Webcrawlers for social sciences — FOSDEM 2021

Benjamin Ooghe-Tabanou · FOSDEM 2021, Open Research

Slides from FOSDEM, used under CC BY 2.0 BE.

Source: https://video.fosdem.org/2021/D.research/open_research_webcrawlers_social_sciences.webm

Converted with Video2Any · video2any.com

Not affiliated with or endorsed by FOSDEM.

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

Benjamin Ooghe-Tabanou

D.research

2021-02-06



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

Benjamin Ooghe-Tabanou

D.research
2021-02-06



SciencesPo

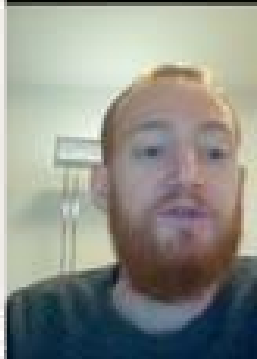
Webcrawlers? What is that?

« 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



SciencesPo

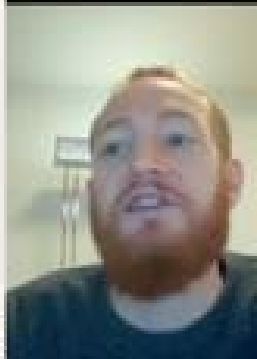
Webcrawlers? What is that?

« 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



SciencesPo

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**. […]*

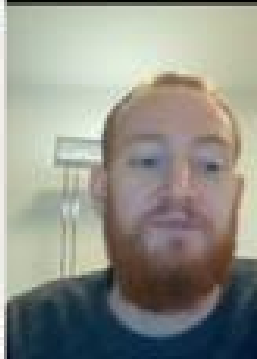
*The texts are known as **nodes**.*

*The process of proceeding from node to node is called **navigation**. »*

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

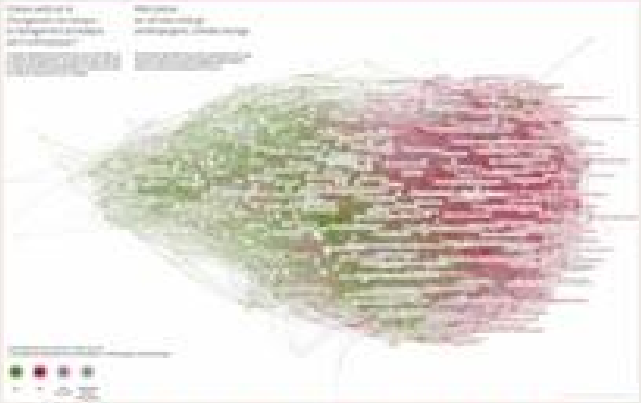
*« A hyperlink is a **manifestation of intention**. **By linking** one page to another, one piece of text to another, **people intend** to do particular things. »*

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

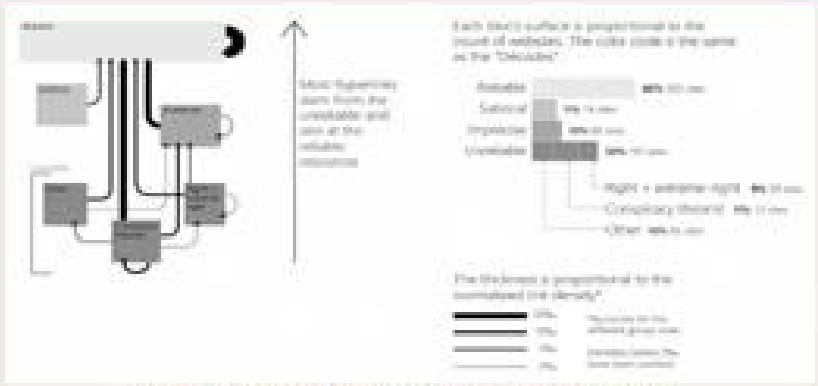


SciencesPo
MÉDIALAB

Some examples of webcrawling research outputs



Double Dating Data: Climate change on the web



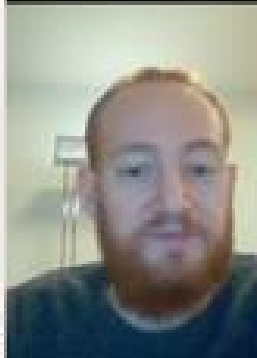
Visual Network Exploration for Data Journalists



Privacy Web Corpus Datascape



Unfolding the multi-layered structure of the French Mediascape



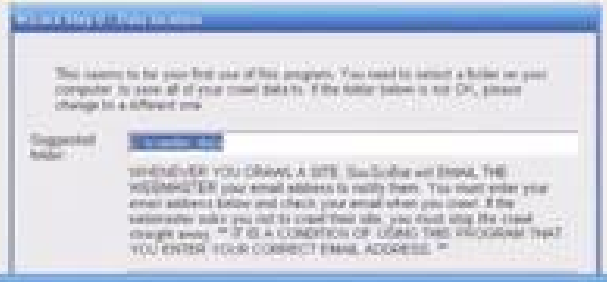
SciencesPo

SocSciBot (by Statistical Cybernetics Research Group, UK)

