Atelier données - Cour des Comptes

Moissonner des données en ligne

quelques éléments méthodologiques et un outil open source pour "scraper" des sites sur internet ...

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

durée de la présentation : 20 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

pourquoi?

- familles de cas d'usages

comment?

- principes de base du "scraping"
- problèmes récurrents
- les outils de scraping
- présentation d'Open Scraper
- démo

quoi?

Quelques définitions

Donnée, modèle de donnée, "scraper", agrégation, 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"

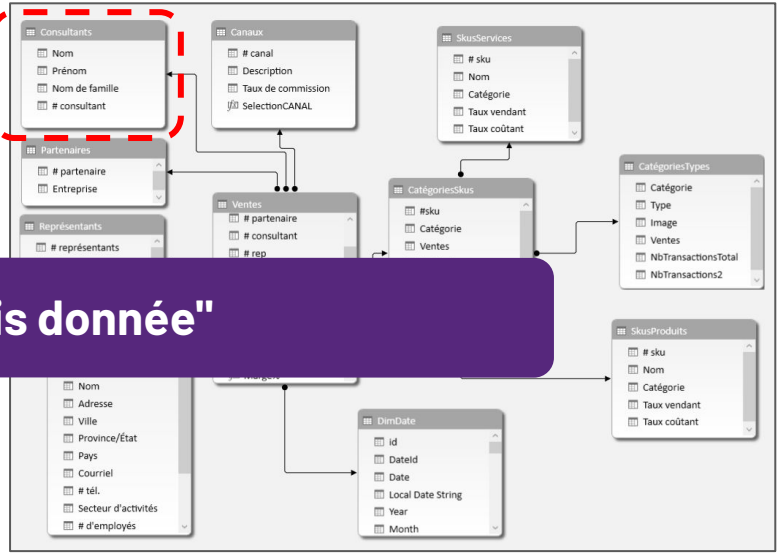
Wikipédia : https://fr.wikipedia.org/wiki/Donn%C3%A9e

Le modèle de donnée : structurer ce qu'on connaît

Modèle de données
tabulaire

	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	B			
8	White	B			
9	Green	W			
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
relationnel (clés)

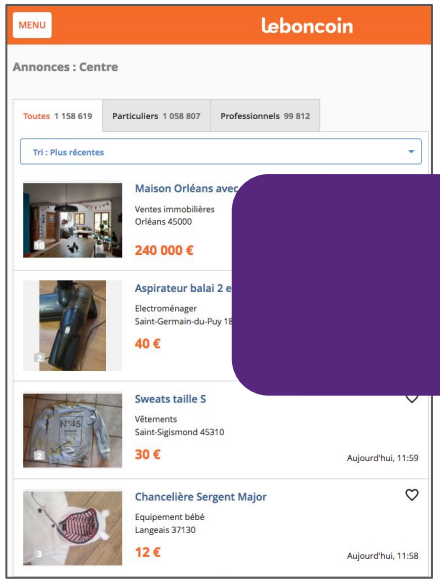


"la donnée n'est jamais donnée"

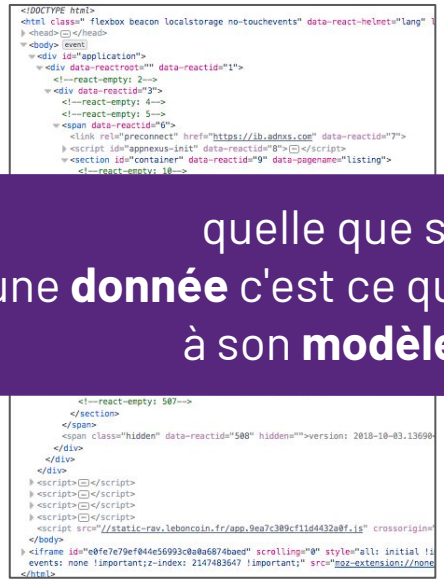
Note : ceci est un exemple, il existe encore beaucoup d'autres modèles de données (graphe, NoSQL, hiérarchique ...)

La donnée en ligne : quelles formes de données ?

site web

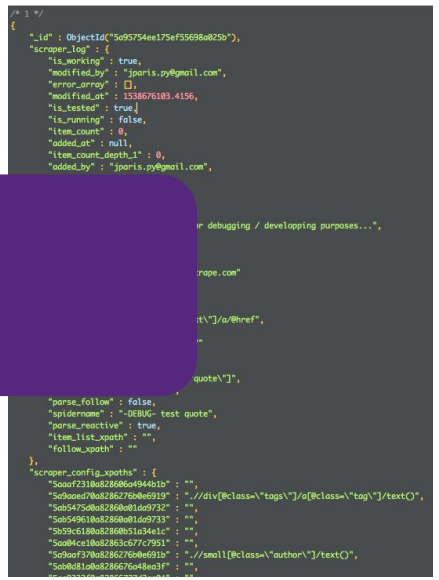


HTML



tableur
csv, xls, tsv, ...

JSON, GeoJSON...
(base de données, API, etc...)



quelle que soit sa forme
une **donnée** c'est ce qu'on cherche à **intégrer**
à son **modèle de données**

pourquoi?

Quelques cas d'usages

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

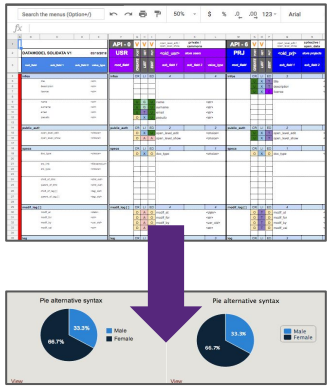
Pourquoi faire du *scraping* de données ?

POUR LES STOCKER

rendre disponible et requêtable la donnée moissonnée

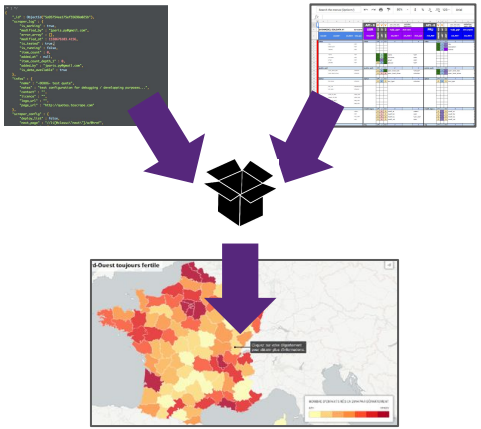
POUR LES ANALYSER

transformer la donnée en information intelligible



POUR LES TRAVAILLER

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



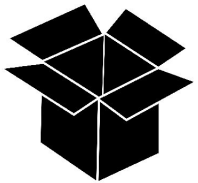
comment?

Principes de base du "scraping" de sites internet

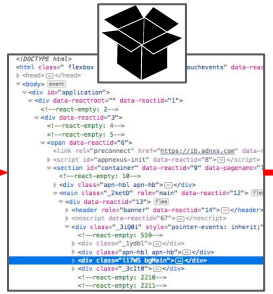
Ce que veut dire en pratique "moissonner de la donnée en ligne"

"Scraping" : moissonner de la donnée publiée en ligne (1/2)

site web
à moissonner



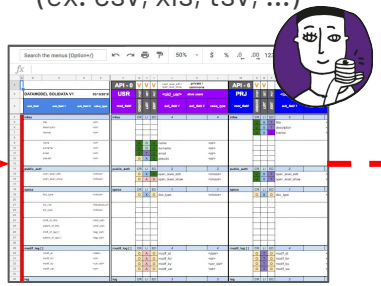
"black box"
automatisant le
moissonnage :
"**scraper**"



lire le HTML

copier/coller
la donnée

vers un tableur
(ex. csv, xls, tsv, ...)

