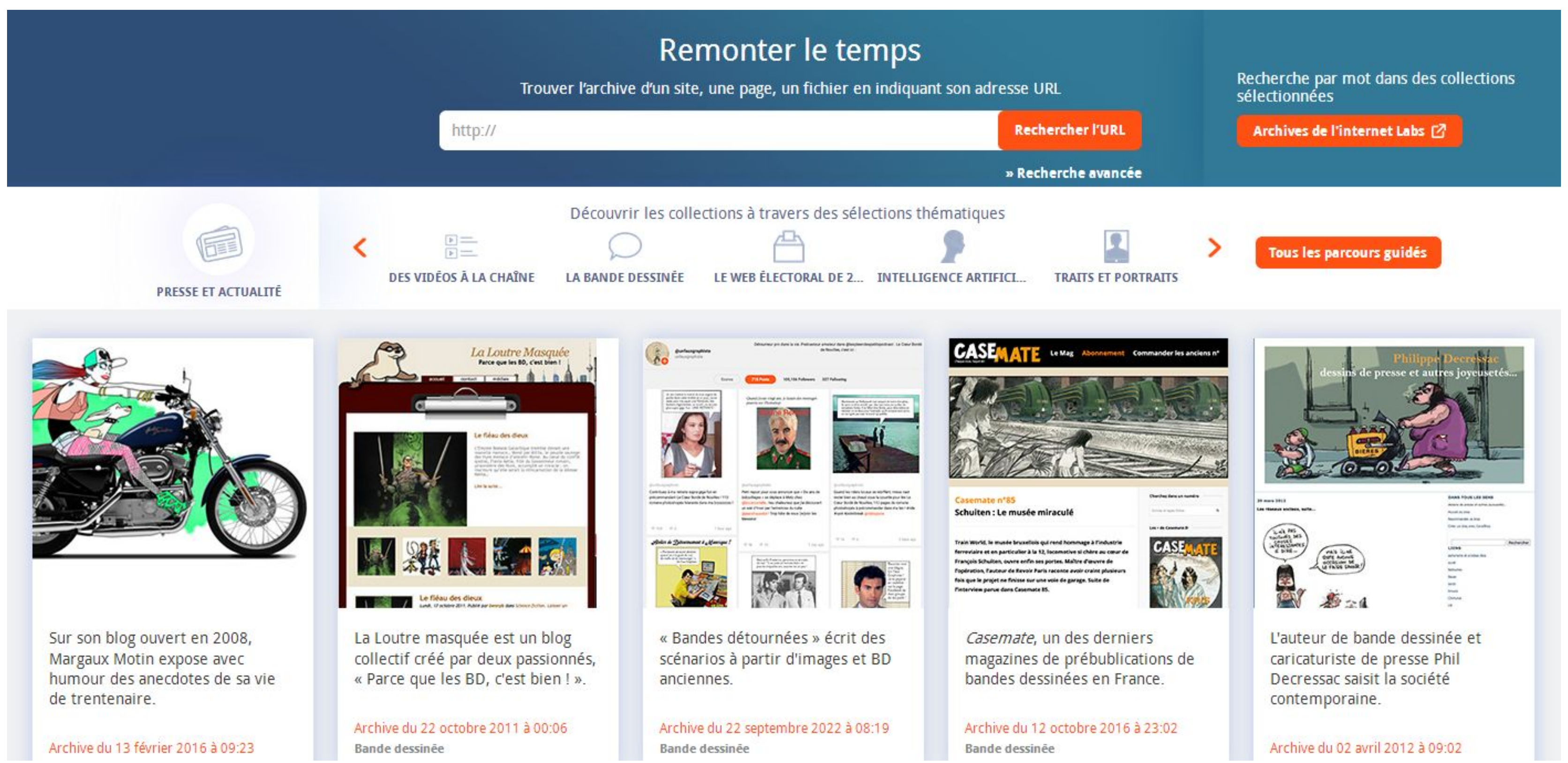
Exploring web archives using Hyphe, a research oriented web crawler

