



HAL
open science

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

Benjamin Ooghe

► **To cite this version:**

Benjamin Ooghe. L'archivage piloté par la recherche : l'exemple du médialab. Master II ENSSIB - Archivage, ENSSIB, Jan 2014, Villeurbanne, France. hal-03631534

HAL Id: hal-03631534

<https://sciencespo.hal.science/hal-03631534v1>

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

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

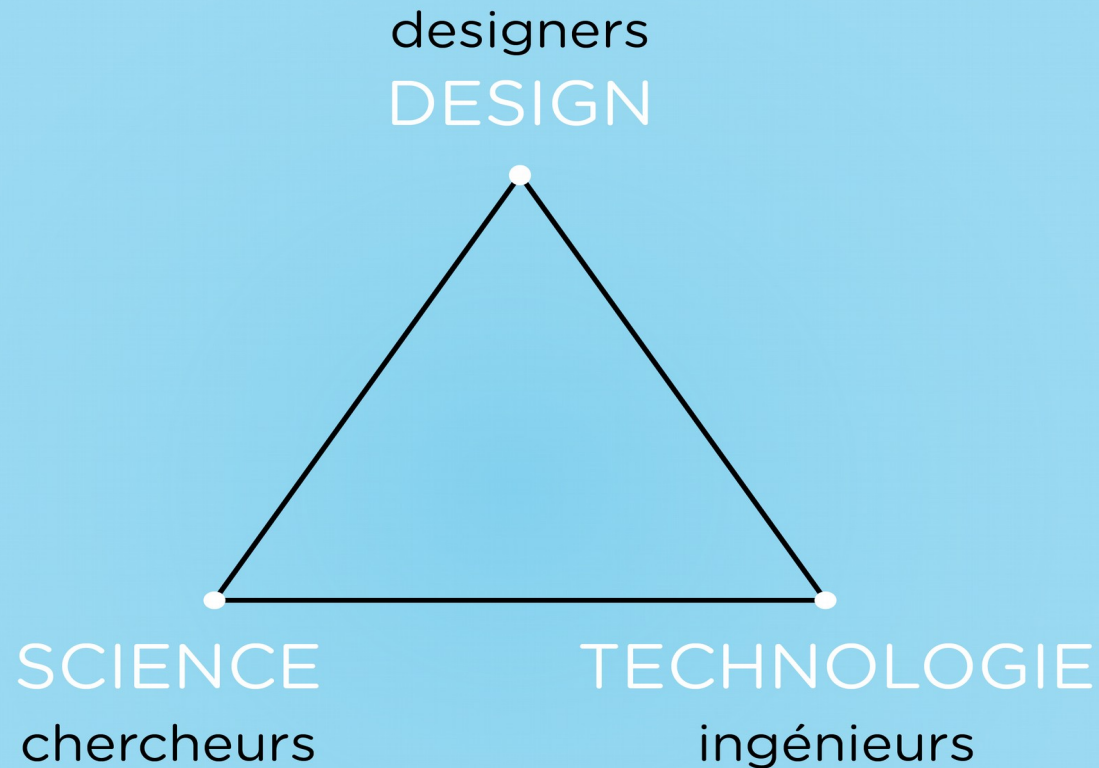
médialab : un lieu



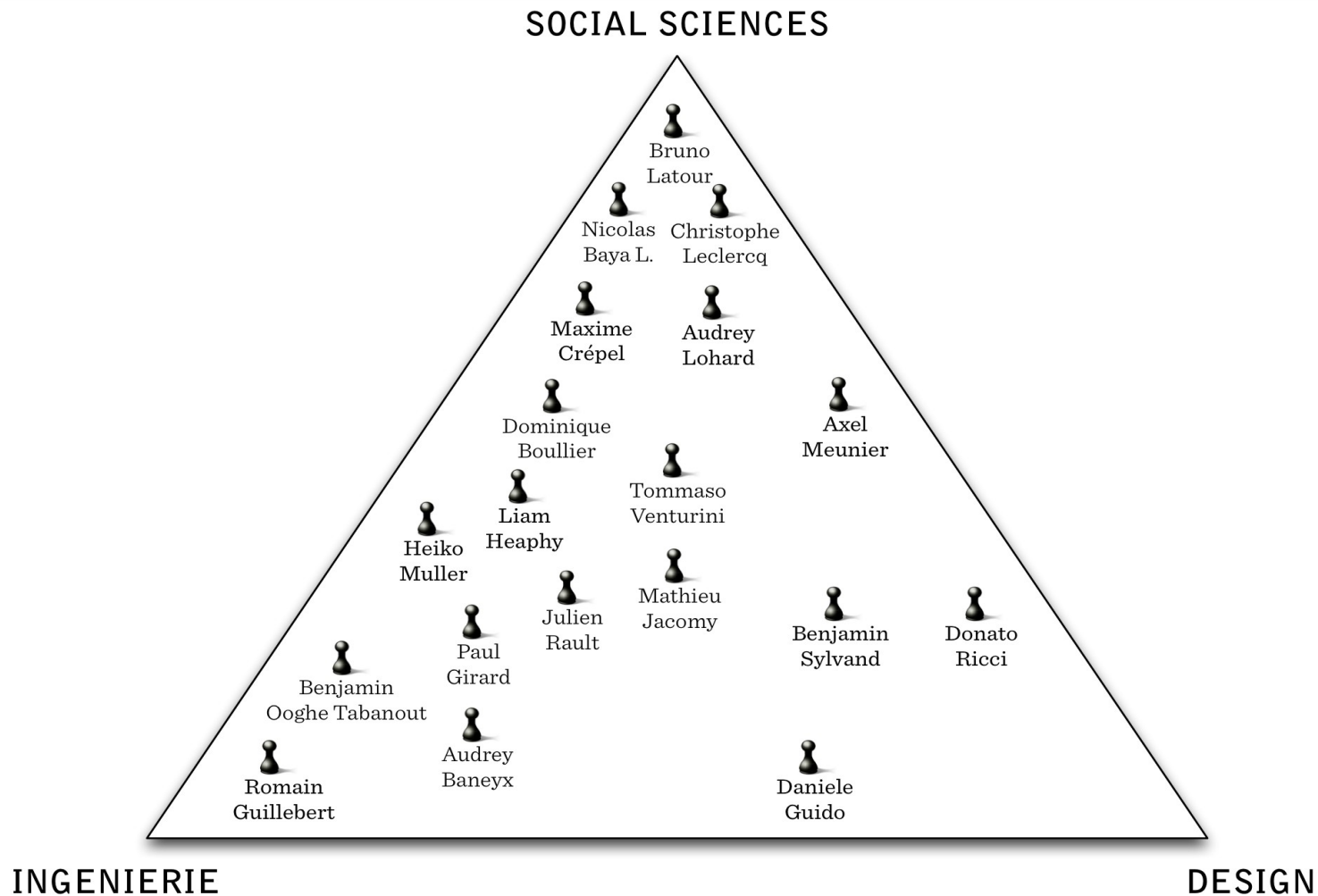
médialab : une équipe



médialab : une combinaison de compétences



médialab : une combinaison de compétences



Notre projet : quel ancrage disciplinaire ?

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...



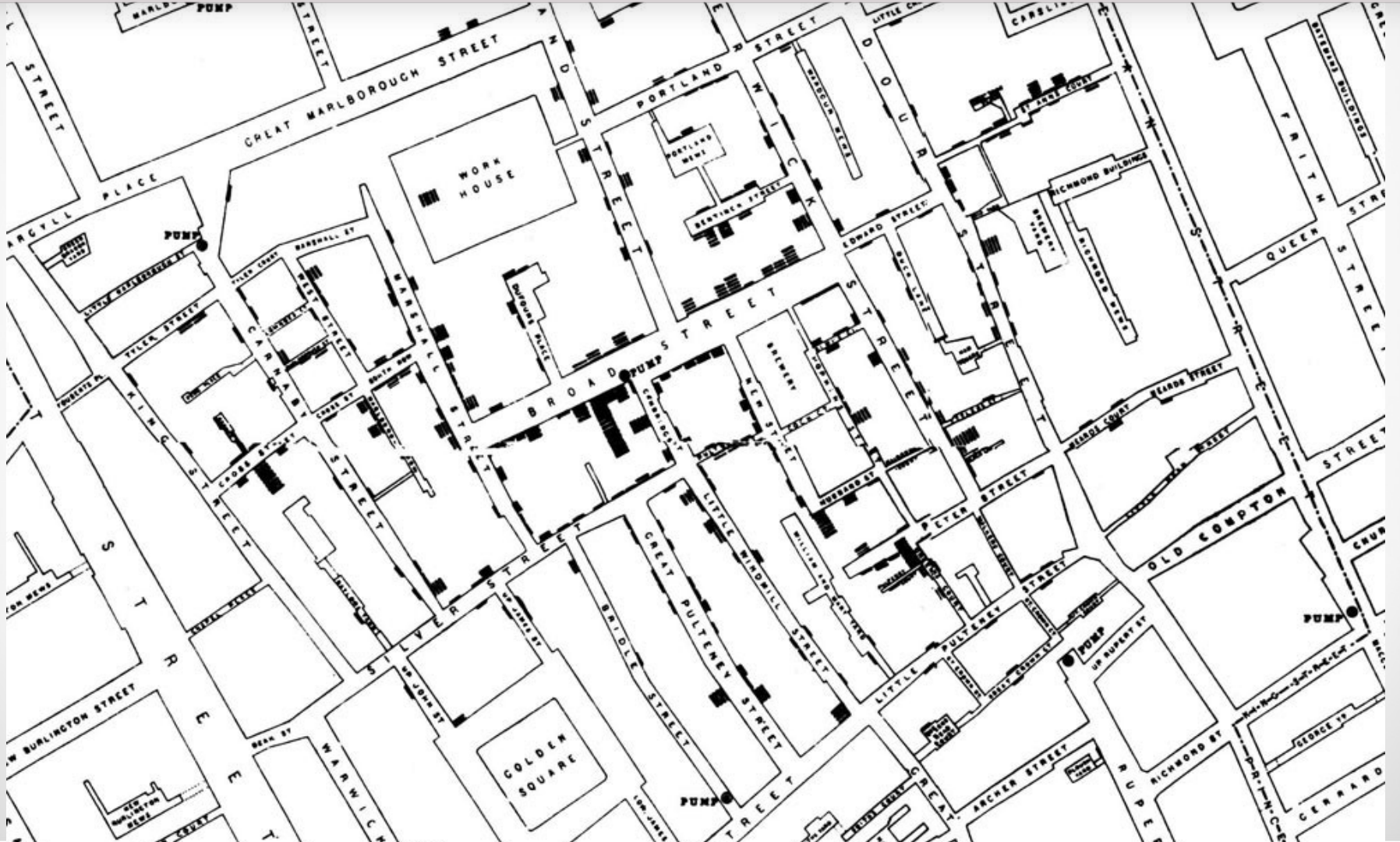
... à un support de traces



Une discipline naissante



*On the mode of
Communication of cholera
John Snow, 1855*



Data deluge

The Economist
25 February 2010



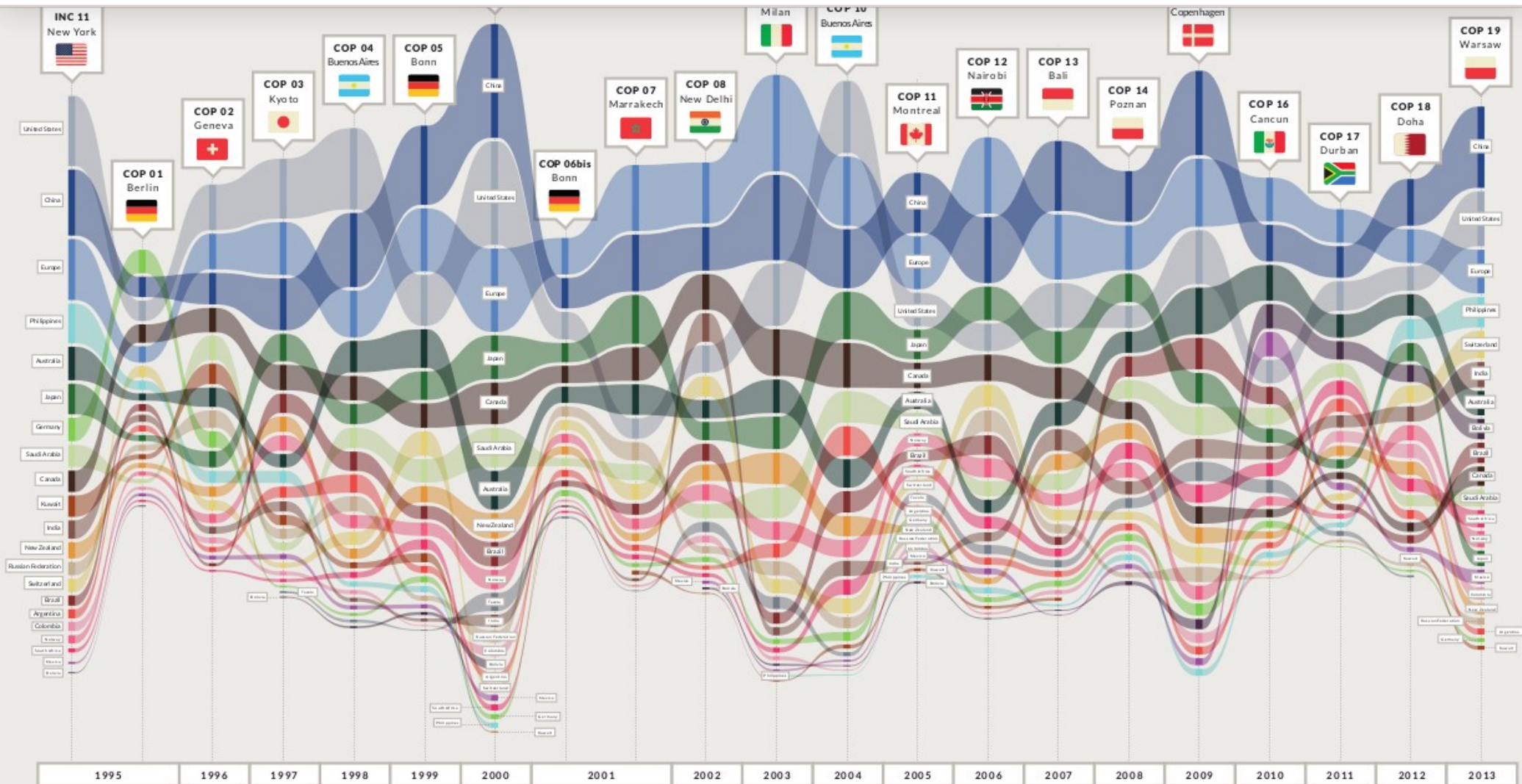
Théorie acteurs-réseaux, B. Latour



Biennale
Venice
2009

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



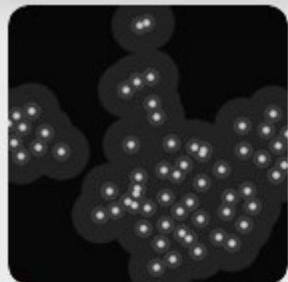
SciencesPo.

