L'archivage piloté par la recherche : l'exemple du médialab

Benjamin Ooghe

SciencesPo. | médialab

L'archivage piloté par la recherche

L'exemple du médialab

ENSSIB – Villeurbanne, 13 janvier 2014

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

Sommaire

1) Le médialab de SciencesPo : historique, problématiques et méthodes

2) L'analyse visuelle de réseau, nouvel outil des sciences sociales

3) L'ingénierie au médialab : conception et design d'outils

4) Equipex DIME-SHS Web : le crawler Hyphe et au-delà

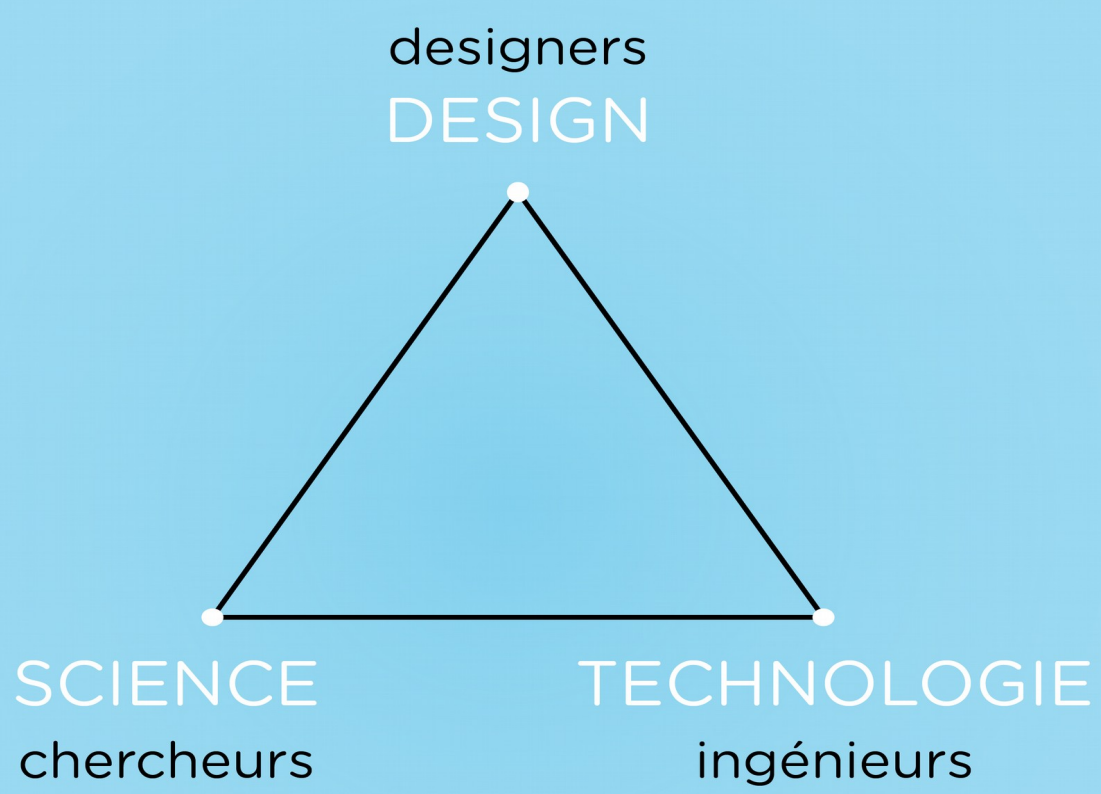
1) Historique du médialab

- Fondé en mai 2009
- Comme centre de recherche numérique au service de Sciences Po et des sciences sociales.
- Pour rassembler instruments et compétences nécessaires à la maîtrise d’une nouvelle source de données : les traces numériques
- Et dépasser la distinction entre méthodes quantitatives et méthodes qualitatives

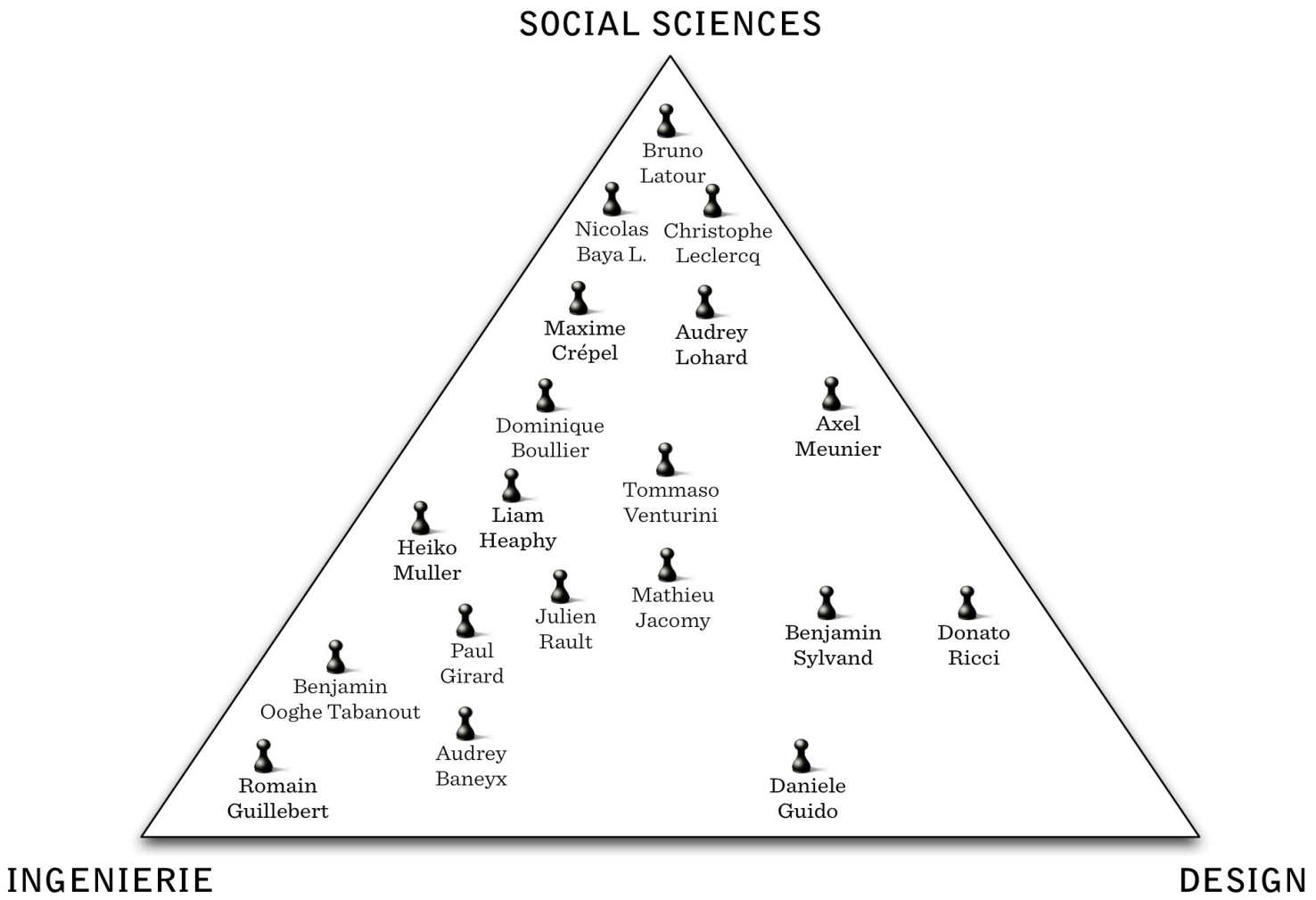
médialab : un lieu

médialab : une équipe

médialab : une combinaison de compétences



médialab : une combinaison de compétences



Notre projet : quel ancrage disciplinaire ?

SciencesPo. | médialab

Double ancrage :

Sciences Humaines et Sociales
(plus particulièrement Sociologie des Sciences)

Sciences de l'information et de la communication
(Media Studies, TIC, design de l'information)

Proche des *Humanités numériques* mais pour les sciences sociales : « Méthodes numériques en Sciences Sociales »

Articulation Quali/Quanti – micro/macro

Nombreuses personnes, données pauvres

Données riches, peu de personnes

Les médias :
d'un objet d’étude...

Photo credit – Brandon Doran via Flickr - ©

… à un support de traces

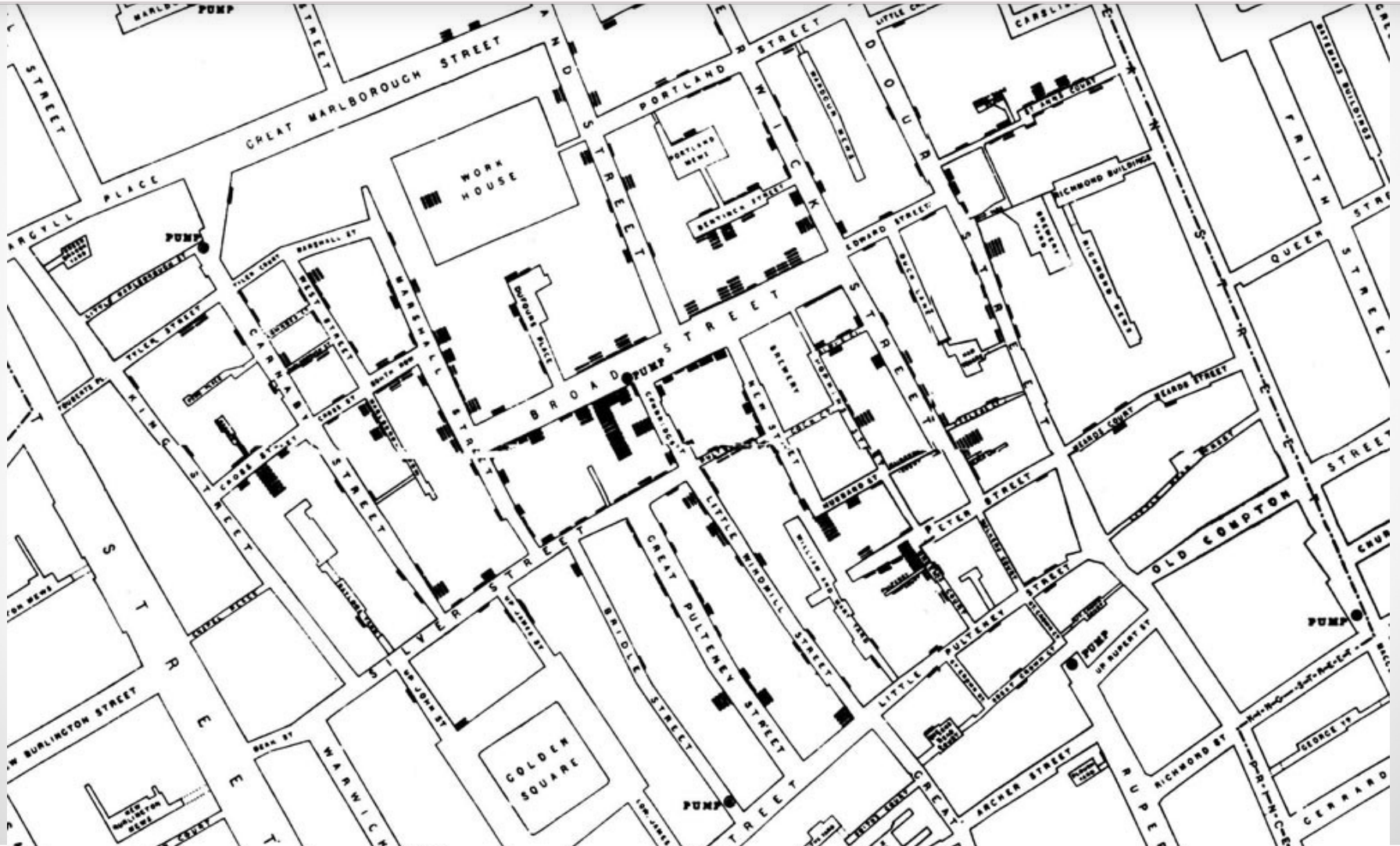
SciencesPo. | médialab

facebook

December 2010

Une discipline naissante

On the mode of
Communication of cholera
John Snow, 1855



Data deluge

The Economist
25 February 2010

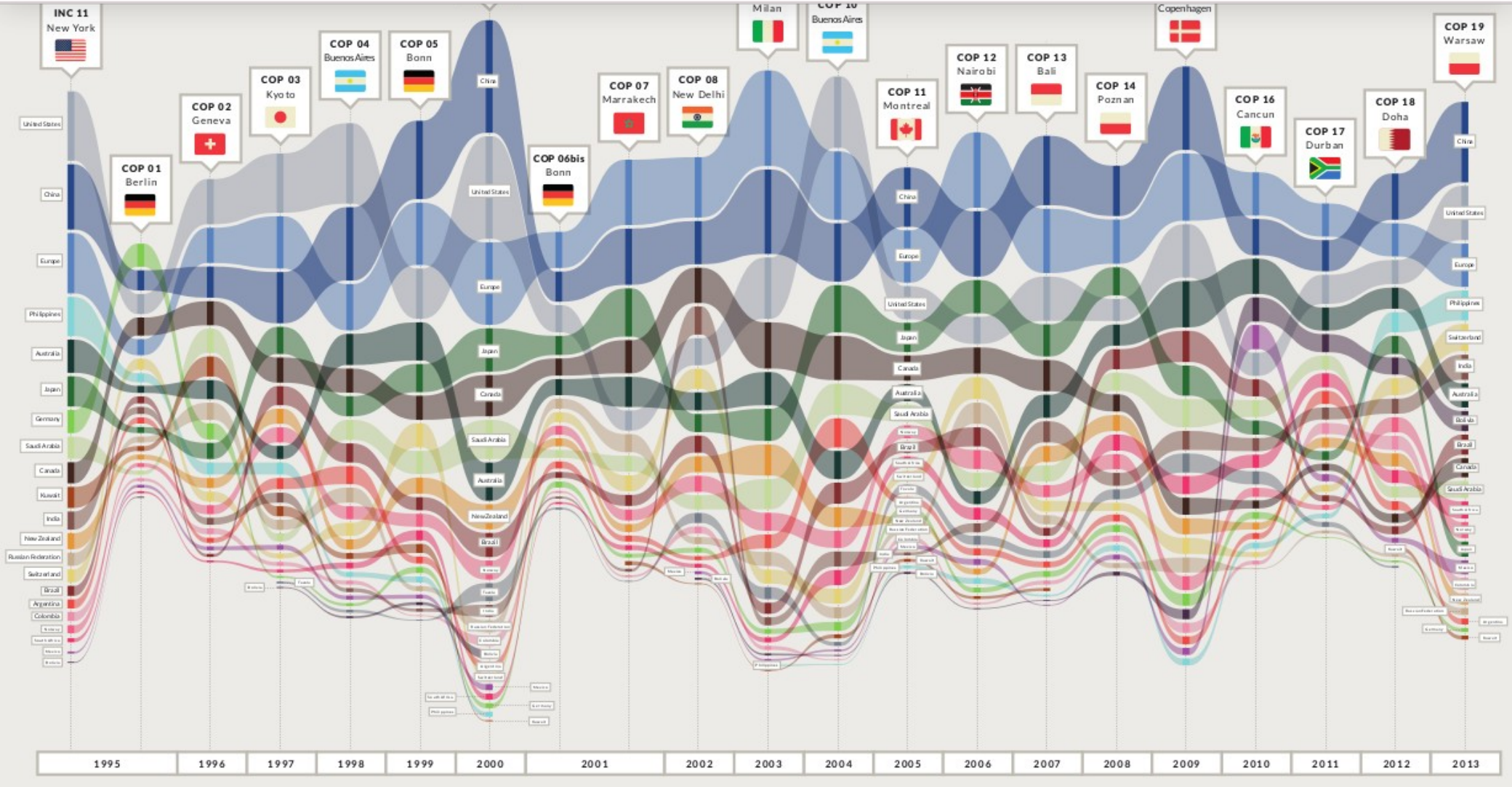
Théorie acteurs-réseaux, B. Latour

Photo credit – Alessandro Coco via bonnierskonsthall.se - CC

SciencesPo. | médialab

Design, interactivité et interprétabilité

EMAPS Sprint
Janvier 2014



L'exploitation des traces : un double retard des sciences sociales

- Par rapport aux sciences de la nature, de l’ingénieur et de l'information
- Par rapport aux secteurs du marketing et de la surveillance
- Vers des méthodes pour exploiter les traces numériques pour les sciences sociales

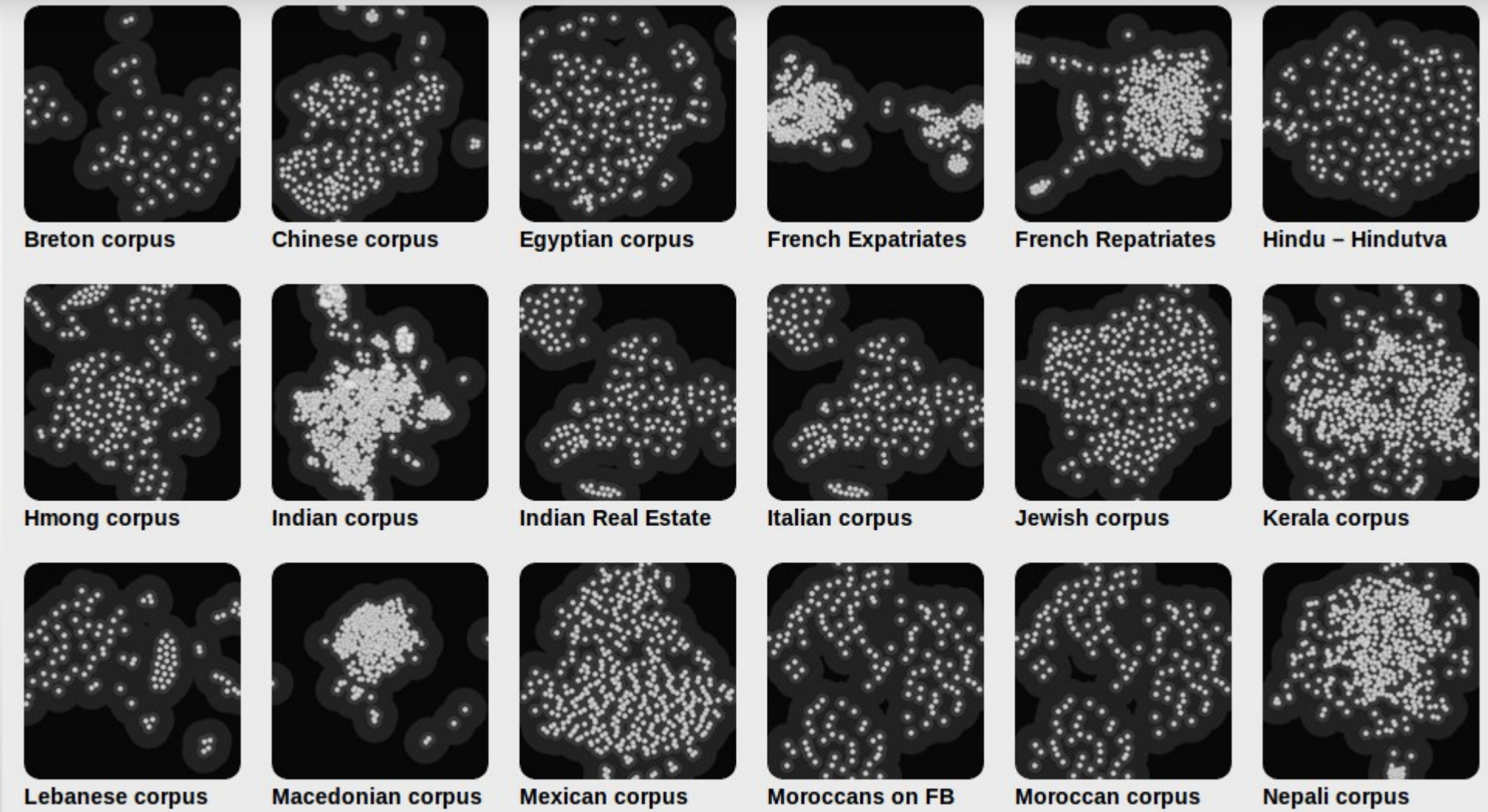
Les missions du médialab

- Recherche méthodologique et développement d’outils numériques pour les sciences sociales
- Participation aux projets de recherche en tant qu’expert méthodologique
- Sensibilisation de la communauté des sciences sociales aux méthodes numériques
- Tête de réseau connectant Sciences Po à l’innovation numérique

