Atelier données - Cour des Comptes

Moissonner des données en ligne

quelques éléments méthodologiques et pratiques avant de "scraper" des sites...

intervenant : Julien Paris (jparis.py@gmail.com)

durée de la présentation : 15 minutes

10/10/2018

Entrepreneur.e d'Intérêt Général

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE LA COHÉSION DES TERRITOIRES

cget
Commissariat général à l'égalité des territoires

PROBLÉMATIQUE GÉNÉRALE :
Récupérer des données déjà publiées sur Internet...

quoi?

- une donnée c'est quoi ?
- le modèle de données
- le "scraping"

pourquoi?

- cas d'usages
- problèmes récurrents

comment?

- problèmes récurrents
- les outils de scraping
- présentation d'Open Scraper

quoi?

Quelques définitions

Données, modèle de donnée, "scraper", consolidation...

Une donnée c'est quoi ?

" ce qui est **connu** et qui sert de **point de départ à un raisonnement** ayant pour objet la détermination d'une **solution à un problème** en relation avec cette donnée"

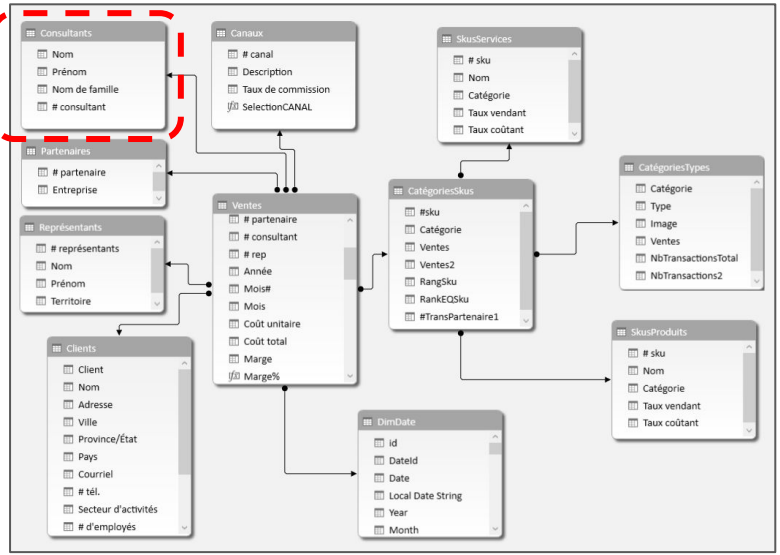
Wikipédia : https://fr.wikipedia.org/wiki/Donn%C3%A9e

Le modèle de donnée : la structure de ce que l'on sait

Modèle de données tabulaires

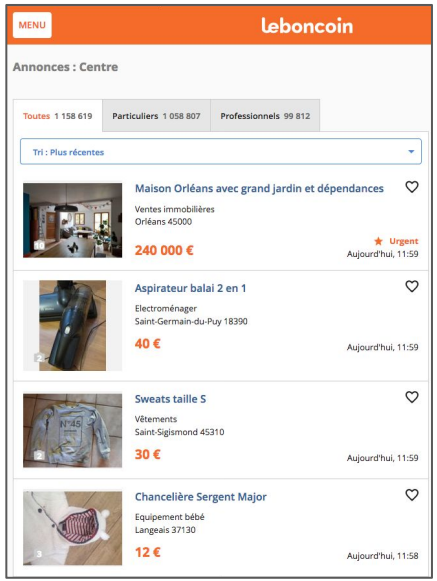
	A	B	C	D	E
1	Customer	City	Region	Product	Quantity
2	Orange	Big Town	West	Milk Chocolate	125
3	Red	Big Town	West	Dark Chocolate	210
4	Pink	Medium Town	East	Milk Chocolate	145
5	Grey	Big Town	West	Chocolate Hazelnut	21
6	Blue	Small Town	South	Dark Chocolate	50
7	Dark	Big Town	West	Chocolate Hazelnut	65
8	White	Big Town	West	Milk Chocolate	40
9	Green	Village	South	Chocolate Hazelnut	122
10	Yellow	Medium Town	East	Dark Chocolate	60
11	Silver	Medium Town	East	Extra Dark Chocolate	30
12	Gold	Medium Town	East	Chocolate Hazelnut	56

Modèle de données relationnelles



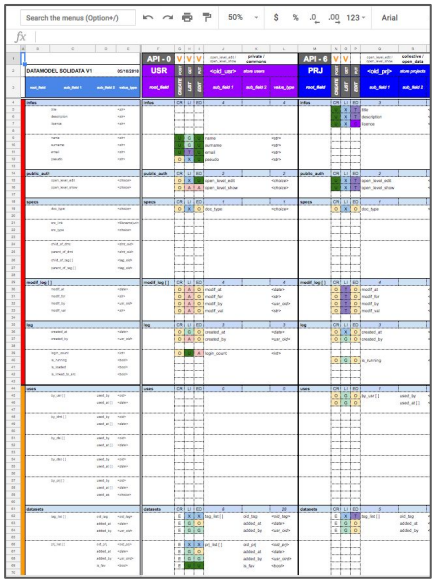
Récupérer une donnée en ligne : quelles formes de données ?

Site web



HTML

Tableur
csv, xls, tsv, ...

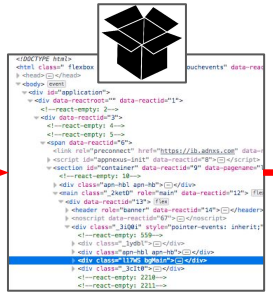


JSON
(base de données, API, etc...)

