Open Data Quality

Qualité des Données, semestre 1, année universitaire 2021-2022

Elise Chin, Télio Cropsal, Théo Quémener
Université Paris-Dauphine | PSL

8 novembre 2021

Vos présentateurs

À propos de nous

Elise Chin

Research Apprentice, ENGIE Lab CRIGEN

Télio Cropsal

Research Apprentice, Huawei Technologies France

Théo Quémener

Apprenti, Renault, Systèmes avancés d'aide à la conduite

Plan

Quelle application des principes de la qualité des données pour l'open data ?

1. Présentation de l'open data et des enjeux associés
2. Mesurer la qualité des données ouvertes
3. Une analyse pratique
    a. Qualité des données relatives à la COVID-19
    b. Analyse du format des jeux de données de la plateforme ouverte des données publiques françaises

1. Présentation de l’open data et des enjeux associés

Open data : une première définition

Définition :

L’open data désigne l’ensemble des données accessibles gratuitement dont l’usage, l’exploitation et le partage sont totalement libres.

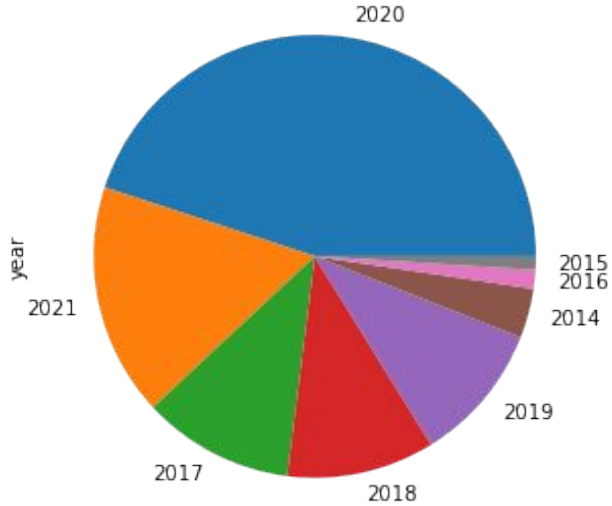
Open data : une ouverture en essor

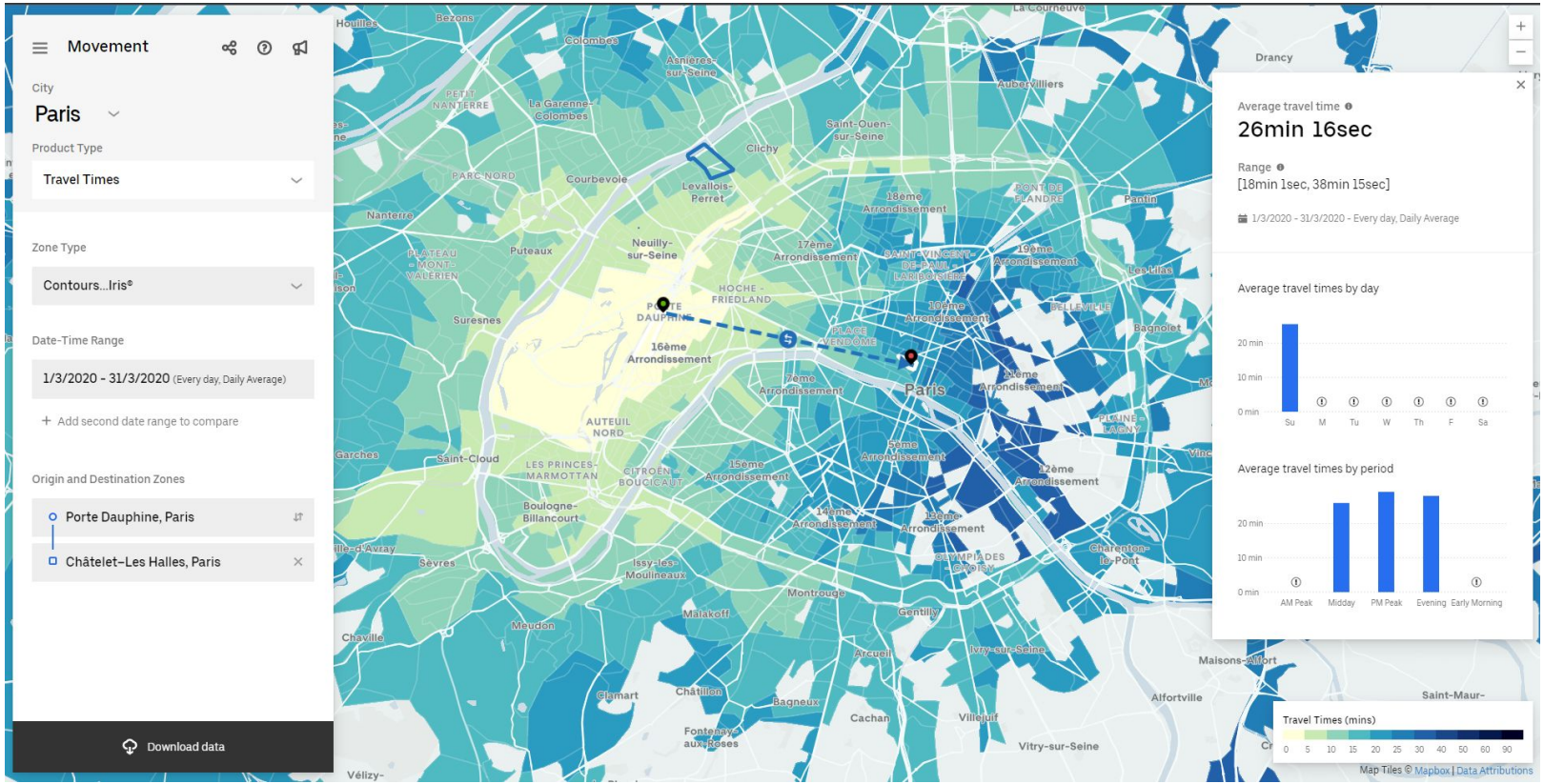
- **Les acteurs :**
  - Le secteur publique : *Agences régionales de santé, les régions, les administrations publiques, etc.*
  - Le secteur privé : *Uber Movement, Inside Airbnb, etc.*