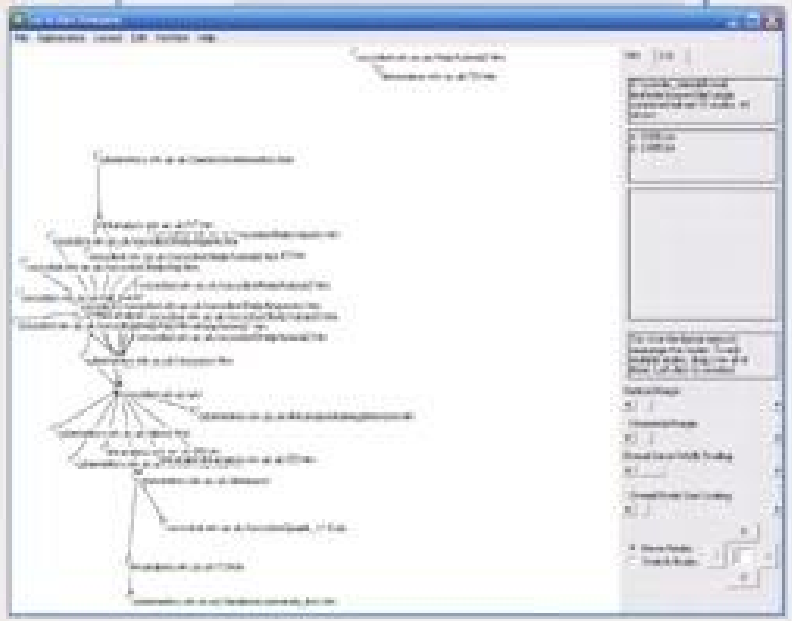
Windows desktop application

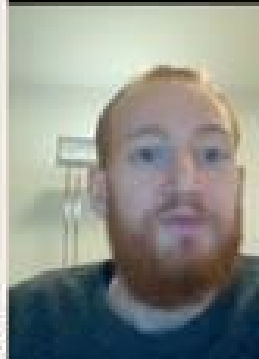
Latest version in 2016

http://socscibot.wlv.ac.uk







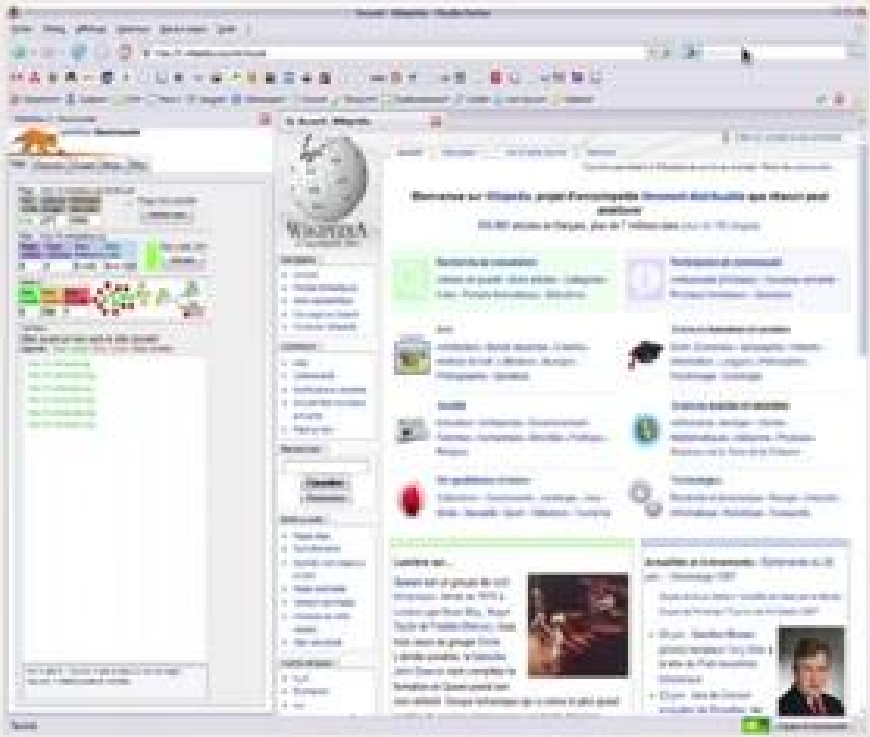


SciencesPo
MÉDIALAB

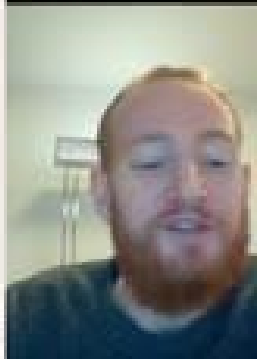
Navicrawler (by WebAtlas & eDiasporas, France)

Firefox (≤ v3.5.1) addon

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





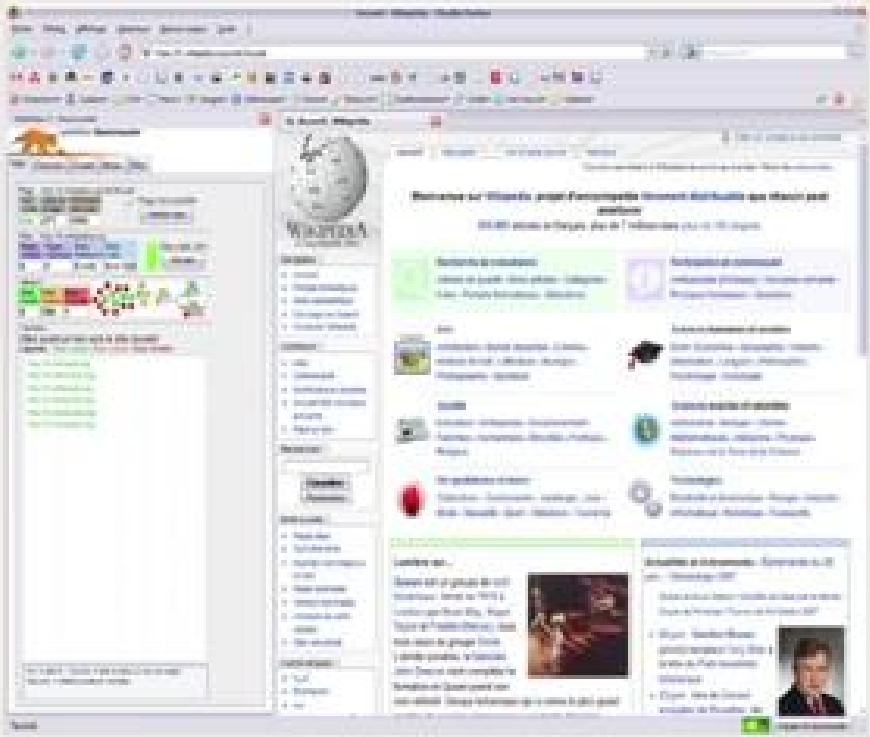


SciencesPo
MÉDIALAB

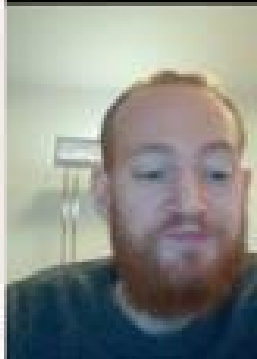
Navicrawler (by WebAtlas & eDiasporas, France)

Firefox (≤ v3.5.1) addon

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







SciencesPo

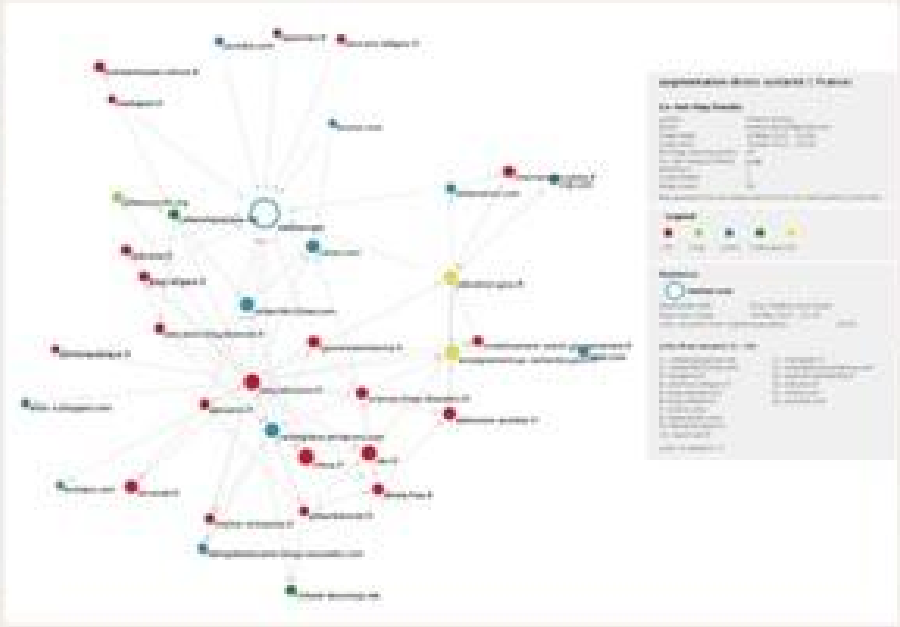
IssueCrawler (by Digital Methods Initiative, Netherlands)

Web application

Different crawling strategies offered

https://www.issuecrawler.net







SciencesPo MÉDIALAB

Hyphe (by médialab Sciences Po, France)

