

From Navicrawler to HyBro: a brief history of webcrawlers for social sciences

FOSDEM'21

supposedly in Brussels, actually on the web!

February 6th 2021

Benjamin Ooghe-Tabanou (@boogheta)

Sciences Po médialab (@medialab_ScPo)

SciencesPo
MÉDIALAB



**DIME-SHS
FORCCAST**

Webcrawlers? What is that?

« Web crawling is the process of building a collection of webpages by starting with an initial set of URLs and recursively traversing the corresponding pages to find additional links. »

Gabe Ignatow & Rada Mihalcea, in: *Text Mining: A Guidebook for the Social Sciences*

But why would anyone want to do that?

- create an index for a search engine (like... you know...)
- extract contents or structured data (like scraping)
- perform some text mining
- study network communities through hyperlinks

Crawling the web for social sciences?

« **Hyperlink** » is at the core of the Web's architecture
→ charged with meaning and structure



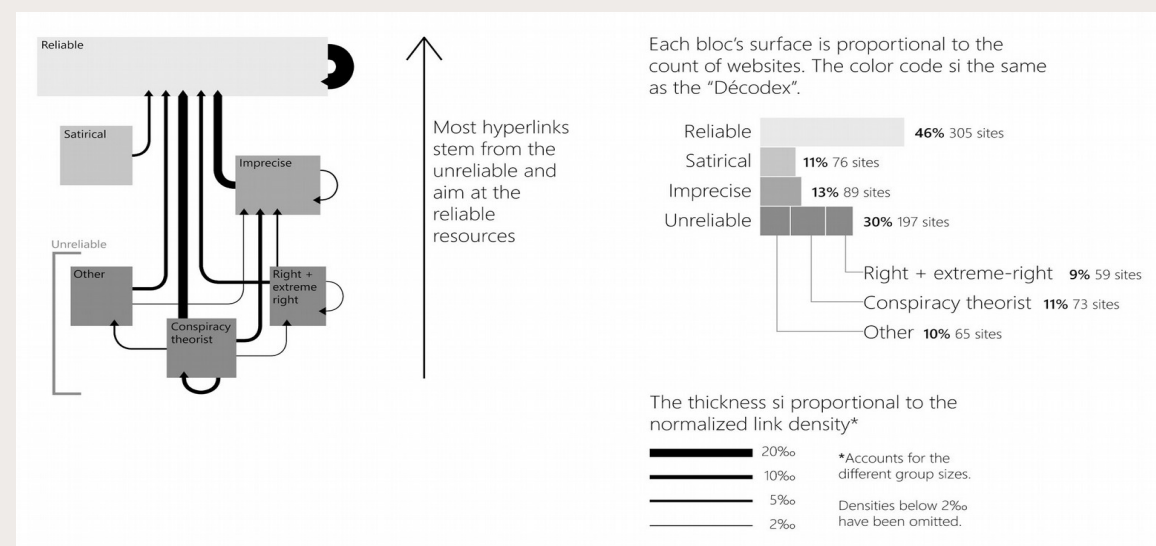
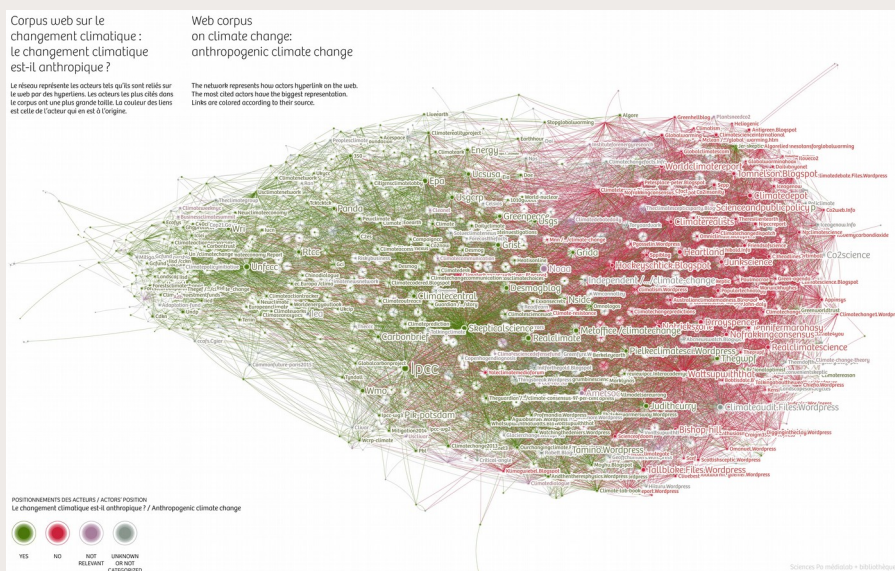
*« The texts are **linked together** in a way that one can go from one concept to another to find the information one wants. The network of links is called a **web**. [...] The texts are known as **nodes**. The process of proceeding from node to node is called **navigation**. »*