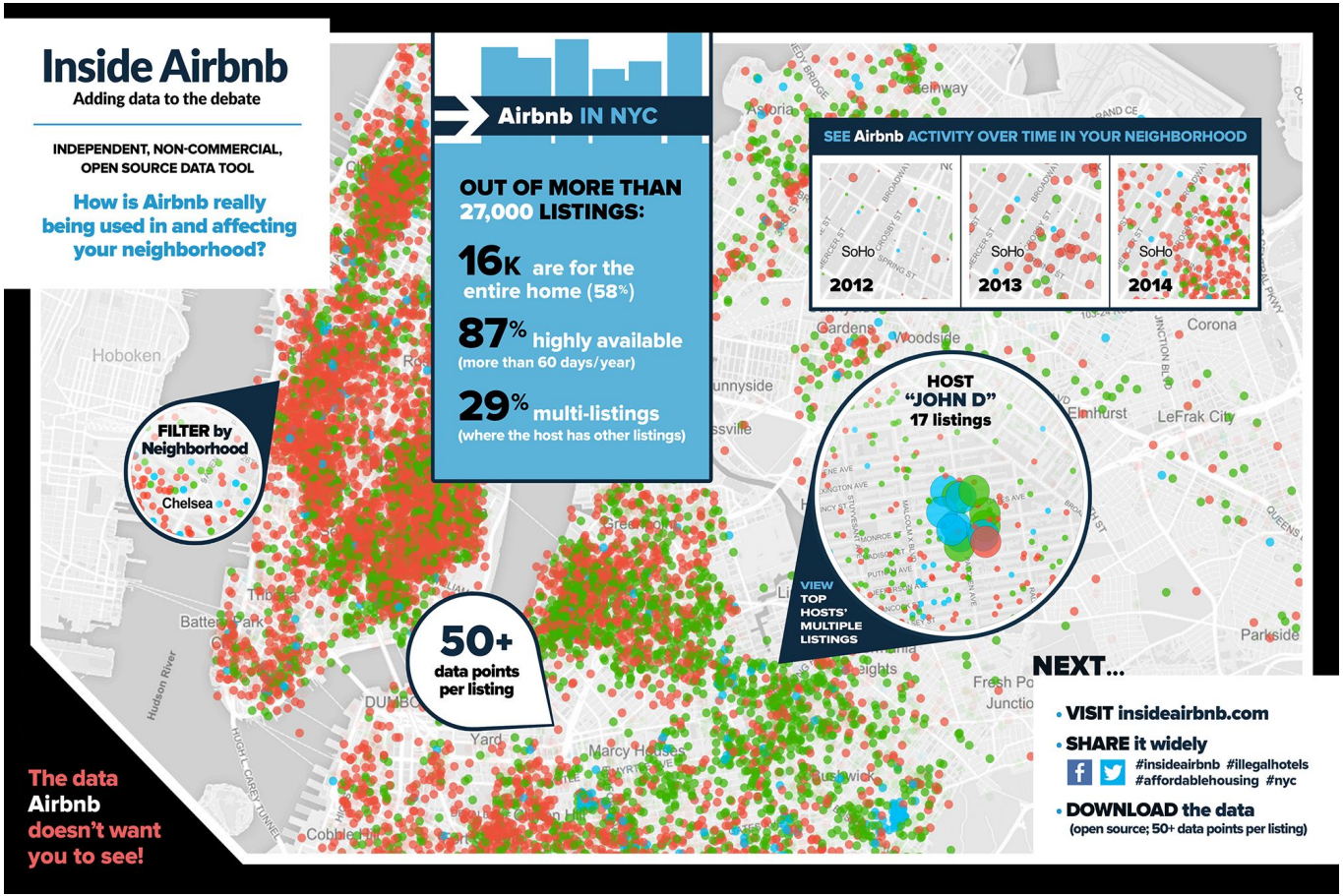
- **Une quantification de l'essor :**

  Plateforme :
  - data.gouv.fr.

  Chiffres :
  - 120382 datasets publiés en 2020.
  - 2613 datasets publiés en 2015.




