

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



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 ?

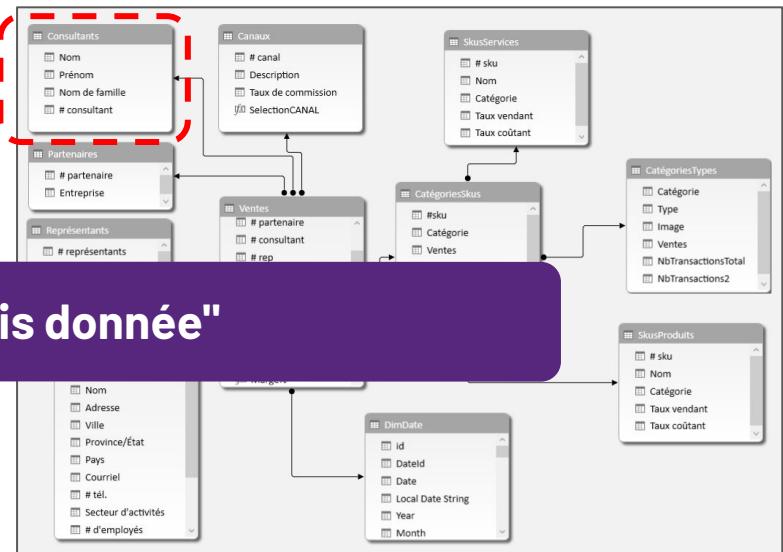
" ce qui est **connu** et qui sert de
point de départ à un raisonnement
ayant pour objet la détermination d'une
solution à un problème en relation avec
cette donnée"

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		
8	White	Big Town	West		
9	Green	Very Big	South		
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

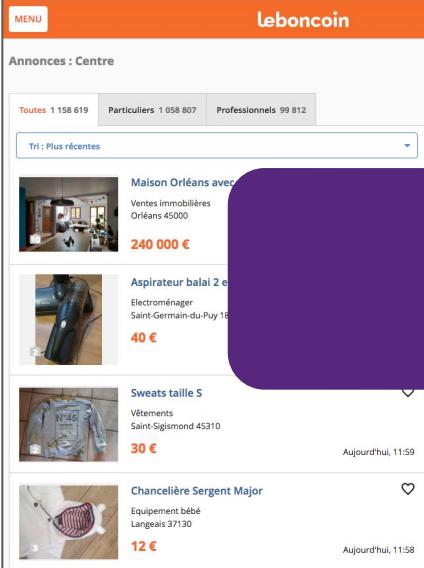
Modèle de données
relationnel (clés)



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

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

quelle que soit sa forme

tableur
csv, xls, tsv, ...

A screenshot of the SAP Fiori Launchpad. At the top, there's a search bar with the placeholder 'Search the menus (Optional)'. Below the search bar, there are several application cards. One card for 'DATAMODEL, SOLIDATA V1' is open, showing its internal structure with sub-cards for 'MM', 'CO', 'FI', and 'ME'. Other cards visible include 'API - O', 'USR', 'API - S', and 'PRU'. The SAP logo and the text 'SAP Fiori Launchpad' are at the bottom.

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

```

    "parse_for": "text",
    "xpath": "/@text",
    "parse_reactive": true,
    "item_list_xpath": "",
    "follow_xpath": ""
  },
  "scraper_config_xpath": [
    "Sod123108a2869ca6944b1b": "",
    "Sod200d5f82862760e6919": "<div[@class='tags']>/text()",
    "Sod545d508d860d10972": "",
    "Sod552d508d860d10973": "",
    "Sod553d508d860d10974": "",
    "Sod4e108b2863c97951": "",
    "Sod5a0f37a082862760e6919": "<small[@class='author']>/text()",
    "Sod5ab108b28676e48e03f": ""
  ]
}

```

pourquoi?

Quelques cas d'usages

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

Pourquoi faire du scraping de données ?