Méthodes numériques,
Méthodes quali-quantitatives

SciencesPo. | médialab

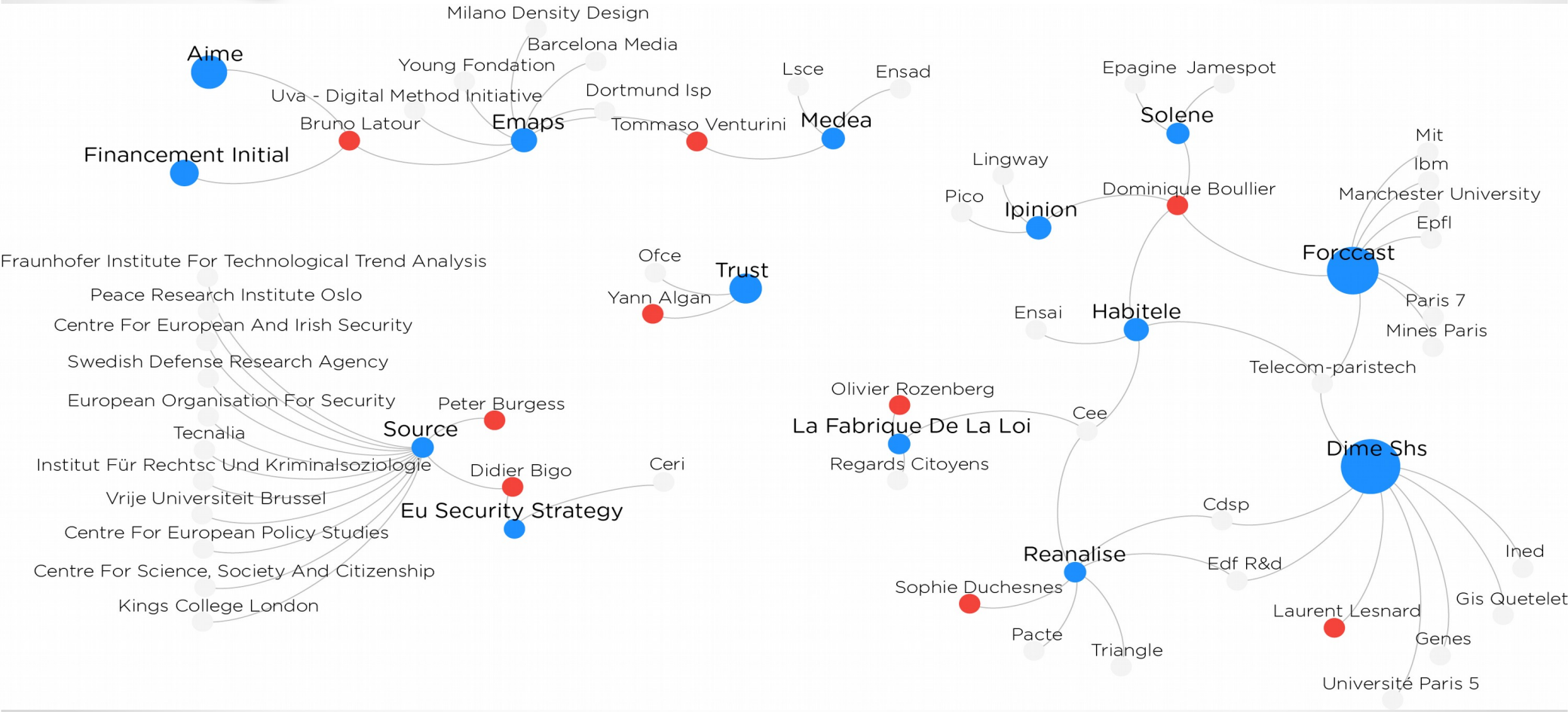
e-Diasporas Atlas
e-diasporas.fr
Dana DIminescu



Une multiplicité de projets et partenaires

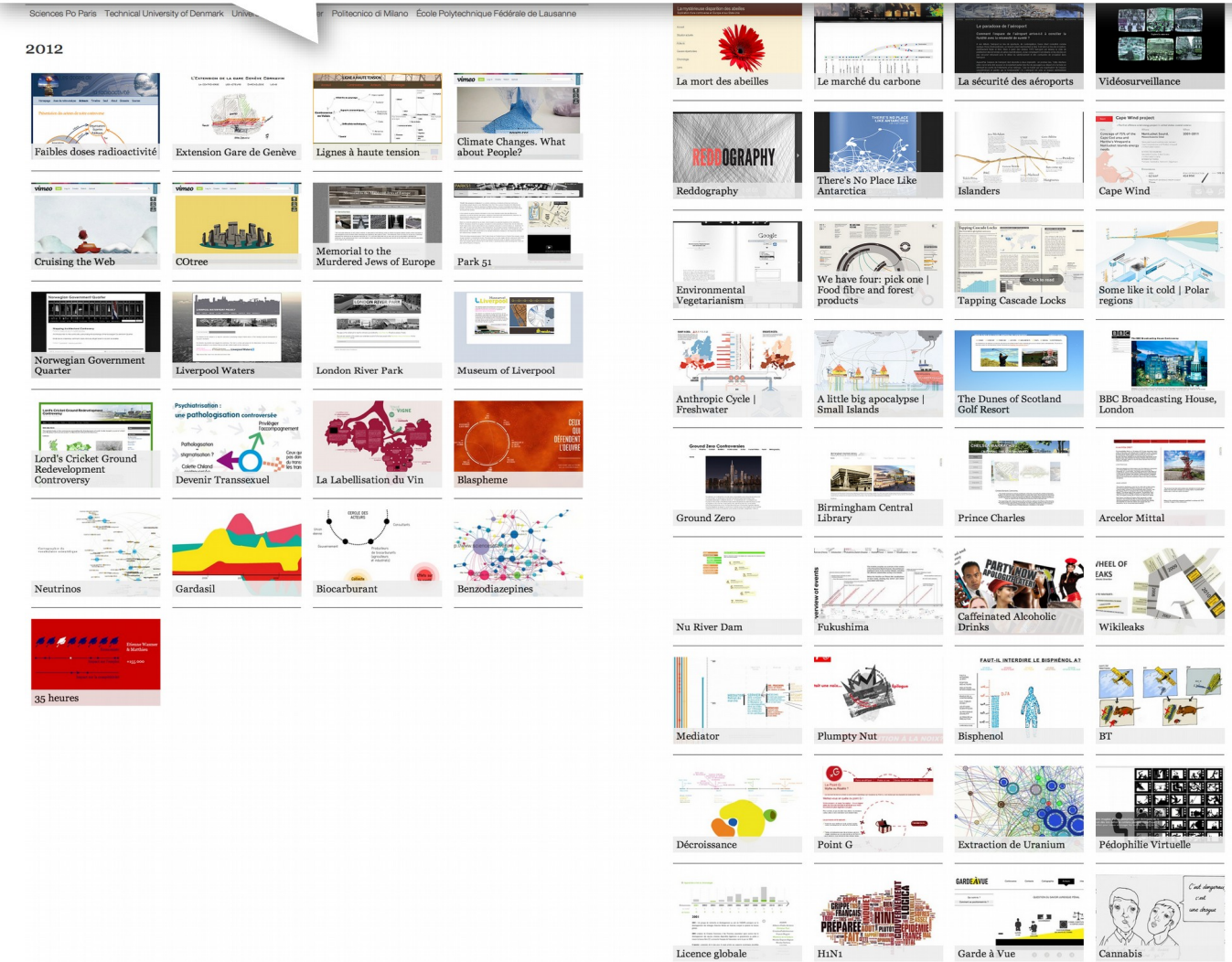
5 projets en tant que *Leader* : ERC, IDEFI, FP7, ANR, MINEFI

6 projets en tant que *Participant* : EQUIPEX, ERC, ANR (2), PICRI, MINEFI

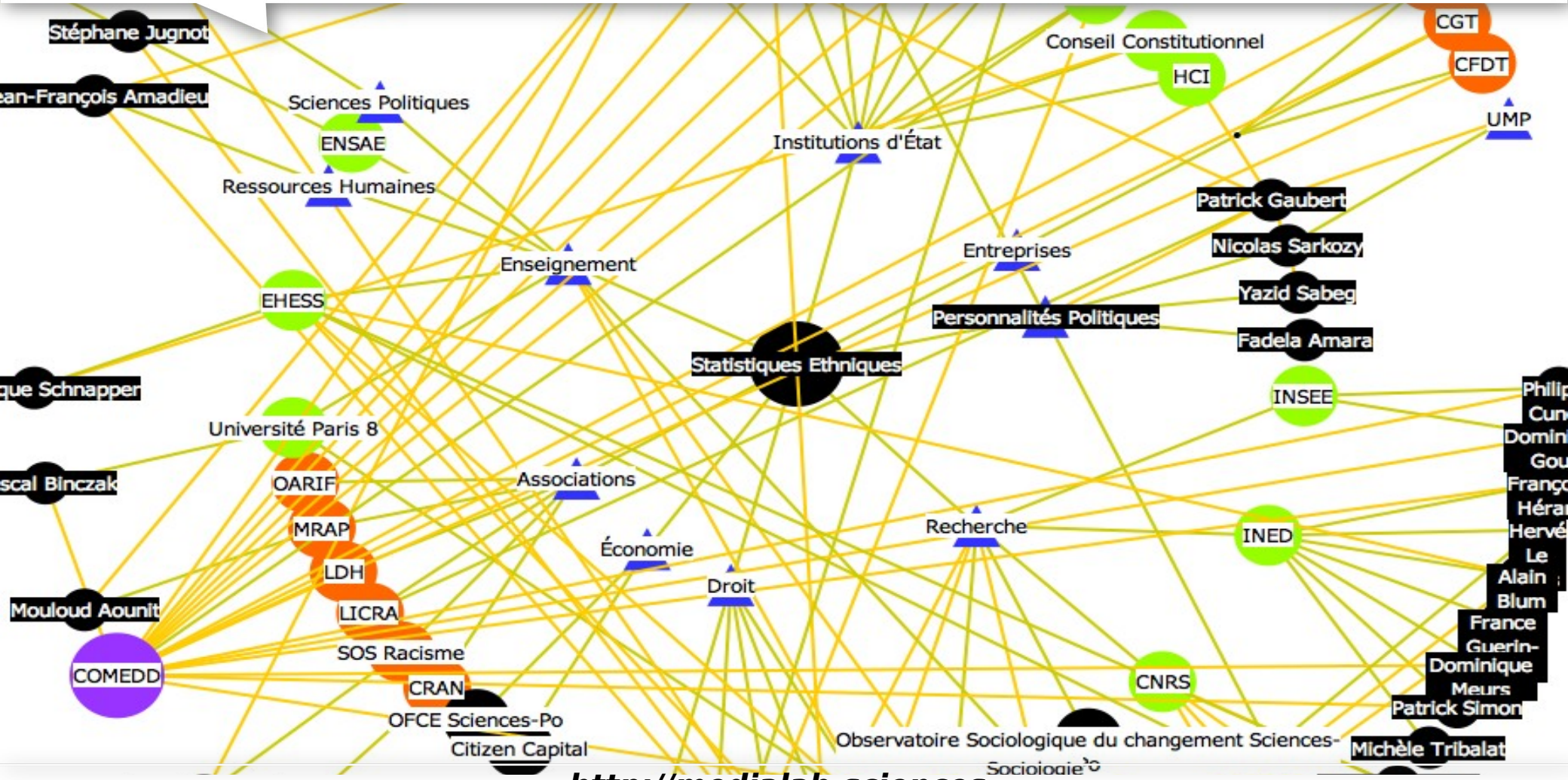


2) L'analyse visuelle de réseaux

Cartographie des controverses

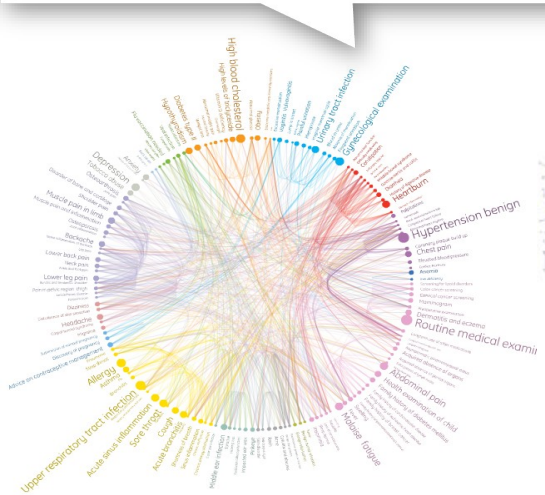


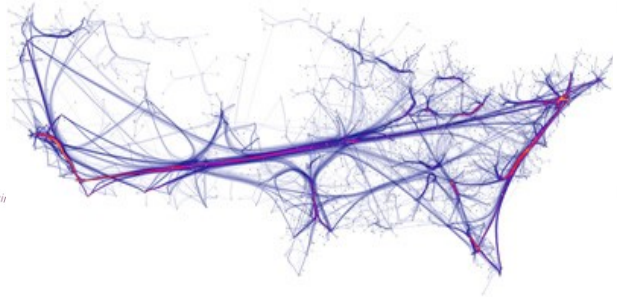
Graphes acteurs-réseaux

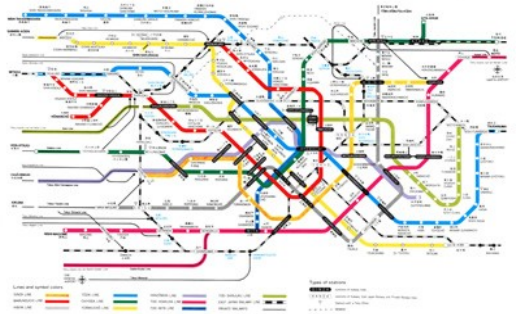


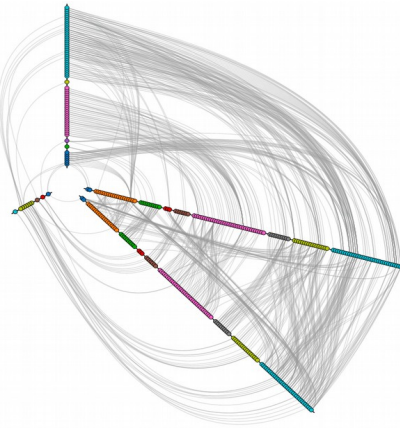
http://medialab.sciences-po.fr/controversies/2010/StatistiquesEthniques/acteurs_implication.php

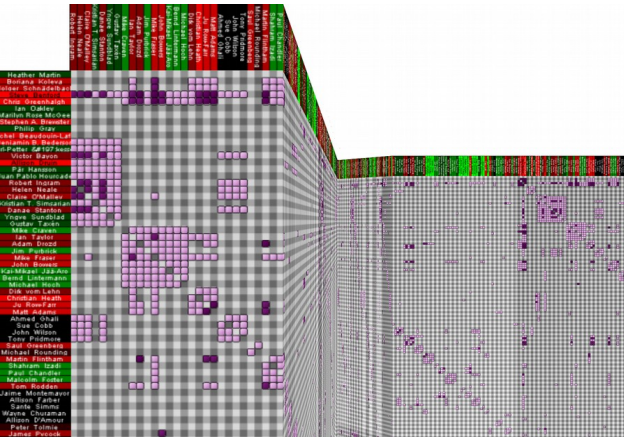
Visualiser des liens, mais comment ?

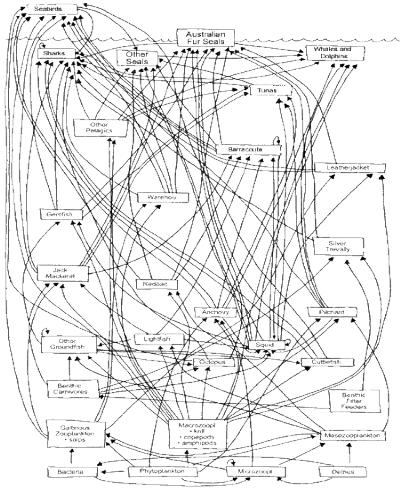


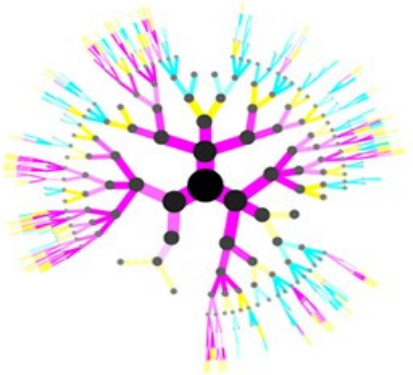












Des réseaux-graphes...

