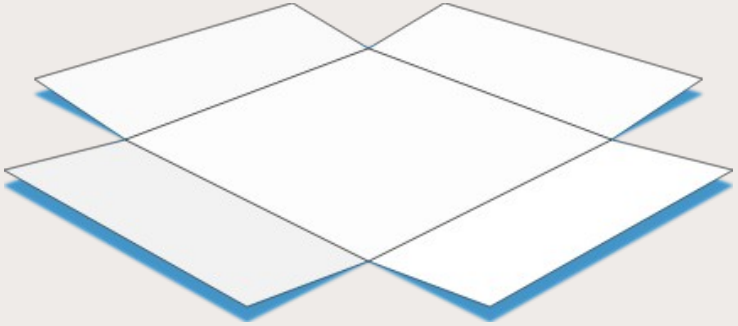
Hyperlink is not dead!

Benjamin Ooghe-Tabanou, Mathieu Jacomy, Paul Girard, Guillaume Plique



Equipex DIME-SHS
ANR-10-EQPX-19-01

Hyperlink is not dead!

Digital Tools & Uses Congress

MSH Paris Nord - October 4th, 2018

Benjamin Ooghe-Tabanou – Mathieu Jacomy – Paul Girard – Guillaume Plique
Sciences Po médialab (@medialab_ScPo)

SciencesPo
MÉDIALAB

Hyperlink is not dead!

Benjamin Ooghe-Tabanou
Sciences Po, médialab
Paris, France
benjamin.ooghe@sciencespo.fr

Mathieu Jacomy
Sciences Po, médialab
Paris, France
mathieu.jacomy@gmail.com

Paul Girard
Sciences Po, médialab
Paris, France
paul.girard@sciencespo.fr

Guillaume Plique
Sciences Po, médialab
Paris, France
guillaume.plique@sciencespo.fr

ABSTRACT

The emergence and success of web platforms nurtured a trend within social studies: "Hyperlink is dead!". Capturing their users into mobile applications and specialised web interface to propose them a specific user experience (and business model), the platforms indeed created new information silos in the open World Wide Web space. The simplified availability of user behavioural data through these platforms APIs reinforced this idea in academic communities by providing scholars with an easy way to collect rich user centric data for their research. After discussing the methodological aspects of the web divide between platforms and classical websites, we will argue that although it becomes more and more invisible, the hyperlink, modern incarnation of intertextual links between documents, is still a central and structural element of the web. Hyperlinks remain an invaluable resource to turn the web into a research field in spite of the complexity to collect, manipulate and curate them. We will illustrate those methodological challenges by describing the choices we made in designing Hyphe, a tool dedicated to the creation of web corpora tailored for mining hypertexts.

CCS CONCEPTS

• **Information systems** → **Web mining**; **Web applications**; **Internet communications tools**;

KEYWORDS

Hyperlink, hypertext, web mining, crawler, corpus, curation, network analysis.

ACM Reference Format:
Benjamin Ooghe-Tabanou, Mathieu Jacomy, Paul Girard, and Guillaume Plique. 2018. Hyperlink is not dead!. In *International conference on Web Studies (WS.2 2018), October 3–5, 2018, Paris, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3240431.3240434

1 ARE HYPERTEXT STUDIES OUTDATED?

The World Wide Web's original design as a vast open documentary space built around the concept of hypertext made it a fantastic research field to study networks of actors. As a literary technology, hyperlinks concepts are anything but new: "*Links are intrinsic to documents, and have been for millennia*" [27]. According to Bardini [5], two main concepts are embedded in hypertexts: association and connection. Hypertexts allow to create conceptual associations between documents - a powerful function when used in a free and creative context - but they can also be very efficient in connecting documents to communicate ideas (i.e. conceptual links) within a community.

When Tim Berners-Lee proposed his World Wide Web (WWW) project, hypertexts were presented more as a way to connect documents to enhance communication through navigation than as form of conceptual associations: "*The texts are linked together in a way that one can go from one concept to another to find the information one wants. (…) The process of proceeding from node to node is called navigation*" [6]. But following Bardini [5], if associative hyperlinks are created freely by authors for their own use, connective ones which have a value within a community are more likely to be moderated. Although the goal of the WWW is to enhance information flows across communities, connections between documents are not controlled, as each individual website's author is responsible only for the connections from his website to the rest of the WWW.

1.1 Hyperlinks directionality: a bottom-up hierarchy

This directionality of the links reveals asymmetrical associations between the linked documents: the referer knows the referee but not necessarily the other way around. Considering hyperlinks as references provides powerful insights on the distribution of influence on the web. As the study of complex networks has demonstrated, online connections are not randomly distributed across the web. According to the principle known as the "Matthew effect" or "preferential attachment" [4], new web documents tend to cite the already most cited documents, reinforcing the concentration of links to a small fraction of pages. A hierarchy naturally emerges from this pattern across all scales of the web: it can be observed locally (eg. inside Wikipedia) as well as in its general structure. This hierarchy is bottom-up because it emerges spontaneously rather than by design, but also because hyperlinks tend to flow from a metaphorical bottom to a metaphorical top. Actors with high visibility drawing most of citations are a handful compared to the mass of low visibility actors who cite them. The structure emerging from the direction of hyperlinks was famously leveraged by Google's

WS.2 2018, October 3–5, 2018, Paris, France

ACM ISBN 978-1-4503-6438-6/18/10.
https://doi.org/10.1145/3240431.3240434

http://hyphe.medialab.sciences-po.fr/docs/20181004-ACM-WebStudies-HyperlinkIsNotDead.pdf
https://doi.org/10.1145/3240431.3240434

1. From the web to platforms:
   a (not so) brief history of hyperlinks
2. médialab: enabling digital field work
   through design & engineering
3. Hyphe: curate hypertexts into
   web corpora
4. Many angles of hypertext studies

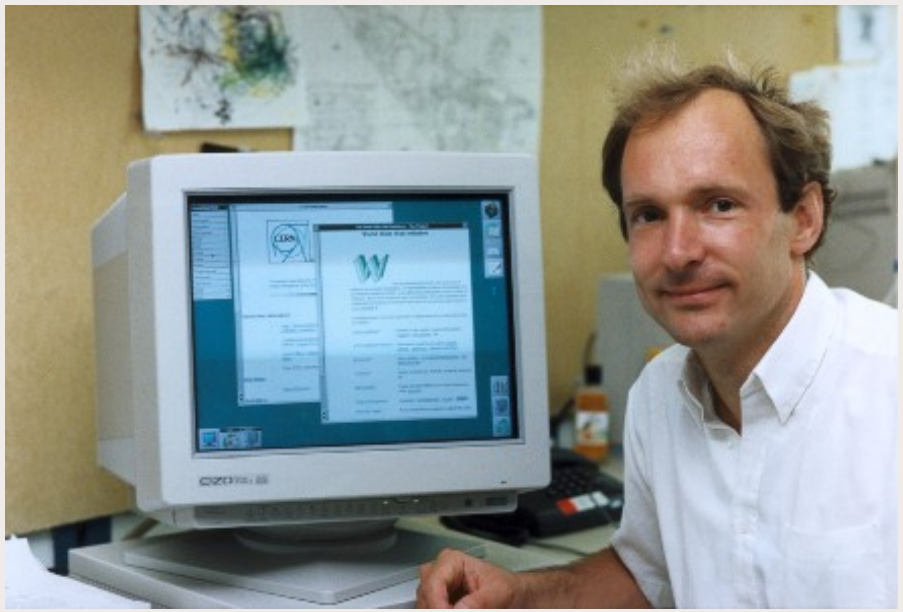
1. From the web to platforms:
   a (not so) brief history of hyperlinks
2. médialab: enabling digital field work through design & engineering
3. Hyphe: curate hypertexts into web corpora
4. Many angles of hypertext studies

The Web: a vast open documentary space

« Links are intrinsic to documents,
and have been for millennia. »

Nelson, T. H. (1992). Literary Machines 93.1. Sausalito, CA: Mindful Press

Hyperlinks: the backbone of the world wide web



*« 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

Hyperlinks: a familiar & transparent feature

« But from the very beginning […]
the web has had ***one defining feature*** *that*
we tend to overlook today, because it has become
so intuitive and natural that it goes unnoticed*. »*

Brügger, 2017, Connecting textual segments: A brief history of the web hyperlink

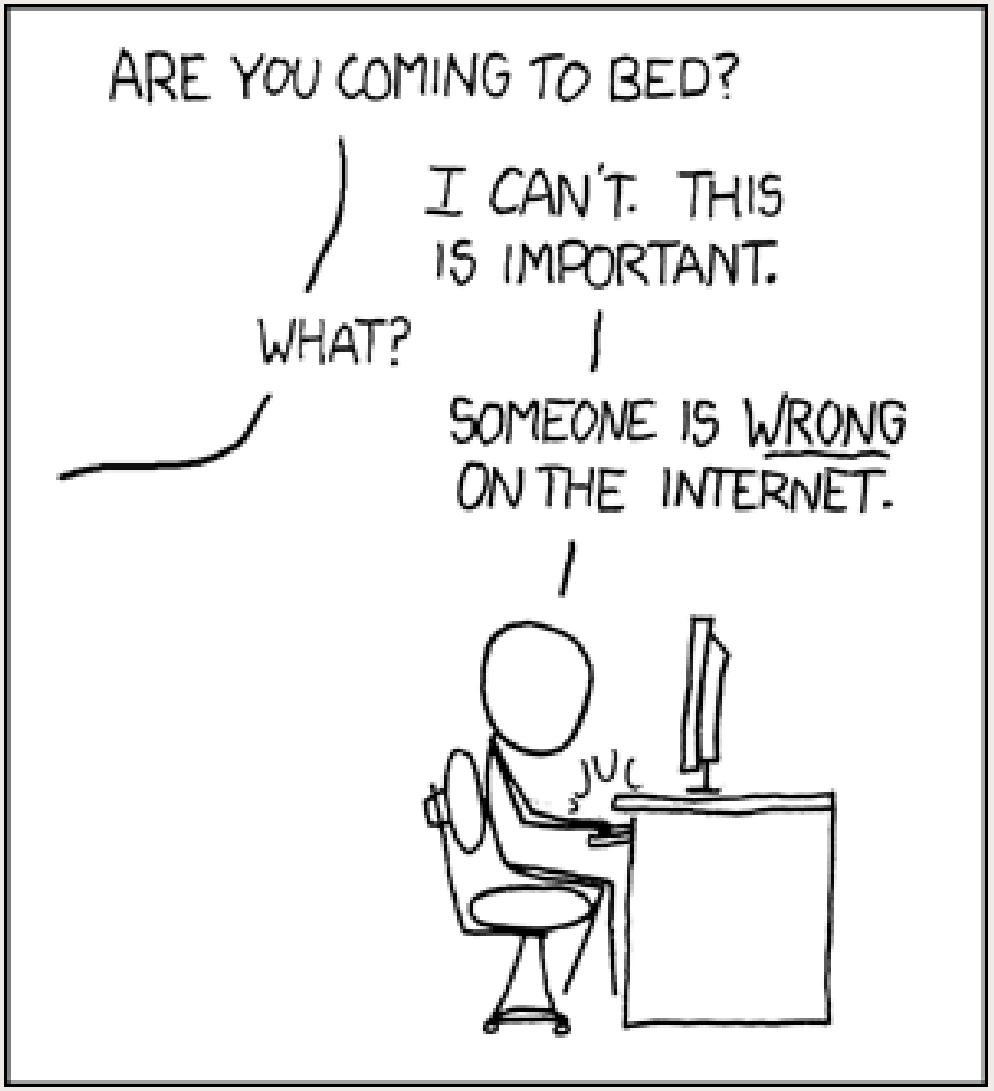
Hyperlinks reflect assymetrical associations

*« 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?