Sara Aubry, BnF & Benjamin Ooghe-Tabanou, médialab Sciences Po

IIPC Web Archiving Conference, April 25-26th 2024, Paris

BnF Archives de l'internet



Remonter le temps

Trouver l'archive d'un site, une page, un fichier en indiquant son adresse URL

http://

Recherche par mot dans des collections sélectionnées

» Recherche avancée

Découvrir les collections à travers des sélections thématiques

Sur son blog ouvert en 2008, Margaux Motin expose avec humour des anecdotes de sa vie de trentenaire.
Archive du 13 février 2016 à 09:23

La Loutre masquée est un blog collectif créé par deux passionnés, « Parce que les BD, c'est bien ! ».
Archive du 22 octobre 2011 à 00:06
Bande dessinée

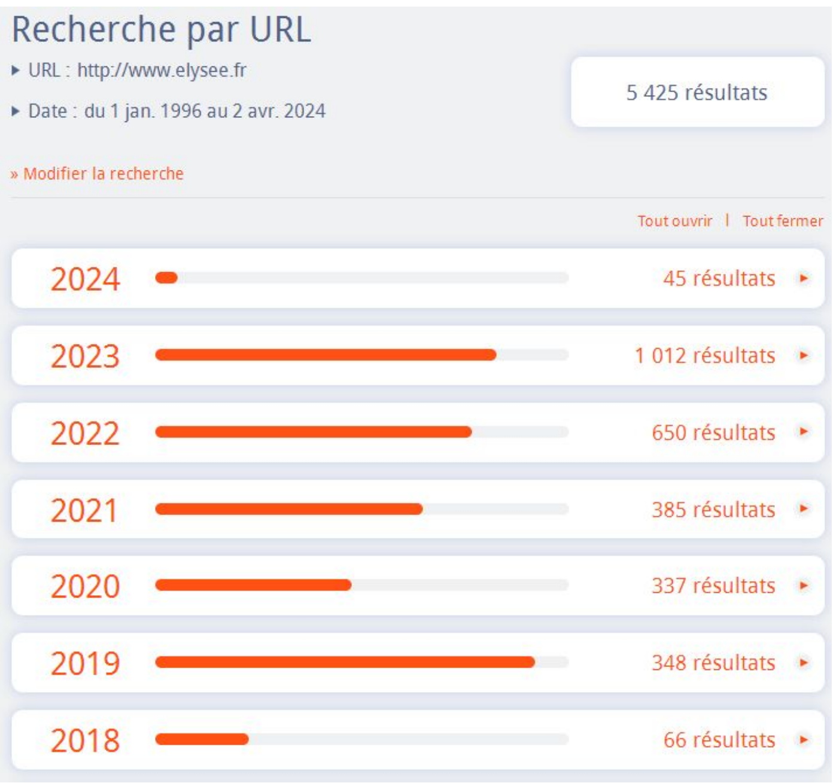
« Bandes détournées » écrit des scénarios à partir d'images et BD anciennes.
Archive du 22 septembre 2022 à 08:19
Bande dessinée

Casemate, un des derniers magazines de prébublications de bandes dessinées en France.
Archive du 12 octobre 2016 à 23:02
Bande dessinée

L'auteur de bande dessinée et caricaturiste de presse Phil Decressac saisit la société contemporaine.
Archive du 02 avril 2012 à 09:02

National Library of France web archives

- From 1996 to now
- 52 billion URLs, 2 PB of data
- Growing with annual crawls of million websites (5.6 mio in 2023) and regular crawls of thematic, event and media focused websites (currently 70,000)