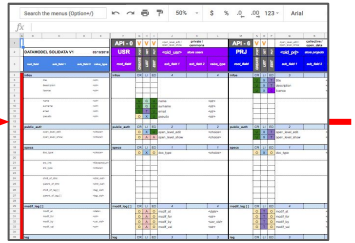
"Scraping" : moissonner de la donnée publiée en ligne

Site web

"black box"
(automatisation du
moissonnage :
"scraper")



lire le HTML



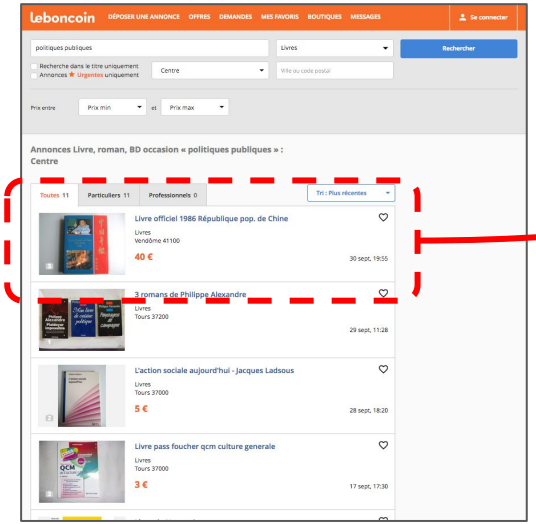
Tableur
csv, xls, tsv, ...

Base de donnée
(ex MongoDB)

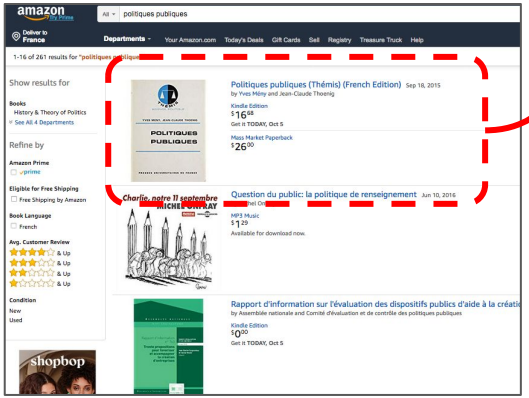
autres
réutilisations

"Scraping" : moissonner de la donnée publiée en ligne

Site web A
(modèle de données A)



Site web B
(modèle de données B)



Données moissonnées
(modèle de données C)

source	titre	auteur	prix
Bon coin	---	---	10€
Bon coin	---	---	12€
Amazon	---	---	16€
Amazon	---	---	9€

pourquoi?

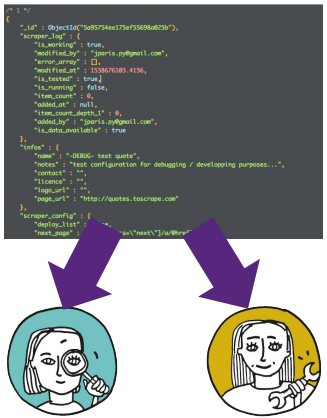
Quelques cas d'usages

A quoi et à qui ça peut servir de moissonner des données ?

Les familles d’utilisations

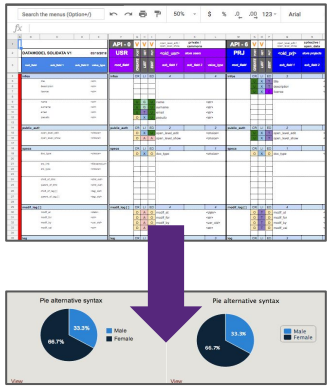
STOCKER LA DONNÉE

rendre disponible et requêtable
la donnée moissonnée



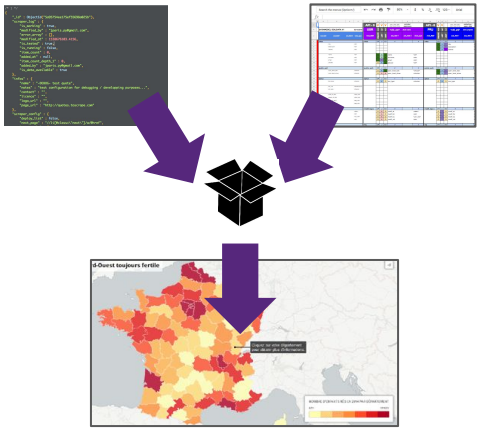
ANALYSER LA DONNÉE

transformer la donnée en
information intelligible



TRAVAILLER LA DONNÉE

enrichir la donnée en la croisant
avec d'autres jeux de données



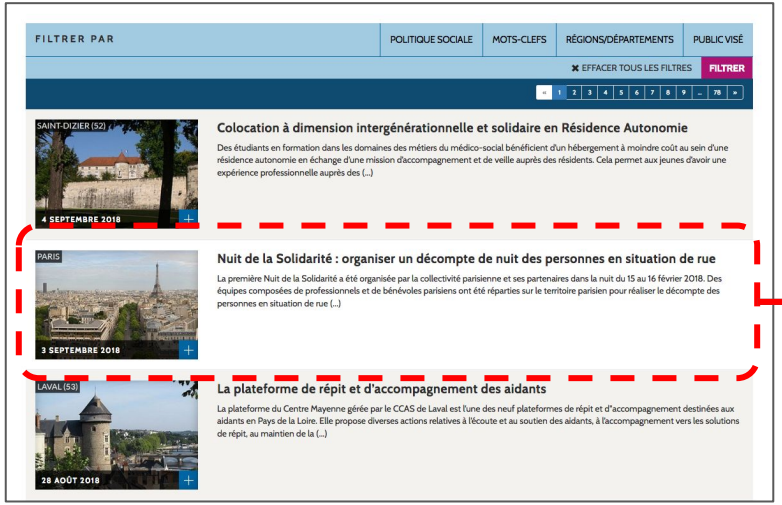
comment?

