



HAL
open science

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

Benjamin Ooghe

► **To cite this version:**

Benjamin Ooghe. médialab Tools : de petits outils simples et complexes au service de la DataScience. Action Nationale de Formation - Collecter et produire des données pour la recherche en SHS, Réseau MATE-SHS, Nov 2016, Fréjus, France. hal-03631539

HAL Id: hal-03631539

<https://sciencespo.hal.science/hal-03631539v1>

Submitted on 5 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

médialab Tools

De petits outils simples et complexes au service de la DataScience

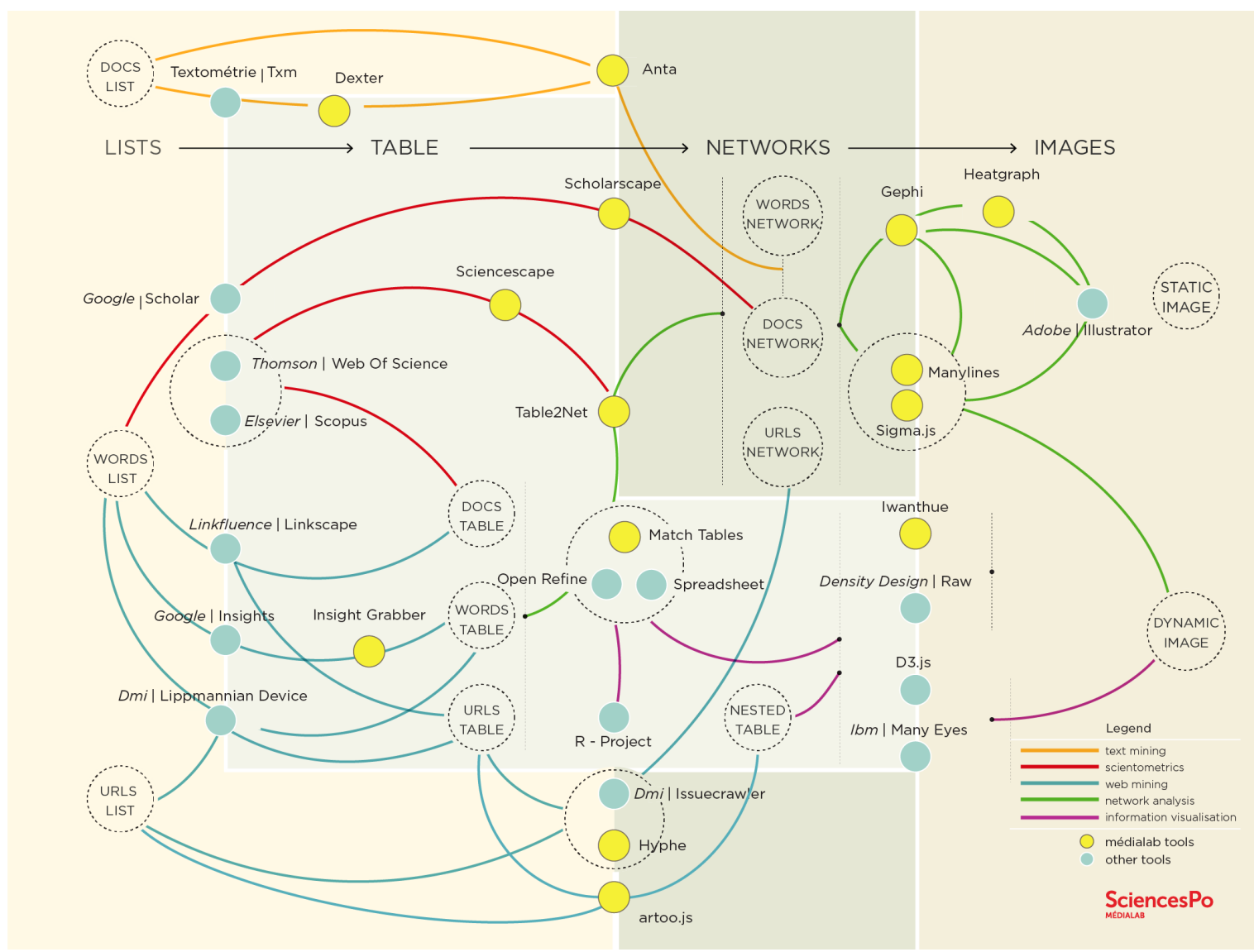
Benjamin Ooghe-Tabanou, Sciences Po, médialab, Paris, France

