

**Day 3 & 4 : Datathon**

**Coding Week Sciences Po**

**Timothée Gidoin - Nicolas Vogtenberger**

**January 2019**

# Before we start : reminder

Find those slides online : <https://github.com/Gidoin/Codingweek>

Notepad beginners coding session : <https://hackmd.io/Fg2Np-hSTZuUuCD1-okgjQ?both>

Notepad advanced coding session : <https://hackmd.io/N13Qq0-gQyGUeeuf3QNeiQ?both>

Dataactivist productions are freely reusable under the terms of the licence [Creative Commons 4.0 BY-SA](#).

[Joel Gombin email address](#) // [Sylvain Lapoix email address](#)

# Before we start : feedbacks

Did you enjoy the first 2 days ?



# Before we start : feedbacks

What could be improved ?



# **Before we start : feedbacks**

**What could be improved ?**

Apart from getting schwifty...

# Before we start : who we are



**Timothée Gidoin**



**Nicolas Vogtenberger**

# **Let's get started**

# Datathon organisation

## 3 goals :

- 1) Make you work, by team, on real public policies issues **through a data point of view**

# Datathon organisation

## 3 goals :

- 1) Make you work, by team, on real public policies issues **through a data point of view**
- 2) **Design a possible solution** to the issue you analysed

# Datathon organisation

## 3 goals :

- 1) Make you work, by team, on real public policies issues **through a data point of view**
- 2) **Design a possible solution** to the issue you analysed
- 3) **Share your results** both online (Github & Medium) and offline

# Datathon organisation

## 3 goals :

- 1) Make you work, by team, on real public policies issues **through a data point of view**
- 2) **Design a possible solution** to the issue you analysed
- 3) **Share your results** both online (Github & Medium) and offline



# Datathon organisation

## Day 1 : data analysis

- 9h : Let's get started (*30 minutes*)
- 9h30 : Introduction to open data (*30 minutes*)
- 10h : Launch of the datathon : unveiling the topics & constitution of the groups (*30 minutes*)
- 10h30 : Looking for sources & datasets (*2 hours*)
- 12h30 : 1 hour Break
- 13h30 : Brainstorming & Data analysing (*2 hours*)
- 15h30 : Graphical data representation (*2 hours*)

**17h30 - 20h** : *Optional* further session for volunteers

# Datathon organisation

## Day 2 : designing a solution

- **9h15** : Introduction to design thinking / workshop (2,75 *hours*)
- 11h : Sprint to formalize both your analysis and solution (4,5 *hours* including 1 hour break)
  - Prepare few slides,
  - Write a Medium article,
  - Tidy and put your code on Codingweek repo Github
- 15h30 : Official restitution in front of the jury (5 minutes per project + 5 minutes Q/A) (2 *hours*)
- 16h45 : Grand Jury decision and prize giving (15 *minutes*)

# **Introduction to open data**

# Open data principles

## 1. Completeness

Datasets released by the government should be as complete as possible, reflecting the entirety of what is recorded about a particular subject.

All raw information from a dataset should be released to the public, except to private information and information that may be sensitive for national safety

# Open data principles

## 2. Primacy / Raw data

Datasets released by the government should be **primary source data**

# Open data principles

## 3. Timely data

Datasets released by the government should be available to the public **as soon as possible**

# Open data principles

## 4. Ease of Physical and Electronic Access

Datasets released by the government should be as accessible as possible, with accessibility defined as **the ease with which information can be obtained**, whether through physical or electronic means

# Open data principles

## 5. Machine readability

Machines can handle certain kinds of inputs much better than others. Information shared in the widely-used PDF format, for example, is very difficult for machines to parse

Thus, information should be stored in widely-used file formats that easily lend themselves to machine processing.

# Open data principles

## 6. Non-discriminatory access to data

“Non-discrimination” refers to who can access data and how they must do so

non-discriminatory access to data means that **any person can access the data at any time without having to identify him/herself or provide any justification for doing so.**

# Open data principles

## 7. Open standards

Open standards refer to who owns the format in which data is stored

# Open data principles

## 7. Open standards

Open standards refer to who owns the format in which data is stored

Microsoft **Excel** is a fairly commonly-used spreadsheet program which costs money to use. Freely available alternative formats often exist by which stored data can be accessed without the need for a software license

# Open data principles

## 8. Open Licence

Maximal openness includes clearly **labeling public information** as a work of the government and **available without restrictions on use** as part of the public domain

In France two type of licences : **Licence Ouverte (CC-BY)** ou **ODBL (CC-BY-SA)**. Do you know the difference ?

