Processing and consolidation of open data on public procurement in France (2015-2023)

Adrien Deschamps, Lucas Potin

Data in Brief
Open access
Article template

ARTICLE INFORMATION

Article title

Processing and consolidation of open data on public procurement in France (2015-2023)

Authors

Adrien Deschamps[1] *

Lucas Potin [2]

Affiliations

[1] Avignon Université, Laboratoire LBNC, 74 Rue Louis Pasteur, 84029 Avignon, France

[2] Avignon Université, Laboratoire Informatique d'Avignon, 339 Chemin des Meinajaries, 84000 Avignon

Corresponding author's email address and Twitter handle

[1] adrien.deschamps@univ-avignon.fr

[2] lucas.potin@univ-avignon.fr

Keywords

Public procurement ; open data ; e-procurement ; corruption ; green public procurement

Abstract

The dataset covers all public procurement contracts published in the dedicated official journal in France from 2015 to 2023. A collection script first reads the raw data from the online notices and processes them into a cleaned table. Then, we use the same machine learning algorithm as in Potin et al. (2023) [1] to identify public authorities and companies so that the notice data can be merged with individual information on the contracting parties. We obtain about one million contractual relationships covering more than 300,000 public contracts from all sectors and institutions. This comprehensive dataset is also quite accurate, as it potentially contains about 100 variables for each observation. These variables relate to contract characteristics (procedure, subject matter, award criteria, clauses...), award outcomes (award price, number of bids...), public authorities (type, geolocation, main activity...) and companies (size, legal status, main activity, age, geolocation...). The

dataset is unprecedented in its accuracy and scope, providing reliable and detailed information on every advertised contract in France for nearly a decade.

SPECIFICATIONS TABLE

Subject	Microeconomics
Specific subject area	The dataset provides comprehensive and accurate information on public contracts in France by processing raw data from online notices and merging it with data on economic agents (contracting authorities and firms).
Type of data	Table Filtered, Processed, Enriched
Data collection	The data collection process begins by downloading contract notices and award notices from the BOAMP website in JSON format. Then the award notices and contract notices are linked together. For each award notice, we divide the contract into lots and winners, so that we have a first table where each row is a contractual relationship between a contracting agency and a company. Each row contains relevant and processed information from the award notice and the associated contract notice. After that, missing identifiers are estimated using the same machine learning algorithm as in Potin et al. (2023) [1]. The next step is to import additional information about national and foreign economic agents.
Data source location	BOAMP : https://www.boamp.fr/pages/donnees-ouvertes-et-api/ Data.gouv : https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret/ https://www.data.gouv.fr/fr/datasets/geolocalisation-des-etablissements-du-repertoire-sirene-pour-les-etudes-statistiques/ INSEE : https://www.insee.fr/fr/information/2510634

Data accessibility	Repository name: "BeauAMP : processing and consolidation of open data on public procurement in France (2015-2023)" Data identification number: 10.5281/zenodo.11001277 Direct URL to data: https://zenodo.org/records/11001277 Instructions for accessing these data: the dataset can be freely downloaded on the Zenodo repository. We recommend to open the Pickle file with Python. In addition, an equivalent CSV file is available, as well as CSV files for each year. The associated GitHub repository can be found at the following address : https://github.com/AdrienDeschampsAU/BeauAMP

VALUE OF THE DATA

- Our dataset not only centralizes a large number of online notices but also processes this textual data to provide reliable and filtered information. Specifically, we propose a categorization of the stated award criteria and harmonize their weights. The raw data from BOAMP is impractical because the contracting authorities write the text freely. We also retain the original description of the award criteria to allow for alternative categorizations.

- Using machine learning, our dataset combines open data on public procurement with information on the characteristics of public and private organizations. We are able to link previously isolated sources of information to provide the largest and most accurate dataset available for public procurement in France. To our knowledge, there is no similarly comprehensive dataset on public procurement in other countries. The dataset by Fazekas et al. (2024) [2] covers many countries but it is less accurate due to the fact that it does not identify companies and contracting authorities. In addition, our data provides more information than the EU Tenders Electronics Daily data can because we process notices directly rather than using the extracts the EU publishes online.

- After estimating the SIRETs of companies and government agencies, we can accurately geolocate these entities. We also use a geolocation API to estimate the GPS location of foreign companies. Our data can be used to create graphs and maps, making it valuable to economists, geographers, and political science researchers alike.

- Our dataset can be connected with countless sources of information. The SIRET is a widespread key for identifying agents in official data, such as companies in the official trade register, the European Union Emissions Trading System, data or gender equality in companies... The SIRET key allows for diverse research topics and enrichments.