Recherche par URL

- URL : http://www.elysee.fr
- Date : du 1 jan. 1996 au 2 avr. 2024

5 425 résultats

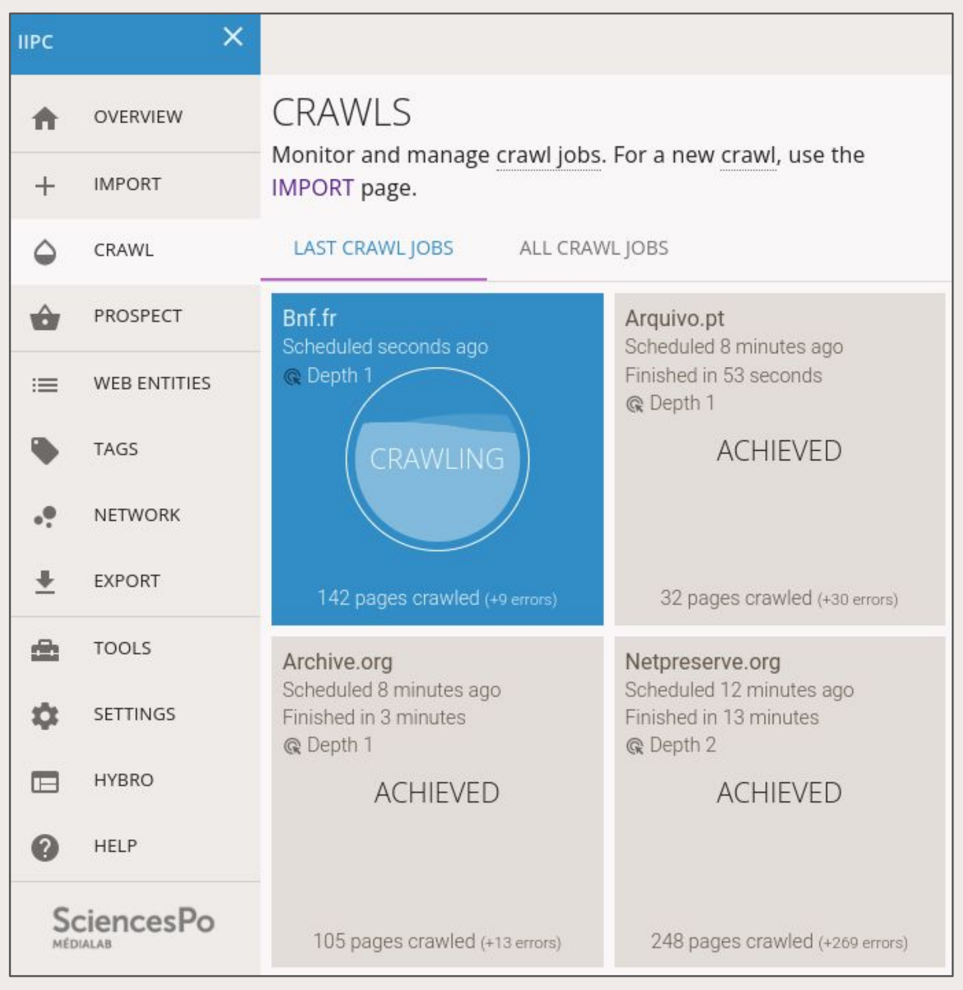
» Modifier la recherche

Tout ouvrir | Tout fermer

2024	45 résultats
2023	1 012 résultats
2022	650 résultats
2021	385 résultats
2020	337 résultats
2019	348 résultats
2018	66 résultats

hyphe

An open source tool for social scientists to create corpuses of web actors and map the hyperlinks between them



CRAWLS

Monitor and manage crawl jobs. For a new crawl, use the IMPORT page.

