

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



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 ?

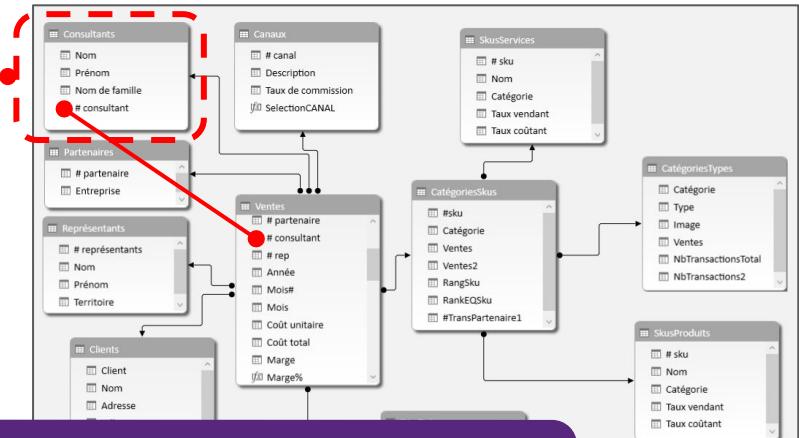
" ce qui est **connu** et qui sert de
point de départ à un raisonnement
ayant pour objet la détermination d'une
solution à un problème en relation avec
cette donnée"

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				
12	Gold				

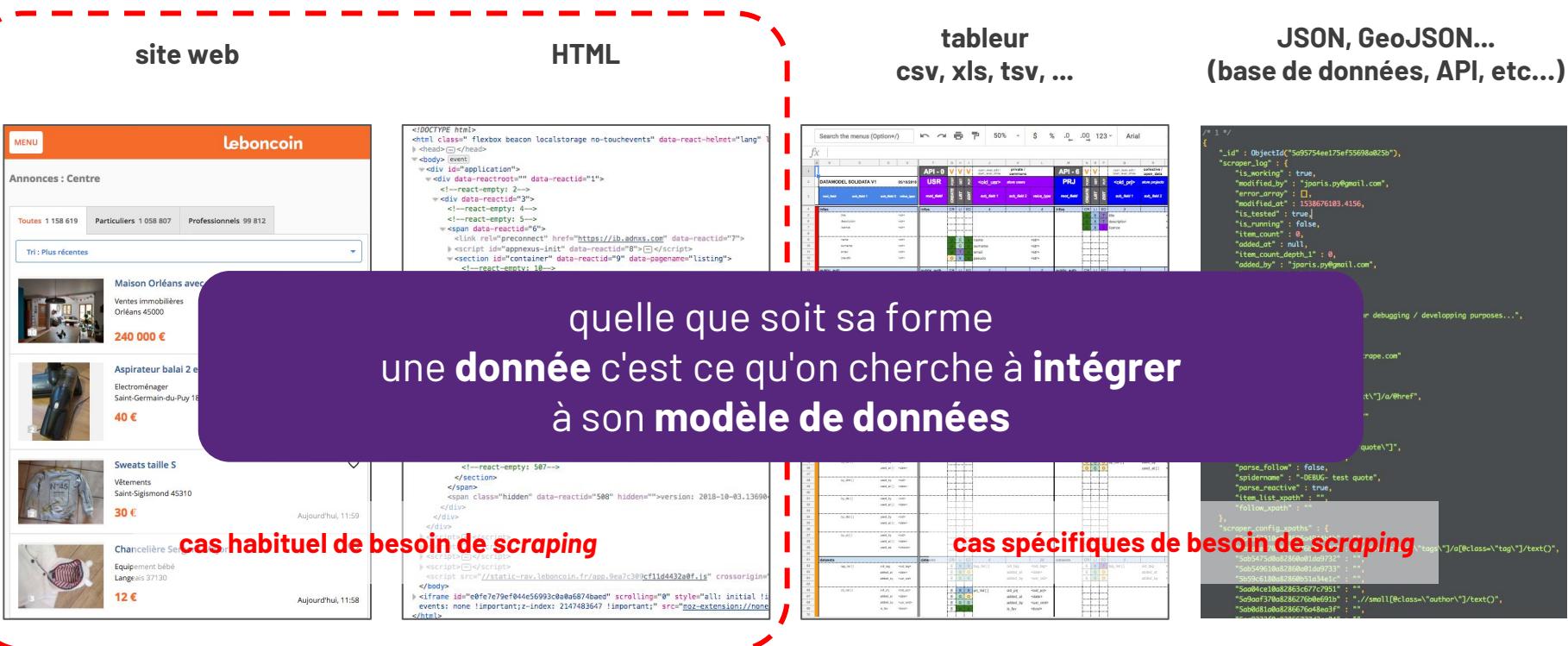
Modèle de données
relationnel (clés étrangères, index)



