Atelier données - Cour des Comptes

Moissonner des données en ligne

quelques éléments méthodologiques et un nouvel outil open source *pour* "scraper" *des sites sur Internet ...*

intervenant : Julien Paris (jparis.py@gmail.com) - dev fullstack - EIG - défi SocialConnect

durée de la présentation : 20 minutes
durée de la démo & questions : 10 minutes
framapad : https://annuel2.framapad.org/p/atelier_scraping_CDC_10102018

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 d'Open Scraper

quoi?

Quelques définitions pratiques

Donnée, modèle de donnée, agrégation, consolidation...

"*Scraper*", "*spider*", "crawler"...

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"

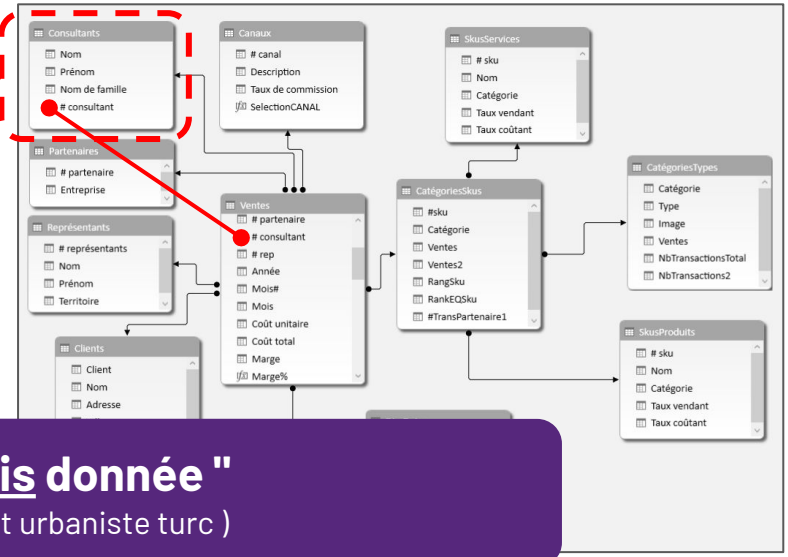
source : Wikipédia, https://fr.wikipedia.org/wiki/Donn%C3%A9e

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	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	M			
12	Gold	M			

Modèle de données
relationnel (clés étrangères, index)



" la donnée n'est jamais donnée "
(*dixit* Mürat Güvenç, sociologue et urbaniste turc)

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

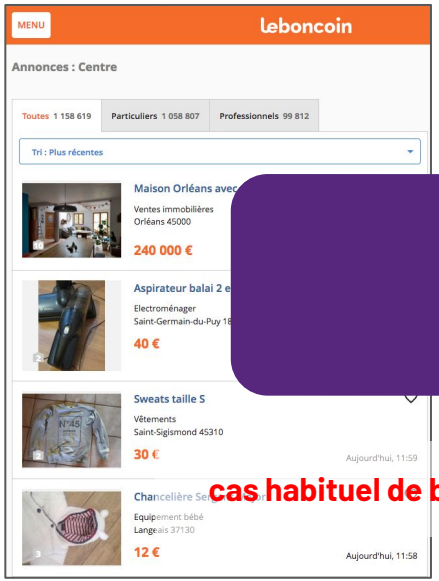
La donnée en ligne : quelles formes de données ?

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**

cas habituel de besoin de *scraping*

cas spécifiques de besoin de *scraping*

pourquoi?

Quelques cas d'usages

A quoi et à qui cela peut-il servir de moissonner des données ?

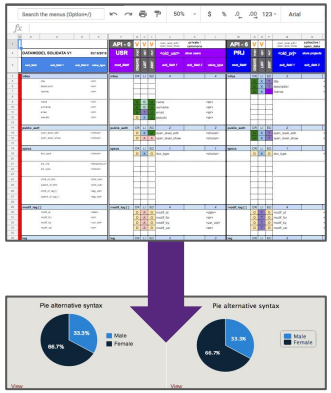
Pourquoi faire du *scraping* de données ?

POUR LES STOCKER

rendre disponible et requêtable la donnée moissonnée

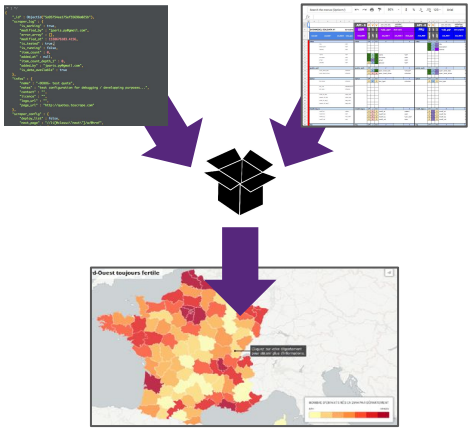
POUR LES ANALYSER

transformer la donnée en information intelligible



POUR LES TRAVAILLER

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



comment?

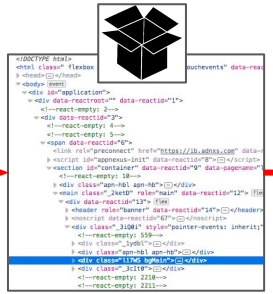
Automatiser tout ce qui peut l'être : le principe de base du "*scraping*" de sites internet

Mise en pratique de la "moisson de données en ligne" dans un cas habituel

"*Scraping*" (1/2) : moissonner un site internet

site web A
à moissonner

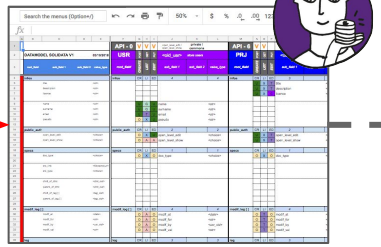
"black box"
automatisant le
moissonnage :
le "***scraper***"



le *scraper* va
lire le HTML,
sélectionner la donnée

puis copier/coller
la donnée ...