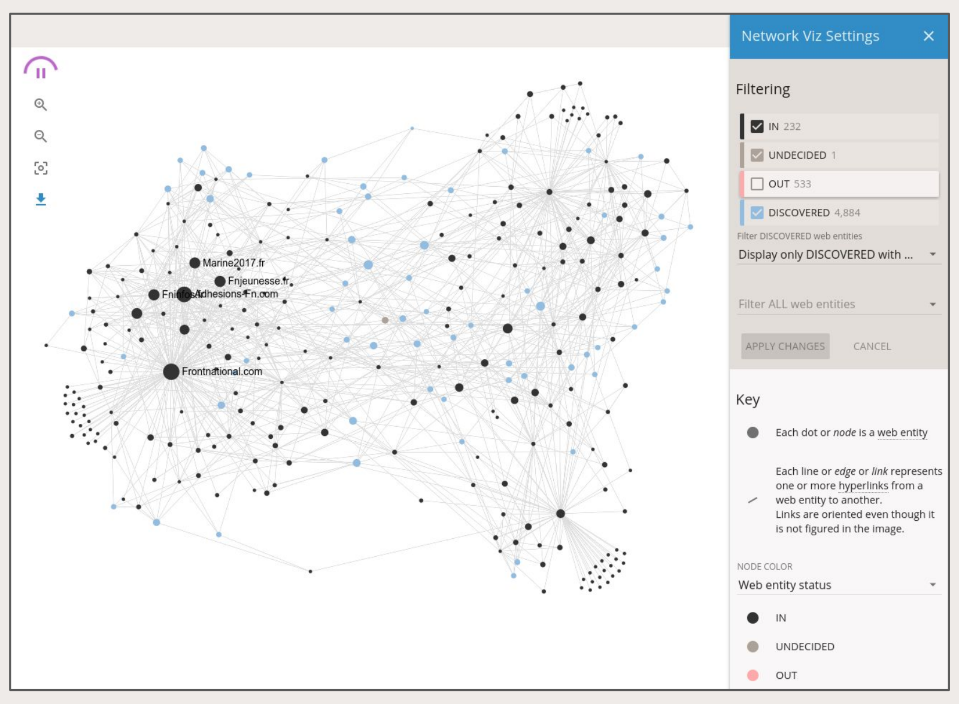
LAST CRAWL JOBS | ALL CRAWL JOBS

Bnf.fr
Scheduled seconds ago
Depth 1
CRAWLING
142 pages crawled (+9 errors)

Arquivo.pt
Scheduled 8 minutes ago
Finished in 53 seconds
Depth 1
ACHIEVED
32 pages crawled (+30 errors)

Archive.org
Scheduled 8 minutes ago
Finished in 3 minutes
Depth 1
ACHIEVED
105 pages crawled (+13 errors)

Netpreserve.org
Scheduled 12 minutes ago
Finished in 13 minutes
Depth 2
ACHIEVED
248 pages crawled (+269 errors)



Network Viz Settings

Filtering

- ☑ IN 232
- ☑ UNDECIDED 1
- ☐ OUT 533
- ☑ DISCOVERED 4,884

Filter DISCOVERED web entities: Display only DISCOVERED with ...

