



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-03631536v1>

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 médiablab

- Fondé en mai 2009

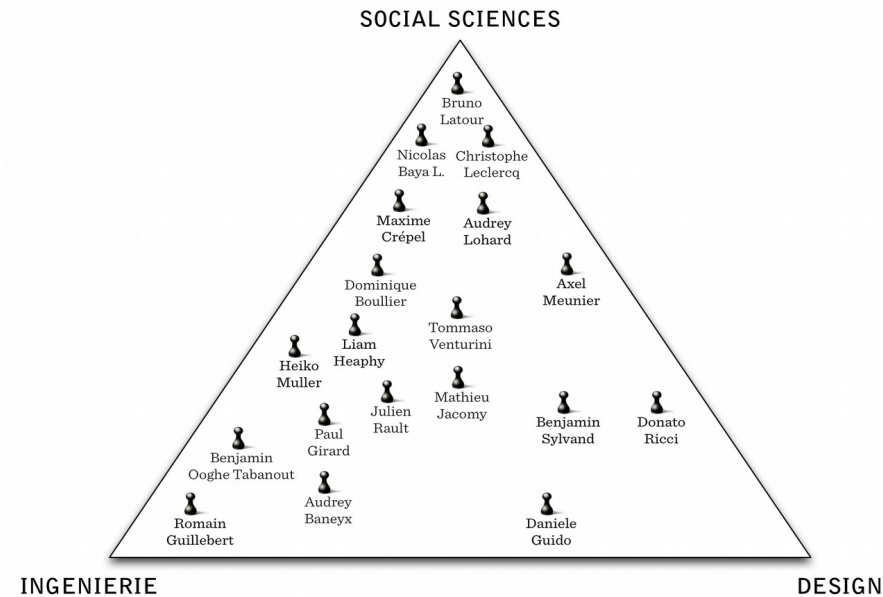
Centre de recherche numérique
au service de SciencesPo et des

- sciences sociales.

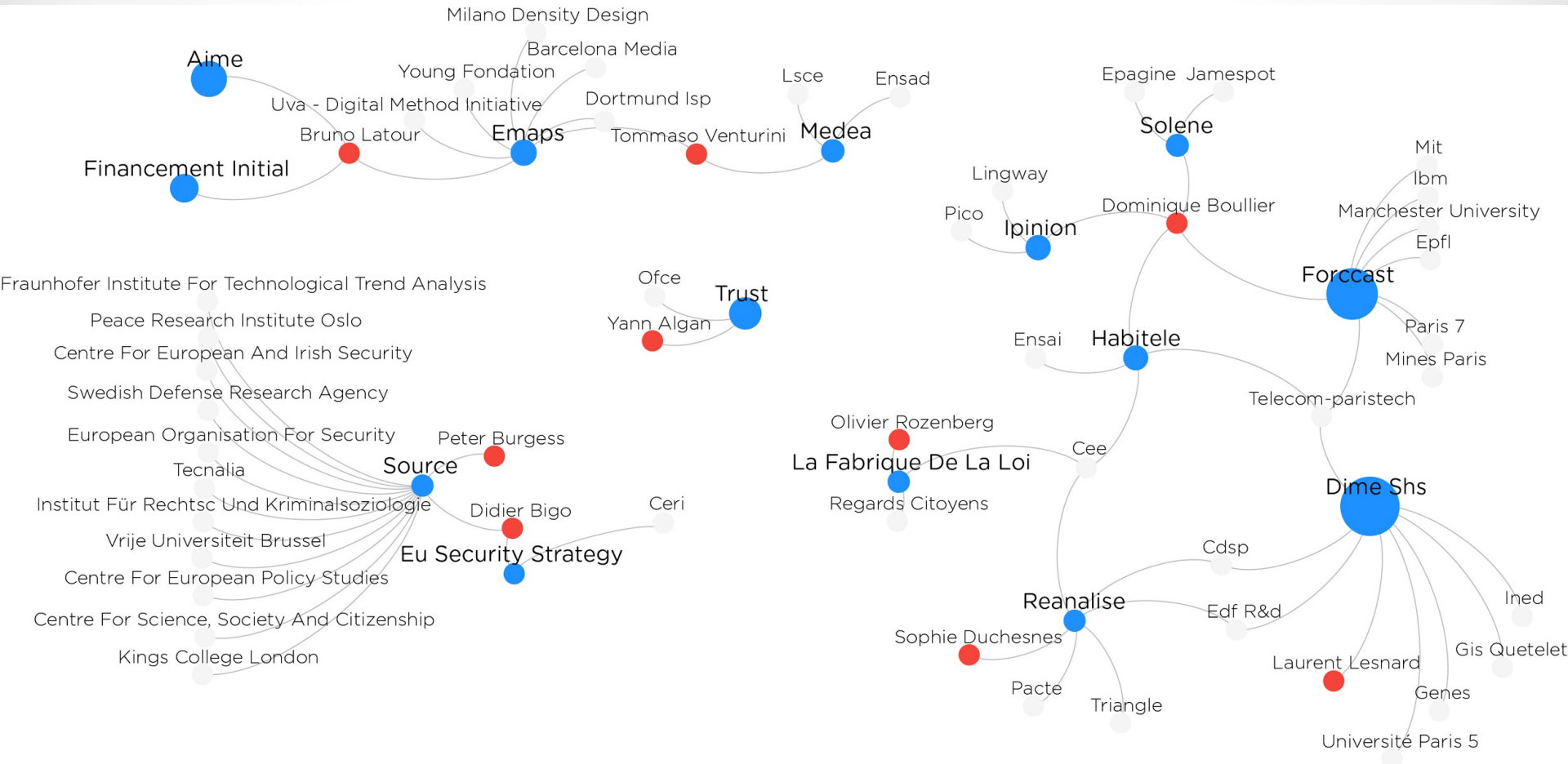
Étude des traces numériques :
articuler les méthodes