médialab

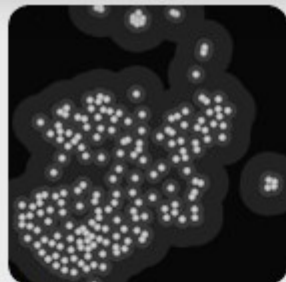
e-Diasporas Atlas

e-diasporas.fr

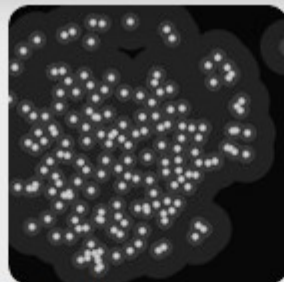
Dana Diminescu



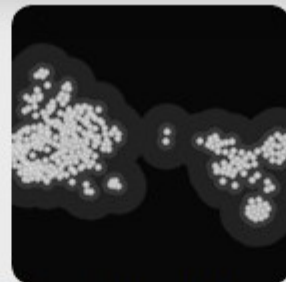
Breton corpus



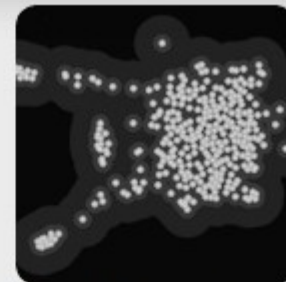
Chinese corpus



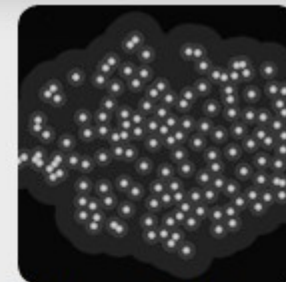
Egyptian corpus



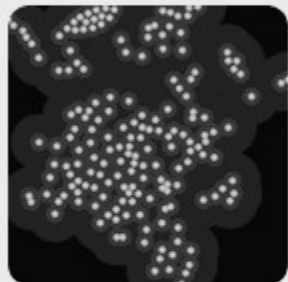
French Expatriates



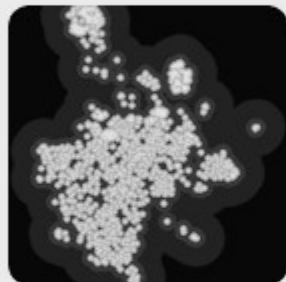
French Repatriates



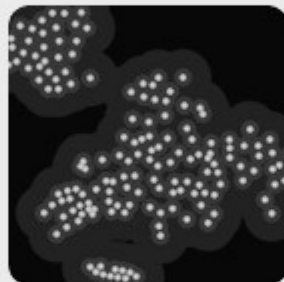
Hindu - Hindutva



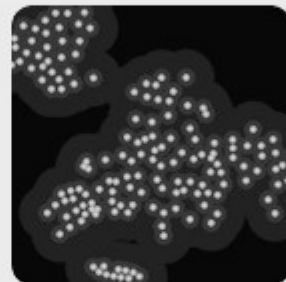
Hmong corpus



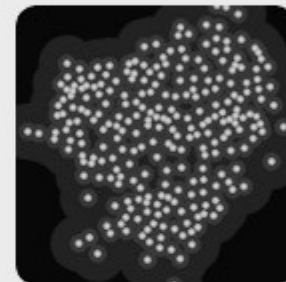
Indian corpus



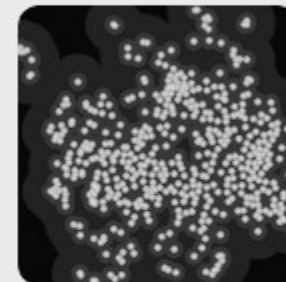
Indian Real Estate



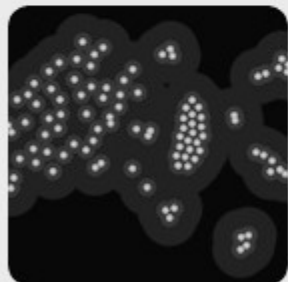
Italian corpus



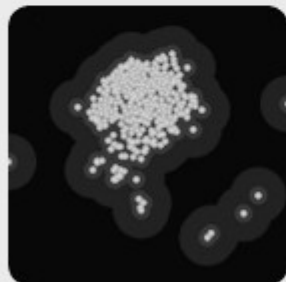
Jewish corpus



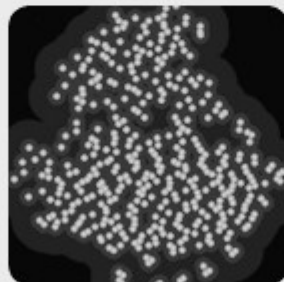
Kerala corpus



Lebanese corpus



Macedonian corpus



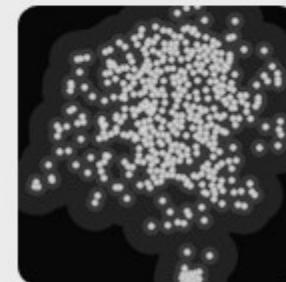
Mexican corpus



Moroccans on FB



Moroccan corpus

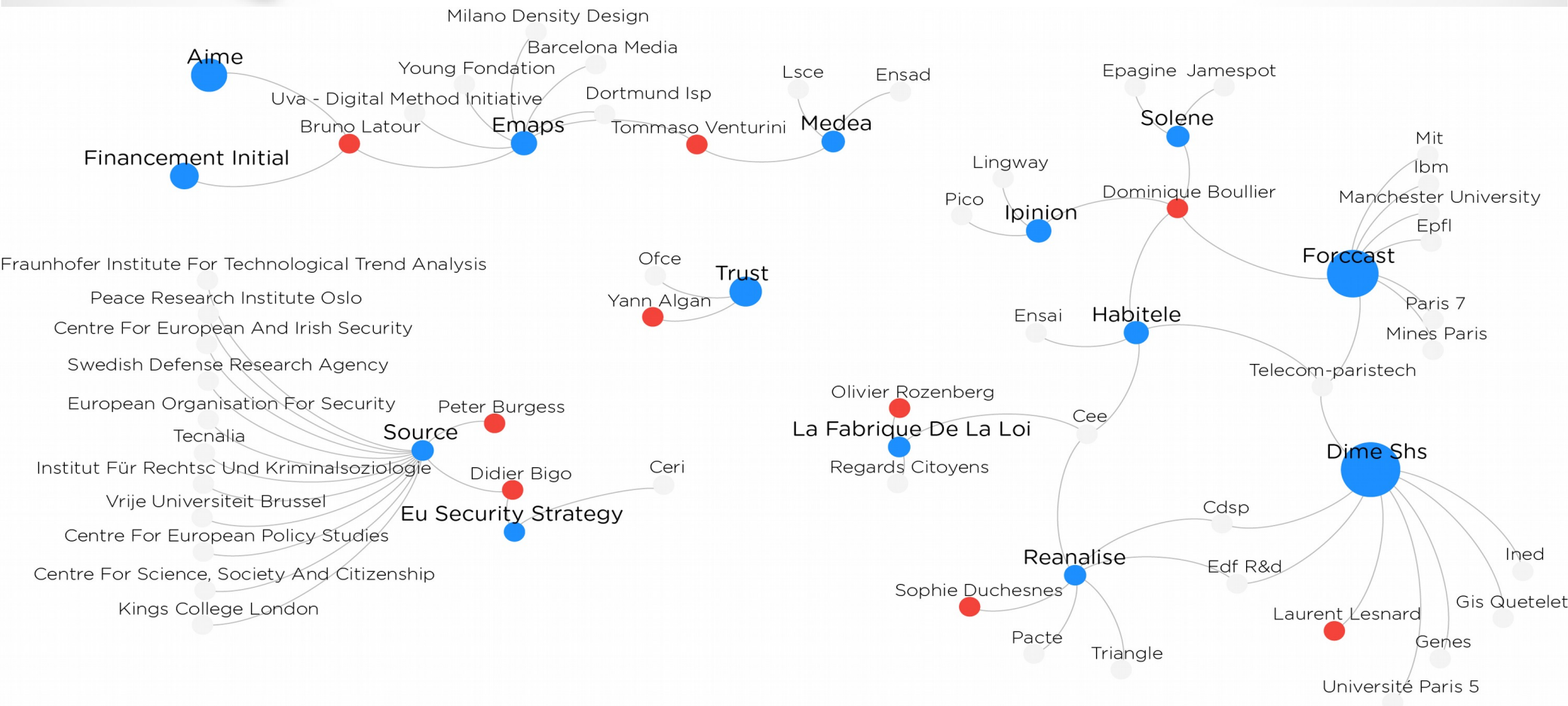


Nepali corpus

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

2012

<p>Faibles doses radioactivité</p>	<p>Extension Gare de Genève</p>	<p>Lignes à haute tension</p>	<p>Climate Changes. What about People?</p>	<p>La mort des abeilles</p>	<p>Le marché du carbone</p>	<p>La sécurité des aéroports</p>	<p>Vidéosurveillance</p>	<p>Enfouissement des déchets nucléaires</p>	<p>THEEYE</p>	<p>High Speed Rail</p>	<p>HOME</p>
<p>Cruising the Web</p>	<p>COTree</p>	<p>Memorial to the Murdered Jews of Europe</p>	<p>Park 51</p>	<p>Reddography</p>	<p>There's No Place Like Antarctica</p>	<p>Islanders</p>	<p>Cape Wind project</p>	<p>The Shard, London</p>	<p>Beijing Olympic Green</p>	<p>Faibles Doses</p>	<p>Fertilisation Océans</p>
<p>Norwegian Government Quarter</p>	<p>Liverpool Waters</p>	<p>London River Park</p>	<p>Museum of Liverpool</p>	<p>Environmental Vegetarianism</p>	<p>We have four: pick one Food fibre and forest products</p>	<p>Tapping Cascade Locks</p>	<p>Some like it cold Polar regions</p>	<p>Hadopi</p>	<p>Kilogramme</p>	<p>Modélisation Finance</p>	<p>Aide au Développement</p>
<p>Lord's Cricket Ground Redevelopment Controversy</p>	<p>Psychotisation - une pathologisation controversée</p>	<p>La Labellisation du Vin</p>	<p>Blasphème</p>	<p>Anthropic Cycle Freshwater</p>	<p>A little big apocalypse Small Islands</p>	<p>The Dunes of Scotland Golf Resort</p>	<p>BBC Broadcasting House, London</p>	<p>Statistiques Ethniques</p>	<p>Versailles Controversé</p>	<p>Mont Saint Michel</p>	<p>Hypersensibilité aux Ondes</p>
<p>Neutrinos</p>	<p>Gardasil</p>	<p>Biocarburant</p>	<p>Benzodiazepines</p>	<p>Ground Zero</p>	<p>Birmingham Central Library</p>	<p>Prince Charles</p>	<p>Arcelor Mittal</p>	<p>Lampes Basse Consommation</p>			
<p>35 heures</p>											

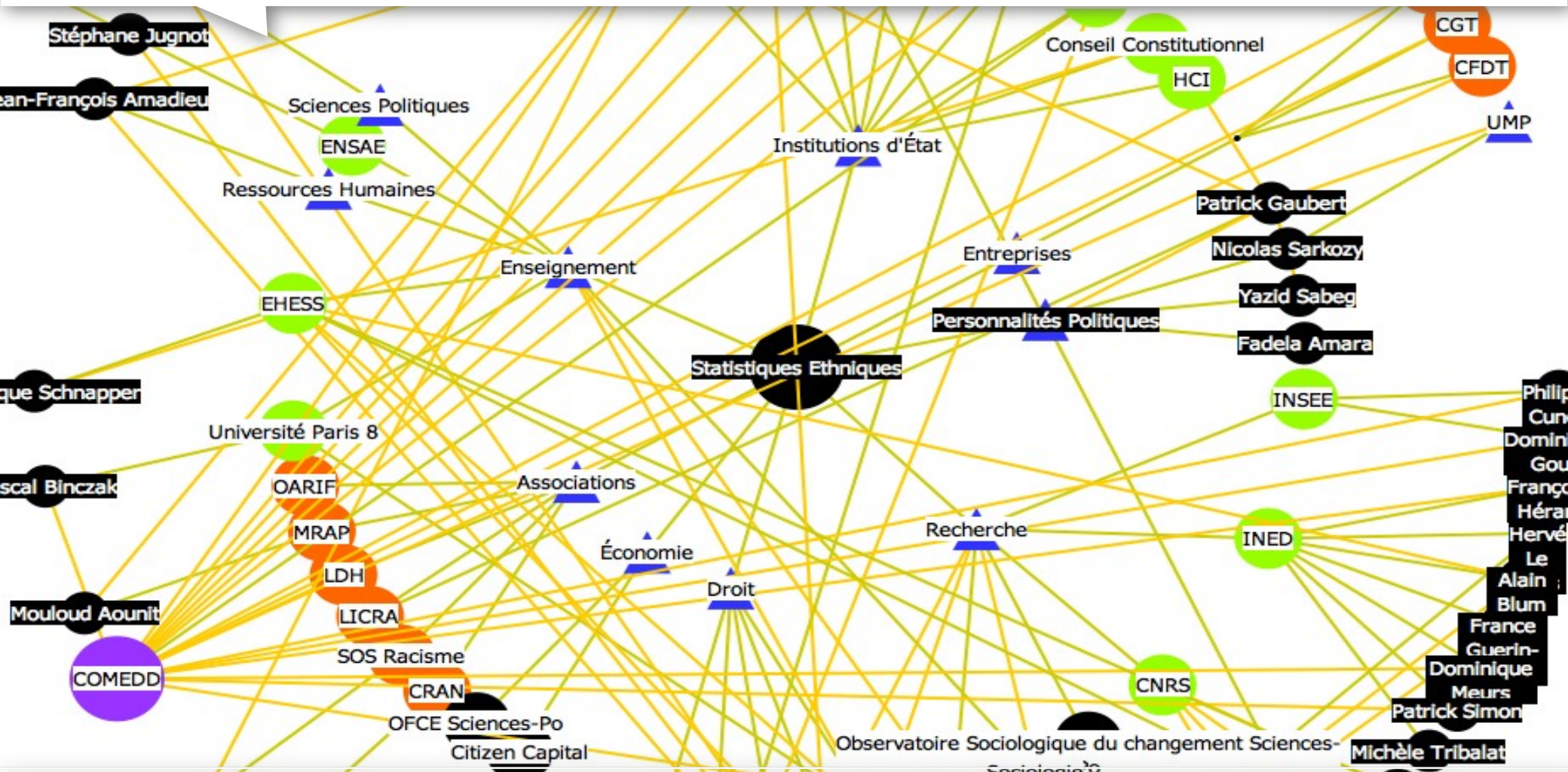
2009

<p>Sydney Opera House</p>	<p>London Olympics 2012 Legacy</p>	<p>Geoffrey Chaucer School</p>	<p>Arctique</p>
<p>LES ENJEUX DE LA CONSERVATION DU GANG DE CORDON</p>	<p>Street Art</p>	<p>Vaccin Hépatite</p>	<p>Biocarburants</p>

2008

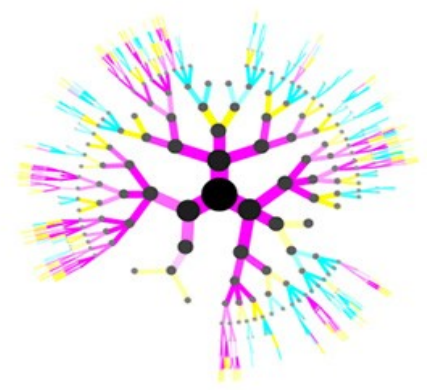
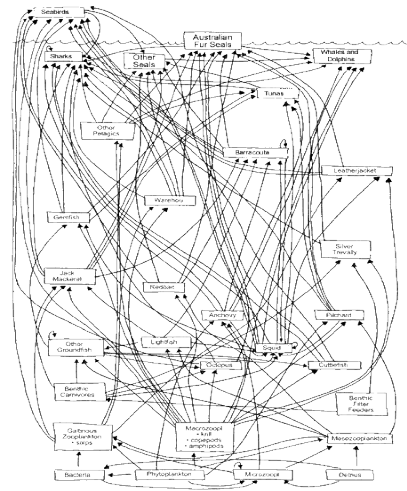
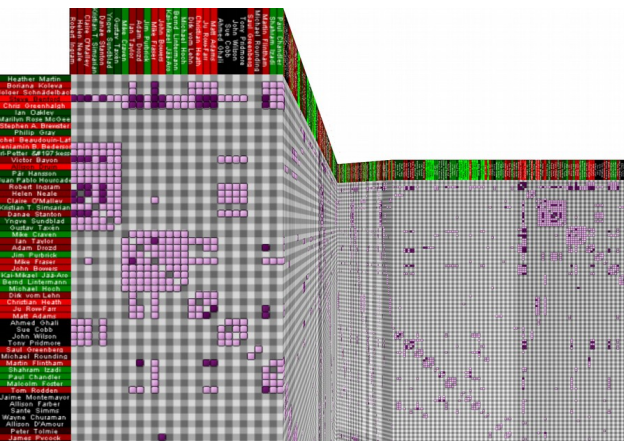
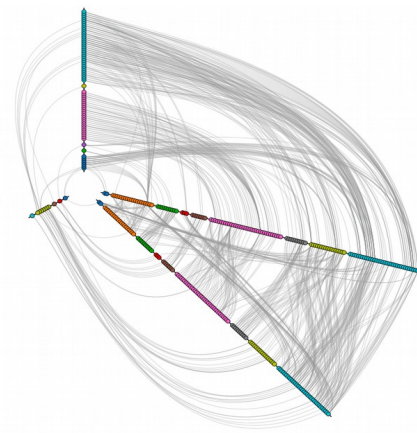
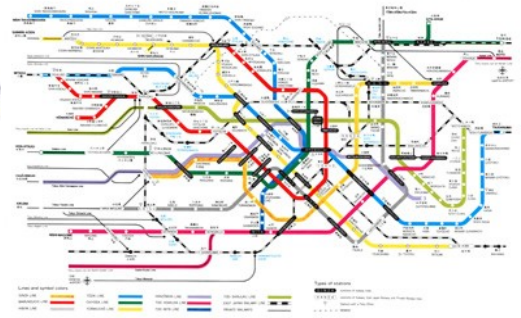
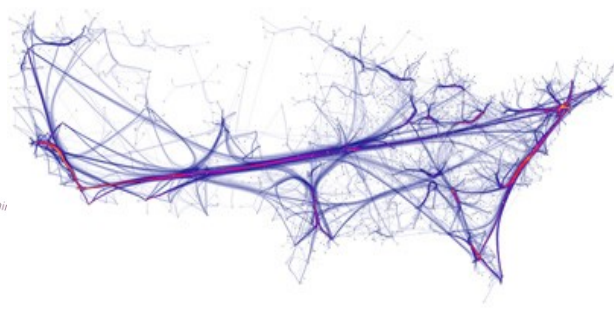
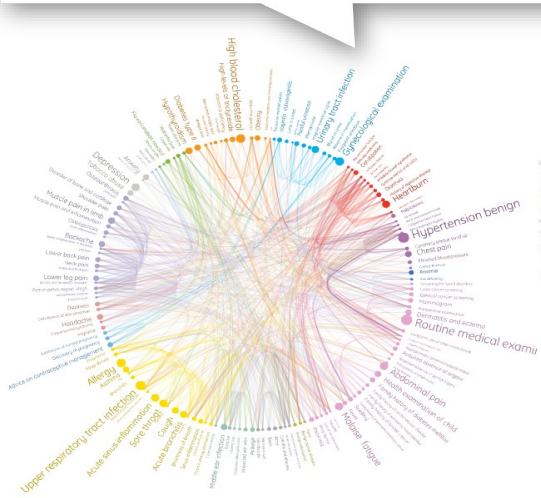
<p>Abeilles et Pesticides</p>	<p>Alzheimer</p>	<p>Hypothèse Gaïa</p>	<p>Classements Universités</p>
-------------------------------	------------------	-----------------------	--------------------------------

