Outils et méthodes pour créer, traiter et analyser des corpus web

Benjamin Ooghe

Outils & méthodes pour créer, traiter et analyser des corpus web

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

medialab.sciences-po.fr

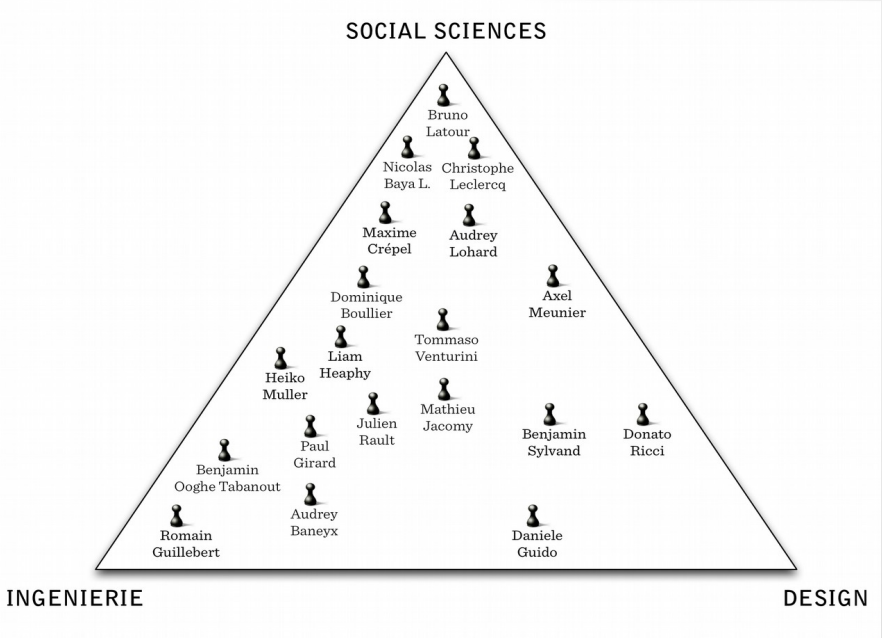
Atelier INA #3 Saison 6 - 17/04/15

Le médialab

- Fondé en mai 2009

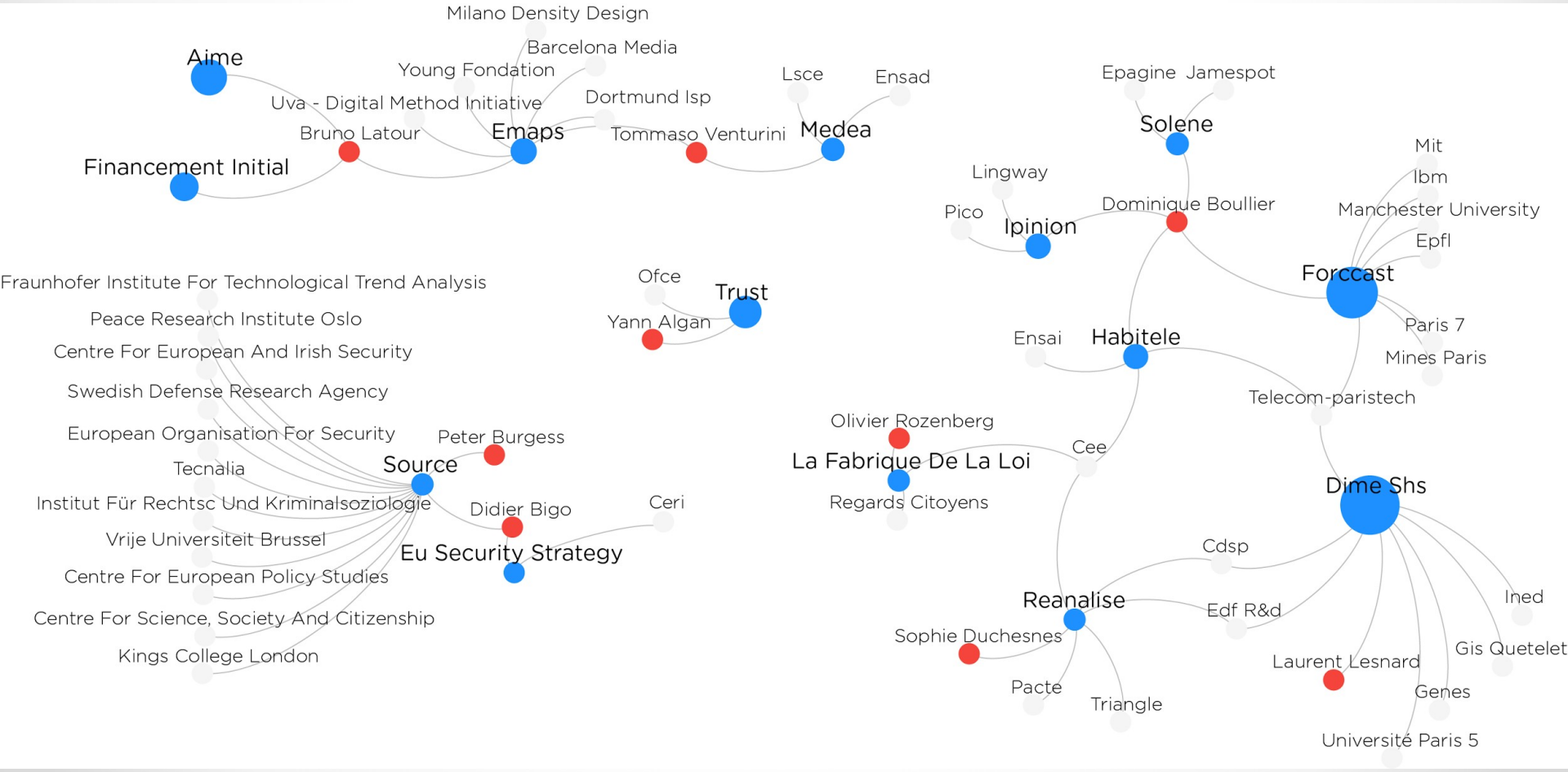
Centre de recherche numérique au service de SciencesPo et des
- sciences sociales.

Étude des traces numériques : articuler les méthodes
- quantitatives et qualitatives



- Pluridisciplinarité : Sciences sociales + Ingénierie + Design

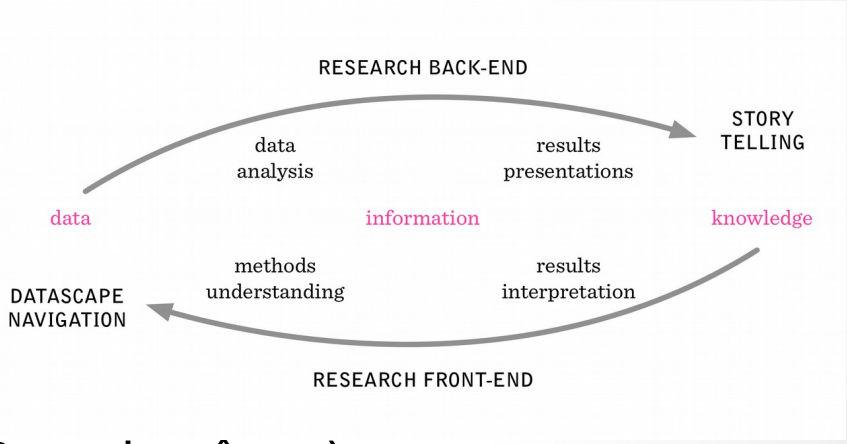
Une multiplicité de projets et partenaires



