médialab Tools : de petits outils simples et complexes au service de la DataScience

Benjamin Ooghe

MATE-SHS

cnrs
dépasser les frontières

médialab Tools

De petits outils simples et complexes au service de la DataScience

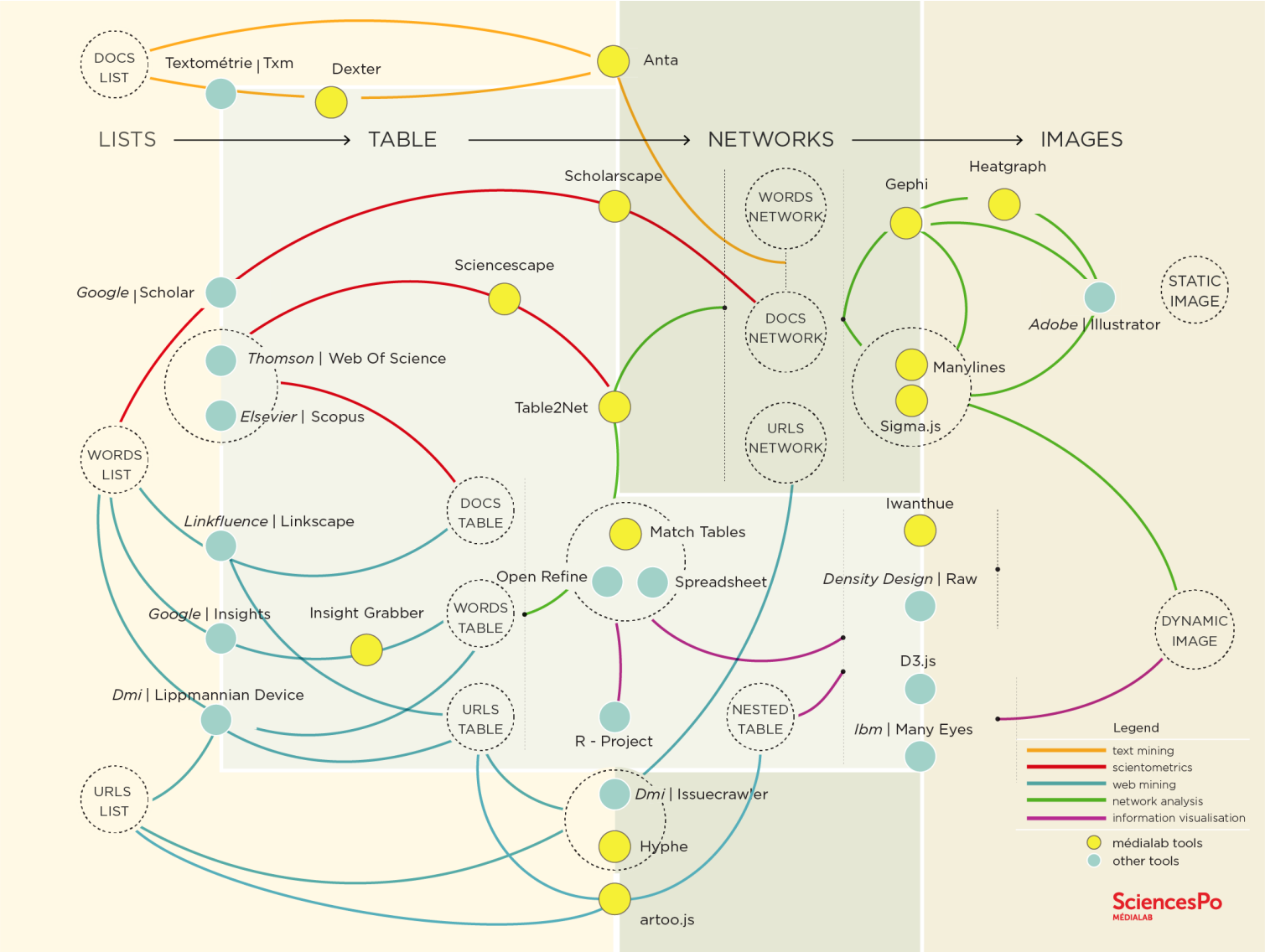
Benjamin Ooghe-Tabanou, Sciences Po, médialab, Paris, France

SciencesPo
MÉDIALAB

Collecter et produire des données pour la recherche en SHS
Fréjus, 15-18 novembre 2016