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

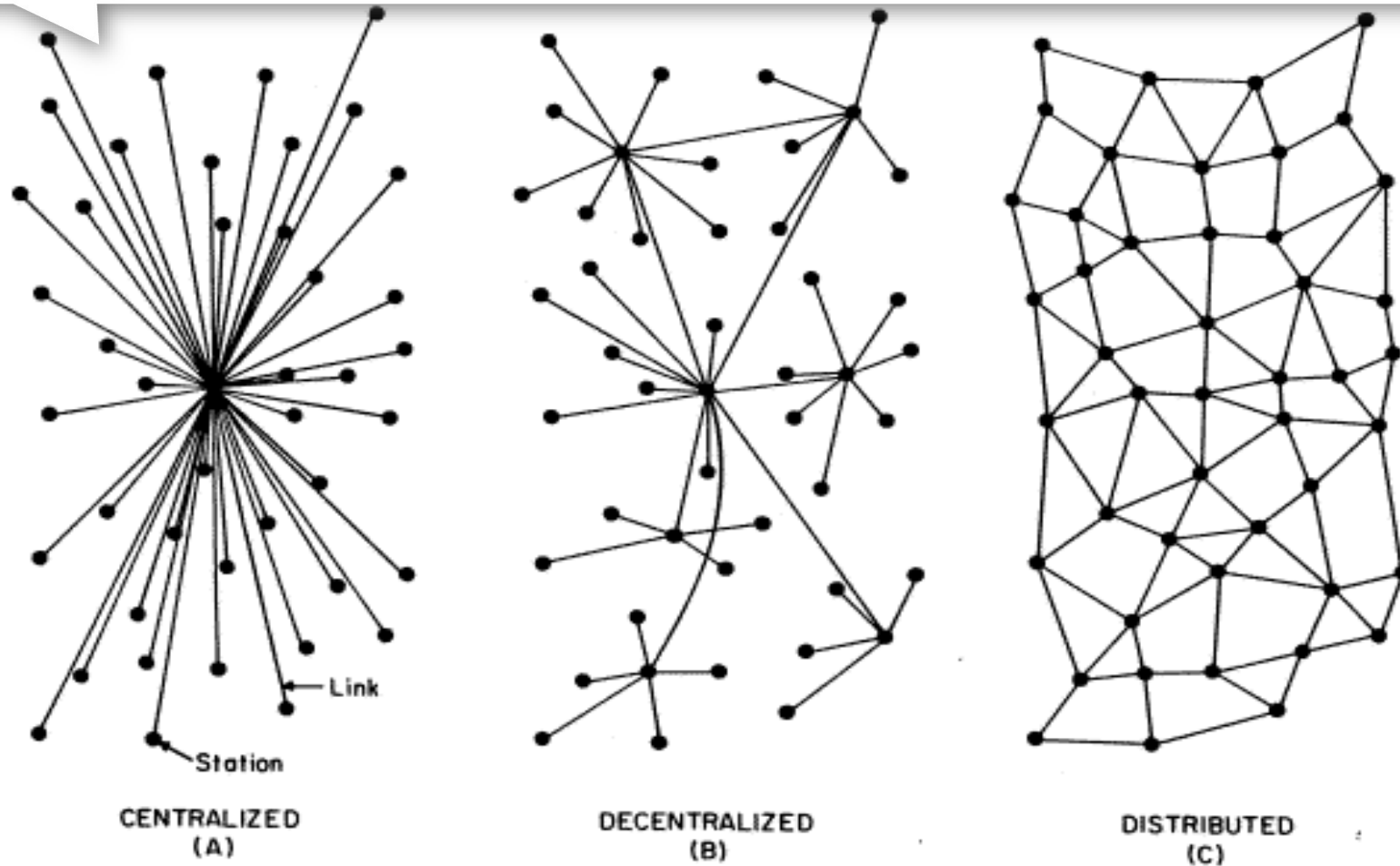


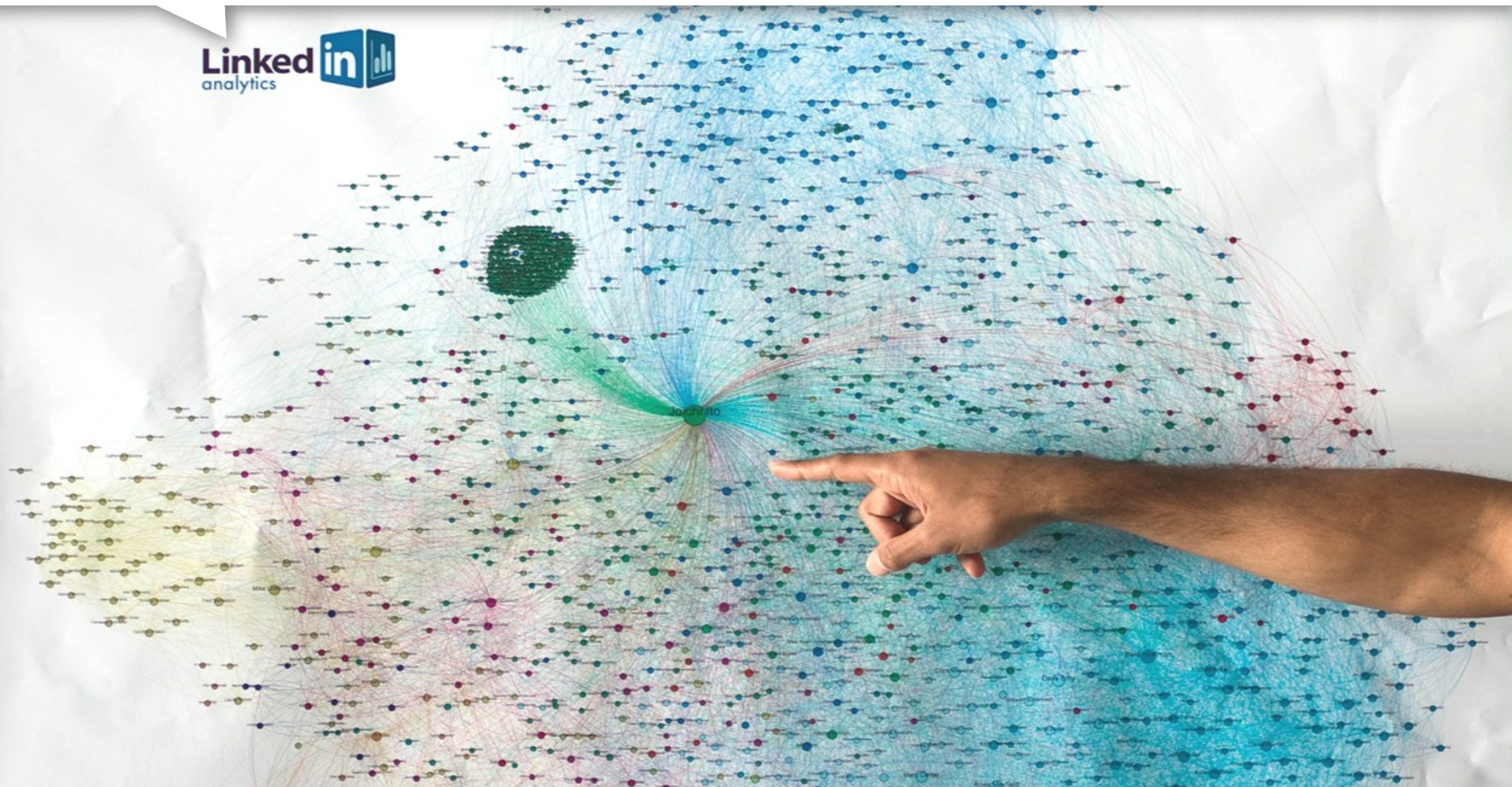
FIG. 1 – Centralized, Decentralized and Distributed Networks

...aux réseaux-cartes

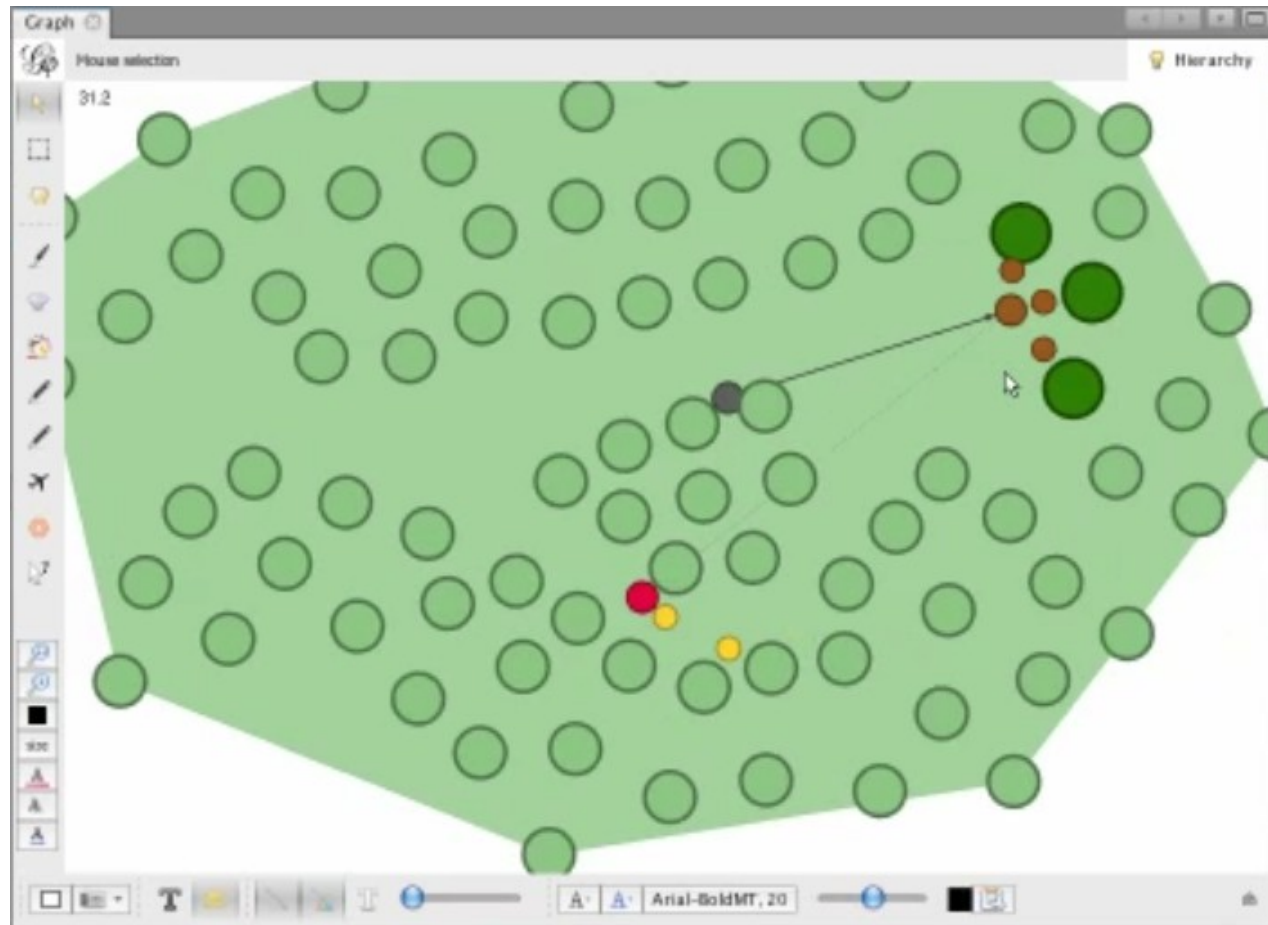
London Underground 1933 Map (Harry Beck)



Les réseaux : outils de navigation



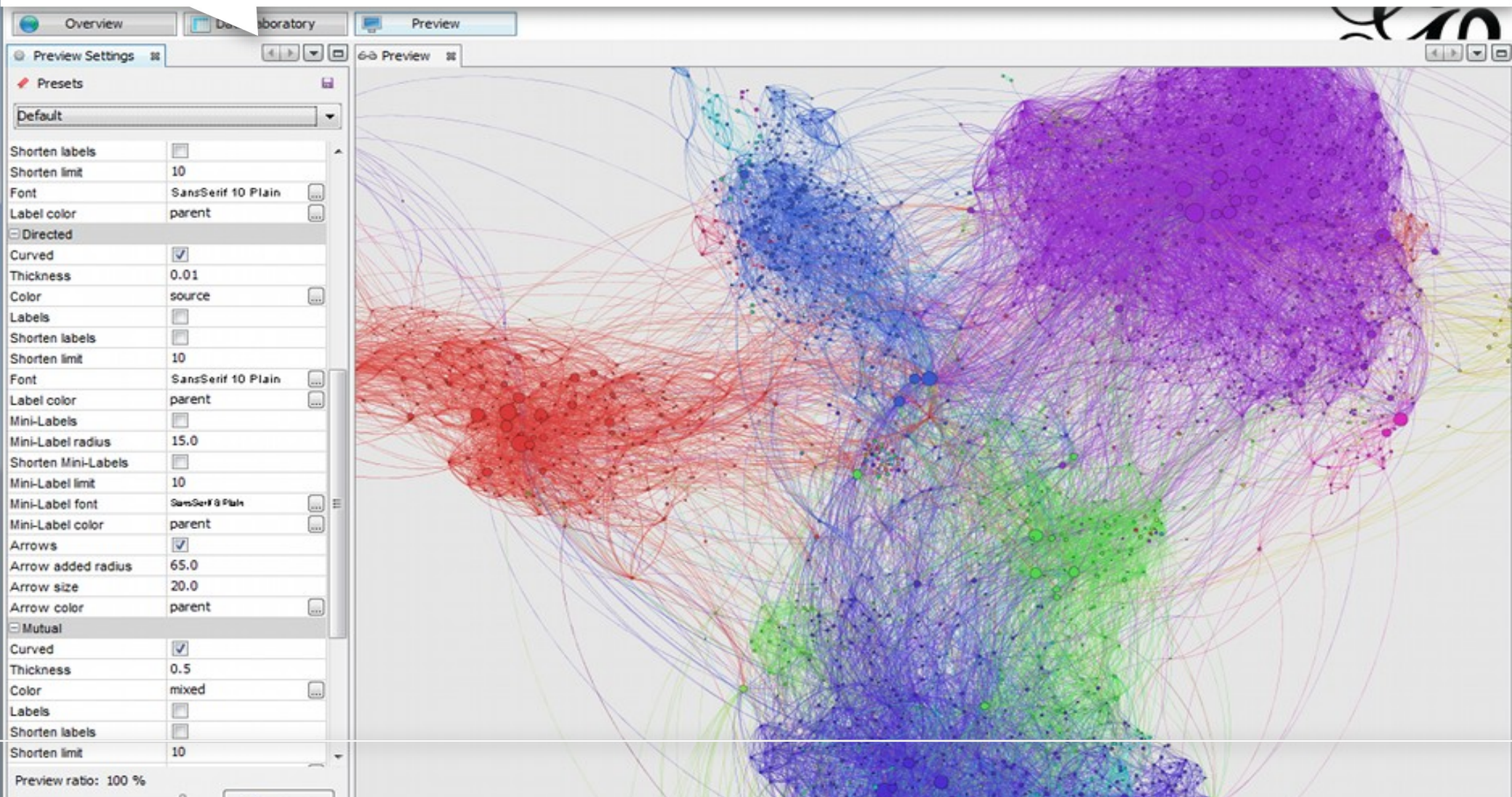
Les réseaux : outils de narration



Le petit chaperon rouge

Les réseaux comme interface

Gephi
gephi.org

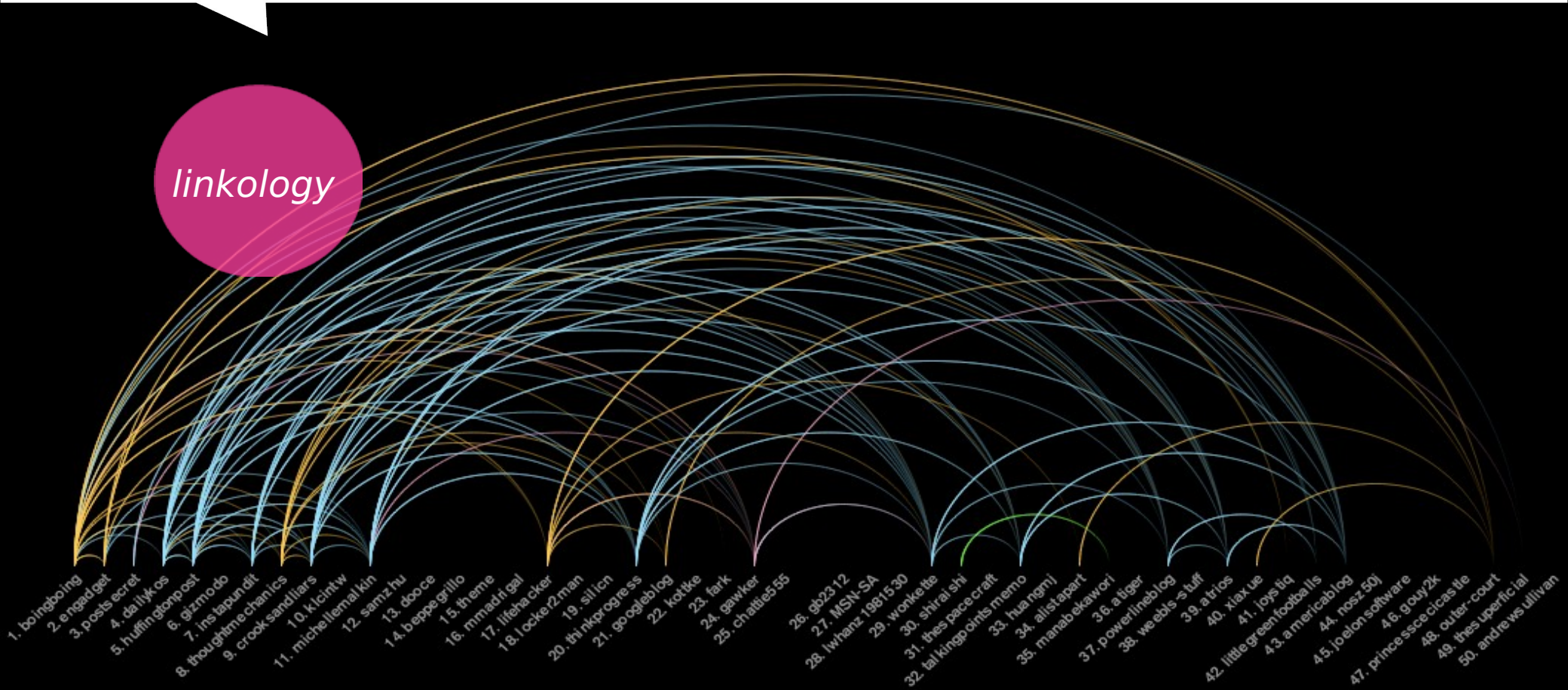


Cartographier le web...