L'instrument DIME Web

- Equipex support aux Sciences Humaines et Sociales

Accompagnement numérique
- et méthodologique



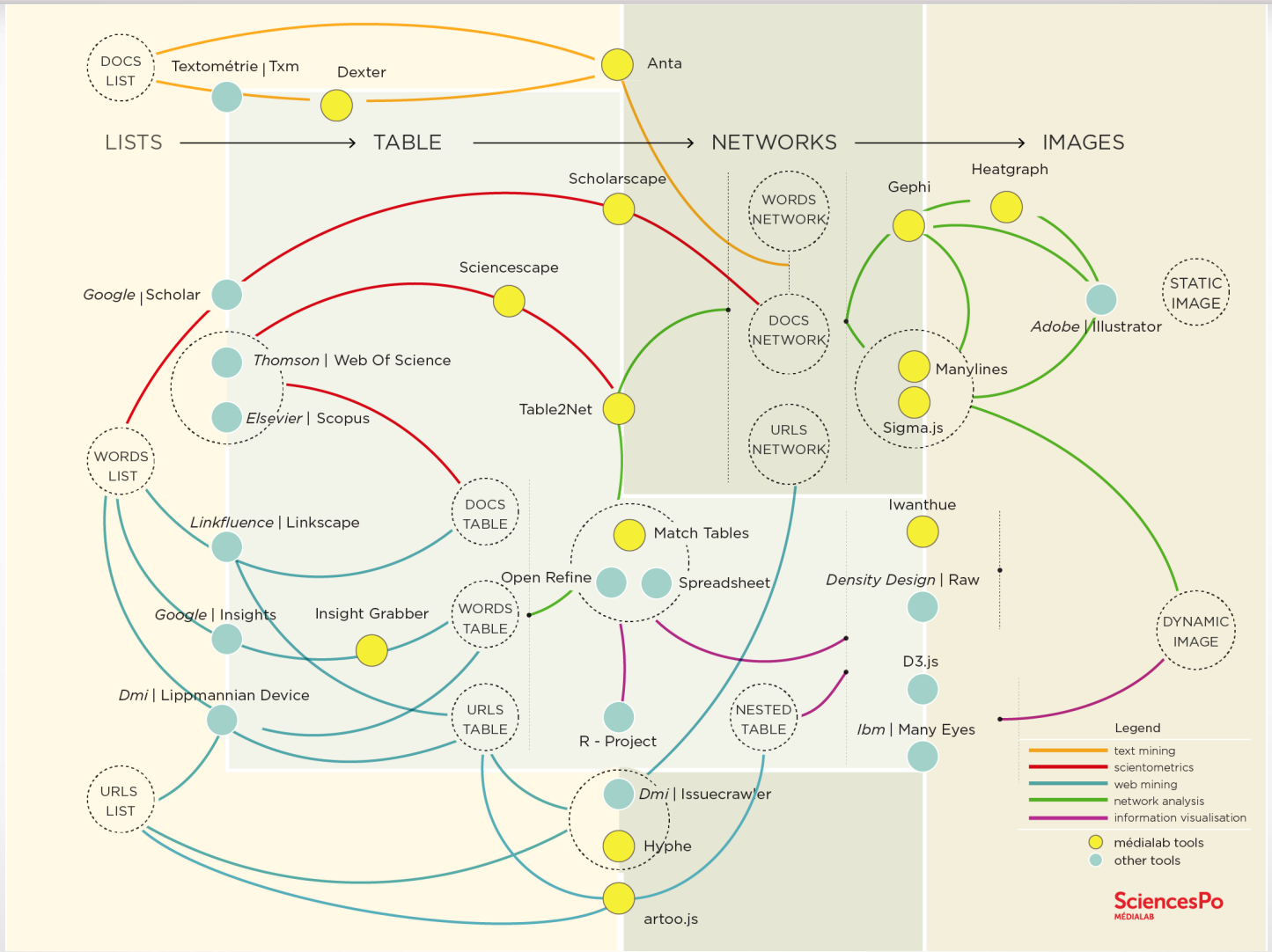
- Méthodologie itérative

- 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



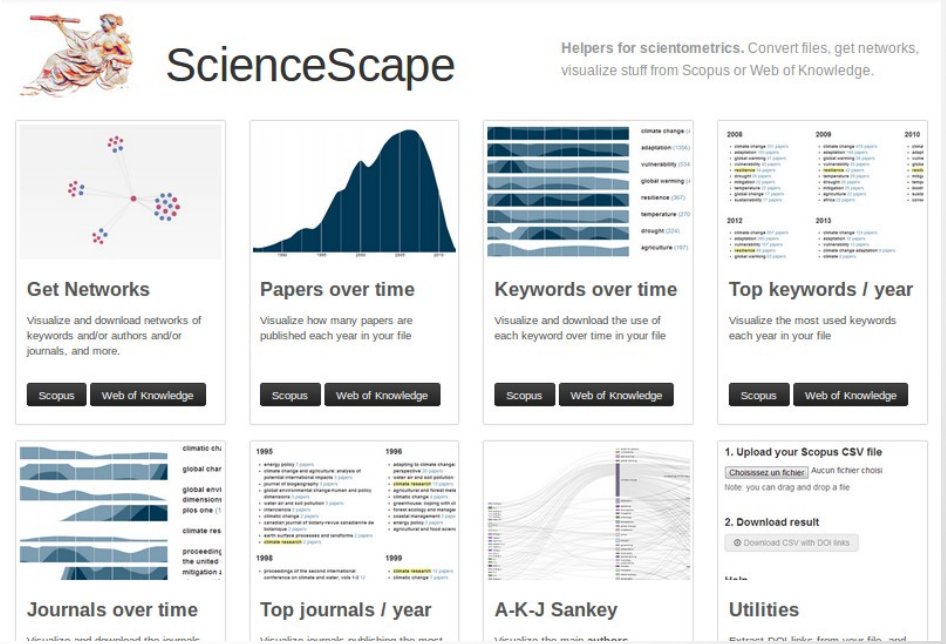
Collectes ciblées (scraping)

- gazouilloire : collecte Twitter programmée
- ScienceScape : construction de corpus scientométriques

artoo.js & sandcrawler.js :

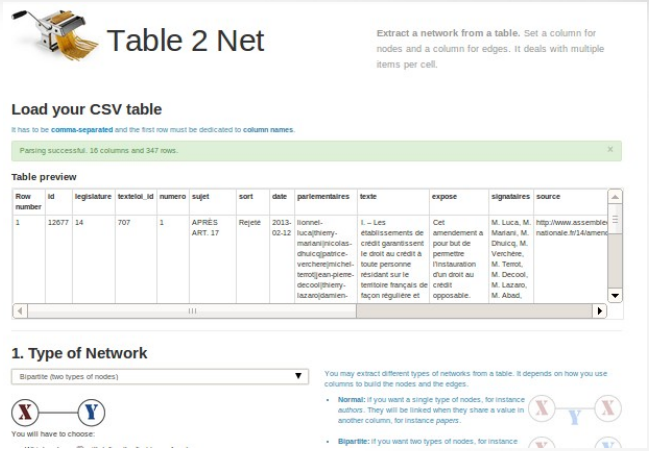
- simuler le navigateur web

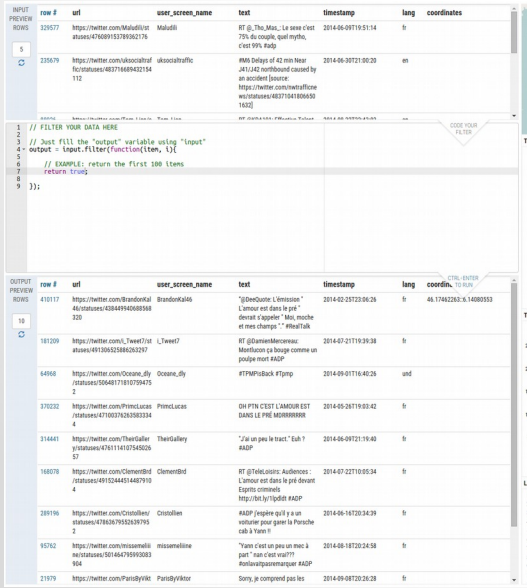
- Scripts dédiés :
- - commentaires LeMonde.fr
- - métadonnées FlickR
- - forum M6.fr
- - …