" 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 ?



pourquoi?

Quelques cas d'usages

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

Pourquoi faire du scraping de données ?

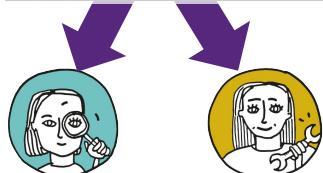


POUR LES STOCKER

rendre disponible et requêtable
la donnée moissonnée

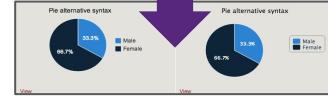
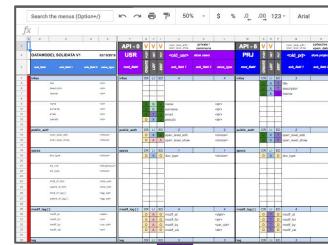
```
... "id": ObjectID("5d95754e17fe73568ab2fb"),
  "scraped_log": [
    {
      "url": "http://www.google.com",
      "date": "2019-01-01T00:00:00Z"
    }
  ],
  "scraped_config": {
    "test": true,
    "log_level": "INFO",
    "log_file": "scraper.log",
    "use_proxy": false,
    "next_page": "/<1/Bla bla</next>/<where*>"
  }
}
```

BESOIN PRIMAIRE :
constituer un ou plusieurs jeux de données



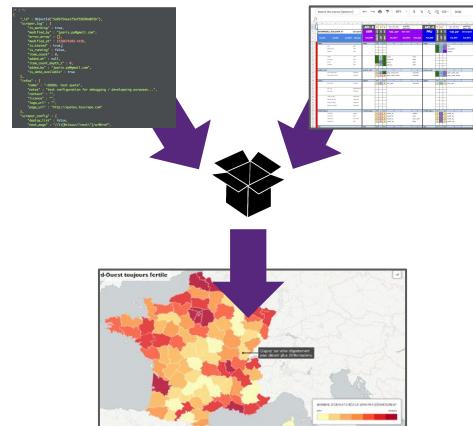
POUR LES ANALYSER

transformer la donnée en information intelligible



POUR LES TRAVAILLER

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



comment?

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

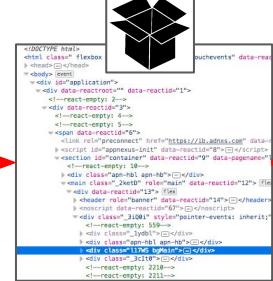
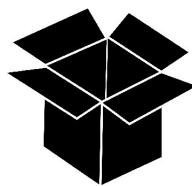
Mise en pratique de la "moisson de données en ligne" dans un cas habituel

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



site web A
à moissonner

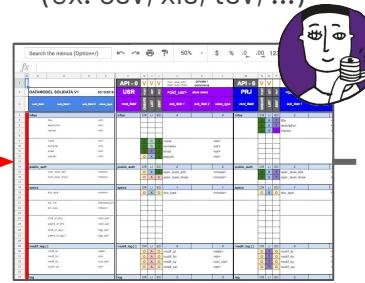
"black box"
automatisant le
moissonnage :
le "**scraper**"



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



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



puis copier/coller
la donnée ...



