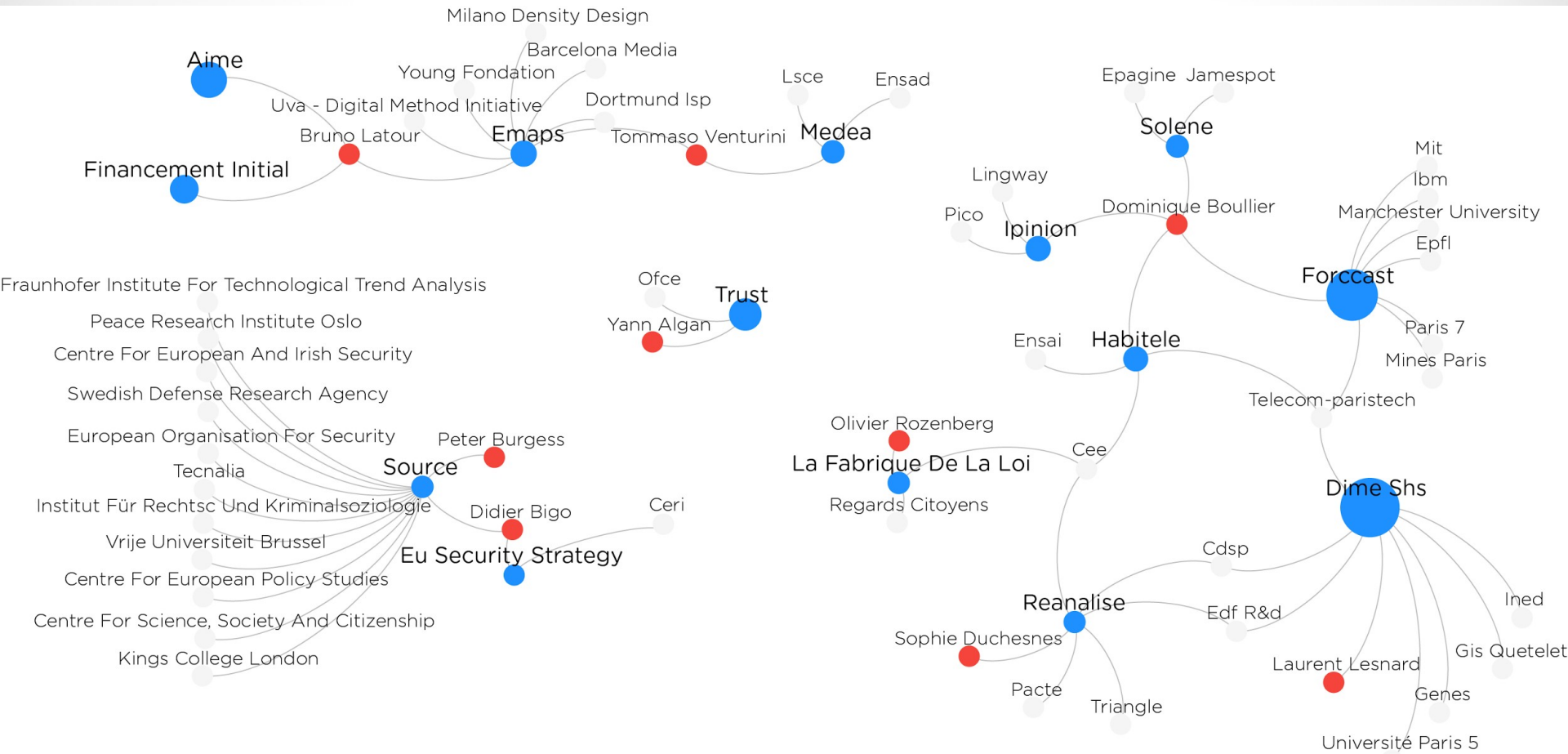
Étude des traces numériques :  
articuler les méthodes

- quantitatives et qualitatives

- Pluridisciplinarité : Sciences sociales + Ingénierie + Design



# □ Une multiplicité de projets et partenaires



# □ L'instrument DIME Web

- Equipex support aux Sciences Humaines et Sociales

## Accompagnement numérique

- et méthodologique

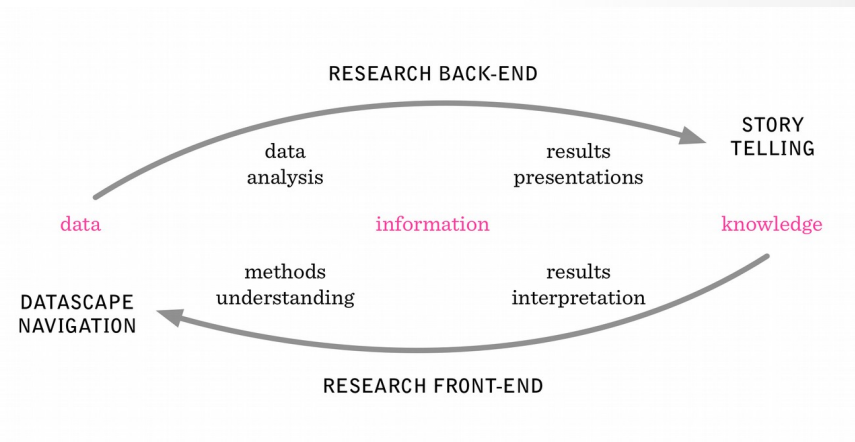
- Méthodologie itérative

- 2 personnes (Mathieu Jacomy & moi-même)