- The dataset is easily accessible to policy makers and practitioners due to its tabular format, which is user-friendly even for those who are not accustomed to working with data. This data is crucial for public services and public spending, and it is important that it is readily available to everyone. Until now, public institutions have not had access to such centralized and comprehensive data. By identifying public entities clearly, national and local governments can evaluate their contracts' characteristics and compare themselves with similar entities. In contrast to the FOPPA database by Potin et al. (2023) [1], our dataset includes environmental and social clauses in addition to award criteria, which is extremely valuable information for analyzing the implementation of green public procurement.

BACKGROUND

The dataset is a component of the primary author's Ph.D. thesis on the empirical evaluation of green public procurement in France. Green public procurement can be implemented through the subject matter of the contract, green award criteria, and green standards. To collect and process data from the official journal for public procurement, we processed data on award criteria and their related weights. The original raw data only contains unusable text. Furthermore, identifying agents provides more information about their characteristics. For instance, studying the impact of green public procurement on SMEs requires merging public procurement data with information on the characteristics of economic agents. Although the dataset can be used for research topics independent of sustainability, it can be particularly useful for the emerging interest in empirical analysis of green public procurement.

DATA DESCRIPTION

The data set consists of 1,162,969 rows and 113 columns. The structure of the rows is described first. Then the variables contained in the columns are described.

Structure of the dataset :

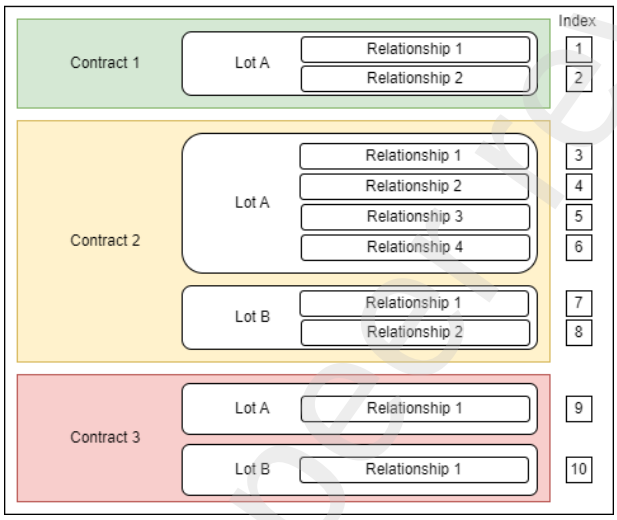

Figure 1 : structure of the dataset

The dataset contains observations of contractual relationships between contracting authorities and enterprises that have been awarded a lot, or a description of the lot if it has not been awarded. Public contracts are typically divided into multiple lots to encourage competition and SME access to public procurement. Figure 1 illustrates that 'Contract 2' and 'Contract 3' were divided into two lots, while 'Contract 1' was not. Additionally, it is important to note that one lot may be represented by multiple rows in the dataset if multiple companies were awarded the same lot, either as part of a business association or as independent entities. For example, in our dataset, there are two observations for the single lot labeled as 'Contract 1', as two distinct contractual relationships were established for this lot. Conversely, each lot that comprises 'Contract 3' was awarded to a single firm, resulting in only one row in the dataset for each lot. The dataset contains the total number of rows, which is the sum of contractual relationships and failures for each lot of each contract mentioned in the award notices on the BOAMP website between 2015 and 2023.

Variables :

The variables can be divided into four groups: information on contracts, information on outcomes of procedures, information on contracting authorities and information on enterprises.

Information on the contract :

Variable name	Variable type	Variable description	Availability (% of observations)
ID_BOAMP_AWARD	String	The identifier of the award notice on the BOAMP website.	100%
ID_BOAMP_CONTRACT	String	The identifier of the contract notice on the BOAMP website.	92%
AWARD_ NOTICE_DATE	Date	The day the award notice was published online.	100%
CONTRACT_ NOTICE_DATE	Date	The day the contract notice was published online.	92%
ID_LOT	String	The identifier of the lot (i.e. contract subdivision) where the contract relationship takes place.	99%
THRESHOLD	Categorical	The advertising threshold of the contract, depending on its value and if it is a work, a service or a supply contract. It can be : - JOUE (*Journal Officiel de l'Union Européenne*) for contracts that must be advertised on the national and European levels. - FNS (*Formulaire National Standard*) for contracts that must be advertised on the national level only. - MAPA (*Marchés A Procédure Adaptée*) when advertising is optional.	100%
CONTRACT_TYPE	Categorical	The type of contract. It can be : - works - supplies - services	83%
PROCEDURE_TYPE	Categorical	The procedure used by the contracting authority for the award of the contract. It can be : - open procedure - negotiated procedure	100%