Paul Baran, 1960

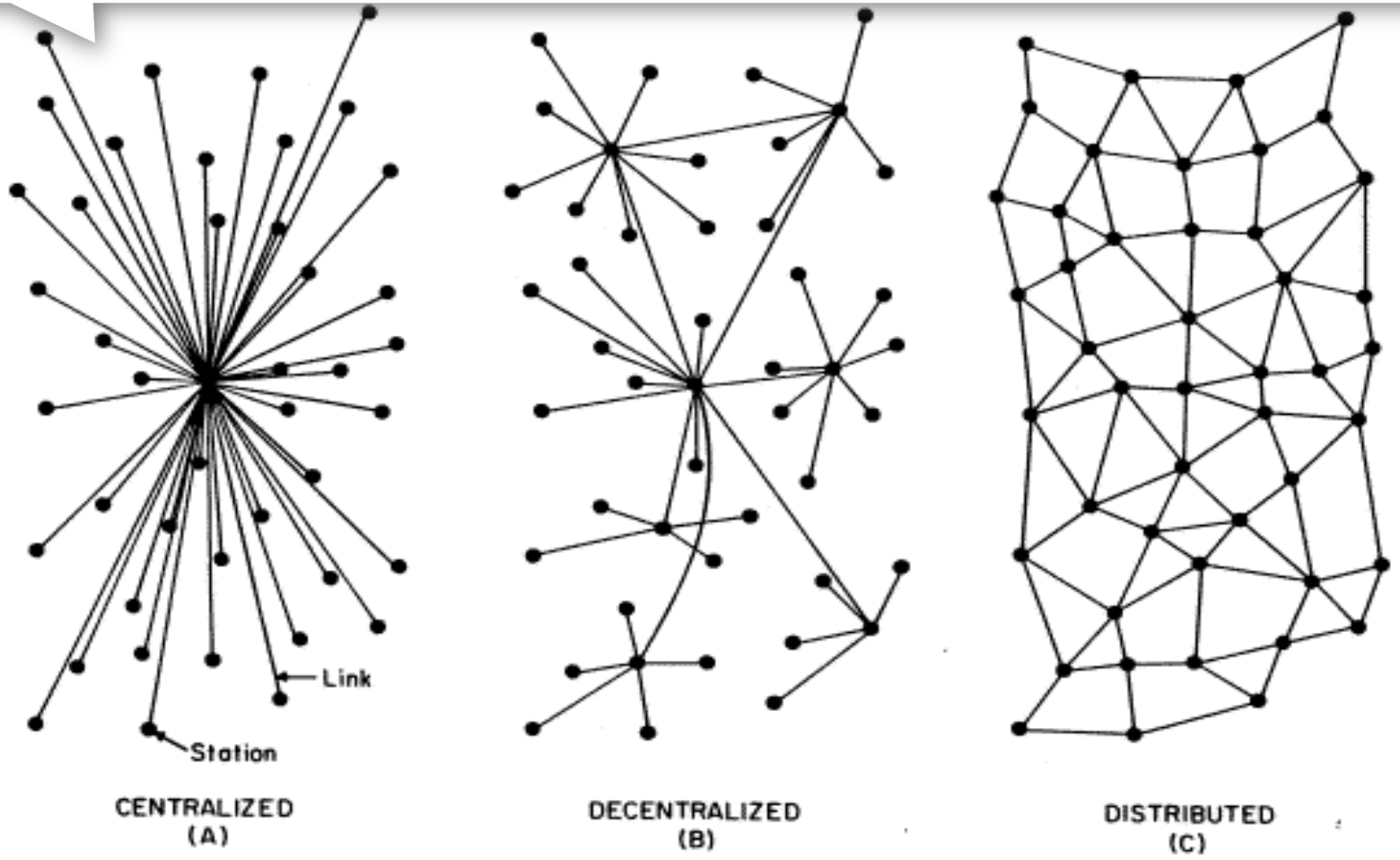


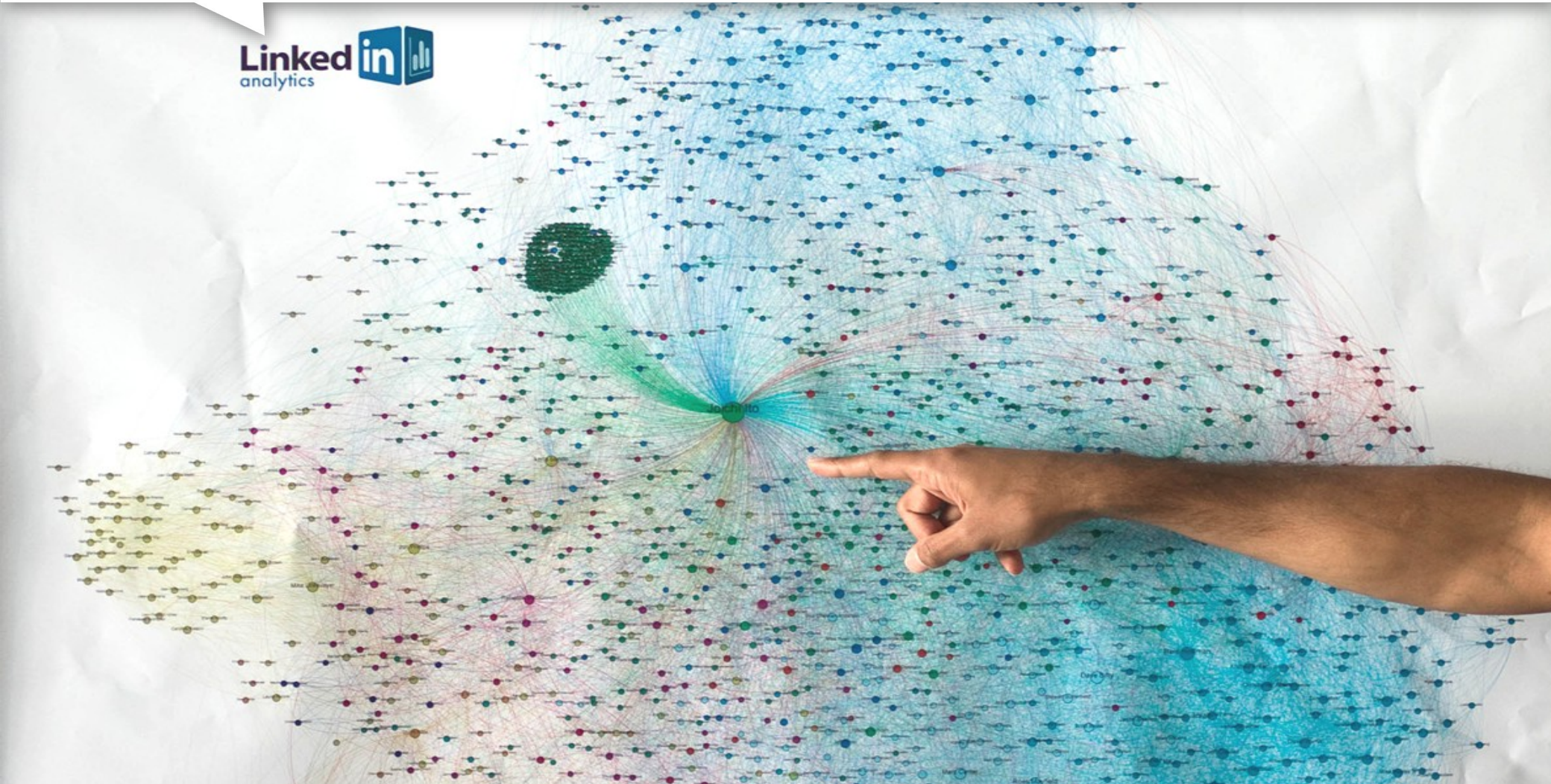
FIG. 1 — Centralized, Decentralized and Distributed Networks

...aux réseaux-cartes

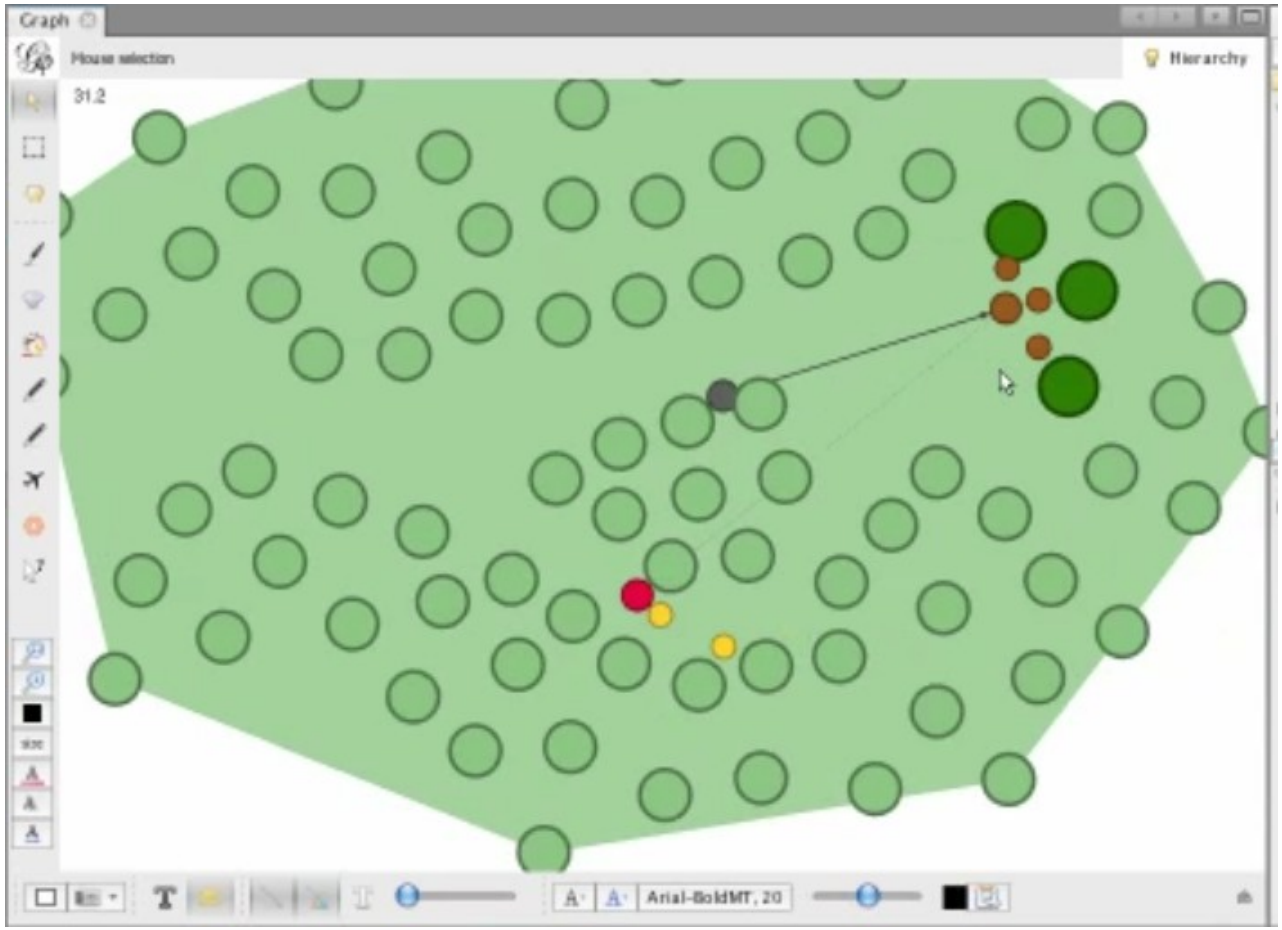
London Underground
1933 Map (Harry Beck)

homepage.ntlworld.com/clivebillson/tube/tube.html - www.fourthway.co.uk/tfl.html

Les réseaux : outils de navigation



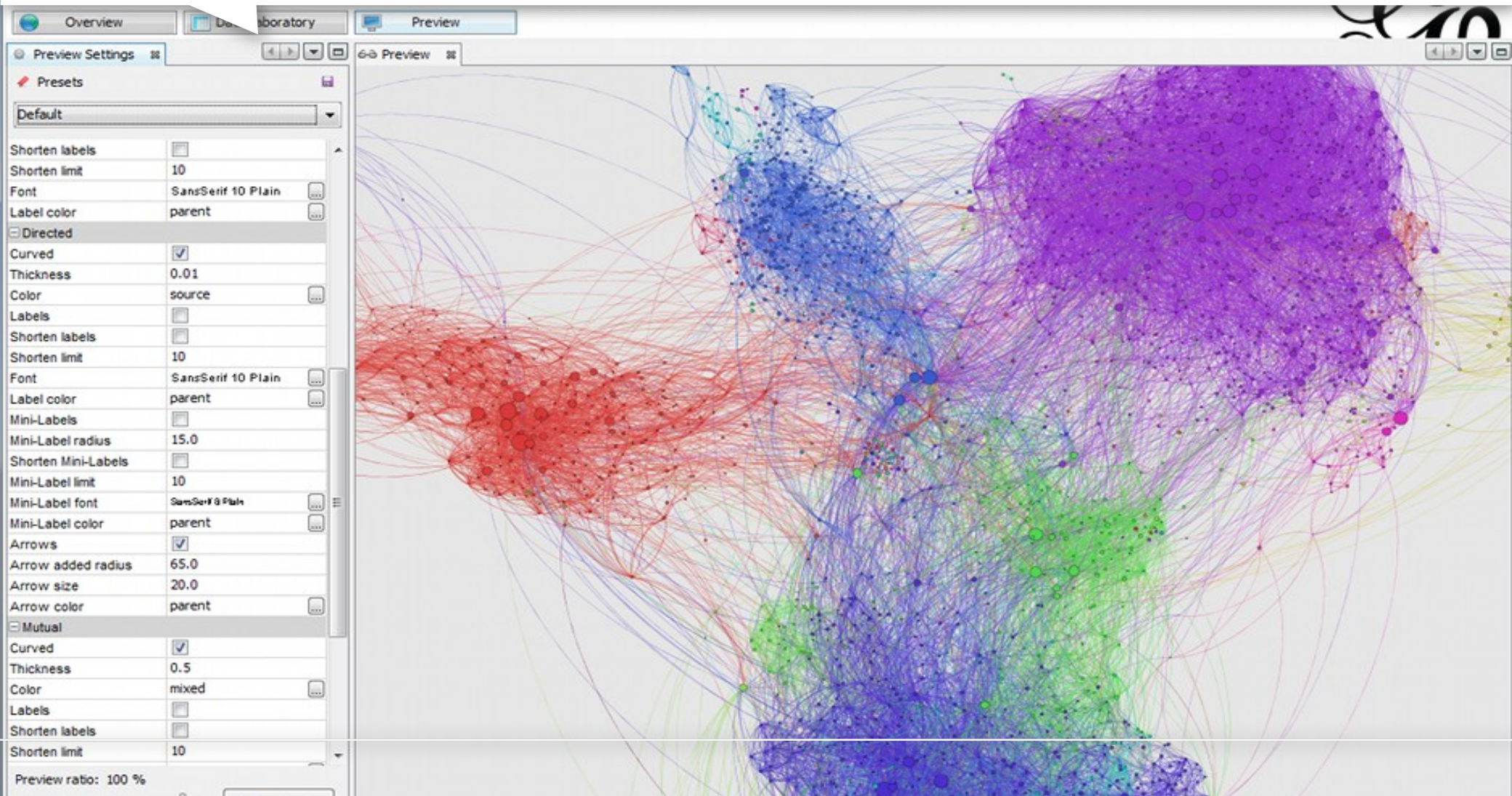
Les réseaux : outils de narration



Le petit chaperon rouge

Les réseaux comme interface

Gephi
gephi.org

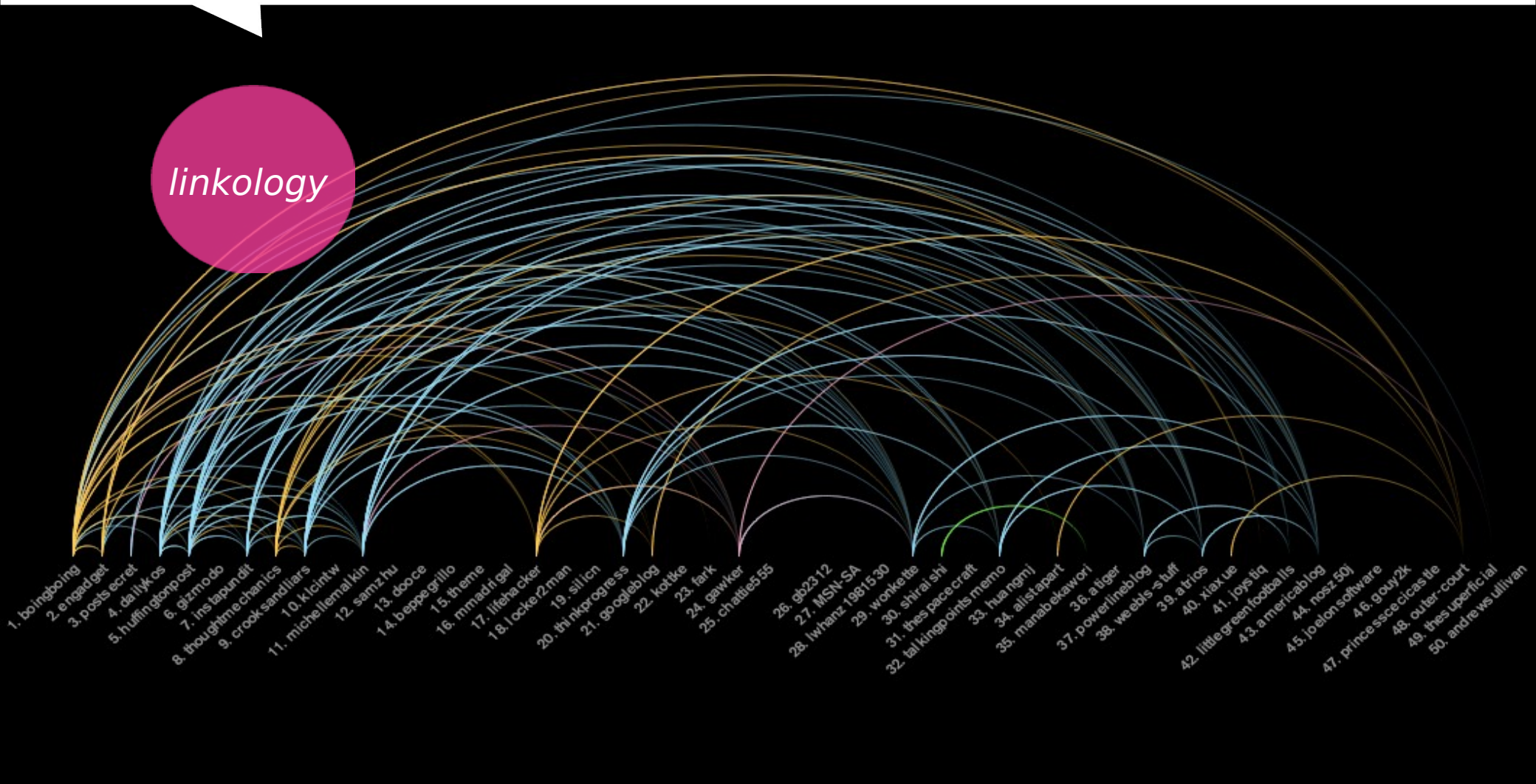


SciencesPo. | médialab

Cartographier le web...