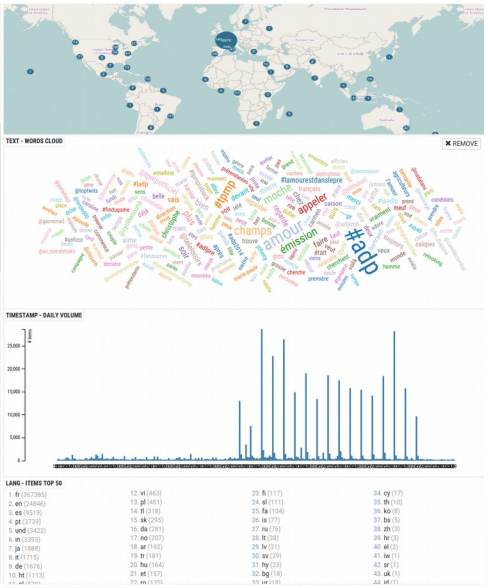


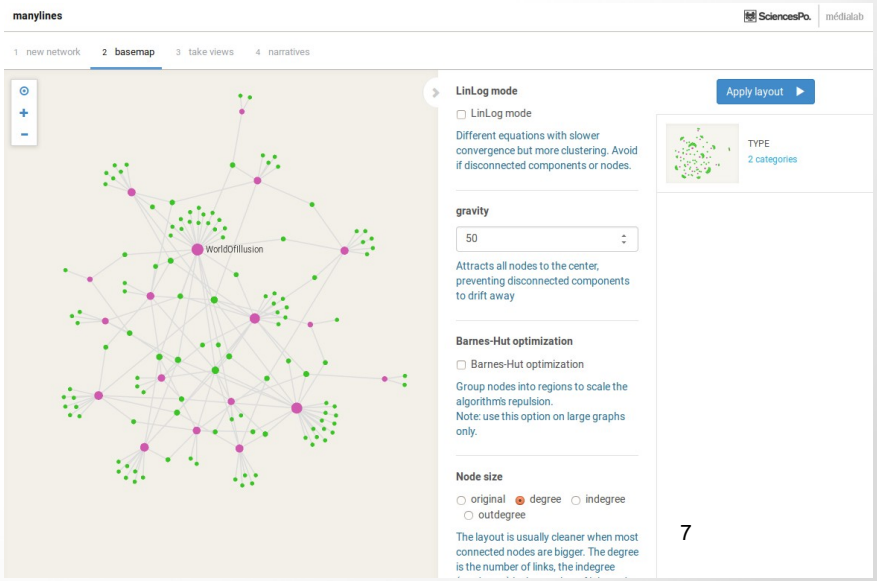
Explorer les données collectées

- Exploration visuelle (raw, rerere)
- Analyse visuelle de réseaux :
- - Gephi, Table2Net, sigma.js, ...
- - ManyLines : storytelling (exemple)





