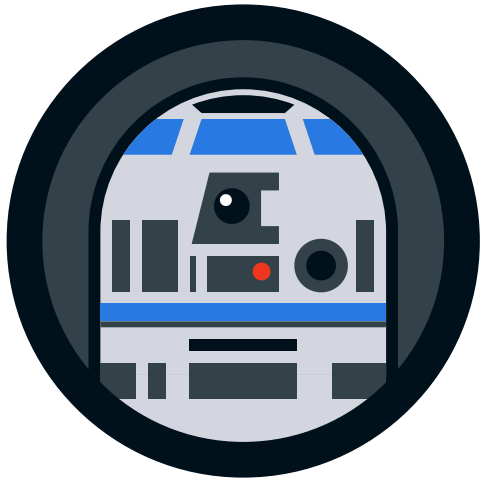


Empowering social scientists with web mining tools

FOSDEM 2020

Open Research Tools and Technologies Devroom

Guillaume Plique, SciencesPo médialab

Why and how to enable researchers to perform complex web mining tasks?

Guillaume Plique, a.k.a. Yomguithereal

SciencesPo
MÉDIALAB

What is web mining?

Scraping

Echo JS 0.11.0 **top latest random submit replies** Yomguithereal (846) | logout

Site News : Follow Echo JS on Twitter, our official account is : @echojs

Top news

▲ Understanding any and unknown in TypeScript. Difference between never and void at wanago.io▼
1 up and 0 down, posted by mwanago 15 minutes ago discuss

▲ Act now to make your React website accessible at blog.logrocket.com▼
1 up and 0 down, posted by blukaterade 46 minutes ago discuss

▲ Is there any free open source code for mobile apps like Uber? at medium.com▼
1 up and 0 down, posted by JohnLee1 2 hours ago discuss

▲ Top VScode Shortcuts For Mac and Windows at codersera.com▼
1 up and 0 down, posted by jasonrees91 3 hours ago discuss

▲ Url knife at github.com▼
1 up and 0 down, posted by strictparser 4 hours ago discuss

▲ Flutter Tutorial- Create Your First Flutter App at codersera.com▼
1 up and 0 down, posted by jonwalterc46 6 hours ago discuss

▲ Flutter ope-source UI library at bit.ly▼
1 up and 0 down, posted by navin10sharma 8 hours ago discuss

▲ 6 Superb Apps Made with Flutter at eluminoustechnologies.com▼
1 up and 0 down, posted by Kumar_Rokade 9 hours ago discuss

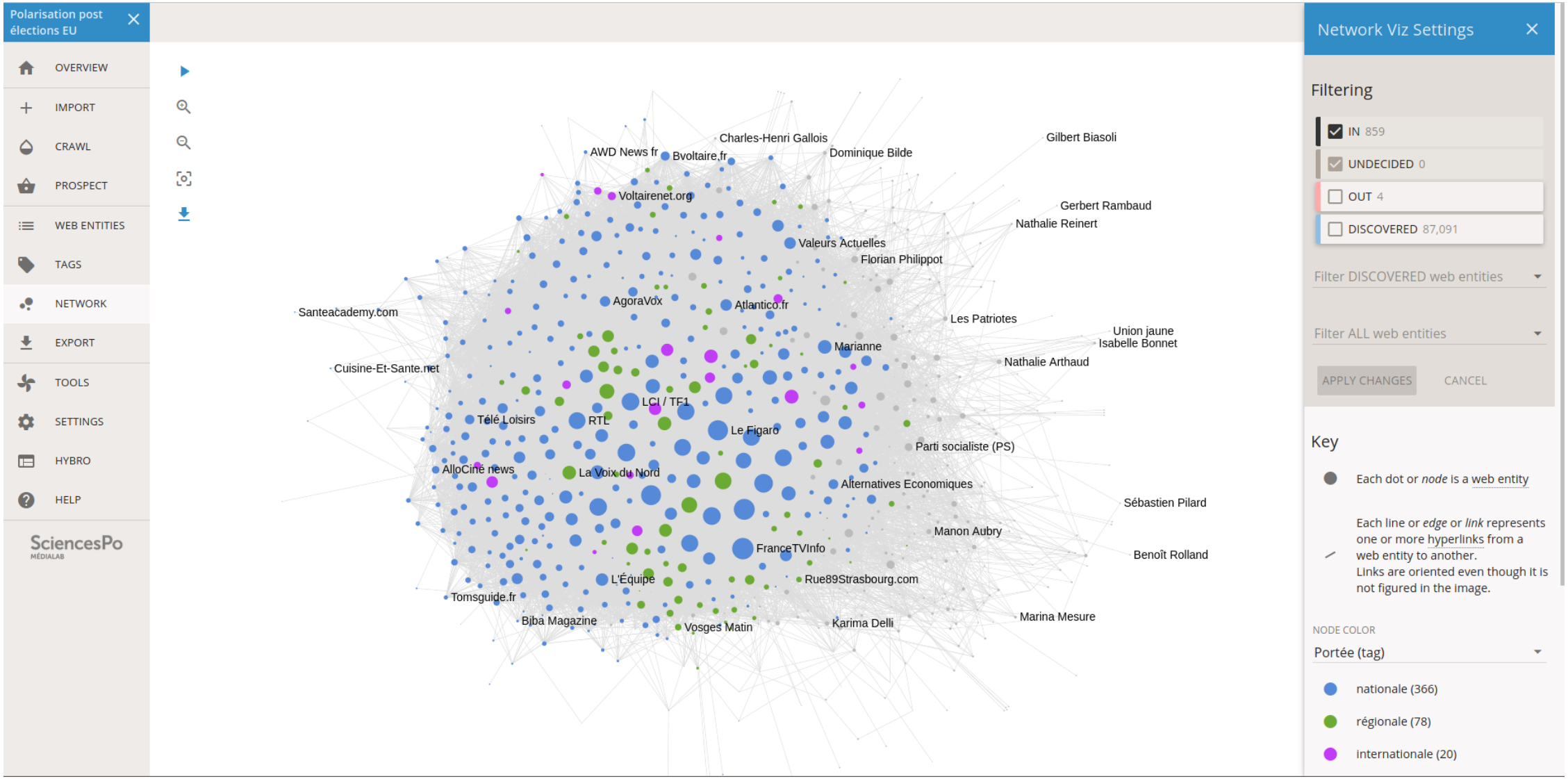
▲ PrimeNG – Why is Necessary for Angular UI Components at bit.ly▼
1 up and 0 down, posted by wuschoools 11 hours ago discuss

