



HAL
open science

Utiliser Hyphe en sciences sociales

Benjamin Ooghe

► **To cite this version:**

Benjamin Ooghe. Utiliser Hyphe en sciences sociales. Colloque DIME-SHS "Des instruments au service de la recherche en sciences sociales", Sciences Po, Sep 2018, Paris, France. hal-03621701

HAL Id: hal-03621701

<https://sciencespo.hal.science/hal-03621701v1>

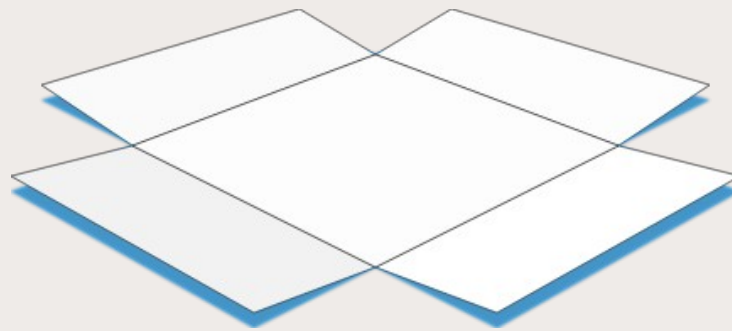
Submitted on 28 Mar 2022

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



Equipex DIME-SHS
ANR-10-EQPX-19-01

DIME Web

Des outils et méthodes numériques pour
exploiter le Web comme terrain d'enquête

Utiliser Hyphe en sciences sociales

28 septembre 2018

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

DIME Web

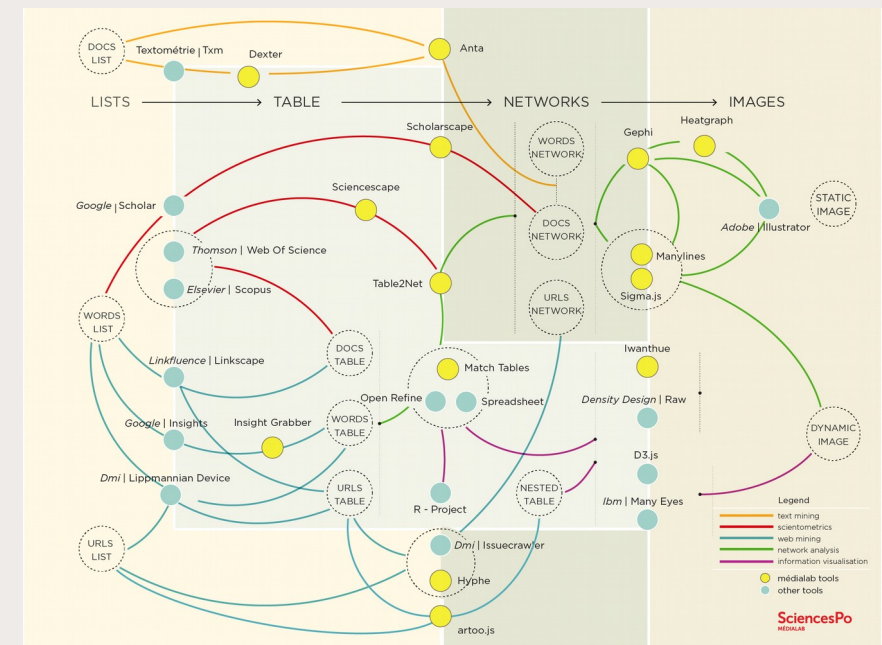
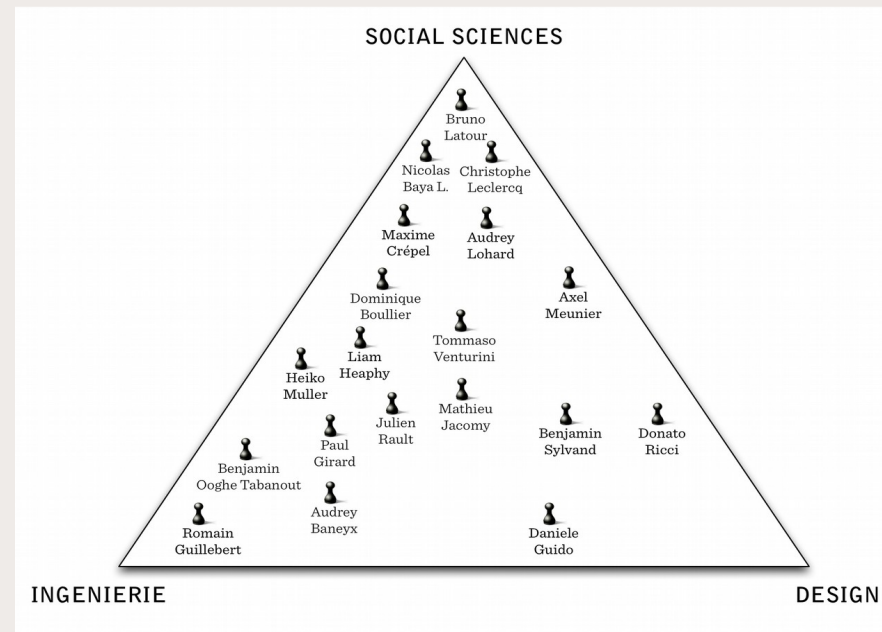
SciencesPo
MÉDIALAB