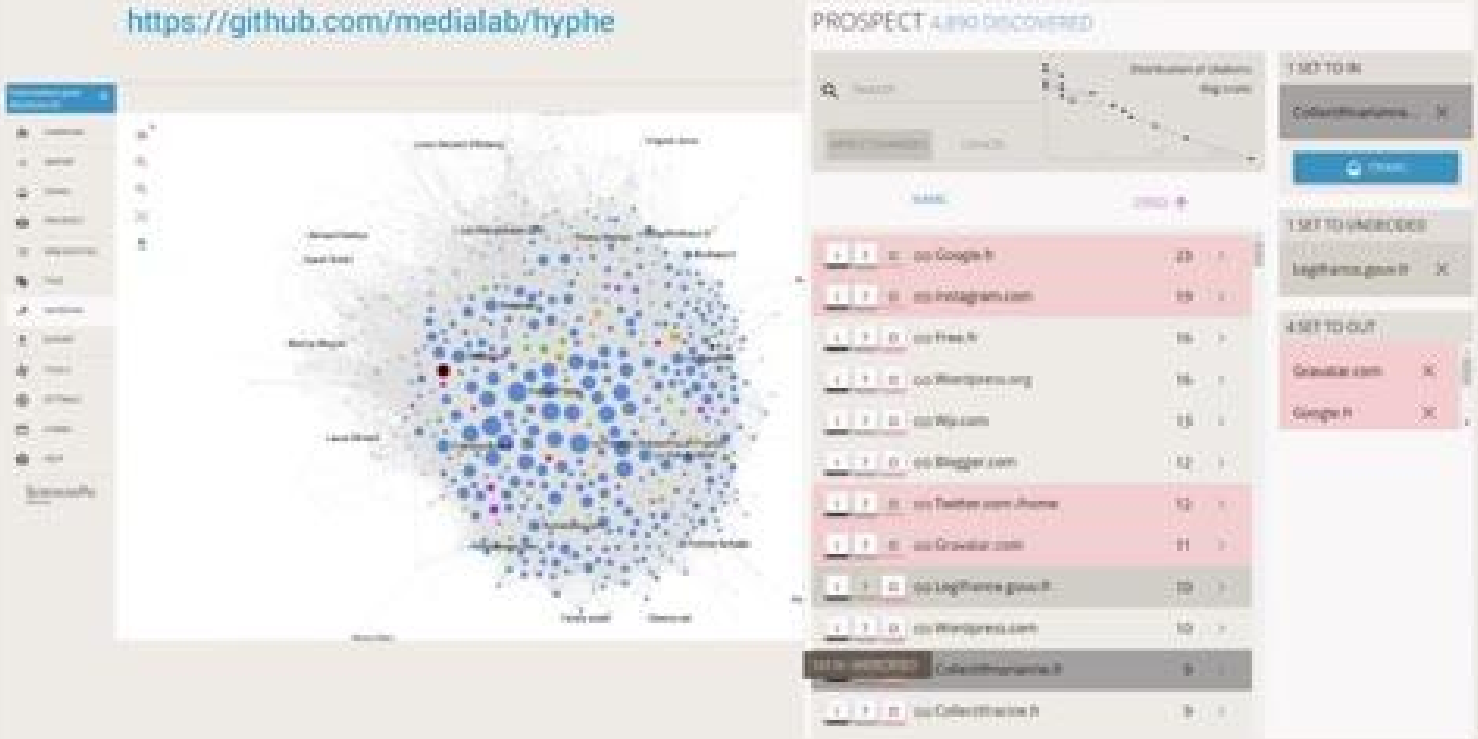
hyphe

Web application

Dynamic web entities + Iterative focused crawling

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

https://github.com/medialab/hyphe



SciencesPo
MÉDIALAB

Hyphe (by médialab Sciences Po, France)

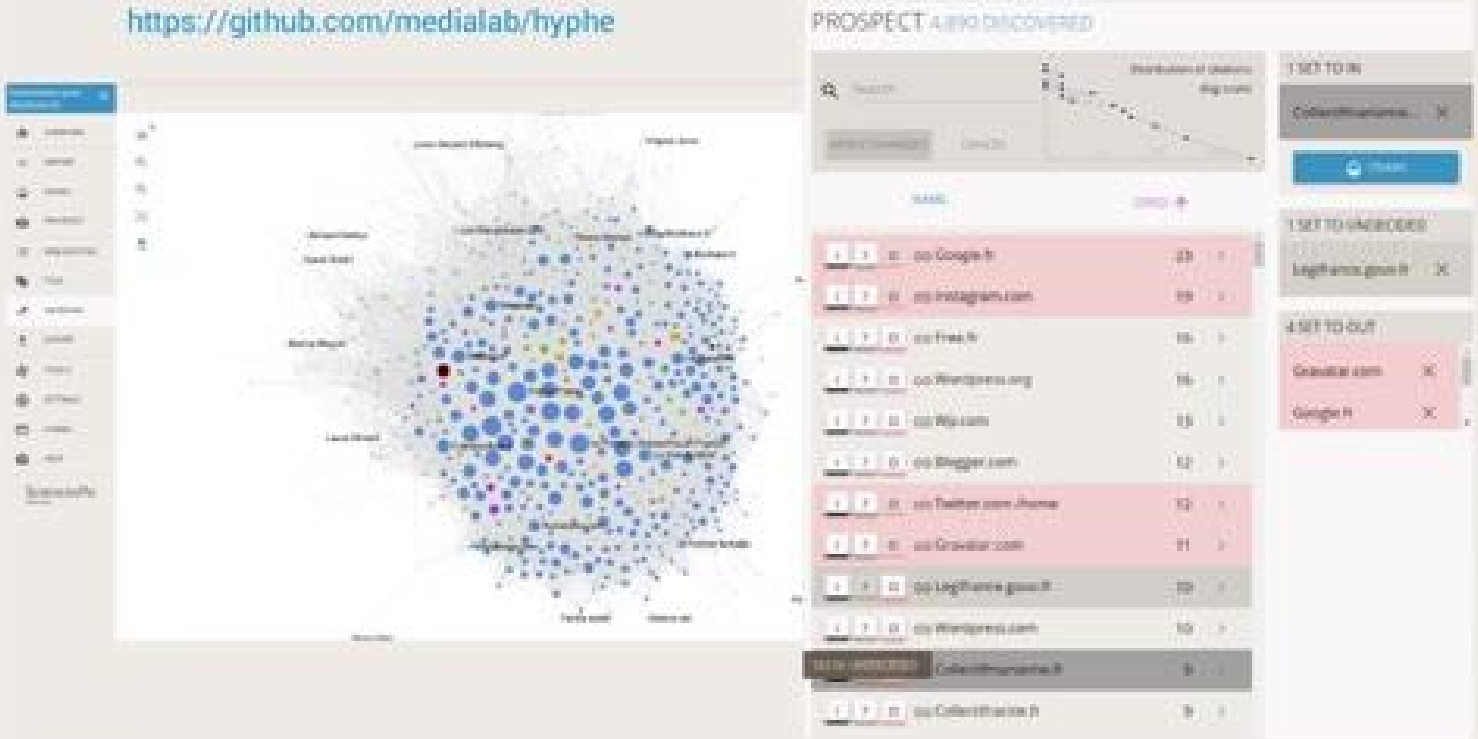
hyphe

Web application

Dynamic web entities + Iterative focused crawling

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

https://github.com/medialab/hyphe



SciencesPo
médialab

HyBro (by médialab Sciences Po, France)

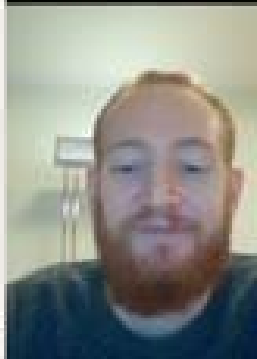
hyphe

Hyphe Browser → crawl with Hyphe, but while browsing like NaviCrawler

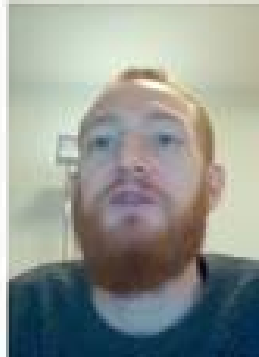
Linux / MacOS / Windows desktop application (based on electronJs)