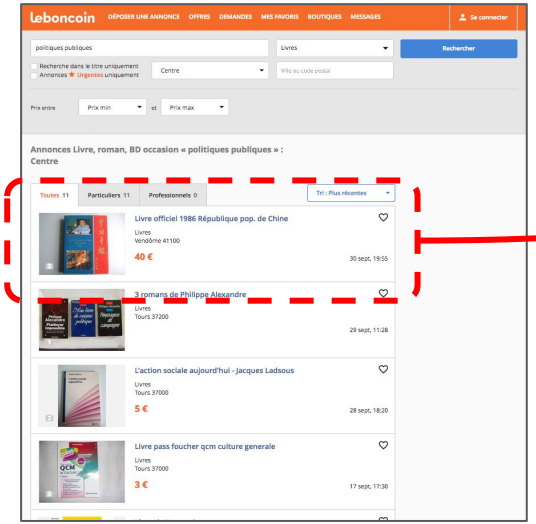


vers une base de donnée
(ex. MongoDB)

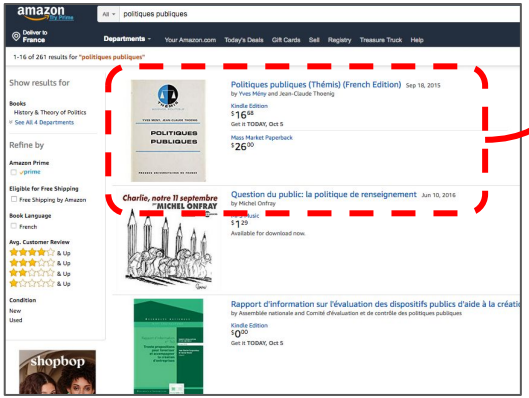
autres
réutilisations

"Scraping" : moissonner de la donnée publiée en ligne (2/2)

site web A
(modèle de données A)



site web B
(modèle de données B)



données moissonnées, structurées et agrégées
(modèle de données C)

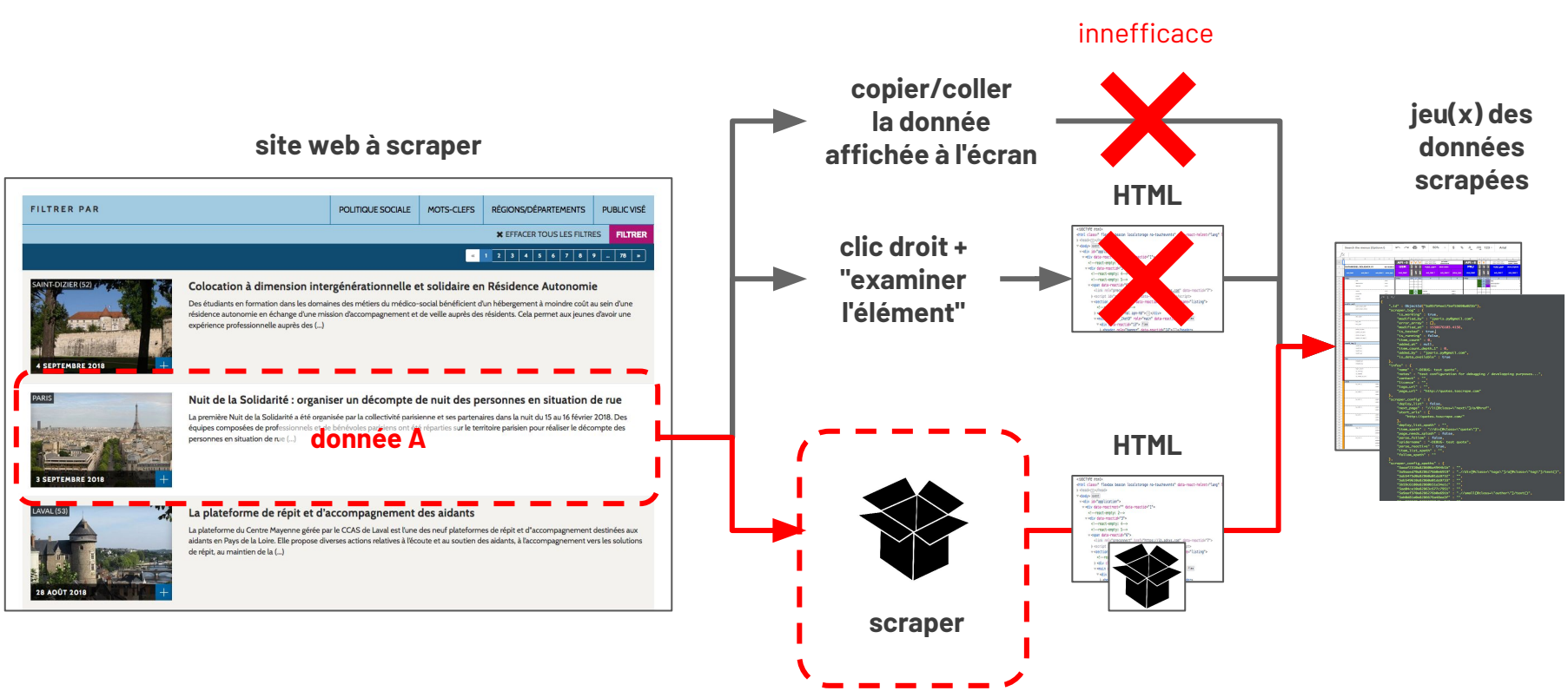
source	titre	auteur	prix
Bon coin	---	---	10€
Bon coin	---	---	12€
Amazon	---	---	16€
Amazon	---	---	9€

comment?