SciencesPo
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 : le METAT
<http://www.medialab.sciences-po.fr/atelier/>

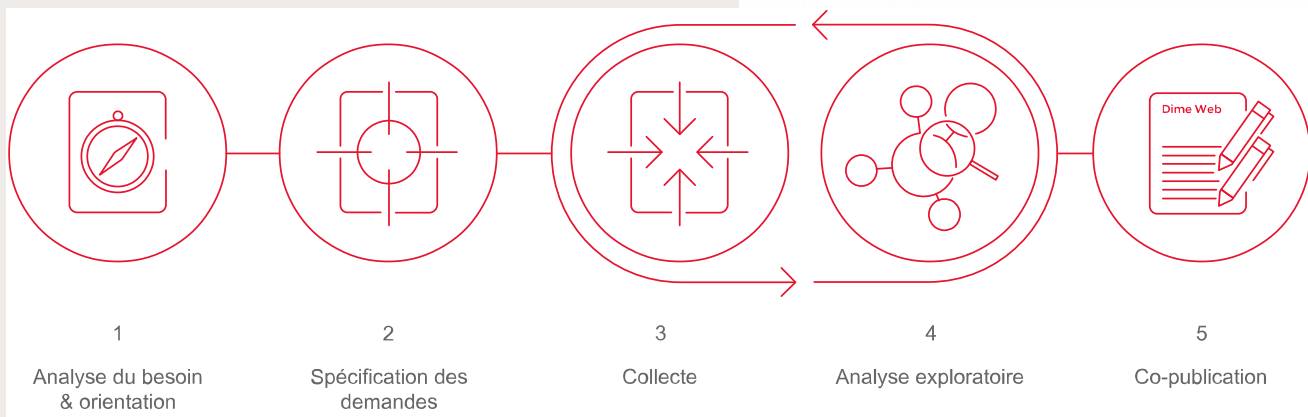
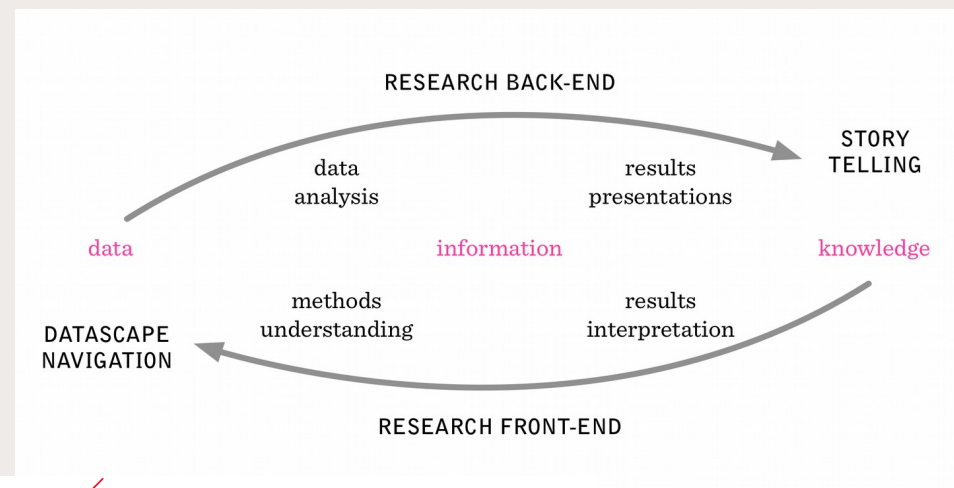


L'instrument DIME Web

<http://dimeweb.dime-shs.sciences-po.fr>

Accompagnement technique et méthodologique à l'utilisation du Web comme nouveau terrain d'enquêtes

- Collecter, enrichir, nettoyer, visualiser et analyser des traces numériques
- Analyse de réseaux, archivage du web, analyse de controverses (ANT)
- Développement d'outils génériques
- Extraction ciblée de contenus
- Publication d'outils Open Source
- Méthodes numériques & itératives
≠ tout automatique



Le Web : une source de données « sales »

Collection de documents web (pages) sur une thématique

→ 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...

Table2Net : faire un réseau à partir d'un CSV

<http://tools.medialab.sciences-po.fr/table2net/>

- Générer un réseau de liens entre éléments à partir des données d'un fichier tableur





Table 2 Net

Load your CSV table

It has to be **comma-separated** and the first row must be dedicated to **column names**.

1. Type of Network

Normal (one type of node) ▼



You will have to choose:

- Which column **X** will define the nodes
- Which column **Y** will define the links

2. Nodes

Which column defines the nodes?

hashtags ▼

Pipe-separated "|" ▼

Sample of nodes extracted with these settings: (🔄 sample)

#goweser #adventurebike #kotaposo #dh #xcbike

3. Links

Which column defines the links?

