Square

GIVING A FUTURE TO TALENT

DATA

# OPEN DATA AS A BUSINESS ACCELERATOR

SAMY FRACSO, ADRIEN AUBERT,
MICHAËL TOUZE, ALEXANDRE ZAVAGNO,
NICOLAS DERRIEN, HUGO BATACHE

FOCUS I MARCH 2022

Square

GIVING A FUTURE TO TALENT

DATA

# OPEN DATA AS A BUSINESS ACCELERATOR

SAMY FRACSO, ADRIEN AUBERT, MICHAËL TOUZE,
ALEXANDRE ZAVAGNO, NICOLAS DERRIEN, HUGO BATACHE

# TABLE OF CONTENTS



Square

GIVING A FUTURE TO TALENT

# FOREWORD

While data unequivocally represents a new form of assets for organizations and constitutes as such a way to activate levers of growth, it is also generally admitted that its potential has yet to be fully grasped, owing to multiple difficulties of empirical and historical natures: technical infrastructures not always updated, cultural gaps and scarcity of skills, etc.

Open DataData is a lever to improve this state of play. It refers to all the digital data that is free. It relies on two key principles: sharing (data must be accessible and free of cost to anyone) and reusing (data must be easily exploitable and reusable).

In the banking sector, Open Data remains scarcely used although it constitutes an indisputable source of information that helps deliver key analyses as a complement to internal data. It is true that multiple uncertainties and constraints frame the use of Open Data: unproven performance gains, significant load of process adaptation and upskilling, transparency and quality of data, etc.

Hurdles to be navigated can seem numerous; the objective of this new Square Focus is to illustrate the possible benefits of resorting to Open Data in operational projects led by banking institutions. After a brief overview of Open Data, we will focus on the advantages and difficulties frequently encountered in their use, after which we intend to overtake these hurdles by demonstrating the value of Open Data through 2 use cases.

# 1.

# OPEN DATA: AN OVERVIEW

Before diving into the details of the advantages and limitations of Open Data in relation to their practical use, we would like to restate in this section the definition and context of development of Open Data; we will also cover the different public initiatives that support Open Data.

## 1.1. GENERAL FRAMEWORK

Open Data encompasses all digital data that is free of rights. It relies on two key principles: sharing (data must be accessible, free of charge, by anyone) and reusing (data must be easily exploitable and reusable). If multiple types of open licenses co-exist at the international level, two licenses are currently applicable to public information in France:

- **The Open License (Licence Ouverte, LO):** a French specificity that was designed by Etalab, this is the most liberal license, granting the following rights on the data source to the user:
  - reproduce and copy it
  - adapt, modify, extract, and transform it, in order to create "derivative information", products and services,
  - transmit, diffuse, redistribute, publish and disseminate it,
  - exploit it commercially,

  The only compensation being to acknowledge the original author of the database.

- **The Open DataBase License (ODbL):** an international license derived from the Anglo-Saxon common law, granting reproduction, copy, adaptation and modification rights; the main difference with the Open License lies in the condition to share data under the same ODbL license (except in the case of financial compensation).

To learn more: Licences - data.gouv.fr

## 1.2. FRANCE'S HEAD START

Open Data is made available both by private and public institutions. In 2011, The French government has created Etalab, an administration[1] that "coordinates the design and implementation of the strategy of the State in the field of data (https://www.etalab.gouv.fr/qui-sommes-nous/). That same year, Etalab launched the data.gouv.fr website, which lists all the public data made available by various French administrations. Data from a large number of sectors can be retrieved: culture, health, economy, etc.

In 2016, the Law No. 2016-1321 of October 7 for a Digital Republic has confirmed the effort and the lead of France in the field of opening the public data. The law has indeed modified the conditions to access public data, with Open Data becoming the standard rather than the exception. All public services are affected: State administration, local authorities for greater than 3,500 inhabitants, public institutions, and private organizations in charge of public service.

Among others, the article 14 of the Digital Republic Act promotes the creation of a set of data said of reference, identified as the data having the greatest economic and social impact. Out of the 9 currently listed sources, two will mainly be used here as a framework for the development and promotion of the economic potential of Open Data:

- The Sirene database of enterprises and their establishments
- The Official Geographic Code database (Code Officiel Géographique, COG)

The full list of reference data can be accessed at the following URL: https://www.data.gouv.fr/fr/reference

Another breakthrough brought upon by the Digital law, the intake of new data said of public interest:

- Public service delegation (energy, waste);
- Case law;
- Mobility & transportation (confirmed in the law of 2019 on the mobility orientation, currently entering into force).