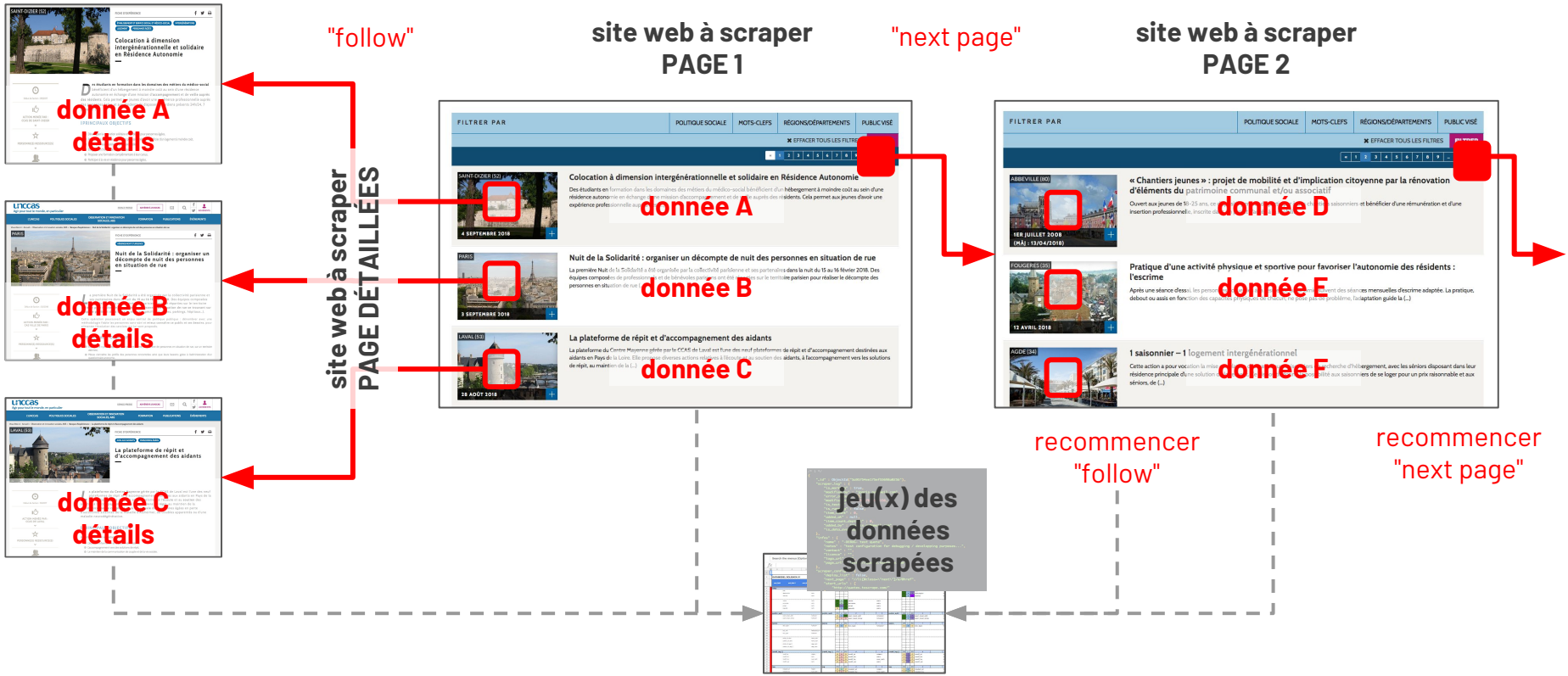
Les problèmes récurrents

Les principales familles de problèmes techniques, méthodologiques et humains

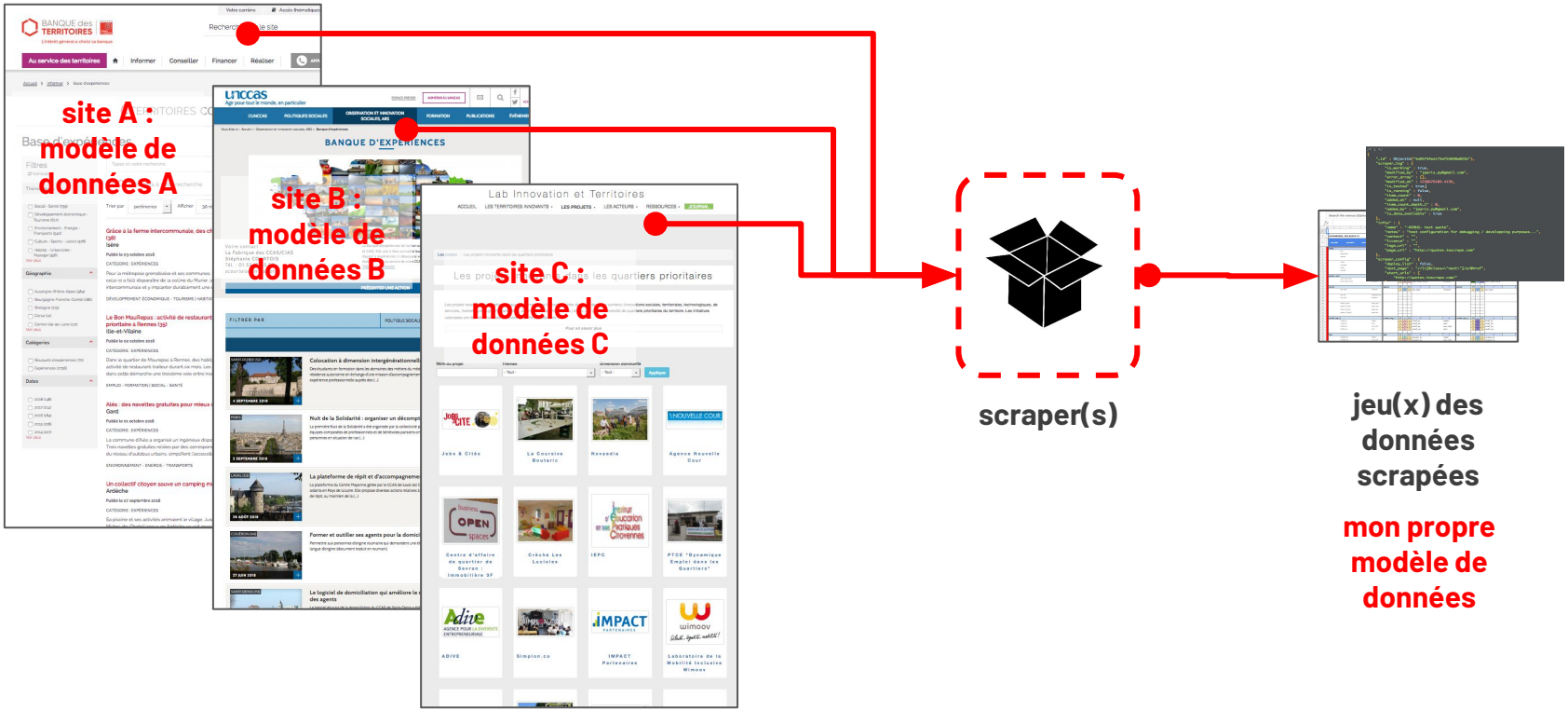
Problème 1 : **Automatiser le "copier-coller"**



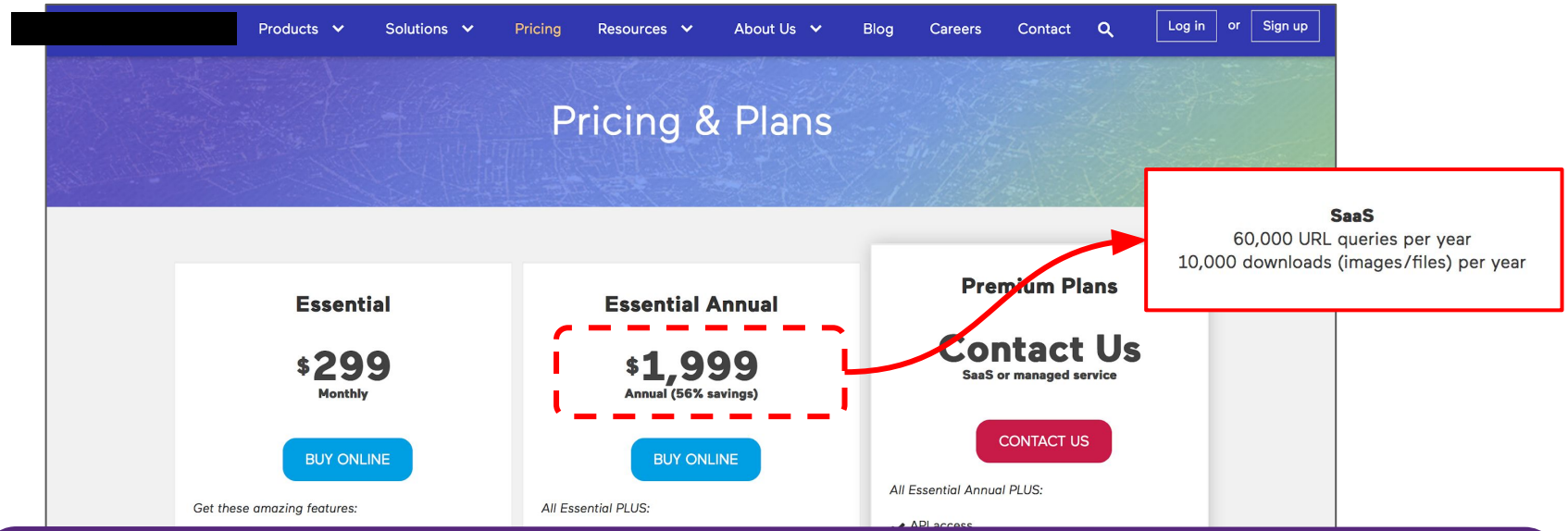
Problème 2 : **Automatiser la navigation dans un site**