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

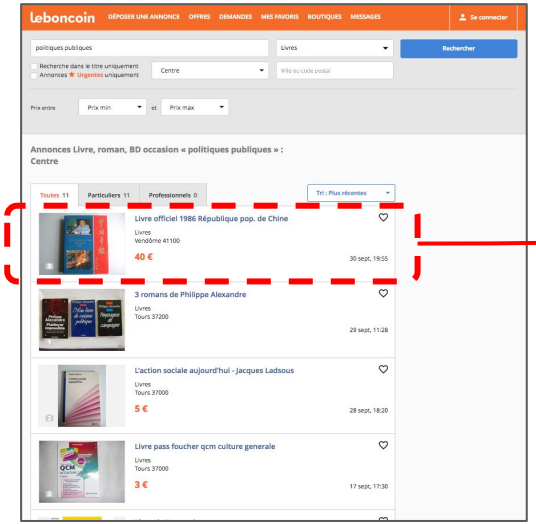


... vers une base de donnée
(ex. MongoDB)

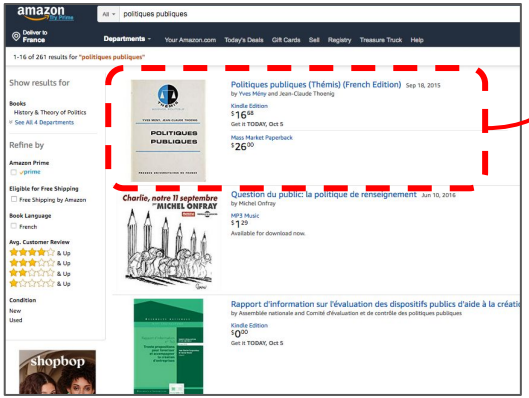
autres
réutilisations

"*Scraping*" (2/2) : moissonner plusieurs sites internet

site web A
(modèle de données A)



site web B
(modèle de données B)



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

source	titre	auteur	prix
Bon coin	livre officiel [...]	---	10€
Bon coin	3 romans de [...]	---	12€
Amazon	Politiques publiques [...]	---	16€
Amazon	Question du public [...]	---	9€

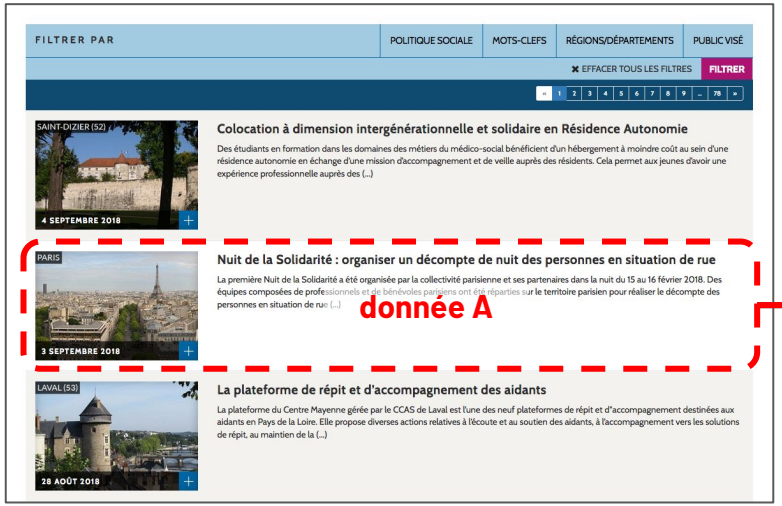
comment?

Les problèmes récurrents

Les principales familles de problèmes techniques, méthodologiques et humains

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

site web à *scraper*



donnée A

simple mais
innefficace

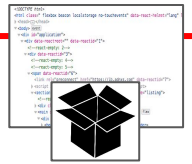
copier/coller à la main la donnée affichée à l'écran

clic droit + "examiner l'élément"

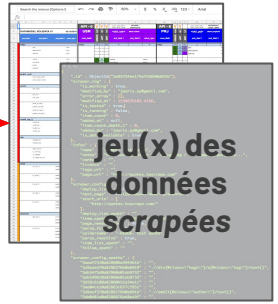
HTML

scraper automatisant le moissonnage

HTML + XPath



efficace
mais complexe

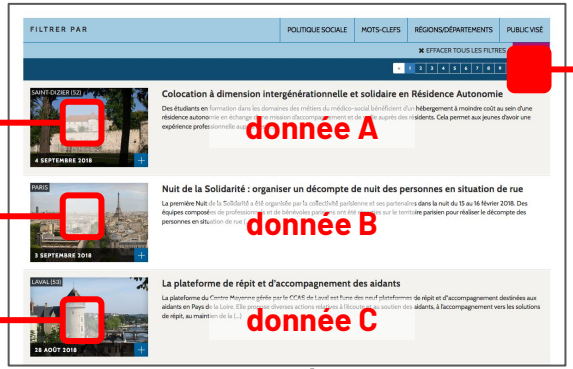


Problème 2 : **Automatiser la navigation dans un site web**

"follow"

site web à scraper
PAGE DÉTAILLÉES

site web à *scraper*
PAGE LISTE - 1



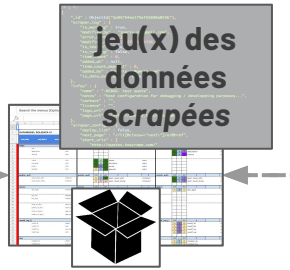
"next page"

site web à *scraper*
PAGE LISTE - 2

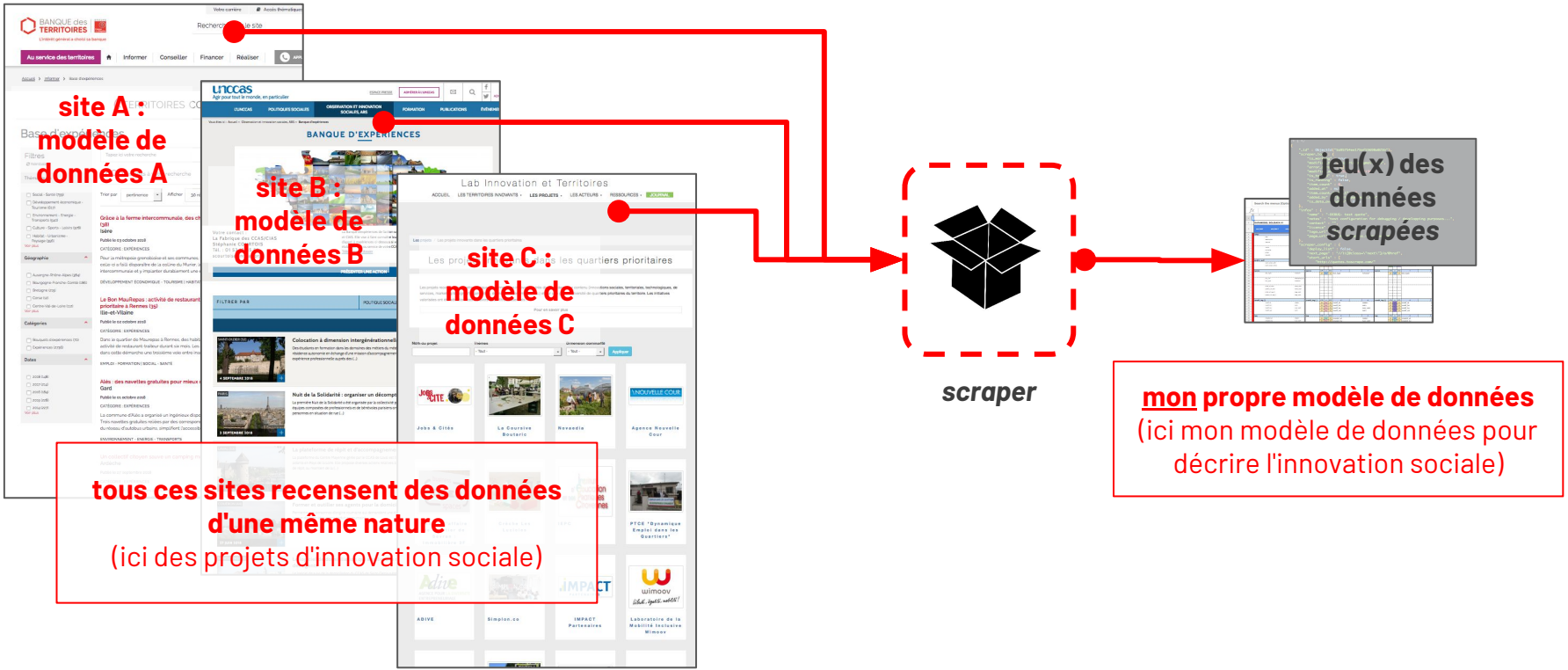
certains sites
sont "réactifs"

