



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

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

I Le médialab

- Fondé en mai 2009

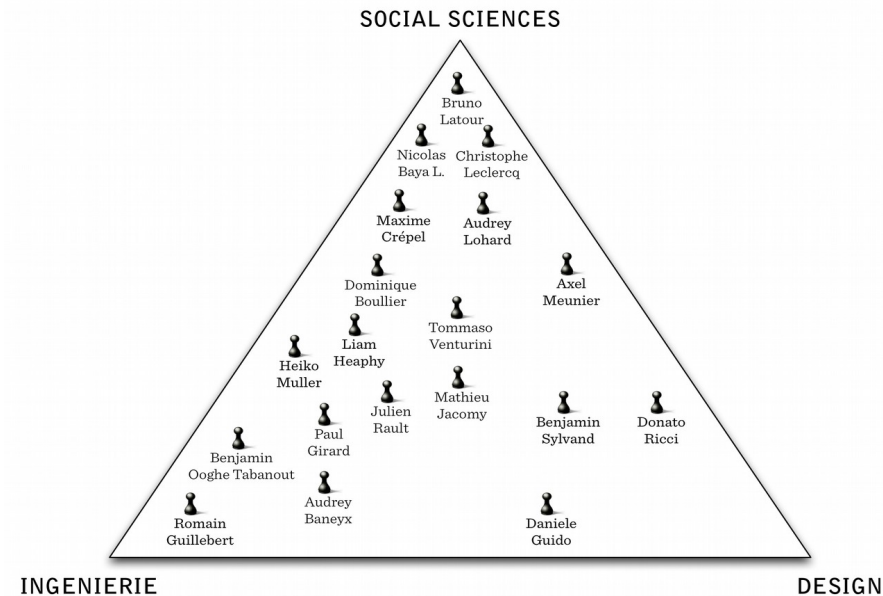
Centre de recherche numérique
au service de SciencesPo et des

- sciences sociales.

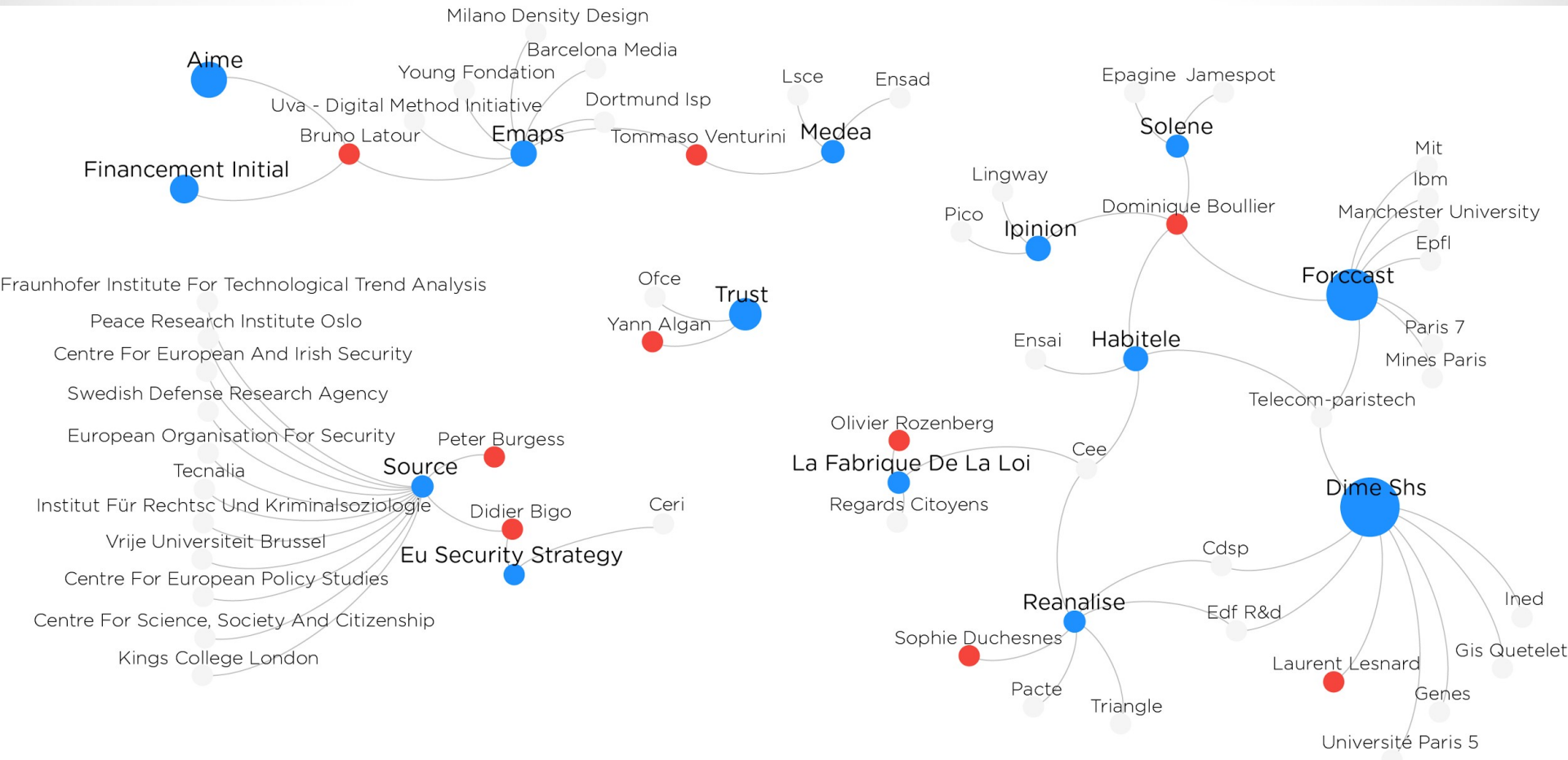
Étude des traces numériques :
articuler les méthodes

