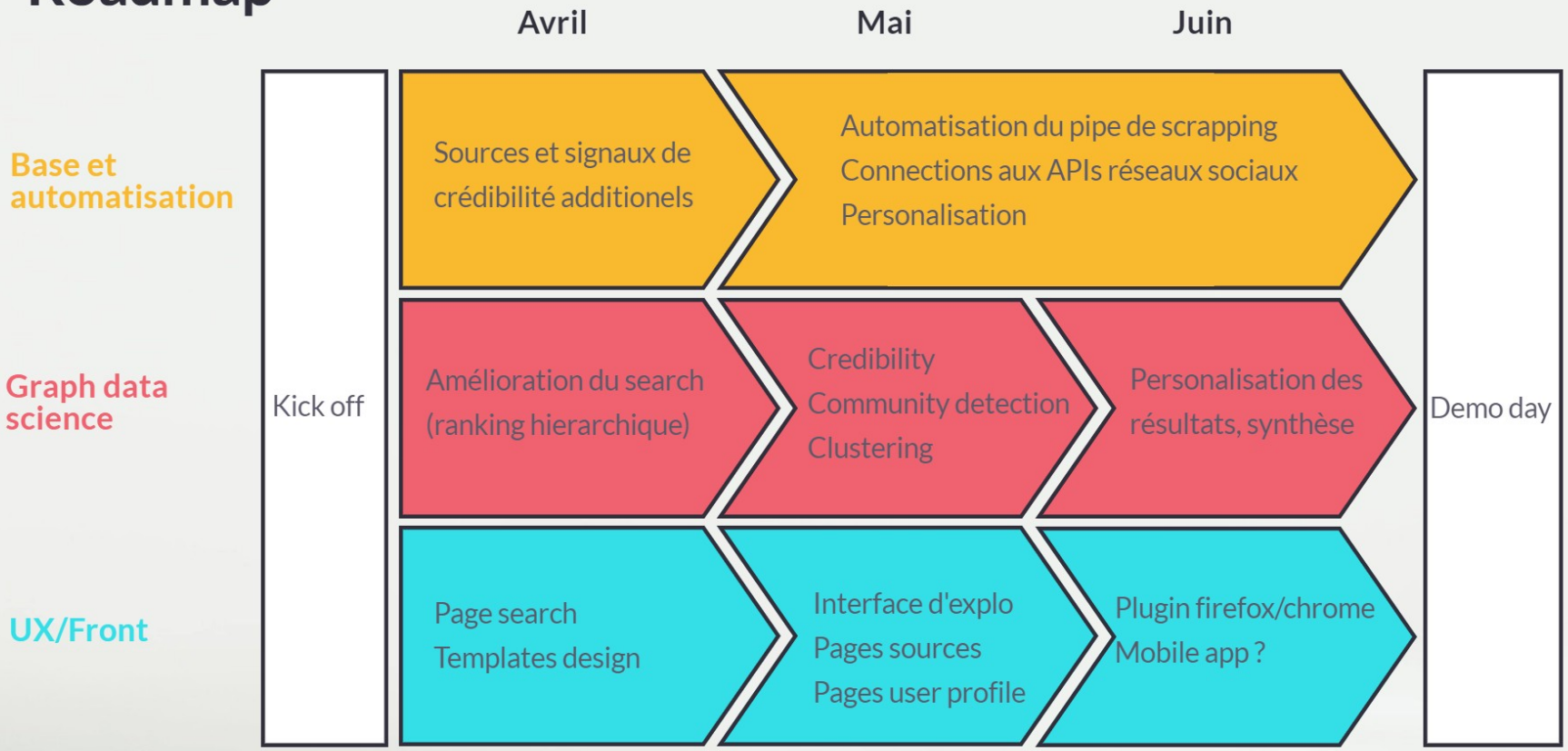
# Validalab

s’informer en confiance

# Roadmap




**Validalab**

**Rejoignez le réseau !**

# « L’alcool tue le coronavirus »


Est-ce que cette information est sérieuse ?

# « L'alcool tue le coronavirus »

THE Sun

**HOT TODDY 'REMEDY' Coronavirus UK – First Brit known to catch virus 'beat deadly flu with glass of hot whisky and honey'**

EXCLUSIVE

Nick Pisa
3 Feb 2020, 9:15 | Updated: 6 Feb 2020, 9:02


NEWS / CORONAVIRUS PANDEMIC

**Iran: Over 700 dead after drinking alcohol to cure coronavirus**

*More than 700 people died in Iran after ingesting toxic methanol, erroneously thinking it can cure the new coronavirus.*

27 Apr 2020

## Tracer l'origine et le résultat de cette information...

# « L’alcool tue le coronavirus »


Organisation mondiale de la Santé

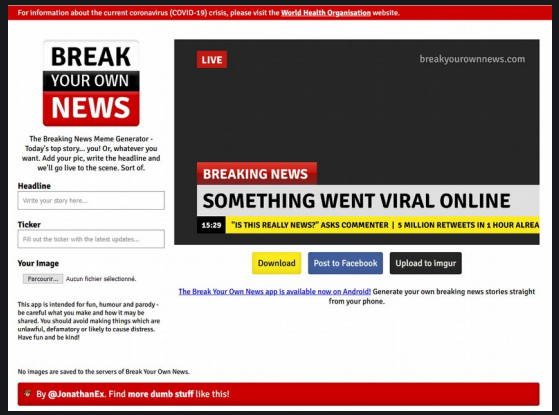
**Boire du méthanol, de l’éthanol ou de l’eau de Javel NE PERMET PAS de prévenir ou de guérir la COVID-19 et peut être extrêmement dangereux**

Le méthanol, l’éthanol et l’eau de Javel sont des poisons. Leur ingestion peut entraîner des lésions voire la mort. Le méthanol, l’éthanol et l’eau de Javel sont parfois utilisés dans des produits d’entretien qui servent à détruire le virus sur les surfaces, mais il ne faut jamais en boire. Ils ne détruisent pas le virus dans l’organisme et entraînent des lésions des organes internes.

Pour se protéger de la COVID-19, il faut désinfecter les objets et les surfaces, en particulier ceux que vous touchez régulièrement. Lavez-vous les mains souvent et soigneusement et évitez de vous toucher les yeux, la bouche et le nez.

## … Accéder à de meilleurs points de vue…

# « L’alcool tue le coronavirus »





… qualifier la source (un site parodique) …

# « L’alcool ne tue pas le coronavirus »


20 minut

**Boire de l’alcool « tue le coronavirus » ? Attention à cette fausse capture d’écran**

**FAKE OFF** Lever le coude n’est pas recommandé pour soigner cette maladie

*M.Co.* | Publié le 28/02/20 à 17h54 — Mis à jour le 28/02/20 à 17h54

… c’est construire le savoir autour de l’information.
c’est la valider ou la rejeter.

Et c’est **automatisable** !

# Un déluge d’information…

Aujourd’hui, **tout le monde** peut publier du contenu en ligne et être lu/vu.

Mais l’économie de l’attention favorise les contenus émotionnels (au dépend des contenus d’experts).

Seuls 24 % des Français interrogés indiquent faire **confiance** aux médias (TV, presse papier et en ligne) – chiffre le plus bas en Europe.