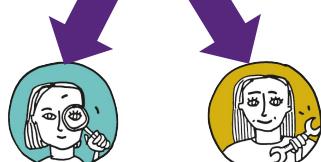


POUR LES STOCKER

rendre disponible et requêtable
la donnée moissonnée

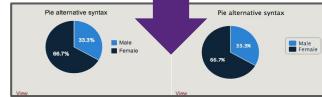
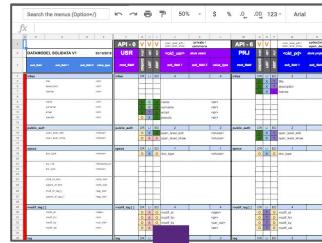
```
... "id": ObjectID("5d95754e17fe73568ab2fb"),
  "scraped_log": [
    {
      "url": "http://www.google.com",
      "status": 200,
      "content_type": "text/html; charset=UTF-8"
    }
  ],
  "scraped_config": [
    {
      "url": "http://www.google.com",
      "method": "GET"
    }
  ]
}
```

BESOIN PRIMAIRE :
constituer un ou
plusieurs jeux de
données



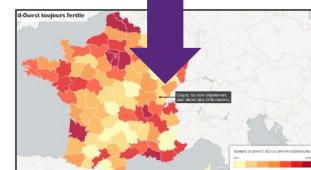
POUR LES ANALYSER

transformer la donnée en
information intelligible



POUR LES TRAVAILLER

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

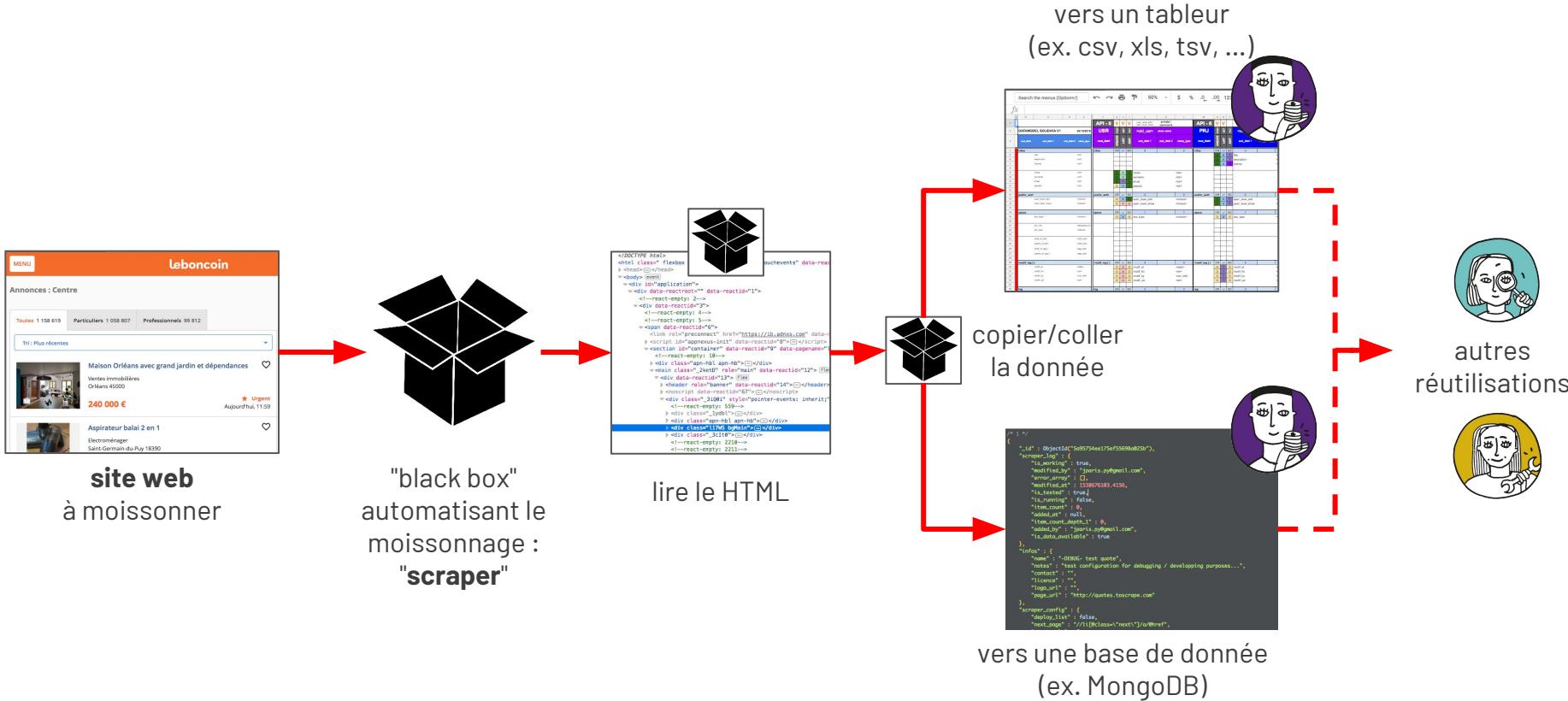


comment?