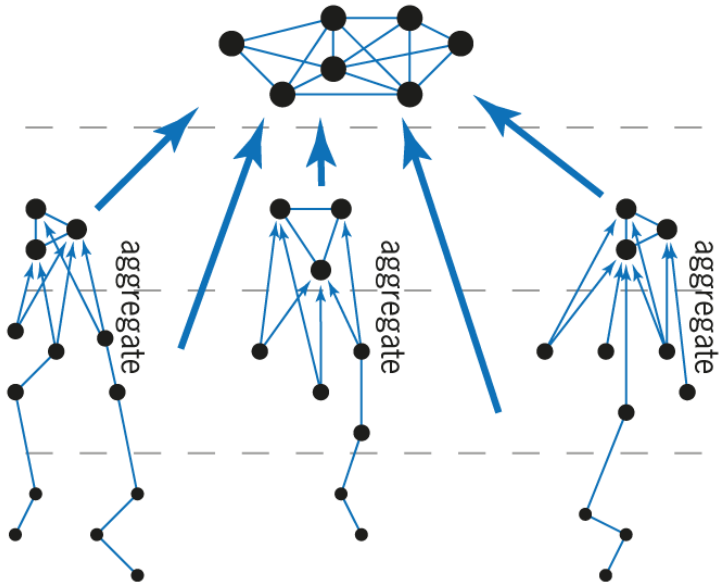




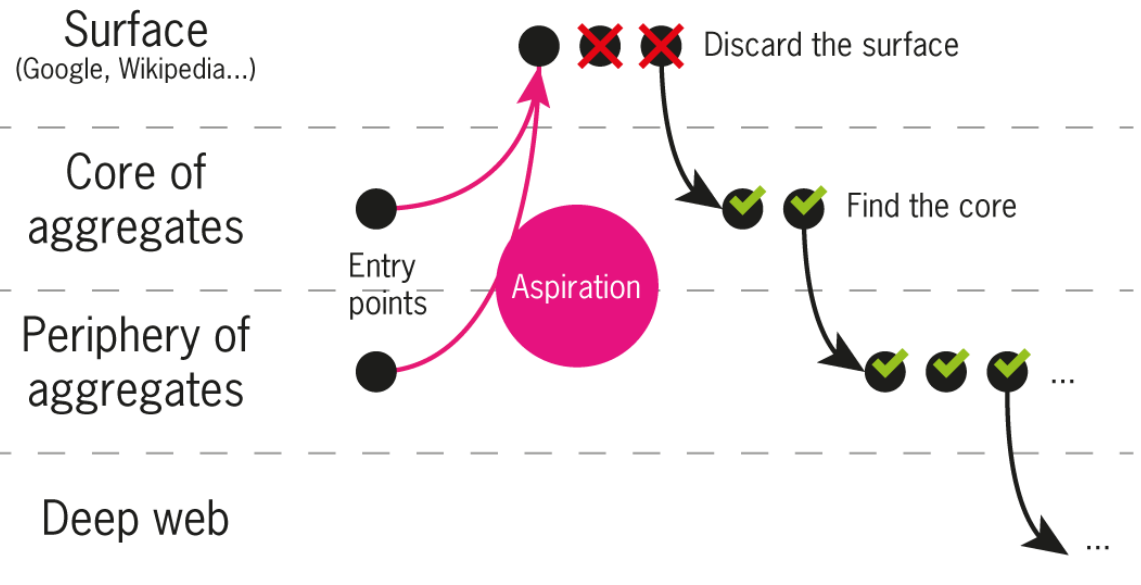
Hyphe : crawler orienté recherche

- 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