# … qui favorise la diffusion des fausses informations

Baisse de la **confiance** générale

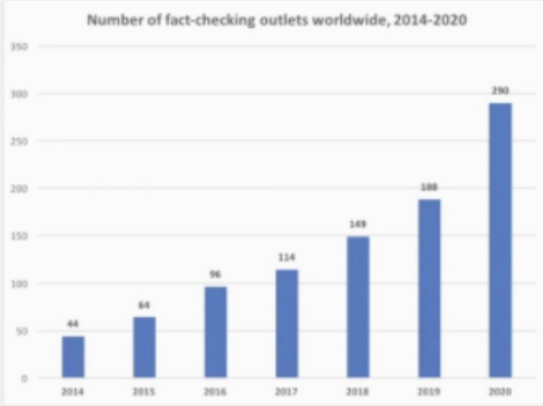
**Polarisation** de la société

**Défiance** envers les marques, les institutions et les médias

Coût humain, politique et financier ($78Milliards/an)

# Qui lutte contre les fausses infos ?

- Des communautés dédiées s’organisent
- Mais les fact checkers pros sont débordés
- Difficultés pour le grand public qui n’a pas les méthodes
- Pas d’outils dédiés, surtout pour le monde francophone



# l’équipe

**Jean-Marc Guerin**, fondateur

13 ans d’expérience à concevoir des solutions d’Intelligence Artificielle dans des startups

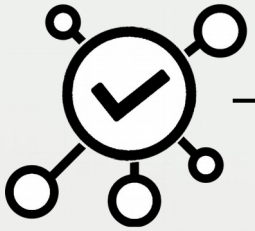
Passionné par le fact checking et l’intelligence collaborative

## Validalab

s’informer en confiance

# L’app Validalab

C’est le moteur de recherche qui **accélère le fact checking** :



**Q**ualifier les sources

**A**ccéder à de meilleurs points de vue

**T**racer les citations

Validalab trouve

- ce qui fait qu’une information est **fiable** ou
- s’il s’agit d’une **fausse rumeur**

# Information **fiable** ou **fausse rumeur** ?

## Pour **tout type** de contenu

- Web et réseaux sociaux
- Vidéos, tweets, articles, et documents
- Spécialisés ou grand public

## Pour **tout profil** d'utilisateur

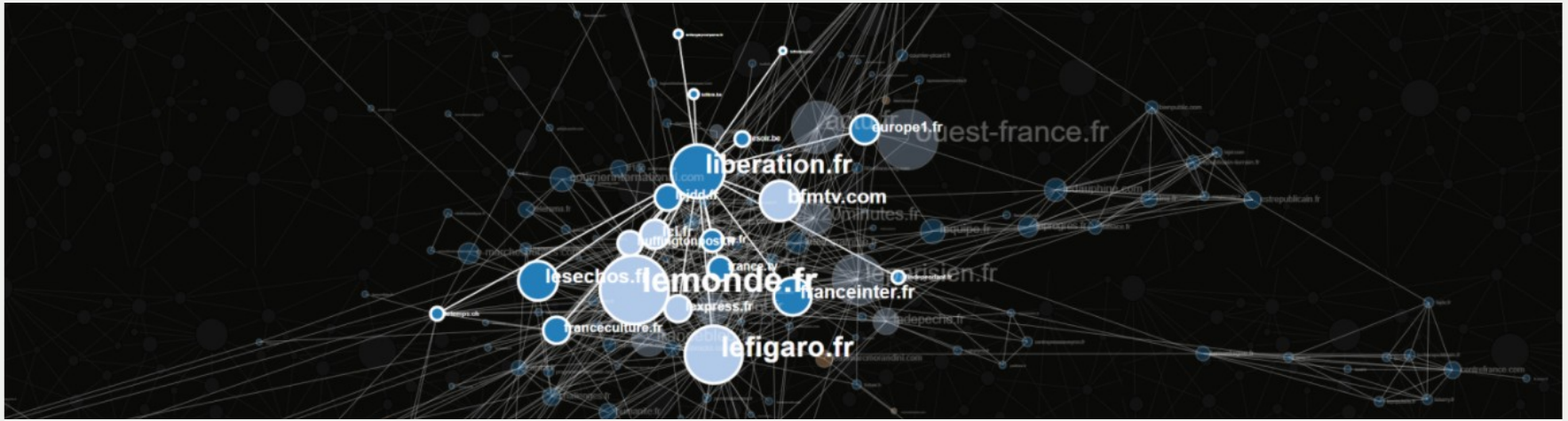
- Grand public
- Factcheckers
- Producteur de contenu
- Experts reconnus

Collaboratif et **totalement scalable** – ready for infowar

# Le Validagraph : concept


**L'information est fiable si toute la chaîne d'information est traçable et crédible**

# Le Validagraph : pratique




### Recontextualisation

Validalab scanne et recontextualise les sites internets, blogs, twitter, youtube sous forme de graph de connaissance : le Validagraph.

### Validation

L'objectif du Validagraph est de vous fournir simplement et rapidement toutes les informations dont vous avez besoin pour comprendre et valider ou non votre information.

### Participation

Vous pouvez participer à la construction du Validagraph en partageant vos résultats de façon anonyme ou en créant votre profil pour suivre votre crédibilité.

# Validalab, c’est

- Un moteur de recherche basé sur la **crédibilité**
- Des outils de **contextualisation** pour mieux comprendre
- Des résultats **adaptés** à l’utilisateur
- Une plateforme de **collaboration** pour amateurs et professionnels