Such initiatives from the French government place France at the 4th rank (and 1st European country) of the Open Data Index, an index created in 2013 by the Open Knowledge Foundation, which measures and ranks countries based on their data openness and accessibility. The English association was founded in 2004, with the aim of promoting the dissemination and accessibility of information, for all and for free. The full ranking can be found here: https://index.okfn.org/place/

Two areas negatively affect the French performance (after Taiwan, Australia, and United Kingdom): data openness on State spendings (at a buyer / vendor granularity), as well as accessibility of full land registry, taxes and selling price. On this last point, the release to the public in 2019 of the Demande de Valeur Foncière (DVF) database should help a positive re-evaluation of the ranking of France.

1. Etalab est une direction de la DINUM (Direction Interministérielle du Numérique)

### CASE LAW OPEN DATA

The data intake in the field of case law, initially planned as part of the Digital Republic Act, has been confirmed by the law of 23 March 2019 on the programming and reform of the justice system.

The main impact of the law is to make all proceedings accessible free-of-charge, on the provision that it does not impinge on stakeholders' privacy. A true game-changer for a field in which the recovery of the history of proceedings and sentences constituted a time-consuming task that could not be shortened, a Square publication on this topic can be accessed here: Justice et intelligence artificielle : les algorithmes sont-ils nos juges de demain ? (journaldunet.com).

## 1.3. OPEN DATA IN USE

The latest major success in the use of Open Data is CovidTracker.fr. In the wake of the Covid-19 health crisis, the website aggregates statistics and charts on the evolution of the pandemic, in France and worldwide. In 2021, a section on vaccination tracking was added, which shows the distribution of vaccinated individuals for each region, as well as the localization of vaccination centers. Of an intuitive use, with clear and dynamic charts, the website has since become a reference in France for the monitoring of the pandemic and is also being used by the government.

Entirely built from Open Data, the data used originates, at a national level, from Santé Publique France (France Public Health) and INSEE (Institut National de la Statistique et des Etudes Economiques, National Institute for Statistics and Economic Studies). At the international level, analyses are derived from data made available by the CSEE (the Centre for Systems Science and Engineering) from the Johns Hopkins University. Following an open source rationale, the author has also made public the full set or programs, which were mostly developed in Python.

# 2.

# OPEN DATA: A PERFORMANCE FACTOR FOR FINANCIAL INSTITUTIONS

## 2.1. CONTEXT

Data is ubiquitous to the activities of organizations, and it is now being used throughout the lifecycle of the usual transactions carried out with customers and suppliers. Such data helps financial institutions develop recommendation systems, models for performance rating, financial risk evaluation, pricing systems, commercial appeal or fraud. A large part of the performance resides precisely in the exploitation of such data, mainly or even exclusively of an internal origin.
In this context, Open Data unlocks new possibilities and innovative applications for refining and diversifying the services provided to clients.

## 2.2. MULTIPLE OPPORTUNITIES FOR THE FINANCIAL SECTOR

In an increasingly data-centric economy, financial institutions can be viewed as pioneers for their use of data sciences, with the integration of scoring tools as early as in the 1950s.

The French Digital Law has made available the Sirene, INSEE and Banque de France (Bank of France) databases. These represent a wealth of information that helps derive new analyses and a better vision of an organization's financial health. Among others, the data helps the study of insolvency (refer the first use case presented here) and the performance of in-depth analyses of the sectorial activity considered from many aspects: through geographic data, the growth and decline of specific fields of activity can be observed at several levels (region, city, etc.).

Much information also originates from such public data: the Sirene database lists the all the establishments connected to an organization, which helps monitor the opening and closing dates of the latter, as well as their addresses.

The aggregation and cross-analysis of such information then constitutes a body of converging indicators for the detection of atypical behaviors, from which the fight against different forms of fraud arises.

Furthermore, if the Open Data phenomenon initially emerged in the wake of climate studies and the publication of information used for such analyses, structured data are now available, which helps introduce new concepts in rating models. To cite only one of these, real estate professionals in Chicago are now sharing data on energy consumption, so that the organizations the most prone to carbon tax in these areas can be identified.

## 2.3. IMPROVED REGULATORY COMPLIANCE

In an increasingly strict regulatory context, the ECB published in May 2020 a guide on the management of climate risks, which specifies the supervisory requirements pertaining to the conditions of capture of such risks, in governance as well as in risk management. A key takeaway of the initial studies (mainly through the ACPR pilot exercise published in 2021) is that the major challenge of the analysis lies in the collection of the data necessary. It is a safe bet that this gap will be partly filled by Open Data, since some of this data is readily available and structured.

Considered from another perspective, external data can also improve studies on data quality thanks to the implementation of stabilization controls. In the case of the Sirene database, it becomes possible to compare the APE activity code collected by the institution with that available in public databases.