Les problèmes récurrents

Les principales familles de problèmes techniques et méthodologiques

Automatiser le "copier-coller"

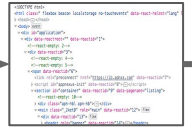
Site web à scraper



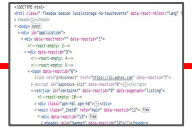
copier/coller

clic droit + "examiner l'élément"

HTML

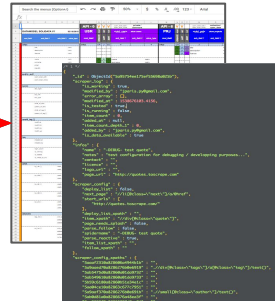


HTML

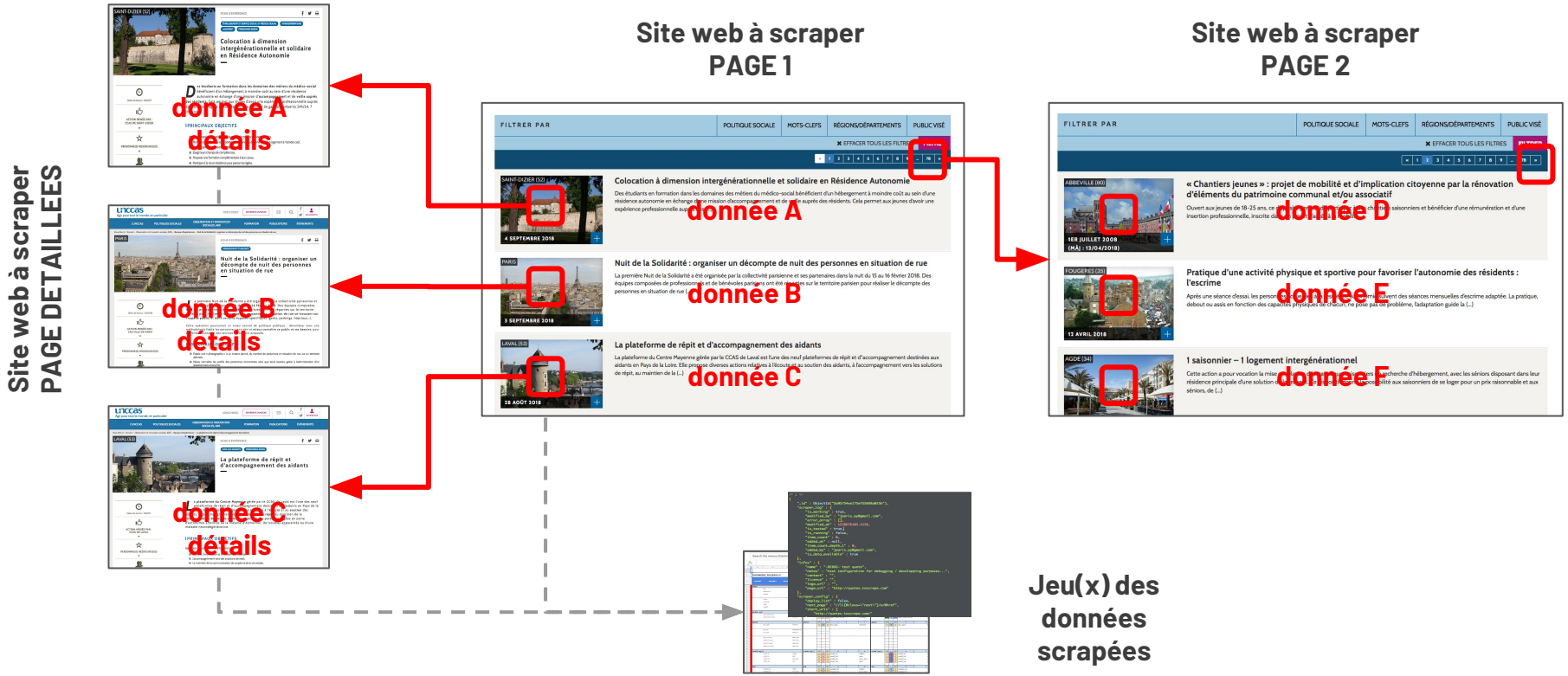


scraper

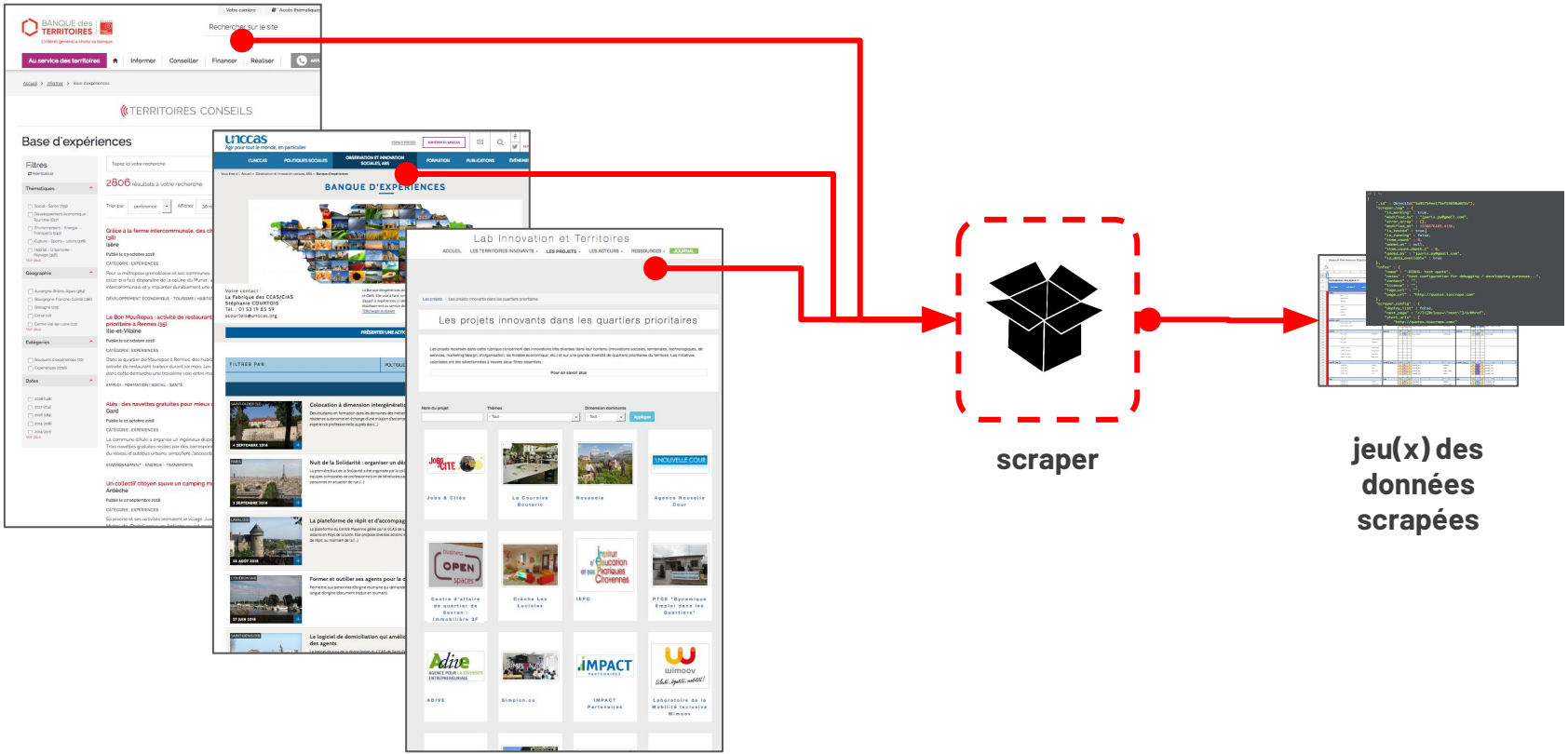
Jeu(x) des données scrapées



Naviguer dans un site peut être compliqué à automatiser



Scraper plusieurs sites autour d'un même modèle de données



comment?

Open Scraper

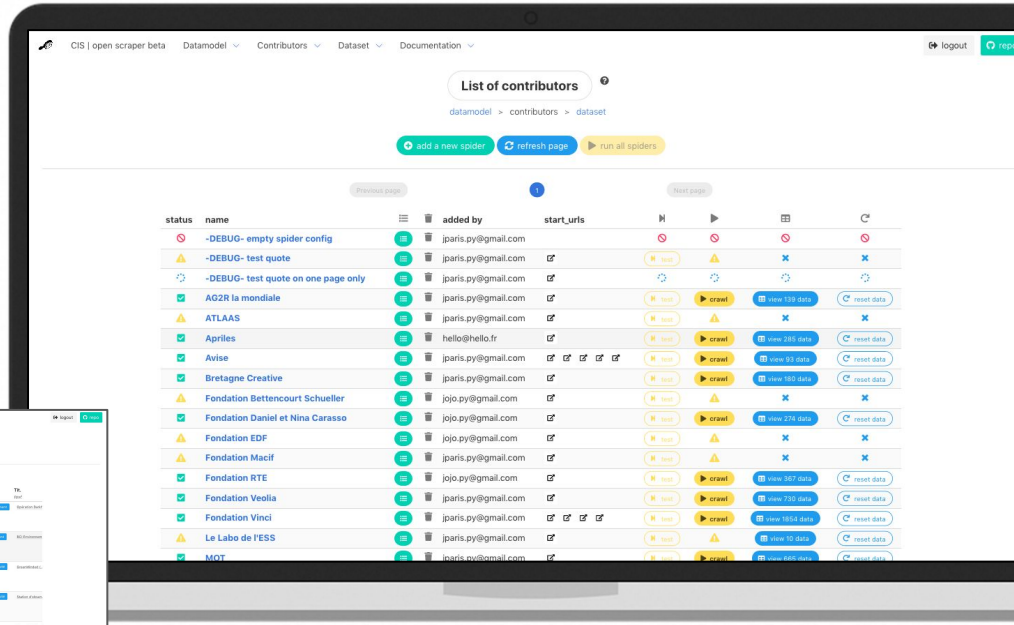