- Objectif ANR d'auto-financement

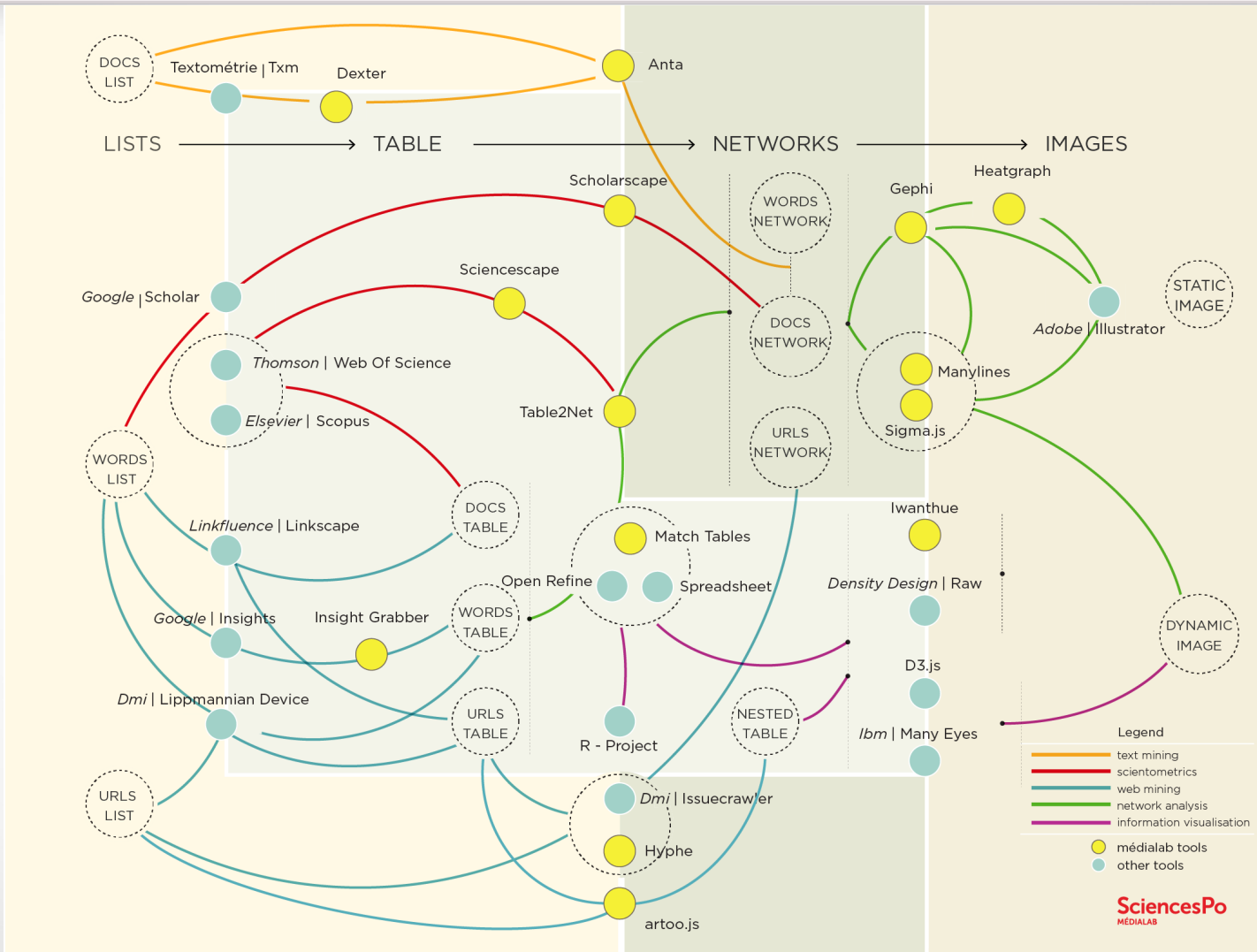
- ⇒ offre de service payant

- ⇒ mutualisation (logiciels libres/OpenSource)



# Un écosystème d'outils

[tools.medialab.sciences-po.fr](http://tools.medialab.sciences-po.fr)



# Collectes ciblées (scraping)

- gazouilloire : collecte Twitter programmée
- ScienceScape : construction de corpus scientométriques

## artoo.js & sandcrawler.js :

- simuler le navigateur web
- Scripts dédiés :
- - commentaires LeMonde.fr
- - métadonnées FlickrR
- - forum M6.fr
- - ...

**ScienceScape**  
Helpers for scientometrics. Convert files, get networks, visualize stuff from Scopus or Web of Knowledge.

The interface features several interactive panels:

- Get Networks:** Visualize and download networks of keywords, authors, and journals.
- Papers over time:** Visualize the number of papers published each year.
- Keywords over time:** Visualize the use of each keyword over time.
- Top keywords / year:** Visualize the most used keywords each year.
- Journals over time:** Visualize and download the journals.
- Top journals / year:** Visualize the journals publishing the most.
- A-K-J Sankey:** Visualize the main authors.
- Utilities:** Upload Scopus CSV files and download results.

# Explorer les données collectées

- Exploration visuelle (raw, rerere)
- Analyse visuelle de réseaux :
  - Gephi, Table2Net, sigma.js, ...
  - ManyLines : storytelling (exemple)

**Table 2 Net**

Extract a network from a table. Set a column for nodes and a column for edges. It deals with multiple items per cell.

**Load your CSV table**

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

Parsing successful. 10 columns and 347 rows.

Row number	id	legislature	textorial_id	numero	suget	sort	date	parlementaires	texte	expose	signataires	source
1	12677	14	707	3	APRES ART 17	Rejeté	2013-02-12	laronel-luca@henry-mariani@proclais-dhauc@gaucque-verche@em@schel-tem@jean-pierre-deco@p@henry-lazar@ed@emsen-	l - Les établissements de crédit garantis sent le droit au credit a toute personne resident sur le territoire francais de facon reguliere et	Cet amendement a pour but de permettre l'instauration d'un droit au credit opposable.	M. Luca, M. Mariani, M. Dhruac, M. Verche, M. Temot, M. Decool, M. Laronel, M. Abad,	http://www.assemblee-nationale.fr/14/amend-

**1. Type of Network**

Bipartite (two types of nodes)

You may extract different types of networks from a table. It depends on how you use columns to build the nodes and the edges.

- **Monadic**: If you want a single type of nodes, for instance authors. They will be linked when they share a value in another column, for instance papers.
- **Bipartite**: If you want two types of nodes, for instance

**INPUT PREVIEW**

row #	url	user_screen_name	text	timestamp	lang	coordinates
32527	https://twitter.com/MahdiElV...	MahdiElV	RT @The_Mac_: Le sexe c'est 75% de coup de coeur mytho, c'est 90% de...	2014-06-09T19:51:14	fr	

**OUTPUT PREVIEW**

row #	url	user_screen_name	text	timestamp	lang	coord.
410117	https://twitter.com/BrandonKalé	BrandonKalé	"DieuQuinte, Lénosse". L'émor est des le p...	2014-02-23T23:06:26	fr	46.1746236;5.1400033

**TEXT - WORDS CLOUD**

**TEXT - WORDS CLOUD**

**TIMESTAMP - DAILY VOLUME**

**LANG - ITEMS TOP 20**

Lang	Count	Lang	Count	Lang	Count
fr	321785	en	117	cy	17
en	10346	es	111	th	10
es	1010	pt	104	ba	10
pt	779	de	177	ba	10
de	5422	ru	176	th	10
ru	209	fr	168	ba	10
fr	118	en	161	ba	10
en	115	en	157	ba	10
de	113	en	157	ba	10
en	113	en	157	ba	10
en	113	en	157	ba	10

**manylines**

1 new network 2 **basemap** 3 take views 4 narratives

**LinLog mode**

LinLog mode

Different equations with slower convergence but more clustering. Avoid if disconnected components or nodes.