autres
réutilisations



```
> 1 />
  "_id" : ObjectId("5e95754eef75ef5569ab250"),
  "scraper_log" : [
    {
      "is_error" : true,
      "modified_by" : "portis.py@gmail.com",
      "error_array" : [
        {
          "id" : "5e95754eef75ef5569ab250",
          "is_crawled" : true,
          "is_parsed" : false,
          "item_count" : 0,
          "added_at" : null,
          "item_id" : 4,
          "added_by" : "portis.py@gmail.com",
          "is_auto_available" : true
        }
      ],
      "info" : [
        {
          "name" : "HOMME test portis",
          "notes" : "test configuration for debugging / developing purposes...",
          "contact" : "+33666666666",
          "location" : "Paris",
          "log_url" : "",
          "page_url" : "http://quotes.toscrape.com"
        }
      ],
      "scraper_config" : [
        {
          "display_label" : false,
          "next_page" : "/(10class)\\" next\"]/o/ref"
        }
      ]
    }
  ]
```

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



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

site web A

(modèle de données A)

The screenshot shows a search results page for "politiques publiques" on the Leboncoin website. The interface includes a header with navigation links like "DÉPOSER UNE ANNONCE", "OFFRES", "PRIMANCES", "MES BIENS", "BOUTIQUES", and "MESSAGES". A search bar at the top has dropdowns for "Recherche dans le titre uniquement" and "Annonces Urgentes uniquement", and a field for "Centre" with a placeholder "Write ou code postal". Below the search bar are filters for "Prix entre" with "Prix min" and "Prix max" dropdowns. The main content area displays a grid of items under the heading "annonces Livre, roman, BD occasion « politiques publiques » : Centre". Each item card includes a thumbnail, the title, location (e.g., Tours), price (e.g., 40€), and a timestamp (e.g., 30 sept, 19:55). A red dashed box highlights the first item in the list.

site web B
(modèle de données B)

The screenshot shows a search results page for "politiques publiques" on the Amazon website. The interface includes a header with "Amazon Books" and a search bar with the query "politiques publiques". Below the search bar are filters for "Show results for" (Books, Kindle Edition, Free Shipping, etc.) and "Refine by" (Amazon Prime, Book Language, Condition). The main content area displays a grid of items under the heading "1-16 of 261 results for "politiques publiques"". Each item card includes a thumbnail, the title, author (e.g., Yves Ménny and Jean-Claude Theisig), price (e.g., \$26.00), and a timestamp (e.g., Sep 16, 2015). A red dashed box highlights the first item in the list.

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

A table titled "données moissonnées, structurées et agrégées (mon modèle de données C)" showing aggregated data from two sources. The columns are "source", "titre", "auteur", and "prix". Red dashed boxes highlight specific entries from the previous screenshots.