▲ Advanced Node.Js: A Hands on Guide to Event Loop, Child Process and Worker Threads in Node.Js at blog.soshace.com▼
5 up and 0 down, posted by vorontsova_mi@soshace.com 3 days ago discuss

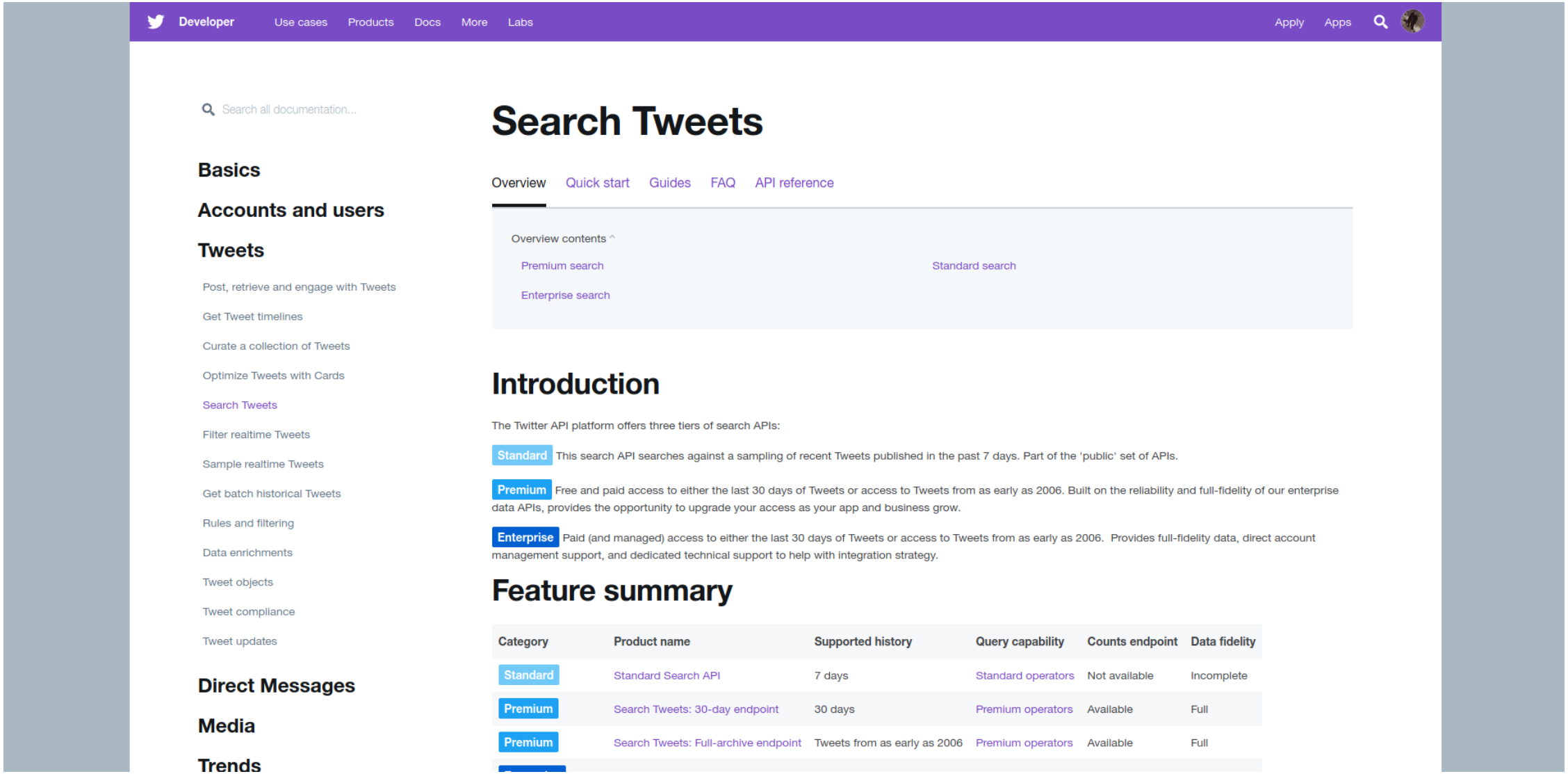
▲ Story behind Micro Frontends in Sabre at medium.com▼
4 up and 0 down, posted by navvn 3 days ago 1 comment

```
<!DOCTYPE html><html>
<head>
<script src="https://www.googletagmanager.com/gtag/js?id=UA-144247239-1" async="true"></script><script>window.dataLayer = window.dataLayer || [];function gtag(){dataLayer.push(arguments);}gtag('js', new Date());gtag('config', 'UA-144247239-1');</script><meta
<title>
Echo JS - JavaScript News
</title>
<meta content="index" name="robots">
<meta content="width=device-width, initial-scale=1, maximum-scale=1" name="viewport">
<link href="/css/style.css?v=10" rel="stylesheet" type="text/css">
<link href="/favicon.ico" rel="shortcut icon">
<script src="/js/jquery.1.6.4.min.js"></script><script src="/js/app.js?v=10"></script>
</head>
<body>
<div class="container">
<header><h1><a href="/">Echo JS</a> <small>0.11.0</small></h1><nav><a href="/">top</a>
<a href="/latest/0">latest</a>
<a href="/random">random</a>
<a href="/submit">submit</a><a href="/replies" class="replies">replies</a></nav> <nav id="account"><a href="/user/Yomguithereal">Yomguithereal (846)</a> | <a href="/logout?apisecret=f259c441ff5de09ec124c8a20e2f42a54cd2689b">logout</a></nav> <a href="#" id="l
<div id="sitenews">
Site News : Follow Echo JS on Twitter, our official account is : <a href="https://twitter.com/echojs">@echojs</a>
</div>
<h2>Top news</h2><section id="newslist"><article data-news-id="35074"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://wanago.io/2020/01/27/understanding-any-and-unknown-in-typescript-difference-between-never-and-void/" rel="nofollow">Understan
<article data-news-id="35073"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://blog.logrocket.com/make-your-react-website-accessible/" rel="nofollow">Act now to make your React website accessible</a></h2> <address>at blog.logrocket.com</address
<article data-news-id="35072"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://medium.com/@keerthanapandian/is-there-any-free-open-source-code-for-mobile-apps-like-uber-1837214825c7" rel="nofollow">Is there any free open source code for mobile
<article data-news-id="35071"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://codersera.com/blog/top-vscode-shortcuts-for-mac-and-windows/" rel="nofollow">Top VScode Shortcuts For Mac and Windows</a></h2> <address>at codersera.com</address><a
<article data-news-id="35070"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://github.com/Andrew-Kang-G/url-knife" rel="nofollow">Url knife</a></h2> <address>at github.com</address><a href="#down" class="downarrow">&#9660;</a><p><span class="up
<article data-news-id="35069"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://codersera.com/blog/first-flutter-app/" rel="nofollow">Flutter Tutorial- Create Your First Flutter App</a></h2> <address>at codersera.com</address><a href="#down" cla