recommencer
"follow"

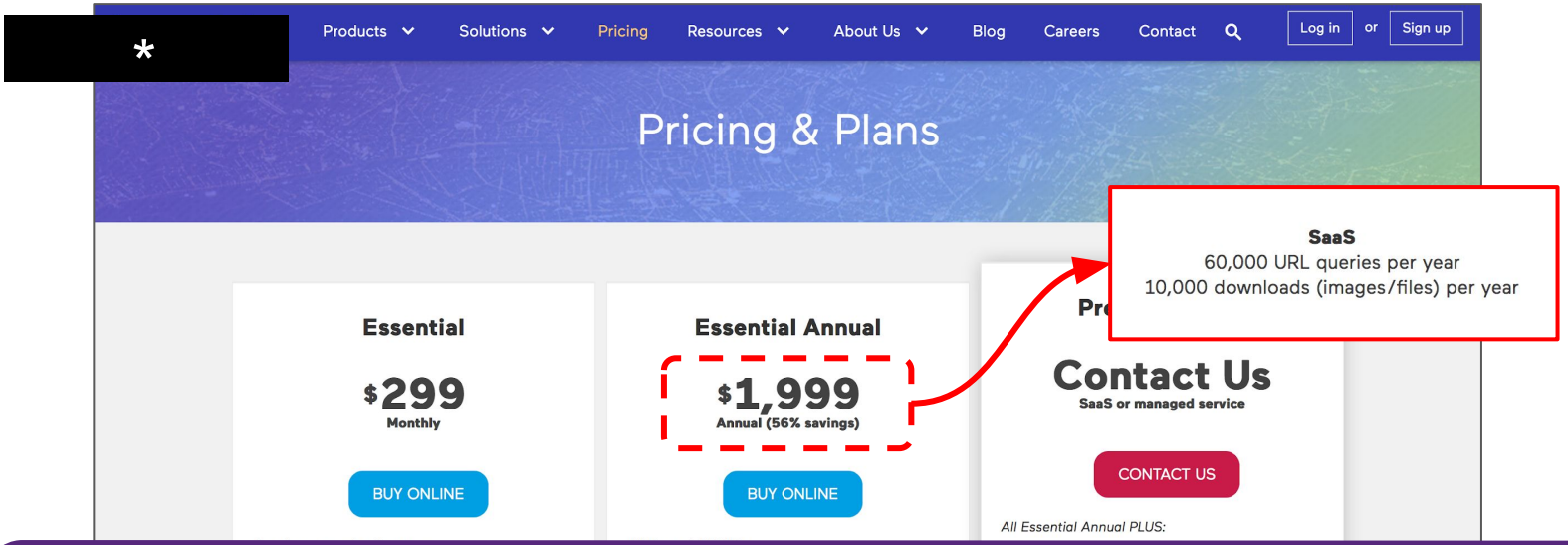
recommencer
"next page"
(etc...)



Problème 3 : **Adapter le *scraping* à des <u>sites web différents</u> 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* (*software as a service*) :
prix annuel pour instancier **un seul *scraper*** (hors API),
c'est-à-dire un site à *scraper* autour d'un modèle de données

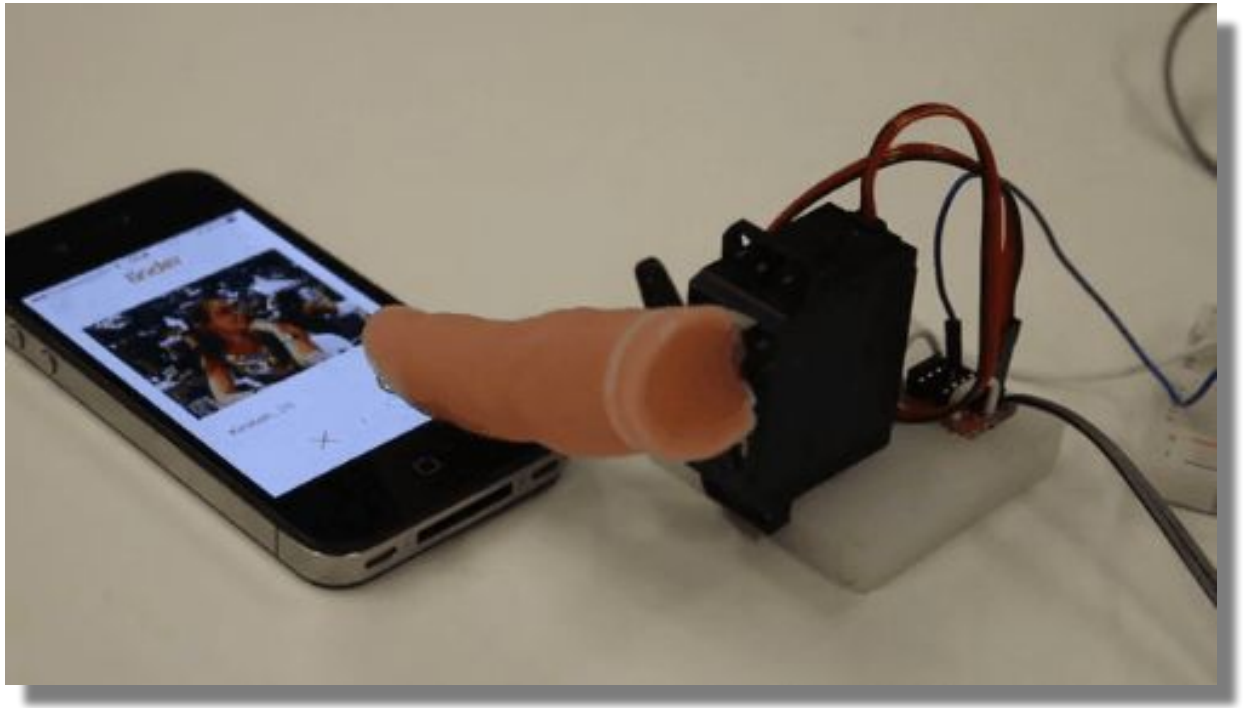
* étant donnés les tarifs pratiqués la décence et la pudeur nous obligent à préserver l'anonymat du service susmentionné

comment?

