

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.* »

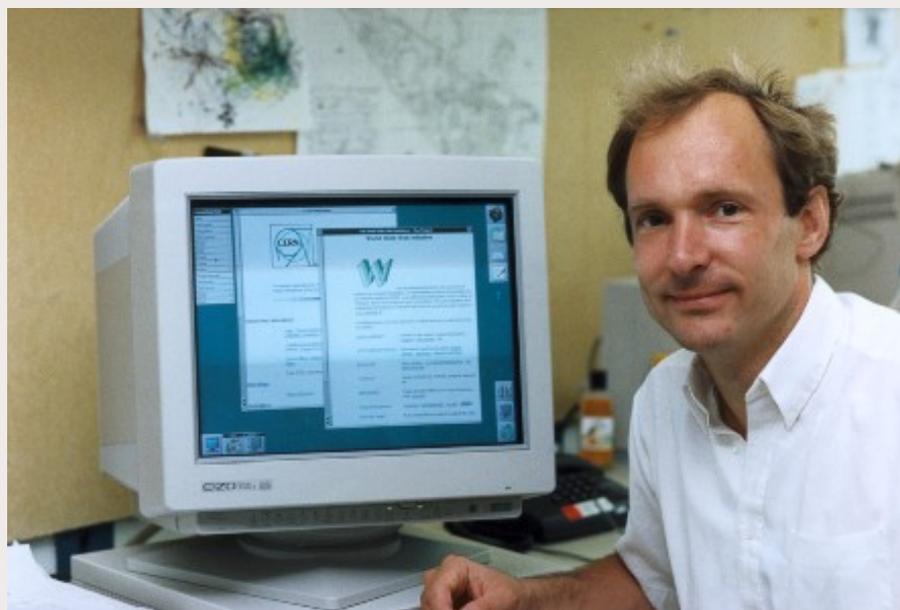
Gabe Ignatow & Rada Mihalcea, in: *Text Mining: A Guidebook for the Social Sciences*

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



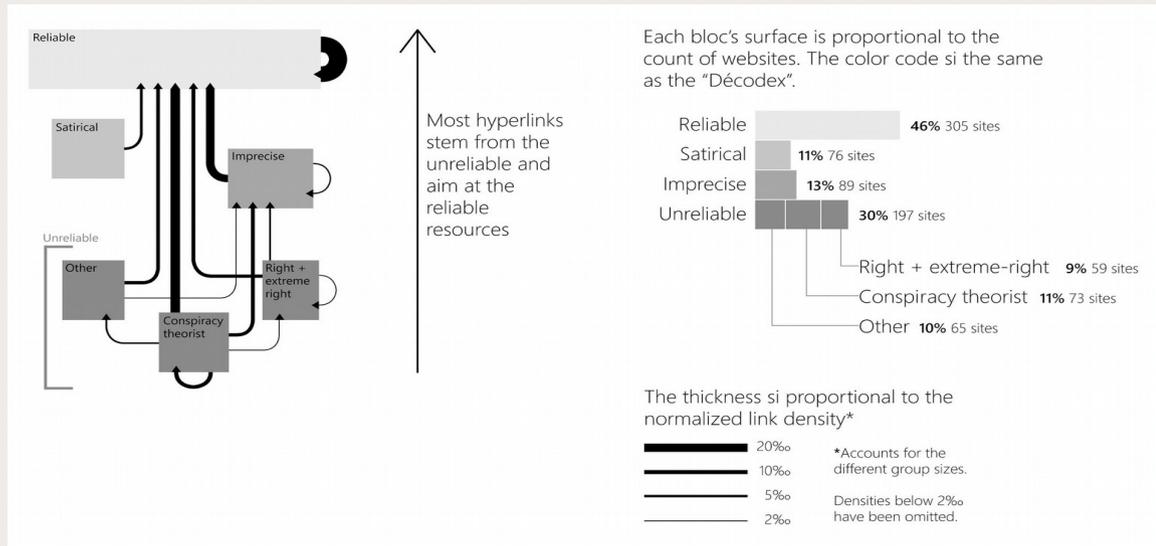
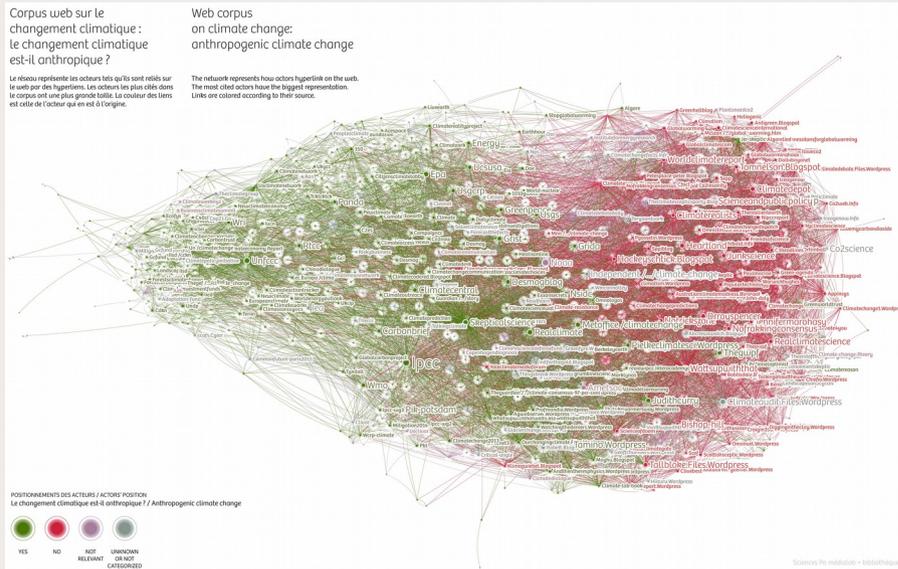
« The texts are **linked together** in a way that one can go from one concept to another to find the information one wants. The network of links is called a **web**. [...] The texts are known as **nodes**. The process of proceeding from node to node is called **navigation**. »

