Workshop : manipulating data 1/2

Data and Algorithms for Public Policy

Timothée Gidoin

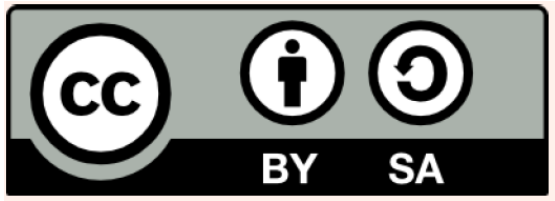
Sciences Po, 2019-09-27

Before we start

Slides : https://gidoin.github.io/sciencespodata/lesson4_dataworkshop.html

Sources : https://github.com/Gidoin/sciencespodata/

The content of this presentation is partly inspired by other presentations made by Datactivist team. I warmly thank them and notably Joël Gombin for his help.



Before we start: reminder

Date	Session #		Teacher
05/09	1	Introduction to data policy	Simon Chignard
12/09	2	Open Data (data as a policy)	Timothée Gidoin
19/09	3	Public sector algorithms: « with great power comes great responsibility »	Simon Chignard
26/09	4	Workshop data manipulation #1: data preparation and analysis	Timothée Gidoin
03/10	5	Workshop data manipulation #2: datavisualization	Timothée Gidoin
10/10	6	Workshop Machine-learning #1 : introduction	Jean-Marie John Mathews
17/10	7	Use Case #1: Predisauvetage (Guest: Antoine Augusti)	Simon Chignard
24/10	8	Explainable algorithms: why ? how ?	Simon Chignard
07/11	9	Use case #2: Predictive policing (Guest: Bilel Benbouzid)	Simon Chignard
14/11	10	Workshop Machine-learning #2 : Fairness	Jean-Marie John Mathews
21/11	11	Workshop Machine-learning #3 : Explicability	Jean-Marie John Mathews
28/11	12	Oral presentations of coursework	Simon Chignard

Before we start: reminder

- **Midterm Exam** :
  - **25% of your total grade**
  - By group of 2 students
  - Data manipulation, analysis visualisation exercise based on open data
  - Instructions will be given end of next workshop (03/10)
  - To be submitted before 25/10 11:59 pm
- **Final Exam** :
  - **75% of your total grade**
  - By group of 4 students
  - 10-pages paper on the analysis of 3 uses cases (outside France) in one of the following topics: social benefits, police / justice, education, public sector human resources
  - 1 oral presentation (15 min) during the last session
  - 1 Medium blog post (1,5 pages) to present your findings
  - Evaluation of final exam: 50% quality of the analysis, 25% oral presentation, 25% quality of Medium blogpost

Before we start: Final exam

5 topics :

- **Social action / social benefits** (unemployment, housing, family benefits, ...)
- **Police, justice, law enforcement**
- **Education** (schools, universities, ...)
- **Public sector human resources management** (recruitment, professional mobility, career management, liabilities and discipline of employees)
- **Health**

1 group per topic, 5 students per group (25 students)

But some parameters to take into account : personal preferences, master speciality, subject difficulty + request to have at least 1 French and foreign student in each group

Before we start: Final exam

Few options, let's chose the algorithms together :

1/ We constitute the groups randomly and assign them randomly to topics

2/ You constitute the group and we assign you randomly to topics

3/ You decide by yourself the group constitution and the topic

but what if more than one group per topic ?

First come first served basis ?

Random assignation next course ?

=> **Groups have to be constituted and assigned with one topic on 03/10 at latest**

Let's go back to open data

What are the 2 main challenges regarding Open Data ?