Row number ▼

One expression per cell ▼

Sample of items extracted with these settings: (🔄 sample)

5252 3621 1816 1847 4562

4. Additional settings

Optional: time series

No temporal data ▼

Select only a column containing integers.

Optional: edge weight

Weight the edges ▼

5. Build the network

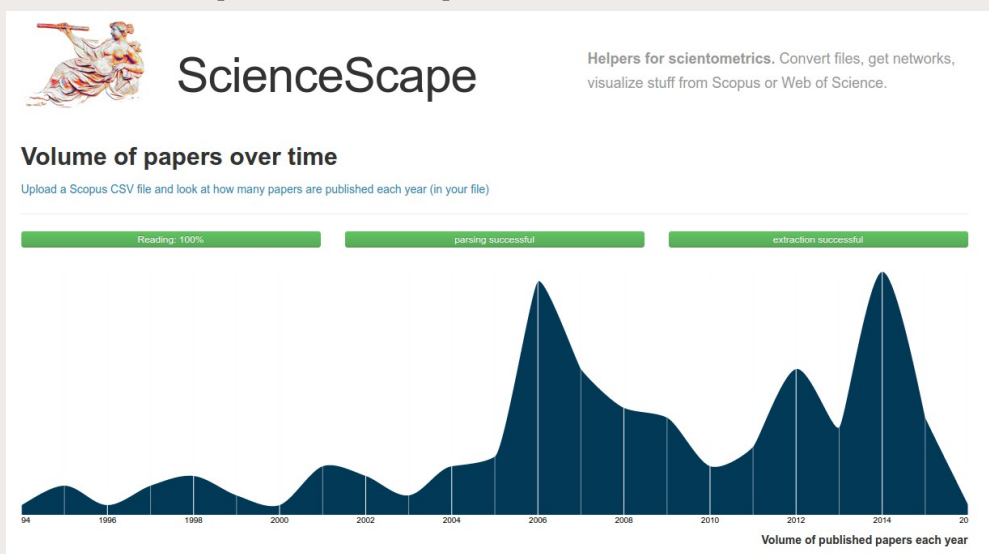
Build and download the network (GEXF)

NB: this may take a while, please be patient.

ScienceScape : scientométrie en quelques clics

<http://tools.medialab.sciences-po.fr/sciencescape/>

- Explorer et analyser les auteurs, mots-clés et revues d'un corpus de publications issu de Scopus ou WebOfScience

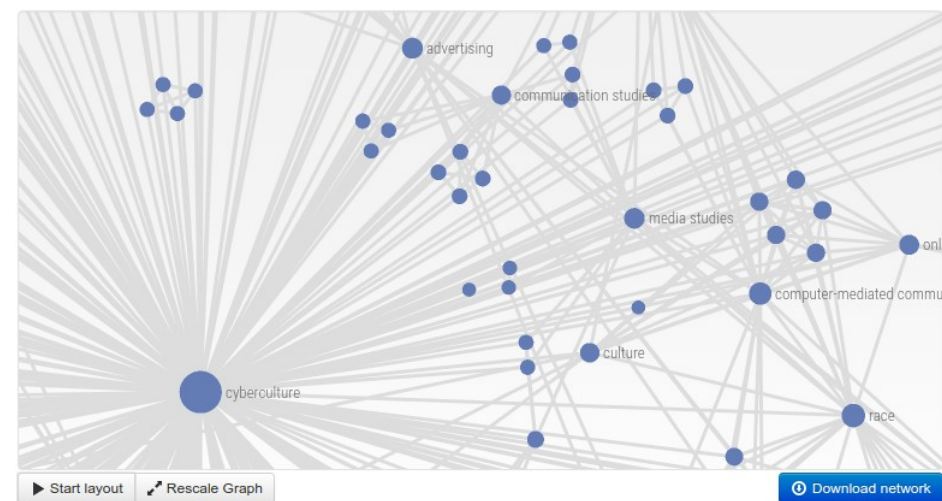


Settings

Type of network

Filtering

Network preview



SeeAlsology : exploration sémantique rapide

<http://tools.medialab.sciences-po.fr/seealsology/>

- Explorer le réseau des liens présents dans les sections « Voir aussi », « Articles connexes » des pages Wikipedia

The screenshot shows the Wikipedia article for 'Humanités numériques'. The page includes a sidebar with navigation options, a main content area with a definition and a list of software tools, and a 'Voir aussi' section. A small image of a desk with a laptop and books is also visible.

Humanités numériques

Les **humanités numériques** (ou *digital humanities*, abrégées "DH", voire **humanités digitales**²) sont un domaine de recherche, d'enseignement et d'ingénierie au croisement de l'informatique et des arts, lettres, sciences humaines et sciences sociales.