Tim Berners-Lee, 1990, *WorldWideWeb: Proposal for a HyperText Project*

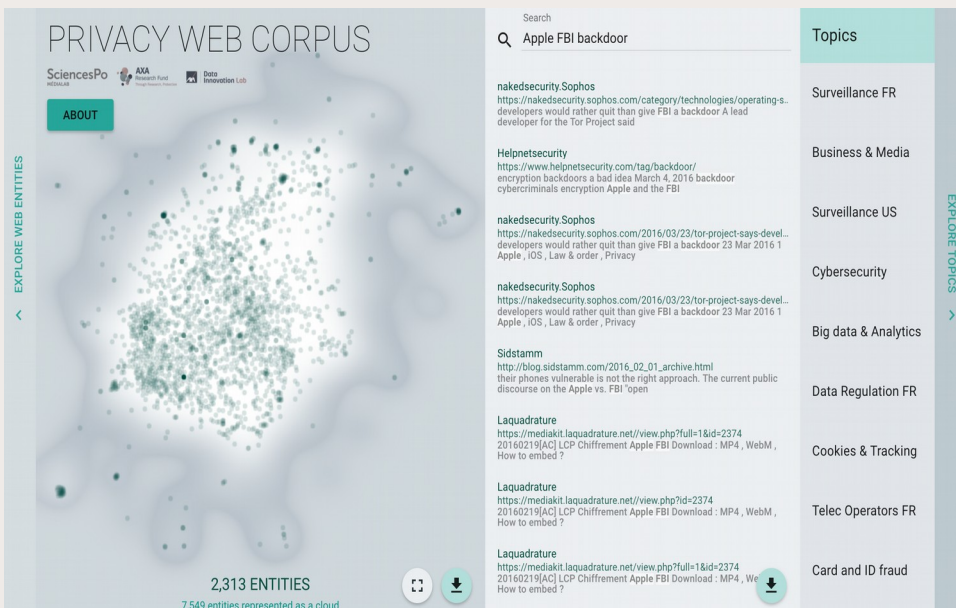
*« A hyperlink is a **manifestation of intention**. By **linking** one page to another, one piece of text to another, **people intend** to do particular things. »*

Ryfe, Mensing, & Kelley, 2016, *What is the meaning of a news link?*