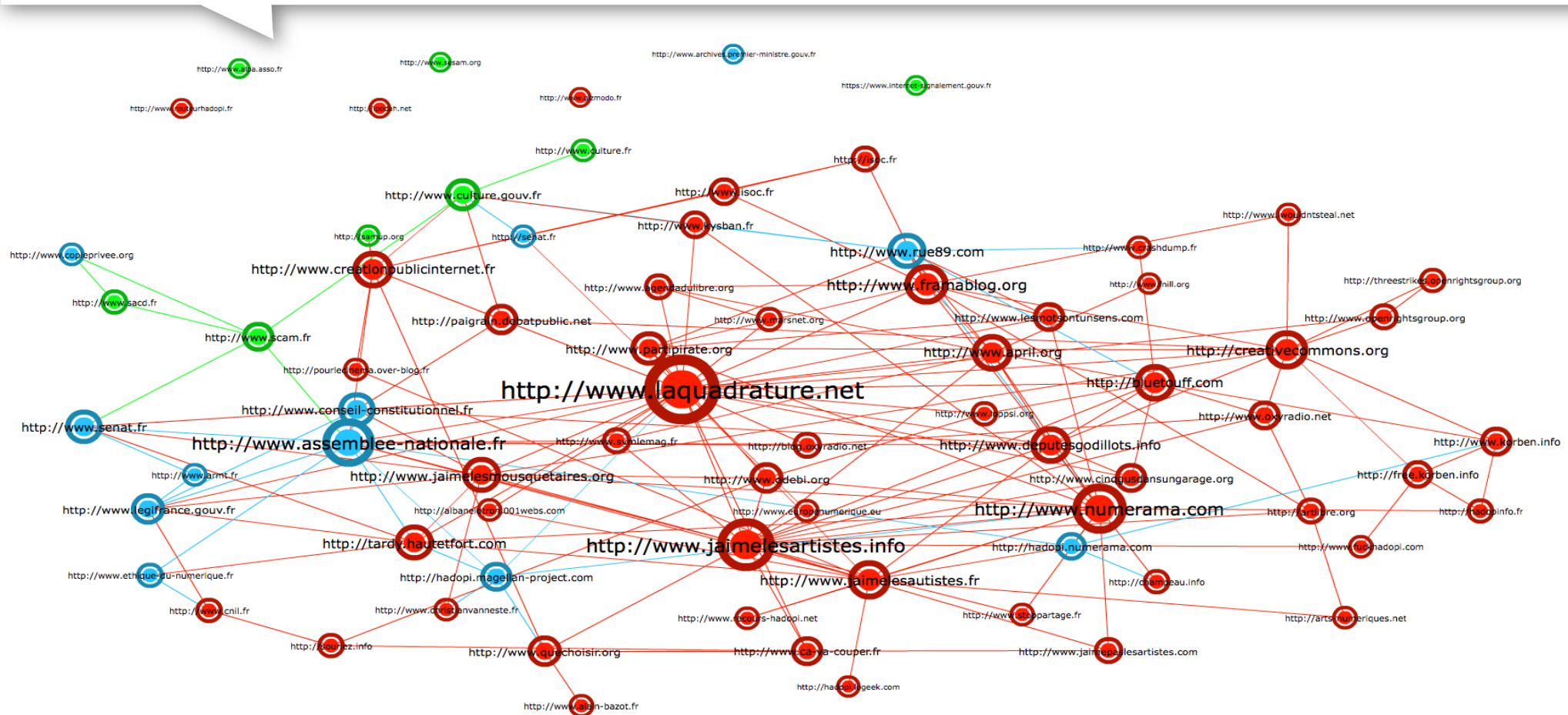
Top 50 US blogs
Ben Fry, 2006

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

linkology



Cartographie des liens hypertextes

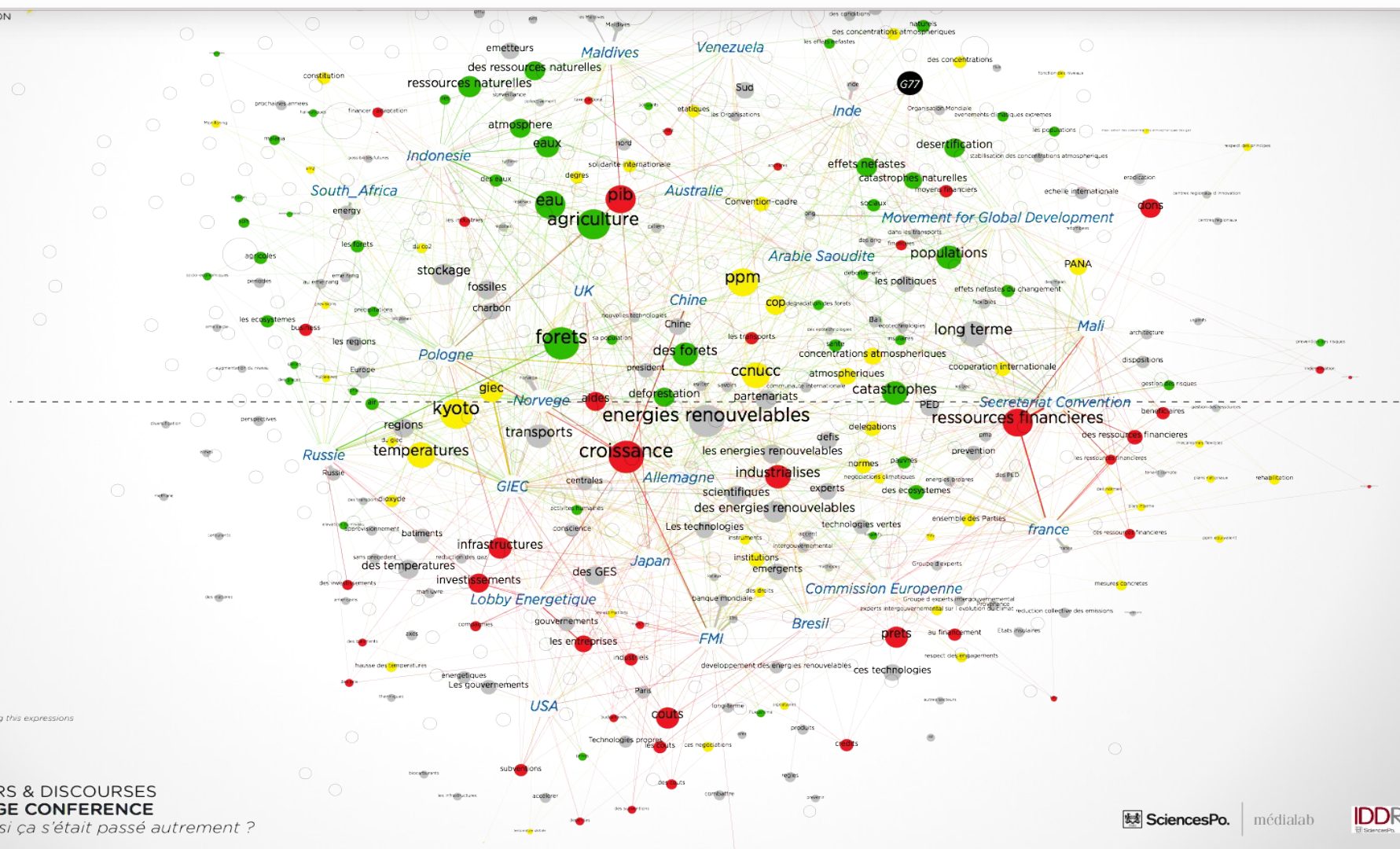


<http://www.medialab.sciences-po.fr/controversies/2010/Hadopi2/index.php?cat=ondaweb&subcat=carto>

Cartographie des liens sémantiques

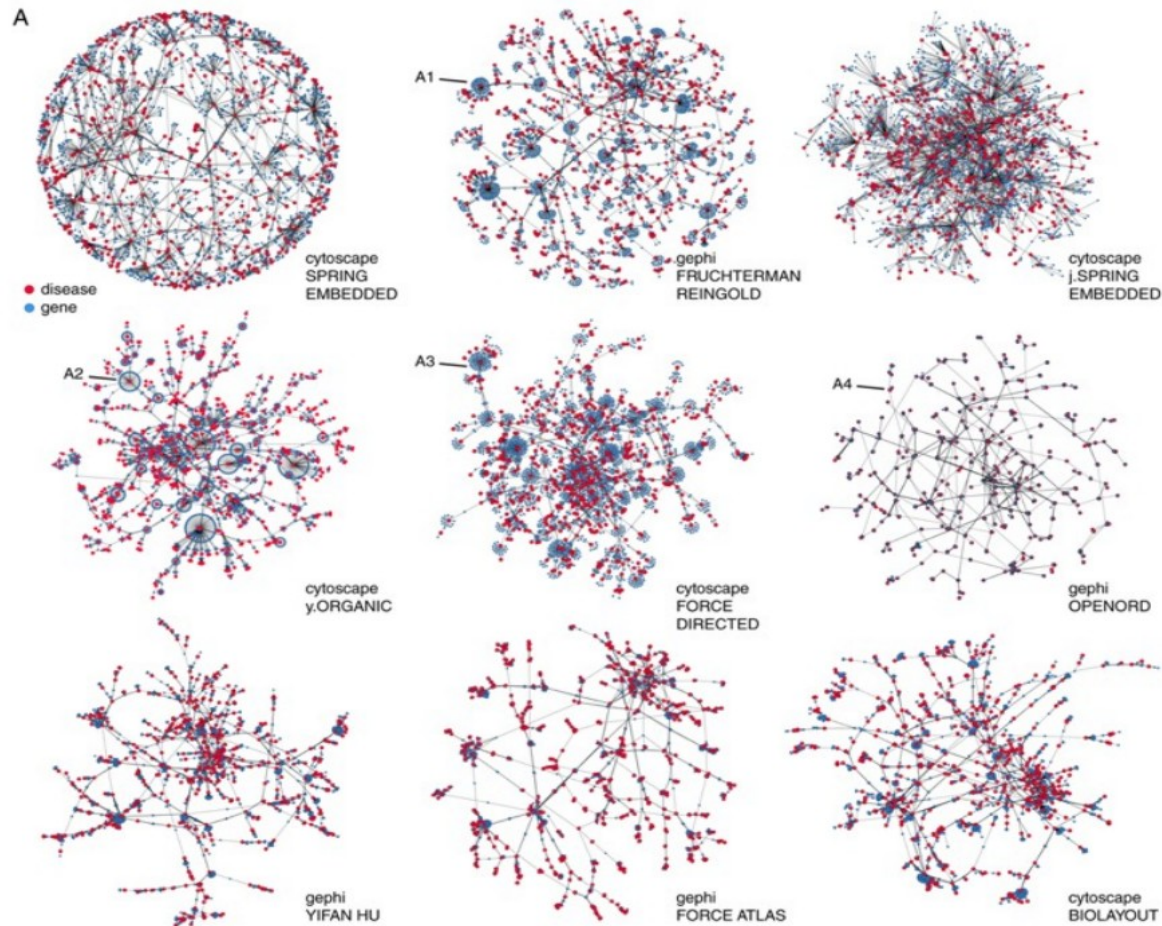
SCIENTIFIC NEGOTIATION

OTHERS



Quel type de spatialisation

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

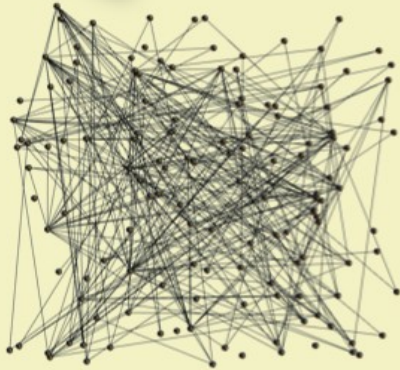


Algorithmes Force-vecteur



Algorithmes Force-vecteur ForceAtlas2

1



2



3



4



5

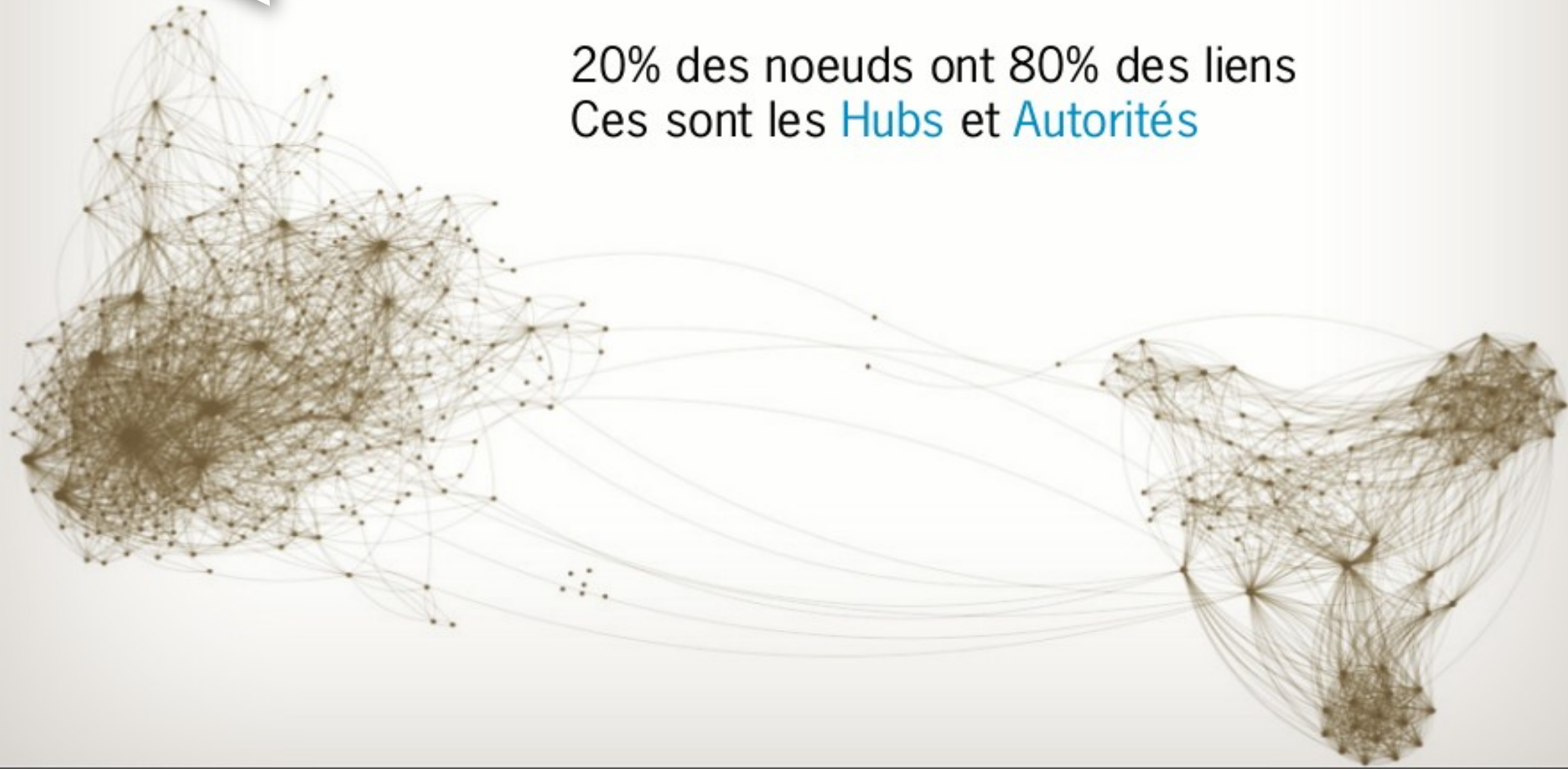


6



Analyse visuelle et clusterisation

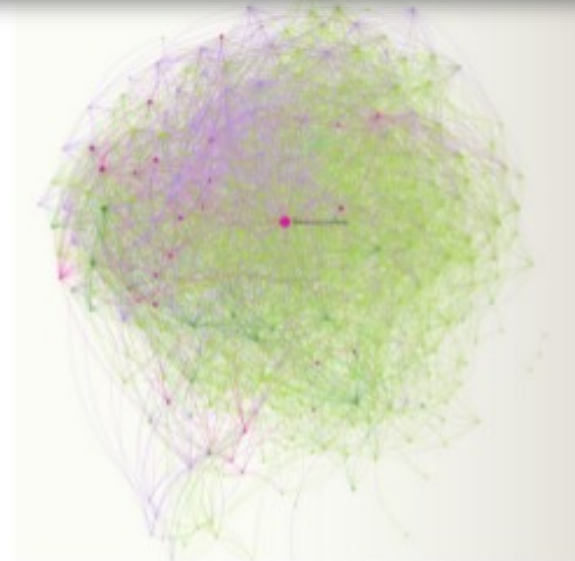
20% des noeuds ont 80% des liens
Ces sont les **Hubs** et **Autorités**



Analyse visuelle et clusterisation

Hubs et Autorités :