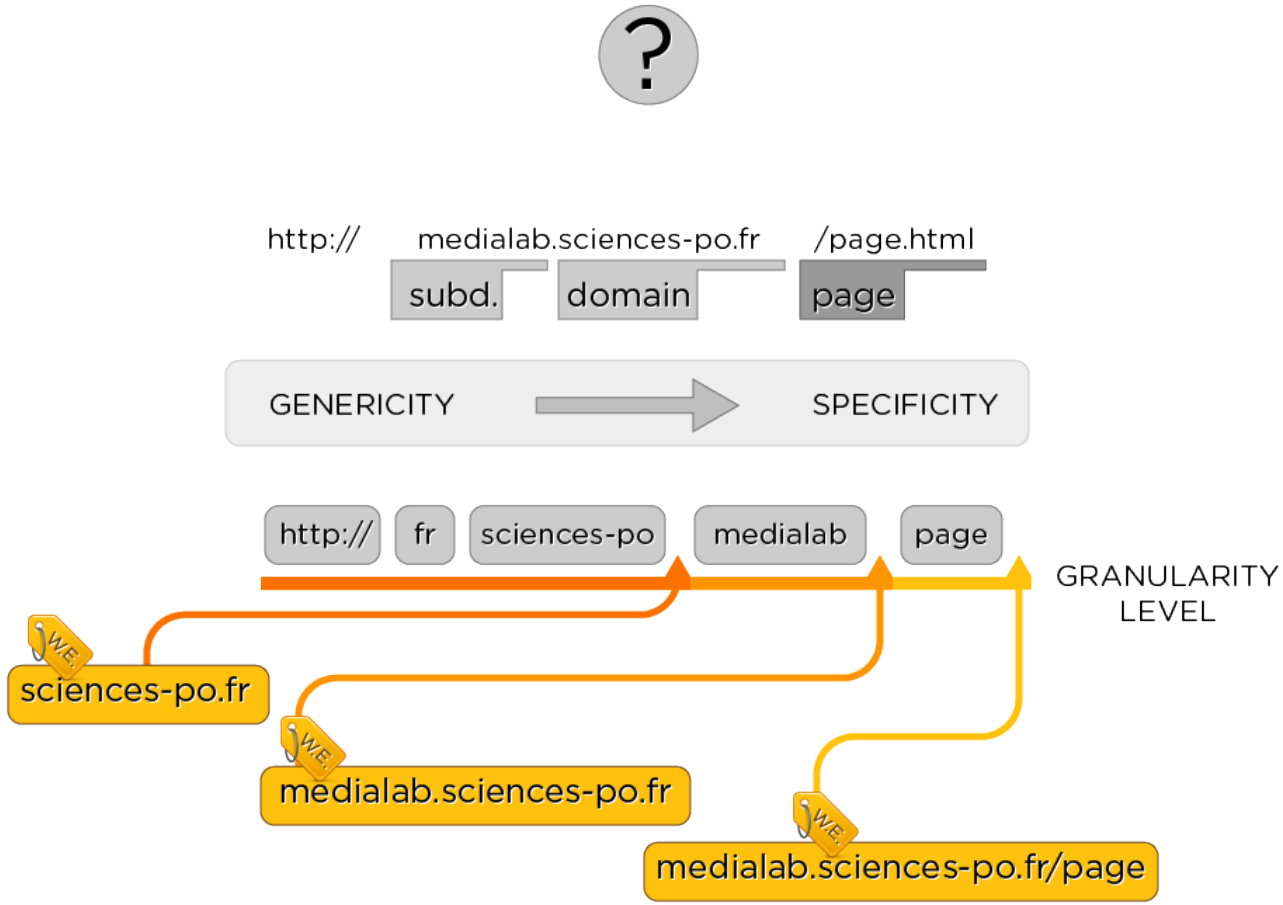


Layers of the web

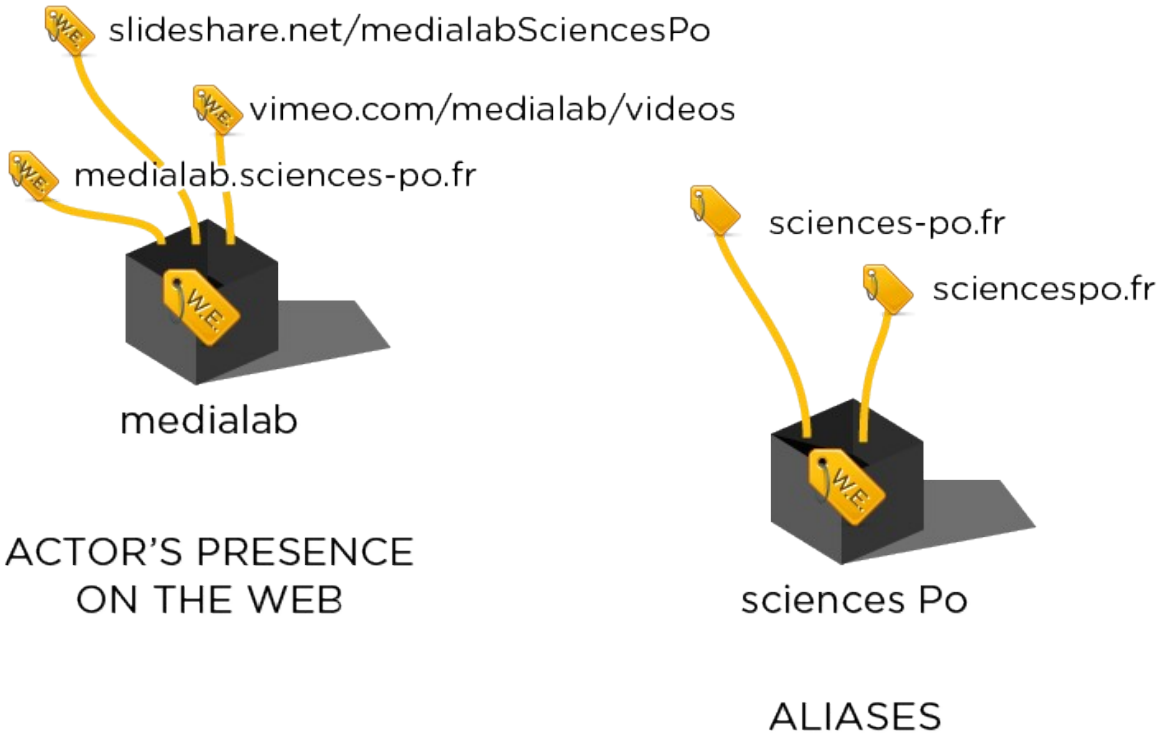


Corpus building scenario

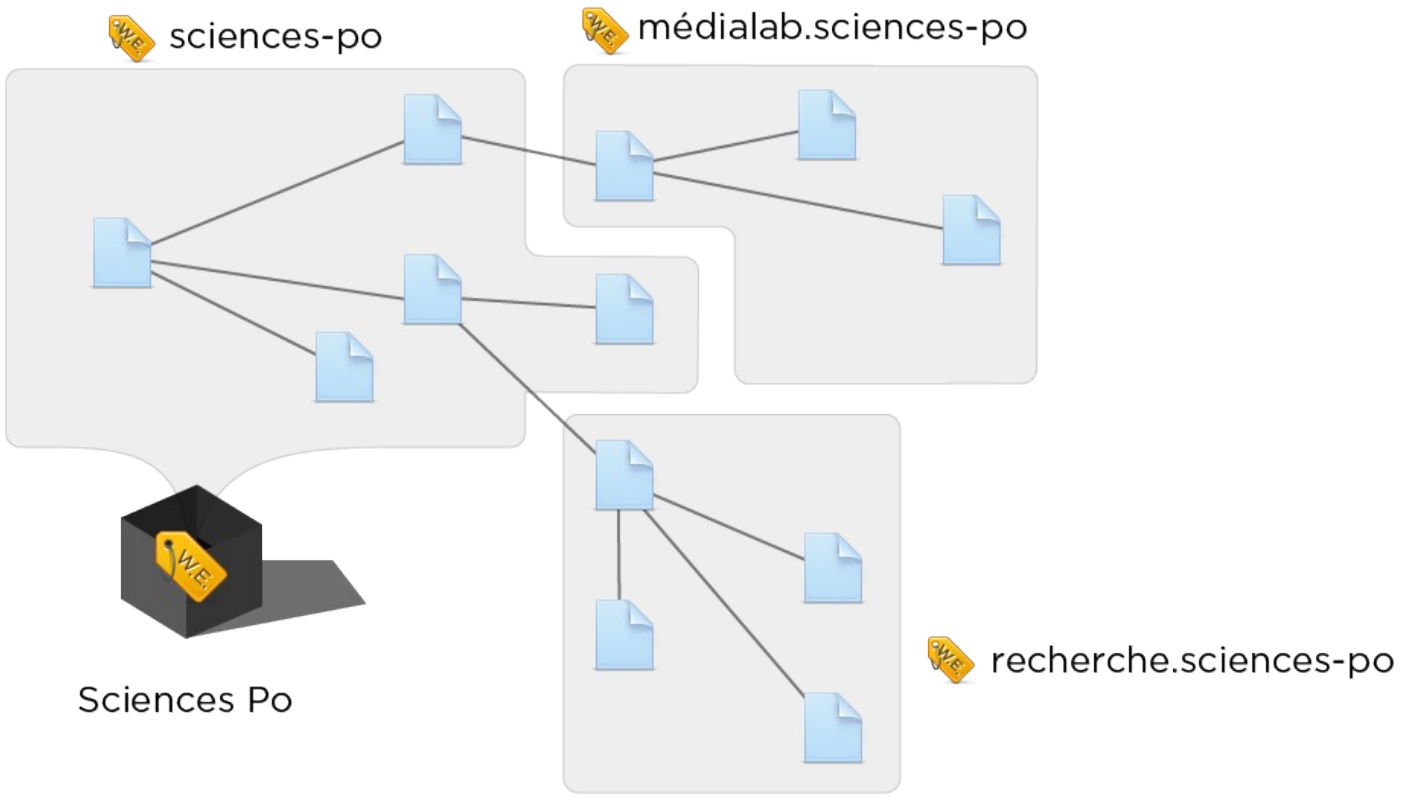
Définir des points d'ancrage précis (LRUs)