Some examples of webcrawling research outputs

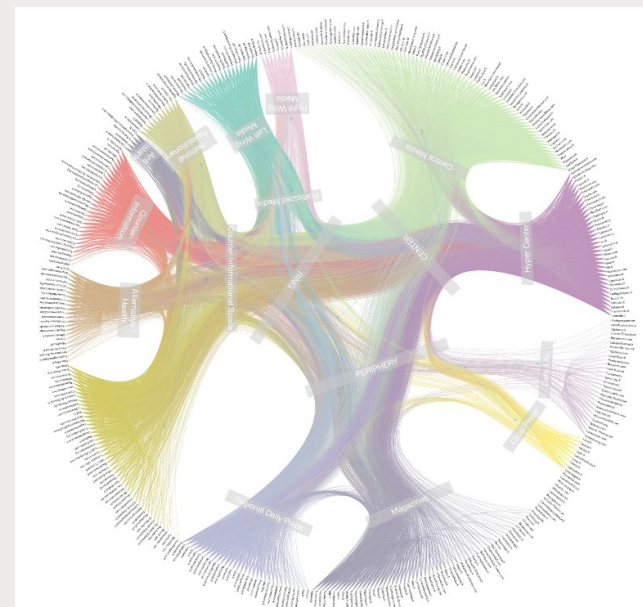


Double Dating Data: Climate change on the web



Privacy Web Corpus Datascape

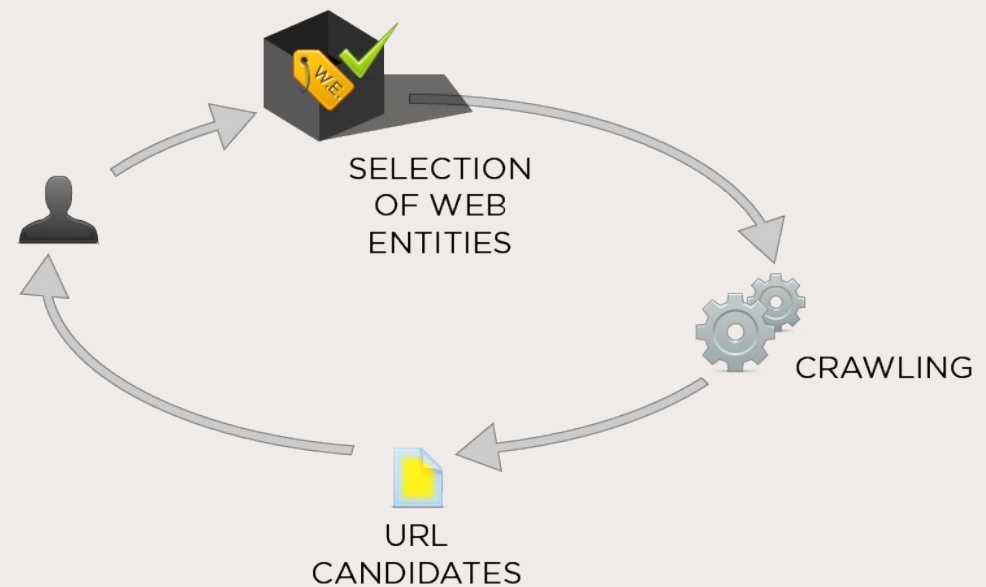
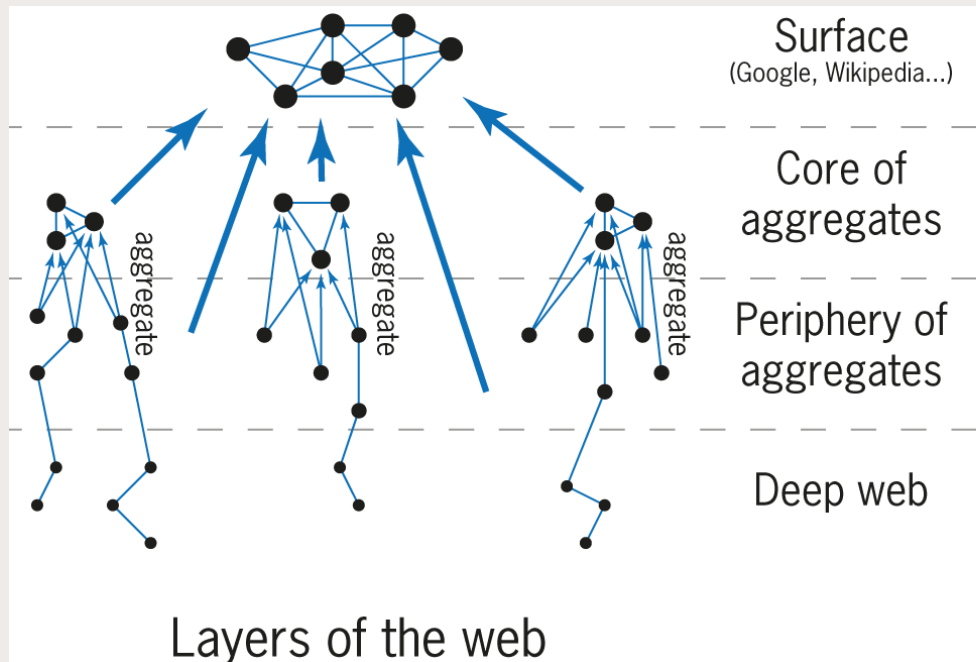
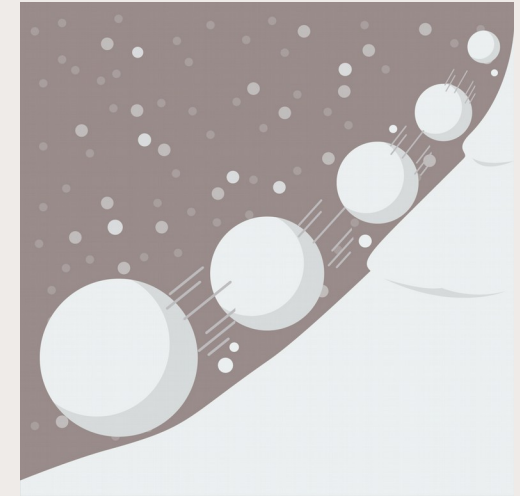
Visual Network Exploration for Data Journalists



