

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



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 ?

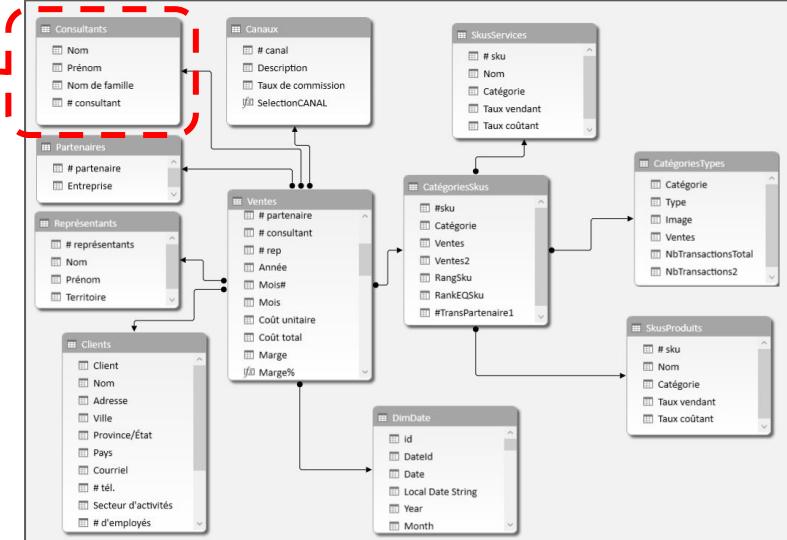
" ce qui est **connu** et qui sert de
point de départ à un raisonnement
ayant pour objet la détermination d'une
solution à un problème en relation avec
cette donnée"

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

Modèle de données
tabulaires

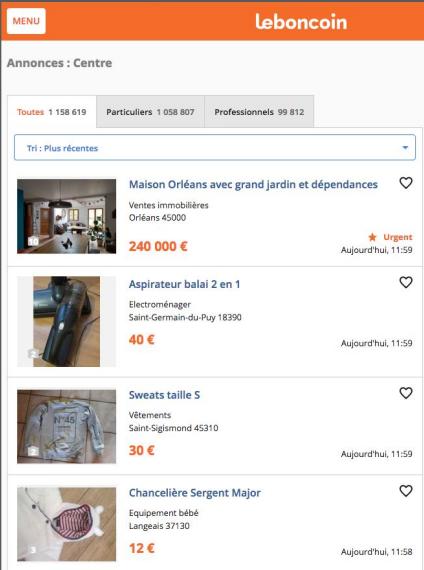
	A	B	C	D	E
1	Customer	City	Region	Product	Quantity
2	Orange	Big Town	West	Milk Chocolate	125
3	Red	Big Town	West	Dark Chocolate	210
4	Pink	Medium Town	East	Milk Chocolate	145
5	Grey	Big Town	West	Chocolate Hazelnut	21
6	Blue	Small Town	South	Dark Chocolate	50
7	Dark	Big Town	West	Chocolate Hazelnut	65
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...)**