## 2.4. NEW DATA SOURCES FOR BETTER-PERFORMING MODELS

Although statistical models for credit risk evaluation can generally be considered as correctly sized, the introduction of types of information that had rarely been used is almost always an indicator of improved performance. Returning to the example of climate data, the literature[2] has repeatedly shown the relationship between financial risks and carbon footprint. In this respect, freely usable climate data constitutes a little-exploited resource that can help improving the performance of current models.

## 2.5. BETTER CUSTOMER KNOWLEDGE– PERSONALIZED SERVICES

Open Data also helps improve customer knowledge: since financial institutions have at their disposal the payment information for their customers, they can now categorize those using the Sirene database, by associating the company to which the payment was made to its field of activity (see the second use case described in this book). Such study opens multiple possibilities, including the detection of early signs of degraded financial health through a relevant view of income/expenses ratios or the detection of a more subtle customer profile through the analysis of their transactions.

Categorizing payments also makes it possible to suggest a relevant synthesis of transactions and even extrapolate an optimization of expenses and income: it is for example absolutely possible to plan the payment of bills based on financial flows, or even by carrying out automated wire transfers between savings and checking accounts. Many other customer service improvements can be envisioned to better customize the relationship.

2. https://www.sciencedirect.com/science/article/abs/pii/S0959652620316814?via%3Dihub

## 2.6. DEVELOPING NEW SERVICES FOR PROFESSIONAL AND ENTERPRISE CUSTOMERS

Examples cited above also help anticipate new services for professionals: sectorial analysis helps identify the areas most likely to generate growth and employment. It is indeed possible to identify areas in which some services are missing and to connect various actors to optimize production lines, sales, transportation, etc.

It is also possible to combine such sectorial analysis with a refined model of cessation of activity (field-specific) in order to improve the studies on land development. By crossing such information with data visualization tools, additional decision-making tool can be built and provide interactive dashboards to business experts.

Altogether, with data being already at the heart of the lifecycle of a financial product, enriching it helps improve current processes by providing new possibilities of analysis. Nonetheless, it is clear that the financial system struggles to implement solutions using these new resources. This can be partly explained by the challenges intrinsic to the exploitation of such information.

# 3.

# OPEN DATA: BARRIERS TO BE OVERCOME BY FINANCIAL INSTITUTIONS

Despite the development of Open Data in the recent years, several hurdles prevent banks and insurance companies from fully adopting their use. Indeed, some characteristics intrinsic to Open Data present a barrier to its exploitation.

## 3.1. MAIN BARRIERS PERTAINING TO THE NATURE OF OPEN DATA

Because Open Data is disseminated by third-party organizations, the quality and the reliability of information delivered are crucial elements for the development of its use. Issues with veracity, uncertainty with regards to the information source and lack of data consistency across time can restrict private players from generalizing their use and have them favor the use of internal information only, on which they can have a higher trust.

Furthermore, although the quality of data can be ensured by additional analyses and controls, the understanding of data delivered can prove to be complex. While the value of Open Data lies in their variety and diversity, the description and characteristics of each variable (metadata) provided by the source can be insufficient or ambiguous, becoming a source of interpretation errors and affecting the expected results in fine. Resorting to Open Data requires an acute knowledge of the data exploited.

Open Data can also suffer from a lack of uniformity and standardization. Indeed, the formats, refresh frequencies, and structures used vary depending on the source, which can represent an obstacle to the aggregation of the different information pieces. Interoperability between Open Data sources and data internal to the organization, or even between Open Data sources themselves complexifies the operational management of projects. Such lack of standardization may result in additional efforts and technical costs to adapt existing information systems to different Open Data sources. Productization of such data can thus also become a true hurdle.

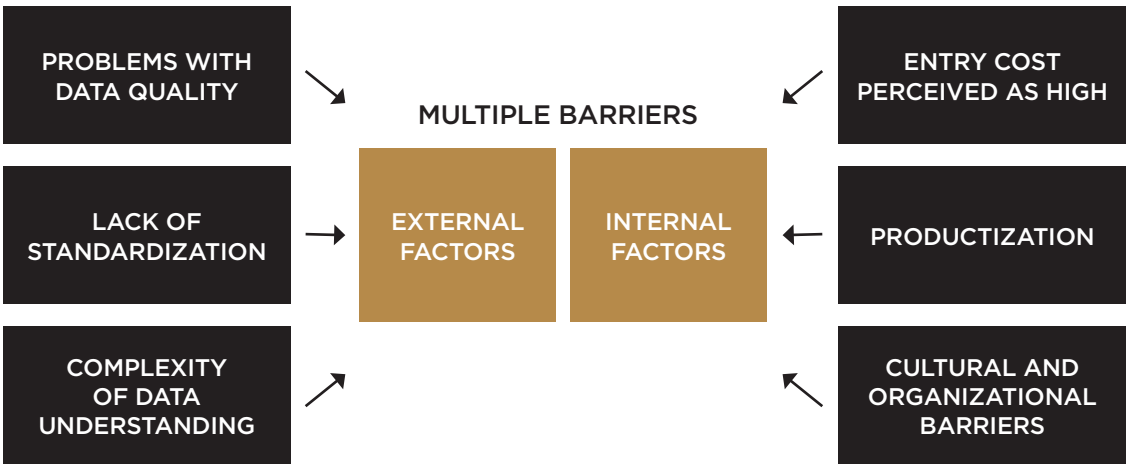
## 3.2. A CULTURE THAT NEEDS TO BE DEVELOPED