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

FILTRER PAR

POLITIQUE SOCIALE MOTS-CLEFS RÉGIONS/DÉPARTEMENTS PUBLIC VISÉ

EFFACER TOUTES LES FILTRES FILTRER

SAINTE-DUZIER (52) 4 SEPTEMBRE 2018 +

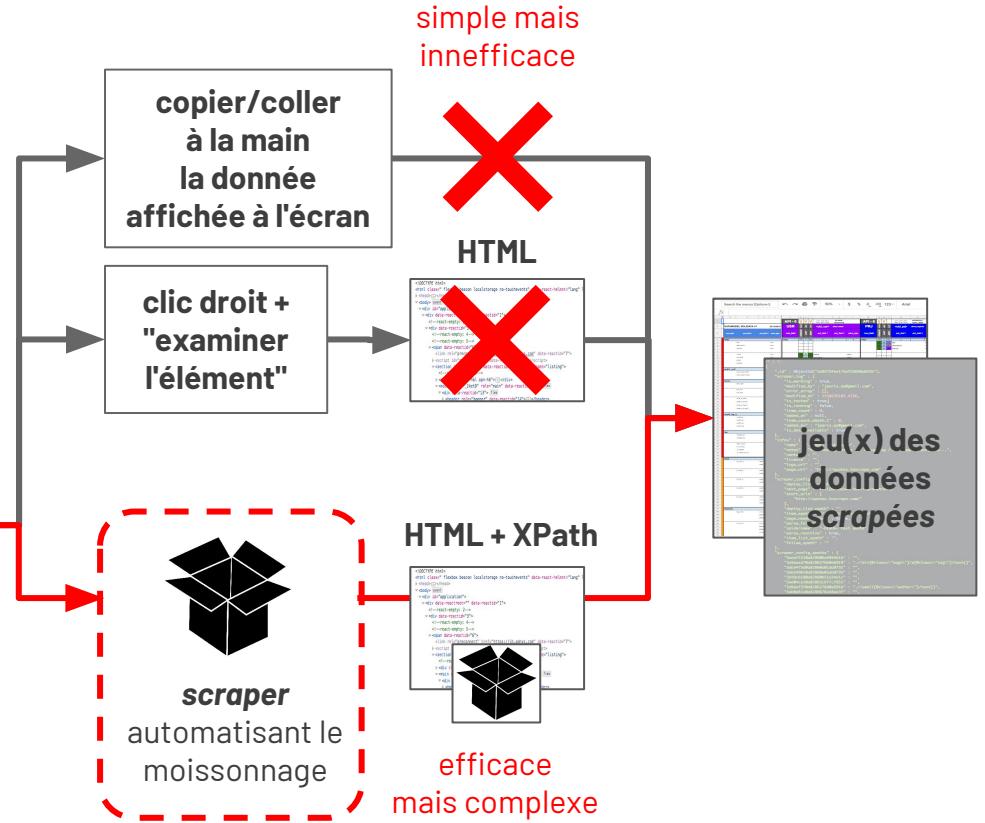
Nuit de la Solidarité : organiser un décompte de nuit des personnes en situation de rue

La première Nuit de la Solidarité a été organisée par 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 (...) **donnée A**