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)



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

site web A

(modèle de données A)

The screenshot shows a search results page for "politiques publiques" on the Leboncoin platform. The interface includes a header with navigation links like "Déposer une annonce", "Offres", "Demandes", "Mes favoris", "Boutiques", and "Messages". A search bar is at the top right. Below it, there are filters for "Prix entre" (Price range) and a dropdown for "Centre". The main content area displays several classified ads:

- Livre officiel 1986 République pop. de Chine (40€)
- 3 romans de Philippe Alexandre (5€)
- L'action sociale aujourd'hui - Jacques Ladsous (5€)
- Livres pass foucher qcim culture générale (3€)

Each listing includes a thumbnail, title, price, location, and date.

site web B
(modèle de données B)

The screenshot shows a search results page for "politiques publiques" on Amazon. The interface includes a header with "Amazon Books" and a search bar. The main content area displays several book listings:

- Politiques publiques (Thémis) (French Edition) by Yves Mény and Jean-Claude Theisig (Kindle Edition, 14€)
- Charlie, notre 11 septembre - MICHEL ONFRAY (Kindle Edition, 1€)
- Question du public: la politique de renseignement by Michel Onfray (Kindle Edition, 1€)
- Rapport d'information sur l'évaluation des dispositifs publics d'aide à la création by Assemblée nationale and Comité d'évaluation et de contrôle des politiques publiques (Kindle Edition, 1€)

Each listing includes a thumbnail, title, author, price, and availability information.

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"