S’informer en confiance

# Validalab

## s’informer en confiance

Jean-Marc Guerin, fondateur
jm@validalab.fr

#GUADELOUPETECH

Mouvement IMPACT FRANCE

# Présentation technique

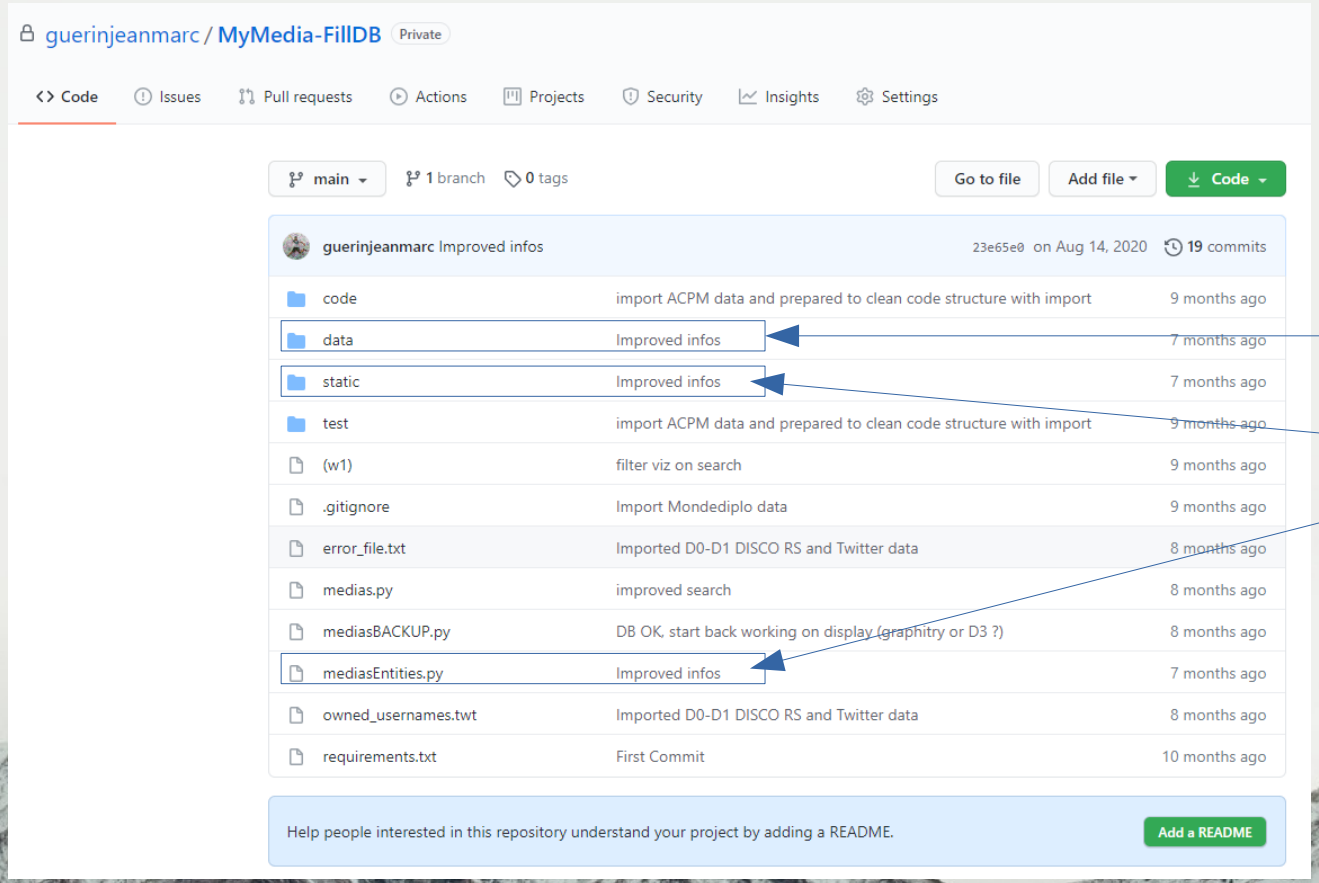
1. L’app Validalab, lancement et demo
2. La base Neo4j
3. Le code pour remplir la base
4. Le code de l’app