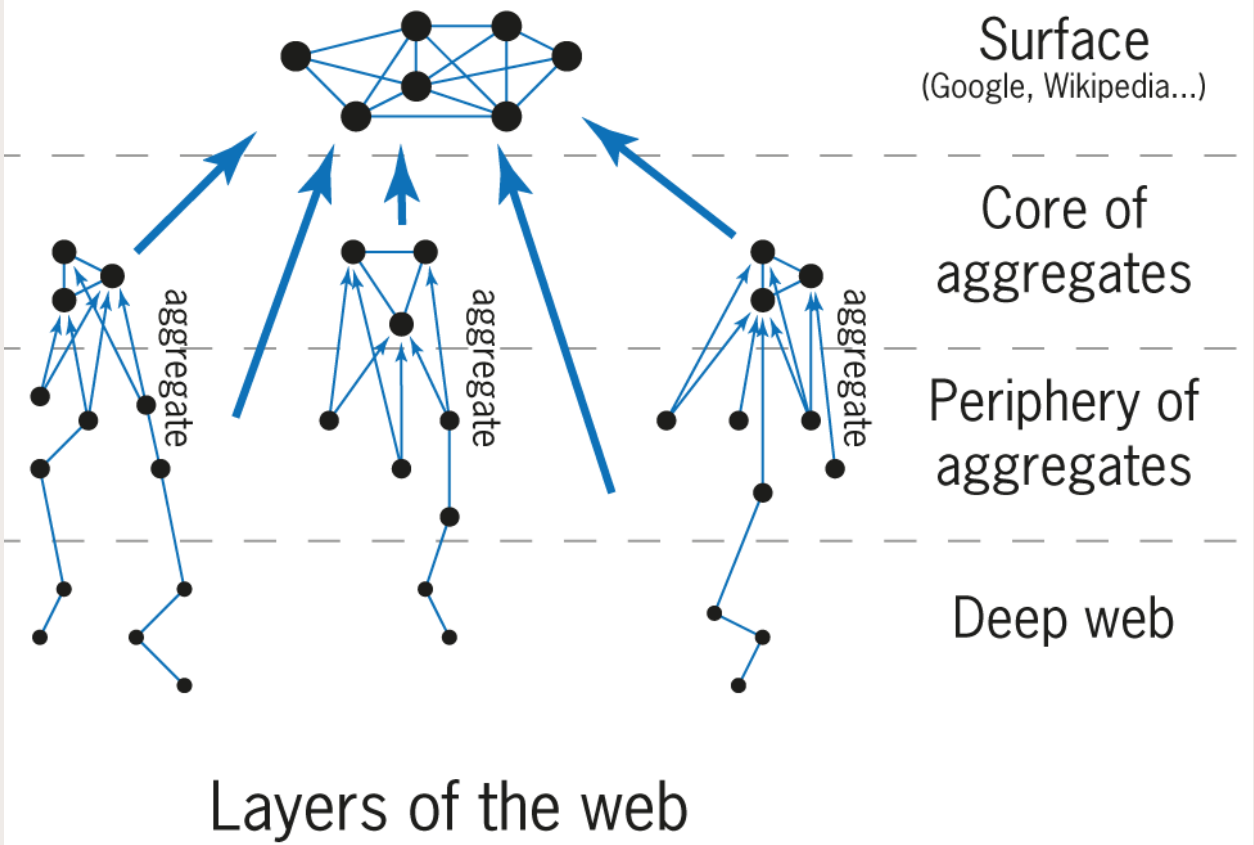
The Web: a place of dialogue and debates



A bottom-up hierarchy emerged from hyperlinks

« Matthew effect » : preferential attachment

→ new web pages tend to cite the already most cited ones



The 2010's divide: the Web Vs. gated platforms

- 1999: Google's « PageRank » algorithm
  L. Page, S. Brin & al. 1999. ThePageRank Citation Ranking: Bringing Order to the Web.
  → development of Search Engines Optimization
  → hyperlinks at the core of online visibility
- 2004: Facebook
- 2006: Twitter
- 2010's: « platformization » of the web, smartphones
  → capturing users with sharing buttons behind closed gates
  → increasing importance of likes & retweets
- 2018: legal definition of platforms in EU copyright reform

*« The definition of an online content sharing service provider under this Directive shall cover **information society service providers** one of the main purposes of which is to **store and give access to the public or to stream significant amounts of copyright protected content** uploaded / made available by its users, and that **optimise content**, and **promote for profit** making purposes, including amongst others displaying, tagging, curating, sequencing, the uploaded works or other subject-matter, irrespective of the means used therefor, and therefore act in an active way. »*

Are hyperlinks studies outdated?

- A new trend in Social Sciences: will LIKES replace LINKS?

- The rise of APIs in SHS:

  → simple access to massive and structured user centric data
  (Twitter, Facebook…)

  → development of Single platform studies

  → risks of commercial & ethical close-back
  (LinkedIn, Cambridge Analytica…)

- Working with the web otherwise can be hard:
  heterogeneous formats, unstructured messy data...

Bruno Latour, médialab founder

« Google is nice, but we need something better »

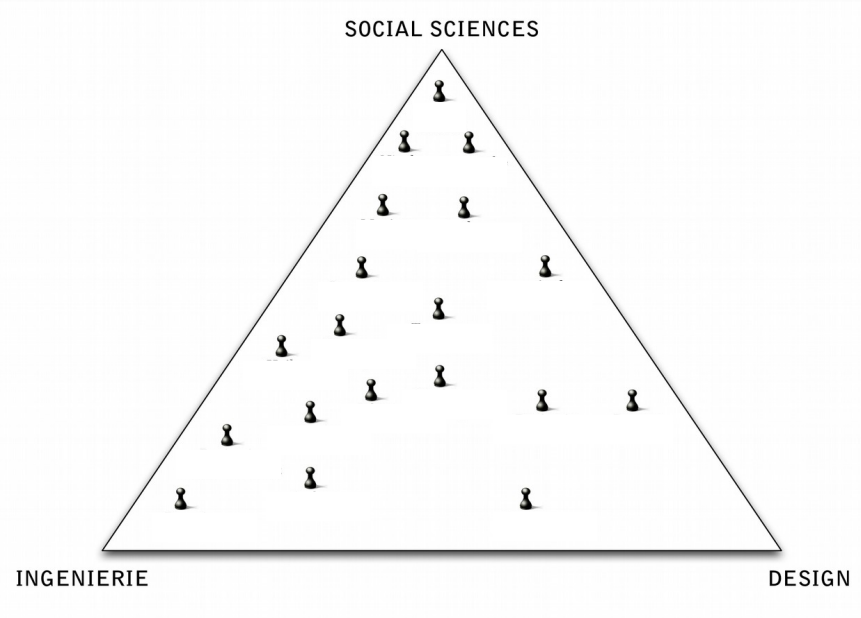
The Indian Express, 2011

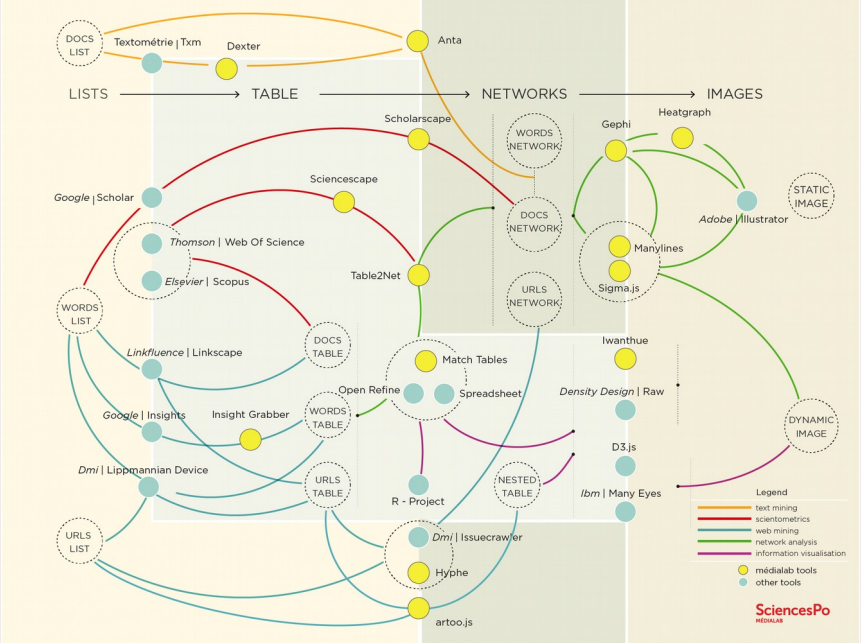
1. From the web to platforms:
   a (not so) brief history of hyperlinks
2. médialab: enabling digital field work through design & engineering
3. Hyphe: curate hypertexts into web corpora
4. Many angles of hypertext studies

médialab @ Sciences Po

https://medialab.sciencespo.fr

- Pluridisciplinary Research Lab created by Bruno Latour in May 2009, led by Dominique Cardon since 2017
- Social Sciences, Engineering & Design
- Articulate qualitative & quantitative methods through a digital approach
- Work with digital traces
- Deploy an ecosystem of tools
  http://tools.medialab.sciences-po.fr
- METAT: a monthly Open Support Workshop
  https://www.sciencespo.fr/recherche/fr/content/metat-latelier-de-methodes





The Web as a research field

DIME Web: one of three instruments of EQUIPEX DIME SHS

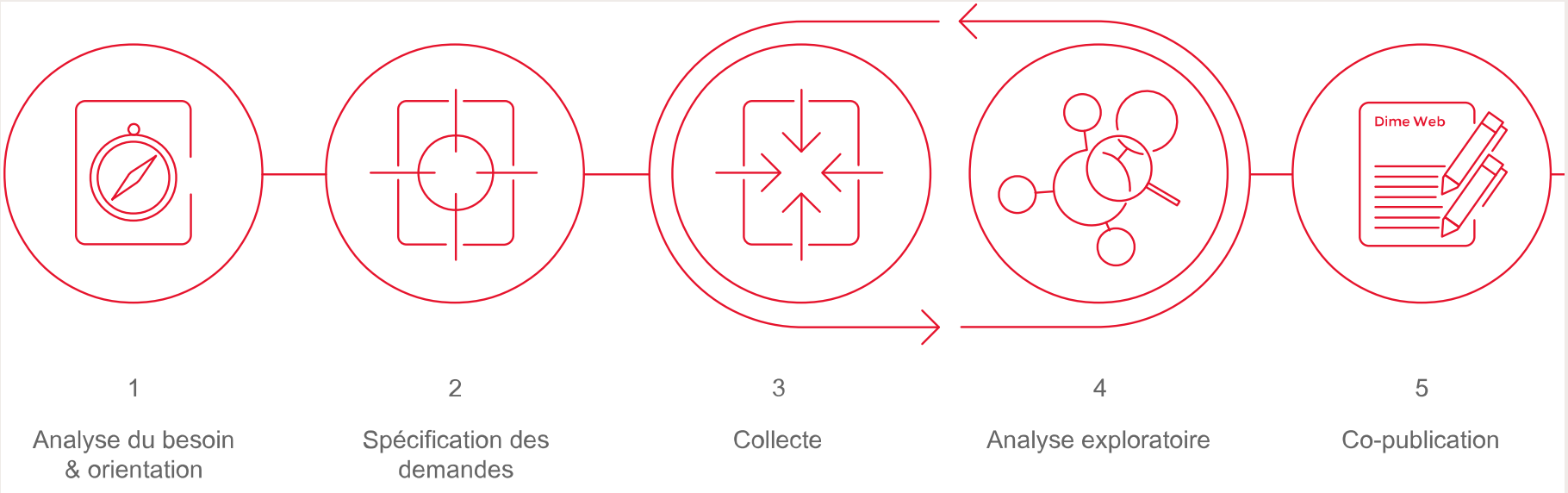
Methodological research and support

Network analysis, Controversies studies

Collect, enrich, clean, visualize & analyze digital traces:

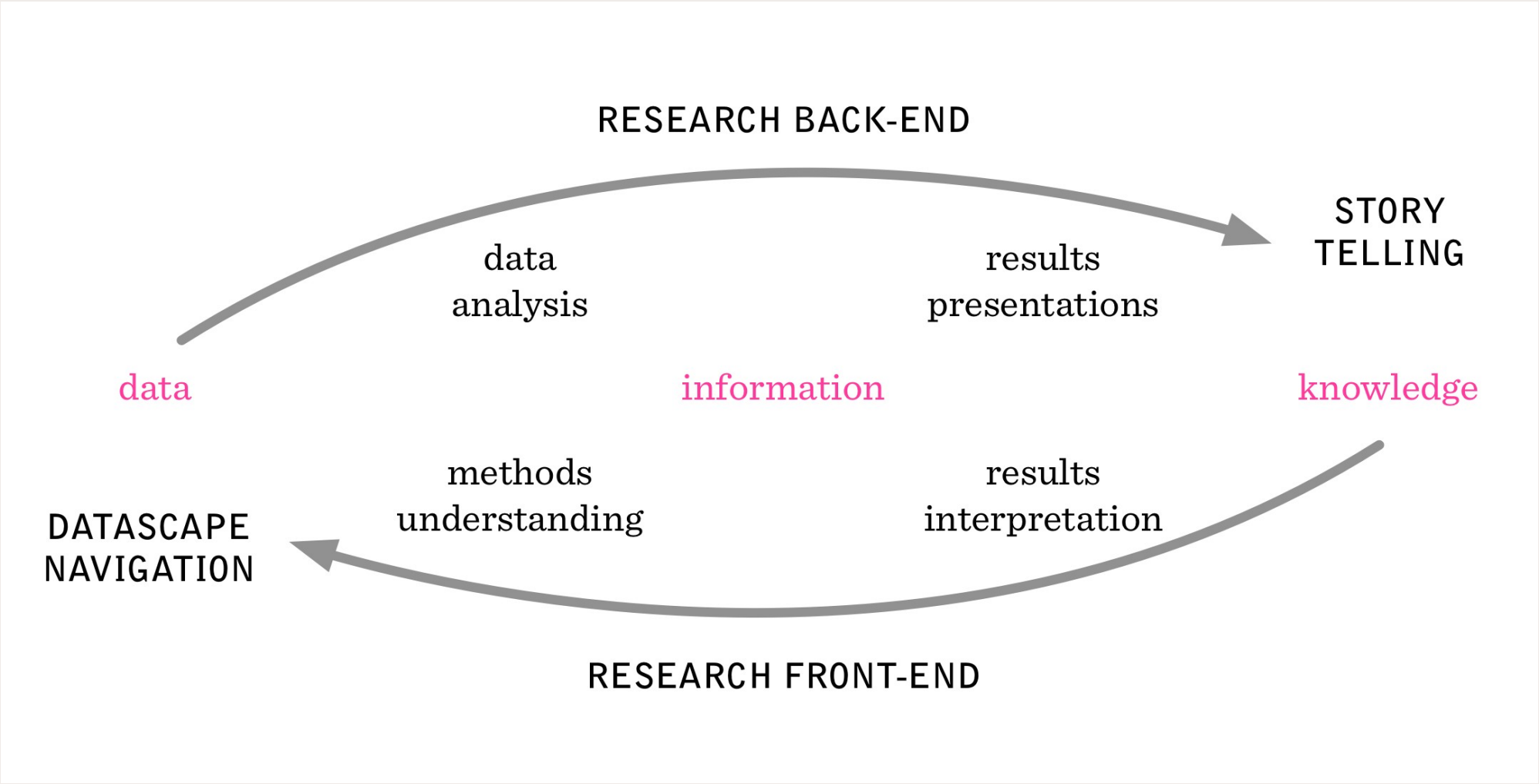
→ **Generic** collection & exploration tools