Un écosystème de petits outils pour la science

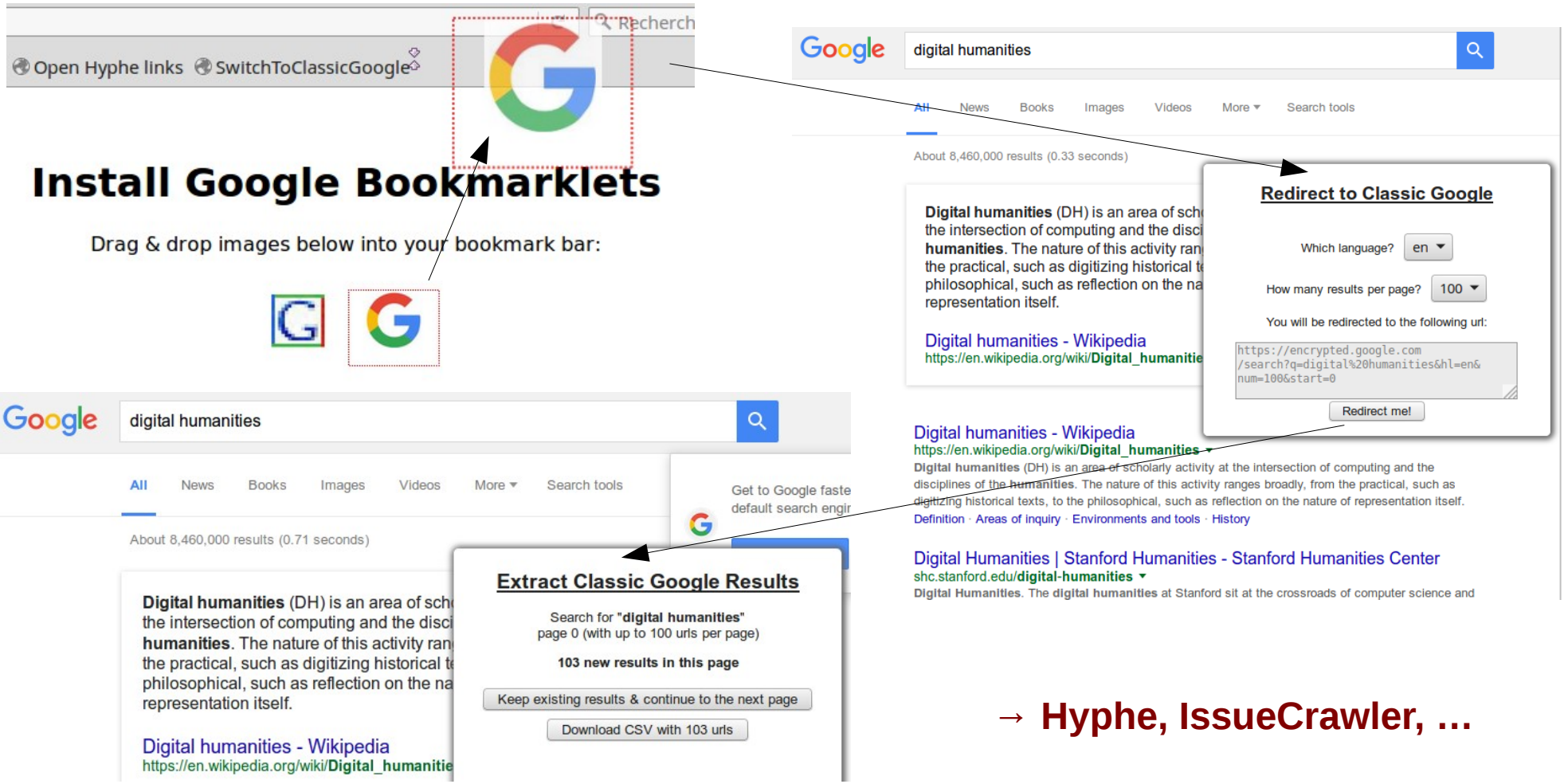
http://tools.medialab.sciences-po.fr



Google bookmarklets : les résultats en CSV

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

- Des boutons dans vos favoris pour récupérer simplement au format tableur les résultats d'une recherche Google