**gravity**

50

Attracts all nodes to the center, preventing disconnected components to drift away

**Barnes-Hut optimization**

Barnes-Hut optimization

Group nodes into regions to scale the algorithms repulsion. Note: use this option on large graphs only.

**Node size**

original  degree  indegree  outdegree

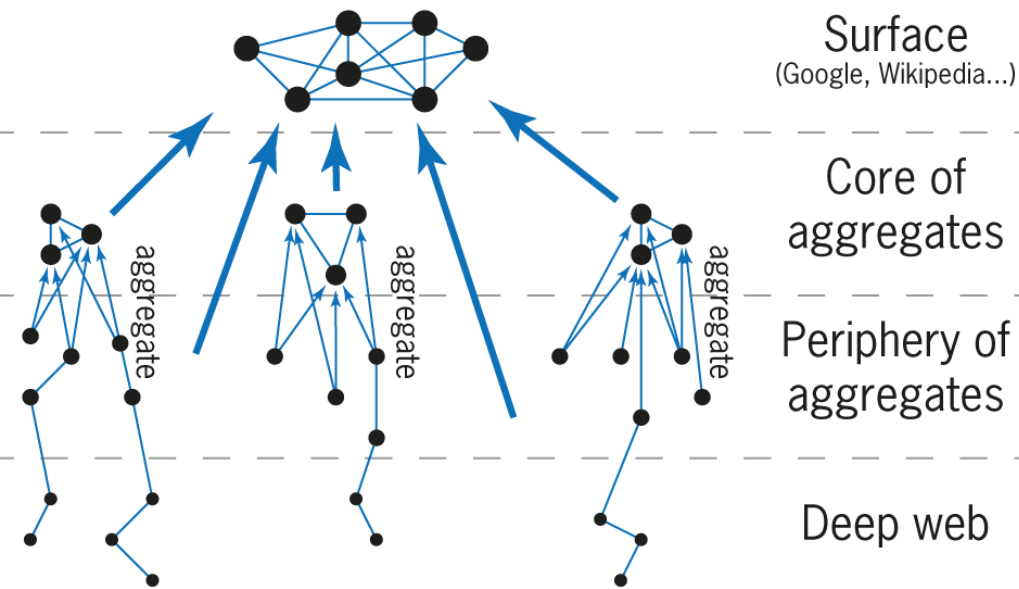
The layout is usually cleaner when most connected nodes are bigger. The degree is the number of links, the indegree



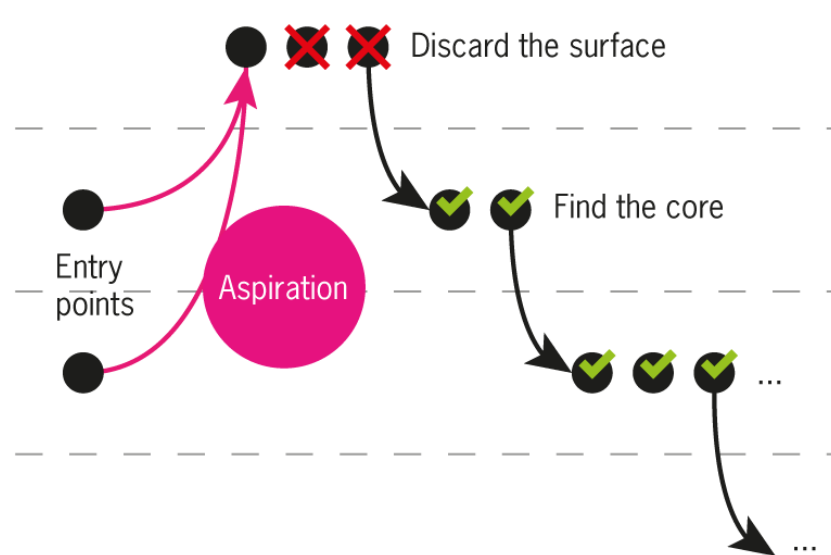
# □ Hyphe : crawler orienté recherche

- Collecte de données web pour construction itérative d'un réseau entre ressources web
- Fins de recherche : études de communautés, controverses...
- Ergonomie : interface web pour utilisateurs compréhensible par des non-informaticiens (chercheurs en SHS)
- Efficacité : gestion simultanée de plusieurs corpus de grande taille pour les SHS (milliers de sites crawlés)
- Rapidité : crawl et indexation en temps réel