As discussed, sustainably integrating Open Data beyond a simple prototype (proof of concept) requires the contribution of a variety of competencies pertaining to automated collection, harmonization and homogenization within a shared data model, quality control and analysis of the information. All the roles associated with data, from IT specialists to data scientists, data managers and business analysts are impacted, which may be perceived as a costly and time-consuming investment.

Yet, such skills are often already available in organizations and issues specific to Open Data are relatively similar to those existing for internal data. Indeed, financial institutions have recently established data hubs that are responsible for data governance, reporting, data quality management, and development of infrastructures for all the organization's stakeholders.

Financial institutions rely mainly on data collected internally and on their mature information systems. Because of the availability of such internal data, resorting to Open Data is not systematically considered during the early stages of data projects. The organization and existing processes in place in these entities do not encourage the search for and the exploration of external and free data sources. Furthermore, the segmentation of skills in data projects, in which each participant has a specific expertise and does not participate in all the project's phases (teams devoted to collection, focusing on analysis, or devoted to business), leads to using the already existing and does not favor the search for new data sources.

The use of Open Data thus necessitates a prospection and monitoring effort that must be placed at the center of the data strategy and culture. Eventually, intrinsic constraints may present a barrier to their exploitation but, more than anything, it is the cultural and organizational habits of companies that are the first hurdle for the exploitation of Open Data.



# 4.

# USE CASE: DETECTING THE CESSATION OF BUSINESS ACTIVITY

## 4.1. CONTEXT & STAKES

Detecting the early signs of business failure remains a major challenge for many economical players such as financing organizations, private and institutional investors as well as public entities. It now represents a regulatory requirement for banks with the "Loan Origination & Monitoring" requirement issued by the European Banking Authority (EBA), which promotes the implementation of Early Warning Indicators (EWI), i.e. the processing of weak signals likely to indicate the degradation of the solvency of a counterparty.

SMEs are the key network of a country's economic tissue. Assessing the risk of business failure through rating models is broadly used by credit institutions and rating agencies. It usually relies on performing extrapolations based on descriptive, behavioral and financial information obtained from the different businesses in order to establish a probability model for global failure.

Since Open Data WebStat (Bank of France), Sirene (INSEE), and Greffe des Tribunaux de Commerce (Data Infogreffe) have been made available, an exploratory field has opened in this area, with the possibility of exploiting both data that is characteristic of businesses and an heterogenous set of behavioral and financial information that were aggregated by field of activity, type of organization and geographical area.

## 4.2. OBJECTIVE

The goal of this use case lies in exploiting financial and behavioral data aggregated by field of activity, mainly to establish a ranking model capable of reliably detecting the risk of business cessation. The model helps improve the performance of existing risk models, especially for financial structures with too few Professional and Enterprise customers to design a relevant model based on internal data only.

## 4.3. METHODOLOGY & RESULTS

The concept of cessation is taken in its broad meaning, i.e. as businesses stopping their activity and their legal existence. Cessation of a business corresponds to the end of life of a legal entity. To account both for the legal requirements and the economic realities, two categories of cessation are considered in the frame of this use case:

- Legal cessation of business: the organization is ceased in case of dissolution when it involves a moral person, or in case of death or upon cessation of all activities when it involves an individual entrepreneur;
- Cessation of business activity: the organization stops its activities. It is also termed economic cessation when all branches of an organization are closed.

Therefore, such a situation arises as the outcome of a procedure of cessation of activity for an organization, and results in the closing and stopping of its activity. It can be either bankruptcy or voluntary cessation. In any case, the organization must have gone through the steps of dissolution, liquidation, tax payments and social statement.

Cessation should not be mistaken for failure of the organization. Failure of an organization is defined by the opening of insolvency proceedings against the entity. Not all failures result in cessation.

Different algorithms were tested:

- An expert model based on logistic regression, in which all the modelling selected and metrics observed follow business consistency and a financial logic (for example by ensuring that the cessation rate increases along with the rise of cumulative debt).
- Several machine learning models (Decision Trees, Random Forest, GBM, etc.) that do not account for business considerations in the features engineering phase, for benchmarking and exploratory goals.