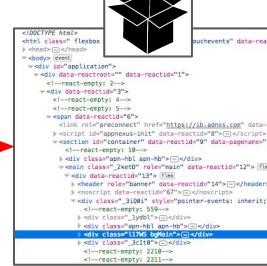
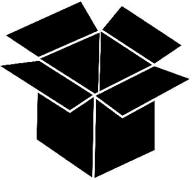
```
    "is_scraping": true,
    "is_working": true,
    "modified_by": "jparis.py@gmail.com",
    "error_array": [],
    "modified_at": 1538676103.4156,
    "is_tested": true,
    "is_stopped": false,
    "item_count": 0,
    "added_at": null,
    "item_depth_1": 0,
    "added_by": "jparis.py@gmail.com",
    "is_data_available": true
  },
  "infos": [
    {
      "name": "DEBUG-test quote",
      "notes": "test configuration for debugging / developing purposes...",
      "contact": "",
      "license": "+",
      "logo_url": "",
      "page_url": "http://quotes.toscrape.com"
    }
  ],
  "scrapers_config": [
    {
      "deploy_list": false,
      "next_xpath": "/@*[name()=next]/a/@href",
      "start_urls": [
        "http://quotes.toscrape.com"
      ],
      "deploy_list_xpath": "",
      "item_xpath": "/@*[name()=quote]",
      "deploy_reactive": false,
      "parse_follow": false,
      "spidername": "DEBUG-test quote",
      "parse_reactive": true,
      "item_list_xpath": "",
      "follow_xpath": ""
    }
  ],
  "scrapers_config_xpaths": [
    {"$odaf2310a82866ea944b1b": ""},
    {"$odkoenf38286286d996f919": "/div[@class='tags']/a[@class='tag']/text()"},
    {"$od54759d86386d169d732": "/div[@class='text']/p/text()"},
    {"$od29188086386d169d732": "/div[@class='text']/p/text()"},
    {"$od4ac1e08a863c3d797931": ""},
    {"$od9af38286286d996f919": "//small[@class='author']/text()"},
    {"$od8d1010a82867676483e0f": ""}
  ]
}
```

"Scraping" : moissonner de la donnée publiée en ligne



Site web

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



lire le HTML



A screenshot of a Microsoft Excel spreadsheet titled "DATA". It contains several columns with data, such as "ID", "Title", "Price", and "Location". The cells are color-coded, and there are formulas and functions visible in the cells.

Tableur
csv, xls, tsv, ...

```
* 1 */
{
  "_id": ObjectId("5e0d754ee775ef5569eb20"),
  "scraper": "leboncoin",
  "is_working": true,
  "modified_by": "joris.py@gmail.com",
  "modified_at": 1538670804156,
  "is_tested": true,
  "is_stable": true,
  "item_count": 0,
  "readable_at": null,
  "is_auto_update": false,
  "added_by": "joris.py@gmail.com",
  "is_auto_updateable": true
},
"info": {
  "name": "leboncoin test quota",
  "notes": "test configuration for debugging / developing purposes...",
  "version": "1.0.0",
  "licens": "MIT",
  "log_url": "",
  "page_url": "http://quotes.toscrape.com"
},
"scraper": {
  "url": "https://www.leboncoin.fr/annonces/centre/158619/particuliers/158807/professionnels/99812/tri_plus_recentes",
  "next_page": "/fl1(klass)>next">/a@ref",
  "max_page": 1
}
```

Base de données
(ex MongoDB)

autres
réutilisations

"Scraping" : moissonner de la donnée publiée en ligne

Site web A
(modèle de données A)

The screenshot shows a search results page on the Leboncoin website for the term "politiques publiques". The results are categorized into "Toutes" (11), "Particuliers" (11), and "Professionnels" (0). The first result is a book titled "Livre officiel 1986 République pop. de Chine" by Yves Ménny and Jean-Claude Theisig, listed at 40€. The second result is a book titled "3 romans de Philippe Alexandre" by Philippe Alexandre, listed at 5€. The third result is a book titled "L'action sociale aujourd'hui - Jacques Ladsous" by Jacques Ladsous, listed at 5€. The fourth result is a book titled "Livre pass foucher qcm culture générale" by Various, listed at 3€.

Site web B
(modèle de données B)

The screenshot shows a search results page on the Amazon website for the term "politiques publiques". The results are categorized into "Books" (1-16 of 261 results) and "Departments" (Your Amazon.com, Today's Deals, Gift Cards, Sell, Registry, Treasure Truck, Help). The top result is a book titled "Politiques publiques (Thémis) (French Edition)" by Yves Ménny and Jean-Claude Theisig, Kindle Edition, \$16.99, Get it TODAY, Oct 5. The second result is a book titled "Question du public: la politique de renseignement" by Various, Kindle Edition, \$2.99, Available for download now. The third result is a book titled "Rapport d'information sur l'évaluation des dispositifs publics d'aide à la création" by Assemblée nationale et Comité d'évaluation et de contrôle des politiques publiques, Kindle Edition, \$0.99, Get it TODAY, Oct 5.