Problème 3 : **Adapter le *scraping* à des sites différents pour agréger autour d'un seul modèle de données**

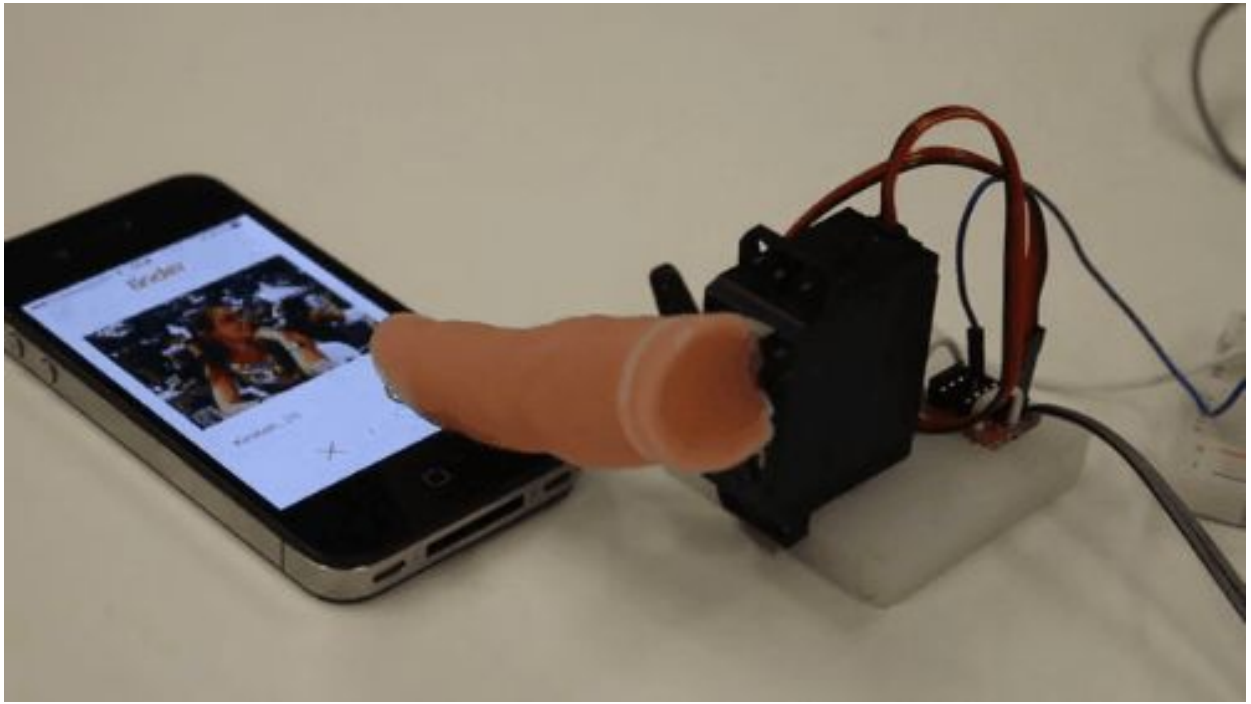


Problème 4 : **Avoir les moyens financiers ou humains…**



exemple de service commercial de *scraping* en SaaS :
prix annuel pour utiliser un seul *scraper*,
c'est-à-dire un site à scraper autour d'un modèle de données

Bref, un *scraper* idéal (efficace, adaptable, pas cher) c'est une version améliorée d'un outil du genre...



comment?

Open Scraper (version 1.0)

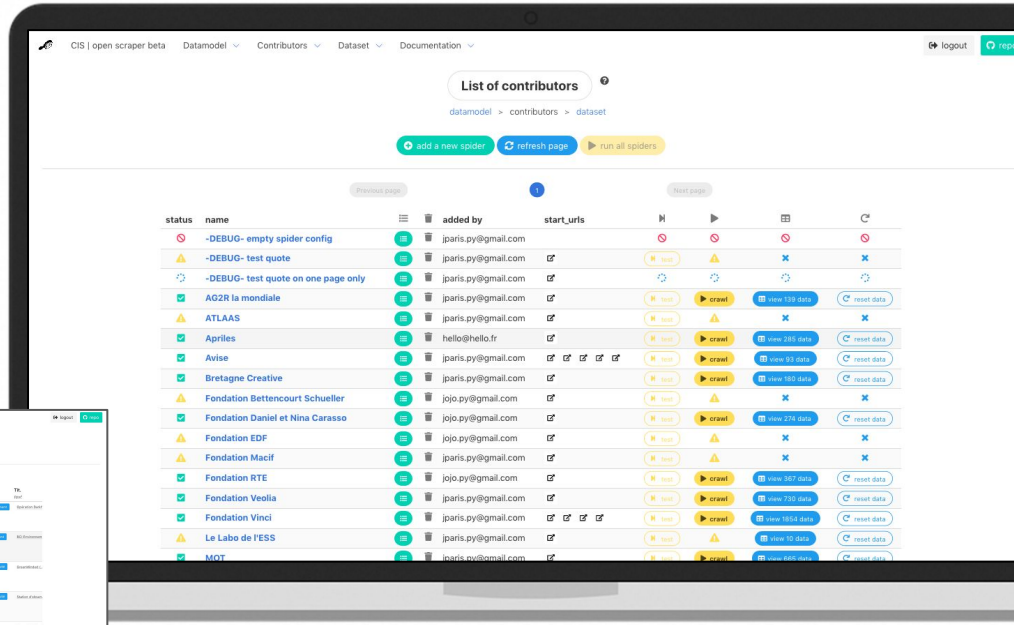
un web application **open source** en développement continu
pour moissonner des données sur différents sites Internet

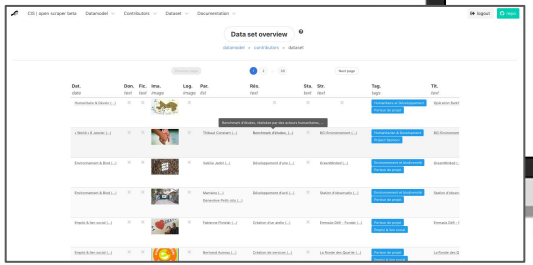
OPEN SCRAPER : agréger des données publiées sur des sites

