Cover page

Ironhack logo

Data analysis

Title of project : Who’s who in civic tech ?

Tatiana de Feraudy

June 2023

Table of contents

**8 Data cleaning and Exploratory data analysis**

**10**

Highest-rated book 11

Language distribution of books 12

Distribution of ratings 13

Book Counts by Publication Year 14

Top 10 Authors with most titles 15

Top 10 Publishers 16

Correlation between Number of Text Reviews and Number of Pages vs Average Rating 17

Genre Analysis, Top 5 Genres by Average Rating and Distribution of Average Ratings 18

**SQL or NoSQL 20**

**ERD**

**MySQL Queries 22**

Query 1) Top 10 books with the most pages 23

Query 2) Who wrote the most books? 23 Query

3) How many books are there without a genre? 24

Query 4) Percentage of books without genre in the dataset? 24

Query 5) Favorite genre of authors (random section taken from the query result) 25

**Conclusion**

Requirements

C1. Qualifier les données grâceà des outils d’analyse et de visualisation de données en vue de vérifier leur adéquation avecle projet

Dans le cadre d’un projet, à partir d’un besoin exprimé, le/la candidat(e) doit :

valider les sources de données enfonction des résultats de l’analyse exploratoire,

produire les visualisations synthétisantles caractéristiques du jeu de données.

Critères :

Les données présentées répondent au besoinfonctionnel et sont disponibles,

les données retenues suite à l’analyse exploratoire sont pertinentes : l’utilité desattributs est démontrée,

les visualisations rendent compte de l’analyseeffectuées,

les visualisations sont cohérentes etexplicitées.

C2. Concevoir une base de données analytique avec l’approche orientée requêtes en vue de la mise à disposition des données pour un traitement analytique ou d’intelligence artificielle

Dans le cadre d’un projet, à partir d’un besoin exprimé, le/la candidat(e) doit :

décrire la stratégie de nettoyage desdonnées définie,

produire les éléments de modélisationdes données,

identifier le ou les systèmes de gestion de bases de données analytique adaptésà la modélisation des données,

décrire la procédure de mise en place dusystème de gestion de base de données

la stratégie de nettoyage présentée est cohérente avec les résultats de l’analyseexploratoire,

la modélisation des données respecte uneformalisation dans une approche orientée requêtes,

la modélisation comprend : les patterns derequêtes, les clés primaires, les indexes, les entités (collections, ou documents, etc. en fonction de la base de données), dans le cas d’une base de données NoSQL\*, les relations, s’il y en a, respectent les méthodes standards : embedding\* ou encorereferencing\* par exemple,

la base de donnée est choisie au regard de lamodélisation des données et des contraintes du projet,

la procédure de mise en place décrit lesétapes à suivre,

le résultat de l’exécution de la procédure estun système de gestion de base de données conforme à la modélisation.

# Introduction

The word ‘civic tech’ was first used in France in 2013, when the Knight foundation released its report on the investment and political potential of these technologies[[1]](#footnote-1). From that date until the Open Government Summit held in Paris at the end of 2016[[2]](#footnote-2), activists, entrepreneurs, investors and public policy-makers participated in creating a “civic tech” ecosystem in the country and giving meaning to the word. The great diversity of actors involved in the ecosystem (NGOs, companies, researchers, public agencies, local and national governments…) is what makes it interesting, but also what makes it difficult to define what ‘civic tech’ means.

For my PhD research in political science and sociology, I study the people, organizations and networks that have contributed to defining civic tech in France and making digital citizen participation a public problem[[3]](#footnote-3). This research is conducted using mainly qualitative methods, i.e. semi-structured interview and participant observation, as well as through the collection of “grey” literature (reports, information and communication material produced by governmental entities, NGOs, companies, etc.) and media on the topic.

The goal of the project presented in this report was therefore to use data analytics to enrich my research and better understand the civic tech ecosystem in France. This general goal was divided into two more precise objectives:

1. Use quantitative methods to ‘decenter’ my approach and gain a broader, more holistic understanding of civic tech. Since data analysis methods require a large amount of data, this also implied thinking about what “big data” could be collected on this topic, and how it could be analyzed.
2. Develop a framework that could be shared and replicated with other researchers or analysts. From data collection and cleaning to database construction and data analysis, I made an effort to think about how the approach could be replicated. This would namely allow draw comparisons with other related ecosystems (for instance the edtech, health-tech, fintech, and more generally the different start-up and tech for good ecosystems in France).

# Project planning

Building a database and analyzing data on the civic tech ecosystem in France required different types of tasks. These can be divided into data collection, data cleaning, data analysis, data visualization and data modeling activities. Although we have separated them in this planning and in the report, they were not conducted in a strict chronological manner.

In my opinion, this is one of the key lessons learned through this project (and more broadly during the bootcamp). Data collection is informed by the expected design of the database, data cleaning is conducted in several steps based on the requirements for data analysis and modeling, and data visualization happens at every stage of the project, although it takes different shapes.