### FEATURES ENGINEERING

Features engineering represents the step during which raw data is selected and transformed into information in order to maximize its relevance during the modelling process.

This stage is crucial in any data science project:

- It brings new dimensions to analysis and explicability,

- It is key for the model performance and robustness, similar to the model refinement phase.

It is however sometimes overlooked by data scientist, at the expense of purely technical aspects. This is the project's most time-consuming phase and the one requiring most business knowledge.

## 4.4. DATA USED

INSEE / SIREN: data informed at the "Legal Entities" level along with the whole history (creation date, administrative status including cessation, main activity, etc. ).

Bank of France (Webstat): information related to the balance sheet and income statement, made available at multiple aggregation levels such as field of activity, geographical location and frequency. For example: debt capacity, share of equity listed in assets, share of liabilities in financial debt, etc.

## 4.5. MODELLING – EXPERT MODEL

Logistic regression: after categorization and application of variable selection methods, the expert model is built around the following inputs:

- Share of liabilities in financial debt
- Workforce productivity (apparent yield: added value expressed as volume over the number of worked hours)
- Share of equity listed in assets (Variance between two statements N/N-1)
- Margin rate (Variance between two statements N/N-1)
- Variation of the generational regional failure rate (Year N/N-1)
- Years of existence since last change of activity
- Number of closed establishments associated to the legal entity.

Result: the model presented excellent performance on the metrics for precision (95%) and recall (96%) for active establishments. For legal entities undergoing cessation of business activity, recall (76%) is also fairly good, considering the data structure and the low number of characteristic data available.

This model thus demonstrates its learning capability for detecting the cessation of activity based on financial information aggregated by field of activity.

The characteristics selected are highly significant especially from a financial point of view. Indeed, the risk of activity cessation of a legal entity within the year following the latest statement publication is all the higher when:

- Share of liabilities in financial debt is greater than 35%;
- Apparent work productivity (apparent yield: added value expressed as volume over the number of worked hours) is lower than 66%;
- Annual variation of equity is lower than -2%;
- Annual variation of margin rate is lower than 1%.

Controlling liabilities guarantees an appropriate management of the net financial debt in the balance sheet and helps organizations protect against the rise of short-term payments and the increasing weight of financial interests in the income. Furthermore, margin rate management, production capacity management, and equity variation all have a significant impact on the sustainability of the activity of SMEs.

## 4.6. MODELLING – CONCURRENT MODELS

In order to identify the classification methods best able to optimize the capability of detecting business cessation risks, results of 9 classification methods were evaluated (K-Nearest Neighbors, Linear SVM, RBF, SVM, Decision tree, Random Forest, Neural Net, Gradient Boosting, Naive Bayes, Quadratic classifier).

The K-Nearest Neighbors, Decision Tree, Random Forest and Gradient Boosting Tree methods delivered the most satisfactory results, with Random Forest and Gradient Boosting Tree showing a slight advantage with recalls (legal entities in business cessation) respectively reaching 92% and 93%, as compared to 76% shown by the logistical regression model.

There is thus a significant difference between the results obtained from an expert model (logistic regression) and those obtained from competing models. Such difference mainly lies in features selection. If the share of liabilities in financial debt, the margin rate variation and apparent work productivity are common factors to the two methods, most descriptive variables, except the number of closed branches, are excluded from benchmark models.

## MAIN TAKEAWAYS

Discriminating risk factors, as defined after the step of supervised features selection for activity cessation, helped obtain segmentations of relevant factors similar to those commonly used in financial analysis.

Regarding challengers models, analyses have shown that a significant difference lies in the assignment of cessation risk depending on the classification method and the field of activity especially. Consequently, the analysis of data coming from Open Data could help identify relevant axes for the analysis of external sectorial risk. This method may be used mainly when determining the acceptance risk, to characterize the risk related to a file, for which no risk observation is available internally.

Although the Sirene data include few behavioral risk factors, the evaluation of a change in field of activity, of the employer nature of the legal entity, or even of a change in legal status, all translate structural changes in organizations, and are therefore signs of possible improvements or degradations of the quality of organizations.

# 5.

# USE CASE: LABELLING OF BANK TRANSACTIONS

## 5.1. CONTEXT AND STAKES

All of us have already been in a situation, where, upon reading our bank statement, a credit card purchase leaves us doubting: unclear labelling, unrecognizable retailer name, non-matching geographical area. Most often, these are truncated names or acronyms that do not allow to immediately recognize the name of the retailer involved. Upon the customer befalls the task to use suitable tools to help identify the payment's beneficiary in case of doubt on a specific transaction, which most often results in a loss of time.

From the financial establishment's perspective, such uncertainties and lack of clarity do not represent an advantage regarding their customers: the lack of structure of the information contained in these transactions does not allow any refined analysis and impedes the improvement of customer knowledge in any significant manner.