Movement
City
Paris
Product Type
Travel Times
Zone Type
Contours...Iris®
Date-Time Range
1/3/2020 - 31/3/2020 (Every day, Daily Average)
+ Add second date range to compare
Origin and Destination Zones
Porte Dauphine, Paris
Châtelet–Les Halles, Paris
Download data
Average travel time
26min 16sec
Range
[18min 1sec, 38min 15sec]
1/3/2020 - 31/3/2020 - Every day, Daily Average
Average travel times by day
20 min
10 min
0 min
Su
M
Tu
W
Th
F
Sa
Average travel times by period
20 min
10 min
0 min
AM Peak
Midday
PM Peak
Evening
Early Morning
Travel Times (mins)
0
5
10
15
20
25
30
40
50
60
90
Map Tiles © Mapbox | Data Attributions
Paris
Boulogne-Billancourt
Saint-Cloud
Issy-les-Moulineaux
Montrouge
Malakoff
Meudon
Clamart
Châtillon
Bagneux
Cachan
Arcueil
Gentilly
Ivry-sur-Seine
Vitry-sur-Seine
Alfortville
Maisons-Alfort
Saint-Maur-
Charenton-le-Pont
Bagnolet
Pantin
Aubervilliers
Drancy
Clichy
Levallois-Perret
Neuilly-sur-Seine
Courbevoie
Puteaux
Suresnes
Nanterre
Colombes
Bezons
Asnières-sur-Seine
Saint-Ouen-sur-Seine
16ème Arrondissement
17ème Arrondissement
18ème Arrondissement
19ème Arrondissement
20ème Arrondissement
15ème Arrondissement
14ème Arrondissement
13ème Arrondissement
12ème Arrondissement
7ème Arrondissement
5ème Arrondissement


Inside Airbnb
Adding data to the debate
INDEPENDENT, NON-COMMERCIAL, OPEN SOURCE DATA TOOL
How is Airbnb really being used in and affecting your neighborhood?
Airbnb IN NYC
OUT OF MORE THAN 27,000 LISTINGS:
16K are for the entire home (58%)
87% highly available (more than 60 days/year)
29% multi-listings (where the host has other listings)
SEE Airbnb ACTIVITY OVER TIME IN YOUR NEIGHBORHOOD
SoHo 2012
SoHo 2013
SoHo 2014
HOST "JOHN D" 17 listings
VIEW TOP HOSTS' MULTIPLE LISTINGS
FILTER by Neighborhood
Chelsea
50+ data points per listing
NEXT...
• VISIT insideairbnb.com
• SHARE it widely
#insideairbnb #illegalhotels #affordablehousing #nyc
• DOWNLOAD the data
(open source; 50+ data points per listing)
The data Airbnb doesn't want you to see!
Hoboken
Astoria
Woodside
Corona
Elmhurst
LeFrak City
Parkside
Hudson River
DUMBO

Open data : source de la dynamique

- Loi République Numérique : Janvier 2016.
  - Mise en ligne par défaut des données publiques dans des formats favorisant l'interopérabilité.
  - Accessibilité aux services publics numériques/droit d'accès à internet : messagerie et service publique, etc.

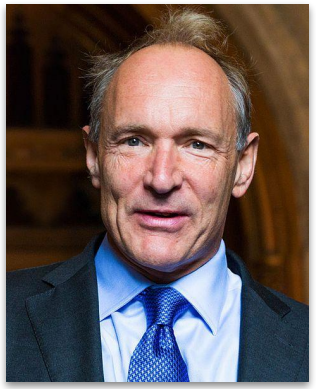
- Deux principales directions :
  - Économique : En 2014 : 4.4% du PIB australien soit 67B$-(*Deciding with data-How data-driven innovation is fuelling Australia's economic growth, PwC Australia 2014*)
  - Sociétale : Handimap, Inside Airbnb, transparence (faire le lien avec loi Répu Numérique)

Open data : nécessité d'implémenter un axe qualité

- **Objectif :** *Libérer le potentiel d'innovation des projets data based avoir un impact économique et sociétal positif fort.*
- **Question :** *Comment l'accélérer, l'améliorer, le rendre plus efficace ?*
  - Principal frein : nettoyage de la donnée avec incertitude sur la viabilité de l'outcome : GIGO.

2. Mesurer la qualité des données ouvertes

Un cas introductif



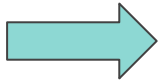
Tim-Berners Lee, fondateur du web et fervent défenseur d'un Internet plus ouvert

Les cinq "étoiles" de Tim-Berners Lee :

- ★ Disponible sur le Web
- ★ Facilement traitable par une machine
- ★ Non-utilisation de formats propriétaires
- ★ Utilise les standards du W3C (RDF, SPARQL, URI)
- ★ Fait des liens avec d'autres sources de données (pour plus de contexte)

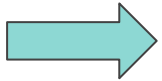
Deux majeurs défauts

- Considère seulement les formats des données et non la signification intrinsèque



Changer de granularité, c'est à dire évaluer la qualité en fonction de différentes échelles