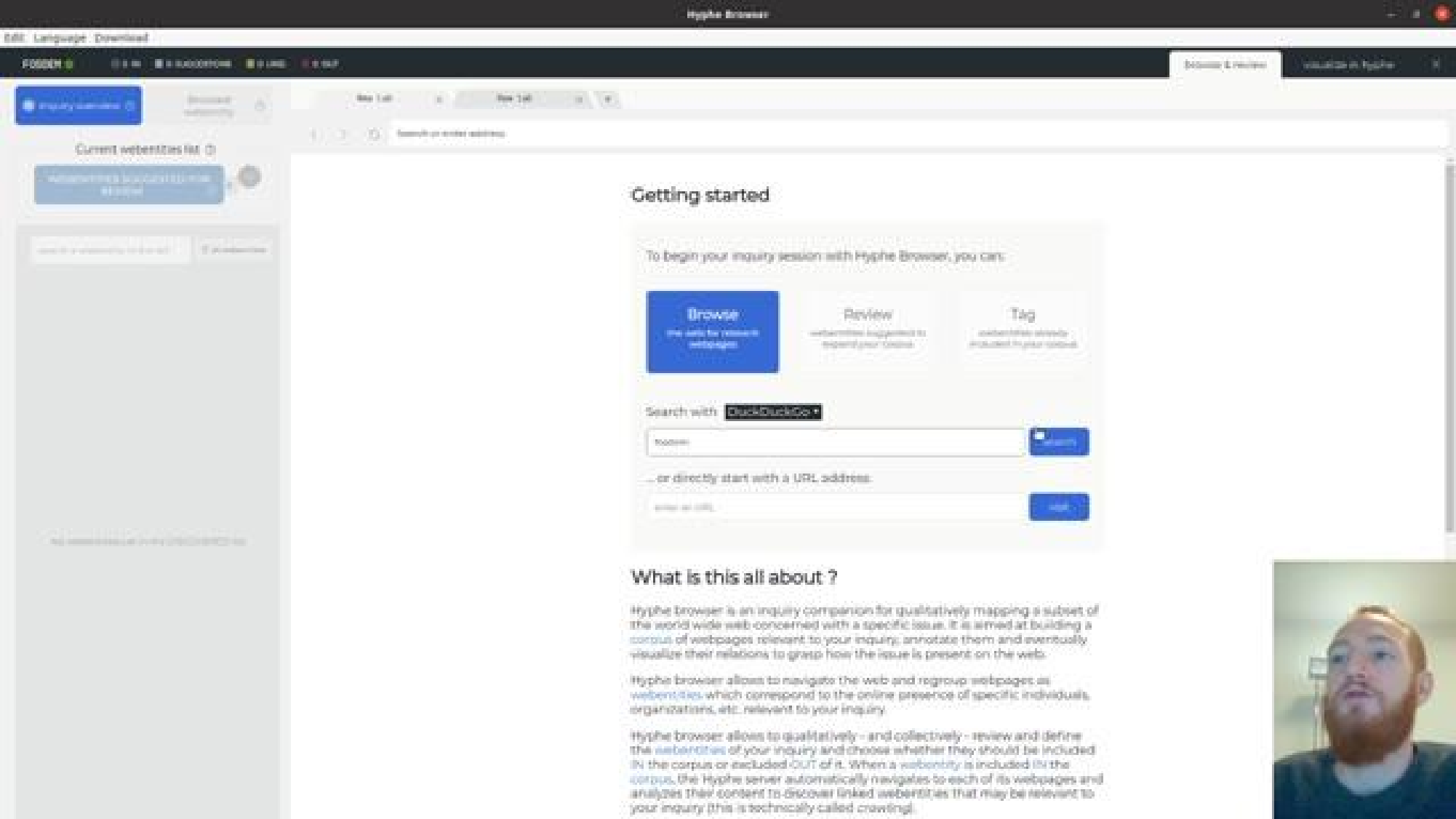
v2.0 just released with complete redesign! https://github.com/medialab/hyphe-browser/releases/