- quantitatives et qualitatives

- Pluridisciplinarité : Sciences sociales + Ingénierie + Design



□ Une multiplicité de projets et partenaires

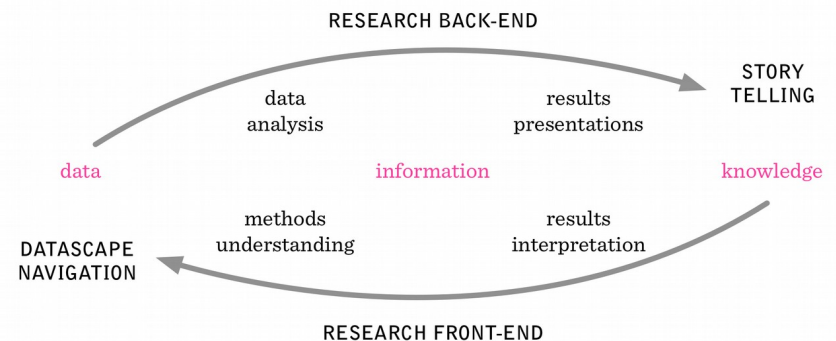


□ L'instrument DIME Web

- Equipex support aux Sciences Humaines et Sociales

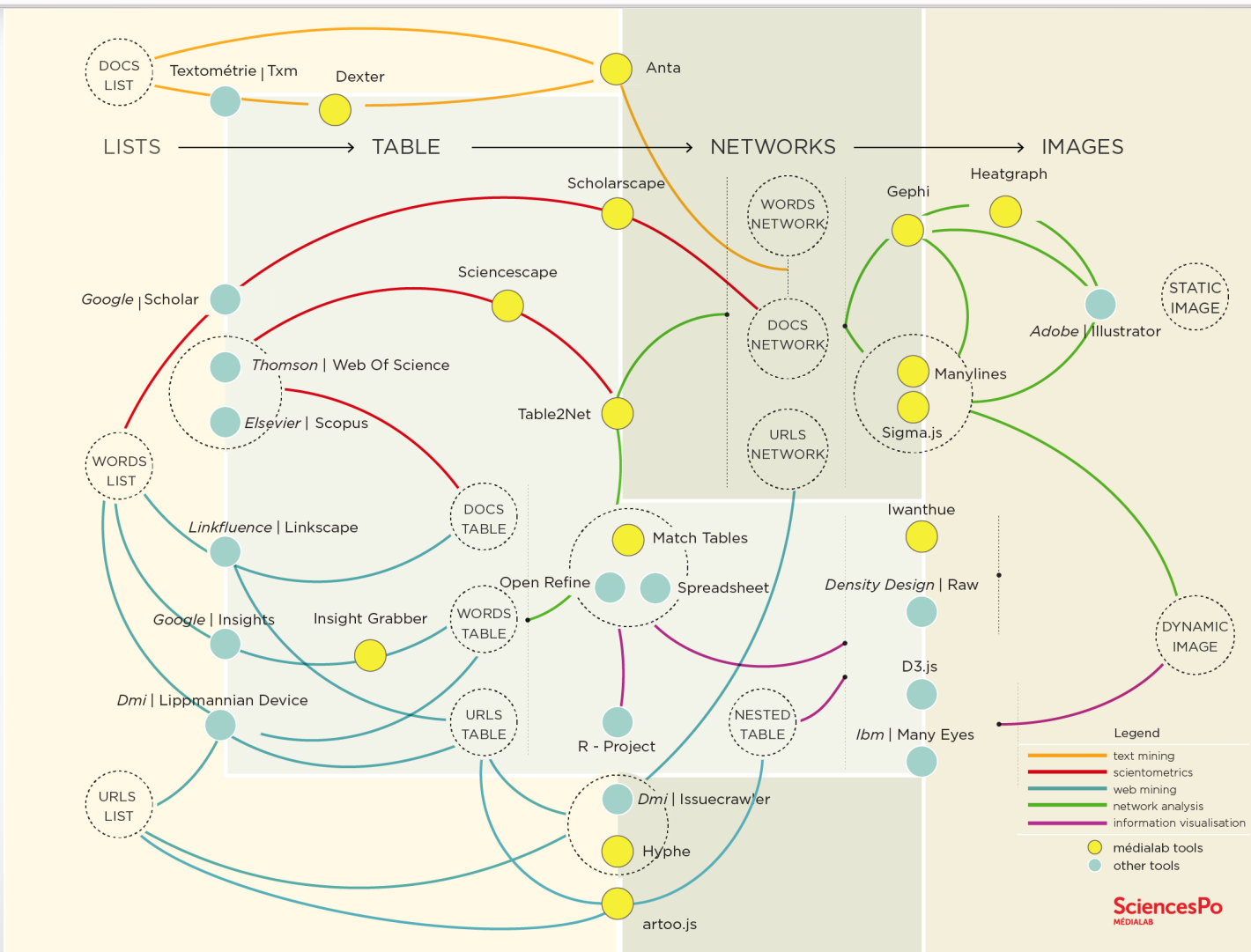
Accompagnement numérique

- 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

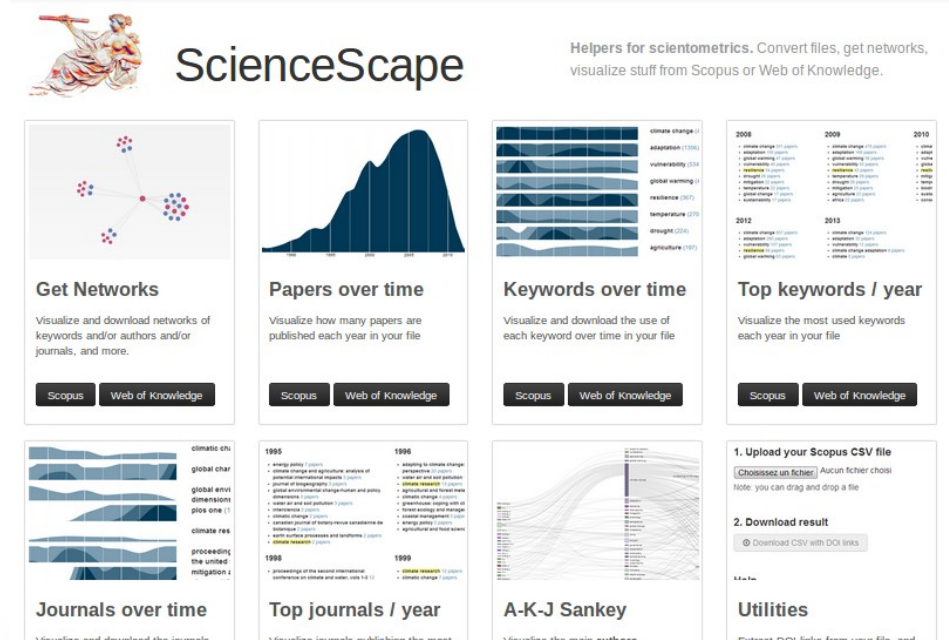


Collectes ciblées (scraping)

- 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
 - - ...



□ 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.