- Ne vérifie pas d'autres caractéristiques souhaitables pour le jeu de données



Utiliser d'autres métriques pour d'autres caractéristiques

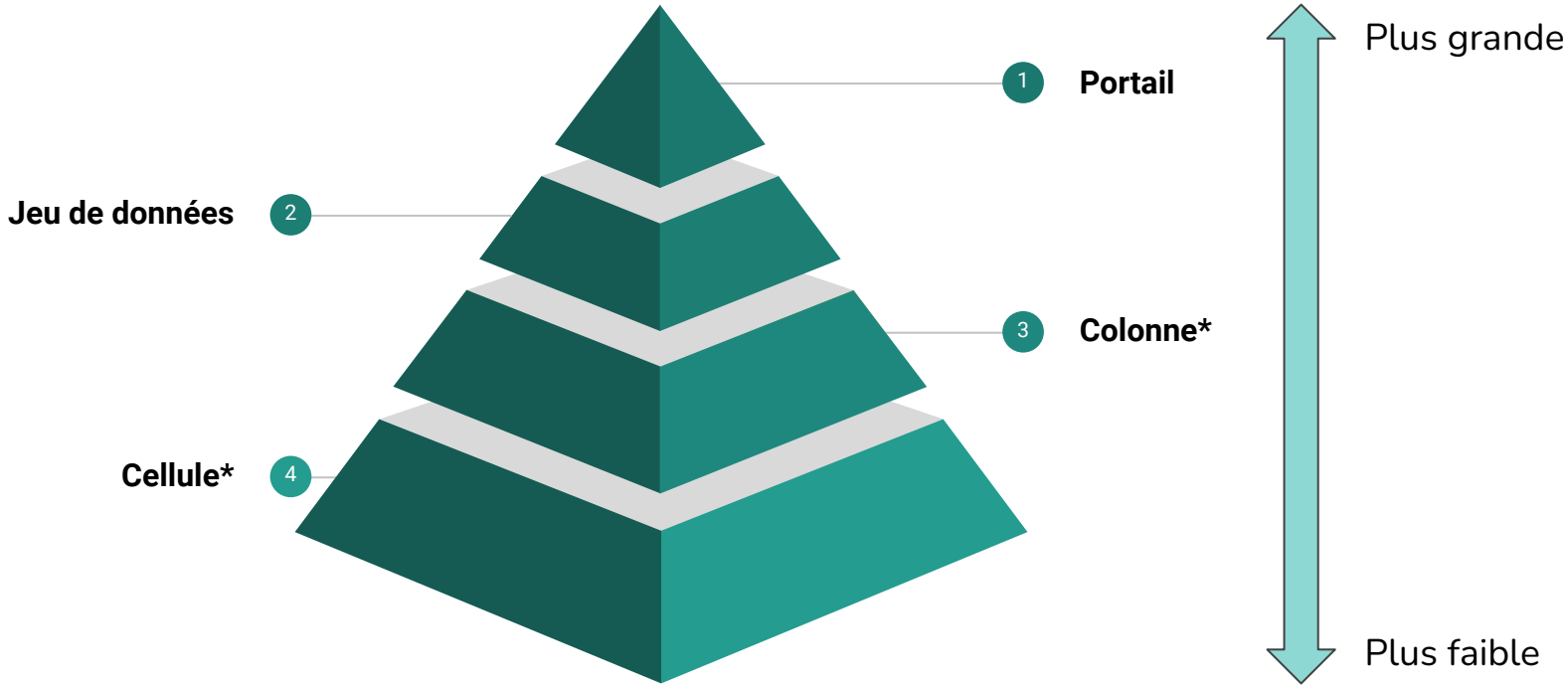
Formalisation

Caractéristiques	Métriques	Données
Idéaux souhaitables pour les données	Cherchent à êtres les plus fidèles aux caractéristiques, selon un certain niveau de granularité	Données au sens large du terme avec les métadonnées

Cadre de travail usuel :

- Collecte des informations contextuelles à propos des données (métadonnées)
- **Evaluation des données à différents niveaux de granularité**
- Amélioration pour atteindre une plus grande qualité des données

Granularité



*Si en présence de données tabulaires

Quelques caractéristiques et métriques pour l'open data

Spécificité de l'open data : ouverture et multitude de "référentiel client"

- Maurino et al. ont calculé **complétude, disponibilité** et **accessibilité** de documents à travers les liens internes et externes, la **précision** en fonction des formats et la **fraîcheur** en fonction de la présence ou non de MÀJ.
- Atz a calculé la **fraîcheur** avec une métrique qui calcule le pourcentage de jeu de données à jour sur un portail
- **Traçabilité, Conformité, Intelligibilité**

“The 8 Principles of Open Government Data” du livre sur l’OGD par Tauberer

- Complète
- Primaire
- “En temps voulu”
- Accessible
- Traitable par une machine
- Non discriminatoire
- Non propriétaire
- Sans licence
- Consultable

Tableau 3 de Vetrò, A., et al., Open data quality measurement framework: Definition and application to Open Government Data, Government Information Quarterly (2016)