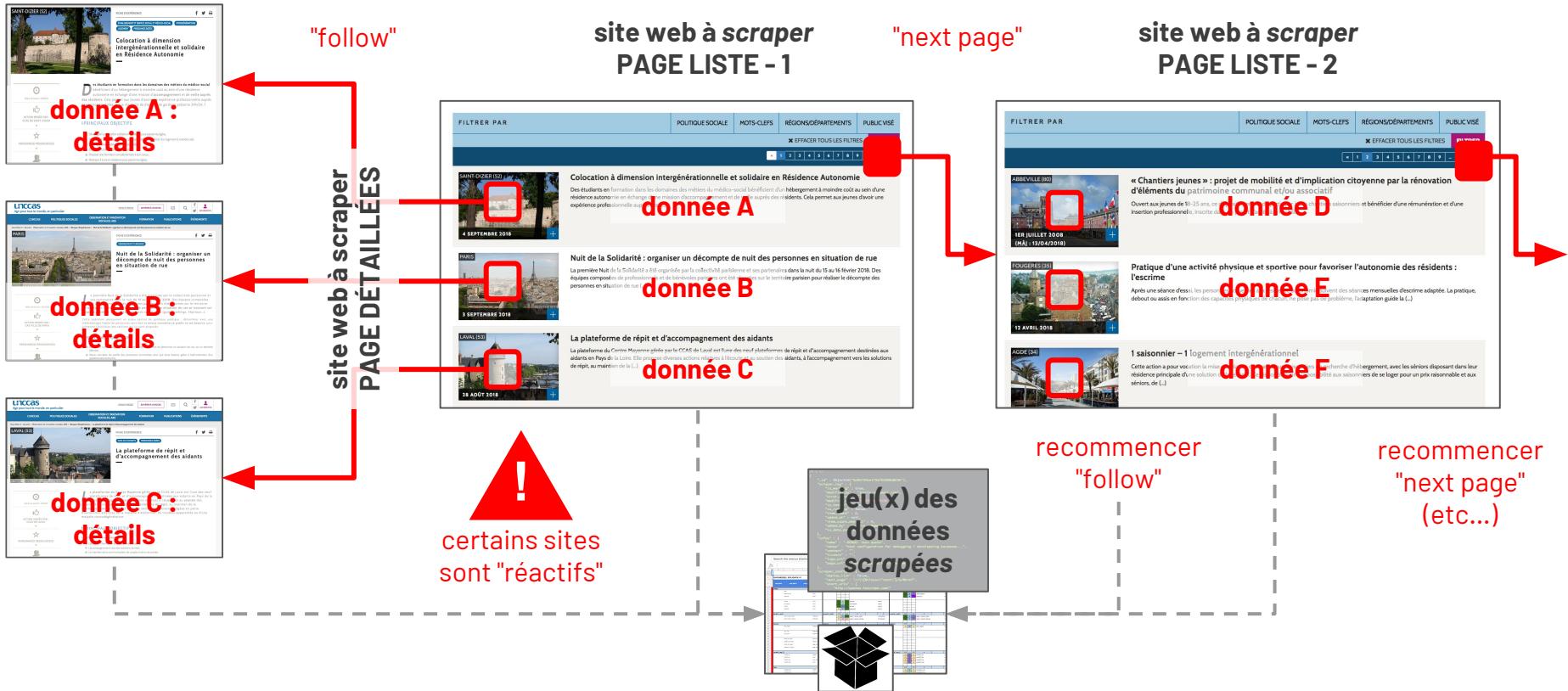
PARIS 3 SEPTEMBRE 2018 +

La plateforme de répit et d'accompagnement des aidants

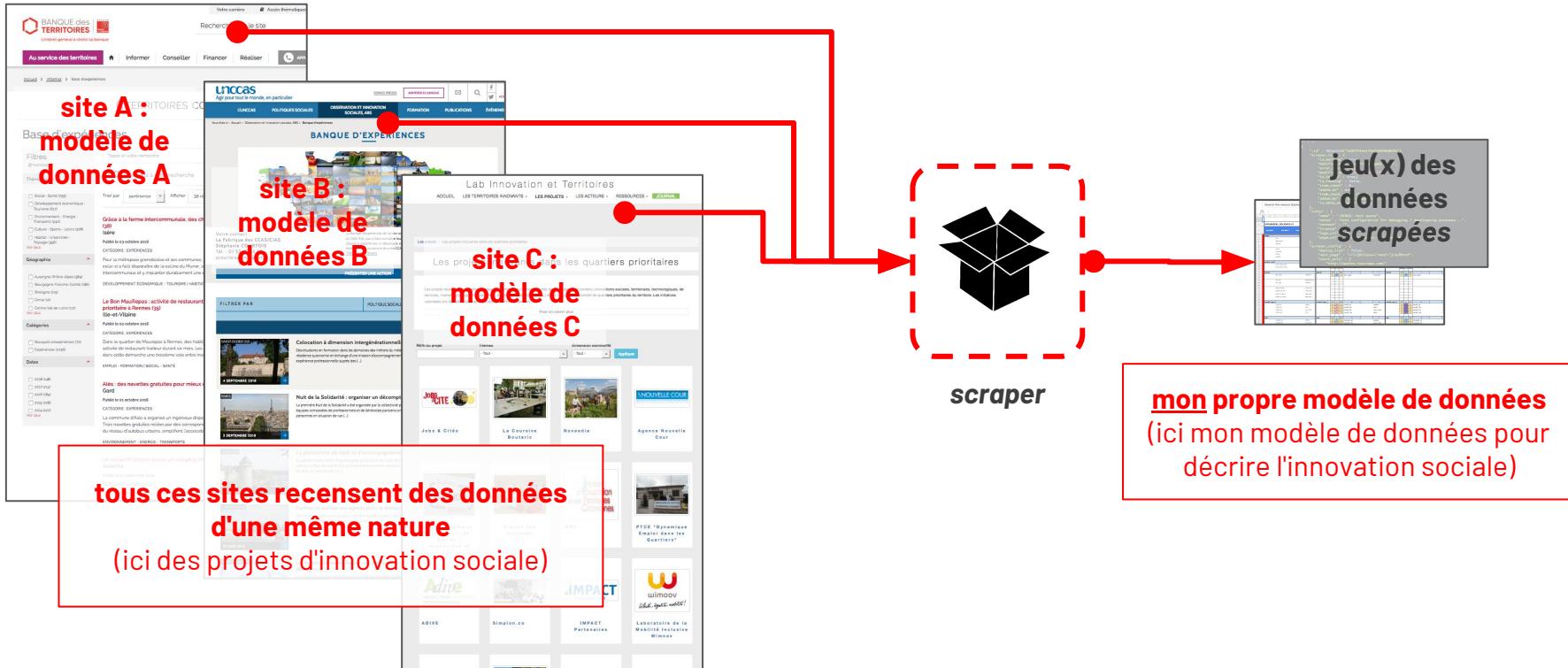
La plateforme du Centre Mayenne gérée par le CCAS de Laval est l'une des neuf plateformes de répit et d'accompagnement destinées aux aidants en Pays de la Loire. Elle propose diverses actions relatives à l'écoute et au soutien des aidants, à l'accompagnement vers les solutions de répit, au maintien de la (...) 28 AOÛT 2018 +