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

Install Google Bookmarklets

Drag & drop images below into your bookmark bar:

[\[G\]](#) [\[G\]](#)

Redirect to Classic Google

Which language?

How many results per page?

You will be redirected to the following url:

```
https://encrypted.google.com/search?q=digital%20humanities&hl=en&num=100&start=0
```

Extract Classic Google Results

Search for "digital humanities" page 0 (with up to 100 urls per page)

103 new results in this page

Keep existing results & continue to the next page

Download CSV with 103 urls

Digital humanities - Wikipedia
https://en.wikipedia.org/wiki/Digital_humanities

Digital Humanities | Stanford Humanities - Stanford Humanities Center
shc.stanford.edu/digital-humanities

Digital humanities - Wikipedia
https://en.wikipedia.org/wiki/Digital_humanities

Digital humanities (DH) is an area of scholarship at the intersection of computing and the disciplines of the humanities. The nature of this activity ranges broadly, from the practical, such as digitizing historical texts, to the philosophical, such as reflection on the nature of representation itself.

Digital humanities - Wikipedia
https://en.wikipedia.org/wiki/Digital_humanities

Digital humanities (DH) is an area of scholarly activity at the intersection of computing and the disciplines of the humanities. The nature of this activity ranges broadly, from the practical, such as digitizing historical texts, to the philosophical, such as reflection on the nature of representation itself.

Digital Humanities | Stanford Humanities - Stanford Humanities Center
shc.stanford.edu/digital-humanities

Digital Humanities. The digital humanities at Stanford sit at the crossroads of computer science and

→ **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>

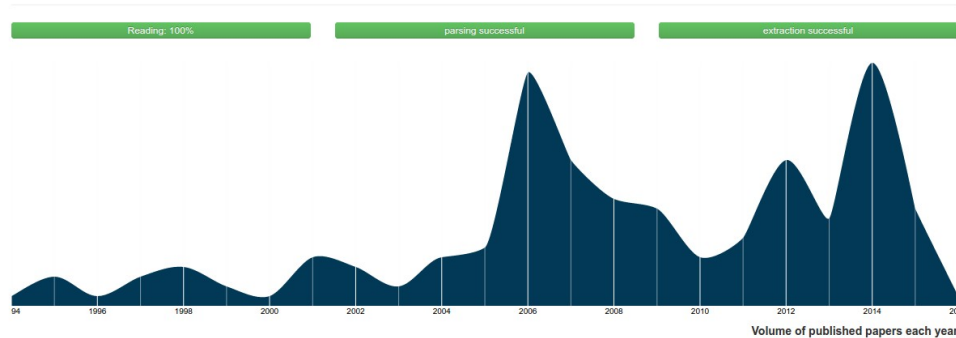


ScienceScape

Helpers for scientometrics. Convert files, get networks, visualize stuff from Scopus or Web of Science.

Volume of papers over time

Upload a Scopus CSV file and look at how many papers are published each year (in your file)