# Le web en couches

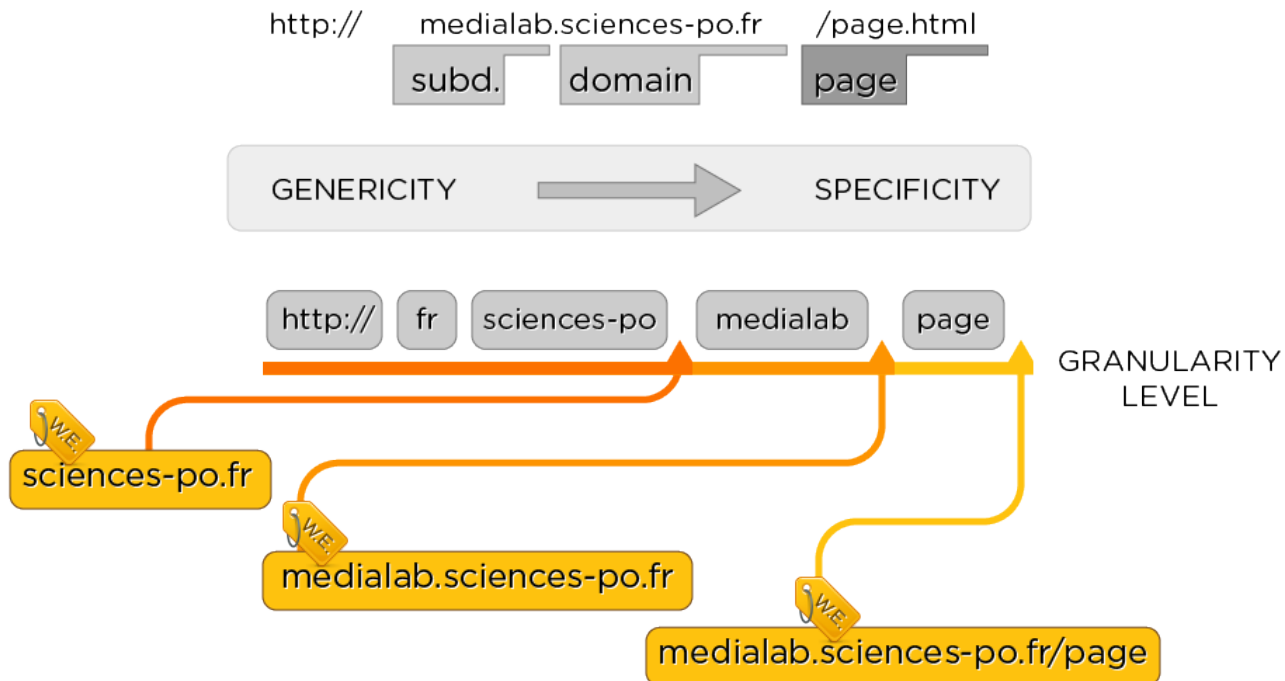


Layers of the web

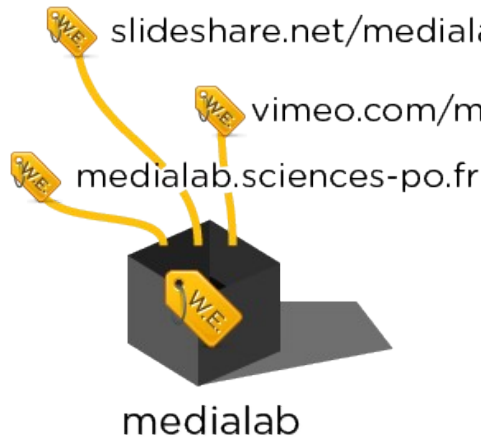


Corpus building scenario

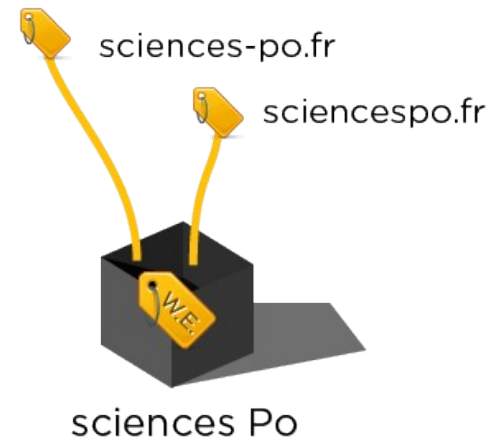
# Définir des points d'ancrage précis (LRUs)