un web applications open source
pour moissonner des données sur Internet

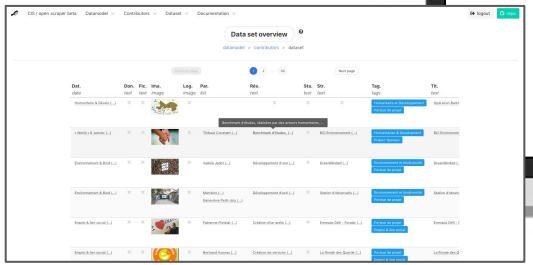
OPEN SCRAPER : agréger des données publiées sur des sites

www.cis-openscraper.com

Problème à résoudre

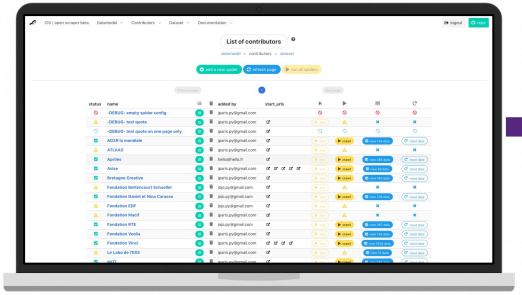
- Récupérer / **scraper** les données publiées sur les sites des partenaires du collectif
- **Homogénéiser** des données disparates par essence
- **Ouvrir** les données en maîtrisant leur degré d'ouverture : opendata / commons / collective / private





Une suite logicielle type ETL en développement continu et en open source :

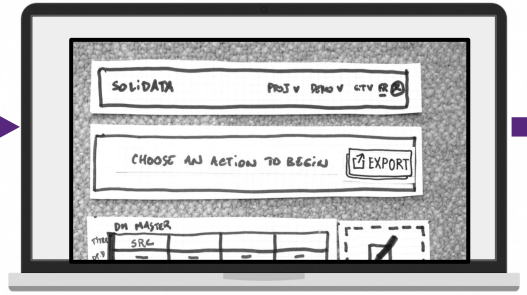
TADATA!