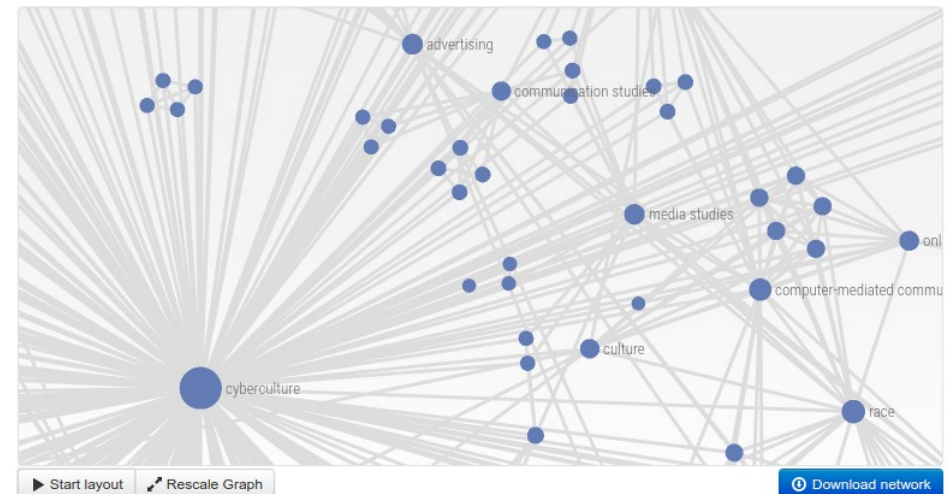


Filtering

Remove disconnected nodes

Build network

Network preview



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

Paste your list of wikipedia articles here or [try an example](#)

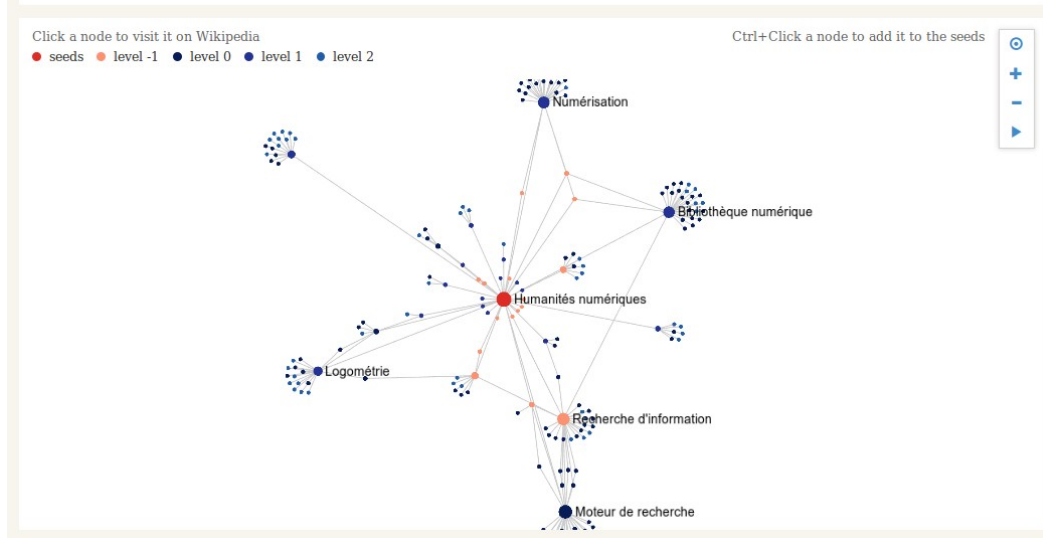
https://fr.wikipedia.org/wiki/Humanit%C3%A9s_num%C3%A9riques

Stop words (press enter or separate the works with a comma)

Wikipedia: x Category: x File: x wikisource: x Commons: x
 liste d: x index d: x catégories d: x portail: x désambiguïsation: x
 résumé d: x Catégorie: x Fichier: x add a word and press Enter

Distance: 2 Parent links

START CRAWLING DOWNLOAD CLEAR CACHE



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 Ratinand	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 Ratinand
- IRaMuTeQ
- Voyant Tools
- e-Diaspora