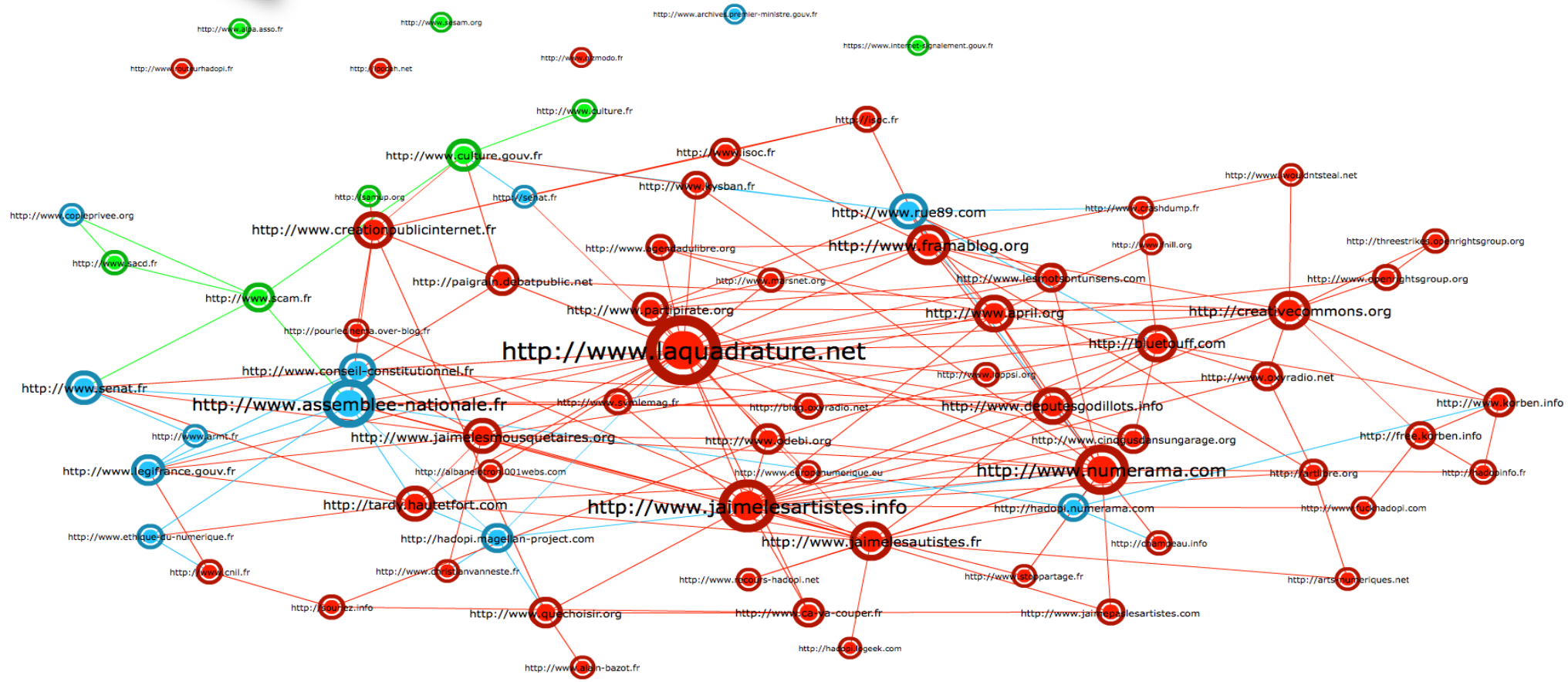
Top 50 US blogs
Ben Fry, 2006

http://nymag.com/news/media/15972/



SciencesPo. | médialab

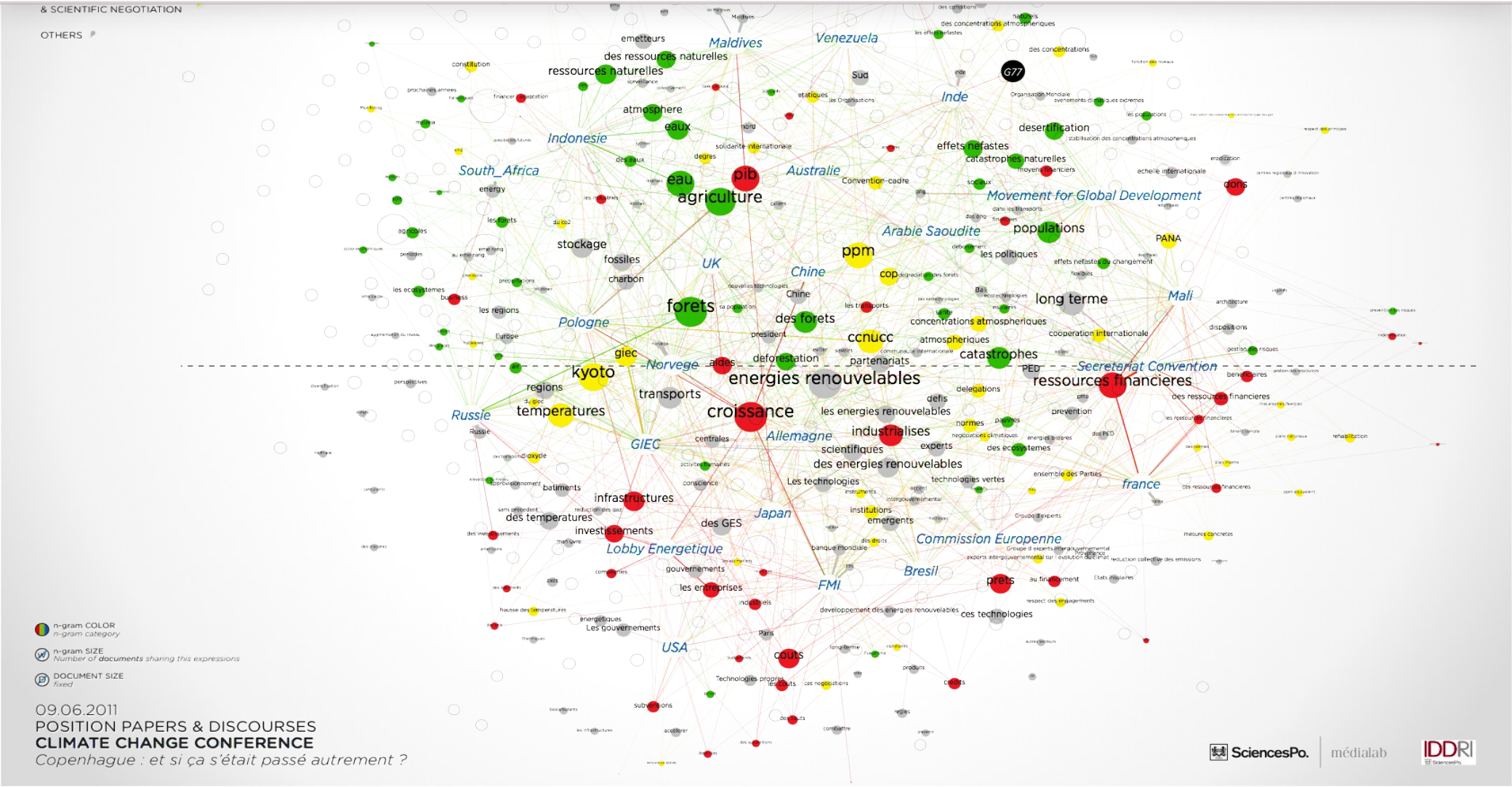
Cartographie des liens hypertextes



http://www.medialab.sciences-po.fr/controversies/2010/Hadopi2/index.php?cat=ondaweb&subcat=carto

SciencesPo. | médialab

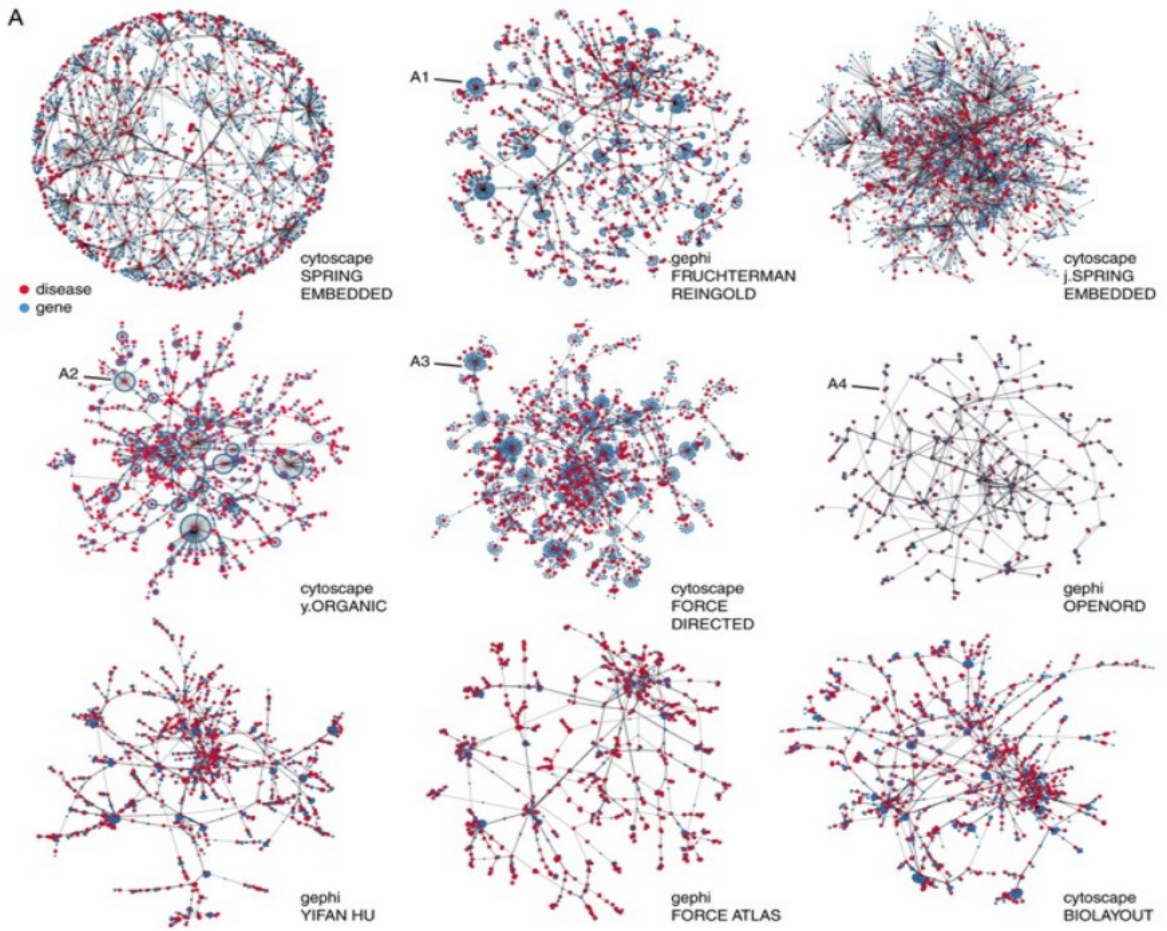
Cartographie des liens sémantiques



SciencesPo. | médialab | IDDRI

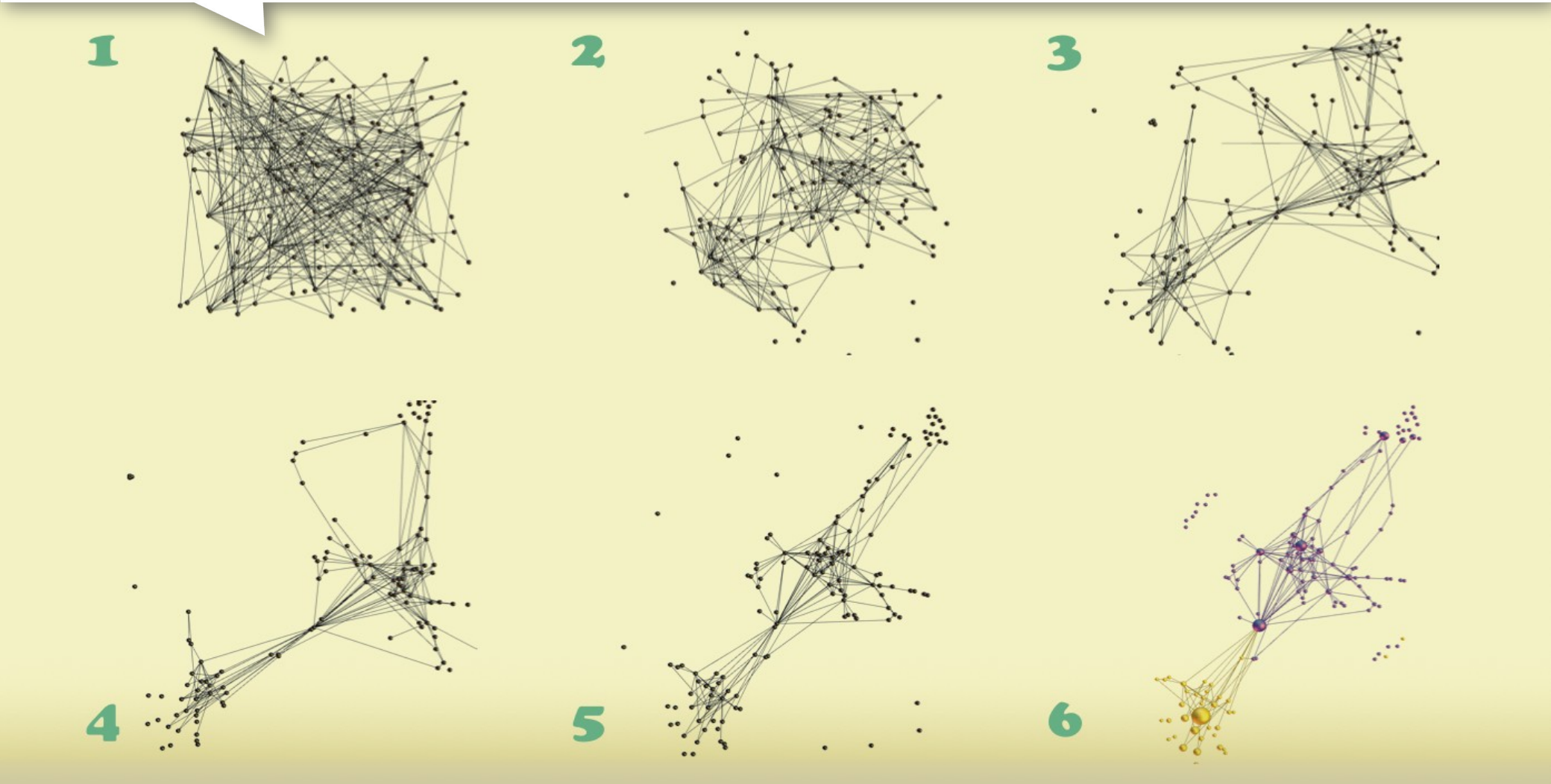
Quel type de spatialisation

Tukey, J. W. (1977)
Exploratory Data Analysis



Algorithmes Force-vecteur

Algorithmes Force-vecteur ForceAtlas2

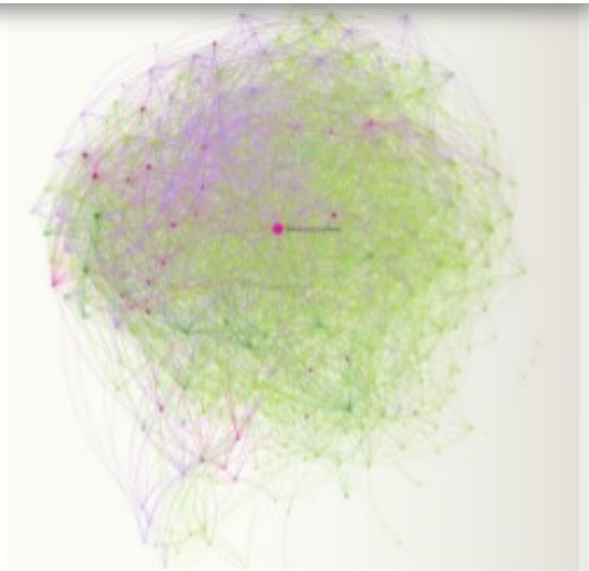


Analyse visuelle et clusterisation

Analyse visuelle et clusterisation

Hubs et Autorités :

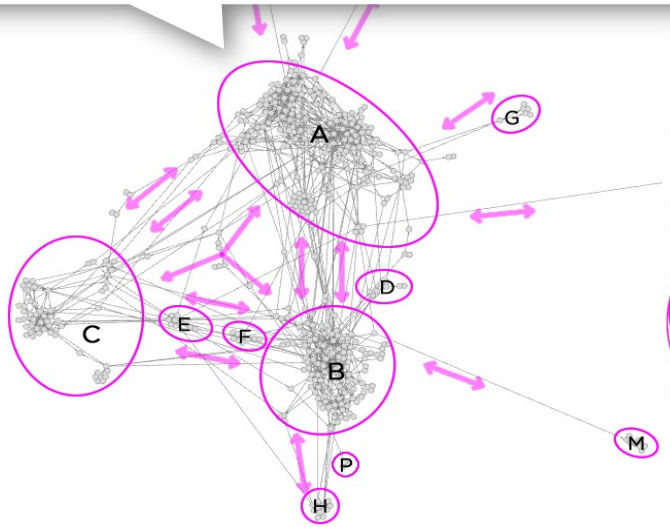
Raccourcis mais aussi carrefours
Ils causent des «boules de cheveux»

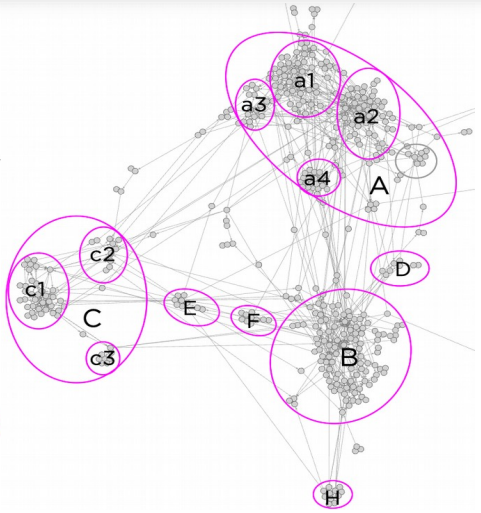


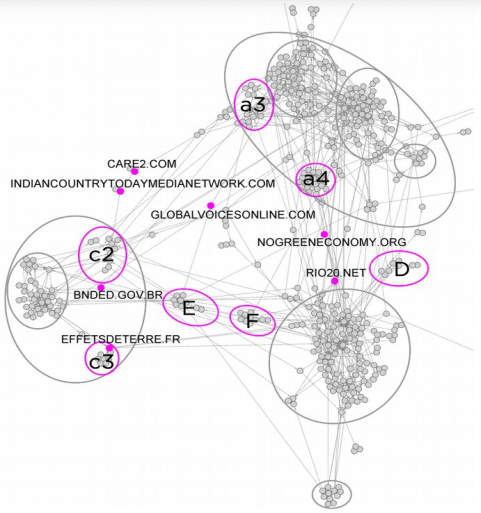
Les clusters contiennent d'autres clusters :

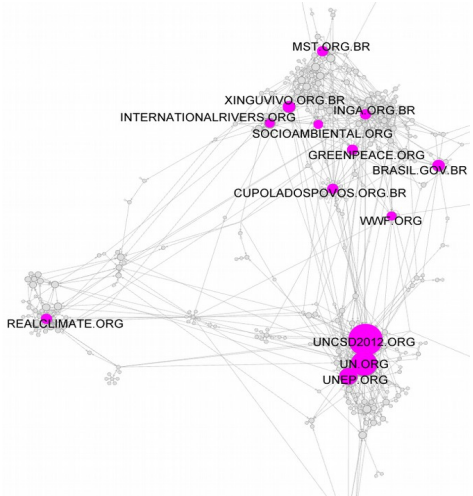
Quels sont les clusters pertinents ?
Où sont les bonnes frontières ?