# Des sites ou... des entités web

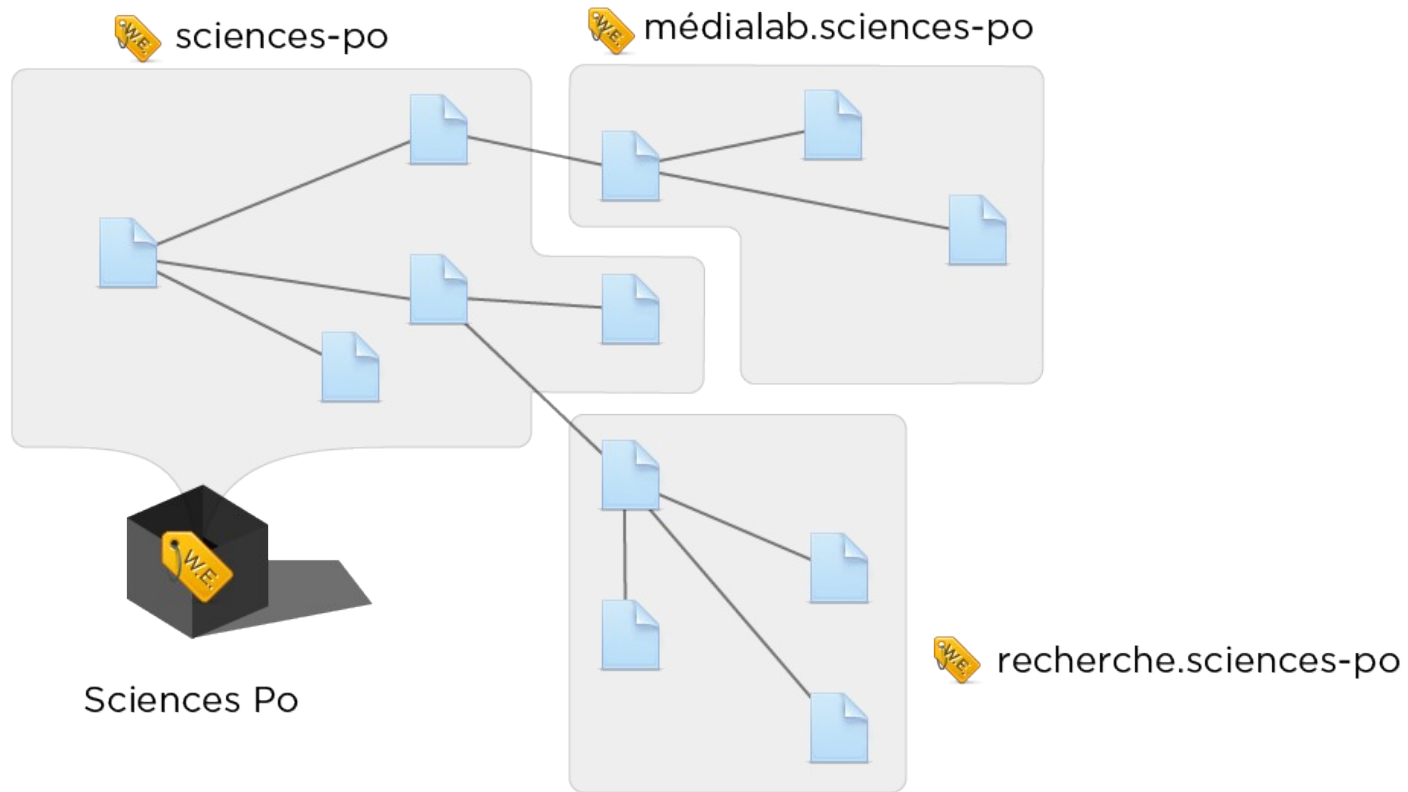


ACTOR'S PRESENCE  
ON THE WEB



ALIASES

# Préserver la complexité

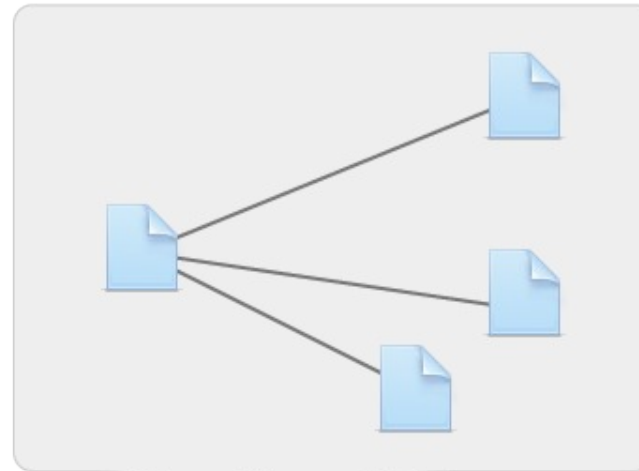


# Le crawl dirigé par la recherche

1

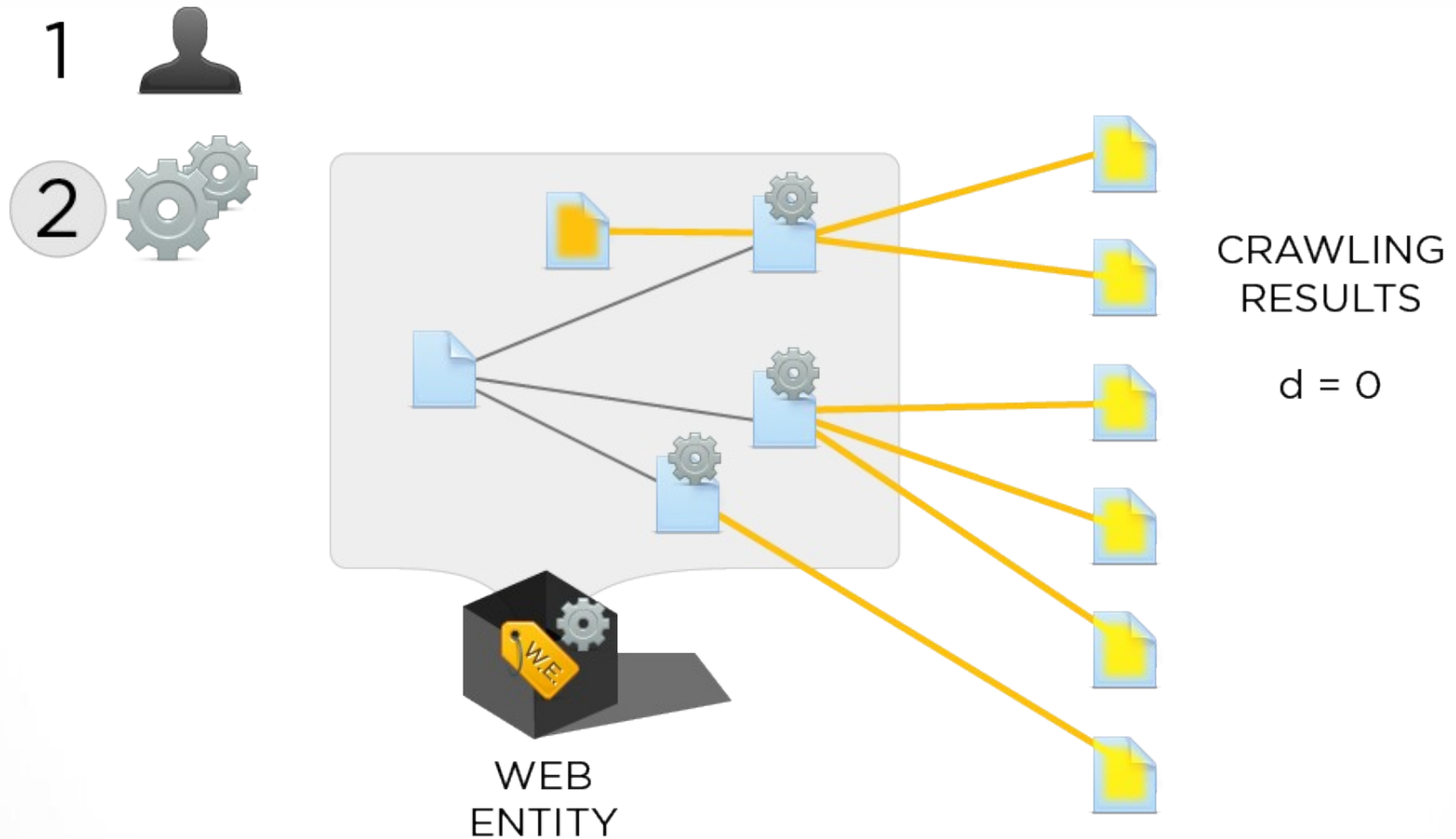


RESEARCHER  
SELECT  
*STARTING*  
*ENTITIES*

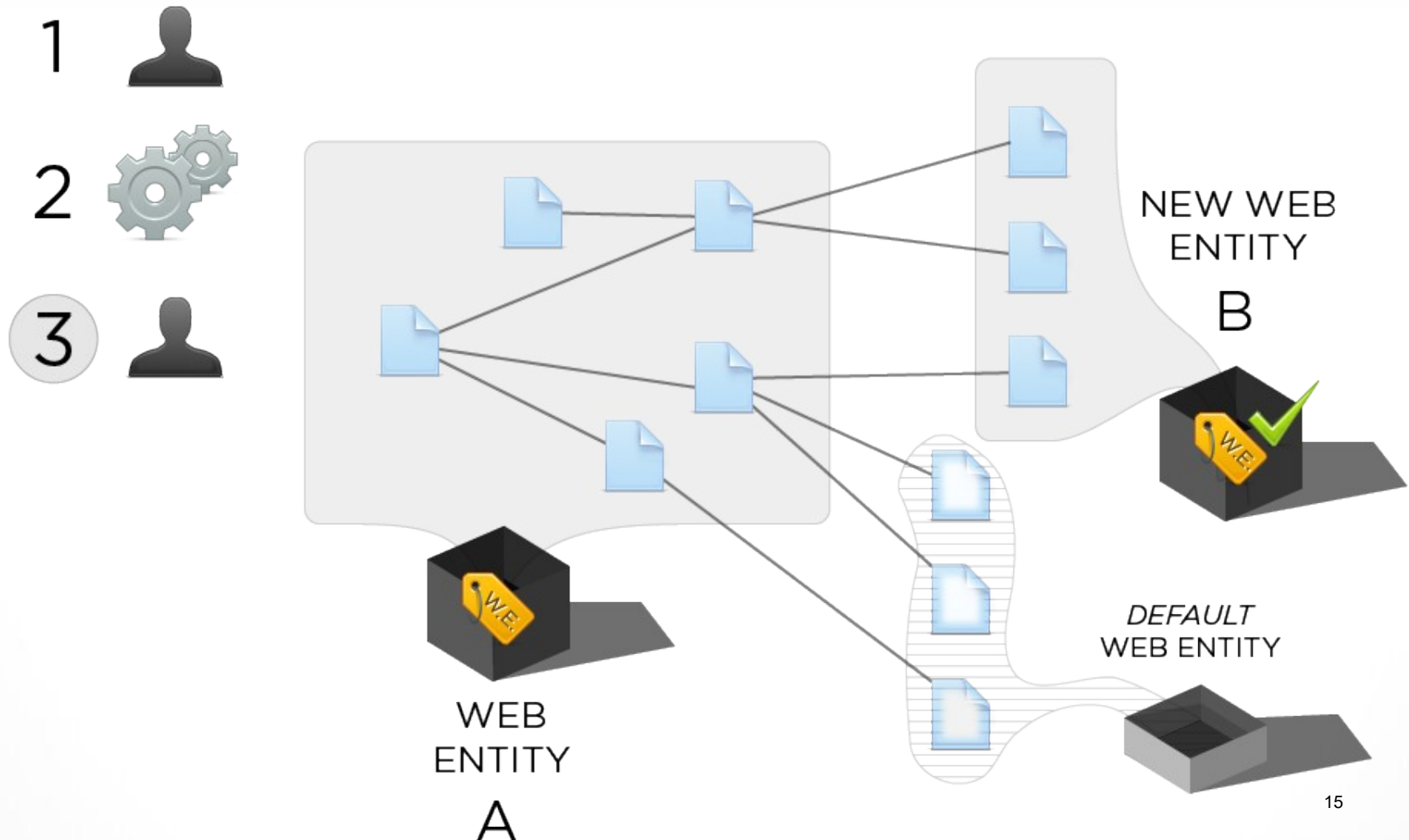


WEB  
ENTITY

# Le crawl dirigé par la recherche

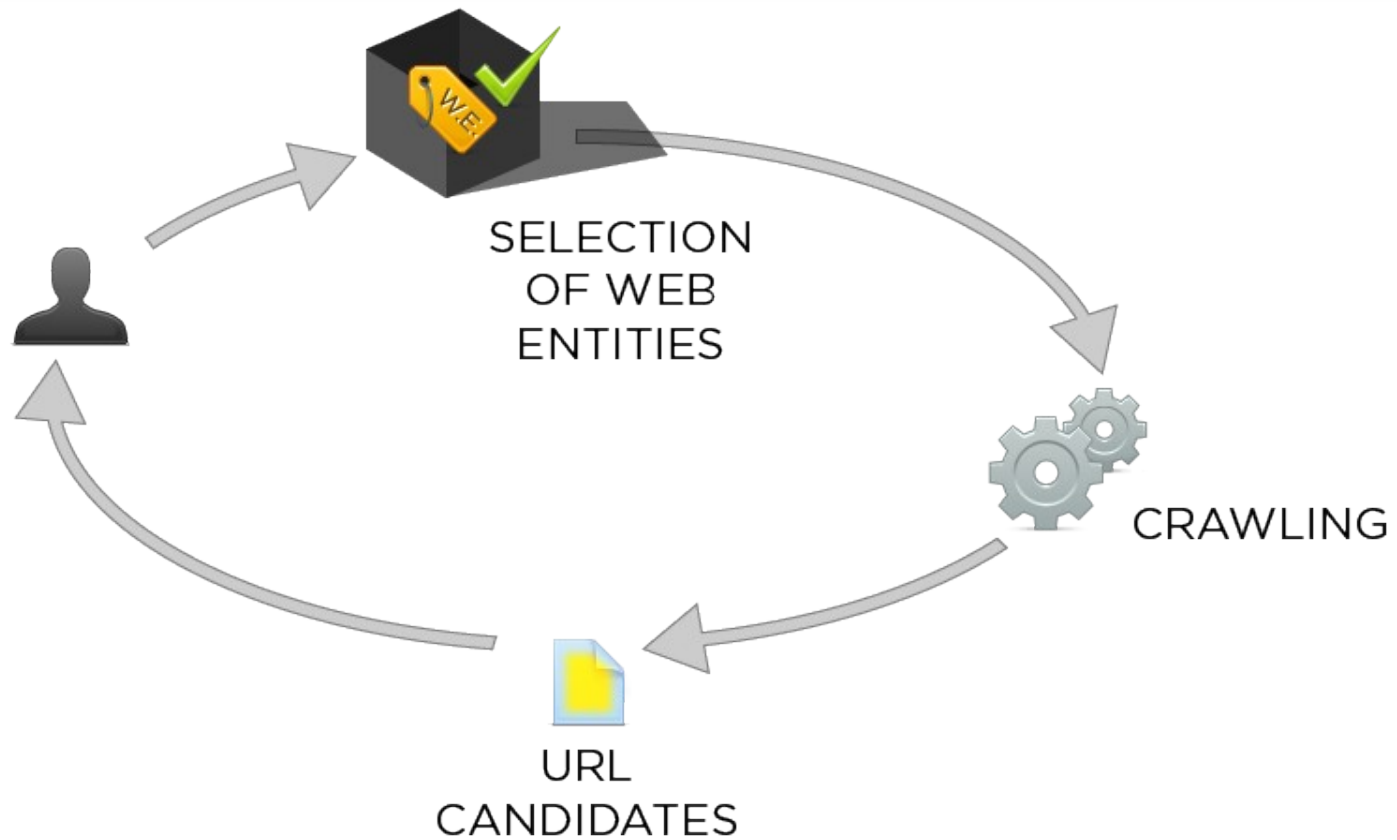


# Le crawl dirigé par la recherche





# Le crawl dirigé par la recherche



# L'interface de Hyphe 2nde version

Démo : [hyphe.medialab.sciences-po.fr/demo](http://hyphe.medialab.sciences-po.fr/demo)

Code source : [github.com/medialab/hyphe](https://github.com/medialab/hyphe)

## Gérer plusieurs corpus

The screenshot shows the Hyphe interface for managing multiple corpora. At the top, there is a logo consisting of three overlapping squares and the word "hyphe" below it. Below the logo, the text "Select your project" is displayed, followed by a status message: "6 projects can be put into operation (6/12 slots occupied)". A search bar labeled "Search a project" is present. Below the search bar, there is a list of projects:

- 2°C**: 112 web entities, Created last month - Used 2 weeks ago
- ASPIRES2**: 381 web entities, Created 2 months ago - Used 2 weeks ago
- Barrage (bis)**: 132 web entities, Created 2 weeks ago - Used 2 weeks ago

At the bottom, there is a separator "- or -" and a button labeled "NEW PROJECT".