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

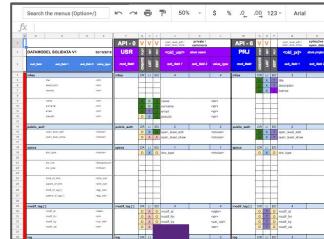
rendre disponible et requêtable
la donnée moissonnée

```
/* ... */  
{  
  "id": ObjectID("5d95754ee175ef55698a023b"),  
  "scraper_log": {  
    "url": "http://www.firebaseio.com",  
    "modified_by": "jports.py@gmail.com",  
    "modified_dt": "2019-07-03T14:41:56",  
    "is_tested": true,  
    "is_trusted": false,  
    "item_count": 1,  
    "added_dt": null,  
    "item_url": "http://www.firebaseio.com",  
    "added_by": "jports.py@gmail.com",  
    "is_item_available": true  
  },  
  "info": {  
    "notes": "DNNB test notes",  
    "notes_c": "test configuration for debugging / developing purposes...",  
    "version_c": "+",  
    "version": "+",  
    "log_url": "+",  
    "page_url": "http://quotes.toscrape.com"  
  },  
  "scraper_config": {  
    "url": "http://www.firebaseio.com/.json",  
    "next_page": "http://www.firebaseio.com/.json"  
  }  
}
```