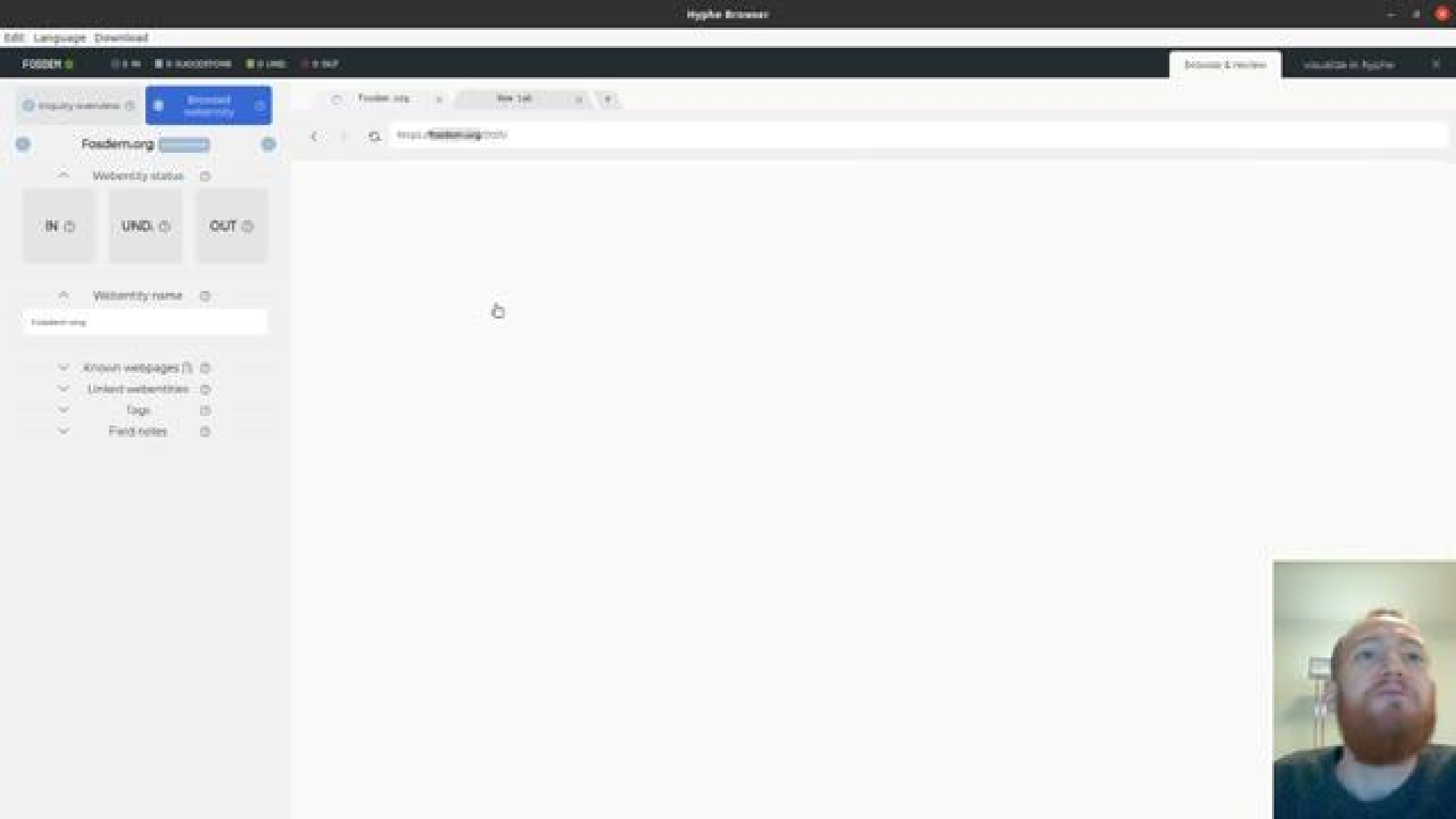


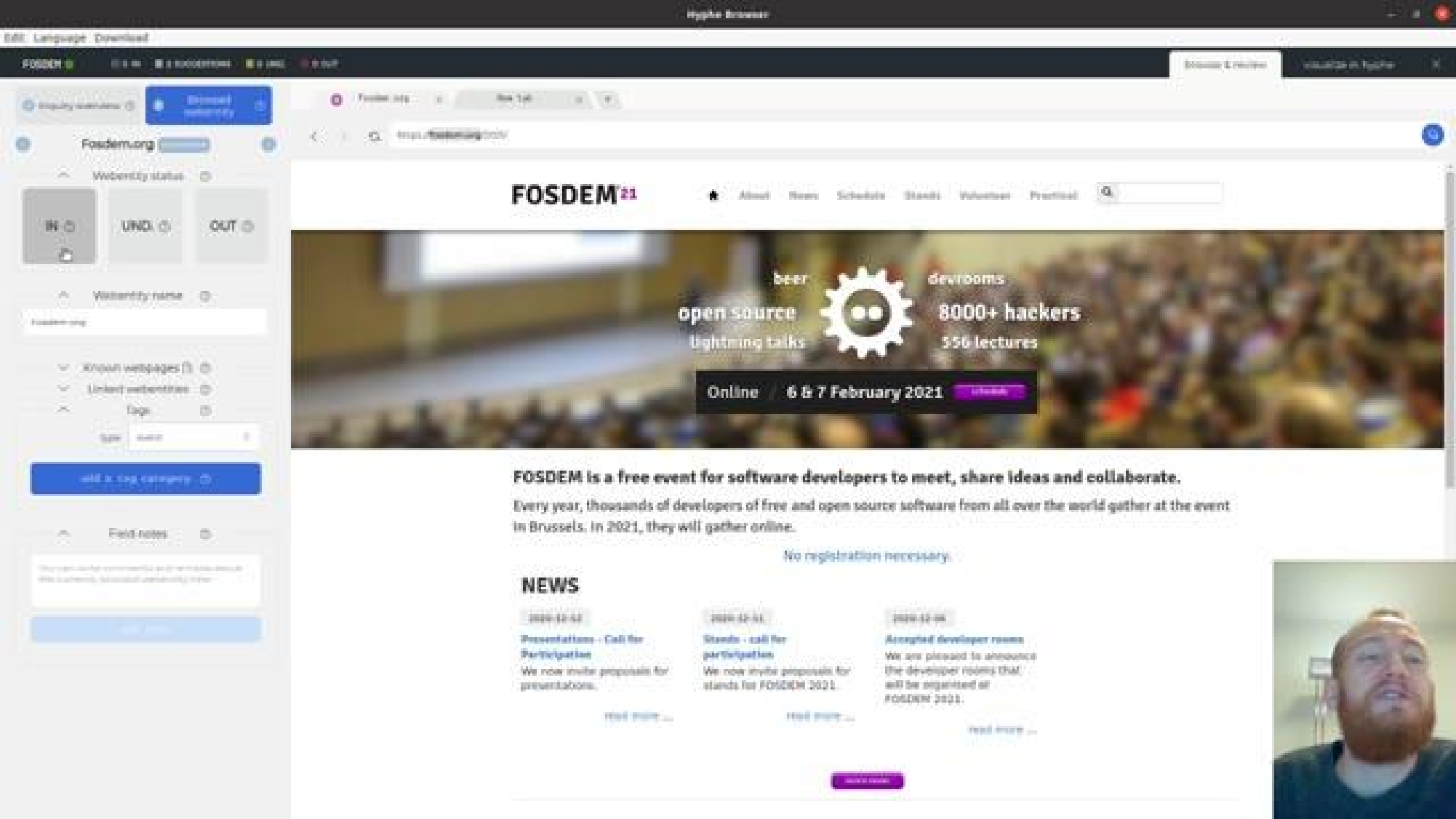


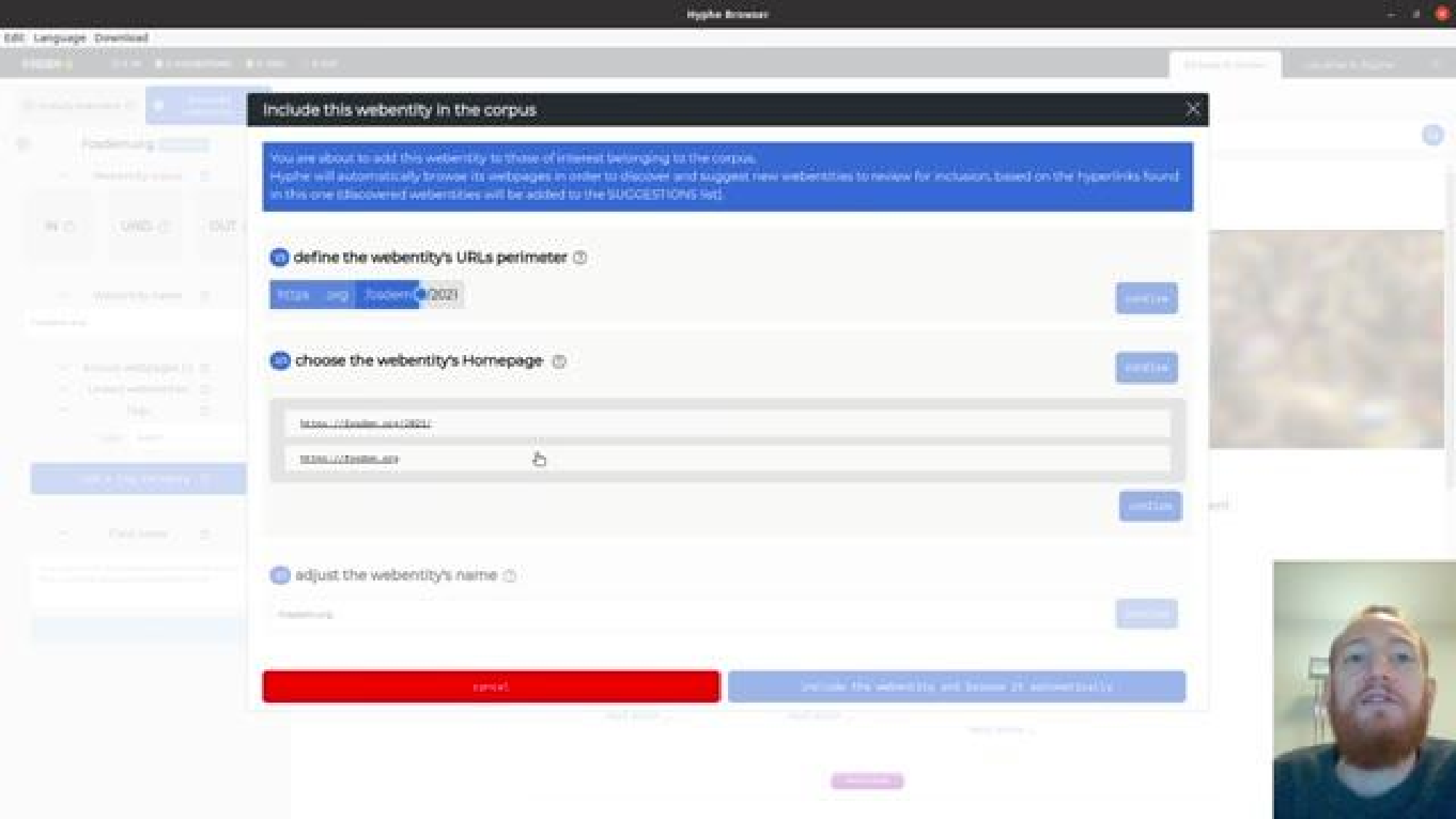

Hyphe browser
Create a corpus

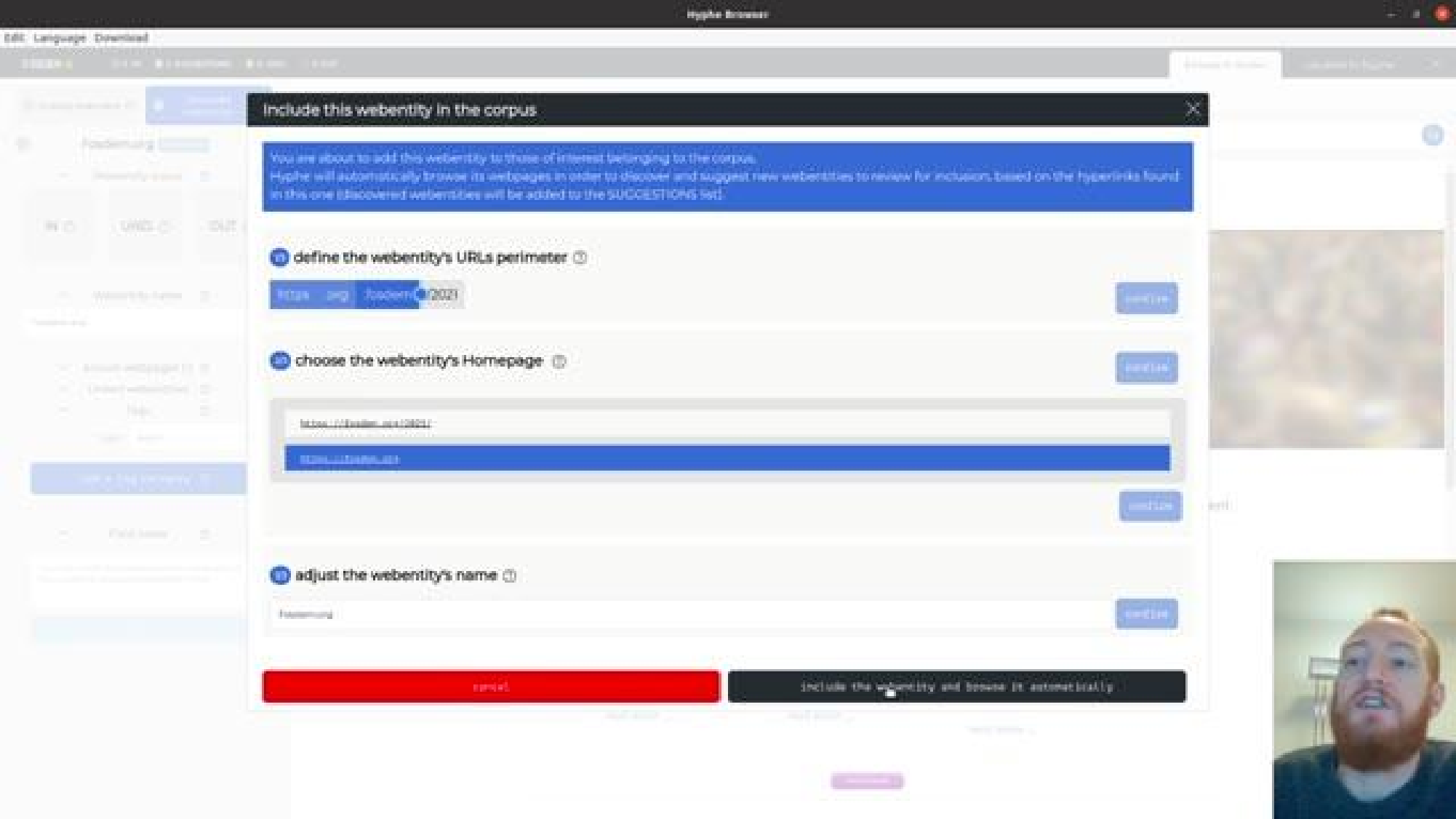


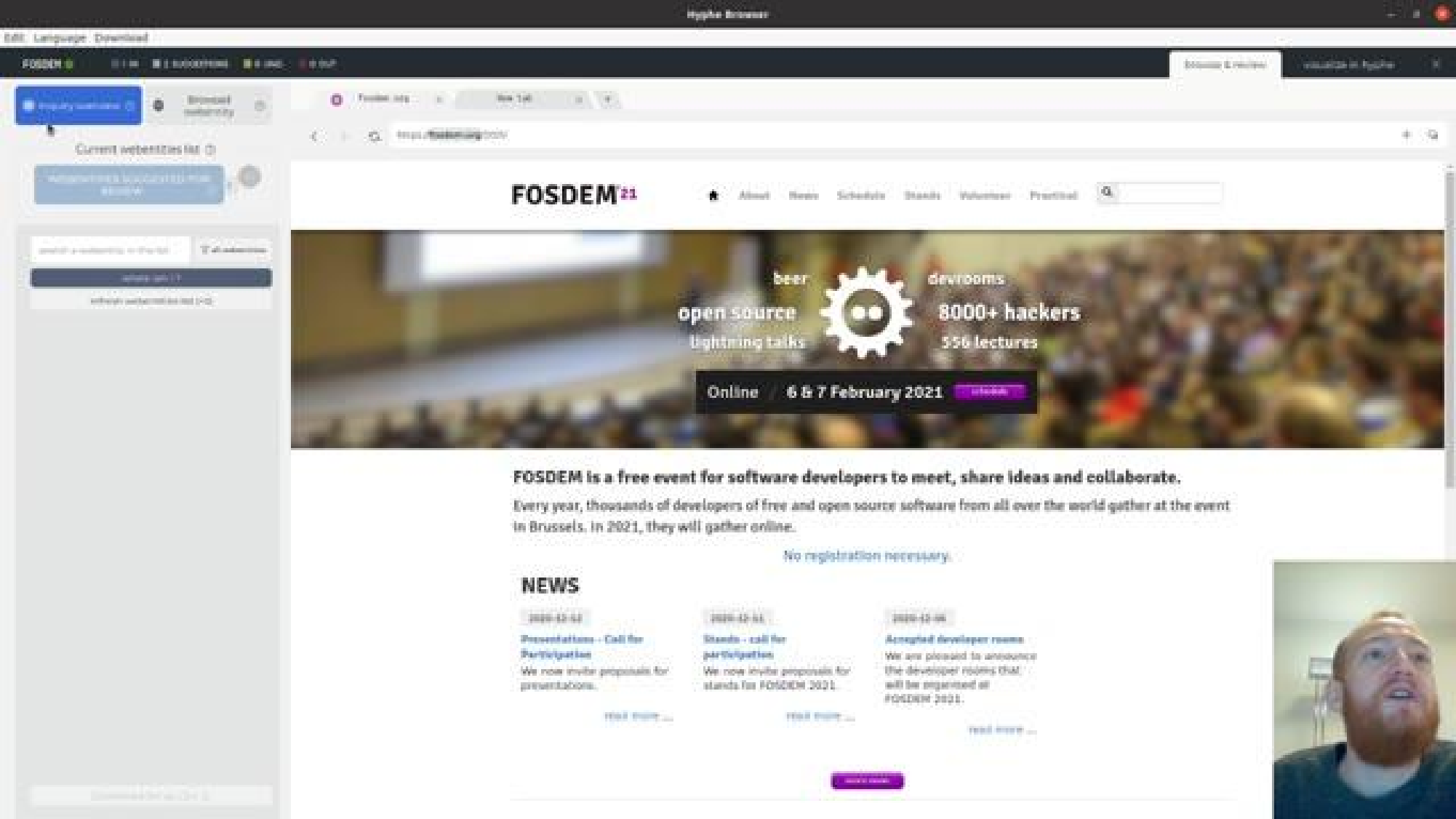

Hyphe Browser
Edit Language Download
Current webentities list
Getting started
To begin your inquiry session with Hyphe Browser, you can:
Browse
Review
Tag
Search with
...or directly start with a URL address
What is this all about ?
Hyphe browser is an inquiry companion for qualitatively mapping a subset of the world wide web concerned with a specific issue. It is aimed at building a corpus of webpages relevant to your inquiry, annotate them and eventually visualize their relations to grasp how the issue is present on the web.
Hyphe browser allows to navigate the web and regroup webpages as webentities which correspond to the online presence of specific individuals, organizations, etc. relevant to your inquiry.
Hyphe browser allows to qualitatively - and collectively - review and define the webentities of your inquiry and choose whether they should be included IN the corpus or excluded OUT of it. When a webentity is included IN the corpus, the Hyphe server automatically navigates to each of its webpages and analyzes their content to discover linked webentities that may be relevant to your inquiry (this is technically called crawling).