Unfolding the multi-layered structure of the French Mediascape

Different possible crawling strategies

- Focus crawling
- In depth snowballing
- Selective crawling with iterative curation
- etc.

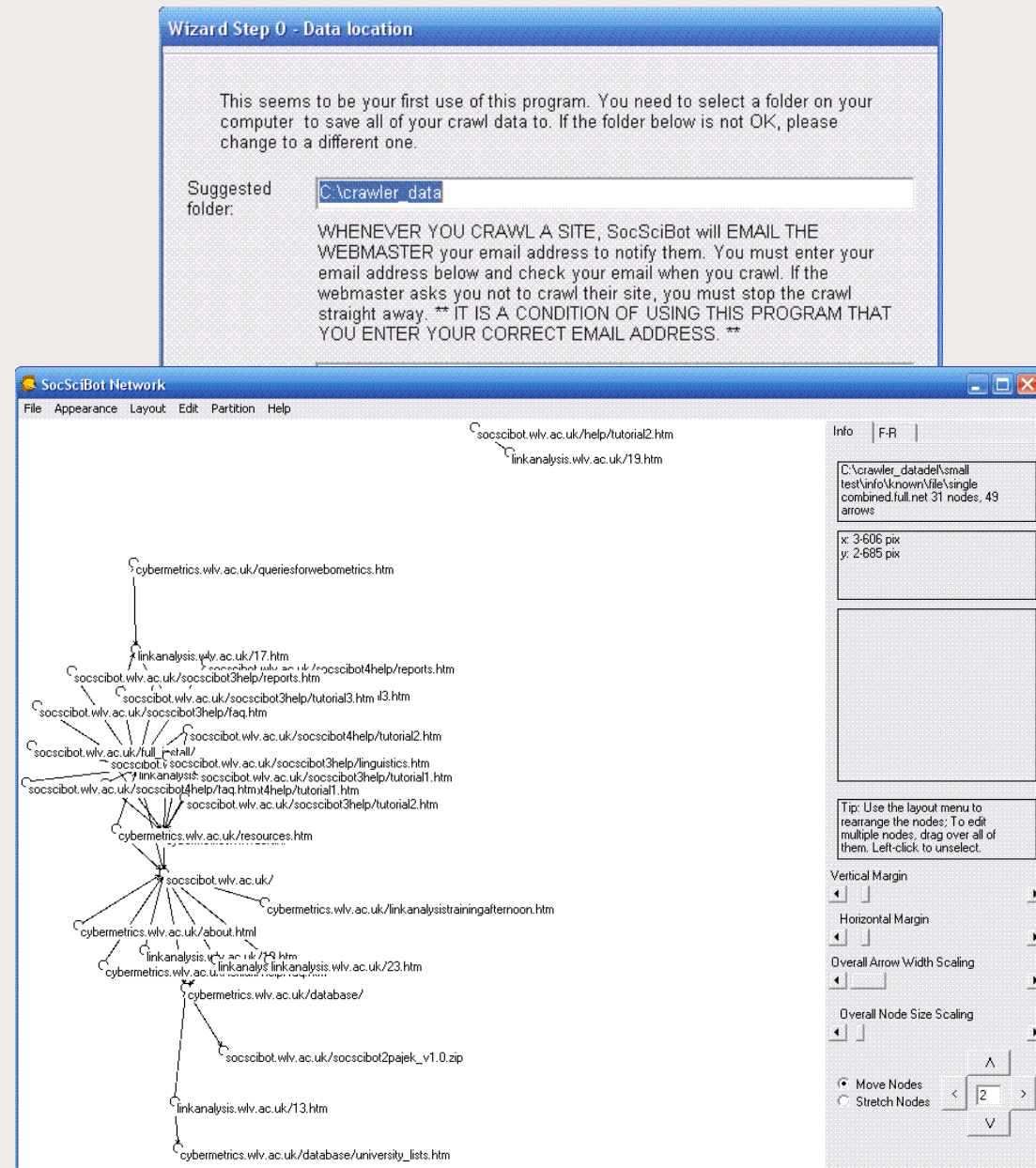
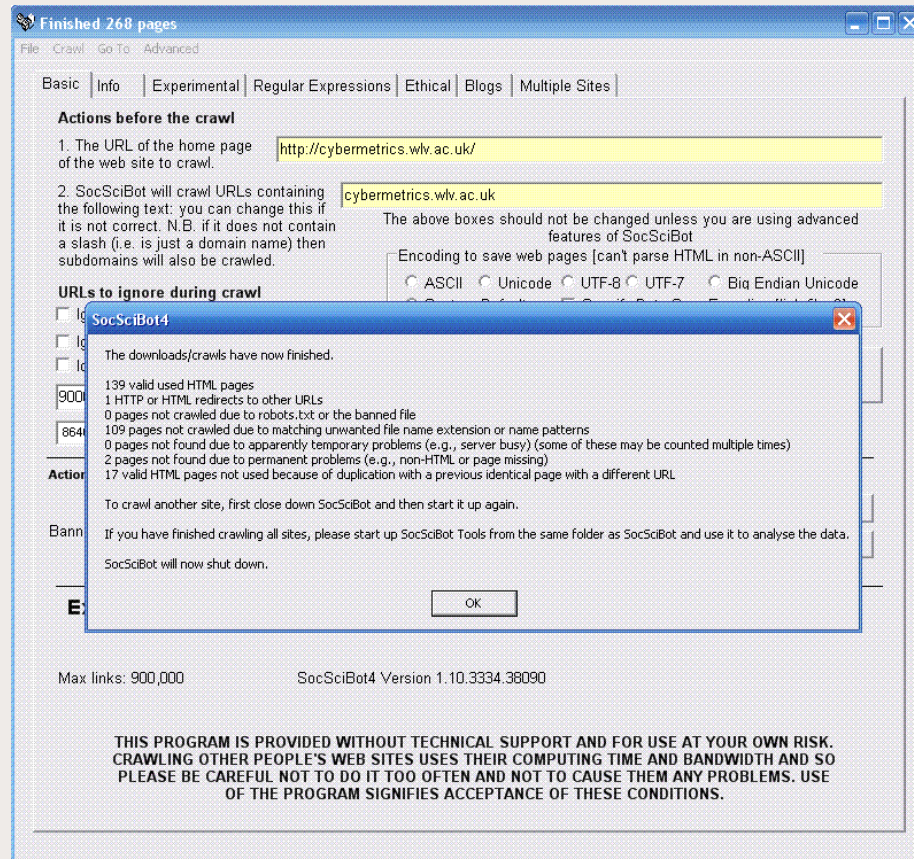


SocSciBot (by Statistical Cybernetics Research Group, UK)