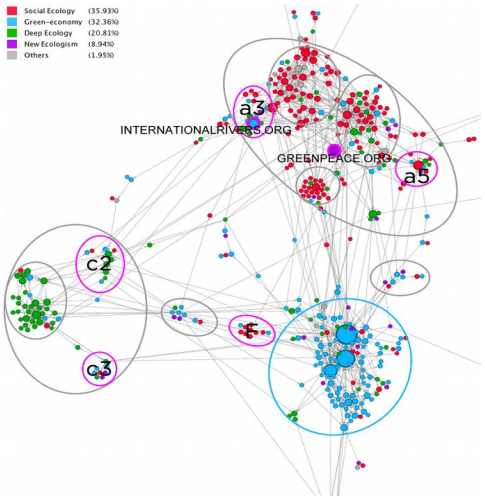
Analyse visuelle et clusterisation

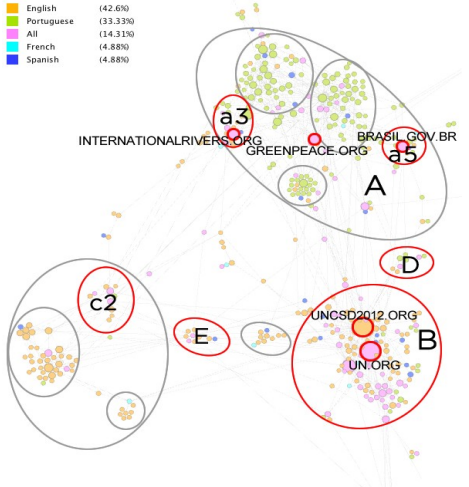












3) Conception et design d'outils

médialab Tools

Tools we develop, and tools we use

I Want Hue

Colors for data scientists. Generate and refine palettes of optimally distinct colors.

Source Documentation > Use it !

Hyphe

Hyphe is a webcrawler made to help scientists

I Want Hue by Mathieu Jacomy

Source Documentation > Use it !

iWantHue

Colors for data scientists. *Generate and refine palettes of optimally distinct colors.*

iWantHue allows you to generate palettes of colors. It is about mastering the properties of a palette by setting a range of Hue, Chroma (unbiased saturation) and Lightness. You can generate palettes of any size or just get the generator for a javascript project. The algorithm optimizes the perceptive distance in the color subspace, ensuring an optimal readability.

How it works

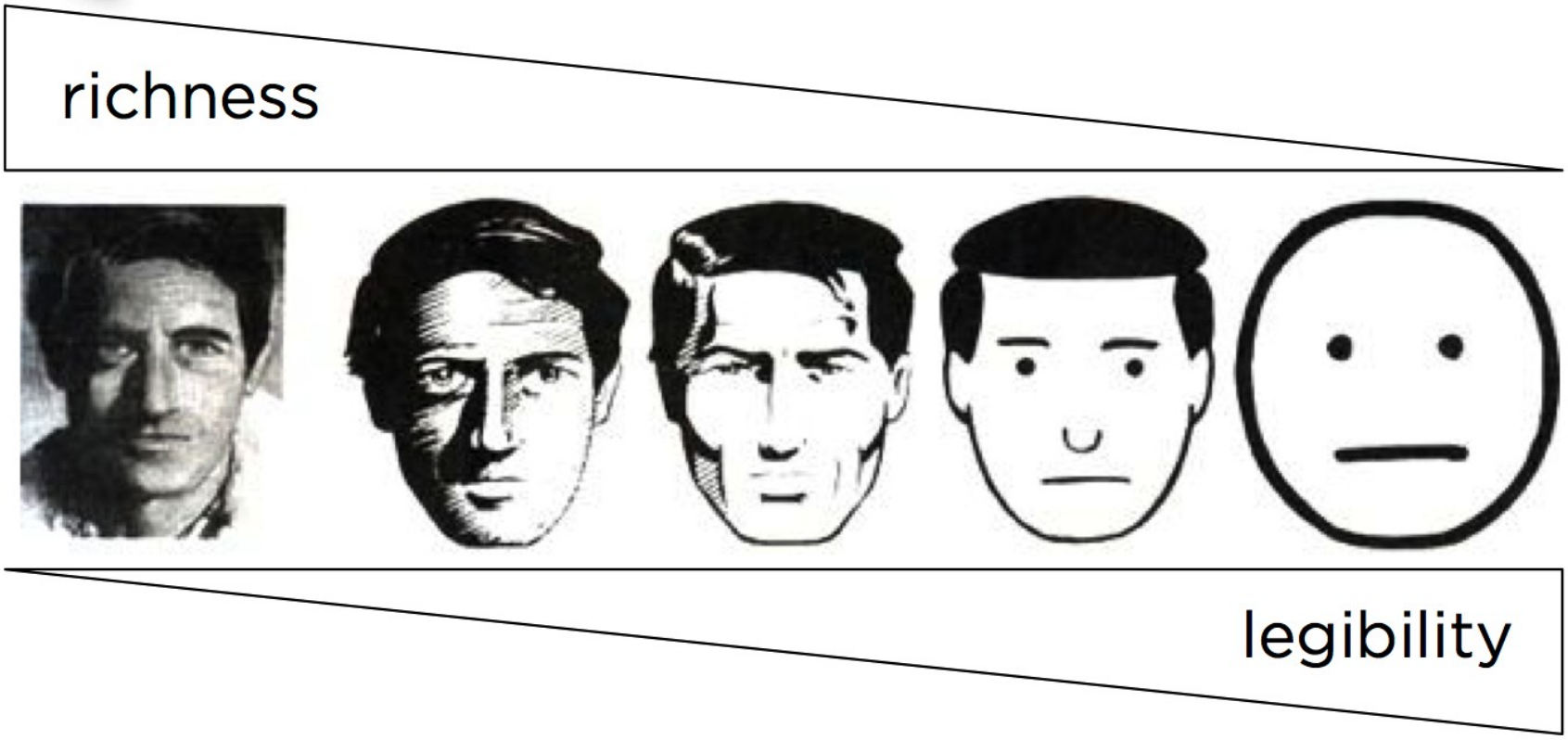
1. **K-means** or **force vector repulsion** algorithms ensure an even distribution of colors
2. The **CIE Lab** color space is used for computation, since it fits human perception
3. The **Hue/Chroma/Lightness** color space is used to set constraints, since it is user-friendly

Examples and a tutorial

Idea

Appréhender la donnée

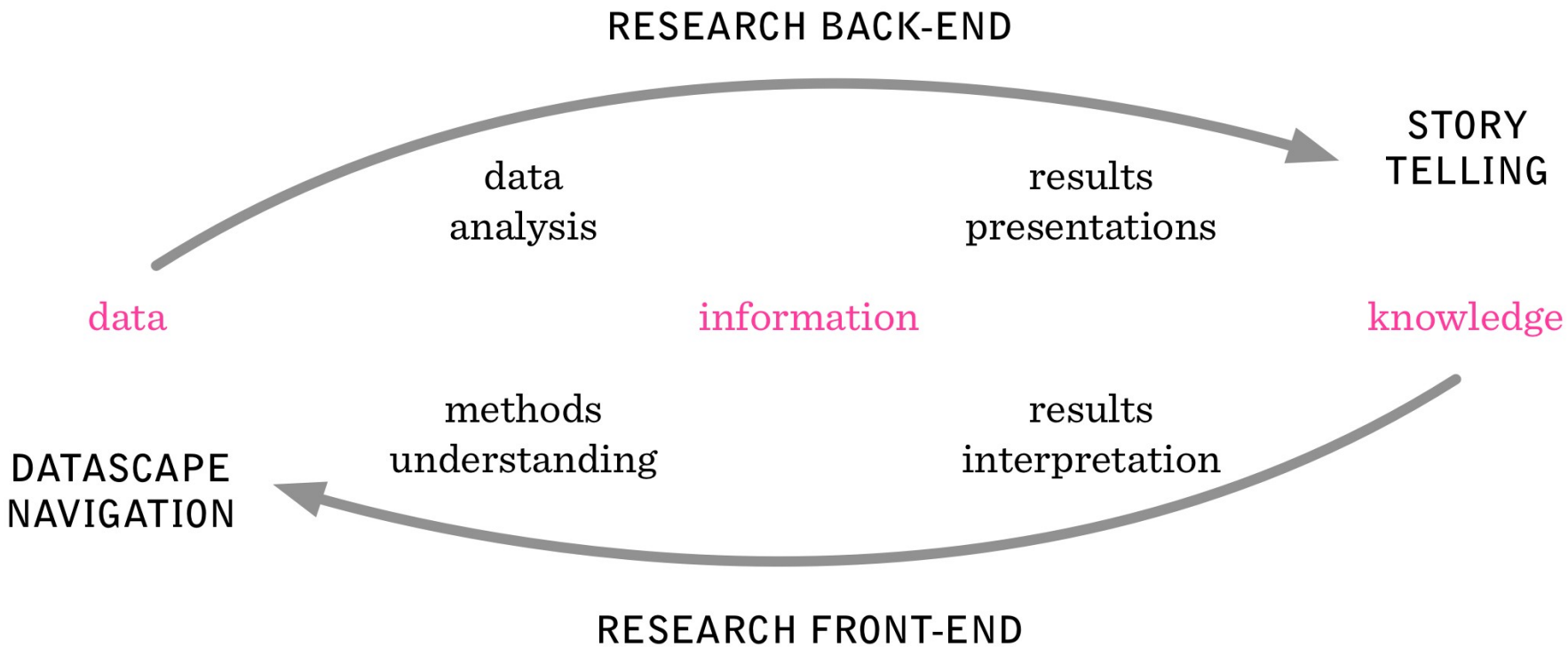
Understanding Comics
Scott McCloud (1993)



knowledge ⟶ information ⟵ data

Le cycle d'exploration – narration

“An essential property of this chain is that it must remain reversible. The succession of stages must be traceable, allowing to travel in both directions” (Latour, 1995)



Le protocole attendu

médialab
(Sciences Po Paris)

Institute of Spatial Planning
Dortmund University

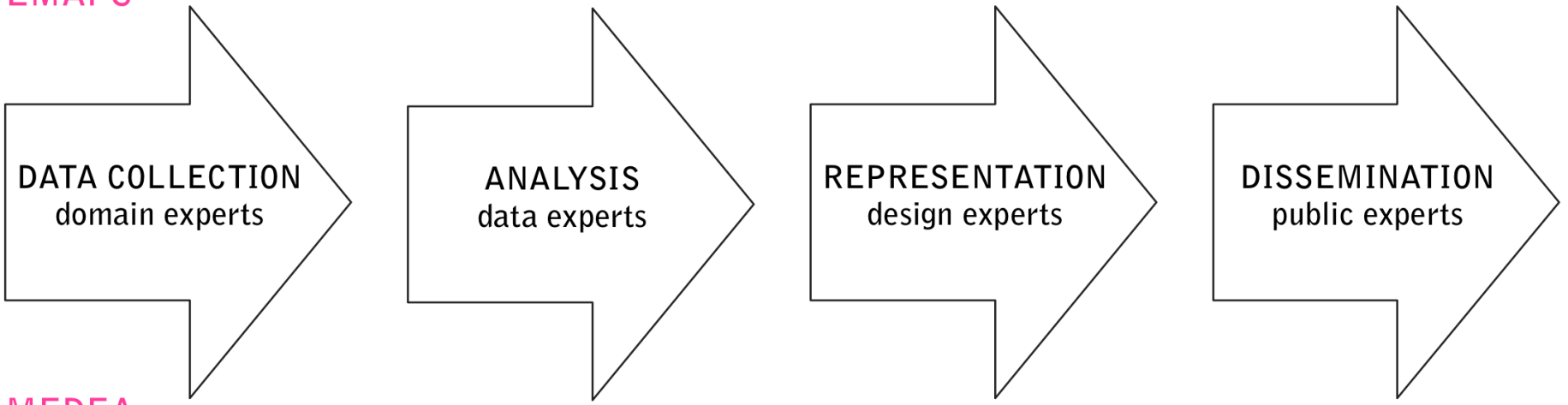
Digital Methods Initiative
(Amsterdam University)

Barcelona Media
(Barcelona)

Density Design
(Milan Politecnico)

Young Foundation
(London)

EMAPS



MEDEA

Laboratoire des Sciences
du Climat et Environnement
(Paris)

médialab
(Sciences Po Paris)

ENSADlab
Ecole National Supérieure
d'Arts Decoratifs (Paris)

La spirale d'interaction

1. HYPOTHESIS

alpha users suggest relevant questions ▶ *data experts* suggest feasible answers

2. SKETCHING

data experts propose operationalisation ▶ *design experts* sketch mockups ▶ *alpha users* evaluate mockups

3. DATA COLLECTION

alpha users provide existing data ▶ *data experts* make old data compatible
alpha users suggest new data sources ▶ *data experts* extract new data

4. ANALYSIS

data experts explore data ◀ ▶ *design experts* visualize data
data experts propose interpretation ▶ *design experts* render intepretation

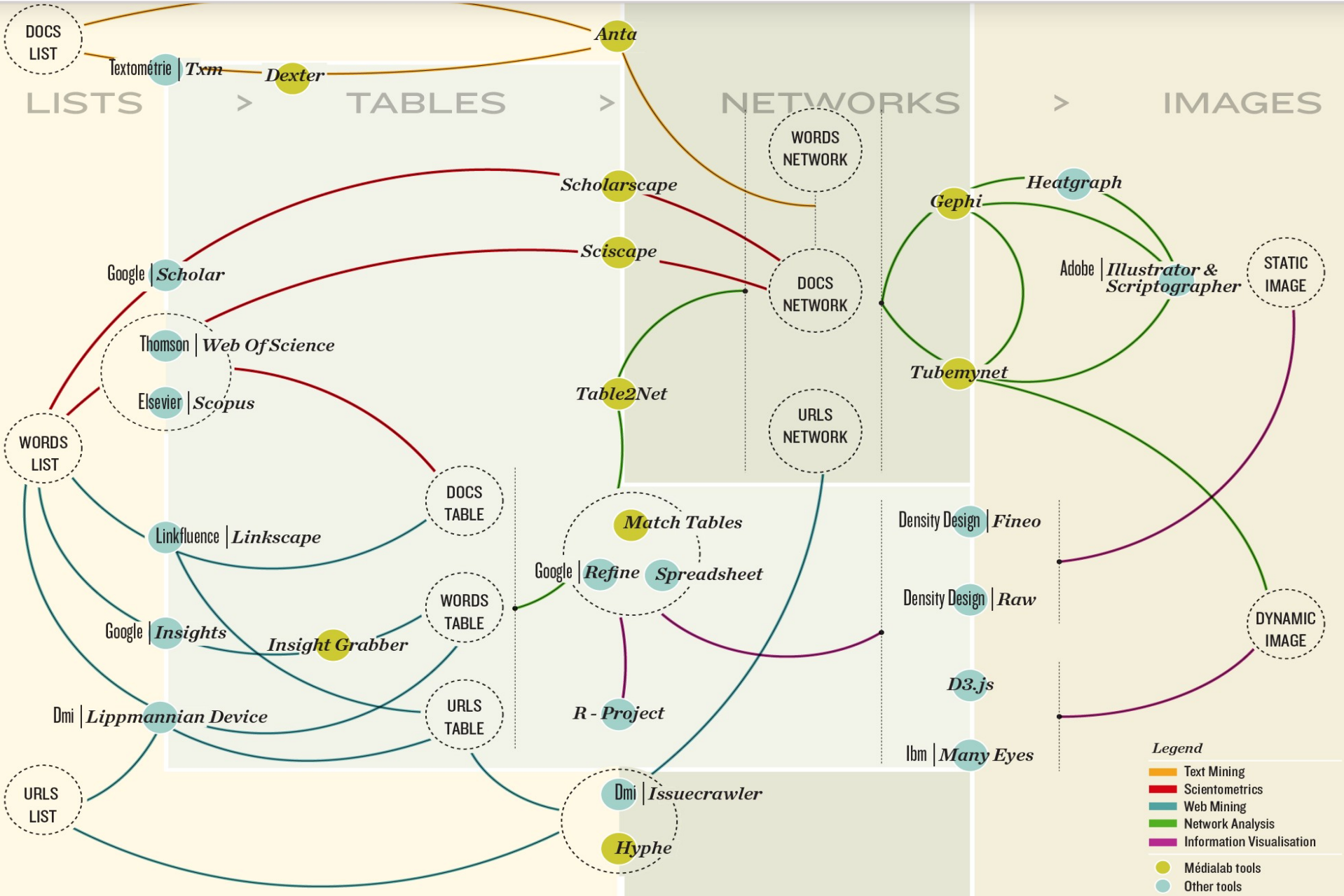
5. PUBLICATION

design experts publish maps ▶ *beta users* react to maps

Une chaîne d'outils

SciencesPo. | médialab

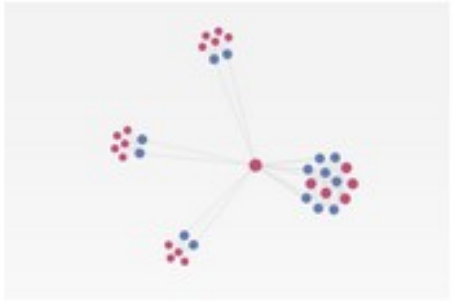
http://tools.medialab.sciences-po.fr



ScienceScape : la scientométrie simplifiée

ScienceScape

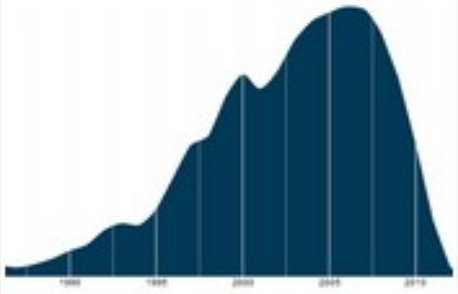
Helpers for scientometrics. Convert files, get networks, visualize stuff from Scopus or Web of Knowledge.