Characteristic	Metric	Level	Description
Traceability	Track of creation	Dataset	Indicates the presence or absence of metadata associated with the process of creation of a dataset.
	Track of updates	Dataset	Indicates the existence or absence of metadata associated with the updates done to a dataset.
Currentness	Percentage of current rows	Cell	Indicates the percentage of rows of a dataset that have current values, it means that they don't have any value that refers to a previous or a following period of time.
	Delay in publication	Dataset	Indicates the ratio between the delay in the publication (number of days passed between the moment in which the information is available and the publication of the dataset) and the period of time referred by the dataset (week, month, year).
Expiration	Delay after expiration	Dataset	Indicates the ratio between the delay in the publication of a dataset after the expiration of its previous version and the period of time referred by the dataset (week, month, year).
Completeness	Percentage of complete cells	Cell	Indicates the percentage of complete cells in a dataset. It means the cells that are not empty and have a meaningful value assigned (i.e. a value coherent with the domain of the column).
	Percentage of complete rows	Cell	Indicates the percentage of complete rows in a dataset. It means the rows that don't have any incomplete cell.
Compliance	Percentage of standardized columns	Cell	Indicates the percentage of standardized columns in a dataset. It just considers the columns that represent some kind of information that has standards associated with it (i.e. geographic information).
	eGMS Compliance	Dataset	Indicates the degree to which a dataset follows the e-GMS standard (as far as the basic elements are concerned, it essentially boils down to a specification of which Dublin Core metadata should be supplied)
	Five star Open Data	Dataset	Indicates the level of the 5 star Open Data model in which the dataset is and the advantage offered by this reason.
Understandability	Percentage of columns with metadata	Cell	Indicates the percentage of columns in a dataset that has associated descriptive metadata. This metadata is important because it allows to easily understanding the information of the data and the way it is represented.
	Percentage of columns in comprehensible format	Cell	Indicates the percentage of columns in a dataset that is represented in a format that can be easily understood by the users and it is also machine-readable.
Accuracy	Percentage of accurate cells	Cell	Indicates the percentage cells in a dataset that has correct values according to the domain and the type of information of the dataset.
	Accuracy in aggregation	Cell	Indicates the ratio between the error in aggregation and the scale of data representation. This metric only applies for the datasets that have aggregation columns or when there are two or more datasets referring to the same information but in a different granularity level.

Observations sur les propriétés des métriques

- En général, l'évaluation des données gouvernementales se fait au niveau des portails
- Les métriques peuvent dépendre de la structure des données, pour les données semi-structurées : Open Linked Data quality par Behkamal et al.
- Kaiser et al. ont défini des principes à respecter pour les métriques de **Traçabilité**, **conformité** et **intelligibilité** : mesurabilité, interprétabilité, agrégation, faisabilité
- Les métriques peuvent être "quantitatives" ou "subjectives"

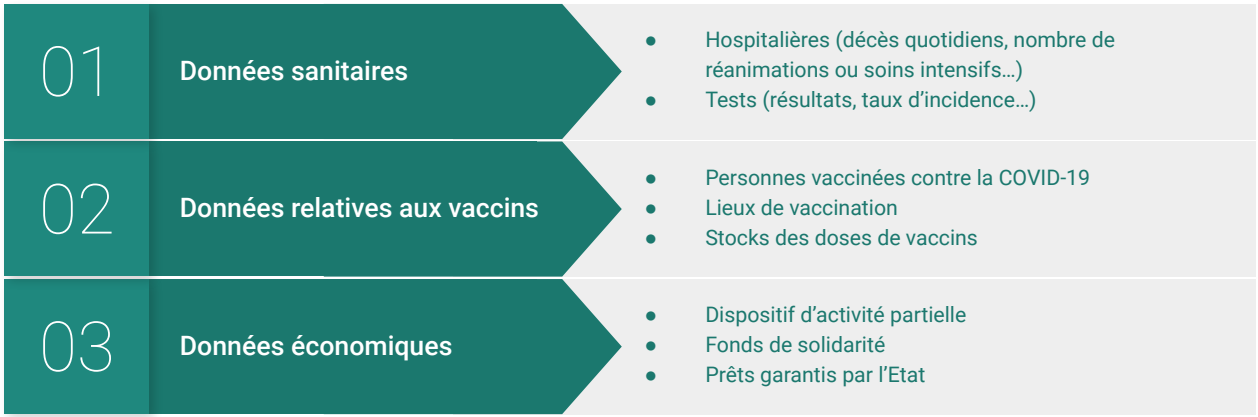
3. Mise en application

3.1 Qualité des données relatives à la COVID-19

Les données ouvertes de la COVID-19

Forte demande d'ouverture des données relatives à la Covid-19, accélérée par la pandémie.