		- competitive dialogue - innovation partnership - adapted procedure - restricted procedure - no competition - open design contest - restricted design contest	
ACCELERATED	Boolean	Indicates whether the contract was made under the accelerated procedure in an emergency context.	99%
AGP	Boolean	Indicates the contract falls under the WTO "Agreement on Government Procurement".	82%
ADVERTISING	Integer	The number of days potential candidates have to submit an offer.	87%
CORRECTIONS	Integer	The number of modifications of the contract notice between the original one and the award decision.	93%
CPV	Categorical	The sector affected by the contract. The Common Procurement Vocabulary is a 9-digit code through which we can identify different accuracy levels, for example the two first digits correspond to a "division".	95%
OBJECT	String	The object of the contract, as stated by the contracting authority.	100%
EXECUTION_SITE	String	The performance site of the contract as stated by the contracting authority for the contract as a whole or for the lot. When information is given on both levels, we keep the lot information.	91%
RENEWABLE	Boolean	Indicates whether the contract notice states that the contract can be renewed under the same conditions.	59%
DURATION	Integer	The duration of the contract, in months.	56%
NUMBER_LOTS	Integer	The number of lots in the contract.	100%
ON_BEHALF	Boolean	Indicates whether the contract was made on behalf of another institution.	72%
CENTRAL_PROCUREMENT	Boolean	Indicates whether the contract was made by a central procurement agency, i.e. a platform that groups purchases.	71%

FRAMEWORK_AGREEMENT	Boolean	Indicates whether the contract consists in a commitment by the purchaser to contract with the holder of the agreement, for a given period and for specified services.	100%
ESTIMATED_PRICE	Float	The estimation of the contract value made by the contracting authority.	13%
ENVIRONMENTAL_CLAUSE	Boolean	Indicates whether the contract includes a clause on environmental matters.	100%
SOCIAL_CLAUSE	Boolean	Indicates whether a contract includes a clause on social matters.	100%
Q_CRITERIA_TEXT	String	The award criterion or a list of award criteria used by the contracting authority on qualitative aspects of the offers (other than price).	72%
Q_CRITERIA_TYPE	Categorical	Suggested natures of qualitative award criteria based on keywords (cf. next section). It can be : - environmental - social - technical - delay - quality - other	72%
Q_CRITERIA_WEIGHTS	Float	The weight of the qualitative criterion (list of weights when there are several criteria, ranked in the same order as the names of criteria).	71%
P_CRITERION_WEIGHT	Float	The weight assigned to the award criterion. The sum of this weight and the total weight given to the qualitative aspects equals 100.	71%
RESERVED_CONTRACT	Categorical	Indicates the target when the contract is reserved for SSE companies. It can be : - workshop - jobs - workshop & jobs - False	100%

Information on the outcome of the procedure :

144

Variable name	Variable type	Variable description	Availability (% of observations)
OUTCOME	Categorical	The outcome of the procedure. It can be : - awarded : the lot has been awarded to one or several companies. - cancelled : the lot has been canceled for general interest reasons. - unsuccessful : the contracting authority hasn't received any satisfactory offer. - not_found : we couldn't find the corresponding award decision for a lot that was described in the notice.	100%
AWARD_DATE	Date	The award date of the lot.	96% *
AWARD_PRICE	Float	The award price of the lot, as stated by the contracting authority.	84% *
MIN_OFFER	Float	The lowest bid the contracting authority received.	3% *
MAX_OFFER	Float	The highest bid the contracting authority received.	3% *
NUMBER_OFFERS	Integer	The number of offers received by the contracting authority received for a given lot.	85% *
NUMBER_OFFERS_SME	Integer	The number of offers received by the contracting authority for a given lot.	8% *
NUMBER_EU_OFFERS	Integer	The number of offers from companies located in EU countries.	3% *
NUMBER_NON_EU_OFFERS	Integer	The number of offers from companies located outside the EU.	3% *
MULTI_WIN	Boolean	Indicates whether the lot was awarded to several independent companies.	99% *

BUSINESS_ASSOCIATION	Boolean	Indicates whether the lot was awarded to several cooperating companies.	98% *
SUBCONTRACTING	Boolean	Indicates whether the lot will be executed by a subcontractor.	76% *