# Open data principles

## 8. Open Licence

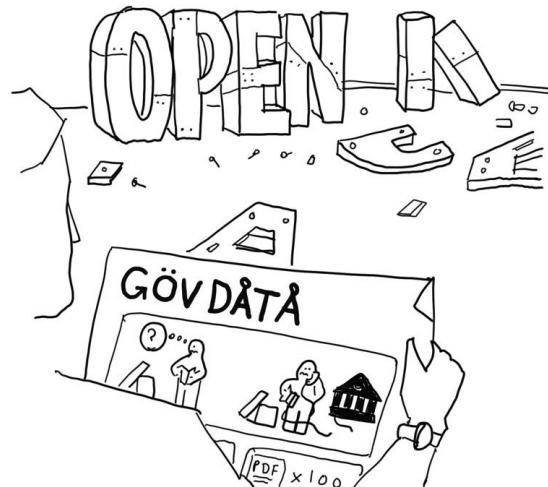
Maximal openness includes clearly **labeling public information** as a work of the government and available without restrictions on use as part of the public domain

In France two type of licences : **Licence Ouverte (CC-BY) ou ODBL (CC-BY-SA)**. Do you know the difference ?

- **LO (from Etalab) / ODBL** : with both you can share, edit the database, create derived products and had a commercial usage
- **LO** : more "permissive" : you just have to mention the source and the date of update
- **ODBL** : you have to share and open your database at the same conditions

# Challenge 1 : data findability

Data findability is a major challenge. We have data portals and registries, but government agencies under one national government still publish data in different ways and different locations. (...) **Data findability is a prerequisite for open data to fulfill its potential and currently most data is very hard to find.**



<https://index.okfn.org/insights/>

# Challenge 1 : data findability



The image shows a composite screenshot of a Twitter search results page and a data.gov.fr dataset page.

**Twitter Search Results:** The search term is "comptage rer". The results page includes a header for "Jules Grandin" (@JulesGrandin) with an "Abonné" (Subscribed) button. The search results list the following tweets:

- 1. Samuel Goëta (@samgoeta)**  
Rechercher des données : une déception en 4 actes
- 2. Jules Grandin (@JulesGrandin)**  
comptage rer
- 3. Jules Grandin (@JulesGrandin)**  
résultats (0,26 secondes)
- 4. Jules Grandin (@JulesGrandin)**  
yageur sur RER - Data.gouv.fr  
gouv.fr/fr/datasets/comptage-voyageur  
» contient le **comptage voyageur effectué**