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

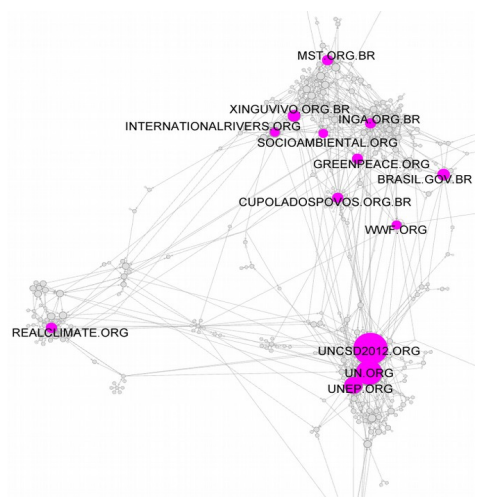
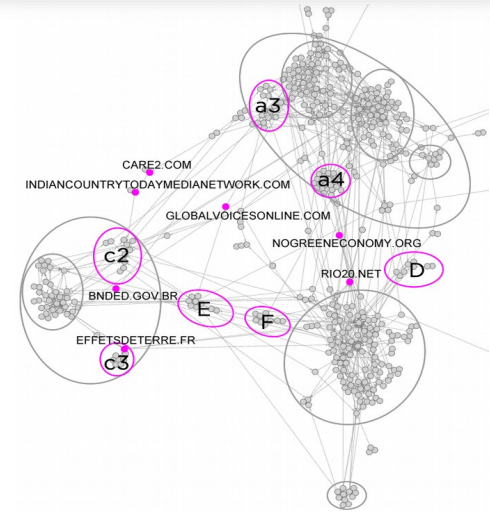
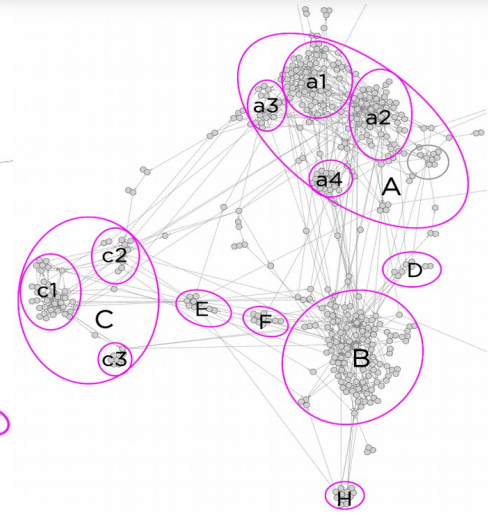
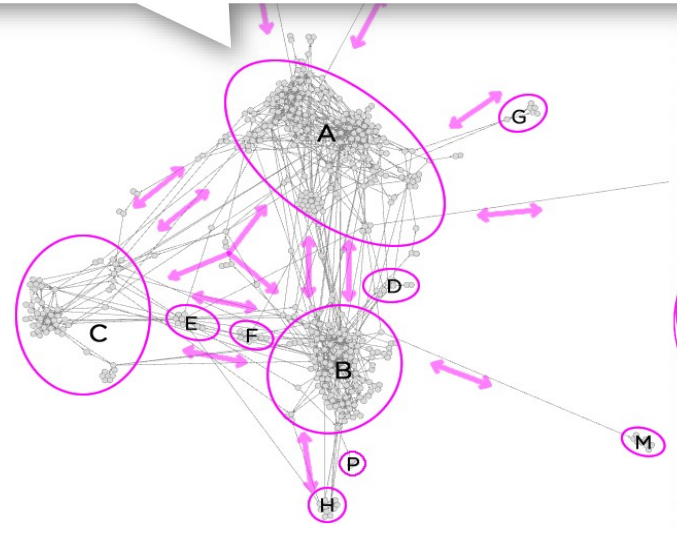


Les clusters contiennent d'autres clusters :

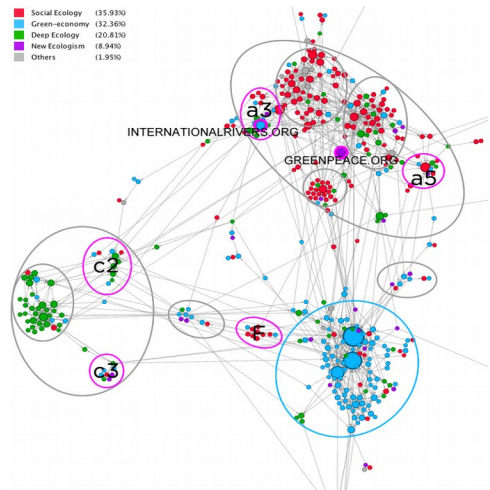
Quels sont les **clusters pertinents** ?
Où sont les bonnes frontières ?



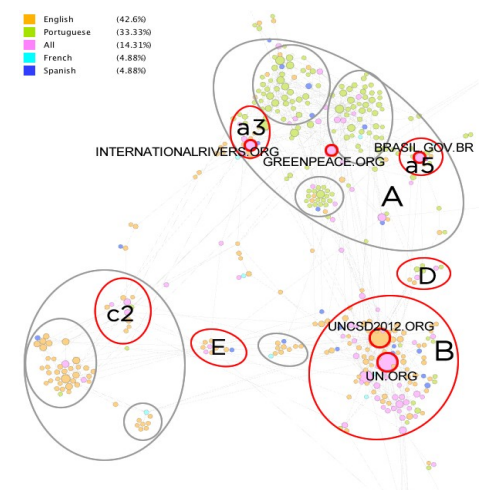
Analyse visuelle et clusterisation



Social Ecology	(35.93%)
Green-economy	(32.36%)
Deep Ecology	(20.81%)
New Ecologism	(8.94%)
Others	(1.95%)



English	(42.6%)
Portuguese	(33.3%)
All	(14.1%)
French	(4.88%)
Spanish	(4.88%)



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!](#)

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

Hyphe

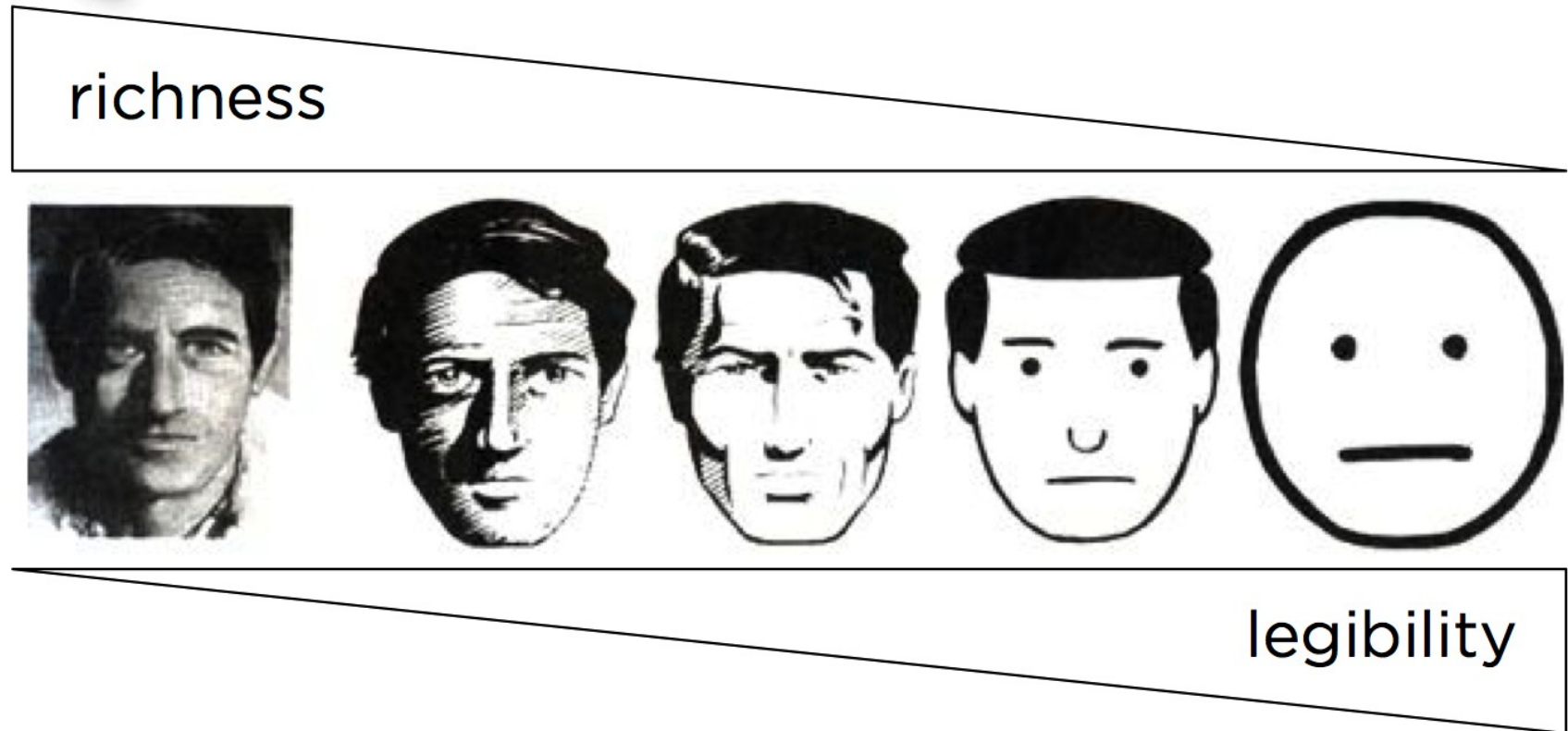
Hyphe is a webcrawler made to help scientists



Crawl and explore a web corpus. progress. Help us by reporting your

Appréhender la donnée

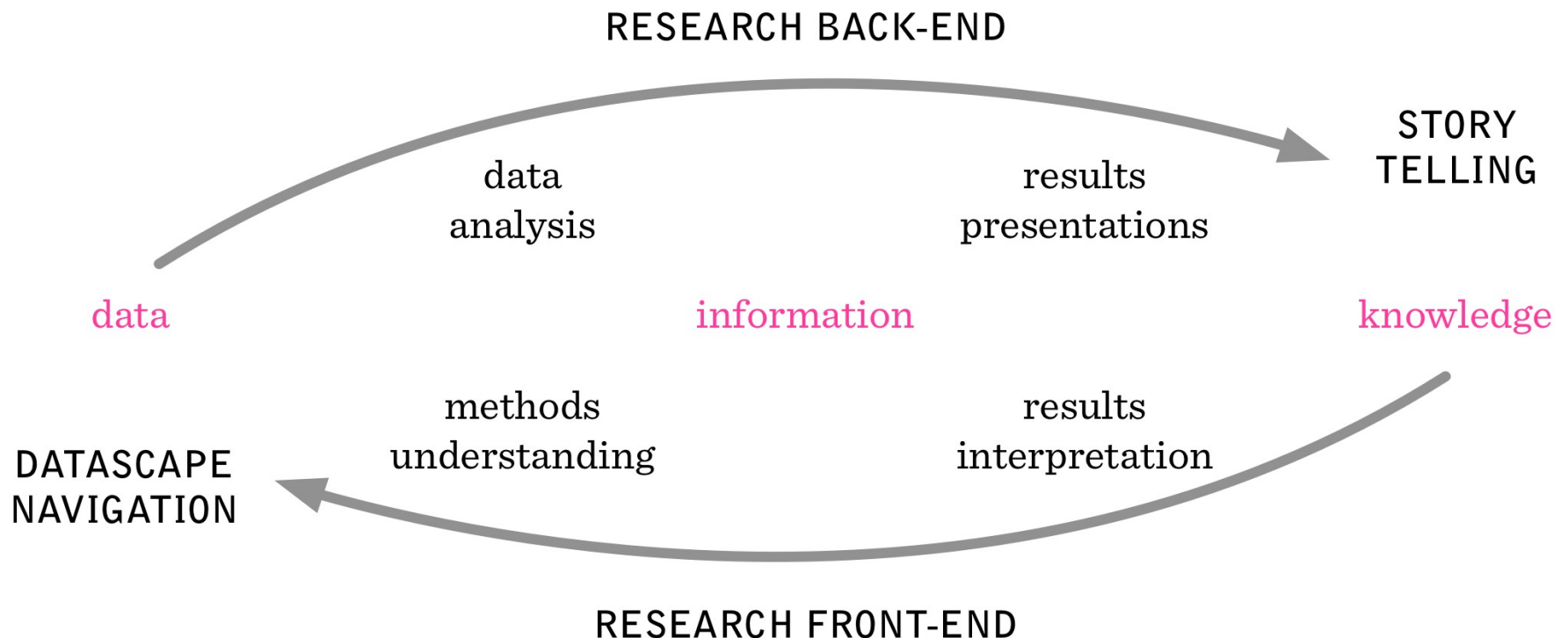
Understanding Comics
Scott McCloud (1993)