145

146 * Among observations where the lot was awarded

147

148

149

150 *Information on the contracting authority :*

151

Variable name	Variable type	Variable description	Availability (% of observations)
CAE_STATED_NAME	String	The name of the contracting authority as mentioned in the award notice.	100%
CAE_SIREN_NAME	String	The name of the contracting authority as mentioned in the official registers.	93%
CAE_SIRET	String	The national identifier (*SIRET*) of the contracting authority. It can be mentioned by the contracting authority or estimated by a machine learning algorithm if it is unknown or if it doesn't correspond to the regular format.	94%
CAE_SIRET_KNOWN	Boolean	Indicates whether a satisfactory national identifier of the contracting authority was mentioned in the award notice at first.	100%
CAE_TOWN	String	The name* of the town the contracting authority is located.	100%
CAE_ZIP_CODE	String	The zip code* of the contracting authority location.	99%

CAE_STATED_TYPE	String	The type of the authority, as stated in the award notice.	91%
CAE_STATED_ACTIVITY	String	The main activity of the contracting authority, as stated in the award notice.	93%
CAE_SIREN	String	The *SIREN* (i.e. the nine first digits of the *SIRET*) of the contracting authority.	94%
CAE_LEGAL_STATUS	Categorical	The code of the official contracting authority legal status. The different legal statuses are detailed in the file 'legal_status' on the GitHub repository.	93%
CAE_LEGAL_STATUS_NAME	Categorical	The name of the official contracting authority legal status. The different legal statuses are detailed in the file 'legal_status' on the GitHub repository.	93%
CAE_EMPLOYER	Boolean	Indicates whether the contracting authority is an employer.	93%
CAE_STAFF	Categorical	The size of the contracting authority staff. Categories are : - 1 or 2 - 3 to 5 - 6 to 9 - 10 to 19 - 20 to 49 - 50 to 99 - 100 to 199 - 200 to 249 - 250 to 499 - 500 to 999 - 1000 to 1999 - 2000 to 4999 - 5000 to 9999 - 10000 and more	82%
CAE_MAIN_ACTIVITY	Categorical	The name of the official main activity of the contracting authority. See the file starting with 'main_activity' for the year of the version of the code on the GitHub repository.	93%
CAE_MAIN_ACTIVITY_CODE	Categorical	The code of the official main activity of the contracting authority.	93%

		See the file starting with 'main_activity' for the year of the version of the code on the GitHub repository.	
CAE_ACTIVITY_VERSION	Categorical	The version of the code for the main activity of the contracting authority. There are four versions : - 1973 (NAP) - 1993 (NAF1993) - 2003(NAFRev1) - 2008 (NAFRev2)	93%
CAE_SSE	Boolean	Indicates whether the contracting authority belongs to the social and solidarity economy.	70%
CAE_CREATION	Date	The date of creation of the contracting authority	93%
CAE_AGENCY_EMPLOYER	Boolean	Indicates whether the establishment that made the contract is an employer.	93%
CAE_AGENCY_STAFF	Categorical	The size of the establishment that made the contract (same code as for CAE_STAFF).	75%
CAE_AGENCY_MAIN_ACTIVITY	Categorical	The name of the official main activity of the establishment. See the file starting with 'main_activity' for the year of the version of the code on the GitHub repository.	93%
CAE_AGENCY_MAIN_ACTIVITY_CODE	Categorical	The code of the official main activity of the establishment. See the file starting with 'main_activity' for the year of the version of the code on the GitHub repository.	93%
CAE_AGENCY_ACTIVITY_VERSION	Categorical	The version of the code for the main activity of the establishment (same versions as for CAE_ACTIVITY_VERSION).	93%
CAE_AGENCY_CREATION	Date	The date of creation of the establishment.	93%
CAE_HEADQUARTERS	Boolean	Indicates whether the establishment where the contract was made is the	93%

		contracting authority's headquarters.	
CAE_ADDRESS	String	The stated address of the contracting authority.	98%
CAE_CITY_CODE	String	The official city code of the town the establishment is located.	93%
CAE_GPS	GPS	The GPS position of the establishment.	93%
CAE_DEPARTEMENT	String	The identifier of the *département* where the contracting authority is located (it corresponds to the two first digits of the postal code).	99%
CAE_REGION	String	The name of the *region* where the contracting authority is located.	99%
CAE_EPCI	String	The identifier of the *EPCI* (i.e. federation of municipalities) where the contracting authority is located.	82%
CAE_EPCI_TYPE	Categorical	The status of the EPCI. See the file 'legal_status' on the GitHub repository.	82%

* Official when the national identifier is known or could be estimated, stated otherwise

Information on companies :