Sur la plateforme des données publiques françaises (data.gouv.fr) :

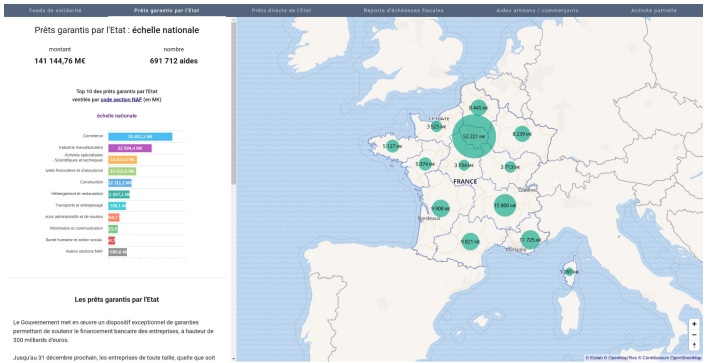


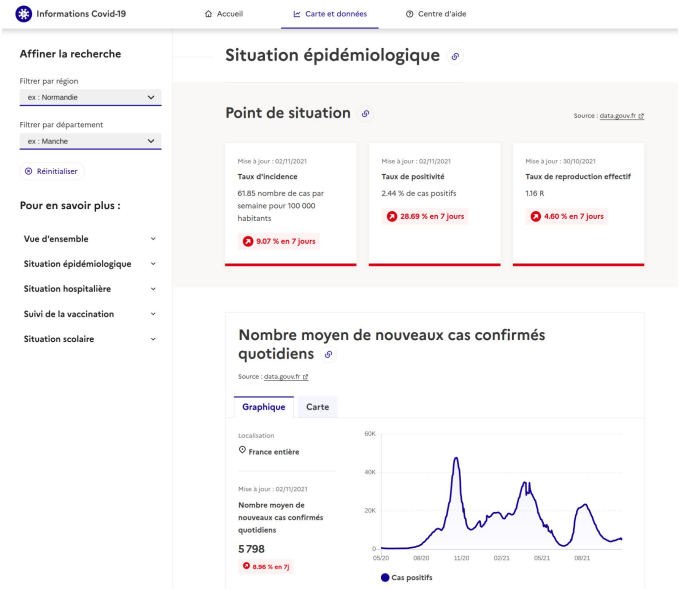
Source : https://www.data.gouv.fr/fr/pages/donnees-coronavirus/

Les réutilisations de données (1/2)

Par différents acteurs : administrations, scientifiques, initiatives citoyennes, privées…

Tableaux de bords, e.g. sur les prêts garantis par l'Etat ou bien la situation épidémiologique





Sources : aides-entreprises.data.gouv.fr (haut) ; gouvernement.fr/info-coronavirus (droite)

Les réutilisations de données (2/2)

Modélisations statistiques, e.g. modèles explicatifs et prédictifs de la probabilité d'entrée en réanimation

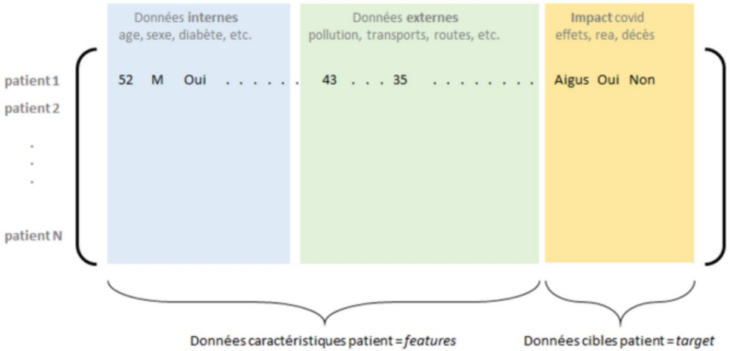
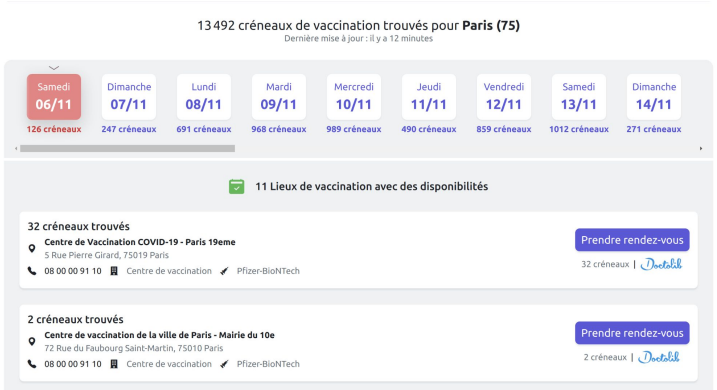


Figure 4: Prédiction de l'impact covid d'un nouveau patient hospitalise

Source : *addactis.com/fr/impact-sanitaire-covid-19-modelisation-epidemiologique-machine-learning/*

Initiative citoyenne comme ViteMaDose pour réserver un rendez-vous de vaccination



Source : *vitemadose.covidtracker.fr*

Défis soulevés par la qualité des données ouvertes de la COVID-19

Format

- Variété des formats car différentes sources de données
  => Problèmes de validité
- Nombreuses disciplines dont proviennent les sources de données
  => Problème d'exploitation due à la compréhension des formats