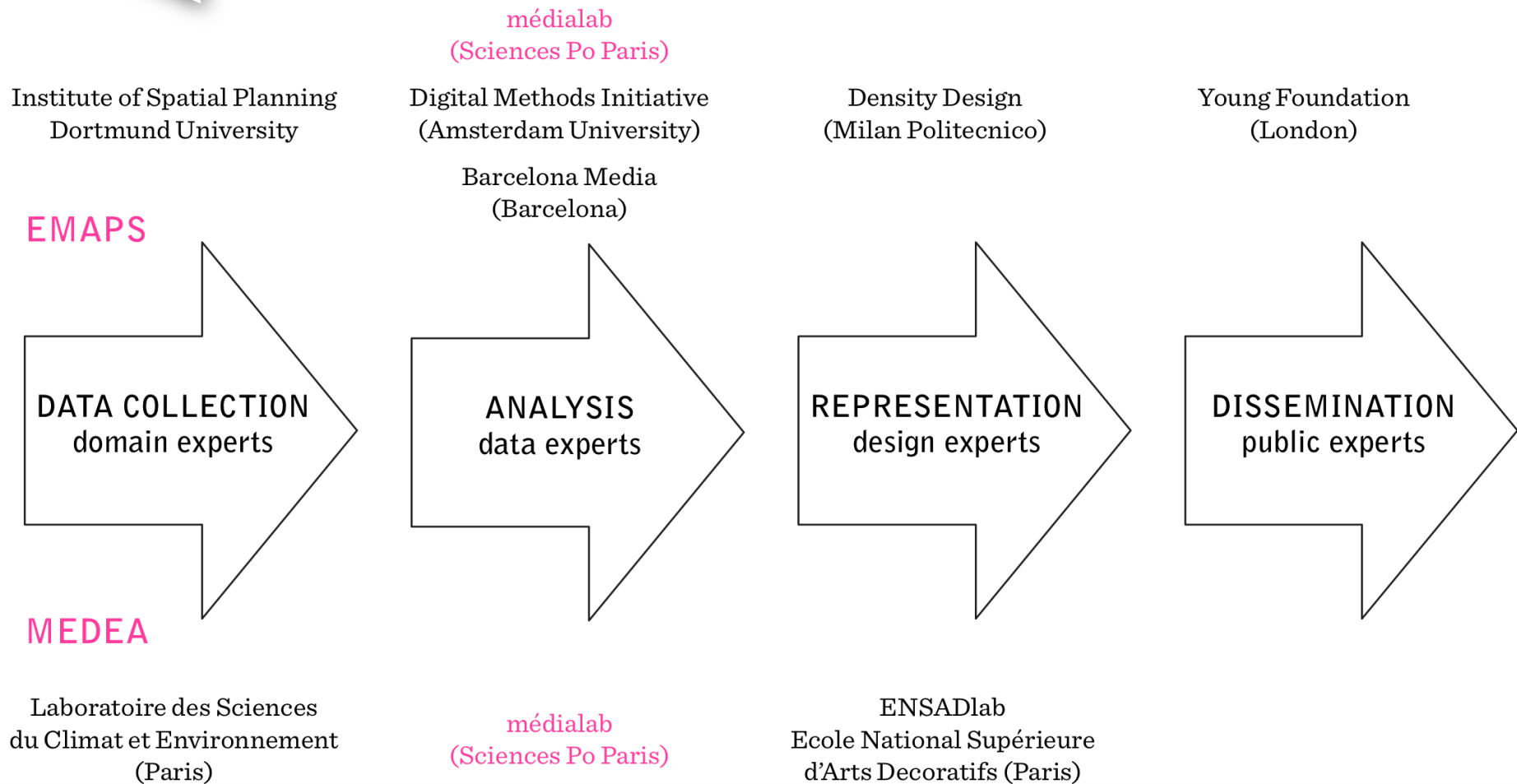
knowledge → information ← data

Le cycle d'exploration - narration

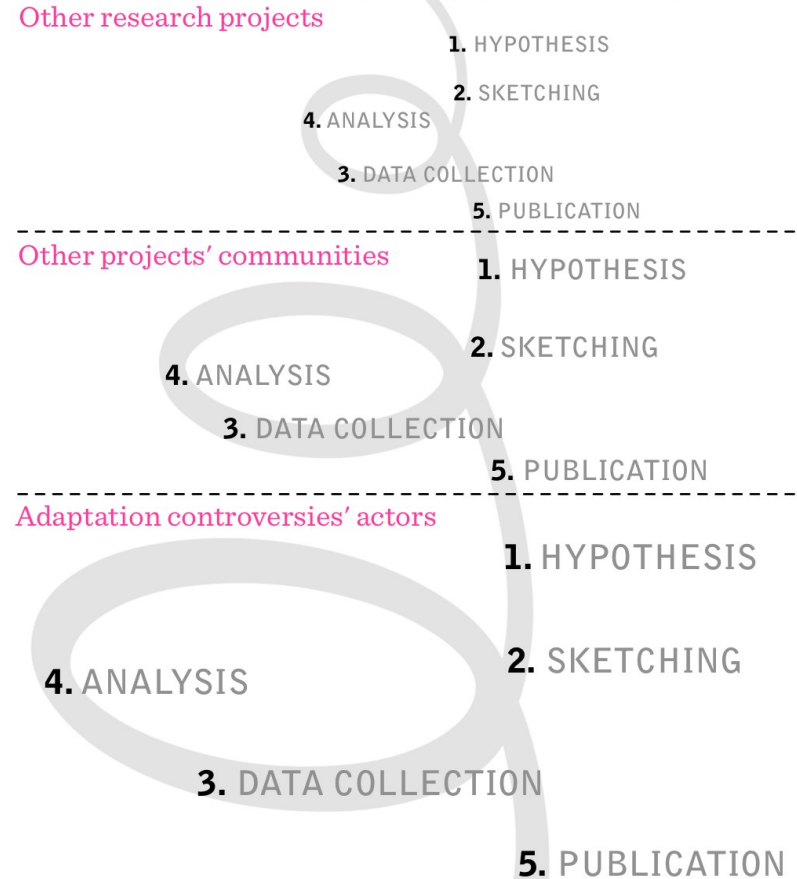
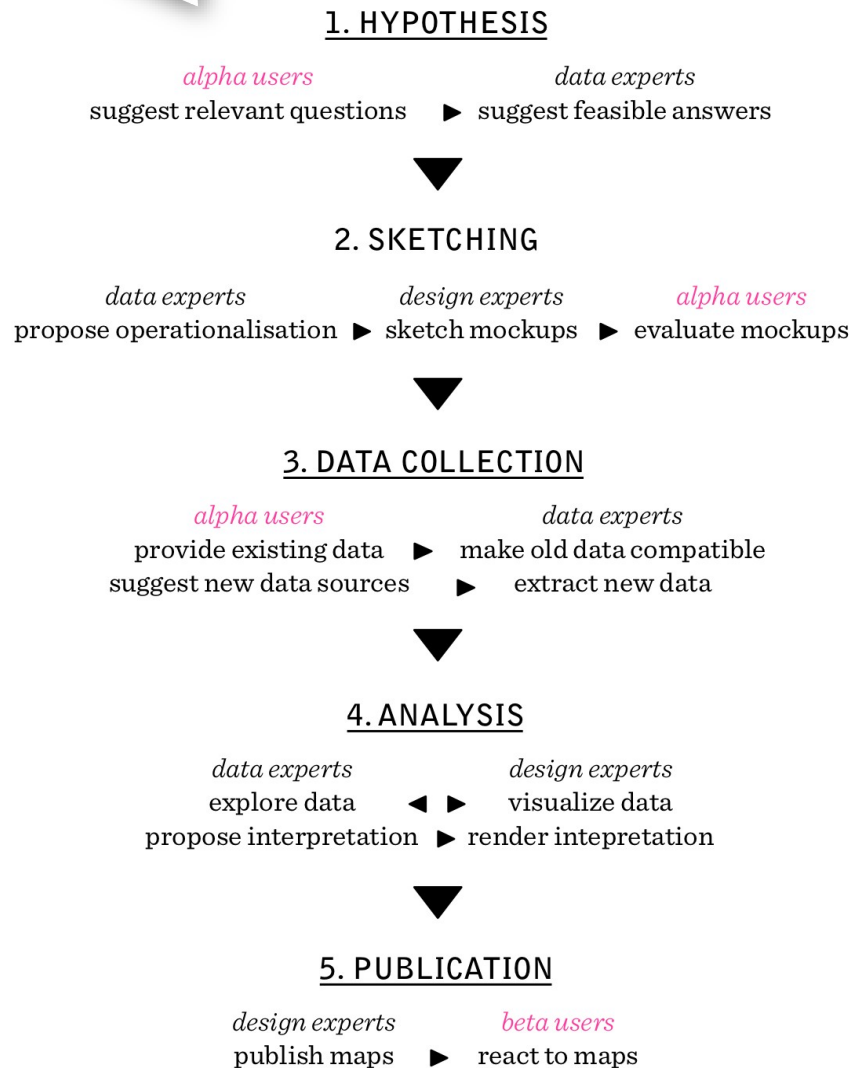
“An essential property of this chain is that it must remain reversible. The succession of stages must be traceable, allowing to travel in both directions” (Latour, 1995)



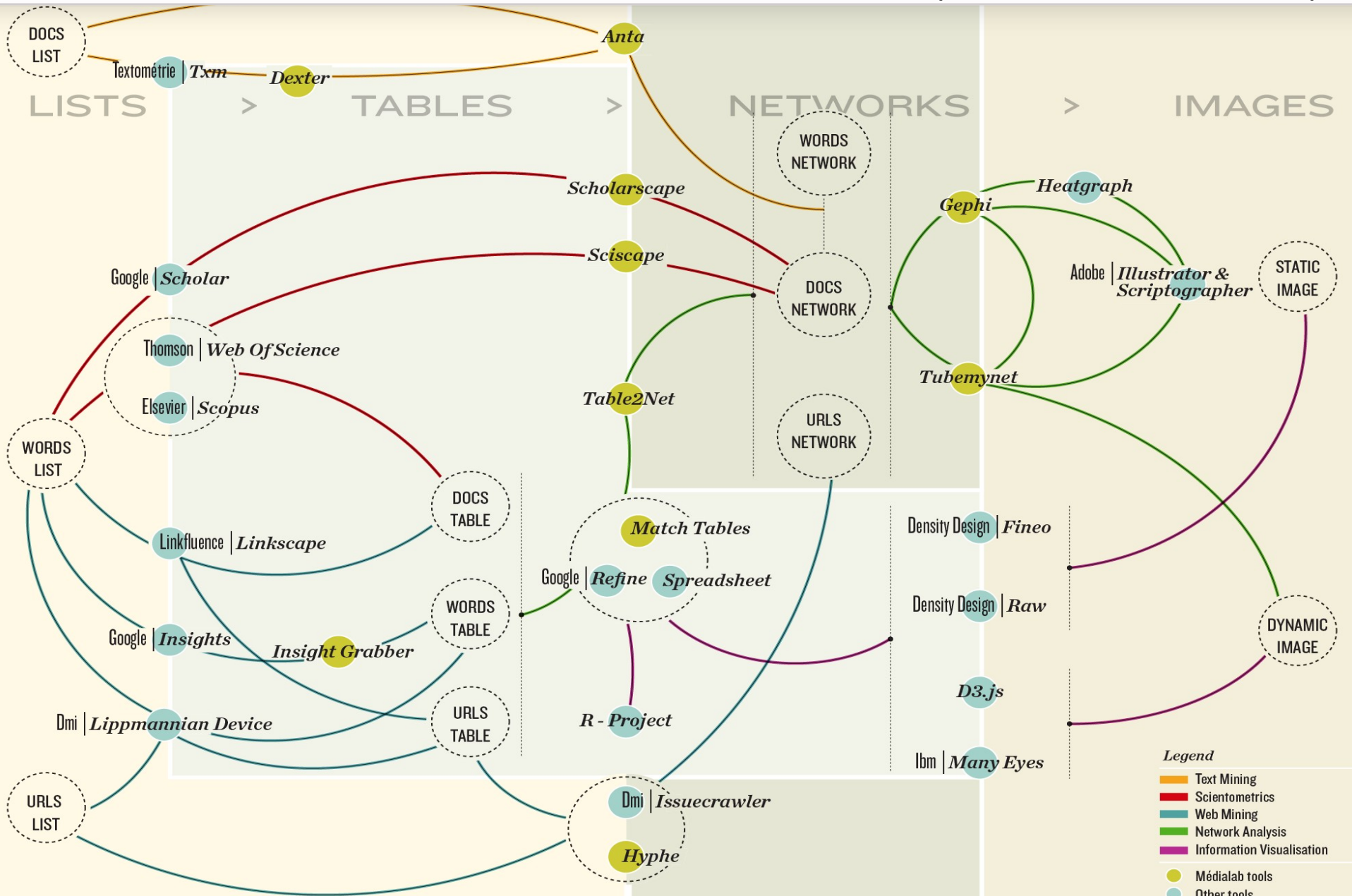
Le protocole attendu



La spirale d'interaction



Une chaîne d'outils



ScienceScape : la scientométrie simplifiée



ScienceScape

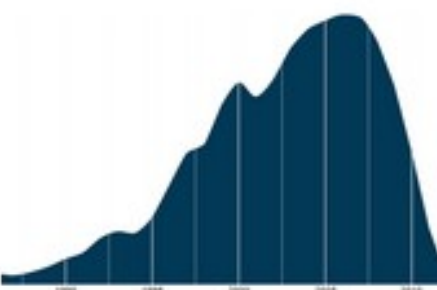
Helpers for scientometrics. Convert files, get networks, visualize stuff from Scopus or Web of Knowledge.



Get Networks

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

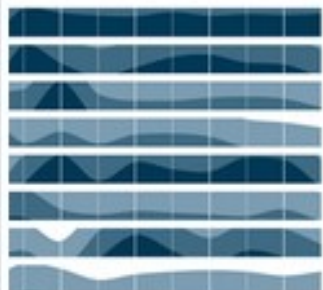
Scopus Web of Knowledge



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

2008	2009	2010
<ul style="list-style-type: none"> climate change 317 papers adaptation 103 papers global warming 47 papers vulnerability 31 papers resilience 34 papers drought 26 papers mitigation 22 papers temperature 22 papers global change 17 papers sustainability 17 papers 	<ul style="list-style-type: none"> climate change 473 papers adaptation 160 papers global warming 34 papers vulnerability 31 papers resilience 42 papers temperature 25 papers drought 25 papers mitigation 25 papers agriculture 22 papers climate 22 papers 	<ul style="list-style-type: none"> climate adapt mitig temp agri agri agri agri agri agri
2012	2013	
<ul style="list-style-type: none"> climate change 357 papers adaptation 107 papers vulnerability 31 papers resilience 35 papers global warming 32 papers 	<ul style="list-style-type: none"> climate change 334 papers adaptation 82 papers vulnerability 31 papers climate change adaptation 3 papers climate 3 papers 	

Top keywords / year

Visualize the most used keywords each year in your file

Scopus Web of Knowledge



climatic ch
global char

1995

- energy policy 7 papers
- climate change and agriculture: analysis of

1996

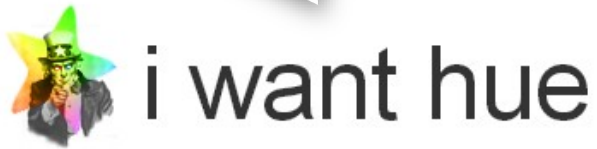
- adapting to climate change: perspective 20 papers



1. Upload your Scopus CSV file

Choisissez un fichier: Aucun fichier choisi


I Want Hue : du choix des couleurs





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

Color space

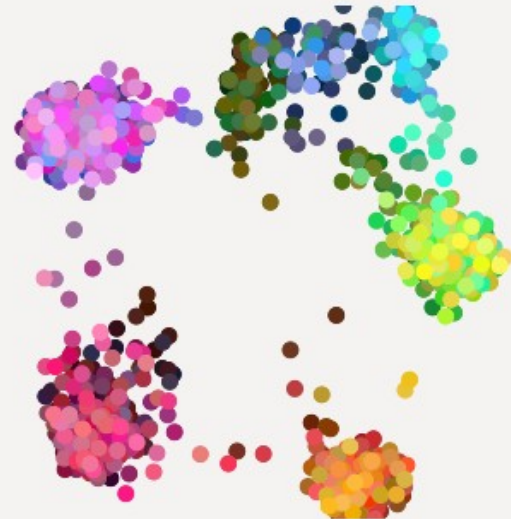
Presets... ▾

H 0  360

C 0.97  3


L 0  1.5

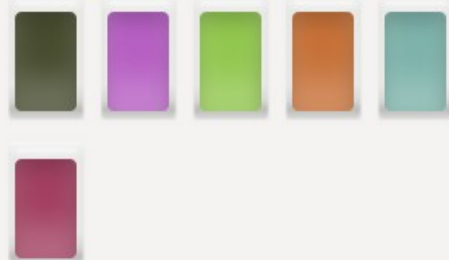
Dark background



Palette

6 colors soft (k-Means) ▾





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	textelol_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-dhucq patrice-verchère 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. Dhucq, M. Verchère, M. Terrot, M. Decool, M. Lazaro, M. Abad,	http://www.assemblee-nationale.fr/14/amendement

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)

Hyphe (Hypertext Corpus Initiative)

Hyphe

1 - Crawl ▾

2 - Classify

3 - Qualify

4 - Explore ▾

CSV ▾

Other ▾

GitHub

Issues

+ Médialab Tools



Crawl and explore a web corpus. Hyphe is still a work in progress. Help us by [reporting your issues](#).

Welcome

Hyphe does not manage different corpora or users at the moment. All the data is stored as a single corpus summarized here.

Status

49019 web entities

50163 pages crawled

No crawl scheduled

Last memory activity 4 hours ago

Last content indexation 2 days ago

Last link built 3 hours ago

Tasks

1. [Crawl](#)
2. [Classify discovered web entities](#)
3. [Qualify](#)
4. [Network of web entities](#)

Working with a CSV

- [Define web entities \(csv\)](#)
- [Diagnostic \(csv\)](#)

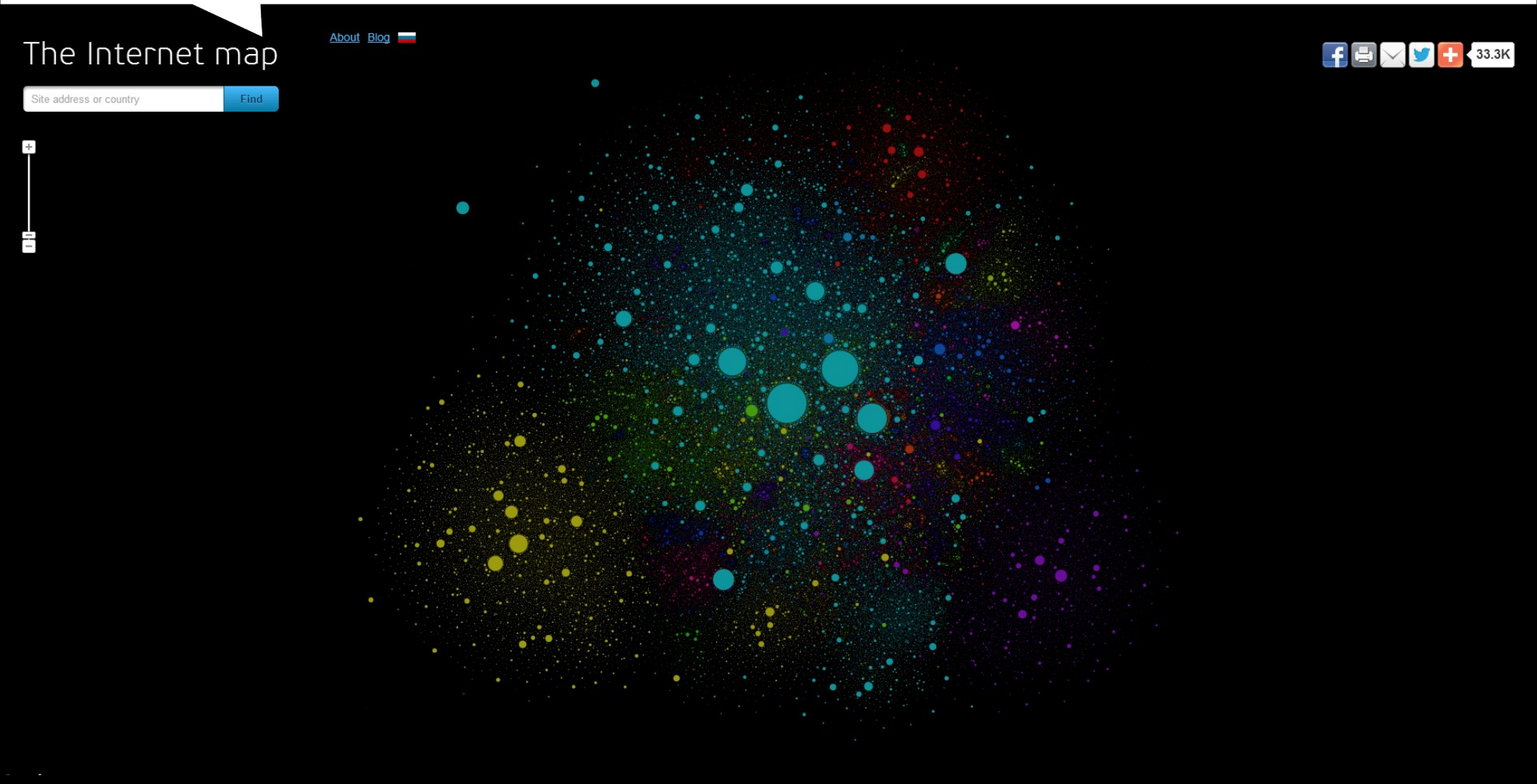
Monitoring

- [List of web entities](#)

Administration

- [Reset all](#)

À quoi ressemble le web ?