Variable name	Variable type	Variable description	Availability (% of observations*)
WIN_STATED_NAME	String	The name of the awarded company as mentioned in the award notice.	99%
WIN_SIREN_NAME	String	The name of the awarded company as mentioned in the official registers.	85%
WIN_SIRET	String	The national identifier of the awarded company. It can be	88%

		mentioned by the contracting authority or estimated by a machine learning algorithm if it is unknown or if it doesn't correspond to the regular format.	
WIN_SIRET_KNOWN	Boolean	Indicates whether a satisfactory national identifier of the awarded company was mentioned in the award notice at first.	100%
WIN_TOWN	String	The name** of the city the awarded company is located.	98%
WIN_ZIP_CODE	String	The postal code** of the awarded company.	93%
WIN_COUNTRY_CODE	String	The country code of the awarded company.	30%
CONTRACTOR_SME	Boolean	Indicates whether the awarded company is a SME (as stated in the award notice).	63%
WIN_SIREN	String	The *SIREN* (i.e. the nine first digits of the *SIRET*) of the awarded company.	88%
WIN_LEGAL_STATUS	Categorical	The code of the legal status of the awarded company.	88%
WIN_LEGAL_STATUS_NAME	Categorical	The official name of the legal status of the awarded company.	88%
WIN_EMPLOYER	Boolean	Indicates whether the awarded firm is an employer.	88%
WIN_STAFF	Categorical	The size of the awarded company's staff (same values as 'CAE_STAFF').	60%
WIN_SSE	Boolean	Indicates whether the awarded firm belongs to the social and solidarity economy.	70%
WIN_MISSION	Boolean	Indicates whether the awarded company is a "Société à Mission", i.e. a company that includes sustainability in its objectives.	46%

WIN_STILL_ACTIVE	Boolean	Indicates whether the awarded company is still active.	88%
WIN_ACTIVITY_VERSION	Categorical	The version of the code for the main activity of the awarded company. Sames values as for 'CAE_ACTIVITY_VERSION'.	88%
WIN_ACTIVITY_LEVEL_1	Categorical	The code of the main activity of the awarded company on level 1 (least accurate).	88%
WIN_ACTIVITY_LEVEL_2	Categorical	The code of the main activity of the awarded company on level 2.	88%
WIN_ACTIVITY_LEVEL_3	Categorical	The code of the main activity of the awarded company on level 3.	88%
WIN_ACTIVITY_LEVEL_4	Categorical	The code of the main activity of the awarded company on level 4.	88%
WIN_ACTIVITY_LEVEL_5	Categorical	The code of the main activity of the awarded company on level 5 (the most accurate).	88%
WIN_MAIN_ACTIVITY	Categorical	The name of the main activity of the awarded company on level 5. See the file starting with 'main_activity' for the year of the version of the code on the GitHub repository.	88%
WIN_CREATION_DATE	Date	The date of creation of the awarded company.	88%
WIN_AGENCY_ACTIVITY_LEVEL_1	Categorical	The code of the main activity of the awarded establishment on level 1.	88%
WIN_AGENCY_ACTIVITY_LEVEL_2	Categorical	The code of the main activity of the awarded establishment on level 2.	88%
WIN_AGENCY_ACTIVITY_LEVEL_3	Categorical	The code of the main activity of the awarded establishment on level 3.	88%
WIN_AGENCY_ACTIVITY_LEVEL_4	Categorical	The code of the main activity of the awarded establishment on level 4.	88%

WIN_AGENCY_ACTIVITY_LEVEL_5	Categorical	The code of the main activity of the awarded establishment on level 5.	88%
WIN_AGENCY_ACTIVITY_VERSION	Categorical	The version of the code for the main activity of the awarded establishment. Same values as for 'CAE_ACTIVITY_VERSION'.	88%
WIN_AGENCY_MAIN_ACTIVITY	Categorical	The name of the main activity of the awarded establishment on level 5. See the file starting with 'main_activity' for the year of the version of the code on the GitHub repository.	88%
WIN_AGENCY_CREATION_DATE	Date	The date of creation of the awarded company's establishment.	88%
WIN_HEADQUARTERS	Boolean	Indicates whether the establishment that was awarded the contract is the company's headquarters.	88%
WIN_ADDRESS	String	The stated address of the company.	88%
WIN_CITY_CODE	String	The official city code of the town the awarded establishment is located.	88%
WIN_GPS	GPS	The GPS position of the awarded establishment.	88%
WIN_DEPARTEMENT	Categorical	The identifier of the *département* where the awarded establishment is located (it corresponds to the two first digits of the postal code).	93%
WIN_REGION	Categorical	The name of the *region* where the awarded establishment is located. The statuses are detailed in the file 'regions'.	93%

159

160 * Observations for which the lot has been awarded

** Official when the national identifier is known or could be estimated, stated otherwise

EXPERIMENTAL DESIGN, MATERIALS AND METHODS

In France, the BOAMP (*Bulletin Officiel des Annonces des Marchés Publics*) is the official journal where Contracting Authorities or Entities (CAEs) publish notices of their procurement contracts when the expenditure exceeds a threshold set by law. Although advertising notices is not mandatory, contracting authorities may choose to do so. The purpose of these advertising restrictions is to prevent favoritism and ensure accountability of contracting authorities in the expenditure of public funds.

