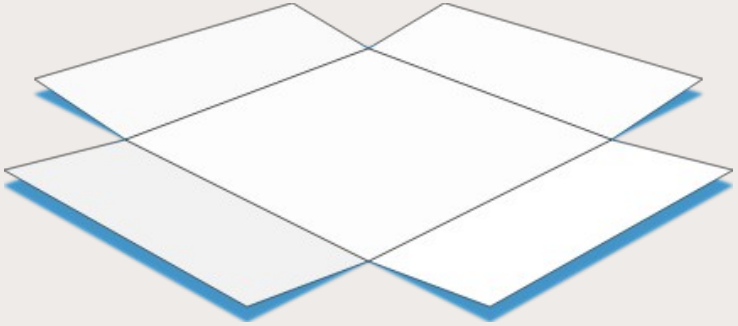
Hyphe : utiliser le Web comme terrain d'enquête

Benjamin Ooghe, Maxime Crépel



hyphe

Utiliser le Web comme terrain d'enquête

Ateliers DATA SHS

MESHS Lille – 15 décembre 2017

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

DIME SHS Web

SciencesPo MÉDIALAB

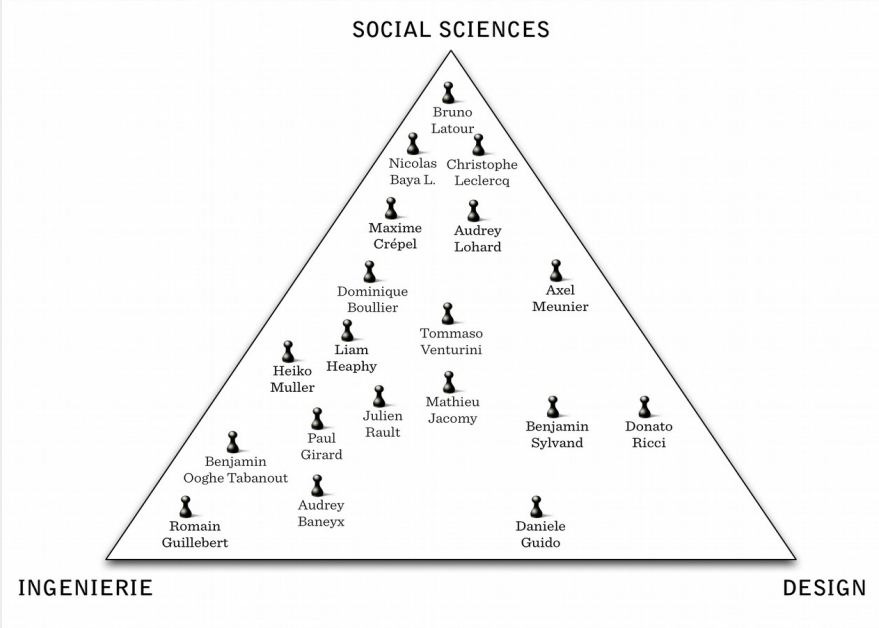
INVESTISSEMENTS D'AVENIR

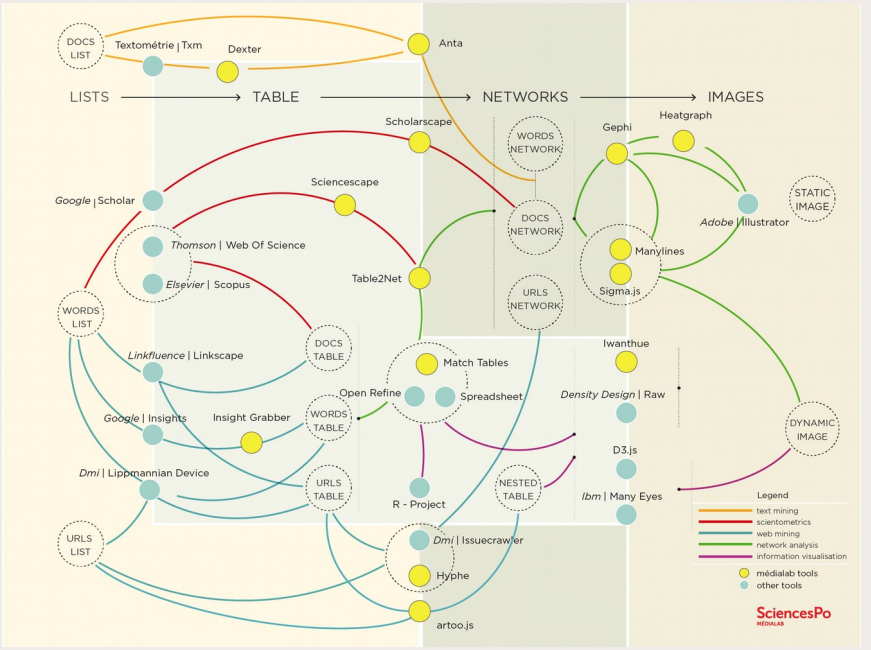
DIME-SHS

Le médialab de Sciences Po

- Centre de recherche de Sciences Po, fondé par Bruno Latour en mai 2009, dirigé par Dominique Cardon depuis 2017
- Numérique, sciences sociales et design

  → Interdisciplinarité
- Articulation des méthodes quali & quanti
- Étude des traces numériques
- Un écosystème d'outils
  http://tools.medialab.sciences-po.fr
- Un atelier ouvert mensuel
  http://www.medialab.sciences-po.fr/atelier/

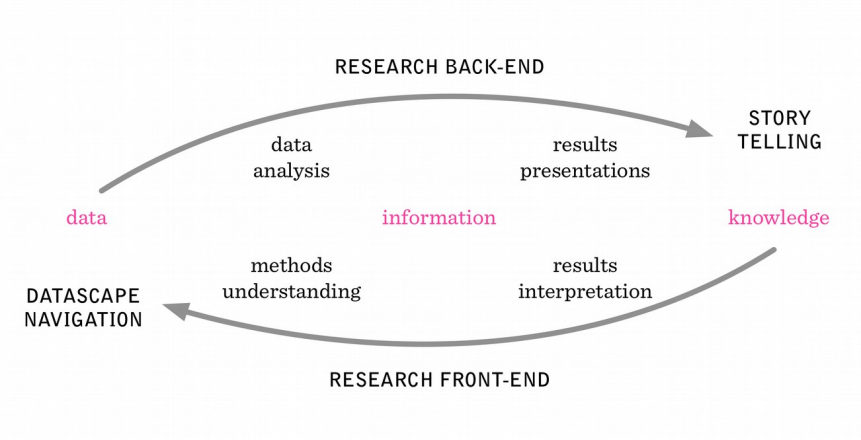




L'instrument DIME-Web (Equipex DIME-SHS)