Open Scraper (version 1.0)

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

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



source : https://giphy.com/gifs/tinder-automation-tinderbot-3rgXByefj5zvCcod0M

OPEN SCRAPER : une *webapp* de *scraping* pour ...

- ☑ **Automatiser le "copier-coller"**
- ☑ **Automatiser la navigation dans un site** (y compris "réactif")
- ☑ **Adapter le *scraping* à des sites différents**
- ☑ **Avoir les moyens financiers ou humains de *scraper***

  - **100% *open source*** : n'importe qui peut voir le code et réutiliser gratuitement l'application
  - **une approche UX*** : pas de scripts, pas forcément besoin d'un.e ingénieur.e
  - **une API** pour rendre la donnée requêtable à distance
  - **une gestion du degré d'ouverture des données**, champ par champ

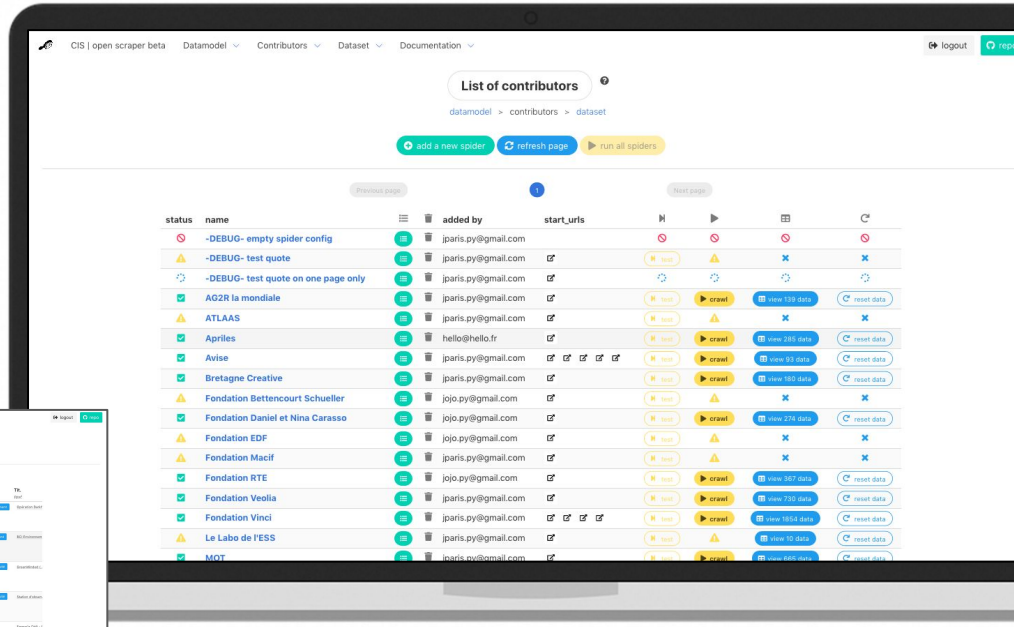
*UX pour "User Experience", soit une approche utilisateur centrée sur son expérience de navigation

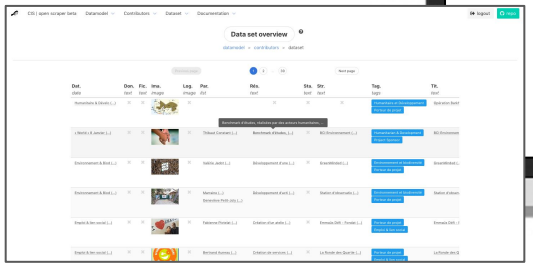
OPEN SCRAPER : agréger des données publiées sur des sites

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 : *open data / commons / collective / private*

www.cis-openscraper.com

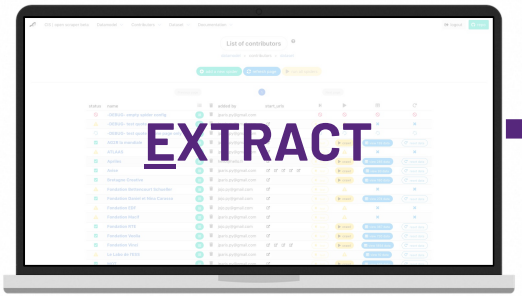


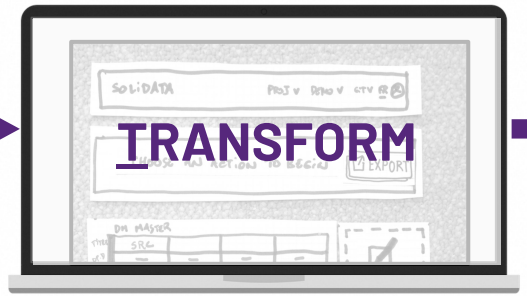


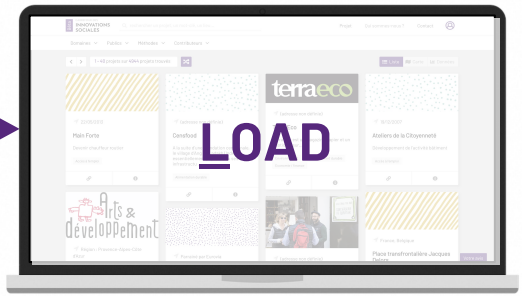
OPEN SCRAPER : une webapp de *scraping* en *open source*

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 (back + front)