# 1. L’app Validalab

# L’app Validalab

- Base Neo4j
- Code python + flask
- Code html + CSS

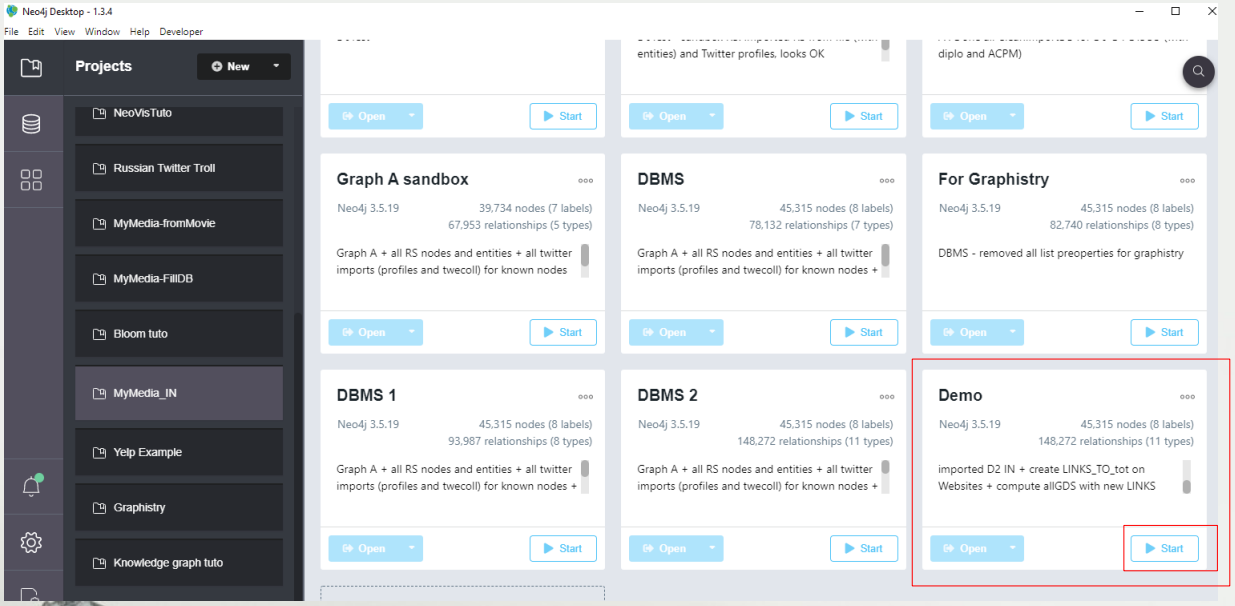
# Repo Github




- Code python pour remplir Neo4j
- HTML +CSS
- App

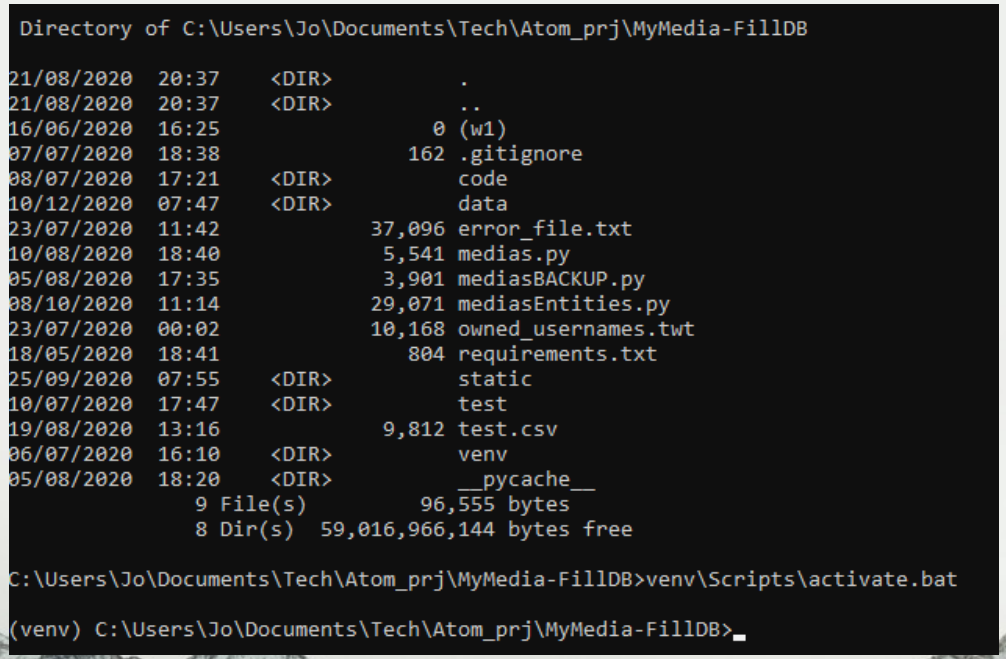
# Pour lancer l’app sur windows

- Lancer la base Neo4j



# Pour lancer l’app sur windows

- Activer l’environnement virtuel venv



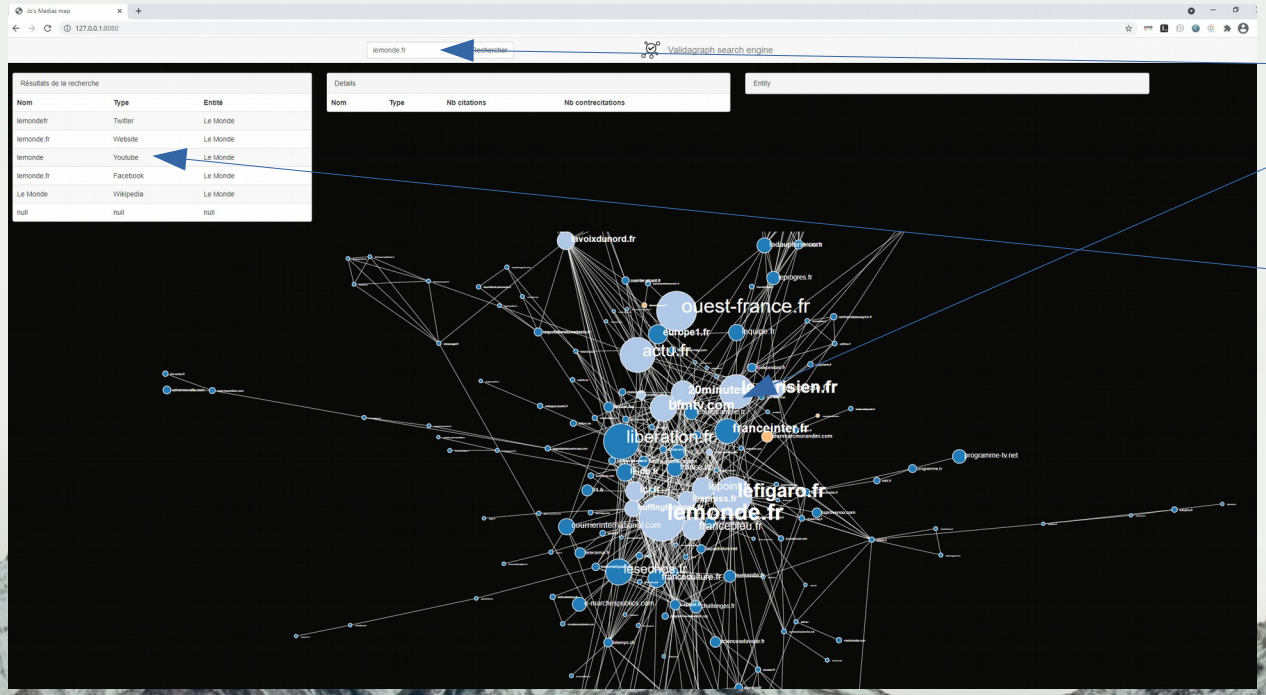
# Pour lancer l’app sur windows

- Executer mediasEntities.py

```
(venv) C:\Users\Jo\Documents\Tech\Atom_prj\MyMedia-FillDB>python mediasEntities.py
 * Serving Flask app "mediasEntities" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: off
 * Running on http://127.0.0.1:8080/ (Press CTRL+C to quit)
```

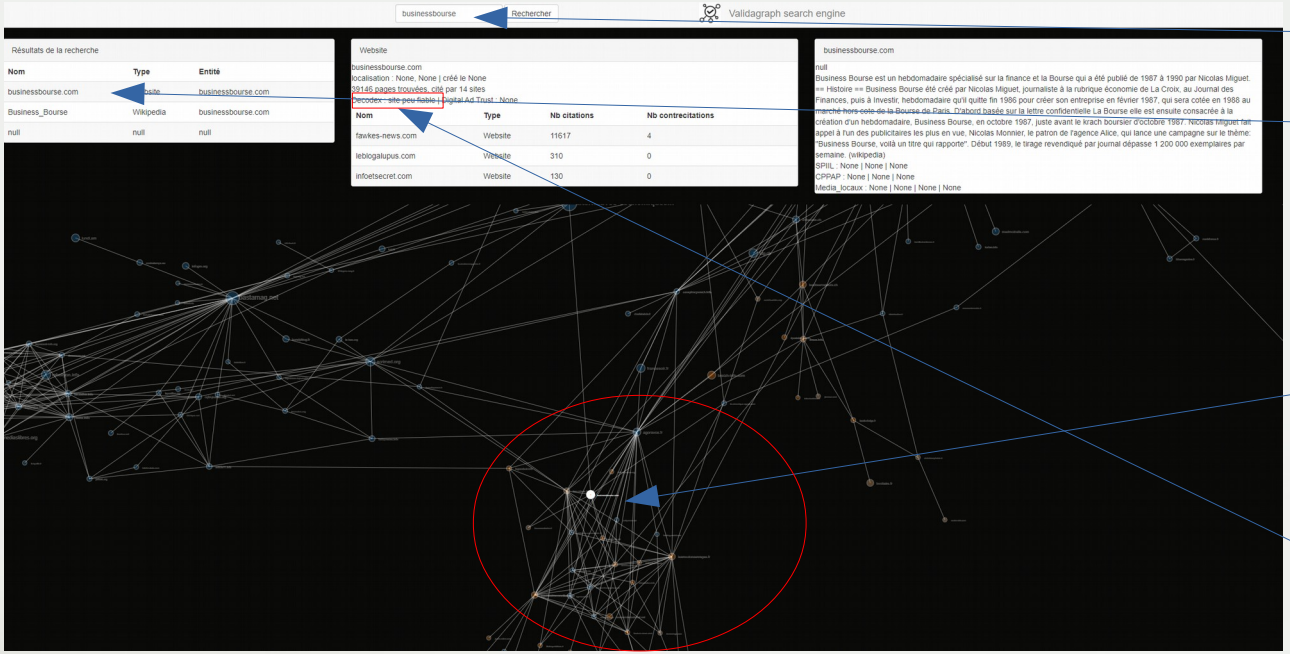
# Pour lancer l’app sur windows

- Dans chrome, aller sur http://127.0.0.1:8080/




- Barre de recherche
- Graph interactif (drag/ drop , click, zoom avec roulette…)
- Cliquer sur les résultats de la recherche permet d’afficher les infos sur les autres paneaux

# Exemple




- Chercher « businessbourse »
- cliquer sur le résultat de recherche « businessbourse.com »
- =>
- positionne le nœud dans la branche liée à la complosphere
- affiche l'info suivantes : site peu fiable pour Decodex

# Pour visualiser d'autres graphs

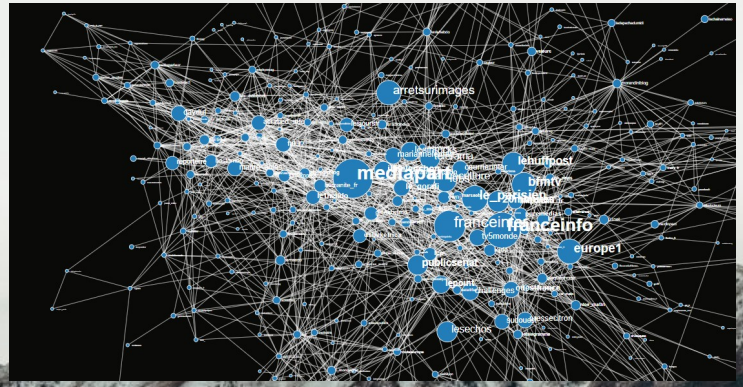
- Dans static/index_medias_entities.html :
- Chercher la ligne 240

```
231  var catColScale = d3.scaleOrdinal(d3.schemeCategory20);
232  var contColLinkScale = d3.scaleLinear()
233      .domain([2, 100])
234      .range(["white", "red"]);
235  var contLinkScale = d3.scaleLinear()
236          .domain([2, 10000])
237          .range([1, 10]);
238
239
240  d3.json("/web_graph", function(error, graph){
241
242  var baseNodes = graph.nodes
243  var baseLinks = graph.links
244
245  console.log(graph.nodes)
246  console.log(graph.links)
247
248  var nodes = [...baseNodes]
249  var links = [...baseLinks]
```