- Accompagnement technique et méthodologique à l'utilisation du Web comme nouveau terrain d'enquête
  - Support aux Sciences Humaines et Sociales
  - Extraction ciblée de contenus/discussions/traces
  - Création de corpus documentaire
  - Méthodes numériques, itératives

    ≠ tout automatique



- Equipex (+ Ellips + beQuali = DIME-SHS)
  - 2 personnes (Mathieu Jacomy et moi-même)
  - Objectif ANR d'auto-financement

    → offre de service payant avec sélection
    → mutualisation (logiciels libres)

Le Web : une source de données « sales »

Collection de documents web (pages) sur un sujet en SHS

→ très grande hétérogénéité (type de contenu & forme)

CRAWLING ≠ SCRAPING

(fouille systématique) (extraction ciblée)

contenus textuels & hyperliens — données structurées

traitement automatique du langage

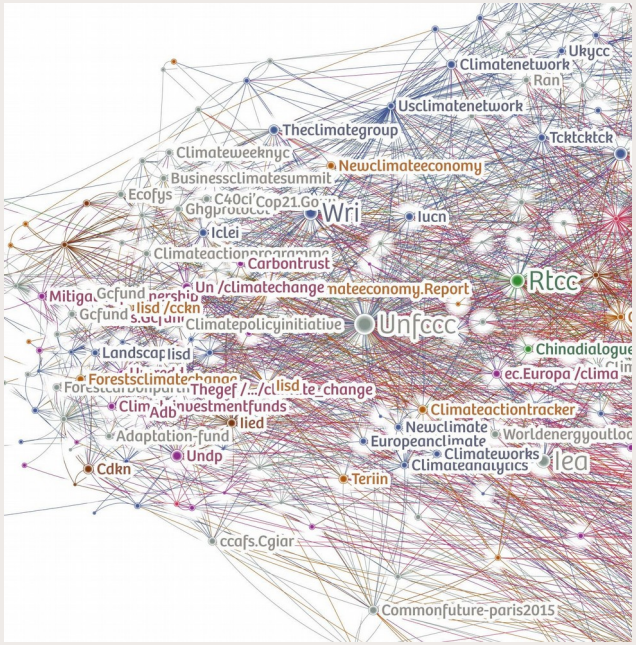
analyse de réseau (effets de communauté)

méthodes quantitatives, statistiques, …

redirections, liens erronés, liens morts et sites disparus, encodage mal indiqué…

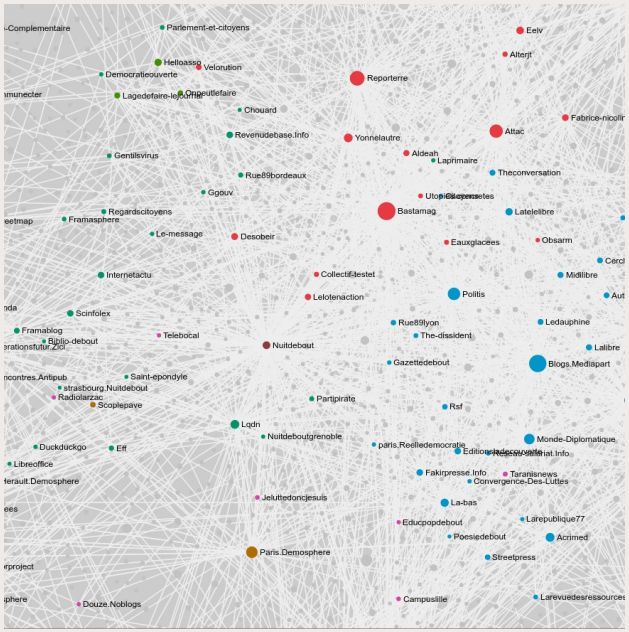
Hyphe : un crawler orienté par la recherche

- Les liens hypertextes : nouveaux révélateurs de relations entre acteurs d'une thématique
- Créer un corpus documentaire
  - « acteurs web » & contenus textuels respectifs
  - liens hypertextes entre ces acteurs
- Études exploratoires ou de controverses dans tous les domaines



http://medialab.github.io/double-dating-data/

COP 21
Vie privée
Extrême droite
Tissu associatif
Produits laitiers
Cellules souches
Administrations culturelles
…