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

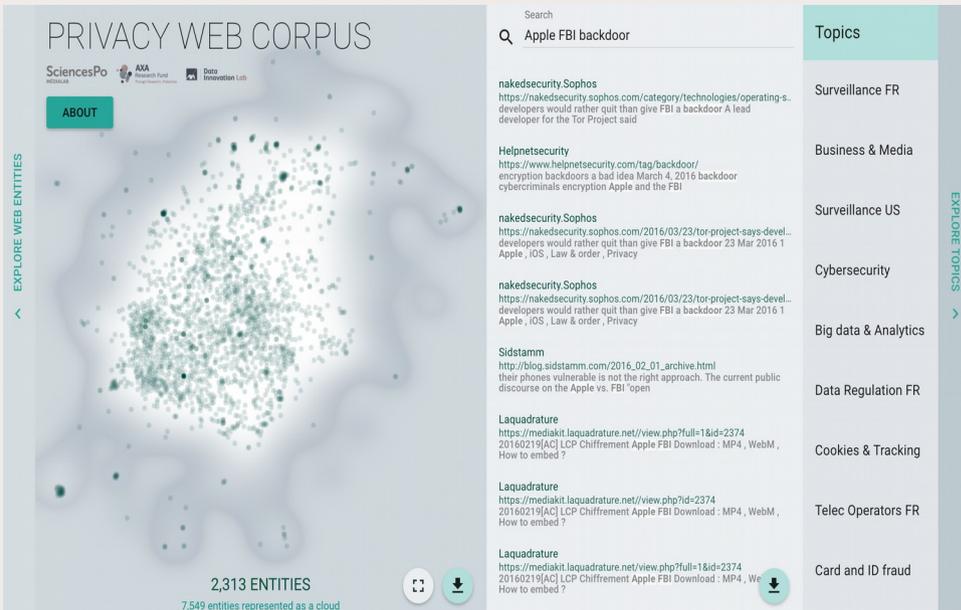
« A hyperlink is a **manifestation of intention**. By linking one page to another, one piece of text to another, **people intend** to do particular things. »