Contracting authorities typically begin by publishing contract notices to inform potential bidders about the upcoming contract and award process. The result is then announced in an award notice. Our dataset was generated using these two types of notices downloaded from the BOAMP website between January 1, 2015 and December 31, 2023, as shown in Figure 2. A Python script is used to process the information and generate a dataset where each row represents a contractual relationship. The missing national identifiers for both buyers and companies need to be identified to merge this dataset with other information from the SIRENE data. This will result in enriched and comprehensive data on public procurement. Since foreign companies can't be found in national registers, we use a geolocation API to estimate their geolocation. The following sections mention the title of different Python files. They can be found on the aforementioned GitHub repository.

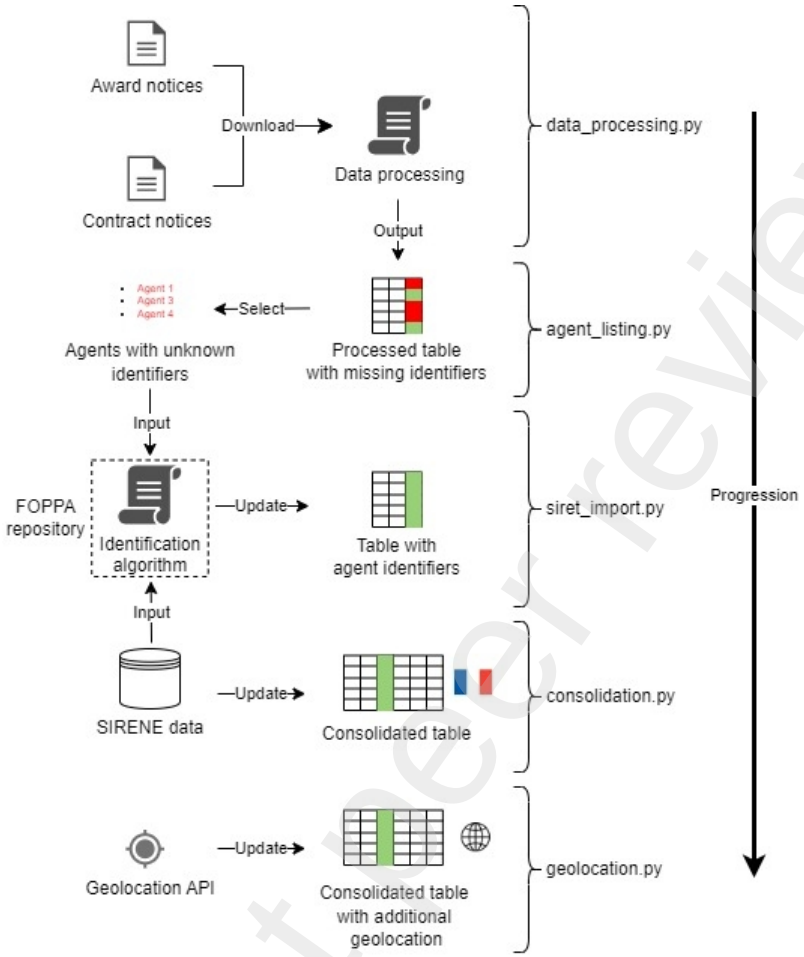

Figure 2 : summary of the process

Step 1 : download the data from the BOAMP website

Very interestingly, we have free access to contract and award notices on the BOAMP website. We can read the PDF of the notices or download them with an integrated API. The file "data_processing" first downloads contract notices for each of the 101 French *départements*. Since notices can be indexed for several *départements*, we then delete duplicates. The same process is repeated for contract notices. We obtain two dataframes with the data of the contract and award notices, respectively.

Step 2 : process contract data

This section summarizes the way we process the data to create a table from the raw notices. A more detailed description of the various processing steps is available in the file “data_processing”.

