

Hyphe

curation-oriented
web crawler for
the social sciences

Mathieu Jacomy

Paul Girard

Benjamin Ooghe-Tabanou

Tommaso Venturini



Supported by
Equipex DIME-SHS

ANR-10-EQPX-19-01

SciencesPo
MÉDIALAB

The screenshot shows two windows of the Hyphe software. The top window is titled 'Import URLs' and shows a list of URLs being imported. The bottom window is titled 'Define Web Entities' and shows a list of entities being defined with their URLs and names.

Import URLs

Define entities

The screenshot shows a grid of crawling tasks. Some tasks are labeled 'PENDING', while others like 'Web Entity Loading' and 'Crawling' are in progress. A task for 'L3s' is shown as 'ACHIEVED'.

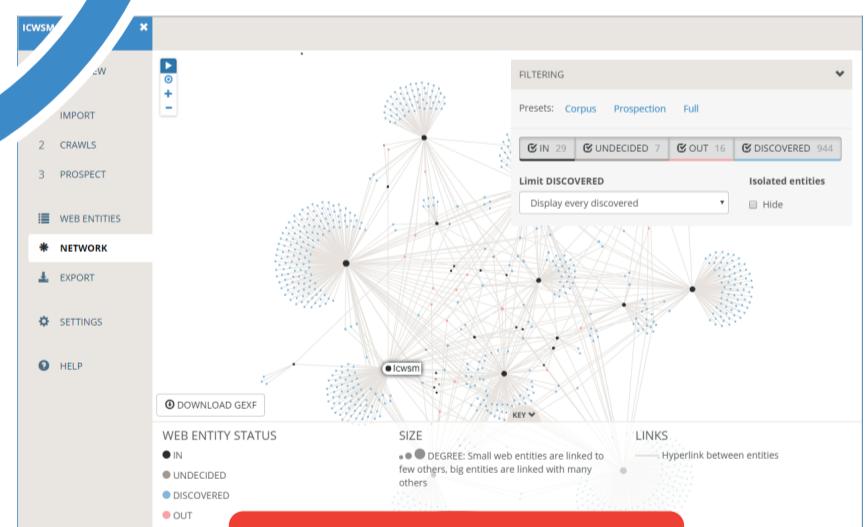
Crawl

Prospect

iterative curation

Monitor corpus

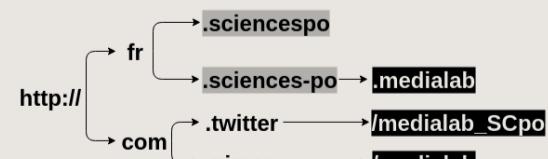
The screenshot shows a list of discovered web entities under the 'PROSPECT' tab. Entities are categorized as IN, UND., OUT, or DISCOVERED. A search bar at the top allows users to type a query.



Example corpus

Custom
Web-entity
Granularity

tree of URL stems

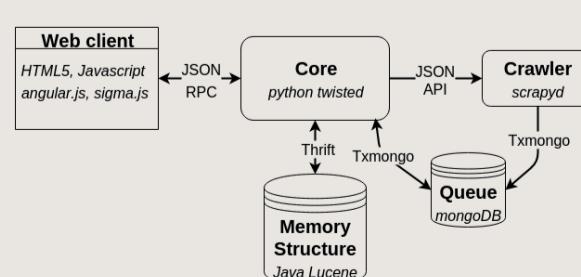


2 web entities
defined

Sciences Po
university

médialab
research lab

Hyphe's
Software
Architecture



DEMO

<http://hyphe.medialab.sciences-po.fr/demo>

The web is a field of investigation for social sciences, and platform-based studies have long proven their relevance. However the generic web is rarely studied in itself though it contains crucial aspects of the embodiment of social actors: personal blogs, institutional websites, hobby-specific media... We realized that some sociologists see existing web crawlers as "black boxes" unsuitable for research though they are willing to study the broad web.

Hyphe is a crawler developed with and for social scientists, with an innovative "curation-oriented" approach. It solves web-mining related problems thanks to specific features such as iterative corpus building or a memory structure allowing researchers to redefine dynamically the granularity of their "web entities".