Problème 2 : Automatiser la navigation dans un site web



Problème 3 : Adapter le scraping à des sites web 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. At the top, there's a navigation bar with links for Products, Solutions, Pricing, Resources, About Us, Blog, Careers, Contact, and a search icon. Below the navigation is a banner with the text "Pricing & Plans". The main content area displays three pricing plans:

- Essential**: \$299 Monthly. A "BUY ONLINE" button is below it.
- Essential Annual**: \$1,999 Annual (56% savings). A "BUY ONLINE" button is below it. This plan is highlighted with a red dashed box.
- Premium**: (partially visible)

To the right of the plans, there's a "Contact Us" section for "SaaS or managed service" with a "CONTACT US" button. Above this, a red box encloses the following text:

SaaS
60,000 URL queries per year
10,000 downloads (images/files) per year

A red arrow points from the "Annual" price to the SaaS information.

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

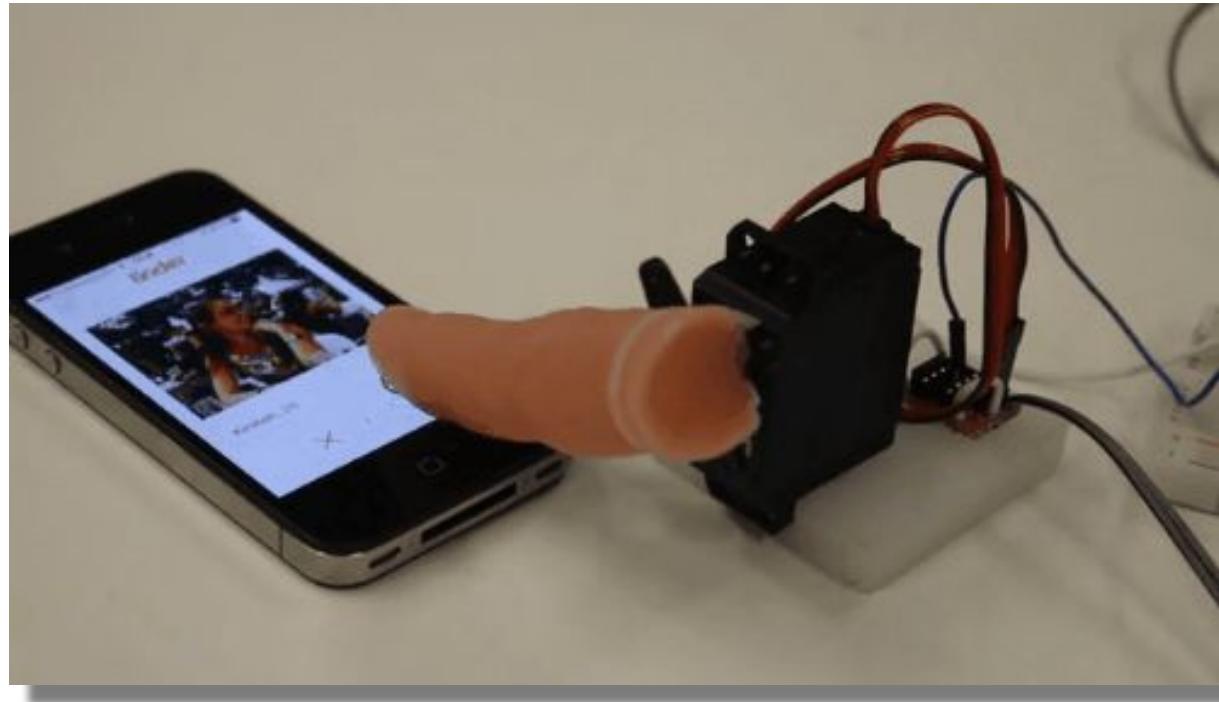
comment?