Get Networks

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

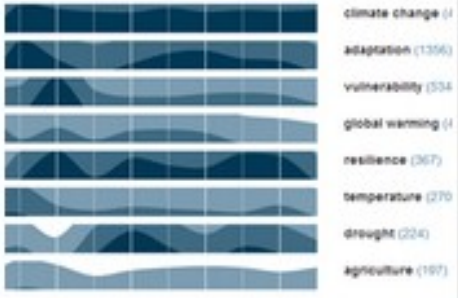
Scopus | Web of Knowledge



Papers over time

Visualize how many papers are published each year in your file

Scopus | Web of Knowledge



Keywords over time

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

Scopus | Web of Knowledge

Top keywords / year

Visualize the most used keywords each year in your file

Scopus | Web of Knowledge

I Want Hue : du choix des couleurs

i want hue

Colors for data scientists. Generate and refine palettes of optimally distinct colors.

Color space

Presets...

H 0 360

C 0.97 3

L 0 1.5

Dark background

Palette

6 colors | soft (k-Means)

Reroll palette

Fit to color space

Table2Net : du CSV au réseau

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. 16 columns and 347 rows.

Table preview

Row number	id	legislature	texteloi_id	numero	sujet	sort	date	parlementaires	texte	expose	signataires	source
1	12677	14	707	1	APRÈS ART. 17	Rejeté	2013-02-12	lionnel-luca|thierry-mariani|nicolas-dhuicq|patrice-verchere|michel-terrot|jean-pierre-decool|thierry-lazaro|damien-	I. – Les établissements de crédit garantissent le droit au crédit à toute personne résidant sur le territoire français de façon régulière et	Cet amendement a pour but de permettre l'instauration d'un droit au crédit opposable.	M. Luca, M. Mariani, M. Dhuicq, M. Verchère, M. Terrot, M. Decool, M. Lazaro, M. Abad,	http://www.assemblee-nationale.fr/14/amend

4) Equipex DIME-SHS

Équipement numérique pour recherche en SHS

- **QUALI** :
  enquêtes panel représentatif INSEE sur tablettes
- **QUANTI :**
  numérisation d'enquêtes
- **WEB :**
  accompagnement méthodo et collecte de données

DIME-SHS Web

un équipement :

- architecture technique : serveurs de calcul et de stockage
- outils et méthodes : Hypertext Corpus Initiative

=> un crawler orienté recherche : Hyphe
=> développement spécifique ponctuel

des services associés :

- accompagnement méthodologique et formation : création, exploration et analyse de corpus web
- hébergement et archivage de corpus

deux personnes

Accompagnement à la recherche

2 Appels à Projets par an

Exemples :

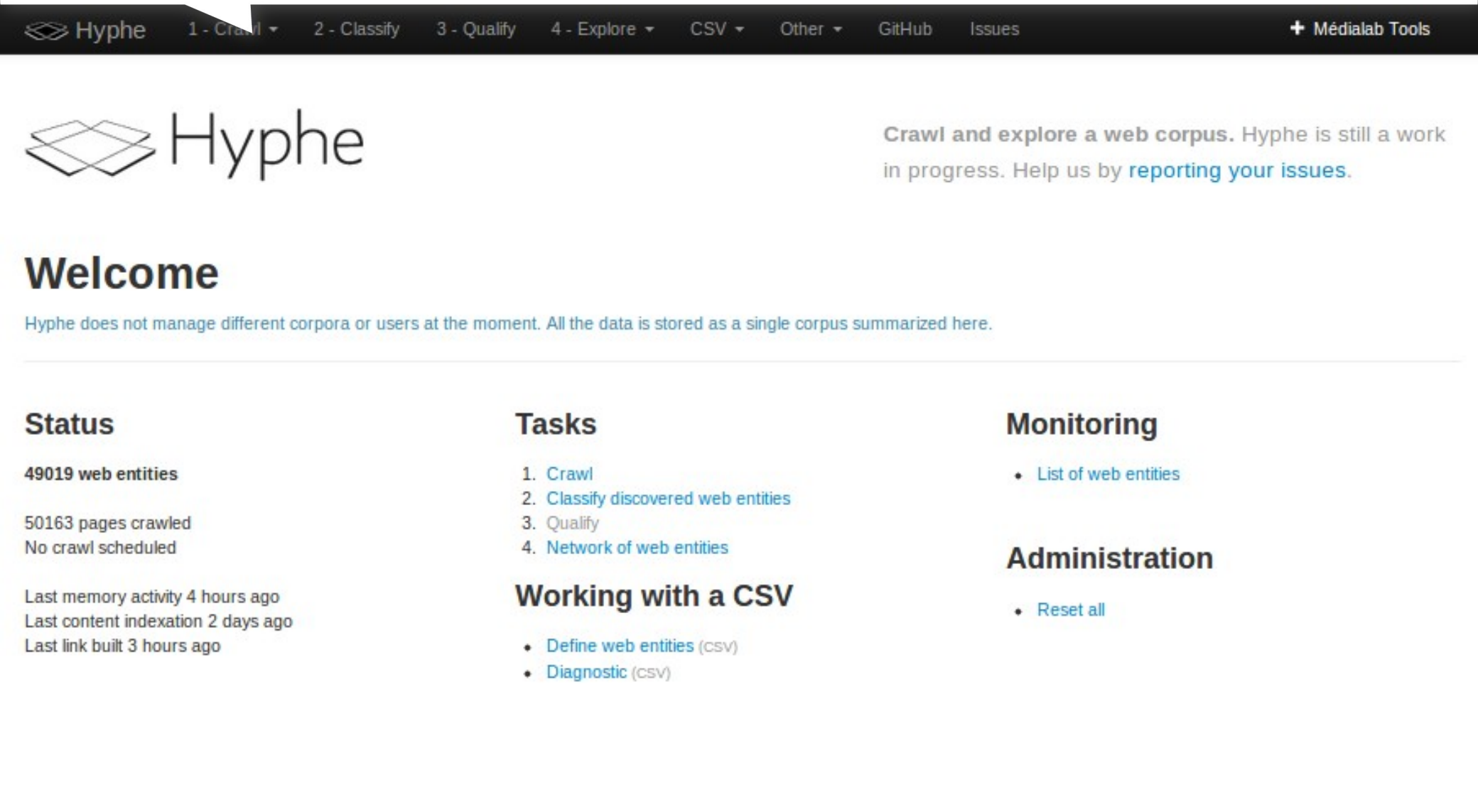
- OpenMarriage (scraping commentaires LeMonde.fr, …)
- SitPol (cartographie
- L'amour est dans le pré (forum, tweets, …)
- …

Divers :

- Cartographie du web de Sciences-Po
- Création de métrique à partir de tweets (Y. Algan)

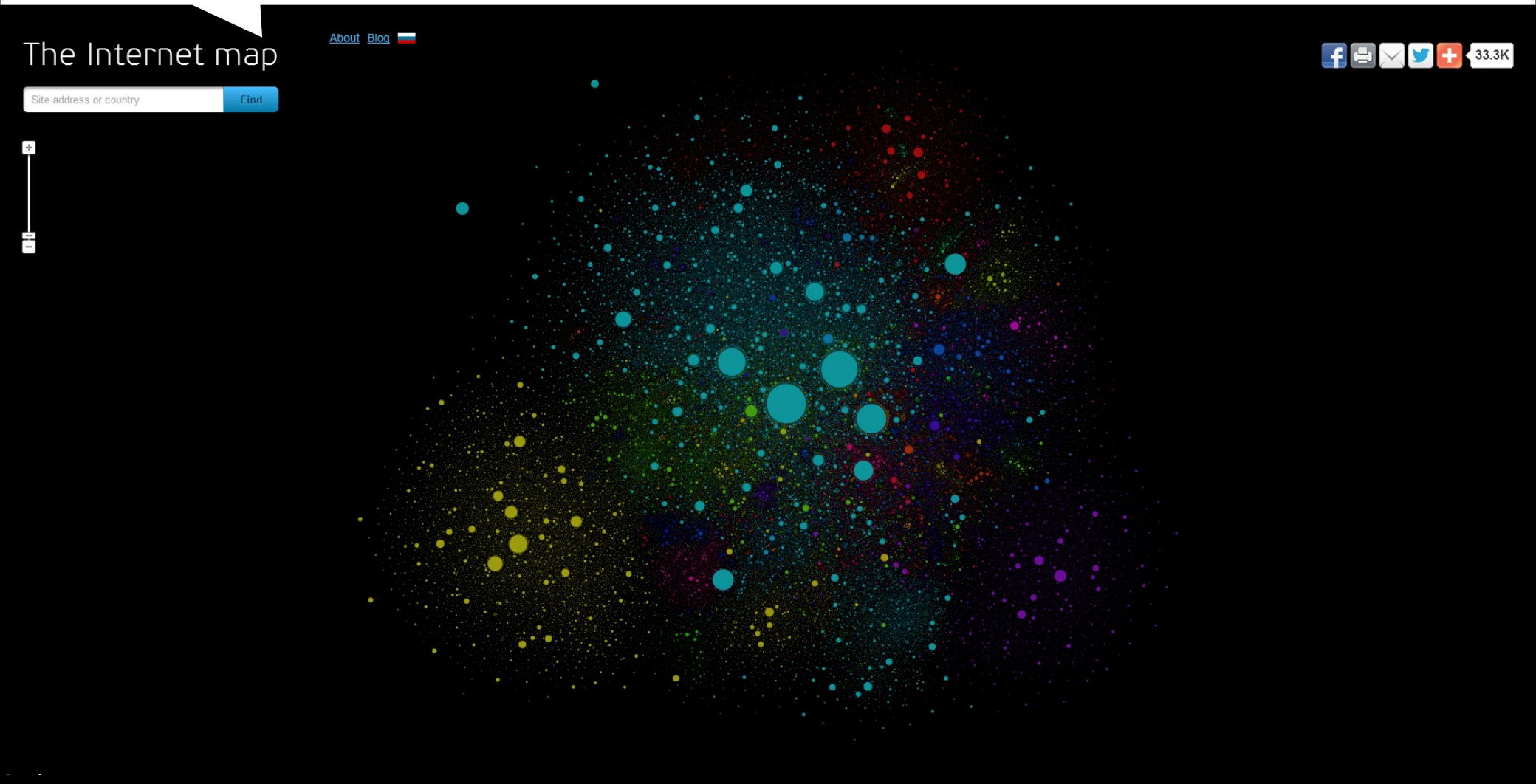
SciencesPo. | médialab

Hyphe (Hypertext Corpus Initiative)



SciencesPo. | médialab

À quoi ressemble le web ?



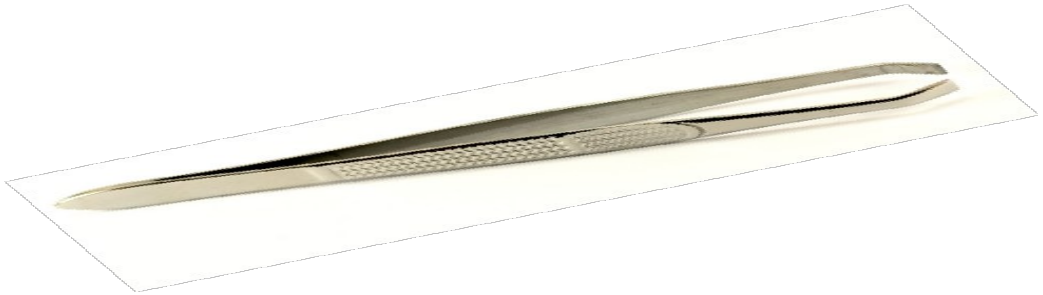
Outils existants pour construire un corpus web

Navicrawler :
créer son corpus manuellement

issuecrawler

Issuecrawler :
créer son corpus automatiquement

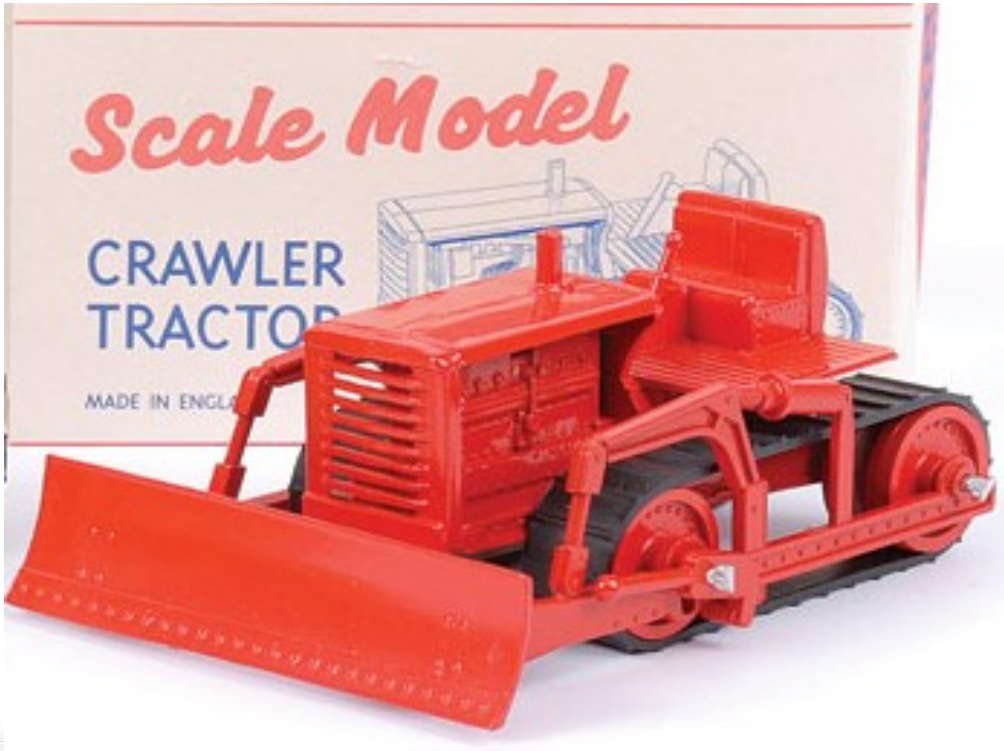
Deux outils non adaptés



une pince à épiler

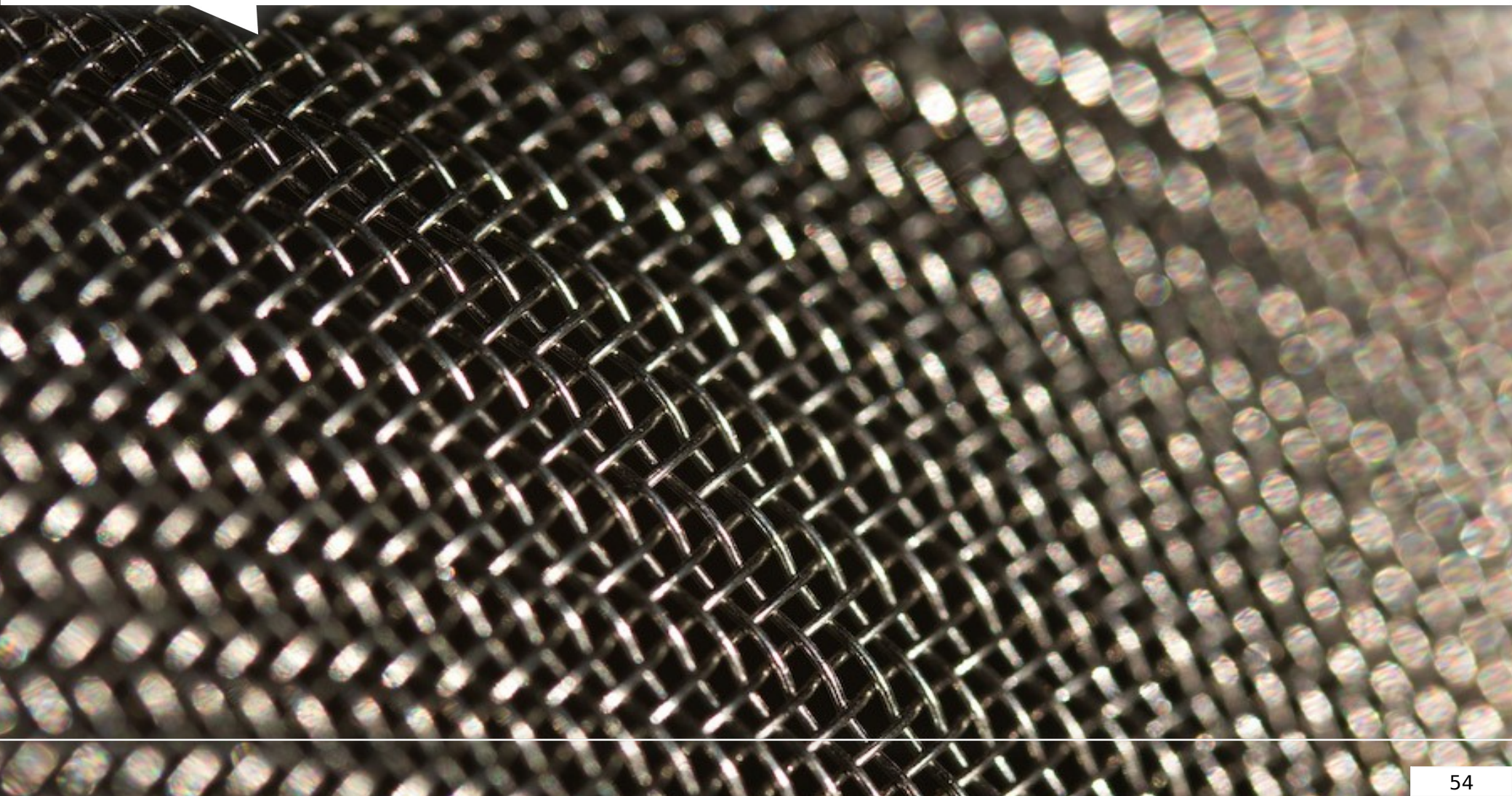
ou

un buldozer

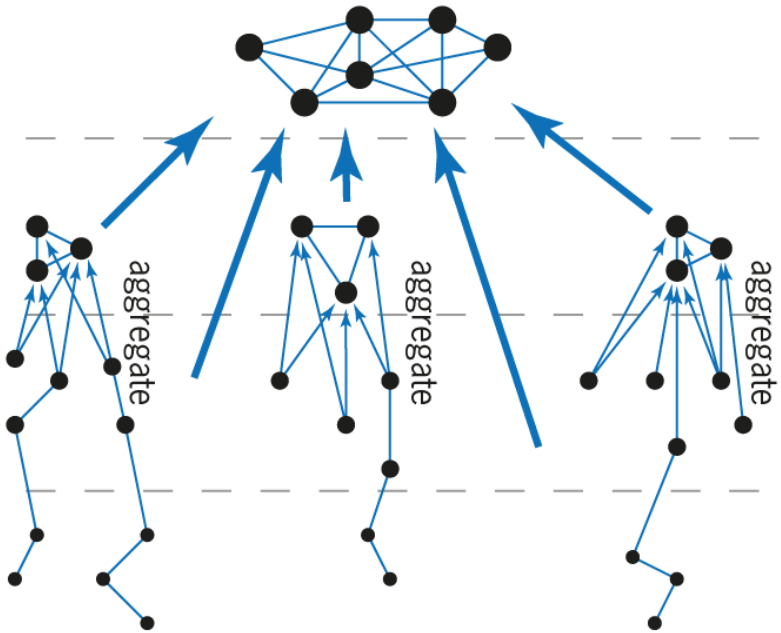


Comment tamiser le web ?