Outils existants pour construire un corpus web



WebAtlas Navicrawler

Navicrawler :
créer son corpus manuellement

issuercrawler

Issuercrawler :
créer son corpus automatiquement

Deux outils non adaptés



une pince à épiler

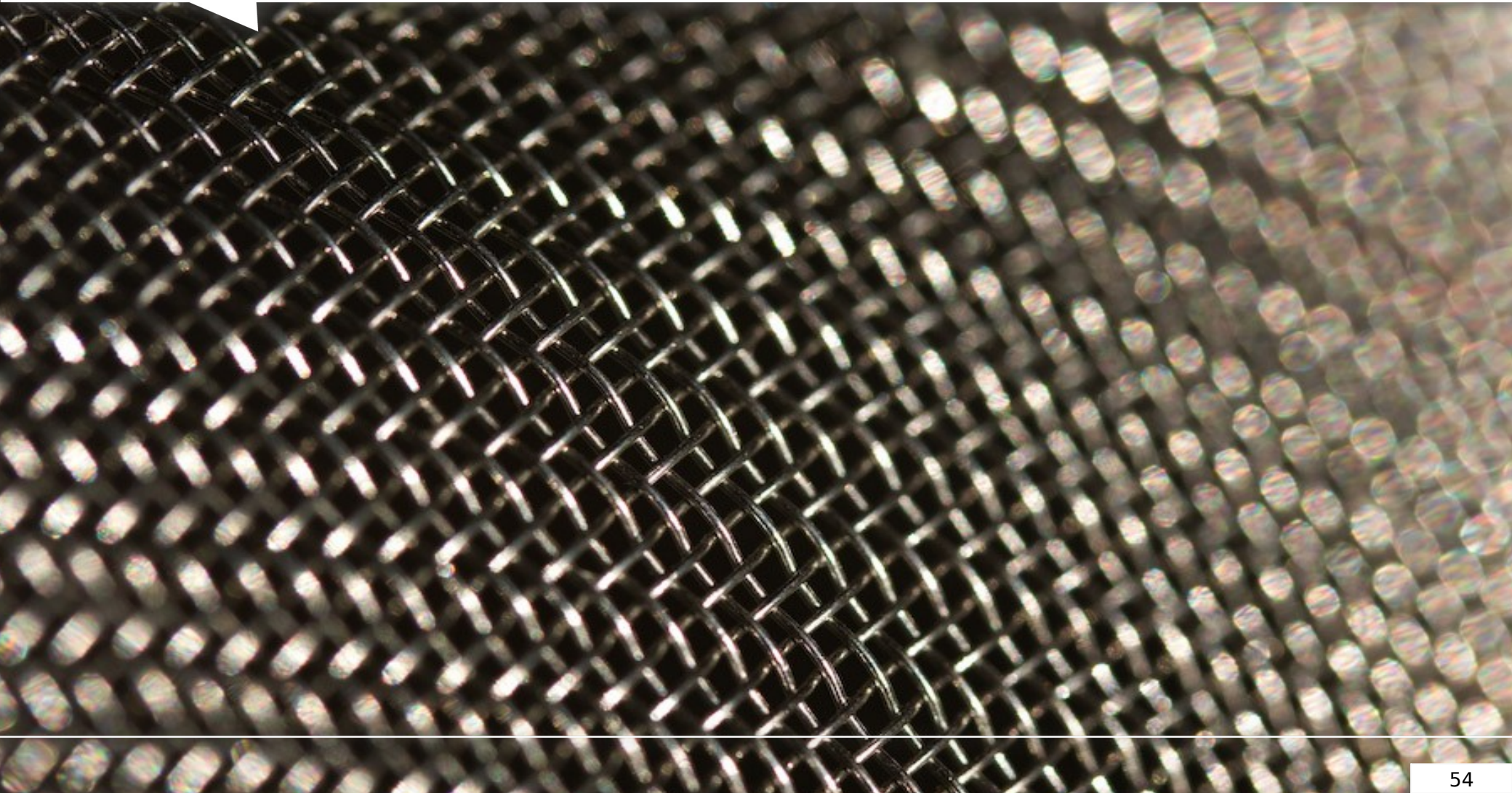
ou

un bulldozer

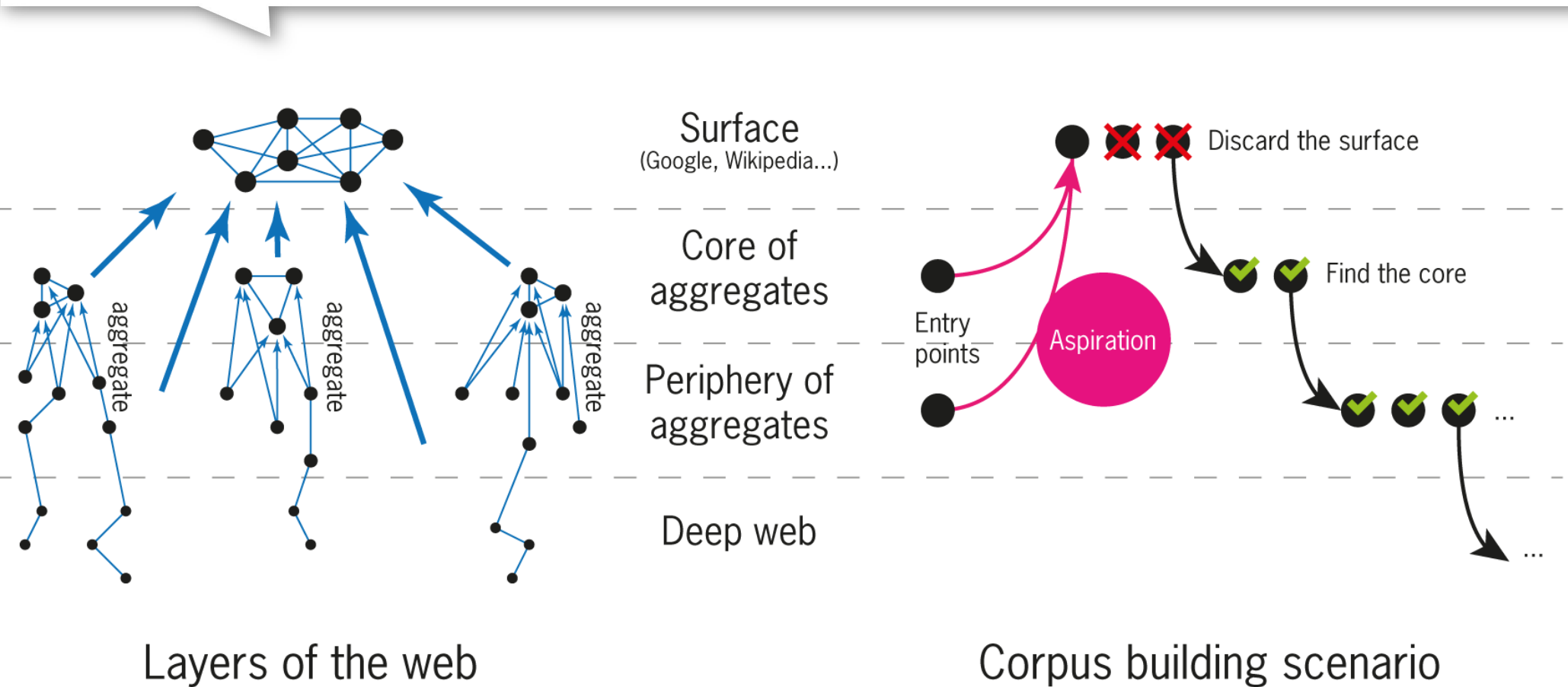


Comment tamiser le web ?

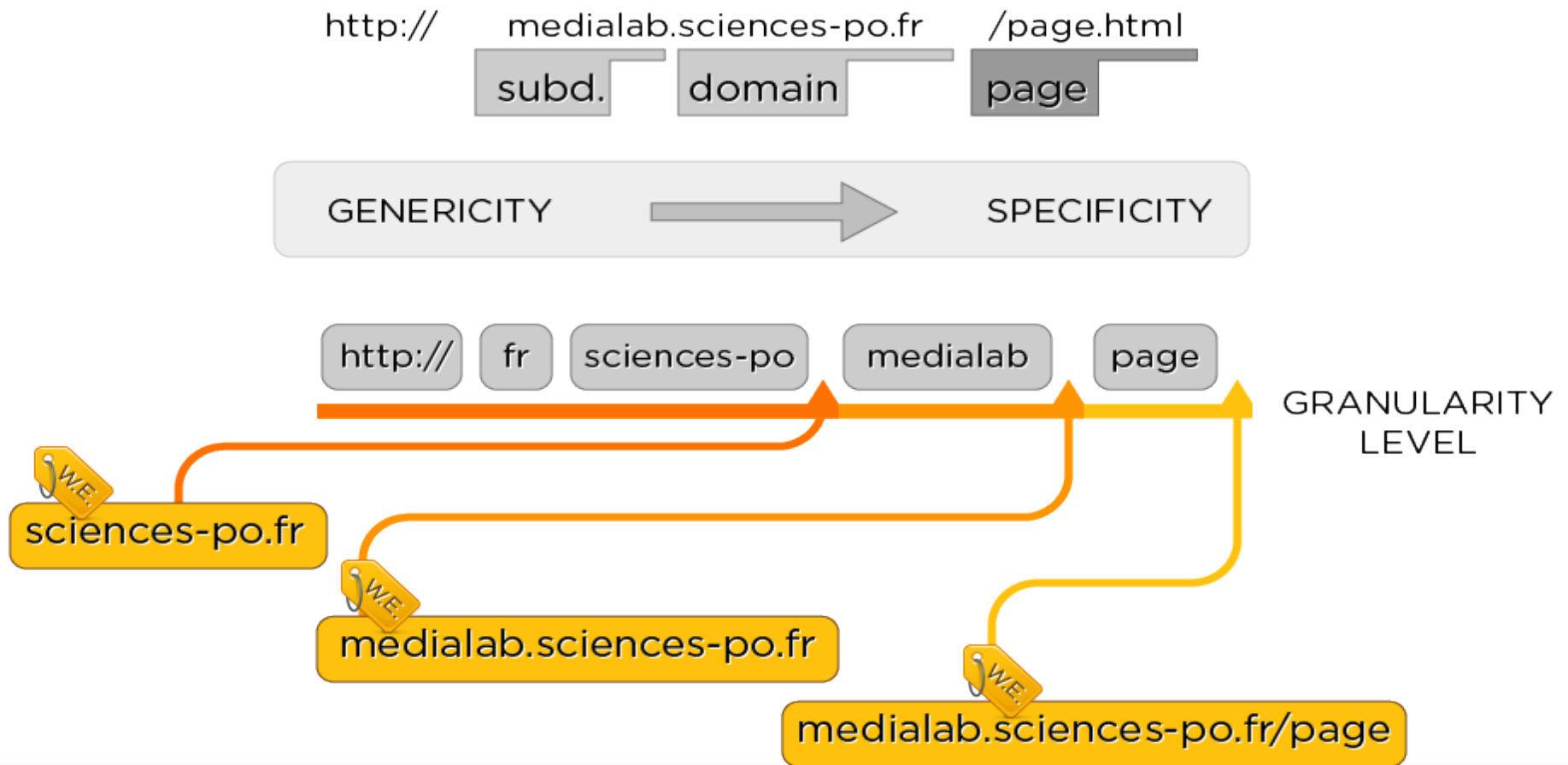
© *Mayhem Chaos*



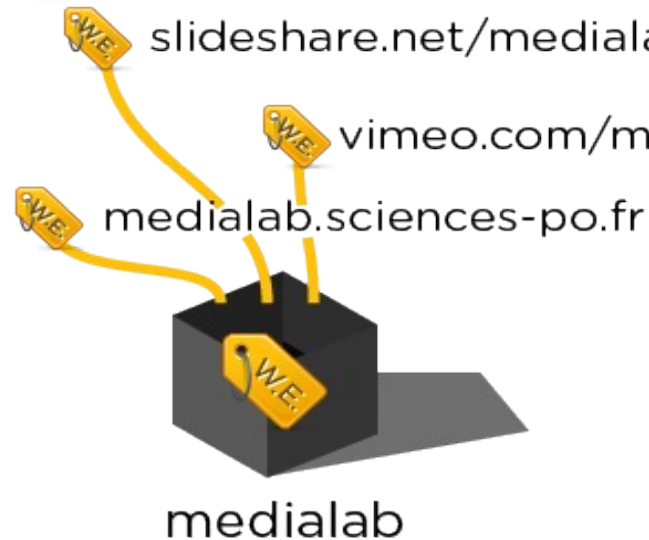
Les couches du web



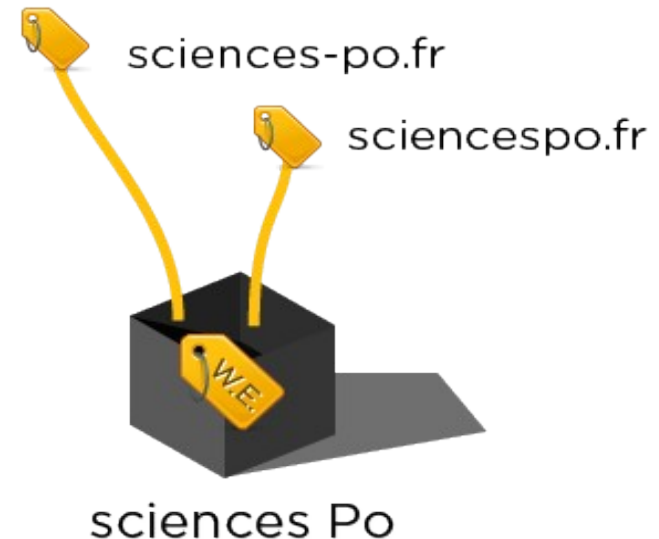
Définir des points d'ancrage précis (LRUs)