<article data-news-id="35068"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://bit.ly/2tS05yN" rel="nofollow">Flutter ope-source UI library</a></h2> <address>at bit.ly</address><a href="#down" class="downarrow">&#9660;</a><p><span class="upvote
<article data-news-id="35067"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://eluminoustechnologies.com/blog/top-apps-made-with-flutter/" rel="nofollow">6 Superb Apps Made with Flutter</a></h2> <address>at eluminoustechnologies.com</address><a
<article data-news-id="35066"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="http://bit.ly/2RuuADP" rel="nofollow">PrimeNG — Why is Necessary for Angular UI Components</a></h2> <address>at bit.ly</address><a href="#down" class="downarrow">&#9660;</a>
<article data-news-id="35048"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://blog.soshace.com/advanced-node-js-a-hands-on-guide-to-event-loop-child-process-and-worker-threads-in-node-js/" rel="nofollow">Advanced Node.Js: A Hands on Guide to E
<article data-news-id="35046"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://medium.com/@elastofragmentoplast/story-behind-micro-frontends-in-sabre-9c776861433d" rel="nofollow">Story behind Micro Frontends in Sabre</a></h2> <address>at medium
<article data-news-id="35060"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://dev.to/florianrappl/5-reasons-for-doing-microfrontends-1mba" rel="nofollow">5 Reasons for Doing Microfrontends</a></h2> <address>at dev.to</address><a href="#down" c
<article data-news-id="35065"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://github.com/IsAmrish/flex-banner" rel="nofollow">Fully responsive react banner for websites</a></h2> <address>at github.com</address><a href="#down" class="downarrow"
<article data-news-id="35018"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://blog.soshace.com/create-simple-pos-with-react-node-and-mongodb-2-auth-state-logout-update-profile/" rel="nofollow">Create simple POS with React, Node and MongoDB #2:
<article data-news-id="35032"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://medium.com/javascript-scene/land-your-dream-javascript-job-with-a-better-resume-beda92bcbed6" rel="nofollow">Land Your Dream JavaScript Job with a Better Resume</a><
<article data-news-id="35064"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://github.com/Volmarg/personal-management-system" rel="nofollow">Personal Management System</a></h2> <address>at github.com</address><a href="#down" class="downarrow">&
<article data-news-id="35040"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://medium.com/@alexewerlof/my-guiding-principles-after-20-years-of-programming-a087dc55596c" rel="nofollow">My guiding principles after 9 of JavaScript development</a><
<article data-news-id="35057"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://github.com/jalal246/packageSorter" rel="nofollow">Sorting packages for monorepos production</a></h2> <address>at github.com</address><a href="#down" class="downarrow
<article data-news-id="35062"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://github.com/victorqribeiro/plot" rel="nofollow">Plot any equation with a few lines of JavaScript</a></h2> <address>at github.com</address><a href="#down" class="downa