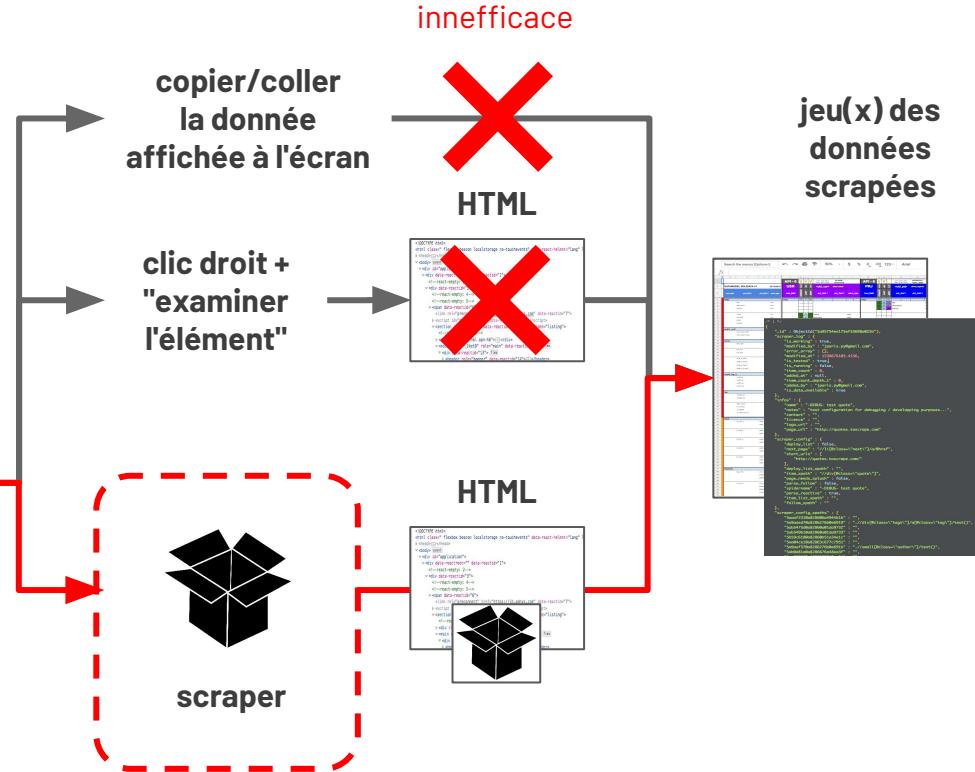
site web à scraper

FILTRER PAR POLITIQUE SOCIALE MOTS-CLEFS RÉGIONS/DÉPARTEMENTS PUBLIC VISÉ

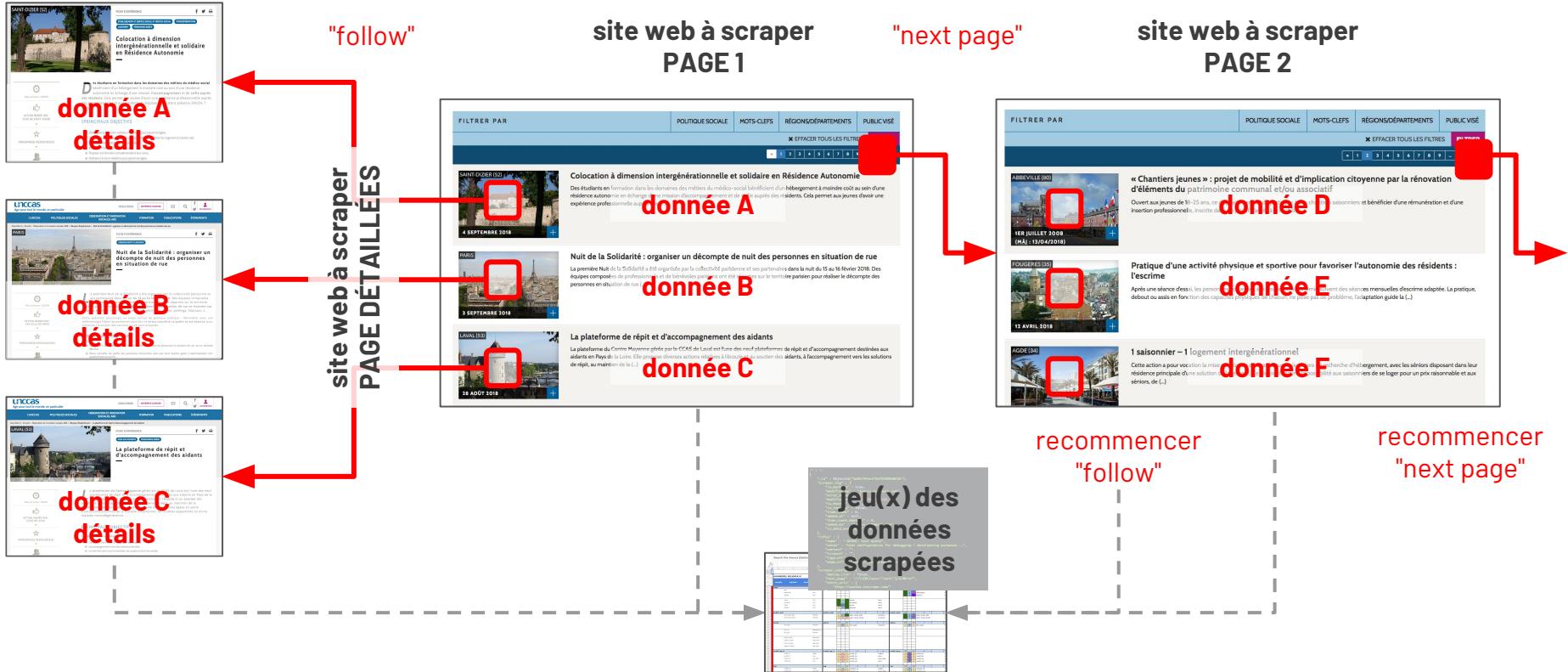
SAINT-DIZIER (52) Colocation à dimension intergénérationnelle et solidaire en Résidence Autonomie
Des étudiants en formation dans les domaines des métiers du médico-social bénéficient d'un hébergement à moindre coût au sein d'une résidence autonomie en échange d'une mission d'accompagnement et de veille auprès des résidents. Cela permet aux jeunes d'avoir une expérience professionnelle auprès des (...)

PARIS 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 la collectivité parisienne et ses partenaires dans la nuit du 15 au 16 février 2016. Des équipes composées de professionnels et de bénévoles parisiens ont été réparties sur le territoire parisien pour réaliser le décompte des personnes en situation de rue (...)