Définition [modifier | modifier le code]

Les humanités numériques peuvent être définies comme l'application du « savoir-faire des technologies de l'information [et de l'informatique/infosciences] aux questions de sciences humaines et sociales »³.

Logiciels [modifier | modifier le code]

- Gephi est un logiciel libre *open source*, issu du projet *e-Diaspora*, permettant la visualisation, l'analyse et l'exploitation en temps réel de données relationnelles ou réseaux.
- IRaMuTeQ est un logiciel libre d'analyse de texte, développé par **Pierre Ratinaud**.
- **Voyant Tools** permet de visualiser et d'explorer des textes
- Prospero (PROgramme de Sociologie Pragmatique, Expérimentale et Réflexive sur Ordinateur - © Doxa) est un logiciel d'analyse de données textuelles qualifié par ses concepteurs de technologie littéraire pour les sciences humaines. Le logiciel a été conçu par le sociologue Francis Chateauraynaud et l'informaticien Jean-Pierre Charriau.
- **Phlcarto** est un logiciel de cartographie. Le code n'en est pas libre, mais le logiciel est gratuit (freeware). Il fonctionne sur Windows.
- OpenRefine est un logiciel libre et gratuit de lissage de données (anciennement nommée Google refine).
- Le projet DIRT recense de très nombreux logiciels: DIRT@ [archive] (*Digital Research Tools - en Anglais*).

Articles connexes [modifier | modifier le code]

- Bibliothèque numérique
- Fouille de textes
- Littérature numérique
- Logométrie
- Moteur de recherche

The screenshot shows the SeeAlsology tool interface. It includes a text input field for pasting a list of Wikipedia articles, a 'START CRAWLING' button, and a network graph visualization. The graph shows a central node 'Humanités numériques' connected to other nodes like 'Bibliothèque numérique', 'Logométrie', and 'Moteur de recherche'. A legend indicates node levels: seeds (red), level -1 (orange), level 0 (blue), level 1 (dark blue), and level 2 (light blue).

Paste your list of wikipedia articles here or [try an example](#)

https://fr.wikipedia.org/wiki/Humanit%C3%A9s_num%C3%A9riques

Stop words (press enter or separate the works with a comma)

Wikipedia: x Category: x File: x wikisource: x Commons: x
liste d x index d x catégories d x portail x désambiguisation x
résumé d x Catégorie: x Fichier: x add a word and press Enter

Distance Parent links

START CRAWLING **DOWNLOAD** **CLEAR CACHE**

Click a node to visit it on Wikipedia

● seeds ● level -1 ● level 0 ● level 1 ● level 2

Ctrl+Click a node to add it to the seeds

Humanités numériques

Bibliothèque numérique

Logométrie

Moteur de recherche

Recherche d'information

Manifeste des Digital Humanities (2010)

CatWalk : sélection qualitative de tweets

<https://medialab.github.io/catwalk/>

- Passer en revue rapidement « à la Tinder » tous les tweets d'un CSV pour décider de les inclure / exclure d'un corpus

The screenshot displays the CatWalk web interface. At the top left, the text 'CATWALK' is visible. To its right are navigation buttons: 'prev', a central input field containing '0', and 'next'. Further right is a 'Download' button with a counter showing '0', '434', and '2'. Below this is a large green button labeled 'IN'. The main content area features a tweet from 'RE-WORK @teamrework' with a 'Follow' button. The tweet text reads: 'Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free ow.ly/4nfo2S #AI @open_ai' and is dated '7:15 PM - 29 Apr 2016'. It includes a photo of Elon Musk and a link to a Wired article. To the right of the tweet is a vertical control panel with buttons: 'previous', 'next', 'IN' (highlighted in green), 'OUT' (highlighted in red), 'UNDECIDED' (highlighted in grey), and 'save'. At the bottom of the interface, there is a footer with the user's profile picture and name '@teamrework', and a summary of the tweet content: 'Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free <http://ow.ly/4nfo2S> #AI @open_ai'.

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

Install Google Bookmarklets
Drag & drop images below into your bookmark bar:

Redirect to Classic Google
Which language? en
How many results per page? 100
You will be redirected to the following url:
`https://encrypted.google.com/search?q=digital%20humanities&hl=en&num=100&start=0`
Redirect me!

Extract Classic Google Results
Search for "digital humanities"
page 0 (with up to 100 uris per page)
103 new results in this page
Keep existing results & continue to the next page
Download CSV with 103 uris

→ « Import urls » dans Hyphe

Un projet logiciel phare : Hyphe

- 10 releases entre juillet 2013 (v.0.0.0) et août 2018 (v1.0.3)