Stopped from stopWord list

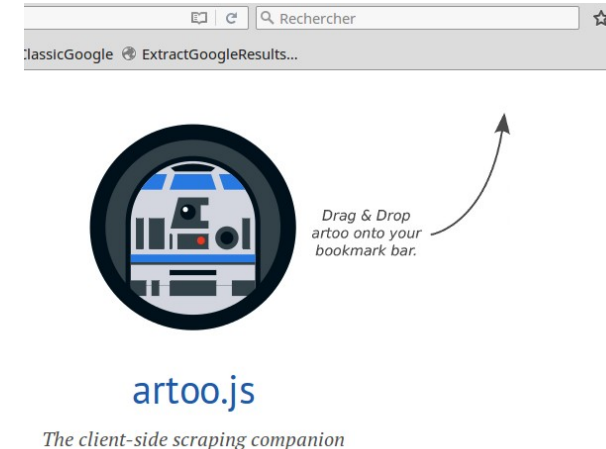
- Catégorie:Société de l'information
- Persée (portail)
- Modèle:probabiliste de pertinence
- Liste de personnalités liées à Périgueux
- Liste de sociologues
- Catégorie:Enseignement supérieur en France
- Catégorie:Centre national de la recherche scientifique
- Catégorie:Université d'Aix-Marseille
- 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://jimony.medialab.sciences-po.fr/data/tools-demo/CO2-world-emission>

Sélectionner un fond de carte

Monde > pays (2016)

iso_a2	iso_a3	iso_n3	name_EN	name_short_EN	name
AW	ABW	533	Aruba	Aruba	Aruba
AF	AFG	4	Afghanistan	Afghanistan	Afgha
AO	AGO	24	Angola	Angola	Angol
AI	AIA	660	Anguilla	Anguilla	Anguil
AX	ALA	248	Aland Islands	Aland	Aland
AL	ALB	8	Albania	Albania	Albani
AD	AND	20	Andorra	Andorra	Andor
AE	ARE	784	United Arab Emirates	United Arab Emirates	United

Télécharger le modèle (.csv)

Importer un fichier csv

Country,CO2 emissions (kt) in 2014[2], % CO2 Emissions by Country,Emission per capita (t) in 2014[3]	projection	surface	distance	angle	Longitude
World,"35,669,000",100%,5.0	Natural Earth	●●●	●●●	●○○	0°
China,"10,540,000",29.5%,7.6	Atlantis	●●●	●○○	○○○	
United States,"5,334,000",15.0%,16.4	Briesemeister	●●●	●○○	○○○	
European Union,"3,415,000",9.6%,6.7	Goode H.	●●●	●○○	○○○	
India,"2,341,000",6.6%,1.8	Mollweide	●●●	●○○	○○○	
Russia,"1,766,000",5.0%,12.4	Natural Earth	●●○	●●●	●○○	
Japan,"1,278,000",3.6%,10.1	Waterman	●●○	●●○	●●○	
Germany,"767,000",2.2%,9.3	Orthographique	●○○	●○○	●○○	
	Plate carrée	●○○	●●●	●○○	
	Mercator	○○○	○○○	●●●	

Khartis fr

données visualisations export

La transformation des données en du dessin constitue l'étape clé. Mais selon les données, certaines visualisations sont plus indiquées que d'autres. Il convient donc de faire le bon choix afin d'éviter les cartes peu lisibles voire celles qui produiraient des contresens.

Choix d'une visualisation

valeurs > symboles
les symboles sont proportionnels aux valeurs

Diagramme de fréquences

classe cumulée

fréquences

valeurs

Symboles

proportionnels regroupés en classes

Forme Cercle Carré

Taille 19

Contraste

Valeur de rupture ? aucune

Couleur

Contour

Opacité

Absence de données

Natural Earth surface distance angle Longitude -88°

Emission per capita (t) in 2014[3]

CO2 emissions (kt) in 2014

0.00 1.278 1.522 11.450799999999998 109.85714285714286 254.93333333333334 502.3333333333333 767

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": " ",
    "secret": " ",
    "oauth_token": " ",
    "oauth_secret": " "
  },
  "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://jimini.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

CATWALK


prev 0 next

Download | 0 434 2

IN

 **RE-WORK**
@teamrework [Follow](#)

Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free ow.ly/4nfo2S
#AI @open_ai
7:15 PM - 29 Apr 2016