Open Scraper

agrégation de données publiées sur des sites web

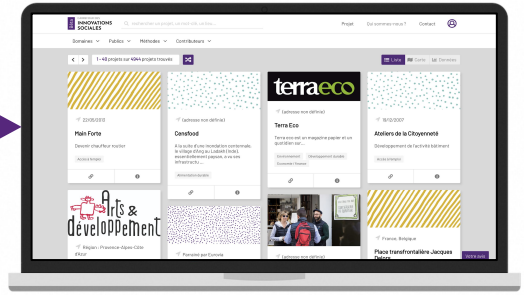
www.cis-openscraper.com



Solidata

agrégation, consolidation et enrichissement des données (en cours de conception)

backend API + front
(en développement)

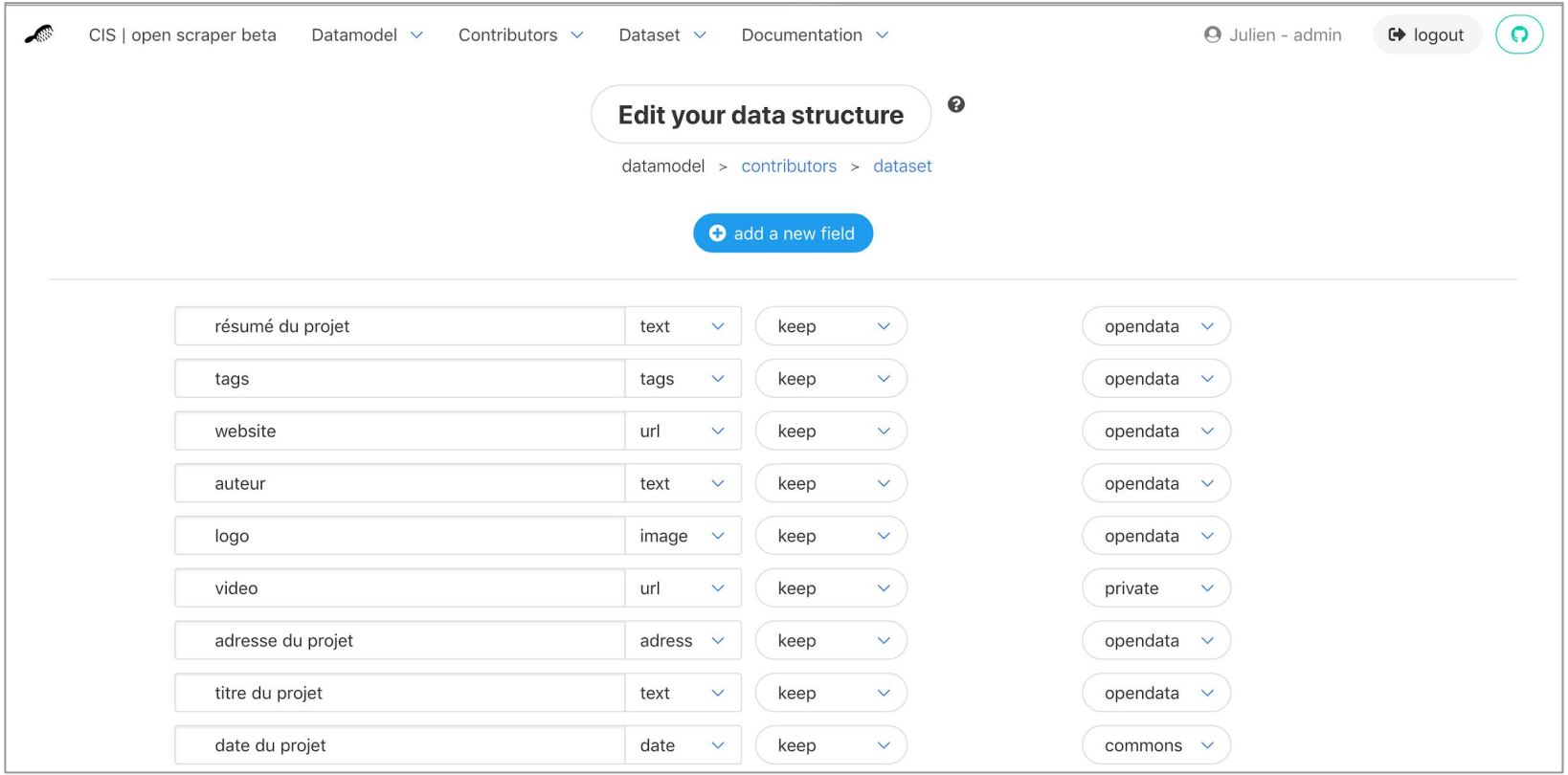


Apiviz

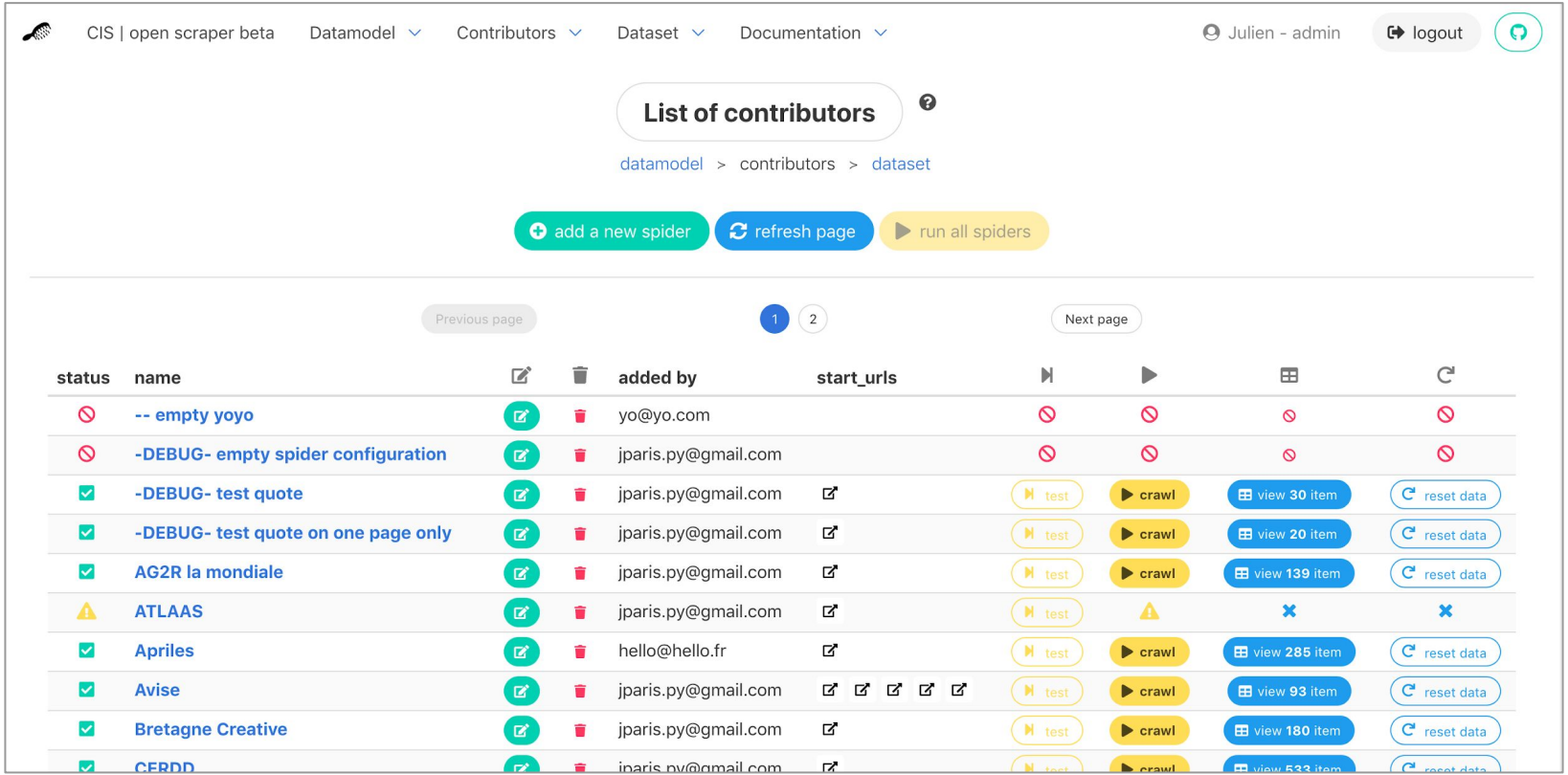
moteur de recherche et visualisation des flux de données

www.carrefourdesinnovationssociales.fr

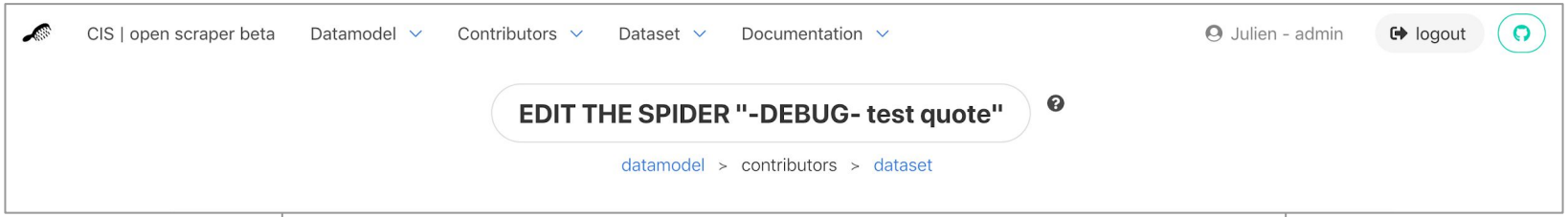
OPEN SCRAPER : définir un modèle de données