Fraîcheur

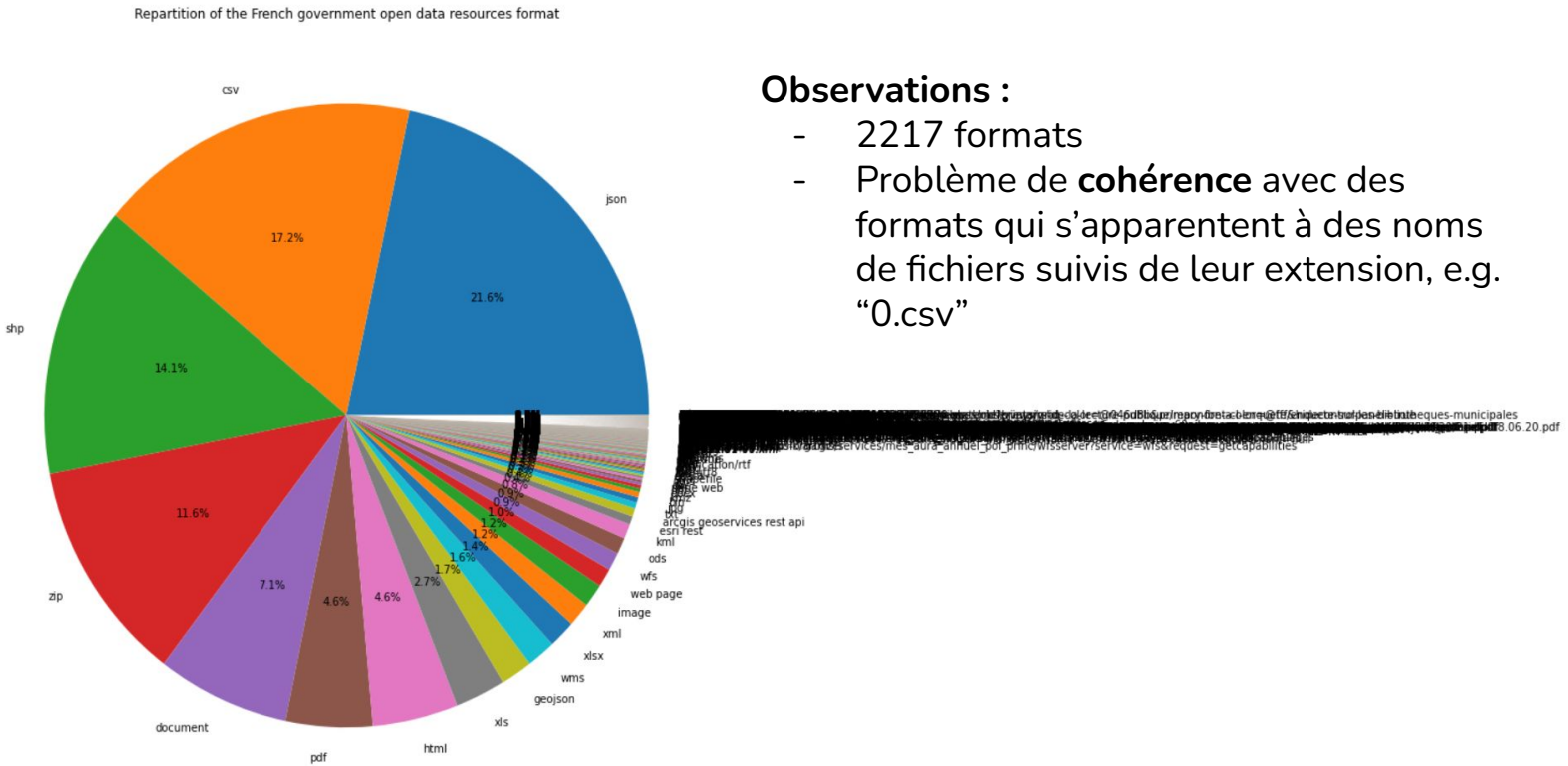
- Pandémie exigence une réponse immédiate des gouvernements et de la population
  => Nécessaire d'avoir les dernières données disponibles
- Problème lié à la disponibilité quotidienne des données couplée à la variété des sources de données

Traçabilité

- Introduction de nouvelles variables pour une information plus précise
  => Effort de vigilance et de surveillance pour suivre les modifications