Des sites ou... des entités web

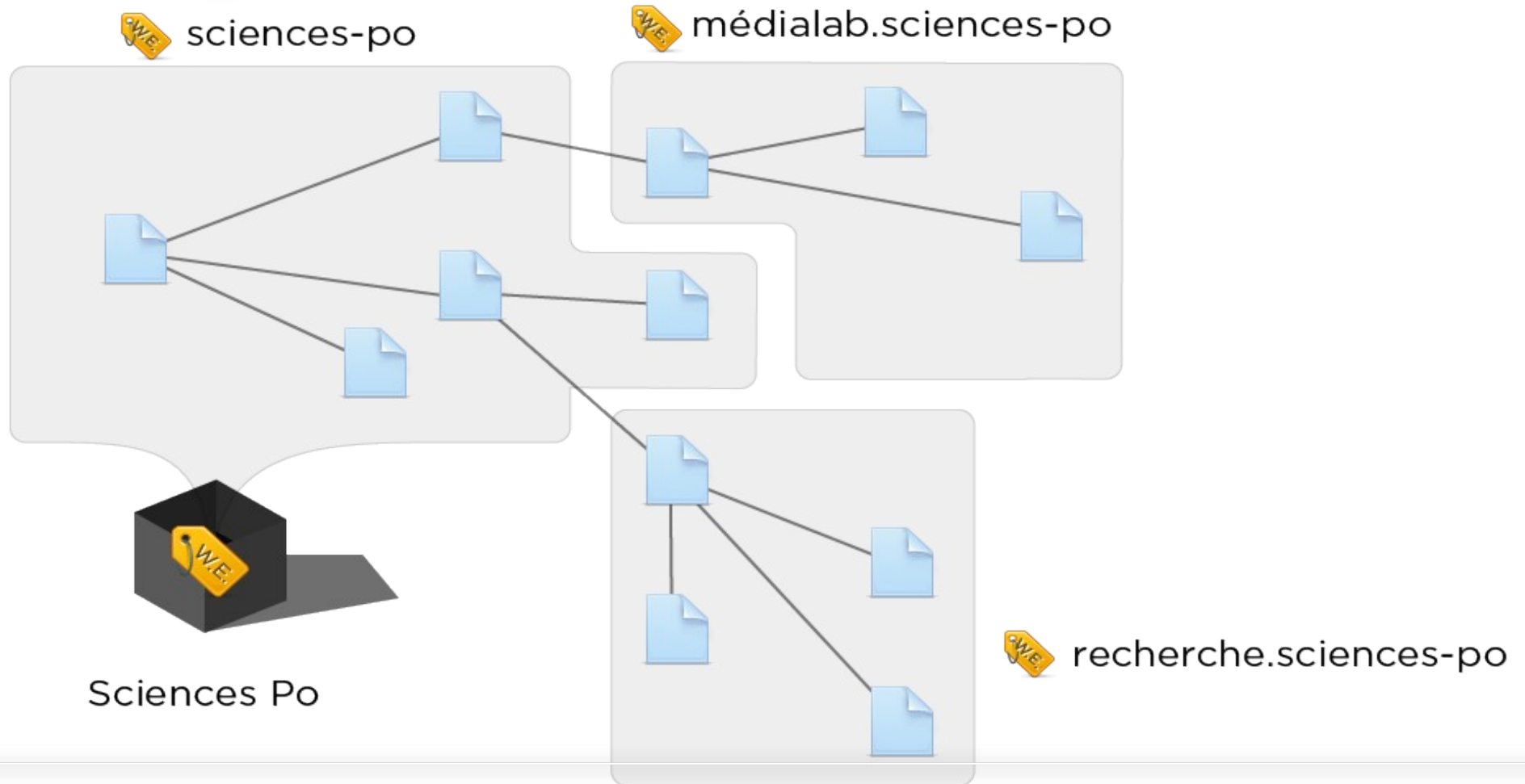


ACTOR'S PRESENCE
ON THE WEB



ALIASES

Préserver la complexité

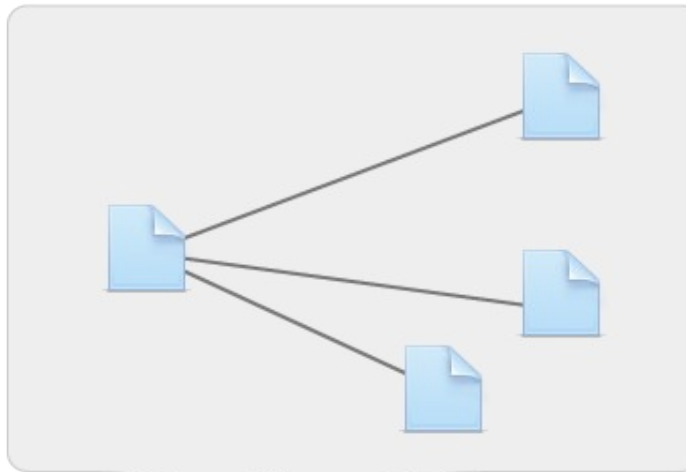


Le crawl dirigé par la recherche

1

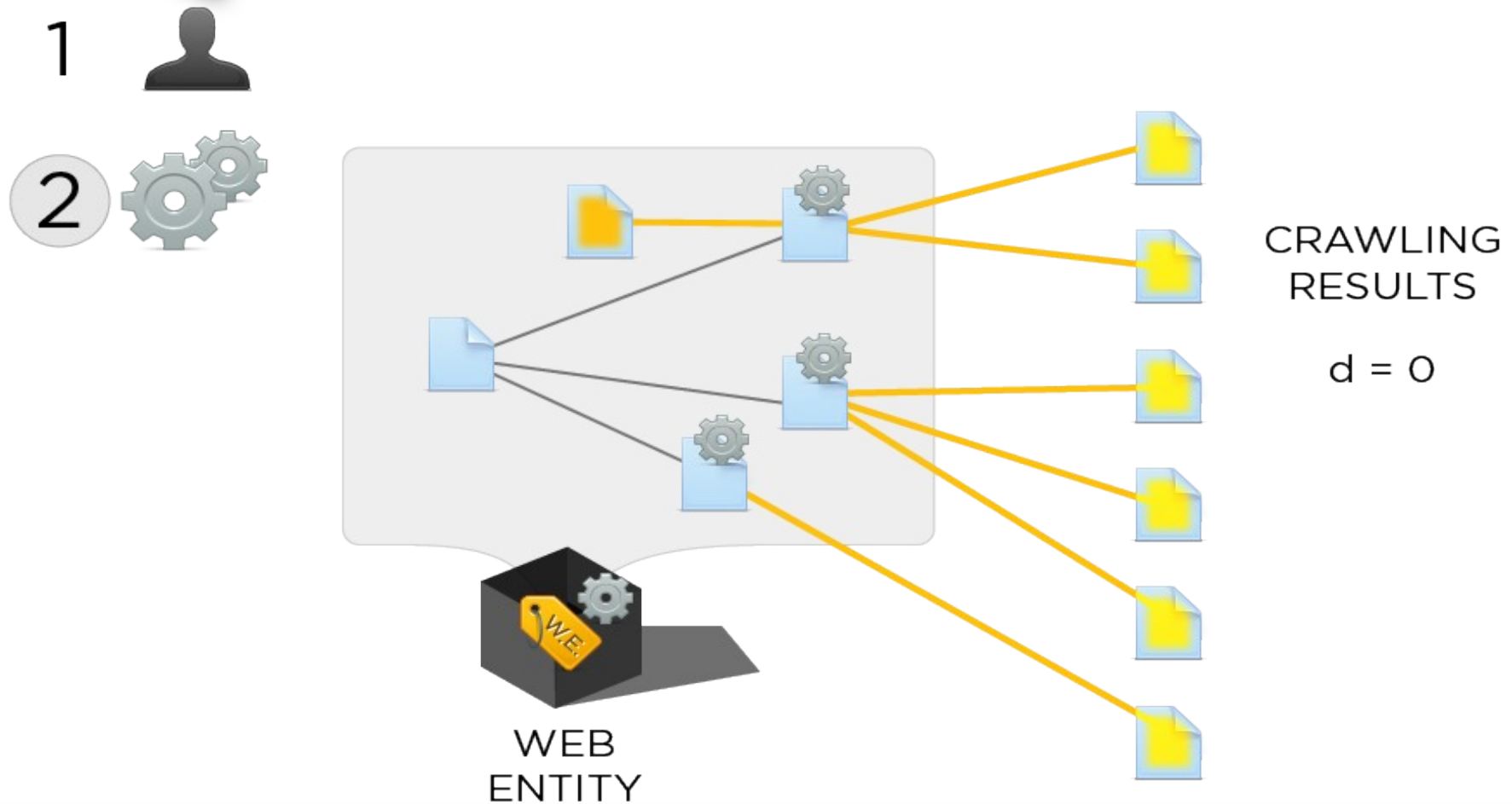


RESEARCHER
SELECT
STARTING
ENTITIES

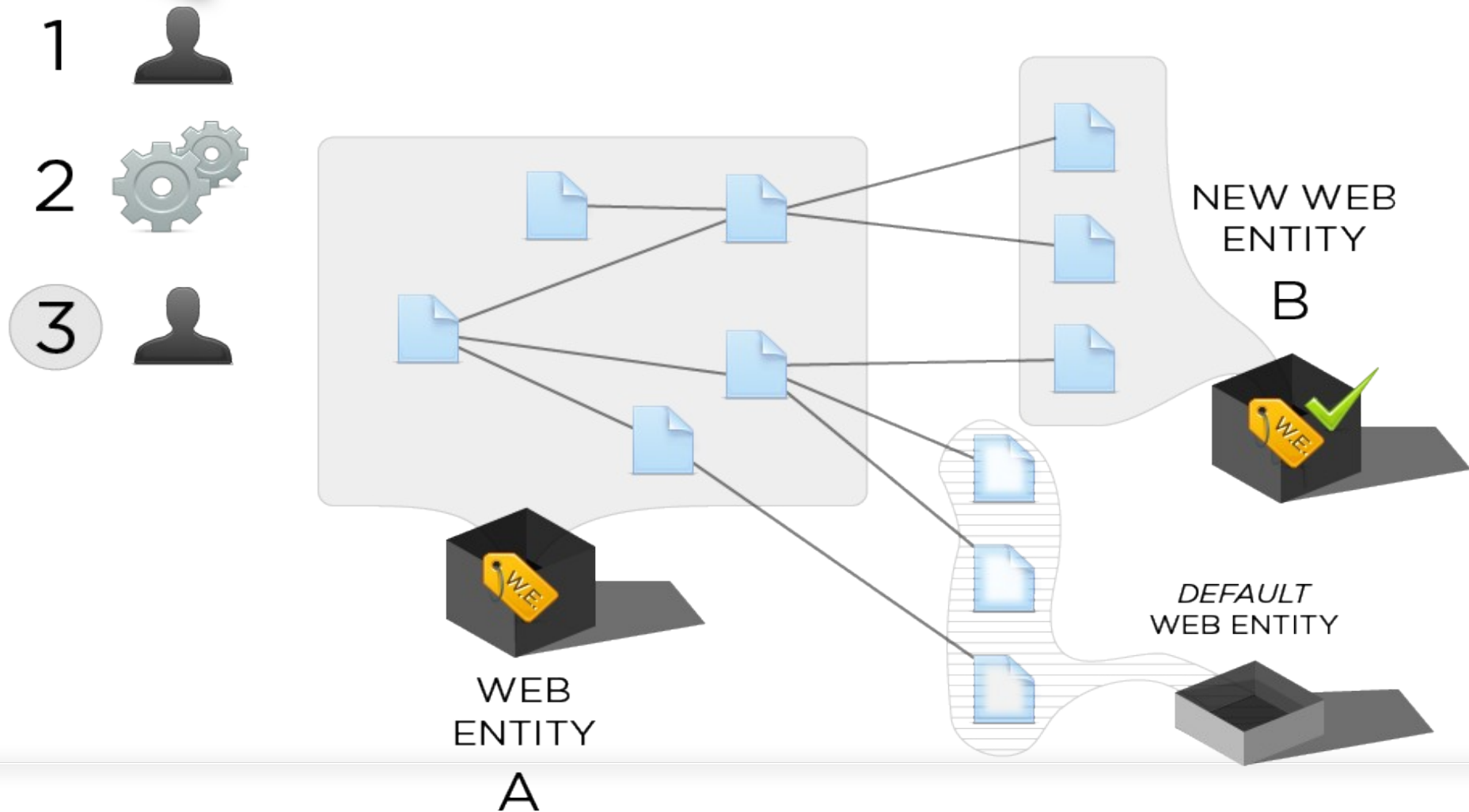


WEB
ENTITY

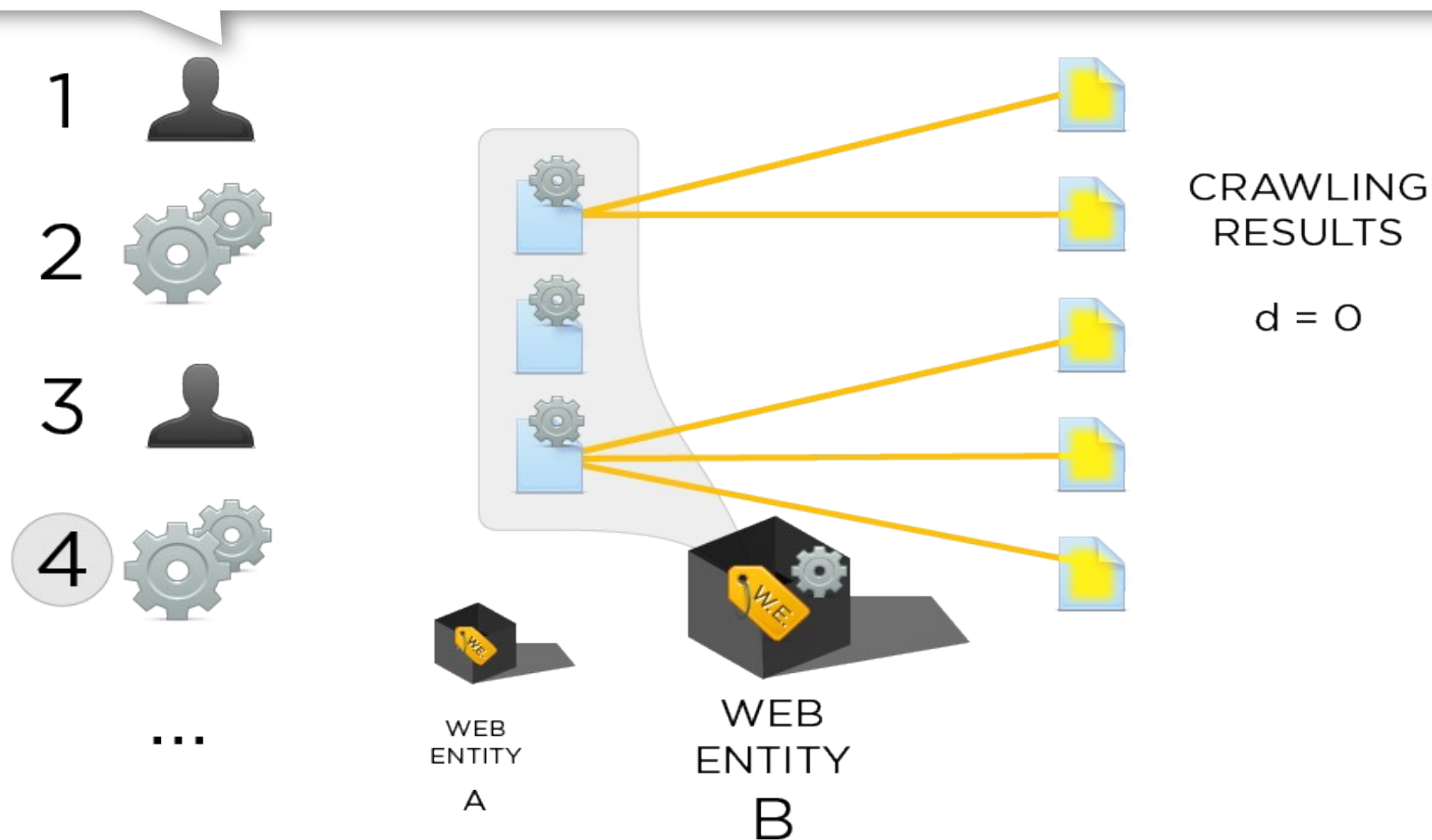
Le crawl dirigé par la recherche



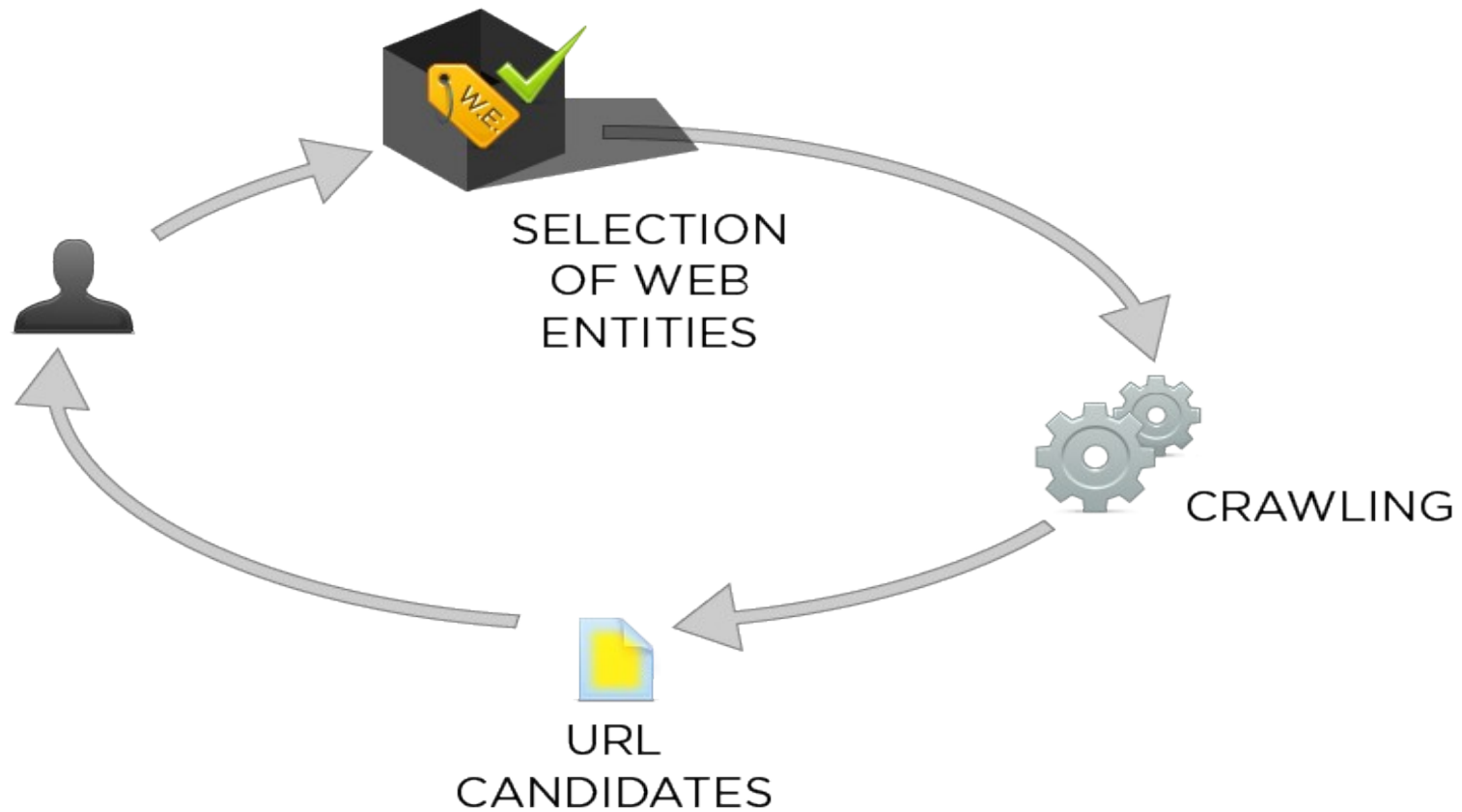
Le crawl dirigé par la recherche



Le crawl dirigé par la recherche



Le crawl dirigé par la recherche



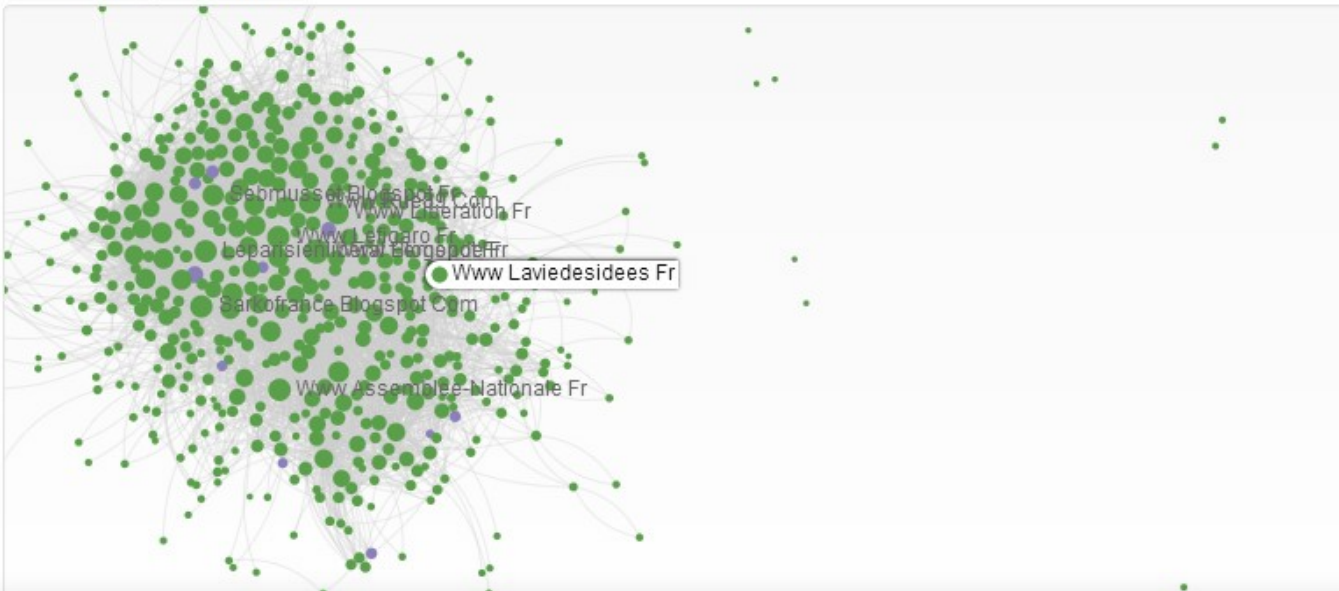
Exemple : SITPOL

Network of webentities

Do shit n' stuff

Preview

▶ Start layout ↗ Reset zoom



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