To convert the downloaded texts into a processed dataset, we first remove notices that do not relate to public contracts (such concession contracts). Then we keep only contract awards published between January 1, 2015 and December 31, 2023. Since contract notices are published before the corresponding award notices, we cannot ignore any contract notice published before January 1, 2015. However, to speed up the process, we only retain contract notices published after January 1, 2013. Finally, we capitalize notice text and remove misleading backslashes.

After defining the variables we want to extract, the data is ready to be processed. We create a list that will be converted to a data frame at the end of the process. We iterate through the rows of the data frame containing the award notices. Using different parts of the award notice, we can identify the common information for all the lots of a given contract. We also try to link the award notice to the corresponding contract notice, which provides new information and sometimes compensates for the missing information in the award notice.

Interestingly, award notices contain a description of each lot's features and a description of the award decisions for each lot. If we can match the lot description with the lot award (directly or after looking for strong similarities in their identifiers), we can have both the lot features and its outcome. We start by extracting the information about each lot's characteristics. Then we do the same for the lot results. If the award decision mentions several companies, we split the information as if there were different lots (but with a common lot identifier). Since we want our dataset to be structured around contractual relationships, we then export the lot features for each award decision by matching the identifiers. If we can't find the corresponding award decision for a described lot (either because the identifiers can't be matched, or because the lot couldn't be awarded but the contracting authority didn't mention it), the lot appears in the dataset with 'not_found' in the 'OUTCOME' column. We finish this process by adding the general contract data to this contractual relationship level data. We find the desired structure for our dataset from the raw notice texts.

The way award criteria are processed is a key feature of our dataset. The first step is to check if the award criteria are mentioned in the notice. However, sometimes the contracting authorities refer to other documents to describe the award criteria, making it difficult to locate them. If the award criteria are found in the notice, they may be detailed in the description of each lot or in the procedure for the entire contract.

To begin, we can standardize the weights of the award criteria. These weights are typically specified by the contracting agency and should align with the scoring rule used to make the award decision.

However, the stated weights are often non-standardized, with some adding up to 1 and others to 100. Contracting authorities may include subcriteria within the main criteria. Therefore, a simple sum of the stated weights may result in misleading weights in our data. To avoid this, we sum up the different weights and scale them to percentages if the result is a consistent number (e.g., 1, 10, 20, 100). It is important to note that we mention the names of the award criteria even if we cannot normalize the weights.

In addition, in the variable 'Q_CRITERIA_TYPE', we suggest the types of qualitative award criteria (as opposed to the price criterion) based on the following list of keywords. If one of the keywords is found in the text of a given award criterion, it will be assigned to the corresponding category. Note that the search follows the order of the lists :

environmental : ['ENVIRONNEM', 'ENVIRONEM', 'ECOLO', 'ÉCOLO', 'ÉCOSYST', 'ECOSYST', 'ÉCO-SYST', 'ECO-SYST', 'RECYCL', 'RECICL', 'SOUTENABI', 'DURAB', 'CLIMAT', 'CARBO', 'DUREE DE VIE', 'DURÉE DE VIE', 'POLLUT']

social : ['SOCIA', 'SOCIÉT', 'SOCIETA', 'ÉTHIQUE', 'ETHIQUE', 'TRACABILI', 'TRAÇABILI', 'INSERTION', 'HUMAIN', ' RSE', 'PERSONNEL']

delay : ['DELAI', 'DÉLAI', 'DURÉE', 'DUREE', 'PÉRIODE', 'PERIODE', 'TEMPS', 'PLANING', 'PLANNING']

quality : ['QUALIT']

technical : ['TECHNIQUE', 'TECHNOLO', 'METHOD', 'MÉTHOD', 'QUALIT', 'FONCTION', 'EXECUTION', 'EXÉCUTION', 'ÉXÉCUTION', 'ÉXECUTION', 'OPÉRAT', 'OPERAT']

If “environmental” and “social” keywords are simultaneously found, then the criterion is considered as “socio_environmental”. If none of the previous keywords is found, the criterion is labelled “other”.

Step 3 : agent identification

In France, both public authorities and companies have national identifiers, the SIRET (*Système d'identification du répertoire des établissements*) and the SIREN (*Système d'identification du répertoire des entreprises*). The SIREN corresponds to the first nine digits of the 14-digit SIRET. The former refers to the company or public institution as a whole, whereas the latter refers to a specific establishment of the agent, thus introducing a geographical dimension.

In an ideal scenario, the contract notices should mention the SIRETs of both contracting authorities and companies. However, in the processed dataset, only 25% and 6% of the observations, respectively, have this information. To consolidate this dataset, we need to merge the table we have so far with individual data on public and private agents. To estimate the SIRETs, we will use the information we have on the contracting authorities and firms, including their status (company vs. public institution), the name of the agent, and their location (address, city, postal code).