Des sites ou... des entités web



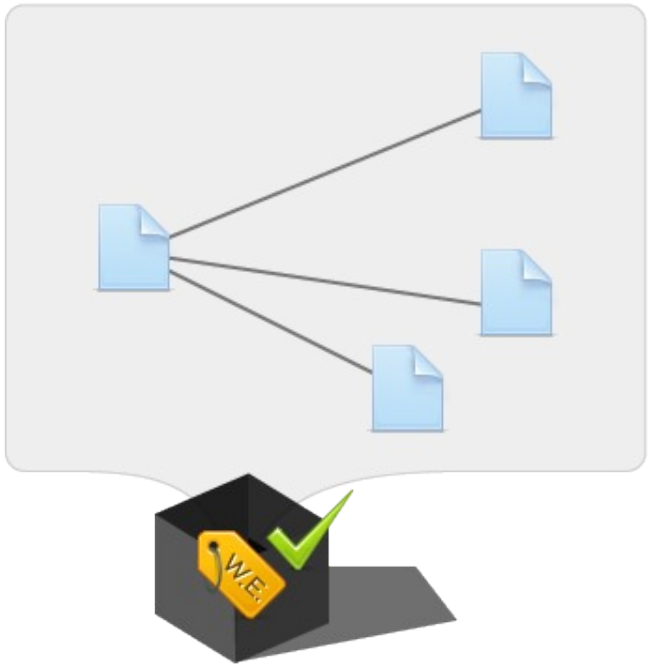
Préserver la complexité



Le crawl dirigé par la recherche

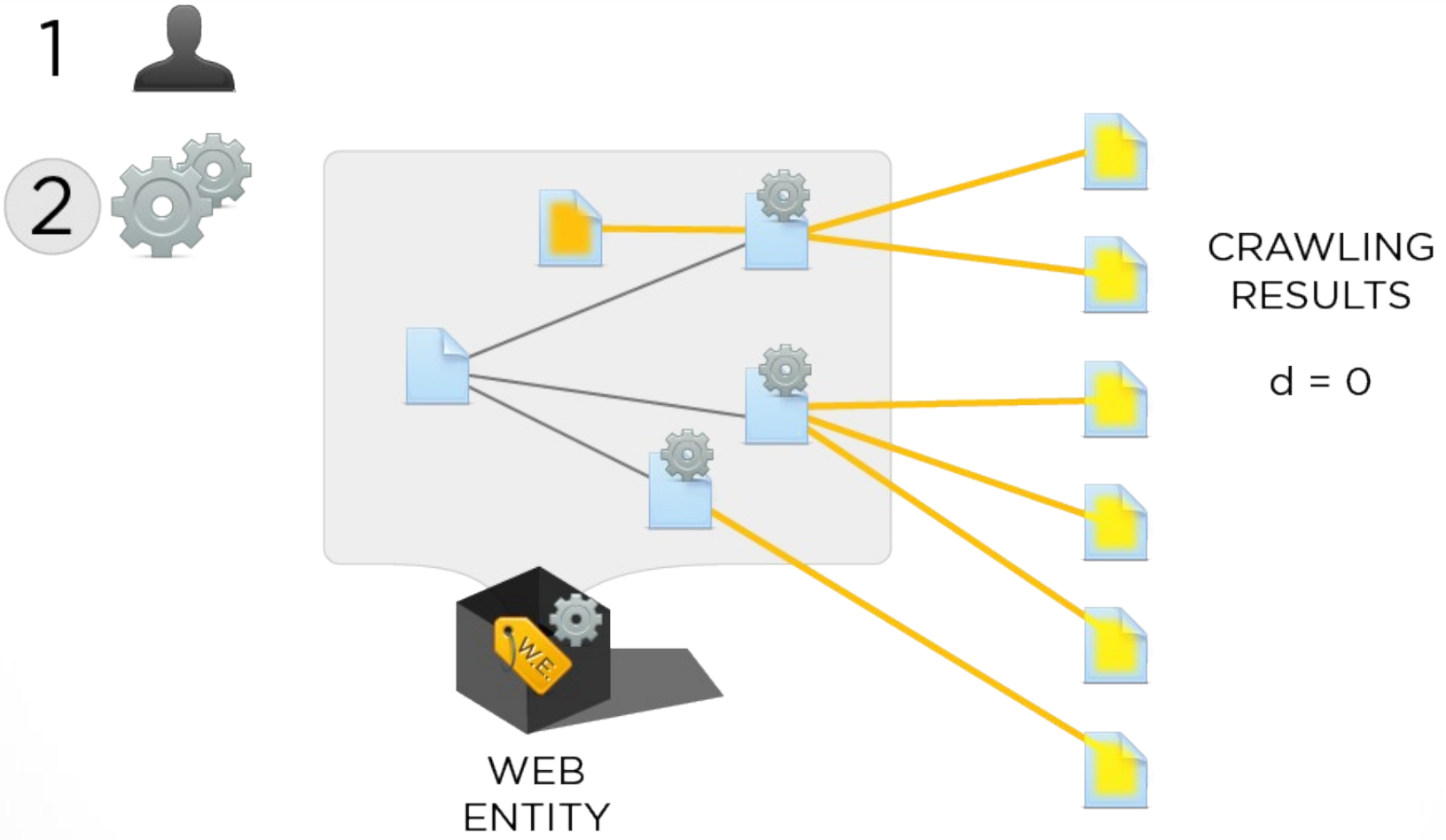
1

RESEARCHER
SELECT
STARTING
ENTITIES

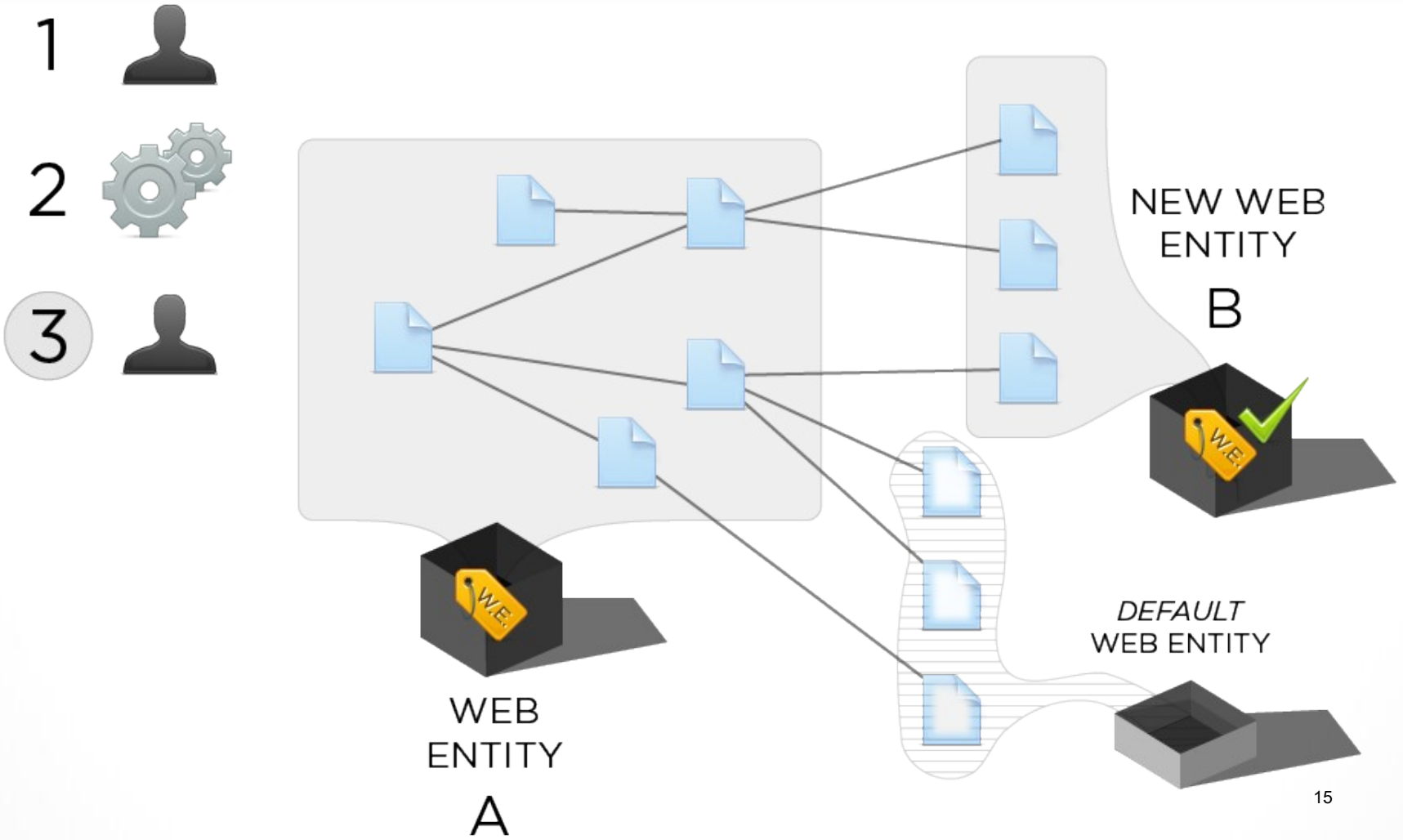


WEB
ENTITY

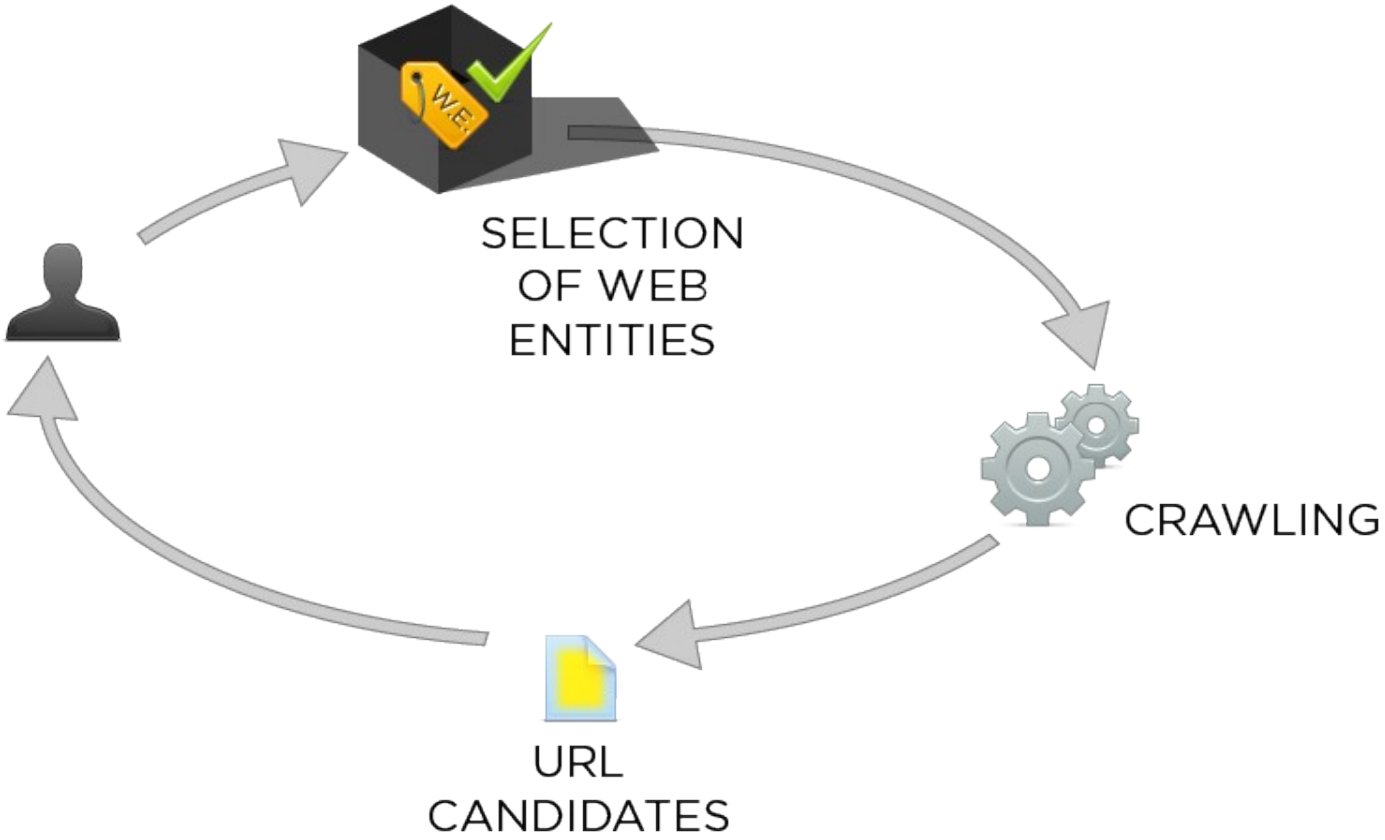