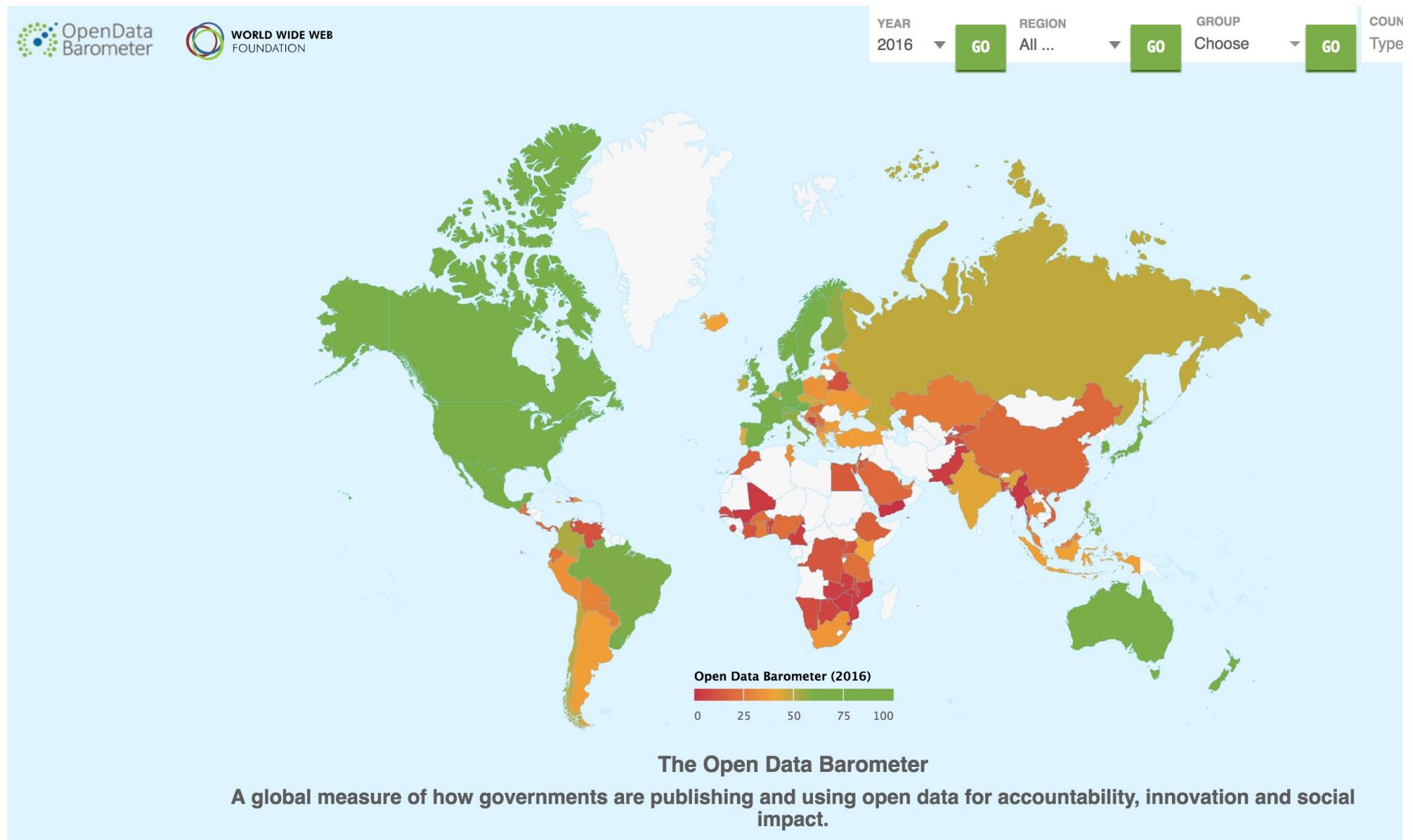
**data.gov.fr Dataset:** The dataset is titled "comptage voyageur sur RER" and is last checked on 29/11/2017. It has 58 Retweets and 156 likes. The dataset page includes a "CSV" download link and a detailed description table.

# Défi 2 : data quality

**Government data is usually incomplete, out of date, of low quality, and fragmented.** In most cases, open data catalogues or portals are manually fed as the result of informal data management approaches. **Procedures, timelines, and responsibilities are frequently unclear among government institutions tasked with this work.**

<http://opendatabarometer.org/4thedition/>

# Défi 2 : data quality



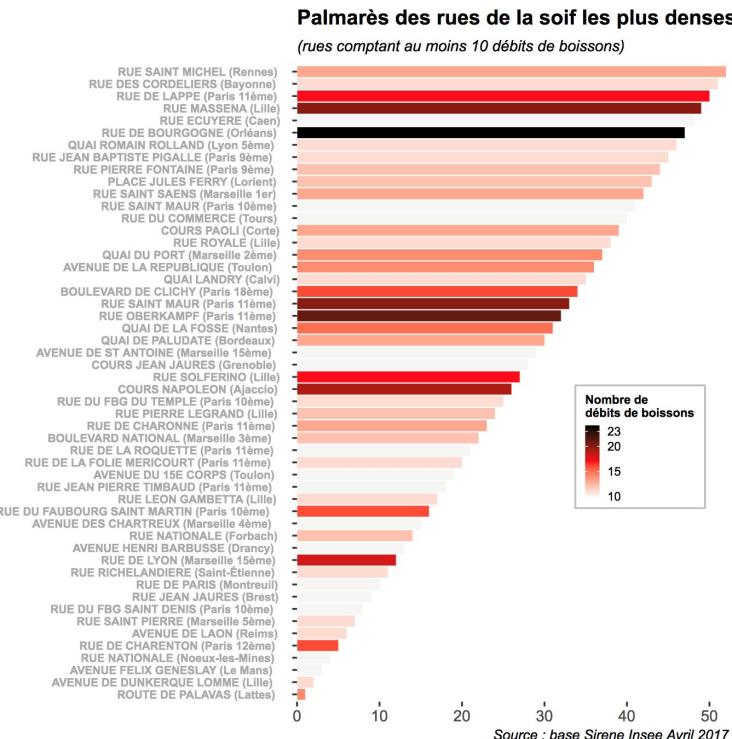
# Open Data example : base SIRENE



The screenshot shows the data.gouv.fr homepage with the following details:

- Header:** data.gouv.fr, Plateforme ouverte des données publiques françaises, Découvrez l'OpenData Données Tableau de bord, Connexion / Inscription.
- Search Bar:** Recherche
- Thematic Filter:** Thématiques
- Contribution Button:** CONTRIBUEZ !
- Dataset Title:** Base Sirene des entreprises et de leurs établissements (SIREN, SIRET)
- Certification:** Ce jeu de données provient d'un service public certifié (DONNÉE DE RÉFÉRENCE)
- Description:** Ce jeu de données permet d'accéder aux 9 millions d'entreprises et 10 millions d'établissements actifs du répertoire Sirene de l'Insee qui enregistre quotidiennement leur état civil :
- List of Features:**
  - quelle que soit leur forme juridique ;
  - quel que soit leur secteur d'activité (industriels, commerçants, artisans, professions libérales, agriculteurs, collectivités territoriales, banques, assurances, associations...);
  - situés en France métropolitaine, ainsi qu'en Guadeloupe, Martinique, Guyane, La Réunion, Mayotte, Saint-Barthélemy, Saint-Martin et Saint-Pierre-et-Miquelon. Les organismes publics ou privés et les entreprises étrangères qui ont une représentation ou une activité en France y sont également répertoriés.
- Text:** Le répertoire Sirene est ainsi la principale source exhaustive sur l'ensemble des entreprises et des établissements actifs.
- Warning:** AVERTISSEMENT
- Legal Note:** La base Sirene contenant des données à caractère personnel, l'Insee attire votre attention sur les obligations légales qui en découlent:
  - Le traitement de ces données relève des obligations de déclaration de la Loi 78-17 du 6 janvier 1978 modifiée, dite Loi CNIL : <https://www.cnil.fr/fr/loi-78-17-du-6-janvier-1978-modifiee>
  - Selon votre usage du jeu de données, il est de votre responsabilité de tenir compte du statut de diffusion le plus récent de chaque personne physique. En effet, l'article A123-96 du code de commerce dispose que : "Toute personne physique peut demander soit directement lors de ses formalités de création ou de modification, soit par lettre adressée au directeur général de l'Institut national de la statistique et des études économiques, que les informations du répertoire la concernant ne puissent être utilisées par des tiers autres que les organismes habilités au titre de l'article R. 123-224 ou les administrations, à des fins de prospection, notamment commerciale."
- Insee Logo:** Producteur, Insee, Mesurer pour comprendre, L'Institut national de la statistique et des études économiques (Insee) collecte, produit, analyse et diffuse des informations sur l'économie et la société françaises. Ces ... +
- Follow Button:** SUIVRE
- Information Section:** Informations, Donnée de référence, Licence Ouverte / Open Licence, Quotidienne