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

Some examples of webcrawling research outputs

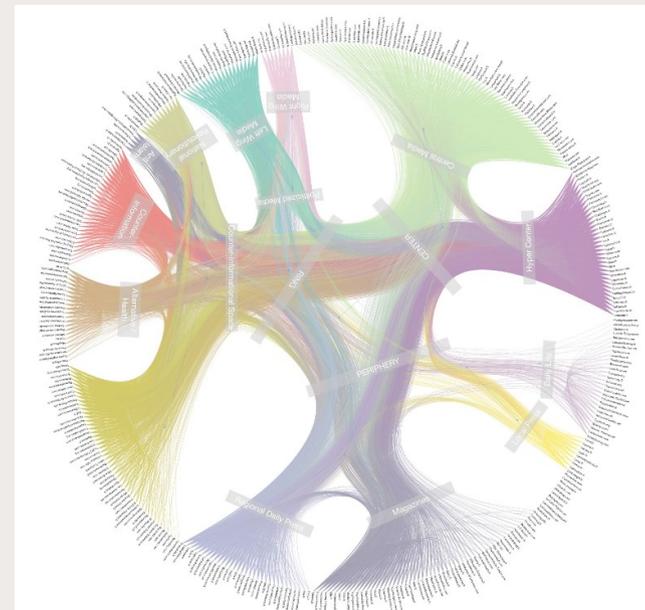


Double Dating Data: Climate change on the web



Privacy Web Corpus Datascape

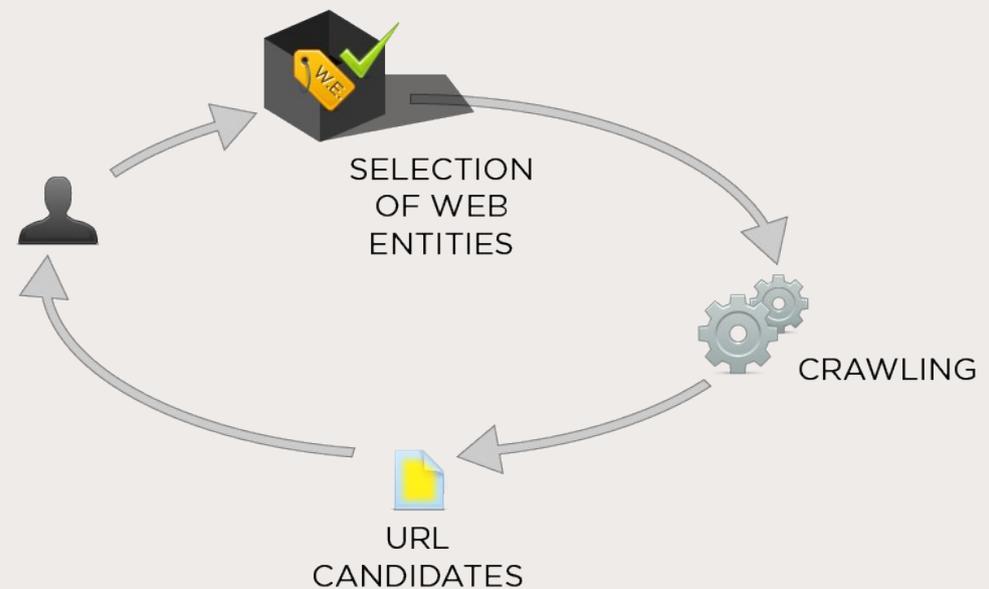
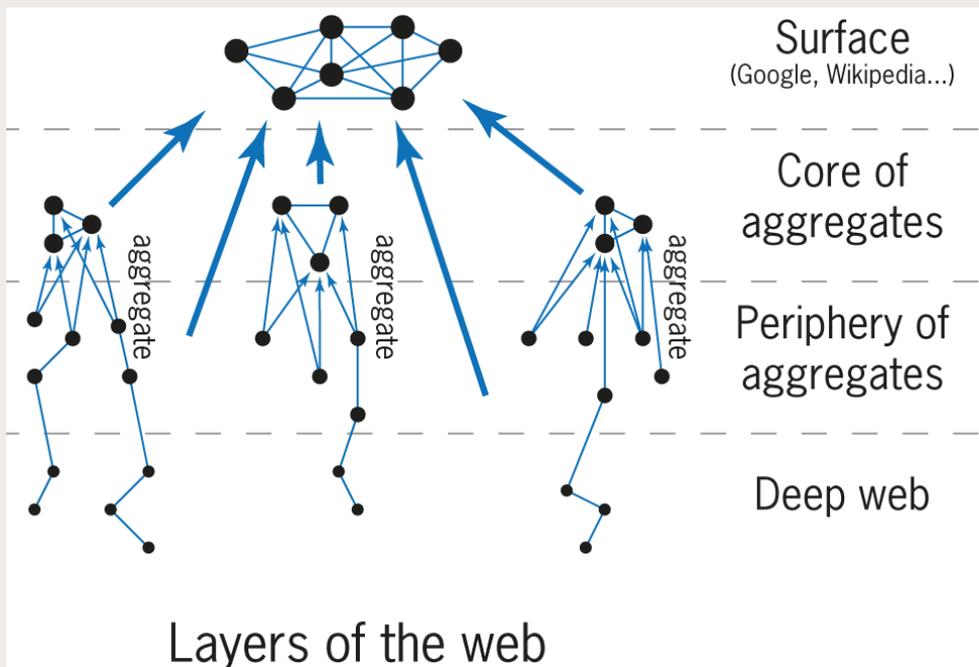
Visual Network Exploration for Data Journalists