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

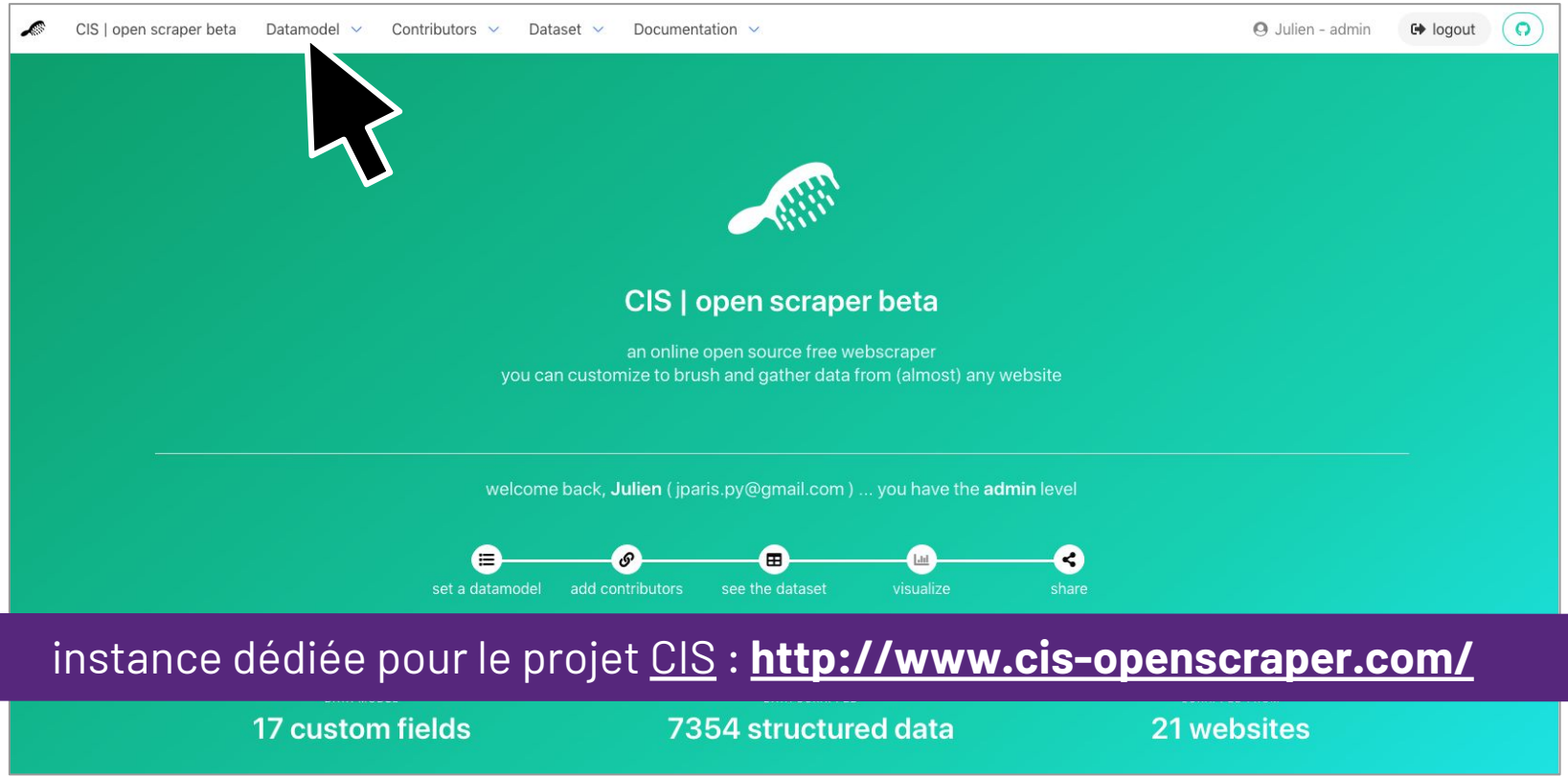
(en développement)

Apiviz (aka CIS_front)