- 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

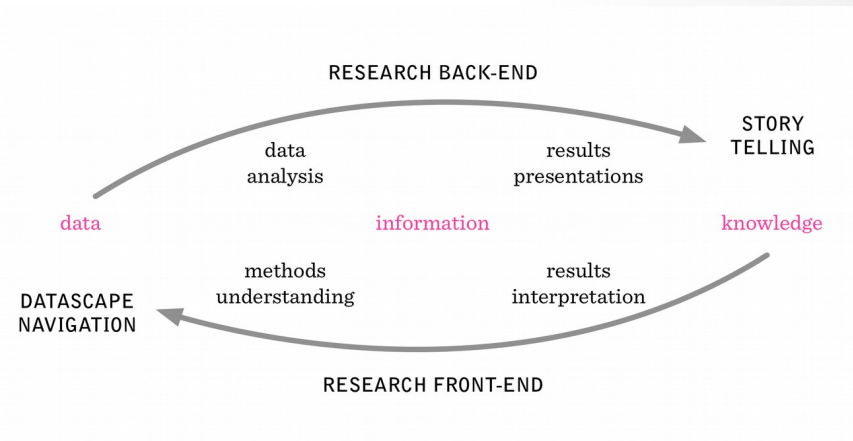
- Méthodologie itérative

- 2 personnes (Mathieu Jacomy & moi-même)

- Objectif ANR d'auto-financement

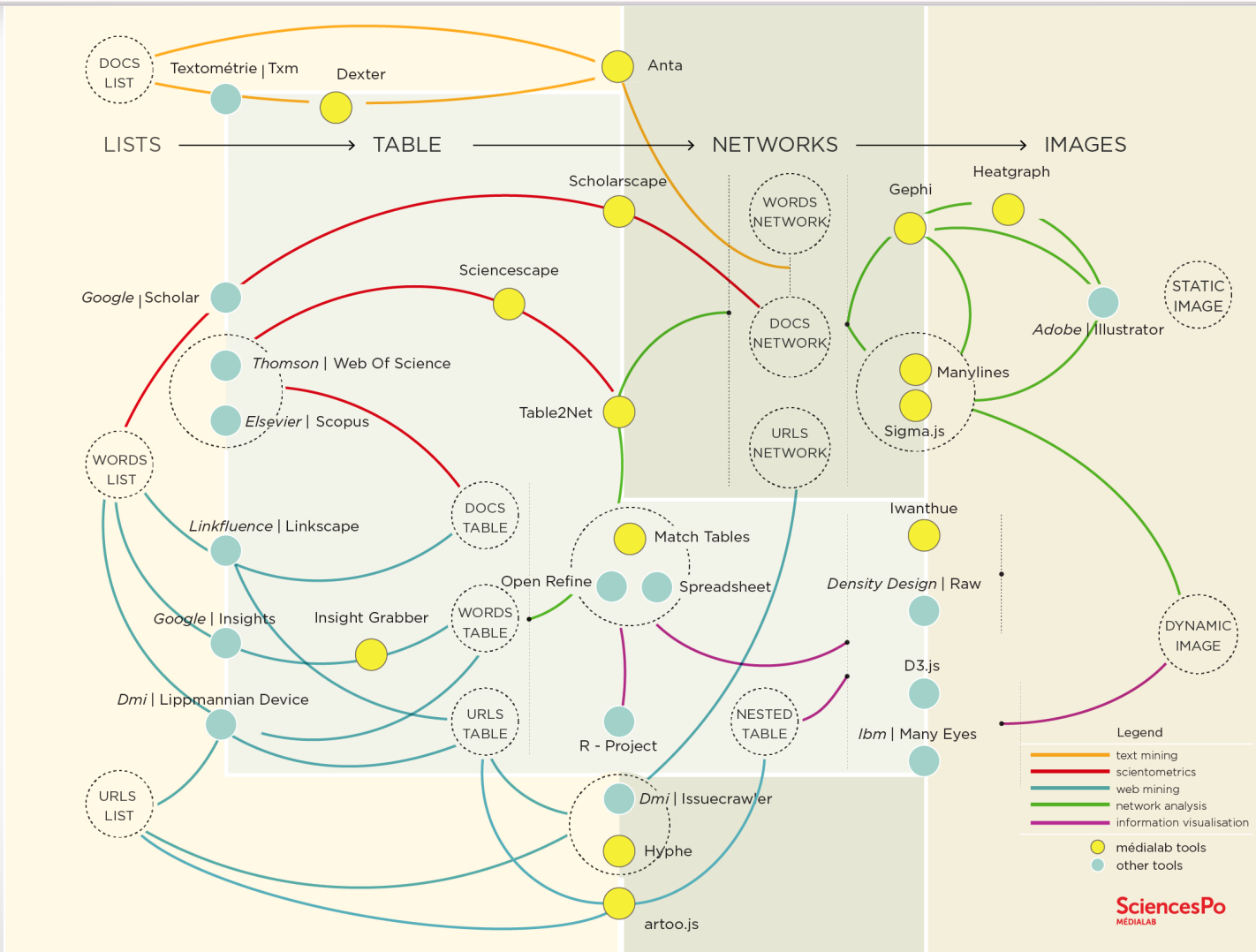
- ⇒ offre de service payant

- ⇒ mutualisation (logiciels libres/OpenSource)



Un écosystème d'outils

tools.medialab.sciences-po.fr

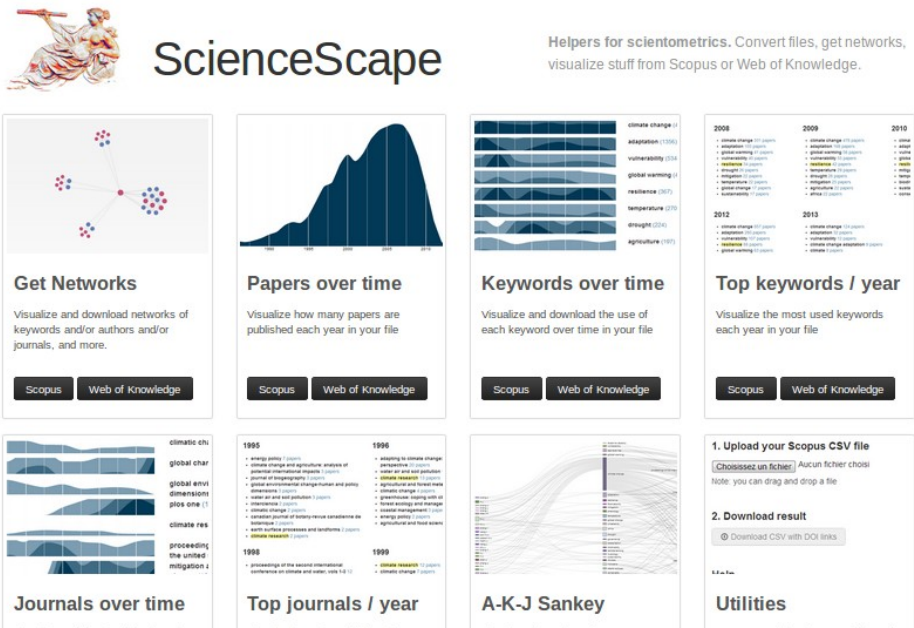


Collectes ciblées (scraping)

- 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
 - - ...