<article data-news-id="35050"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://github.com/pioardi/poolifier" rel="nofollow">Node js Thread Pool Implementation</a></h2> <address>at github.com</address><a href="#down" class="downarrow">&#9660;</a
<article data-news-id="35047"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://efficientcoder.net/angular-tutorial-build-an-example-app-with-angular-cli-router-httpclient-and-angular-material/" rel="nofollow">Angular 9 Tutorial: Build an Exampl
<article data-news-id="35061"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://github.com/ammarahm-ed/react-native-actions-sheet" rel="nofollow">Cross Platform, highly customizable native performance ActionSheet for react native.</a></h2> <addr
<article data-news-id="34803"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://slicker.me/javascript/curves.htm" rel="nofollow">Wanna wholla lotta curves? - easy tutorial</a></h2> <address>at slicker.me</address><a href="#down" class="downarrow
<article data-news-id="35030"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://flatlogic.com/blog/5-famous-apps-built-with-react-native/" rel="nofollow">5 Famous Apps Built With React Native</a></h2> <address>at flatlogic.com</address><a href="
<article data-news-id="34842"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://dev.to/pretaporter/power-in-tiny-libraries-2m57" rel="nofollow">Power in tiny libraries</a></h2> <address>at dev.to</address><a href="#down" class="downarrow">&#9660
<article data-news-id="35027"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://dev.to/antonioru/i-m-sharing-a-collection-of-hopefully-useful-react-hooks-26bn" rel="nofollow">A collection of React hooks to speed up your development process 🍾</a
<article data-news-id="34969"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://appleple.github.io/smartblock/" rel="nofollow">Modern block styled editor powered with React and ProseMirror</a></h2> <address>at appleple.github.io</address><a href
<article data-news-id="34963"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://orizens.com/blog/how-to-not-have-a-mess-with-react-hooks-and-redux/" rel="nofollow">How To Not Have A Mess with React Hooks &amp;amp; Redux</a></h2> <address>at orizens.
<article data-news-id="34863"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://yvonnickfrin.dev/shutdown-correctly-nodejs-apps" rel="nofollow">Shutdown correctly Node.js apps</a></h2> <address>at yvonnickfrin.dev</address><a href="#down" class=
<article data-news-id="35058"><a href="#up" class="uparrow">&#9650;</a> <h2><a href="https://owlypixel.com/build-serverless-writing-pad/" rel="nofollow">Build Your Own Serverless Writing Pad with Gatsby, Netlify, and FaunaDB</a></h2> <address>at owlypixel.co
</section>
</div>
<footer><a href="/about">about</a> | <a href="https://github.com/echojs/echojs">source code</a> | <a href="/rss">rss feed</a> | <a href="https://twitter.com/echojs">twitter</a></footer><script>var apisecret = 'f259c441ff5de09ec124c8a20e2f42a54cd2689b';</scri
<div class="keyboard-help-banner banner-background banner">

</div>
 <div class="keyboard-help-banner banner-foreground banner">
<div class="primary-message">
Keyboard shortcuts
</div>
 <div class="secondary-message">
```