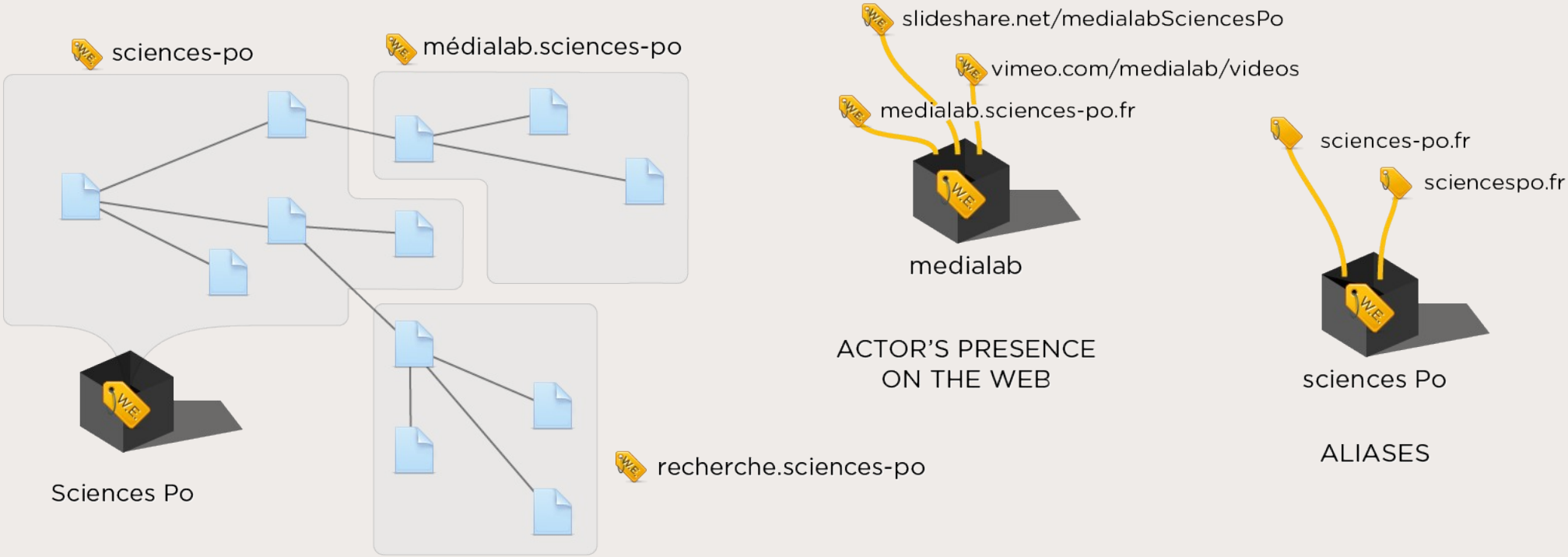


http://utopies-concretes.org/

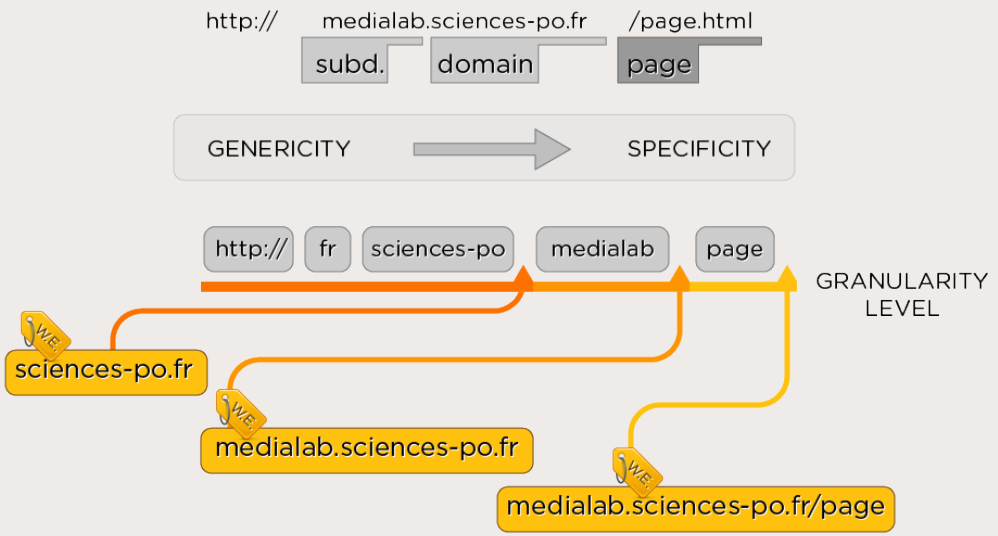
Principes méthodologiques : « WebEntités »

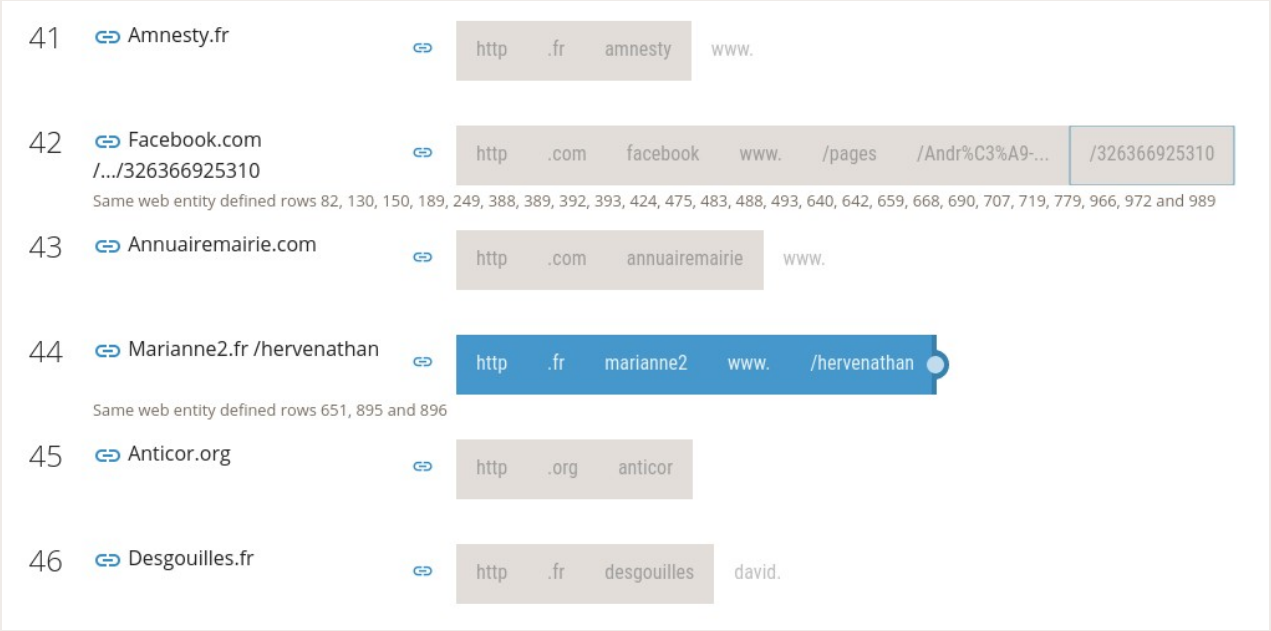
Comment gérer la diversité de granularité des sites web ?

→ « WebEntités » : agrégats reflétant des entités documentaires cohérentes du point de vue du chercheur



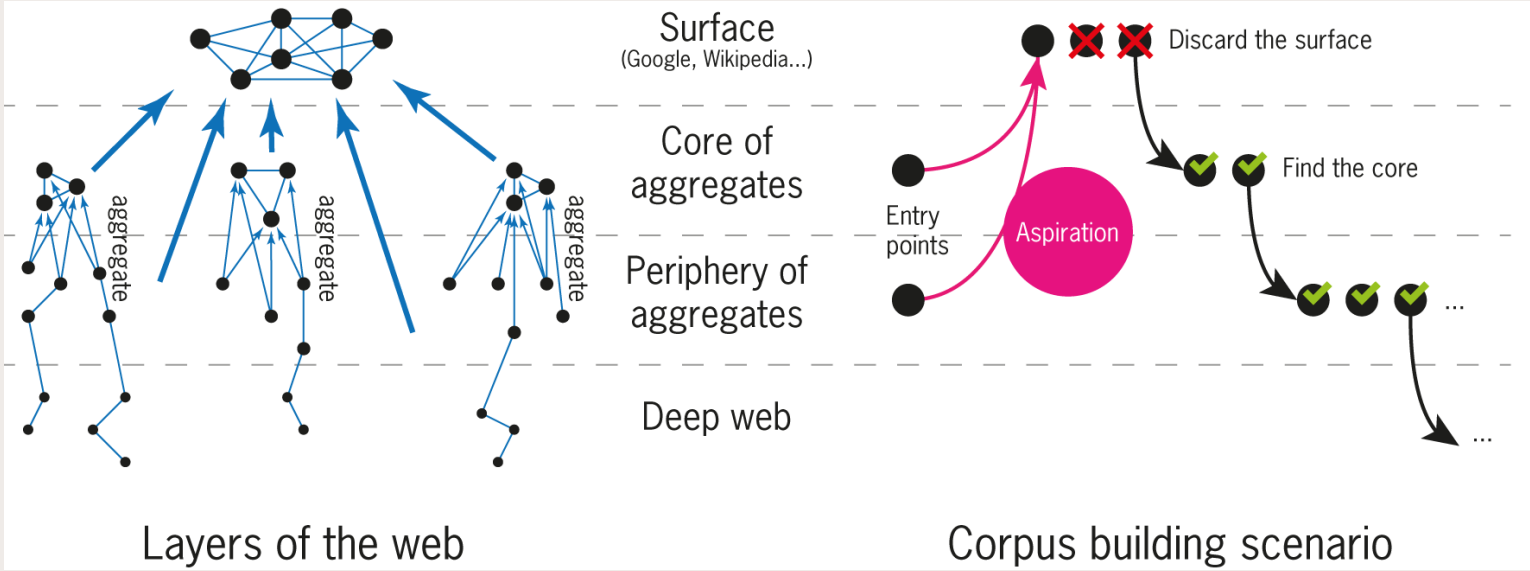
Principes méthodologiques : « WebEntités »