Inside OpenAI, Elon Musk's Wild Plan to...
OpenAI wants to give away the 21st century's most transformative technology. In wired.com

← ↻ 7 ❤ 15

↑ — previous
↓ — next
→ — IN
← — OUT
u — UNDECIDED
s — save

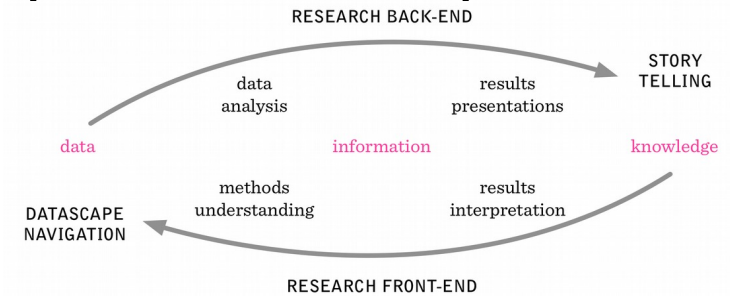


Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free <http://ow.ly/4nfo2S> #AI @open_ai

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



INPUT 22999 r

time	created_at	from_user_name	text	
1589312	1478882712.0	2016-11-11T16:45:12	mwrightc	"If we lose the humanities we will be just like Pontius Pilate, following the crowd."

PREVIEW 3

OUTPUT 12587 r

links	medias_urls	medias_files	quoted_usernames	hashtags
https://pbs.twimg.com/media/CxNO-QnWgAAtb1K.jpg	798068409783160833	zmonnagotla	#rcc2016 #innovation #opendata #impact	

PREVIEW 3

CREATED AT - DAILY VOLUME

Visualiser la sélection

CODE YOUR FILTER

```

1 // FILTER YOUR DATA HERE
2
3 // Just fill the "output" variable using "input"
4 output = input.filter(function(item, i){
5
6 // EXTRACT QUOTED USERS
7 item.quoted_usernames = (item.text.match(/@\w+/g) || []).join('|').replace(/@/g, '').toLowerCase();
8
9 // EXTRACT HASHTAGS
10 item.hashtags = (item.text.match(/#\w+/g) || []).join('|').toLowerCase();
11
12 // FILTER Retweets
13 return ! item.text.match(/^RT @\w+:/);
14
15
16 });
17
18 // Hit CTRL + ENTER to run the code