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

Get Networks
Visualize and download networks of keywords, and/or authors and/or journals, and more.

Papers over time
Visualize how many papers are published each year in your file

Keywords over time
Visualize and download the use of each keyword over time in your file

Top keywords / year
Visualize the most used keywords each year in your file

Journals over time
Visualize and download the journals

Top journals / year
Visualize journals publishing the most

A-K-J Sankey
Visualize the main authors

Utilities
Extract CSV index from your file, and

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@procolias@cauc@spacence-verche@em@schel-tem@jean-pierre-deco@p@henry-lazar@od@sen-	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. Dhruaj, M. Verche, M. Temst, M. Decool, M. Lacroix, 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 et des le p..."	2014-02-23T23:06:26	fr	46.17462263;5.14000353

TEXT - WORDS CLOUD

TIMESTAMP - DAILY VOLUME

LANG - ITEMS TOP 20

Lang	Count	Lang	Count	Lang	Count
fr	12,46 (60)	en	23,8 (117)	cy	34 (17)
es	13 (61)	pt	24 (111)	th	35 (17)
de	14 (71)	it	25 (124)	ko	36 (18)
pt	15 (74)	ru	26 (127)	ba	37 (18)
nl	16 (78)	tr	27 (132)	zh	38 (19)
ru	17 (83)	uk	28 (137)	sv	39 (19)
uk	18 (88)	pl	29 (142)	vi	40 (20)
pl	19 (93)	fr	30 (147)	id	41 (21)
fr	20 (98)	nl	31 (152)	de	42 (21)
nl	21 (103)	en	32 (157)	uk	43 (21)
en	22 (108)	pt	33 (162)	ru	44 (21)
pt	23 (113)	de	34 (167)	fr	45 (21)
de	24 (118)	it	35 (172)	en	46 (21)
it	25 (123)	tr	36 (177)	pt	47 (21)
tr	26 (128)	uk	37 (182)	it	48 (21)
uk	27 (133)	pl	38 (187)	nl	49 (21)
pl	28 (138)	ru	39 (192)	uk	50 (21)
ru	29 (143)	pl	40 (197)	de	51 (21)
ru	30 (148)	nl	41 (202)	en	52 (21)
nl	31 (153)	fr	42 (207)	pt	53 (21)
fr	32 (158)	en	43 (212)	ru	54 (21)
en	33 (163)	de	44 (217)	it	55 (21)
de	34 (168)	it	45 (222)	nl	56 (21)
it	35 (173)	tr	46 (227)	uk	57 (21)
tr	36 (178)	pt	47 (232)	pl	58 (21)
pt	37 (183)	uk	48 (237)	ru	59 (21)
uk	38 (188)	nl	49 (242)	en	60 (21)
uk	39 (193)	fr	50 (247)	pt	61 (21)
fr	40 (198)	en	51 (252)	ru	62 (21)
en	41 (203)	de	52 (257)	it	63 (21)
de	42 (208)	it	53 (262)	nl	64 (21)
it	43 (213)	tr	54 (267)	uk	65 (21)
tr	44 (218)	pt	55 (272)	pl	66 (21)
pt	45 (223)	uk	56 (277)	ru	67 (21)
uk	46 (228)	nl	57 (282)	en	68 (21)
uk	47 (233)	fr	58 (287)	pt	69 (21)
fr	48 (238)	en	59 (292)	ru	70 (21)
en	49 (243)	de	60 (297)	it	71 (21)
de	50 (248)	it	61 (302)	nl	72 (21)
it	51 (253)	tr	62 (307)	uk	73 (21)
tr	52 (258)	pt	63 (312)	pl	74 (21)
pt	53 (263)	uk	64 (317)	ru	75 (21)
uk	54 (268)	nl	65 (322)	en	76 (21)
uk	55 (273)	fr	66 (327)	pt	77 (21)
fr	56 (278)	en	67 (332)	ru	78 (21)
en	57 (283)	de	68 (337)	it	79 (21)
de	58 (288)	it	69 (342)	nl	80 (21)
it	59 (293)	tr	70 (347)	uk	81 (21)
tr	60 (298)	pt	71 (352)	pl	82 (21)
pt	61 (303)	uk	72 (357)	ru	83 (21)
uk	62 (308)	nl	73 (362)	en	84 (21)
uk	63 (313)	fr	74 (367)	pt	85 (21)
fr	64 (318)	en	75 (372)	ru	86 (21)
en	65 (323)	de	76 (377)	it	87 (21)
de	66 (328)	it	77 (382)	nl	88 (21)
it	67 (333)	tr	78 (387)	uk	89 (21)
tr	68 (338)	pt	79 (392)	pl	90 (21)
pt	69 (343)	uk	80 (397)	ru	91 (21)
uk	70 (348)	nl	81 (402)	en	92 (21)
uk	71 (353)	fr	82 (407)	pt	93 (21)
fr	72 (358)	en	83 (412)	ru	94 (21)
en	73 (363)	de	84 (417)	it	95 (21)
de	74 (368)	it	85 (422)	nl	96 (21)
it	75 (373)	tr	86 (427)	uk	97 (21)
tr	76 (378)	pt	87 (432)	pl	98 (21)
pt	77 (383)	uk	88 (437)	ru	99 (21)
uk	78 (388)	nl	89 (442)	en	100 (21)