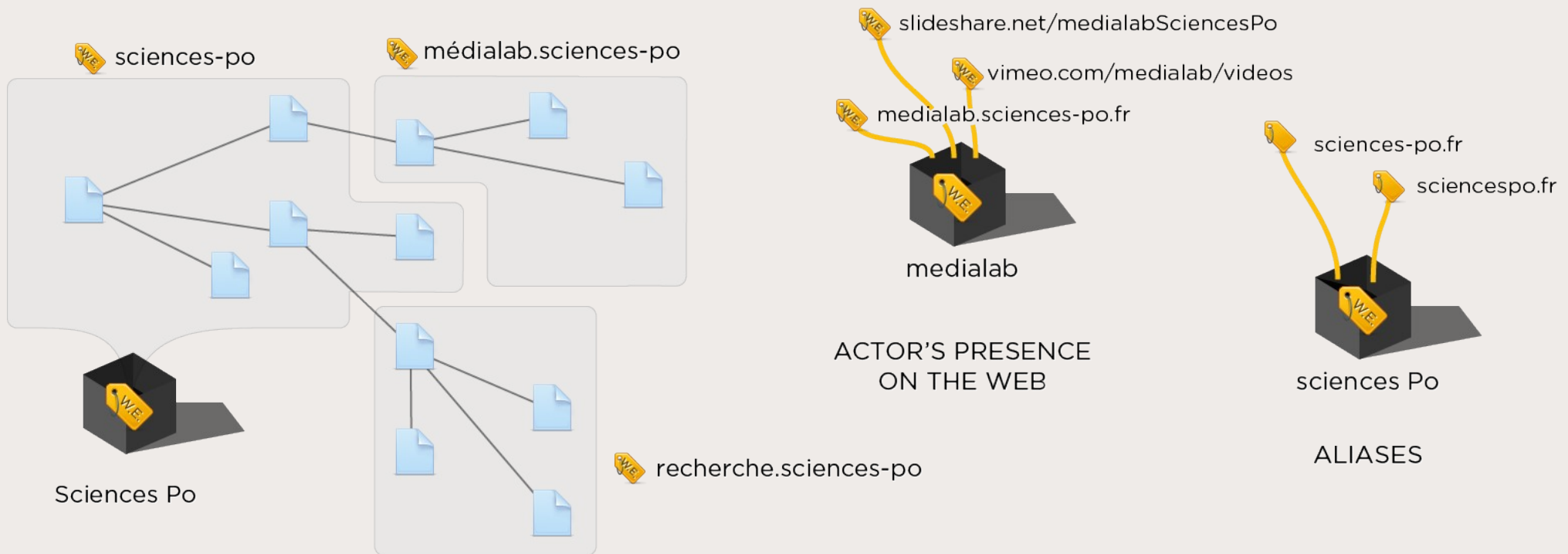


- Plusieurs dizaines de présentations, sessions pratiques et formations assurées
- + de 50 déploiements externes à notre connaissance (France, UK, DK, USA...)
- + de 150 étudiants formés à l'analyse du web chaque année avec Hyphe
- + de 500 utilisateurs testant la démo chaque année

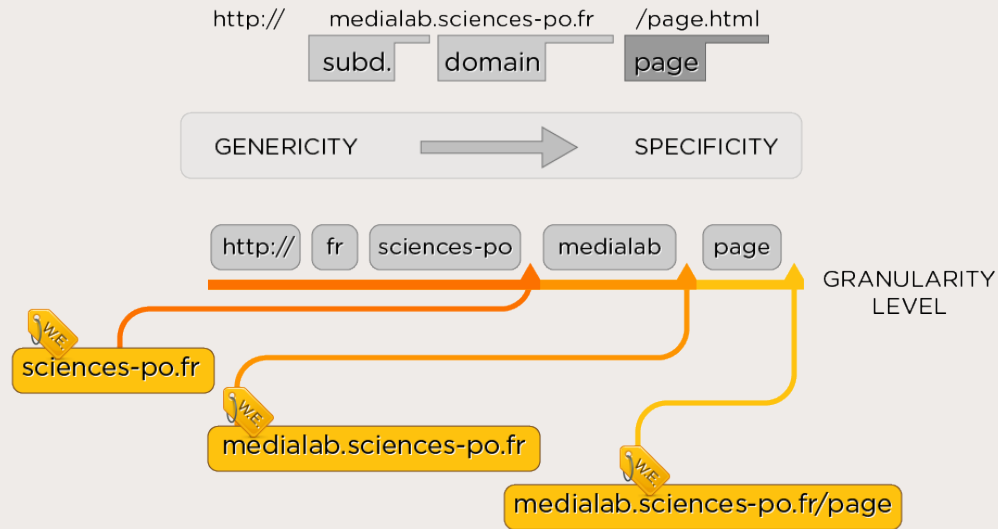
Principes méthodologiques : « WebEntités »

Comment gérer la diversité de granularité des sites web ?