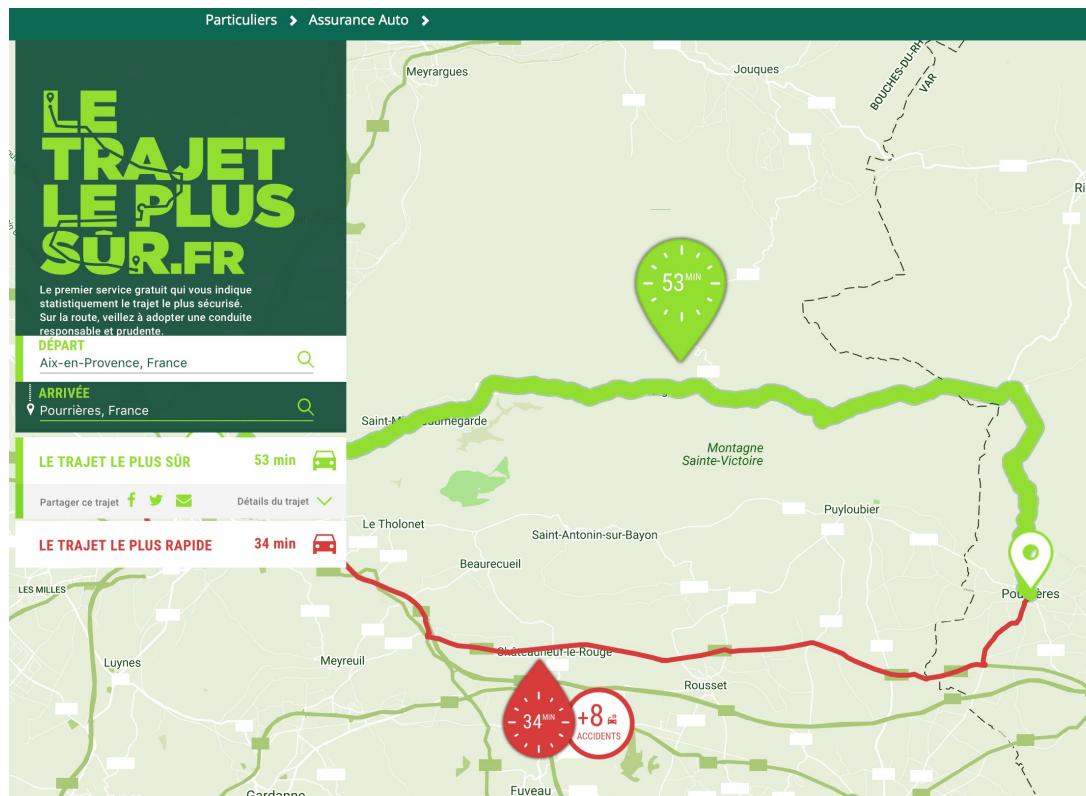
# Open Data reutilisation : base SIRENE



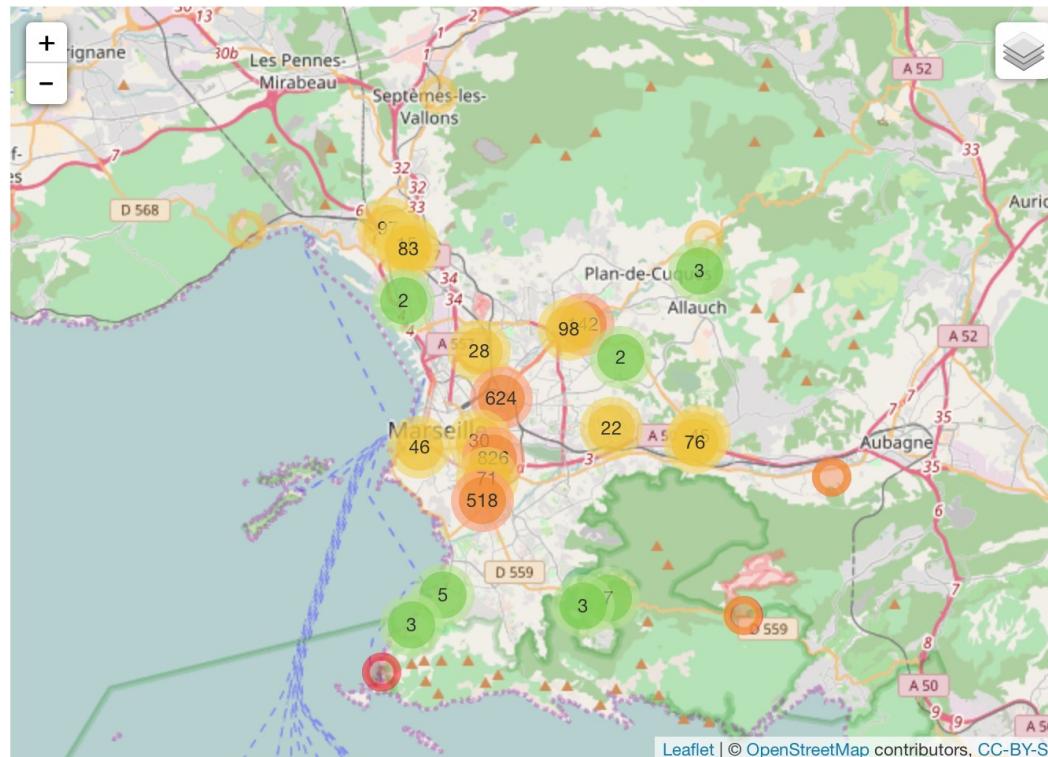
# Open Data example : road accidents

The screenshot shows the data.gouv.fr platform interface. At the top, there is a header with the French flag, the text 'data.gouv.fr', and a 'Connexion / Inscription' button. Below the header, there is a search bar, a 'Thématiques' dropdown, and a 'CONTRIBUEZ!' button. The main content area is titled 'Base de données accidents corporels de la circulation'. It includes a sub-section titled 'Ce jeu de données provient d'un service public certifié' with a 'NEC MERGITUR' badge. The main text describes the dataset as containing information on road accidents, specifically 'accidents corporels de la circulation', with data from 2005 to 2015. It mentions the 'Fichier BAAC' and the 'ONISR'. Below this, there is a section on 'Ressources' with three downloadable files: 'vehicules\_2015.csv', 'Description\_des\_bases\_de\_donnees\_ONISR\_Annees\_2005\_a\_2015.pdf', and 'caracteristiques\_2015.csv'. To the right, there is a 'Producteur' section featuring the French flag, the 'MINISTÈRE DE L'INTÉRIEUR' logo, and a 'SUIVRE' button. Below this, there is an 'Informations' section with a list of details including 'Nec Mergitur', 'Licence Ouverte / Open Licence', 'Annuelle', '8 juillet 2013', '20 octobre 2016', '20 octobre 2016', 'POI', and various tags like 'ACCIDENT', 'ACCIDENTS-DE-L...', 'BAAC', 'CIRCULATION', 'CIRCULATION-RO...', 'PASSAGERS', 'VELO', 'VOITURE', and 'SUUGERER UN MOT-CLÉ'.

# Open Data reutilisation : road accidents



# Open Data reutilisation : road accidents



en cliquant sur le petit pictogramme de calque en haut à droite de la carte, vous pouvez