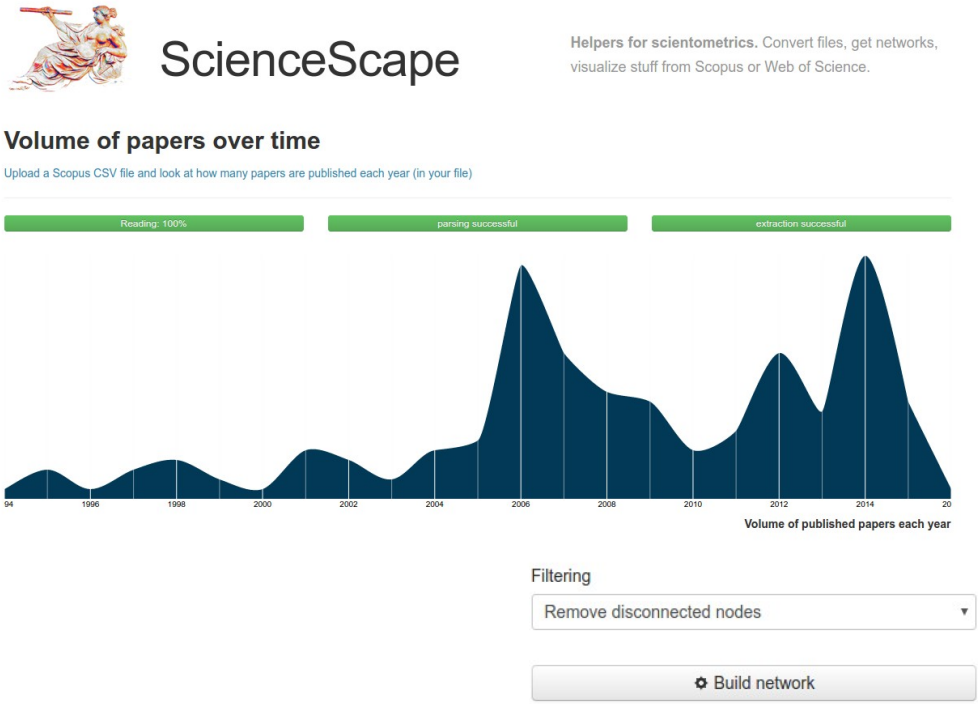
→ **Hyphe, IssueCrawler, …**

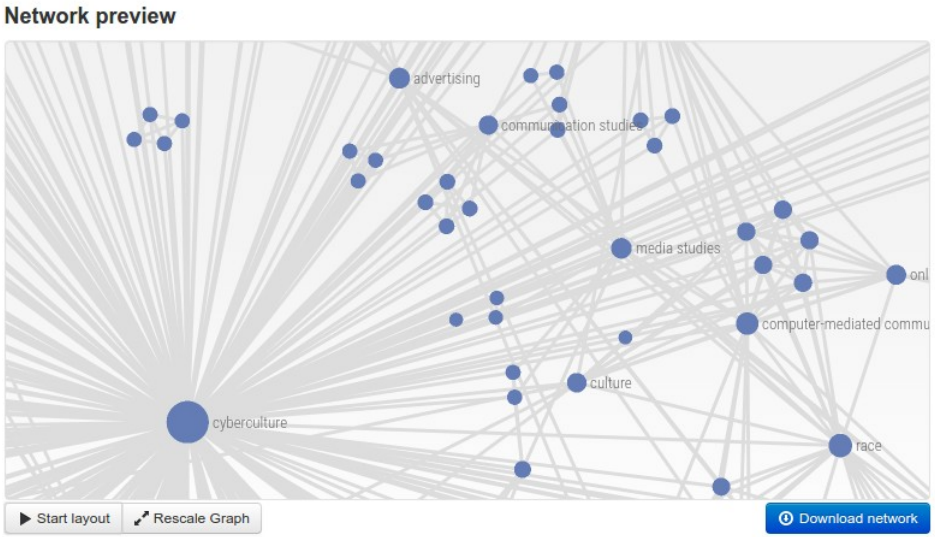
ScienceScape : scientométrie en quelques clics

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

- Étudier les auteurs, mots-clés et revues d'un ensemble de publications exportées depuis Scopus ou WebOfScience

