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

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”.

DEMO
http://hyphe.medialab.sciences-po.fr/demo

Import URLs
Define entities
Crawl
Prospect
Monitor corpus

iterative
curation

Custom
Web-entity
Granularity

Hyphes’s
Software
Architecture

Example corpus

STEM CELLS COMMUNITY
Virginie Tournay
Sciences Po Paris CEVIPOF EUCELLEX

IN 490
UND. 41
DISC. 4564
OUT 647