Table preview

Row number	id	legislature	textuel_id	numero	sujet	sort	date	parlementaires	texte	expose	signataires	source
1	12077	14	707	1	APRES ART. 17	Rejete	2003-02-12	lucac@senyves-marian@jocodes-duc@jacqueline-verheynen@jocodes-terroijean-pierre@duc@senyves-lacaz@jean-damen-	1. – Les établissements de crédit garantis par le pour but de permettre à toute personne résidant sur le territoire français de la manière la plus régulière et	Cet amendement a pour objet de permettre l'instauration d'un droit au crédit opposable.	M. Lucac, M. Mariani, M. Vercheux, M. Terroij, M. Decoud, M. Lacaze, M. Abiad,	http://www.assemblee-nationale.fr/14/amen/

1. Type of Network

Bipartite (two types of nodes)



You will have to choose:

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

- **Normal**: 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



manylines

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

WorldDiffusion

LinLog mode

☐ LinLog mode

Different equations with lower 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 algorithm's repulsion.
Note: use this option on large graphs only.

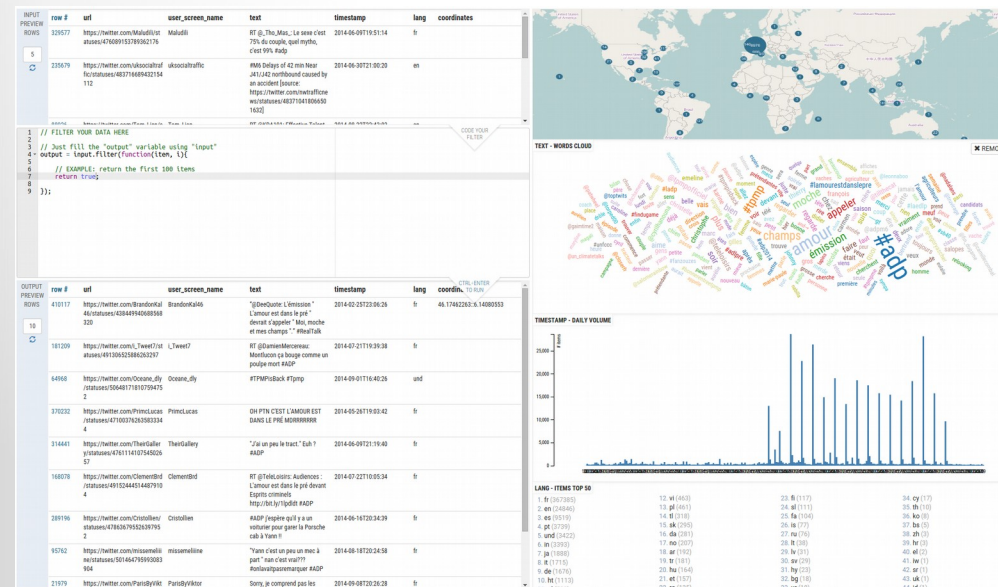
Node size

☐ original ☒ degree ☐ indegree
☐ outdegree

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

Apply layout ▶

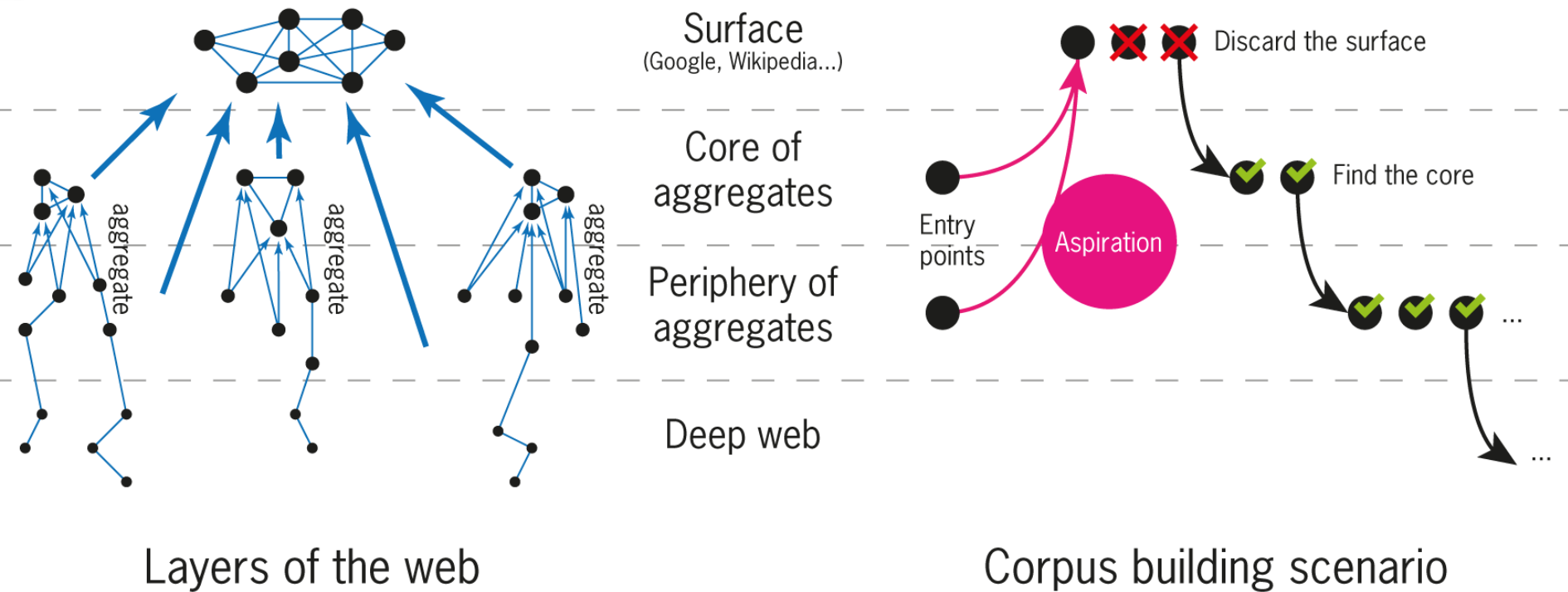
TYPE
2 categories



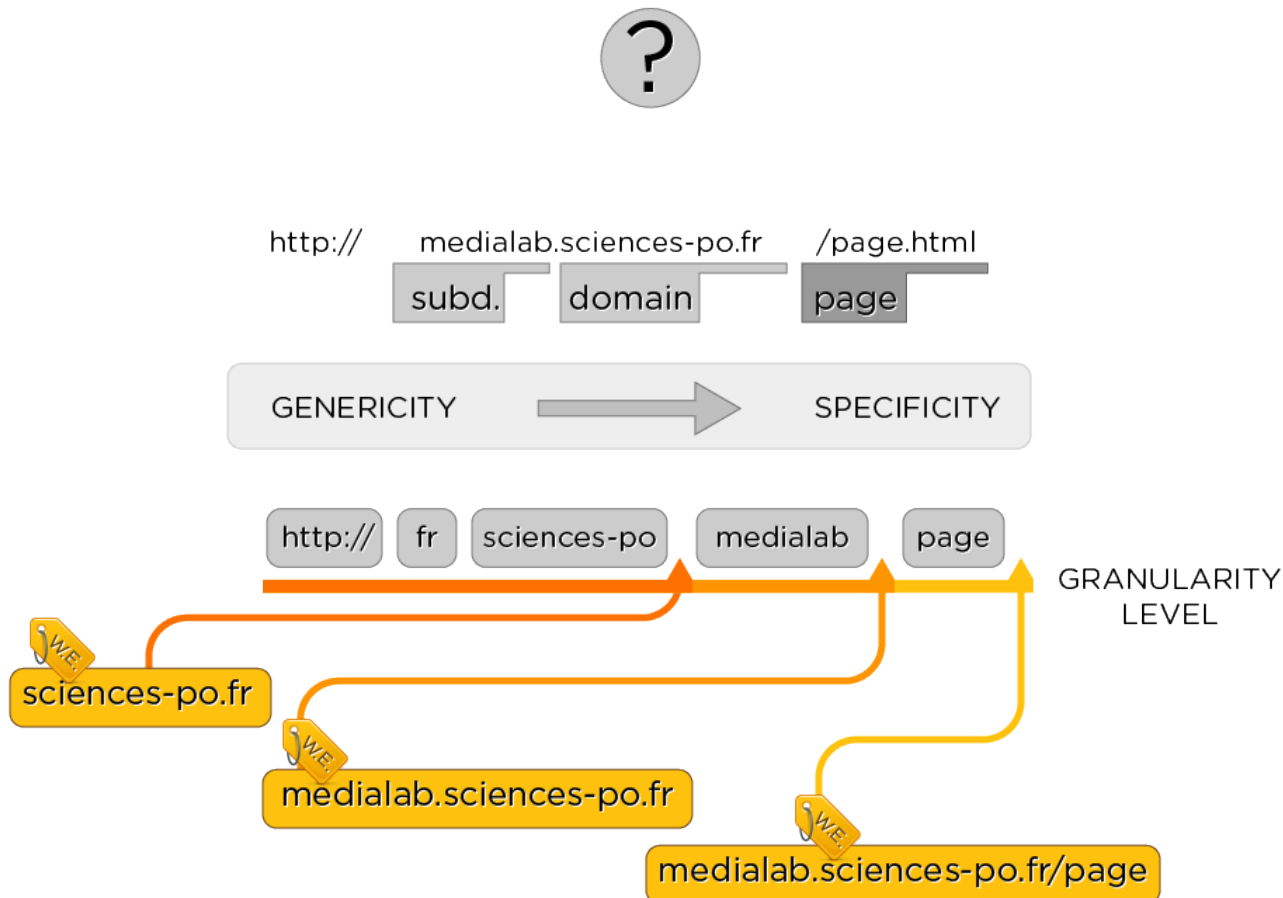
□ Hyphe : crawler orienté recherche

- 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

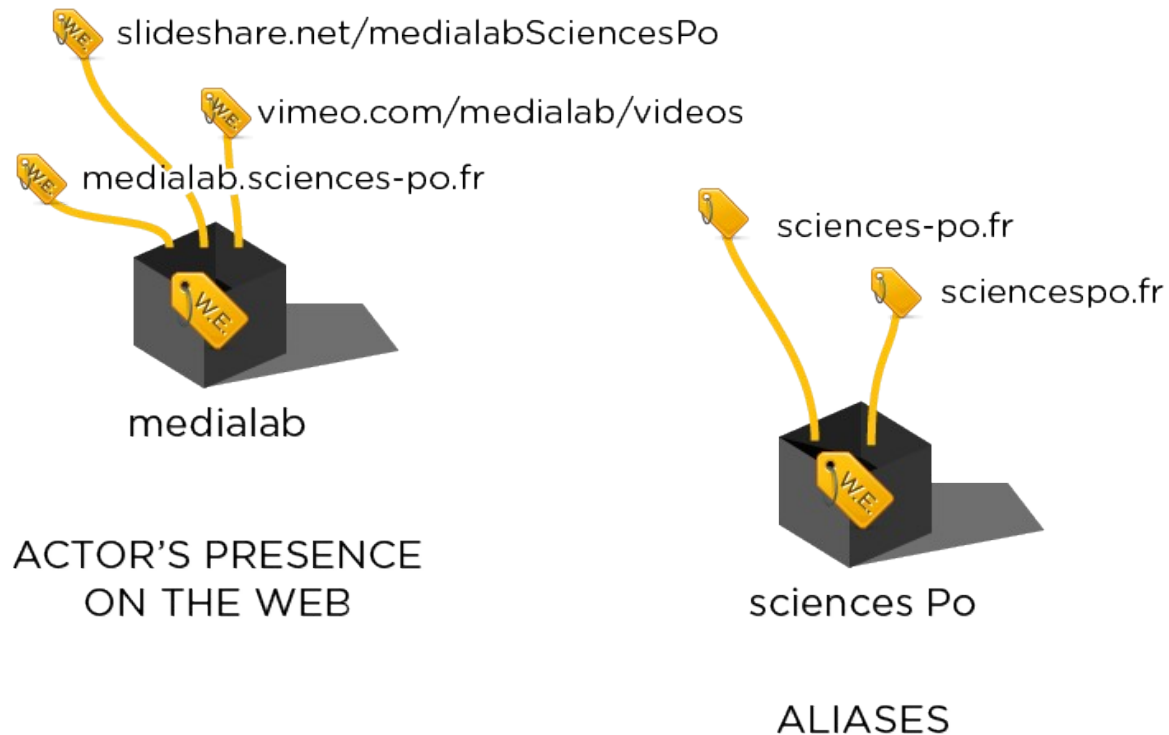
Le web en couches



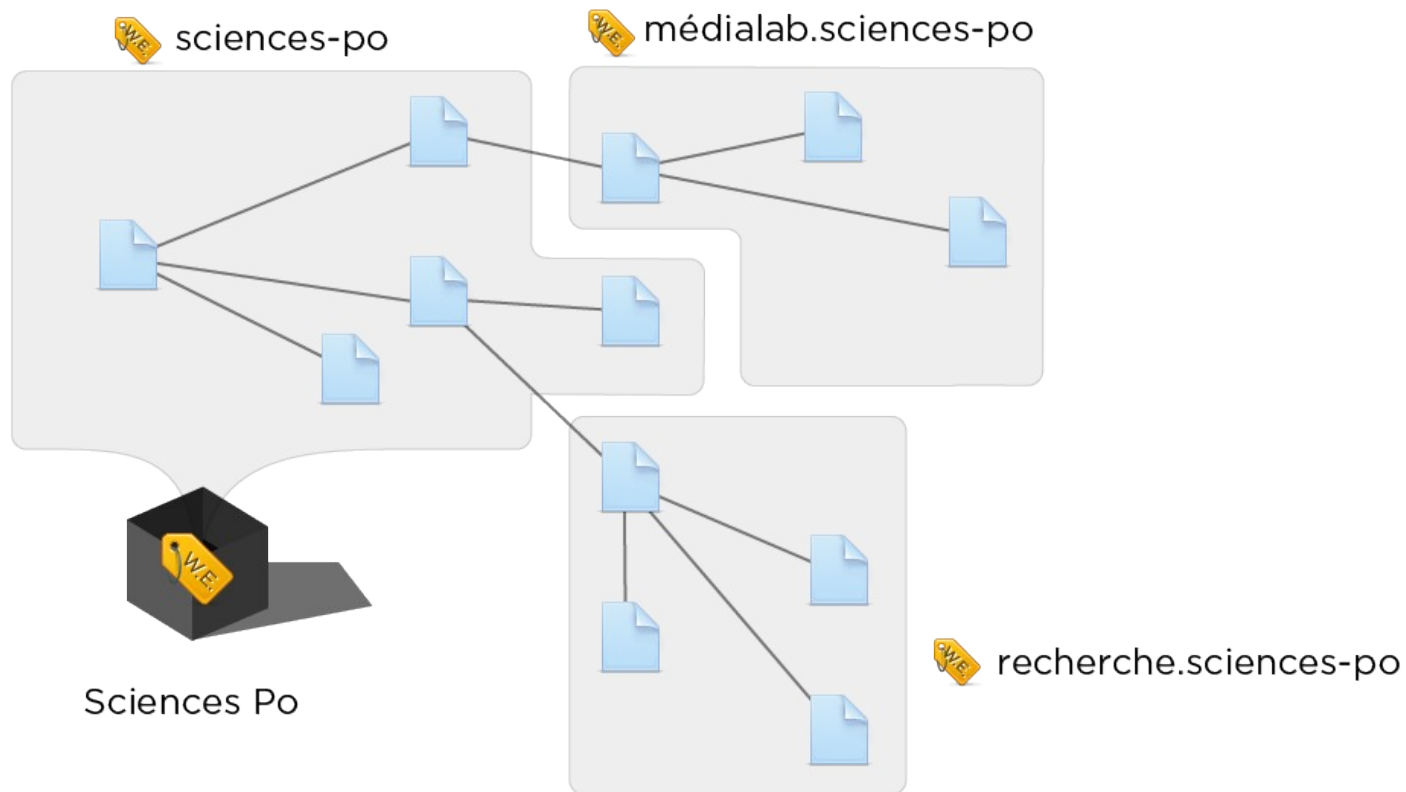
Définir des points d'ancrage précis (LRUs)