Unfolding the multi-layered structure of the French Mediascape

Different possible crawling strategies

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

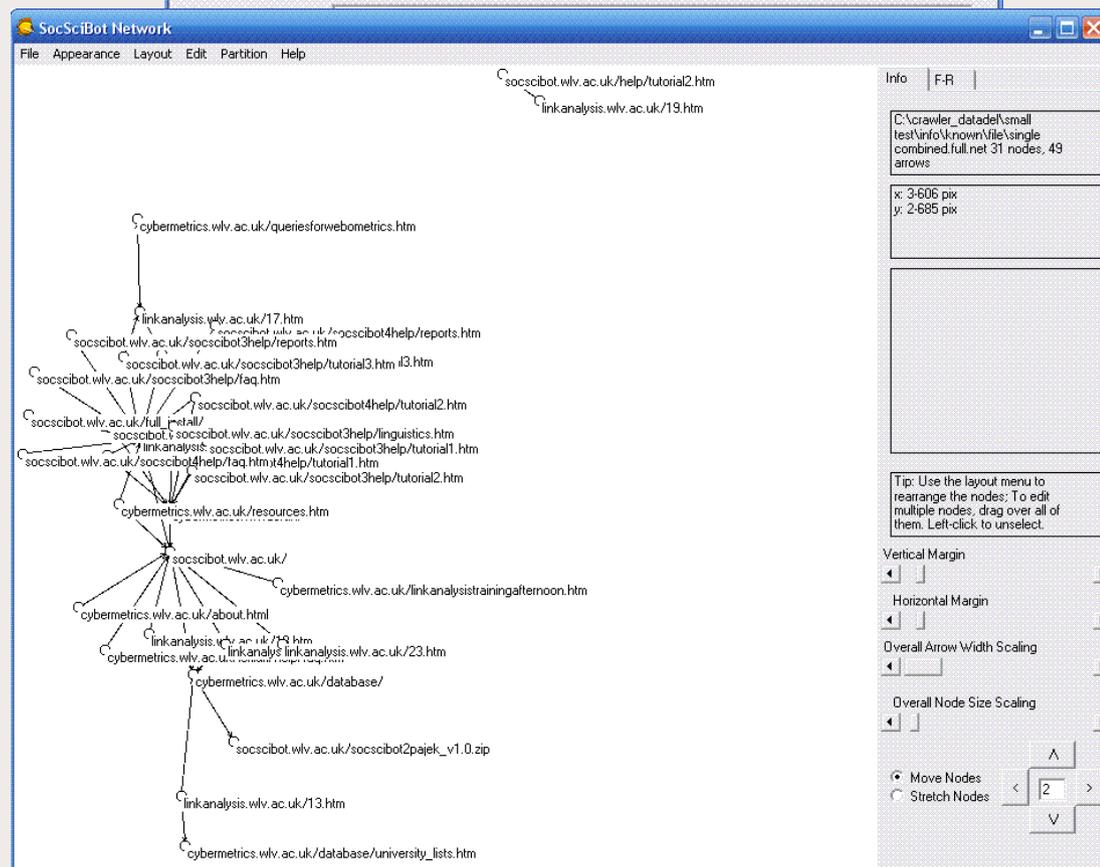
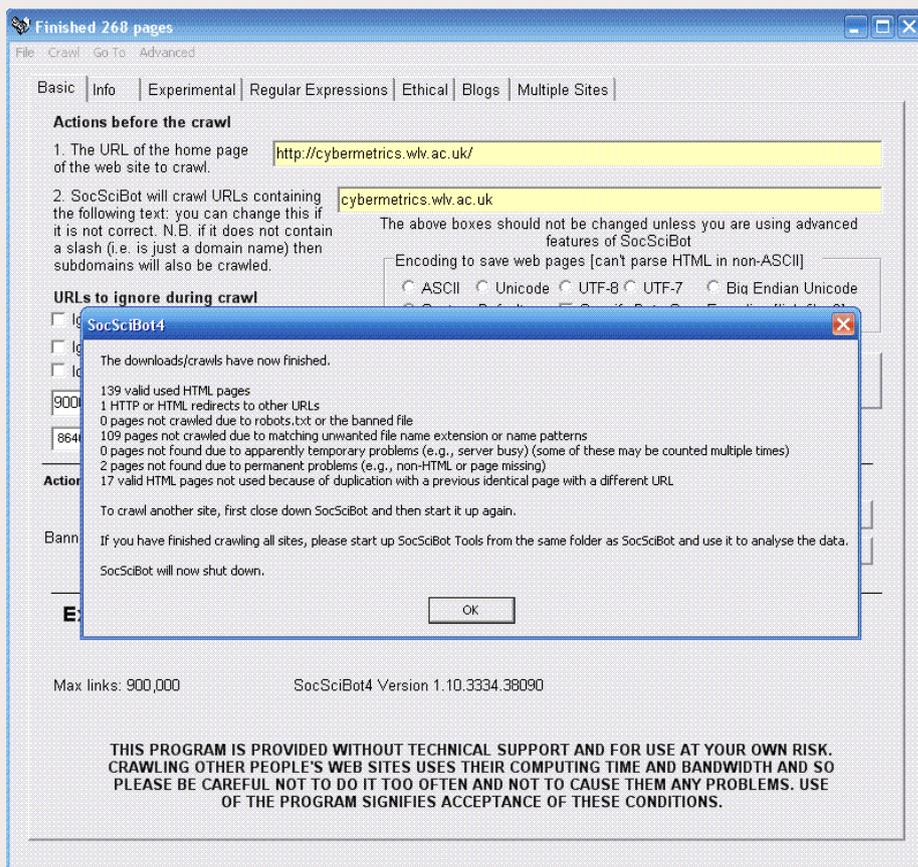
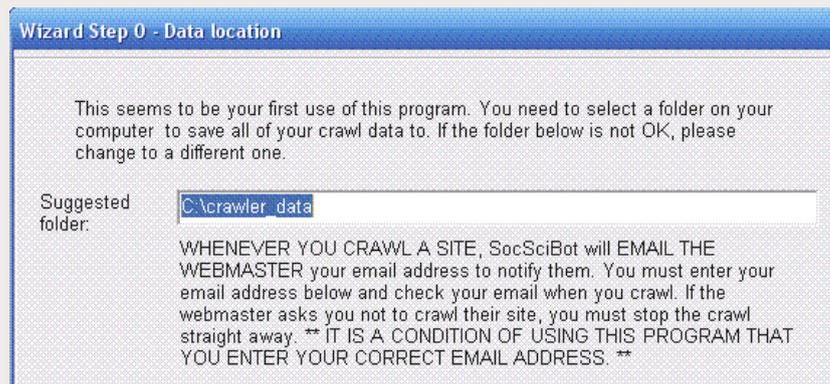