→ **Specific** extraction scripts



A Quali/Quanti approach: exploratory iterations

Digital ≠ Magic



→ always avoid full automation

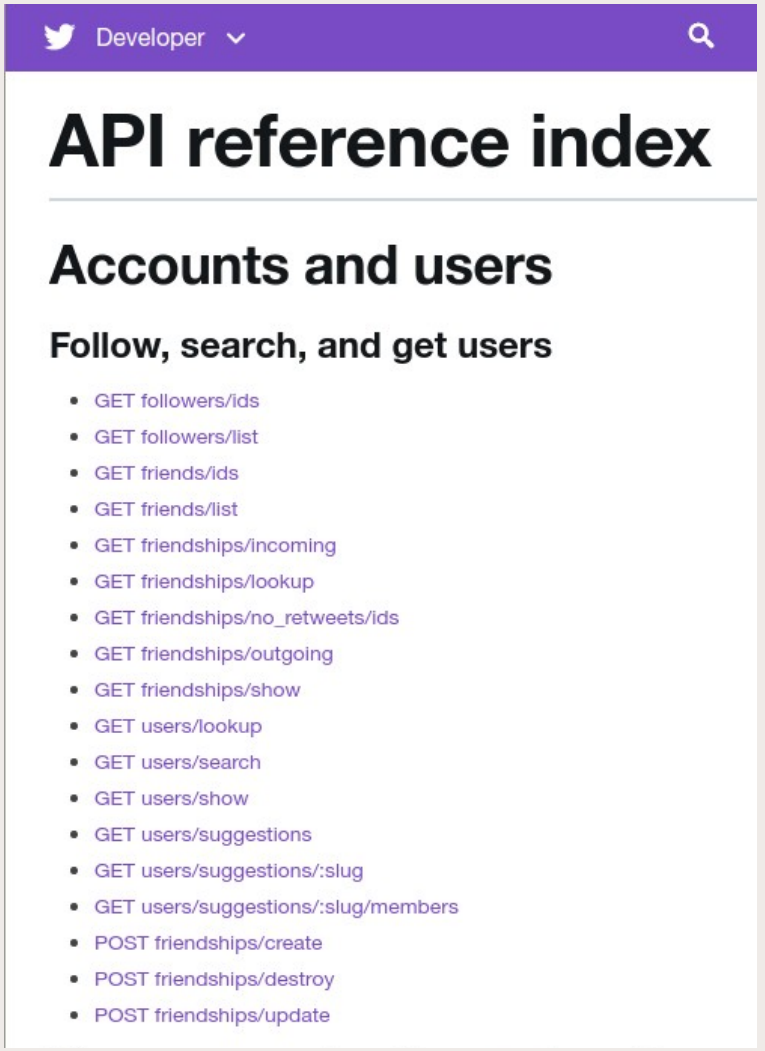
Research Driven Development

- Aim at large **Adoption:**
  - → **build** tools fit for Social Scientists
  - → **design** user centric interfaces
  - → **publish** tools directly usable online

- Ensure maximum **Reusability:**
  - → seize **opportunities** to implement new functionalities
  - → Free Libre **Open Source** Software
    (downloadable, installable, editable, reviewable, improvable)

Playground: working with platforms' APIs

- « Application Programming Interfaces »
- Structured data
- Massive and/or complete
- User behavioral data
- Problems:
  - volume
  - rate limits
  - black boxes
  - platforms' perspectives



Gazouilloire: systematic collection of tweets data

https://github.com/medialab/gazouilloire

Live and continuous collection via Search & Stream:

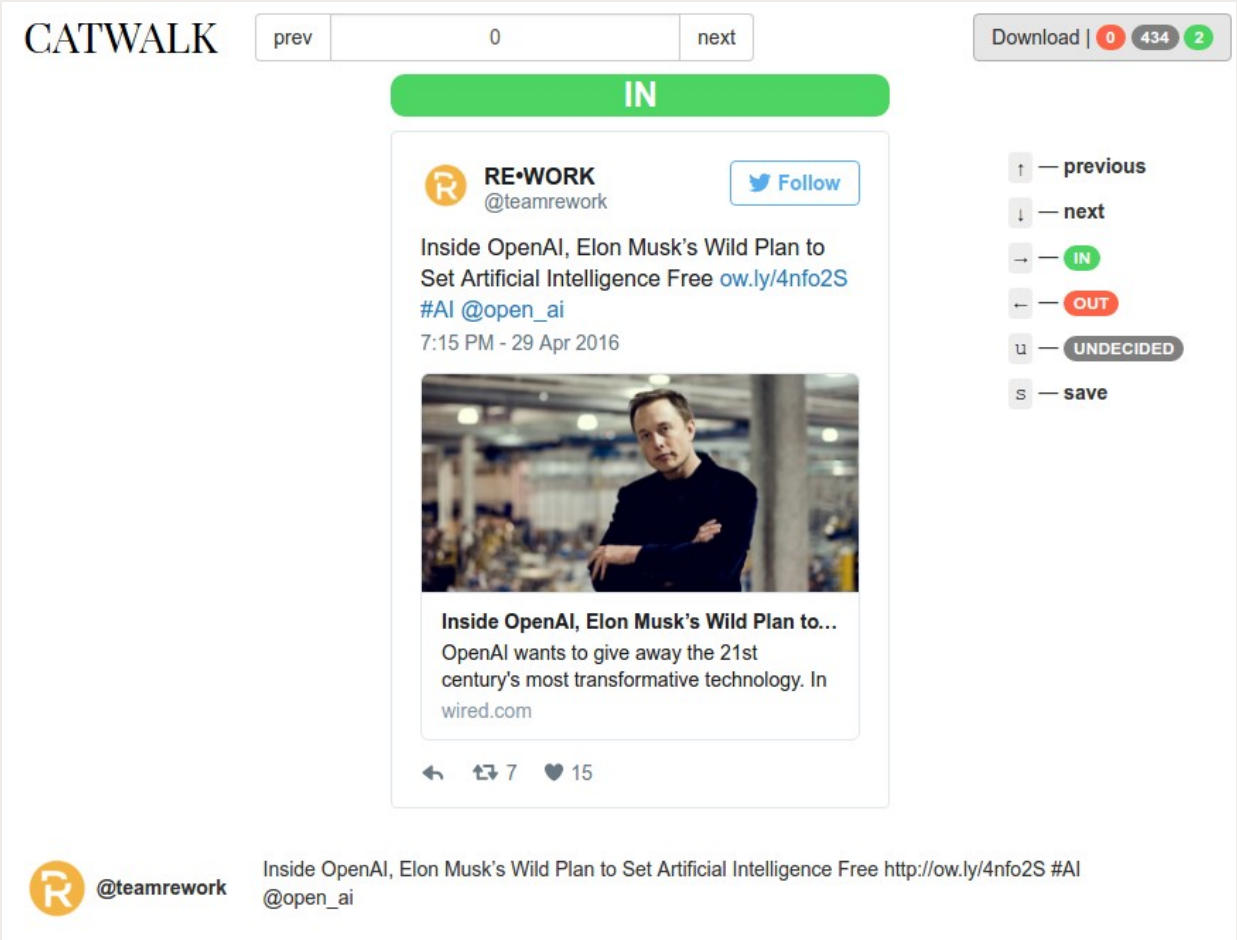
- tweets by keywords, urls, users, places, langages (up to 10 days back)
- threads and attached medias
- user profiles

```
{
    "twitter": {
        "user": "Gazou_medialab2",
        "key": "[redacted]",
        "secret": "[redacted]",
        "oauth_token": "[redacted]",
        "oauth_secret": "[redacted]"
    },
    "mongo": {
        "host": "localhost",
        "port": 27017,
        "db": "tweets-naturpradi"
    },
    "keywords": [
        "écologique Paris",
        "végétation Paris",
        "verger Paris",
        "grenelle environnement Paris",
        "locavore Paris"
    ],
    "time_limited_keywords": {
    },
    "geolocalisation": null,
    "geolocalisation_type": "admin",
    "resolve_redirected_links": true,
    "grab_conversations": true,
    "download_medias": true,
    "medias_directory": "/store/tweets/naturpradi/media/",
    "timezone": "Europe/Paris",
    "debug": true
}
```

```
[2016-11-22 15:23:34.056196] DEBUG: Starting search queries with 328 remaining calls for the next 655 seconds
[2016-11-22 15:23:34.259849] DEBUG: [search] +1 tweets (agriculture%20Paris OR agricultures%20Paris OR agroforesterie%20Paris)
[2016-11-22 15:23:35.807085] DEBUG: Saved 1 tweets in MongoDB
[2016-11-22 15:23:37.358533] DEBUG: [search] +1 tweets (espaces%20verts%20Paris OR ferme%20Paris OR fermes%20Paris)
[2016-11-22 15:23:37.810930] DEBUG: Saved 1 tweets in MongoDB
[2016-11-22 15:23:45.049743] DEBUG: [stream] +1 tweet
[2016-11-22 15:23:45.821150] DEBUG: Saved 1 tweets in MongoDB
[2016-11-22 15:24:51.598045] DEBUG: [stream] +1 tweet
[2016-11-22 15:24:51.893009] DEBUG: Saved 1 tweets in MongoDB
[2016-11-22 15:24:52.401661] DEBUG: [medias] +1 files
[2016-11-22 15:24:58.073013] DEBUG: Starting search queries with 286 remaining calls for the next 571 seconds
[2016-11-22 15:25:00.383614] DEBUG: [stream] +1 tweet
[2016-11-22 15:25:01.905385] DEBUG: Saved 1 tweets in MongoDB
[2016-11-22 15:26:18.060840] DEBUG: Starting search queries with 246 remaining calls for the next 491 seconds
[2016-11-22 15:26:19.922864] DEBUG: [search] +1 tweets (compost%20Paris OR composts%20Paris OR compostage%20Paris)
[2016-11-22 15:26:19.989779] DEBUG: Saved 1 tweets in MongoDB
```

CatWalk: qualitatively pick tweets from a corpus

https://medialab.github.io/catwalk/

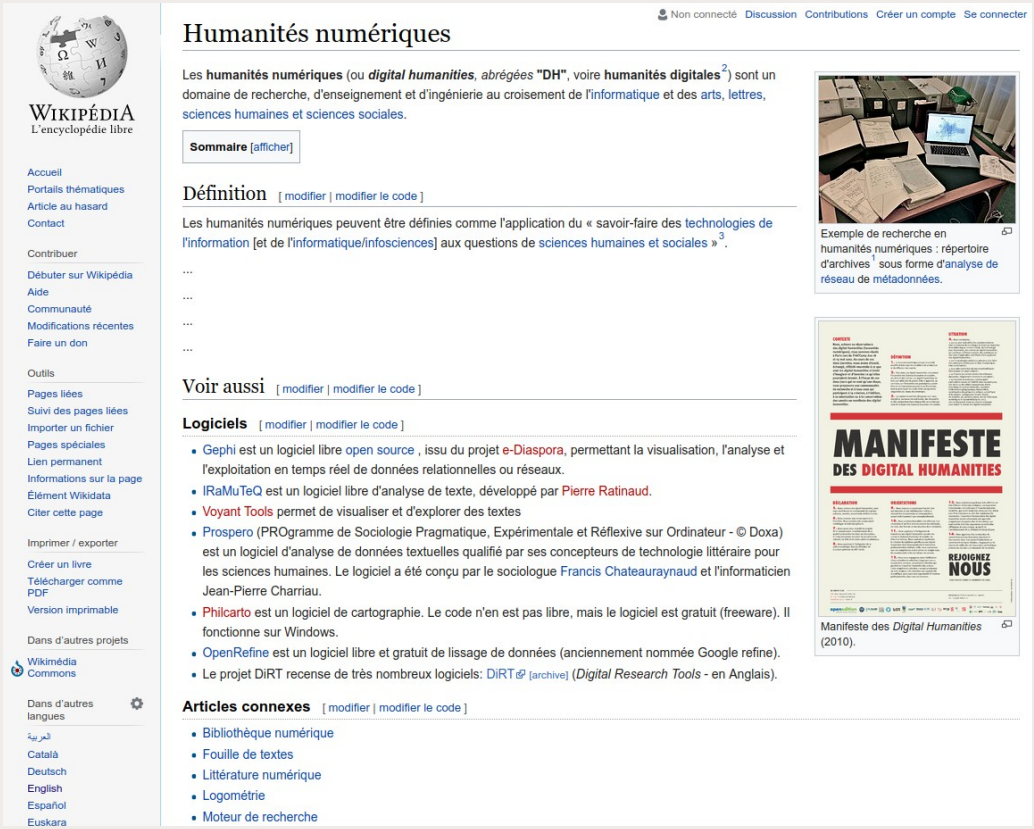


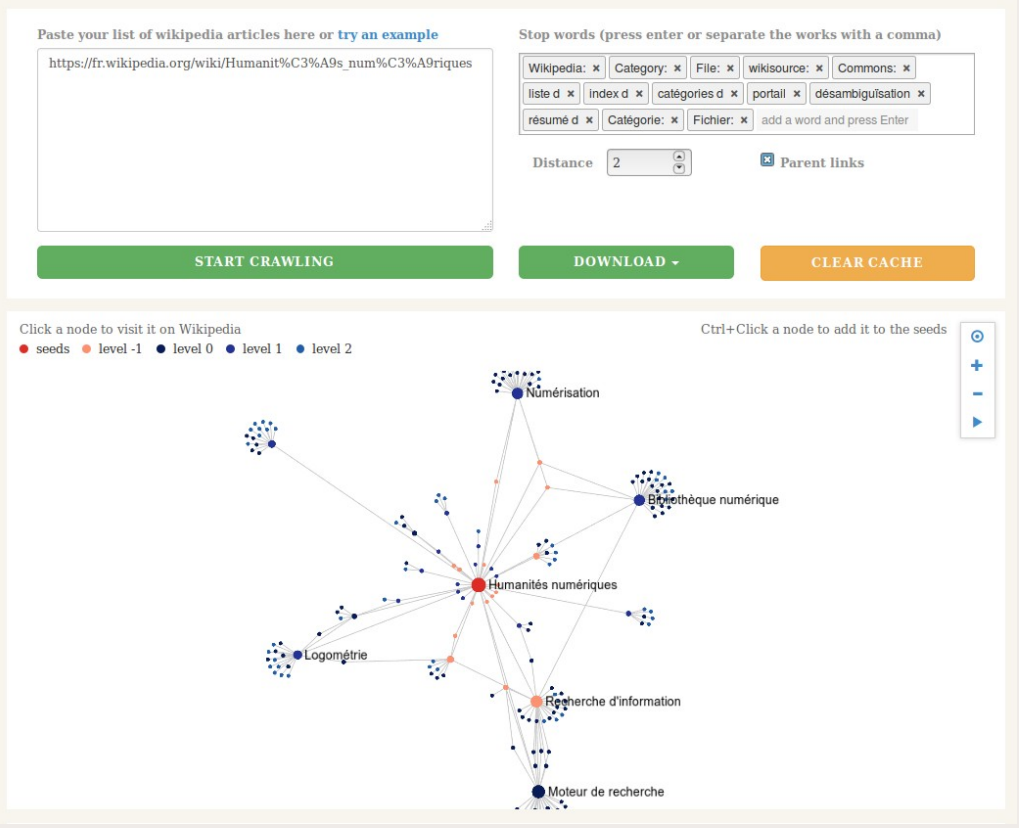
« *Tinder-like* » review of all tweets listed in a CSV to decide to take them IN or OUT