  → exemple : http://jiminy.medialab.sciences-po.fr/data/tools-demo/scopus.csv



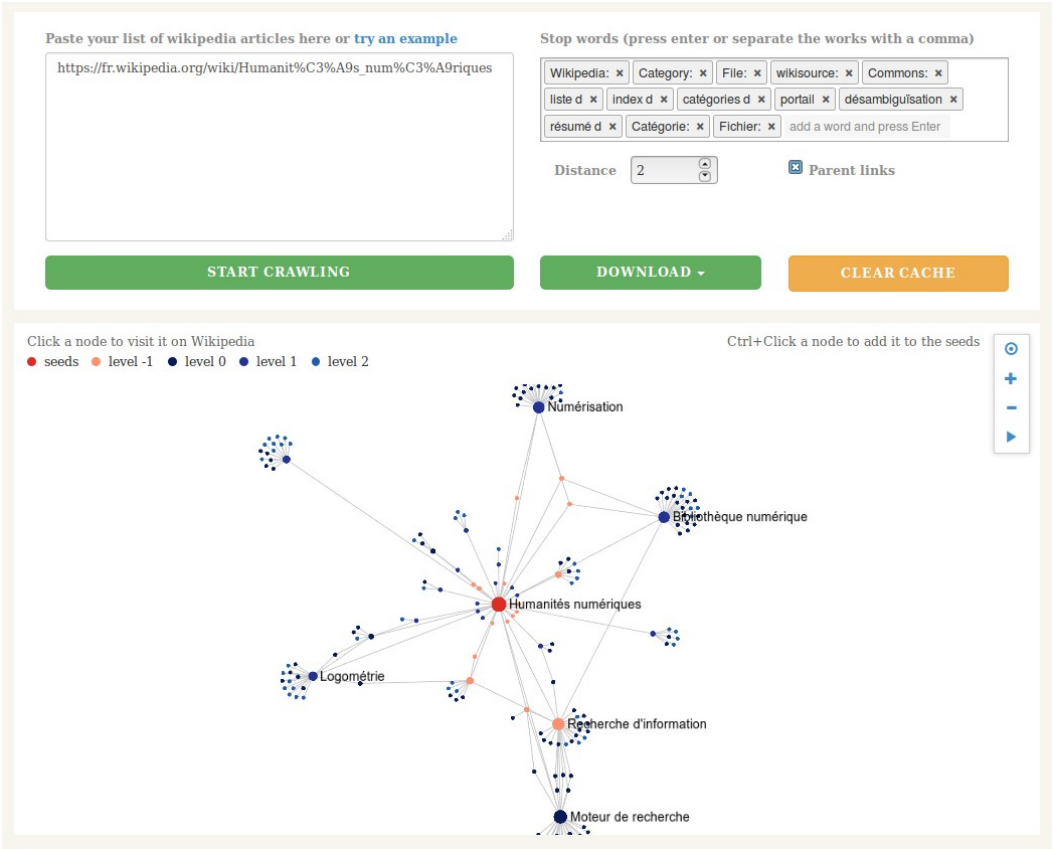


SeeAlsology : exploration sémantique rapide

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

- Explorer le réseau des liens présents dans les sections « Voir aussi », « Articles connexes » des pages Wikipedia

→ exemple : https://fr.wikipedia.org/wiki/Humanit%C3%A9s_num%C3%A9riques



Results preview

Source	Target	Level
Humanités numériques	Bibliothèque numérique	1
Humanités numériques	Fouille de textes	1
Humanités numériques	Littérature informatique	1
Humanités numériques	Logométrie	1
Humanités numériques	Moteur de recherche	1
Humanités numériques	Numérisation	1
Humanités numériques	Recherche d'information	1
Humanités numériques	Système de recherche d'information	1
Humanités numériques	Science de l'information	1
Humanités numériques	Text Encoding Initiative	1
Humanités numériques	Topic model	1
Humanités numériques	Digital Medievalist	1
Humanités numériques	Gephi	1
Humanités numériques	open source	1
Humanités numériques	e-Diaspora	1
Humanités numériques	Hyperbase	1
Humanités numériques	Étienne Brunet (linguiste)	1
Humanités numériques	Université Nice-Sophia-Antipolis	1
Humanités numériques	IRaMuTeQ	1
Humanités numériques	Pierre Ratinaud	1
Humanités numériques	Voyant Tools	1
Humanités numériques	Prospéro (logiciel)	1
Humanités numériques	Francis Chateauraynaud	1
Humanités numériques	Philcarto	1
Humanités numériques	OpenRefine	1
Digital Medievalist	Humanités numériques	0

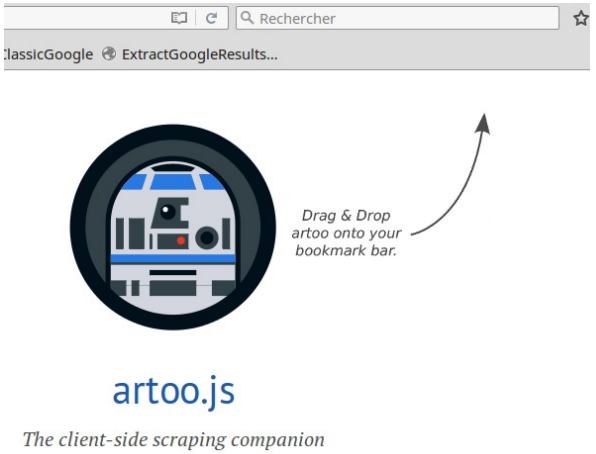
Not found (See also section missing, or bad page name)
- Science de l'information
- Philcarto
- Prospéro (logiciel)
- Littérature informatique
- Pierre Ratinaud
- IRaMuTeQ
- Voyant Tools
- e-Diaspora

Stopped from stopWord list
- Catégorie:Société de l'information
- Persée (portail)
- Modèle probabiliste de pertinence
- Liste de personnalités liées à Périgueux
- Liste de sociologues
- Catégorie:Enseignement supérieur en France
- Catégorie:Centre national de la recherche scientifique
- Catégorie:Université d'Aix-Marseille
- Catégorie:Œuvre philosophique
- Catégorie:Sciences de l'information et de la communication
- Liste de linguistes
- Catégorie:Université de Lausanne
- Catégorie:Édition électronique
- Catégorie:Sciences humaines et sociales
- Catégorie:Libre accès en science
- Catégorie:Universitaire français du XXIe siècle
- Catégorie:Naissance en août 1971
- Catégorie:Naissance à Carpentras
- Catégorie:Informaticien français
- Catégorie:Ingénieur français
- Liste de moteurs de recherche
- Liste de bases de données des ministères français
- Portail du patrimoine oral
- Liste de bibliothèques numériques
- Liste de normes ISO par domaines
- Liste de projets de sciences citoyennes
- Liste de bibliothèques détruites
- Liste de logiciels libres
- Portail web
- Lycos (portail web)
- Liste de frameworks PHP
- Liste de logiciels wiki
- Liste des universités en France
- Catégorie:Traitement automatique du

→ **Gephi, Manylines, …**

artoo : extraire des données du web (avancé)