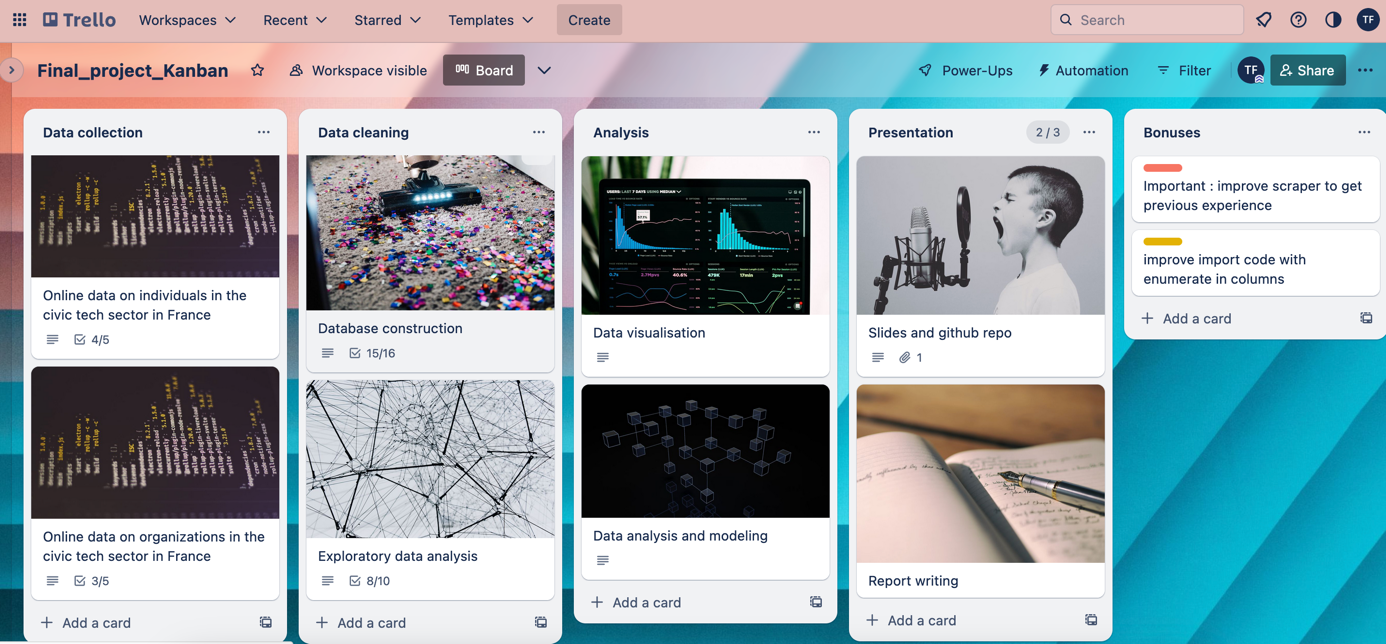


Figure 1. Screenshot of the kanban board for the Who's who in civic tech? Ironhack projet, 2023.

## Data collection – 2 days

Data collection was conducted in several steps:

1. Identifying relevant and available data sources.
2. Scraping LinkedIn profiles and Pappers’ websites to produce individual files (per person, per company, per NGO).
3. Producing a framework and tools to increase the dataset and automate data collection. While we started with a limited amount of data, we progressively increased the size of our dataset, and now have the tools to keep doing it if necessary.

## Data cleaning – 3 days

Data cleaning is by far the most time-consuming step of data analysis. As we mentioned earlier, cleaning data implies thinking ahead and imagining the possible uses of the data. To summarize, different types of activities were conducted to improve different elements of the data.

1. Structure: scraped data is oddly structured, with nested dictionaries and dataframes, repeated information, and some source-specific information that is not useful for analysis. Cleaning it is necessary to identify the ideal structure to keep as much detail as possible while achieving a clear database structure.
2. Readability: data from websites is produced in order to be read by humans. However, we need the data to be as raw as possible to be able to conduct analyses on it. We need to remove special characters, harmonize data types, and more generally make data readable for non-humans.

## Data analysis and visualization – 3 days

These two tasks are described together here as they often go hand in hand. While our analysis code is often divided into “data discovery”, “data cleaning”, and “data analysis”, these often happen at the same time. For instance, data cleaning often implies visualizing the distribution of data, the weight of different categories, or the importance of correlations between features. Our data analysis and visualization steps aimed at answering two main questions.

1. Who are the people in the French civic tech ecosystem?
2. What characterizes the organizations associated with this ecosystem?

The data analysis we conducted is based on the data collected, so it has its limitations. For people, we worked on their education, their latest professional experiences, the languages they speak, and how they define themselves. For organizations, we mainly analyzed information on creation date, city, founders/ directors, whether they are still active, their field of activity and their financial performance (sales, revenue…).

Data analysis and visualization was done through three different tools:

* Jupyter notebooks coded in Python, for table construction, exploratory analysis, basic plotting of relationships and distributions and reformatting/ exporting new datasets (libraries: pandas, json, numpy, matplotlib, seaborn, and sqlalchemy)
* MySQL Workbench, which allowed to organize the tables and their relationships in a database format and to formulate queries which gave new insights on the data.
* Tableau (desktop version) to produce visual renditions of data analysis.

## Project tracking, documentation, presentation and archiving – 2 days

Some of the tasks to produce data analysis projects are hidden behind our Trello board. These include the project tracking activities (creating plans, making to-do lists), discussions with colleagues and teachers to decide on the best options[[4]](#footnote-4), documentation of activities (keeping readme and gitignore files updated for the github repository, reviewing and organizing notebooks for readability), organizing files and repositories, and of course preparing presentation materials (reports and slides). Stating them here seemed important as part of this work is aimed at sharing the process with other researchers who may be interested in these tools.

# Data collection

## Challenge 1. How do we collect data when there is no data?

One of the reasons why I started by PhD a few years ago is precisely that there is very little data about the civic tech ecosystem in France. The information about the topic is mostly produced by the actors themselves, and is mostly promotional material, whether it is in press articles, blogs or conferences. Moreover, since companies in the field are relatively young and unstable, they are not very keen on disclosing information, and neither are individuals.

The qualitative data collected during my PhD research provides a very good understanding of individual trajectories and professions, but only for the 50 individuals with which I was able to conduct semi-directive interviews. I also studied organizations and networks based on participant observation during three years in the field. However, I was often faced with questions that I didn’t have an answer to. One of them was: “what is the best technology?” For this one, my answer was usually that it depends on what you want to do with it.

Another question was: how do we know which company will survive? Although I am still not convinced this is something that can be predicted (otherwise, no company would every fail), or that it is something that should be predicted (cf. self-fulfilling prophecies), I thought it would be interesting to see if we could model whether civic tech (or more general, innovative “for good” startups) would survive or not based on the data we have. We can also ask the question in terms of whether people will stay for a long time in the field, or not.