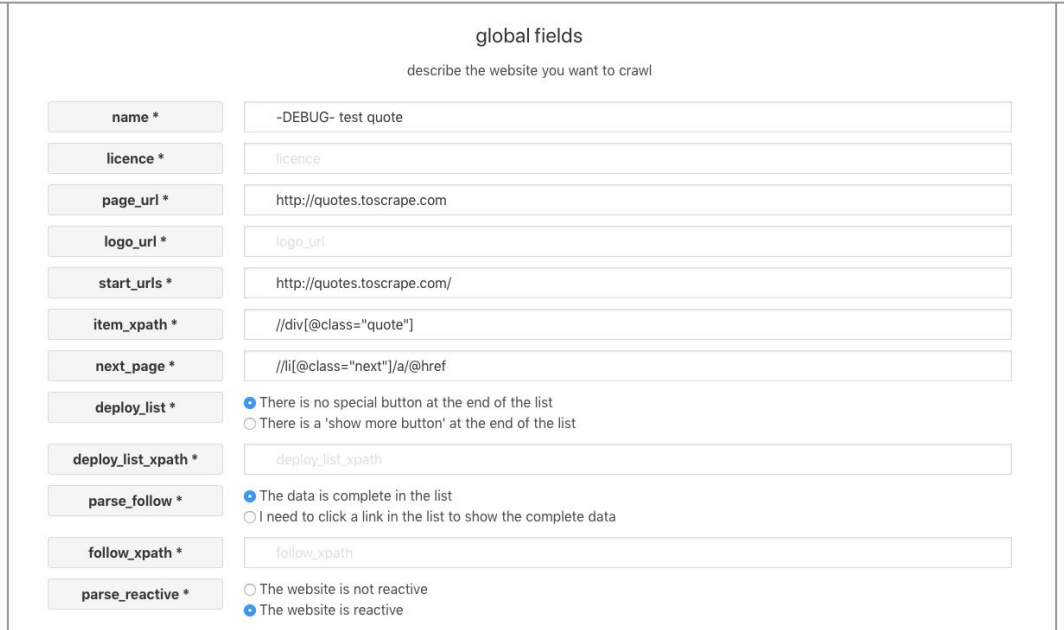


OPEN SCRAPER : ajouter un site à scraper

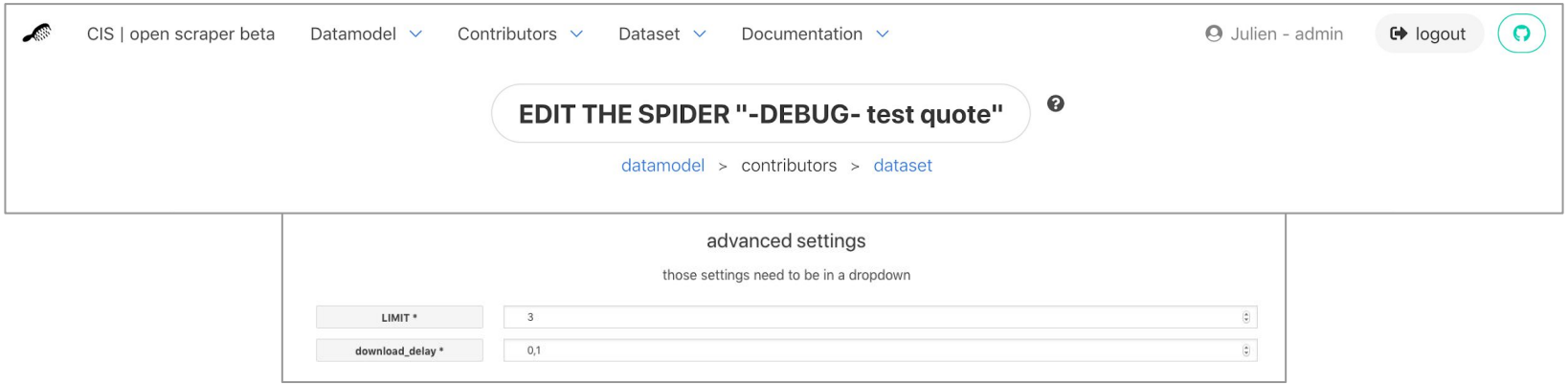


OPEN SCRAPER : éditer un scraper (XPATH)

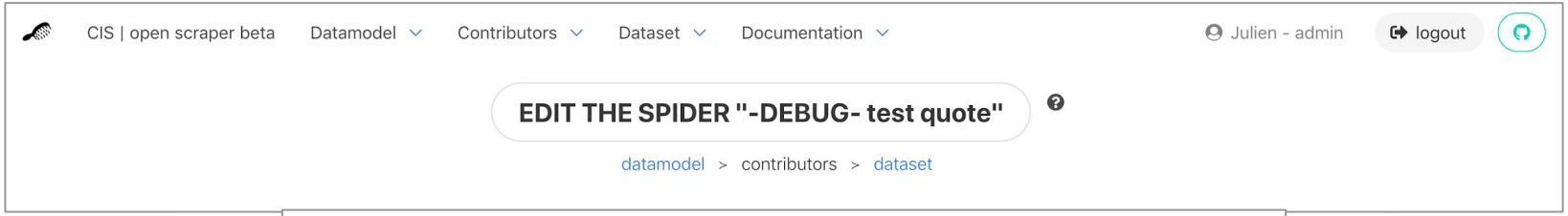


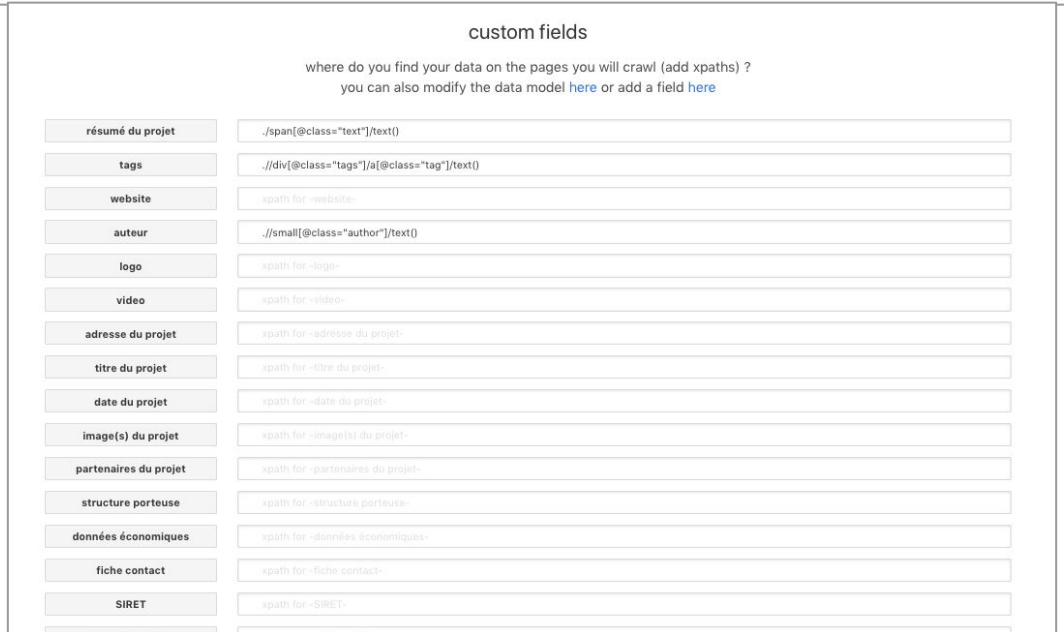


OPEN SCRAPER : éditer un scraper (XPATH)