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



- Un bookmarklet à ajouter dans la barre de favoris du navigateur
- Une librairie JavaScript de fonctions utiles pour le scraping (extraction de données) depuis la console du navigateur (F12)

→ exemple : https://en.wikipedia.org/wiki/List_of_countries_by_carbon_dioxide_emissions

```
> 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
```

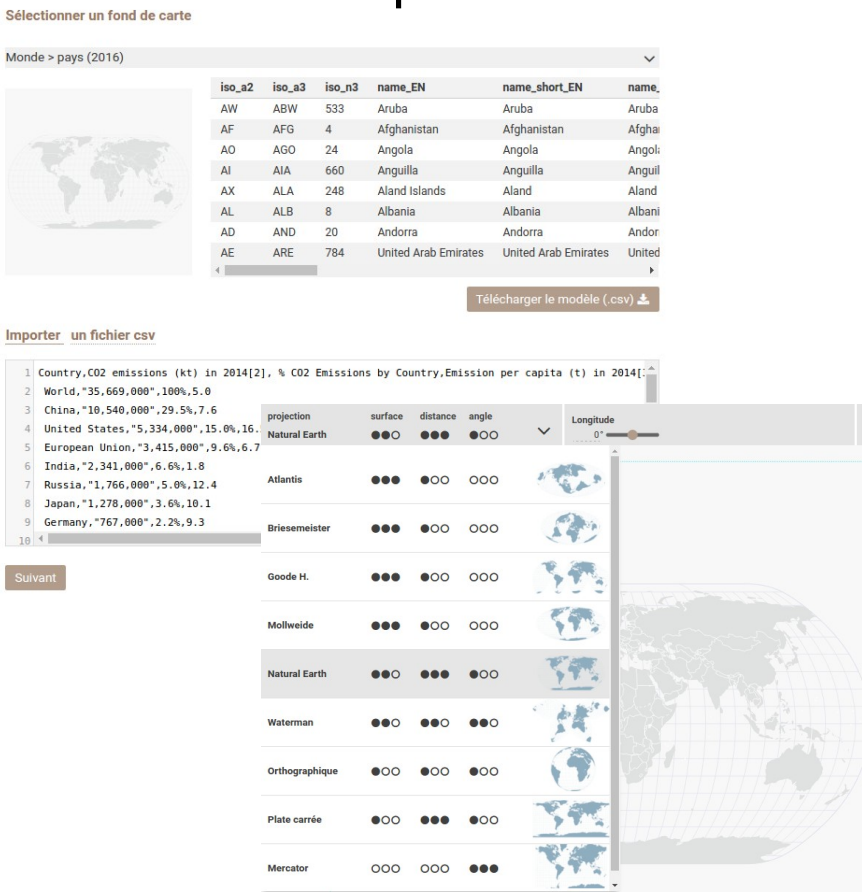
→ CSV-Rinse-Repeat, Table2Net, Khartis, …

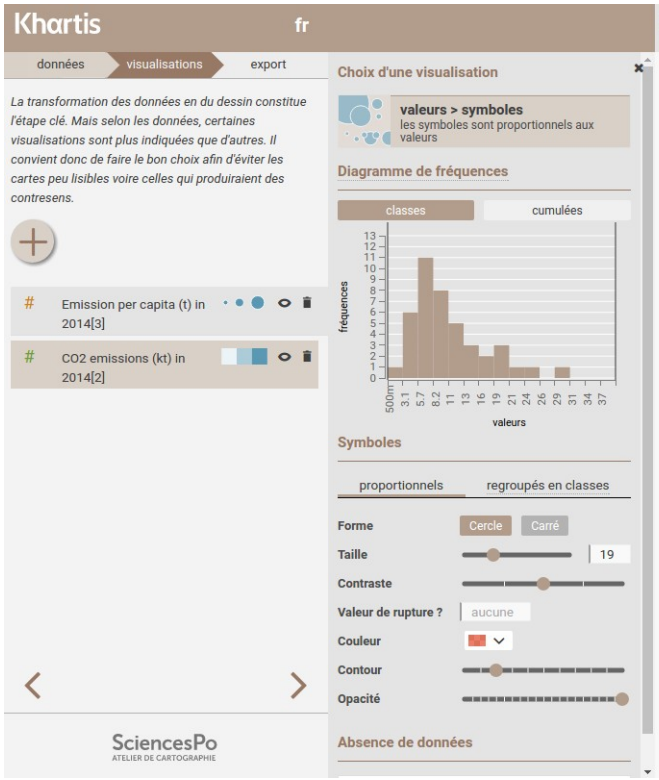
Khartis : cartographier les données d'un CSV