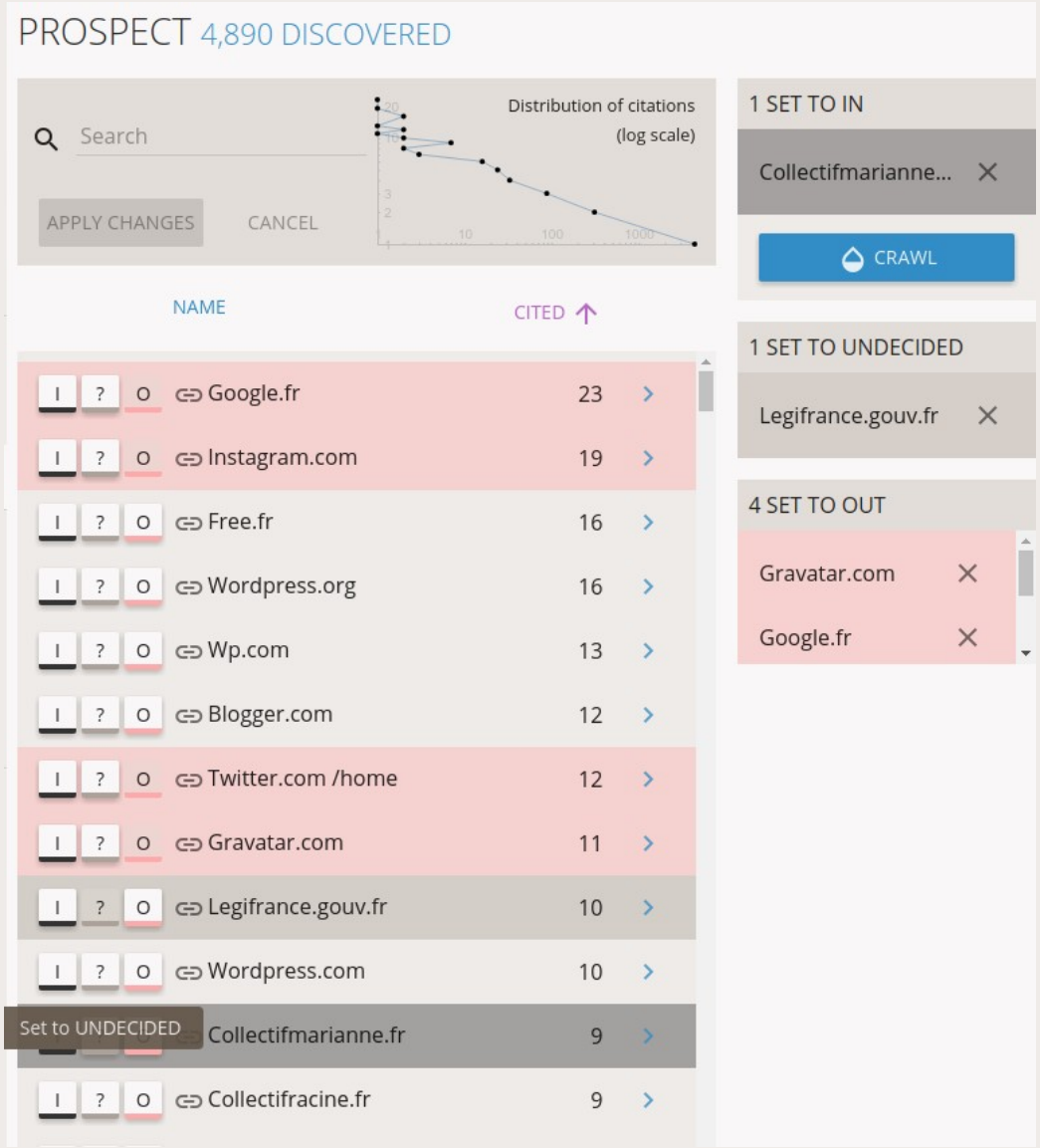
Principes méthodologiques : « Prospection »

- Démarrage : points d'entrées libres (recherche web qualitative, GoogleBookmarklets, annuaire, liste d'acteurs issue d'entretiens…)
- Crawler = robot qui fouille les pages web et clique sur les liens
  - Crawlers classiques : boule de neige (fouille systématique jusque N clics)
    → bruit de la couche haute du web (Google, YouTube, Wikipedia…)
  - Hyphe : crawl ciblé, uniquement les pages internes des WebEntités choisies
    → éditorialisation et contrôle de la construction thématique



Principes méthodologiques : « Prospection »

- Exploitation de la nature hypertextuelle du web
- Identification des acteurs web liés potentiellement pertinents
- Travail de terrain (virtuel)
  → exclure ou inclure
- Décisions éditoriales classiques de type gestion documentaire

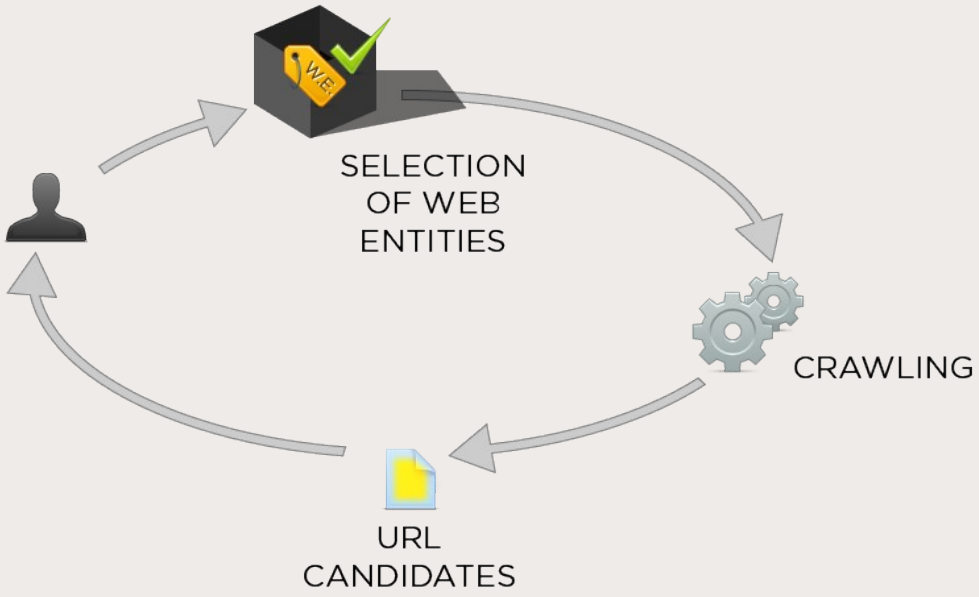


Principes méthodologiques : « Prospection »