Although the modeling part of the project is yet to be done, the data collection was informed by this objective, as well as by the availability of data. There were three sources of data:

* my PhD research provided me with a list of people and organizations that I could focus on. These lists were namely constructed by analyzing press articles and extracting the people who talked about civic tech in the press and the organizations that were mentioned.
* the LinkedIn[[5]](#footnote-5) profiles of individuals related to the civic tech ecosystem,
* the information on companies and NGOs provided by the Pappers[[6]](#footnote-6) website.

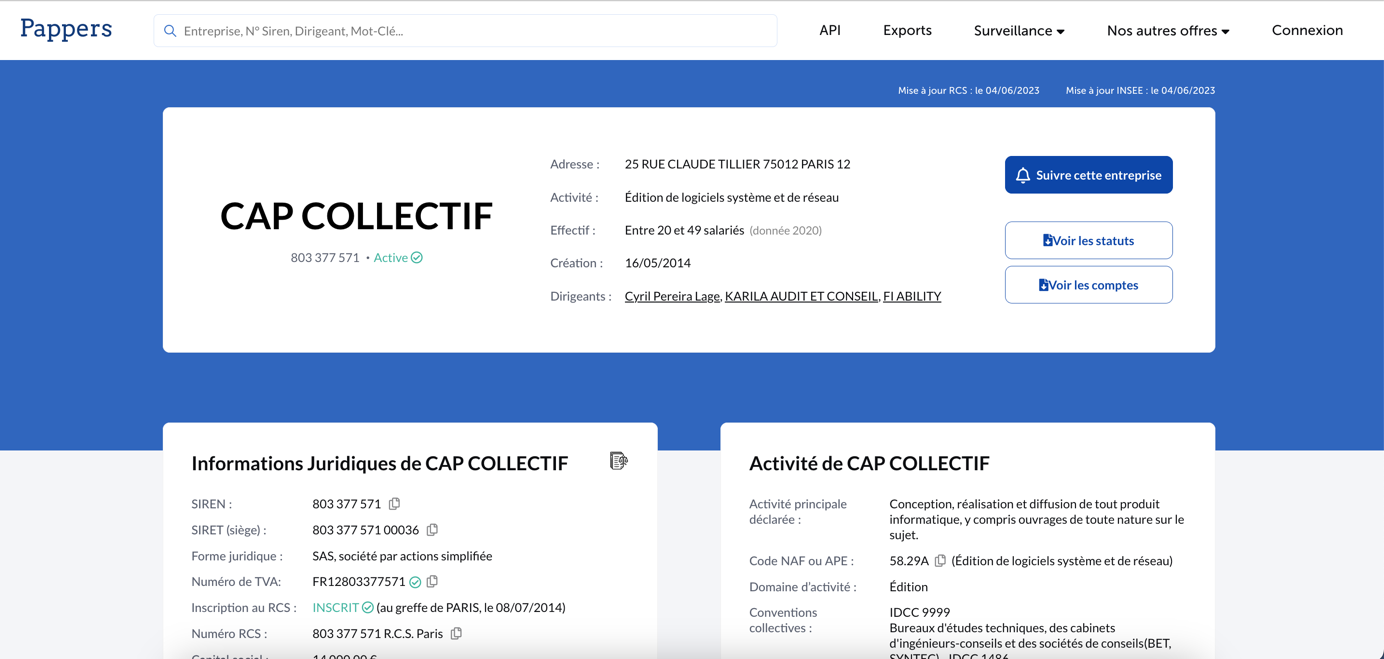


Figure 2. Screenshot of the Pappers page concerning the civic tech company Cap Collectif (june 4th, 2023)

## Challenge 2. How do we collect data when people don’t want us to?

The two online data sources identified don’t allow to download data for free. In addition, the requests library does not function for these websites (it returns a 403 forbidden error), even for single page requests. Fortunately, I had the chance to work on web scraping with research colleagues a while ago, using the R language. For this project, I adapted several different scripts:

* Two scripts written namely by Constantin Brissaud and Aurélien Goutmesdt. The first script, in R, uses the Selenium library on R to automate the process of researching a name on Google and saving the urls that the search returns (for the two first result pages). It produces a csv file with a column of names and a column of urls. The second script (in Python) reads through the csv file and extracts the urls including linkedin.
* The LinkedIn scraper developed by Tom Quirk (<https://github.com/tomquirk/linkedin-api>). Although some of the endpoints have changed and some of the code required adaptation, this repository (recommended by Constantin) provides Python code to extract information from LinkedIn. However, LinkedIn blocked me after a certain number of requests, so it was necessary to adapt the code.
* For the Pappers website, I chose to manually save the source pages of the organizations I wanted to study (the code to automate this is under development).



Figure 3. Code snippet from Tom Quirk's LinkedIn scrapper for profile scraping

## Challenge 3. How do we fix the structure in scraped files?

The LinkedIn scraper returns json files, while the collection of Pappers’ pages returns html files. Both formats need to be transformed to create tables of information that can be used in a database. Our main challenge, in both cases, was to handle nested information (nested dictionaries and lists in json files and nested dataframes in the html files).



Figure 4. Example of a nested dataframe in the dictionary created from the html files.

## 

Figure 5. Excerpt from a json file produced by the LinkedIn scraper (produced with https://codebeautify.org/jsonviewer)

1. **Handling the html structure of nested dataframes**

For the dictionary of companies’ information, the re-structuring was done in an iterative manner. After exploring the layers and sub-layers of the dictionary (cf. figure 4), I used the pandas library to select specific elements in the dictionary and build a new dataframe through a mix of concatenating and merging options (see figure 6).