## Résumé d'un corpus

The screenshot shows the Hyphe interface for a corpus summary. The title bar is "Pharma" with a close button. The interface is divided into two main sections: a sidebar on the left and a main content area on the right.

**Sidebar (Left):**

- OVERVIEW** (selected)
- 1 IMPORT URLs
- 2 CRAWL
- 3 PROSPECT
- 4 EXPORT WEB ENTITIES
- LIST WEB ENTITIES
- VISUALIZE NETWORK
- MONITOR CRAWLS
- SETTINGS

**Main Content Area (Right):**

**OVERVIEW**

IN 346	OUT None
DISCOVERED 14065	UNDECIDED None

# L'interface de Hyphe 2nde version

## Définir précisément les WebEntités

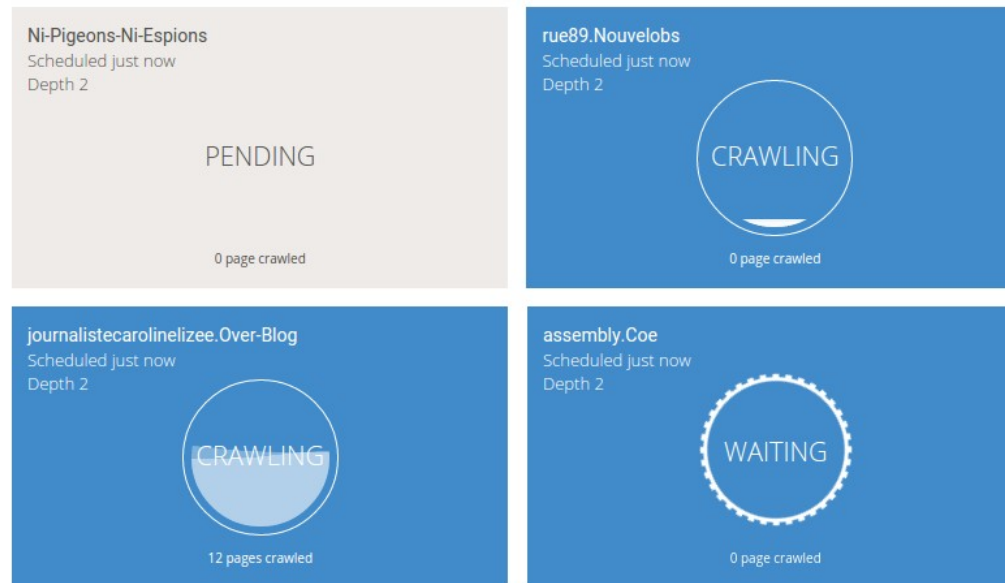
12	Actuchomage New ▲ Same web entity defined row 614	http .org actuchomage www.
13	Afev New	http .fr afev www.
14	Nouvelobs New ▲ Same web entity defined rows 15, 112, 115, 146, 485, 551, 587, 601, 785, 809 and 912	http .com nouvelobs blogs. globe.
15	Nouvelobs New ▲ Same web entity defined rows 14, 112, 115, 146, 485, 551, 587, 601, 785, 809 and 912	http .com nouvelobs blogs. pascalbonifa...
16	www2.Euromemorandum /uploads New	http .eu euromemorandum www2. /uploads /
17	Ademe New	http .fr ademe www2.

# L'interface de Hyphe 2nde version

## Surveiller l'avancement des crawls

### MONITOR CRAWLS

Last Hour Today This Week All Crawls



# L'interface de Hyphe 2nde version

## Identifier d'autres WebEntités à inclure au corpus

PROSPECT

14065 DISCOVERED WEB ENTITIES

Type a query  Search

(Range selector not implemented yet)

Name	Prefixes	Is Cited
Youtube	■■■■■	124 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Google	■■■■■	103 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Googleapis	■■■■■	66 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Legifrance.Gouv	■■■■■	64 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Wikipedia	■■■■■	56 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Twitter /share	■■■■■	55 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Facebook /pages	■■■■■	50 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Cdc	■■■■■	46 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Nytimes	■■■■■	43 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Asso	■■■■■	42 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Twitter	■■■■■	42 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Europa	■■■■■	41 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Adobe	■■■■■	40 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Apple	■■■■■	40 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Ac	■■■■■	39 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Free	■■■■■	39 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Yahoo	■■■■■	34 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Gmpg	■■■■■	33 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Inserm	■■■■■	33 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>
Blogspot	■■■■■	31 <input type="button" value="IN"/> <input type="button" value="OUT"/> <input type="button" value="UND."/>

SET TO: IN (2)