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-3rgXByefj5zvCc0d0M>

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 requêteable 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*

The screenshot shows a dashboard titled "Dataset Overview". It lists several datasets with columns for ID, Name, Description, Status, and Actions. One dataset is highlighted in yellow: "Mémoires & Mémoires" (Status: OK). Other datasets include "Documentaire & Mémoires" (Status: OK), "Environnement & Développement Durable" (Status: OK), and "Environnement & Développement Durable" (Status: OK).

The screenshot shows a "List of contributors" page. It displays a table with columns: status, name, added by, start_urls, and actions. The table lists 18 entries, each with a status icon (green, yellow, red), a name (e.g., "-DEBUG-", "AG2R la mondiale", "Apriles", "Avisé", "Bretagne Creative", "Fondation Bettencourt Schueller", "Fondation Daniel et Nina Carasso", "Fondation EDF", "Fondation Macif", "Fondation RTE", "Fondation Veolia", "Le Labo de l'ESS", "MOT"), an "added by" column (e.g., "iparis.py@gmail.com", "jojo.py@gmail.com", "hello@hello.fr", "iparis.py@gmail.com", "iparis.py@gmail.com", "jojo.py@gmail.com", "jojo.py@gmail.com", "jojo.py@gmail.com", "iparis.py@gmail.com", "iparis.py@gmail.com"), a "start_urls" column (e.g., "http://www.ag2r.com", "http://www.ag2rlamondiale.com", "http://www.apriles.org", "http://avisecitoyen.org", "http://bretagnecreative.org", "http://fondation.bettencourt-schueller.org", "http://fondation-daniel-nina-carasso.org", "http://fondation-edf.org", "http://fondation-macif.org", "http://fondation-rte.org", "http://fondation-veolia.org", "http://lelabo-ess.org", "http://mot-association.org"), and a series of buttons for each row: "view", "crawl", "reset data", and "refresh page".

www.cis-openscraper.com

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

The screenshot shows the GitHub repository page for `entrepreneur-interet-general / OpenScraper`. The page features logos for Python, Tornado, Scrapy, Selenium, and mongoDB. The Python logo is prominently displayed at the top left. To its right is the Tornado logo with the word "Tornado" in large blue letters. Below these are logos for Scrapy (a green circle with a brush icon) and Selenium (a blue square with a green checkmark and a green Python icon). At the bottom is the mongoDB logo (a green leaf). The repository has 18 stars, 19 forks, and 4 issues. It includes sections for code, pull requests, issues, marketplace, and explore. A sidebar on the left lists commits from JulienParis, including merges and fixes. A purple banner at the bottom contains the URL <https://github.com/entrepreneur-interet-general/OpenScraper>.

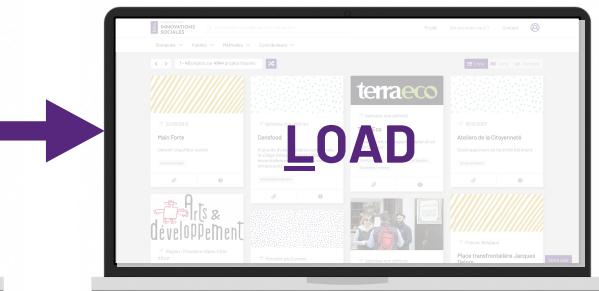
serveur :  python™
An open source library for scraping: tornado's a public service for webscraping <http://www.cs-openscraper.com/>

scrappers :  Scrapy

base de données :  mongoDB

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

OPEN SCRAPPER fait partie d'une suite logicielle type ETL en développement continu et en *open source* : TADATA!