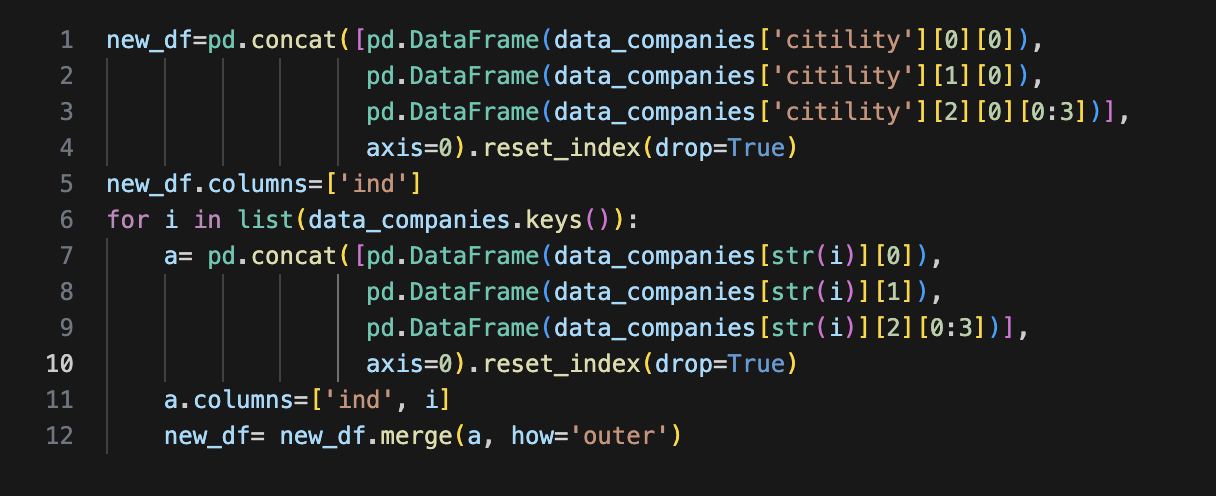


Figure 6. Creating a new dataframe from elements in nested dataframes in a dictionary

1. **Handling the nested dictionaries in the json files**

Json files have the advantage of having a dictionary structure, which makes it easy to navigate in it. However, when the dictionary has a complex structure with many sub-dictionaries, it is quickly easy to lose ourselves in it. In addition, we wanted to be able to read the data to think about how to structure it and organize it.

Our main issue was that for each individual, all his or her work experiences were lumped together in a dictionary that included, for each work experience, the name of the company, the number of employees in the company (in a sub-dictionary), the job title, the start date (with a sub-dictionary of month and year), etc… This was also the case for the education, languages, honors received, publications, certifications, volunteering, and projects sections of the profiles.

After trying to explode all these sections into individual data points, I noticed many ended up producing columns with a lot of missing values, and with little explaining power. Although this detail could be useful later, for specific queries, for this project I chose to focus on education and experience.



Figure 7. Function to import json files in a dataframe while restructuring data organization

I decide to include the data structuring directly in the json import function, for the education, experience, and language sections, as shown in figure 7. For the other sections, I only created a new column that states whether or not the section was filled. For languages, I chose to also keep the information on whether the person had included more than 3 languages, as this seemed to be an important element to study international companies and highly educated individuals.

Later on, I chose to further separate the information (exploding each of the work experiences and of the education programs, for instance), but using the panda Series function, as shown for example in figure 8. More detail of the column construction for each of the scraped sources is available in the commented jupyter notebook files named companies\_data\_scraping and jsons\_to\_dataframe\_cleaning.

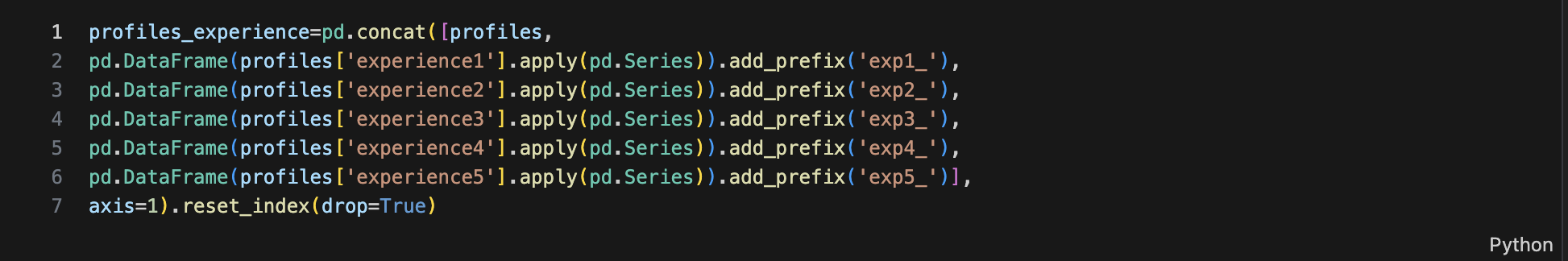


Figure 8. Creating new dataframe columns with the pd.Series method

# Data cleaning

In addition to selecting relevant information and structuring it into tables, data cleaning was necessary to simplify and make the datasets more readable. I am here separating two types of data cleaning operations.

The first part of data cleaning is purely functional. It includes dropping duplicates and columns that don’t have a significance for us (e.g. identifiers from source website), renaming columns to avoid spaces and special characters, managing the index, and dealing with data types. The figure 9, below, provides an example of this type of data cleaning, here replacing numbers formatted as text by only numeric characters, in order to be able to assign a numeric data type to the column and analyze it as such. Figure 10 provides a before/ after screenshot of a table that goes through simple data cleaning.