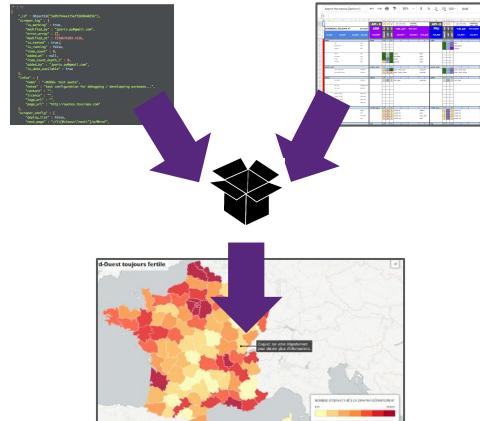
ANALYSER LA DONNÉE

transformer la donnée en
information intelligible



TRAVAILLER LA DONNÉE

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

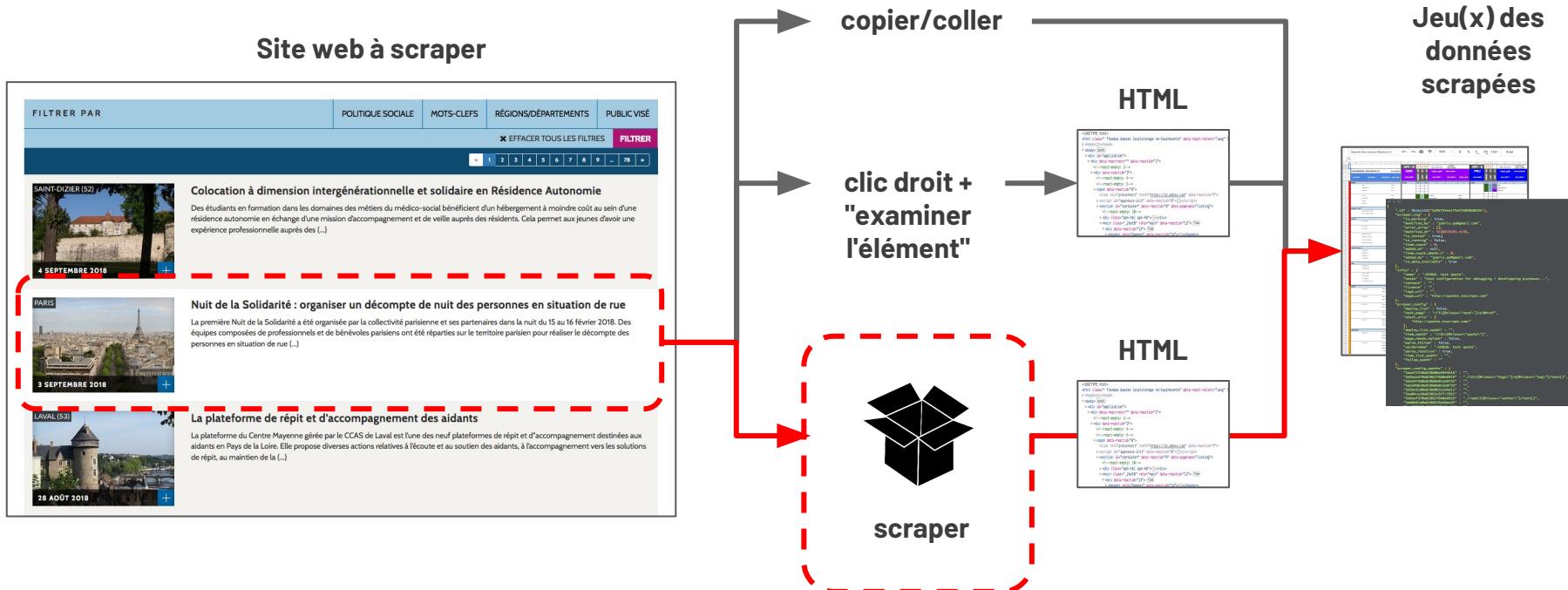


comment?

Les problèmes récurrents

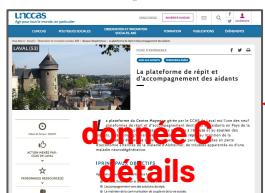
Les principales familles de problèmes techniques et méthodologiques

Automatiser le "copier-coller"



Naviguer dans un site peut être compliqué à automatiser

Site web à scraper
PAGE DÉTAILLÉES



Site web à scraper
PAGE 1

FILTRER PAR

POLITIQUE SOCIALE MOTS-CLEFS RÉGIONS/DÉPARTEMENTS PUBLIC VISE

Colocation à dimension intergénérationnelle et solidaire en Résidence Autonomie

Des résidents en formation dans les domaines du médico-social bénéficient d'un hébergement à moindre coût au sein d'une résidence autonome en échange de missions d'accompagnement et de suivi auprès des résidents. Cela permet aux jeunes d'avoir une expérience professionnelle aux côtés des résidents.

SAINT-DIZIER (52)

4 SEPTEMBRE 2018

PARIS

3 SEPTEMBRE 2018

LAVAL (53)

28 AOÛT 2018

Nuit de la Solidarité : organiser un décompte de nuit des personnes en situation de rue

La première Nuit de la Solidarité a été organisée par le collectif parisien et ses partenaires dans la nuit du 15 au 16 février 2018. Des équipes composées de professionnels et de bénévoles parisiens ont été mobilisées sur le territoire pour réaliser le décompte des personnes en situation de rue.

La plateforme de répit et d'accompagnement des aidants

La plateforme du Centre Mayenne gérée par le CCAS de Laval est l'une des rares plateformes de répit et d'accompagnement destinées aux aidants en Pays de la Loire. Elle propose diverses actions relatives à l'accès au soutien des aidants, à l'accompagnement vers les solutions de répit, au maintien de la (...)

```
l = []
for e in events:
    date = e['date']
    titre = e['titre']
    l.append([date, titre])
```

Site web à scraper
PAGE 2

FILTRER PAR

POLITIQUE SOCIALE MOTS-CLEFS RÉGIONS/DÉPARTEMENTS PUBLIC VISE

« Chantiers jeunes » : projet de mobilité et d'implication citoyenne par la rénovation d'éléments du patrimoine communal et/ou associatif

Ouvert aux jeunes de 18-25 ans, ce chantier permet aux jeunes de faire des bénévoles et de se former, et de bénéficier d'une rémunération et d'une insertion professionnelle, inscrite au titre de l'insertion.