Inserm

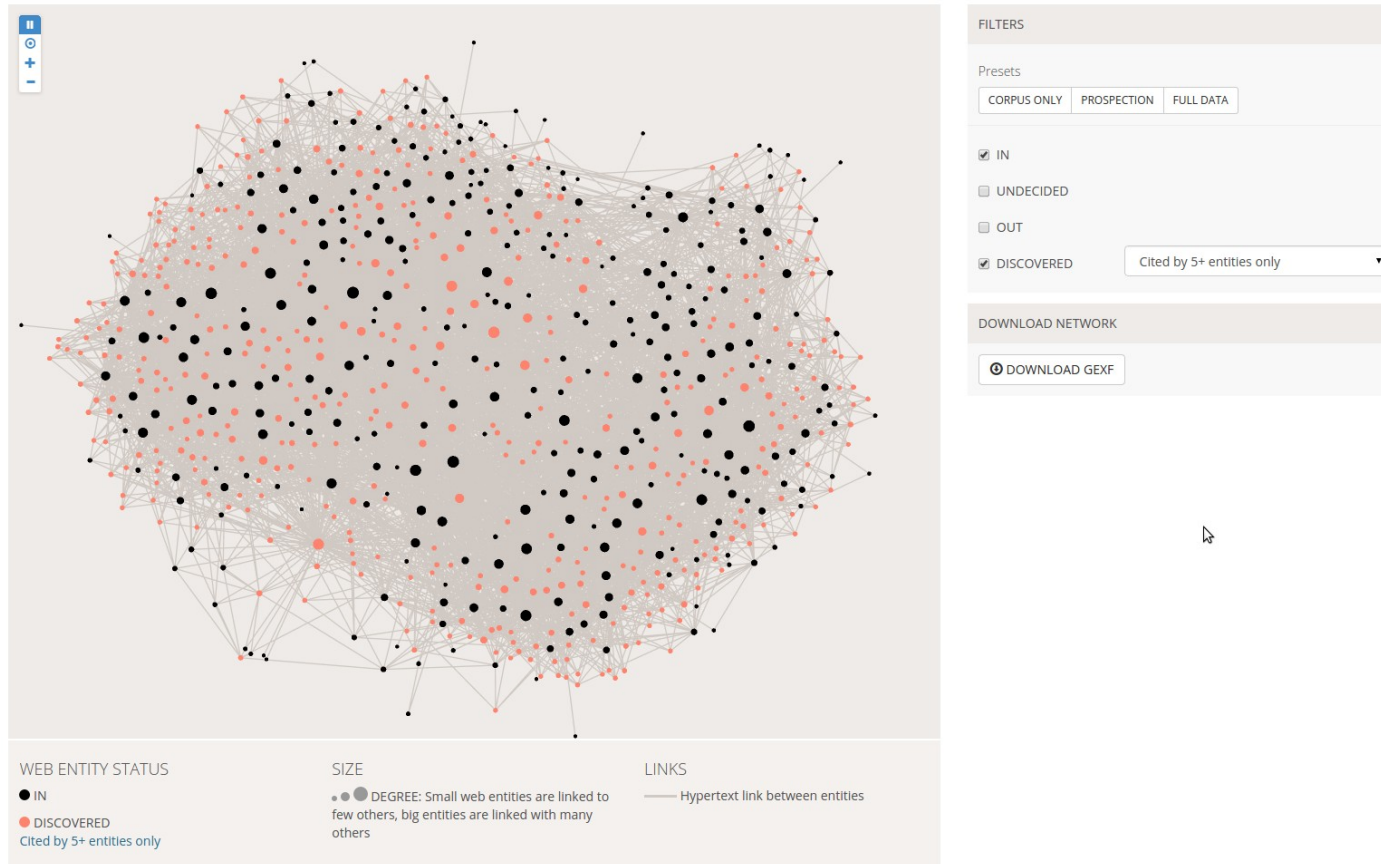
Cdc

SET TO: OUT (16)

SET TO: UNDECIDED (1)

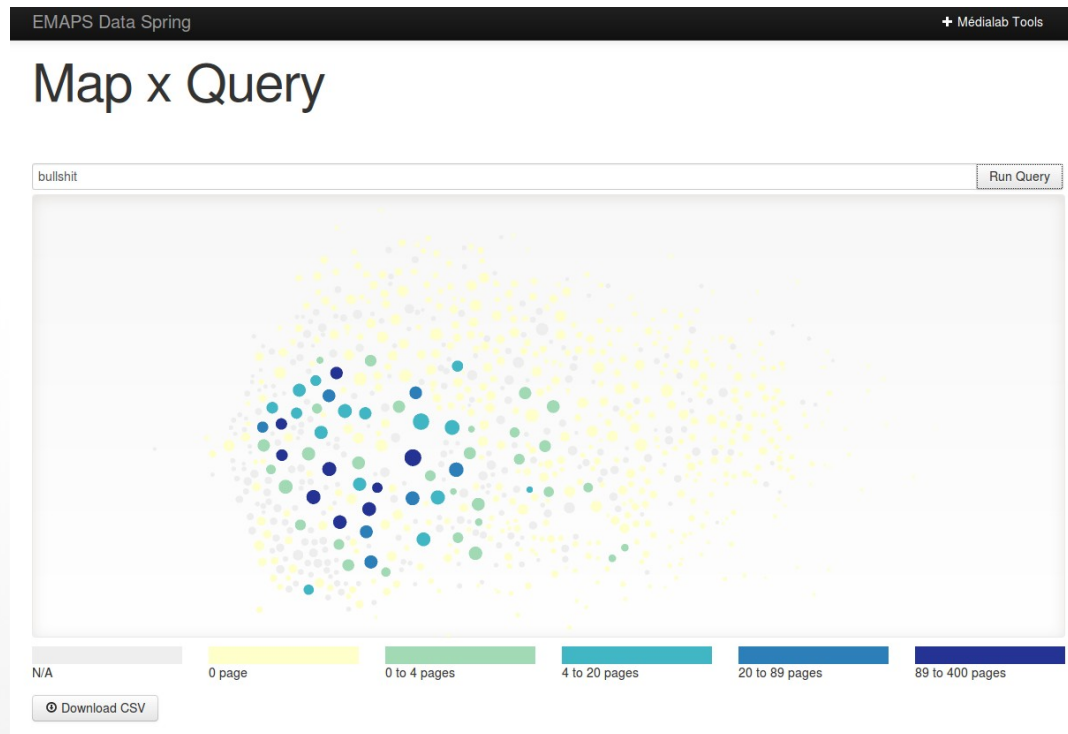
# L'interface de Hyphe 2nde version

## Explorer le réseau des liens entre WebEntités



# □ Analyse de contenus (expérimentation)

- Hyphe collecte également le contenu texte à chaque crawl
- Indexation SolR ([exemple](#) « bullshit » sur un corpus climat)



# ▣ À venir dans Hyphe...

- Import/export & « rebuild corpus » pour exploration temporelle
- Stabiliser PhantomJS pour le crawl browser-like (Facebook, ...)
- Interface de catégorisation (tags)
- Prospection « en contexte »
- Outil de contrôle qualité des crawls et du corpus
- Outil d'archivage et présentation des corpus finalisés
- Hyphe embarqué sur clé USB





□ Merci de votre attention !

**SciencesPo**  
MÉDIALAB

[@medialab\\_ScPo](#)

benjamin.ooghe@sciencespo.fr