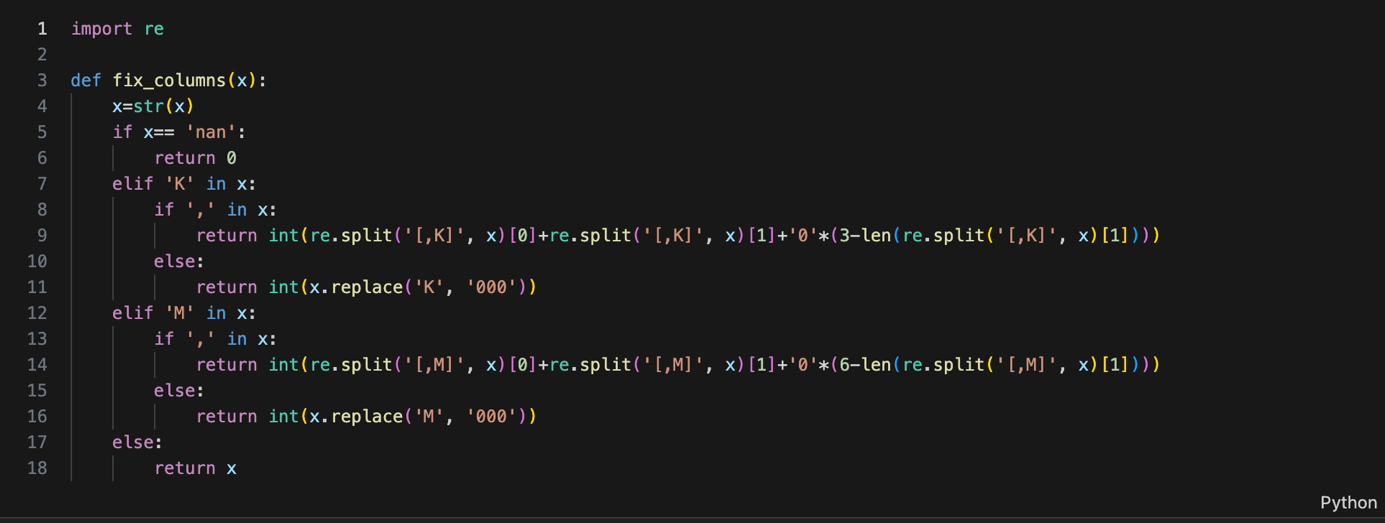


Figure 9. Function to deal with numbers stored as text.





Figure 10. Before and after: functional data cleaning

A second type of data cleaning operations could be defined as more analytical: they imply recoding some of the information, creating new categorical variables, in order to make it possible to analyze a large amount of information. One characteristic of our dataset is that it is mainly text, and although that will be precious when doing NLP treatments, it is not so practical for initial data exploration and database construction. Below are some examples of the recoding operations I performed to have a dataset that would be appropriate for EDA, visualization and direct querying.

Figure 11. Recoding to have a limited number of categories (location) or to create new binary columns that qualify a position (title)

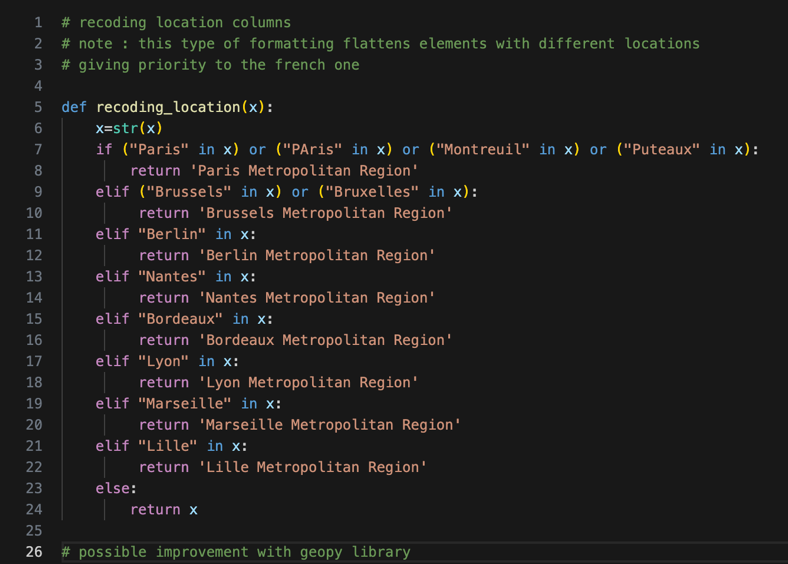




Figure 12. Using regular expressions (regex) to extract meaningful information from columns and create new categories (date, registered or not) of single data type.

# Database construction

## Building separate tables

In order to be able to analyze our data, we want each data point to be unique, and to have a unique identifier. In order to achieve this, I split my datasets into 7 different tables.

The “names” table includes only first and last name as well as individual ID. I decided to separate this information to allow for a minimum of anonymization. This is not enough for a larger scale project dealing with personal data, but for the scope of this experimentation, at least it allows to envision how to separate personal data, specify different authorizations for this part of the database, and respect GDPR requirements.

The “people” table provides basic information about the individuals. It contains a lot of the original information as text, as well as some basic information recoded as binary variables, as explained in the sections above.

The “people\_experience” table repeats the individual IDs for each professional experience of the person (n=5), while also allowing to have an id specific to each experience (or each row).

The “people\_education” considers each education experience as an instance (row), and again relates it to the individual ID so we can connect it to the people table.

The “companies\_info” table provides basic information on companies (commercial, for profit) in the civic tech ecosystem, including whether or not they are still active.

The “companies\_finance” table contains the financial information about the companies in the previous table, with a distinct row for each year for which there is information.