www.cis-openscraper.com

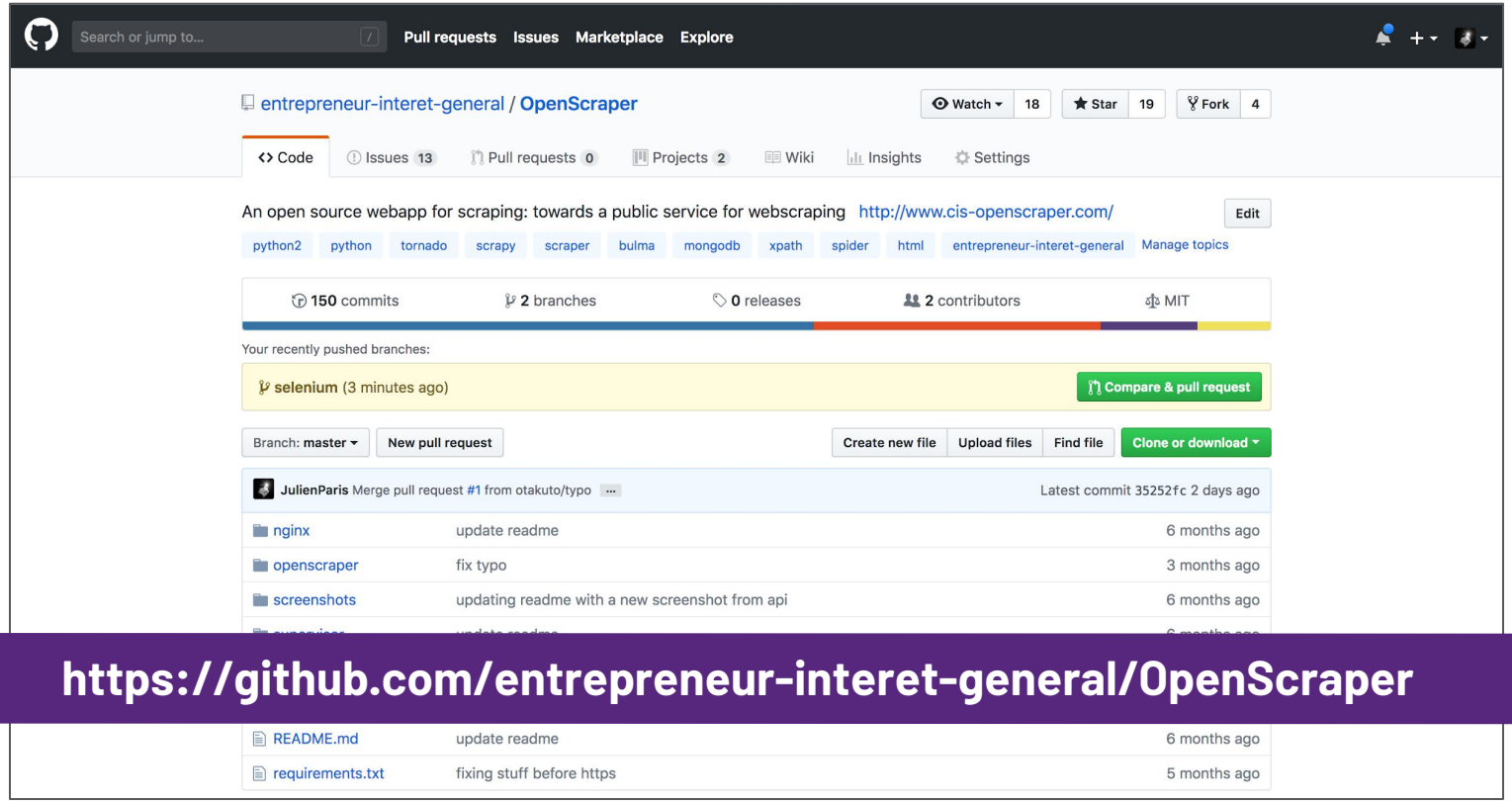
Problème à résoudre

- 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





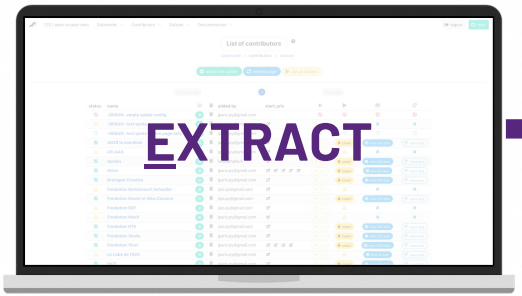
OPEN SCRAPER : une webapp de scraping en *open source*

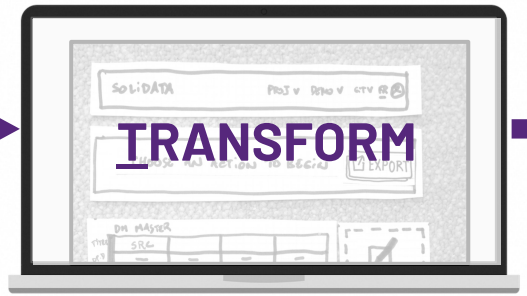


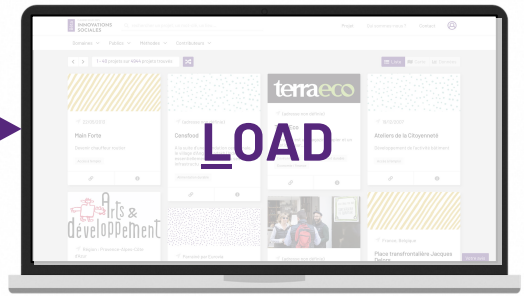
https://github.com/entrepreneur-interet-general/OpenScraper

OPEN SCRAPER fait partie d'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)

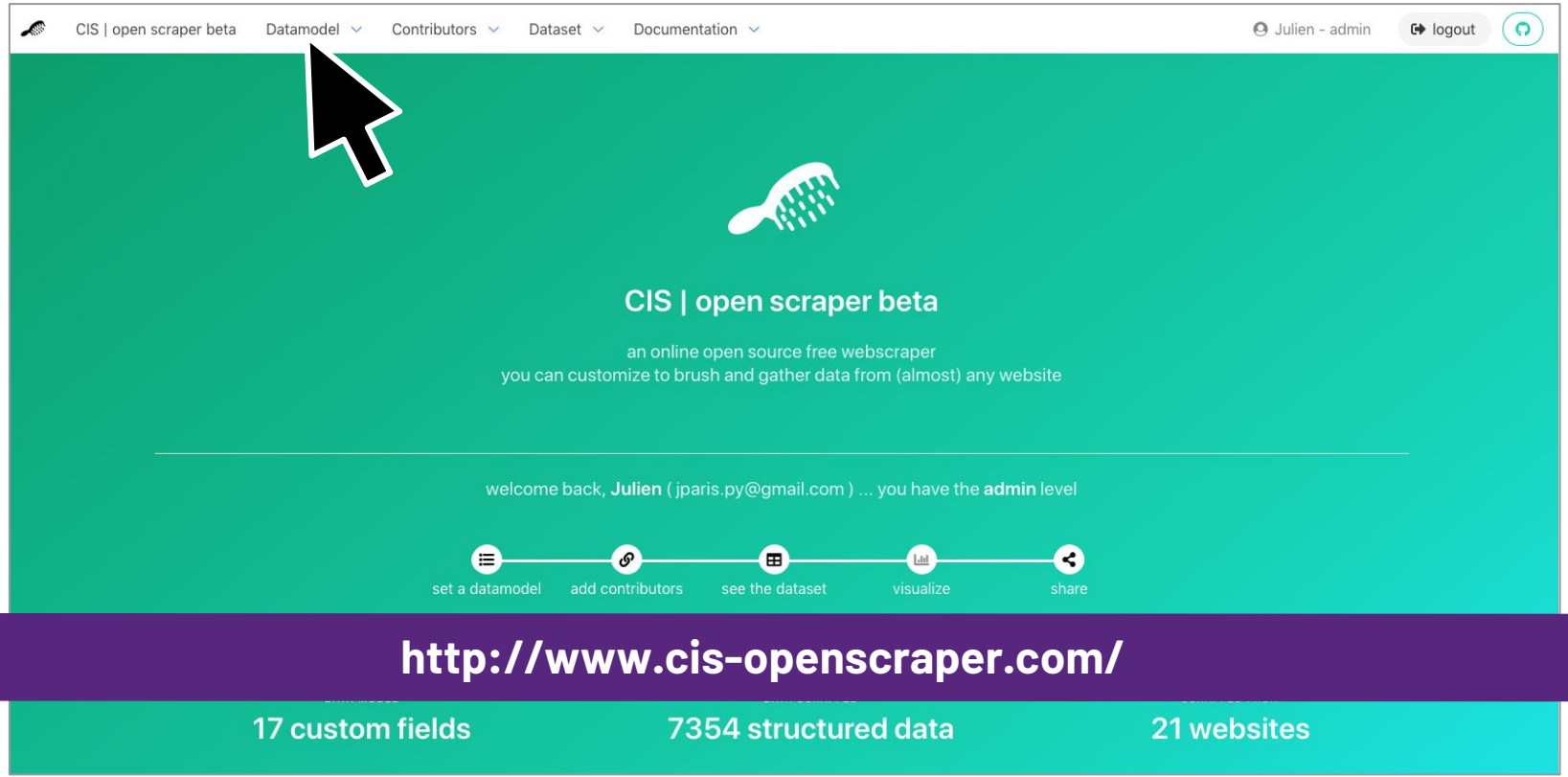
backend API + front
(en développement)