SeeAlsology: semantic exploration from Wikipedia

http://tools.medialab.sciences-po.fr/seealsology/

Build & explore a semantic network of fields linked within « See Also » sections of Wikipedia pages





APIs are rare: data Scraping to the rescue

- Extract information from a specific single source
- Build structured data from all kinds of web contents

  url, date, author, image, description, place, values…

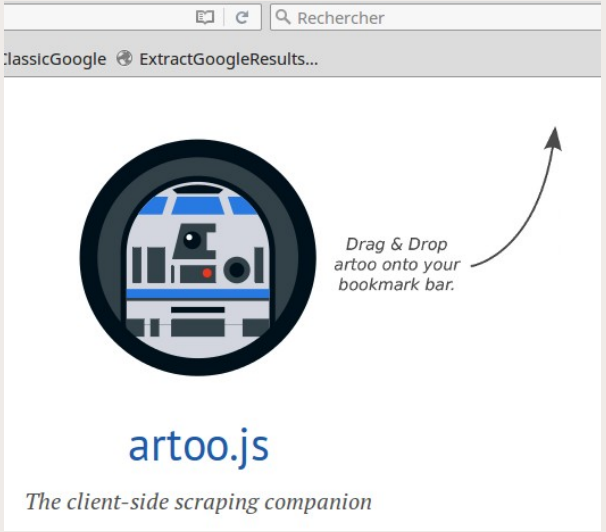
  → Enable statistics & quantitative analysis
- Whatever is public & formatted can be scraped
- Problem: The web constantly changes

  → Maintenance issues

Artoo.js: the browser scraping companion

https://medialab.github.io/artoo/

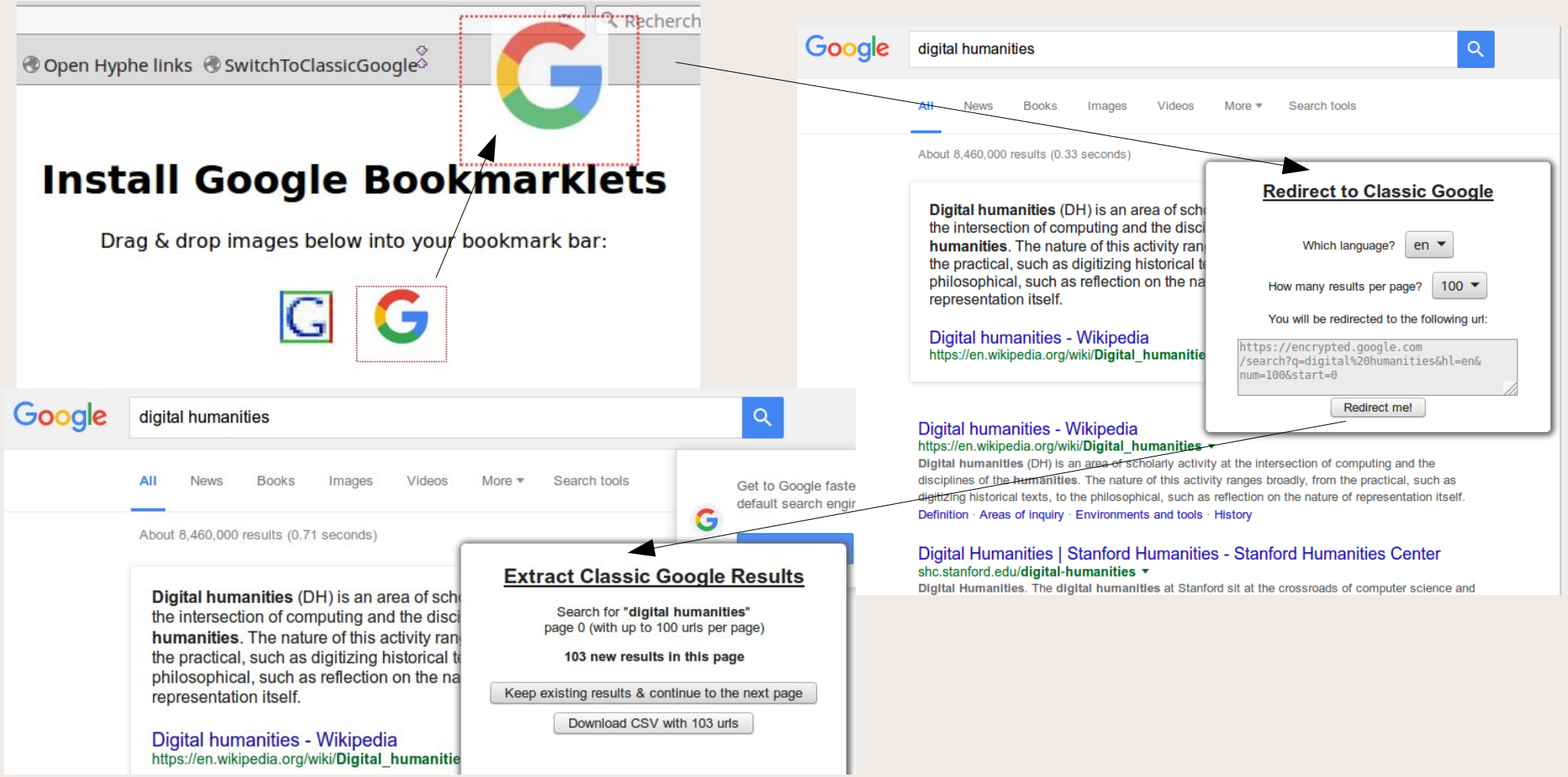
- Common scraping traps:
  - user authentification
  - cookies
  - dynamic pages with JS…
- Avoid simulating browsers: hack them!
- Embed jQuery helpers to scrap directly from within the browser's console (F12)




```
> var data = artoo.scrapeTable( ".wikitable", {headers: 'th'} );
 undefined
> data.length;
 49
> data[0];
 Object {Country: " World", CO2 emissions (kt) in 2014[2]: "35,669,000",
" % CO2 Emissions by Country": "100%", Emission per capita (t) in 2014[3]:
"5.0"}
> artoo.saveCsv(data, "CO2-world-emissions.csv");
 undefined
```

Google bookmarklets: search results as CSV

https://medialab.github.io/google-bookmarklets/



Simple buttons installable into the browser's bookmarks to let one easily scrap search results as tabular data

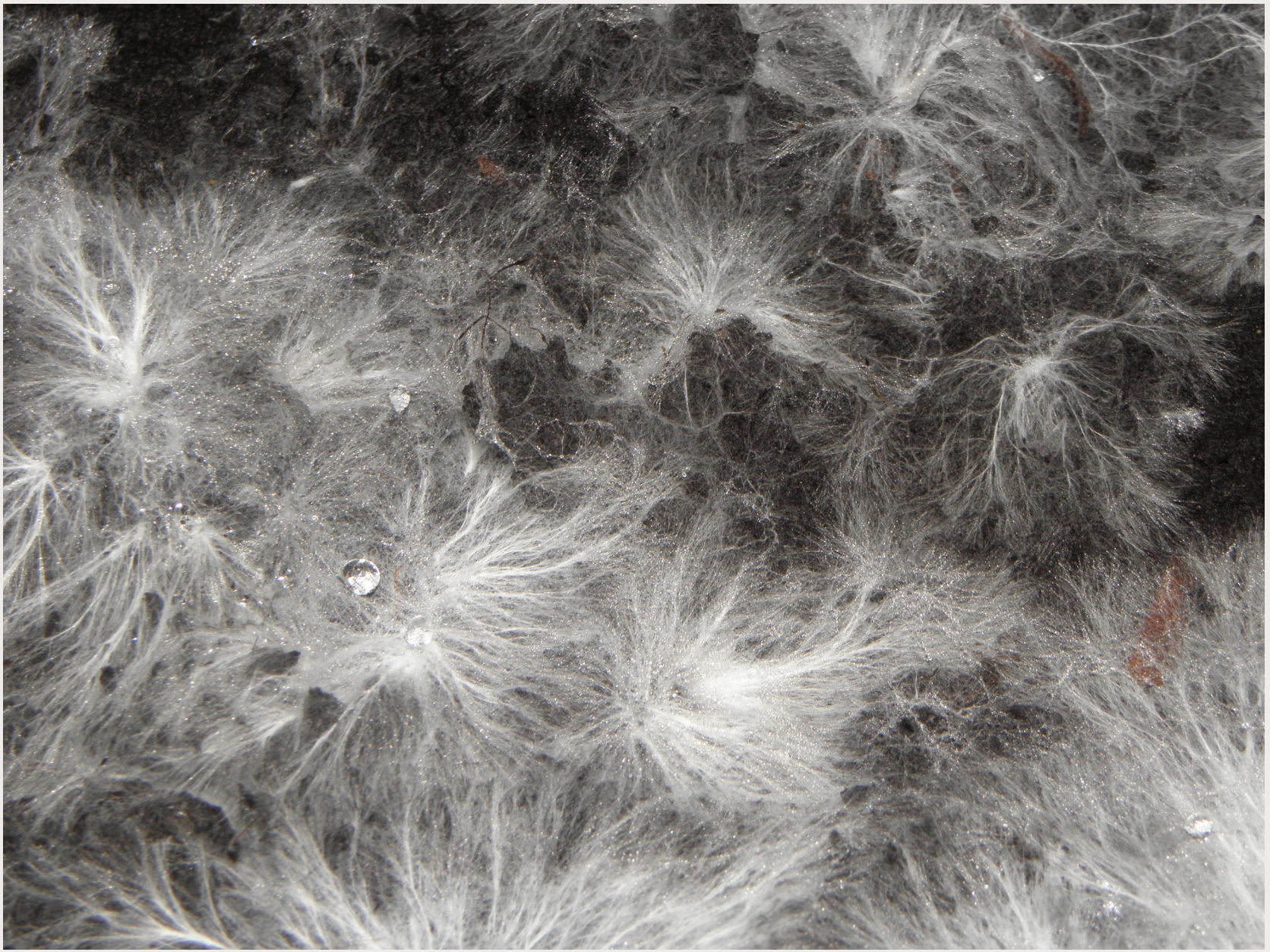
Crawling: harvest data from a variety of sources

- Extract coherent data from heterogeneous sources
- Build loosely structured data from common grounds found within web documents (pages)
  - text content → natural langage processing
  - hyperlinks → network analysis
- Problem: the web is a « dirty mess »

  redirections, erroneous links, dead links, disappeared websites, bad encoding…

1. From the web to platforms:
   a (not so) brief history of hyperlinks
2. médialab: enabling digital field work
   through design & engineering
3. Hyphe: curate hypertexts into
   web corpora
4. Many angles of hypertext studies

The interlinked mycellium: a network of hyphæ



Hyphe: a research directed crawler

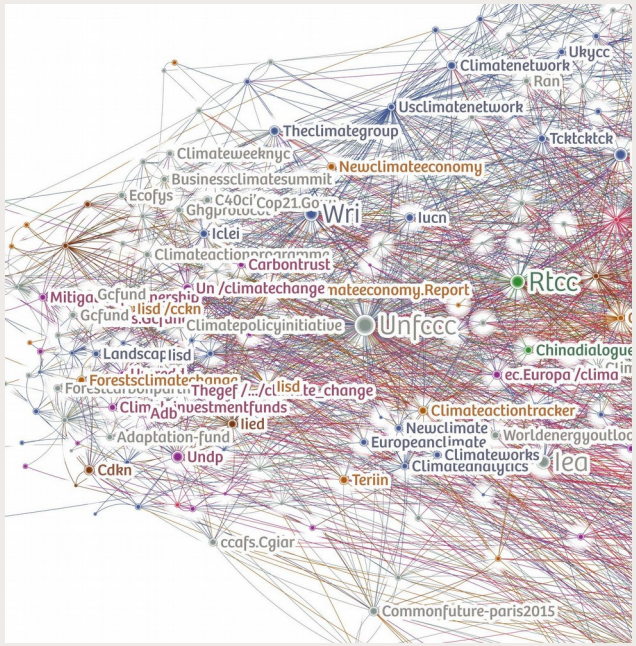
http://hyphe.medialab.sciences-po.fr/demo/