Des sites ou... des entités web



Préserver la complexité

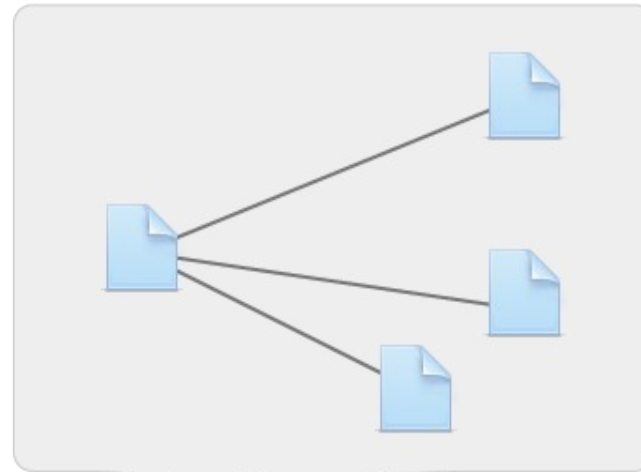


Le crawl dirigé par la recherche

1

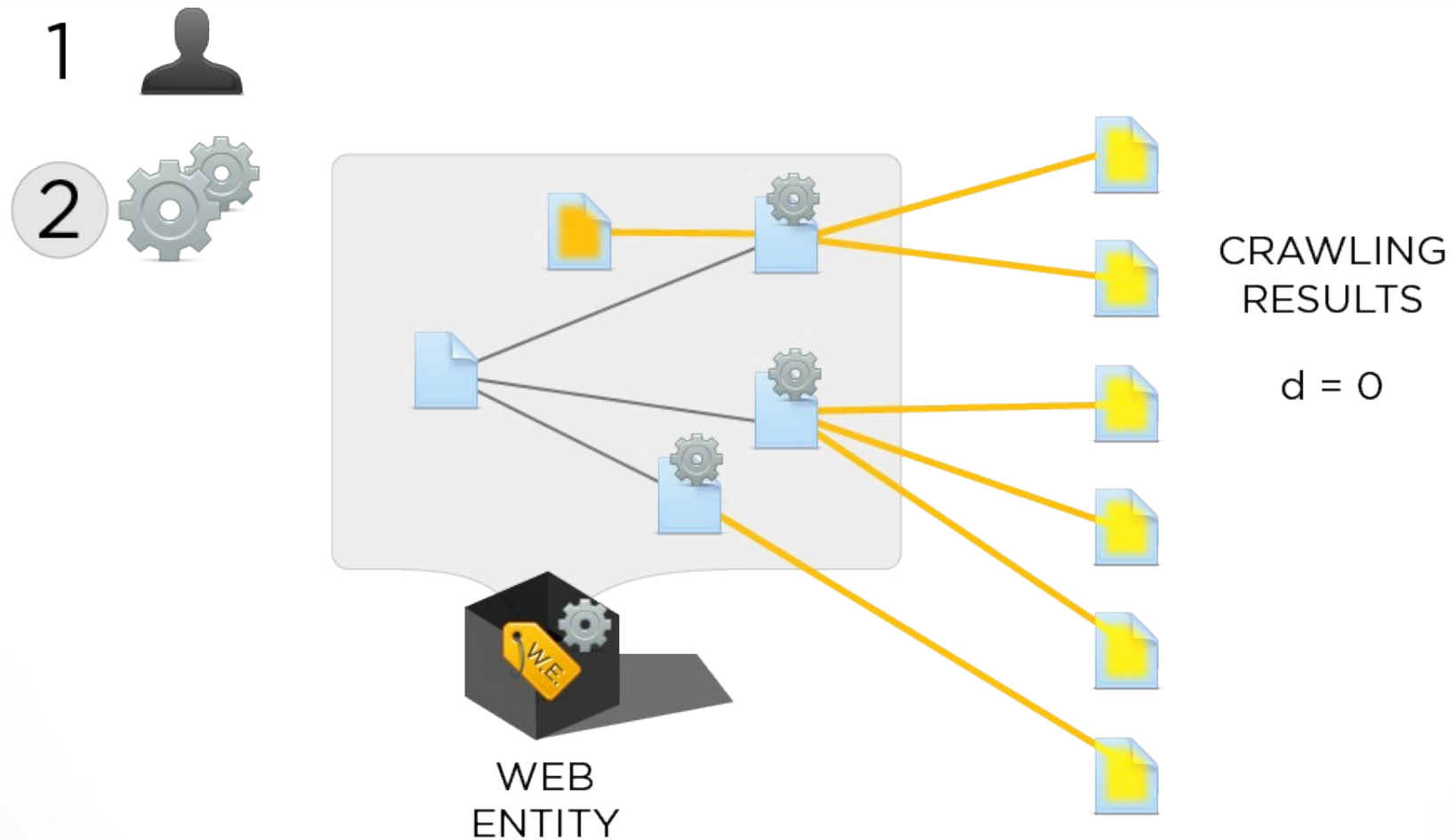


RESEARCHER
SELECT
STARTING
ENTITIES

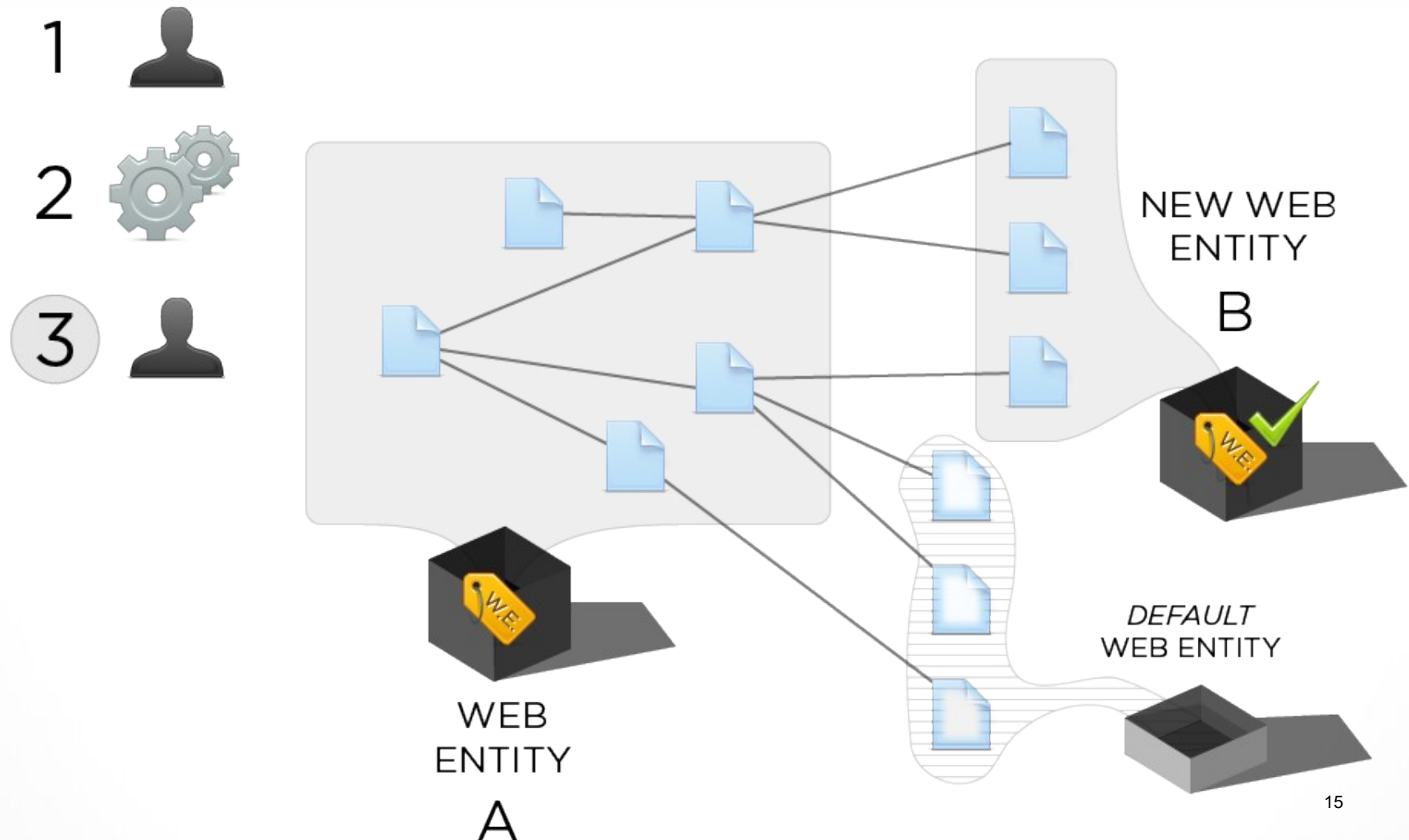


WEB
ENTITY

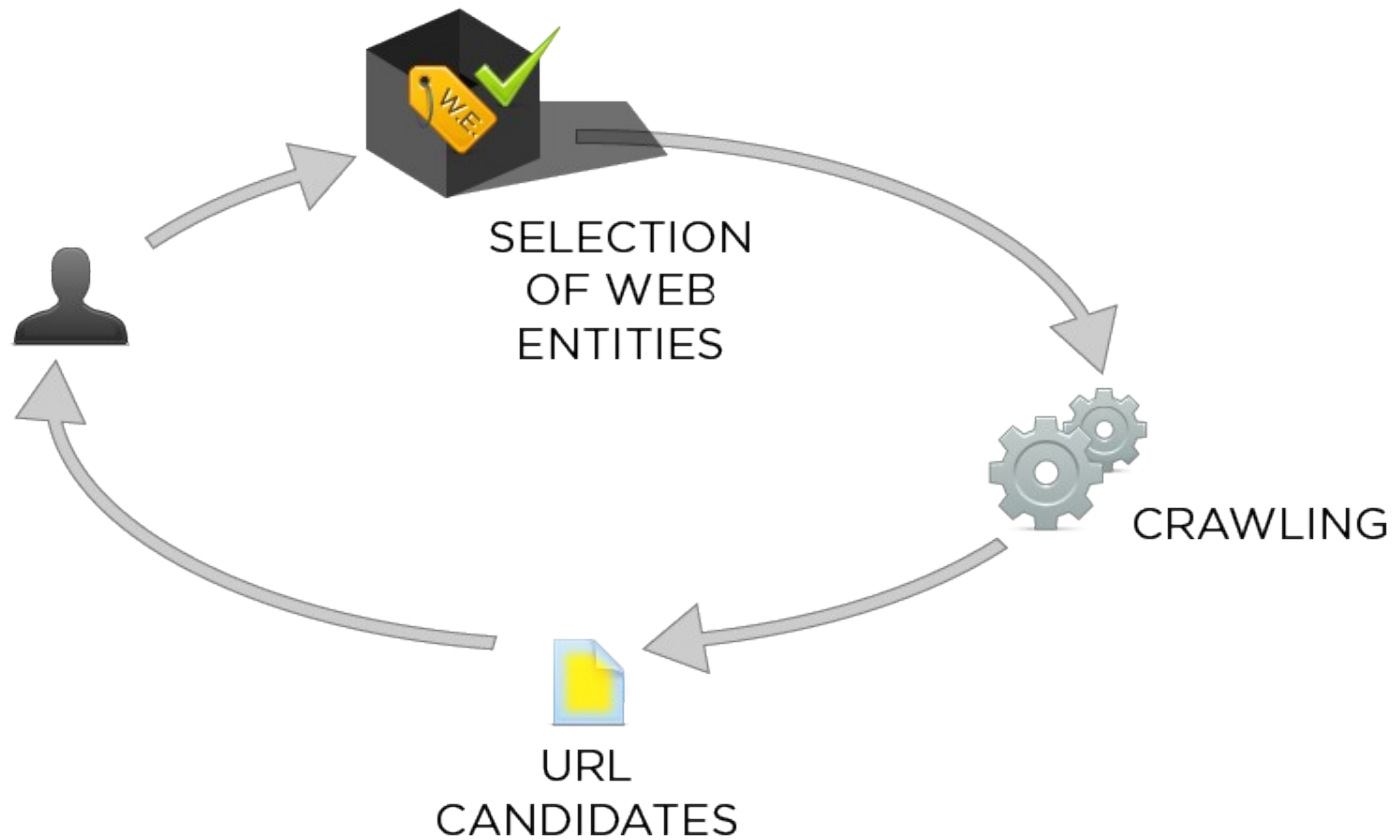
Le crawl dirigé par la recherche



Le crawl dirigé par la recherche



Le crawl dirigé par la recherche