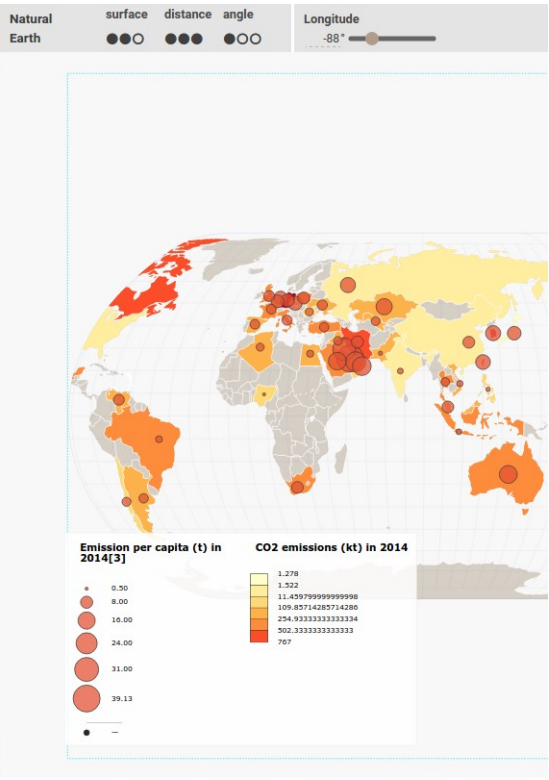
http://www.sciencespo.fr/cartographie/khartis/

- Cartographier des données tabulaires géonommées en quelques clics et exporter une image PNG ou SVG

→ exemple avec le CSV tiré d'artoo : http://jiminy.medialab.sciences-po.fr/data/tools-demo/CO2-world-emissio







Gazouilloire : extraction de tweets (avancé)

https://github.com/medialab/gazouilloire

- Collecter en direct en continu (et jusqu'à 7 jours en arrière)
  - des tweets par mots-clés, utilisateurs, localisation, langue…
  - les conversations et médias associés
  - des profils d'utilisateurs

```
{
    "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
```

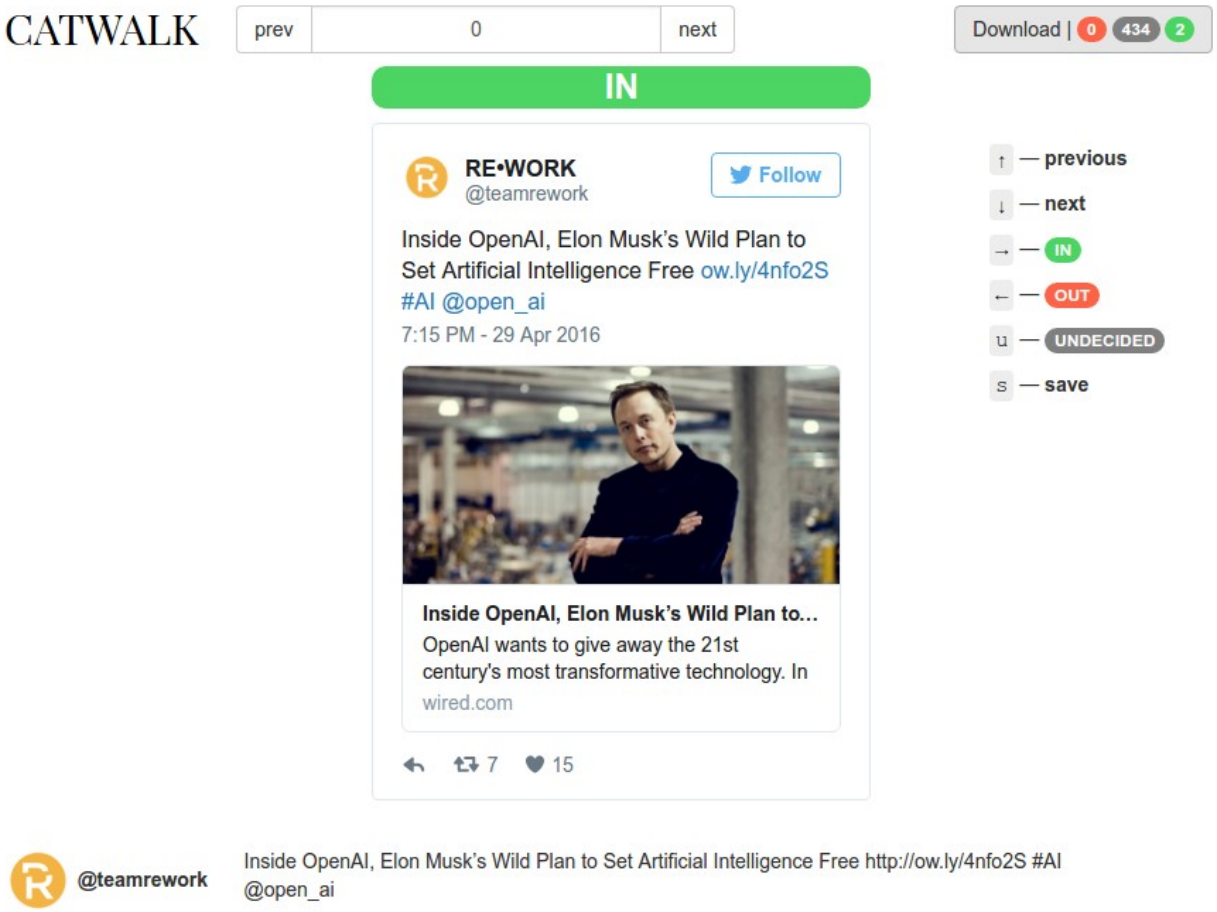
→ exemple d'export CSV : http://jiminy.medialab.sciences-po.fr/data/tools-demo/tweets-data-humanities.csv