Filter ALL web entities

APPLY CHANGES | CANCEL

Key

● Each dot or *node* is a web entity

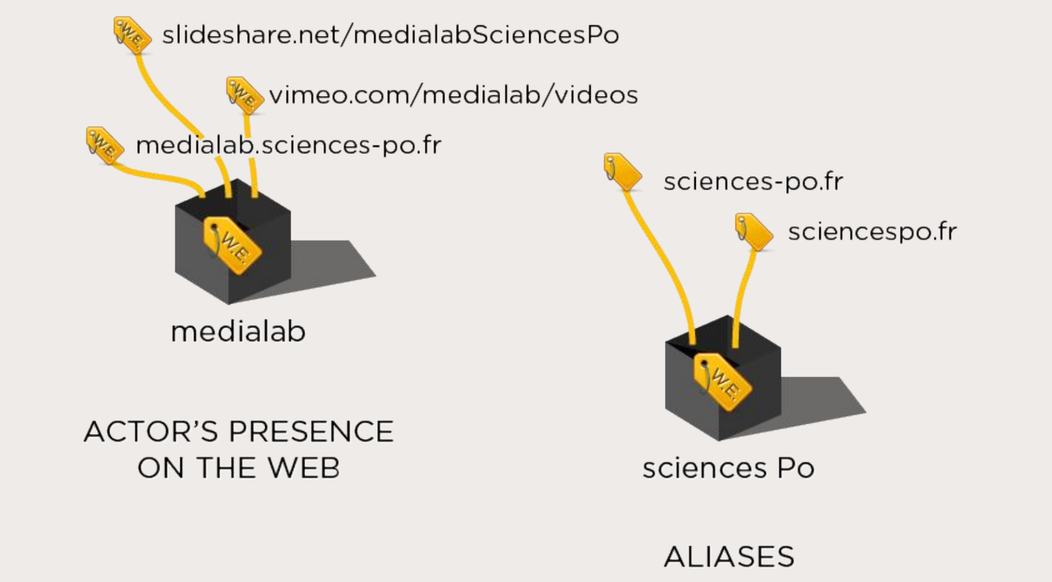
Each line or *edge* or *link* represents one or more hyperlinks from a web entity to another. Links are oriented even though it is not figured in the image.

NODE COLOR: Web entity status

- ● IN
- ● UNDECIDED
- ● OUT

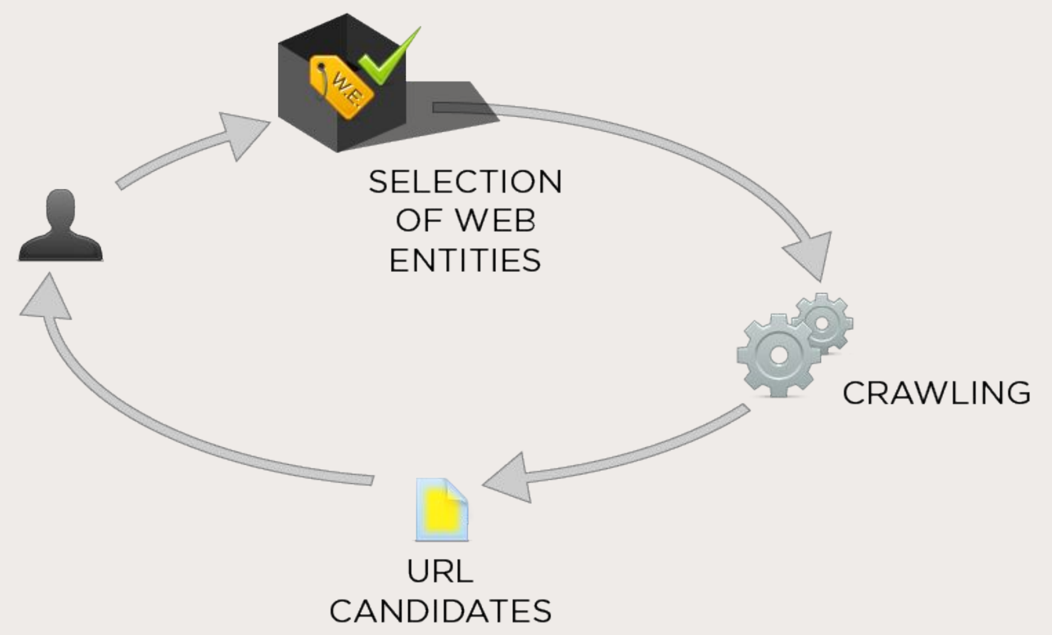
Methodological principles:

- ***Web entities***: fine definition of actors beyond "websites"