Le crawl dirigé par la recherche



Le crawl dirigé par la recherche



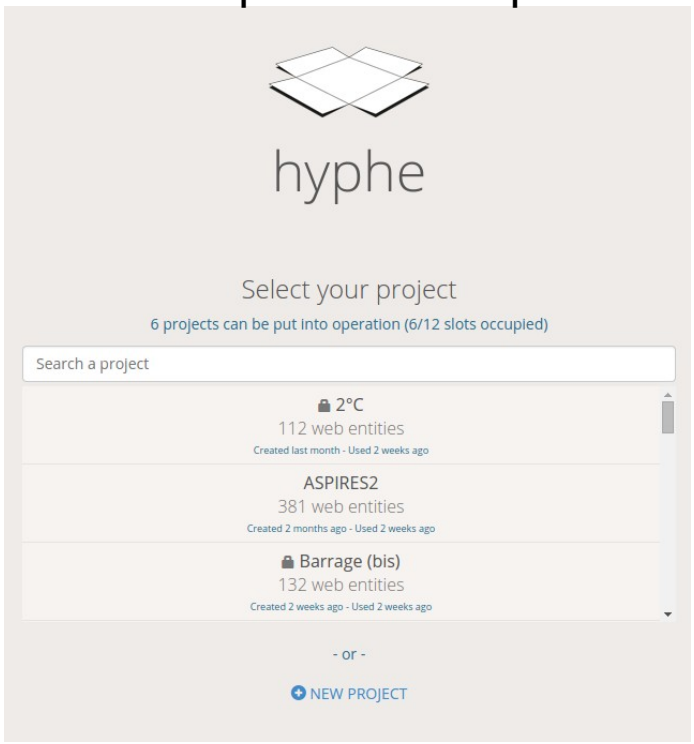
Le crawl dirigé par la recherche



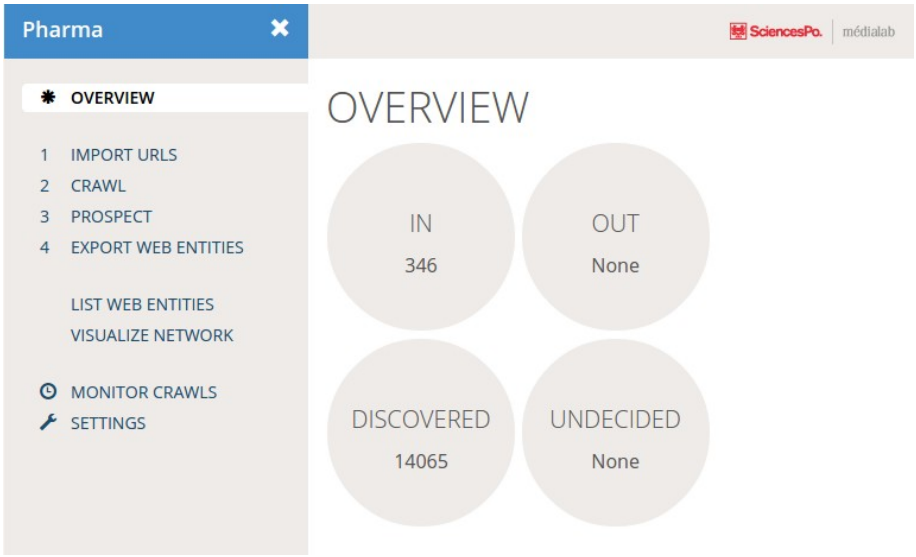
L'interface de Hyphe 2nde version

Démo : hyphe.medialab.sciences-po.fr/demo Code source : github.com/medialab/hyphe