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

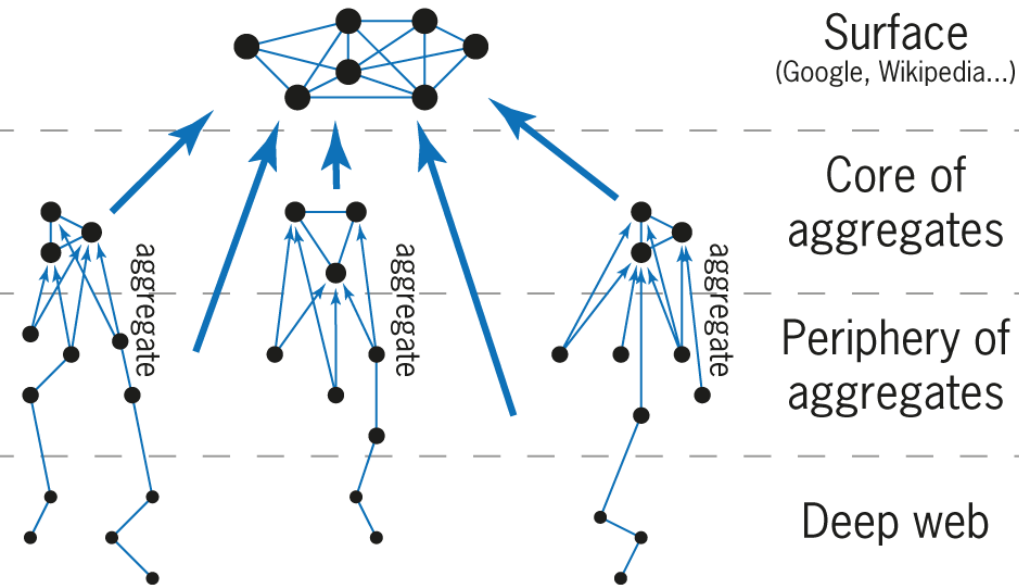
original degree indegree outdegree

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

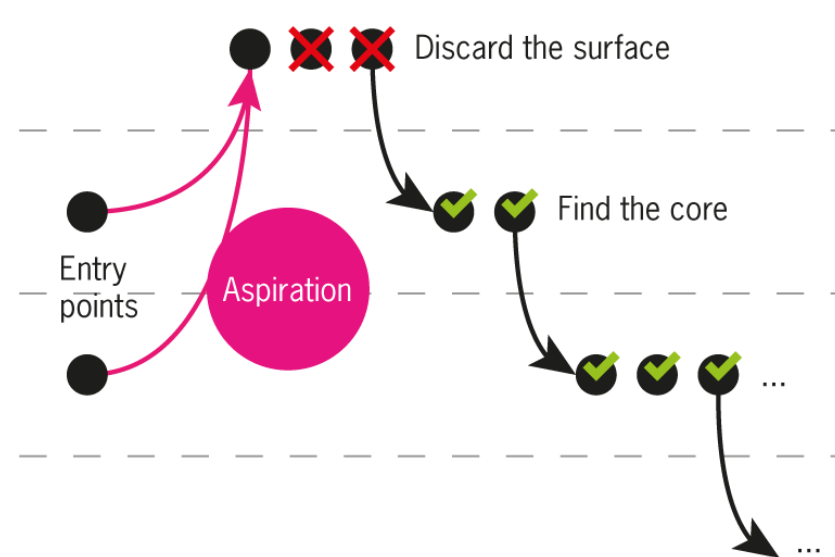
□ Hyphe : crawler orienté recherche

- 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

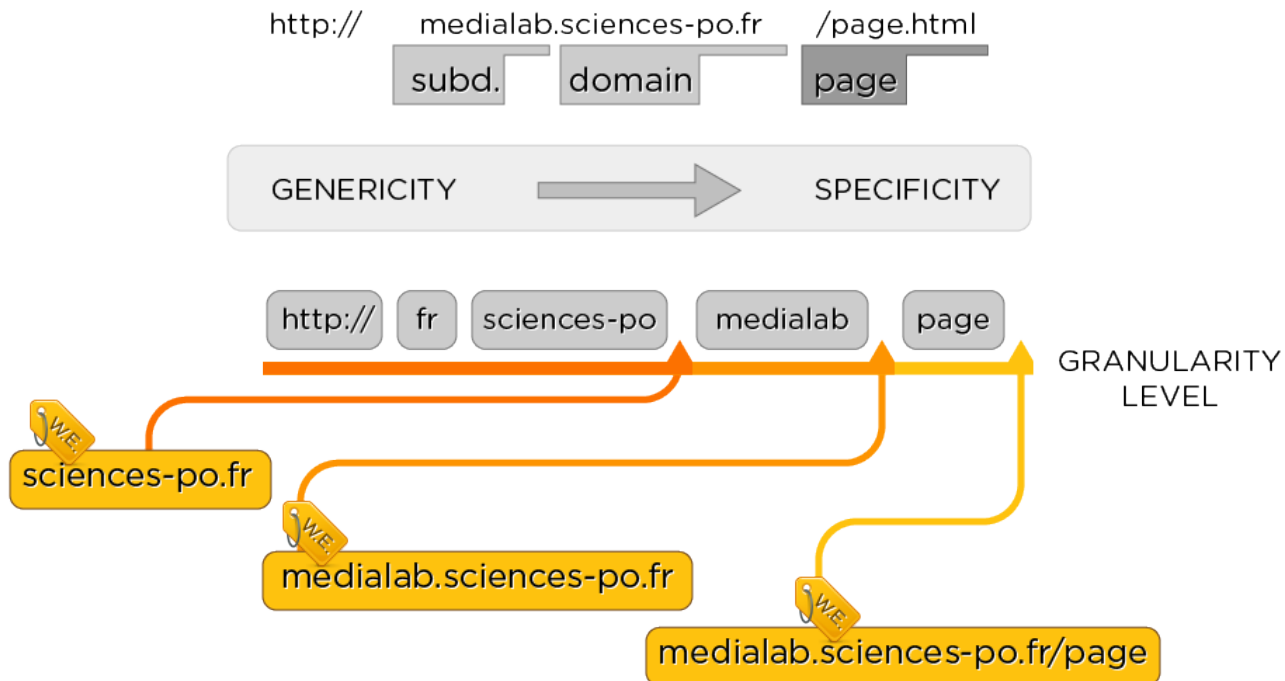


Layers of the web

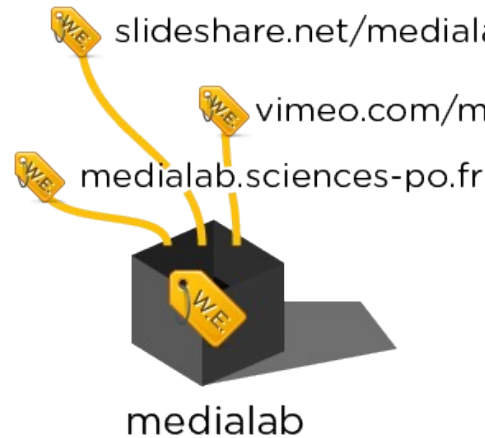


Corpus building scenario

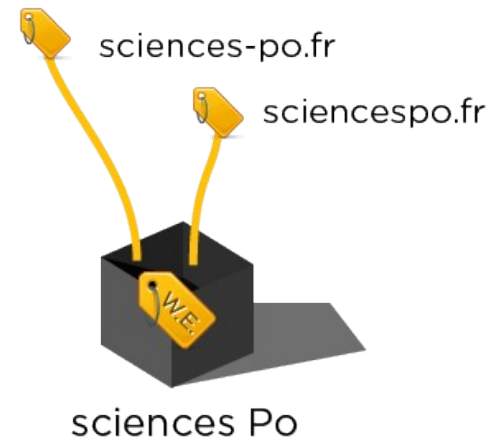
Définir des points d'ancrage précis (LRUs)