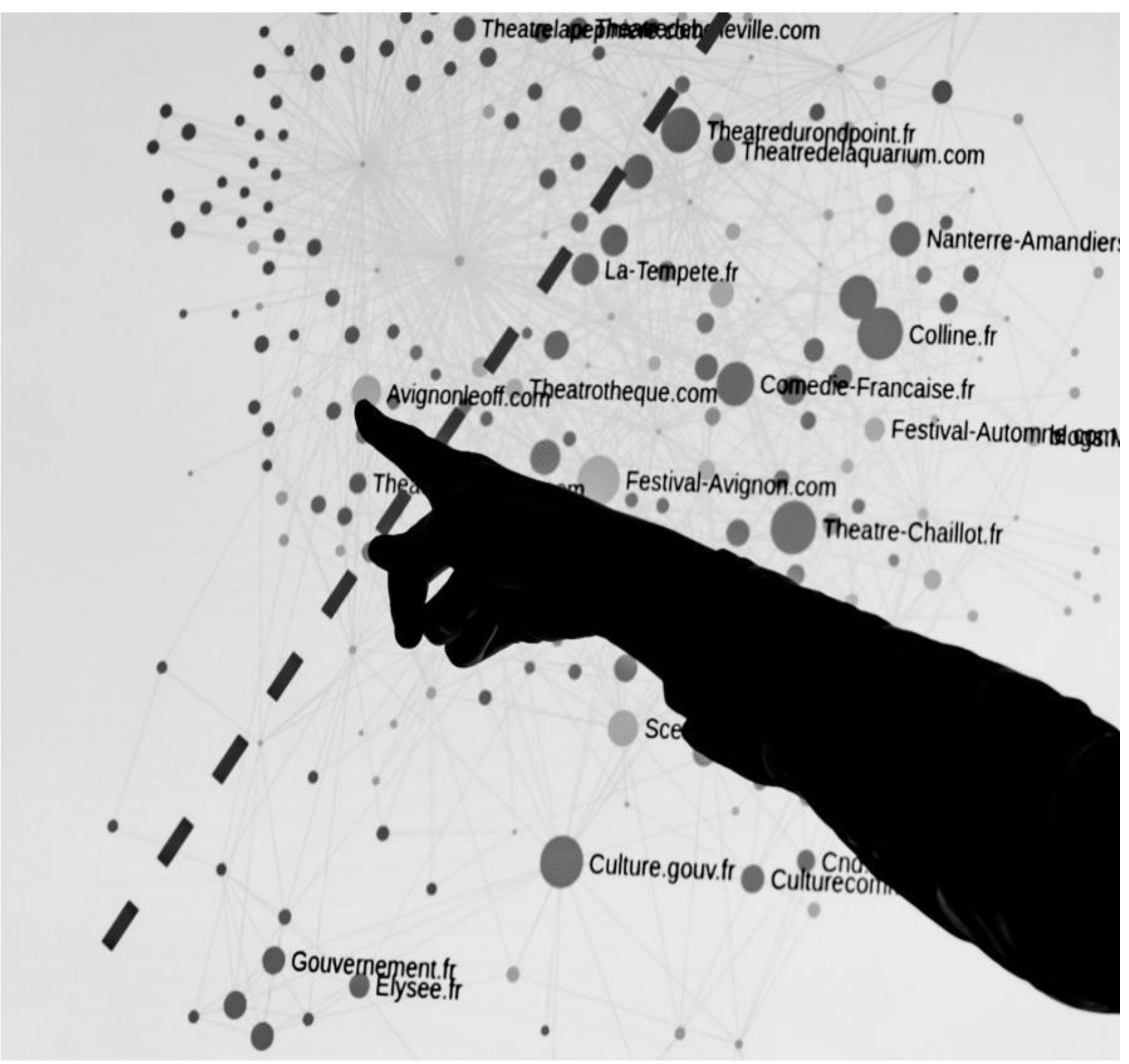


slideshare.net/medialabSciencesPo
vimeo.com/medialab/videos
medialab.sciences-po.fr
medialab
ACTOR'S PRESENCE ON THE WEB
sciences-po.fr
sciencespo.fr
sciences Po
ALIASES