(en développement)

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

instance dédiée pour le projet CIS : <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 application 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 (ajouts de 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 type="radio"/> The website is not reactive <input checked="" type="radio"/> The website is reactive

OPEN SCRAPER : éditer un scraper (ajouts de XPath)

The screenshot shows a web application interface 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 top bar, there is a user profile icon for 'Julien - admin', a 'logout' button, and a green circular icon with a white question mark. Below the navigation 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 the value '3' and 'download_delay *' with the value '0,1'. A large black cursor arrow points towards the bottom left of the 'advanced settings' box.

advanced settings
those settings need to be in a dropdown

LIMIT *	3
download_delay *	0,1

OPEN SCRAPER : éditer un scraper (ajouts de 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 -lo-</code>
video	<code>xpath for -vid-</code>
adresse du projet	<code>xpath for -adres-</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 SCRAPPER : prévisualiser le jeu de données "scrapées"

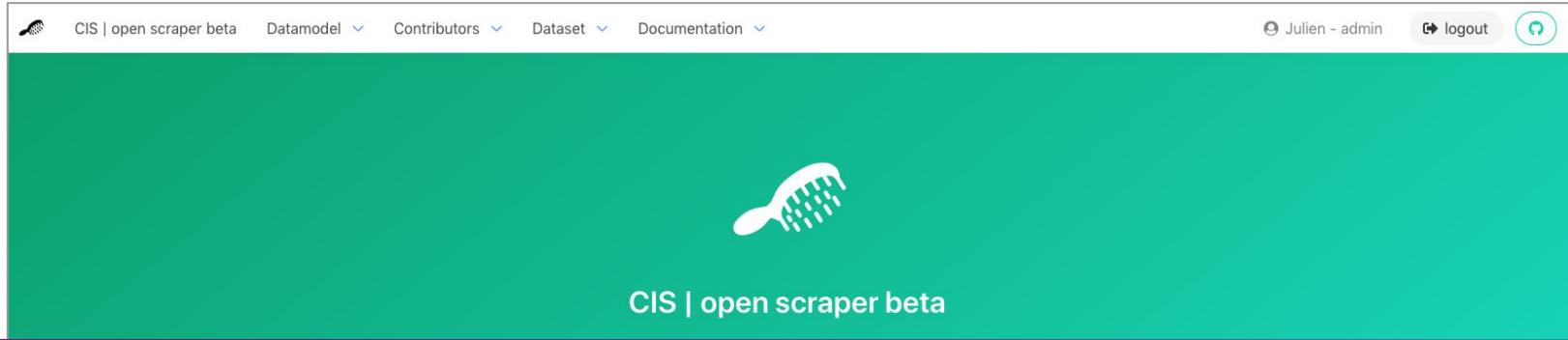
Data set overview								
datamodel > contributors > dataset								
Previous page Next page								
Tag.	Web.	Aut.	Log.	Adr.	Tit.	Ima.	Str.	Sir.
tags	url	text	image	adress	text	image	text	text
-	-	-	-	-	-	-	-	-
opendata	opendata	opendata	opendata	opendata	opendata	opendata	opendata	opendata
Gouvernance, 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 en JSON 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)"
    ],
    "titre du projet": [ { "v": 7 items, 186 bytes
      "Médiation sociale en milieu scolaire en Guyane, une intervention efficace dans un territoire d'exception"
    }
  ]
}
}, { "v": 11 properties, 1 KB
  "website": [ { "v": 2 items, 76 bytes
    "mailto:gilles.perole@mouans-sartoux.net",
    "http://mead-mouans-sartoux.fr"
  ],
  "titre du projet": [ { "v": 7 items, 186 bytes
    "Médiation sociale en milieu scolaire en Guyane, une intervention efficace dans un territoire d'exception"
  }
]
}
```

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

OPEN SCRAPER : démo de la version 1.0



CIS | open scraper beta

Datamodel ▾

Contributors ▾

Dataset ▾

Documentation ▾

Julien - admin

logout



CIS | open scraper beta

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