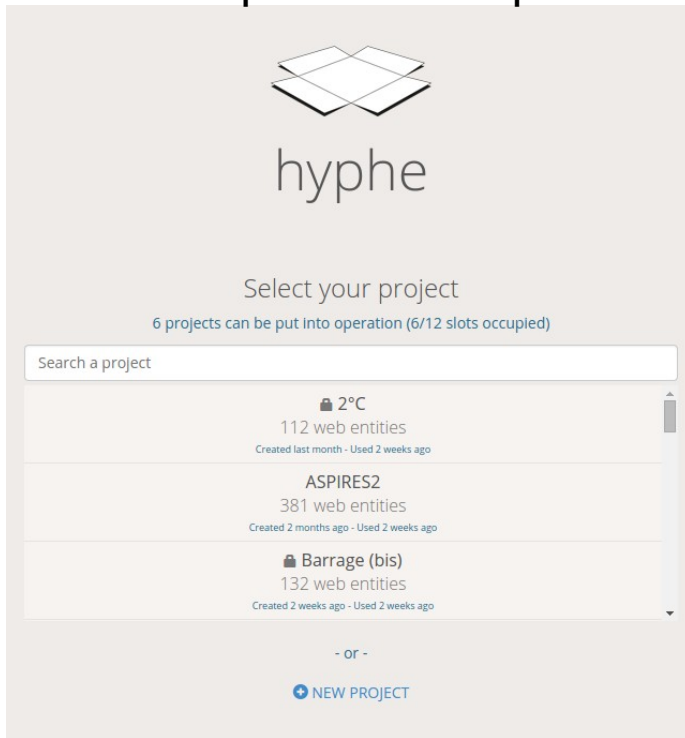


L'interface de Hyphe 2nde version

Démo : hyphe.medialab.sciences-po.fr/demo

Code source : github.com/medialab/hyphe

Gérer plusieurs corpus



The interface shows the Hyphe logo and a 'Select your project' section. Below this, a search bar is followed by a list of projects. Each project entry includes a lock icon, a name, the number of web entities, and creation/usage dates. At the bottom, there is a 'NEW PROJECT' button.