- ***Prospection loop***: iterative curation of discovered actors



Connecting Hyphe to web archives, an outcome of the project RESPADON

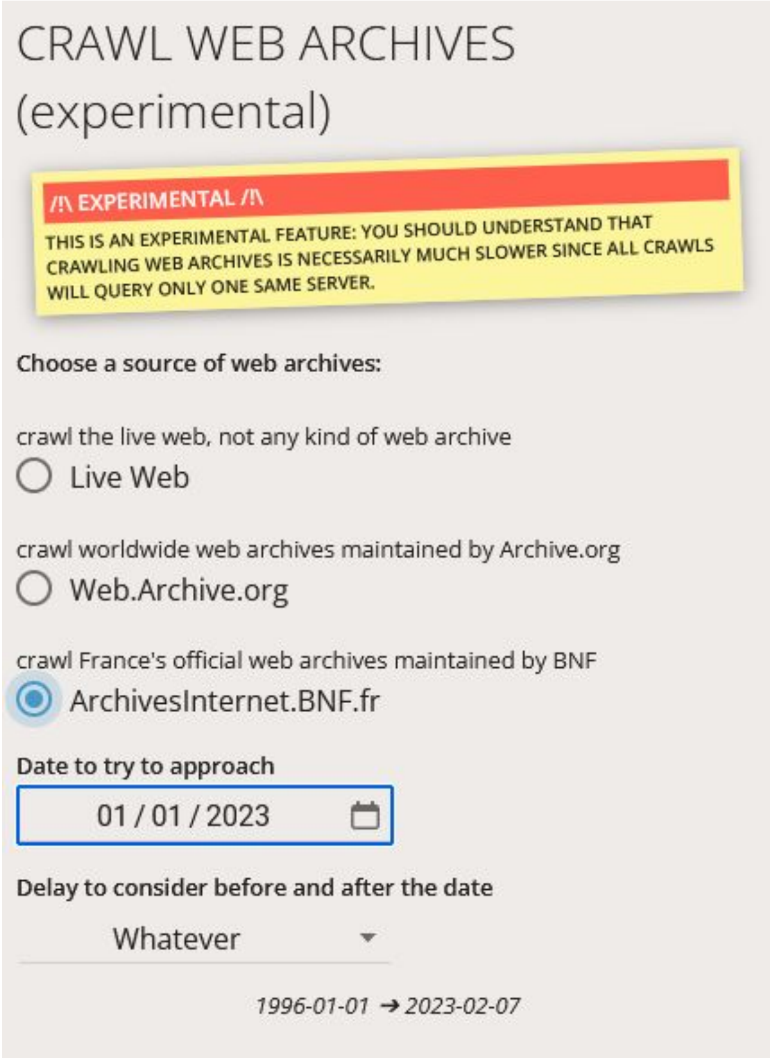


Allowing Hyphe to:

- Target a specific date and a time frame around this date
- Browse the archive via a wayback-like system
- Target precise capture dates to avoid drifting over time
- Support permalinks and archival URLs

Addressing new research methods & use cases:

- Map online communities from the past
- Compare time snapshots & study the structural transformations
- Complete live web with archived material of disappeared websites



CRAWL WEB ARCHIVES (experimental)

/!\ EXPERIMENTAL /!\
THIS IS AN EXPERIMENTAL FEATURE: YOU SHOULD UNDERSTAND THAT CRAWLING WEB ARCHIVES IS NECESSARILY MUCH SLOWER SINCE ALL CRAWLS WILL QUERY ONLY ONE SAME SERVER.