→ Catwalk, CSV-Rinse-Repeat, …

CatWalk : sélection qualitative de tweets

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

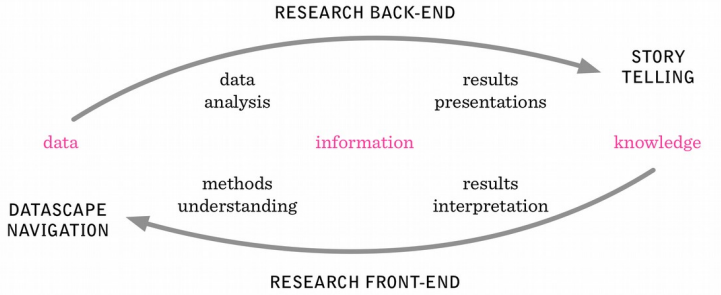
- Passer en revue rapidement « *à la Tinder* » tous les tweets d'un CSV pour décider de les inclure / exclure d'un corpus

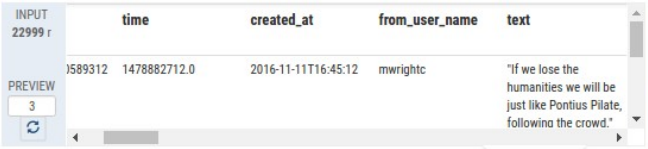


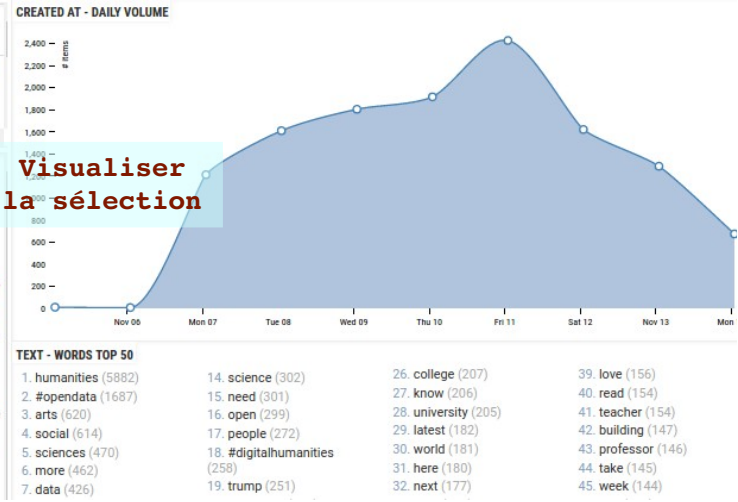
CSV-Rinse-Repeat : exploration de CSV (avancé)

http://tools.medialab.sciences-po.fr/csv-rinse-repeat/

- Itérations successives pour identifier problèmes & questions
- Nettoyer, filtrer, explorer, visualiser, enrichir et exporter en JavaScript le contenu d'un CSV








```
output = input.filter(function(item, i){
  item.quoted_usernames = (item.text.match(/@\w+/g) || [])                  // Extract quoted users
                          .join('|').replace(/@/g, '').toLowerCase();
  item.hashtags = (item.text.match(/#\w+/g) || []).join('|').toLowerCase();  // Extract hashtags
  return ! item.text.match(/^RT @\w+: /);                                    // Filter retweets
});
```

Table2Net : faire un réseau à partir d'un CSV

http://tools.medialab.sciences-po.fr/table2net/

- Générer un graphe de liens entre éléments à partir des données d'un fichier tableur
- Exemples tirés de CSV-Rinse-Repeat : http://jiminy.medialab.sciences-po.fr/data/tools-demo/tweets-data-humanities-rinsed
  - Réseau normal : nodes = hashtags / links = row number
  - Réseau bipartite : nodes 1 = from_user_name / nodes 2 = hashtags
  - Réseau citations : nodes = from_user_name / links = quoted_usernames
  - etc.

Table 2 Net

Load your CSV table

It has to be **comma-separated** and the first row must be dedicated to **column names**.

Parsing successful. 44 columns and 5576 rows.

1. Type of Network

Normal (one type of node)

You will have to choose:

- Which column X will define the nodes
- Which column Y will define the links

2. Nodes

X Which column defines the nodes?

hashtags

Pipe-separated "|"

Sample of nodes extracted with these settings: (sample)

#goweser #adventurebike #kotaposo #dh #xcbike

3. Links

Y Which column defines the links?