- Expansion éditorialisée et itérative du corpus
- Coût en temps humain : travail de curation répétitif
  « crawler orienté par la recherche »
- La liste des WebEntités découvertes s'allonge exponentiellement

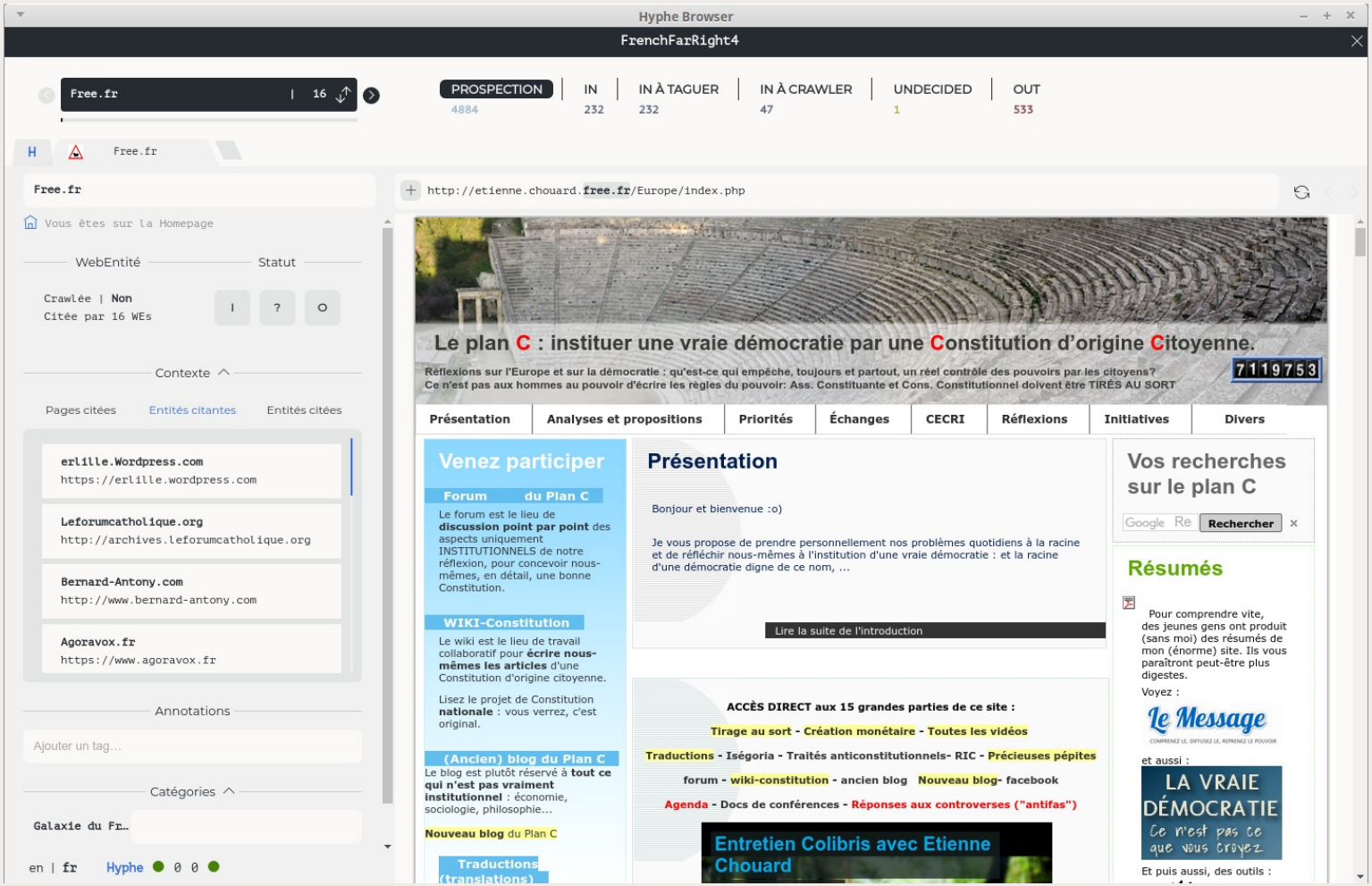
  → Quand s'arrêter ?