Apiviz

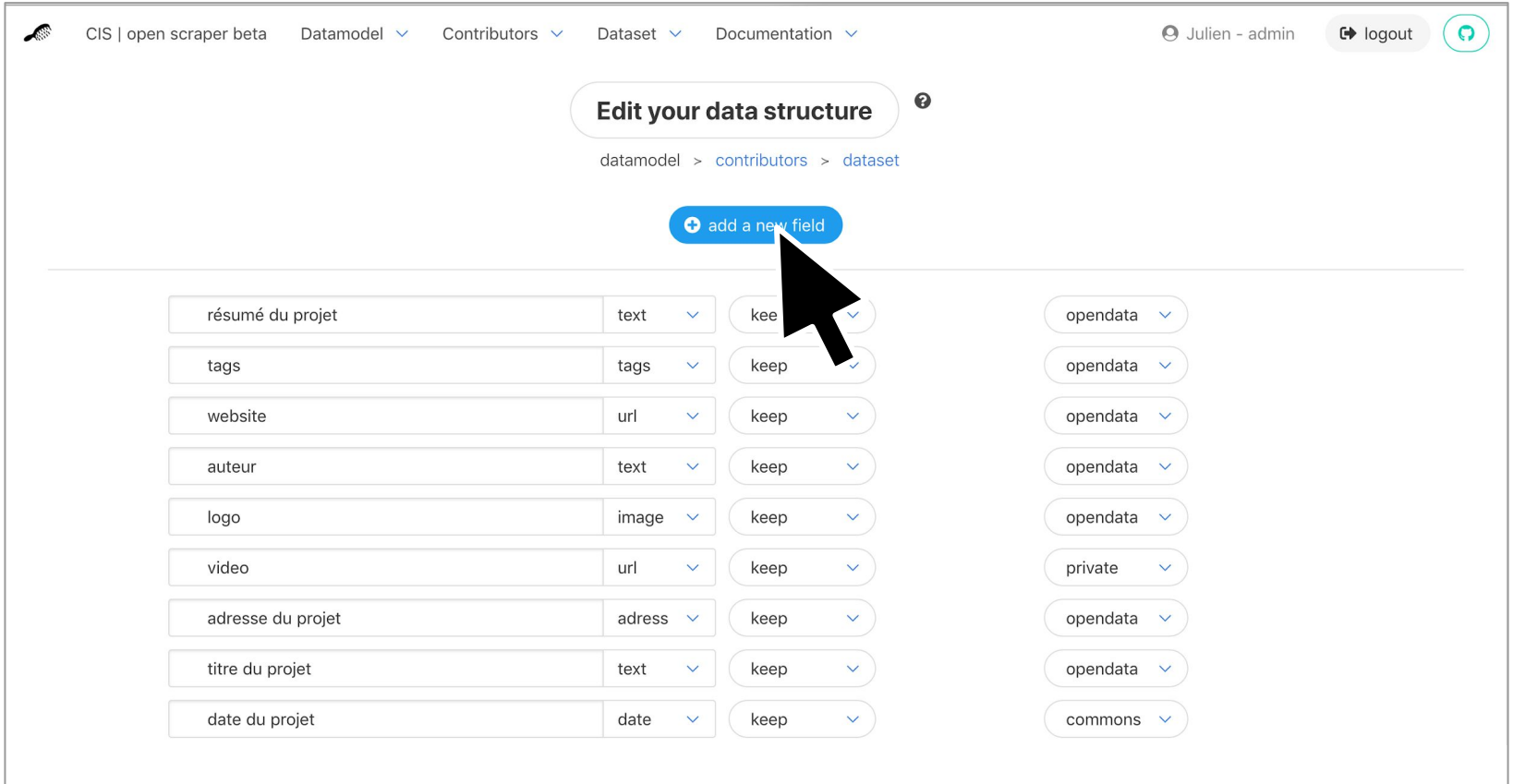
moteur de recherche et visualisation des flux de données

www.carrefourdesinnovationssociales.fr

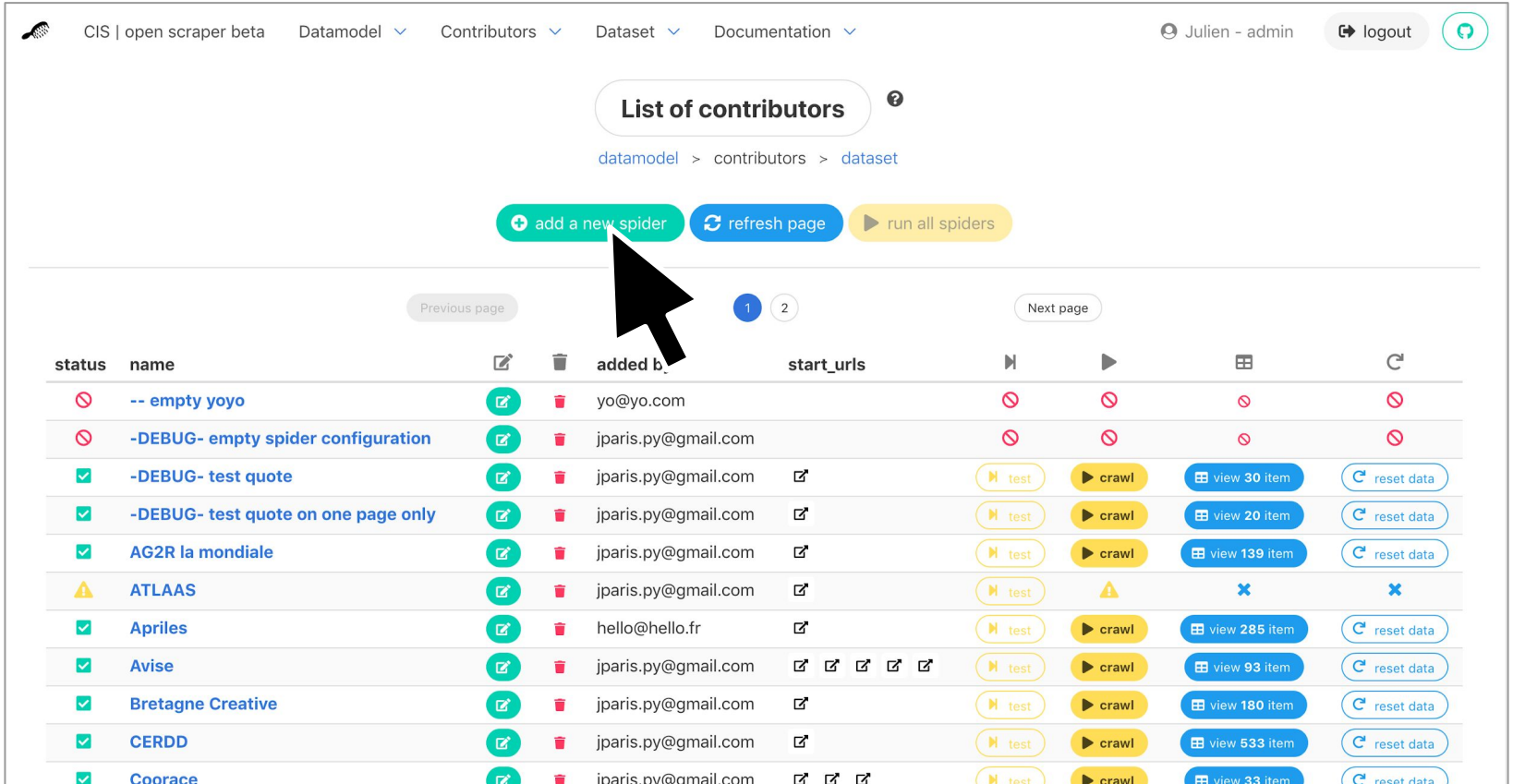
OPEN SCRAPER : configurer et lancer des scrapers en ligne