Crawling


Polarisation post élections EU
OVERVIEW
IMPORT
CRAWL
PROSPECT
WEB ENTITIES
TAGS
NETWORK
EXPORT
TOOLS
SETTINGS
HYBRO
HELP
SciencesPo
MÉDIALAB
Charles-Henri Gallois
AWD News fr
Bvoltaire.fr
Dominique Bilde
Gilbert Biasoli
Voltairenet.org
Gerbert Rambaud
Nathalie Reinert
Valeurs Actuelles
Florian Philippot
AgoraVox
Atlantico.fr
Santeacademy.com
Les Patriotes
Union jaune
Isabelle Bonnet
Marianne
Cuisine-Et-Sante.net
Nathalie Arthaud
LCI / TF1
Télé Loisirs
RTL
Le Figaro
Parti socialiste (PS)
AlloCiné news
La Voix du Nord
Alternatives Economiques
Sébastien Pilard
Manon Aubry
FranceTVInfo
Benoît Rolland
L'Équipe
Rue89Strasbourg.com
Tomsguide.fr
Biba Magazine
Vosges Matin
Karima Delli
Marina Mesure
Network Viz Settings
Filtering
IN 859
UNDECIDED 0
OUT 4
DISCOVERED 87,091
Filter DISCOVERED web entities
Filter ALL web entities
APPLY CHANGES
CANCEL
Key
Each dot or node is a web entity
Each line or edge or link represents one or more hyperlinks from a web entity to another.
Links are oriented even though it is not figured in the image.
NODE COLOR
Portée (tag)
nationale (366)
régionale (78)
internationale (20)

Collecting data from APIs



But why is this useful to [social] sciences?

Bad take

01. Every social sciences data collection is biaised (i.e. observer's paradox)
02. People express themselves without being asked to, on the Internet
03. What's more they are not being observed (lol, I know...)
04. Web mining is therefore a superior source of data for social sciences!

Good take

01. Internet data comes with its own biases that you should be aware of
02. Apply `media studies` and `STS` without moderation
03. Still is another very interesting and large data source!

Web mining is hard

You need to know The Web™:

`DNS` `HTTP` `HTML` `CSS` `JS` `DOM` `AJAX` `SSR` `CSR` `XPATH` ...

How do you teach researchers web technologies

01. The same as anyone else really (CSS as sushi plates anyone?)
02. What most consider as an easy layer of technologies really ISN'T
03. We really are standing on the shoulders of giants

Teaching researchers how to scrape

01. Fighting the platforms and their APIs
02. Legal issues in some countries
03. Sometimes forbidden to teach it (~lock picking)
04. Publication wiggles (the monkey army)

Jupyterizing researchers is not a solution

01. Some researchers don't have the **time** nor the **will** to learn python and web stuff.
02. We should be OK with that!

Web mining is HARD

It really is a craftsmanship.

Internet is a dirty, dirty place

Browsers truly are heuristical wonders!

Multithreading, parallelization, throttling etc.

Once we cut access to Google to our whole university!

Complex spidering, scalability, storage, indexing, recombobulation, steam engines, fancy boats, unionization, agility, upper management, Peters syndrom, eXtreme programming

Most of it is irrelevant and made up but you get the point...