moteur de recherche et visualisation des flux de données

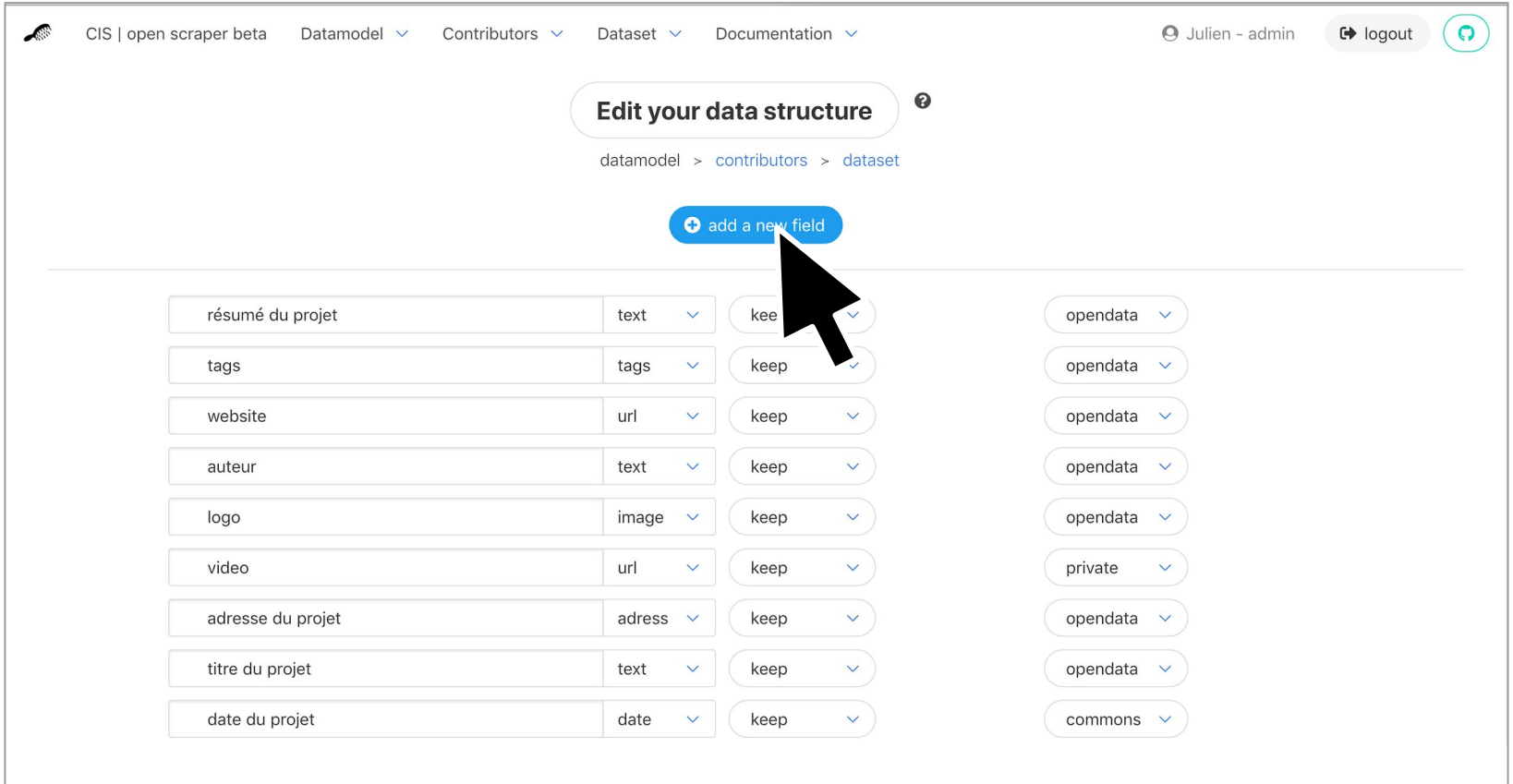
www.carrefourdesinnovationssociales.fr

OPEN SCRAPER : configurer et lancer des *scrapers* en ligne

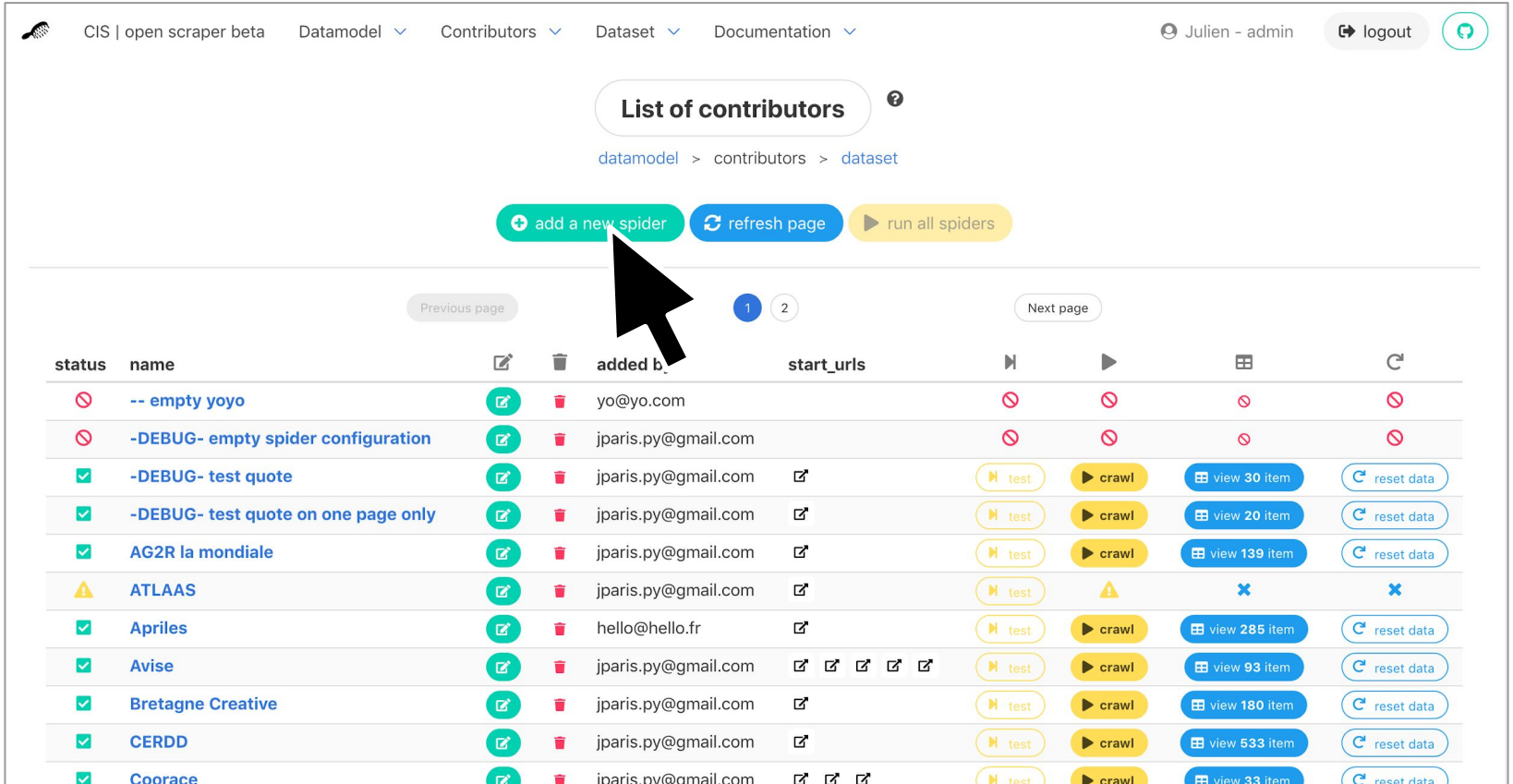