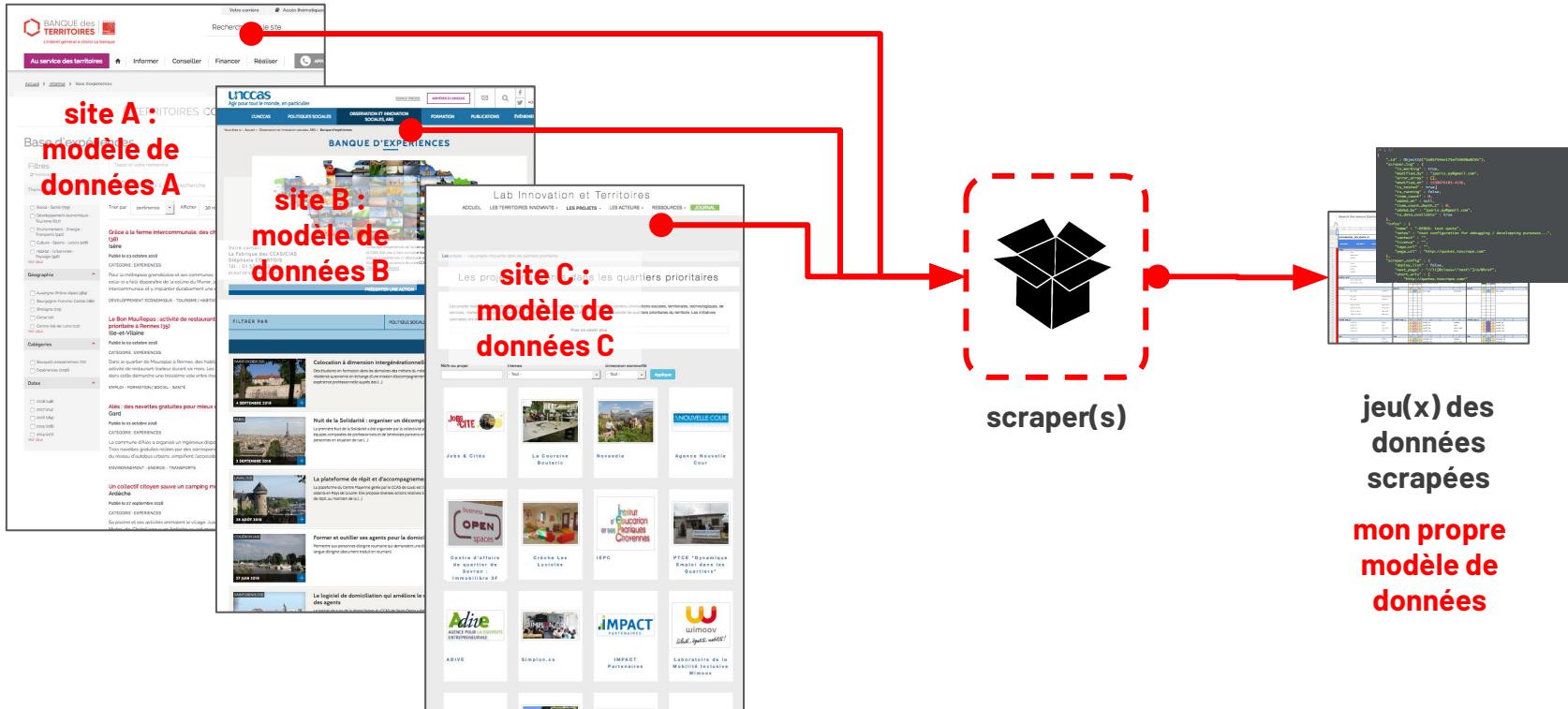
LAVAUL (53) La plateforme de répit et d'accompagnement des aidants
La plateforme du Centre Mayennais gérée par le CCAS de Laval est l'une des neuf plateformes de répit et d'accompagnement destinées aux aidants en Pays de la Loire. Elle propose diverses actions relatives à l'écoute et au soutien des aidants, à l'accompagnement vers les solutions de répit, au maintien de la (...)



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...