SocSciBot (by Statistical Cybernetics Research Group, UK)

Windows desktop application

Latest version in 2016

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



Navicrawler (by WebAtlas & eDiasporas, France)

Firefox (≤ v3.5.1) addon

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



The screenshot shows the Firefox browser window with the Navicrawler extension active. The extension interface includes a navigation menu (Nav, Classer, Crawl, Heur., File) and a main display area with several sections:

- Page:** http://fr.wikipedia.org/wiki/Accueil. Statistics: Déjà visitée la page: Oui (271), Profondeur dans site: 1000. Buttons: Repérer page, Page Non-repérée.
- Site:** http://fr.wikipedia.org. Statistics: Pages Visitées: 6, Pages Repérées: 1, Sites Réfèreurs: 6 => X, Sites Cités: X => 128. Buttons: Site visité (OK), Refuser.
- Session:** Sites Visités: 8, Sites Voisins: 399, Sites Frontières: 1. Includes a network diagram.
- Tableau:** Sites ayant un lien vers le site courant. Légende: Sites visités, Sites voisins, Sites frontières.
- Liens présents dans la page:** A list of links found on the page, including internal and external links.

This screenshot provides a closer look at the Navicrawler extension's data tables and network visualization:

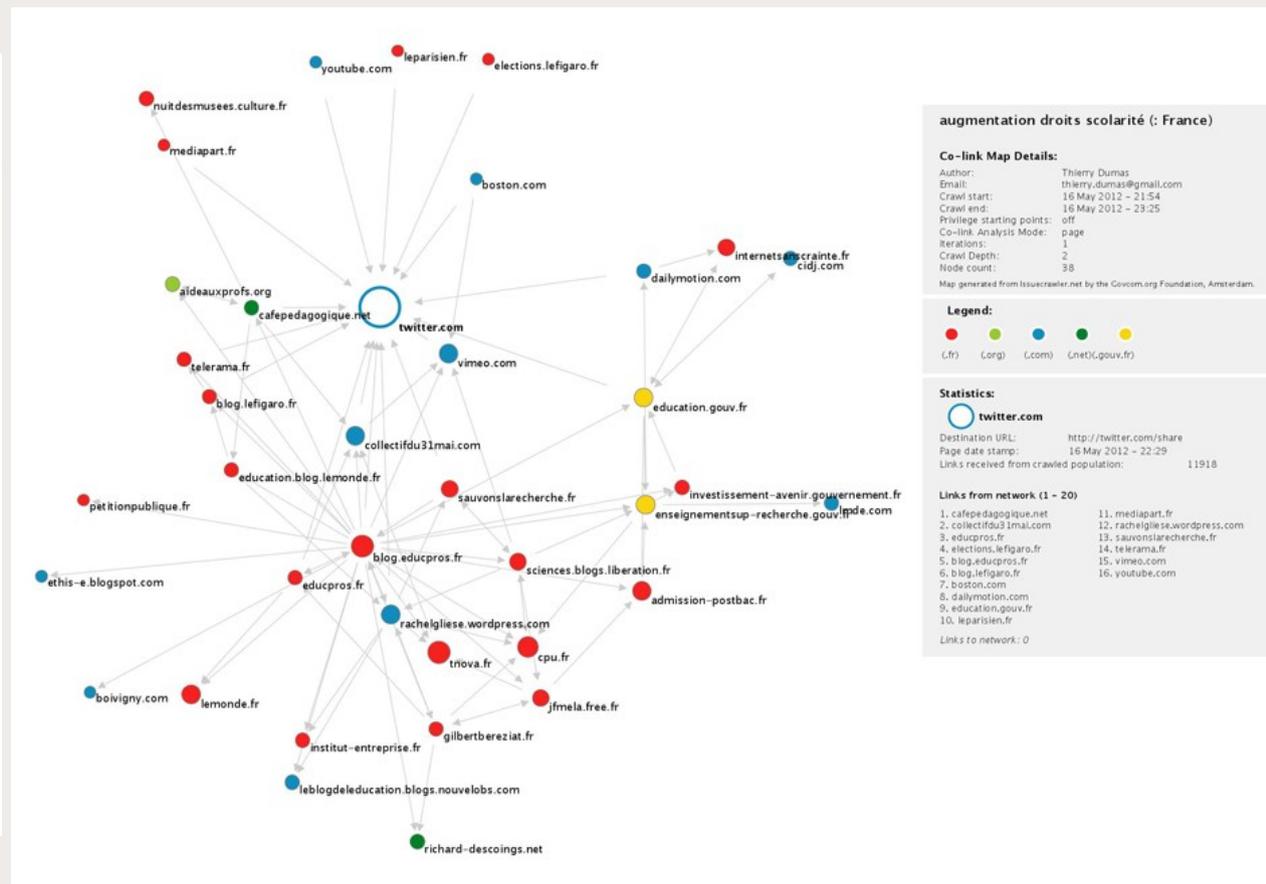
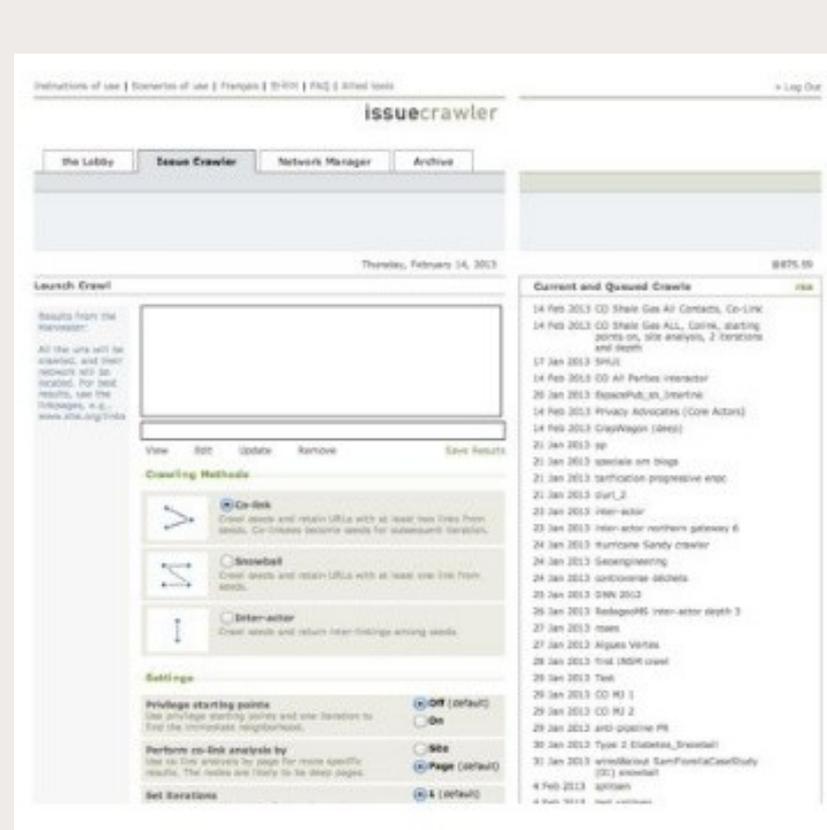
- Page:** http://fr.wikipedia.org/wiki/Accueil. Page Non-repérée. Buttons: Repérer page.
- Site:** http://fr.wikipedia.org. Statistics: Pages Visitées: 6, Pages Repérées: 1, Sites Réfèreurs: 6 => X, Sites Cités: X => 128. Buttons: Site visité (OK), Refuser.
- Session:** Sites Visités: 8, Sites Voisins: 399, Sites Frontières: 1. Includes a network diagram with nodes and edges.
- Tableau:** Sites ayant un lien vers le site courant. Légende: Sites visités, Sites voisins, Sites frontières.
- Liens présents dans la page:** A list of links found on the page, including internal and external links.

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>

The screenshot displays the VOSON 2.5.0 web application interface. At the top, there is a navigation bar with menus for 'Info', 'Data', 'Analysis', 'Preferences', and 'Help', along with a 'Logout' button. The current session is identified as 'Franja electoral' with 708 nodes and an edge type of 'mention'. Below the navigation bar, there are tabs for 'Welcome', 'Show Databases', 'DataBrowser', 'Complete Network', and 'SNA'. The main area is divided into three sections: 'Controls', a network graph, and a 'SNA' statistics table.