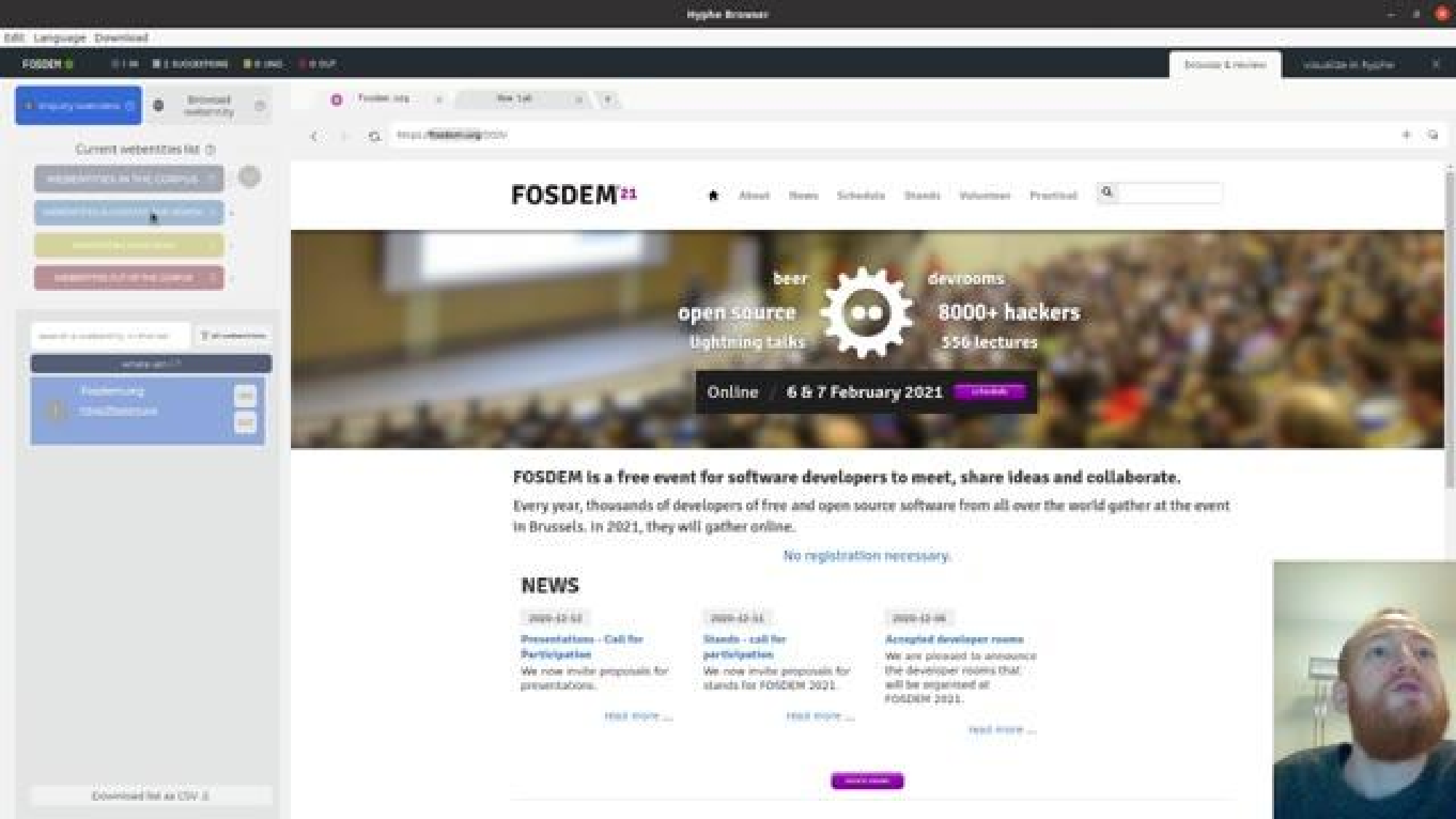

Hyphe Browser
Edit Language Download
FOSDEM
Webentity status
IN
UNDL
OUT
Webentity name
Known webpages
Undesired webentities
Tags
Field notes


Hyphe Browser
Edit Language Download
FOSDEM
Fosdem.org
Webentity status
IN
UND
OUT
Webentity name
Known webpages
Undent webentities
Tags
Field notes
FOSDEM'21
About
News
Schedule
Stands
Volunteer
Practical
beer
devrooms
open source
8000+ hackers
lightning talks
550 lectures
Online
6 & 7 February 2021
FOSDEM is a free event for software developers to meet, share ideas and collaborate.
Every year, thousands of developers of free and open source software from all over the world gather at the event in Brussels. In 2021, they will gather online.
No registration necessary.
NEWS
Presentations - Call for Participation
We now invite proposals for presentations.
Stands - call for participation
We now invite proposals for stands for FOSDEM 2021.
Accepted developer rooms
We are pleased to announce the developer rooms that will be organised at FOSDEM 2021.

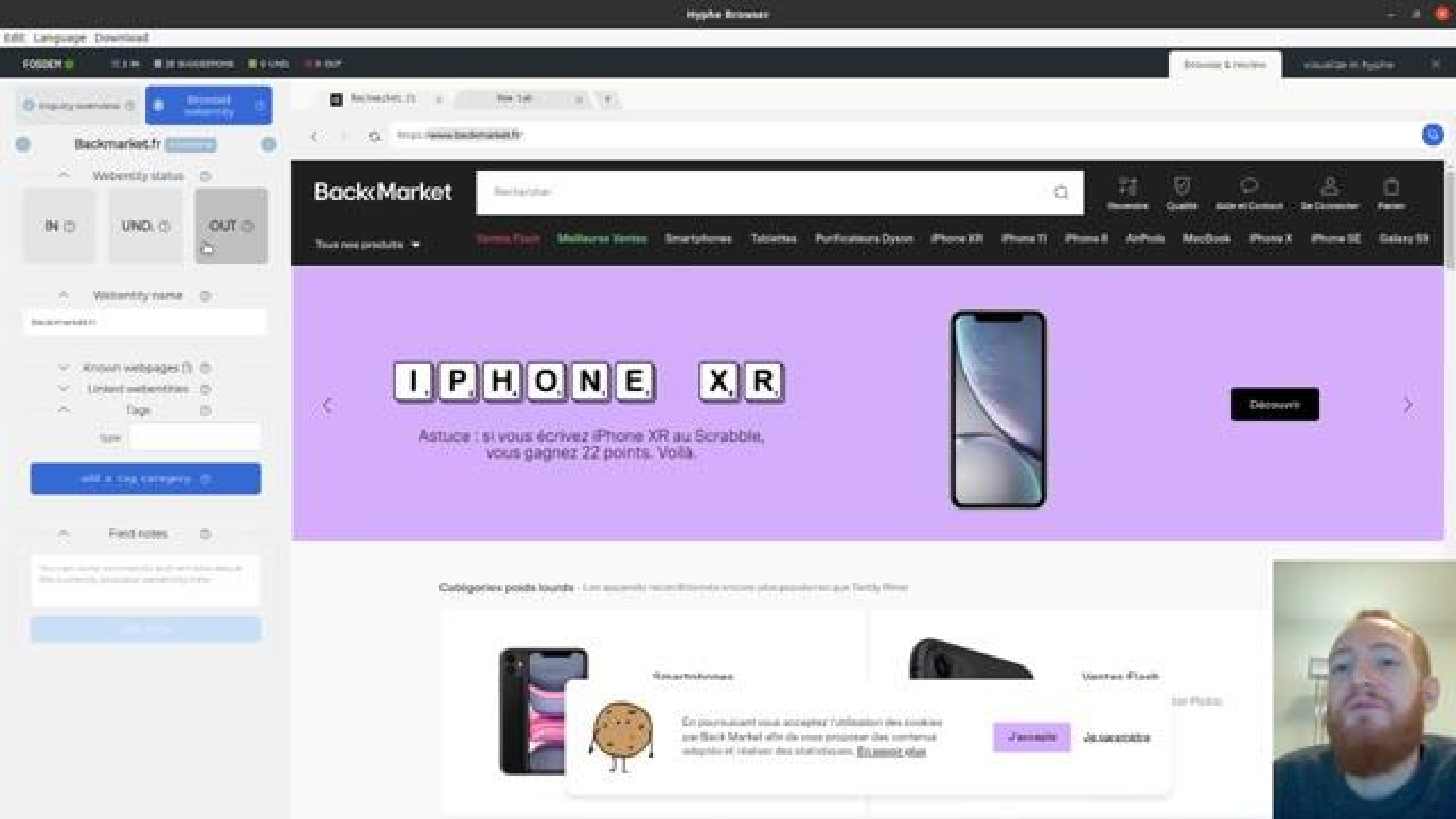

Hyphe Browser
Edit Language Download
Include this webentity in the corpus
You are about to add this webentity to those of interest belonging to the corpus.
Hyphe will automatically browse its webpages in order to discover and suggest new webentities to review for inclusion, based on the hyperlinks found in this one (discovered webentities will be added to the SUGGESTIONS list).
define the webentity's URLs perimeter
choose the webentity's Homepage
adjust the webentity's name
cancel