Gérer plusieurs corpus



Résumé d'un corpus

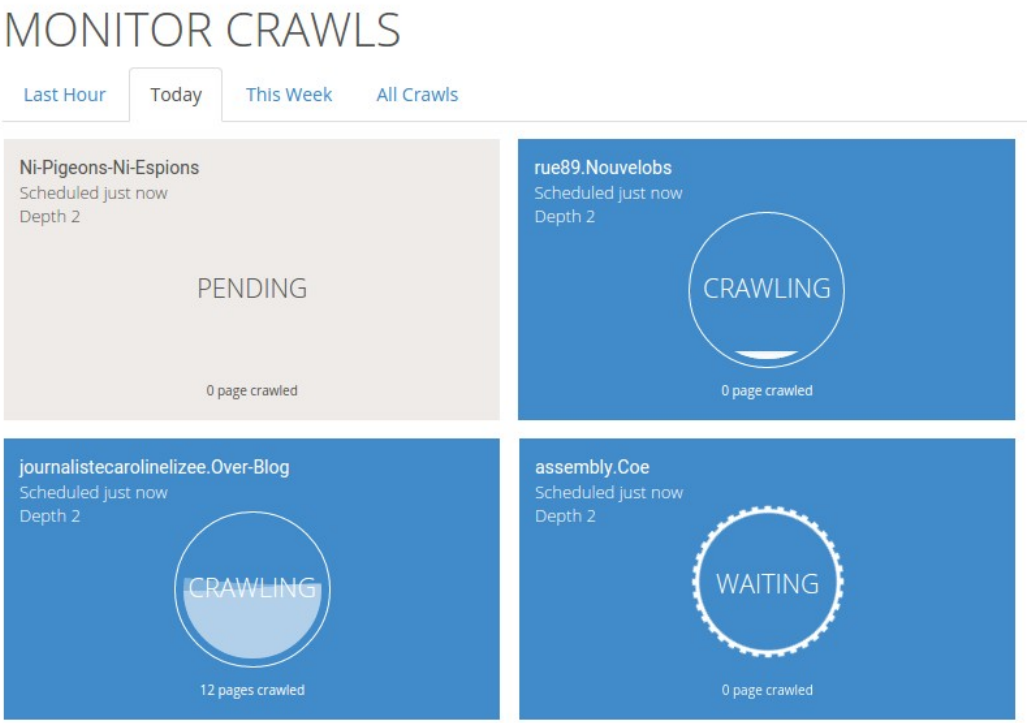


L'interface de Hyphe 2nde version

Définir précisément les WebEntités

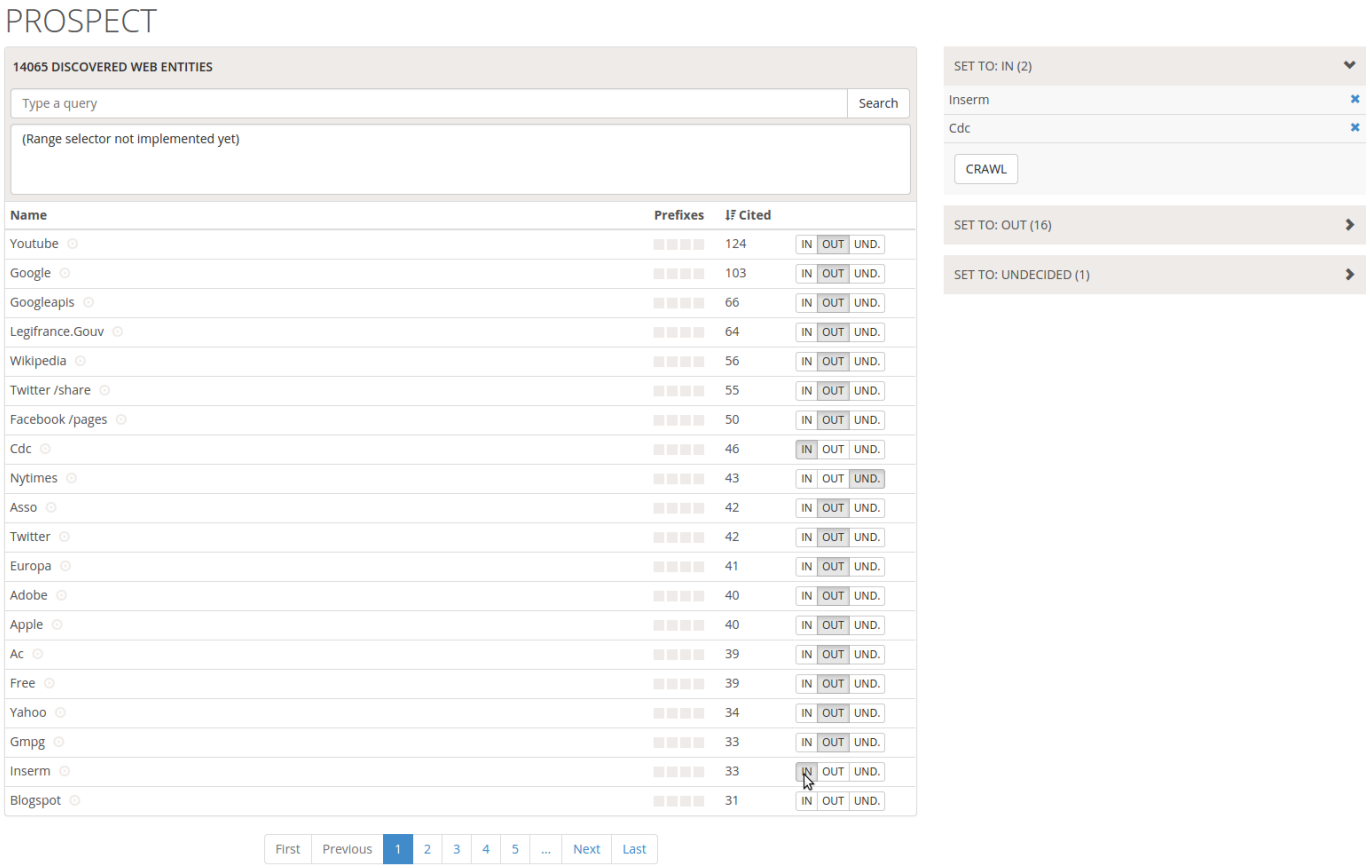
L'interface de Hyphe 2nde version

Surveiller l'avancement des crawls



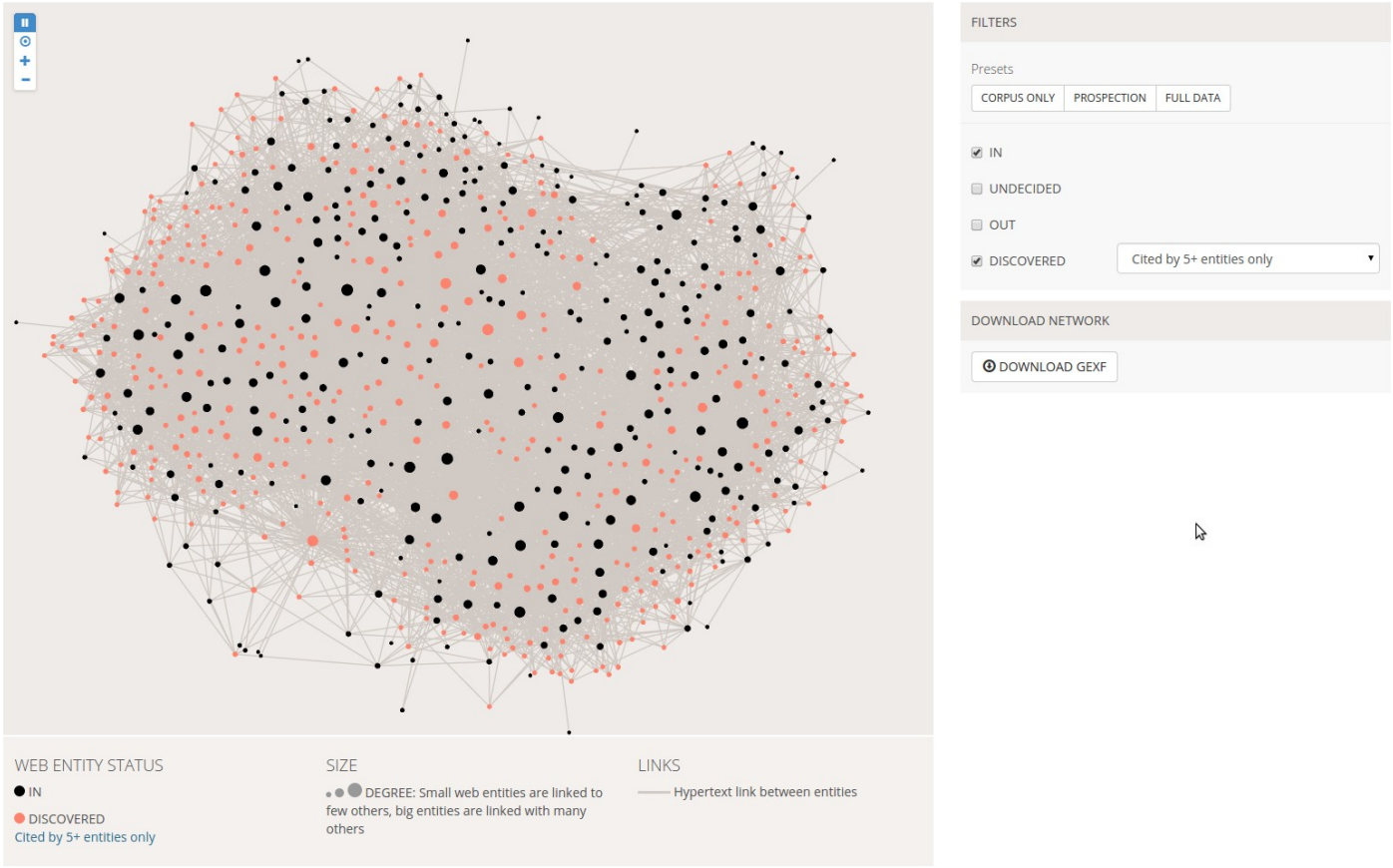
L'interface de Hyphe 2nde version

Identifier d'autres WebEntités à inclure au corpus



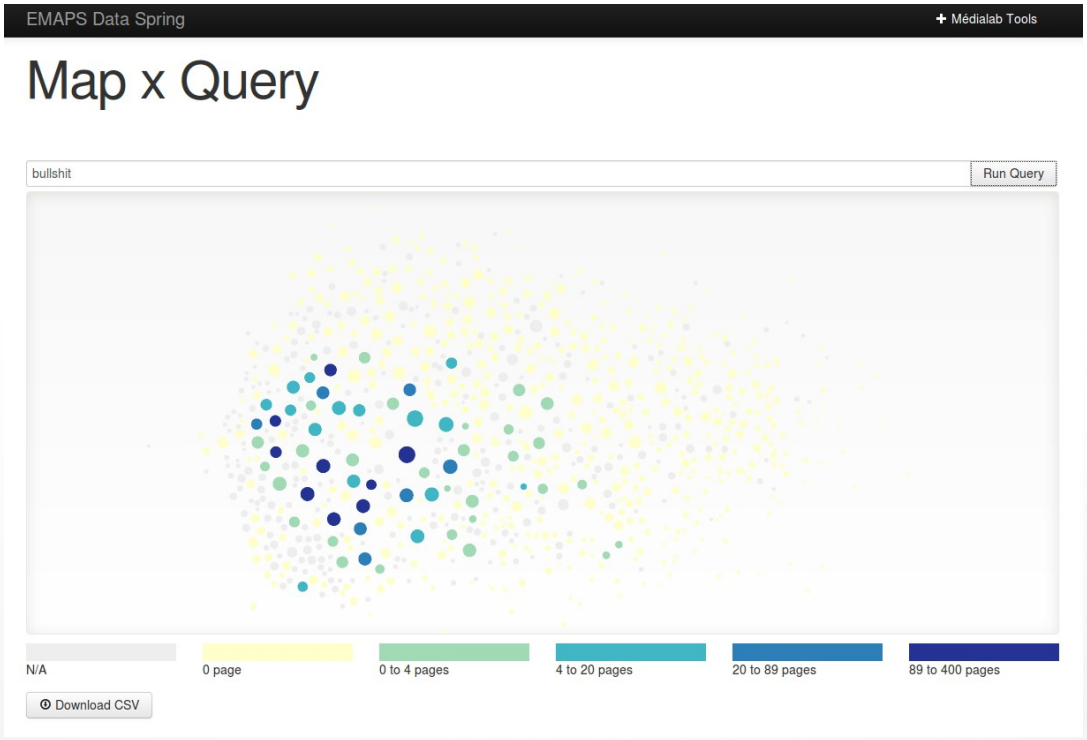
L'interface de Hyphe 2nde version

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