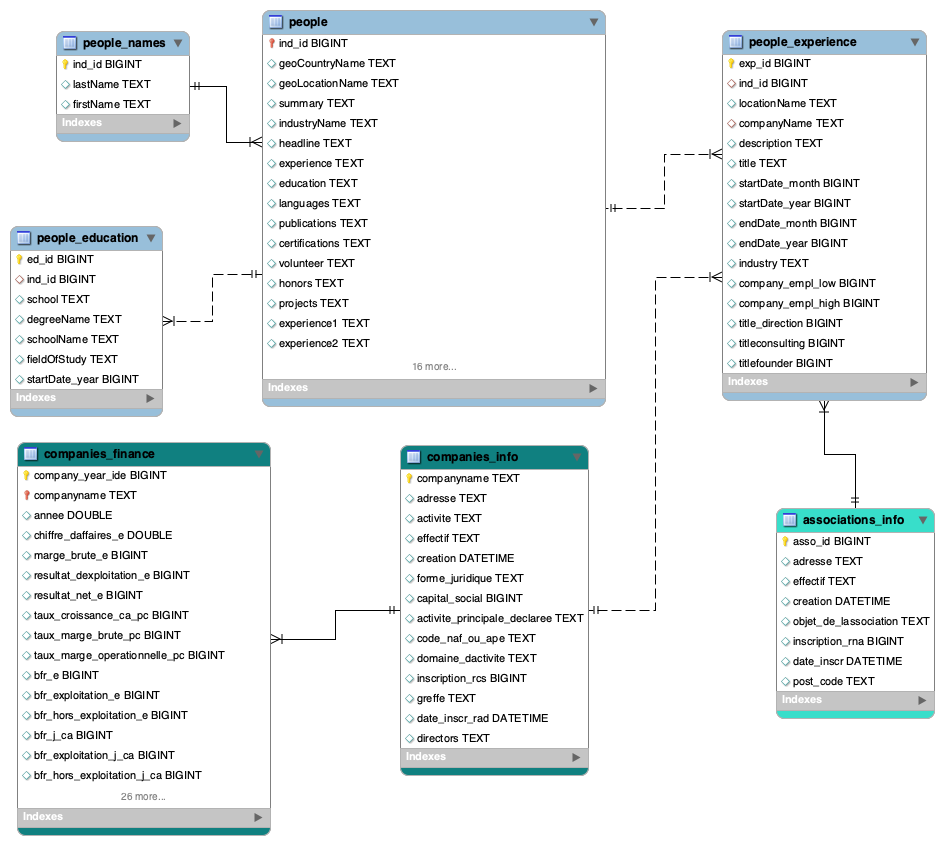
The “associations\_info” table was produced from a subset of the companies\_info table : some of the organizations identified are not for profit. Since these organizations have different data points than the companies, I chose to have a separate table. However, the association name serves as an identifier to relate these organizations to the people\_experience table, as some people have identified experiences in NGOs.

Annex 1 provides screenshots of the first rows of these 7 different tables.

## NoSQL vs. SQL

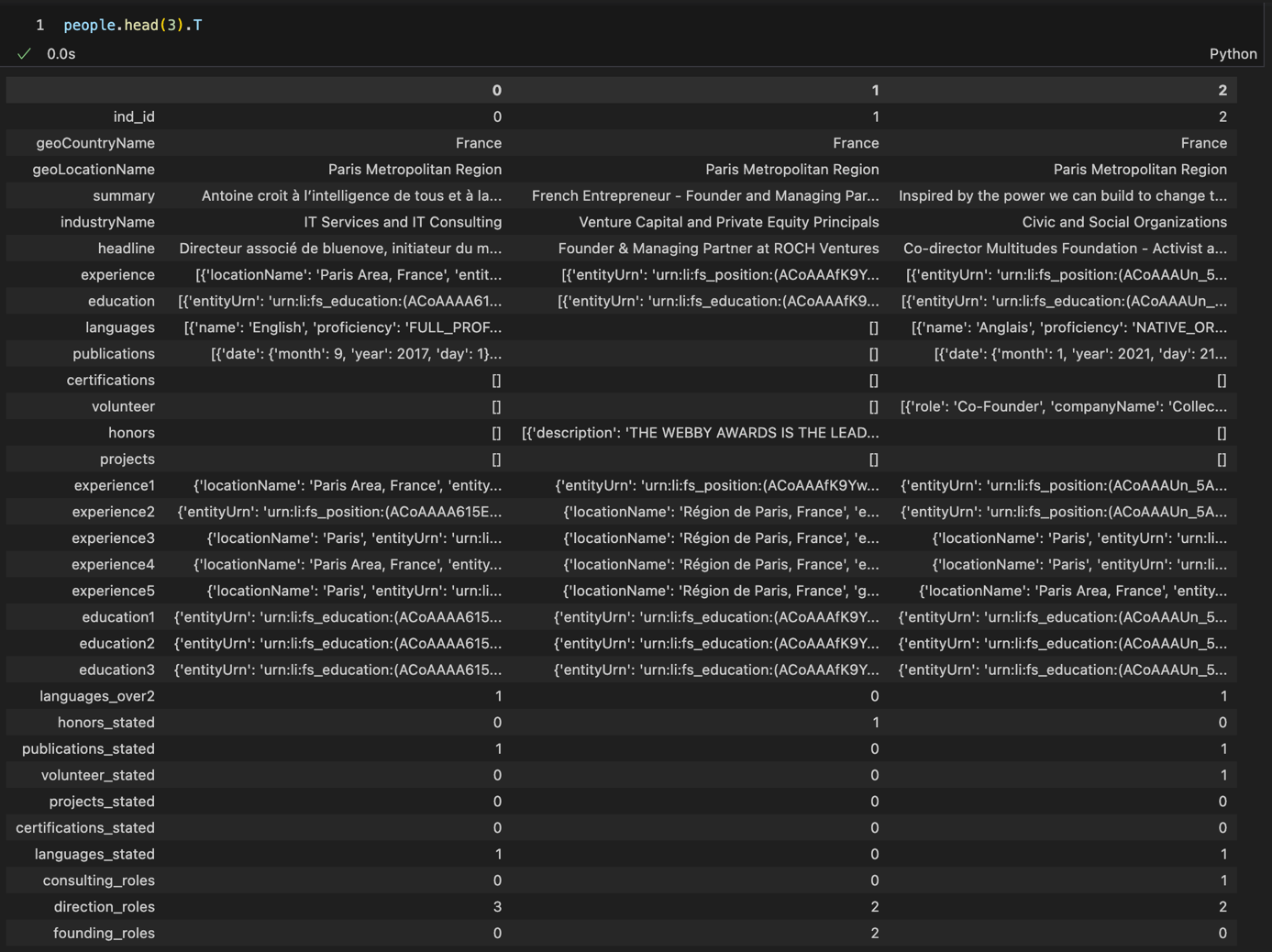
## Entity relationship diagram

Below is the diagram that shows the relationships between the tables in our database.