Challenge 2 : data quality

Challenge 2 : data quality

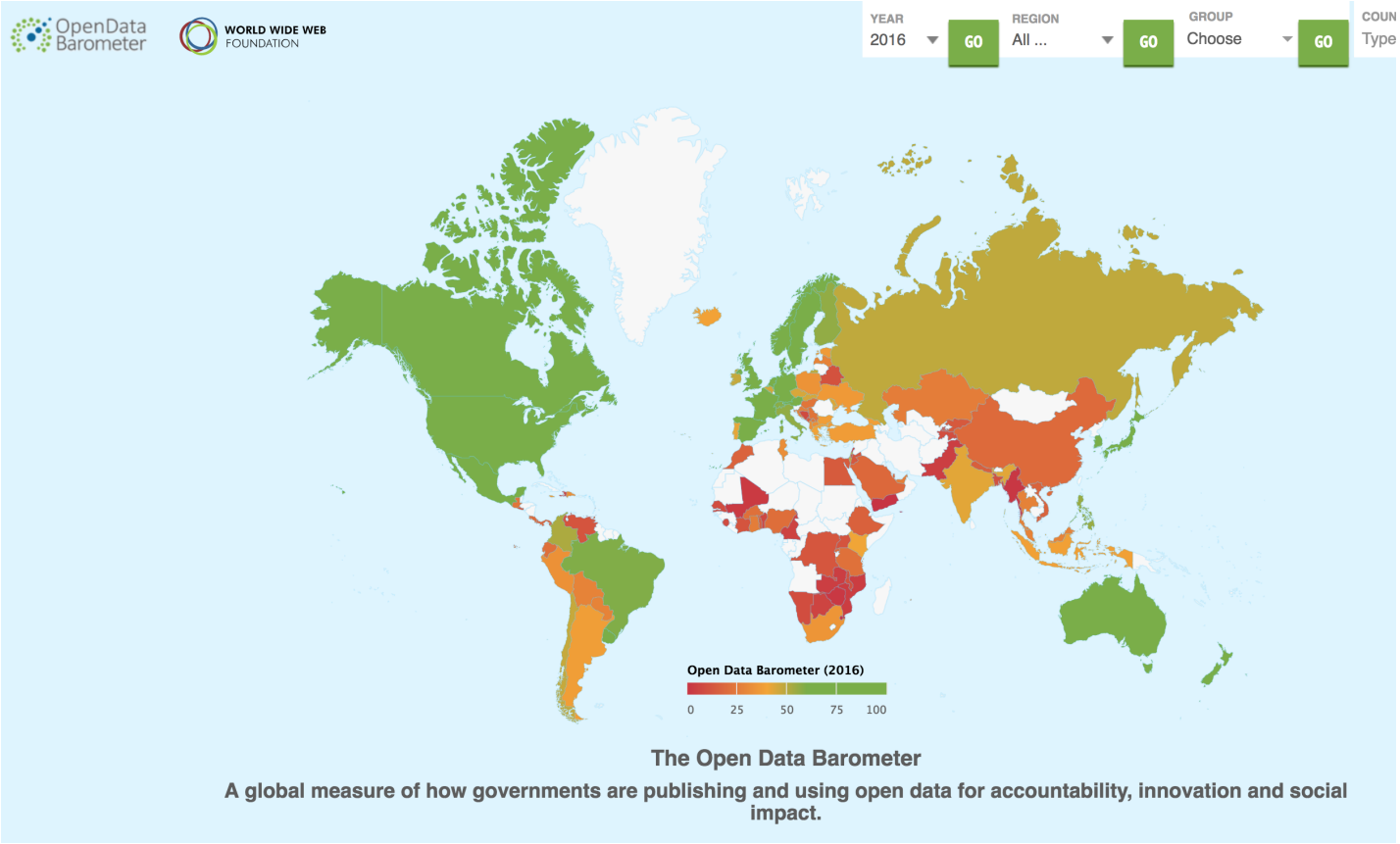
> **Government data is usually incomplete, out of date, of low quality, and fragmented.** In most cases, open data catalogues or portals are manually fed as the result of informal data management approaches. **Procedures, timelines, and responsibilities are frequently unclear among government institutions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

OpenDataBarometer ?

> It's a global measure of how governments are publishing and using open data for accountability, innovation and social impact. The Leaders Edition looks at the 30 governments that have adopted the Open Data Charter and those that, as G20 members, have committed to G20 Anti-Corruption Open Data Principles.

http://opendatabarometer.org/4thedition/report/

Challenge 2 : data quality



Challenge 2 : data quality

Country	Score OUT OF 100	Score Change SINCE FIRST EDITION	Score Trend OVER PAST EDITIONS	Readiness OUT OF 100	Implementation OUT OF 100	Emerging Impact OUT OF 100
Canada See details	76	18 ▲		86	87	55
United Kingdom See details	76	-4 ▼		83	89	57
Australia See details	75	17 ▲		79	84	62
France See details	72	17 ▲		84	77	55
Korea See details	72	25 ▲		82	67	67
Mexico See details	69	33 ▲		79	67	62
Japan See details	68	24 ▲		78	68	58
New Zealand See details	68	5 ▲		79	72	52
United States of America See details	64	-11 ▼		79	76	37
Germany See details	58	2 ▲		76	72	27
Uruguay See details	56	23 ▲		71	70	28
Colombia See details	52	25 ▲		69	60	28
Russia See details	51	10 ▲		62	59	32
Brazil See details	50	15 ▲		63	56	30
Italy See details	50	8 ▲		61	61	27
India See details	48	16 ▲		64	49	32
Argentina See details	47	14 ▲		66	56	20
Ukraine See details	47	25 ▲		60	52	28
Philippines See details	42	19 ▲		54	42	30
Chile See details	40	2 ▲		54	55	12
Indonesia See details	37	17 ▲		49	45	17
South Africa See details	36	14 ▲		50	37	22

OpenDataBarometer 2017 ranking

Challenge 2 : data quality

Sometimes data are well too agregated...