© *Mayhem Chaos*



Les couches du web



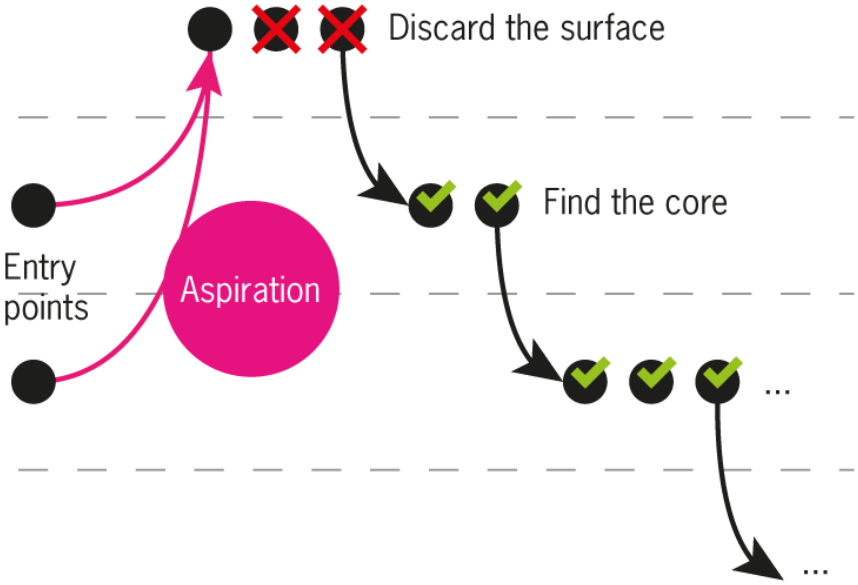
Surface
(Google, Wikipedia...)

Core of aggregates

Periphery of aggregates

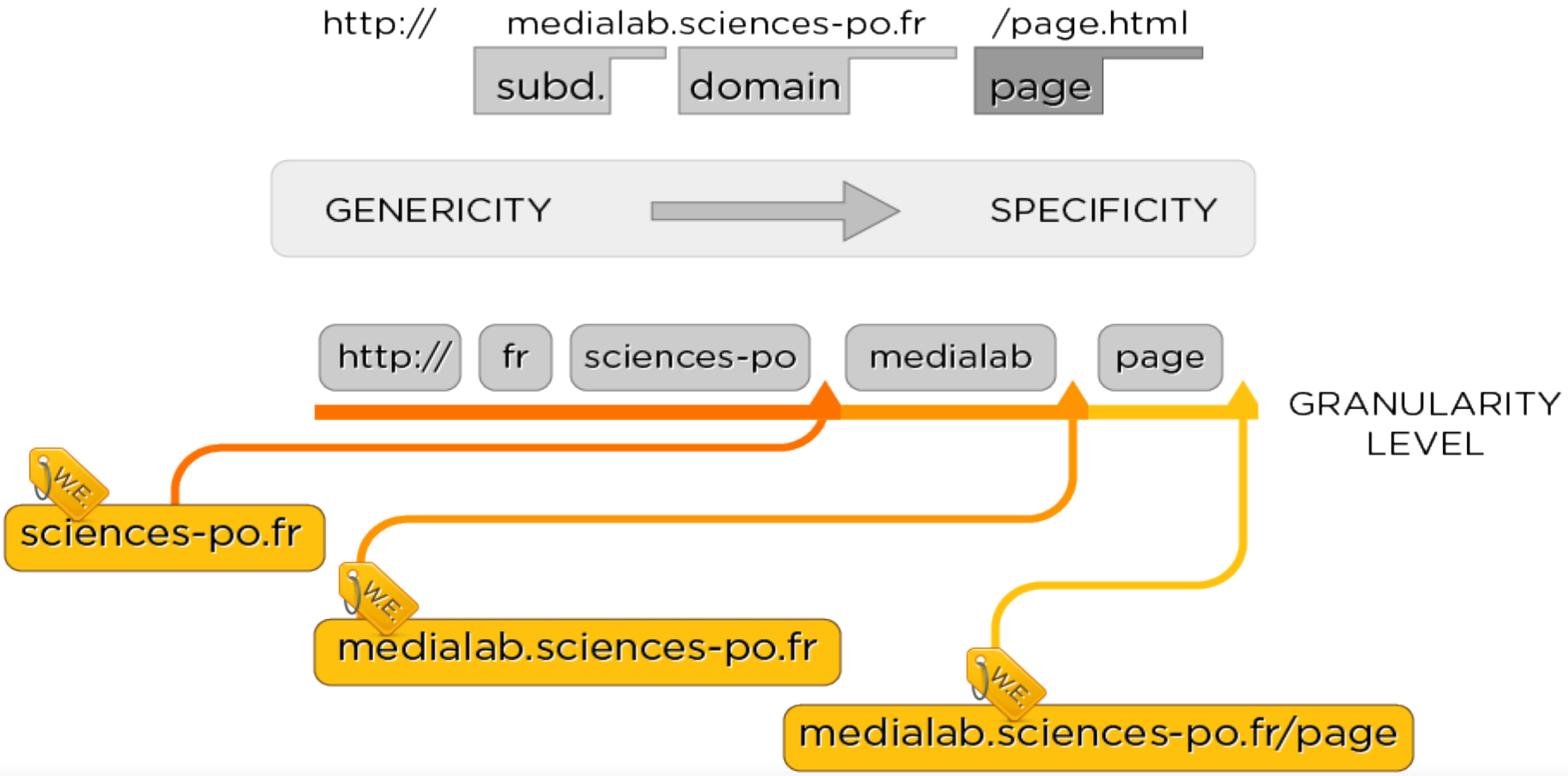
Deep web

Layers of the web

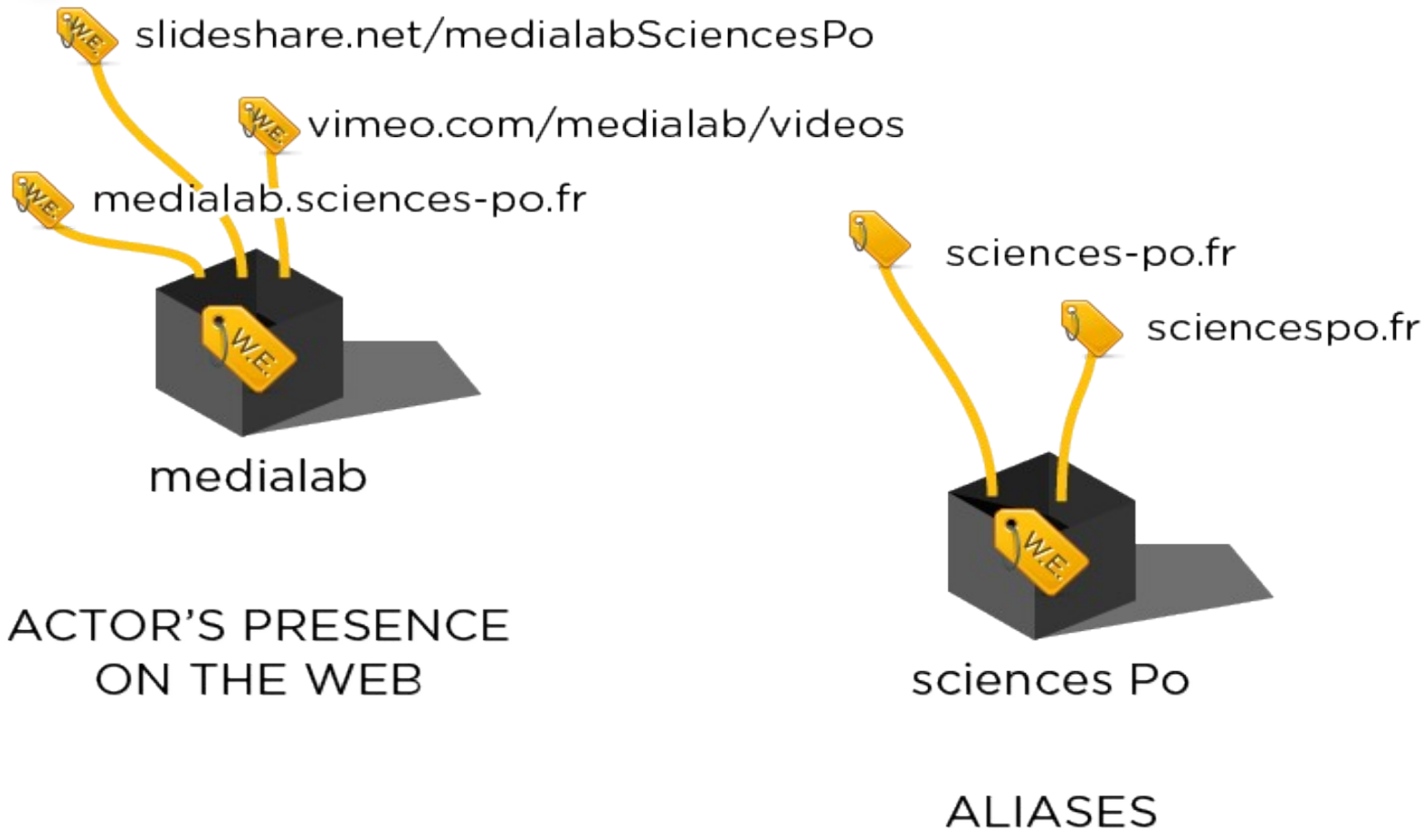


Corpus building scenario

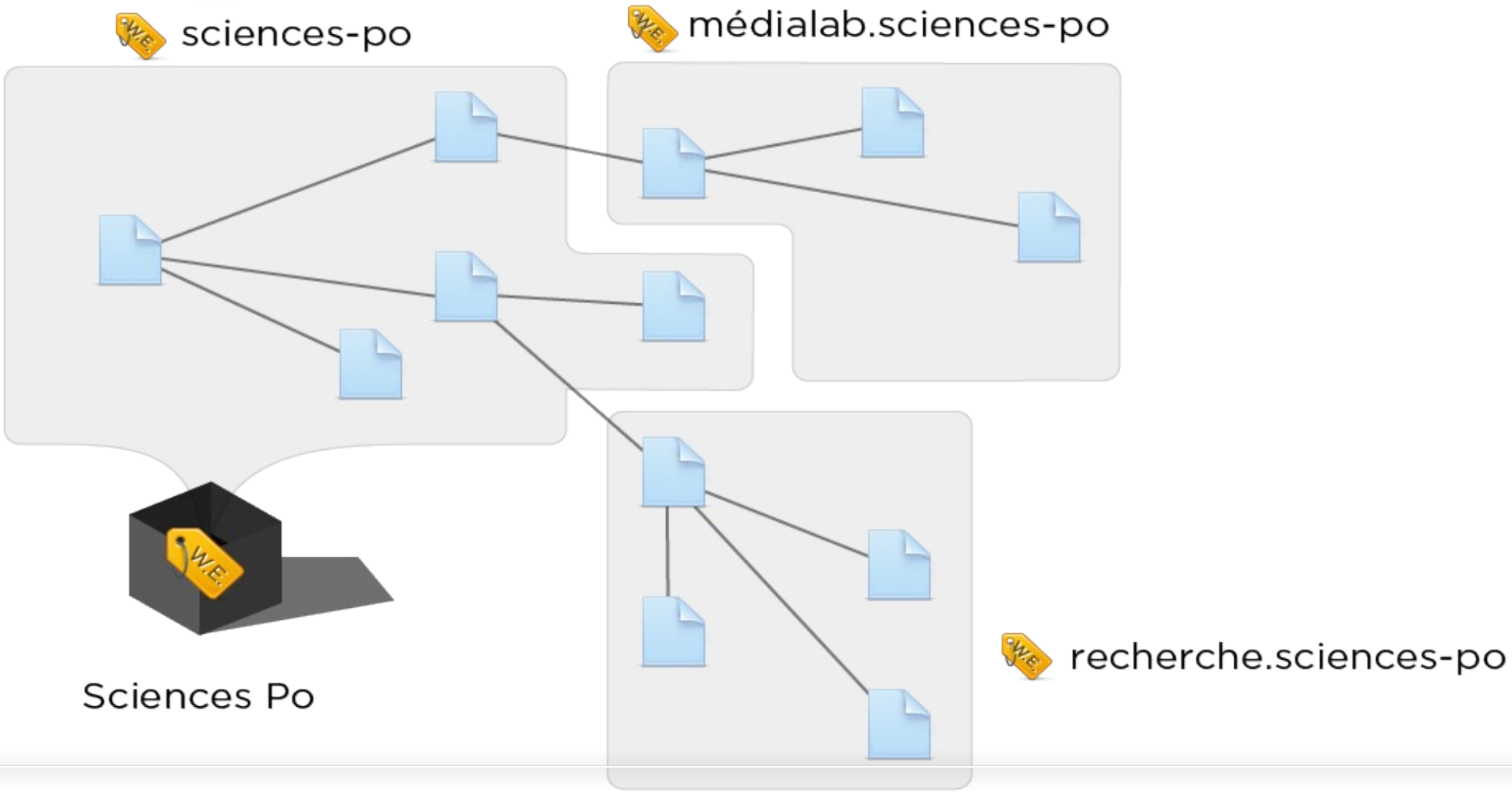
Définir des points d'ancrage précis (LRUs)