ABBEVILLE (60)

16 JUILLET 2018 (M&A - 10/04/2018)

FOUGERES (35)

11 AVRIL 2018

AGDE (34)

Pratique d'une activité physique et sportive pour favoriser l'autonomie des résidents : l'escrime

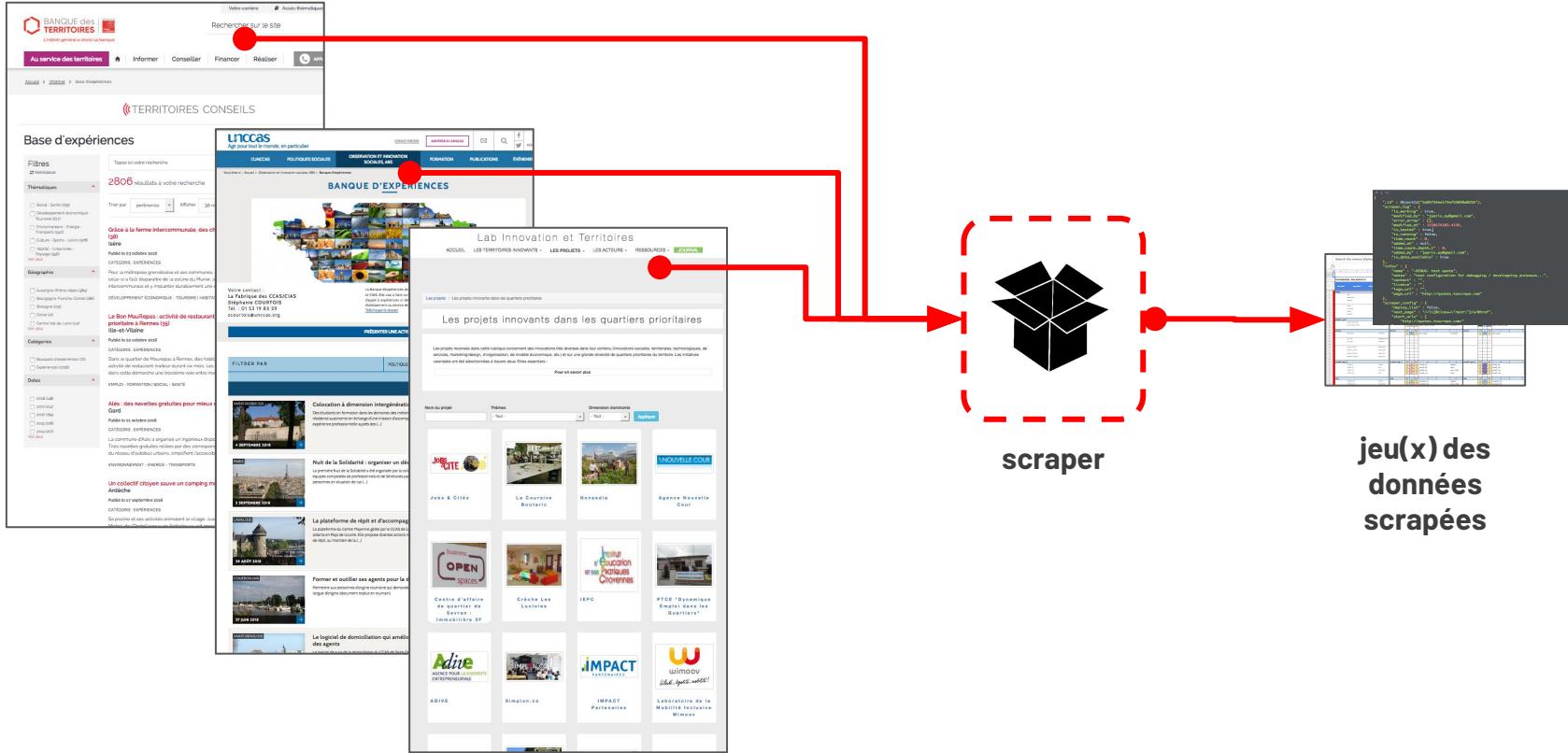
Après une séance d'essai, les personnes peuvent suivre une pratique régulière. Les séances mensuelles d'escrime adaptée. La pratique, début ou assis en fonction des capacités physiques de chacun, ne pose pas de problème, l'adaptation guide la (...)