  → Seuil de citation



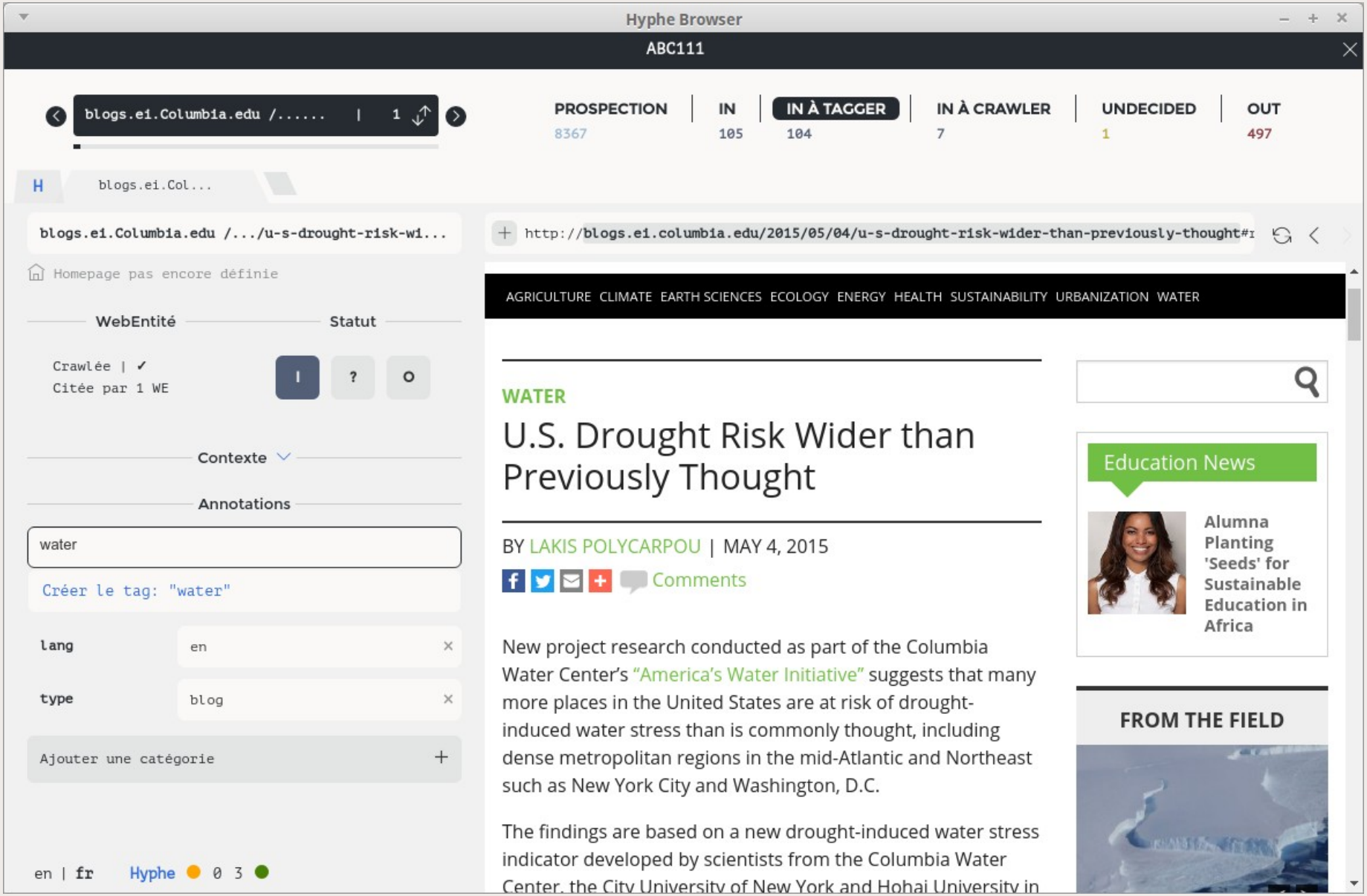
HyBro : un browser pour prospecter in situ

- Hyphe-Browser : héritier du « NaviCrawler »
- Un navigateur web connecté à Hyphe



https://github.com/medialab/hyphe-browser/releases/

Catégoriser les WebEntités avec HyBro



https://github.com/medialab/hyphe-browser/releases/

Gérer ses catégorisations (tags)

TAGS

Filter web entities (status *IN* only). Tag one or a selection of web entities.

439
WEB ENTITIES

TAG FILTERS

439 WEB ENTITIES | WEB ENTITIES NETWORK

Special filters

- *Untagged*
- *Partially untagged*
- *Conflicts*

Free Tags

- *Untagged*

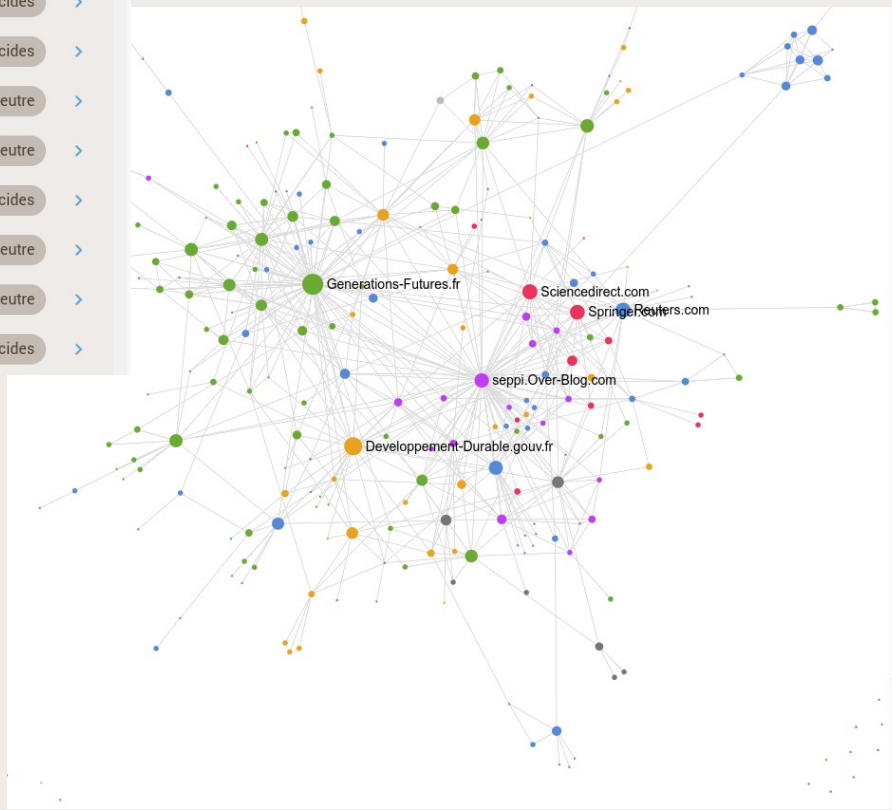
Acteur

- *Untagged*
- Presse *157*
- Association *111*
- Institution *51*
- Blog *56*
- Publication scientifique *23*

Search

Display a category: Point de vue

Futura-Sciences.com /.../biologie-pesticide-9169	Neutre
Lefigaro.fr /.../37002-20170627ARTFIG00002-pesticidepe-sti-sid-n-m-...	Neutre
Parents.fr /.../pesticides-et-grossesse-des-risques-confi...	Contre les pesticides
formulaires.Fondation-Nicolas-Hulot.org /.../stop_pestic...	Contre les pesticides
Contrepoints.org /.../270496-pesticides-lintox-discours-bio	Pour les pesticides
Observatoire-Pesticides.gouv.fr	Neutre
Letemps.ch /.../toxicite-pesticides-tueurs-dabeilles-confirmee-terrain	Neutre
Sciencepresse.qc.ca /.../neonicotinoides-pesticides-tue...	Contre les pesticides
Notre-Planete.info /.../4613-liste-fruits-legumes-pesticides	Neutre
Lepoint.fr /.../pesticides-tueurs-d-abeilles-bayer-interpelle-par-un-mil...	Neutre
Consoglobe.com /abeilles-pesticides-bayer-cg	Contre les pesticides



Key

- Each dot or *node* is a web entity
- Each line or *edge* or *link* represents one or more hyperlinks from a web entity to another. Links are oriented even though it is not figured in the image.