How would this be affected if the information were to become structured? Based on a unique retailer's ID, a field of activity associated to each transaction would help derive a multitude of use cases:

- Customer Marketing: finer clustering of customers in relation with their expense center, their preferential sector, the location of their expenses. More broadly, this could yield a dynamic segmentation, thanks to the identification of new consumer behaviors, reflecting societal transformations (especially following the Covid-19 crisis);
- KYC (Know Your Customer): detecting changes in situation (moving, marriage), or new customer profiles (e.g., environmental-conscious profile);
- Credit risk: personalized threshold for debt rate and balance to cover the cost of everyday life;
- Customer service: improving expense management based on their habits, helping locate each transaction for an always simpler control (or monitoring?).

### WHY SUCH A FUZZY AREA?

First and foremost because banking establishments do not have control on the information included in transaction descriptions, and thus rely on the clearing house. Indeed, banks do not have the initiative of updating transaction descriptions. For example, a description can range from the full name of an organization to a shorter version of it or an acronym. Similarly, they may differ between two payment terminals inside of a same shop. Such difficulties make the automation of transaction categories immediately more complex.

A standardization of transactions should be envisioned, which would result in a certain harmonization of descriptions across all financial establishments and thus help simplify the readability of transactions.

## 5.2. METHODOLOGY & RESULTS

In this case, Open Data represents an avenue to be explored, since most debit card transactions harbor a description that is explicitly related with the merchant's name: its legal name, its used names, all of which are associated with a complete or partial geographical information (city, area, hamlet). Indeed, the INSEE Sirene databases contain such type of information (names of organizations, geographical information).

The logic would be to match, for each banking transaction, an organization, a field of activity and a location.

For this application, only card bills (excluding direct debits, wire transfers, checks, withdrawals) taking place in France would be considered.

## 5.3. DATA USED

We are using a sample of banking transactions originating from different banks (raw .csv format).

Two Open Data reference source were used in the frame of this use case:

- Sirene data;
- Code Officiels Géographiques (Geographical Official Codes) data.

Constitutive attributes are the following: transaction category, transaction date, name (most likely made of a merchant's name and a geographical area), identifier of the payment method, account number, country code, and amount associated with a currency.

| | |
|---|---|
| **DATE OF THE TRANSACTION** | 300121 |
| **NAME** | LEGO STORE |
| **LOCATION** | PARIS |
| **CARD NUMBER** | 0000XXXXXXXX0000 |
| **COUNTRY CODE** | FRA |
| **AMOUNT** | 19,99 |
| **CURRENCY** | EUR |

## 5.4. FUNCTIONAL ANALYSIS OF A TRANSACTION

Parsing of the transaction (action aiming at structuring text) consists in differentiating and isolating the set of elements that constitute a description, which comes down to distinguishing:

- The merchant's name;
- The location (city, hamlet, name, zip code, department);
- Amount;
- Card number;
- Country code for the transaction's location.

Although the parsing method is complex and involves several steps, we will use a simple case as an example here:

PAYMENT CARD FROM 300121 LEGO STORE PARIS CARD CARTE 0000XXXXXXXX0000 FRA 19,99EUR

A functional parsing of the full transaction description helps identify the different attributes that constitute it: transaction category, transaction date, name (most likely made of a merchant's name and a geographical area), identifier of the payment method, account number, country code, and amount associated with a currency.

The logic would be to match, for each banking transaction, an organization, a field of activity and a location.

For this application, only card bills (excluding direct debits, wire transfers, checks, withdrawals, etc.) taking place in France would be considered.

## 5.5. MATCHING OF THE <NAME, LOCATION> PAIR

Sirene databases present several attributes that help characterize an organization: its legal name, up to 3 used names, and even more when the history of organization is included. In addition to these names, the Sirene databases also aggregate the exact location of organizations.

Our approach relies then on a chain of technical algorithms, which help connect the pair <name, location> to the information contained in the Sirene databases. In case of multiple results (in which several establishments seem to match the pair <name, location>), a sequence of expert and functional rules are then applied, helping associate the transaction to a unique establishment (unique SIRET code).

Here are a few examples for the process of result de-duplication:

- Only list establishments that are active at the date of transaction;
- Prioritization of fields of activity: companies in the field of activity 56.10A (Traditional Catering) have a greater probability to be the expected answer, rather then organizations with a code NAF94.2OZ (Activities of worker's unions).