Build your own web documents corpus
to study social phenomena online

→ gather « web actors »

→ explore hyperlinks between them

hyphe



http://medialab.github.io/double-dating-data/

COP 21

Vie privée

Extrême droite

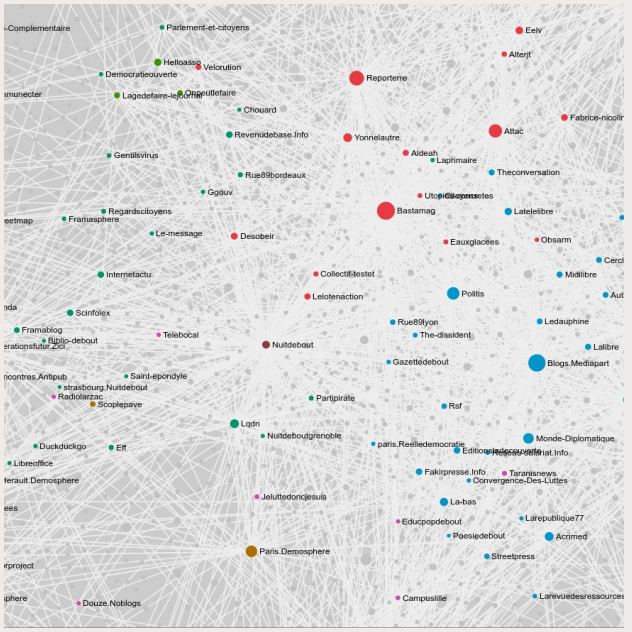
Tissu associatif

Produits laitiers

Cellules souches

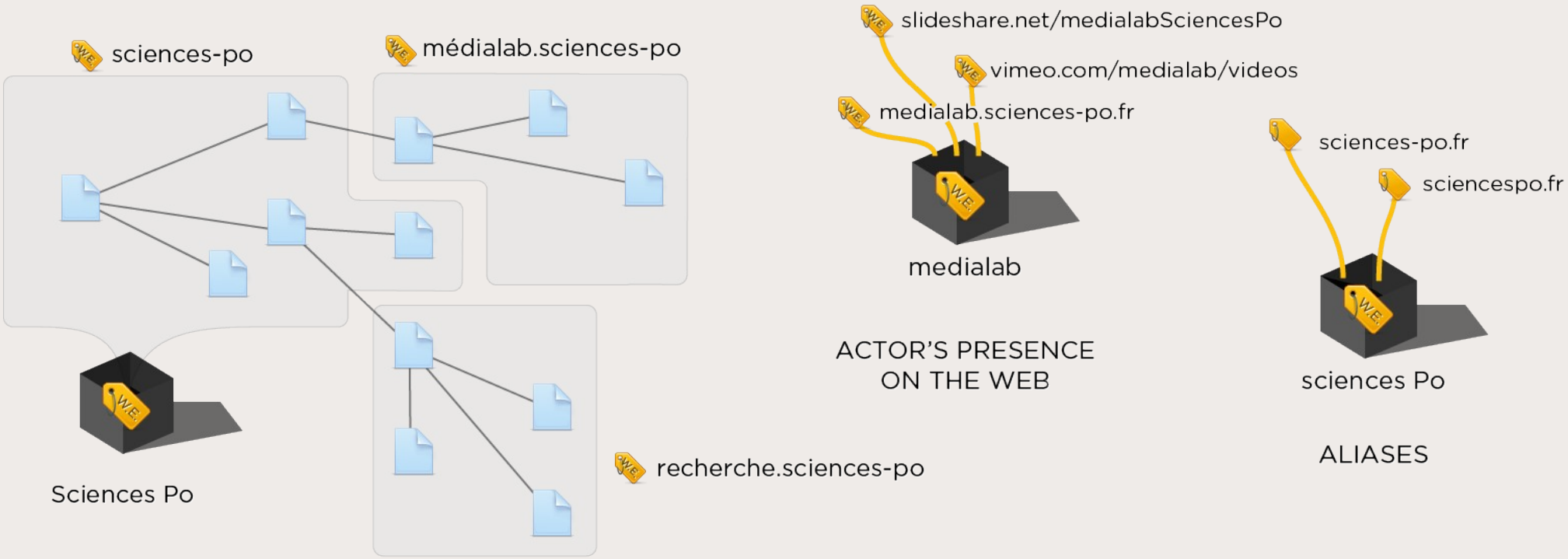
Administrations culturelles

…



http://utopies-concretes.org/

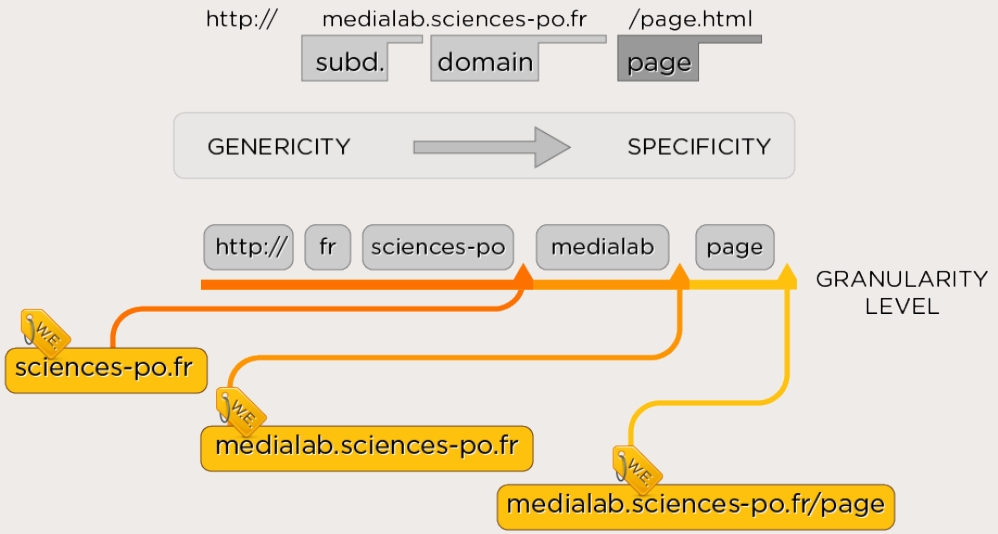
There is no such thing as a « website » !



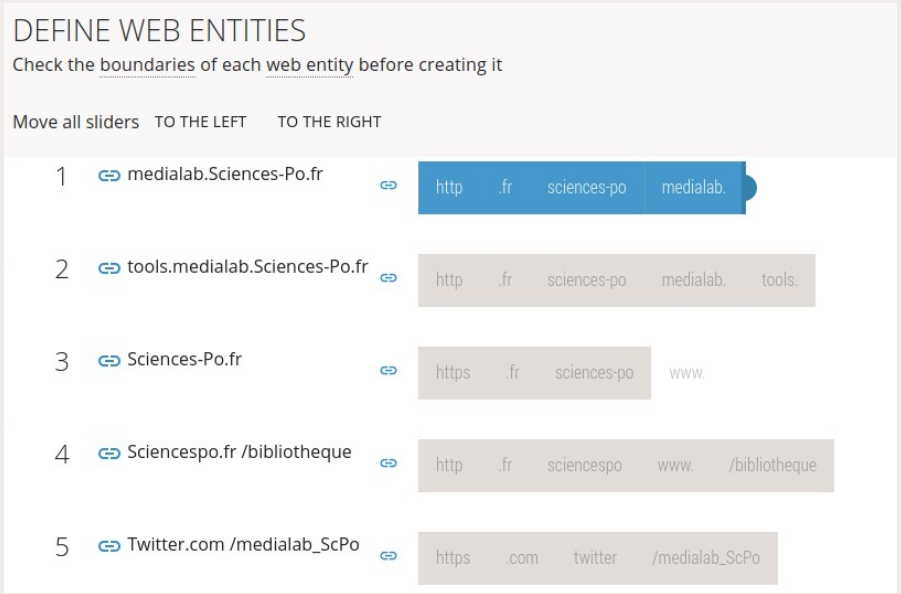
→ « **WebEntities** » : bundles of webpages aggregating coherent actors to answer a specific research question

= set of URL prefixes

Finely delimit the web territories of actors



Manually setup prefix patterns to adjust the cursor of « WebEntities »

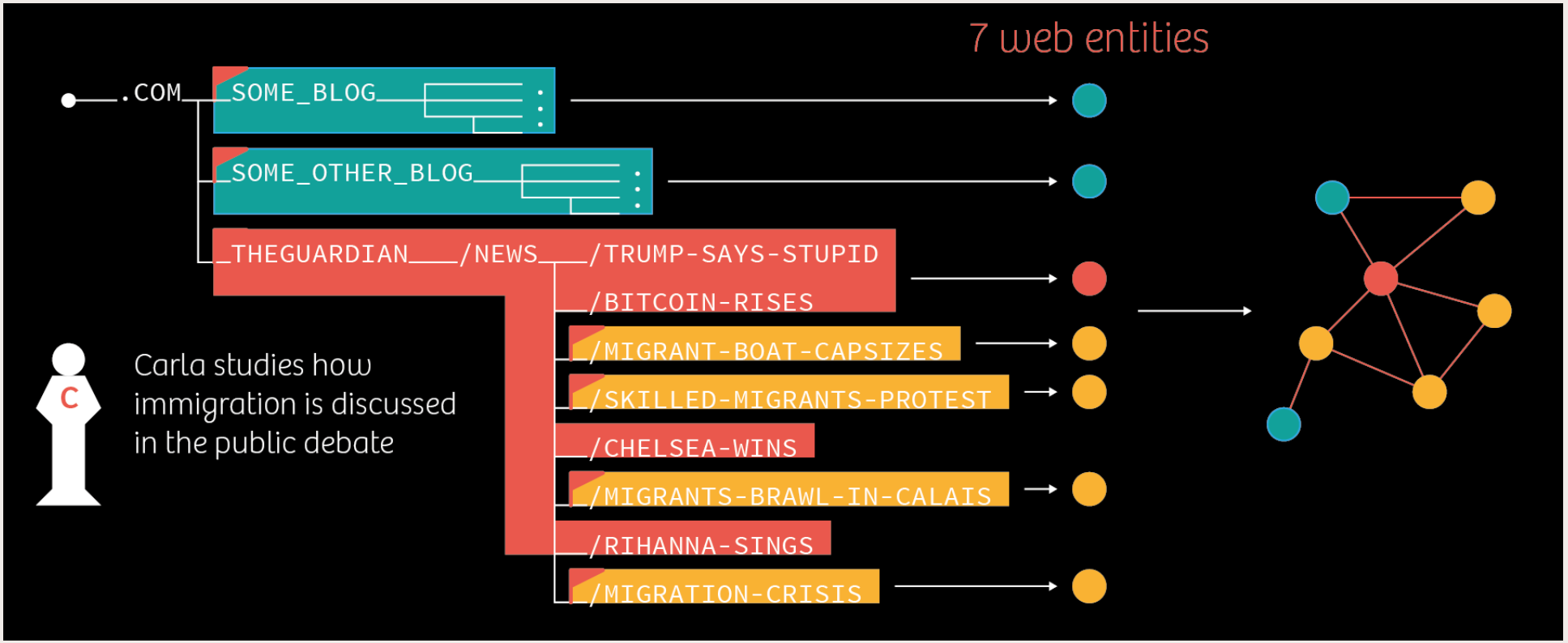


Which data structure to manage hypertexts?