hype

Select your project

6 projects can be put into operation (6/12 slots occupied)

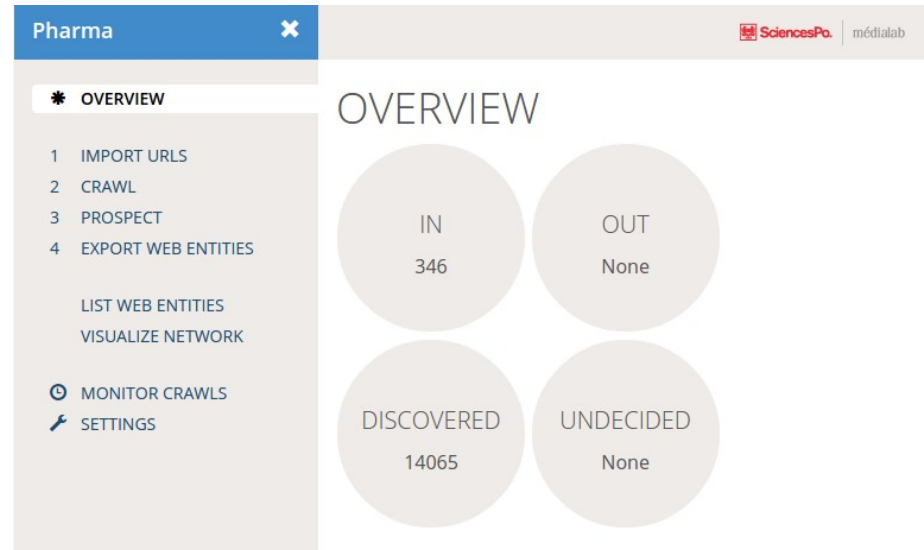
Search a project

- 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

- Or -

NEW PROJECT

Résumé d'un corpus



The interface shows a sidebar with a list of actions and a main area with an overview of the corpus statistics.