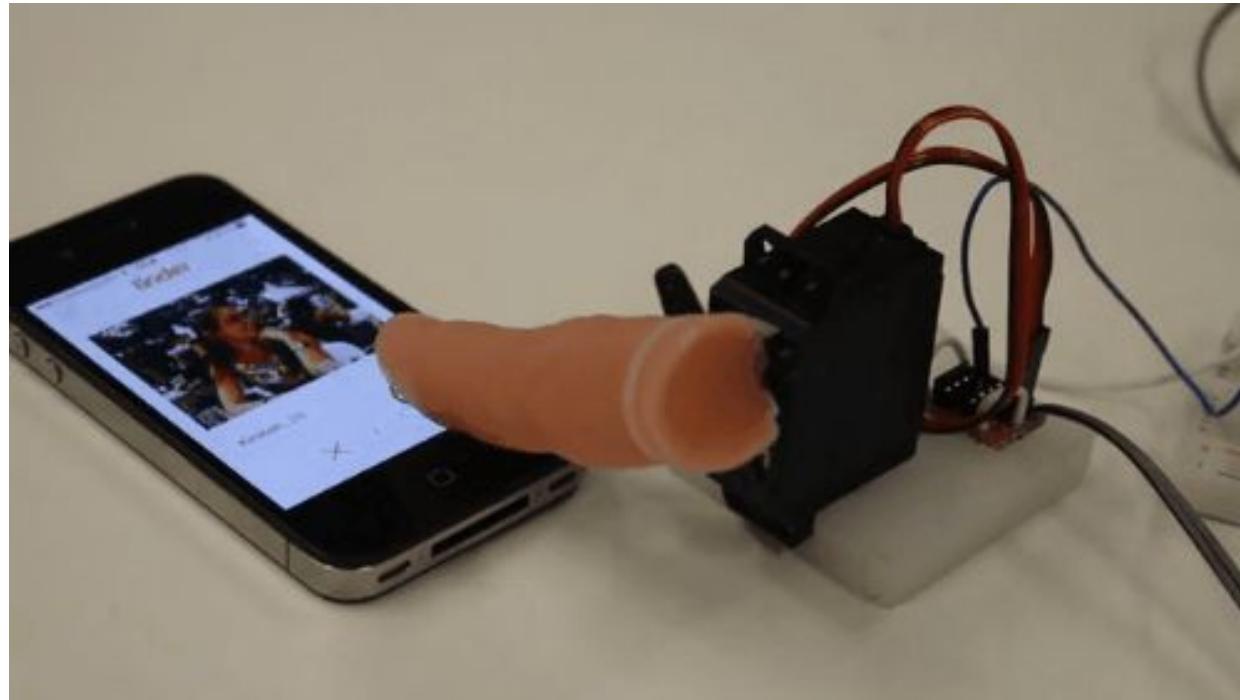
The screenshot shows a website's pricing section titled "Pricing & Plans". It features three main plan options: "Essential", "Essential Annual", and "Premium Plans".

- Essential:** \$299 Monthly. Includes a "BUY ONLINE" button.
- Essential Annual:** \$1,999 Annual (56% savings). Includes a "BUY ONLINE" button.
- Premium Plans:** Contact Us (SaaS or managed service). Includes a "CONTACT US" button. A red box highlights this plan, and a red arrow points from it to a callout box containing the text "SaaS 60,000 URL queries per year 10,000 downloads (images/files) per year".

Below the plans, there are additional sections: "Get these amazing features:" (with a list of bullet points), "All Essential PLUS:", and "All Essential Annual PLUS:" (which includes "API access").

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)

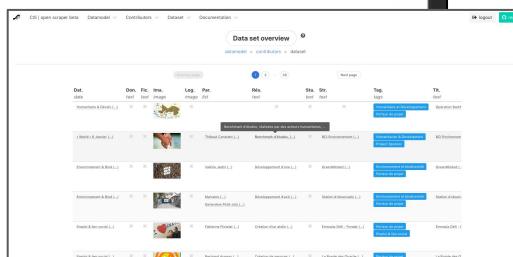


un web application **open source** en développement continu
pour moissonner des données sur différents sites Internet