https://medialab.github.io/hyphe-traph/fosdem2018/#/

- **Tree** of urls
- **Graph** of hyperlinks
- Dynamic branches aggregates

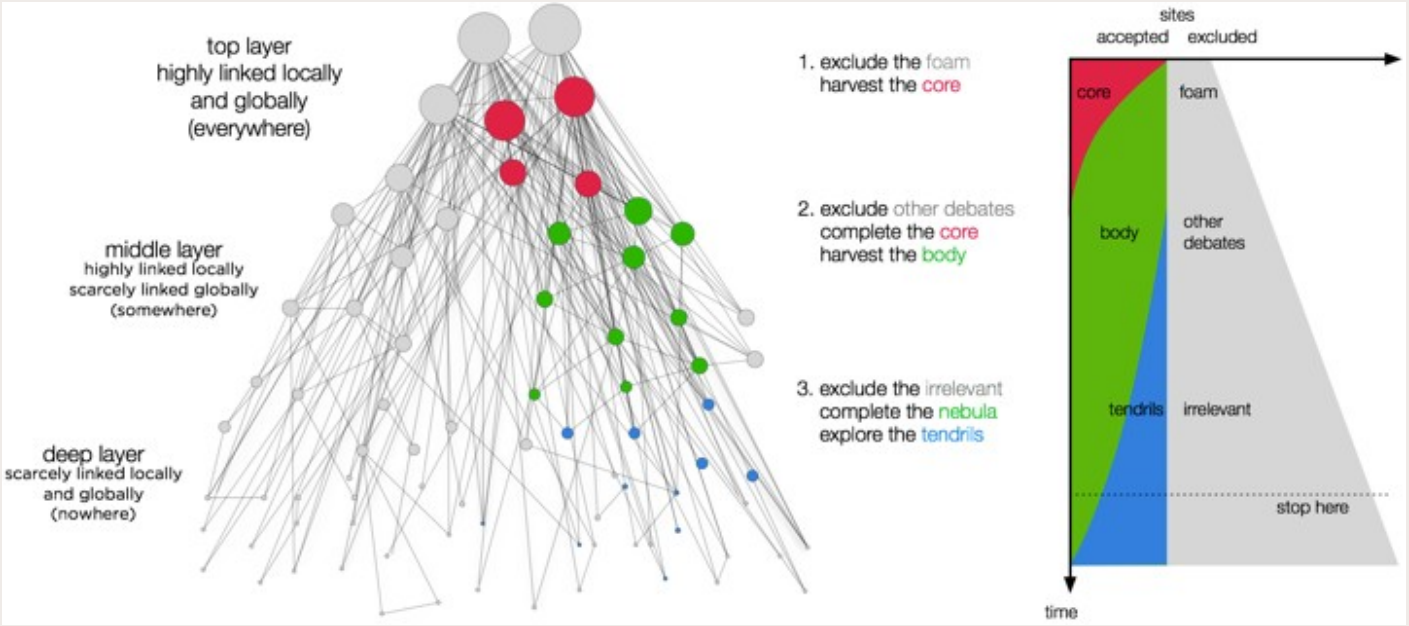
→ Hyphe's Traph



Hyphe's crawling strategy: leverage hyperlinks

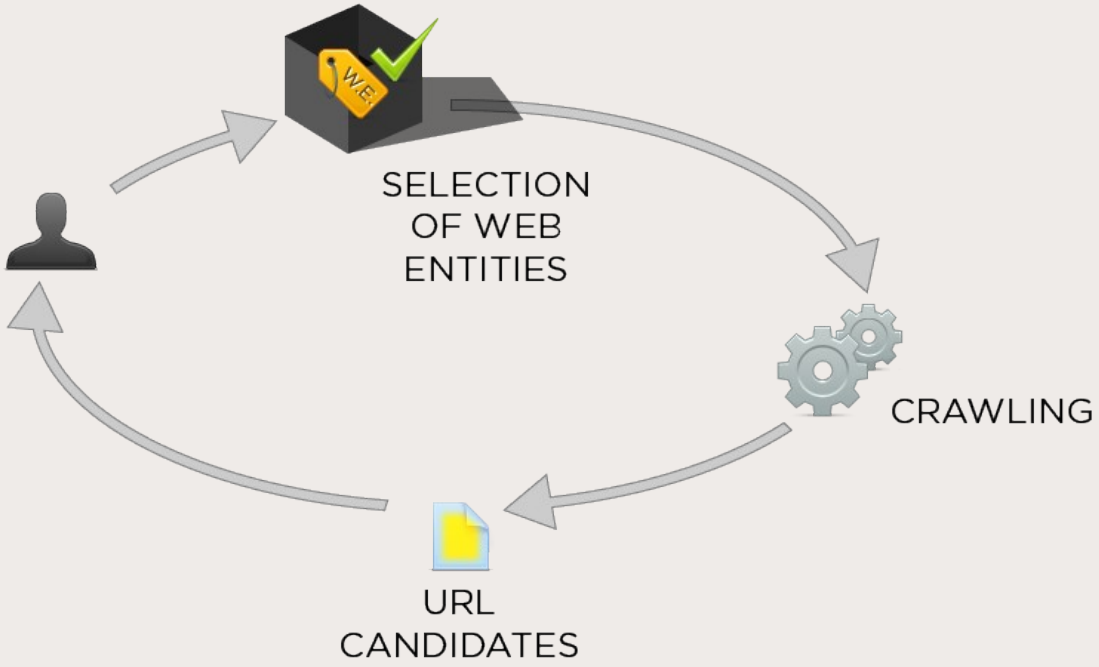
- Classical crawlers (DMI's IssueCrawler): Snowball
  → Top Layer attraction (Google, YouTube, Wikipedia…)
  → Topic drifts

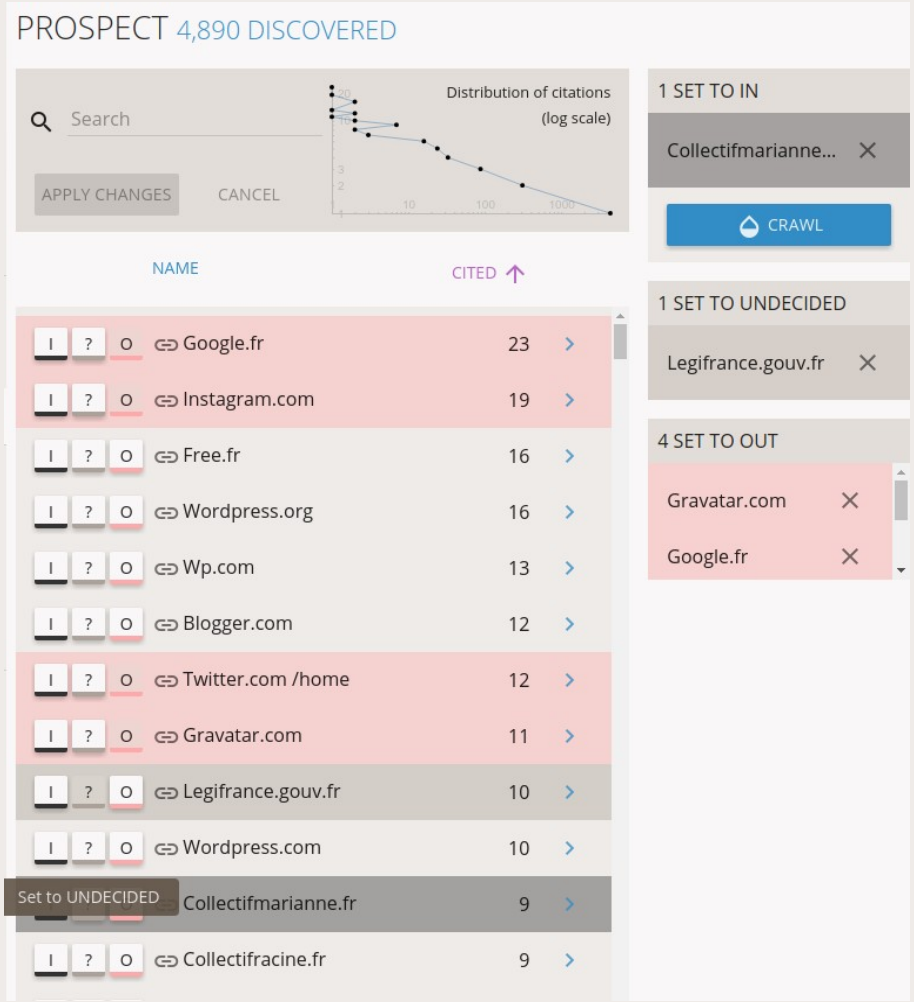
- Hyphe:
  - crawl exclusively pages within the chosen WebEntities
  - sort discovered entities by degree of citation
  - humanly select new entities to include and crawl



Web prospection loop: curating a corpus iteratively

- Step by step iterative expansion & curation of entities



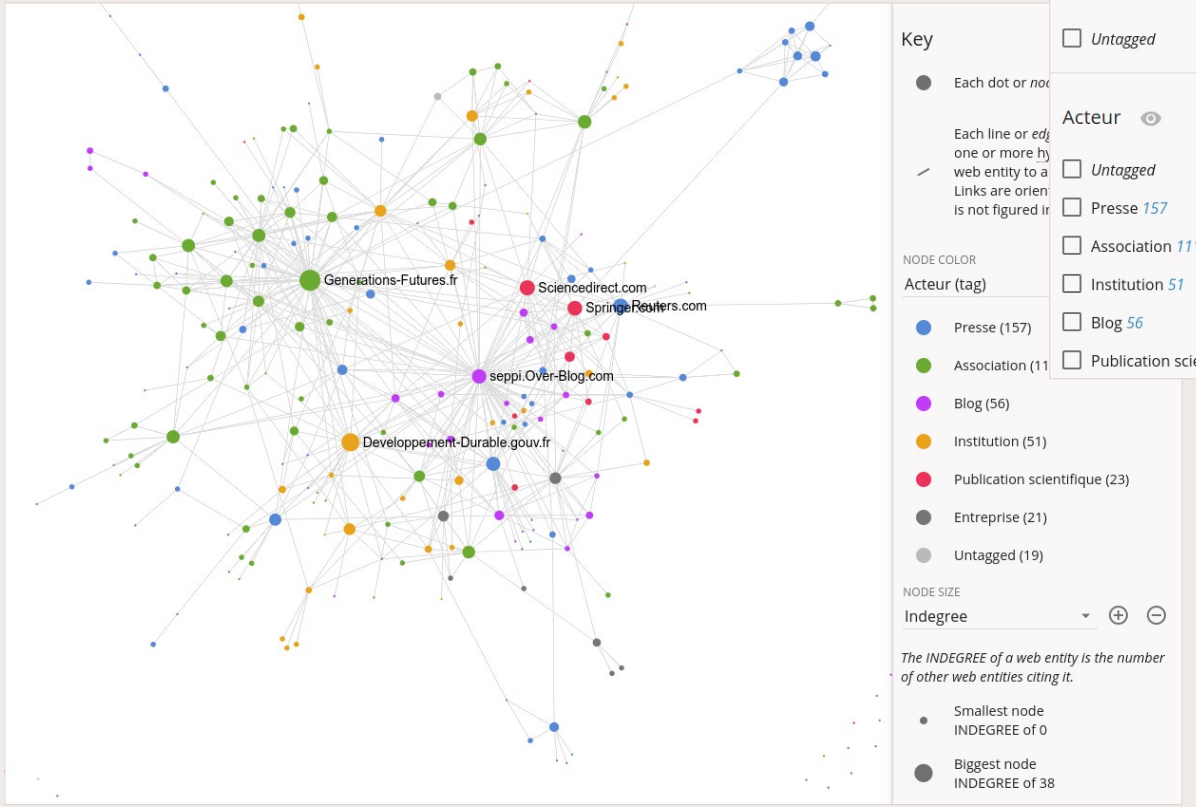


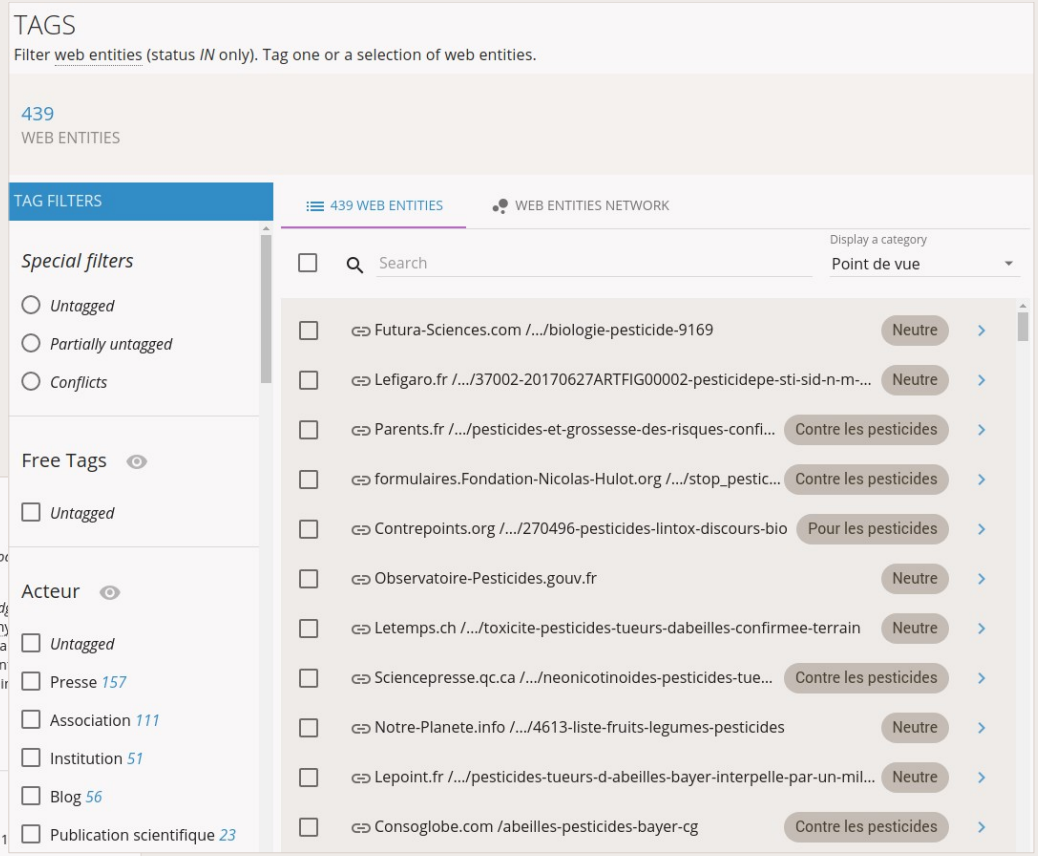
- Human/Time cost
- How to know when to stop?