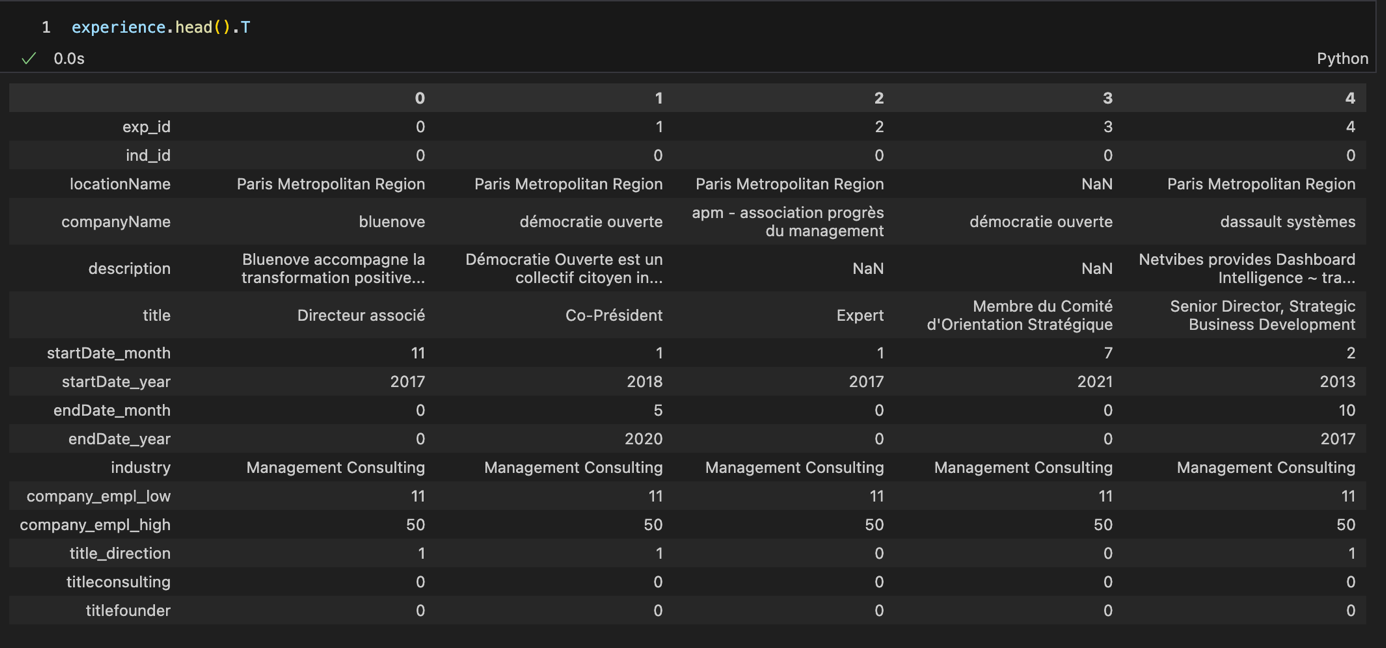


# Annex 1 – Screenshots of database tables (first rows)

1. People (transposed)



1. Experience (transposed)



1. Education (transposed)



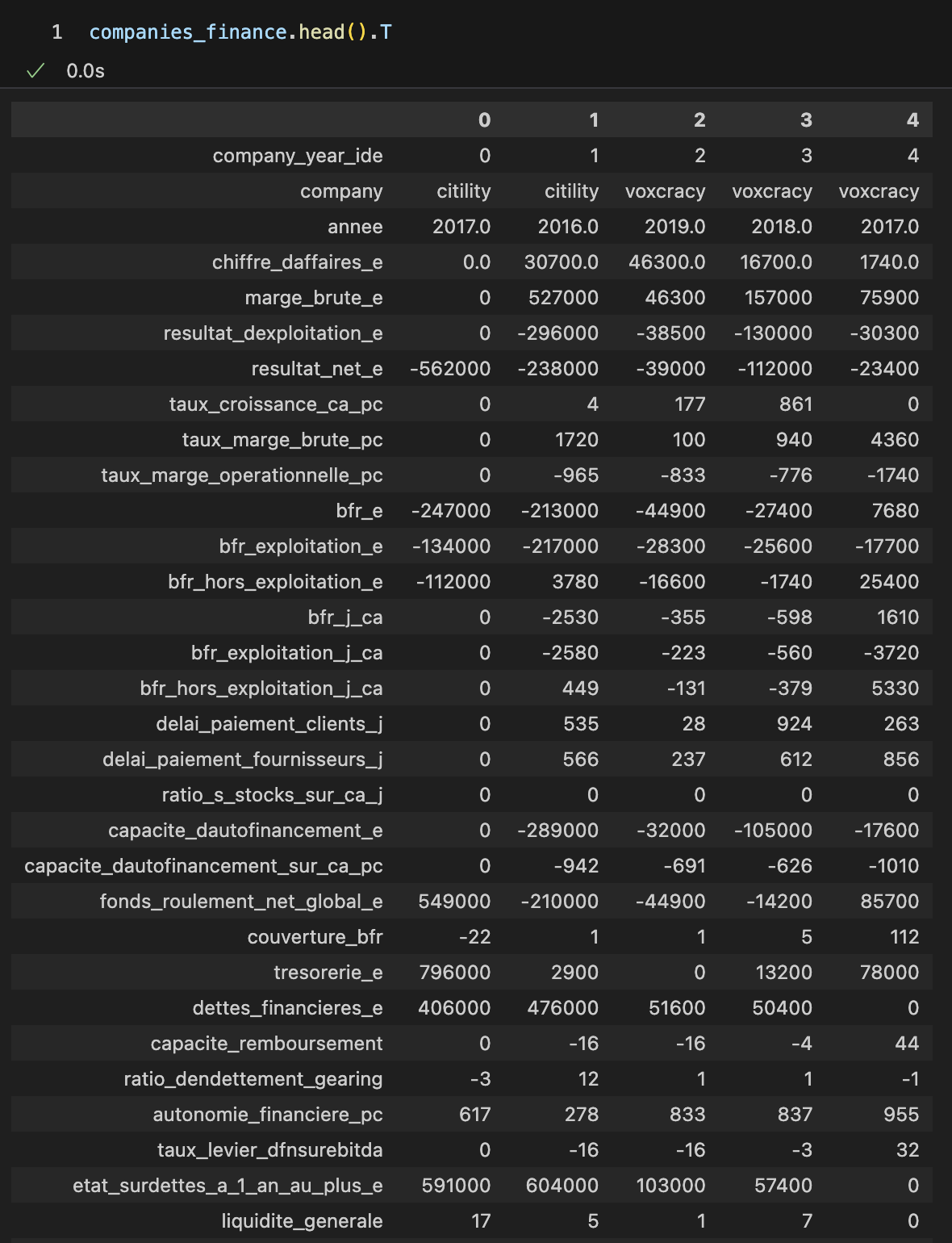
1. Associations

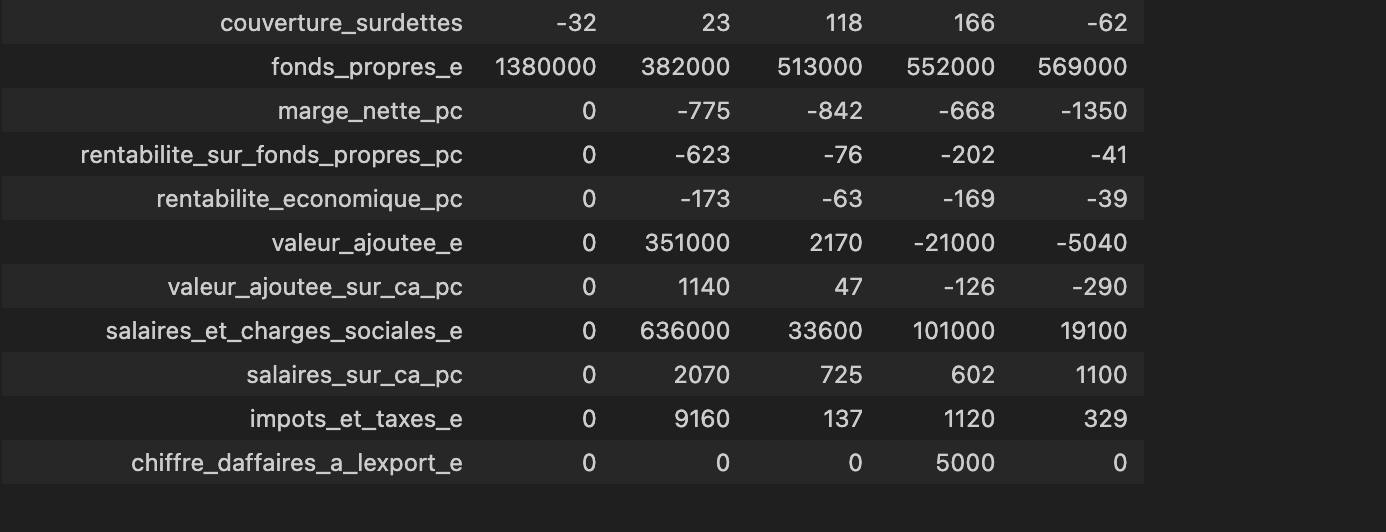


1. Companies information (transposed)



1. Companies financial information





1. Patel Y. et. al., *The Emergence of Civic Tech : Investments in a Growing Field*, Knight Foundation, December 2013, URL : <https://knightfoundation.org/features/civictech/>. [↑](#footnote-ref-1)
2. Barack Obama’s administration was the first to use the terms ‘open government’ to describe a program of governmental actions aimed at making public action more transparent, collaborative (with civil society, including the private sector) and more participative (i.e. involving citizens). This program is detailed in Obama B., *Transparency and Open Government. Memorandum for the heads of executive departments and agencies*, January 21st, 2009, The White House, URL : <https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government>.

   The Open Government Partnership is an international organization that promotes this program around the world. Founded in 2011, it now includes more than 70 countries, whose governments commit to voluntarily improving their practices. Objectives are stated in multi-year plans which are reviewed by independent auditors. In December 2016, the French government, as co-chair of the initiative at the time, hosted the OGP international summit in Paris. [↑](#footnote-ref-2)