Choose a source of web archives:

crawl the live web, not any kind of web archive
○ Live Web

crawl worldwide web archives maintained by Archive.org
○ Web.Archive.org

crawl France's official web archives maintained by BNF
● ArchivesInternet.BNF.fr

Date to try to approach
01 / 01 / 2023

Delay to consider before and after the date
Whatever

1996-01-01 → 2023-02-07

Hyphe is now compatible with web archives from BnF, INA, Arquivo.pt & Internet Archive

Experiment crawling through the archives in a DataSprint

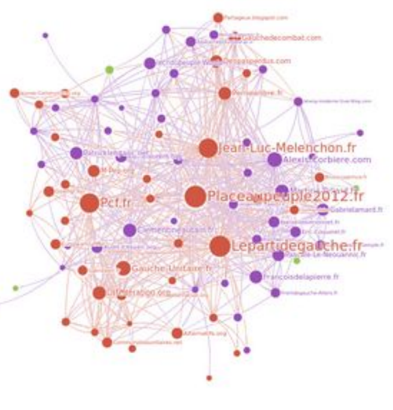
- Hands on workshop during 5 days within BnF DataLab
- Dive into data to explore hypothesis on research questions
- Gather complementary skills & expertises:
  research, web archiving, engineering, design, digital methods
- Experiment with different methodological and exploratory approaches
- Find preliminary insights

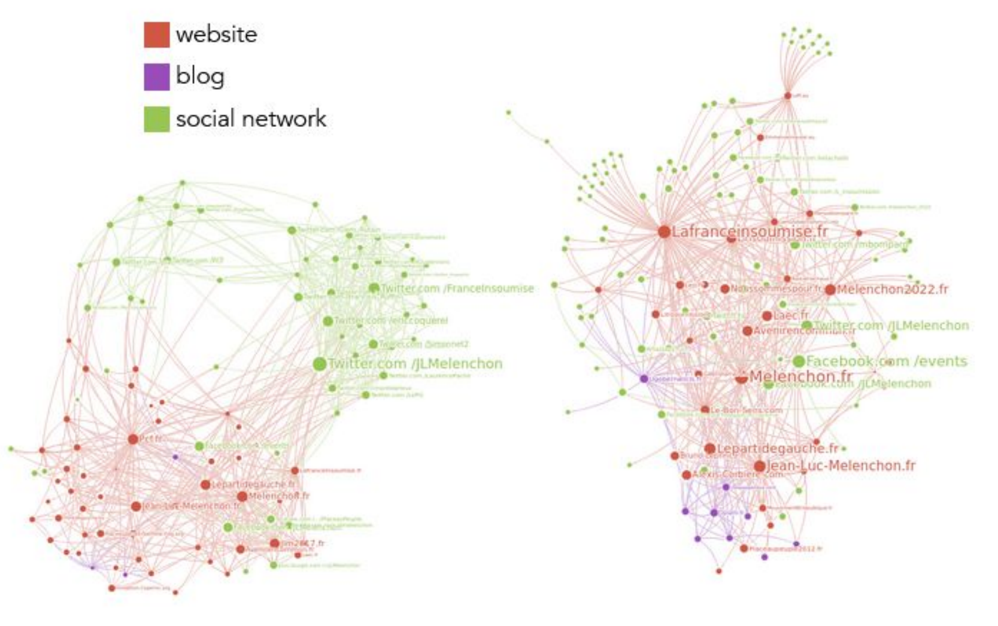
Focus on a presidential candidate's campaigns: Jean-Luc Mélenchon

Hypothesis: online communities changed across 2012, 2017 & 2022 campaigns

Methodological approach:

- Define time windows depending on available archives
- Start coherent crawls from stable sources (Wikipedia)
- Define common rules to select and include actors
- Use ontologies to tag actors (editorial form, nature)
- Produce graphs to visualize differences
- Identify disappeared actors, newcomers and resilient ones





- Hyphe's source code, demo & references: https://github.com/medialab/hyphe
- BnF's Web Archives introduction: https://www.bnf.fr/fr/depot-legal-du-web
- ResPaDon DataSprint results: https://respadon.medialab.sciencespo.fr
- Contacts: sara.aubry@bnf.fr benjamin.ooghe@sciencespo.fr
- Photo credits: Caroline Maufroy / Sciences Po