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



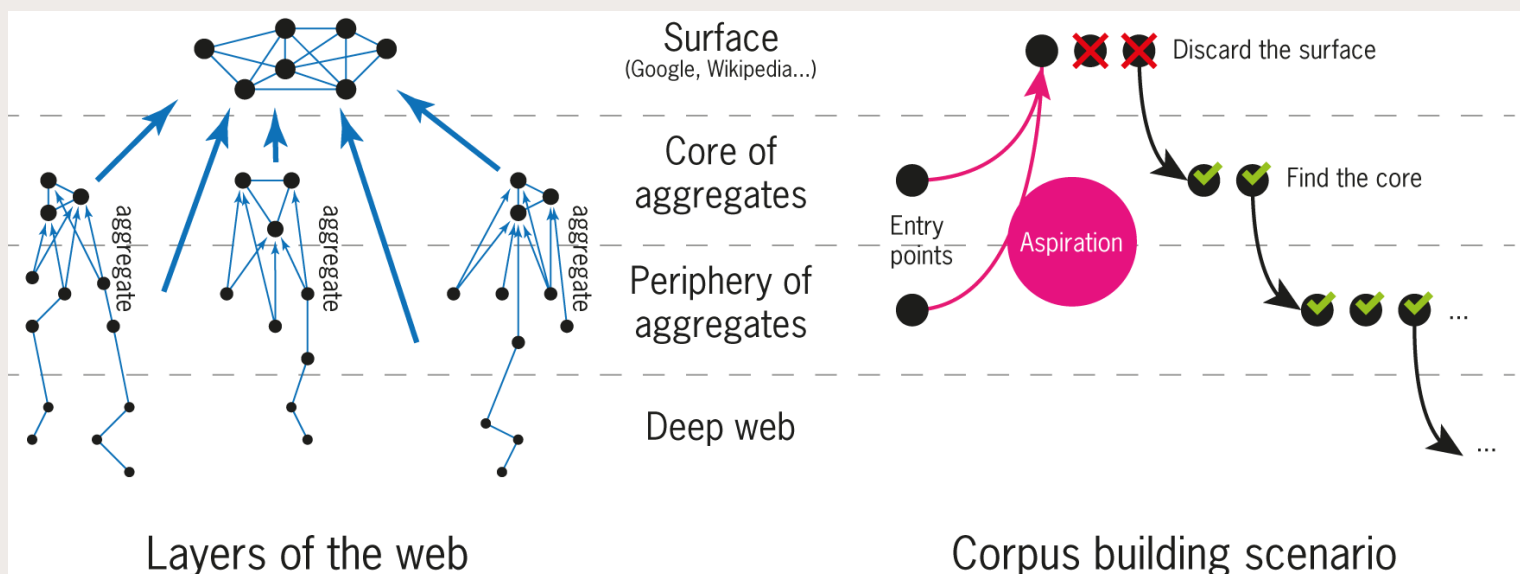
Principes méthodologiques : « WebEntités »



41	Amnesty.fr	http .fr amnesty www.
42	Facebook.com /.../326366925310	http .com facebook www. /pages /Andr%C3%A9-... /326366925310 Same web entity defined rows 82, 130, 150, 189, 249, 388, 389, 392, 393, 424, 475, 483, 488, 493, 640, 642, 659, 668, 690, 707, 719, 779, 966, 972 and 989
43	Annuairemairie.com	http .com annuairemairie www.
44	Marianne2.fr /hervenathan	http .fr marianne2 www. /hervenathan Same web entity defined rows 651, 895 and 896
45	Anticor.org	http .org anticor
46	Desgouilles.fr	http .fr desgouilles david.

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 (numérique) → exclure ou inclure
- Décisions éditoriales classiques de type gestion documentaire

PROSPECT 4,890 DISCOVERED

Search APPLY CHANGES CANCEL

Distribution of citations (log scale)

NAME	CITED ↑
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Google.fr	23
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Instagram.com	19
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Free.fr	16
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Wordpress.org	16
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Wp.com	13
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Blogger.com	12
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Twitter.com /home	12
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Gravatar.com	11
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Legifrance.gouv.fr	10
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Wordpress.com	10
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Collectifmarianne.fr	9
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> Collectifracine.fr	9