Challenge 2 : data quality

Or hardly exploitable…

133			de 5 964 à 11 896	6,90 €	
134			de 11 897 à 26 420	9,40 €	
135			de 26 421 à 70 830	10,90 €	
136			de 70 831 à 150 000	21,00 €	15,70 €
137	Hors barème LEA	Revenus annuels imposables 2012 (loi de Finances 2013)	au delà de 150 000	32,00 €	26,00 €
138	**activités de loisirs sans hébergement*				
139	• **ALSH* SANS RESTAURATION** (tarifs journaliers)				
140			Tranches Quotient Familial / Revenus Imposables	Participations Familiales	
141			de 0 à 300	0,50 €	
142			de 301 à 600	1,25 €	
143	Barème LEA	Quotient Familial	de 601 à 900	2,25 €	
144				1 enfant inscrit	2 enfants et ·
145			de 0 à 5 963	1,30 €	
146			de 5 964 à 11 896	4,60 €	
147			de 11 897 à 26 420	7,10 €	
148			de 26 421 à 70 830	8,70 €	
149			de 70 831 à 150 000	18,60 €	13,40 €
150	Hors barème LEA	Revenus annuels imposables 2012 (loi de Finances 2013)	au delà de 150 000	29,60 €	23,70 €
151	*Paiement dû pour toute journée commencée*				
152	** activités de loisirs sans hébergement*				
153	Dispositif LEA (Loisirs Équitables Accessibles) : il se substitue à celui de délivrance de bons CAF et s'appuie sur une politique tarifaire modulée sur le quotient familial (QF) et non plus sur les revenus imposables.				
154	• **CENTRES DE VACANCES, ECHANGES INTERNATIONAUX, MINI SEJOURS**				
155	Dominante d'activités proposées : sport, culture, loisirs (tarifs par jour et par enfant)				
156	- sensibilisation	15,70 €			
157	- découverte	21,00 €			
158	- initiation	26,20 €			
159	- pratique	31,50 €			
160	- perfectionnement	36,70 €			

Source

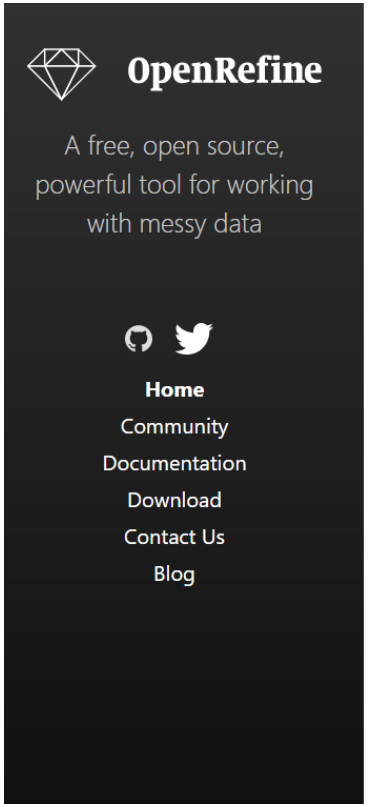
Challenge 2 : data quality

Or hardly exploitable...[2]

	A	B	C	D	E	F	G
13							
14	Nombre de places assises..... 2 306 places						
15	(dont médiathèques centrales 1 665 places)						
16							
17	Entrées..... 982 793						
18							
19	Inscrits..... 56 821						
20							
21	Prêts..... 1 955 381						
22							
23	Ensemble des collections..... plus d'1 000 000 de documents						
24	dont :						
25	disques et textes enregistrés 97 792						
26	vidéocassettes et DVD 42 839						
27	partitions 3 547						
28	documents adaptés aux personnes handicapées..... 10 075						
29	les collections patrimoniales : 321 398						
30	pages numérisés..... 226 300						
31							
32	Services informatiques et numérisations						
33	Accès à la vidéo à la demande,						
34	Accès au téléchargement de musique et de livres						