How do we empower researchers then?

By **designing** tools suited to their research questions

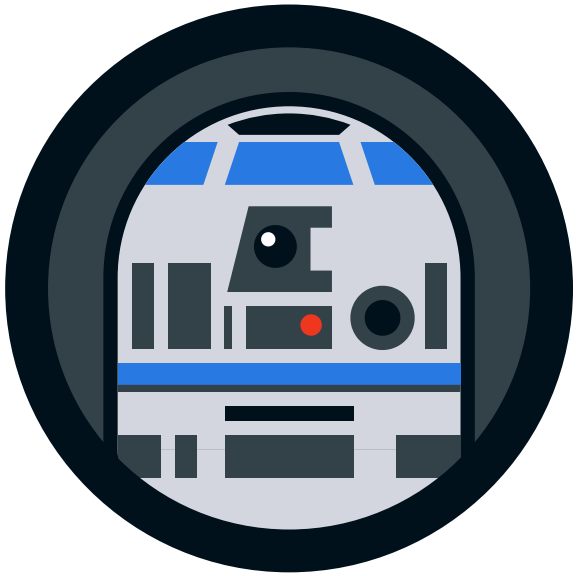
SciencesPo's médialab

01. Social Science Researchers
02. Designers
03. Engineers

A brief guided tour of tools we designed

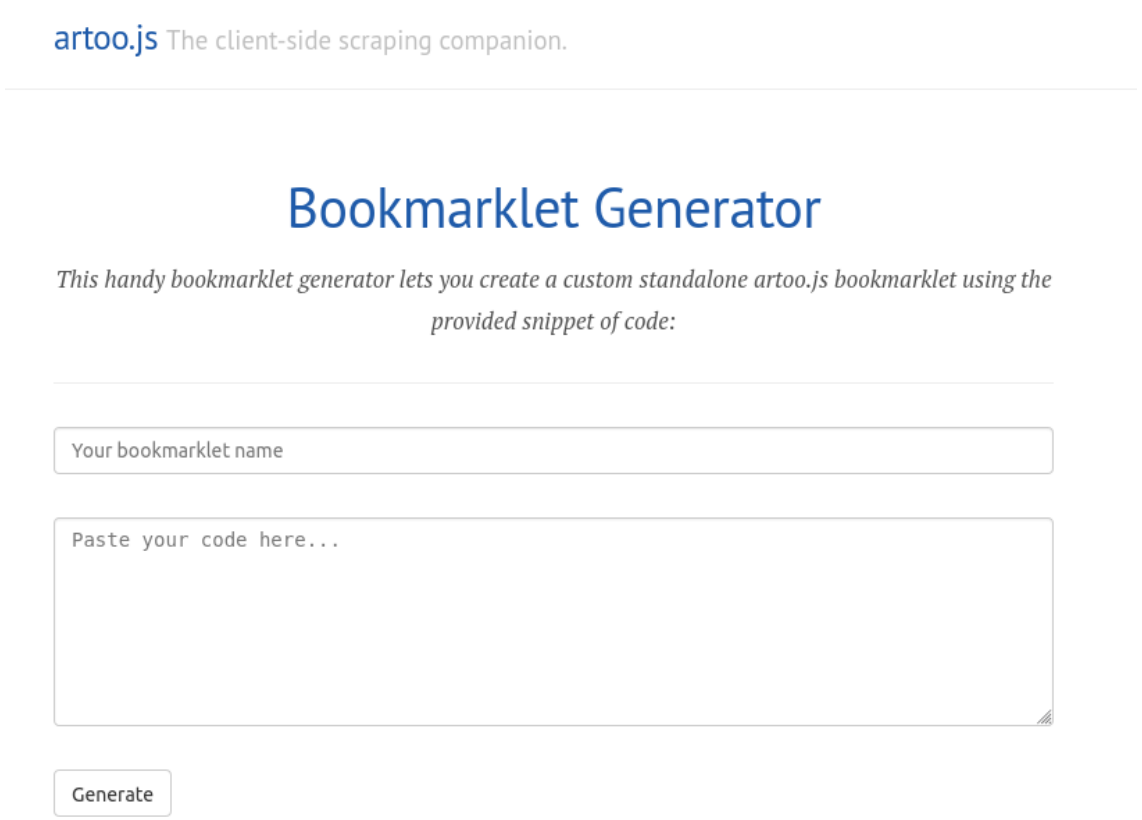
01. artoo.js
02. minet
03. Hyphe
04. (Gazouilloire)

Parasitizing web browsers instead of emulating them!