Remplacer « /web_graph » par :

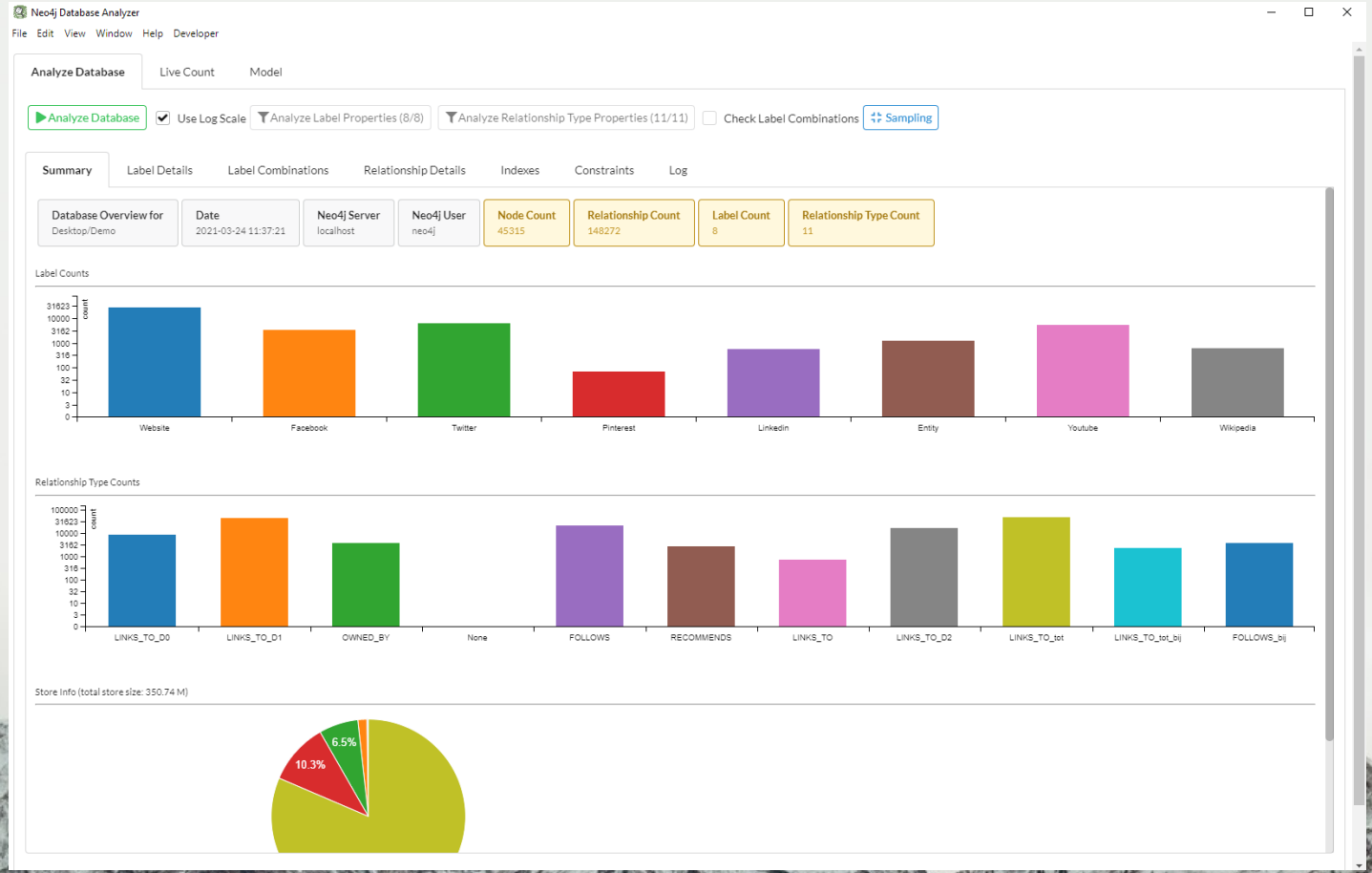
- « ent_graph » pour visualiser les entités
- « twit_graph » pour les comptes twitter
- « yt_graph » pour les comptes youtube

```
239
240  d3.json("/twit_graph", function(error, graph){
241
```