Windows desktop application

Latest version in 2016

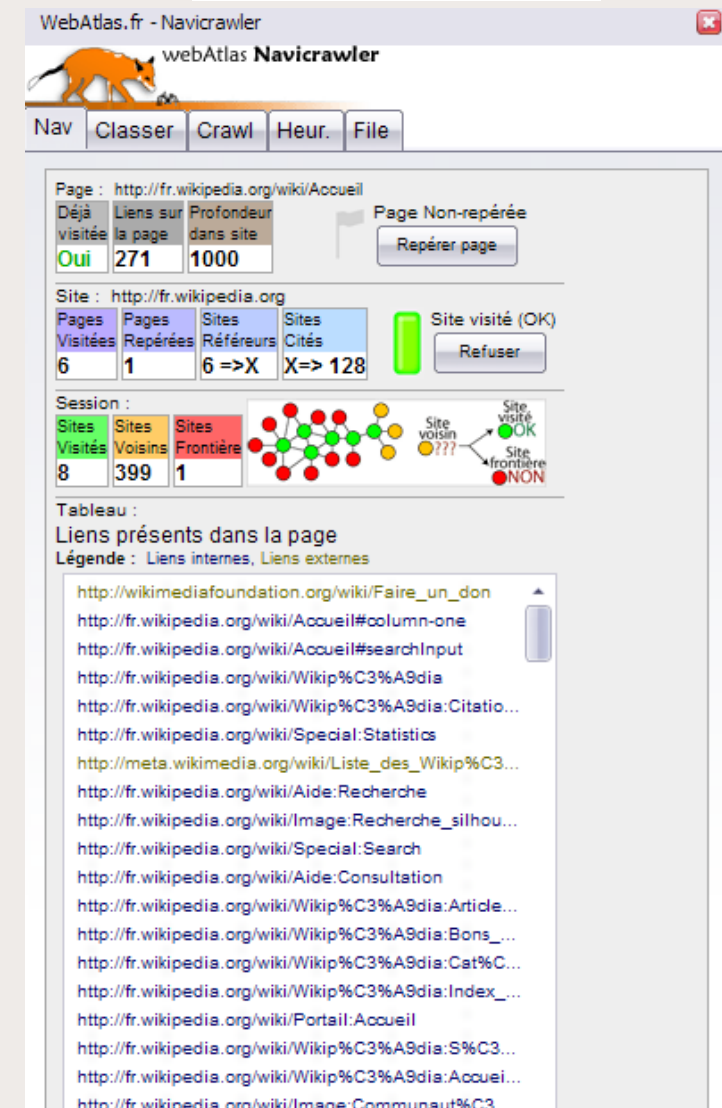
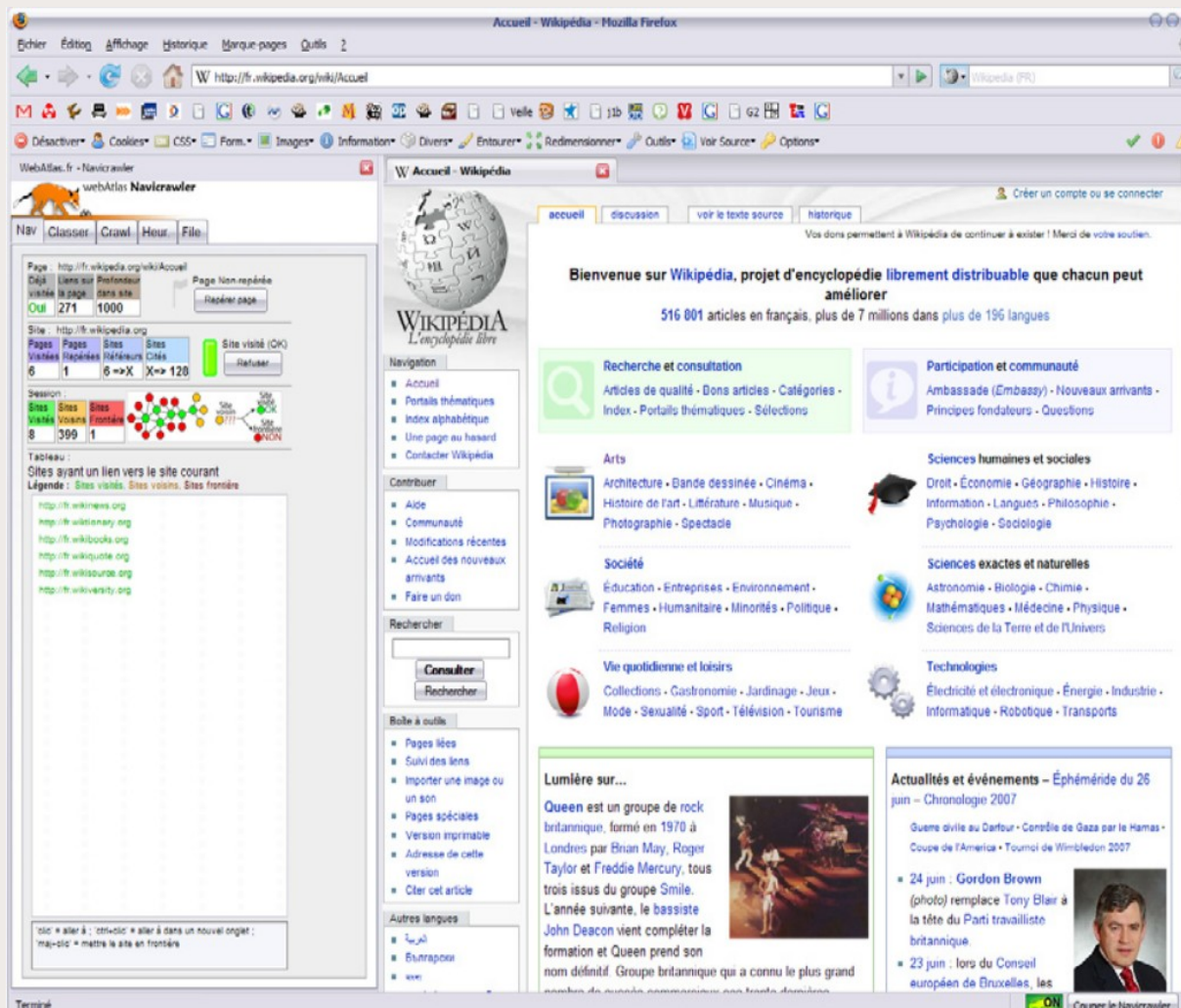
<http://socscibot.wlv.ac.uk>



Navicrawler (by WebAtlas & eDiasporas, France)

Firefox ($\leq v3.5.1$) addon

<https://github.com/medialab/navicrawler> (archived)