Des sites ou... des entités web

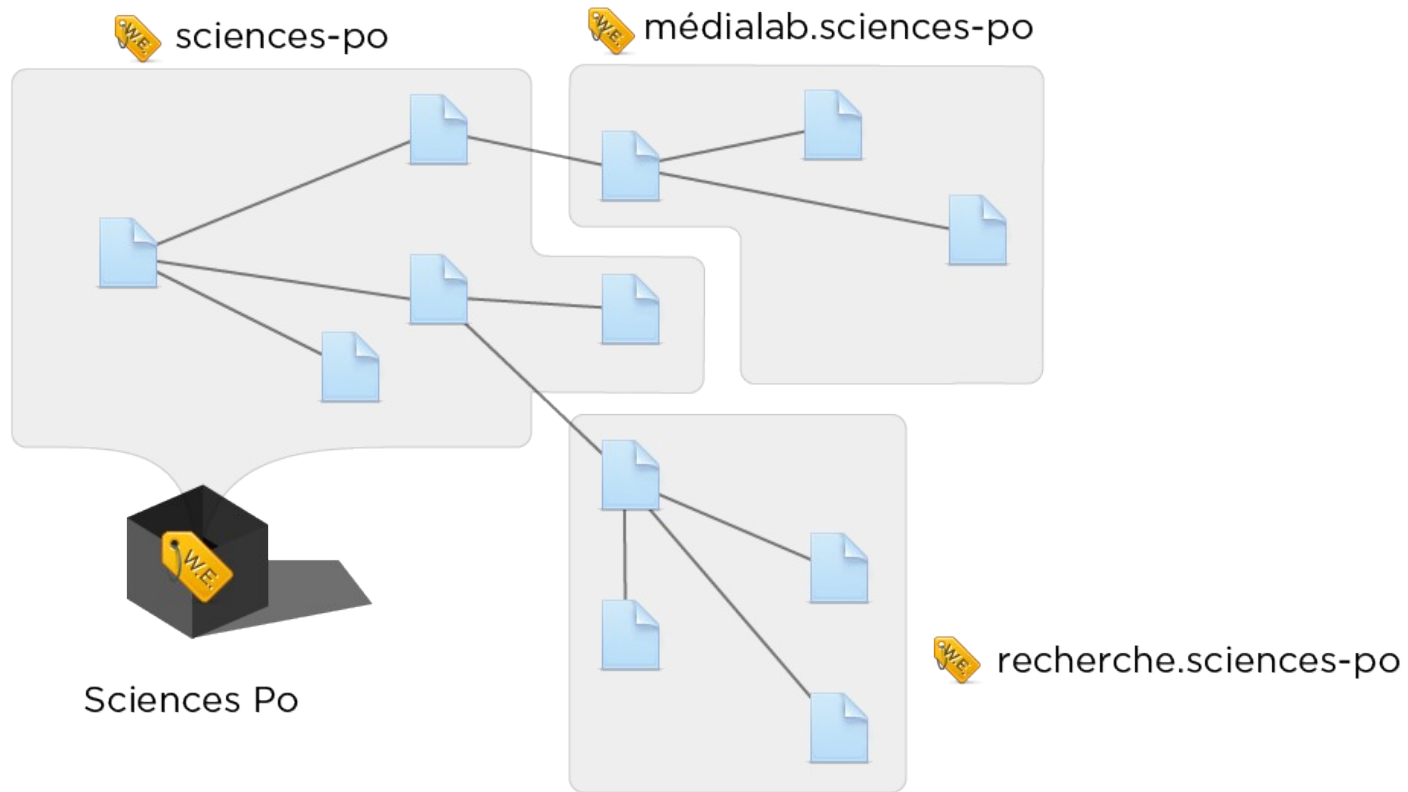


ACTOR'S PRESENCE
ON THE WEB



ALIASES

Préserver la complexité

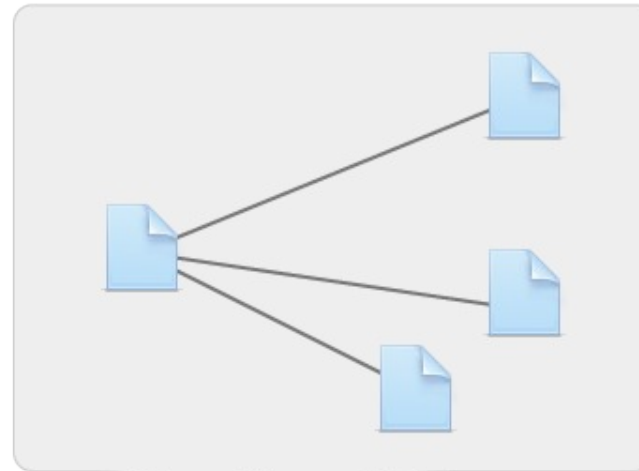


Le crawl dirigé par la recherche

1

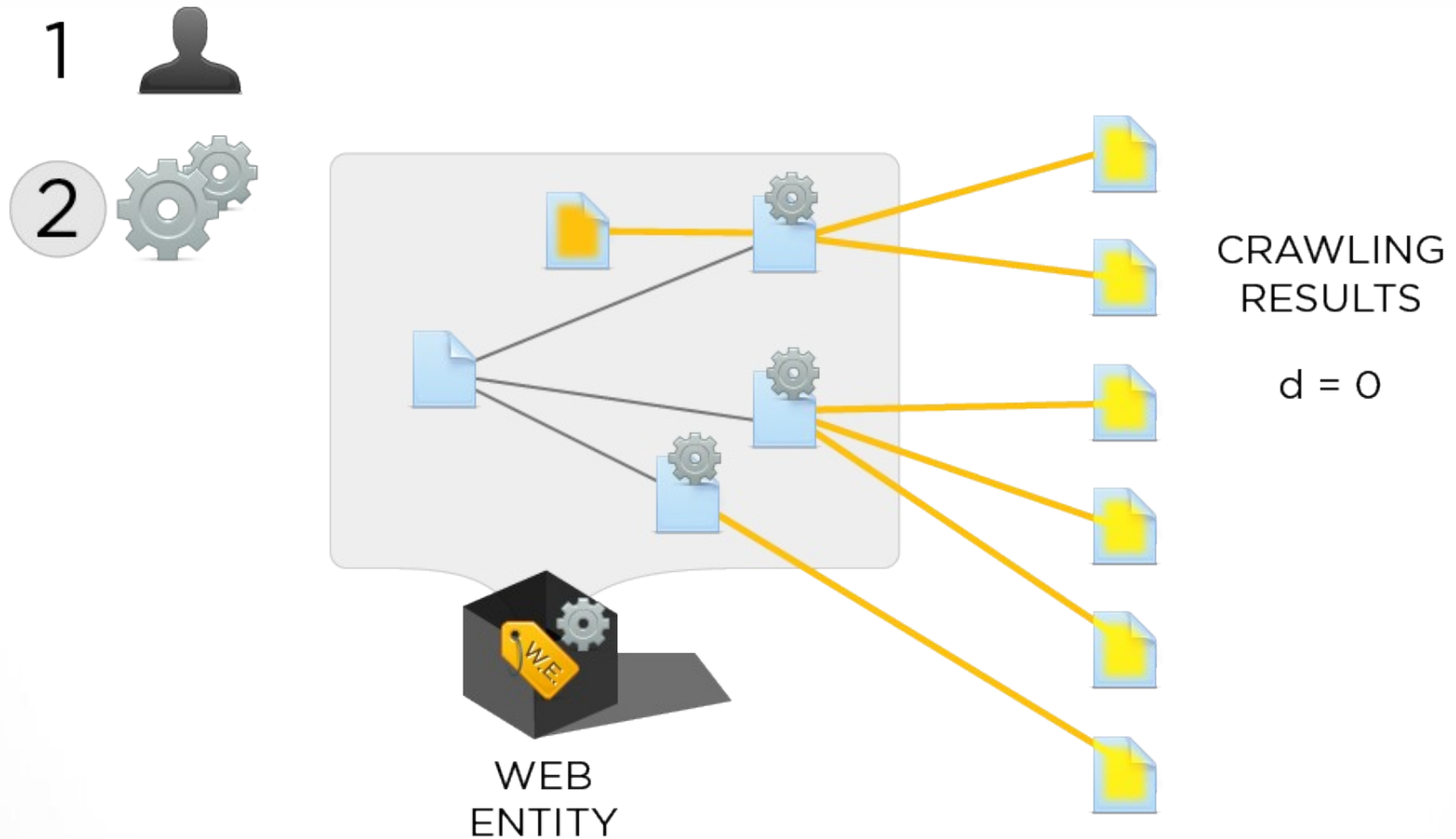


RESEARCHER
SELECT
STARTING
ENTITIES

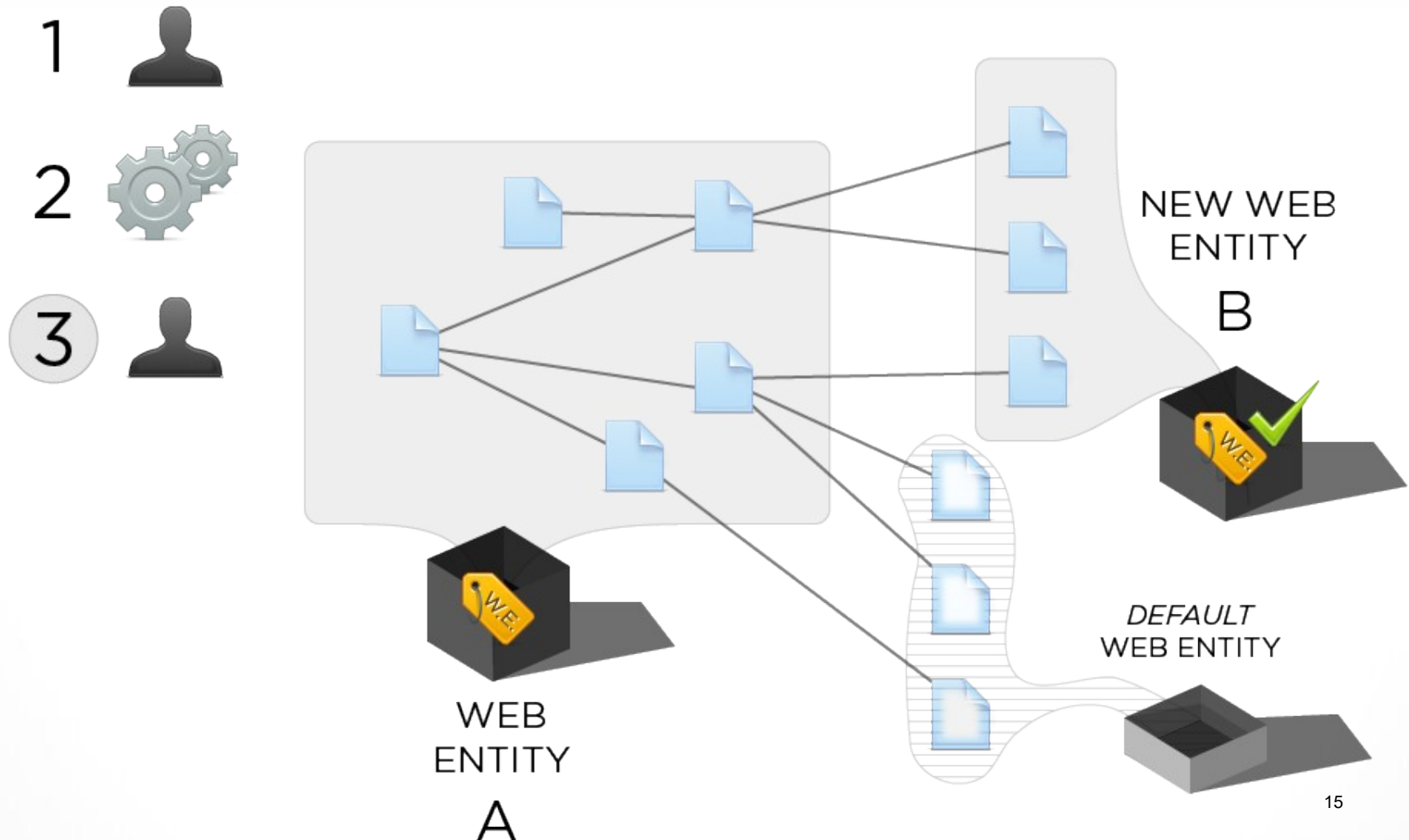


WEB
ENTITY

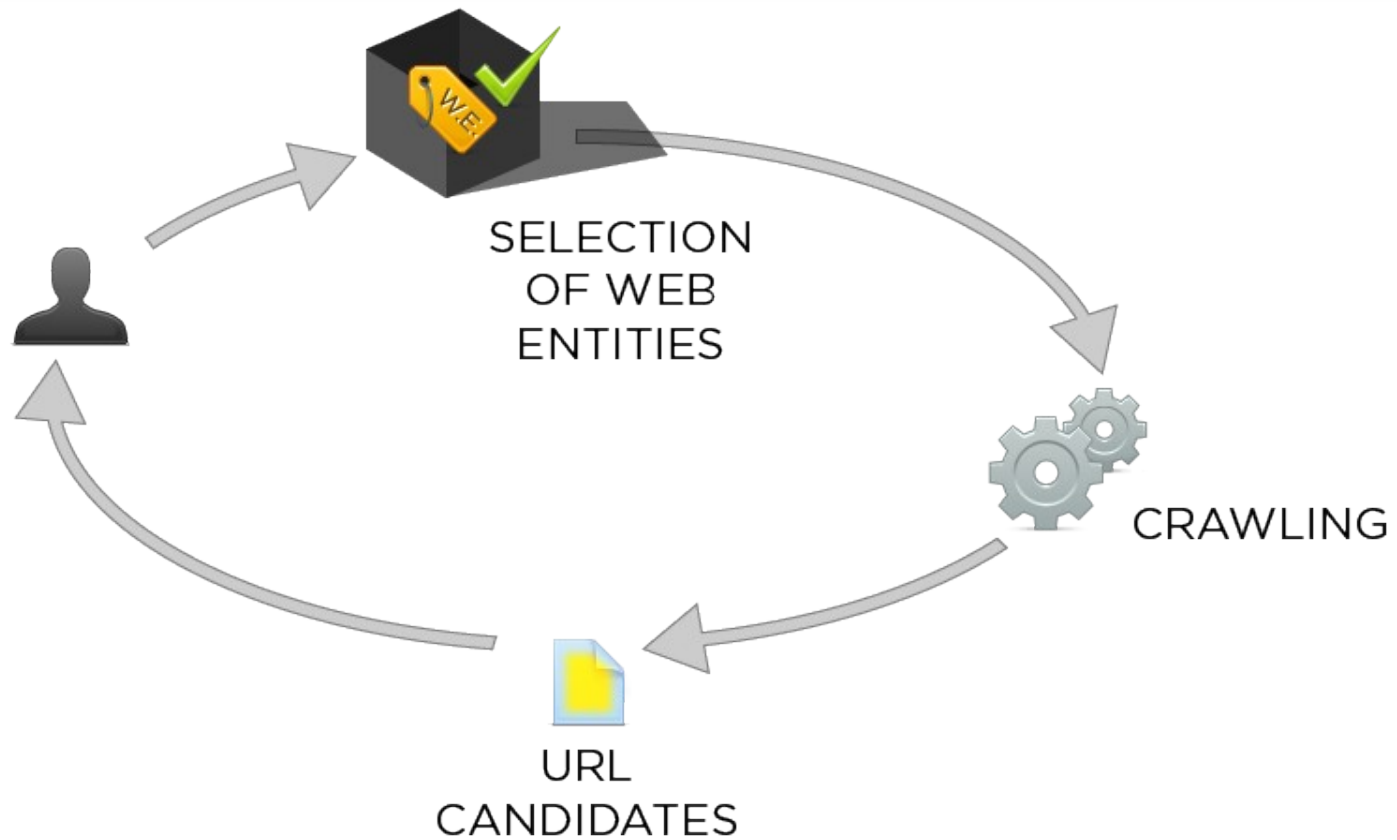
Le crawl dirigé par la recherche



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