! salonsnier – 1 logement intergénérationnel

Cette action a pour vocation la mise en place d'un logement intergénérationnel, en recherche d'hébergement, avec les seniors disposant dans leur résidence principale d'une solution d'hébergement pour les personnes âgées, et aux personnes de se loger pour un prix raisonnable et aux standards de (...)

Jeu(x) des
données
scrapées

Scrapper plusieurs sites autour d'un même modèle de données



jeu(x) des
données
scrapées

comment?

Open Scraper

un web applications open source
pour moissonner des données sur Internet

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 : opendata / commons / collective / private

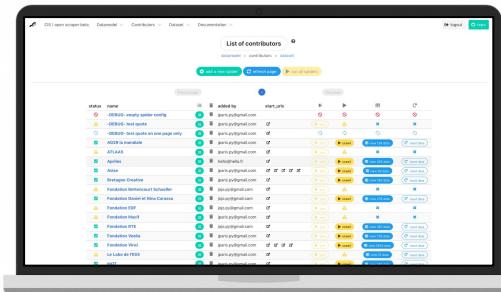
This screenshot shows the 'Dataset Overview' section of the CIS open scraper beta interface. It displays a grid of data items from various partners, each represented by a thumbnail icon and a brief description. The columns include 'Titre', 'Description', 'Type', and 'Actions'. Examples of datasets shown include 'Département de la Vendée', 'Le Labo de l'ESS', and 'Fondation Veolia'.

This screenshot shows the 'List of contributors' page of the CIS open scraper beta. It lists 20 contributors, each with a status icon (red circle with minus, yellow triangle with minus, green circle with plus), name, email, and a row of buttons for managing spiders. The contributors listed include '-DEBUG- empty spider config', '-DEBUG- test quote', '-DEBUG- test quote on one page only', 'AG2R la mondiale', 'ATLAAS', 'Apriles', 'Avisé', 'Bretagne Creative', 'Fondation Bettencourt Schueller', 'Fondation Daniel et Nina Carasso', 'Fondation EDF', 'Fondation Macif', 'Fondation RTE', 'Fondation Veolia', 'Fondation Vinci', 'Le Labo de l'ESS', and 'MOT'.

www.cis-openscraper.com

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

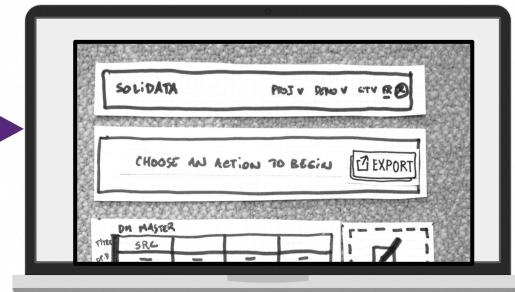
TADATA!



Open Scraper

agrégation de données publiées sur des sites web

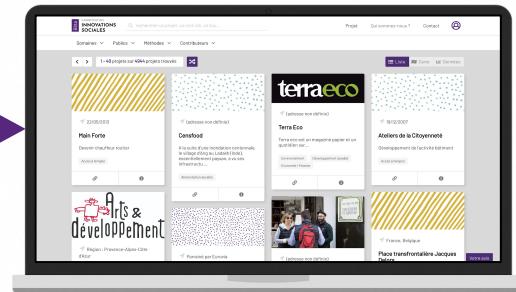
www.cis-openscraper.com



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

CIS | open scraper beta Datamodel ▾ Contributors ▾ Dataset ▾ Documentation ▾  Julien - admin  logout 