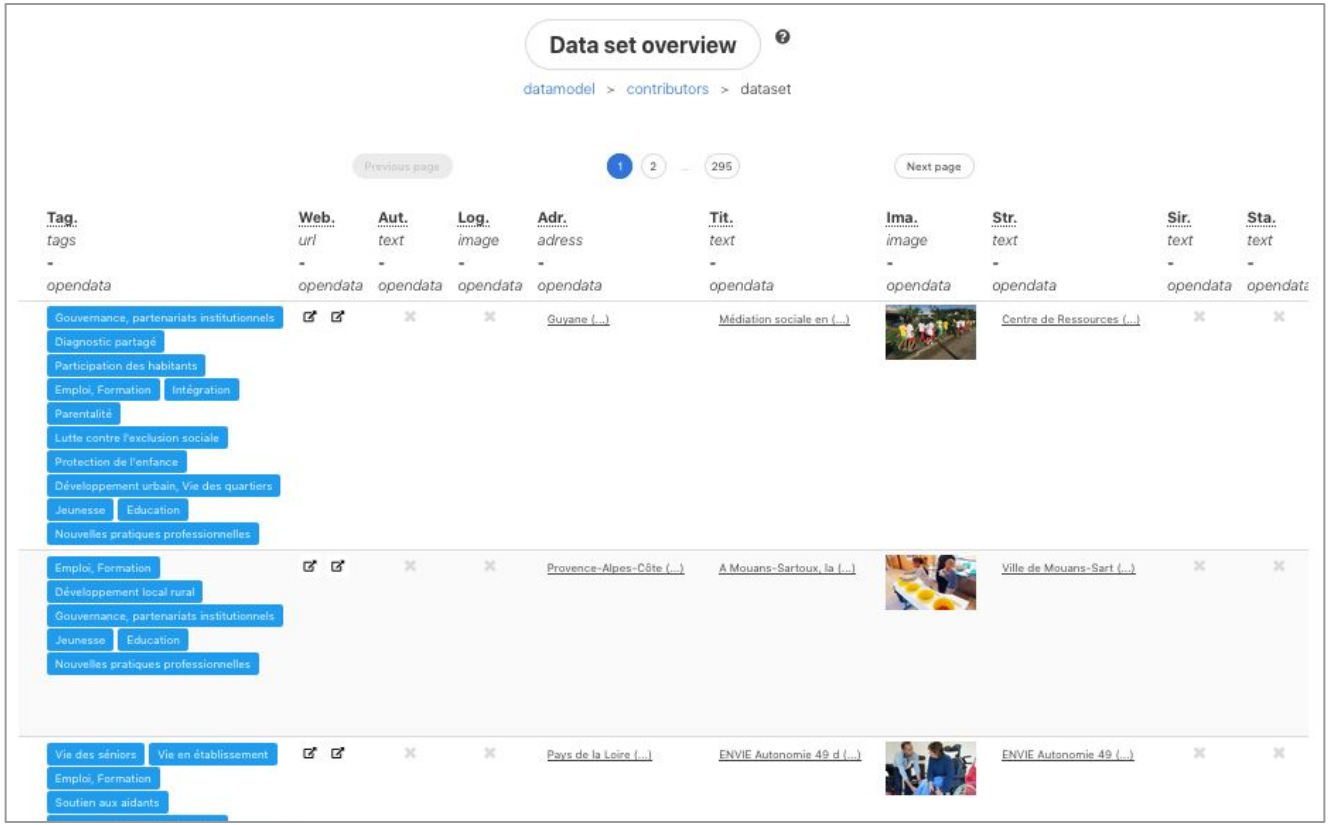


OPEN SCRAPER : éditer un scraper (XPATH)

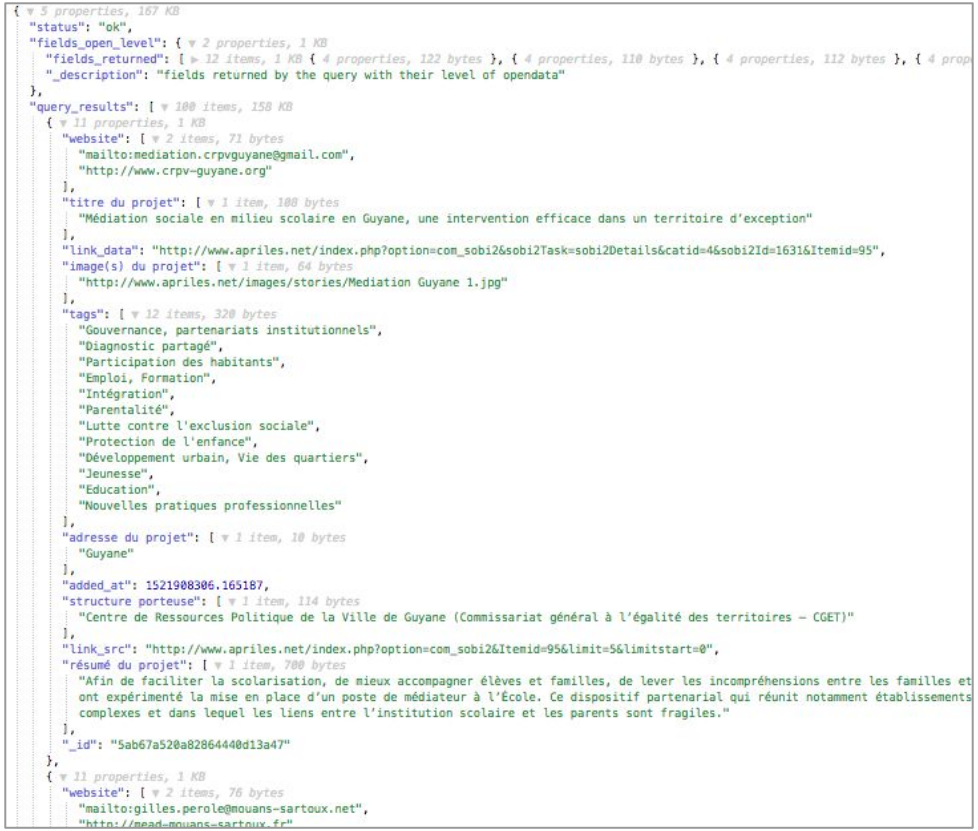




OPEN SCRAPER : prévisualiser le jeu de données "scrapées"



OPEN SCRAPER : ouvrir les données via l'API



Merci !

Association de préfiguration du Carrefour des innovations sociales

Yannick Blanc, Président : blanc.yannick@gmail.com
Emmanuel Dupont, Vice-Président : emmanuel.dupont@cget.gouv.fr

Équipe projet

Elise Lalique, designer UX-UI / CGET / EIG : elise.lalique@cget.gouv.fr
Bénédicte Pachod, coordinatrice / CGET : benedicte.pachod@cget.gouv.fr
Julien Paris, développeur / CGET / EIG : julien.paris@cget.gouv.fr