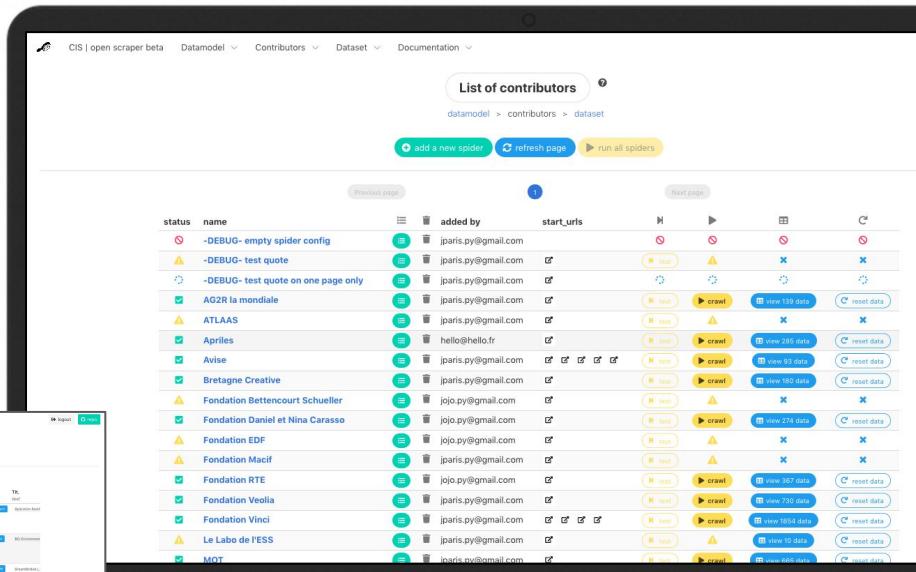
OPEN SCRAPER : agréger des données publiées sur des sites

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



www.cis-openscraper.com



status	name	added by	start_urls	actions
✗	-DEBUG- empty spider config	jparis.py@gmail.com	✗	[crawl] [reset]
⚠	-DEBUG- test quote	jparis.py@gmail.com	✗	[crawl] [reset]
⚠	-DEBUG- test quote on one page only	jparis.py@gmail.com	✗	[crawl] [reset]
✓	AG2R la mondiale	jparis.py@gmail.com	✗	[crawl] [reset]
⚠	ATLAAS	jparis.py@gmail.com	✗	[crawl] [reset]
✓	Aprilès	hello@hello.fr	✗	[crawl] [reset]
✓	Avisé	jparis.py@gmail.com	✗	[crawl] [reset]
✓	Bretagne Creative	jparis.py@gmail.com	✗	[crawl] [reset]
⚠	Fondation Bettencourt Schueller	jojo.py@gmail.com	✗	[crawl] [reset]
✓	Fondation Daniel et Nina Carasso	jojo.py@gmail.com	✗	[crawl] [reset]
⚠	Fondation EDF	jojo.py@gmail.com	✗	[crawl] [reset]
⚠	Fondation Macif	jparis.py@gmail.com	✗	[crawl] [reset]
✓	Fondation RTE	jojo.py@gmail.com	✗	[crawl] [reset]
✓	Fondation Veolia	jparis.py@gmail.com	✗	[crawl] [reset]
✓	Le Labo de l'ESS	jparis.py@gmail.com	✗	[crawl] [reset]
✗	MOT	jparis.py@gmail.com	✗	[crawl] [reset]

OPEN SCRAPER : une webapp de scraping en open source

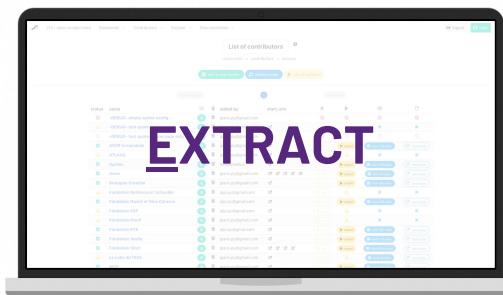
The screenshot shows the GitHub repository page for 'entrepreneur-interet-general / OpenScraper'. The repository has 150 commits, 2 branches, 0 releases, and 2 contributors. It uses the MIT license. A recent commit by 'JulienParis' was a merge pull request from 'otakuto/typo'. The repository has 13 issues, 0 pull requests, 2 projects, and 19 stars. It is public and can be forked. The URL is <http://www.cis-openscraper.com/>.