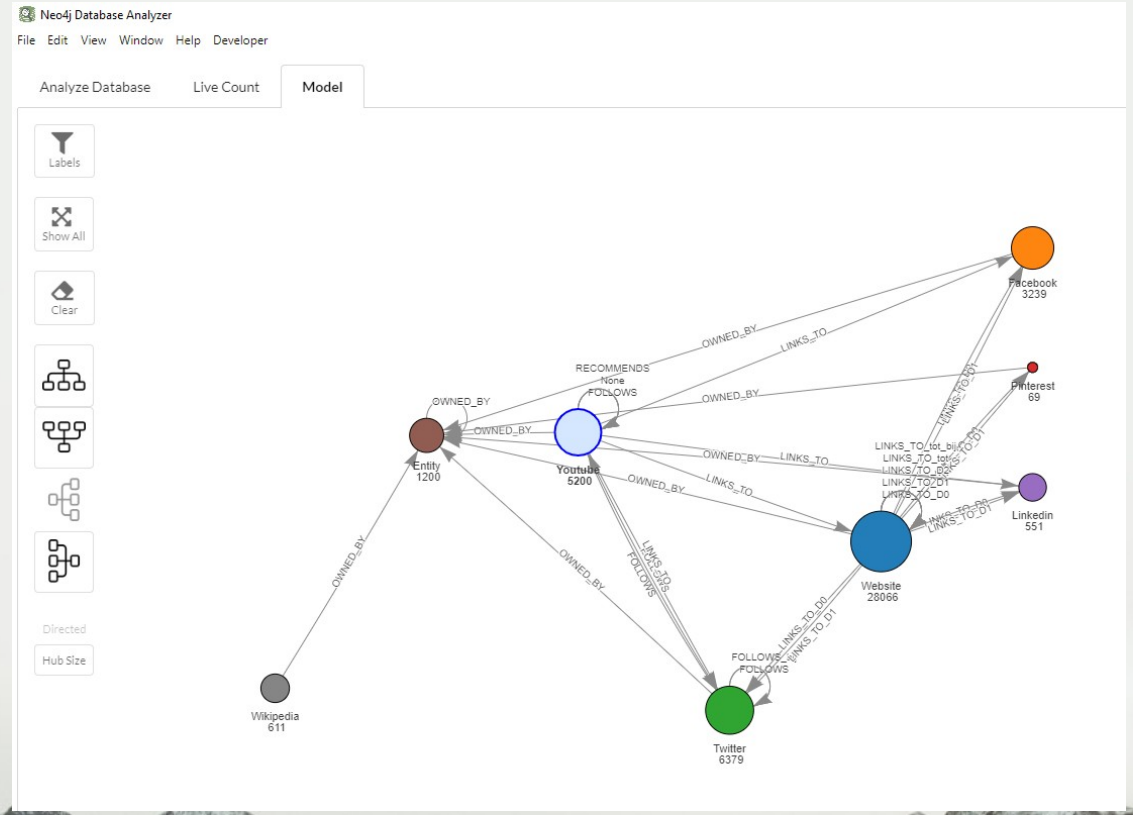
# 2. La base Neo4j

# Statistiques globales (database analyzer)

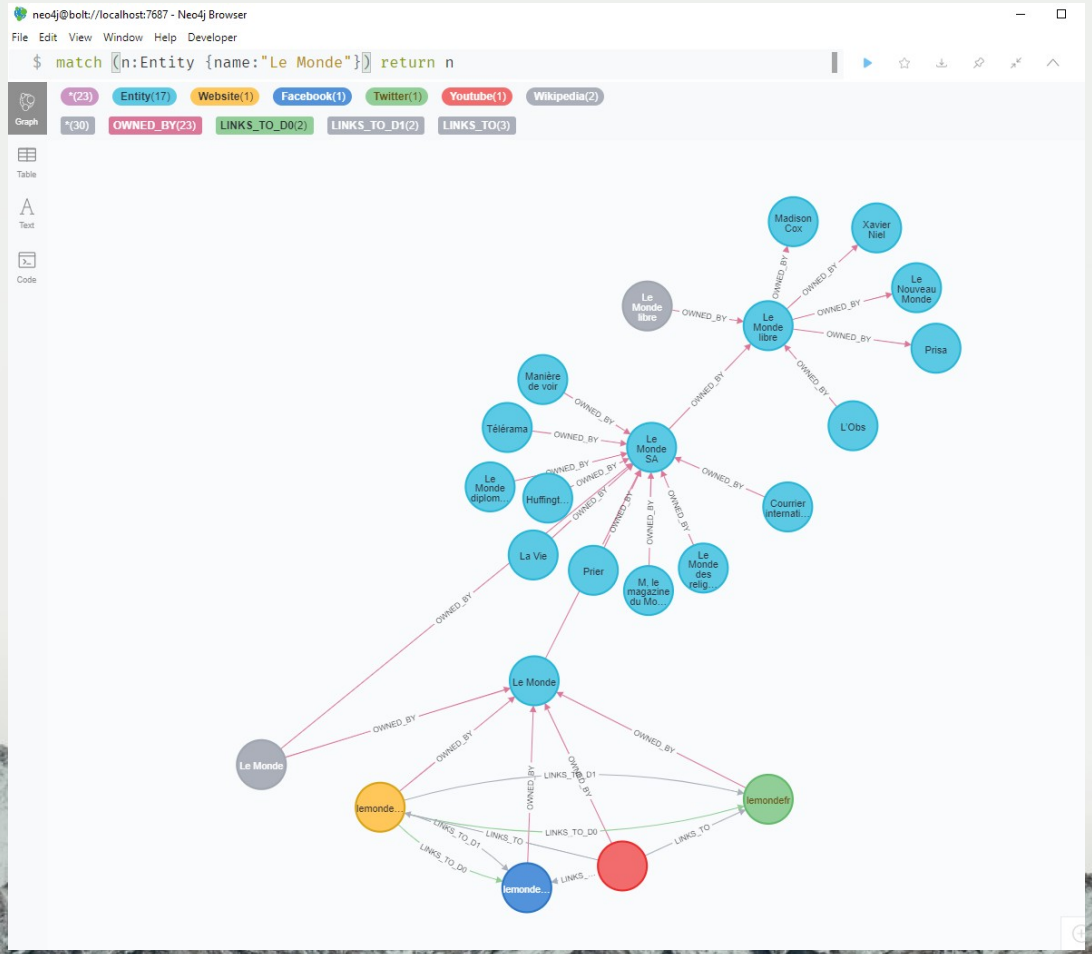



- 45 315 nœuds / 8 labels
- 148 272 relations / 11 types

# datamodel


Neo4j Database Analyzer
File Edit View Window Help Developer
Analyze Database
Live Count
Model
Labels
Show All
Clear
Directed
Hub Size
Facebook
3239
Pinterest
69
Linkedin
551
Website
28066
Youtube
5200
Entity
1200
Wikipedia
611
Twitter
6379
OWNED_BY
LINKS_TO
RECOMMENDS
None
FOLLOWS
LINKS_TO_D0
LINKS_TO_D1

# Exemple : Entity « Le Monde »




Dans la base, l'entité Le Monde possède

- Le site web lemonde.fr
- Le site web links les comptes :
  - Youtube : lemonde
  - Twitter : lemondefr
  - Facebook : lemonde.fr
- La page wikipedia Le Monde

L'entité Le Monde appartient à

- Le Monde SA
- Qui appartient à Le Monde libre
- Qui appartient à Xavier Nieal, Madison Cox, Prisa, Le Nouveau Monde...

# Exemple : le website lemonde.fr

- Links_to 304 nœuds

```
match (n:Website {name:"lemonde.fr"})-[r:LINKS_TO_tot]→(m) return count(m)
```

| count(m) |
| --- |
| 304 |

- Est linké par 270 nœuds

```
match (n:Website {name:"lemonde.fr"})←[r:LINKS_TO_tot]-(m) return count(m)
```

| count(m) |
| --- |
| 270 |

- Dont 44 de façon bijective

```
match (n:Website {name:"lemonde.fr"})←[r:LINKS_TO_tot_bij]-(m) return count(m)
```

| count(m) |
| --- |
| 44 |

- liste

```
match (n:Website {name:"lemonde.fr"})←[r:LINKS_TO_tot_bij]-(m) return (m.name)
```

| | (m.name) |
| --- | --- |
| 1 | "leparisien.fr" |
| 2 | "conspiracywatch.info" |
| 3 | "lavie.fr" |
| 4 | "lesinrocks.com" |
| 5 | "huffingtonpost.fr" |
| 6 | "dreuz.info" |

# 3. Le code pour remplir la base

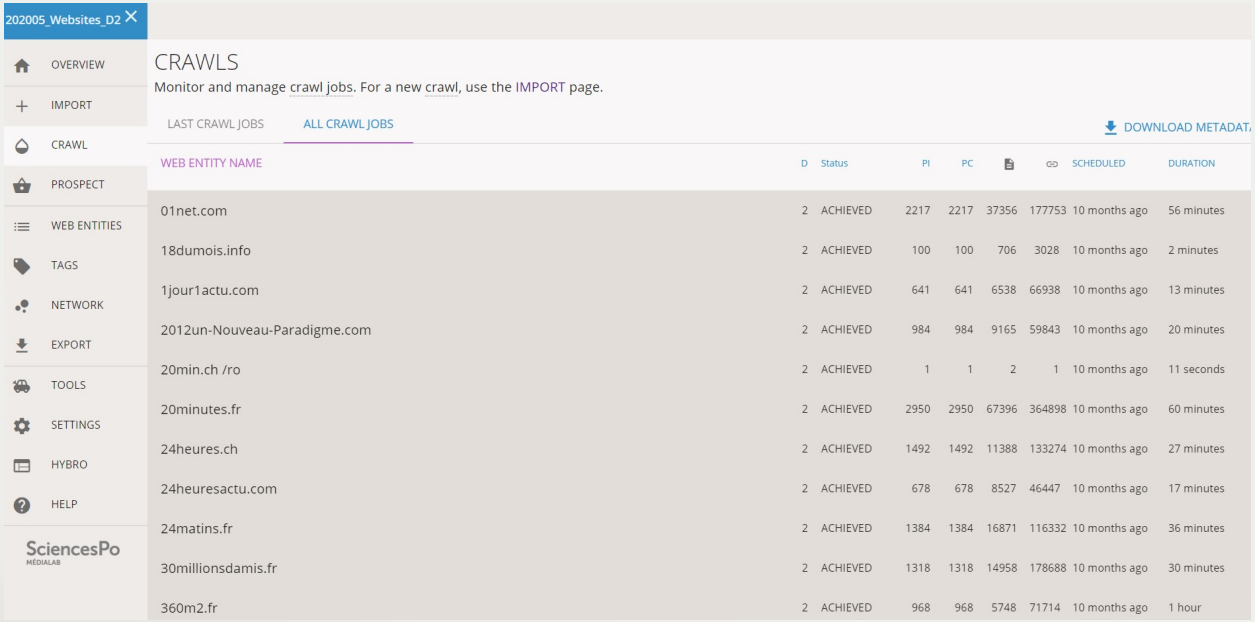
# Résumé : pour créer une database clean