Controls

- Node colour:** Modularity clu
- Link visibility:** directional link
- Label visibility:** no labels
- Node size:** Indegree*
- Highlight nodes:** no

Network Graph: A complex network graph with 708 nodes and 874 edges. Nodes are colored based on modularity clusters and sized based on indegree. The graph shows a dense central cluster with several smaller clusters and isolates.

SNA Statistics Table:

Network size	708
Number of edges	874
Number of components	40
Number of isolates	51
Smallest component size	2
Largest component size	563
Average component size	16.425
Number of connected nodes	657
Inclusiveness	0.927966
Network density	0.00174606
Total number of dyads	250278
...number of mutual dyads	2
...number of asymmetric dyads	870
...number of null dyads	249406
Dyadic reciprocity 1 (ratio mutuals to all)	7.99111e-06
Dyadic reciprocity 2 (ratio mutuals to nonnull)	0.00229358
Edge reciprocity	0.00457666
Centralisation (indegree, unnormalised)	38066
Centralisation (indegree, normalised)	0.076155
Centralisation (outdegree, unnormalised)	13286
Centralisation (outdegree, normalised)	0.02658
Centralisation (degree, unnormalised)	37192
Centralisation (degree, normalised)	0.0372559

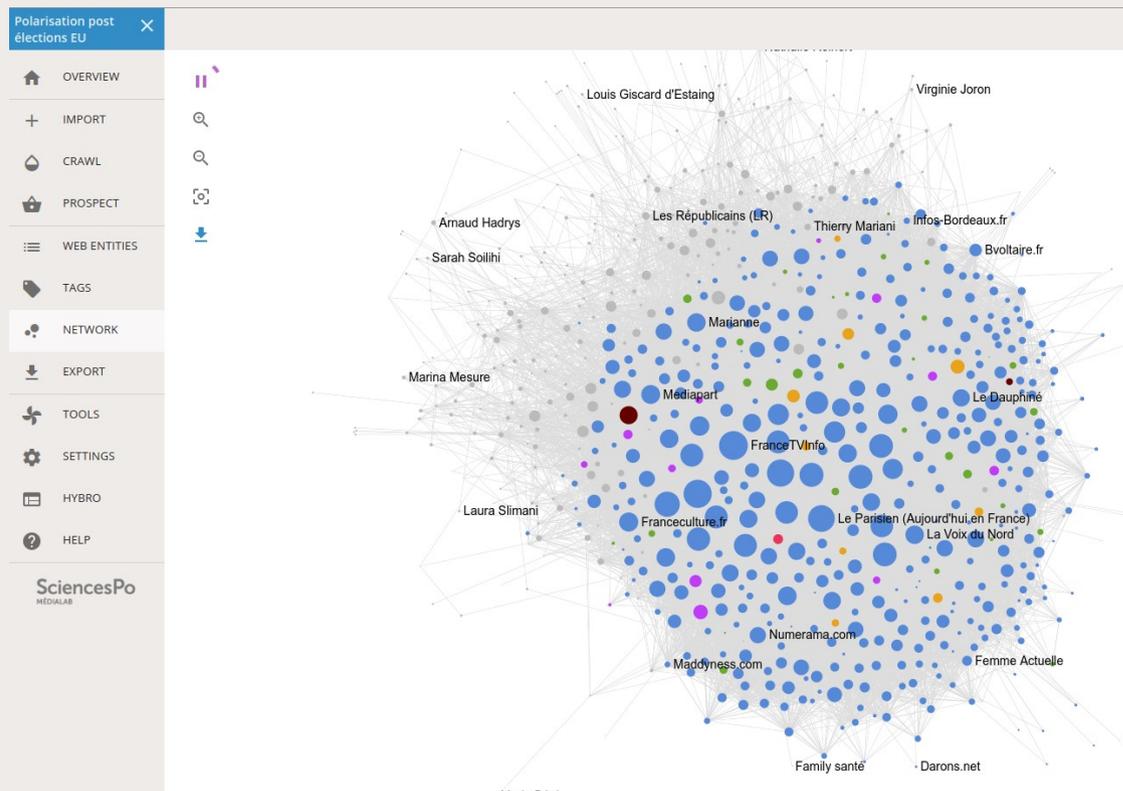
Hyphe (by médialab Sciences Po, France)

Web application

Dynamic web entities + Iterative focused crawling

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

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

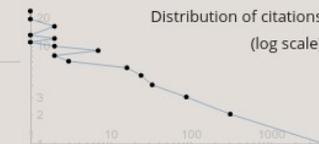


PROSPECT 4,890 DISCOVERED

Search

APPLY CHANGES CANCEL

Distribution of citations (log scale)