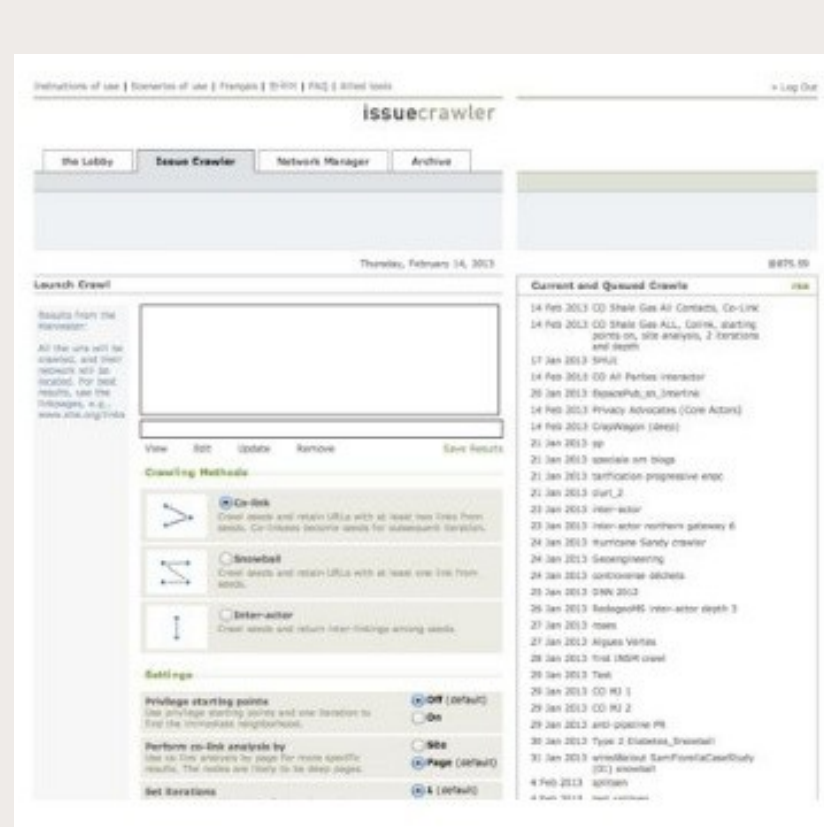


IssueCrawler (by Digital Methods Initiative, Netherlands)

Web application

Different crawling strategies offered

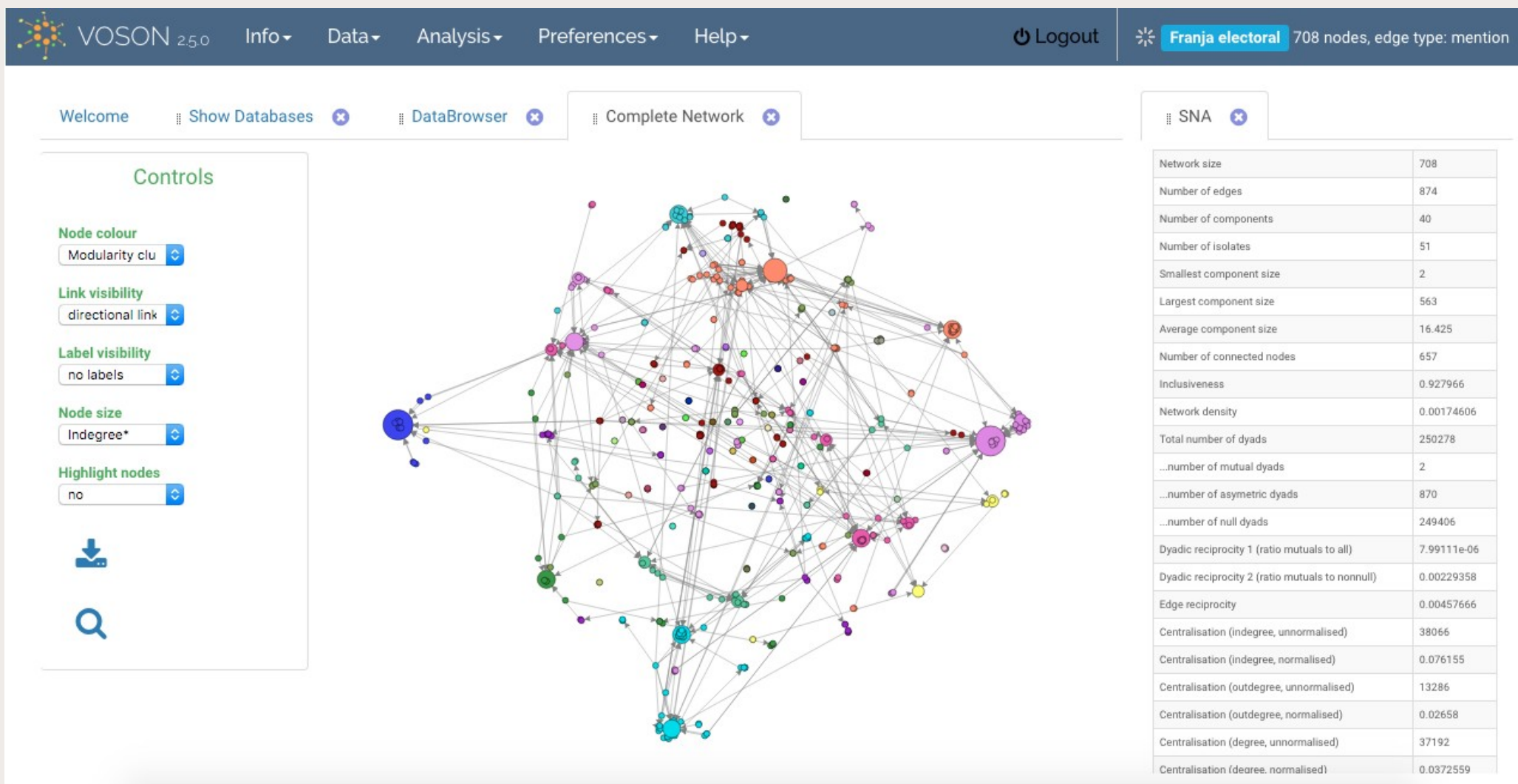
<https://www.issuecrawler.net>



VOSON (by Research School of Social Sciences, Australia)