instance dédiée pour le projet CIS : **http://www.cis-openscraper.com/**

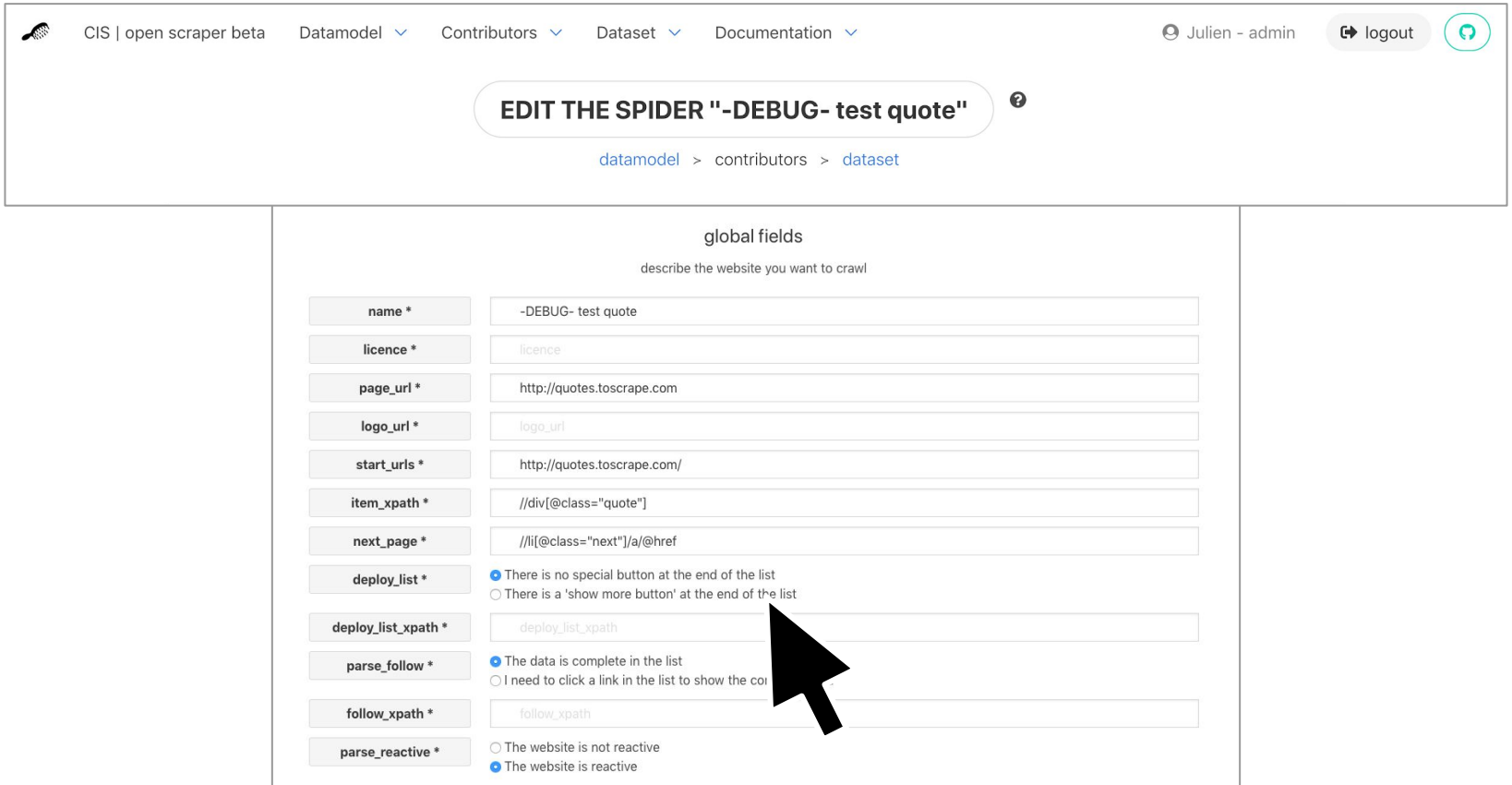
OPEN SCRAPER : définir un modèle de données



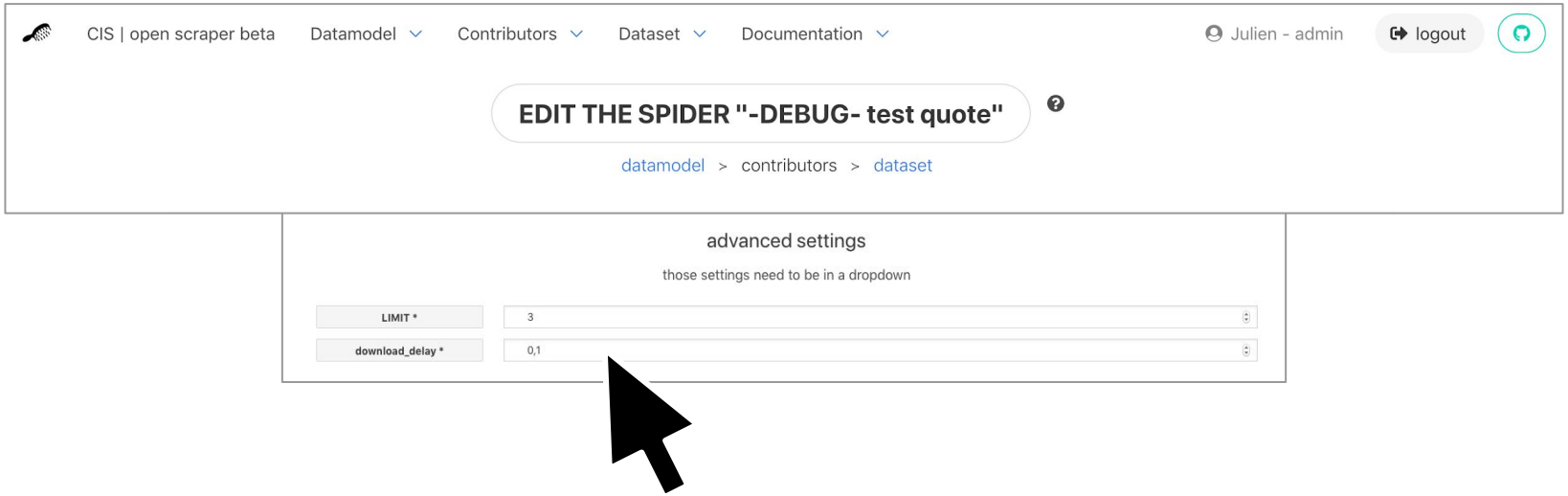
OPEN SCRAPER : ajouter un site à *scraper*