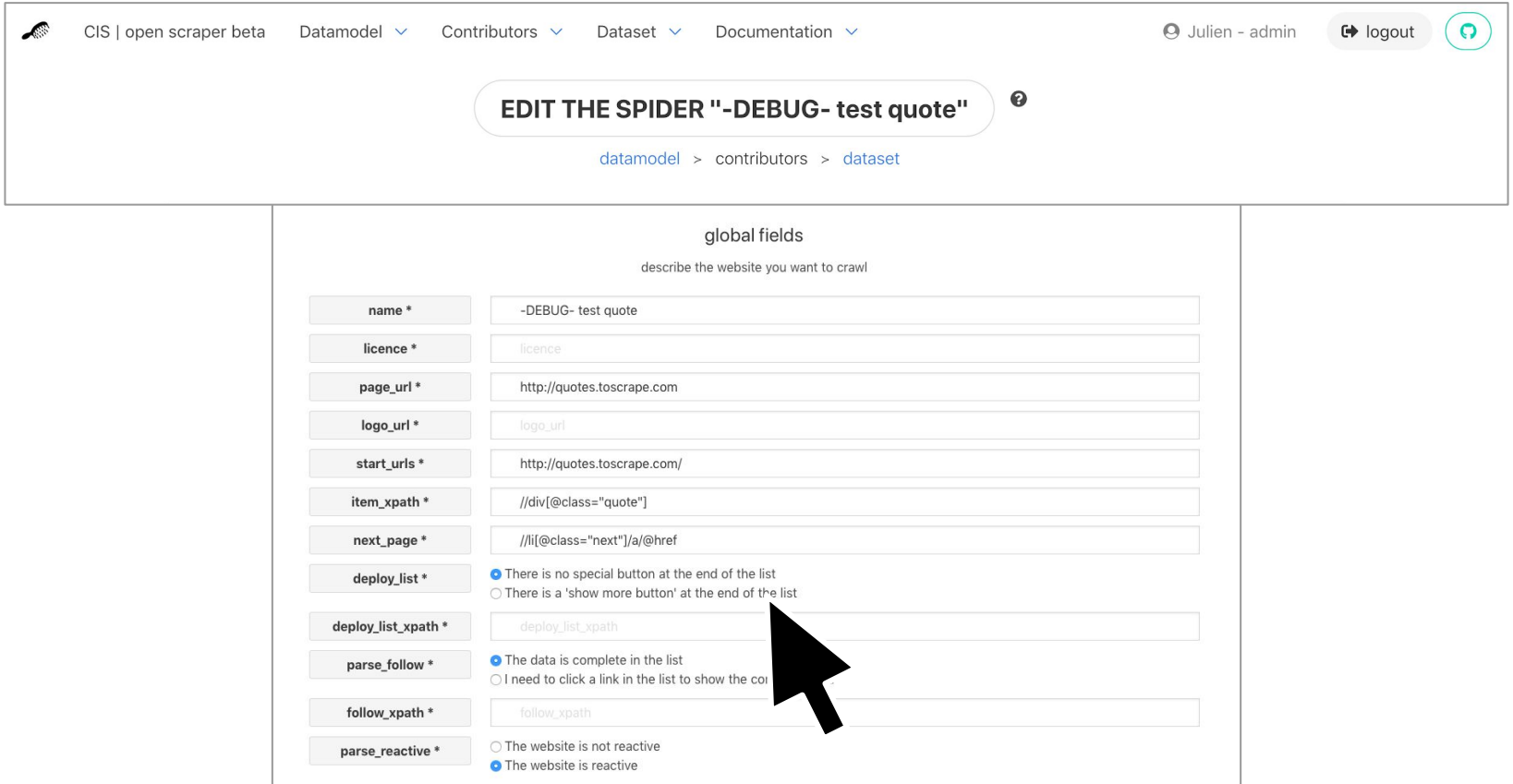
OPEN SCRAPER : définir un modèle de données



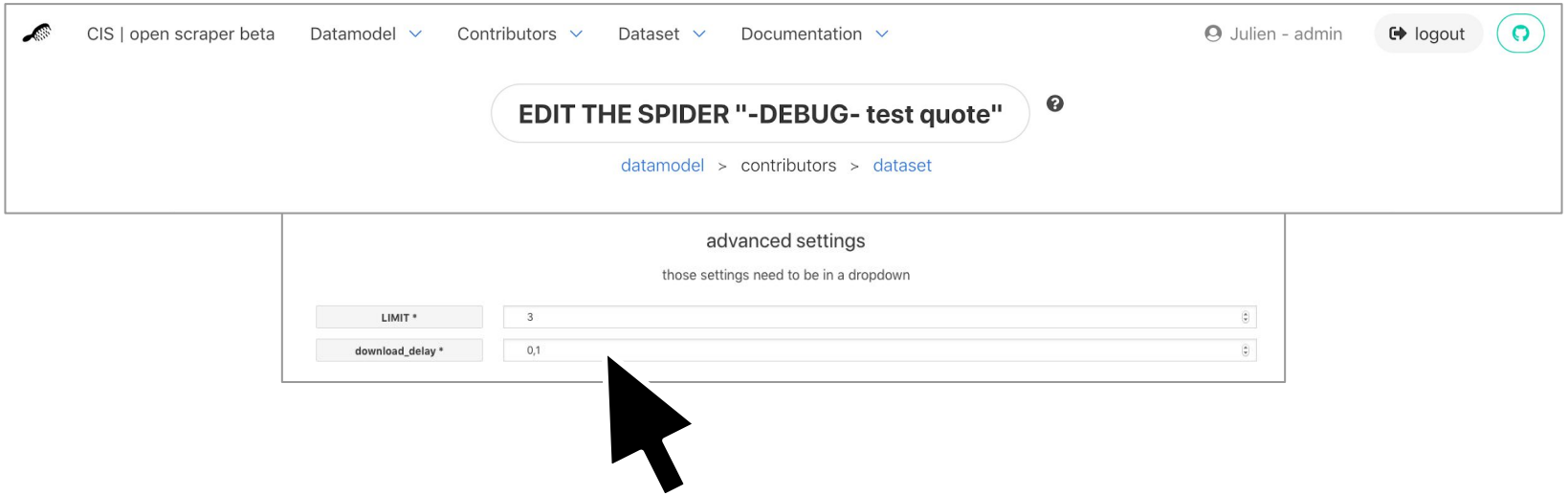
OPEN SCRAPER : ajouter un site à scraper