Demo Time!

Leveraging bookmarklets to empower researchers



But can we scale up?

Not-contractual logo - Jules Farjas ©

Handling the pesky details for you

01. Multithreaded, memory-efficient **fetching** from the web.
02. Multithreaded, scalable **crawling** using a comfy DSL.
03. Multiprocessed raw text **content extraction** from HTML pages.
04. Multiprocessed **scraping** from HTML pages using a comfy DSL.
05. **URL-related heuristics** utilities such as normalization and matching.
06. Data collection from various **APIs** such as CrowdTangle.

The Unix philosophy

Do one thing well

```
xsv search -s url urls.csv | minet fetch url -d html > result.txt
```

Demo time!

The low-fi approach

```
# Yomgui at mbp-de-plique-1.home in ~/code/minet on git:master ✖ [15:54:11]
→ ./ftest/fetch.sh
Fetching pages:   0%|                                                    | 0/100 [00:00<?, ? urls/s]
```

```
# Yomgui at mbp-de-plique-1.home in ~/code/minet on git:master ✖ [15:51:23]
→ ./ftest/fetch.sh
```

Relocalizing data collection

01. Sometimes you don't need a server
02. We are rarely doing BigData™
03. Let's put the researcher at the center so they can control their data

A programmatic API

Jupyter's back y'all!

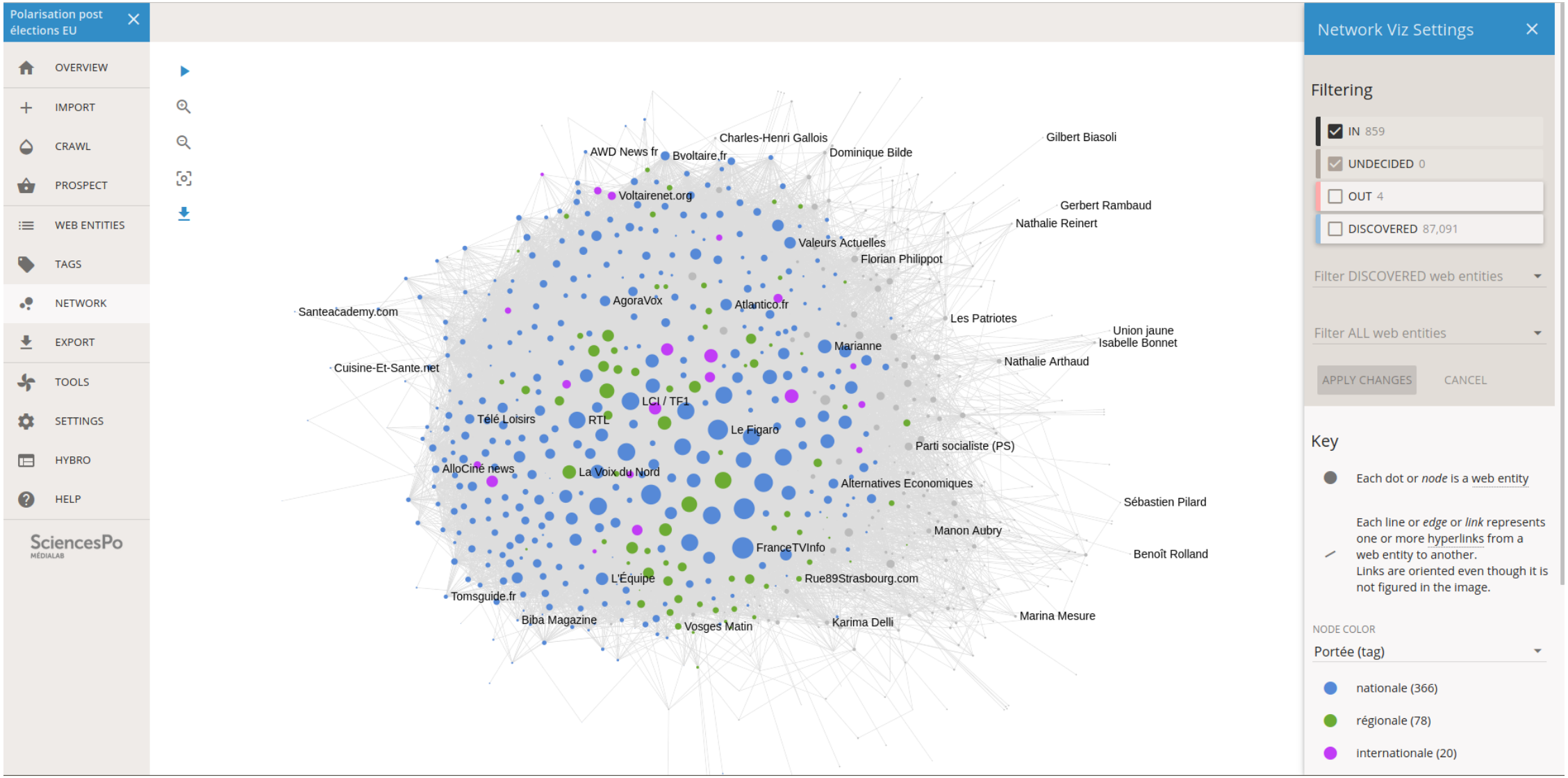
```
from minet import multithreaded_fetch

for result in multithreaded_fetch(urls_iterator):
  print(result.status)
```