Web application

<http://vosonlab.net/VOSON>



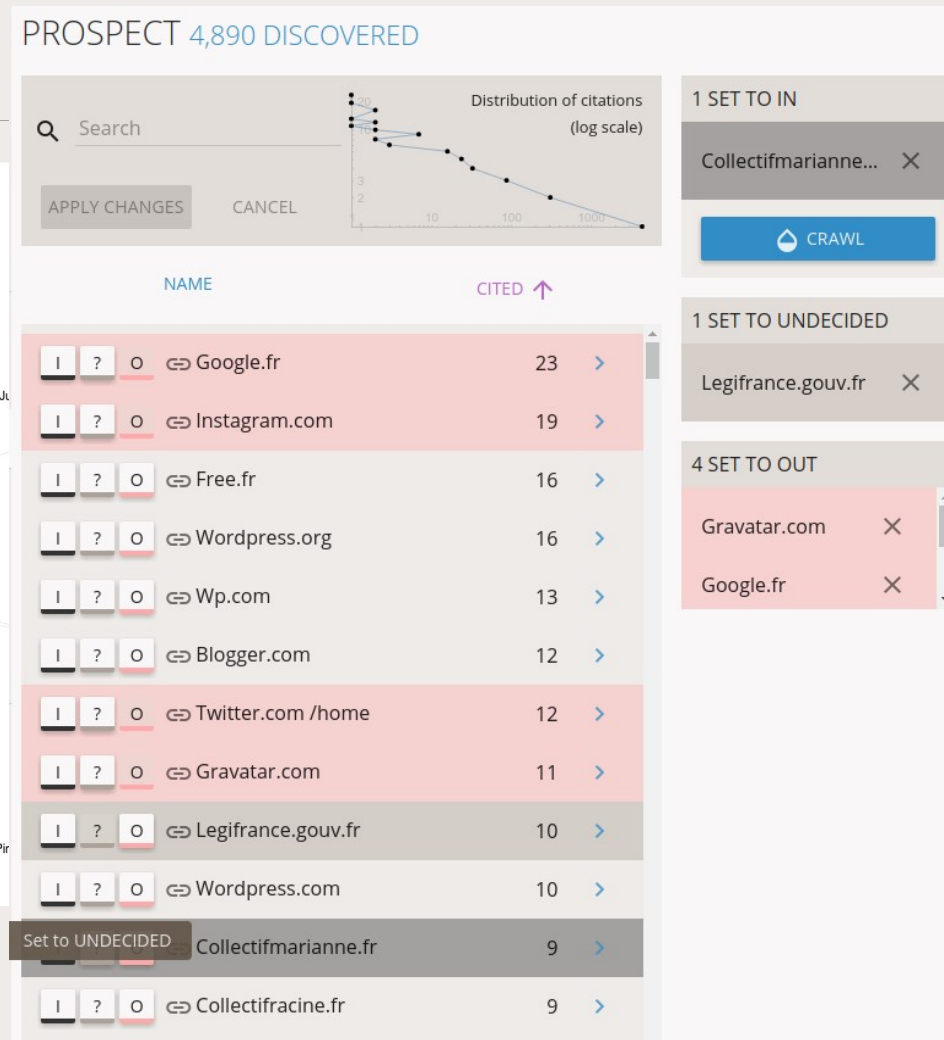
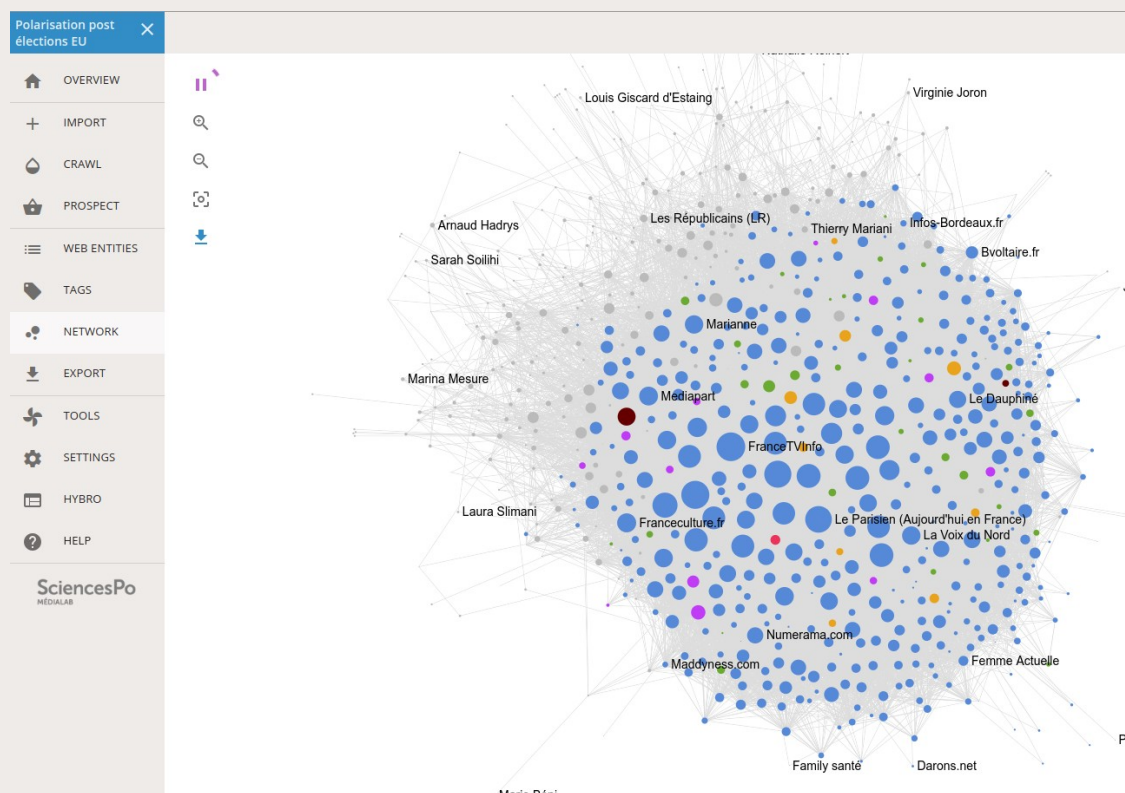
Hyphe (by médialab Sciences Po, France)