Code source : 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 "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 left sidebar and a main content area.

Left Sidebar (Navigation):

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

Main Content Area (OVERVIEW):

The overview section displays four circular statistics:

- IN**: 346
- OUT**: None
- DISCOVERED**: 14065
- UNDECIDED**: None

The top right corner of the interface shows the SciencesPo. and médiablab logos.

L'interface de Hyphe 2nde version

Définir précisément les WebEntités

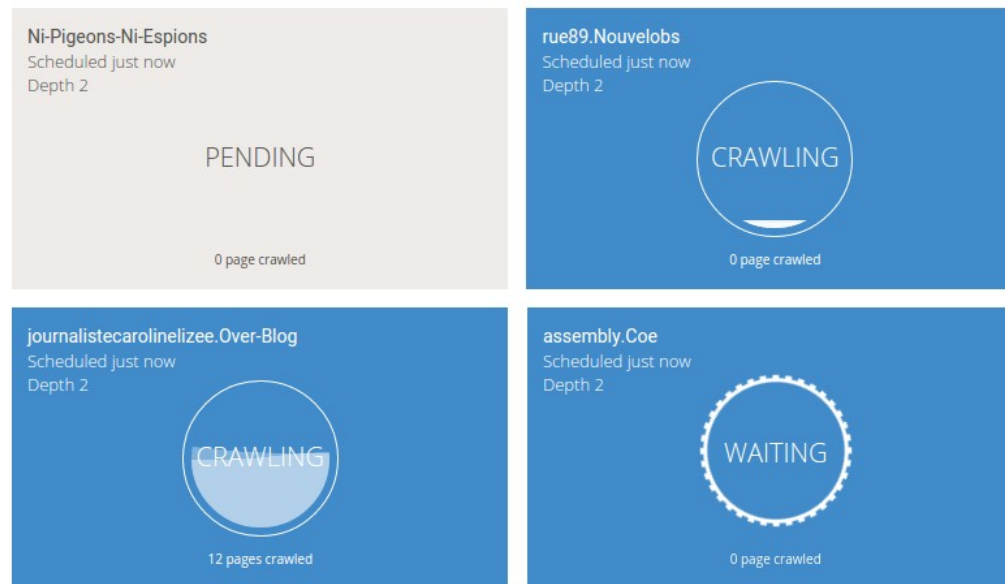
12	Actuchomage New ⚠ Same web entity defined row 614	http .org actuchomage www.
13	Afev New	http .fr afev www.
14	Nouvelobs New ⚠ Same web entity defined rows 15, 112, 115, 146, 485, 551, 587, 601, 785, 809 and 912	http .com nouvelobs blogs. globe.
15	Nouvelobs New ⚠ Same web entity defined rows 14, 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