3. A ‘public problem’ is an issue that is recognized as requiring ‘public’ action (as opposed to an individual or a ‘private’ problem). Unemployment, water pollution, childcare or citizen participation can be defined as public problems. The concept is interesting because it invites us to see the problems that public policies address not as issues that get a response because they are more important than others, but as “constructions”. When studying public problems, we focus on the work done by different actors (activists, companies, lobbies, individual and collective groups in government agencies, researchers…) to define the problem and acceptable solutions, identify who is responsible, and convince the public sector to take action. For more detail, cf. Erik Neveu, « L’analyse des problèmes publics. Un champ d’étude interdisciplinaire au cœur des enjeux sociaux présents. », *Idées économiques et sociales* 2017/4 n°190, p.6-19. [↑](#footnote-ref-3)
4. Many thanks to Thomas, Andy and Elnara for their advice and patient support during the bootcamp, to Lucien, Robin, Katrina, Axèle, Adèle, Matthieu, Ismael, Romain and Louis for the group work, park lunches and laughs, to Emma and Choco for. [↑](#footnote-ref-4)
5. LinkedIn is an online social network focused on professional relationships (cf. <https://www.linkedin.com>) [↑](#footnote-ref-5)
6. Pappers is a French company that makes open public data on businesses available on an online platform (<https://www.pappers.fr/a-propos>). Downloading the data, however, requires paying to use the API. [↑](#footnote-ref-6)