```

Filtrer & enrichir les données en JS

HASHTAGS - WORDS CLOUD

TEXT - WORDS TOP 50

1. humanities (5882)	14. science (302)	26. college (207)	39. love (156)
2. #opendata (1687)	15. need (301)	27. know (206)	40. read (154)
3. arts (620)	16. open (299)	28. university (205)	41. teacher (154)
4. social (614)	17. people (272)	29. latest (182)	42. building (147)
5. sciences (470)	18. #digitalhumanities (258)	30. world (181)	43. professor (146)
6. more (462)	19. trump (251)	31. here (180)	44. take (145)
7. data (426)		32. next (177)	45. week (144)

```

output = input.filter(function(item, i){
  item.quoted_usernames = (item.text.match(/@\w+/g) || []).join('|').replace(/@/g, '').toLowerCase(); // Extract quoted users
  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://jimony.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.

→ **Gephi, Manylines, ...**

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

Which column defines the nodes?

hashtags

Pipe-separated "|"

Sample of nodes extracted with these settings: [sample](#)

[#goweser](#) [#adventurebike](#) [#kotaposo](#) [#dh](#) [#xcbike](#)

3. Links

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)

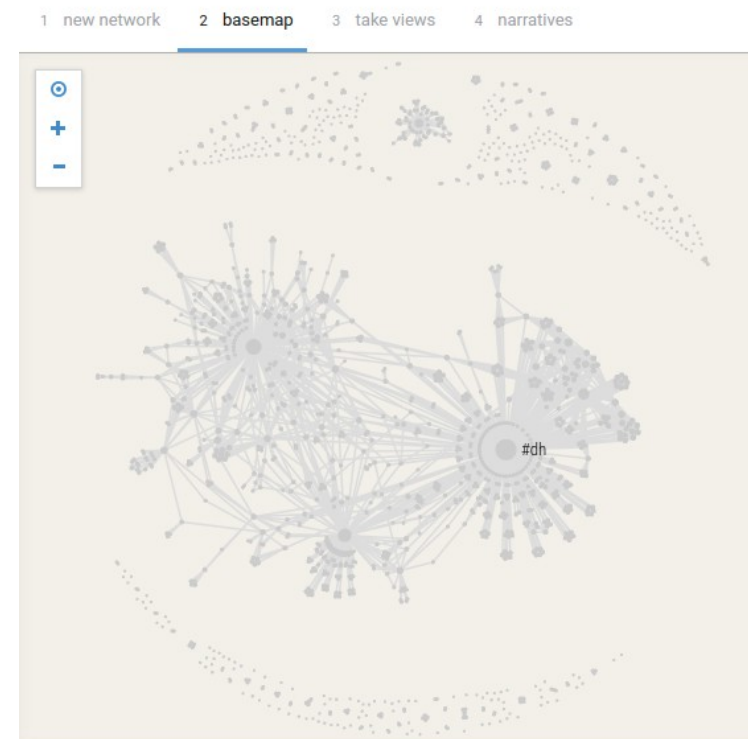
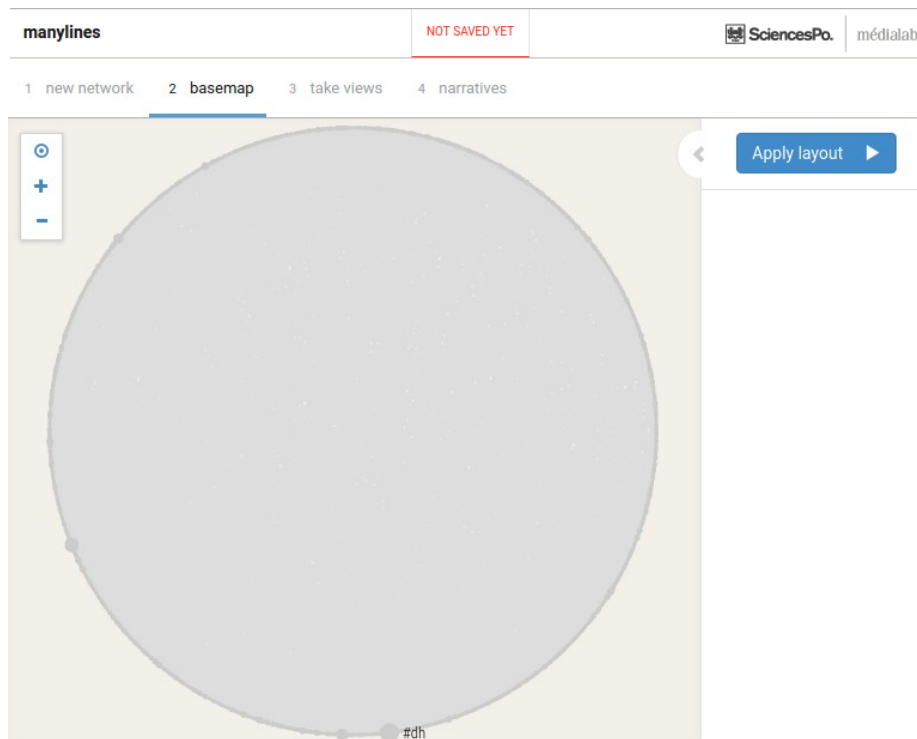
NB: this may take a while, please be patient.

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 !

SciencesPo
MÉDIALAB

[@medialab_ScPo](https://twitter.com/medialab_ScPo)

benjamin.ooghe@sciencespo.fr