NODE COLOR: Acteur (tag)

- Presse (157)
- Association (112)
- Blog (56)
- Institution (51)
- Publication scientifique (23)
- Entreprise (21)
- Untagged (19)

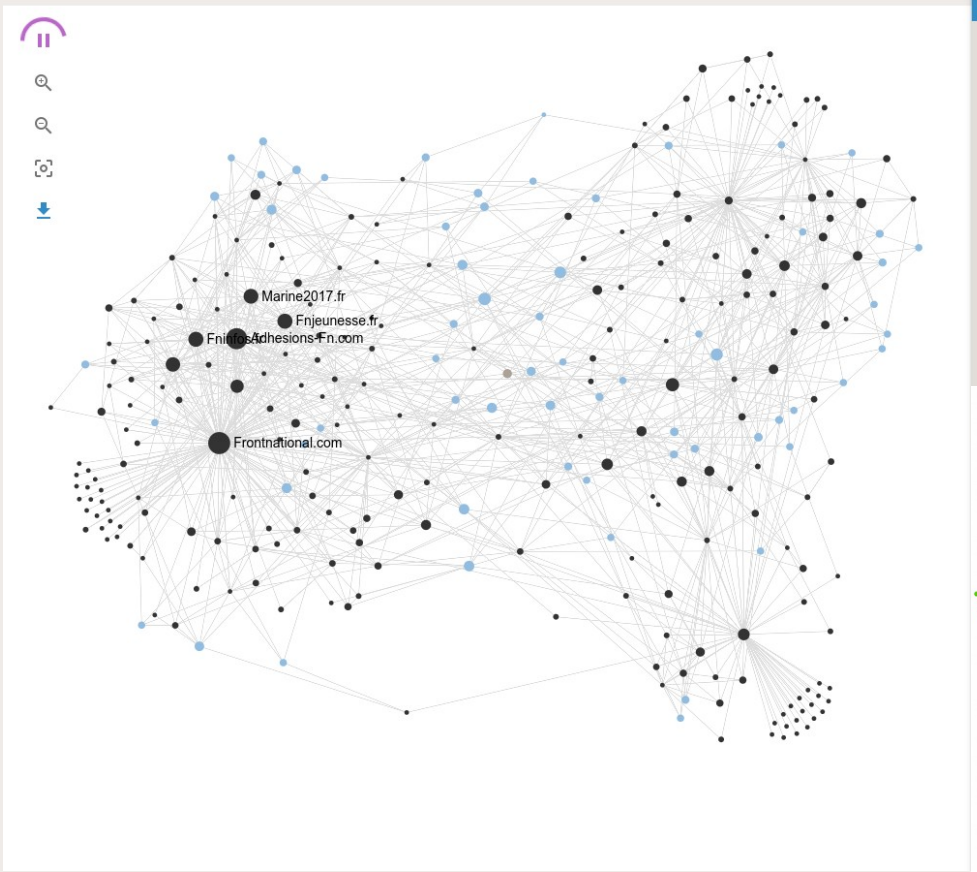
NODE SIZE: Indegree

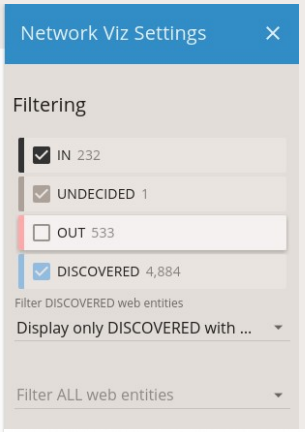
The INDEGREE of a web entity is the number of other web entities citing it.

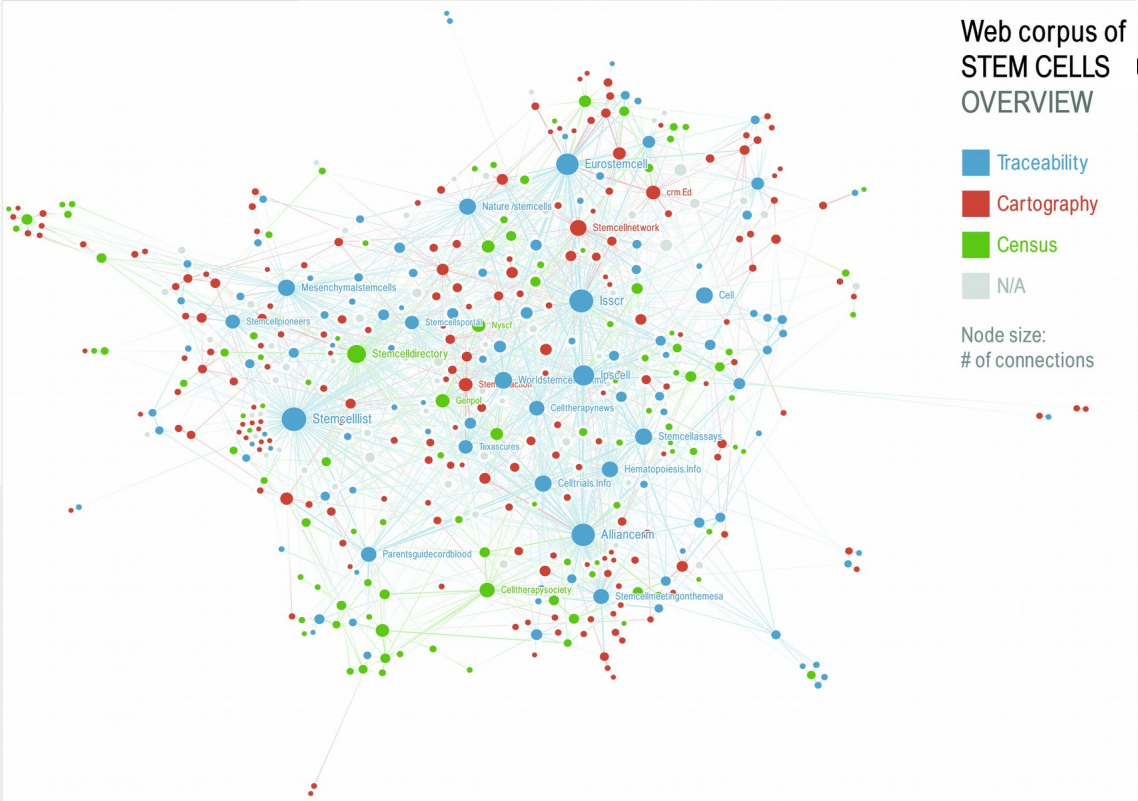
- Smallest node INDEGREE of 0
- Biggest node INDEGREE of 38

SciencesPo
MÉDIALAB

Explorer le réseau des liens entre acteurs



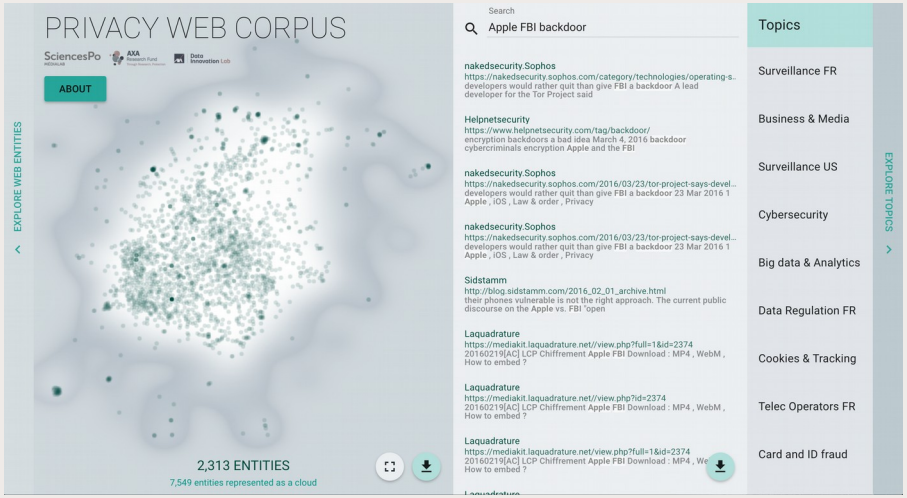




Social Representations of Stem Cells, Virginie Tournay, CEVIPOF, 2016

Et pour la suite ?