In the 'agent_listing' file, we generate two lists of unique agent mentions by selecting observations where the SIRETs are ignored or the SIRET format is not respected. The lists are created by collecting observations with the same set of name, address, city, and postal code. One list is for contracting authorities and the other one is for companies. These lists serve as input for the agent identification algorithm developed by Potin et al. (2023) [1]. The machine learning process estimates the agent in the database that is most similar to the agent we want to identify. First, the algorithm filters the SIRENE database by geographical area, date, and activity domain for each agent. Then, it filters the remaining potential official identities by stated name. Finally, it estimates the most likely official identity based on the stated address of the agent. The same logic applies, except that the European data is replaced with the table obtained at the end of step 2. According to Potin et al., their algorithm was able to identify approximately 80% of the agents, with about 75% for contracting authorities and 81% for firms.

After a few weeks, the algorithm was able to return a list of estimated identifiers based on the input we sent him. This list is contained in the file “estimated_identifiers”. We import the estimated identifiers into our data set in the file "identifier_import".

Step 4 : using the agent identifiers to complete the dataset with the SIRENE database

Once we have a more significant number of official identifiers, especially for companies, we can consolidate our dataset by partially merging it with open data on economic agents (even though this step could be done only with the identifiers mentioned in the notices, i.e. skipping step 3).

In the file “consolidation”, we read the datasets “Sirene : Fichier StockEtablissementHistorique” and “Sirene : Fichier StockEtablissement” for establishments (SIRET) and the datasets “Sirene : Fichier StockUniteLegaleHistorique du 01 Fevrier 2024” and “Sirene : Fichier StockUniteLegale” for legal entities (SIREN). These datasets come from the government repository SIRENE. We merge these files two by two so that we get two tables with variables we need for each SIRET/SIREN. In addition, we import a file containing the geolocation of establishments through their SIRET.

We subset the SIREN and SIRET data to the identifiers we actually need for our dataset. We also make our data easier to read by importing the files containing the names associated with the codes used for legal status and primary activity. We add new variables to our initial dataset of award notices. We consolidate this dataset with the information we have on the characteristics of the agents based on their SIREN and SIRET. We also add information about the city, *département*, and *région* in which they are located.

Step 5 : using the agent address to estimate the geolocation of foreign companies

The final step consists in trying to geolocate foreign companies thanks to their address or country code. Indeed, foreign agents are not mentioned in the SIRENE database, since it only covers French organizations.

In the file "geolocation", we use a two-step process to estimate their geolocation. First, we use the stated address of the company as an input for the API "Nominatim". When this first attempt doesn't work, we assign the average GPS coordinates of the involved foreign country based on the stated ISO country code in the award notice.

LIMITATIONS

The notices published by the contracting authorities are the original source of information for the dataset. Therefore, the data are subject to the imperfections and errors contained in the notices. Some variables are often ignored by the civil servant responsible for filling the notices (e.g. the minimum and maximum values of tenders). Except for SIRETs, this limitation is impossible to overcome. However, the quality of the data tends to improve over time and the larger the contract value, the better the information available. Although our machine learning algorithm is a useful tool for connecting data sources, it can be imprecise when information is limited. While the agent identification algorithm limits the margin of error in geolocation by searching within a given geographic area, it is important to avoid blindly trusting the identity of individual agents if their estimated location seems surprising.

ETHICS STATEMENT

The authors have read and follow the ethical requirements for publication in Data in Brief and confirms that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

CRediT AUTHOR STATEMENT

Adrien Deschamps : Conceptualization, Methodology, Software, Writing - Original Draft, Writing - Review & Editing

Lucas Potin : Resources

ACKNOWLEDGEMENTS

We would like to thank Pierre-Henri Morand for his advice based on his previous experience with the FOPPA database.

This research was partially funded by the French national research agency (Agence Nationale de la Recherche, ANR), within the framework of the following project :

"Détecter la Corruption dans les Marchés Publics
Grant ANR-19-CE38-0004"

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Potin, L., Labatut, V., Morand, PH. *et al.* FOPPA: an open database of French public procurement award notices from 2010–2020. *Sci Data* 10, 303 (2023). https://doi.org/10.1038/s41597-023-02213-z

[2] Mihaly Fazekas , Bence Toth , Aly Abdou , Ahmed Al-Shaibani , ´ Global Contract-level Public Procurement Dataset, Data in Brief (2024), doi: https://doi.org/10.1016/j.dib.2024.110412