1 SET TO IN
Collectifmarianne... ✕

CRAWL

1 SET TO UNDECIDED
Legifrance.gouv.fr ✕

4 SET TO OUT
Gravatar.com ✕
Google.fr ✕

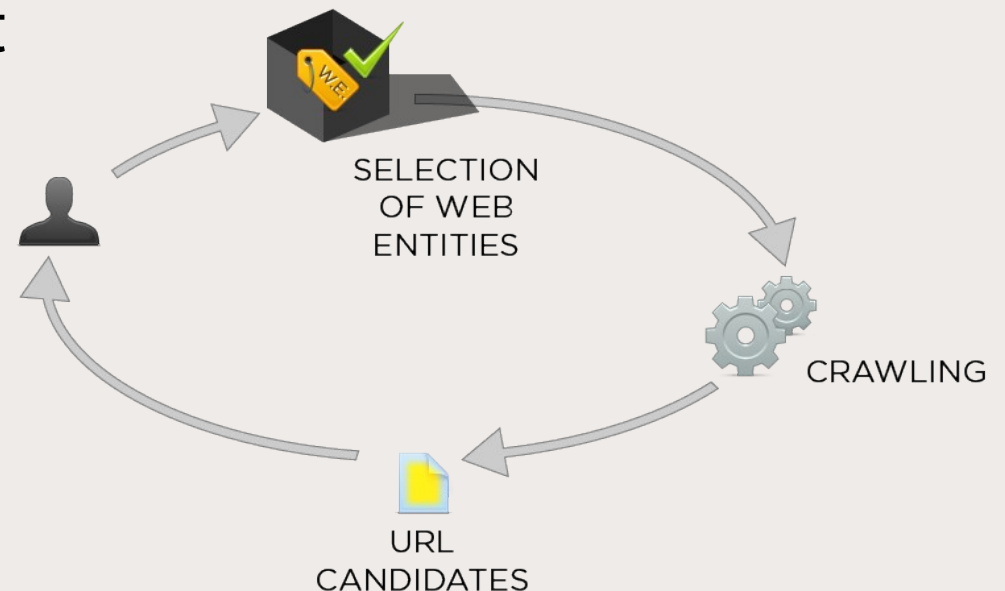
Set to UNDECIDED

Principes méthodologiques : « Itérations quali »

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

→ Quand s'arrêter ?

→ Seuil de citation



HyBro : un browser dédié à la curation de corpus

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

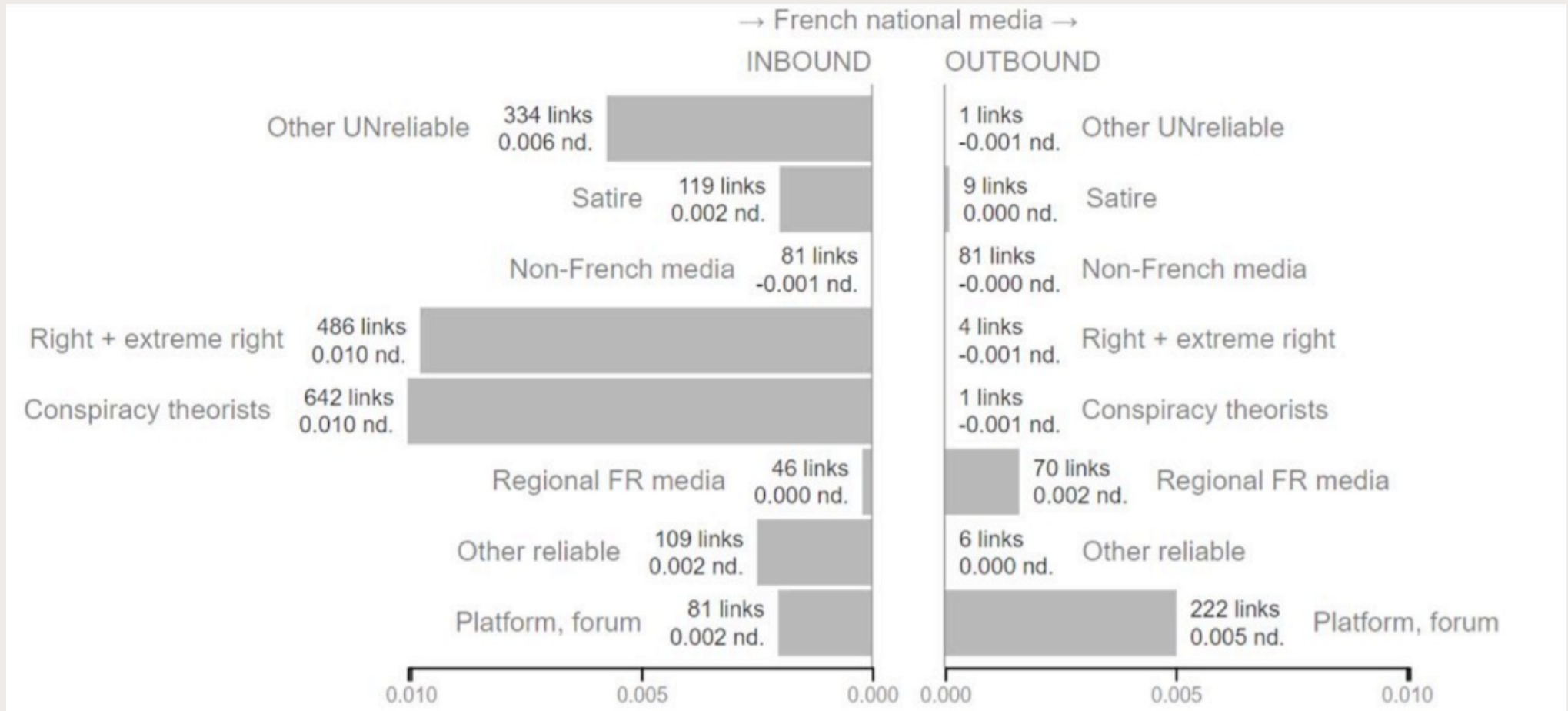
Héritier du « NaviCrawler » : un navigateur web connecté à Hyphe :

→ prospection et catégorisation in-situ (terrain numérique)