  → hyperlink citations threshold

Qualify the corpus entities (tagging)

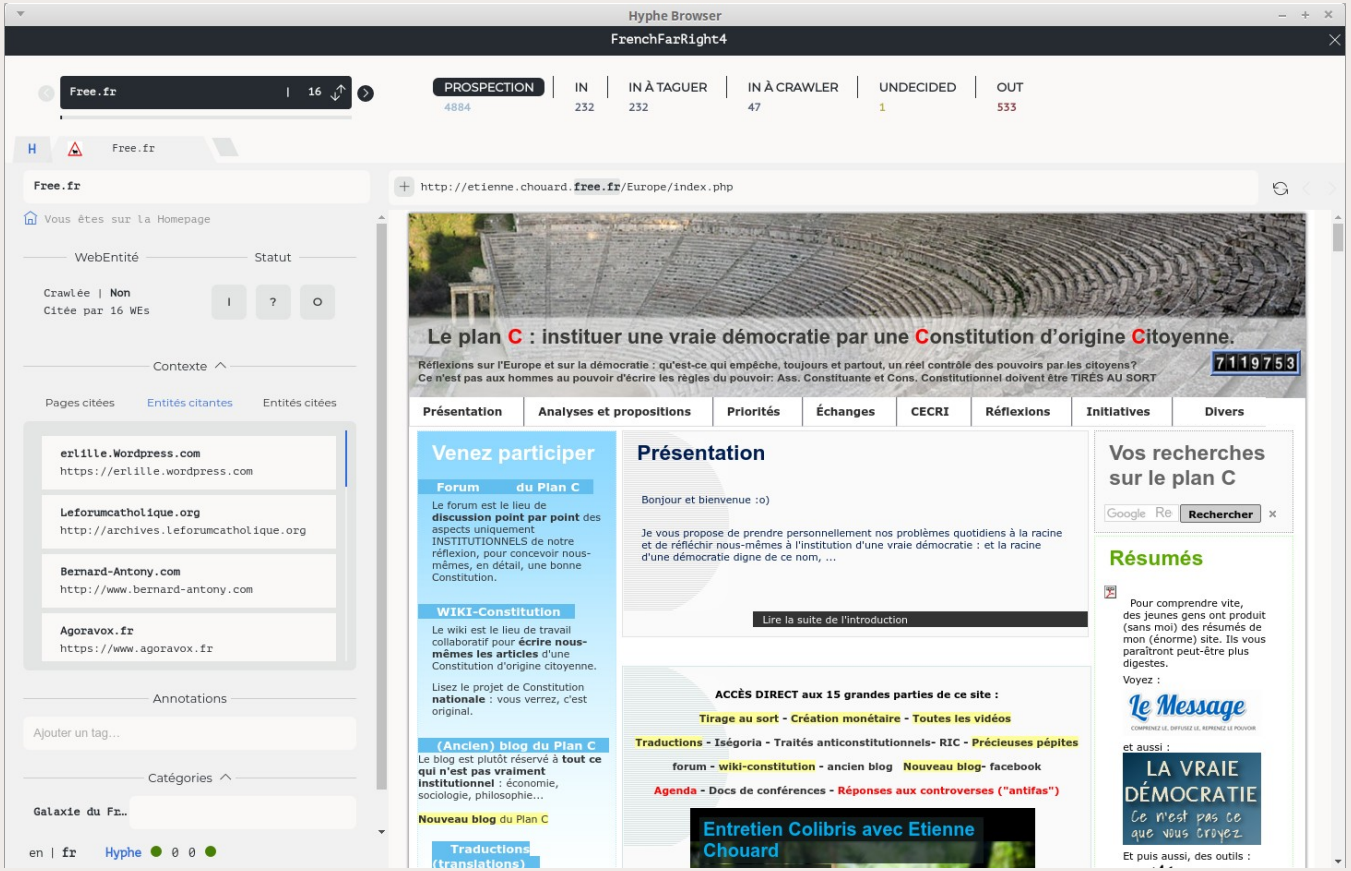
- Free notes
- Categories





HyBro: a web browser designed for corpus curation

https://github.com/medialab/hyphe-browser/releases/



« NaviCrawler » heritage : build a Hyphe web corpus while browsing

→ « in-situ » prospection & tagging (digital field work)

→ teach the web to students (IDEFI FORCCAST)

1. From the web to platforms:
   a (not so) brief history of hyperlinks
2. médialab: enabling digital field work
   through design & engineering
3. Hyphe: curate hypertexts into
   web corpora
4. Many angles of hypertext studies

Many ways to use Hyphe

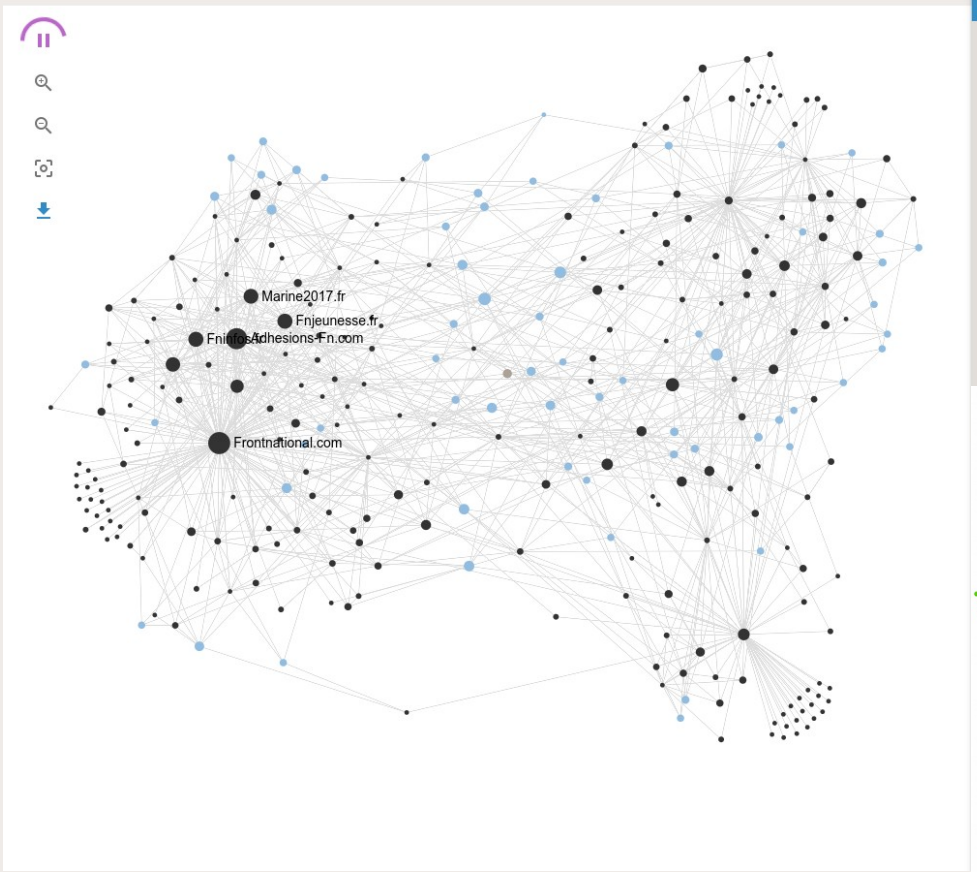
- Complete methodology includes:

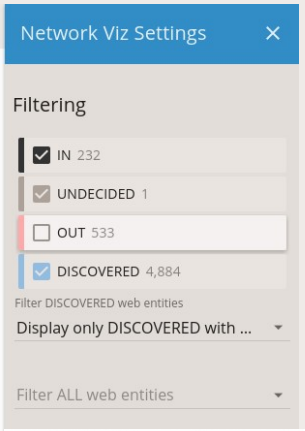
  sourcing, automatized collection, iterative corpus building
  qualitative categorization, exploratory analysis,
  network visualization, quantitative statistical analysis

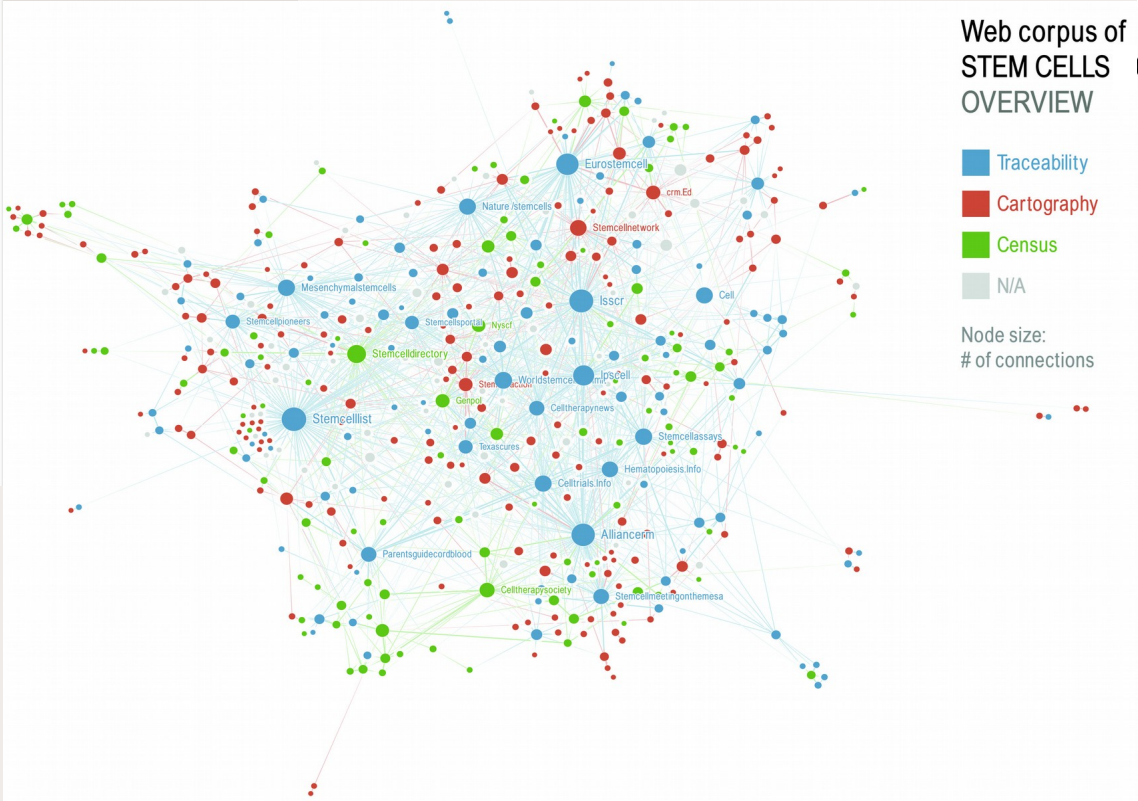
- Diverse audiences:
  - Research: help social scientists work on digital fields
  - Pedagogy: teach students what the web is beyong Google & Facebook

- Possible small & large scale analyses:
  - a website's internal structure
  - a theme's ensemble of actors and their ties
  - a controversy's alliances & oppositions
  - etc.