**For clean fresh new database :**

1. run all cleanImportDB.py (carefull, one procedure by one, connection pb with py2neo.) : tested with D0 and D1 DISCO, missing some websites compared to MyMedia-FillDB\data\202007WebsitesRS_D0
2. run sandboximportRSformSiteList.py, all cells starting with A. This imports file firstPageRS_df_edit.csv
3. Twitter: in twitter.py
   - if not done, download twitter profiles with cell
   - imports all Twitter profile files (with cell)
   - use twecoll for follow network (using `python twecoll init -q NAME` (where NAME is NAME.twt file) + `python twecoll fetch NAME`)
   - import twecoll dat file with import_twecoll_dat
4. Youtube: in youtube.py
   - download yt profiles and subscriptions from cell
   - push profile info in db
   - import youtube recommendations from profile
   - download profile / subscriptions for newly added youtube nodes (from previous step)
   - again : push profile in db + import youtube reco from profile
   - download links from youtube
   - import links in db
   - import subscriptions info
5. wikipedia : wikipedia.py
   - search for wikipedia page corresponding to Entities + create wiki_search.csv
   - Manual QC / Edit of the csv
   - import edited file as wikinodes and owned_by connection
   - Download wikipages as files
   - import info in db : infobox, categories, wikidata ref
   - Import wikipedia summary
6. Whois
   - download whois files
   - TODO : parse and import !
7. CPPAP, Decodex, SPIIL, Mediaslocaux, DigitalAdTrust: classements.py
   - follow script to import as Website properties

+ **computeGDS.py** pour des calculs de data science sur graph (pagerank, centrality, community detection, clustering...)

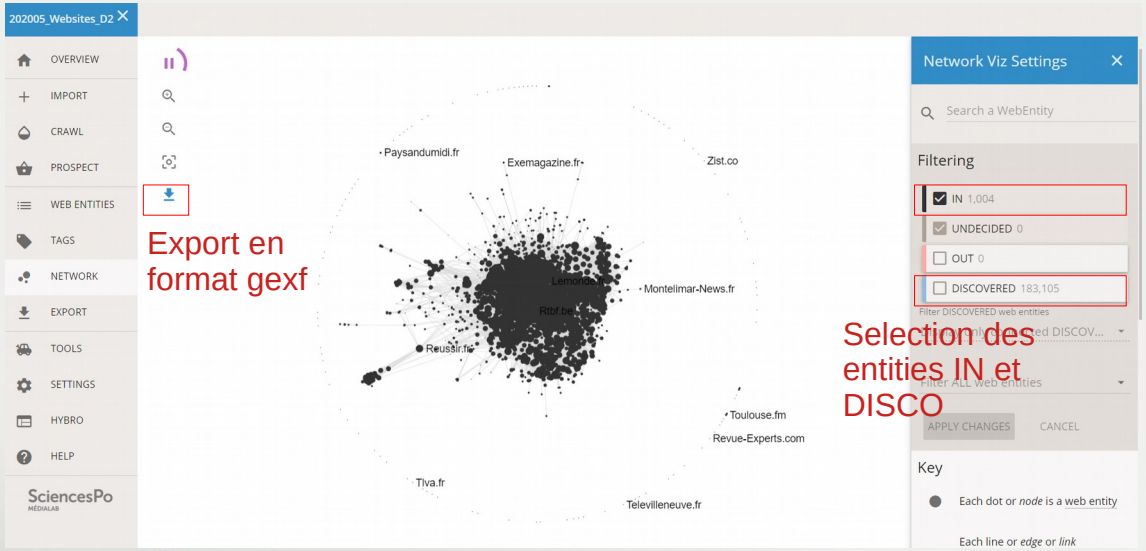
# Le links crawler : hyphe

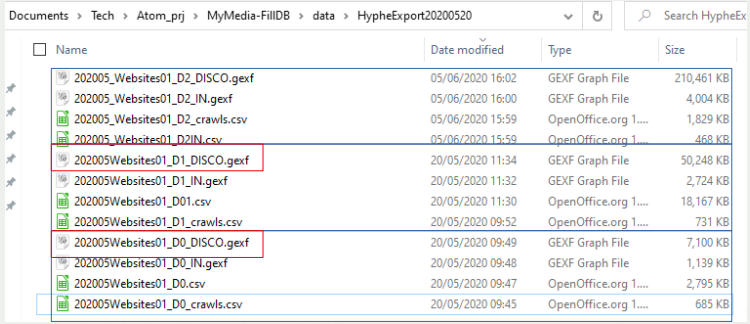



- De sciencepo medialab : site, github et demo
- Plateforme installé via docker en local, interface intuitive et robustesse des crawls
- Crawle une liste d'url et récupère tous les liens hypertextes. Possibilité de choisir la profondeur (0 pour les url, 1 pour les liens dans les url, 2 pour les liens des liens…)
- Itération pour une liste de 1000 sites de medias francophones

# Le links crawler : hyphe

Exports de hyphe (dans network, selection des entities IN et DISCO pour projet crawl depth 0 et crawl depth 1)




Depth 2

Depth 1

Depth 0

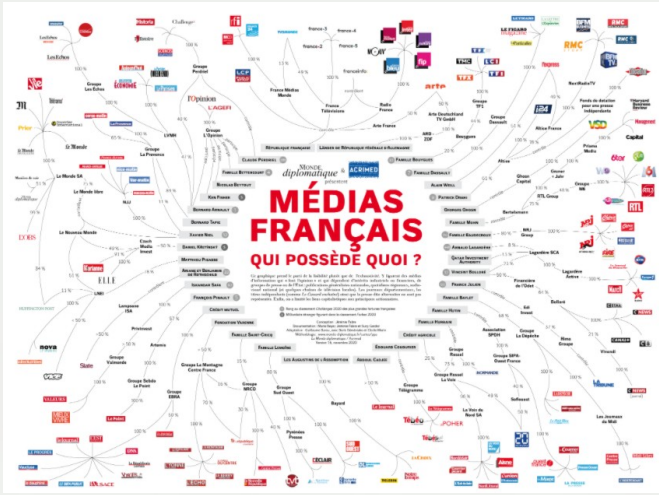
Les fichiers importés dans la base : D0_DISCO.gexf et D1_DISCO.gexf.

D2_DISCO mériterait d'être importé mais le code non optimisé met trop de temps à l'import...