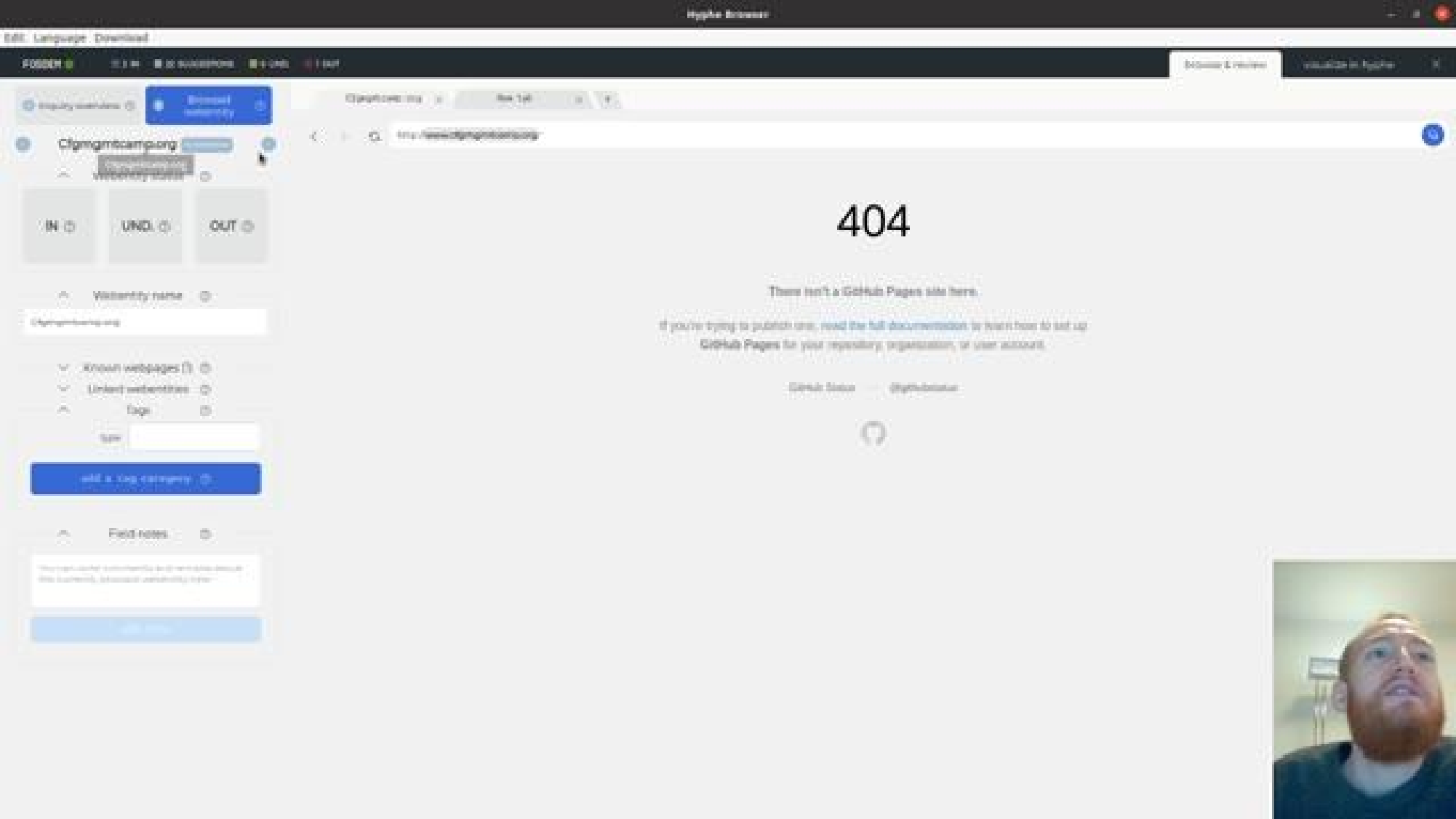

Hyphe Browser
Edit Language Download
Include this webentity in the corpus
You are about to add this webentity to those of interest belonging to the corpus.
Hyphe will automatically browse its webpages in order to discover and suggest new webentities to review for inclusion, based on the hyperlinks found in this one (discovered webentities will be added to the SUGGESTIONS list).
1 define the webentity's URLs perimeter
2 choose the webentity's Homepage
3 adjust the webentity's name
cancel
include the webentity and browse it automatically


Hyphe Browser
Edit Language Download
Current webentities list
FOSDEM 21
About News Schedule Stands Volunteer Practical
beer
devrooms
open source
8000+ hackers
lightning talks
550 lectures
Online 6 & 7 February 2021
FOSDEM is a free event for software developers to meet, share ideas and collaborate.
Every year, thousands of developers of free and open source software from all over the world gather at the event in Brussels. In 2021, they will gather online.
No registration necessary.
NEWS
Presentations - Call for Participation
We now invite proposals for presentations.
Stands - call for participation
We now invite proposals for stands for FOSDEM 2021.
Accepted developer rooms
We are pleased to announce the developer rooms that will be organised at FOSDEM 2021.


Hyphe Browser
Edit Language Download
FOSDEM
Current webentities list
FOSDEM'21
About News Schedule Stands Volunteer Practical
beer
devrooms
open source
8000+ hackers
lightning talks
550 lectures
Online 6 & 7 February 2021
FOSDEM is a free event for software developers to meet, share ideas and collaborate.
Every year, thousands of developers of free and open source software from all over the world gather at the event in Brussels. In 2021, they will gather online.
No registration necessary.
NEWS
Presentations - Call for Participation
We now invite proposals for presentations.
read more ...
Stands - call for participation
We now invite proposals for stands for FOSDEM 2021.
read more ...
Accepted developer rooms
We are pleased to announce the developer rooms that will be organised at FOSDEM 2021.
read more ...
Download list as CSV


Hyphe Browser
Edit Language Download
Apache.org
Webentity status
IN
UNDECIDED
OUT
Webentity name
Known webpages
Undesirable webentities
Tags
Field notes
APACHE CON
Content Available Now!
THE APACHE SOFTWARE FOUNDATION
20TH ANNIVERSARY
COMMUNITY-LED DEVELOPMENT "THE APACHE WAY"
THE WORLD'S LARGEST OPEN SOURCE FOUNDATION
$20B+ worth of Apache Open Source software products are made available to the public at large at 100% no cost, and benefit billions of users around the world.
Make a one-time or recurring donation | Corporate Support and ASF Sponsorship


Hyphe Browser
Edit Language Download
Backmarket.fr
Webentity status
IN
UND.
OUT
Webentity name
Known webpages
Undesired webentities
Tags
Field notes
BackMarket
Tous nos produits
IPHONE XR
Astuce : si vous écrivez iPhone XR au Scrabble, vous gagnez 22 points. Voilà.
Découvrir
Catégories poids lourds
J'accepte


Hyphe Browser
Edit Language Download
Browse & review
Visualize on Hyphe
Browse webentity
Cfgmgmtcamp.org
IN
UND.
OUT
Webentity name
Known webpages
Undent webentities
Tags
Field notes
404
There isn't a GitHub Pages site here.
If you're trying to publish one, read the full documentation to learn how to set up GitHub Pages for your repository, organization, or user account.

Hyphe Browser

Edit Language Download

Crawling this webentity has been programmed !

Creativecommons.org

Webentity status

IN | UND. | OUT

Webentity name

Known webpages

Undent webentities

Tags

Field notes

creative commons

Share your work | Use & remix | What We Do | Blog

Help us build a vibrant, collaborative global commons

Donate Now

Open access to knowledge is critical—especially during a global health emergency.

Creative Commons calls on organizations around the world to make their patents and copyrights freely available in the fight against the COVID-19 pandemic.



CC's Response to COVID-19

WHAT'S


Hyphe Browser
Edit Language Download
OVERVIEW
IMPORT
CRAWL
PROSPECT
WEB ENTITIES
TAGS
NETWORK
EXPORT
TOOLS
SETTINGS
HELP
SciencesPo
Network Viz Settings
Filtering
IN
UNDECIDED
OUT
DISCOVERED
Key