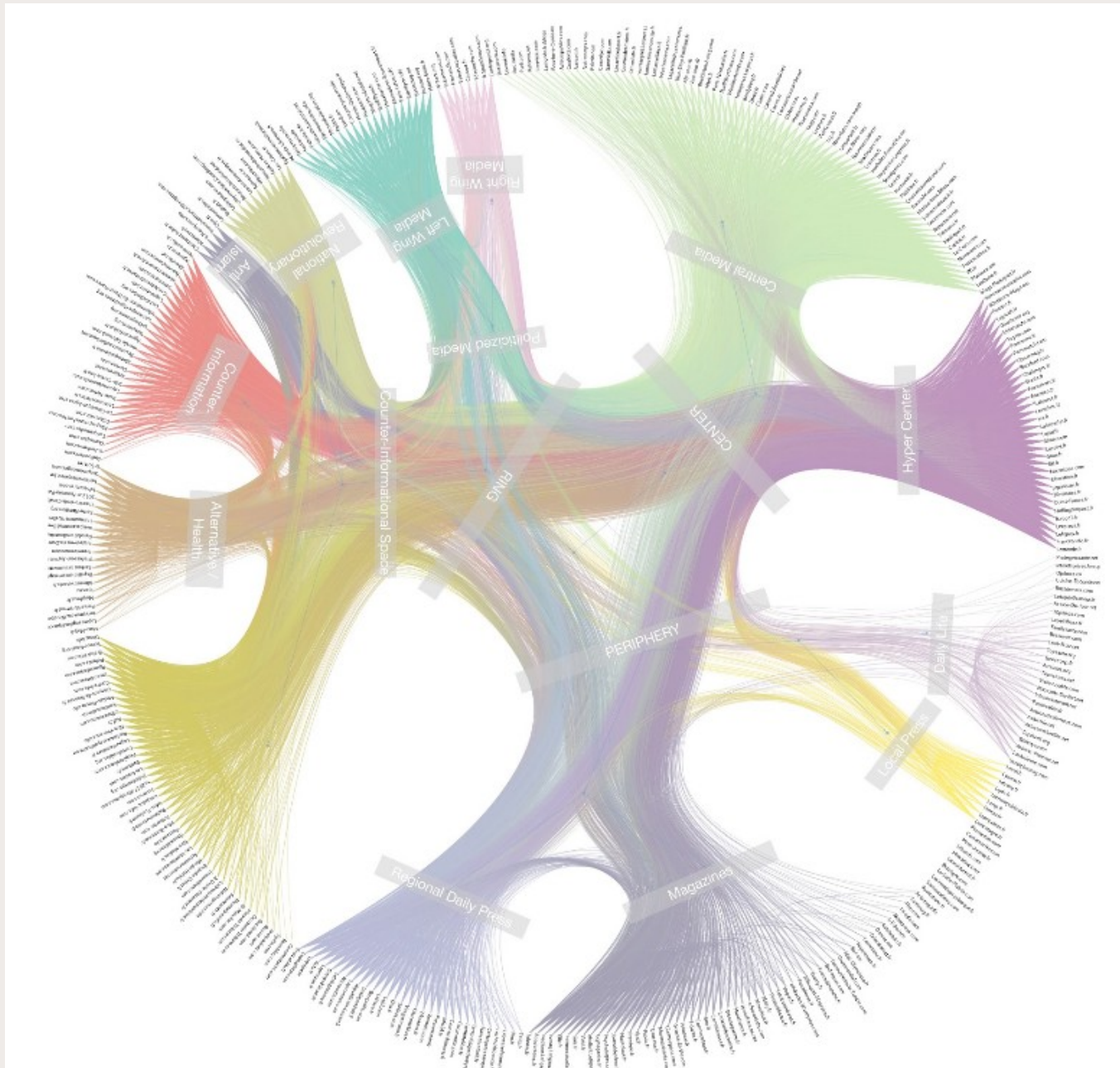
→ formation d'étudiants et lycéens au web (IDEFI FORCCAST)

The screenshot shows the Hyphe Browser interface. At the top, there's a search bar with 'Free.fr' and a status bar with 'PROSPECTION' and other filters. Below the search bar, there's a sidebar with 'Free.fr' and various filters like 'WebEntité', 'Statut', 'Contexte', and 'Pages citées'. The main content area displays a web page from 'http://etienne.chouard.free.fr/Europe/index.php' with a title 'Le plan C : instituer une vraie démocratie par une Constitution d'origine Citoyenne.' and a large image of an amphitheater. The page content includes sections for 'Venez participer', 'Présentation', and 'Vos recherches sur le plan C'. The interface is designed for navigating and analyzing web content.

Analyser la structuration interne des corpus



Explorer les dynamiques de polarisation



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
- Utiliser les technologies modernes pour crawler les sites en JavaScript (Facebook, etc.)
- Outils de contrôle qualité des crawls et du corpus
- Outils d'archivage et présentation des corpus finalisés
- Mise à disposition à la demande automatisée (SAAS)