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

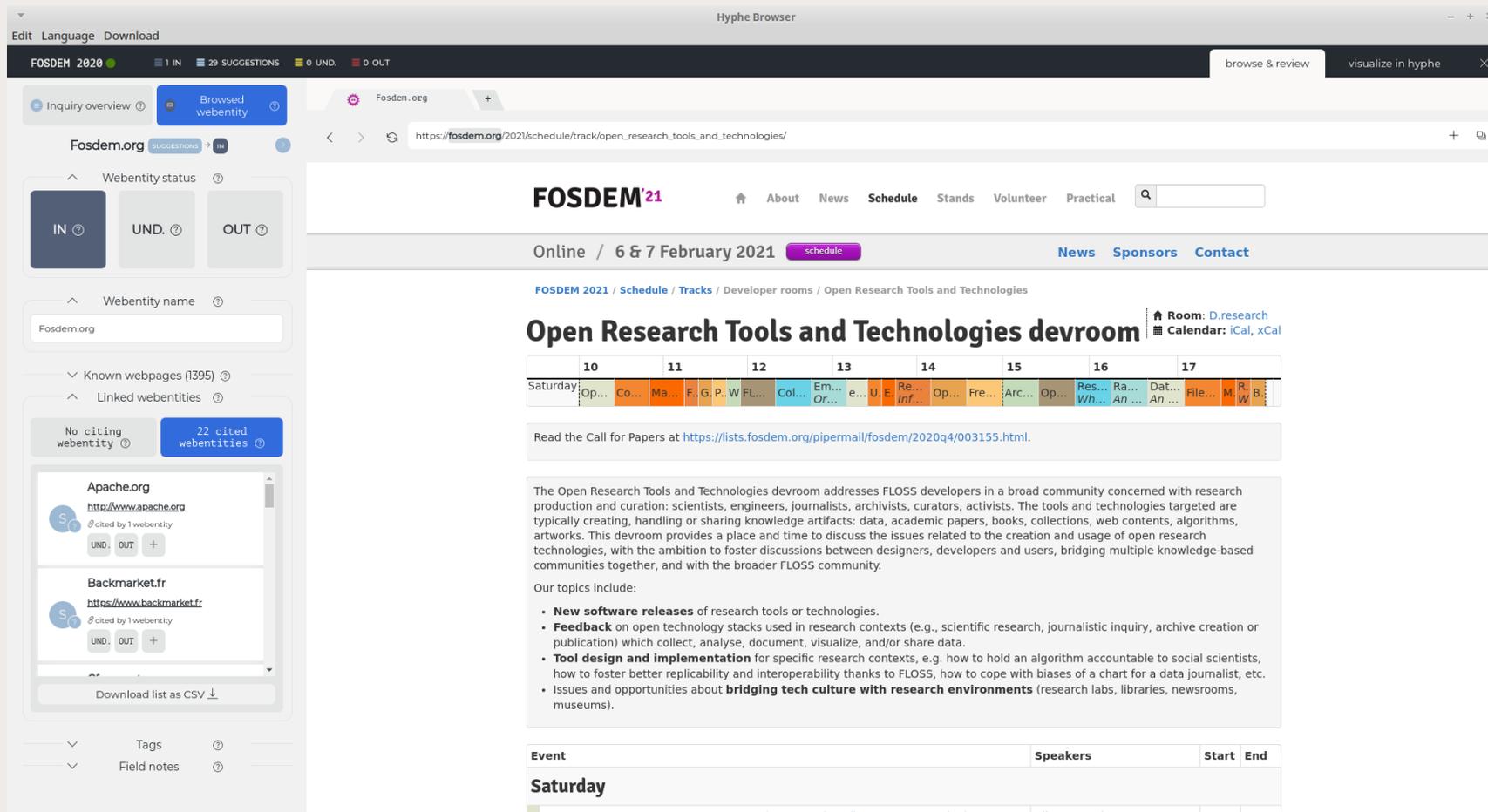
4 SET TO OUT

Gravatar.com ✕

Google.fr ✕

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 main window displays the FOSDEM'21 website, specifically the 'Open Research Tools and Technologies devroom' page. The sidebar on the left contains several sections:

- Inquiry overview**: Shows 'Browsed weblentity' with a count of 29.
- Webentity status**: Includes buttons for 'IN', 'UND.', and 'OUT'.
- Webentity name**: A search field containing 'Fosdem.org'.
- Known webpages (1395)**: A list of webpages, including 'Apache.org' and 'Backmarket.fr', each with a 'cited by 1 weblentity' indicator and 'UND.', 'OUT', and '+' buttons.
- Linked weblentities**: A section for managing linked entities.
- Tags** and **Field notes**: Sections for managing metadata.

The main content area shows the FOSDEM'21 website with a navigation menu, a date range of '6 & 7 February 2021', and a calendar view for the event. The 'Open Research Tools and Technologies devroom' section includes a description of the devroom's focus on FLOSS developers and a list of topics to be discussed.