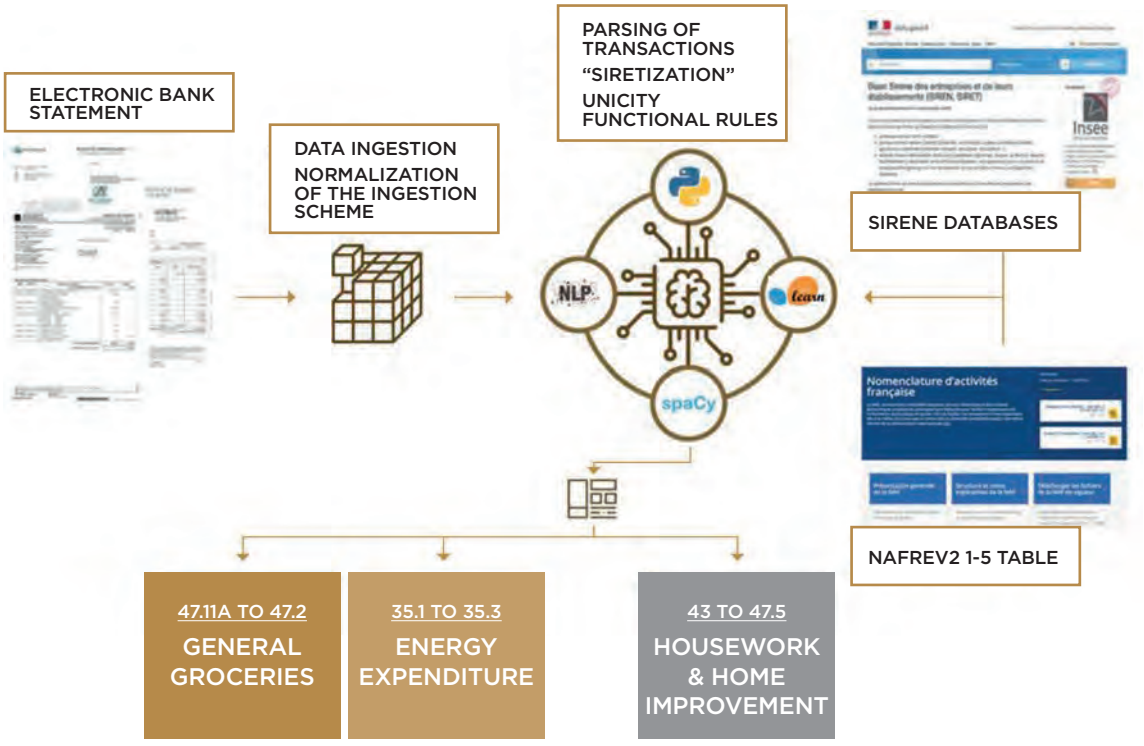
In our example, using the algorithm helps obtain the following information associated:

| | |
|---|---|
| **SIRET CODE** | 75252617800096 |
| **NAME** | PARIS 1ER ARRONDISSEMENT |
| **CITY** | PARIS |
| **ADDRESS** | 101 RUE BERGER |
| **ZIPCODE** | 75001 |
| **NAF CODE** | 47.65Z (RETAIL OF GAMES AND TOYS IN SPECIALIZED STORES) |

The expense has been associated with the category having a NAF code “Retail of games and toys in specialized stores, while the bank classified it within the category “Children – others”.

Another very revealing example of the added value brought by such a labelling step: expenses made at a Décathlon store (sport mass distribution) and Fitness Park (gym) are both placed in the “Sport” category by the bank. Here, the NAF codes for Décathlon and Fitness Park respectively match with 47.64Z “Retail of sport articles in specialized stores” and 93.13Z “Activities in physical culture centers”. An example which demonstrates a distinction between the purchase of sport’s articles and sport practice.

The following chart summarizes the steps necessary to build a “siretization” (Siret-assignment) process and categorize the transactions described previously.



## 5.6. RESULT

81% of the transactions out of the entire transaction list could be associated with a SIRET number. Among these transactions, the accuracy rate of the association of SIRET was 78%. It is indeed 61% of the transactions that could be correctly labelled.

How about the room for improvement? By taking into account, in a real case of a banking establishment, the volume represented by the pair <name, location>, manual associations shall not be excluded to process atypical cases and improve the matching rate of transactions. Additionally, banking establishment have a key information at their disposal, the Merchant Category Code, which characterizes the field of activity of a retailer. A match<name, location> would then greatly help improve the precision of "siretization" operations.

Through this use case, Open Data demonstrates again its relevance and the diversity of needs it can address, this time by enhancing the inputs for customer knowledge improvement. The functionality of associating banking transactions with establishments being sufficiently generic, it can be at the core of new solutions, of a more comprehensive and localized categorization of expenses visible on their personal customer space, and of new, customized financing offers based on the consumer profiles of customers.