# Import des données hyphe, monde diplo et ACPM

1. Le code utilisé est le début de **data/cleanImportDB.py**

- À partir des exports de Hyphe (D0_DISCO.gexf et D1_DISCO.gexf), création des nœuds (Website) et des liens LINKS_TO_D0 (depth 0) et LINKS_TO_D1 (depth D1)
- Import des données mondediplo disponible ici https://github.com/mdiplo/Medias_francais : création des entités et des relations OWNED_BY
- Import des données ACPM : statistiques des grands sites d'information française disponible ici : https://www.acpm.fr/Les-chiffres/Frequentation-internet/Sites-Grand-Public/Classement-unifie
  et ici : https://www.acpm.fr/Les-chiffres/Frequentation-internet/Sites-Pro/Classement-unifie
  importés comme propriété dans les noeuds



# Import des nœuds réseaux sociaux

| | site | fb | twit | yt_user | yt_url | sc | insta | dm | site_name | id | twit_hyphe | fb_hyphe | fb_final | twit_final | pi_hyphe |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | http://agri71.fr | agri71 | agri71 | | | | | | Agri71.fr | 34 | agri71 | agri71 | agri71 | agri71 | |
| 1 | http://axonais.fr | laxonaisofficiel | | | | | | | Axonais.fr | 73 | | laxonaisofficiel | laxonaisofficiel | | |
| 2 | http://bretons.bzh | bretons | magazinebretons | | | | | | Bretons.bzh | 98 | magazinebretons | | bretons | magazinebretons | |
| 3 | http://c4magazine.org | | | | | | | | C4magazine.org | 111 | | | | | |
| 4 | http://cameditsport.com | cameditsport | levegi1 | cameditsport.com | https://www.youtube.com/channel/UCrCQO-y5BgQqH5xVqo47LxQ | | | | Cameditsport.com | 117 | | | cameditsport | levegi1 | |
| 5 | http://chouard.org/blog/ | | | | | | | | Chouard.org /blog | 137 | | juan.branco.98 | | | |
| 6 | http://courrier-francais.com | | | | | | | | Courrier-Francais.com | 165 | | | | | |
| 7 | http://cqfd-journal.org | cqfd-mensuel-16714 | cqfdjournal | | | | | | Cqfd-Journal.org | 177 | cqfdjournal | cqfd-mensuel-16714783 | cqfd-mensuel-1671478326 | cqfdjournal | |
| 8 | http://creuse-agricole.com | creuseagricole | creuseagricole | | | | | | Creuse-Agricole.com | 186 | creuseagricole | creuseagricole | creuseagricole | creuseagricole | |
| 9 | http://darons.net | | parantsmagazine | | | | | | Darons.net | 188 | parantsmagazine | | | | |
| 10 | http://diktacratie.com | diktacratiecom | diktacratie | | | | | | Diktacratie.com | 187 | | | diktacratiecom | diktacratie | |
| 11 | http://entraid.com | entraid | journal_entraid | entraid.com | https://www.youtube.com/channel/UCtk948ORqO0eeEWtgNEAuQ | | | | Entraid.com | 219 | journal_entraid | entraid | entraid | journal_entraid | |
| 12 | http://factuel.info | factuelinfo | factuelinfo | | | | | | Factuel.info | 237 | factuelinfo | | factuelinfo | factuelinfo | |
| 13 | http://haute-loire-paysanne.fr | | | | | | | | Haute-Loire-Paysanne.fr | 305 | | | | | |
| 14 | http://ilfattoquotidiano.fr | | | | | | | | Ilfattoquotidiano.fr | 338 | estoyche | | | estoyche | |
| 15 | http://information-en-direct-france.com | | | | | | | | Information-En-Direct-France.com | 352 | | flashinfogiletsjaunes | | | mickaelmicka |
| 16 | http://kezako.unisciel.fr | unisciel | unisciel | unisciel | http://www.youtube.com/unisciel | | | | kezako.Unisciel.fr | 998 | unisciel | unisciel | unisciel | unisciel | |
| 17 | http://lagauchematuer.fr | la-gauche-ma-tuer | lagauchematuer | | | | | | Lagauchematuer.fr | 416 | lagauchematuer | | la-gauche-ma-tuer | lagauchematuer | |
| 18 | http://lalterego.fr | | | | https://www.youtube.com/channel/UCUXg6UxOBQYejlcSqZvg4wQ | | | | Lalterego.fr | 430 | lalteregofr | lalteregofr | lalteregofr | lalteregofr | |
| 19 | http://larenaissanceduloiretcher.fr | la-renaissance-du-loir-et-cher-111130649012373 | | | | | | | Larenaissanceduloiretcher.fr | 452 | | la-renaissance-du-loir-e | la-renaissance-du-loir-et-cher-111130649012373 | | |
| 20 | http://larevuedesmedias.ina.fr | larevuedesmedias | ina_revuemedias | | https://www.youtube.com/channel/UCD9xpWa_mjbbH0sskVjZxhg | | | | larevuedesmedias.Ina.fr | 340 | ina_revuemedias | larevuedesmedias | larevuedesmedias | ina_revuemedias | |
| 21 | http://latelelibre.fr | | | | | | | | Latelelibre.fr | 470 | | | | | |
| 22 | http://laterredecheznous.com | | | | | | | | Laterredecheznous.com | 471 | | | | | |
| 23 | http://le-crestois.fr | | | | | | | | Le-Crestois.fr | 487 | | journal.lecrestois | journal.lecrestois | | |
| 24 | http://leblogdenestor.com | leblogdenestor | marion_nestor | | https://www.youtube.com/channel/UCj0HNkgz8eikcMM81NA4CEA | | | | Leblogdenestor.com | 497 | leblogdenestor | leblogdenestor | leblogdenestor | leblogdenestor | |
| 25 | http://lechodelaboucle.fr | lécho-de-la-boucle | echodelaboucle | | | | | | Lechodelaboucle.fr | 503 | echodelaboucle | | lécho-de-la-boucle | echodelaboucle | |
| 26 | http://ledemocratevernonnais.fr | democratevernonnais | ldmocrate | | | | | | Ledemocratevernonnais.fr | 514 | | | democratevernonnais | ldmocrate | |
| 27 | http://ledr.fr | ledr.fr | enfantsdurhone | | | | | | Ledr.fr | 518 | | 525389480826266_339 | ledr.fr | enfantsdurhone | |

2. code **data/SanboximportRSfromSIteList.py**

Récupère les noms de comptes réseaux sociaux crawlés par hyphe

Créé un fichier pour QC et edition

Importe les comptes dans Neo4j et créés liens

# Import des données twitter

3. code **data/twitter.py**

Connection à l'API twitter (créer un compte dev et utiliser vos api keys pour le moment : remplacer les xxxxx en début de code)

Si non dispo, télécharge les infos de profile les noeuds twitter de la base comme des propriétés

Importe dans Neo4j

Puis utiliser twecol (code ici : https://github.com/jdevoo/twecoll ) pour télécharger les follow network

Le code importe les dat files de twecol dans Neo4j (relationship FOLLOWS)

# Import des données youtube

- Code **data/youtube.py**

  Connection à l ‘API youtube (se créer un compte dev sur google et utilisez vos propres clefs quand demandées)

  Pour les comptes youtube, télécharge les infos de profil et subscriptions et les importe dans Neo4j

# Import des données wikipedia

- Code **data/wikipedia.py**
- Cherche et importe les infos wikipedia sur les entités de la base Neo4j
- TODO : importer touts les infos wikidata utiles

# Autres données

- Whois : download les infos whois dans des fichiers. Non importé dans Neo4j (TODO : a parser) → adresse des serveurs, personnes responsables…
- Classemen.py : import des groupes de crédibilité