From above: clusters, opposition & affinity

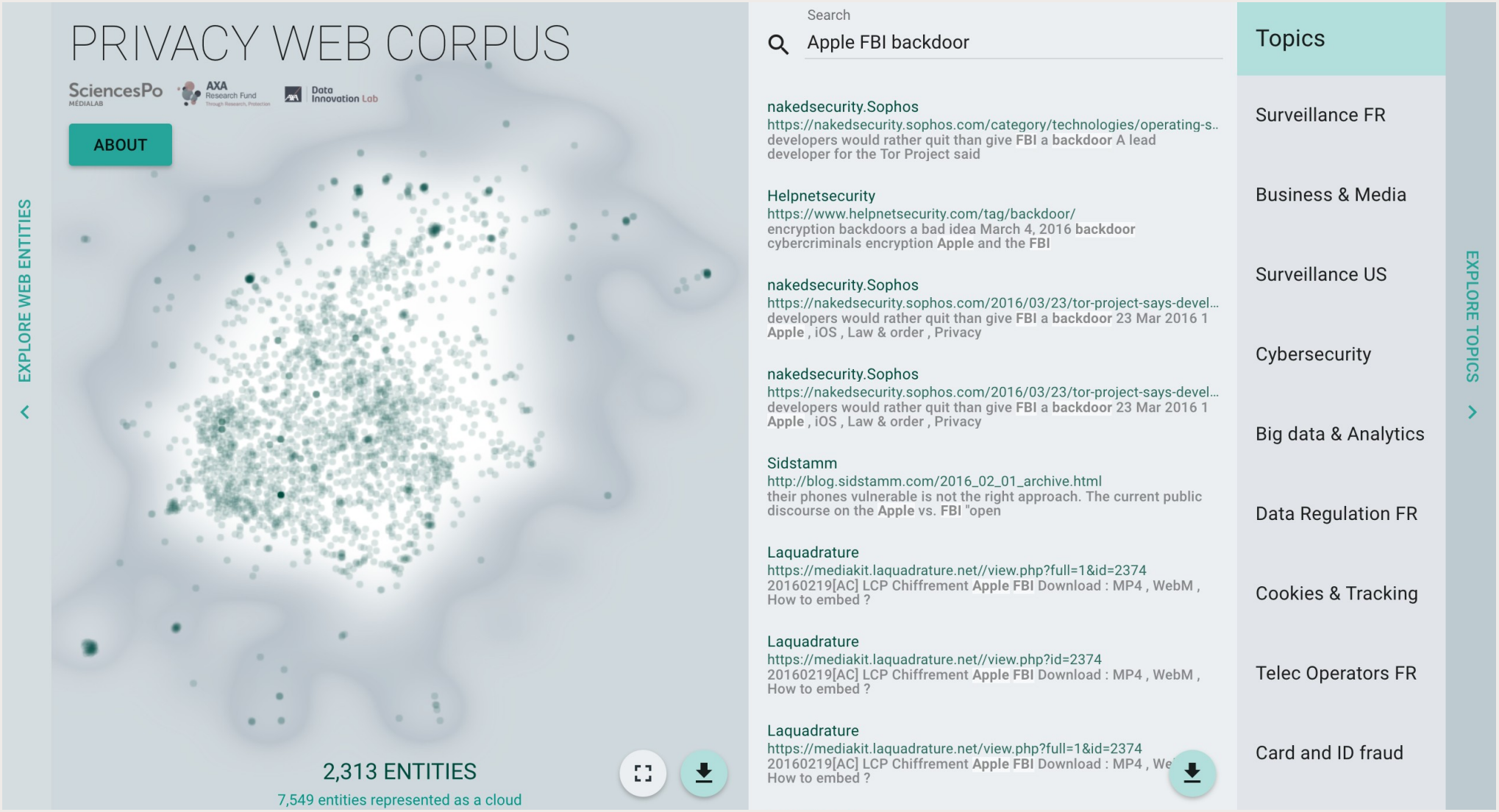






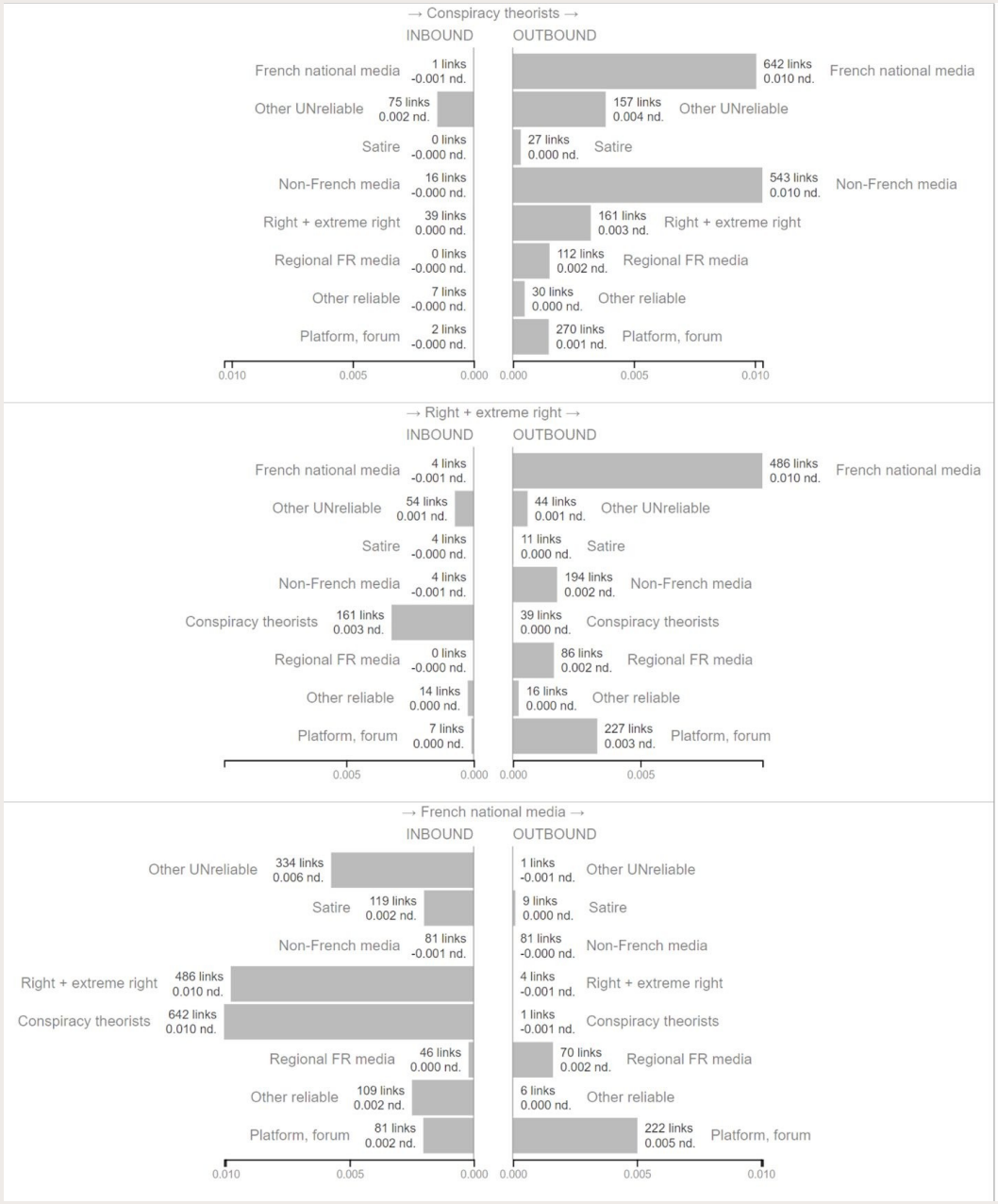
Social Representations of Stem Cells, Virginie Tournay, CEVIPOF, 2016

From within: explore webpages contents



http://tools.medialab.sciences-po.fr/privacy/

From the sides: a hierarchy of directed hyperlinks



Conspiracy theorists
- Not reliable
- Not cited
- Cite many reliable sources

other

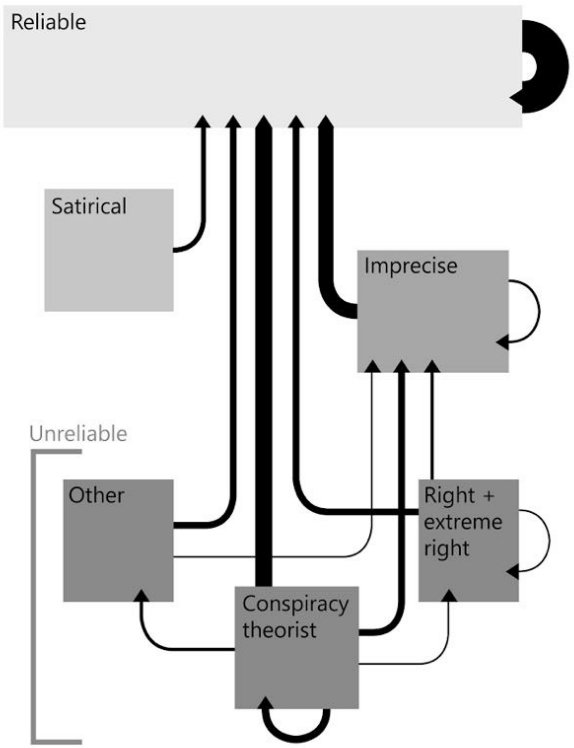
Right and extreme right
- Not reliable
- Cited only by conspirationists
- Cite many reliable sources

other

other

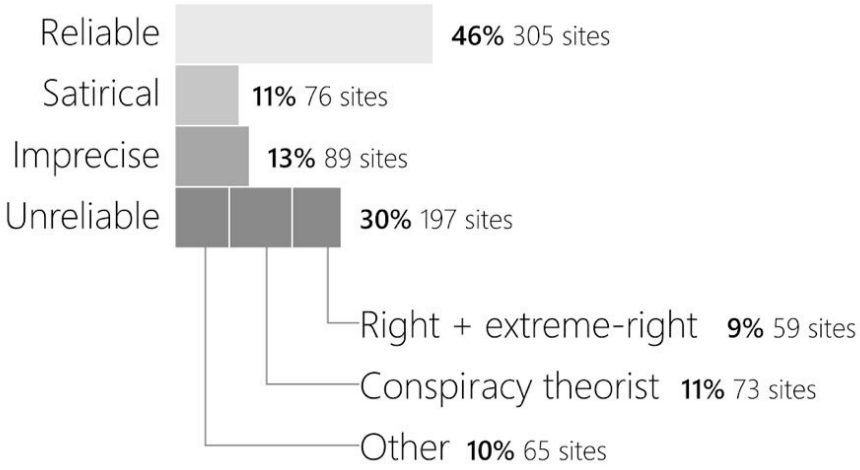
French national media
- Reliable *(*in this qualification)*
- Cited by everyone
- Cite only platforms and regional media

Hyperlinks directionality: a bottom-up hierarchy

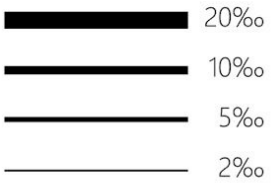


Most hyperlinks stem from the unreliable and aim at the reliable resources

Each bloc's surface is proportional to the count of websites. The color code si the same as the "Décodex".



The thickness si proportional to the normalized link density*

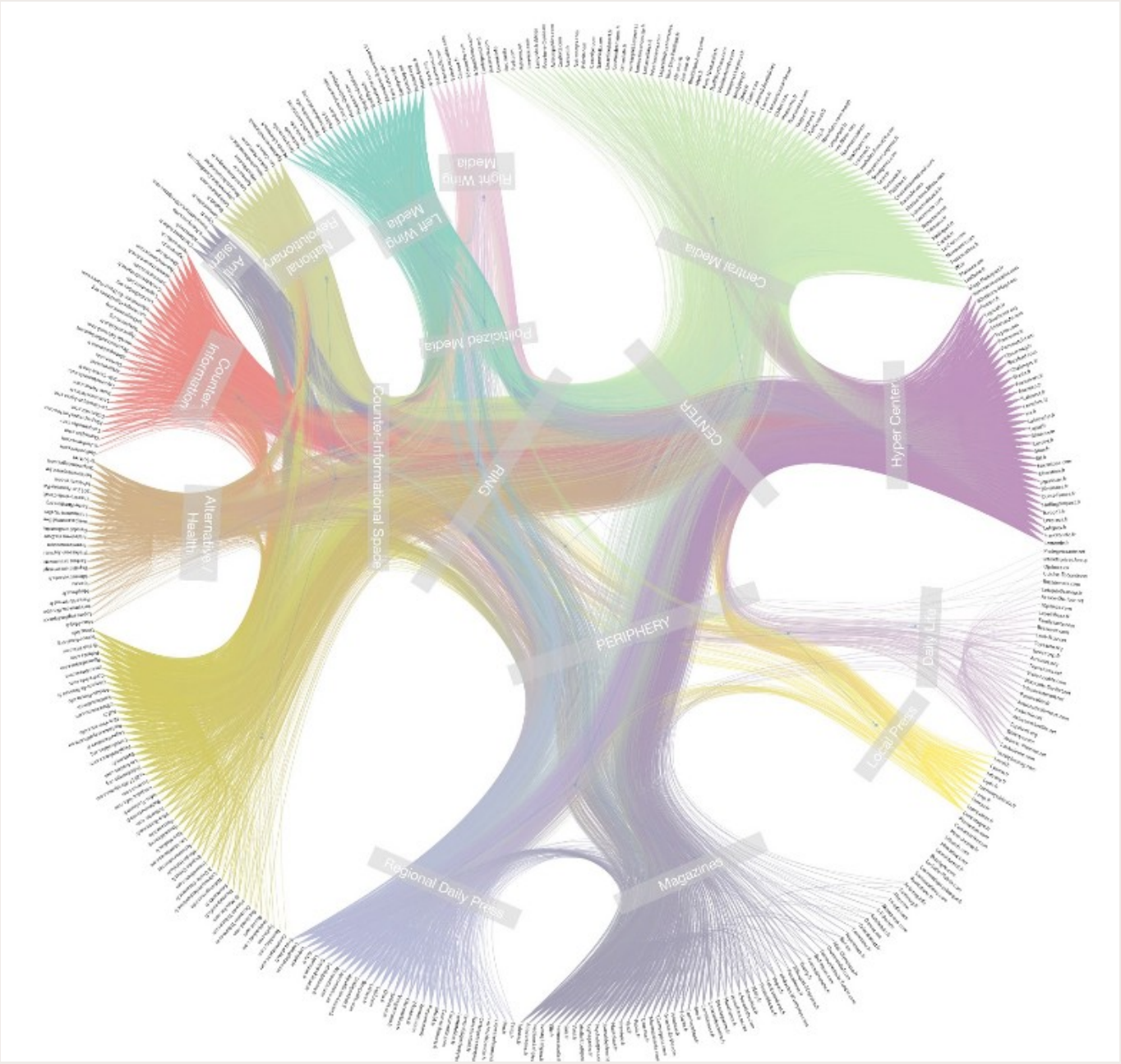


*Accounts for the different group sizes.

Densities below 2‰ have been omitted.

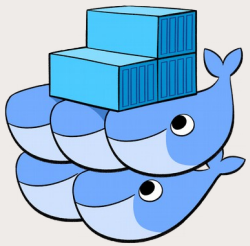
https://www.researchgate.net/publication/320225750_Visual_Network_Exploration_for_Data_Journalists

Explore polarization dynamics



Roadmap: what next?

- Import / export corpora or lists of webentities & crawls:
  - → duplication, reproduction
  - → longitudinal time exploration
- Integrated text processing (NLP) for content analysis
- Use modern web technologies to handle JavaScript powered contents (Facebook, React applications, etc.)
- Quality control tools for crawls
- Archival & exploration tools to publish finalized web corpora
- Propose automatic setups of Hyphe as Software As A Service



Hyperlink is not dead: long live hyperlink studies!

- 10-year financial support of EQUIPEX

  → building cutting-edge research instruments demands long-term support
- Hyphe is Free & Open Source, try it online with the demo!

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

https://github.com/medialab/hyphe

Questions?

benjamin.ooghe@sciencespo.fr

@boogheta **@medialab_ScPo**

Bibliography

Reference publications:

- Jacomy M., Venturini T., Heymann S., Bastian M. (2014), **ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software**, PLos ONE, 9(6), 1-18

  https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679

- Jacomy M., Girard P., Ooghe-Tabanou B., Venturini T. (2016), **Hyphe, a Curation-Oriented Approach to Web Crawling for the Social Sciences**, ICWSM 2016, Cologne

  https://spire.sciencespo.fr/hdl:/2441/6obemb2hsj9pboj9bbvc7sftne

- Plique G., Jacomy M., Ooghe-Tabanou B., Girard P. (2018), **It's a Tree... It's a Graph... It's a Traph! Designing an on-file multi-level graph index for the Hyphe web crawler**, FOSDEM 2018, Bruxelles

  https://medialab.github.io/hyphe-traph/fosdem2018/#/

- Ooghe-Tabanou B., Girard P., Jacomy M., Plique G. (2018), **Hyperlink is not dead!,** ACM Proceedings of the 2nd International Conference on Web Studies (WS.2 2018) Paris.

  http://hyphe.medialab.sciences-po.fr/docs/20181004-ACM-WebStudies-HyperlinkIsNotDead.pdf