Edit your data structure

datamodel > contributors > dataset

 add a new field

résumé du projet	text	keep	opendata
tags	tags	keep	opendata
website	url	keep	opendata
auteur	text	keep	opendata
logo	image	keep	opendata
video	url	keep	private
adresse du projet	adress	keep	opendata
titre du projet	text	keep	opendata
date du projet	date	keep	commons

OPEN SCRAPER : ajouter un site à scraper

CIS | open scraper beta Datamodel ▾ Contributors ▾ Dataset ▾ Documentation ▾ Julien - admin logout 

List of contributors

datamodel > contributors > dataset

+ add a new spider refresh page run all spiders

Previous page 1 2 Next page

status	name	added by	start_urls	▶	▶	grid	edit
🚫	-- empty yoyo	  yo@yo.com	 	 test	 crawl	 view 30 item	 reset data
🚫	-DEBUG- empty spider configuration	  jparis.py@gmail.com	 	 test	 crawl	 view 20 item	 reset data
✓	-DEBUG- test quote	  jparis.py@gmail.com	 	 test	 crawl	 view 139 item	 reset data
✓	-DEBUG- test quote on one page only	  jparis.py@gmail.com	 	 test	 crawl	 view 285 item	 reset data
✓	AG2R la mondiale	  jparis.py@gmail.com	 	 test	 crawl	 view 180 item	 reset data
⚠	ATLAAS	  jparis.py@gmail.com	 	 test	 crawl	 view 93 item	 reset data
✓	Apriles	  hello@hello.fr	 	 test	 crawl	 view 533 item	 reset data
✓	Avise	  jparis.py@gmail.com	    	 test	 crawl	 view 533 item	 reset data
✓	Bretagne Creative	  jparis.py@gmail.com	 	 test	 crawl	 view 180 item	 reset data
✓	CERDD	  inaris.nv@mail.com	 	 test	 crawl	 view 533 item	 reset data

OPEN SCRAPER : éditer un scraper (XPATH)

CIS | open scraper beta Datamodel ▾ Contributors ▾ Dataset ▾ Documentation ▾  Julien - admin  logout 

EDIT THE SPIDER "-DEBUG- test quote"

datamodel > contributors > dataset

global fields

describe the website you want to crawl

name *	-DEBUG- test quote
licence *	licence
page_url *	http://quotes.toscrape.com
logo_url *	logo_url
start_urls *	http://quotes.toscrape.com/
item_xpath *	//div[@class="quote"]
next_page *	//li[@class="next"]/a/@href
deploy_list *	<input checked="" type="radio"/> There is no special button at the end of the list <input type="radio"/> There is a 'show more button' at the end of the list
deploy_list_xpath *	deploy_list_xpath
parse_follow *	<input checked="" type="radio"/> The data is complete in the list <input type="radio"/> I need to click a link in the list to show the complete data
follow_xpath *	follow_xpath
parse_reactive *	<input type="radio"/> The website is not reactive <input checked="" type="radio"/> The website is reactive

OPEN SCRAPER : éditer un scraper (XPATH)

The screenshot shows a web application for managing scrapers. At the top, there is a navigation bar with links for 'CIS | open scraper beta', 'Datamodel', 'Contributors', 'Dataset', and 'Documentation'. On the right side of the nav bar, there is a user profile for 'Julien - admin' and a 'logout' button. A green circular icon with a white question mark is also present.

In the center, a large button labeled 'EDIT THE SPIDER "-DEBUG- test quote"' is visible, along with a small question mark icon. Below this, a breadcrumb navigation shows the path: 'datamodel > contributors > dataset'.

A section titled 'advanced settings' is displayed, with a note stating 'those settings need to be in a dropdown'. It contains two input fields: one for 'LIMIT *' with the value '3' and another for 'download_delay *' with the value '0,1'.