Web application

Dynamic web entities + Iterative focused crawling

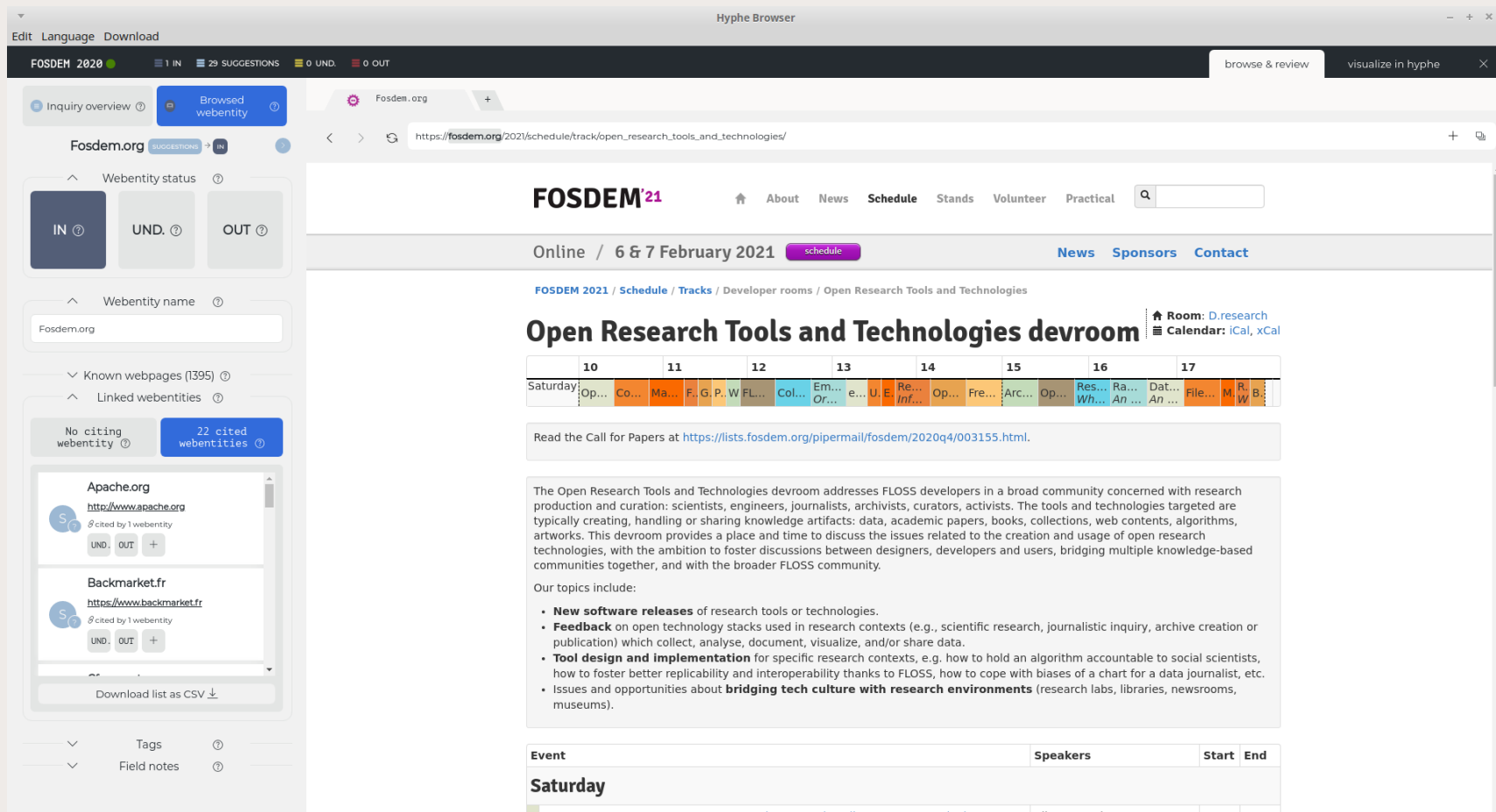
<https://hyphe.medialab.sciences-po.fr/demo/>

<https://github.com/medialab/hyphe>



HyBro (by médialab Sciences Po, France)

Hyphe Browser → crawl with Hyphe, but while browsing like NaviCrawler
Linux / MacOS / Windows desktop application (based on electronJs)
v2.0 just released with complete redesign! <https://github.com/medialab/hyphe-browser/releases/>



The screenshot shows the Hyphe Browser interface. The sidebar on the left contains several sections: 'Webentity status' with buttons for 'IN', 'UND', and 'OUT'; 'Webentity name' with a search bar containing 'Fosdem.org'; 'Known webpages (1395)' and 'Linked webentities' sections; and a 'Download list as CSV' button. The main content area displays the FOSDEM'21 website, which includes a navigation bar with 'About', 'News', 'Schedule', 'Stands', 'Volunteer', and 'Practical'. The main content area features a 'FOSDEM'21' header, a 'Schedule' button, and a section for 'Open Research Tools and Technologies devroom'. This section includes a calendar view for February 6-7, 2021, and a list of topics for the devroom, such as 'New software releases', 'Feedback', 'Tool design and implementation', and 'Issues and opportunities about bridging tech culture with research environments'.