Source

Challenge 2 : data quality

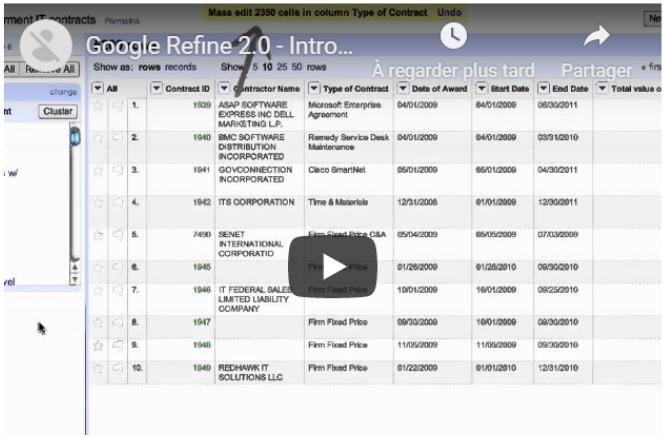


Google News Initiative

Introduction to OpenRefine

1. Explore Data

OpenRefine can help you explore large data sets with ease. You can find out more about this functionality by watching the video below.



Download Open Refine (+ tutorials)

Challenge 2 : data quality

Mode d'emploi

Cet outil vous permet de créer un fichier CSV en vous assurant qu'il est conforme à un schéma, c'est-à-dire que ses données sont complètes, valides et structurées.

1. Sélectionnez le schéma qui vous intéresse dans la liste déroulante, les schémas disponibles ici étant ceux référencés sur schema.data.gouv.fr.
2. Remplissez le formulaire : vous allez ainsi créer la première ligne de votre fichier CSV.
3. L'outil vous prévient d'éventuelles erreurs de validation, le cas échéant vous pouvez les corriger.
4. Une fois votre formulaire valide, les valeurs apparaissent sous la forme d'une ligne dans un tableau récapitulatif.
5. Vous pouvez alors choisir d'ajouter une ou plusieurs lignes (répétez les étapes 2 à 4) ou télécharger le fichier CSV correspondant au tableau récapitulatif.

Choisissez un schéma à utiliser :

csv-gg est un logiciel libre développé par Etalab.
Attention il s'agit d'un projet expérimental. En cas de question ou de problème, vous pouvez ouvrir un ticket ici ou nous envoyer un email.

Have a look to CSV GG (an Etalab initiative)

Tidy data

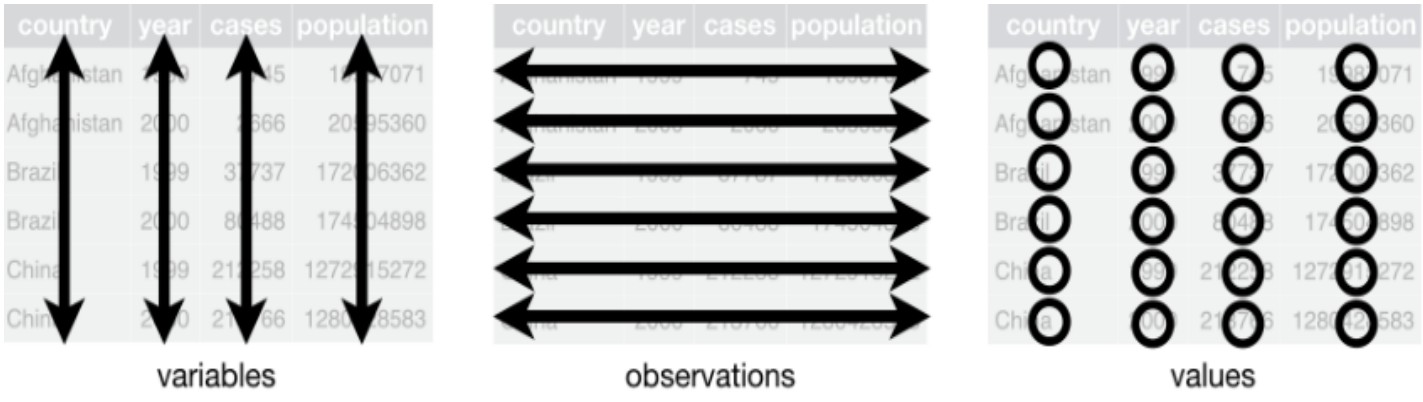
Tidy data Paradigm (Hadley Wickham)

> “All happy families are alike, but every unhappy family is unhappy in its own way” – Leon Tolstoï
>
> “Tidy datasets are all alike, but every messy dataset is messy in its own way.” – Hadley Wickham

Tidy data

Tidy data principles ("données ordonnées")

- Each variable in the data set is placed in its own column
- Each observation is placed in its own row
- Each value is placed in its own cell



- Each type of unit observed is placed in its own table

Manipulating (Open) Data

Le pipeline de données

According to you, what are the steps while manipulating data ?