OPEN SCRAPER : éditer un scraper (XPATH)

CIS | open scraper beta Datamodel ▾ Contributors ▾ Dataset ▾ Documentation ▾  Julien - admin  logout 

EDIT THE SPIDER "-DEBUG- test quote"

datamodel > contributors > dataset

custom fields

where do you find your data on the pages you will crawl (add xpaths) ?
you can also modify the data model [here](#) or add a field [here](#)

résumé du projet	<code>/span[@class="text"]/text()</code>
tags	<code>./div[@class="tags"]/a[@class="tag"]/text()</code>
website	<code>xpath for -website-</code>
auteur	<code>./small[@class="author"]/text()</code>
logo	<code>xpath for -logo-</code>
video	<code>xpath for -video-</code>
adresse du projet	<code>xpath for -adresse du projet-</code>
titre du projet	<code>xpath for -titre du projet-</code>
date du projet	<code>xpath for -date du projet-</code>
image(s) du projet	<code>xpath for -image(s) du projet-</code>
partenaires du projet	<code>xpath for -partenaires du projet-</code>
structure porteuse	<code>xpath for -structure porteuse-</code>
données économiques	<code>xpath for -données économiques-</code>
fiche contact	<code>xpath for -fiche contact-</code>
SIRET	<code>xpath for -SIRET-</code>

OPEN SCRAPER : prévisualiser le jeu de données "scrapées"

OPEN SCRAPER : ouvrir les données via l'API

```
{ "v": 5, "properties": 167 KB
  "status": "ok",
  "fields_open_level": { "v": 2, "properties": 1 KB
    "fields_returned": [ { "v": 12 items, 1 KB { "4 properties", 122 bytes }, { "4 properties", 110 bytes }, { "4 properties", 112 bytes }, { "4 properties", 112 bytes } ],
    "_description": "Fields returned by the query with their level of opendata"
  },
  "query_results": [ { "v": 100 items, 158 KB
    "v": 11 properties, 1 KB
    "website": [ { "v": 2 items, 71 bytes
      "mailto:mediation.crpvguyane@gmail.com",
      "http://www.crpv-guyane.org"
    ],
    "titre du projet": [ { "v": 1 item, 180 bytes
      "Médiation sociale en milieu scolaire en Guyane, une intervention efficace dans un territoire d'exception"
    },
    "link_data": "http://www.apriles.net/index.php?option=com_sobi2&sobi2Task=sobi2Details&catid=4&sobi2Id=1631&Itemid=95",
    "image(s) du projet": [ { "v": 1 item, 64 bytes
      "http://www.apriles.net/images/stories/Mediation Guyane 1.jpg"
    ],
    "tags": [ { "v": 12 items, 320 bytes
      "Gouvernance, partenariats institutionnels",
      "Diagnostic partagé",
      "Participation des habitants",
      "Emploi, Formation",
      "Intégration",
      "Parentalité",
      "Lutte contre l'exclusion sociale",
      "Protection de l'enfance",
      "Développement urbain, Vie des quartiers",
      "Jeunesse",
      "Education",
      "Nouvelles pratiques professionnelles"
    ],
    "adresse du projet": [ { "v": 1 item, 10 bytes
      "Guyane"
    },
    "added_at": 1521908306.165187,
    "structure porteuse": [ { "v": 1 item, 114 bytes
      "Centre de Ressources Politique de la Ville de Guyane (Commissariat général à l'égalité des territoires - CGET)"
    },
    "link_src": "http://www.apriles.net/index.php?option=com_sobi2&Itemid=95&limit=5&limitstart=0",
    "résumé du projet": [ { "v": 1 item, 700 bytes
      "Afin de faciliter la scolarisation, de mieux accompagner élèves et familles, de lever les incompréhensions entre les familles et ont expérimenté la mise en place d'un poste de médiateur à l'école. Ce dispositif partenarial qui réunit notamment établissements complexes et dans lequel les liens entre l'institution scolaire et les parents sont fragiles."
    },
    "_id": "5ab67a520a8286440d13a47"
  },
  "v": 11 properties, 1 KB
  "website": [ { "v": 2 items, 76 bytes
    "mailto:gilles.perole@mouans-sartoux.net",
    "http://mead-mouans-sartoux.fr"
  ]
}
```

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