# 6.
# CONCLUSION

Open Data provides opportunities for increasing the diversity and quality of service offered, independently of the field of activity considered. Its main strength resides in its capability to add context to customer knowledge, especially for the Entreprises, Professionals and VSB segments. It helps access a broad range of information and therefore is of great interest in the acquisition of customer knowledge, most notably for recent customers and prospects. It is an undeniable competitive advantage for banks upon customer onboarding or for detecting development opportunities for still unexploited customer segments. Finally, Open Data may prove essential in the case of resolution of issues such as data quality control, insufficient data or insufficient history depth, as well as for the perspectives offered with feature engineering.

As demonstrated through the use cases presented here, Open Data holds the promise of delivering practical and highly-performing answers to business issues. In the case of the detection of business cessation, several fields of application can be envisioned such as the detection of the risk of supply chain disruption in connection with business cessation in the Procurement/Supply Chain field, the detection of investment opportunities, business takeover, the allocation of development aid for high-risk sectors, or even the use of cessation models in complement to internal models of failure probability for banking establishments.

However, using Open Data requires the deployment of a research effort and its application in operational processes will need to go through an acculturation phase and a demonstration of their capability to solve business problems that internal data only cannot yet inform. From then on, institutions should launch exploratory processes and sort through useful Open Data based on their application field, integrate them into their information system in order to make systematic exploitation possible.

Regardless of the operational and organization constraints that still restrict the use of Open Data, it shall not be avoided and represents a major challenge for institutions in the frame of their climate risk management.

GIVING A FUTURE TO TALENT

**Founded in 2008, Square is a strategy and business consulting group that bring together 9 medium-sized firms in France, Belgium and Luxembourg. dway, Circle, Flow&Co, Forizons, Initio Belgique, Initio Luxembourg, Tallis, Vertuo, Viatys are consulting firms specialized in trade, activity sector or level of intervention.**

**This organization, unique and specific, favours the closeness, commitment, agility and expertise at the heart of each firm. The complementarity of the firms allows Square to address, with more than 700 consultants, the most complex projects of its clients.**

### DATA
Square develops Data strategies and ensures their operational implementation by taking the lead in Data Management, Data Analysis and Data Science projects. Our expert and pragmatic approach aims to enhance and secure companies' data assets.

### DIGITAL & MARKETING
Square assists its clients in the development of their digital strategy, the design and implementation of new digital journeys for their clients or their employees, as well as in all internal transformation projects and support for new design methods.

### INNOVATION
Square supports its clients in the transformation of their innovation dynamics. Our consultants, with their tailor-made approach, help to design, industrialize and govern innovation to ensure the sustainable growth of companies and their transformation into socially and environmentally responsible entities.

### MARKETING
Square supports its clients across the entire marketing spectrum: strategic marketing, relationship marketing, product marketing, communication, pricing, customer satisfaction. Our expertise, initially focused on the banking and insurance sectors, is now aimed at all B2C industries or services.

### PEOPLE & CHANGE
Square helps its clients to acquire, integrate and develop their organization's human capital. In order to create greater commitment within teams, our interventions focus primarily on adapting work methods to operational and cultural changes, the effectiveness of human resources departments and skills development.

### RISK & FINANCE
Square leads the management of financial and non-financial risk control programs, as well as the transformation of the Risk and Finance functions in response to changes in prudential regulations and issues related to data control.

### REGULATORY & COMPLIANCE
Square advises its clients in the rollout of new regulations, as well as in the optimization and enhancement of control systems. This area of excellence is supported by a community of experts of 130 consultants who, in addition to client assignments, conduct major research and publication work.

### SUSTAINABLE ORGANIZATIONS & FINANCE
Square supports its clients in their shift towards a more responsible model. Our guidance includes the strategic definition of a CSR ambition, the transformation of business models, and compliance projects in both their regulatory and their Data Management and Data Science aspects. Square also advises its clients in human and cultural support projects relating their CSR policy.

### SUPPLY-CHAIN
Square supports industrial and service companies in the design, deployment and optimization of their supply chain, from procurement to the last mile. Our experts implement best practices in logistics, digital and data to ensure operational excellence of the supply chain and live up to the promises made to the end customer.

**This new Square Focus book, prepared by the consultants of the Area of Excellence Data, dives into an as yet little explored opportunity for financial institutions: open data. While focusing on the structuration and the exploitation of their own data assets, organizations may tend to overlook the possible gains arising from resorting to public data when triggering initiatives relying on data. In this document, we thus cover the rise of public data and share a number of concrete business use cases brought upon by public data in a Data project.**

# CONTACT

**ADRIEN AUBERT**
Associate partner
adrien.aubert@vertuoconseil.com

Square

*Crédits Unsplash: Laureen Missaire (p. 1), Erol Ahmed (p.8), Artem Zhukov (p.12), Zanr Lee (p.16, 28).*


GIVING A FUTURE TO TALENT

square-management.com