Data manipulation through spreadsheet : filter & sort

Form a group of 2-3 students (please 1 French per group)

- Find and download the dataset with the number of ISF licence payer in 2017
- What is the difference between the tabs *définitif* and *définitif_patrimoine* ?
- Sort data from tab *définitif_patrimoine* so as to be ordered exactly in the same way as those from tab *définitif*
- In column h, compute the total ISF paid *per city*
- Which cities paid the most ISF in 2017 ? (in absolute terms)
- Which cities from "Ile de France" region paid the most ISF in 2017 ? (in absolute terms)
- Which Parisian departments/districts paid the most ISF in 2017 ? (in absolute terms)

Data manipulation through spreadsheet : functions

- What is the total amount of ISF given by those cities ?

Use the function *sommeprod / sumproduct*

=> 2 535 350 477

- Let's pretend that the average ISF paid by Levallois-Perret inhabitants is representative of the average ISF paid by the French licence payers.
- Compute the hypothetic total amount of ISF that it would generate in this scenario. Would the French government be more profitable in that scenario ?

Block cells with **$**

- If you want to visualise at a glance the difference of property/estate, use the **conditionnal formating**

Data manipulation through spreadsheet : functions

Data manipulation through spreadsheet : functions

- You would like to get the number of ISF licence payers in "Ille et Vilaine" department but without filering nor sorting...

Function *somme.si / sum.if*

1 836

- You would like to get the number of ISF licence payers in Paris but only in districts where the average ISF paid is above 10 000€ (and still without filtering nor sorting) Function *somme.si.ens / sumifs*

60 176

Data manipulation through spreadsheet : functions

- Same constraints (no filtering/sorting) - but now we want to know the number of cities in Ille et Vilaine that have ISF taxpayers

Function *nb.si / count.if*

3

- And how many Parisian districts whose average ISF is above 10 000€ are there ?

Function *nb.si.ens / count.ifs*

17

Data manipulation through spreadsheet : Vlookup

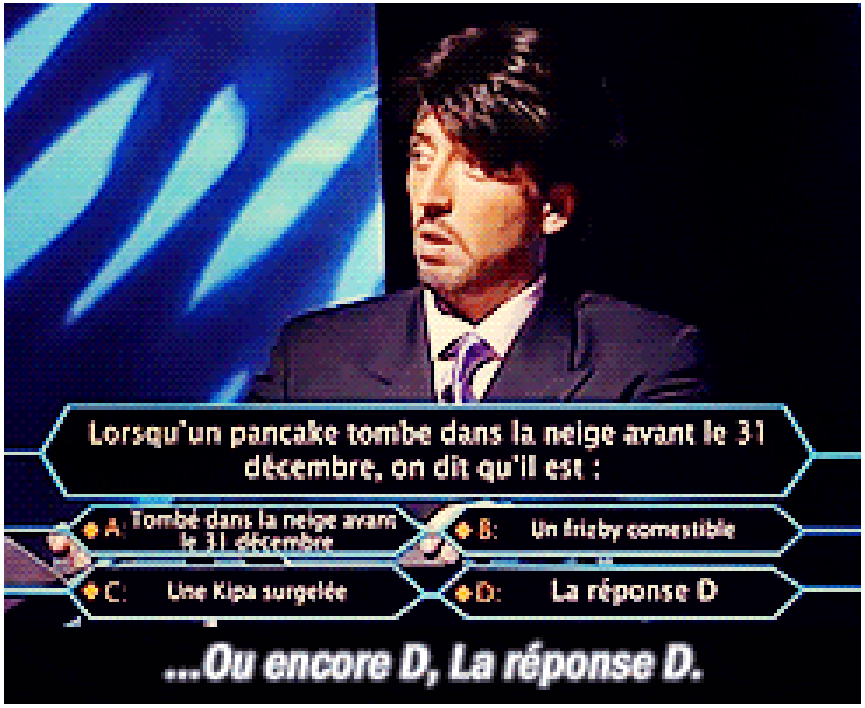
- ISF data are interesting but there is one key variable missing...

--

population !

- Find and download a dataset enabling you to get demographic data given city per city
- Look at its structure and select the variables / columns that you want
- Open a now tab in your ISF spreadsheet, copy/paste the columns from your tab "patrimoine" and add a new column where you will collect the number of inhabitants associated to each city that have ISF taxpayers. For this, you need to use the function *rechercheV* / *Vlookup*

Data manipulation through spreadsheet : Vlookup



See you next week !

So happyyyyy togetheeeeer !

Thank you !

Contact : timothee.gidoin@sciencespo.fr