Pharma

SciencesPo. médiablab

OVERVIEW

- 1 IMPORT URLS
- 2 CRAWL
- 3 PROSPECT
- 4 EXPORT WEB ENTITIES

LIST WEB ENTITIES
VISUALIZE NETWORK

MONITOR CRAWLS
SETTINGS

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](#)

Ni-Pigeons-Ni-Espions
Scheduled just now
Depth 2

PENDING

0 page crawled

rue89.Nouvelobs
Scheduled just now
Depth 2

CRAWLING

0 page crawled

journalistecarolinelizee.Over-Blog
Scheduled just now
Depth 2

CRAWLING

12 pages crawled

assembly.Coe
Scheduled just now
Depth 2

WAITING

0 page crawled

L'interface de Hyphe 2nde version

Identifier d'autres WebEntités à inclure au corpus

PROSPECT

14065 DISCOVERED WEB ENTITIES

(Range selector not implemented yet)

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

SET TO: IN (2)

Inserm

Cdc

CRAWL

SET TO: OUT (16)

SET TO: UNDECIDED (1)

First

Previous

1

2

3

4

5

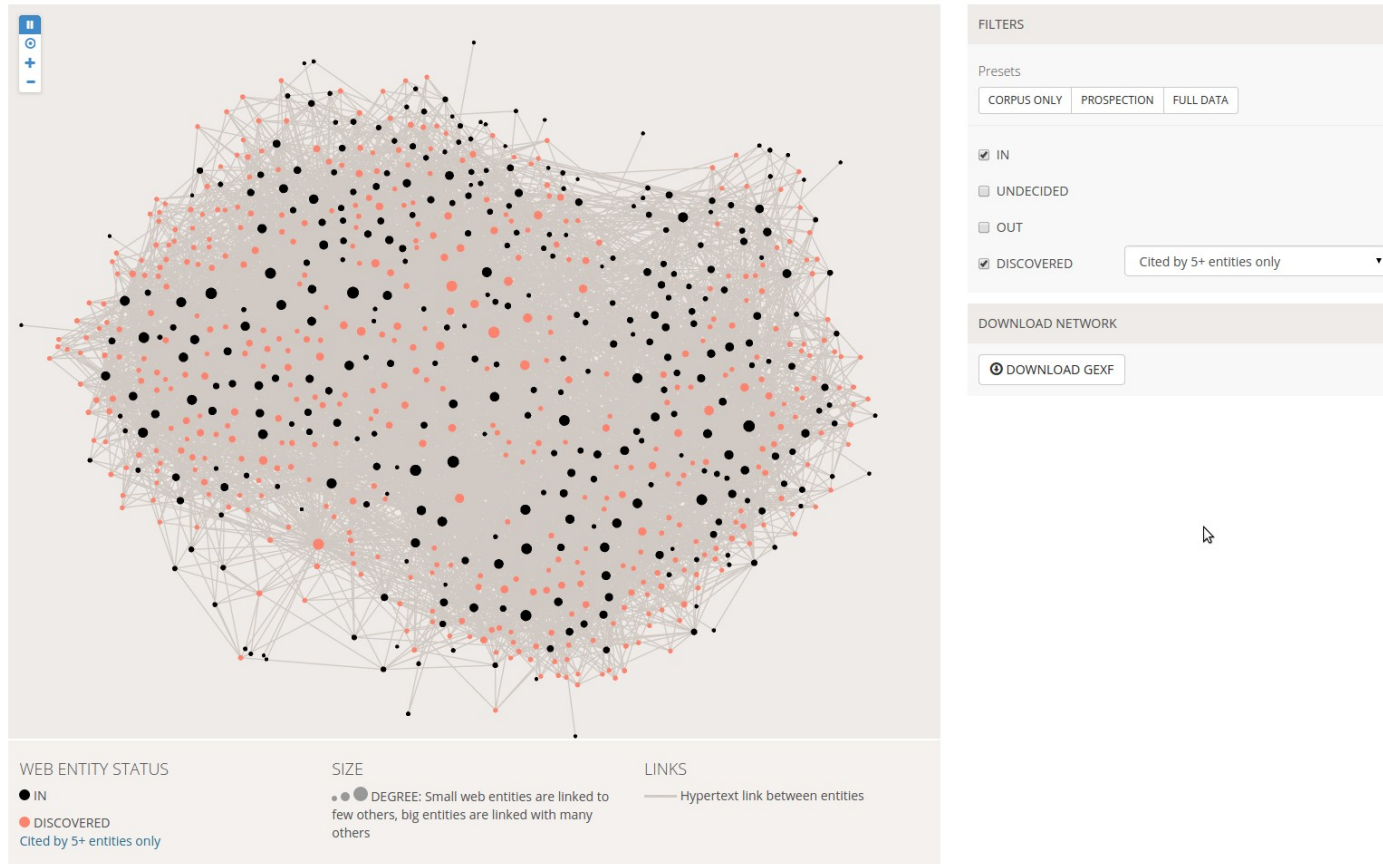
...

Next

Last

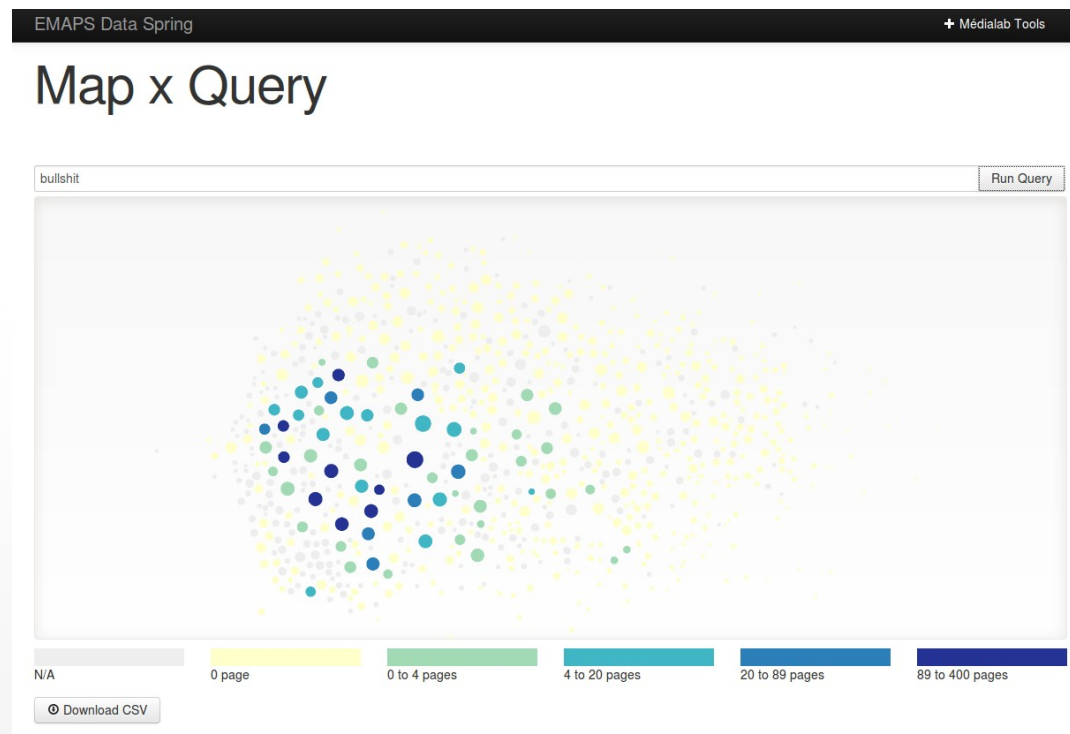
L'interface de Hyphe 2nde version

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