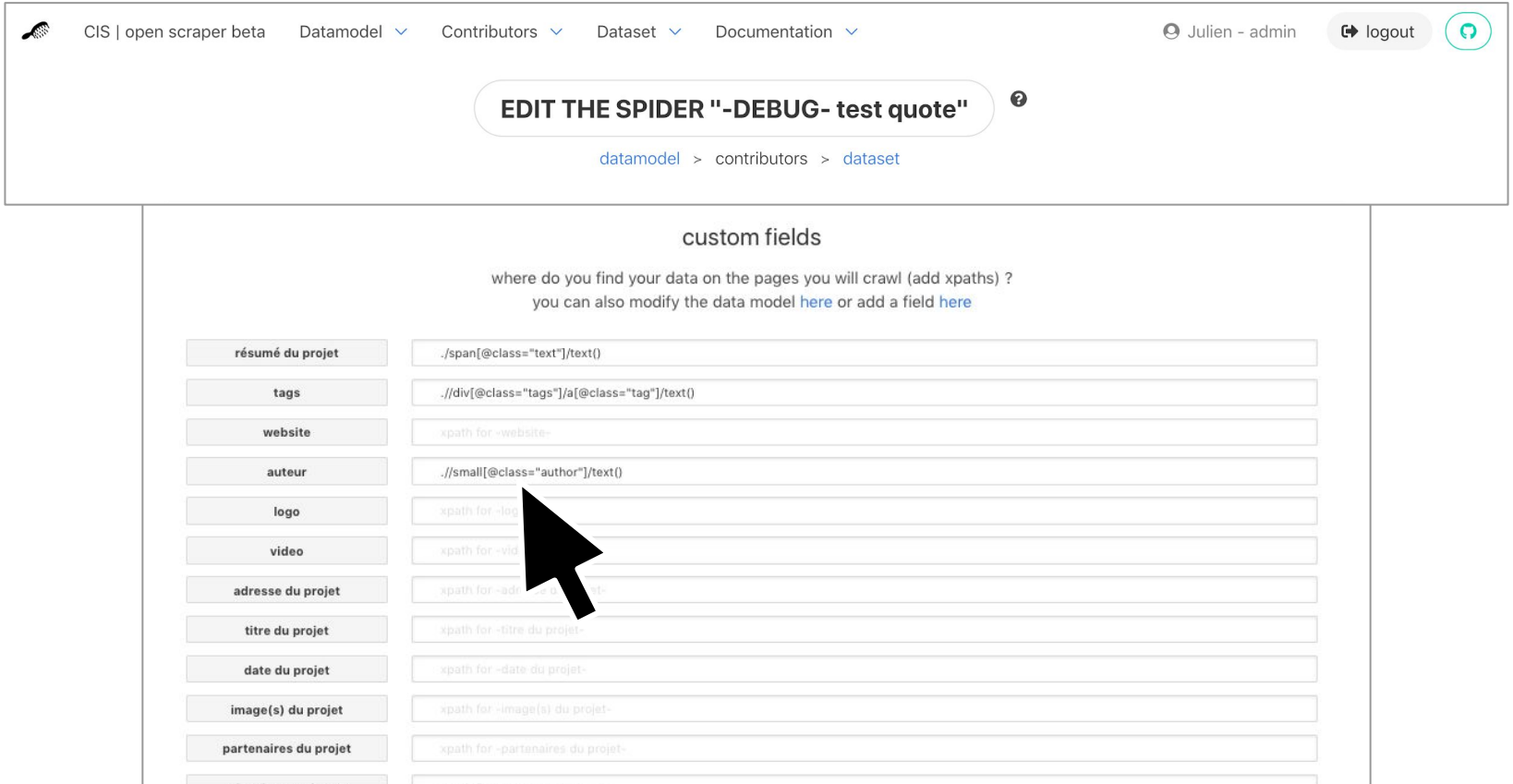
OPEN SCRAPER : éditer un scraper (XPATH)



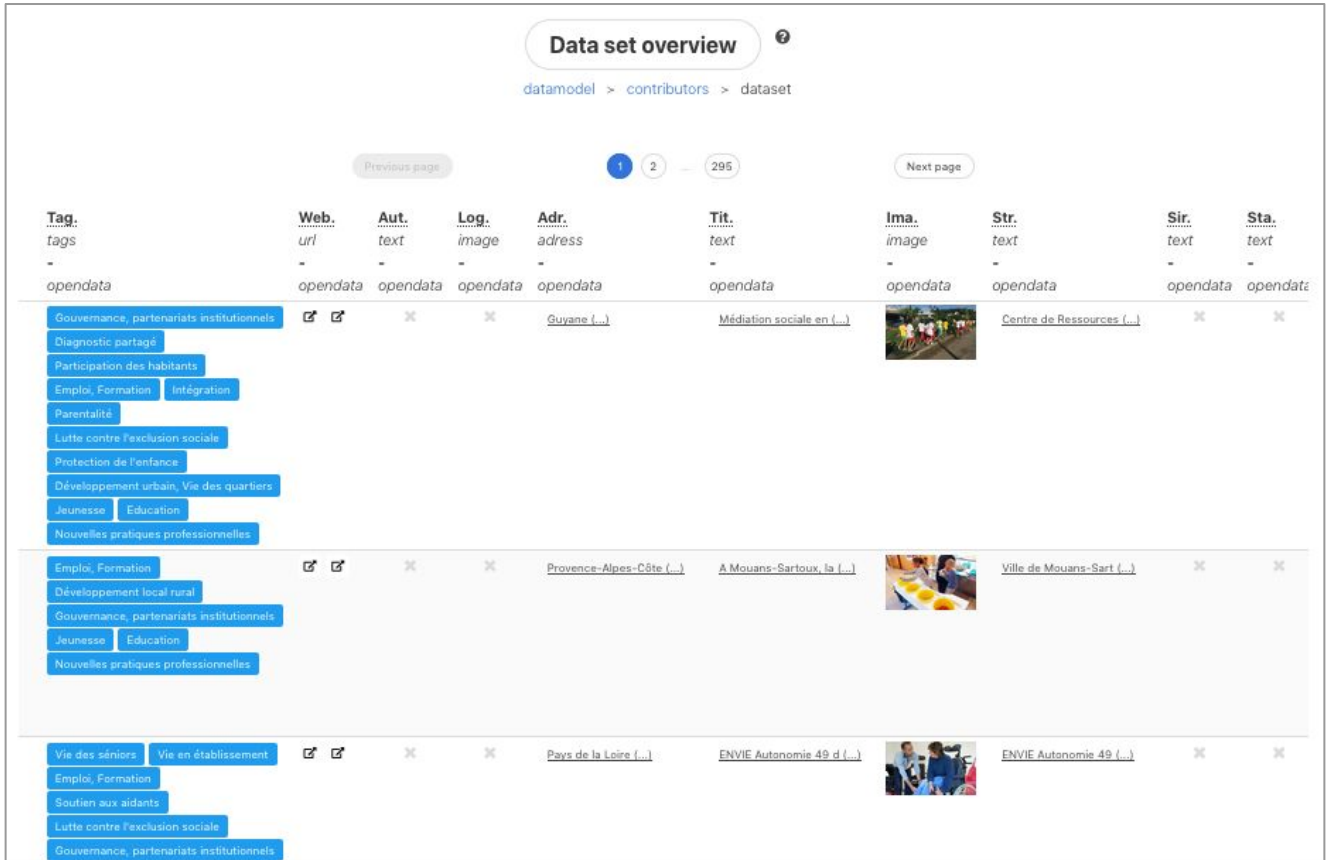
OPEN SCRAPER : éditer un scraper (XPATH)



OPEN SCRAPER : éditer un scraper (XPATH)



OPEN SCRAPER : prévisualiser le jeu de données "scrapées"



OPEN SCRAPER : ouvrir les données via l'API

exemple : http://www.cis-openscraper.com/api/data?search_for=finances

OPEN SCRAPER

Démo sur serveur local
&
questions / réponses

Merci !

Équipe projet Social Connect / Carrefour des innovations sociales (défi EIG 2)

Julien Paris, développeur / CGET / EIG : jparis.py@gmail.com
Elise Lalique, designer UX-UI / CGET / EIG : elise.lalique@cget.gouv.fr
Bénédicte Pachod, coordinatrice / CGET : benedicte.pachod@cget.gouv.fr
Rémy Seillier, coordinateur / CGET : remy.seillier@cget.gouv.fr

Association de préfiguration du Carrefour des innovations sociales

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