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

README.md	update readme	6 months ago
requirements.txt	fixing stuff before https	5 months ago

OPEN SCRAPPER 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

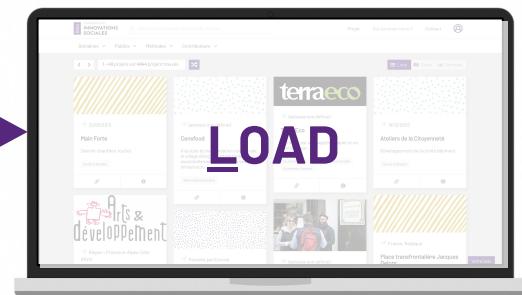
www.cis-openscraper.com



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

CIS | open scraper beta

an online open source free webscraper
you can customize to brush and gather data from (almost) any website

welcome back, Julien (jparis.py@gmail.com) ... you have the **admin** level

set a datamodel add contributors see the dataset visualize share

<http://www.cis-openscraper.com/>

17 custom fields

7354 structured data

21 websites

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
tags	tags	keep
website	url	keep
auteur	text	keep
logo	image	keep
video	url	keep
adresse du projet	adress	keep
titre du projet	text	keep
date du projet	date	keep

opendata

opendata

opendata

opendata

opendata

private

opendata

opendata

commons



OPEN SCRAPER : ajouter un site à scraper

Screenshot of the CIS | open scraper beta interface showing the 'List of contributors' page.

The top navigation bar includes: CIS | open scraper beta, Datamodel, Contributors, Dataset, Documentation, Julien - admin, logout, and a user icon.

The main title is 'List of contributors' with a subtitle 'datamodel > contributors > dataset'.

Buttons at the top: 'add a new spider' (highlighted with a large black arrow), 'refresh page', and 'run all spiders'.

Pagination: Previous page, page 1 (highlighted), page 2, Next page.

status	name	added by	start_urls	test	crawl	view 30 item	reset data
∅	-- empty yoyo		yo@yo.com				
∅	-DEBUG- empty spider configuration		jparis.py@gmail.com				
✓	-DEBUG- test quote		jparis.py@gmail.com				
✓	-DEBUG- test quote on one page only		jparis.py@gmail.com				
✓	AG2R la mondiale		jparis.py@gmail.com				
⚠	ATLAAS		jparis.py@gmail.com				
✓	Apriles		hello@hello.fr				
✓	Avise		jparis.py@gmail.com				
✓	Bretagne Creative		jparis.py@gmail.com				
✓	CERDD		jparis.py@gmail.com				
✓	Coorace		jparis.py@gmail.com				

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 content
follow_xpath *	follow_xpath
parse_reactive *	<input checked="" type="radio"/> The website is not reactive <input 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 'Julien - admin', a 'logout' button, and a green circular icon with a white question mark. Below the nav bar, a main title 'EDIT THE SPIDER "-DEBUG- test quote"' is centered, with a question mark icon to its right. Underneath the title, a breadcrumb navigation shows 'datamodel > contributors > dataset'. The main content area is titled 'advanced settings' and contains two input fields: 'LIMIT *' with value '3' and 'download_delay *' with value '0,1'. A large black cursor arrow points towards the bottom left of the 'advanced settings' box.



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>



OPEN SCRAPER : prévisualiser le jeu de données "scrapées"

Data set overview								
datamodel > contributors > dataset								
Previous page Next page								
Tag.	Web.	Aut.	Log.	Adr.	Tit.	Ima.	Str.	Sir.
tags	url	text	image	adress	text	image	text	text
-	-	-	-	-	-	-	-	-
opendata	opendata	opendata	opendata	opendata	opendata	opendata	opendata	opendata
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	    	    	Guyane (...)	Médiation sociale en (...)		Centre de Ressources (...)	 	
Emploi, Formation Développement local rural Gouvernance, partenariats institutionnels Jeunesse Education Nouvelles pratiques professionnelles	    	    	Provence-Alpes-Côte (...)	A Mouans-Sartoux, la (...)		Ville de Mouans-Sart (...)	 	
Vie des séniors Vie en établissement Emploi, Formation Soutien aux aidants Lutte contre l'exclusion sociale Gouvernance, partenariats institutionnels	    	    	Pays de la Loire (...)	ENVIE Autonomie 49 d (...)		ENVIE Autonomie 49 (...)	 	

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, 110 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, 100 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)"
      },
      "url": "http://www.apriles.net/index.php?option=com_sobi2&sobi2Task=sobi2Details&catid=4&sobi2Id=1631&Itemid=95"
    ]
  }
}
```

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