- Import / export de listes de webentités et crawls ou de corpus :
  - → duplication, reproduction
  - → exploration longitudinale dans le temps



- Exploitation intégrée des contenus textuels issus des pages crawlées et analyse automatique du langage
- Stabiliser PhantomJS pour le crawl browser-like (Facebook, etc.)
- Contrôle qualité des crawls et du corpus
- Outil d'archivage et présentation des corpus finalisés
- Hyphe embarqué sur clé USB

Bibliographie & liens divers

- Concepts et explications :
  http://hyphe.medialab.sciences-po.fr/
- Instance de démo (restreinte) en libre accès :
  http://hyphe.medialab.sciences-po.fr/demo/
- Publications associées :
  - Jacomy M., Girard P., Ooghe-Tabanou B., Venturini T. (2016), **Hyphe, a Curation-Oriented Approach to Web Crawling for the Social Sciences**, ICWSM 2016, Cologne, Allemagne.
    https://spire.sciencespo.fr/hdl:/2441/6obemb2hsj9pboj9bbvc7sftne
  - Jacomy M., Venturini T., Heymann S., Bastian M. (2014), **ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software**, PLoS ONE 9(6): e98679. doi:10.1371/journal.pone.0098679.
    http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679
  - Venturini T., Jacomy M., Pereira D. (2015), **Visual Network Analysis: the Example of the Rio+20 Online Debate**, Working paper.
    http://www.medialab.sciences-po.fr/wp-content/uploads/2015/06/VisualNetwork_Paper-10.pdf

Merci de votre attention !

Et maintenant, à vous de jouer !

SciencesPo
MÉDIALAB

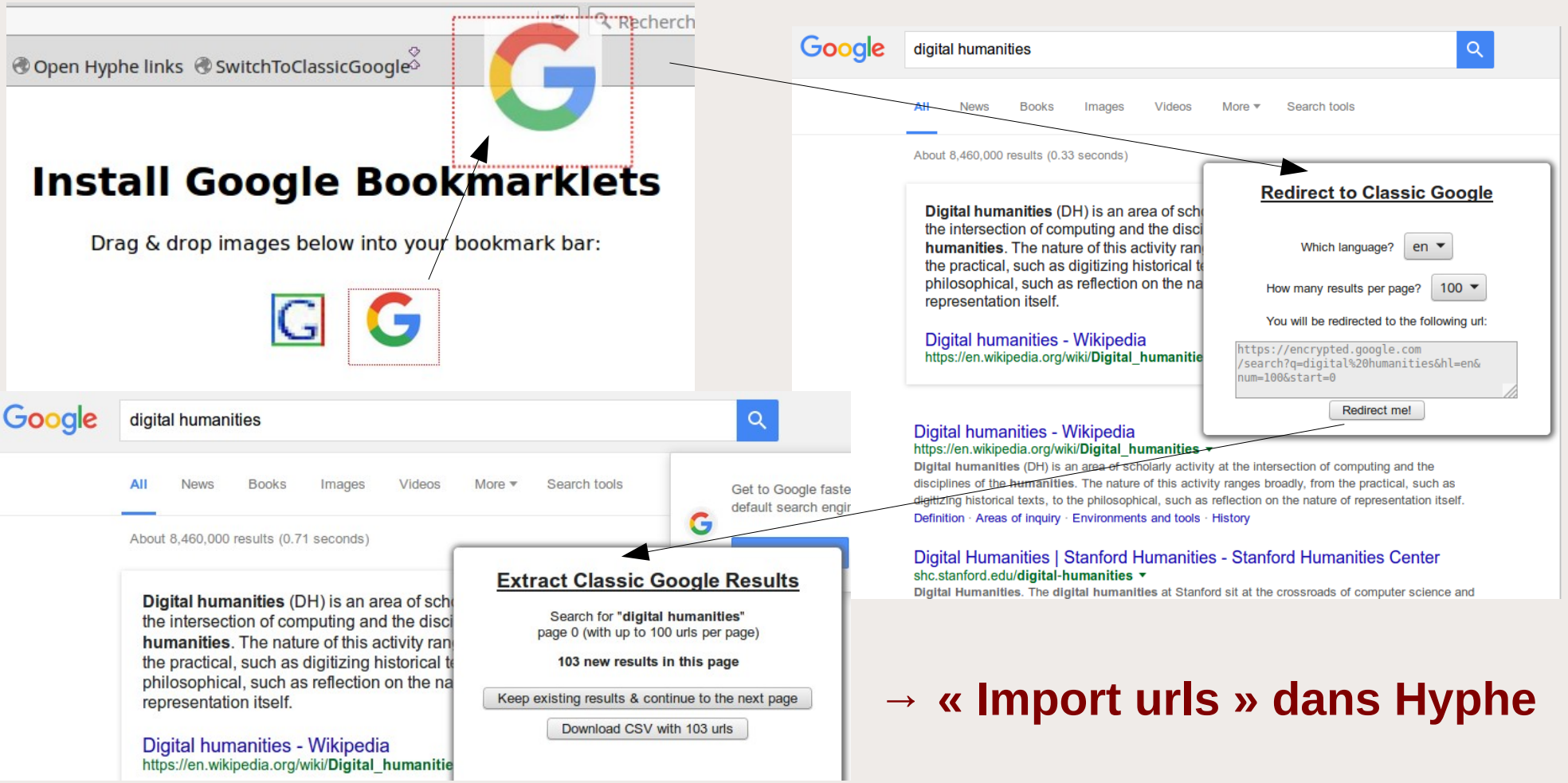
@medialab_ScPo

benjamin.ooghe@sciencespo.fr

Google bookmarklets : résultats Google en CSV

https://medialab.github.io/google-bookmarklets/

Des boutons dans vos favoris pour récupérer simplement au format tableur les résultats d'une recherche Google



→ « Import urls » dans Hyphe