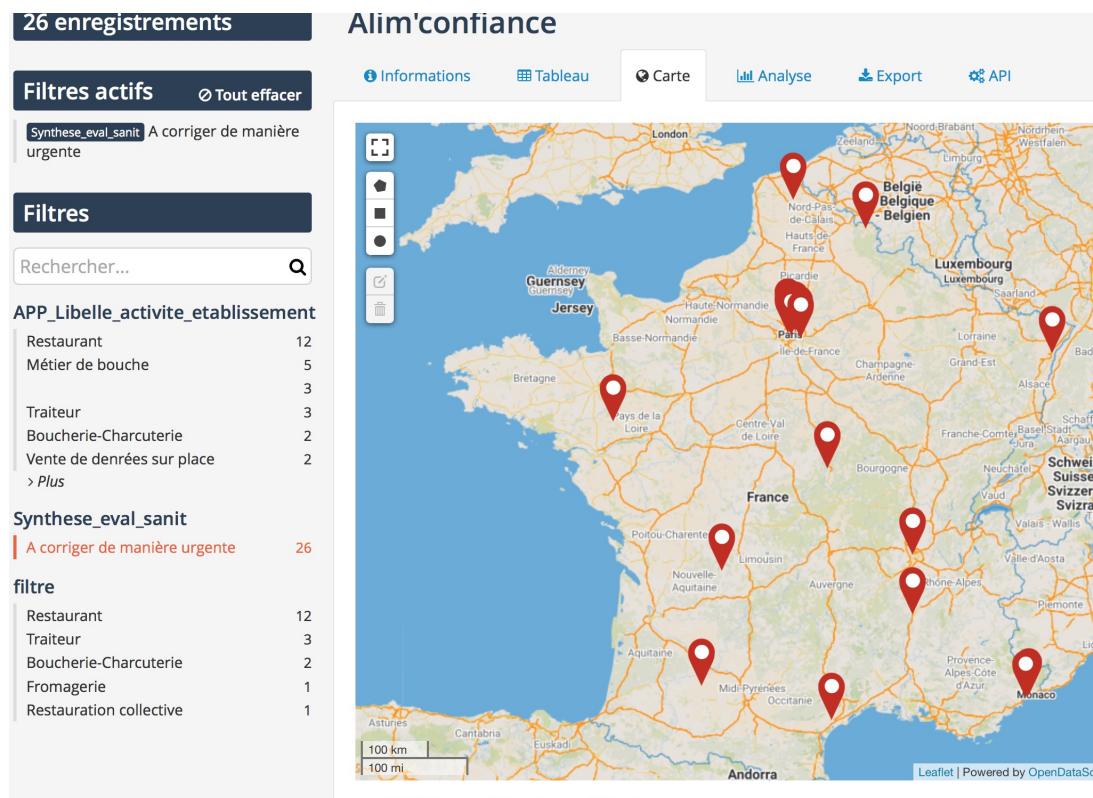
Map by Joël Gombin in Marsactu

# Open Data example : sanitary inspections

The screenshot shows the data.gouv.fr platform interface. At the top, there is a logo for the French Republic and the text "data.gouv.fr". Below the logo, a navigation bar includes "Découvrez l'OpenData", "Données", "Tableau de bord", "Connexion / Inscription", and a "CONTRIBUEZ!" button. The main content area is titled "Résultats des contrôles officiels sanitaires : dispositif d'information « Alim'confiance »". A sub-section title "Ce jeu de données provient d'un service public certifié" is present. The text explains the publication of official sanitary inspection results in the food sector (restaurants, canteens, abattoirs, etc.) as part of the "Alim'confiance" campaign. It mentions the decree of December 2016 and the start of publication in March 2017. A section titled "Quels sont secteurs d'activité concernés ?" describes the sectors involved. Another section, "Dans tous les pays, la mise en place de la mesure s'est toujours accompagnée d'une amélioration du niveau sanitaire des établissements", states that the measure has always been accompanied by an improvement in the sanitary level of establishments. Below this, there is a "Ressources" section with links to "Données brutes CSV" and "Données brutes Excel". To the right, there is a sidebar titled "Producteur" featuring the French Republic logo and the Ministry of Agriculture, Agroalimentary, and Forestry logo. It also includes a "SUIVRE" button and a "Informations" section with a list of dates and a "POI" section with tags like "CONTROLE", "CONTROLE-SANIT...", "HYGIENE", "RESTAURANT", "RESTAURATION", "RESULTATS", "SANITAIRE", and a "SUGGÉRER UN MOT-CLÉ" button.

Résultats des contrôles officiels sanitaires : dispositif d'information « Alim'confiance »

# Open Data reutilisation : sanitary inspections



26 établissements au niveau d'hygiène à corriger de manière urgent

# Open Data example : crime data

Chiffres départementaux mensuels relatifs aux crimes et délits enregistrés par les services de police et de gendarmerie depuis janvier 1996

Ce jeu de données provient d'un service public certifié

NEC MERGITUR

Ces données correspondent aux nombres de crimes et délits enregistrés mensuellement par les services de police et de gendarmerie. Les tableaux mis à disposition forment ce que l'on appelait « l'état 4001 ».

Ils contiennent des informations, de caractère administratif, sur l'activité judiciaire des services de police et de gendarmerie, y compris celles des DOM-COM, depuis janvier 1996. Ces données sont mises à jour mensuellement.

## Ressources

Documentation\_des\_chiffres\_mensuels\_departementaux.docx

docx (40.8Ko) 737 Disponible

TÉLÉCHARGER

Tableaux\_4001\_TS.xlsx

Le fichier Tableaux\_4001\_TS contient les données enregistrées par les unités de la gendarmerie et de la police de chaque département.

xlsx (11.6Mo) 855 Disponible

PRÉVISUALISER TÉLÉCHARGER

Producteur



MINISTÈRE  
DE  
L'INTÉRIEUR

Ministère de l'Intérieur

Placé au cœur de l'État, le ministère de l'Intérieur assure la permanence et la continuité de l'État. Cette fonction régaliennes se concrétise par le rôle majeur et les services...

VOIR LE PROFIL

CONTACTER

# **Launch of the datathon**

# Topics & groups

Here are the 6 topics

- **Do the French students cost too much ?**
- **Do the French public servants cost too much ?** *Grand débat National*
- **So as to reduce the public debt and taxes, which public spendings shall be cut off in priority ?**  
*Grand débat National*
- **Are there too many representatives in France ?**
- **Are French representatives paid too much ?**
- **Are French public services too concentrated in the cities ?** *Grand débat National*

You have 20 minutes to form a group of 5-6 students with at least 2 "advanced" students

# Thank you

Contact : [timothee@datactivi.st](mailto:timothee@datactivi.st)