3.2 Analyse du format des jeux de données de data.gouv.fr

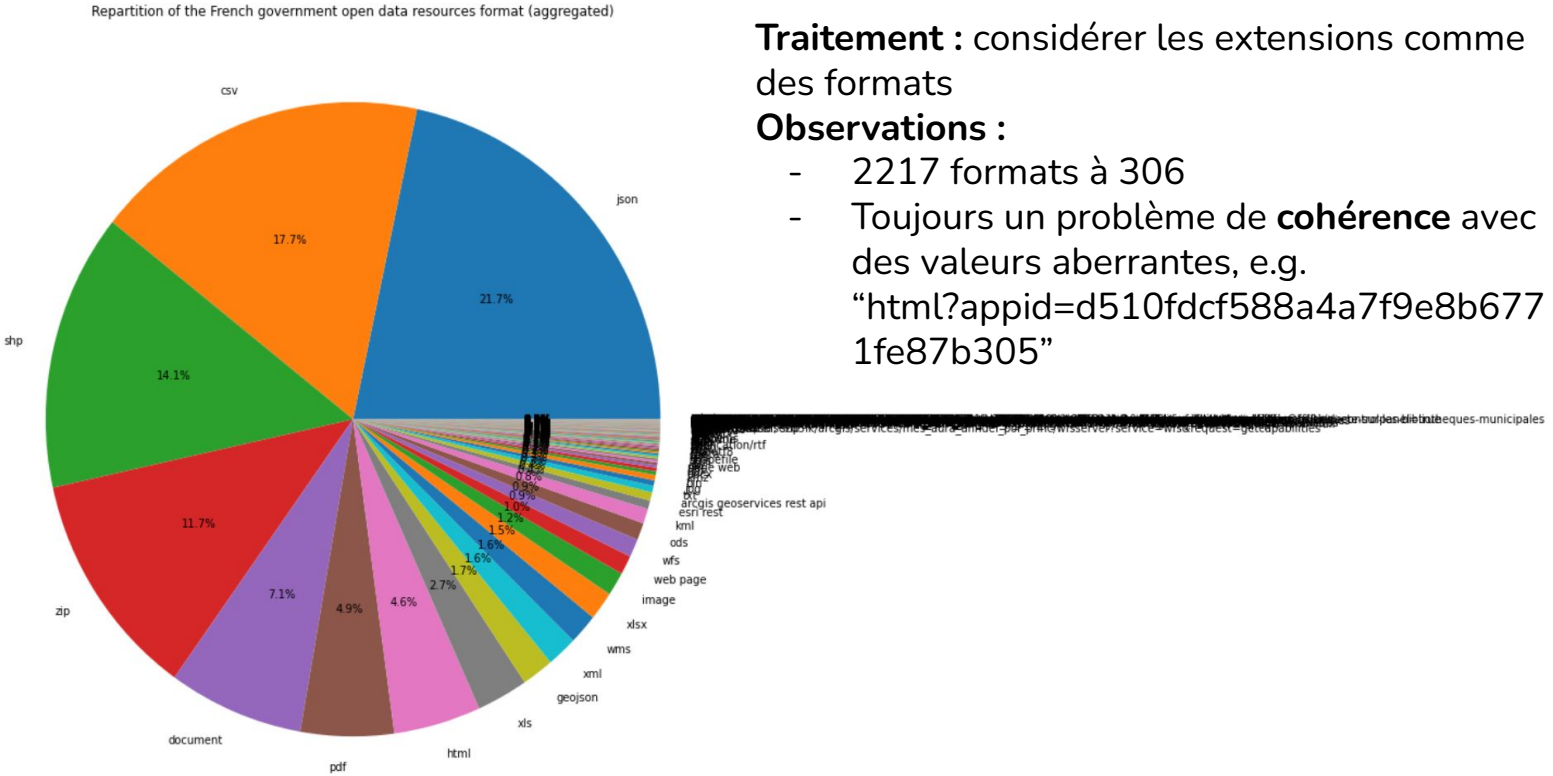
Un premier comptage



Observations :

- 2217 formats
- Problème de **cohérence** avec des formats qui s’apparentent à des noms de fichiers suivis de leur extension, e.g. “0.csv”

Sans les extensions

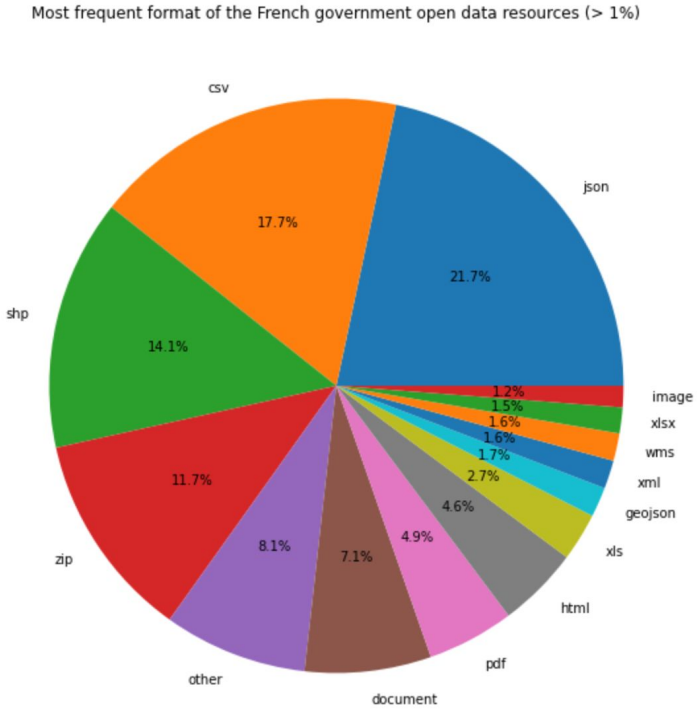


Traitement : considérer les extensions comme des formats

Observations :

- 2217 formats à 306
- Toujours un problème de **cohérence** avec des valeurs aberrantes, e.g. “html?appid=d510fdcf588a4a7f9e8b6771fe87b305”

Conservation des formats avec plus 1% de représentation



Observations :

- Formats très variés → Données variées
- Formats prépondérants sont :
  - JSON (21.7%)
  - CSV (17.7%)
  - SHP (14.1%) → Systèmes d'information géographiques
- JSON et CSV facilement réutilisables car machine-readable. En revanche, PDF et HTML le sont moins.
- Présence de formats propriétaires (XLS, XLSX)
- Dans le champ "other", présence de valeurs aberrantes

Conclusion

Merci !