How to enable researchers to crawl the Web?

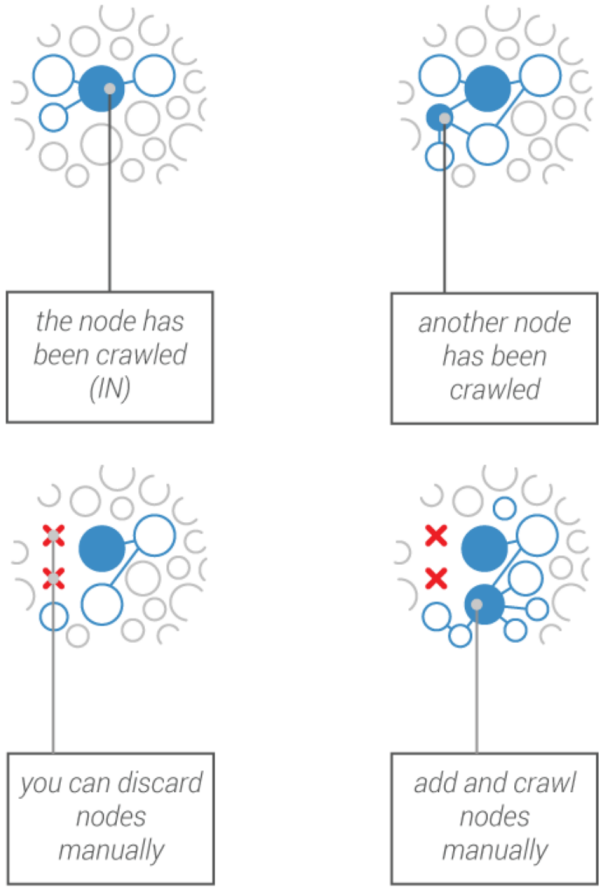
HYPHE

Hyphe is a web corpus curation tool
featuring a research-driven web crawler

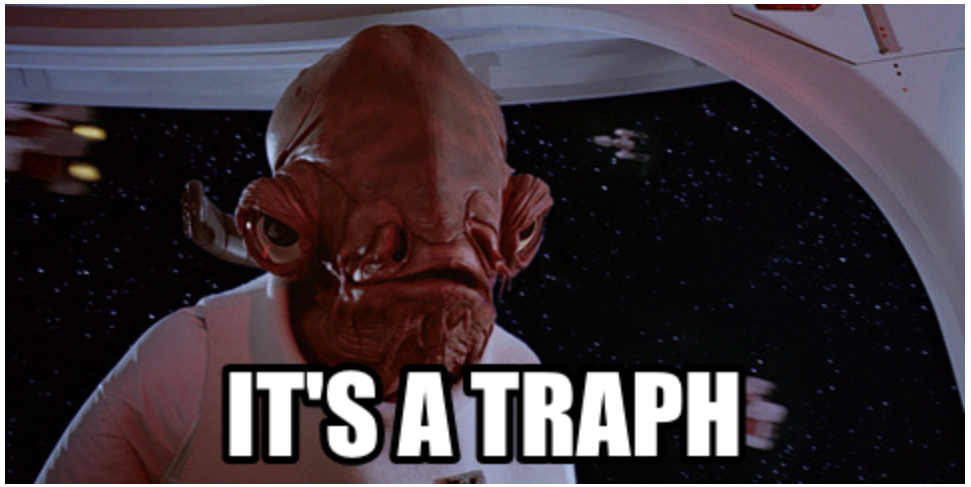
A dedicated interface



Serving a robust methodology



Non-trivial technical challenges



Trade-off between scalability & usability

We need to be able to **design** user paths.

The future!

What about a GUI for minet?

Thank you for listening!