Row number

One expression per cell

Sample of items extracted with these settings: (sample)

5252 3621 1816 1847 4562

4. Additional settings

Optional: time series

No temporal data

Select only a column containing **integers**.

Optional: edge weight

Weight the edges

5. Build the network

Build and download the network (GEXF)

NB: this may take a while, please be patient.

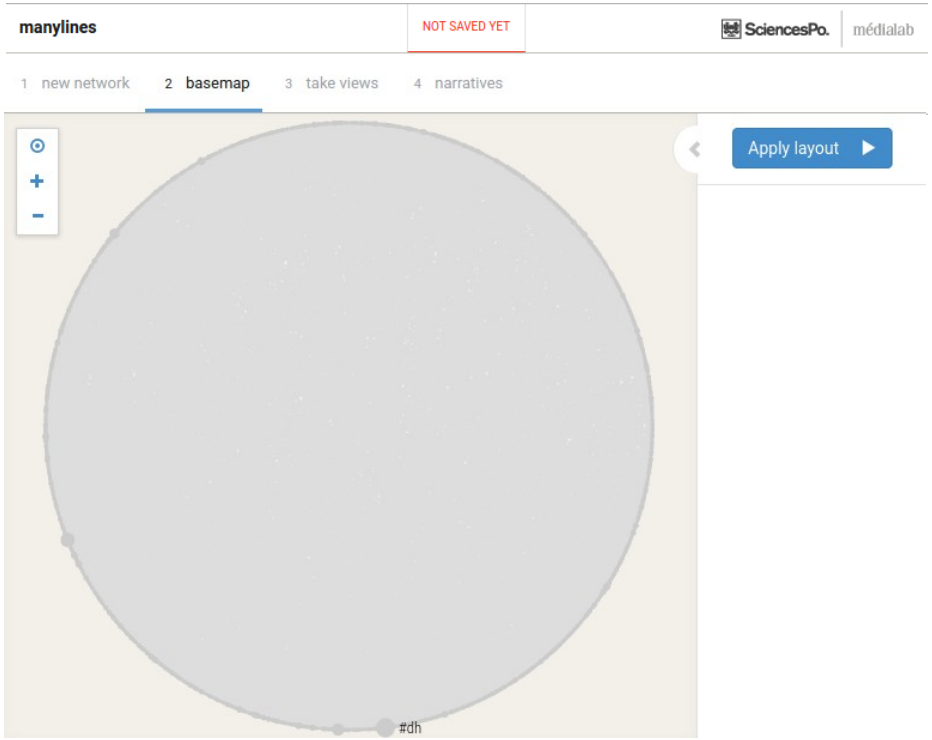
→ **Gephi, Manylines, …**

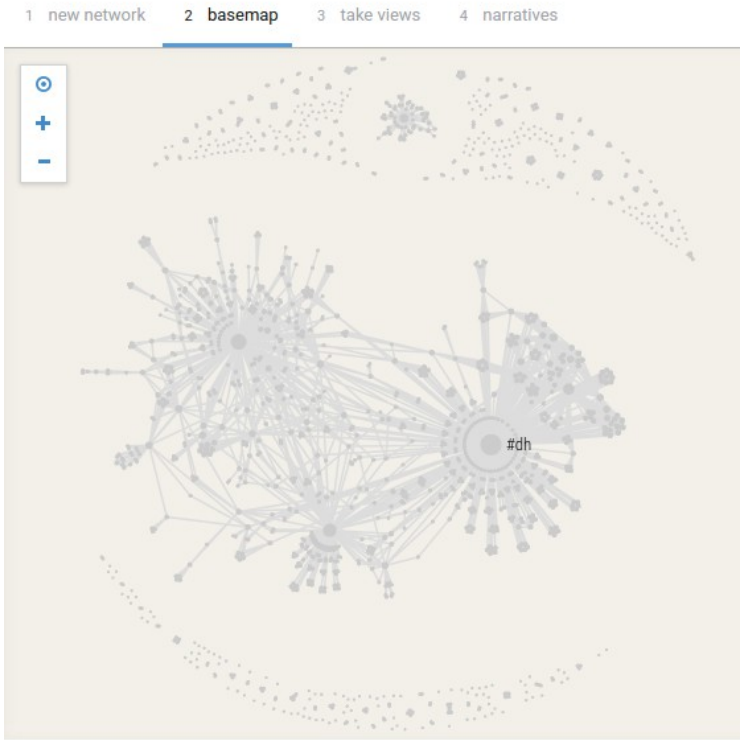
Manylines : publier et documenter un réseau

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

- Explorer et publier rapidement en ligne un réseau
- Raconter le réseau sous la forme de slides de présentation

→ ex exporté de Table2Net : http://jiminy.medialab.sciences-po.fr/data/tools-demo/tweets-data-humanities-hashtags-network.gexf





- Exemple de slides résultants : http://tools.medialab.sciences-po.fr/manylines/embed#/narrative/5aaec64d-96c1-463

Merci de votre attention !

@medialab_ScPo

benjamin.ooghe@sciencespo.fr