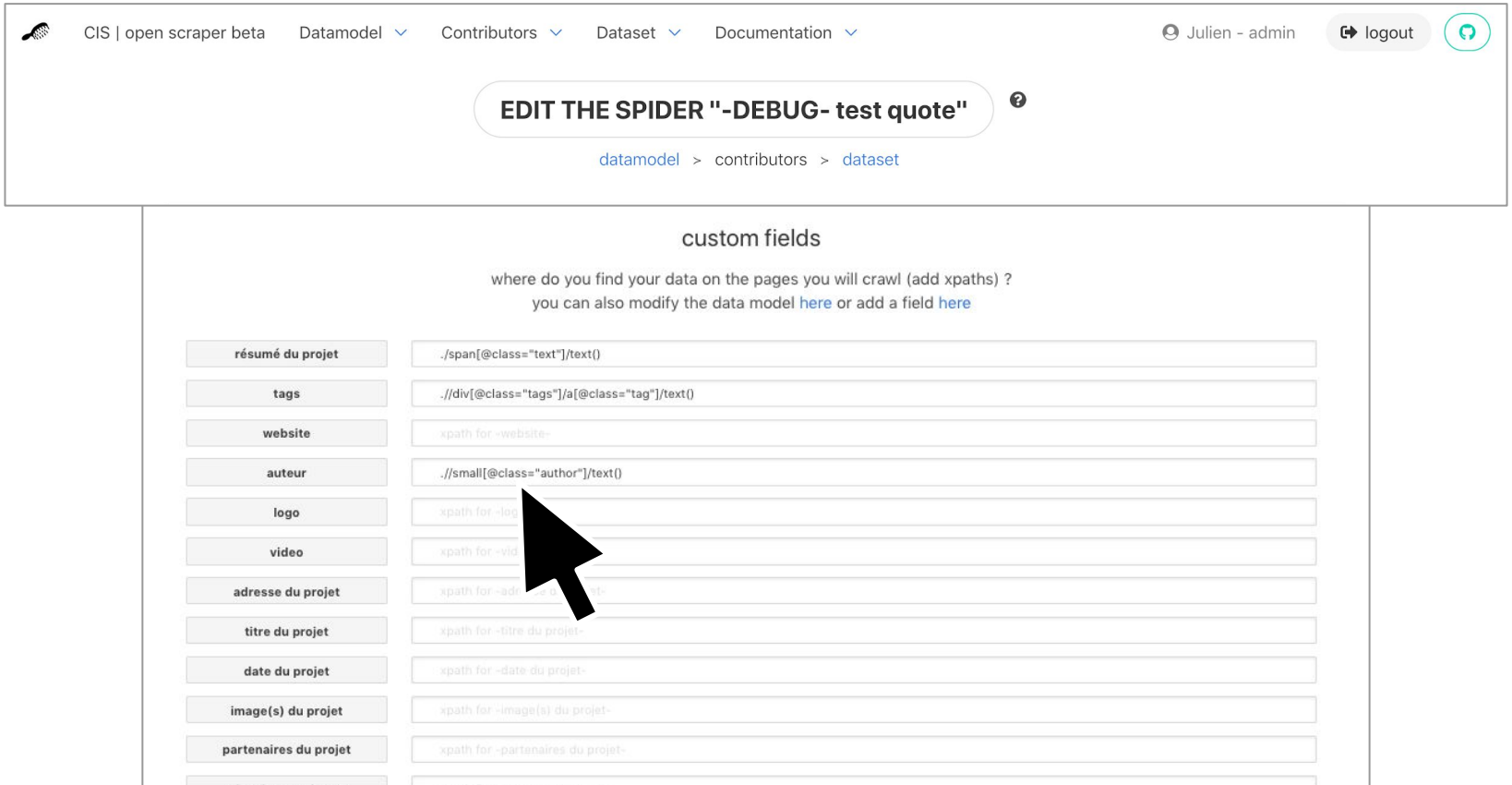
OPEN SCRAPER : éditer un *scraper* (ajouts de *XPath*)



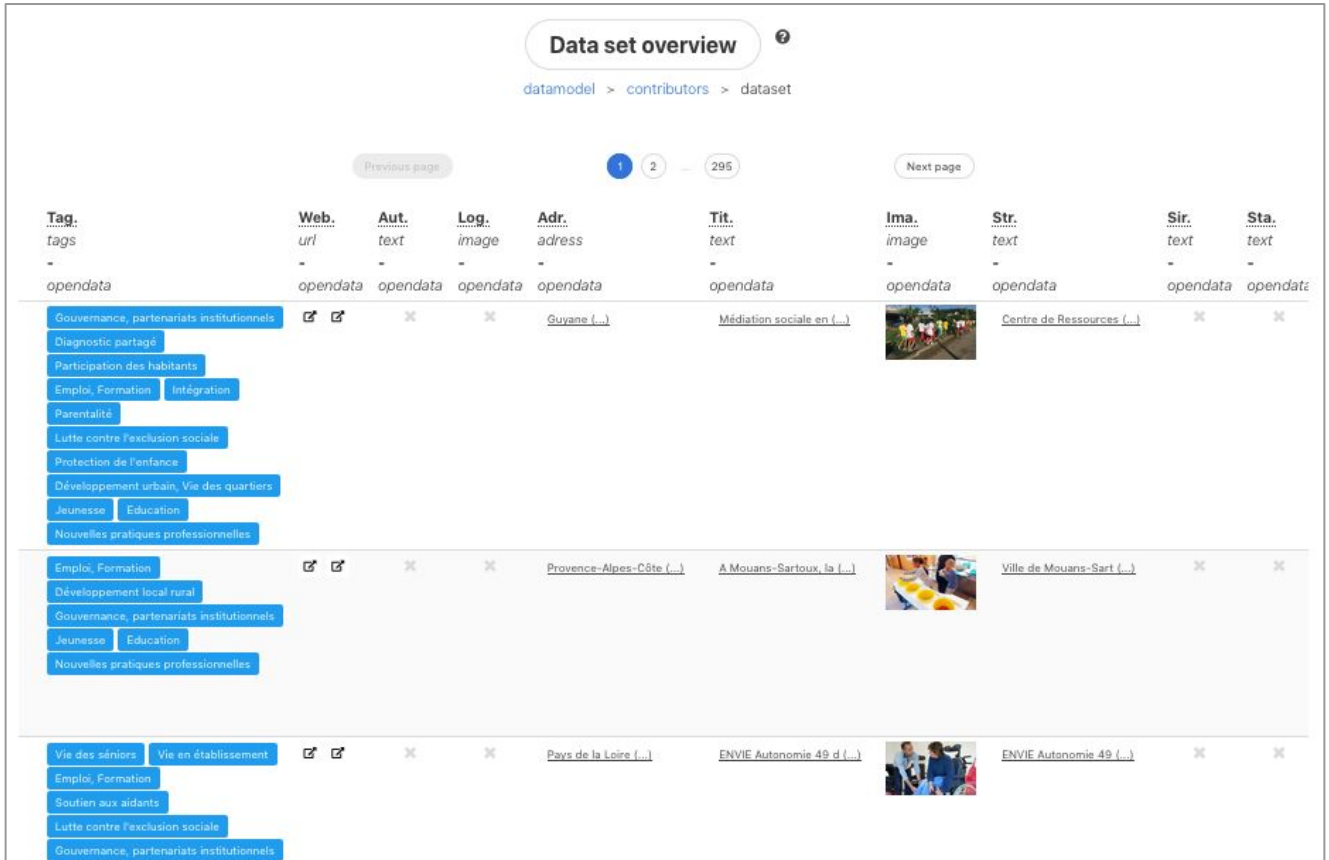
OPEN SCRAPER : éditer un *scraper* (ajouts de *XPath*)



OPEN SCRAPER : éditer un *scraper* (ajouts de *XPath*)



OPEN SCRAPER : prévisualiser le jeu de données "*scrapées*"



OPEN SCRAPER : ouvrir les données en JSON via l'API

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

OPEN SCRAPER : démo de la version 1.0

Démo sur serveur local (pour voir les coulisses) **& questions / réponses**

questions / réponses

Chantiers en cours : dockerisation, gestion de projets de *scraping*, etc...

Merci !

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

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

Pour recevoir le lien vers la présentation et le pdf inscrivez-vous sur le framapad :
https://annuel2.framapad.org/p/atelier_scraping_CDC_10102018