Des sites ou... des entités web



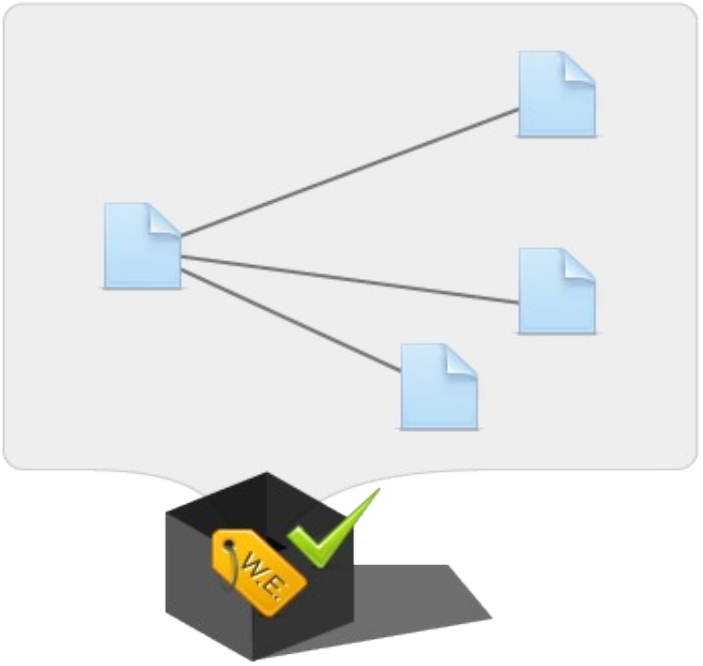
Préserver la complexité



Le crawl dirigé par la recherche

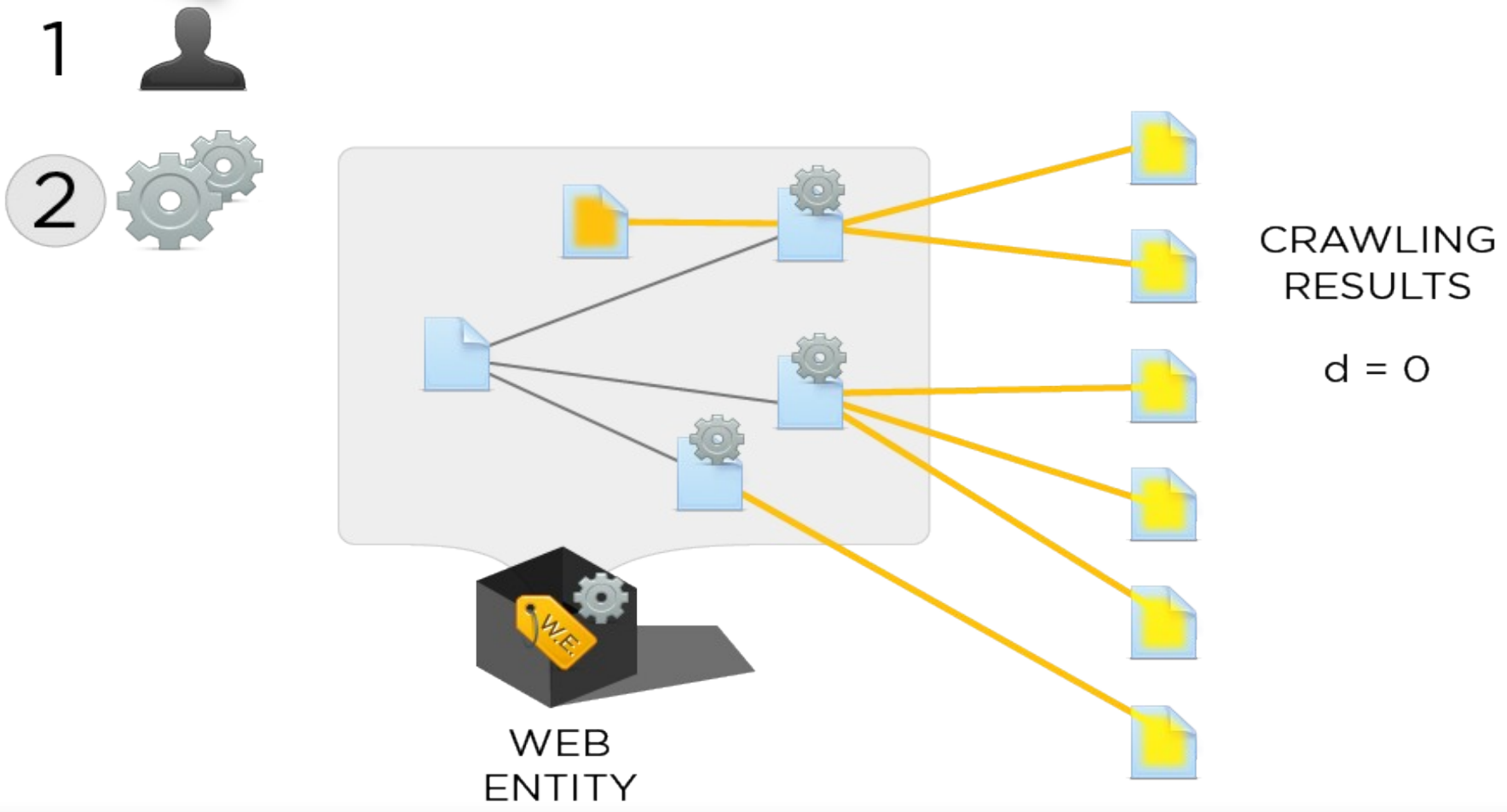
1

RESEARCHER
SELECT
STARTING
ENTITIES

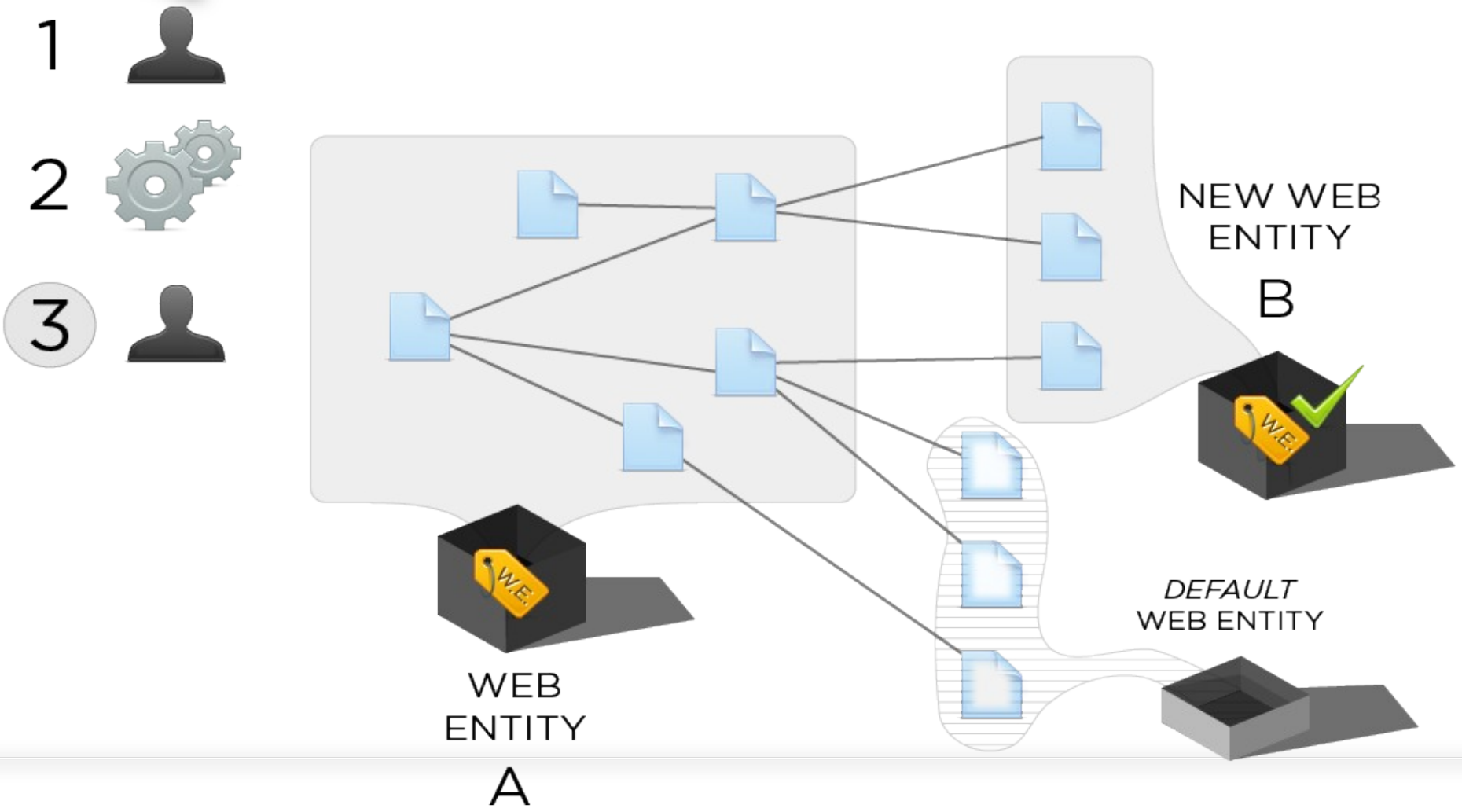


WEB
ENTITY

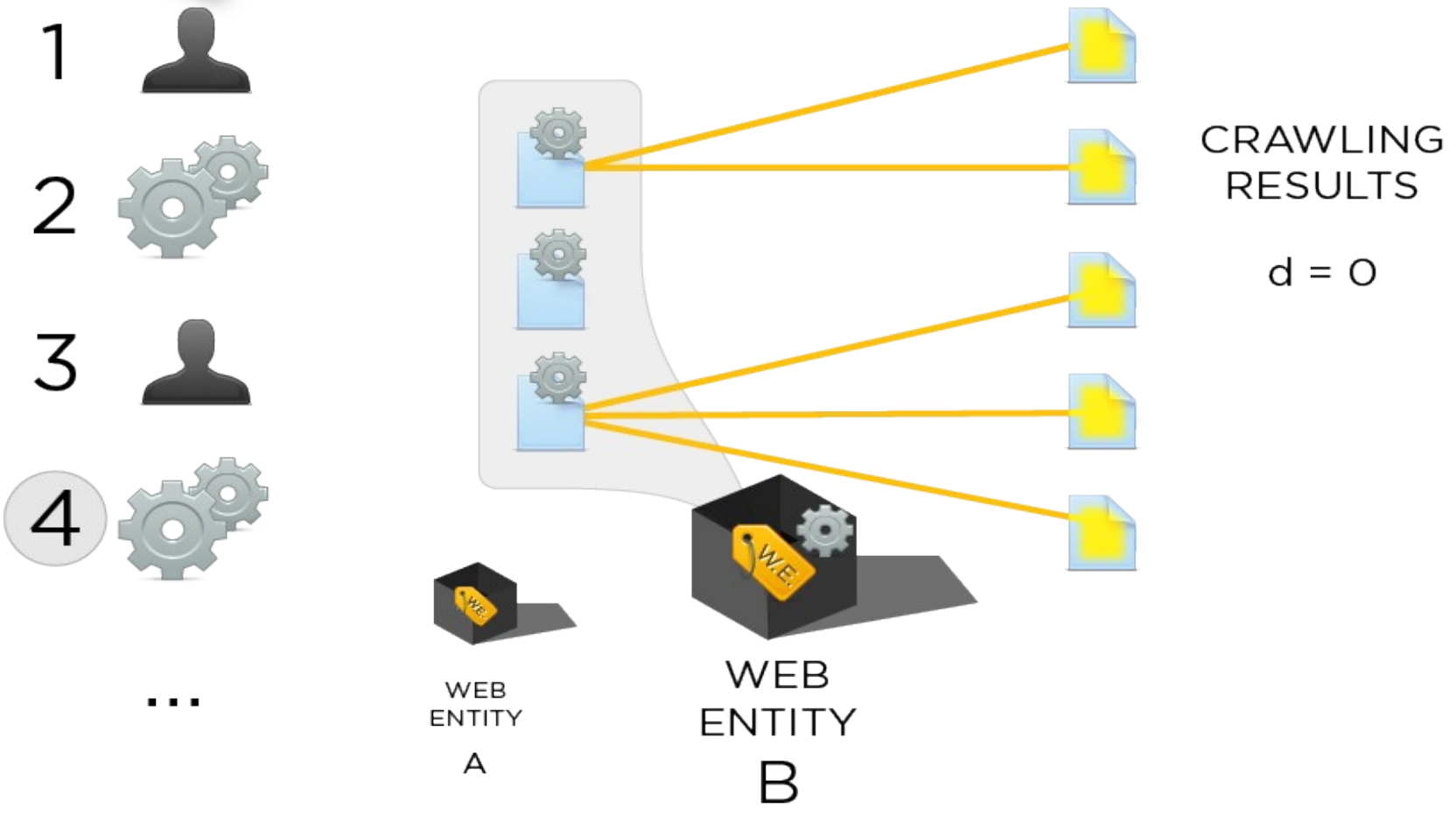
Le crawl dirigé par la recherche



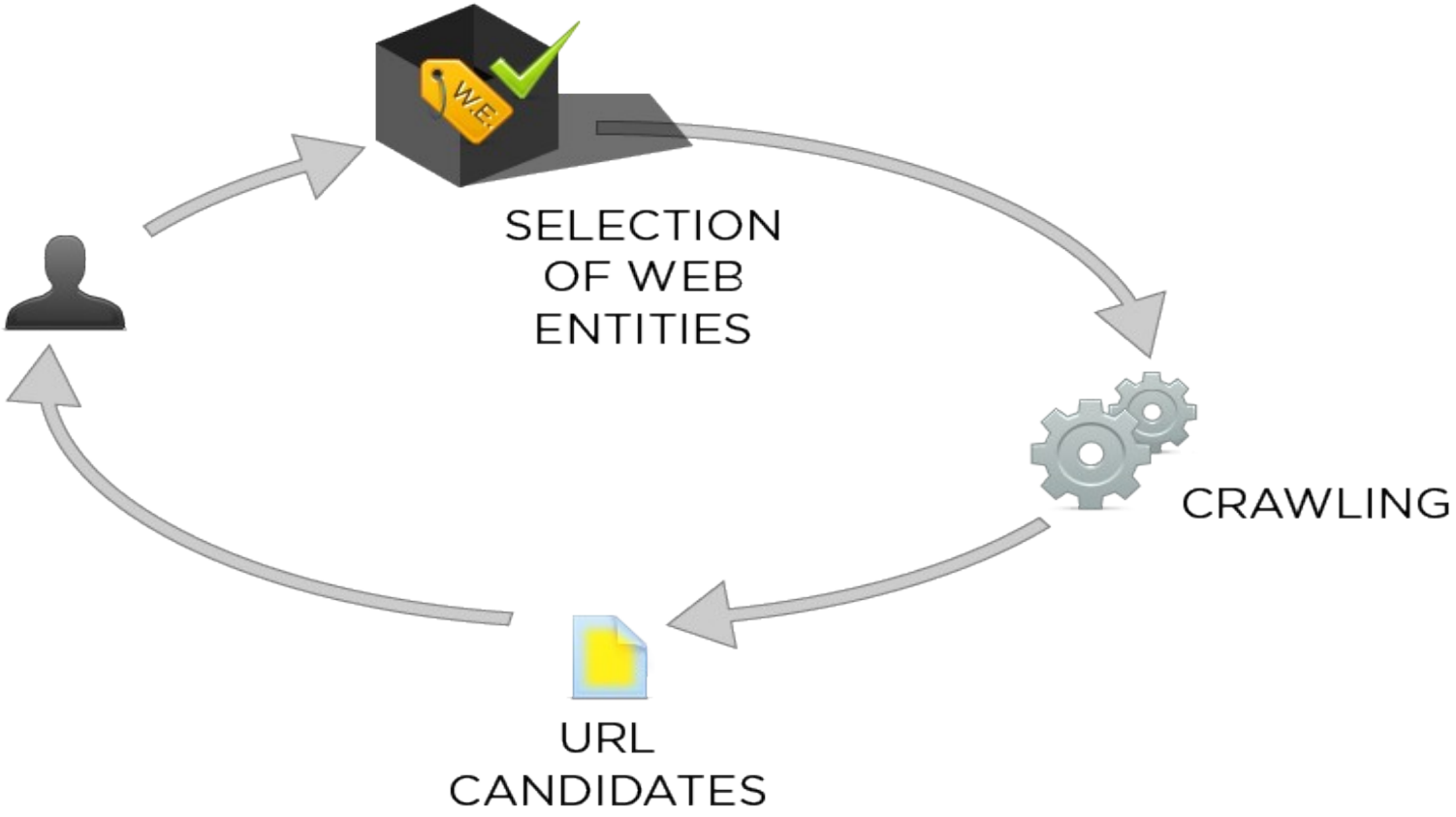
Le crawl dirigé par la recherche



Le crawl dirigé par la recherche



Le crawl dirigé par la recherche

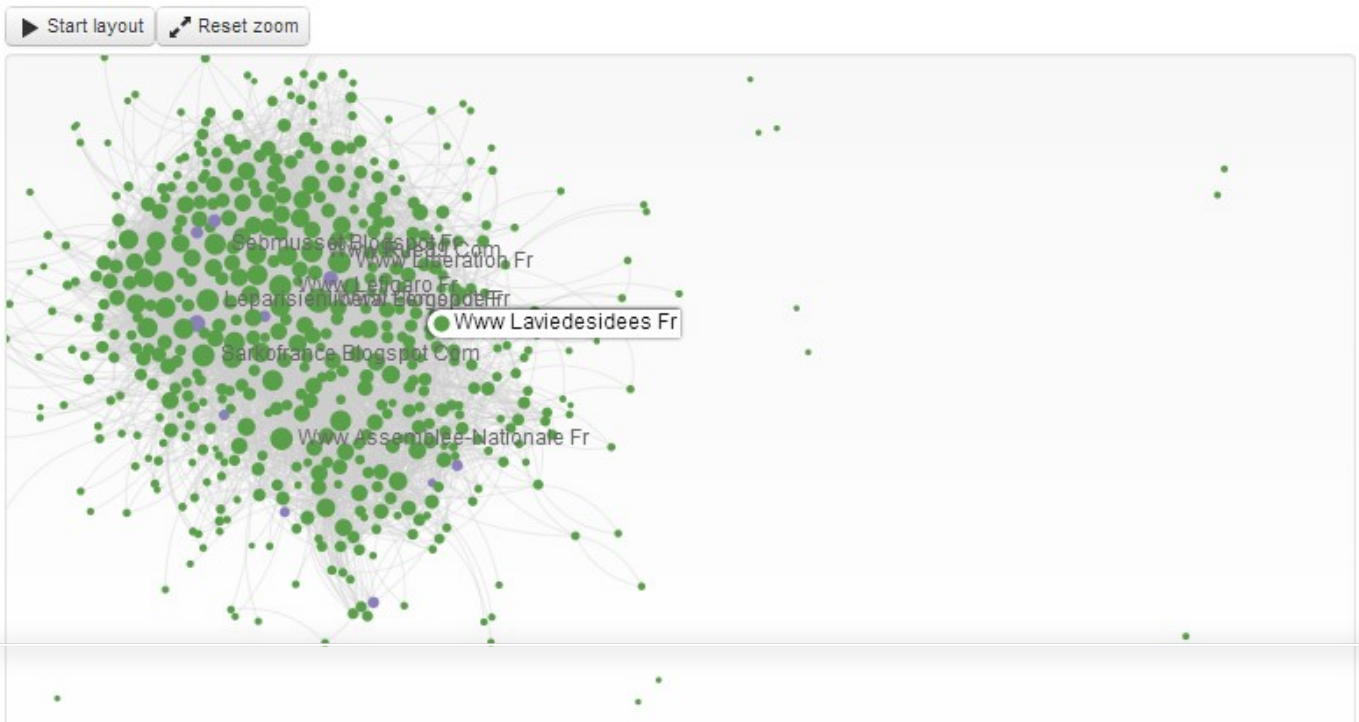


Exemple : SITPOL

Network of webentities

Do shit n' stuff

Preview



Modes

Which web entities to display

- **Corpus in progress** - IN + UNDECIDED
  The corpus including web entities you still have to accept or refuse
- **Top neighbors** - IN + UNDECIDED + top DISCOVERED
  The corpus in progress with neighbors (discovered web entities) cited 3+ times by other web entities
- **Corpus strict** - IN only
  The pure corpus, as the result of selection process
- **Frontier** - IN + UNDECIDED + OUT
  The corpus and its frontier (rejected web entities), for analysis or monitoring the selection process
- **All neighbors** - IN + UNDECIDED + DISCOVERED
  The corpus in progress with all discovered web entities

Hyphe : roadmap

- Interfaces utilisateurs idiot-proof / avancée
- Multicorpus
- Analyse de contenus textes
- Exploration de corpus d'archives existants
  → BNF / INA
- Analyse évolutive dans le temps
- Intégration simplifiée
  → Clé usb, extension navigateur...

Merci de votre attention !

benjamin.ooghe@sciencespo.fr