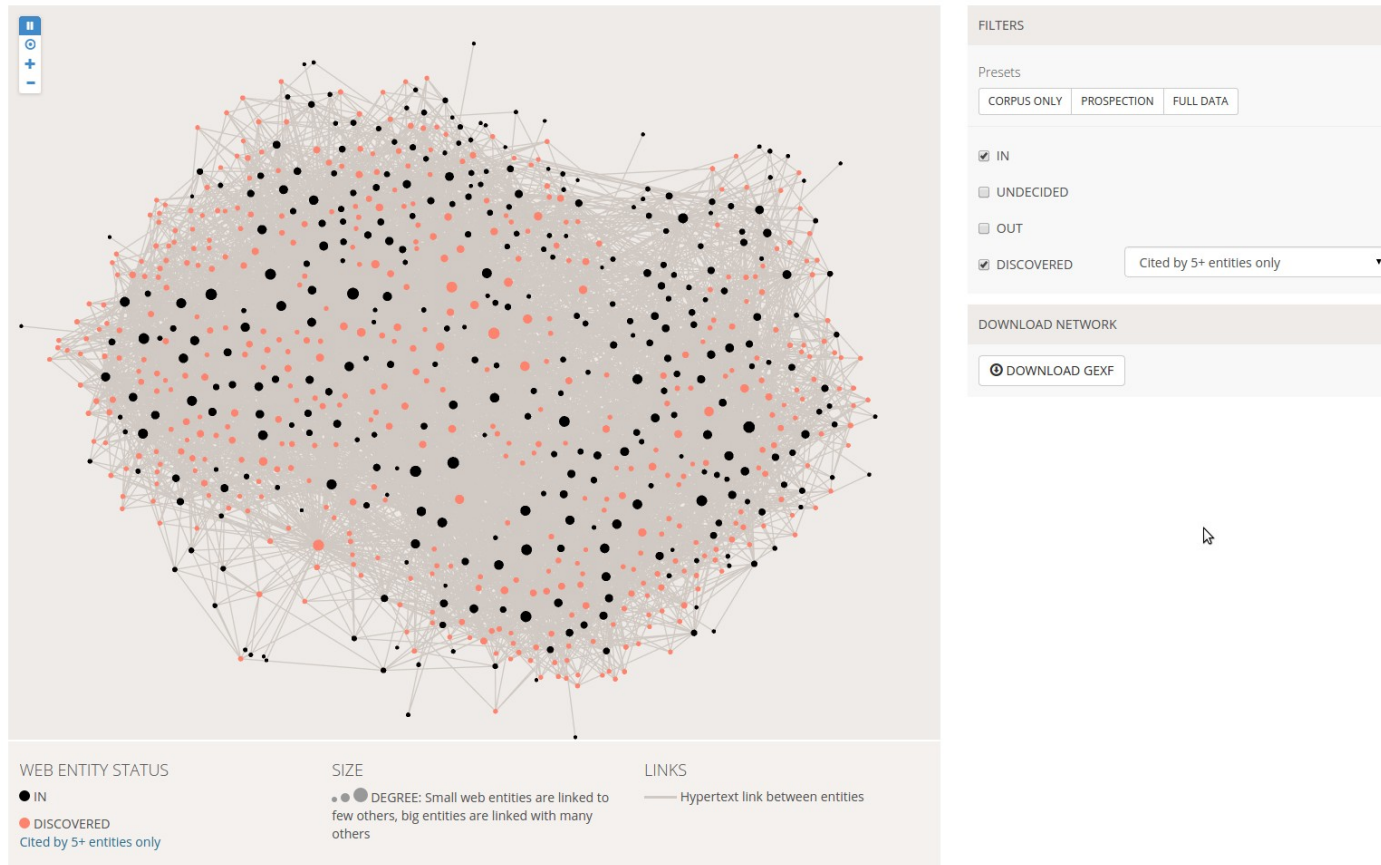
Cdc

SET TO: OUT (16)

SET TO: UNDECIDED (1)

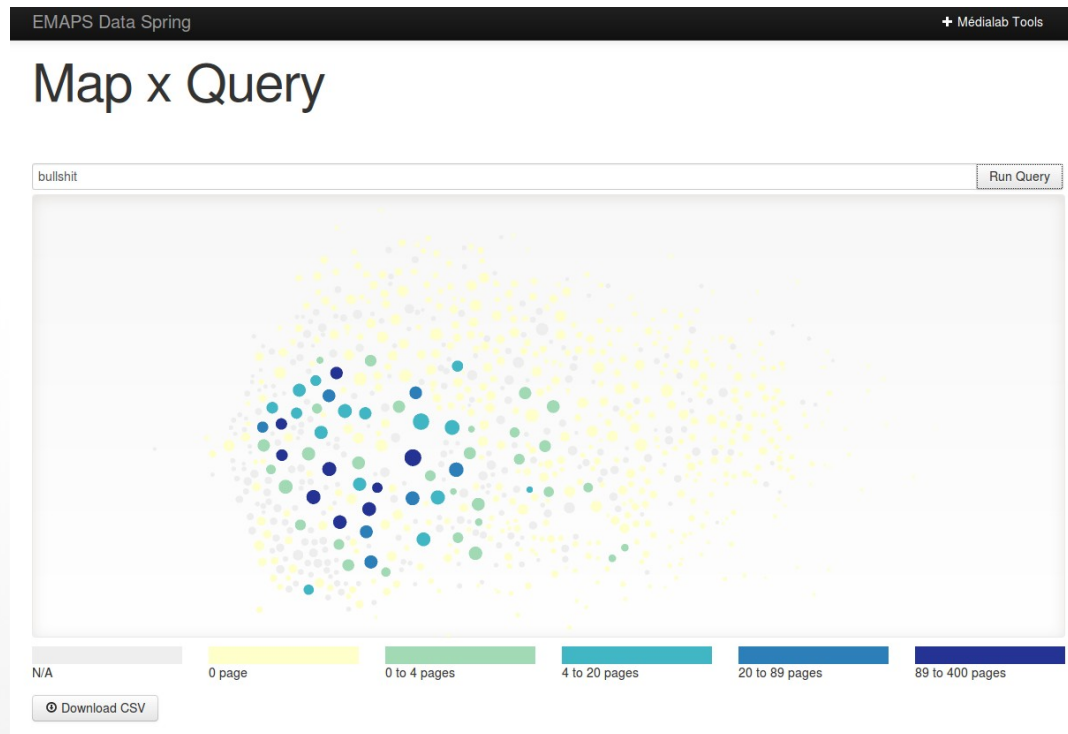
L'interface de Hyphe 2nde version

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](https://twitter.com/medialab_ScPo)

benjamin.ooghe@sciencespo.fr